

Universidad de Almería

FACULTAD DE PSICOLOGÍA

Departamento de Ciencias Humanas y Sociales



MODELOS PREDICTIVOS EN ACTITUDES EMPRENDEDORAS:  
ANÁLISIS COMPARATIVO DE LAS CONDICIONES DE EJECUCIÓN  
DE LAS REDES BAYESIANAS Y LA REGRESIÓN LOGÍSTICA

TESIS DOCTORAL

*Elaboración: Jorge López Puga*

*Dirección: Juan García García*

*Almería, 2009*



MODELOS PREDICTIVOS EN ACTITUDES EMPRENDEDORAS:  
ANÁLISIS COMPARATIVO DE LAS CONDICIONES DE EJECUCIÓN  
DE LAS REDES BAYESIANAS Y LA REGRESIÓN LOGÍSTICA

Tesis Doctoral presentada por  
JORGE LÓPEZ PUGA dentro del  
Programa de Doctorado:  
*«Perspectivas Avanzadas en Investigación e Intervención Psicológica»*  
para aspirar al grado de Doctor  
por la Universidad de Almería  
y dirigida por el Doctor  
JUAN GARCÍA GARCÍA

Almería, a 29 de Octubre de 2009



*A Ana María, Julia y Álvaro*



# Índice general

<b>I</b>	<b>Revisión Teórica</b>	<b>35</b>
<b>1.</b>	<b>Modelado predictivo</b>	<b>37</b>
1.1.	Introducción . . . . .	37
1.2.	Modelo de regresión lineal . . . . .	38
1.2.1.	Asunciones del modelo de regresión lineal . . . . .	41
	Linealidad . . . . .	42
	Colinealidad . . . . .	42
	Media condicional nula para $\epsilon$ . . . . .	42
	Varianza, correlación y distribución de $\epsilon_i$ . . . . .	43
1.2.2.	Interpretación de los coeficientes de regresión . . . . .	43
1.2.3.	Problemas con el Modelo de Probabilidad Lineal . . . . .	44
1.3.	Regresión logística . . . . .	46
1.4.	Redes bayesianas . . . . .	47
<b>2.</b>	<b>Fundamentos de la regresión logística</b>	<b>53</b>
2.1.	Componentes del modelo . . . . .	53
2.1.1.	Función logística . . . . .	53
2.1.2.	Riesgo y ventaja relativa . . . . .	54
2.1.3.	Modelo estadístico . . . . .	56
2.1.4.	Estimación del modelo . . . . .	58
2.1.5.	Interpretación de coeficientes . . . . .	61
2.2.	Bondad de ajuste . . . . .	64
2.2.1.	Ajuste predictivo . . . . .	65
	Tabla de clasificación o matriz de confusión . . . . .	65

## ÍNDICE GENERAL

Parámetros predictivos . . . . .	66
Curvas ROC . . . . .	67
2.2.2. Índices de bondad de ajuste general . . . . .	69
Verosimilitud del modelo . . . . .	69
Test de Hosmer-Lemeshow . . . . .	70
Varianza explicada . . . . .	71
Ajuste residual e influencia . . . . .	72
2.2.3. Ajuste de los coeficientes del modelo . . . . .	73
2.3. Limitaciones y avances en el análisis de regresión logística . . . . .	74
2.3.1. Tamaño de muestra . . . . .	74
2.3.2. Casos perdidos . . . . .	77
2.3.3. Categorización de las variables . . . . .	77
<b>3. Fundamentos de las redes bayesianas</b> . . . . .	<b>81</b>
3.1. Estructura y funcionamiento . . . . .	81
3.1.1. Estructura . . . . .	83
Dimensión cualitativa . . . . .	83
Dimensión cuantitativa . . . . .	86
3.1.2. Funcionamiento . . . . .	89
3.2. Los diagramas de influencia . . . . .	93
3.3. Construcción de redes bayesianas . . . . .	99
3.3.1. Estimación paramétrica . . . . .	99
3.3.2. Estimación estructural . . . . .	100
3.4. Bondad de ajuste . . . . .	106
3.5. Ventajas de las redes bayesianas . . . . .	107
3.6. <i>Software</i> . . . . .	109
3.6.1. TETRAD . . . . .	109
3.6.2. NETICA™ . . . . .	110
3.6.3. ELVIRA . . . . .	111
3.6.4. MBNTx . . . . .	112
3.6.5. ERGO™ . . . . .	112
3.6.6. HUGIN™ . . . . .	113
3.6.7. Otros programas . . . . .	114
3.7. El rol de las redes bayesianas en Psicología . . . . .	117

3.7.1.	La Psicología y las redes bayesianas . . . . .	117
3.7.2.	Aplicaciones en Psicología . . . . .	119
	Sistemas de Tutorización Inteligente . . . . .	119
	Aprendizaje Causal . . . . .	122
	Psicología Forense . . . . .	124
	Psicolingüística . . . . .	125
	Psicodiagnóstico . . . . .	127
	Psicología Organizacional . . . . .	128
<b>4.</b>	<b>Actitudes emprendedoras</b>	<b>131</b>
4.1.	Reseña histórica . . . . .	131
4.2.	Conceptualización y tipologías emprendedoras . . . . .	135
4.3.	Teorías explicativas . . . . .	142
4.3.1.	Dimensión socio-económica . . . . .	142
	Modelo de las turbulencias en el mercado laboral . . . . .	142
	Modelos <i>pull</i> y <i>push</i> . . . . .	143
	Modelo de la marginación social . . . . .	144
	Modelo intergeneracional . . . . .	145
4.3.2.	Dimensión psicológica . . . . .	146
	Modelo psicodinámico . . . . .	146
	Modelo de los rasgos . . . . .	147
	Modelo conductual . . . . .	149
	Modelo de los sesgos . . . . .	149
4.4.	Actitudes emprendedoras en la universidad española . . . . .	151
<b>II</b>	<b>Trabajo Empírico</b>	<b>157</b>
<b>5.</b>	<b>Metodología general</b>	<b>159</b>
5.1.	Contextualización de la investigación . . . . .	159
5.2.	Objetivos . . . . .	160
5.3.	Método . . . . .	163
5.3.1.	Participantes . . . . .	163
	Perfil académico y laboral . . . . .	164
	Perfil familiar . . . . .	166

## ÍNDICE GENERAL

5.3.2. Materiales . . . . .	167
Propiedades psicométricas de las escalas . . . . .	170
5.3.3. Procedimiento . . . . .	174
<b>6. Estudio 1. Construcción y depuración de modelos de red bayesiana y regresión logística</b>	<b>175</b>
6.1. Descripción . . . . .	175
6.2. Método . . . . .	176
6.2.1. Muestra . . . . .	176
6.2.2. Materiales . . . . .	176
6.2.3. Análisis de datos . . . . .	177
6.3. Resultados . . . . .	178
6.3.1. Modelos de regresión logística . . . . .	179
Componente emocional . . . . .	179
Componente cognitivo . . . . .	182
Componente conductual . . . . .	183
6.3.2. Modelos de red bayesiana . . . . .	183
6.3.3. Ajuste predictivo de los modelos seleccionados . . . . .	186
6.3.4. Comparación preliminar entre técnicas . . . . .	190
6.4. Discusión parcial . . . . .	191
<b>7. Estudio 2. Tamaño de muestra y área bajo la curva ROC</b>	<b>195</b>
7.1. Descripción . . . . .	195
7.2. Método . . . . .	196
7.2.1. Muestra . . . . .	196
7.2.2. Análisis de datos . . . . .	196
7.3. Resultados . . . . .	200
7.4. Discusión parcial . . . . .	206
<b>8. Estudio 3. Nivel de categorización y área bajo la curva ROC</b>	<b>209</b>
8.1. Descripción . . . . .	209
8.2. Método . . . . .	210
8.2.1. Muestra . . . . .	210
8.2.2. Análisis de datos . . . . .	210
8.3. Resultados . . . . .	212

8.4. Discusión parcial . . . . .	213
<b>9. Estudio 4. Tipo de categorización y área bajo la curva ROC</b>	<b>219</b>
9.1. Descripción . . . . .	219
9.2. Método . . . . .	220
9.2.1. Muestra . . . . .	220
9.2.2. Análisis de datos . . . . .	220
9.3. Resultados . . . . .	222
9.4. Discusión parcial . . . . .	227
<b>10. Estudio 5. Porcentaje de casos perdidos y área bajo la curva ROC</b>	<b>229</b>
10.1. Descripción . . . . .	229
10.2. Método . . . . .	230
10.2.1. Muestra . . . . .	230
10.2.2. Análisis de datos . . . . .	230
10.3. Resultados . . . . .	232
10.4. Discusión parcial . . . . .	237
<b>III Conclusiones</b>	<b>241</b>
<b>11. Discusión general</b>	<b>243</b>
11.1. Redes bayesianas <i>versus</i> regresión logística . . . . .	243
11.2. Actitudes emprendedoras . . . . .	248
<b>IV Epílogo</b>	<b>251</b>
<b>12. Reflexiones finales</b>	<b>253</b>
<b>V Anexo</b>	<b>261</b>
<b>Referencias</b>	<b>281</b>

## ÍNDICE GENERAL

# Índice de figuras

1.1. El proceso de modelado estadístico . . . . .	38
1.2. Conductas agresivas en función del porcentaje de clases a las que asisten los adolescentes . . . . .	40
1.3. Observaciones y estimación lineal de la relación que se establece entre las conductas agresivas y las faltas a clases . . . . .	41
1.4. Evolución del número de publicaciones en el <i>ISI Web of Knowledge</i> <sup>SM</sup> donde el tópico fue la regresión logística . . . . .	47
1.5. Evolución del número de publicaciones en el <i>ISI Web of Knowledge</i> <sup>SM</sup> donde el tópico fueron las redes bayesianas . . . . .	49
1.6. Árbol de decisión diagnóstica usado en el <i>Great Ormond Street Hospital</i> para la evaluación de neonatos . . . . .	50
1.7. Ejemplo del factor de certidumbre . . . . .	51
2.1. Función logística . . . . .	54
2.2. Transformación del espacio probabilístico al de los <i>odds</i> . . . . .	55
2.3. Función de Máxima Verosimilitud . . . . .	60
2.4. Observaciones, estimación lineal y estimación logística para la relación entre la tasa de absentismo y la probabilidad de mostrar conductas agresivas . . . . .	63
2.5. Representación de dos curvas ROC . . . . .	69
2.6. Relación entre verosimilitud y lejanía . . . . .	70
3.1. Conexión serial en una red bayesiana . . . . .	85
3.2. Conexión divergente en una red bayesiana . . . . .	86

## ÍNDICE DE FIGURAS

3.3. Conexión convergente en una red bayesiana . . . . .	86
3.4. Red <i>Recuerdo</i> . . . . .	89
3.5. Red <i>Recuerdo</i> compilada . . . . .	92
3.6. Red <i>Recuerdo</i> con <i>Número de Palabras</i> instanciado . . . . .	93
3.7. Red <i>Recuerdo</i> con <i>Puntuación en un Examen</i> instanciado . . . . .	93
3.8. Red <i>Recuerdo</i> con <i>Recuerdo Después de 30 m.</i> instanciado . . . . .	94
3.9. Red <i>Recuerdo</i> con <i>Recuerdo Después de 30 m.</i> y <i>Número de Palabras</i> instanciados . . . . .	94
3.10. Red <i>Ansiedad</i> . . . . .	95
3.11. Red <i>Ansiedad</i> compilada . . . . .	98
3.12. Red <i>Ansiedad</i> con evidencia de ansiedad nula . . . . .	98
3.13. Red <i>Ansiedad</i> con evidencia de ansiedad normal . . . . .	98
3.14. Red <i>Ansiedad</i> con evidencia de ansiedad patológica . . . . .	99
3.15. Algoritmo PC implementado en TETRAD . . . . .	101
3.16. Fases para la construcción de un mapa causal . . . . .	104
3.17. Representación de una red orientada a objetos . . . . .	114
4.1. Corrientes de pensamiento surgidas a partir del trabajo de Richard Cantillon . . . . .	134
4.2. Evolución del número de empresas españolas desde 1999 a 2008 . . . . .	140
4.3. Relación entre desempleo y creación de nuevas empresas . . . . .	143
4.4. Frecuencia de alusión a los elementos más importantes en la investigación sobre emprendedores . . . . .	148
4.5. Modelo de la teoría de la acción razonada . . . . .	153
4.6. Modelos de la teoría del comportamiento planeado . . . . .	154
5.1. Taxonomía de clasificadores de red bayesiana . . . . .	162
5.2. Vía de acceso a la universidad de los participantes . . . . .	164
5.3. Puesto ocupado y sector de actividad donde trabajan los participantes . . . . .	166
5.4. Nivel de estudios familiar . . . . .	166
5.5. Ocupación principal de los padres . . . . .	167
5.6. Distribución de frecuencias de las puntuaciones de las escalas . . . . .	171

## ÍNDICE DE FIGURAS

6.1. Comparación de los residuos estudentizados al cuadrado y de las distancias de Cook con la probabilidad pronosticada para el modelo DES . . . . .	182
6.2. Comparación de los residuos estudentizados al cuadrado y de las distancias de Cook con la probabilidad pronosticada para el modelo COG . . . . .	184
6.3. Comparación de los residuos estudentizados al cuadrado y de las distancias de Cook con la probabilidad pronosticada para el modelo CON . . . . .	186
6.4. Probabilidades previas del componente emocional en una red bayesiana . . . . .	187
6.5. Probabilidades previas del componente cognitivo en una red bayesiana . . . . .	188
6.6. Probabilidades previas del componente conductual en una red bayesiana . . . . .	189
7.1. Curvas ROC globales en función de la variable dependiente . . . . .	201
7.2. Curvas ROC en función del número de eventos por variable . . . . .	202
7.3. Evolución del área bajo la curva ROC en función de la variable dependiente y del número de eventos por variable . . . . .	204
8.1. Efecto global de la recategorización sobre la curva ROC . . . . .	212
8.2. Curvas ROC en función del nivel de categorización . . . . .	215
9.1. Efecto global de la categorización sobre la curva ROC en función de la técnica . . . . .	222
9.2. Curvas ROC en función del tipo de categorización . . . . .	225
10.1. Esquema del proceso de simulación de casos perdidos . . . . .	231
10.2. Curvas ROC globales en función de la variable dependiente . . . . .	232
10.3. Curvas ROC en función del porcentaje de casos perdidos . . . . .	235
10.4. Evolución del área ROC en función del porcentaje de casos perdidos y de la variable dependiente . . . . .	238

## ÍNDICE DE FIGURAS

# Índice de tablas

1.1. Número de publicaciones cuyo tópico fueron las redes bayesianas o la regresión logística desde 1998 a 2008 . . . . .	48
2.1. Matriz de confusión o tabla de clasificación . . . . .	65
3.1. Algunas definiciones de red bayesiana . . . . .	82
3.2. Tabla de probabilidad condicional para el nodo <i>Número de Palabras</i> en la red <i>Recuerdo</i> . . . . .	90
3.3. Tabla de probabilidad condicional para el nodo <i>Tipo de Procesamiento</i> en la red <i>Recuerdo</i> . . . . .	90
3.4. Tabla de probabilidad condicional para el nodo <i>Calidad del Sueño</i> en la red <i>Recuerdo</i> . . . . .	90
3.5. Tabla de probabilidad condicional para el nodo <i>Tiempo de Aprendizaje</i> en la red <i>Recuerdo</i> . . . . .	90
3.6. Tabla de probabilidad condicional para el nodo <i>Recuerdo 30 m. después</i> en la red <i>Recuerdo</i> . . . . .	91
3.7. Tabla de probabilidad condicional para el nodo <i>Puntuación en un Examen</i> en la red <i>Recuerdo</i> . . . . .	91
3.8. Tabla de utilidad del nodo <i>Satisfacción</i> . . . . .	96
3.9. Tabla de probabilidad condicional del nodo <i>Grado de Ansiedad</i> . . . . .	97
5.1. Diferencias entre la regresión logística y las redes bayesianas . . . . .	161
5.2. Estudios cursados por los participantes . . . . .	165
5.3. Descriptivos de las distribuciones de las puntuaciones de las escalas . . . . .	172
5.4. Consistencia interna de las escalas . . . . .	172

## ÍNDICE DE TABLAS

5.5. Validez discriminante para las escalas utilizadas . . . . .	173
6.1. Los modelos más predictivos en función de cada una de las técnicas y de las variables dependientes . . . . .	180
6.2. Modelo de regresión logística para explicar el componente emocional de la actitud . . . . .	181
6.3. Test de ajuste Hosmer-Lemeshow para el componente emocional .	181
6.4. Modelo de regresión logística para explicar el componente cognitivo de la actitud . . . . .	183
6.5. Test de ajuste Hosmer-Lemeshow para el componente cognitivo .	184
6.6. Modelo de regresión logística para explicar el componente conductual de la actitud . . . . .	185
6.7. Test de ajuste Hosmer-Lemeshow para el componente conductual	185
6.8. Ajuste predictivo para los modelos seleccionados . . . . .	190
6.9. Comparación preliminar entre técnicas . . . . .	191
7.1. Tamaños de muestra en función del número deseado de eventos por variable y de la tasa de respuestas positivas . . . . .	197
7.2. Áreas bajo la curva ROC en función del tamaño de la muestra y del tipo de técnica . . . . .	203
7.3. Comparación de las áreas bajo la curva ROC en función de la variable dependiente y del tamaño de muestra . . . . .	205
7.4. Tamaño de las diferencias entre áreas bajo la curva ROC en función del número de eventos por variable y la variable dependiente . . .	206
8.1. Valores de corte para las variables independientes . . . . .	211
8.2. Áreas bajo la curva ROC en función del tamaño del nivel de categorización y del tipo de técnica . . . . .	214
8.3. Comparación de áreas bajo la curva ROC en función del tamaño del nivel de categorización y del tipo de técnica . . . . .	216
8.4. Tamaño de las diferencias entre áreas bajo la curva ROC en función del número de categorías en las variables independientes y la variable dependiente . . . . .	216
9.1. Valores de corte para las variables independientes para la recodificación basada en los percentiles . . . . .	221

## ÍNDICE DE TABLAS

9.2. Áreas bajo la curva ROC en función del tipo de recodificación, de la variable dependiente y de la técnica . . . . .	224
9.3. Comparación de las áreas bajo la curva ROC en función de la variable dependiente y del tipo de recodificación . . . . .	226
9.4. Tamaño de las diferencias entre áreas bajo la curva ROC en función del tipo de recategorización en las variables independientes y de la variable dependiente . . . . .	227
10.1. Áreas bajo la curva ROC en función del porcentaje de casos perdidos y del tipo de técnica . . . . .	234
10.2. Comparación de áreas bajo la curva ROC en función del tamaño del porcentaje de casos perdidos y del tipo de técnica . . . . .	236
10.3. Tamaño de las diferencias entre áreas bajo la curva ROC en función del porcentaje de casos perdidos y la variable dependiente . . . . .	237
12.1. Ajuste predictivo de los modelos estimados . . . . .	275

## ÍNDICE DE TABLAS

# Agradecimientos

*"And if it wasn't for  
the music I don't know  
what I do"*

(Indeep, 1982)

Me gustaría agradecer enormemente su apoyo y ayuda a todas las personas que han hecho posible que este texto haya sido editado. Sé que es difícil no olvidarse de las personas que han servido de catalizadores ante la creación de este trabajo; sin embargo, me gustaría dar las gracias a todas y cada una de ellas por haber hecho posible, directa o indirectamente, que este texto vea la luz. Y esto es así porque, contrariamente a lo que piensen otras personas respecto a su propio trabajo, considero que como mucho el 3% de este trabajo, siendo generoso, se debe única y exclusivamente a mi contribución. O lo que es lo mismo, cualquier persona en mi situación lo habría hecho igual o, incluso, mejor. Esto es, que este trabajo no hubiese existido si no hubiese tenido la ayuda de un gran número de personas que me han guiado, moldeado y ayudado a lo largo de mi vida. Sin su soporte nada de esto hubiese sido posible. El número de nombres propios que aparecen en esta sección es muy reducido en comparación a la realidad del asunto. Nombrarlos a todos y todas sería virtualmente inviable pero quiero dejar claro que todos estáis en mi mente.

*Me gustaría aquí, en este sentido, hacer una apreciación particular en relación a los que han sido mis maestros durante esta etapa de mi vida. En primer lugar, querría perfilar ligeramente el concepto que tengo de «maestro». Para mí, un maestro no es solamente el maestro oficial, como por ejemplo fue mi primer maestro José Figueroa. Más bien, considero que han sido mis maestros todos aquellos de los que he aprendido algo. Incluso, en esta definición tan general, creo que cabe considerar a los libros, a los programas informáticos, a los presentadores de radio o televisión y muchos más como mis maestros o maestras. Además, dado que estamos en un contexto académico me voy a permitir el lujo de introducir aquí una idea que me viene rondando últimamente la cabeza: pienso que para salir airoso en este trabajo (me refiero al académico) creo que no puedes olvidar a ninguno de tus maestros y tienes que intentar que tus alumnos se acuerden de ti. Como he dicho anteriormente, tengo a todas las personas que me han ayudado en mi mente (todas son potencialmente mis maestros) aunque no aparezcan aquí con nombres propios. Para organizar un poco esta sección de agradecimientos, y dado que este trabajo se enmarca dentro del área de la metodología de las ciencias del comportamiento, la he dividido en cuatro bloques que se referirían a los agradecimientos familiares, laborales, de soporte logístico y académicos y/o de investigación. Como suele suceder en la mayoría de las clasificaciones esta clasificación no es mutuamente excluyente y algunos elementos podrían estar claramente en varias categorías. No obstante, considero que éste es un método útil para organizar el texto.*

En primer lugar, quiero agradecer a mi familia su papel durante el periodo de tiempo en que he estado desarrollando este trabajo. Pero no agradezco, como suele verse en este tipo de trabajos, el que hayan soportado numerosas horas de trabajo incesante en detrimento de mi vida familiar; sino que les agradezco profundamente el haberme hecho notar cuando tenía que desconectar de este tema y cuando tenía que dedicarme a cultivar mis relaciones familiares. En este sentido, mi mujer, compañera y no siempre bien o suficientemente ponderada amiga ha sido la persona que más me ha ayudado. Asimismo, a mi hija e hijo he de agradecerles que, en cierto modo, me transmitan su energía y me hagan recordar lo que es (y fue) ser niño. Un agradecimiento especial también requieren –adueñándome de la terminología anglosajona – mis madres, la biológica y la «legal»; mis padres, el biológico y el «legal»; mis hermanas «legales» y mis hermanos por haberme apo-

yado y por ayudarme a ser como soy. Ni que decir tiene que estos agradecimientos son extensibles al resto de familiares. Permítaseme aquí hacer un inciso sobre lo que considero como «mis orígenes», en detrimento de nombrar cada uno de los nombres de los integrantes de mi familia y que, en definitiva, son la materia prima de lo que soy en la actualidad.

El caso es que, siguiendo con la analogía anglosajona presentada anteriormente, creo que mi trabajo – este trabajo, por ejemplo – tiene un alto componente de lo que se podrían denominar los pilares de mi *yo*. Por un lado, aludiendo a mi familia legal, creo que mi trabajo se parece mucho a la construcción y más concretamente al gremio del alicatado. En cierto modo, tengo la sensación de utilizar, y de necesitar, la creatividad que tiene que poner en marcha un albañil cuando se enfrenta ante la tarea de alicatar una cocina o un cuarto de baño teniendo en mente el conservar las proporciones óptimas de espacio, geometría y distribución de los azulejos. Es más, igual que el albañil utiliza ladrillos para construir edificios, creo que yo utilizo palabras para construir textos. Creo, por otro lado, que además de todo esto, me he apropiado del afán de mi familia legal por la estética y geometría del ambiente. Por poner un ejemplo, cuando se recubre una piscina con *gresite* (también denominado como *mini-baldosines*) hay que tener un especial cuidado para que las planchas de material sigan la forma geométrica de la superficie y para que quede estéticamente agradable. Yo he tenido una sensación similar de ambición por la belleza geométrica de este texto cuando he trabajado con el editor de textos y cuando he diseñado las figuras que se incluyen en este trabajo.

En el caso de mi familia biológica he de confesar que, pese a que en cierta etapa de mi desarrollo (la adolescencia) lo odiaba, le estoy desmesuradamente agradecido por haberme integrado y familiarizado con la agricultura. En primer lugar, porque como se verá más abajo en esta tesis, la agricultura te genera cierto espíritu emprendedor. Cuando te dedicas a ella tienes que estar continuamente decidiendo si plantar de una cosa o de otra, si utilizar un tratamiento u otro, decidiendo si tienes que encañar los frutos de una manera o de otra; y durante todo ese proceso te ves empujado a emprender acciones que, según tus deseos, han de ir encaminadas a conseguir un éxito económico.

Además, de la agricultura he aprendido los valores del sacrificio y la paciencia. Imagínese que usted planta un terreno, pongamos 1000 metros cuadrados (lo que

son aproximadamente dos *marjales*), de pimientos con el objetivo de obtener un beneficio económico a partir de ello. Pues bien, usted tendrá que dedicar mucho trabajo, esfuerzo y sacrificio a ese pedazo de tierra y a esas plantas antes de obtener alguna recompensa visible; a parte, claro está, de la placentera satisfacción de ver como se desarrollan esas pequeñas plantitas para dar pimientos. Esto es, tendrá que regar, encañar, controlar plagas, solucionar carencias nutritivas de las plantas, soportar altas temperaturas bajo el sol, recolectar o controlar las instalaciones de su terreno y todo ello sin ningún tipo de recompensa económica. No será hasta el momento posterior a la recolección cuando pueda saborear el resultado de tanta paciencia y sacrificio. Así pues, por suerte o por desgracia, me ha tocado vivir esta situación y por ello he asimilado estos valores de paciencia y sacrificio en el trabajo. E indudablemente, esta tesis es fruto de todo esto.

Sin embargo, creo que lo más bonito que me ha aportado la agricultura es esta idea menos mercantilista *del cultivar*, ya que cultivar 2000 plantas de pimientos puede ser análogo a escribir una tesis como esta. El agricultor observa con paciencia el crecimiento de cada una de sus plantas con extrema delicadeza, como si fuesen sus hijas, de manera similar a como un doctorando percibe la evolución de su texto a medida que pasan los meses y que va ampliando su conocimiento sobre el tema de su trabajo. Ante cualquier vicisitud en su plantación el agricultor se preocupa e intenta poner remedio antes que perder a su prole vegetal de igual manera que el doctorando sufre cuando no consigue plasmar con claridad alguna idea en el lienzo de las palabras. En definitiva, y para no dar más rodeos, creo que la persona que escribe una tesis está cultivando el conocimiento de manera parecida a como el agricultor cultiva una plantación de pimientos.

Considero que el bloque relativo a los agradecimientos en el contexto laboral es muy importante desde que me adueñé de una declaración enunciada por el Profesor Nazario Yuste que venía a decir algo así como que *la etapa de la vida que denominamos adultez se caracteriza por dos tareas centrales: amar y trabajar*. En este sentido, creo que las empresas J. Juarez Cabrera S.A. y Paris & Paris S.L. merecen un agradecimiento especial porque compartí uno de los primeros años de mis estudios de doctorado con un equipo de personas sensacional y porque trabajar en esta gran bodega repartiendo, principalmente vino, me hizo rescatar de mi sombría memoria algo de mis raíces. Por otro lado, también tengo que agradecer enormemente la oportunidad que me ofreció la Universidad de Granada para

trabajar en el proyecto *Factores disposicionales y estructurales que influyen en la toma de decisión del jurado acerca del objeto del veredicto y nuevas técnicas de análisis en la toma de decisiones*, subvencionado por el Ministerio de Ciencia y Tecnología (Proyecto BSO2002-03347 del Plan Nacional I+D+I) y cuya investigadora responsable fue la Profesora Emilia Inmaculada de la Fuente. Fue durante este periodo de tiempo cuando adquirí la mayor porción de conocimiento base sobre redes bayesianas que me ha permitido elaborar este trabajo. En tercer lugar, me gustaría agradecer profundamente a Carlos Santamaría (Grupo Jarquil y Ares Andalucía S. L.) haber contado conmigo para colaborar en el desarrollo del proyecto *FIDELIA* y del Cuadro de Mando del grupo. Por último, he de agradecer a mis compañeros y compañeras de área el estar proporcionándome un entorno de trabajo ameno a la vez que enriquecedor. Al Profesor Antonio José Rojas he de darle las gracias por haberme proporcionado un marco general para el óptimo desempeño de mi tarea docente en la asignatura de Psicometría así como por haberme hecho reflexionar frente a temas delicados en nuestras «discusiones de pasillo». Igualmente, tengo que agradecer el mismo apoyo docente a las Profesoras María Leticia de la Fuente y a María Victoria Acosta con la asignatura de Diseños de Investigación en Psicología. Al Profesor Juan Manuel Llopis le agradezco, aunque parezca trivial, el ameno viaje a Oviedo para asistir al III Congreso Europeo de Metodología. Es de obligada referencia mi agradecimiento a la Profesora María Isabel Archilla por haberme dado una lección importante sobre la gestión de las clases y la relación con el alumnado. Por último, Miguel Ángel Sanz (anteriormente Sánchez) también merece una gran porción de mi agradecimiento porque me ayudó en todo momento a funcionar óptimamente en el puesto de trabajo que ocupó en la Universidad de Almería.

Como decía anteriormente, es complicado establecer un sistema de clasificación inequívoco y la dimensión de mis agradecimientos relativa al soporte logístico es un ejemplo de ello ya que algunas referencias que aquí se incluyen podrían estar en otras categorías y todo el conjunto de elementos podrían englobarse bajo este segmento de clasificación. En primer lugar, he de agradecer a la Fundación Mediterránea Universidad-Empresa de la Universidad de Almería, y más concretamente a Carlos Jesús Cano y Ana Belén Egea, la oportunidad que me brindaron para participar en el proyecto de investigación *Estudio Sobre Actitudes y Recursos Emprendedores en el Contexto Universitario* y que fue el germen de este traba-

jo. Asimismo, he de agradecer desmesuradamente la colaboración desinteresada que prestaron tanto los profesores y profesoras que cedieron su tiempo docente para que sus alumnos cumplimentasen la encuesta que se ha utilizado en esta tesis; así como los centenares de estudiantes universitarios que tomaron parte en el estudio. Por otro lado, quiero agradecer abiertamente el inmejorable servicio que me ha facilitado la Biblioteca Nicolás Salmerón de la Universidad a todos los niveles, aunque me gustaría remarcar la excelente labor que llevan a cabo en el servicio de préstamo interbibliotecario. Indudablemente sin su ayuda mucho de este trabajo no sería así. Asimismo, merecen agradecimiento las Bibliotecas Municipales de Adra y Priego de Córdoba. Un agradecimiento especial merece también el Vicerrectorado de Investigación de la Universidad de Almería, que me ha ayudado a gestionar administrativamente los proyectos de investigación que han sufragado el desarrollo de este trabajo. También he de agradecer al Servicio de las Tecnologías de la Información y las Comunicaciones el papel que ha desempeñado como coadyuvante de la génesis de esta tesis. En otros términos, también quiero agradecer el gran servicio prestado por la Guardería CASE-UAL, sin su ayuda mi trabajo se hubiese demorado más en el tiempo. Por último, tengo que reconocer la gran ayuda que ha supuesto para el desarrollo de este trabajo la Escuela Oficial de Idiomas de Almería. He de agradecer, en este sentido, a todos mis profesores y profesoras la profesionalidad que ha propiciado la optimización de mi manejo del inglés y, como consecuencia, ha favorecido al enriquecimiento teórico-práctico de esta tesis. Agradezco a Brent Boerlage su ayuda técnica en el uso de Netica™.

He de dar las gracias también a toda la gente que trabaja en el proyecto L<sup>A</sup>T<sub>E</sub>X, y que de una manera u otra han coadyuvado en la génesis de este trabajo. En especial, me gustaría agradecer al Profesor José Luis Guzmán sus consejos para trabajar con imágenes `postscript`, a Erik Meijer por ayudarme en la adaptación española del paquete `apacite`, a Fernando Reche por ayudarme con el formato del documento, a Antonio Salmerón por sus consejos sobre el uso del paquete `babel` y a Isabel María Ortiz por sus fuentes para crear «supertablas».

El campo de agradecimientos que he definido como académico o de investigación es el más voluminoso. No obstante, todos estos agradecimientos podrían encajar dentro de lo laboral pero me pareció práctico hacer esta diferenciación. En primer lugar me gustaría agradecer a la Profesora María Soledad Navas el

haber delegado en mí parte de la responsabilidad para llevar a cabo un proyecto de investigación propuesto por ASALER (Asociación Almeriense de Emigrantes Retornados). Asimismo, he de agradecer a Óscar Palomino (Innova Humana Consultores) y a Juana María Sánchez (ASALER) su soporte para lo que se ha convertido en el primer libro donde comparto la autoría (Ruiz-Ruano y López, 2008b) y por haberme involucrado en una investigación que me permitió viajar por primera vez en avión y conocer parte de Latinoamérica (Argentina y Venezuela) y Europa (Alemania y Francia), así como Galicia. También tengo que agradecer a la Profesora María Isabel Molina del I.E.S. Turaniana (Roquetas de Mar, Almería) que contase conmigo para analizar los datos del proyecto de investigación «Educación para la Democracia y la Convivencia Pacífica» y a la Profesora Elena Marcos por invitarme a tomar parte en la investigación que llevó a cabo sobre los beneficios de la lactancia materna.

Por otro lado, tengo que mostrar mi más sincero agradecimiento al Departamento de Neurociencias y Ciencias de la Salud de la Universidad de Almería que para mí fue, y sigue siendo, un referente a seguir. Gracias a su subvención parcial pude llevar a cabo el trabajo que me condujo a publicar mi primer artículo experimental (J. López y García, 2007d). Me gustaría mostrar mi agradecimiento al Profesor Francisco Antonio Nieto por, además de haberme tutorizado durante algún tiempo, haberme enseñado a trabajar con animales en el laboratorio de psicobiología. También agradezco enormemente al Profesor Diego Alonso las explicaciones orientadas a clarificarme algunas nociones del cálculo de probabilidades y la comprensión del teorema de Bayes. Los Profesores Luís Fuentes y Juan José Ortells me dieron consejos muy útiles durante la defensa de mi trabajo conducente a la obtención del Diploma de Estudios Avanzados del Programa de Doctorado *Neurociencia Cognitiva* que me han servido para mejorar esta tesis. Entre otros, me gustaría dar las gracias por diferentes motivos a los Profesores José Manuel Cimadevilla, Fernando Sánchez, Roberto Álvarez y a las Profesoras Inmaculada Cubero, María del Carmen Sánchez y Pilar Flores. Por último, un agradecimiento especial para Estrella López por su brillante gestión administrativa.

Una mención especial de agradecimiento merece también el Profesor Dermot Barnes-Holmes de la *National University of Ireland* en Maynooth (NUIM) por haber dado crédito a mis ideas sobre el uso del formalismo de las redes bayesianas

para modelar el aprendizaje causal y por haberme invitado a compartir con él y su equipo una estancia de investigación en la República de Irlanda. Asimismo, agradezco a Derek Walsh y a Anne Dooley el soporte técnico y administrativo que me proporcionaron durante mi estancia en la NUIM. Por otro lado, agradecer a Corinna Stewart y Nigel Vahey su cálida acogida en Irlanda.

Por último pero no menos importante, e indudablemente, quiero dar las gracias al Profesor Juan García por su apoyo incondicional, por sus sabios consejos, por haber confiado en mí y por hacerme sentir como el protagonista de la historia «El verdadero valor del anillo» tras la enseñanza que le proporciona su maestro (Bucay, 1999). El hecho de que el agradecimiento a Juan aparezca aquí no es casual sino deliberado. Y esto ha sido así porque confío en la concepción contemporánea que se tiene sobre el funcionamiento de la memoria (p.e., J. M. Ruiz, 1994) y me gustaría producir un efecto de recencia para acentuar el papel de Juan en esta tesis. Si al principio decía que mi contribución a esta tesis había sido de alrededor de un 3%, creo que en el caso de Juan la contribución, y posiblemente me quede corto, es de treinta veces más. Esto es así porque fue él el que me planteó el problema para que fuese tratable desde un punto de vista práctico, el que medió para que pudiese recoger los datos y el que me ha estado apoyando laboral y emocionalmente. Sin embargo, he de confesar que este camino que ahora ve con más nitidez el horizonte no ha sido fácil y el sufrimiento ha estado presente de principio a fin. Las numerosas correcciones y las propuestas de modificación que necesariamente Juan me ha sugerido durante todo el proceso han llegado a, casi, hacerme sucumbir frente a la indefensión aprendida; aunque, sin embargo, creo que son el fruto natural y esperable de una relación académica de este tipo. Así pues, creo que el verdadero artífice de este trabajo es Juan García y creo que él se merece mi más cercano, cordial, franco e hiperbolizado agradecimiento. GRACIAS.

# Prefacio

*"It's nice to be important  
but it's more important to be nice"*

(Scooter, 1995)

Me gustaría, en primer lugar, hacer una apreciación aquí sobre la razón que me ha llevado a introducir, así como la topografía que las caracteriza, citas al principio de cada capítulo. La mayoría de las citas que aparecen de este modo están en inglés y se podrían achacar varios motivos a este hecho. Por ejemplo, si estuviésemos hablando del mercado musical podríamos decir que así se «vende más». O también se me podría acusar de vanguardista porque el inglés es un idioma que «está de moda». Incluso se me podría tachar de imperialista, capitalista o *globalista* porque el inglés es la lengua que «domina» el planeta. Para ser sinceros no sé si alguna de ellas, una combinación concreta de algunas, todas, ninguna u otras razones que no he contemplado aquí son las razones que me han empujado a redactar las citas de cada capítulo en inglés. Pero lo que sí tengo claro es que casi todas ellas pertenecen a obras de arte, y eso sí ha sido decidido por mí deliberadamente. Asimismo, he intentado que la mayor parte de las referencias que cito sean fácilmente accesibles desde Internet principalmente, dado que muchas son fragmentos de canciones, a través de [www.youtube.com](http://www.youtube.com).

Este hecho puede parecer descontextualizado ya que el objetivo de este texto es una descripción científica de un fenómeno y el arte y la ciencia se pueden

considerar campos contrapuestos. No obstante, da la casualidad, que desde mi punto de vista la ciencia es entendida como un tipo particular de arte. Siendo consciente de que si hubiese dicho esto públicamente en una reunión científica me estarían dando palos a diestro y siniestro, explicaré brevemente a lo que me refiero con un ejemplo concreto relativo a las gráficas estadísticas. Las gráficas estadísticas son elementos muy útiles para resumir la información contenida en un grupo de datos. En este trabajo he tenido especial cuidado para introducir gráficos de alta calidad. De hecho, todos los gráficos incluidos en este trabajo son imágenes digitales vectorizadas. A su vez, suelo ser muy escrupuloso con el diseño de los gráficos ya que me gusta añadirles o quitarles elementos para que resulten en un aspecto «bello». De este modo, cuando me dedico a personalizar los gráficos que resumen mis datos me siento, en cierto modo, como el pintor que plasma un paisaje en su lienzo o como el compositor que desafía las fracciones de tiempo en el pentagrama.

El trabajo que aquí presento es el reflejo del flagrante éxito y fracaso de mi temprana trayectoria profesional. Evidentemente, como la mayoría de las personas, considero que este trabajo – mi trabajo – es el mejor de todos; de igual manera que un padre o una madre considera que su hijo o hija es un cielo, un encanto o la mejor persona del mundo cuando en realidad – o para el resto de la sociedad – es un o una sinvergüenza de mucho cuidado. No obstante, creo que esta tesis denota puntos negativos y positivos en mi corta experiencia académica.

En primer lugar, considero que este trabajo es un fracaso para mí porque cuando empecé a estudiar Psicología, a finales de 1998, me encontraba entusiasmado con la idea de aproximarme al estudio de la psique humana en términos biológicos, a nivel neural, y me fascinaba la idea de entender como trabaja nuestro cerebro para dar lugar a una concepción unificada de la realidad. Sin embargo, además de saciar mi hambre intelectual relativa a la mecánica bio-físico-química que hace posible nuestras sensaciones, nuestros procesos cognitivos, nuestra conducta y nuestras emociones; me fui dando cuenta, poco a poco, del amplio espectro de conocimientos que la disciplina psicológica podía aportarme. Antes de acabar el primer ciclo de la titulación ya había establecido contacto directo con el laboratorio de psicobiología y había colaborado en el desarrollo de estudios empíricos con animales experimentales. Los años pasaron y la perspectiva con la que empecé a estudiar psicología fue madurando con el tiempo. La gran fascinación con que me

aproximé a la disciplina psicológica se fue complementando a medida que conocía las grandes áreas que integran la materia y continué formándome con una gran tendencia al ámbito de la investigación.

Cuando llegué al último curso de la carrera tenía una visión general de lo que era la Psicología y mi interés por las bases biológicas de nuestra psique continuaba martilleando en mi cabeza. De este modo, me decidí a cursar la asignatura más práctica de la licenciatura (el *Practicum*) en el área de conocimiento que más me había motivado hasta entonces, la Psicobiología. Sin embargo, durante ese año, y con el objetivo de completar el número de créditos de libre configuración del programa de estudios; me matriculé en una asignatura de Ingeniería Informática titulada *Neurocomputación*. Quizá este último evento sea el responsable, aunque en parte o indirectamente, de que hoy esté escribiendo estas palabras. El tópico en el que me centré durante este año académico fue el de las Redes Neuronales Artificiales (RNA). Y tras un, relativamente, largo periodo de estudio, elaboré un monográfico donde pretendí plasmar los conocimientos y las ideas que había incubado en relación a la simbiosis que se establecía entre las Ciencias de la Computación y la Psicología en los últimos tiempos. En el citado informe, mencioné una idea de la Inteligencia Artificial orientada a la predicción de eventos naturales exacerbados tildándola con un componente social. Y aunque mi aproximación al tema fue únicamente exploratoria, me satisfizo profundamente.

Asimismo, un punto arquimédico en mi corta trayectoria profesional lo supuso la asignatura optativa *Aprendizaje y Cognición*. El caso fue que cuando intenté matricularme en esta asignatura – con el deseo de saciar mi apetito por este tipo de conocimiento – el cupo ya estaba cerrado con lo que solicité, formalmente y por escrito, que se me aceptase como alumno en términos administrativos en aquella asignatura. Dado que el tiempo pasaba y no recibía respuesta administrativa ante mi reclamación decidí asistir a las clases de esta materia. El tiempo pasó, el cuatrimestre avanzó y cursé la asignatura como si hubiese estado matriculado en ella formalmente sin noticias sobre mi reclamación. Mi logro académico en la asignatura fue marcadamente satisfactorio pero la «tiranía» burocrática segó mis ilusiones de ver reflejado en mi expediente tan brillante resultado curricular. Desde el principio, las Profesoras de la asignatura, Pilar Flores y María de los Ángeles Fernández, me apoyaron en mi reclamación; incluso el Decano de la Facultad de Humanidades y Ciencias de la Educación, el Profesor Antonio Daniel Fuentes,

y la persona competente en el asunto, la Profesora Carmen Pozo, mediaron en mi causa pero nuestra insistencia fue infructífera. Finalmente, exigí poder reunirme con el Rector de la universidad, el Profesor Alfredo Martínez, para tratar este tema pero su postura y actitud ante mi problema volvieron a hundirme en la desilusión y acabé aceptando, por suerte o por desgracia, que había perdido virtualmente el tiempo que dediqué a trabajar dicha asignatura.

Por lo tanto, dado lo expuesto en estos últimos párrafos, considero que este trabajo – que se desvía claramente de estos intereses académicos e investigativos primigenios – representa un nítido fracaso en mi corta carrera profesional. Sin embargo, también puedo considerarlo como un éxito cuando reflexiono sobre las consecuencias éticas, análogamente a Sharkey (2008), del trabajo que podía haber llevado a cabo en otra situación. Concretamente, me hubiese gustado poder haber desarrollado un modelo formal de computación neural utilizando el concepto de *computación local* que aprendí estudiando sobre las redes bayesianas. Así, podría haberme dedicado a trabajar en el campo del aprendizaje y esta tesis versaría sobre modelos probabilísticos de aprendizaje. Más específicamente, me hubiese gustado haber podido indagar en el problema que yo denomino *paradoja de la integración neural*. Permítame el lector que le exponga brevemente mis ideas al respecto. Consideremos **una** neurona del cerebelo, la cual puede recibir información de, aproximadamente, 130.000 neuronas (Kandel, Schwartz, y Jessell, 1995/1997). Consideremos también, simplificando tremendamente el funcionamiento neural, que la comunicación entre neuronas está mediada por un lenguaje de todo-o-nada (potencial de acción sí o potencial de acción no, esto es básicamente lo que plantea la Ley de la Tasa). En este caso concreto, una neurona purkinje del cerebelo tendría que evaluar una tabla de verdad con  $2^{130000}$  casos antes de poder emitir una respuesta. Teniendo en cuenta que, como decía anteriormente, el lenguaje de las neuronas ha sido simplificado, por ejemplo, obviando los procesos de neuromodulación y los mecanismos de adaptación en los receptores pre y post-sinápticos; el número de casos que una neurona tendría que evaluar antes de desencadenar, o no, un potencial de acción se baraja en términos astronómicos. Aproximadamente, nuestra amiga la neurona tendría que revisar una tabla con  $7,9 \times 10^{39133}$  filas. Es difícil para mí pensar en lo que supone esta cifra en términos concretos pero el lector puede hacerse una idea de lo que sería este número sin estar expresado en modo científico teniendo en cuenta que para

escribirlo en modo convencional tendríamos que escribir un 79 seguido de más de 11 páginas de ceros en un documento de Microsoft Word con fuente Times New Roman a tamaño 12.

Este problema está parcialmente solucionado. Por ejemplo, se sabe que las neuronas no tienen que llevar a cabo esa desmesurada tarea de evaluación antes de emitir una respuesta. Más bien, las neuronas utilizan sistemas heurísticos que les permiten dar una respuesta aproximada o borrosa ante un conjunto de estimulaciones (Jastrow, 1981/1993). En cualquier caso, el problema de la integración neural que acabo de comentar podría encuadrarse dentro de lo que la Teoría de la Complejidad Computacional denomina como problemas NP-Complejo dado el dramático incremento de casos que se producen a medida que aumenta el número de elementos eferentes en la red neural simple que vengo utilizando de ejemplo. En este contexto es donde creo que el concepto de computación local podría aplicarse con relativo éxito dado que, opino, los actuales modelos de RNA se encuentran en su «edad de bronce» y que antes que otra cosa tendríamos que desarrollar un modelo de neurona satisfactorio que soportase trabajos de procesamiento más complejos.

Visto lo expuesto en las líneas anteriores podría decirse que mi actividad, y el trabajo que aquí presento, es un nefasto negocio. No obstante, pese a mis fracasos también he conseguido algunos éxitos en mi carrera profesional y personal. Por ejemplo, pese a que mis motivaciones estaban orientadas al campo de la neurociencia, bien es cierto que mis resultados académicos saturaban en el Área de Psicología Social<sup>1</sup>. En este sentido, el trabajo que aquí presento se basa en un tópico eminentemente social (las actitudes emprendedoras) y se ha usado una técnica estadística relativamente novedosa en las ciencias sociales. Si se me permite, puedo presumir de haber sido la persona que introdujo la primera definición de red bayesiana en la versión española de Wikipedia el 10 de febrero de 2005 a las 11:10 horas. Y aunque hay discusión de la relevancia científica de este portal de Internet (Giles, 2005a), me siento orgulloso del trabajo que había hecho hasta el momento y hasta lo que ha llegado. Por otro lado, también me siento orgulloso de haber podido colaborar en la adaptación del paquete `apacite` para L<sup>A</sup>T<sub>E</sub>X diseñado por Erik Meijer para crear referencias según el estilo APA

---

<sup>1</sup>Permítaseme citar un par de versos de una canción de Duncan Dhu (1989) al hilo de esta situación: *No tengas miedo de bajarte, en la primera estación, no hay nada como equivocarse, si es con el corazón.*

(American Psychological Association, 2001).

Por último, pero no menos importante, considero que es para mí un éxito el poder ser Profesor Asociado y poder presentar esta tesis en la ciudad que me vio nacer, que me ayudó a enamorarme de mi esposa, que me abrigó para concebir mi hija e hijo, que me formó como profesional, que me dio trabajo y que me colma de felicidad con su sol incansable.

*Almería, 29 de octubre de 2009*

*Jorge López Puga*

**Parte I**

**Revisión Teórica**



# Capítulo 1

## Modelado predictivo

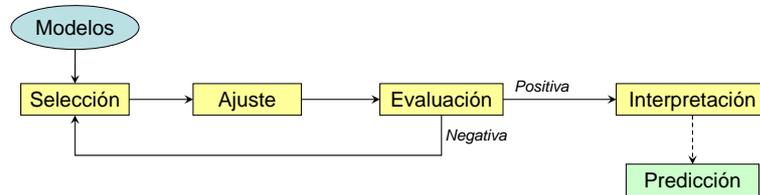
*"[...] all we are is dust in the wind,  
[...] everything is dust in the wind."*

(Kansas, 1977)

### 1.1. Introducción

Como señalan Ato y López (1996) el procedimiento de modelado estadístico consiste en la aplicación de una serie de procesos con el fin de conseguir una explicación apropiada respecto al comportamiento de un conjunto de variables. En su forma genérica, la metodología de modelado estadístico se compone de una serie de toma de decisiones que van desde la selección de un modelo hasta la interpretación de su utilidad dentro del ámbito sustantivo de la investigación, pasando por una fase iterativa de evaluación de la bondad de ajuste (Figura 1.1). Asimismo, el proceso de modelado estadístico está orientado, en última instancia, a predecir y controlar el fenómeno natural que representa (Ríos, 1995).

Existe una amplia variedad de modelos estadísticos que pueden utilizarse para representar o modelar una misma realidad, cada uno de ellos con unas peculiaridades concretas acompañadas de sus ventajas y desventajas. Este capítulo tiene como objetivo introducir las dos estrategias de modelado que se han utilizado



**Figura 1.1:** El proceso de modelado estadístico. Adaptado de Ato y López (1996).

en esta tesis: la regresión logística y las redes bayesianas. En primer lugar, se introducirán los principios básicos de la regresión lineal sobre los que se sustenta el análisis de regresión logística, tanto desde un punto de vista formal como histórico.

Por su parte, en la sección dedicada a las redes bayesianas se destacará como estas herramientas estadísticas pueden considerarse técnicas surgidas en lo que ha sido denominado como *filosofía de red* en el campo del análisis de datos multivariante (J. López, García, De la Fuente, y De la Fuente, 2007).

## 1.2. Modelo de regresión lineal

El concepto de *regresión* es frecuente en los informes y tratados científicos ya que alude a un procedimiento estadístico orientado a detectar las relaciones de dependencia que se establecen entre variables. Esta necesidad de establecer o detectar las relaciones que se establecen entre ciertas variables es crucial para la investigación en todos los ámbitos. Por ejemplo, por medio de estudios de regresión podríamos intentar responder a preguntas como ¿por qué algunos individuos padecen alteraciones en sus niveles de tensión arterial?, ¿cuál es la relación que se establece entre el nivel educativo y la probabilidad de que una persona muestre actitudes racistas?, ¿cómo se relaciona el nivel de ingresos con la posibilidad de comprar cierto producto?, ¿qué factores determinan el riesgo de padecer infartos de miocardio?, ¿de qué manera están relacionadas las calificaciones académicas de un adolescente con el número de horas que dedica a jugar a videojuegos?, ¿qué relación existe entre los niveles de triglicéridos y los problemas arteriales?, ¿cómo se relaciona la tendencia de voto en las próximas elecciones con el número de hijos? o ¿cuál es el riesgo de padecer algún tipo de alergia dependiendo de la

alimentación que se ha seguido durante la infancia? Así pues, el análisis de regresión se ha convertido en una herramienta indispensable para tratar de responder preguntas de este tipo en campos como la medicina, psicología, política, o en el campo educativo.

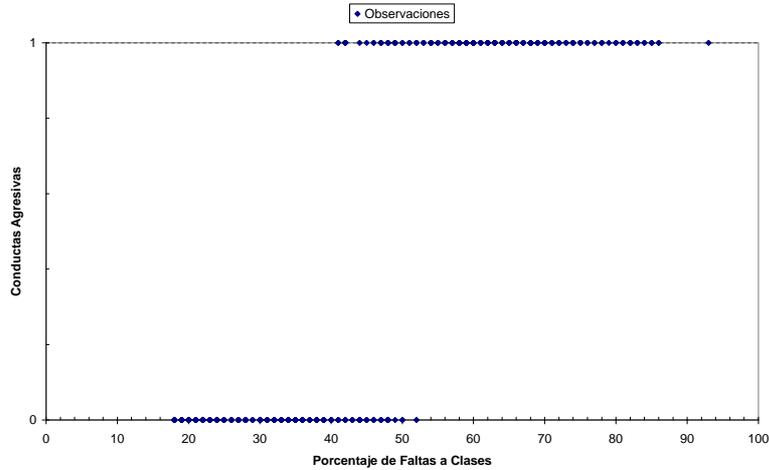
Antes de describir los fundamentos de la regresión logística es conveniente comenzar señalando los principios del modelo de regresión lineal ya que comparten aspectos comunes (Jovel, 1995) y porque dio lugar al *Modelo de Probabilidad Lineal* que es una traslación literal del modelo lineal al caso de variables dicotómicas (Long, 1997). Procediendo de este modo también podremos señalar las limitaciones de este modelo para trabajar con variables dependientes discretas.

Supongamos que estamos interesados en estudiar la relación que existe entre el porcentaje de clases a las que asiste un adolescente y la probabilidad de mostrar conductas agresivas. Para que nuestro estudio sea válido lo primero que tendríamos que hacer es seleccionar una muestra representativa de la población y recoger alguna información referente a la tasa de asistencia a clase de los adolescentes y algún indicador/es de conductas violentas. Imaginemos que hemos seleccionado una muestra aleatoria de personas, que hemos recopilado información sobre su porcentaje de asistencia a clases y que hemos observado sus comportamientos en el patio del centro escolar durante algún tiempo. Ahora supongamos que representamos los datos en un gráfico codificando la variable *Conductas Agresivas* de forma dicotómica donde el valor 0 indica ausencia de agresividad en los comportamientos y que el valor 1 representa la presencia de conductas violentas<sup>1</sup>. Una hipotética distribución de los datos sería la que aparece en la Figura 1.2.

Como se puede ver en la gráfica de la Figura 1.2 hasta llegar alrededor del 40 % de faltas a clase las conductas agresivas son inexistentes y al pasar del, aproximadamente, 50 % de ausencias en clase este tipo de problemas están siempre

---

<sup>1</sup>Esta es la forma más recomendable de codificar este tipo de variables, las que desde un punto de vista del diseño se denominan *variables dependientes, resultado* o *explicadas*, cuando construimos ficheros de datos para programas estadísticos que van a ser analizados por medio de regresión logística. Por lo general se recomienda asignar el valor 1 a las respuestas positivas y el valor 0 a las negativas para que el programa use la categoría etiquetada como 1 como la categoría de prueba y la categoría 0 como la categoría de contraste. También es recomendable asignar el valor 1 a los niveles o estados de variables que se consideran riesgos relacionados con algún tipo de respuesta o resultado en las variables que desde un punto de vista del diseño se denominan *variables independientes, criterio, predictivas* o *explicativas*. No obstante, siempre hay que comprobar cual es la codificación interna que hace el *software* que estamos utilizando ya que puede variar de unos programas a otros.



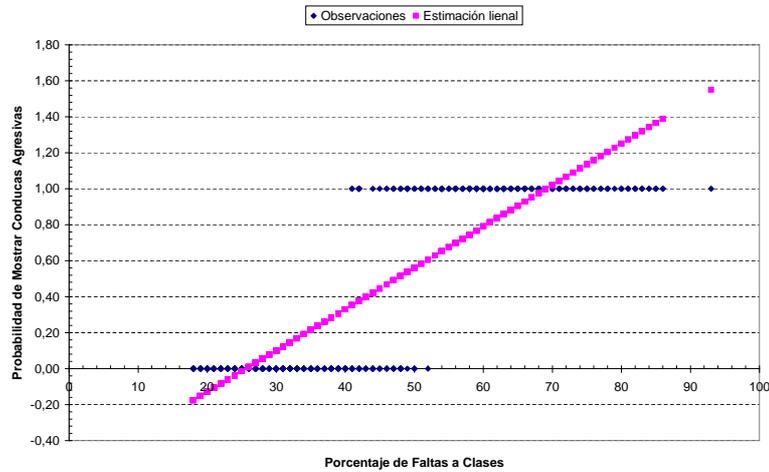
**Figura 1.2:** Conductas agresivas en función del porcentaje de clases a las que asisten los adolescentes.

presentes. Hay un sector comprendido entre el 40 % y el 50 % donde coexisten personas que muestran conductas agresivas con las que no. Esta representación gráfica puede ser de mucha utilidad descriptiva (E. King y Ryan, 2002), principalmente cuando tratamos de detectar situaciones deletéreas para el análisis de regresión logística como la separación o la cuasi-separación (p. e., DeMaris, 2002; Irala, Fernández-Crehuet, y Serrano, 1997), pero carece de valor cuando necesitamos hacer estimaciones o predicciones relativas a la relación que se establece entre la tasa de absentismo escolar y el riesgo de mostrar conductas agresivas. Para ir un poco más allá de la mera descripción de los datos tendríamos que realizar algún tipo de análisis de *regresión*. ¿Qué pasaría si realizamos un análisis de regresión lineal tomando el *Porcentaje de Faltas a Clase* como variable independiente y la variable *Conductas Agresivas* como variable dependiente?

En la Figura 1.3 aparece representada la recta ajustada de regresión lineal de la relación entre el porcentaje de faltas a clases y la probabilidad estimada de mostrar conductas agresivas. Esta generalización simple del modelo de regresión lineal aplicado a la estimación de probabilidades ha sido denominado como *Modelo de Probabilidad Lineal* (MPL) (Long, 1997). El modelo de regresión lineal, y por ende el MPL, se define como

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \dots + \beta_K x_{iK} + \epsilon_i, \quad (1.1)$$

donde  $y$  es la variable dependiente, las  $x$  son las variables independientes y  $\epsilon$



**Figura 1.3:** Distribución de los casos que muestran conductas agresivas en función de las faltas a clases y la estimación lineal de la relación.

representa un error estocástico. Por su parte, cada  $\beta_k \in \beta_K$  es un parámetro que indica el efecto de una variable independiente  $x_i$  sobre  $y$ . Por último,  $\beta_0$  es el *intercepto* u *origen* que representa el valor que toma  $y$  cuando todas las variables independientes son igual a cero.

Usando nomenclatura matricial podemos representar el modelo como

$$\mathbf{y} = \mathbf{X}\beta + \epsilon,$$

donde  $\mathbf{y}$  es un vector de valores para cada  $y_i \dots y_N$ ,  $\mathbf{X}$  es una matriz  $N \times K$  que contiene en sus columnas el conjunto de variables predictoras  $x_{i1} \dots x_{iK}$  y en sus filas el conjunto de observaciones  $x_{1i} \dots x_{Ni}$ ,  $\beta$  es el vector de parámetros asociados a cada variable explicativa más el intercepto  $\beta_0 \dots \beta_K$  y  $\epsilon$  es el vector de errores asociados a cada observación  $\epsilon_1 \dots \epsilon_N$ .

Si definimos  $\mathbf{x}_i$  como la  $i$ -ésima fila de  $\mathbf{X}$ , la ecuación 1.1 puede escribirse como

$$\mathbf{y} = \mathbf{x}_i\beta + \epsilon_i.$$

### 1.2.1. Asunciones del modelo de regresión lineal

Para garantizar la validez del modelo de regresión lineal se han de cumplir una serie de supuestos. Los supuestos del modelo se refieren, como señala Long (1997), a

dos grupos de asunciones: un primer grupo relacionado con las variables independientes, y un segundo grupo respecto a la distribución de los errores del modelo. En primer lugar se tratarán brevemente los supuestos referentes a las variables independientes y a continuación los referidos a la distribución de los errores.

### **Linealidad**

De acuerdo con la ecuación 1.1, la variable dependiente  $y$  está relacionada linealmente con el conjunto de predictores del modelo por medio del conjunto de parámetros  $\beta$ . No obstante, el modelo puede incluir variables relacionadas no linealmente con la variable respuesta por medio de transformaciones en las variables originales.

### **Colinealidad**

Las variables del modelo de regresión lineal deben ser linealmente independientes. Esto es, que ninguna de las variables predictoras puede ser una combinación lineal exacta del conjunto restante de variables explicativas. Cuando alguna variable predictora es una combinación lineal exacta de las demás, se dice que existe *colinealidad*. A su vez, cuando todas o algunas variables independientes están muy relacionadas linealmente entre sí, se dice que existe *multicolinealidad*.

La multicolinealidad es un problema ya que tiene como consecuencia estimaciones muy poco precisas de los coeficientes del modelo de regresión, sus niveles de significación son sesgados y, como resultado, la validez del modelo queda en entredicho (Juan, 2003).

En lo que respecta a los supuestos relacionados con la distribución de los errores o residuos del modelo estarían las siguientes asunciones.

### **Media condicional nula para $\epsilon$**

La expectativa, o promedio, condicional del error es cero; esto es

$$E(\epsilon_i | \mathbf{x}_i) = 0.$$

Esto significa que para un conjunto dado de predictores, el error esperado es cero. Esta asunción implica que la expectativa condicional de  $y$  dado  $\mathbf{x}$  es una combinación lineal de las variables independientes.

**Varianza, correlación y distribución de  $\epsilon_i$** 

Como señalan I. M. Ortiz y Oña (2003), la varianza de los errores es constante. Esto es,

$$\text{Var}(\epsilon_i) = \sigma^2.$$

Además, no existe correlación entre los errores del modelo.

$$\text{Cov}(\epsilon_i, \epsilon_j) = 0.$$

Por su parte, la distribución de los errores del modelo de regresión lineal debe seguir una distribución normal. Esto es,

$$\epsilon_i \sim N(0, \sigma) \quad \text{o} \quad Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma)$$

**1.2.2. Interpretación de los coeficientes de regresión**

Dado que la *derivada parcial* de  $y$  respecto a  $x_i$  está definida como

$$\frac{\partial E(y|\mathbf{x})}{\partial x_k} = \frac{\partial \mathbf{x}\beta}{\partial x_k} = \beta_k,$$

la interpretación de los coeficientes del modelo de regresión lineal se hace relativamente sencilla. Esto significa que el resultado de la derivada parcial de  $y$  respecto a  $x$  es la pendiente de la recta que relaciona a  $x$  e  $y$ , manteniendo el resto de las variables del modelo constantes. Por tanto, la interpretación de los coeficientes  $\beta_k$  se haría del siguiente modo:

- Por cada unidad de aumento en  $x_k$ , el cambio esperado en  $y$  equivale a  $\beta_k$ , manteniendo el resto de variables del modelo constantes.

Siguiendo con el ejemplo que se introdujo anteriormente sobre la relación que se establece entre el porcentaje de asistencia a clases y la probabilidad de mostrar conductas agresivas, si tenemos en cuenta que la ecuación de la recta que define al modelo es

$$\widehat{PCA} = -0,59 + 0,023 \times PFC, \tag{1.2}$$

donde  $\widehat{PCA}$  se refiere a la probabilidad estimada de que un adolescente muestre conductas agresivas y  $PFC$  representa el porcentaje de faltas a clases; podríamos decir que por un aumento del 1 % en las faltas a clase la probabilidad de mostrar conductas agresivas aumentará en 0,023. Como se puede observar en la Figura 1.3 existe una relación directamente proporcional entre el aumento del porcentaje de faltas y la probabilidad de mostrar conductas agresivas. Sin embargo, pese a que la bondad de ajuste del modelo es buena ( $R^2 = 0,74$ ), se observa que esta estimación de la probabilidad de mostrar conductas agresivas a partir del porcentaje de faltas a clase no es apropiada porque las predicciones que hace el modelo se salen de los límites que habíamos definido en nuestra variable dependiente (0 y 1). Por ejemplo, ¿cuál sería la probabilidad de mostrar conductas agresivas para un adolescente que falta a clase un 93 % de las veces? El modelo de regresión lineal nos diría que la probabilidad estimada de mostrar agresividad es de 1,55 ( $-0,59 + 0,023 \times 93$ ), lo cual cae fuera de los límites de la variable tal y como la hemos definido.

### 1.2.3. Problemas con el Modelo de Probabilidad Lineal

Como señalan Silva y Barroso (2004) el modelo de regresión lineal ha sido objeto de una amplia difusión en el entorno científico, pero sólo tiene sentido cuando las relaciones que se establecen entre la/las variable/es independiente/es y la dependiente es de tipo lineal. Cuando la relación funcional que existe entre los predictores y la variable explicada no es de naturaleza lineal tenemos que prescindir de este modelo y utilizar otra alternativa.

Para Jovel (1995) el problema principal del modelo de probabilidad lineal es su reducida sensibilidad para predecir correctamente los valores de la variable dependiente. Long (1997) se refiere a esta limitación como la *predicción sin sentido* ya que las estimaciones del modelo lineal aplicadas al caso de variables dicotómicas genera valores que, como se ha mostrado anteriormente, se salen del rango aceptable de la probabilidad. Como se puede observar en la Figura 1.3, cuando el porcentaje de absentismo baja del 25 % o cuando pasa por encima del 70 % las estimaciones de la probabilidad de que un adolescente muestre conductas agresivas salen fuera de los límites permitidos de un valor probabilístico.

Otro conjunto de limitaciones del modelo de probabilidad lineal se deriva del restringido número de valores que puede llegar a tomar la variable dependiente

(Jovel, 1995). En primer lugar, tenemos el problema de la *normalidad* ya que para un valor dado  $x_* \in x$  el error de la estimación lineal viene dado por  $\epsilon_1 = 1 - E(y|x_*)$  o  $\epsilon_0 = 0 - E(y|x_*)$  dado que  $y$  sólo puede tomar los valores 0 y 1. Claramente este tipo de errores no se distribuyen siguiendo una distribución normal y por consiguiente el modelo viola uno de los supuestos del modelo de regresión lineal. Así pues, aunque los parámetros del modelo no están sesgados (p. e., Jovel, 1995; Long, 1997), los errores estimados del modelo sí lo están y, por tanto, los intervalos de confianza junto a los estadísticos que dependen de ellos mostrarán altos niveles de sesgo.

Como consecuencia de la limitación anterior no podemos asumir que la distribución de los errores tenga una varianza constante como cabría esperar en una distribución normal ( $N(0, 1)$ ), luego se produce otra trasgresión flagrante del modelo de regresión lineal. A esta limitación se le suele denominar *heterocedasticidad de los errores*.

Para terminar, el principal problema que presenta el modelo lineal de probabilidad en opinión de Long (1997) es la *forma funcional*. Esto es, dado que el modelo es lineal, cada unidad de incremento en  $x_k$  produce un cambio constante en la probabilidad de un evento de la magnitud  $\beta_k$ . Así pues, este incremento es independiente del valor de  $x$ , lo que es irreal en muchas situaciones prácticas. Por ejemplo, consideremos un modelo de probabilidad lineal que estima una reducción del 0,295 en la probabilidad de que una mujer consiga un trabajo por cada hijo adicional. Esto supone que una mujer con cuatro hijos tendrá una probabilidad 1,18 veces más pequeña de encontrar trabajo que una mujer con un único hijo. Por el contrario, sería más realista que el primer hijo disminuyese la probabilidad de encontrar trabajo en un 0,3, el segundo en un 0,2, el tercero en un 0,08 y así sucesivamente. Esto es, que el modelo debería ser no-lineal y que el cambio en las probabilidades estimadas fuese disminuyendo a medida que la probabilidad tiende a 0 y a 1.

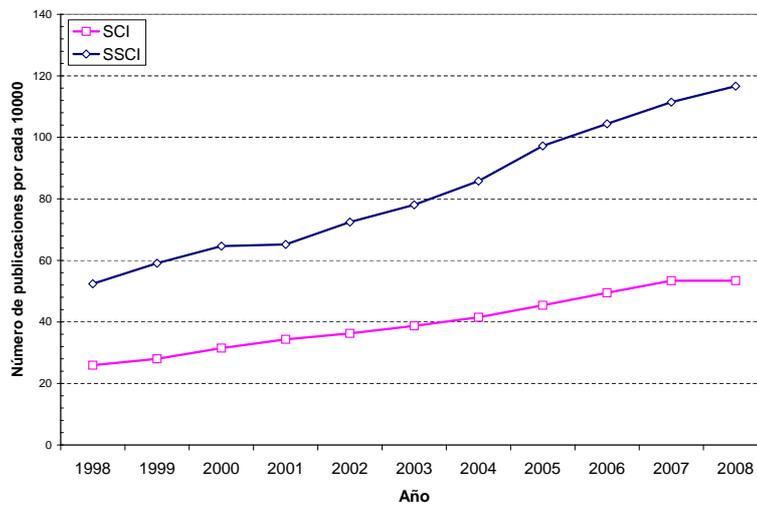
El análisis de regresión logística surgió para subsanar este fallo en las probabilidades predichas cuando la variable dependiente que estamos utilizando es dicotómica.

### 1.3. Regresión logística

La regresión logística surgió durante la década de 1960 como alternativa al procedimiento de estimación de los mínimos cuadrados ordinarios (OLS, *Ordinary Least Squares*) tradicionalmente usado en el modelo de regresión lineal (Long, 1997; Ortega y Cayuela, 2002) y su objetivo era estimar la probabilidad de ocurrencia de un evento como función de un conjunto de variables independientes (Walker y Duncan, 1967). No obstante, su uso empezó a popularizarse a partir de los años setenta cuando comenzó a implementarse en el *software* estadístico (Peng y So, 2002). El nacimiento de la regresión logística, por otro lado, estuvo condicionado por un contexto histórico dominado por una crisis en las herramientas estadísticas orientadas a la predicción ya que las herramientas de predicción discreta como los antecedentes de las Redes Neuronales Artificiales, los *perceptrones*, mostraban serias limitaciones para hacer predicciones ante problemas relativamente simples (Cowell, Dawid, Lauritzen, y Spiegelhalter, 1999; Delgado, 2003; Quinlan, 1991; SPSS y Recognition Systems, 1997).

Aunque existen procedimientos orientados a la obtención de modelos logísticos donde las variables dependientes constan más de dos niveles (regresión logística multinomial) o valores ordinales (regresión logística ordinal) (p.e., Long, 1997; Pérez, 2005), este trabajo se centrará en el modelo de regresión logística binaria por ser el más clásico, el método más usado en la investigación y porque los modelos multinomiales son una generalización del modelo binario para el caso en que la variable dependiente tiene más de dos estados.

La regresión logística es uno de los procedimientos estadísticos que más se utilizan en la investigación científica. Por ejemplo, E. King y Ryan (2002) hicieron notar que en 1999 se publicaron 2770 artículos, la mayor parte de ellos relacionados sustancialmente con el ámbito bio-médico, en los que el término «*regresión logística*» aparecía en el título o en alguna de sus palabras clave. Por su parte, el uso de la regresión logística en las ciencias sociales ha sido mayor, en proporción, que en resto de las áreas científicas. Por ejemplo, si tomamos el periodo de tiempo que va desde 1998 a 2007 y buscamos el número de publicaciones en las que el tópico haya sido la regresión logística en las bases de datos del SCI (*Science Citation Index*) y del SSCI (*Social Sciences Citation Index*) obtendríamos una



**Figura 1.4:** Evolución del número de publicaciones en el *ISI Web of Knowledge*<sup>SM</sup> donde el tópico fue la regresión logística.

tabla similar a 1.1<sup>2</sup>. Como se puede observar en la Tabla, el número de trabajos publicados cuyo tópico fue la regresión logística es considerablemente menor en las ciencias sociales en toda la serie temporal. No obstante, si se relativizan estas cantidades en relación al total de publicaciones por año y base de datos podemos observar que la regresión logística es más popular en las ciencias sociales (Figura 1.4).

## 1.4. Redes bayesianas

Siguiendo con la discusión planteada anteriormente para el caso de la regresión logística, en la Tabla 1.1 se puede observar que el número de publicaciones registradas en el ISI durante los últimos diez años en relación a las redes bayesianas es sustancialmente menor que el número de publicaciones cuyo tópico fue la regresión logística. No obstante, hay que hacer notar que el patrón de evolución proporcional es contrario en el caso de las redes bayesianas (Figura 1.5). Mientras que la regresión logística es un procedimiento que ha sido más tratado en el campo de las ciencias sociales, las redes bayesianas no lo han sido tanto en comparación con la investigación general.

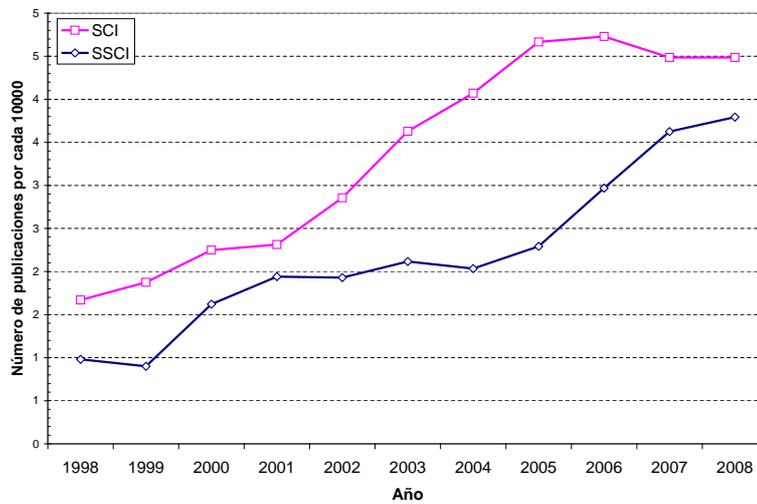
El nacimiento de las redes bayesianas tuvo lugar en la Universidad de Stan-

<sup>2</sup>Datos actualizados a 8 de diciembre de 2008.

**Tabla 1.1:** Número de publicaciones cuyo tópico fueron las redes bayesianas (RB) o la regresión logística (RL) desde 1998 a 2008.

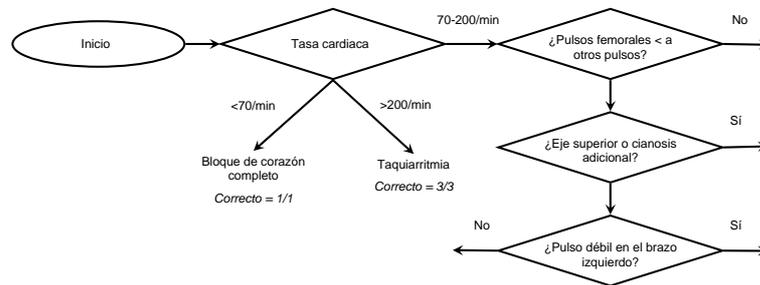
BASE DE DATOS	AÑO	RL	RB	TAP
SCI	1998	5.013	323	1.934.454
	1999	5.409	362	1.931.240
	2000	6.160	440	1.956.809
	2001	6.785	457	1.975.308
	2002	7.585	597	2.090.285
	2003	8.470	794	2.187.826
	2004	9.907	971	2.385.027
	2005	11.512	1183	2.535.492
	2006	12.326	1178	2.490.800
	2007	13.924	1169	2.606.697
2008	13.923	1169	2.606.682	
SSCI	1998	1.497	28	285.577
	1999	1.708	26	288.948
	2000	1.914	48	296.155
	2001	1.981	59	303.909
	2002	2.217	59	305.873
	2003	2.361	64	302.414
	2004	2.740	65	319.431
	2005	3.435	81	353.253
	2006	3.656	104	350.233
	2007	4.089	133	366.869
2008	4.089	133	350.569	

Fuente: *Science Citation Index* (SCI) y *Social Sciences Citation Index* (SSCI) de la *ISI Web of Knowledge*<sup>SM</sup>. TAP: total de artículos publicado en cada base de datos por año.



**Figura 1.5:** Evolución del número de publicaciones en el *ISI Web of Knowledge*<sup>SM</sup> donde el tópico fueron las redes bayesianas.

ford durante el periodo temporal comprendido entre 1970 y finales de la década de los 80. El desarrollo estuvo condicionado por dos razones: en primer lugar, se produjo una profunda crisis en la investigación relativa a las Redes Neuronales Artificiales (RNA) tras el famoso informe (*Perceptrons*) de Marvin Minsky y Seymour Papert (Quinlan, 1991; SPSS y Recognition Systems, 1997); y en segundo lugar, porque surgió gran interés por la toma de decisiones y la predicción en condiciones de incertidumbre (Cowell et al., 1999; Oatley y Ewart, 2003). Como se comentó anteriormente, las redes bayesianas se basan en la misma filosofía de red sobre la que se construyeron las RNA. La filosofía de red, surgida en el contexto del conexionismo y el asociacionismo se inicia en los años 50, cuando el neurobiólogo Warren McCulloch y el estadístico Walter Pitts desarrollaron la primera neurona artificial (Quinlan, 1991). Esta iniciativa fue evolucionando hasta que, coincidiendo con el auge conexionista en Psicología (Hothersall, 1995/1997), se forjó el paradigma de las redes neuronales artificiales (RNA). Desde esta perspectiva se trató de desarrollar sistemas inteligentes que modelizasen los procesos de cómputo que tienen lugar en el sistema neural humano. La base principal de las RNA es el *principio de Hebb*, *principio de estabilización selectiva* o *principio de convergencia sincrónica* (Carlson, 1993/2000), que viene a ser una generalización del condicionamiento clásico (Pavlov, 1927) aplicado a la estructura química y anatómica de las sinapsis.



**Figura 1.6:** Árbol de decisión diagnóstica usado en el GOSH. Adaptado de Cowell et al. (1999).

Los antecedentes más directos de las redes bayesianas son los árboles de decisión diagnóstica (también conocidos como árboles de clasificación, diagramas de flujo, flujogramas o algoritmos). Un árbol de decisión diagnóstica es un conjunto de preguntas o frases que se relacionan secuencialmente, de tal modo que la respuesta a una pregunta (o la culminación de una fase) determina la siguiente cuestión a resolver. Este tipo de sistemas orientado a la toma de decisiones, aunque no necesitan implementarse en computadoras ya a que son sencillos de utilizar y permiten rectificar ante una decisión tomada erróneamente, tienen una serie de inconvenientes que los hace problemáticos. Por ejemplo, son sistemas rígidos que no toleran la presencia de datos perdidos. En la Figura 1.6 se representa un árbol de decisión diagnóstica que fue diseñado en el *Great Ormond Street Hospital* (GOSH) para la detección de problemas cardiovasculares congénitos en bebés.

Como se puede ver en la ilustración, lo primero que se hace cuando nace el niño es medir la tasa cardiaca. En este punto de decisión hay tres posibles caminos a tomar. Supongamos que las pulsaciones son mayores a 200, en este caso se diagnostica taquiarritmia (pérdida del ritmo cardíaco por encima de las pulsaciones normales). La inscripción *Correcto = 3/3* indica que en la base de datos disponible 3 de los casos diagnosticados de este modo, según estas condiciones, fueron clasificados correctamente. Si, por el contrario, las pulsaciones se encuentran entre 70 y 200 se pasa al siguiente punto de decisión y así sucesivamente.

Posteriormente aparecieron los sistemas de producción, que también se les ha denominado como sistemas-basados en reglas. Un sistema de este tipo se vale de las reglas de la lógica para dar lugar a conclusiones válidas. Para ello se utilizan las típicas conjunciones lógicas como «SI», «ENTONCES», «O», «NO» o «Y».

- **SI** hay dolor de cabeza **Y** fiebre **ENTONCES** hay gripe (factor de certidumbre 0.7)
- **SI** hay gripe **ENTONCES** se estornuda (factor de certidumbre 0.9)
- **SI** hay gripe **ENTONCES** hay debilidad (factor de certidumbre 0.9)

**Figura 1.7:** Ejemplo del factor de certidumbre. Adaptado de Cowell et. al. (1999).

Una muestra de regla de producción sería la siguiente declaración: - SI un animal tiene plumas ENTONCES es un pájaro. Por lo tanto, aseveraciones de este tipo se irían combinando de un modo eficiente para dar lugar al razonamiento simbólico y obtener así conclusiones verdaderas. Sin embargo, este tipo de mecanismos también se enfrentan a una serie de debilidades. Por ejemplo, tienden a generar afirmaciones o declaraciones de verdad en vez de solucionar o responder cuestiones. Además, algunas cadenas de razonamiento complejas serían muy difíciles de interpretar.

Una técnica que se utilizó para superar las deficiencias observadas en los sistemas de producción fue el adjuntar un factor de certidumbre (FC) a las declaraciones o premisas del sistema (Figura 1.7). Este avance y la consideración de las relaciones de independencia condicional entre variables podrían considerarse como unos de los elementos más importantes en el desarrollo de las redes bayesianas. En este sentido, como señalan Castillo, Gutiérrez, y Hadi (1998), las redes bayesianas pueden considerarse como una generalización de los sistemas expertos basados en reglas donde el motor de inferencia son reglas estadísticas.

Un ejemplo de este tipo de sistemas orientados al psicodiagnóstico es DAI (Adarraga y Zaccagnini, 1992), un programa informático basado en conocimiento que trabaja con grados de certidumbre y que está orientado al diagnóstico de diferentes tipos de autismo infantil.

Modelado predictivo

# Capítulo 2

## Fundamentos de la regresión logística

*"[...] No one may have remarked on my good housekeeping before, but everyone noticed how careless I was now [...]."*

(Chevalier, 1999)

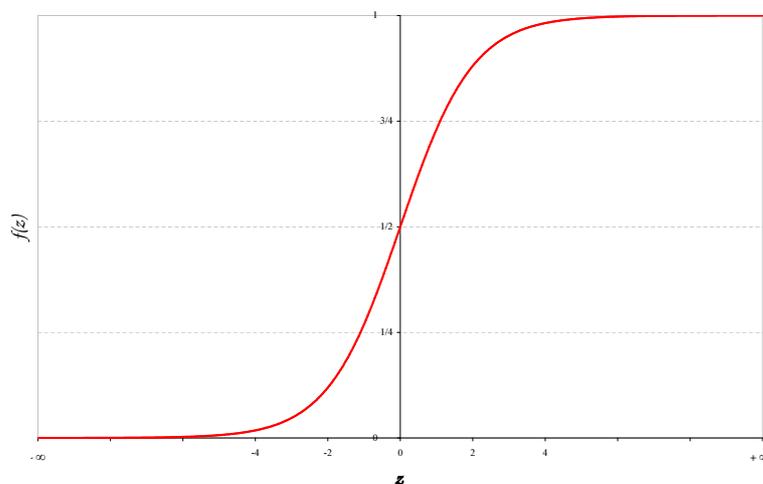
### 2.1. Componentes del modelo

#### 2.1.1. Función logística

Para explicar el modelo estadístico que subyace en la regresión logística hay que hacer referencia a la función que le da nombre: la *función logística*. La función logística viene dada por la ecuación

$$f(z) = \frac{1}{1 + e^{-z}} \quad (2.1)$$

donde  $e \approx 2,718$ , se refiere a la *constante de Neper* (en honor al matemático escocés John Napier) o *número de Euler* (en honor al matemático suizo Leonhard Paul Euler) que es la base de los logaritmos naturales o neperianos. Como se puede apreciar en la Figura 2.1 la función logística está comprendida entre 0 y 1



**Figura 2.1:** Función logística.

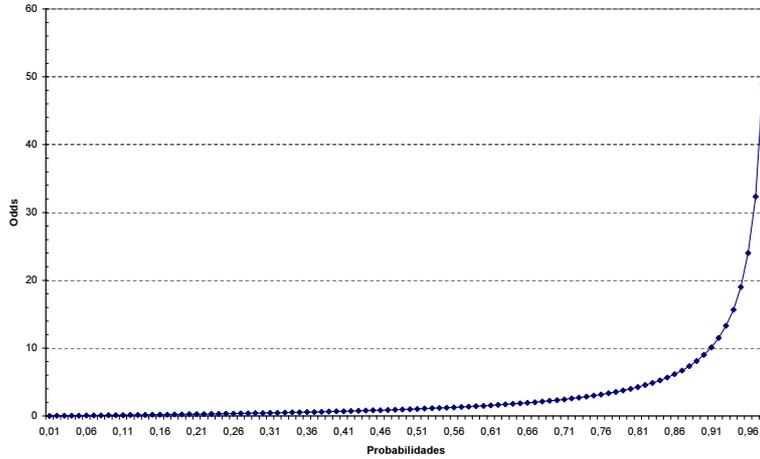
independientemente del valor que tome  $z$ ; luego es igual a 0 cuando  $z$  vale  $-\infty$  y a 1 cuando vale  $+\infty$ .

Como señala Kleinbaum (1994), hay dos aspectos que hacen que la función logística se haya popularizado, principalmente, dentro del campo de las ciencias de la salud: 1) el hecho de que sus valores estén comprendidos dentro del rango  $0 \leq f(z) \leq 1$ , lo cual vendría a paliar el problema de las estimaciones sin sentido que hacía el modelo lineal; y 2) la forma funcional, que solucionaría el problema planteado por Long (1997) ya que el valor de la función tiende a aproximarse a sus límites superior e inferior asintóticamente.

### 2.1.2. Riesgo y ventaja relativa

El modelo de regresión logística se basa en el concepto de *odds* o *ventajas*. Este término es muy común en el mundo anglosajón en el contexto de las apuestas (p. e., Jovel, 1995; Silva y Barroso, 2004; Yarandi y Simpson, 1991), aunque no es más que una forma diferente de expresar la probabilidad. Técnicamente, si consideramos un suceso  $S$  y denotamos la probabilidad de que ocurra ese suceso por  $p(S)$ , entonces la *odds* ( $O(S)$ ) correspondiente a ese suceso se obtendría al resolver la ecuación:

$$O(S) = \frac{p(S)}{1 - p(S)} . \quad (2.2)$$



**Figura 2.2:** Transformación del espacio probabilístico al de los *odds*.

Por ejemplo, supongamos que la probabilidad de que una persona sana desarrolle un trastorno psicótico es igual a  $\frac{1}{200}$ ; esto es, que existe una probabilidad de  $5 \times 10^{-3}$  de que una persona desarrolle el trastorno en condiciones normales. En este caso, la *odds* de que una persona sana desarrolle un episodio psicótico es de  $\frac{0,005}{1-0,005}$  que es igual a  $5,03 \times 10^{-3}$ . O lo que es lo mismo, que existe una *odds*, *ventaja* o *posibilidad* de 1 : 199<sup>1</sup> de que una persona sana desarrolle un trastorno psicótico. En la Figura 2.2 se puede observar como van cambiando las *odds* a medida que varía la probabilidad.

Un concepto relacionado con las *odds* es la idea de *riesgo relativo* (*RR*). El riesgo relativo expresa el grado en que es más probable que se produzca un suceso *S* dada una condición concreta (por ejemplo *A*) frente a cuando se está en otra condición (por ejemplo *B*). Expresado en términos de probabilidad condicional, el riesgo relativo se podría representar con la ecuación

$$RR = \frac{p(S|A)}{p(S|B)}, \quad (2.3)$$

esto es, se indica cuanto más probable es que se produzca el suceso *S* cuando se da la situación *A* que cuando se da la situación *B*. Siguiendo con el ejemplo anterior, si sabemos que la probabilidad de que se produzca un trastorno psicótico cuando se ha consumido cierta droga es de  $\frac{3}{200}$ , tendremos que el riesgo relativo

<sup>1</sup>Esta nomenclatura se suele utilizar en el mundo anglosajón para expresar apuestas.

de padecer un trastorno psicótico bajo los efectos de la droga es de  $\frac{0,015}{0,005} = 3$ . O lo que es lo mismo, existe el triple de riesgo de desarrollar un trastorno psicótico cuando se toma la droga frente a cuando no se toma.

También podríamos expresar la relación entre el consumo de la droga y la probabilidad de desarrollar trastorno psicótico en términos de *odds*. En este caso, usando la nomenclatura de la probabilidad condicional, tendríamos que la *odds ratio* (*OR*)<sup>2</sup> de que se produzca el suceso o efecto *S* en personas expuestas a un tratamiento *T* (como puede ser el consumo de la droga en nuestro ejemplo) en relación a las personas que no han recibido el tratamiento ( $\bar{T}$ ) vendría expresado por la siguiente igualdad

$$OR = \frac{p(S|T)}{1 - p(S|T)} : \frac{p(S|\bar{T})}{1 - p(S|\bar{T})} . \quad (2.4)$$

En nuestro ejemplo la *odds* de padecer un trastorno psicótico en condiciones normales es de 1 : 199 mientras que la *odds* de padecerlo al estar bajo los efectos de la droga es de 3 : 197; luego la *oportunidad relativa* de padecer un trastorno psicótico asociado al consumo de la droga sería de  $\frac{3/197}{1/199} = 3,03$ . Esto significa que es 3,03 veces más probable desarrollar un trastorno psicótico cuando se toma la droga que cuando no se toma.

### 2.1.3. Modelo estadístico

El análisis de regresión logística generaliza el modelo de regresión lineal tomando como variable dependiente el logaritmo neperiano de la *odds*, conocido como *logit*, correspondiente a la variable dicotómica. Así pues, si la estimación lineal vista en la ecuación 1.2 viene definida por el modelo

$$y = \alpha + \beta x , \quad (2.5)$$

---

<sup>2</sup>Nosotros, aunque el término *odds ratio* es el más usado en la bibliografía en castellano sobre el tema, somos afines a la traducción hecha por Barón y Téllez (2005) cuando lo llaman *ventaja relativa* u *oportunidad relativa*. En cualquier caso, el término *odds ratio* ha sido traducido de muchas maneras (p. e., razón de productos cruzados, razón de desigualdades, desigualdad relativa, razón de la diferencia, razón de ventaja, razón relativa, razón de posibilidades o razón de momios) y no está exento de polémica (Bautista, 1995; Irala et al., 1997; Tapia, 1995).

## 2.1 Componentes del modelo

donde  $y$  representa a la variable dependiente,  $\alpha$  se refiere a la constante poblacional del punto de corte de la recta con el eje  $y$  u origen,  $\beta$  denota la pendiente poblacional de la recta y  $x$  representa la variable independiente; en el modelo de regresión logística simple la ecuación sería

$$y = \ln \left( \frac{p(y = 1)}{1 - p(y = 1)} \right) = \alpha + \beta x . \quad (2.6)$$

Lo que significa que la técnica de regresión logística modela linealmente el cambio que se produce en la oportunidad relativa de la variable dependiente como resultado de la influencia de la variable independiente. O lo que es lo mismo, generaliza el modelo lineal para el caso de una variable dependiente dicotómica usando una función de enlace *logit* (Ato y López, 1996).

La expresión de la ecuación 2.6 se podría enunciar también del siguiente modo,

$$\frac{p(y = 1)}{1 - p(y = 1)} = e^{\alpha + \beta x} . \quad (2.7)$$

Por su parte, a partir de la ecuación 2.7 podemos despejar el valor de la probabilidad realizando los siguientes pasos

$$p(y = 1) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}} , \quad (2.8)$$

donde,

$$p(y = 1) = \frac{1}{1 + e^{-\alpha - \beta x}} , \quad (2.9)$$

que es, como se mostró en la ecuación 2.1, la función que define la función logística.

En el caso multivariado, podemos generalizar la ecuación 2.6 y tendríamos que

$$y = \ln \left( \frac{p(y = 1)}{1 - p(y = 1)} \right) = \alpha + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \dots + \beta_K x_{iK} , \quad (2.10)$$

lo que equivale a

$$y = \frac{p(y = 1)}{1 - p(y = 1)} = e^{\alpha + \sum_{k=1}^K \beta_k x_{ik}} ,$$

donde

$$p(y = 1) = \frac{e^{\alpha + \sum_{k=1}^K \beta_k x_{ik}}}{1 + e^{\alpha + \sum_{k=1}^K \beta_k x_{ik}}} ,$$

o lo que es lo mismo expresado en los términos de la función logística

$$p(y = 1) = \frac{1}{1 + e^{-\alpha - \sum_{k=1}^K \beta_k x_{ik}}} ,$$

donde las  $x$  son las variables independientes. Por su parte, cada  $\beta_k \in \beta_K$  son los parámetros que indican el efecto de una variable independiente  $x$  sobre el *logit* de  $y$ . Por último,  $\alpha$  es el equivalente al intercepto de la regresión lineal. Este parámetro hace que la función logística se desplace (en una gráfica del tipo 2.1) a la izquierda a medida que aumenta su valor y hacia la derecha a medida que disminuye.

#### 2.1.4. Estimación del modelo

En el análisis de regresión logística la estimación de los parámetros se realiza por el procedimiento de máxima verosimilitud (MV), frente al análisis de regresión lineal donde lo más común es utilizar el procedimiento de *Minimos Cuadrados Ordinarios* (MCO) <sup>3</sup>. El procedimiento de máxima verosimilitud se introdujo con posterioridad al método MCO dado que, entre otras cosas, requería un mayor volumen de recursos de cómputo para ser ejecutado eficientemente por los ordenadores (Kleinbaum, 1994).

Mientras que en el procedimiento MCO el vector  $\beta$  de los parámetros del modelo es aquel  $\hat{\beta}$  que reduce la sumatoria de los residuos cuadráticos

$$\sum_{i=1}^N (y_i - \mathbf{x}_i \hat{\beta})^2 ,$$

---

<sup>3</sup>Si asumimos que los errores del modelo están normalmente distribuidos, también podemos utilizar el procedimiento de máxima verosimilitud para el modelo de regresión lineal como sucede en los Modelos de Ecuaciones Estructurales (Batista y Coenders, 2000).

## 2.1 Componentes del modelo

el modelo de máxima verosimilitud estima el conjunto de valores (parámetros) que maximizan la probabilidad de observar los datos muestrales.

Consideremos, por ejemplo, el problema de estimar la probabilidad de obtener un cierto número de personas con el rasgo  $\phi$  en una muestra dada. Si sabemos que la prevalencia del rasgo  $\phi$  en la población es del 50% podríamos utilizar la distribución binomial para estimar la probabilidad de que  $s$  personas en una muestra de tamaño  $N$  manifiesten el rasgo  $\phi$ , donde el parámetro poblacional  $\pi$  ( $p(\delta) = 0,5$ ) representa la proporción de personas con el rasgo (Dixon, 1964/1970; Uña, Tomeo, y San Martín, 2003):

$$p(s|\pi, N) = \frac{N!}{s!(N-s)!} \pi^s (1-\pi)^{N-s}, \quad (2.11)$$

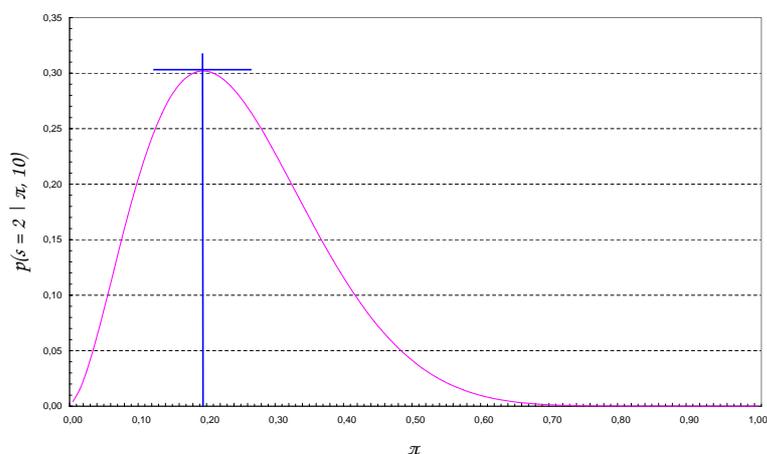
donde  $k! = k \times (k-1) \dots 2 \times 1$  representa el factorial de un número. Por ejemplo, la probabilidad de encontrar 2 personas con el rasgo  $\phi$  que se distribuye poblacionalmente con  $\pi = 0,5$  en una muestra de 10 personas es:

$$p(s=2|\pi=0,5, N=10) = \frac{10!}{2! \times 8!} \times 0,5^2 \times (1-0,5)^8 = 0,044.$$

Este tipo de problema, donde conocemos la fórmula de la distribución de probabilidades, los valores de los parámetros  $\pi$  y  $N$  y queremos conocer la probabilidad de ocurrencia asociada a un caso particular  $s$ , es típico en la estimación de probabilidades. Lo que hace el procedimiento de máxima verosimilitud en la regresión logística es, tomando el caso de la distribución binomial como una analogía sensiblemente más sencilla, es estimar  $\pi$  cuando se conocen los parámetros  $s$  y  $N$ . Así pues, la *estimación máximo-verosímil es aquel valor del parámetro, o conjunto de valores para los parámetros, que hace que los valores observados sean más probables*.

Supongamos que conocemos  $s=3$  y  $N=10$  pero que no conocemos el valor del parámetro  $\pi$ . ¿Cuál sería el valor de  $\pi$  que hubiese generado  $s=3$  con mayor probabilidad? Dado que la ecuación 2.11 computa la probabilidad de ocurrencia  $s$  eventos como función de los parámetros  $\pi$  y  $N$  se le denomina *función de probabilidad*, ya que  $\pi$  y  $N$  son constantes mientras que  $s$  varía. Cuando utilizamos esta misma ecuación y la condicionamos al valor que toma  $\pi$  la llamamos *función*

## Fundamentos de la regresión logística



**Figura 2.3:** Función de Máxima Verosimilitud. Probabilidad de que  $s = 3$  para diferentes valores de  $\pi$ .

de verosimilitud, donde los valores de  $N$  y  $s$  se mantienen constantes mientras que  $\pi$  varía. En el caso de nuestro ejemplo la función de verosimilitud sería

$$p(\pi | s = 2, N = 10) = \frac{10!}{2! \times 8!} \times \pi^2 \times (1 - \pi)^8 .$$

Así, el valor estimado máximo-verosímil del parámetro  $\pi$  será aquél valor  $\hat{\pi}$  que maximice la verosimilitud de observar el conjunto de datos que fue realmente observado. En este ejemplo concreto, y como se puede observar en la Figura 2.3, esto sucede en el punto de la tangente de la función que corresponde con el valor más alto de la curva. Luego la estimación máximo-verosímil para nuestro ejemplo es  $\hat{\pi} \approx 0,19$ .

Existen dos tipos, genéricamente, de estimaciones de máxima-verosimilitud para el análisis de regresión logística: la condicionada y la no condicionada. En la estimación *MV no condicionada* se tiene en cuenta la probabilidad conjunta de los datos del estudio entendida como el producto de la probabilidad conjunta para los casos positivos (casos que muestran una respuesta positiva en la variable de respuesta) y la probabilidad conjunta para los casos negativos. Por su parte, la función de verosimilitud para la estimación *MV condicionada* representa la probabilidad de los datos muestrales dadas todas las posibles configuraciones del conjunto de variables.

Kleinbaum (1994) recomienda utilizar la versión no condicionada cuando el número de parámetros es pequeño en relación al número de observaciones, mien-

tras que recomienda usar el modelo condicionado cuando es grande en relación al número de casos. En general la estimación condicional consume más recursos computacionales pero tiende a generar parámetros más ajustados.

Silva y Barroso (2004) describen el procedimiento de estimación de máxima verosimilitud no condicionada a través de los mínimos cuadrados del siguiente modo. Consideremos que  $V$  representa la verosimilitud del modelo, que viene definida por

$$V = \hat{p}_1 \times \hat{p}_2 \times \cdots \times \hat{p}_d \times (1 - \hat{p}_{d+1}) \times (1 - \hat{p}_{d+2}) \times \cdots \times (1 - \hat{p}_n), \quad (2.12)$$

donde  $\hat{p}_d$  se refiere a la probabilidad de respuesta positiva predicha por el modelo para el sujeto  $d$  que presenta el rasgo y  $(1 - \hat{p}_n)$  representa la probabilidad de respuesta positiva estimada por el modelo para una persona  $n$  que no presenta el rasgo. En la medida que la verosimilitud del modelo se acerca a 1 el modelo se entiende más apropiado para modelar la realidad. La estimación máximo verosímil intenta, por tanto, conseguir el conjunto de coeficientes que maximicen esta función. El proceso para hallar los parámetros que maximicen la función de verosimilitud es complicado y costoso computacionalmente; por ejemplo, uno de los métodos iterativos para encontrar las estimaciones máximo verosímiles implica el uso del algoritmo Newton-Ralphson.

### 2.1.5. Interpretación de coeficientes

Cuando realizamos un análisis de regresión logística sobre una muestra de datos obtenida a partir de una población obtenemos una ecuación muestral  $\ln\left(\frac{p}{1-p}\right) = a + b_1x_1$  donde  $a$  es una estimación del parámetro poblacional  $\alpha$  y  $b_1$  es una estimación de  $\beta$  en la ecuación 2.5. En general, la interpretación más común de los resultados pasa por analizar los coeficientes del modelo en términos de ventajas relativas. Por ejemplo, siguiendo con el ejemplo de la relación entre la tasa de asistencia a clase y la manifestación de conductas agresivas visto anteriormente, si aplicamos a los datos un análisis de regresión logística obtendríamos el siguiente resultado

$$\ln \left( \frac{p(PCA = 1)}{1 - p(PCA = 1)} \right) = -18,68 + 0,40 \times PFC . \quad (2.13)$$

En este caso, el coeficiente  $b_1$  de 0,40 indica que por cada punto porcentual que aumente el porcentaje de faltas a clase el logaritmo de la *odds* de mostrar conductas agresivas aumenta en 0,40. Cuando el logaritmo de la *odds* aumente también lo hará la probabilidad. O dicho de otro modo, por cada aumento de un punto porcentual en las faltas a clase, la *odds* predicha de mostrar conductas agresivas se incrementa en  $e^{0,40} = 1,49$ . El modelo anterior nos permite estimar la probabilidad de que una persona muestre conductas agresivas cuando conocemos tasa de asistencia a clases. Por ejemplo, supongamos que queremos conocer la probabilidad que tiene una persona de mostrar conductas violentas dado que falta al 49% de las clases. Como

$$\ln \left( \frac{p}{1 - p} \right) = -18,68 + 0,40 \times 49 = 0,92 ,$$

tenemos que

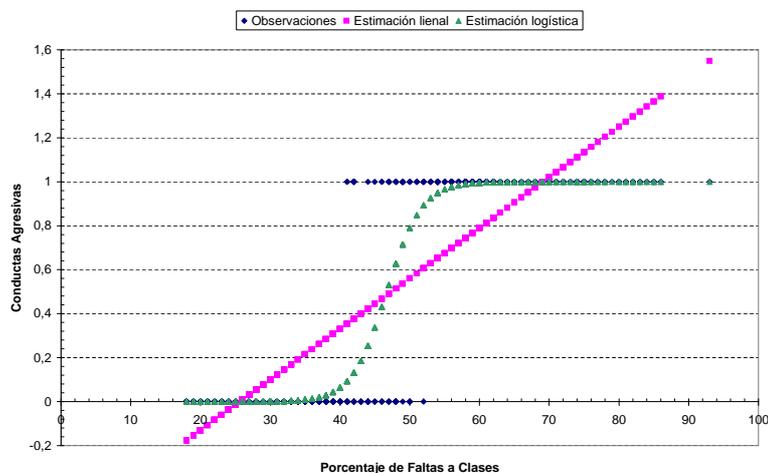
$$\frac{p}{1 - p} = e^{0,92} = 2,51 ,$$

con lo que

$$p = \frac{2,51}{1 + 2,51} = 0,72 .$$

Esto quiere decir que la probabilidad de que un adolescente que falta al 49% de las clases muestre conductas agresivas es de 0,72. En la Figura 2.4 se puede observar la representación gráfica de la relación que se establece entre la tasa de absentismo en las clases y la probabilidad de mostrar conductas agresivas basándonos en el modelo logístico representado por la ecuación 2.13. Como se puede observar en la gráfica el trazo definido por el modelo logístico no cae fuera de los límites (0 y 1) establecidos por los supuestos de la teoría de probabilidades. De esta manera, el

## 2.1 Componentes del modelo



**Figura 2.4:** Datos observados, estimación lineal y estimación logística de la relación que se establece entre la tasa de absentismo y la probabilidad de mostrar conductas agresivas.

modelo de regresión logística soluciona el problema que tenía la regresión lineal cuando sus estimaciones *escapaban* del intervalo permisible en el caso de trabajar con variables dicotómicas.

No obstante, en la práctica es más útil interpretar los coeficientes del análisis de regresión logística en términos de oportunidades relativas ( $e^{b_n}$ ). En términos epidemiológicos una oportunidad relativa **igual a 1** indica que la variable independiente a la que se asocia no tiene relación con la variable de respuesta. Esto es, se interpretaría como que la *odds* u oportunidad para las personas expuestas al tratamiento es la misma que la de las personas no expuestas. Cuando obtenemos coeficientes que dan lugar a oportunidades relativas **mayores que 1** se interpreta que la variable independiente asociada a ese coeficiente es un factor de riesgo para la aparición de un valor positivo en la variable dependiente. Por último, cuando una variable independiente está asociada a una oportunidad relativa **inferior a 1** podremos interpretar que la variable asociada al parámetro reduce la aparición de un resultado positivo en la variable dependiente (por ejemplo, cuando un medicamento reduce el riesgo de desarrollar cierta enfermedad). Cuando se usan intervalos de confianza para evaluar la influencia de las variables independientes del modelo sobre la variable dependiente hay que tener en cuenta que si el intervalo del exponencial del parámetro no incluye al valor 1, entonces podemos concluir que existe relación estadísticamente significativa entre las va-

riables (Sofroniou y Hutcheson, 2002). En caso contrario, podemos concluir que la relación no es estadísticamente significativa.

Por ejemplo, consideremos el siguiente modelo de regresión logística donde se ha estimado la probabilidad de padecer trastorno cardiovascular (TC) en función de si la persona fuma y hace deporte

$$\ln \left( \frac{p(TC)}{1 - p(TC)} \right) = 1,01 + 2,72 \times Alcohol - 5,93 \times Deporte , \quad (2.14)$$

donde *Alcohol* y *Deporte* son dos variables dicotómicas que valen 1 cuando tienen valor positivo (esto es, cuando se toma alcohol o se hace deporte) y 0 en caso contrario. Como se puede ver en la ecuación 2.14 el coeficiente asociado al consumo de alcohol es positivo mientras que el de la práctica deportiva es negativo lo que significa que el consumo de alcohol es un factor de riesgo para sufrir trastornos cardiovasculares mientras que el deporte es una actividad que protege frente a ellos. Ya que  $e^{2,72} = 15,18$  podríamos decir que el beber alcohol aumenta en más de 15 veces el riesgo de padecer trastornos cardiovasculares. Es más, si tomamos como referencia el intervalo de confianza para el exponencial del parámetro al 95 % de confianza podríamos decir que el riesgo de padecer trastornos cardiovasculares si se consume alcohol aumenta entre 6 y 40 veces. Por su parte, la práctica de deporte reduce las posibilidades de padecer trastorno cardiovascular en más de 300 veces ya que  $e^{-5,93} = 0,0027$ .

## 2.2. Bondad de ajuste

Existen varios procedimientos que se utilizan para determinar la bondad de ajuste del modelo de regresión logística; esto es, análisis que comprueban como los datos observados son explicados a partir del modelo estadístico (Hosmer, Taber, y Lemeshow, 1991). En este texto, por claridad expositiva, se han agrupado los procedimientos de evaluación de la bondad de ajuste en tres bloques: 1) ajuste predictivo, relacionado con la habilidad que tiene el modelo para predecir (o clasificar) el valor de la variable dependiente en función de los predictores; 2) ajuste general, que indica de modo generalizado la validez del modelo; y 3) ajuste individual, que indica el ajuste individualizado de los parámetros del modelo.

**Tabla 2.1:** Matriz de confusión o tabla de clasificación.

		<b>Predicho</b>	
		NO	SI
<b>Observado</b>	NO	$a$	$b$
	SI	$c$	$d$

Esta tesis se centra en el estudio del primer tipo de evidencias de ajuste. Así pues, en primer lugar se describirán los procedimientos diseñados para valorar el ajuste predictivo en la regresión logística<sup>4</sup> y a continuación se comentarán brevemente el resto de procedimientos que existen para testar la bondad de ajuste del modelo.

### 2.2.1. Ajuste predictivo

#### Tabla de clasificación o matriz de confusión

En las tablas de clasificación se comparan los resultados predichos por el modelo frente a los datos observados. Esto es, una vez creado el modelo de regresión logística se va tomando cada caso del conjunto de observaciones y se va contrastando la predicción que hace el modelo frente al verdadero estado del caso en relación a la variable dependiente. El resultado de este proceso se refleja en lo que se conoce como *tabla de clasificación o matriz de confusión* (Tabla 2.1), donde aparece el número de casos en que el modelo ha llevado a cabo clasificaciones correctas e incorrectas.

En la Tabla 2.1 aparecen los posibles valores que puede tomar la variable dependiente en un análisis de regresión logística dicotómica en términos de Sí (1) o No (0). Así mismo, se hace una diferenciación entre lo que se ha observado en el conjunto de datos (Observado) y lo que ha predicho el modelo. Para que un modelo de regresión logística esté bien ajustado desde un punto de vista predictivo los valores de la diagonal ( $a$  y  $d$ ) han de ser mayores que el resto de valores de la tabla, ya que estos valores representan la coincidencia entre lo observado y lo predicho por el modelo. Lo ideal sería encontrar ceros en las casillas  $b$  y  $d$  mientras

<sup>4</sup>Dado que en las redes bayesianas también pueden testarse con este tipo de procedimientos, las explicaciones introducidas aquí pueden trasvasarse directamente al estudio de la bondad de ajuste de este tipo de herramientas.

que la diagonal contuviese valores diferentes de cero. Cada una de estas casillas se denomina en función de la fracción de personas a las que se refiere. Así, la casilla  $d$  representa lo que se denomina como *Fracción de Positivos Correctos*, la  $a$  es la *Fracción de Negativos Correctos*, la  $b$  sería la *Fracción de Falsos Positivos* y la  $c$  se refiere a la *Fracción de Falsos Negativos* (p. e., Hopley y van Schalkwyk, 2001; Johnson y Albert, 1999).

### Parámetros predictivos

A partir de la tabla de confusión se puede extraer información muy valiosa respecto a los modelos de regresión logística en relación a sus propiedades predictivas. En términos cuantitativos podemos obtener múltiples indicadores que nos informarían de lo bien o lo mal que nuestro modelo es capaz de clasificar casos dentro de las categorías que describe la variable dependiente. Entre los valores que podemos calcular destacan<sup>5</sup>:

- Sensibilidad del modelo:  $S = \frac{d}{c+d}$ ,
- Especificidad del modelo:  $E = \frac{a}{a+b}$ ,
- Proporción de falsos positivos:  $FPF = \frac{b}{a+b}$ ,
- Proporción de falsos negativos:  $PFN = \frac{c}{c+d}$ ,
- Valor predictivo positivo:  $VPP = \frac{d}{b+d}$ ,
- Valor predictivo negativo:  $VPN = \frac{a}{a+c}$ ,
- Tasa de clasificaciones correctas:  $TCC = \frac{a+d}{a+b+c+d}$ .

La *sensibilidad* ( $S$ ), tasa de verdaderos positivos o tasa de aciertos representa la proporción de casos positivos (cuando  $y = 1$ ) clasificados correctamente por el modelo respecto al total de casos positivos observados en los datos. Por su parte, la *especificidad* ( $E$ ) o tasa de verdaderos negativos representa la proporción de casos negativos (cuando  $y = 0$ ) correctamente clasificados como negativos por el modelo. La *Proporción de Falsos Positivos* ( $FPF$ ) o tasa de falsas alarmas

---

<sup>5</sup>Tomando de base la Tabla 2.1.

indica la proporción de casos clasificados incorrectamente como positivos, mientras que la *Proporción de Falsos Negativos (PFN)* representa el número de casos que han sido clasificados como negativos pese a que han sido observados como positivos. Por otro lado, el *Valor Predictivo Positivo (VPP)* mide la proporción de casos clasificados como positivos en relación al total de casos que el modelo clasificó como positivos, mientras que el *Valor Predictivo Negativo (VPN)* señala el porcentaje de casos negativos frente al total de casos clasificados como negativos por el modelo. Por último, el la *Tasa Global de Clasificaciones Correctas (TCC)* indica el número de casos clasificados correctamente en función del total de casos observados.

### Curvas ROC

La curva ROC<sup>6</sup> (*Receiver Operating Characteristic*) es una técnica que proporciona una forma refinada de representar parte de la información contenida en las matrices de confusión. En concreto, la curva ROC evalúa gráficamente la habilidad del modelo para discriminar entre casos positivos y negativos a medida que se hace variar el umbral de corte en la probabilidad predicha por el modelo para realizar la clasificación.

Las curvas ROC surgieron en el contexto de la optimización de señales de radar (Burgeño, García-Bastos, y González-Buitrago, 1995) pero rápidamente fueron trasladadas al campo de la investigación biomédica donde son utilizadas como herramientas de gran valor heurístico ante la toma de decisiones diagnóstica y terapéutica (Hanley y McNeil, 1982; I. López y Píta, 1998). Las curvas ROC fueron utilizadas en el campo de la psicofísica bajo la óptica de la Teoría de Detección de Señales para modelar los procesos de decisión que ponen en marcha los individuos ante la percepción de diferentes estímulos sensoriales (p. e., Fernández y Pinillos, 1992; Monserrat, 1998). A su vez, las curvas ROC se utilizan como herramienta para evaluar la bondad predictiva de modelos estadísticos (p. e., Baesens et al., 2004; Concejero, 2004; DeMaris, 2002; Johnson y Albert, 1999; Jovel, 1995; SPSS, 2006).

Técnicamente, las curvas ROC representan el cambio que se produce en los niveles de sensibilidad y en el opuesto de la proporción de falsos positivos (especi-

---

<sup>6</sup>Denominada también como curva de Características Operativas del Receptor (COR) en algunos textos (Franco y Vivo, 2007).

ficidad) a medida que se va variando los umbrales de corte para las clasificaciones de los casos del conjunto de datos. La sensibilidad y la especificidad en este contexto pueden interpretarse en términos de probabilidad condicional. Así, la sensibilidad ( $S$ ) sería:

$$S = p(\theta_E|E), \quad (2.15)$$

esto es, representa la probabilidad de clasificar como positivo, o enfermo en términos epidemiológicos, un caso seleccionado aleatoriamente del conjunto de datos ( $\theta_E$ ) cuando realmente presenta el rasgo ( $E$ ). En cuanto a la sensibilidad, representaría la probabilidad de clasificar como caso negativo un registro seleccionado aleatoriamente ( $\theta_{\bar{E}}$ ) que realmente no presenta el rasgo ( $\bar{E}$ ):

$$S = p(\theta_{\bar{E}}|\bar{E}). \quad (2.16)$$

La curva ROC queda comprendida en un cuadrante de  $1 \times 1$  y el área comprendida bajo ella es usada como un indicador de la bondad predictiva de la técnica o método diagnóstico (Hanley y McNeil, 1982). En términos probabilísticos, el área bajo la curva ( $ABC$ ) se define como la probabilidad de clasificar correctamente un par de casos (positivo y negativo) seleccionados aleatoriamente tal y como les corresponde. Esto es, el área bajo la curva ROC se define como:

$$ABC = p[p(\theta_E = E) > p(\theta_{\bar{E}} = E)]; \quad (2.17)$$

o lo que es lo mismo, indica la probabilidad de que el resultado de la prueba resulte más positivo cuando un caso seleccionado al azar posee el rasgo criterio. Así, a mayor área bajo la curva mayores cualidades predictivas tiene el modelo. A este valor diagnóstico de la ejecución predictiva también se le ha llamado *concordancia* ( $c$ ) o *componente de discriminación* (Harrel, Lee, Matchar, y Reichert, 1985). En la Figura 2.5 se puede observar que la curva de color rojo está por debajo de la de color azul lo que indica que el modelo estadístico o de diagnóstico que la ha generado es menos válido, en términos predictivos, que el modelo que ha

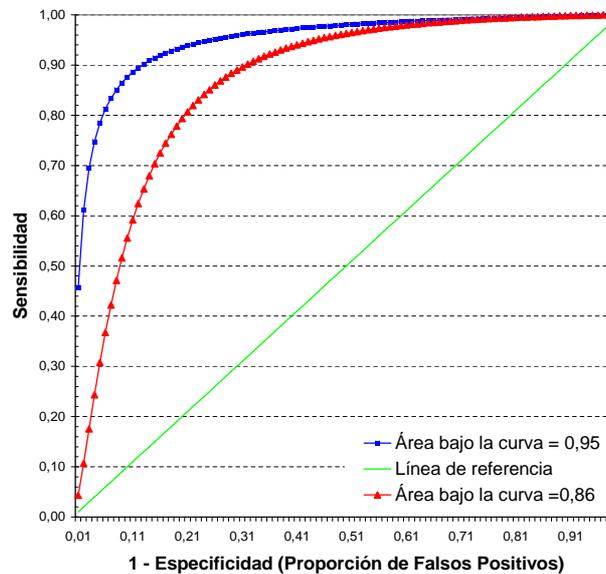


Figura 2.5: Representación de dos curvas ROC.

generado la curva azul (Hanley y McNeil, 1983). La línea verde que se ubica en la diagonal de gráfico representa el punto donde un modelo sería malo clasificando, o lo que es lo mismo, que clasifica aleatoriamente a los elementos del conjunto de datos. Así pues, una curva que apareciese sobre esta diagonal, o muy cerca de ella, representaría a un modelo pobremente predictivo.

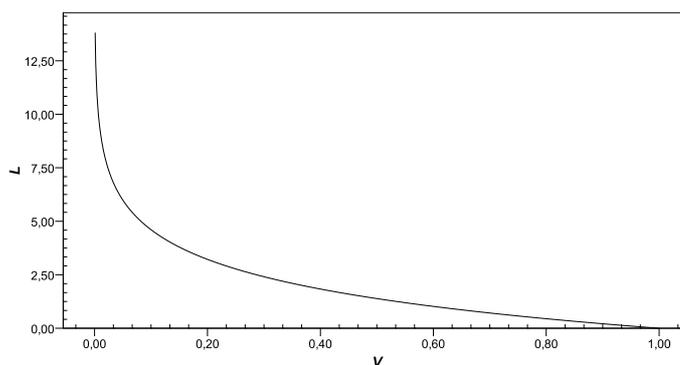
### 2.2.2. Índices de bondad de ajuste general

#### Verosimilitud del modelo

Como se vio en la sección 2.1.4 la estimación del modelo de regresión logística se basa en una función que maximiza la verosimilitud de las probabilidades predichas. Así pues, una forma de evaluar el grado en que el modelo estimado describe los datos observados se basa en  $V$ . La *lejanía* ( $L$ ) de un modelo de regresión logística se evalúa con la siguiente expresión:

$$L = -2 \ln V. \quad (2.18)$$

La lejanía (del inglés *deviance*) también se suele denominar  $-2LL$ ,  $-2\log$  de la verosimilitud o  $-2\text{LogLikelihood}$  en algunos textos y programas estadísticos.



**Figura 2.6:** Relación entre verosimilitud ( $V$ ) y lejanía ( $L$ ).

Dado que  $V < 1$ , su logaritmo siempre será negativo. Sin embargo, al multiplicarla por  $-2$  se obtiene un número positivo de magnitud doble, en valor absoluto, respecto a la verosimilitud. Para que un modelo sea apropiado, este debe de atribuir una alta probabilidad a los casos en que está presente el rasgo y una baja probabilidad en los que no está presente. Así pues, un modelo óptimo tendrá una verosimilitud muy cercana a 1 y, como consecuencia, una lejanía muy pequeña. Por su parte, un modelo pésimo tendrá una verosimilitud muy próxima a cero y una lejanía grande. En la Figura 2.6 se representa el cambio que experimenta la lejanía a medida que aumenta la verosimilitud del modelo.

Para comparar el aporte que proporciona la inclusión de una variable en un modelo de regresión logística se utiliza la *razón de verosimilitudes*. Esto es, lo que se pretende es comparar la verosimilitud de un modelo  $V$  que contiene  $k$  variables con otro modelo  $V^*$  que incluye sólo un subconjunto  $k^*$  de las variables originales. En este caso se utiliza el estadístico

$$RV = L^* - L = (-2 \ln V^*) - (-2 \ln V) = -2 \ln \left( \frac{V^*}{V} \right), \quad (2.19)$$

que sigue una distribución  $\chi^2$  con  $k - k^*$  grados de libertad y mide el aporte que suponen las  $k - k^*$  variables anidadas a la verosimilitud del modelo.

### Test de Hosmer-Lemeshow

El estadístico de Hosmer-Lemeshow es útil para evaluar el ajuste global del modelo, particularmente cuando se dispone de muchas variables independientes, o

cuando algunas de las variables independientes son continuas. Para calcular su valor hay que dividir la muestra en diez grupos compensados de acuerdo con sus probabilidades pronosticadas. A continuación se computa

$$G_{HL}^2 = \sum_{j=1}^{10} \frac{(O_j - E_j)^2}{E_j \left(\frac{1-E_j}{n_j}\right)} \sim \chi_8^2, \quad (2.20)$$

donde  $n_j$  se refiere al número de observaciones en el grupo  $j$ -ésimo,  $O_j = \sum_i y_{ij}$  se refiere al número de casos en el grupo  $j$ -ésimo y  $E_j = \sum_i \hat{p}_{ij}$  representa el número esperado de casos en el grupo  $j$ -ésimo.

La hipótesis nula que se testa en este contraste es que el modelo de regresión logística es bueno, mientras que la hipótesis alternativa indica que es malo. En este caso, la significación del estadístico se interpreta de manera análoga a como se interpretaría un coeficiente de determinación en regresión logística. Para que nos podamos fiar de este estadístico las frecuencias esperables de la tabla de contingencia generada con este procedimiento tienen que ser mayores que cinco, en caso contrario tenemos que dudar de la fiabilidad del estadístico.

### Varianza explicada

De manera análoga a la regresión lineal, en la regresión logística existe el coeficiente de determinación ( $R^2$ ) que indica la proporción de varianza de la variable dependiente que es explicada a partir de la variable/s independiente/s. Existen diferentes tipos de estadísticos  $R^2$  para la regresión logística pero su interpretación no es tan clara y directa como en el análisis de regresión lineal ya que su interpretación puede hacerse entendiendo que la variable dependiente representa los puntos de corte para una variable cuantitativa continua latente o una variable dicotómica observada (p. e. DeMaris, 2002; Long, 1997).

El primer, y más intuitivo, índice de determinación se calcularía estimando el cuadrado de la correlación entre la puntuación observada para la variable dependiente ( $y$ ) y las probabilidades estimadas ( $\hat{\pi}$ ) por el modelo:  $R^2 = r_{y,\hat{\pi}}^2$ . Por otro lado,  $R^2$  también puede entenderse como una reducción proporcional en la medida del error; esto es, la estimación de la proporción de error que se reduce para predecir  $y$  cuando usamos el modelo. Así,

$$R^2 = \frac{\sum(y - \bar{y})^2 - \sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}, \quad (2.21)$$

donde  $\sum(y - \bar{y})^2$  se referiría al error total y  $\sum(y - \hat{y})^2$  representaría el error cometido por el modelo. El *índice de la razón de verosimilitud* se basa en la misma idea pero usa la lejanía como índice de comparación (Ato y López, 1996; Hair, Anderson, Tatham, y Black, 1998):

$$R_L^2 = \frac{-2 \ln L_0 - (-2 \ln L_1)}{-2 \ln L_0}, \quad (2.22)$$

donde  $L_0$  es la lejanía del modelo que incluye sólo el parámetro constante mientras que  $L_1$  se refiere a la lejanía del modelo hipotetizado.

Por otro lado, la  $R^2$  generalizada tiene la ventaja de que es idéntica al modelo de regresión lineal cuando los errores se distribuyen normalmente. En el caso de la regresión logística se ha denominado como  $R^2$  de Cox y Snell y viene definida por la siguiente expresión

$$R_G^2 = 1 - \left( \frac{L_0}{L_1} \right)^{\frac{2}{N}}, \quad (2.23)$$

donde  $N$  se refiere al tamaño de la muestra. No obstante, la  $R^2$  generalizada tiene el problema de tomar un punto máximo inferior a 1 por lo que se ha propuesto la  $R^2$  *escalada generalizada* (también conocida como  $R$  de Nagelkerke) para solventar este problema (p. e., Alderete, 2006):

$$R_{EG}^2 = \frac{R_G^2}{1 - (L_0)^{\frac{2}{N}}}. \quad (2.24)$$

### Ajuste residual e influencia

Para terminar, hay que destacar otro conjunto de análisis utilizados para estudiar la bondad de ajuste de los modelos de regresión logística que se encuadran bajo lo que se ha denominado como parámetros de *diagnóstico de la regresión* (Hosmer

et al., 1991). Podríamos decir que existen dos formas de estudiar el ajuste en este sentido, por medio de los residuos y estudiando los casos con gran influencia en el modelo.

Los *residuos* de un modelo de regresión son, para cada caso, la diferencia entre el valor observado y el valor estimado para la variable dependiente según el modelo. El estudio de los residuos sirve para analizar el grado en que cada caso de los datos es predicho con más o menos error. Existen distintos tipos de residuos entre los que cabe destacar los residuos no tipificados, los tipificados, los estudentizados o las desviadas.

En cuanto al estudio de los puntos de influencia, el análisis se puede enfocar desde dos perspectivas diferentes: estudiando la influencia que tienen ciertas configuraciones de variables sobre la probabilidad estimada por el modelo o comprobando la influencia que tiene cada caso en la estimación de los parámetros del modelo. En cualquier caso, es útil presentar y estudiar los gráficos donde relacionen las probabilidades estimadas por el modelo y los estadísticos de influencia.

### 2.2.3. Ajuste de los coeficientes del modelo

La significación estadística de los coeficientes del modelo se basa en el estadístico de Wald que cuando hay un grado de libertad se calcula aplicando la fórmula

$$\text{Estadístico de Wald} = \left( \frac{b_n}{ET_{b_n}} \right)^2, \quad (2.25)$$

donde  $b_n$  se refiere al coeficiente de la regresión logística y  $ET_{b_n}$  se refiere al error típico de la distribución muestral del estadístico. Por ejemplo, siguiendo con el ejemplo presentado anteriormente (ecuación 2.14), el valor del estadístico Wald asociado al coeficiente de regresión para la variable *Alcohol* es de 33,27. Dado que el contraste de hipótesis que se hace con el estadístico de Wald es

$$H_0 : \beta_n = 0 \quad H_1 : \beta_n \neq 0 ,$$

y teniendo en cuenta que la significación de este contraste es inferior a 0,001, podemos concluir (con poca probabilidad de equivocarnos) que el parámetro asociado al consumo de alcohol es diferente de cero. O lo que es lo mismo, que

existe relación estadísticamente significativa entre el fumar y el padecer trastorno cardiovascular.

## 2.3. Limitaciones y avances en el análisis de regresión logística

Pese a que el análisis de regresión logística tiene una serie de ventajas frente a, por ejemplo, el análisis de regresión lineal (por ejemplo, no ha de cumplir supuestos distribucionales en los datos), existen algunos puntos en los que el análisis de regresión logística tiene problemas. Aunque se han identificado varios aspectos que influyen negativamente sobre el análisis de regresión logística (por ejemplo, el nivel de *separación* entre los valores de la variable dependiente, problemas numéricos derivados de la frecuencia observada en las combinaciones entre los niveles de la/s variable/s independiente/s y la dependiente, la colinealidad, el tamaño de muestra, el número de casos perdidos o la categorización de las variables) esta sección se dedicará a destacar las limitaciones que tiene el modelo logístico en el contexto de la investigación de esta tesis. A su vez, se perfilarán algunos de los desarrollos introducidos en esta técnica de análisis para solventar estos problemas.

### 2.3.1. Tamaño de muestra

El tamaño de la muestra es posiblemente uno de los aspectos más preocupantes en el diseño de estudios que implican un análisis de regresión logística (p. e., Bull, Mak, y Greenwood, 2002; Firth, 1993; Harrel et al., 1985; Jovel, 1995; Long, 1997; Ortega y Cayuela, 2002; Silva y Barroso, 2004). Ello se debe al propio procedimiento de estimación del modelo. Dado que el método de estimación del modelo de regresión logística es máximo-verosímil, a medida que aumenta el tamaño de la muestra las estimaciones son más estables y fiables.

La discusión relativa al efecto que tiene el tamaño de la muestra sobre los resultados del análisis de regresión logística ha girado en torno a la estimación de los parámetros del modelo así como en la determinación de las relaciones entre variables. En este sentido, se ha identificado tres tipos de errores que pueden producirse en la estimación de modelos logísticos (Irala et al., 1997; Ortega y Cayuela, 2002; Peduzzi, Concato, Kemper, Holford, y Feinstein, 1996): el *sobre-*

### 2.3 Limitaciones y avances en el análisis de regresión logística

*ajuste* o *error tipo I* se produce cuando se mantienen muchas variables que sólo aportan «ruido» en el modelo final; el *sub-ajuste* o *error tipo II* se da cuando no han sido incluidas en el modelo final variables relevantes; y el *ajuste paradójico* o *error tipo III* que aparece cuando un factor concreto ha sido asociado en la dirección opuesta al efecto real que tiene la variable independiente sobre la variable dependiente.

Lilienfeld y Pyne (1984) observaron que el tamaño de la muestra influía en la fiabilidad de las estimaciones de los parámetros, en la exactitud de la estimación de  $\beta$  y en la exactitud del contraste de hipótesis  $H_0 : \beta = 0$ . En concreto, encontraron que la desviación típica para las estimaciones de  $\beta$  en muestras simuladas se reducía a medida que aumentaba el tamaño de la muestra. La exactitud de esta estimación también se veía afectada a medida que se reducía el tamaño de la muestra, aunque en este caso la dirección del sesgo dependía de la distribución muestral de las observaciones (esto es, si los datos se distribuían exponencial, normal o binomialmente). En cuanto a la exactitud de contraste de hipótesis, también encontraron que a medida que aumentaba la muestra los resultados eran más válidos.

En lo que respecta al porcentaje de varianza de la variable dependiente explicada por el conjunto de variables independientes, DeMaris (2002) encontró que las muestras que contenían menos de 200 observaciones solían mostrar problemas numéricos en las estimaciones de los parámetros pero no encontró ningún patrón significativo en relación al coeficiente de determinación.

Con el objetivo de minimizar el impacto del tamaño de la muestra, Whittermore (1981) desarrolló un procedimiento para calcular tamaños muestrales óptimos en el análisis de regresión logística usando una estimación máximo-verosímil asintótica de la matriz de covarianza a partir de la matriz de Hess, la cual es válida como representante de la matriz de información de Fisher cuando la probabilidad de respuesta es baja. Con posterioridad, Hsieh (1989) publicó un conjunto de tablas de tamaños muestrales para regresión logística simple y múltiple basados en la propuesta de Whittermore. Mostraron que sus tablas son útiles para estudios epidemiológicos en los que la prevalencia de los factores es alta o baja y que, aunque las tablas no son del todo precisas para ciertos factores de riesgo, son razonablemente apropiadas para distribuciones en las que los factores de riesgo se distribuyen exponencial y normalmente.

Recientemente se han desarrollado métodos más sencillos orientados a estimar el tamaño muestral necesario para estudios que implican análisis de regresión teniendo en cuenta la tasa de prevalencia de las variables independientes, la potencia estadística así como su nivel de medida (Hsieh, Bloch, y Larsen, 1998). También se han diseñado programas estadísticos para automatizar estos cálculos (p. e., nQuery, SSIZE o EGRET SIZ).

La recomendación general que se hace respecto al tamaño de la muestra es que exista un ratio entre el número de observaciones y el número de parámetros mayor que diez (Bull et al., 2002); esto es, cuando

$$\frac{\text{Número de Observaciones}}{\text{Número de parámetros}} \geq 10,$$

podemos decir que tenemos un tamaño muestra adecuado para el modelo.

Otro aspecto relacionado con el tamaño de la muestra que ha acaparado gran interés es el número de eventos por variable (EPV); esto es, la relación que existe entre el número de veces en que la variable dependiente tiene resultados positivos y el número de variables del modelo de regresión (Concato, Peduzzi, Holford, y Feinstein, 1995; Harrel et al., 1985; G. King y Zeng, 2001a, 2001b; Peduzzi, Concato, Feinstein, y Holford, 1995; Peduzzi et al., 1996). En concreto, se estima que en la medida en que

$$\frac{f(y = 1)}{\text{Número de Variables Independientes}} \geq 10,$$

el porcentaje de sesgo relativo en la estimación de los coeficientes así como sus errores absolutos se reduce, la varianza de las estimaciones se minimiza, se reduce la tasa de ajustes paradójicos y aumenta la potencia estadística (Peduzzi et al., 1995, 1996).

Una de las estrategias que se ha seguido para paliar el problema del tamaño de la muestra ha sido el desarrollo de lo que se conoce como *regresión logística exacta* (Hirji, Mehta, y Patel, 1987; E. King y Ryan, 2002; Mehta y Patel, 1995). La regresión logística exacta se basa en cálculos condicionales y permutacionales. No obstante, tiene el problema de basarse en procedimientos de cómputo intensivos que no han podido ser implementados en los programas informáticos hasta hace relativamente poco tiempo.

Otra técnica que se ha seguido para intentar solucionar los problemas que

## 2.3 Limitaciones y avances en el análisis de regresión logística

surgen de un reducido tamaño de muestra es la optimización de las funciones que hacen mínimo el error asintótico (Firth, 1993). Por ejemplo, (Bull et al., 2002) desarrollaron una función de penalización que reducía los problemas numéricos en la regresión logística multinomial. El *software* LogXact desarrollado por Cytel Software (www.cytel.com) es un ejemplo del intento de superar esta limitación de la regresión logística.

### 2.3.2. Casos perdidos

El problema de la presencia de datos perdidos está muy relacionado con el problema del tamaño de la muestra descrito anteriormente. Ello es debido a que cuando existe un caso con un dato perdido en alguna de sus variables en una muestra, este caso suele ser excluido del análisis y, por consiguiente, el tamaño de la muestra se reduce. Aunque se han desarrollado procedimientos de imputación orientados a reducir el impacto de la tasa de valores perdidos en las muestras (p. e. Allison, 2002; Hair et al., 1998), lo más recomendable es poner medidas para que se produzca la menor tasa posible de datos perdidos.

No obstante, también se han desarrollado métodos para paliar la desventaja que supone la presencia de datos perdidos en las muestras. Un ejemplo de paquete estadístico que incorpora este nuevo tipo de procedimientos es LogXact.

### 2.3.3. Categorización de las variables

Para terminar, hay que destacar que la discretización a la que son sometidas algunas variables puede afectar a los resultados de los análisis de regresión logística especialmente, como señalan (Reade y Kupper, 1995), en los estudios sociales.

La categorización de variables continuas tiene como ventaja el evitar asunciones distribucionales sobre las variables, permite realizar interpretaciones más simples de los resultados, proporciona mayor robustez en el caso de que el modelo haya sido mal especificado o cuando existen datos extremos (Cusmille y Bangdiwala, 2000; Zhao y Kolonel, 1992). No obstante, también supone una pérdida de información importante que produce serios sesgos en los resultados estadísticos.

En primer lugar, las estimaciones de las medidas de asociación entre las variables pueden variar dependiendo del punto de corte tomado para hacer la clasificación (Altman, Lausen, Sauerbrei, y Schumacher, 1994; Flegal, Keyl, y Nieto,

1991). Por ejemplo, Ragland (1992) observó que al recodificar dicotómicamente una variable dependiente continua distribuida normalmente, la función que la relaciona con la *odds ratio* toma forma de U a medida que aumentamos el valor de corte.

El error de clasificación se da cuando los sujetos o casos de una base de datos son asignados a categorías particulares aunque deberían haber sido asignados a otras categorías (Reade y Kupper, 1995). Se han identificado dos tipos importantes de errores de clasificación: las clasificaciones no diferenciales y las clasificaciones diferenciales. En los errores de clasificación no diferencial la probabilidad de mala clasificación de un sujeto o caso es independiente de si el sujeto o dato posee el rasgo o no (p. e., una enfermedad). Por su parte, en los errores de clasificación diferencial existe una probabilidad diferencial de ser mal clasificado para sujetos o casos que poseen o no poseen el rasgo. Flegal et al. (1991) encontraron que los errores de clasificación en las variables dependientes de un modelo de regresión logística tenían un efecto diferente dependiendo de si se cometían clasificaciones diferenciales o no diferenciales. En concreto, observaron que en el caso de clasificaciones diferenciales se produce un sesgo en la *odds ratio* que tendía a subestimar la relación entre variable dependiente e independiente. Por otro lado, cuando la mala clasificación era de tipo no diferencial el sesgo en el nivel de riesgo relativo tendía a la nulidad o se alejaba de ella.

Por su parte, Fung y Howe (1984) estudiaron los efectos de la mala clasificación no diferencial conjunta de una variable independiente y una variable confusora sobre la estimación del riesgo relativo. Un ejemplo del efecto que tienen las variables confusoras sobre la estimación de los parámetros de un modelo logístico lo podemos encontrar en el trabajo de Chuang (1997); donde se encontró que la tendencia a abandonar los estudios en estudiantes de secundaria no estaba relacionada con la raza de los mismos, sino que, más bien, existían un conjunto de variables confusoras (como haber sido expulsado del colegio o haber estado implicado en actividades ilegales) que habían estado enmascarando la relación entre variable dependiente e independiente en estudios anteriores. Así pues, cuando la variable independiente y la variable confusora están asociadas positivamente y el factor de riesgo es el que está mal clasificado, las estimaciones de la *odds ratio* tiende a uno. Sin embargo, cuando la variable independiente y la confusora estaban relacionados negativamente, tanto la mala clasificación de una como de otra

### 2.3 Limitaciones y avances en el análisis de regresión logística

sesgan las estimaciones hacia la unidad.

Otro efecto importante que tiene la categorización de variables continuas sobre el análisis de datos es la reducción de la potencia estadística (Cusmille y Bangdiwala, 2000; Fung y Howe, 1984; Ragland, 1992). A su vez, la mala categorización de las variables también puede aumentar el riesgo del Error Tipo I (Altman et al., 1994; Reade y Kupper, 1995).

Para terminar, también hay que tener en cuenta el efecto que tienen los errores de medida en relación a los errores de clasificación. Flegal et al. (1991) hicieron notar que cuando una variable continua independiente medida con error era dicotomizada tomando un punto de corte arbitrario, la probabilidad de mala clasificación podía no ser uniforme a lo largo de los valores de exposición. Más concretamente, cuando las categorías de exposición habían sido formadas a partir de una variable continua con errores distribuidos normalmente, la probabilidad de mala clasificación es mayor cerca del punto de corte usado para formar las categorías, independientemente del valor usado para hacer la clasificación.

## Fundamentos de la regresión logística

# Capítulo 3

## Fundamentos de las redes bayesianas

*"[...] que le murasen y tapiasen el aposento de los libros, [...] quizá quitando la causa, cesaría el efecto [...]"*

(De Cervantes, 1605/2004)

### 3.1. Estructura y funcionamiento

Las redes bayesianas (también conocidas como redes causales probabilísticas, sistemas expertos bayesianos, sistemas expertos probabilísticos, redes causales, redes de creencia o diagramas de influencia) son herramientas estadísticas que representan un conjunto de incertidumbres asociadas en base a las relaciones de independencia condicional que se establecen entre ellas. Pertenecen al conjunto de técnicas orientadas a la modelización gráfica (I. Martínez y Rodríguez, 2003) y forman parte de la familia de los *Sistemas Estocásticos Altamente Estructurados* (Cowell et al., 1999). En la Tabla 3.1 aparecen diferentes definiciones de red bayesiana.

Por tanto, una red bayesiana es un modelo probabilístico multivariante que representa las variables del modelo gráficamente; siguiendo a Kadie, Hovel, y Horvitz (2001) diríamos que una red bayesiana es un conjunto de variables, una estructura gráfica conectando estas variables y un conjunto de distribuciones de

probabilidad condicional. Codifica la incertidumbre asociada a cada variable por medio de probabilidades y, gracias al teorema de Bayes, esta incertidumbre es susceptible de ser modificada en base a observaciones (o evidencias) sobre el modelo.

**Tabla 3.1:** Algunas definiciones de red bayesiana

DEFINICIÓN	AUTOR/ES	AÑO
<i>Una red bayesiana es una representación para una distribución de probabilidad conjunta en la que se puede asignar la probabilidad para todas las posibles combinaciones de valores para todas las variables</i>	Martin y VanLehn	1995
<i>Una red bayesiana es un modelo gráfico para las relaciones probabilísticas entre un conjunto de variables</i>	Heckerman	1995
<i>Una red bayesiana es un modo de representar incertidumbres relacionadas</i>	Edwards	1998
<i>Una red bayesiana es la especificación de la distribución de probabilidad conjunta de varias variables en términos de las distribuciones condicionales de cada variable</i>	Nadkarni y Shenoy	2001
<i>Las redes bayesianas son un método de inferencia donde la incertidumbre es manipulada de un modo lógicamente riguroso pero simple</i>	Garbolino y Taroni	2002
<i>Una red bayesiana es un triplete <math>(\mathcal{V}, \mathcal{G}, \mathcal{P})</math>. <math>\mathcal{V}</math> es un conjunto de variables, <math>\mathcal{G}</math> es un grafo dirigido acíclico cuyos nodos corresponden uno-a-uno a los miembros de <math>\mathcal{V}</math> de tal modo que cada variable es condicionalmente independiente de sus no descendentes dados sus padres. Denotando los padres de <math>v \in \mathcal{V}</math> en <math>\mathcal{G}</math> por <math>\pi(v)</math>; <math>\mathcal{P}</math> es un conjunto de distribuciones de probabilidad: <math>\mathcal{P} = \{P(v \pi(v))   v \in \mathcal{V}\}</math></i>	Xiang	2002
<i>Las redes causales bayesianas representan hipótesis causales como grafos dirigidos</i>	Glymour	2003
<i>Un red bayesiana es u par <math>(\mathcal{G}, \mathcal{P})</math> donde <math>\mathcal{G}</math> es un grafo dirigido acíclico en el que cada vértice es una variable aleatoria, <math>\mathcal{P} = \{p(x_i \pi_i), \dots, p(x_n \pi_n)\}</math> es un conjunto de <math>n</math> funciones de probabilidad condicionada, una para cada variable, y <math>\pi_i</math> es el conjunto de padres del nodo <math>X_i</math> en <math>\mathcal{G}</math></i>	I. Martínez y Rodríguez	2003

Las variables del modelo pueden ser probabilísticas o determinísticas. Cuando una variable es probabilística los valores que puede tomar están determinados con una probabilidad en función de los valores que toman las otras variables de las que depende. Por el contrario, cuando una variable es determinística su estado, en relación a las variables de las que depende, se expresa en términos absolutos. Por ejemplo, una variable que representa una compuerta lógica tipo «Y» u «O» sería una variable determinística. A continuación se verá más detalladamente en qué consiste la estructura y el funcionamiento de una red bayesiana.

### 3.1.1. Estructura

Un criterio que se ha utilizado para describir la estructura de una red bayesiana ha sido el grado en que sus elementos son «visibles» para el usuario (Edwards, 1998; Edwards y Fasolo, 2001). En el nivel más molar estarían las variables del modelo representadas por nodos y un conjunto de flechas que relacionan estas variables en términos de influencia. A un nivel inferior estarían los niveles o estados, también conocidos como *espacio de estados* (Nadkarni y Shenoy, 2001, 2004), de cada una de las variables del modelo. En tercer lugar, tendríamos un conjunto de tablas de probabilidad condicional, una para cada nodo, donde se representaría la probabilidad de cada estado de la variable condicionado a los posibles valores de las variables que les envían una flecha. Por último, y en el nivel más bajo, tendríamos un conjunto de algoritmos que permiten que la red recalcule las probabilidades cuando recibe algún tipo de evidencia.

Sin embargo, la descripción más frecuente de la estructura de una red bayesiana se refiere a sus dos dimensiones, una cualitativa y otra cuantitativa (p. e., I. Martínez y Rodríguez, 2003), que se sustentan en dos grandes pilares de la modelización matemática: la teoría de grafos y la teoría de la probabilidad, respectivamente (Ríos, 1995). Veamos en que consiste cada una de estas dimensiones.

#### Dimensión cualitativa

Una red bayesiana es un *grafo*, esto es; una representación gráfica de un problema. Aunque la definición de grafo, y la terminología que lo acompaña, varía en función de los autores (p. e., Harary, 1969; Gould, 1988; Spirtes, Glymour, y Scheines, 2000; Tutte, 1984; Xiang, 2002); podemos definir grafo como *un par  $G = (V, E)$ , donde  $V$  es un conjunto finito de vértices, nodos o variables y  $E$  es un subconjunto del conjunto  $V \times V$  de pares ordenados de vértices llamados enlaces o aristas.* Además, una red bayesiana es un tipo particular de gráfico que se denomina *grafo dirigido acíclico*. Dirigido hace referencia a que los enlaces entre los vértices del grafo están orientados. Por ejemplo, si  $(A, B) \in E$  pero  $(B, A) \notin E$ , diremos que hay un *enlace dirigido* o *arco* entre los nodos y lo representaremos como  $A \rightarrow B$ . Acíclico se refiere a que no pueden existir ciclos, bucles o *loops* en el grafo, esto es; que si comenzamos a recorrer un camino desde un nodo, nunca

podríamos regresar al punto de partida.

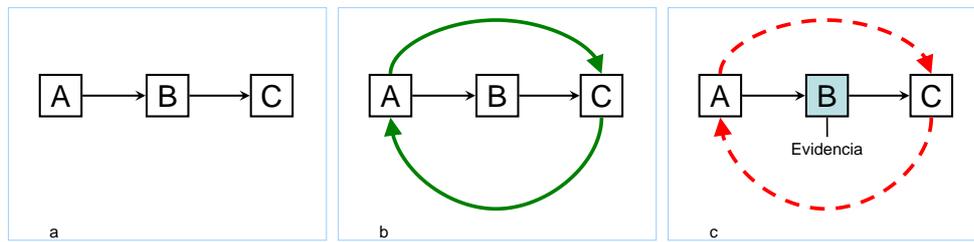
Una conexión tipo  $A \rightarrow B$  indica dependencia o relevancia directa, o sea; que  $B$  depende de  $A$ . En términos causales diríamos que  $A$  es la causa de  $B$ , y  $B$  el efecto de  $A$ . También se dice que  $A$  es *padre* (o la variable *madre*) de  $B$  y que  $B$  es el *hijo* (o la variable hija) de  $A$ . Aunque la presencia de arcos entre nodos codifica información esencial sobre el modelo representado en la red, la ausencia de arcos entre nodos aporta una valiosísima información ya que representa independencia condicional (Edwards, 1998).

El *principio de independencia condicional* puede enunciarse del siguiente modo: sean tres conjuntos  $X, Y$  y  $Z$  de variables;  $X$  e  $Y$  son (condicionalmente) independientes dado  $Z$  si  $p(x|z) = p(x|yz)$ . Otro modo de enunciar el principio de independencia condicional es el siguiente: dos variables,  $X$  e  $Y$ , son independientes en probabilidad de una tercera  $Z$  si y solamente si  $p(xy|z) = p(x|z) \times p(y|z)$ . La consecuencia fundamental de este principio es que la probabilidad de  $X$  es la misma condicionándola a  $Z$  que condicionándola a  $Z$  e  $Y$ . O lo que es lo mismo, que si conocemos el estado de  $Z$ , el conocimiento del estado de  $Y$  es irrelevante para saber algo de  $X$ . Por último, una propiedad importante del principio de independencia condicional es que es simétrico; esto es, cuando  $X$  es independiente de  $Y$ ,  $Y$  es independiente de  $X$ . Esta propiedad de las redes bayesianas permite factorizar la distribución conjunta de las variables del modelo de tal manera que la propagación de evidencias a través del modelo se realiza eficientemente ya que

$$p(x) = \prod_{v \in V} p(x_v | x_{pa(v)}), \quad (3.1)$$

donde  $x_v$  se refiere a un nodo concreto del modelo y  $x_{pa(v)}$  se refiere a sus padres (p. e., Cowell et al., 1999; Gámez, 1998; Hoeting, Madigan, y Raftery, 1999; Huete, 1998; J. López et al., 2007; Nadkarni y Shenoy, 2001).

En una red bayesiana pueden existir tres tipos de conexiones básicas (seriales, convergentes y divergentes), cada una con propiedades cualitativas diferentes y que favorecen matizaciones en la propagación de probabilidades ante una nueva evidencia sobre el modelo. Estas estructuras son bien conocidas en estadística (p.e., I. Martínez y Rodríguez, 2003), se tienen en cuenta a la hora de aprender estructuras causales (p. e., Glymour, 2001, 2003) y han sido propuestas como



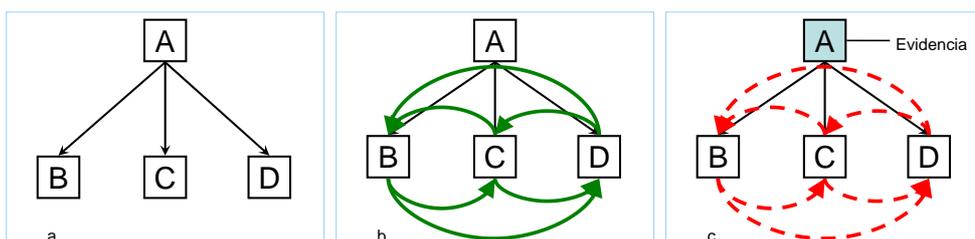
**Figura 3.1:** Conexión serial (a), su flujo de información (b) y la alteración de la propagación ante una evidencia (c).

una herramienta valiosa en la evaluación de evidencias científicas en el contexto forense (p. e., Garbolino y Taroni, 2002; Oatley y Ewart, 2003).

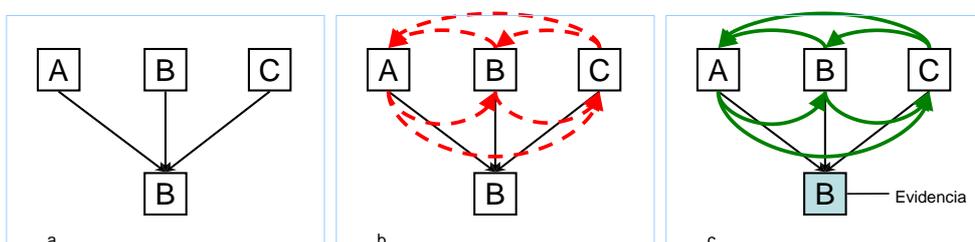
Las *conexiones en serie*, también llamadas *cadena causal*, representan a un conjunto de variables asociadas linealmente que denota dependencia entre las variables (Figura 3.1). En este caso,  $B$  depende de  $A$  y  $C$  de  $B$ . En términos causales diríamos que  $A$  es causa de  $B$  y que  $B$  es causa de  $C$ . En este caso, dada la dependencia entre las variables, cuando conocemos información sobre  $A$  podemos modificar nuestra certeza sobre el estado de  $C$ ; y a la inversa, cuando sabemos algo sobre el estado de  $C$  la creencia sobre el estado de  $A$  se altera. Sin embargo, si el estado de  $B$  es conocido; conocer alguna información sobre  $A$  o  $C$  no modificará nuestra creencia sobre el estado de estas variables. Podríamos decir que la propagación de la información se bloquea y se dice que  $A$  y  $C$  se tornan condicionalmente independientes dado  $B$  (Figura 3.1 derecha).

En las *conexiones divergentes* tenemos un nodo *padre* que proyecta sus arcos hacia varios hijos; o lo que es lo mismo, las flechas que salen de él y divergen hacia sus hijos (Figura 3.2 izquierda). Este tipo de conexión es el más apropiado para representar enfermedades o trastornos que son detectables por sus síntomas (Herskovits y Dagher, 1997) y en términos causales se interpretaría que una causa (p. e. un resfriado) genera varios síntomas (p. e. tos, congestión, dolor de cabeza, etc.). Cuando no conocemos el estado de la variable *madre* existe una dependencia entre las variables análoga a la que se da en la conexión serial. Sin embargo, cuando el estado de la variable *madre* se conoce las variables hijo se tornan independientes (Figura 3.2 derecha) y la información no se propaga si añadimos evidencias sobre los nodos hijos. En nuestro ejemplo,  $B$ ,  $C$  y  $D$  serían independientes dado  $A$ .

Por último, en las *conexiones convergentes* (llamadas también *cabeza a cabeza*)



**Figura 3.2:** Conexión divergente (a), su flujo de información (b) y la alteración de la propagación ante una evidencia (c).



**Figura 3.3:** Conexión convergente (a), su flujo de información (b) y la alteración de la propagación ante una evidencia (c).

varias variables apuntan con sus arcos hacia una variable de convergencia (Figura 3.3), esto es, convergen en un mismo nodo hijo. En este tipo de conexiones las variables *madre* son independientes entre sí, pero la variable de convergencia puede influir sobre sus nodos *padre* (Figura 3.3 centro). La propiedad importante de este tipo de conexión a la hora de propagar la información es que cuando se tiene evidencia sobre la variable de convergencia, los nodos *padre* se vuelven dependientes y la evidencia sobre el estado de uno de ellos se propaga por los demás (Figura 3.3 derecha). Se diría, en nuestro ejemplo, que *A*, *B* y *C* son dependientes dado *D*.

### Dimensión cuantitativa

Hay tres elementos esenciales que caracterizan la dimensión cuantitativa de una red bayesiana: el concepto de probabilidad como un grado de creencia subjetiva relativa a la ocurrencia de un evento, el teorema de Bayes como heurístico actualizador de creencias y un conjunto de funciones de probabilidad condicionada.

Existen, al menos, cuatro formas de entender la probabilidad: clásica, empírica, axiomática y subjetivamente. Desde la concepción clásica, introducida por Laplace, la probabilidad de que ocurra un evento de un espacio muestral viene

dado por la razón que se establece entre el número de casos favorables y el número de casos posibles (3.2).

$$p(A) = \frac{n_A}{n}, \quad (3.2)$$

donde  $n_A$  representa el número de casos favorables a  $A$  y  $n$  representa el número total de casos.

Sin embargo, este principio conduce a conclusiones falaces en determinadas situaciones (p. e. cuando la muestra de sucesos es muy pequeña). Para solventar este problema se propuso otra definición de probabilidad, compatible con la clásica, donde se entiende que la probabilidad de un evento esta en función del número de ensayos favorables al suceso cuando estos tienden a infinito (3.3).

$$p(A) = \lim_{n \rightarrow \infty} \frac{n_A}{n}. \quad (3.3)$$

Pero esta definición de probabilidad también plantea problemas porque, por ejemplo, hay situaciones en las que no es posible repetir un experimento infinitas veces. Por lo tanto se hizo necesario desarrollar otra forma de entender la probabilidad. Fue Kolmogoroff, entre 1930 y 1950, quien propuso un conjunto de axiomas que hicieran matemáticamente tratable el concepto de probabilidad; a saber: que la probabilidad está comprendida entre 0 y 1, que la probabilidad de un suceso seguro es 1 y que la probabilidad de la unión de sucesos incompatibles es la suma de sus probabilidades.

Por último, la probabilidad puede ser entendida como algo subjetivo (Cox, 1946), como el grado de creencia que tenemos sobre la ocurrencia de un suceso. La probabilidad sería entendida como «algo que está en nuestra mente y no en el mundo físico real» (Dixon, 1964/1970, p. 5.2). A esta concepción de la probabilidad se le suele denominar bayesiana y para clarificar esta nueva concepción de la probabilidad se suele aludir a la dicotomía frecuentista/bayesiana (p.e., Alonso y Tubau, 2002; Cowell et al., 1999; De la Fuente, García, y De la Fuente, 2002; Heckerman, 1995; Serrano, 2003). Por otro lado, la concepción bayesiana de la probabilidad está lejos de ser unitaria; más bien, como señala Neapolitan y Morris (2004) pueden identificarse tres grupos de perspectivas bayesianas que difieren entre sí en la concepción de probabilidad física: a) una rama que la considera inexistente, b) otra que la rechaza y c) un tercer grupo la usa como si

existiese aunque no la justifica filosóficamente.

De una manera u otra, la probabilidad es una forma de cuantificar la incertidumbre asociada a la ocurrencia de eventos. En el caso de las redes bayesianas la probabilidad es entendida en sentido subjetivo, es decir, como la creencia subjetiva que tenemos de la ocurrencia de un suceso.

El teorema de Bayes (3.4) es una regla que se deriva del concepto de probabilidad condicional aplicado a la intersección de sucesos. Es una herramienta que ayuda a actualizar la creencia que tenemos sobre un evento cuando conocemos información relacionada con el mismo.

$$p(B|A) = \frac{p(A|B) \times p(B)}{p(A)}. \quad (3.4)$$

Por ejemplo, supongamos que la prevalencia del Trastorno Depresivo Mayor en la población general es del 25 % ( $p(D) = 0,25$ ) y que el 60 % de las personas que sufren un Trastorno Depresivo Mayor tienen insomnio ( $p(I|D) = 0,6$ ). Si una amiga nuestra nos dijese que ha estado experimentando serios problemas para dormir y que se ha sentido muy afligida durante los últimos meses, podríamos calcular la probabilidad de que estuviese sufriendo un Trastorno Depresivo Mayor usando la regla de Bayes:

$$p(D|I) = \frac{p(I|D) \times p(D)}{p(I)} = \frac{0,6 \times 0,25}{1} = 0,15.$$

Pues bien, las redes bayesianas usan el teorema de Bayes y los principios de independencia condicional entre las variables de un modelo gráfico para propagar probabilidades en base a evidencias de este tipo. Aunque estos principios en los que se basa la estadística bayesiana fueron producidos hace alrededor de 200 años, no fue hasta la década de 1930 cuando se pusieron en práctica (SPSS y Recognition Systems, 1997).

Por último, las tablas de probabilidad condicional (TPC) son los elementos sobre los que opera el teorema de Bayes. En una red bayesiana cada variable trabaja como una función probabilística definida por una tabla de probabilidad condicional que representa los valores (o niveles) de la variable condicionados a cada una de las posibles combinaciones que se establecen entre los niveles o valores de las variables *madre*. A todas estas combinaciones entre los valores de la variable y los valores de sus variables *madre* se les asigna una probabilidad y

estas probabilidades son sobre las que trabaja el teorema de Bayes.

### 3.1.2. Funcionamiento

Como se ha señalado, una red bayesiana es un modelo probabilístico multivariado que trabaja tanto con información cuantitativa como cualitativa. Esto es una gran ventaja ya que podemos tener una idea general del modelo con un simple *vistazo* y porque los algoritmos de propagación de probabilidades nos indican el estado más probable de una o varias variables teniendo en cuenta un conjunto de evidencias. Para mostrar como funciona una red bayesiana en su modo de inferencia vamos a utilizar la red *Recuerdo*<sup>1</sup> que ha sido diseñada específicamente para este texto (Figura 3.4). Podemos suponer que para desarrollarla hemos llevado a cabo un conjunto de estudios experimentales con los que hemos obtenido registros de las variables que aquí aparecen reflejadas.

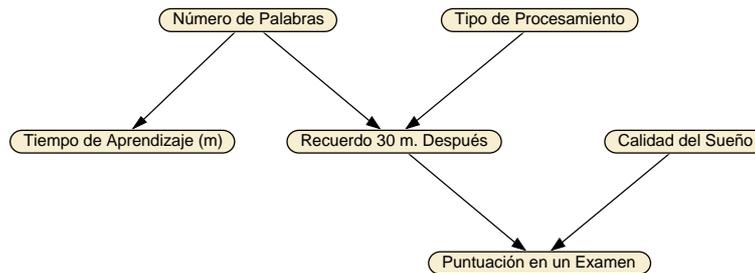


Figura 3.4: Red *Recuerdo*.

Este sería el aspecto cualitativo, o el nivel más molar (Edwards, 1998; Edwards y Fasolo, 2001), de nuestra red. La descripción de las variables y sus posibles valores, o sus espacios de estados (Nadkarni y Shenoy, 2001, 2004), serían los siguientes<sup>2</sup>:

- *Número de palabras (NP)*: número de palabras de una lista que puede tomar los valores *Muchas* y *Pocas*.
- *Tipo de Procesamiento (TP)*: que hace referencia al tipo de estrategia llevada a cabo durante la fase de aprendizaje y que puede tomar los valores

<sup>1</sup>Los datos presentados en esta red son ficticios y su valor es únicamente ilustrativo.

<sup>2</sup>Aunque las redes bayesianas pueden representar variables cualitativas de múltiples niveles y variables cuantitativas, aquí vamos a suponer que todas las variables son cualitativas y dicotómicas con fines ilustrativos.

**Tabla 3.2:** TPC para el nodo *NP*.

$p(NP)$	Muchas	Pocas
	0,50	0,5

**Tabla 3.3:** TPC para el nodo *TP*.

$p(TP)$	Profundo	Superficial
	0,5	0,5

**Tabla 3.4:** TPC para el nodo *SU*.

$p(SU)$	Buena	Mala
	0,45	0,55

**Tabla 3.5:** TPC para el nodo *TA*.

$p(TA NP)$	Alto	Bajo
Muchas	0,75	0,25
Pocas	0,25	0,75

*Profundo o Superficial.*

- *Tiempo de Aprendizaje (TA):* tiempo necesario para aprender una lista de palabras con los valores *Alto* y *Bajo*.
- *Recuerdo Después de 30 m. (RE):* indica como ha sido el recuerdo de palabras después de la fase de aprendizaje, sus valores son *Buena* y *Mala*.
- *Puntuación en un Examen (EX):* representa la puntuación en un examen dos días después del aprendizaje y puede tomar los valores de *Alta* o *Baja*.
- *Calidad del Sueño (SU):* referida a la calidad con que ha dormido una persona la noche anterior al examen que puede ser *Buena* o *Mala*.

Con lo anterior tendríamos definida la estructura cualitativa de la red pero para poder verla funcionar necesitaríamos un conjunto de funciones de probabilidad, una por variable, codificadas mediante tablas de probabilidad condicional (TPC). Las Tablas de la 3.2 a la 3.7 representan las tablas de probabilidad condicional contenidas en cada variable.

La interpretación de las probabilidades de las tablas es la siguiente. En las tablas que pertenecen a variables que no tienen *padres* aparecen las probabilidades previas asignadas a cada nivel de la variable. Por ejemplo, en la Tabla 3.4 se indica que la probabilidad de tener una noche de sueño placentero antes del examen es de 0,45; mientras que la probabilidad de que sea mala es de 0,55. En las variables que tienen *padres* habrá tantas probabilidades como permitan todas las posibles

**Tabla 3.6:** TPC para el nodo *RE*.

$p(RE NP, TP)$	Bueno	Malo
Muchas, Superficial	0,25	0,75
Muchas, Profundo	0,75	0,25
Pocas, Superficial	0,55	0,45
Pocas, Profundo	0,85	0,15

**Tabla 3.7:** TPC para el nodo *EX*.

$p(EX RE, SU)$	Alta	Baja
Bueno, Buena	0,95	0,05
Bueno, Mala	0,70	0,30
Malo, Buena	0,50	0,50
Malo, Mala	0,25	0,75

combinaciones entre los niveles de las variables *padre* y la variable en cuestión. Por ejemplo, en el cruce de la segunda fila y la segunda columna de la Tabla 3.6 aparece la probabilidad (0,25) de que el recuerdo después de 30 minutos sea bueno *dado que* el número de palabras de la lista sea elevado y se haya hecho una codificación superficial. A la derecha de este valor aparecería la probabilidad de que el recuerdo fuese malo en el caso de que el número de palabras de la lista fuese grande y se hubiese realizado un procesamiento superficial. Sin embargo, aunque la red *Recuerdo* necesita todas estas probabilidades para funcionar, no son visibles para el usuario que pretende realizar inferencias derivadas de una evidencia sobre el modelo.

Lo primero que tenemos que hacer cuando tratamos de utilizar una red bayesiana es *compilarla*. Cuando compilamos una red bayesiana generamos una representación interna del conjunto de funciones de probabilidad codificadas mediante tablas de probabilidad condicional. Esta representación interna da como resultado el conjunto de probabilidades previas asociadas a cada nivel de cada una de las variables del modelo (Figura 3.5)<sup>3</sup>. Las barras que aparecen a la derecha de la probabilidad previa para cada nivel de la variable representa el tamaño de la probabilidad proporcionalmente. En lo que respecta a la factorización del modelo, basándonos en la ecuación 3.1, la probabilidad conjunta de las variables viene dada por

$$p(X) = p(NP) \times p(TP) \times p(TA|NP) \times p(RE|NP, TP) \times p(SU) \times p(EX|RE, SU).$$

En este momento podemos *instanciar*<sup>4</sup> alguna de las variables del modelo.

<sup>3</sup>En los gráficos las probabilidades se expresarán como porcentajes.

<sup>4</sup>Instanciar es suministrar una evidencia/ejemplo a la red que genera cambios en algunas

## Fundamentos de las redes bayesianas

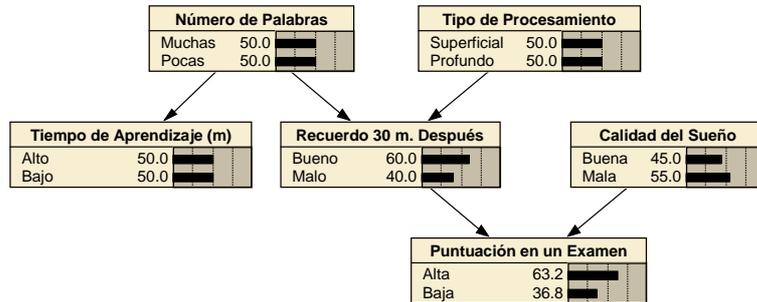


Figura 3.5: Red *Recuerdo* compilada.

Cuando hacemos esto, esta información se propaga por la red gracias al teorema de Bayes y algunas probabilidades son recalculadas, obteniéndose lo que se denomina *distribución posterior o a posteriori*. Hay dos modos de instanciar una variable: con evidencia *dura*, cuando la probabilidad del estado de la variable es conocido al 100%; o con evidencia *suave*, que serían las situaciones en las que tenemos un conocimiento incerto sobre el estado de la variable.

Como se puede ver en la red, la presencia de un arco que une dos nodos representa relación de relevancia o dependencia directa. Por ejemplo, *Número de Palabras* → *Tiempo de Aprendizaje* indica que el tiempo necesario para aprender todas las palabras de una lista depende del número de palabras que contiene.

Consideremos el subgrafo *Número de Palabras* → *Recuerdo Después de 30 m.* → *Puntuación en un Examen*. Según esta representación el número de palabras de una lista influye en el recuerdo y este, a su vez, sobre la puntuación en un examen. Además, cuando no se sabe nada sobre el recuerdo, la puntuación en un examen nos permite estimar el número de palabras que había en la lista y viceversa. Por ejemplo, imaginemos que el número de palabras fue pequeño para un participante del estudio (Figura 3.6). Si instanciamos la red haciendo que el valor de la variable *Número de Palabras* en su nivel *Pocas* valga 1 veremos que la probabilidad de tener un buen recuerdo pasados 30 minutos aumenta desde 0,6 hasta 0,7. Por otro lado, la probabilidad de que esa persona obtuviese una puntuación elevada en el examen aumenta de 0,63 a 0,68.

Imaginemos ahora que nos encontramos a un participante del estudio por la calle y nos comenta que en el examen obtuvo una puntuación muy baja. Si introducimos esa información en la red (Figura 3.7) podemos observar como la

---

variables.

### 3.2 Los diagramas de influencia

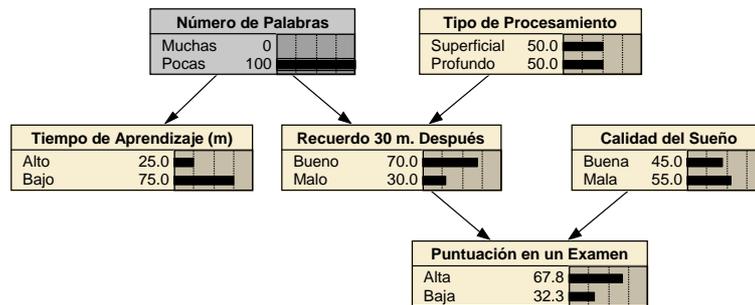


Figura 3.6: Red *Recuerdo* con *Número de Palabras* instanciado.

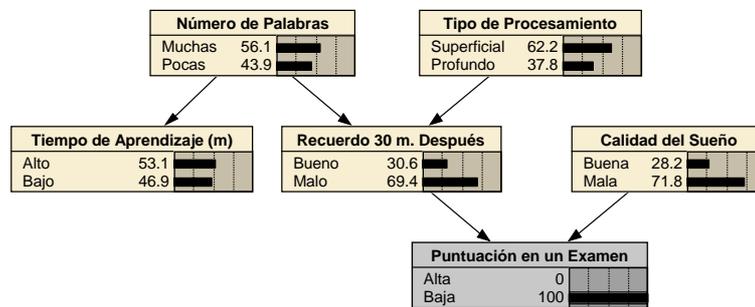


Figura 3.7: Red *Recuerdo* con *Puntuación en un Examen* instanciado.

información se propaga en sentido contrario a los arcos y la probabilidad de un mal recuerdo asciende de 0,4 hasta 0,69. A su vez, la probabilidad de haber estudiado una lista con muchas palabras aumenta de 0,50 a 0,56.

Supongamos ahora que el recuerdo de una persona tras 30 minutos fue malo (Figura 3.8), en este caso la probabilidad de que puntúe bajo en el examen pasa a 0,64. Si, además, nos dicen que esa persona perteneció al grupo que estudió la lista de pocas palabras, las probabilidades de *Puntuación en un Examen* no sufren ningún cambio (Figura 3.9) ya que la puntuación en el examen es independiente del número de palabras dado el recuerdo tras 30 minutos (puede que el participante del experimento no estuviese motivado lo suficiente).

## 3.2. Los diagramas de influencia

Existe otro tipo de redes denominadas diagramas de influencia que son un híbrido de las redes bayesianas y de la teoría de la utilidad (Edwards, 1998; Edwards y Fasolo, 2001). La característica esencial de este tipo de redes es que añaden una dimensión de utilidad y están orientadas a la toma de decisiones. En este tipo de

## Fundamentos de las redes bayesianas

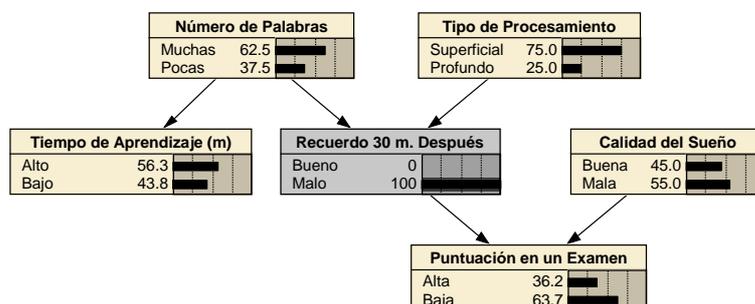


Figura 3.8: Red *Recuerdo* con *Recuerdo Después de 30 m.* instanciado.

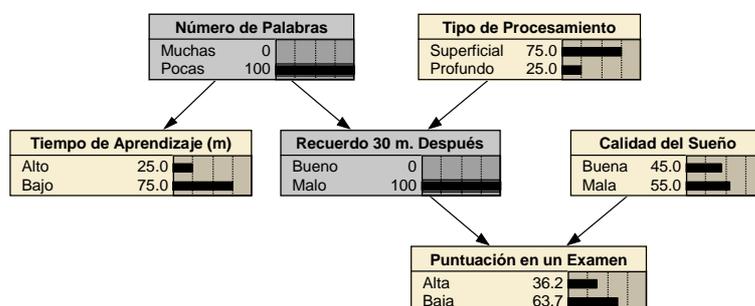


Figura 3.9: Red *Recuerdo* con *Recuerdo Después de 30 m.* y *Número de Palabras* instanciados.

redes hay dos tipos nuevos de nodos: los nodos de utilidad y los nodos de decisión.

Los *nodos de utilidad* o *satisfacción* son tablas de utilidad, donde aparecen las posibles combinaciones de los estados de los nodos *padre*. Sin embargo, en vez de llevar asociada una probabilidad a cada una de las posibles combinaciones de los estados de los nodos *padre*, aparece un valor de *utilidad* para cada una de las combinaciones. Por su parte, en los *nodos de decisión* aparece la *utilidad esperada* de cada una de las posibles alternativas de decisión.

Los conceptos de utilidad y utilidad esperada provienen de la teoría de la decisión aplicados al campo de la economía y se desarrollaron a partir de la idea de valor esperado o esperanza matemática (Edwards y Fasolo, 2001; Esteve, 1996b; Hothersall, 1995/1997; Maciá, Barbero, Pérez-Llantada, y Vila, 1990; Valencia, 1994).

Para ver como funciona un diagrama de influencia vamos a utilizar la red *Ansiedad*<sup>5</sup> (Figura 3.10) y vamos a considerar el siguiente ejemplo:

«Serafín es un psicólogo clínico especializado en trastornos psicológi-

<sup>5</sup>Los datos presentados en esta red son ficticios y su valor es únicamente ilustrativo.

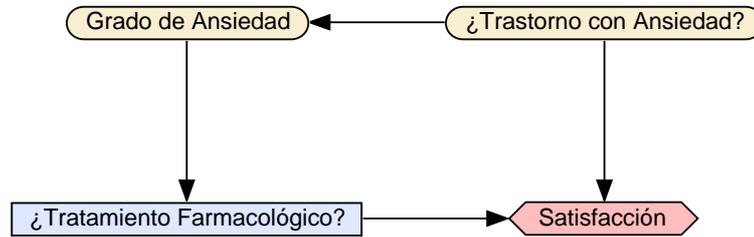


Figura 3.10: Red *Ansiedad*.

cos relacionados con la ansiedad. Por experiencia propia sabe que el 70 % de las personas que acuden a su consulta no adolecen de un trastorno de ansiedad, sino que a la base de sus trastornos hay otros fenómenos. Las pruebas de que dispone Serafín indican si una persona que acude a su consulta carece de ansiedad, tiene niveles normales de ansiedad o muestra ansiedad patológica. Dado que Serafín es un profesional conocedor de los últimos avances psicofarmacológicos, cuando detecta que una persona puede tener un problema grave de ansiedad se pone en contacto con un colega psiquiatra para trabajar en el diseño de un posible tratamiento farmacológico. Lo que más le molesta a Serafín es que una persona tenga un trastorno y no lo detecte o que, en menor medida, una persona no tenga un trastorno pero él piense que sí. Por otro lado, se pone muy contento cuando una persona no tiene trastorno relacionado con ansiedad ni necesita fármacos para su recuperación. También se alegra, aunque menos, cuando detecta que una persona sufre un trastorno de ansiedad grave y necesita medicamentos».

Como se puede apreciar en el grafo, hay tres tipos de nodos. El nodo hexagonal es un nodo de utilidad y el nodo rectangular es un nodo de decisión. Veamos como se reparten los datos de la narración por los nodos de la red y como se llega a esta estructura.

1. Serafín sabe que 0,7 es la probabilidad de que una persona que llega a su consulta tenga un trastorno psicológico relacionado con ansiedad, mientras

que 0,3 es la probabilidad de que no lo sufra. Estos datos se representarían en una tabla de probabilidad análoga a las presentadas en las Tablas 3.2, 3.3 y 3.4 de la red *Recuerdo*.

2. El ejemplo dice que Serafín está muy contento cuando a su consulta llega una persona que no sufre un trastorno de ansiedad y no necesita tratamiento farmacológico (p. e. 100 unidades de satisfacción). También le alegra, pero menos (p. e. 70 unidades de satisfacción), cuando detecta que una persona tiene ansiedad patológica y necesita ayuda farmacológica ya que sus principios éticos le hacen pensar que así puede colaborar con su colega para solucionar los problemas de las personas. Sin embargo, lo que no le gusta nada (p. e. 0 unidades de satisfacción) es no detectar a una persona con ansiedad que necesita un tratamiento farmacológico. Por su parte, también le molesta (p. e. 20 unidades de satisfacción) cuando una persona que no tiene un problema de ansiedad considerable es tratada con psicofármacos. Estos datos quedarían representados en la Tabla 3.8.
  
3. La base de datos de los pacientes de Serafín indica que el 70% de las personas que no tienen ansiedad grave puntúan «Nada» en sus pruebas psicométricas que miden el grado de ansiedad, el 20% puntúan «Normal» y el 10% puntúan «Patológica». Por su parte, el 15% de las personas que tienen un problema severo con la ansiedad puntúan «Nada», el 25% puntúan «Normal» y el 60% puntúan «Patológica» (Tabla 3.9).

En la Figura 3.10 se puede apreciar que hay un arco desde el nodo *¿Trastorno con Ansiedad?* hasta el nodo *Grado de Ansiedad*, representando que el grado de ansiedad que manifiesta una persona depende (es un efecto, un síntoma o un indicio) de la existencia de una psicopatología. Por otro lado, hay una flecha desde el

**Tabla 3.8:** Tabla de utilidad del nodo *Satisfacción*.

<i>¿Trastorno con ansiedad?</i>	<i>¿Tratamiento farmacológico?</i>	Satisfacción
No	Si	20
No	No	100
Si	Si	70
Si	No	0

**Tabla 3.9:** Tabla de probabilidad condicional del nodo *Grado de Ansiedad*.

Trastorno	$p(\text{Ansiedad Nula})$	$p(\text{Ansiedad Normal})$	$p(\text{Ansiedad Patológica})$
No	0,70	0,20	0,10
Si	0,15	0,25	0,60

nodo *Grado de Ansiedad* hasta el nodo de decisión *¿Tratamiento Farmacológico?*, lo que indica que Serafín tendrá que conocer el grado de ansiedad que manifiestan sus pacientes antes de decidir si necesitan tomar psicofármacos o no. Sin embargo, no existe un arco que salga desde el nodo *¿Trastorno con Ansiedad?* y que vaya a parar a *¿Tratamiento Farmacológico?* ya que si el psicólogo conociese esa relación no necesitaría medir el grado de ansiedad ni diseñar esta red bayesiana. Una vez definidos los nodos y establecidas las relaciones de independencia condicional entre las variables, tenemos que compilar el diagrama de influencia para que podamos mostrarle evidencias (de un modo análogo a como lo hacíamos con las redes bayesianas). El aspecto que tendría la red Ansiedad compilada sería el que aparece en la Figura 3.11.

Como se puede observar en la Figura 3.11, la mejor decisión a tomar cuando llega una nueva persona a la consulta de Serafín sería no pensar en el tratamiento farmacológico (utilidad de 70). Eso se explica porque lo más probable (0,7) es que la persona no tenga problemas serios de ansiedad. Supongamos que Serafín lleva a cabo algunas de las pruebas diagnósticas que le permiten determinar si el problema de la persona es debido a la ansiedad. Tras realizarle tests psicológicos y psicofisiológicos descubre que sus niveles de ansiedad son muy bajos. En este caso (Figura 3.12), la probabilidad de que la persona sufra un trastorno donde subyacen problemas de ansiedad ha disminuido hasta 0,09 y la utilidad de prescindir de un tratamiento farmacológico ha aumentado hasta el 91,589.

Imaginemos lo que pasaría si nuestro psicólogo ficticio encontrase que un nuevo paciente obtiene puntuaciones normales de ansiedad (Figura 3.13). En este caso, la probabilidad de que la persona sufriese un trastorno generado por la ansiedad aumentaría hasta 0,35 y la utilidad de prescribir psicofármacos aumenta hasta 37,442. No obstante, lo más sensato sería pensar, hasta no obtener más evidencias, que la persona no sufre ansiedad severa (probabilidad de 0,65) y que algunas puntuaciones relacionadas con el grado de ansiedad son el reflejo de la activación

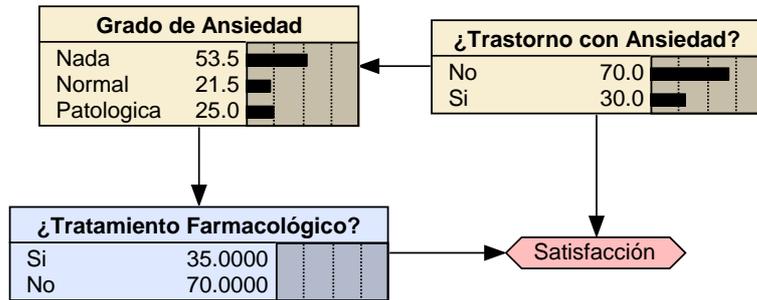


Figura 3.11: Red *Ansiedad* compilada.

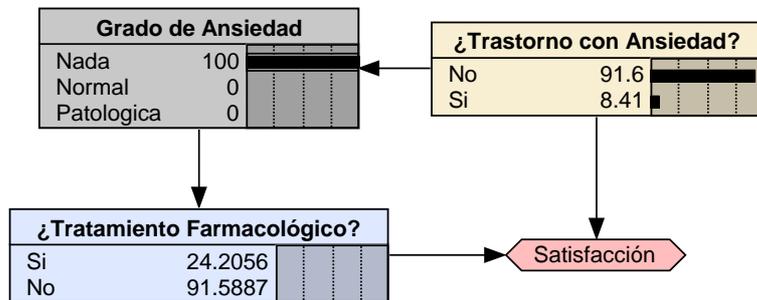


Figura 3.12: Red *Ansiedad* con evidencia de ansiedad nula.

natural que necesitamos para desenvolvemos adaptativamente en nuestro entorno.

Pensemos, para terminar, que las pruebas de Serafín indican que la persona parece estar más ansiosa de lo que cabría esperar (Figura 3.14). En este caso la probabilidad de que exista un problema latente de ansiedad aumenta sustancialmente hasta 0,72 y la utilidad de acudir a un tratamiento farmacológico se eleva hasta el 56,000.

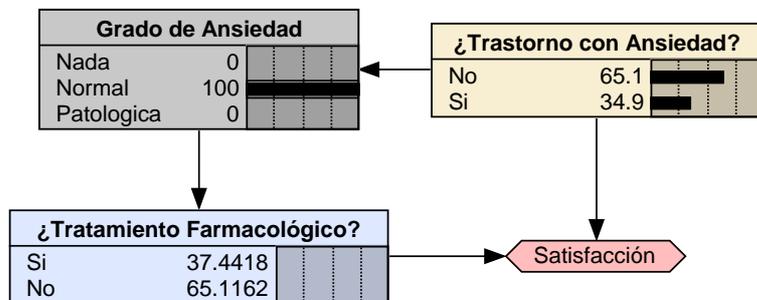


Figura 3.13: Red *Ansiedad* con evidencia de ansiedad normal.

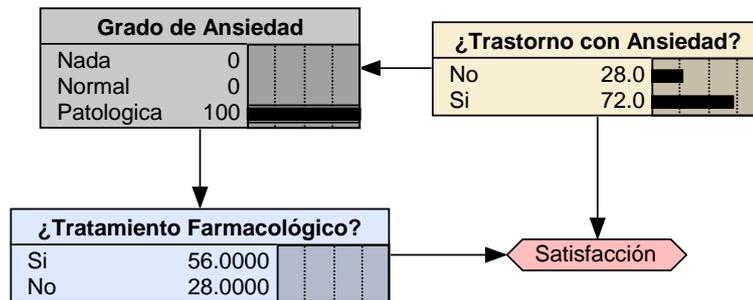


Figura 3.14: Red *Ansiedad* con evidencia de ansiedad patológica.

### 3.3. Construcción de redes bayesianas

La construcción de redes bayesianas implica determinar las dos dimensiones que las componen; esto es, la cuantitativa y la cualitativa. A su vez, existen dos procedimientos generales para crear redes bayesianas (Mani, McDermott, y Valtorta, 1997): a) *construcción automática*, que consiste en desarrollar una red bayesiana a partir de una base de datos por medio de algoritmos que decodifican el grafo subyacente en los datos y los parámetros del modelo; y b) *construcción a partir del conocimiento de expertos*, donde son los expertos quienes facilitan el conocimiento que tienen sobre un área para que sea transformado en una red bayesiana.

#### 3.3.1. Estimación paramétrica

La dimensión cuantitativa<sup>6</sup> de una red bayesiana puede ser construida automáticamente por, entre otros, el algoritmo CB (p. e., Mani et al., 1997), el algoritmo ELR (Greiner, Su, Shen, y Zhou, 2005; Greiner y Zhou, 2002; Shen, Su, Greiner, Musilek, y Cheng, 2003) o por el algoritmo EM (p. e., Cowell et al., 1999; Hugin Expert, 2004; Norsys Software, 2004; Oatley y Ewart, 2003). No obstante, el método clásico orientado a estimar los parámetros de una red bayesiana se basa en una aproximación maximo-verosimil o de *estimación basada en la frecuencia* que se podría expresar como:

$$p(x_i | x_{\pi(i)}) = \frac{n(x_i, x_{\pi(i)})}{n(x_{\pi(i)})}, \quad (3.5)$$

<sup>6</sup>Aunque, por claridad expositiva, se comience tratando la estimación paramétrica o cuantitativa de una red bayesiana; hay que tener en cuenta que lo primero que se establece al construir un modelo de red bayesiana es su dimensión cualitativa. Una vez que la estimación estructural ha sido hecha se estiman los parámetros de la red.

donde:

- $n(x_{\pi(i)})$  es el número de casos en la base de datos en que las variables  $X_{\pi(i)}$  toman el valor  $x_{\pi(i)}$ ,
- y  $n(x_i, x_{\pi(i)})$  es el número de casos en la base de datos en los que  $X_i = x_i$  y  $X_{\pi(i)} = x_{\pi(i)}$ .

Sin embargo, este modo de estimar los parámetros del modelo puede conducir a estimaciones *no definidas* (cuando alguna combinación de valores de las variables no está presente en la base de datos) y *sobreajustadas* que proucirían a parámetros sesgados en casos en los que ciertas combinaciones de variables están sub-representadas o sobre-representadas. Por ello, se introdujo un factor de corrección en la ecuación 3.5 basado en la Ley de la Sucesión de Laplace que corrige este defecto (M. E. Morales, 2006; Ng y Jordan, 2002):

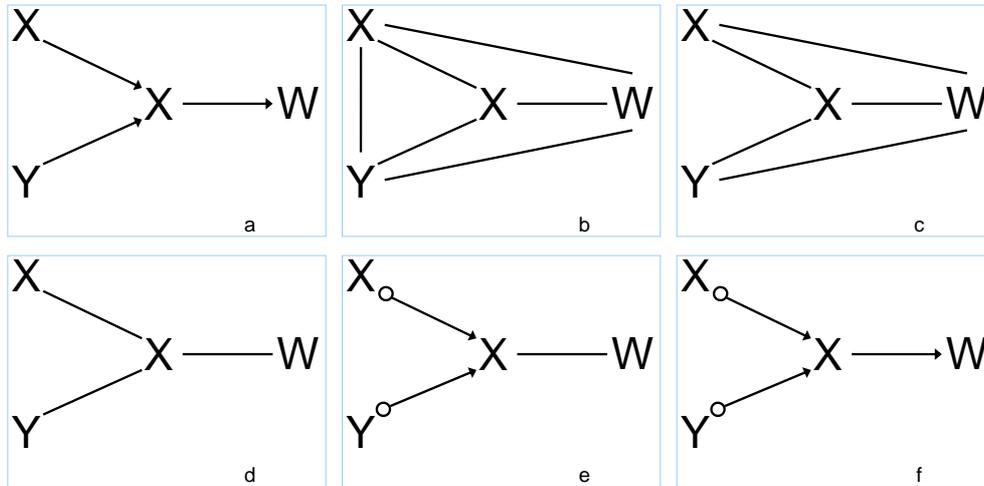
$$p(x_i|x_{\pi(i)}) = \frac{n(x_i, x_{\pi(i)}) + 1}{n(x_{\pi(i)}) + |X_i|}, \quad (3.6)$$

donde:  $|X_i|$  es el número de estados de la variable  $X_i$ .

### 3.3.2. Estimación estructural

Por otro lado, la dimensión cualitativa puede generarse con programas como COGITO™. Por ejemplo, Herskovits y Dagher (1997) proponen un procedimiento para desarrollar redes bayesianas a partir de bases de datos orientadas al cuidado de la salud. El algoritmo PC implementado en TETRAD (Cowell et al., 1999; Glymour, 2001; Gopnik et al., 2004; Scheines, Spirtes, Glymour, Meek, y Richardson, 2005; Spirtes et al., 2000) y el algoritmo K2 junto al algoritmo de búsqueda estocástica con vecindad variable (Céspedes, Rumí, Salmerón, y Soler, 2003; M. E. Morales, 2006) son ejemplos de este tipo de procedimientos.

El algoritmo PC implementado en TETRAD es especialmente relevante para el conocimiento psicológico por dos razones: 1) porque se construyó teniendo en cuenta los fundamentos del aprendizaje humano (Glymour, 2001, 2003), y 2) porque ha servido para generar modelos teóricos que dan una explicación al aprendizaje causal (p. e., Gopnik et al., 2004; Gopnik y Schulz, 2004).



**Figura 3.15:** Algoritmo PC usado en TETRAD. Adaptado de Gopnik et al. (2004).

TETRAD es un sistema de búsqueda de estructuras causales basado en la *constricción*<sup>7</sup> (*constraint-based method*) por medio de técnicas estadísticas estándares que testan las relaciones de dependencia e independencia condicional entre las variables de un modelo (Spirtes et al., 2000). Para ejemplificar el funcionamiento del algoritmo PC propondremos un ejemplo que consta de cinco pasos (Figura 3.15). Supongamos que partimos de una estructura gráfica determinada (Figura 3.15a), que *a priori* no conocemos, y recogemos un conjunto de datos para cada una de sus variables.

El primer paso que da TETRAD es crear un grafo no dirigido que une todas las variables (Figura 3.15b). A continuación se realizan tests de independencia entre cada par de variables y se eliminan las aristas que hay entre las variables supuestamente independientes (Figura 3.15c). Para cada par  $U, V$ , de variables conectados por una arista, y por cada variable  $T$  conectada por una arista a una o a ambas  $U, V$ , se comprueba si  $U$  y  $V$  son independientes dado  $T$ . En el caso de que se encuentre independencia se elimina la arista entre  $U$  y  $V$  (Figura 3.15d). Seguidamente, para cada trío de variables  $T, V, R$  semejante a  $T - V - R$  que no tienen arista entre  $T$  y  $R$ , orientar como  $T \circ \rightarrow V \leftarrow \circ R$  si y solamente si  $V$  no fue condicionada sobre la eliminación de la arista  $T - R$  (Figura 3.15e). Las marcas « $\circ$ » indican que el procedimiento no puede determinar si la asociación es

<sup>7</sup>Los métodos basados en la constricción llevan a cabo un procedimiento de estimación que restringe cierto tipo de resultados de acuerdo a un conjunto de postulados teóricos.

producida por una flecha causal desde una variable a otra o por una causa común no observada entre las variables. Finalmente, para cada trío de variables  $T, V, R$  en que  $T$  tenga una arista con punta dirigida hacia  $V$  y  $V - R$ , y  $T$  no tenga arista conectándola a  $R$ , orientar  $V - R$  como  $V \rightarrow R$ . El resultado final del proceso aparece en la Figura 3.15f.

Sin embargo, no siempre es posible construir una red bayesiana a partir de datos porque no siempre hay una base de datos disponible que recoja las variables de interés o porque algunas mediciones pueden ser éticamente inviables (p. e. una mutación genética). Por esta razón cobra sentido el desarrollo de redes bayesianas a partir del conocimiento de expertos.

Nadkarni y Shenoy (2004) siguieron cuatro fases para desarrollar un mapa causal bayesiano a partir del conocimiento de un experto: 1) elicitación de los datos, 2) derivación del mapa causal, 3) modificación del mapa causal para obtener un mapa causal bayesiano, y 4) estimación de los parámetros del mapa.

En la primera fase se capturó el conocimiento que tiene un experto sobre el área que domina. Tradicionalmente se han usado dos tipos de técnicas para extraer información del conocimiento experto: estructuradas y no estructuradas. Las *técnicas estructuradas* consisten en la presentación de conceptos predefinidos sobre los que el experto ha de indicar la dirección y el signo (positivo o negativo) de la relación que se establece entre ellos. Esta aproximación es útil cuando se pretenden validar o confirmar modelos y no es preferible cuando los dominios de conocimiento no están bien definidos. Por su parte, en las *técnicas no estructuradas*, lo que se pretende es extraer información de un dominio de conocimiento nuevo o no familiar. Para ello se van planteando preguntas del estilo: *¿cuáles son los factores relevantes en la decisión?*

En el trabajo de Nadkarni y Shenoy (2004) se lleva a cabo un procedimiento híbrido, donde se aprovecha la potencia exploradora de las técnicas no estructuradas y el poder confirmatorio de las técnicas estructuradas. Para ello someten a los expertos a una entrevista semiestructurada que da lugar a un conjunto narrativo del que se pueden extraer, por medio del análisis textual, las variables y las relaciones que integrarán el mapa causal. Por ejemplo, imaginemos que estamos intentando construir una red para el diagnóstico de los trastornos psicóticos. Una posible entrevista podría tener la siguiente forma:

**Pregunta 1:** ¿Qué factores considerarías para decidir si una persona sufre un

trastorno psicótico?

**Respuesta del experto clínico:** Consideraría el **grado de deterioro funcional**;... el **tipo, la forma y la cantidad de alucinaciones**; y, por supuesto, la **duración**...

**Pregunta 2:** Mencionaste el **grado de deterioro funcional**. ¿Qué factores específicos determinan el grado de deterioro funcional?

**Respuesta del experto clínico:** Las **ideas delirantes** y el **grado en que el comportamiento es desorganizado** pueden afectar al trastorno... Creo que las **concentraciones de ciertas sustancias en el cerebro** también pueden afectar al deterioro funcional de los trastornos psicóticos.

**Pregunta 3:** Mencionaste las **ideas delirantes**. ¿Qué tipo de ideas delirantes son las más importantes para ti en el diagnóstico de un trastorno psicótico?

**Respuesta del experto clínico:** ...las **de persecución y las autorreferenciales** son las más importantes... Bueno, las **ideas delirantes de control** son también bastante importantes.

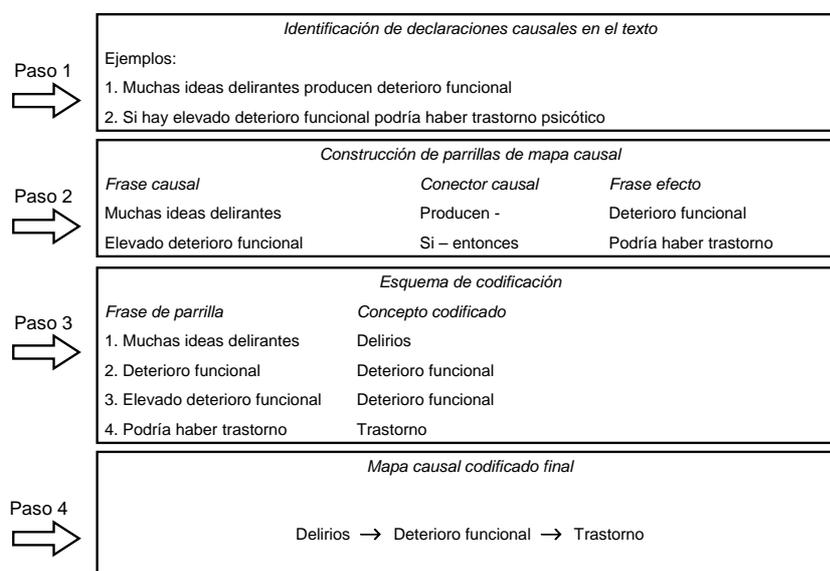
La siguiente fase sería derivar los mapas causales. En esta fase se obtienen los mapas causales a partir de los textos obtenidos en las narraciones de la entrevista (Figura 3.16). Para ello, lo primero que hay que hacer es localizar las declaraciones causales, esto es, referencias explícitas por parte del entrevistado a relaciones causales. Para ello sería importante desarrollar algún tipo de guía o diccionario de declaraciones causales donde se defina su significado y reglas para reconocer conectores causales.

El siguiente paso sería construir parrillas de mapas causales. Así, una vez que se han aislado las declaraciones causales, se disgregan en frases causales conectores causales y frases efecto. A continuación se ha de desarrollar un esquema de codificación en el que se extraiga la información relevante de las frases contenidas en la parrilla de mapa causal. Esto es llevado a cabo por medio de un proceso conocido como *filtrado* o *agregación*. Lo que se pretende en este punto es eliminar la redundancia de palabras y extraer los conceptos clave que pasarán a formar parte del mapa causal. En esta fase de construcción es muy importante contar con la colaboración del experto ya que puede aportar información muy importante para eliminar inconsistencias y errores.

Por último, se han de transformar las declaraciones causales en una red que capture las relaciones causales entre los conceptos claves definidos en el esquema de codificación.

En la tercera fase hay que transformar el mapa causal en un mapa bayesiano.

## Fundamentos de las redes bayesianas



**Figura 3.16:** Fases para la construcción de un mapa causal. 1. Identificación de manifestaciones causales. 2. Construcción de parrillas. 3. Determinación del esquema de codificación. 4. Trazado del mapa causal.

Aunque un mapa causal capta el entramado de relaciones causales que se presentan en el ambiente, podría no ser informativo a la hora de realizar inferencias, lo que nos conduciría a errores. Ello es debido a que los mapas causales son mapas de dependencia (mapa-D), mientras que las redes bayesianas son mapas de independencia (mapa-I) (I. Martínez y Rodríguez, 2003; Nadkarni y Shenoy, 2001, 2004; Spirtes et al., 2000). Técnicamente, *un grafo  $\mathcal{G}$  se dice que es un **mapa de dependencia** (mapa-D) de un modelo de dependencia  $\mathcal{M}$  si cada relación de dependencia deducida de  $\mathcal{G}$  es también deducida de  $\mathcal{M}$ , y un grafo  $\mathcal{G}$  se dice que es un **mapa de independencia** (mapa-I) de un modelo  $\mathcal{M}$  si cada relación de independencia deducida de  $\mathcal{G}$  también puede ser deducida de  $\mathcal{M}$ . En la práctica, dos conceptos unidos por un arco en un mapa-D tendrían relación; sin embargo, la ausencia de un arco que une dos conceptos no implica, necesariamente, independencia. Por el contrario, en un mapa-I, la ausencia de una unión entre dos nodos implica que estos elementos son condicionalmente independientes dadas las demás variables del modelo.*

Por tanto, cuando transformamos el mapa causal en un mapa causal bayesiano obtenemos una estructura que tiene las propiedades de las redes bayesianas y los mapas causales. Se consigue lo que técnicamente se conoce como **mapa perfecto**

(I. Martínez y Rodríguez, 2003; Nadkarni y Shenoy, 2001, 2004; Spirtes et al., 2000).

Para obtener la estructura bayesiana definitiva los autores proponen dos métodos: una entrevista estructurada o matrices adyacentes. En la entrevista estructurada se presentan preguntas donde se relacionan las variables del modelo de dos en dos. La tarea del experto es señalar el sentido de la relación entre cada par de variables y el signo de la misma (o sea, si es directa o inversa). En las matrices adyacentes se presentan todas las variables del modelo en la primera fila y en la primera columna de una tabla. De esta manera tendríamos una matriz en que las casillas de la diagonal representarían la convergencia de una misma variable y el resto de las celdas las posibles combinaciones causales entre los pares de variables. La tarea del experto en esta ocasión es rellenar las celdas con los signos «+» (relación positiva), «-» (relación negativa) o «0» (ausencia de relación).

Hay que poner especial atención para eliminar inconsistencias estructurales que hayan podido sobrevenir en la construcción de la red. Hay que, por ejemplo, distinguir claramente lo que son relaciones causales directas de las indirectas y eliminar relaciones circulares. A este proceso de modificación de elementos estructurales se le denomina *poda* (Edwards, 1998) y puede consistir en eliminar arcos, nodos, crear nodos o cambiar sentidos de influencia.

La última fase de construcción propuesta por Nadkarni y Shenoy consiste en estimar los parámetros de la red bayesiana; esto es, las probabilidades asociadas a cada variable de la red. El proceso de estimación de probabilidades es un proceso subjetivo en el que el experto estima la probabilidad de cada variable del modelo condicionada al resto de variables. Este proceso puede descomponerse en dos pasos: a) determinación del número de estados de cada variable y, b) asignación de probabilidades a cada uno de los posibles estados de la variable.

La generación de distribuciones de probabilidad por expertos es un elemento fundamental de la determinación cuantitativa de un sistema experto y tenemos que tener en cuenta que las personas no solemos hacer estimaciones de probabilidad que se ajustan fehacientemente a las reglas de la probabilidad (p. e., Kahneman, 2003; Kahneman, Slovic, y Tversky, 1982; Kahneman y Tversky, 1973; Tversky y Kahneman, 1974). Por lo general, los sesgos que guían nuestro razonamiento tienden a distorsionar nuestras estimaciones de probabilidad de una manera predecible y replicable que pueden generar inconsistencias en las tablas

de probabilidad condicional de los sistemas expertos bayesianos.

Das (2004) ha desarrollado un procedimiento orientado a *poblar*<sup>8</sup> las TPCs de una red bayesiana que tiene en cuenta los sesgos de estimación de probabilidades en que puede incurrir un experto. La técnica consiste en pedirle al experto que genere valores de probabilidad para configuraciones determinadas de las variables del modelo (aquellas que son más frecuentes, o más representativas/accesibles en términos heurísticos, en la experiencia del experto). A partir de ellas se obtienen, a través de un algoritmo que pondera la relación entre las variables, las probabilidades de las TPCs. El problema del aumento exponencial del número de variables de un modelo se simplifica a un aumento lineal del número de probabilidades en función del número de nodos de la red. El inconveniente de esta técnica es que las variables del modelo tienen que tener el mismo número de estados.

### 3.4. Bondad de ajuste

Los estadísticos de bondad de ajuste para una red bayesiana son, en algunos casos, los mismos que han sido desarrollados para otros modelos probabilísticos. Por ejemplo, la tabla de confusión o matriz de clasificación vista para el caso de la regresión logística también es aplicable al análisis de la bondad de ajuste de una red bayesiana. A su vez, los estadísticos de sensibilidad, especificidad, valor predictivo positivo, valor predicativo negativo, proporción de falsos positivos, así como la proporción de falsos negativos también se pueden utilizar para estimar la bondad predictiva de una red bayesiana.

Por otro lado, existen tres parámetros (la pérdida logarítmica, la pérdida cuadrática y la compensación esférica) que se utilizan para comprobar el grado en que la red es consistente con un conjunto de datos que no ha sido utilizado para estimar los parámetros ni la estructura de la red. El uso de estos estadísticos cobra especial relevancia en estudios de validación cruzada (p. e., en J. García, López, Cano, Gea, y De la Fuente, 2006; J. García, López, De la Fuente, Cano, y Gea, 2007; J. López, De la Fuente, García, y Ortega, 2008; J. López, Ruiz-Ruano, y García, 2008a).

---

<sup>8</sup>El objetivo del procedimiento sería completar las tablas en función de un conjunto de parámetros estimados por el experto. En este sentido el proceso podría definirse como un procedimiento de *imputación*.

La *pérdida logarítmica* ( $\lambda$ ) que oscila entre 0 e infinito, donde 0 indica el mejor ajuste, vendría definida por la siguiente ecuación:

$$\lambda = \frac{1}{N} \sum_{i=1}^N -\ln(P_c), \quad (3.7)$$

donde

- $P_c$  es la probabilidad predicha para el estado correcto del nodo,
- y  $N$  representa el número de casos.

La *pérdida cuadrática* ( $\varsigma$ ) (también conocida como *brier score*) que puede estar comprendida entre 0 y 2, indicando 0 la mejor ejecución de la red, se define por la siguiente expresión:

$$\varsigma = \frac{1}{N} \sum_{i=1}^N \left( 1 - 2 \times P_c + \sum_{j=1}^n P_j^2 \right), \quad (3.8)$$

donde

- $P_j$  es la probabilidad predicha para el estado  $j$
- y  $n$  es el número de estados del nodo.

Por último, la *compensación esférica* ( $\epsilon$ ) que puede variar entre 0 y 1, donde 1 se refiere a la mejor ejecución, queda determinada por la siguiente fórmula:

$$\epsilon = \frac{1}{N} \sum_{i=1}^N \left( \frac{P_c}{\sqrt{\sum_{j=1}^n P_j^2}} \right). \quad (3.9)$$

### 3.5. Ventajas de las redes bayesianas

La ventaja fundamental de las redes bayesianas es representar, al unísono, la dimensión cuantitativa y cualitativa de un modelo matemático (p. e., Edwards, 1998; Heckerman, 1995). Esta ventaja ha sido propiciada por el aumento de la potencia de cómputo que han experimentado las computadoras en los últimos años y gracias al desarrollo de algoritmos de propagación de probabilidades que optimizan los recursos (Heckerman, Mamdani, y Wellman, 1995).

Una de las ventajas más importante que tienen las redes bayesianas es que pueden trabajar con datos perdidos (p. e., destacado por Nadkarni y Shenoy, 2004; Jansen et al., 2003). Esta cualidad es muy importante en el contexto de la investigación aplicada porque, en la mayoría de los casos, en nuestras investigaciones no contamos con una base de datos «pulcra» y muchos de los análisis que realizamos sobre ella pueden no ser relevantes por la presencia de datos perdidos. Permiten también reducir el sobreajuste de los datos (Heckerman, 1995) y combinar el conocimiento previo que tenemos respecto al problema de estudio con los datos experimentales (Nadkarni y Shenoy, 2004). A su vez, F. Ruiz, García, y Pérez (2005) destacan que las redes bayesianas no suponen un determinado modelo subyacente en los datos, por lo que no necesitaríamos justificar ningún supuesto de partida en nuestros datos para realizar los análisis (como normalidad u homocedasticidad). Por su parte, S. M. Lee, Abbott, y Johantgen (2005) destacan la facilidad que tienen las redes bayesianas para identificar efectos de interacción y para modelar relaciones no lineales entre variables.

Como se ha comentado anteriormente, las redes bayesianas son herramientas que permiten determinar la estructura causal de un conjunto de datos (p. e., Céspedes et al., 2003; Heckerman, 1995; Hugin Expert, 2004; Norsys Software, 2004; Spirtes et al., 2000). Así, a partir de un fichero de datos, las redes bayesianas son capaces de construir un modelo gráfico que represente las relaciones probabilísticas contenidas en la matriz *variables*  $\times$  *observaciones*.

Existen ventajas prácticas para los usuarios de las redes bayesianas. Por ejemplo, el sistema representa todo el conocimiento en un único formato (probabilístico y gráfico) lo que hace sencillas las interpretaciones (Huete, 1998). Por otro lado, la asignación de probabilidades, o su generación, es mucho más asequible cuando se pretende representar el conocimiento de un experto en una red bayesiana (Edwards, 1998; I. Martínez y Rodríguez, 2003).

En el contexto de la inferencia, las redes bayesianas permiten realizar inferencias bidireccionales; esto es, desde los efectos a las causas y desde las causas a los efectos. Y lo que es más interesante, permiten llevar a cabo inferencias abductivas; o sea, encontrar la mejor explicación para un conjunto de datos (Gámez, 1998; Huete, 1998). Sin embargo, la ventaja más importante en la inferencia con este tipo de herramientas está en la posibilidad que tienen de realizar *computaciones locales* (p. e., Pearl, 2001; Xiang, 2002). Esta propiedad permite que la

actualización de probabilidades se lleve a cabo eficientemente sin tener que calcular todas las posibles combinaciones entre todos los niveles de las variables (que supondría un incremento exponencial de los cálculos a medida que aumenta el número de variables del modelo) cuando se incorpora determinado conocimiento a la estructura gráfica.

### **3.6. *Software***

En esta sección se van a presentar algunos de los programas informáticos más importantes que trabajan con redes bayesianas. Esta sección tiene sentido dado que el uso de redes bayesianas no está ampliamente extendido en el contexto de la investigación social, y particularmente en el campo de la Psicología. De todas formas, para encontrar un catalogo más detallado de programas informáticos diseñados para diseñar y usar redes bayesianas se recomienda el trabajo de Cowell et al. (1999) o de Korb y Nicholson (2004).

En primer lugar destacaremos las características de programas, tanto comerciales como libres, que permiten trabajar con modelos gráficos probabilísticos. Para terminar, se comentarán algunos de los procedimientos implementados en otros entornos estadísticos que, aun no habiendo sido concebidos para trabajar con este tipo de herramientas, ofrecen la posibilidad de crear y usar redes bayesianas.

#### **3.6.1. TETRAD**

El objetivo de este programa es proporcionar una herramienta de cálculo sofisticado a los usuarios con nociones básicas de estadística sin conocimiento de programación informática. Aunque no pretende suplantar a otros programas comerciales (como LISREL, EQS u otros desarrollados para trabajar con modelos gráficos estadísticos), TETRAD usa algunos de sus procedimientos e incluye otros diferentes. La orientación a la búsqueda («exploración», «descubrimiento») de estructuras causales, con o sin variables latentes, es el rasgo que lo distingue de otros programas similares. TETRAD describe los modelos causales en tres partes o etapas distintas (Scheines et al., 2005): gráficamente, representando un grafo causal especificando relaciones causales hipotéticas entre variables; especificando

una familia de distribuciones de probabilidad y tipos de parámetros asociados con la estructura gráfica; y una especificación numérica de estos parámetros. El programa ha sido desarrollado durante varios años al amparo de la *National Aeronautics and Space Administration* y de la *Office of Naval Research*.

TETRAD no es un programa de redes bayesianas *per se*, ya que no consta de una *interface* gráfica donde aparezca, al unísono, la estructura gráfica y cuantitativa de un modelo. Además, no tiene la capacidad de trabajar con datos perdidos como hacen las redes bayesianas estándares. Sin embargo, TETRAD ha sido diseñado para que sea capaz de construir una red bayesiana que pueda usarse como sistema experto a partir de una base de datos (Spirtes et al., 2000). Así pues, TETRAD es capaz de extraer la estructura gráfica causal subyacente en una matriz de datos para que pueda ser implementada en una red bayesiana. Todo ello es posible gracias al algoritmo de aprendizaje automático que se ha descrito anteriormente para la construcción de redes bayesianas. TETRAD también puede usarse para trabajar con una red bayesiana en la predicción y clasificación. TETRAD puede descargarse de internet sin ningún coste ni limitación desde la dirección electrónica <http://www.phil.cmu.edu/projects/tetrad/index.html>.

### 3.6.2. NETICA™

Netica™ es un producto desarrollado por *Norsys Software Corporation*, una compañía canadiense afincada en Vancouver dedicada al desarrollo de programas informáticos orientados a resolución de problemas mediante el uso de redes de creencia bayesianas y diagramas de influencia ([www.norsys.com](http://www.norsys.com)). Norsys trabaja con una gran variedad de organismos públicos y privados dedicados a la investigación, educación o a la comercialización de productos. Norsys Software trabaja tanto con clientes en el plano comercial (entre los que podemos destacar Toshiba, US Environmental Protection Agency (EPA), Siemens Medical Solutions, NASA, US Air Force, Exxon - Mobil, CRC for Meteorology, Expert Microsystems, Barclays Bank, Nokia y el Educational Testing Service) como en el plano académico donde colabora con más de 150 organizaciones educativas que incluyen universidades distribuidas por todo el mundo.

Netica™ es capaz de representar tanto redes bayesianas como diagramas de influencia basados en la teoría clásica de la maximización de la utilidad esperada. Permite la creación *rápida* de tablas de probabilidad condicional a partir de

ficheros de datos y da la posibilidad de realizar la operación en sentido inverso; esto es, a partir de una red bayesiana simular un conjunto de datos que se ajusten al modelo gráfico y probabilístico contenidos en la red.

Este programa ofrece la posibilidad de evaluar la sensibilidad relativa de un nodo en relación a las evidencias que podamos encontrar en otras variables del modelo. Además, permite evaluar el ajuste del modelo con un conjunto de datos procedentes de la realidad. Otras opciones que ofrece Netica™ son el evaluar la conveniencia de postular variables ocultas entre conjuntos de variables, usar el algoritmo EM y aplicar el gradiente de descenso en el aprendizaje como en una red neural artificial.

### 3.6.3. ELVIRA

ELVIRA (el nombre oficial es *Entorno de Desarrollo para Modelos Gráficos Probabilísticos*) es un programa experimental que funciona en entorno Java y que está siendo desarrollado por el *Grupo de Investigación del Tratamiento de la Incertidumbre en Inteligencia Artificial* (Universidad de Granada), el *Grupo de Sistemas Inteligentes* (Universidad del País Vasco), el *Laboratorio de Sistemas Inteligentes* (Universidad de Castilla La Mancha), el *Departamento de Inteligencia Artificial* (Universidad Nacional de Educación a Distancia), el *Departamento de Estadística y Matemática Aplicada* (Universidad de Almería) y el *Decision Support System Group* (Aalborg University, Dinamarca).

ELVIRA, al igual que Netica™, permite representar redes bayesianas y diagramas de influencia. Sin embargo, tiene una propiedad que lo hace bastante atractivo: la representación cualitativa de la fuerza asociativa entre cada par de variables. Hay dos aspectos en ELVIRA que representan el grado de asociación entre variables: el color y el grosor del arco. Mediante el color del arco se representa el sentido o tipo de relación causal que se establece entre las variables conectadas por el arco; de tal modo que: las flechas que son de color rojo denotan una relación positiva, el color azul indica una relación negativa, el color púrpura señala una relación indeterminada y cuando no hay relación entre los nodos la flecha será negra. A su vez, el grosor del arco indica el grado o la intensidad de la relación que se establece entre dos nodos de tal modo que las flechas más gruesas indican una relación más fuerte.

Otra ventaja que presenta este programa es la posibilidad de marcar varias

evidencias en una misma variable, lo que facilita los procesos de inferencia y optimiza la toma de decisiones en algunos casos. Por último, como buen sistema experto (Adarraga y Zaccagnini, 1988), ELVIRA facilita una explicación verbal para el efecto que producen las evidencias que mostremos a la red en términos de las distribuciones posteriores que genera.

### 3.6.4. MBNTx

El *Bayesian Network Editor and Tool Kit* (MBNTx) es un conjunto de herramientas diseñadas por *Microsoft* para construir, evaluar y utilizar redes bayesianas en diferentes contextos. Microsoft lleva algún tiempo usando e investigando esta tecnología; de hecho, la ayuda o asistente de *Microsoft Word* que aparece en el paquete *Office 97* caricaturizada con un clip (o cualquier otro personaje) es una aplicación de las redes bayesianas en la resolución de problemas.

El MBNTx está orientado principalmente al diagnóstico y a la resolución de problemas. El programa usa el análisis de costos-beneficios para ir proporcionando recomendaciones dinámicamente. Además, si no se le da prioridades sobre los costes, el programa ordena automáticamente las recomendaciones. A su vez, este programa tiene sus propias herramientas estadísticas y algoritmos de aprendizaje automático de redes bayesianas.

Como contrapartida hay que destacar que el programa sólo trabaja con variables discretas. Por otro lado, aunque permite manipular información referente a variables ocultas por medio de modelos de Markov, estos procedimientos se muestran ineficientes en modelos que contienen muchos nodos (Kadie et al., 2001).

### 3.6.5. ERGO™

*Noetic Systems Incorporated* ([www.noeticsystems.com](http://www.noeticsystems.com)) se fundó en 1989 con el objetivo de transferir a la sociedad tecnologías de razonamiento probabilístico basadas en redes de creencias bayesianas. Su objetivo era superar las dificultades en el tratamiento de la incertidumbre que mostraron los sistemas expertos tradicionales basados en reglas.

ERGO™ es el primero de una serie de productos que esta compañía pone a disposición de los decisores que quieran beneficiarse de los algoritmos de propagación de probabilidades y de las prestaciones que ofrecen las redes bayesianas. El

diseño de ERGO™ está pensado para que la asignación de probabilidades sea lo menos tedioso posible. Podríamos decir que es un programa orientado al usuario.

ERGO™ usa el algoritmo optimizado de Lauritzen y Spiegelhalter para la propagación de probabilidades, que combinado con la teoría de grafos utilizada en su diseño garantizan la consistencia interna de su funcionamiento. Aunque no contiene ningún procedimiento de aprendizaje automático de parámetros o estructuras, su uso combinado con otros programas puede suplir esta ausencia. Por ejemplo, Herskovits y Dagher (1997) utilizaron COGITO™ para obtener una estructura causal a partir de diferentes bases de datos y posteriormente trasladaron esta información a una red bayesiana creada con ERGO™.

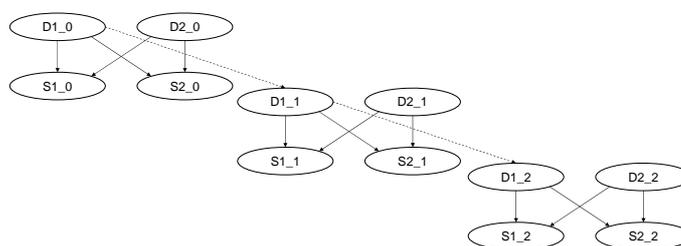
Una ventaja de este programa es que puede vincularse a formularios de hipertexto, de tal manera que rellenando las preguntas de un formulario la red bayesiana que opera sobre las respuestas del mismo nos daría la respuesta (en términos probabilísticos) que andamos buscando. También puede incorporarse a bases de datos de 4ª dimensión en computadoras *Macintosh*, o en aplicaciones *Visual Basic* en una computadora personal.

### 3.6.6. HUGIN™

El programa HUGIN™ ha sido desarrollado por la compañía danesa *Hugin Expert A/S* ([www.hugin.com](http://www.hugin.com)). El nacimiento de esta compañía se produjo en 1989 asociado al desarrollo de técnicas orientadas al diagnóstico de enfermedades musculares y nerviosas. En 1998 *Hewlett-Packard* adquirió el 45 % de las acciones de Hugin Expert con motivo de su expansión comercial por Dinamarca. Hoy día HUGIN™ es uno de los programas de redes de redes bayesianas más famosos tanto en el entorno comercial como en el ámbito académico.

HUGIN™ dispone de las características exigibles a un buen programa de redes bayesianas (tolerancia a los datos perdidos, representación de diagramas de influencia, editor inteligible de tablas de probabilidad condicional, procedimientos de aprendizaje automático de probabilidades, módulo de simulación de casos, etc.). Sin embargo, tiene dos características que lo hacen destacar sobre los demás.

En primer lugar HUGIN™ posee un procedimiento de aprendizaje estructural (Hugin Expert, 2004). Esto quiere decir que HUGIN™, al igual que TETRAD, puede obtener una estructura cualitativa de red bayesiana a partir de un conjunto de datos por medio del algoritmo PC. Por medio de este algoritmo se obtiene un



**Figura 3.17:** Representación de una red orientada a objetos con 4 nodos. Adaptada de Hugin Expert (2004).

grafo consistente con las relaciones de independencia condicional que se establecen en una matriz de datos usada como *input*. A continuación, los parámetros de cada nodo pueden obtenerse por medio del algoritmo EM que es facilitado por el programa.

Otra característica que hace destacar a HUGIN™ sobre otras herramientas que realizan trabajos similares es la posibilidad de crear redes bayesianas orientadas a objetos. La *programación orientada a objetos* se ha derivado de lo que se ha venido a denominar la cuarta generación de programación informática que considera un programa como un conjunto de objetos abstractos que interaccionan entre sí. En el contexto que nos ocupa, una red bayesiana orientada a objetos sería un modelo gráfico donde uno o varios nodos del modelo podrían ser entendidos como otras redes bayesianas. Esto es, imagine una red bayesiana con cuatro variables. Ahora piense que uno de los nodos representa a otro conjunto de nodos que forman una red bayesiana de otros cuatro vértices, donde uno de esos nodos (nuevamente) puede entenderse como una red bayesiana integrada por cuatro variables. En la Figura 3.17 aparece la representación esquemática de una red bayesiana orientada a objetos con esta estructura. Los arcos discontinuos indican que un nodo particular representa un grafo. Este tipo de redes bayesianas son oportunas para representar relaciones causales que implican, por ejemplo, series temporales de observaciones.

### 3.6.7. Otros programas

Además de estos programas específicos orientados a trabajar con redes bayesianas existen compañías que han implementado ciertos procedimientos en sus productos que permiten utilizar algunas opciones para trabajar con redes de este tipo. Por

ejemplo, SPSS y Recognition Systems (1997) implementaron un procedimiento de entrenamiento y validación de redes bayesianas en *Neural Connection*<sup>®</sup>.

En *Neural Connection*<sup>®</sup> las redes bayesianas son tratadas como una adaptación del *perceptrón multicapa* a la estadística bayesiana y están centradas en la estimación de los parámetros de la red. No obstante, mientras que en el perceptrón multicapa se busca la minimización del error cuadrático dado por la ecuación:

$$E_D = \sum_x E_x = \frac{1}{2} \sum_x \sum_j (t_{xj} - a_{xj})^2,$$

donde

- $x$  es un ejemplo del conjunto de datos,
- $j$  es una salida del modelo,
- $t_{xj}$  es un valor objetivo,
- y  $a_{xj}$  es la salida de la red;

el procedimiento de aprendizaje implementado en este programa para redes bayesianas tiende a minimizar una *función de penalización* dada por:

$$C = \beta E_D + \alpha E_W.$$

Donde  $E_W$  es una función ponderada de penalización,  $\beta$  y  $\alpha$  son *hiperparámetros* y el objetivo del entrenamiento es estimar valores apropiados para estos parámetros a partir del conjunto de entrenamiento. El algoritmo de aprendizaje aplica el procedimiento del error de retropropagación hacia atrás (*backpropagation*) para minimizar la función de coste de la red.

Otro programa comercial que permite trabajar con redes bayesianas es Matlab<sup>®</sup>. Matlab<sup>®</sup> es la abreviatura de *Matrix Laboratory* y es un programa de matemáticas creado por *The MathWorks* en 1984. La primera versión del programa surgió con la idea de integrar paquetes de subrutinas utilizadas en Fortran para trabajar álgebra lineal y análisis numérico. Matlab<sup>®</sup> utiliza un lenguaje de programación propio creado en 1970 como alternativa a Fortran para trabajar con el *software* de matrices *LINPACK* y *EISPACK*. Aunque se trata de un programa comer-

cial, la comunidad científica desarrolla *paquetes*<sup>9</sup> que se distribuyen libremente en internet desde la página oficial de *The MathWorks* y que permiten ampliar las prestaciones originales del programa.

El *Bayes Net Toolbox* (BNT) para Matlab<sup>®</sup> es uno de estos paquetes diseñado para trabajar con redes bayesianas en este entorno. BNT ha sido desarrollado por Kevin P. Murphy desde el *Department of Computer Science and Statistic* en la *University of British Columbia* (Vancouver, Canadá). Este paquete soporta inferencia exacta y aproximada, aprendizaje estructural y paramétrico, permite trabajar con diferentes tipos de distribuciones (multinomiales, gaussianas, etc.) y proporciona una herramienta estructurada que puede manejar *Clasificadores Naive de Bayes, Hidden, Coupled, Input-output* o *Factorial Markov Models*, entre otros.

Por último, existe un lenguaje y entorno de programación estadística libre<sup>10</sup> llamado R que permite trabajar con redes bayesianas usando el paquete *deal*. R es la versión libre de S que fue creado en los *Bell Laboratories* (formalmente llamados AT&T, y ahora *Lucent Technologies*) por John Chambers y colaboradores. Proporciona una amplia variedad de técnicas estadísticas (lineales, no-lineales, clásicas, bayesianas, etc.) y gráficas. Una de sus mayores ventajas es su versatilidad y sus posibilidades de ampliación; a su vez, permite trabajar simultáneamente con lenguajes de programación de cuarta generación como C o C++.

El paquete *deal* es un software de disponibilidad libre desarrollado por Susanne G. Bøttcher y Claus Dethlefsen con el apoyo de *Nordisk A/S*. Los autores han restringido el programa para que opere con redes gaussianas y permite tanto el aprendizaje estructural como el paramétrico. Además, incorpora un módulo de simulación de bases de datos y permite leer y guardar archivos procedentes de HUGIN<sup>™</sup>. Tiene implementados una serie de procedimientos orientados a evaluar el ajuste de una red a un conjunto de datos y permite usar tanto variables continuas como discretas.

---

<sup>9</sup>Líneas de código fuente que permiten controlar un programa informático para realizar tareas concretas.

<sup>10</sup>Referido a las Licencias Públicas Generales en el contexto del proyecto GNU (*GNU is Not Unix*).

## 3.7. El rol de las redes bayesianas en Psicología

### 3.7.1. La Psicología y las redes bayesianas

En el capítulo dedicado a la regresión logística se ha obviado una sección análoga a esta porque la regresión logística es una técnica más consolidada en Psicología frente a lo que ocurre con las redes bayesianas. Así pues, las redes bayesianas aportan una doble ventaja, practico-teórica, al quehacer científico de la Psicología (J. López, García, y De la Fuente, 2005; J. López et al., 2007). En primer lugar, dado que son herramientas de modelado estadístico, permiten ejecutar un tipo de análisis de datos que no han sido tratados en esta disciplina. Las redes bayesianas podrían, en este sentido, servir como herramienta de contraste de hipótesis referidas a modelos empíricos. Por otro lado, las redes bayesianas podrían servir para clarificar algunas cuestiones teóricas de la disciplina psicológica.

Ward Edwards (1998) destaca que las redes bayesianas son importantes para la Psicología en dos sentidos: económica y científicamente. A nivel económico las redes bayesianas podrían generar un mercado de «elicitación de probabilidades» orientado al desarrollo de sistemas expertos donde la Psicología podría jugar un papel fundamental. La tarea a realizar por la Psicología en este sentido sería orientar a los expertos sobre como hacer juicios de probabilidad apropiados intentando sortear los sesgos en que solemos incurrir las personas cuando realizamos tareas de este tipo. En el plano científico las redes bayesianas no pueden perderse de vista si la Psicología pretende conocer los mecanismos por los cuales las personas evaluamos, decidimos y realizamos inferencias; ya que pueden servir de referencia analítica y teórica en el sentido que ya se ha sugerido.

Sin embargo, la propuesta más llamativa del uso de las redes bayesianas en Psicología proviene de la mano de Clark Glymour (2001) en su libro *The mind's arrows. Bayes nets and graphical causal models in psychology*. La obra de Glymour gira en torno al fenómeno de la causalidad y al proceso de construcción de conocimiento por medio de la observación y la manipulación de nuestro entorno. El aprendizaje causal es punto arquimédico con el que inicia su trabajo, proponiendo que en nuestra interacción con el entorno aplicamos una serie de procesos (análogos a los algoritmos desarrollados para estudiar las relaciones causales estadísticas) que conducen a una representación interna de la estructura causal. Sin embargo, su propuesta no se limita al desarrollo ontogenético del aprendizaje

causal humano, sino que traslada esta filosofía al campo de la Neuropsicología (reentendiendo los diagramas de flujo de la Psicología Cognitiva clásica en un sentido bayesiano) o a la Psicometría (criticando el análisis factorial y el análisis de regresión como métodos fiables para extraer el número correcto de variables latentes y para representar la estructura causal subyacente).

La propuesta teórica de Glymour ha sido criticada por ser demasiado especulativa y por carecer de una sólida evidencia experimental (Borsboom, 2002). En primer lugar, los pocos experimentos usados para justificar sus razonamientos parecen seleccionados adrede para soportar su teoría y los resultados son interpretados sobre consideraciones *post hoc*. Por otro lado, Glymour parece no compensar equitativamente los *pros* y los *contras* de su propuesta. Por ejemplo, si su argumentación es correcta y cada uno de nosotros llevamos implementados en nuestro cerebro un sistema que nos permite «desenmascarar» la estructura causal del ambiente en términos de relaciones de dependencia e independencia probabilística, ¿cómo es posible que el aprendizaje formal de estos conceptos sea tan difícil?

Estas críticas serían consistentes con los trabajos clásicos que informan de dificultades en la estimación y/o comprensión de probabilidades condicionales en términos causales (p. e., Kahneman, 2003; Kahneman y Tversky, 1973; Tversky y Kahneman, 1974, 1982). Sin embargo, este hecho podría deberse a los formatos de presentación de la información o a artefactos metodológicos. Así pues, la presentación de información probabilística de una manera más comprensiva para los participantes experimentales podría mejorar su aprendizaje/razonamiento y acercarlo más a lo establecido por la estadística (p. e., Gigerenzer, 1996; Seldmeier y Gigerenzer, 2001).

En el resto de esta sección se destacará la utilización de este tipo de herramientas en la construcción de Sistemas de Tutorización Inteligente orientados a la enseñanza y la medición del aprendizaje. A continuación, se abordará el aspecto que más trabajos de investigación está generando actualmente como consecuencia del acercamiento de las redes bayesianas a la Psicología: el aprendizaje causal. Por último, se verá la aproximación que se ha hecho desde la Psicología Forense a estas herramientas.

### 3.7.2. Aplicaciones en Psicología

#### Sistemas de Tutorización Inteligente

El desarrollo de las ciencias de la computación, el abaratamiento de programas y de la maquinaria informática están dando lugar a una revolución educativa (Urretavizcaya, 2001). Hoy en día existen diferentes sistemas informáticos que permiten tutorizar o formar a personas a través de internet (p. e. WebCT) y programas educativos que permiten aprender casi cualquier materia académica (p. e. idiomas, matemáticas, ciencias naturales, etc.). Sin embargo, tenemos un problema a este nivel: hacer que el aprendizaje sea tan apropiado como lo sería cuando es conducido por una persona. Los Sistemas de Tutorización Inteligente (STI) son herramientas informáticas que han sido desarrolladas para funcionar como un mecanismo inteligente orientado a la enseñanza de cualquier tipo de contenido, y en este contexto es donde las redes bayesianas han mostrado ser útiles (J. López y García, 2007a, 2007b, 2008; Millán, Perez-de-la-Cruz, y Suárez, 2002).

Según Xiang (2002), los STI son un tipo de *agente* inteligente caracterizado por la toma de decisiones y la realización de acciones sin la intervención de ninguna persona que monitoriza el funcionamiento del sistema. Esto supondría que el sistema tiene que, en primer lugar, evaluar el conocimiento que el aprendiz tiene sobre la materia. A continuación, ha de presentar información acorde a ese nivel (p. e. definiciones, material gráfico o auditivo, etc.). Después ha de evaluar si la persona ha asimilado este nuevo conocimiento y presentar nueva información acorde con lo que ha aprendido. En general, un STI genera un modelo del conocimiento que tiene el alumno y le presenta la información acorde a el (p. e., Conejo, Millán, Perez de la Cruz, y Trella, 2001; Martin y VanLehn, 1995; Mislevy y Gitomer, 1996; Padilla, Lara, y Márquez, 2001).

En cierto modo, los STI irían más allá que los tests adaptativos computerizados (Olea y Ponsoda, 1996) ya que, además de evaluar adaptativa y dinámicamente el conocimiento sobre una materia, han de proveer material de trabajo en cada fase del programa formativo. Las redes bayesianas, dado que tienen la capacidad de generar una representación probabilística de algún aspecto de la realidad, pueden utilizarse como mecanismo intermedio entre el conocimiento que tiene el educando y la materia de aprendizaje a facilitar en base a lo asimilado.

Una de las primeras aproximaciones orientadas a evaluar el aprendizaje por medio de redes bayesianas se llevó a cabo con estudiantes de secundaria en la asignatura de física (Martin y VanLehn, 1995). OLAE (*Off-Line Assessment of Expertise*) fue un sistema que evaluaba diferentes aspectos clave en la resolución de problemas de física. Constaba de una interface gráfica dividida en cuatro secciones <sup>11</sup>: una barra de tareas para seleccionar categorías de problemas, un recuadro que presentaba el enunciado del problema junto a una representación gráfica del mismo, una representación gráfica donde el aprendiz tenía que señalar los fenómenos físicos implicados en el problema y una sección de respuestas donde el aprendiz tenía que expresar las fórmulas a utilizar y donde tenía que dar su respuesta. El sistema iba registrando las respuestas de los alumnos y generaba una red bayesiana que representaba las probabilidades asociadas a la utilización de cada fórmula física.

El sistema OLAE, sin embargo, no funcionaba como un tutor inteligente ya que se limitaba a estimar la probabilidad relativa de utilización correcta de fórmulas y principios físicos. Además, no realizaba estos cálculos dinámicamente, o sea, a medida que el aprendiz daba una respuesta; si no que lo hacía *a posteriori*. OLAE supuso un avance en el campo de los STI porque mostró que el conocimiento que tiene una persona sobre un dominio puede evaluarse usando redes bayesianas de un modo estructurado.

Otro sistema de tutorización inteligente es *Andes* cuya característica principal es que no reduce la iniciativa del usuario estableciendo vías estrictas en el aprendizaje (Conati, Gertner, y VanLehn, 2002). Más bien, es un sistema abierto que permite aprender una habilidad o concepto por medio de diferentes procedimientos. Otra característica importante de *Andes* es que puede suministrar ayuda dinámica cuando el/la estudiante se encuentra en un callejón sin salida y no sabe como continuar en la resolución de un problema. Se desarrolló a partir de OLAE y su objetivo es la enseñanza de física newtoniana a nivel universitario.

*Andes* tiene dos módulos principales, uno para resolución de problemas y otro para el estudio de ejemplos. El módulo de resolución de problemas es parecido al de OLAE pero incorporando una ayuda dinámica mas eficiente. En el módulo de estudio de ejemplos los estudiantes van siguiendo la resolución de los problemas

---

<sup>11</sup>El sistema OLAE permitía recoger información en cinco tipos de actividades por medio de una computadora. Aquí se describe la actividad de *resolución de problemas*.

### 3.7 El rol de las redes bayesianas en Psicología

de manera análoga a como lo harían en una clase impartida por un tutor humano. Así, el desarrollo de la explicación del problema va adaptándose a lo que la persona sabe a lo largo del proceso formativo.

El sistema hace una serie de asunciones acerca del proceso de generalización del aprendizaje; esto es, el grado en el que los usuarios son capaces de trasladar lo aprendido en un contexto a otro. Esto es un problema crucial en el proceso de enseñanza-aprendizaje ya que uno de los objetivos de la enseñanza es aplicar lo que se ha aprendido en clase a problemas de la vida cotidiana. Para modelar el proceso de generalización utiliza una estructura de red bayesiana divergente donde el nodo de divergencia representa la probabilidad de uso correcto de una determinada ecuación física y los nodos hijos representan diferentes contextos.

Mislevy y Gitomer (1996) trabajaron con el programa HIDRIVE desarrollado por los Laboratorios Armstrong de las Fuerzas Aéreas de los Estados Unidos. HIDRIVE es un STI desarrollado para simular el funcionamiento de un avión de combate F-15. Un problema comienza con una secuencia de video donde un piloto describe algunas deficiencias en el funcionamiento de un aparato que está a punto de aterrizar o que ya ha aterrizado (por ejemplo, el chequeo rutinario del timón de aterrizaje no responde correctamente). La interface gráfica permite al estudiante llevar a cabo una tarea de solución de problemas revisando videos del aparato y permitiéndole actuar sobre ellos. La red bayesiana generada a partir de HIDRIVE consta de 22 nodos organizados jerárquicamente en cuatro capas. En la cúspide de la jerarquía estaría la ejecución global. En un nivel inferior estarían los tres tipos de conocimientos evaluados: conocimiento de sistema, conocimiento estratégico y conocimiento procedimental. En la tercera capa estarían los subcomponentes de cada tipo de conocimiento. La tercera capa sería una capa oculta donde se integra la información de las capas precedentes. La cuarta capa es donde se interpretan las acciones del estudiante en relación a los posibles escenarios que plantea el sistema. Todos los arcos se orientan desde los niveles jerárquicos más altos hacia los más bajos.

Uno de los avances más importantes que se han realizado en este campo ha sido el desarrollo e implementación de STI en internet. Conejo et al. (2004, 2001) han desarrollado un sistema (SIETTE) que lleva a cabo procesos de diagnóstico y evaluación del aprendizaje por medio de internet. Otro aspecto interesante de SIETTE es que se basa en la Teoría de Respuesta al Ítem para escalar el material

didáctico y a los usuarios, mientras que el motor de inferencia que diagnóstica el estado de conocimiento de aprendiz lo lleva a cabo una red bayesiana.

Las redes bayesianas se perfilan como una herramienta de utilidad para la construcción de sistemas de tutorización inteligente. Glymour (2001) diría que éstas son capaces de representar el conocimiento latente del educando de una manera más eficiente a como lo representan los análisis factoriales o las técnicas de regresión. No obstante, los desarrollos teóricos y prácticos en la Teoría de Respuesta al Ítem como modelo para la estimación de conocimiento o habilidad latente, así como en los procesos de evaluación secuencial podrían optimizar los sistemas de autorización virtual inteligentes habidos hasta ahora (J. López y García, 2007b, 2008; J. López, Ruiz-Ruano, y García, 2008a).

### Aprendizaje Causal

El mayor volumen de estudios donde la Psicología hace uso de las redes bayesianas se ha generado investigando el aprendizaje causal. La obra de Glymour es, nuevamente, un referente clave en este contexto. A grandes rasgos, Glymour propone que las personas construimos una representación causal de la realidad por medio de la observación y la manipulación de ciertos eventos que la componen. Esta interacción observacional y manipulativa con el ambiente nos va dando pistas para ir generando una idea sobre la estructura causal del mundo; la cual, a su vez, es comparable, por lo menos fenomenológicamente, a una red bayesiana (Glymour, 2001, 2003).

Esta argumentación filosófica ha recibido evidencia experimental y se ha afianzado en el contexto científico como teoría del aprendizaje en niños pequeños (Gopnik et al., 2004; Gopnik y Schulz, 2004). La idea que subyace a la propuesta proviene de la idea de Heider al considerar que las personas somos como *científicos ingenuos* que llevamos a cabo procesos de generación, contraste y modificación de hipótesis sobre la realidad de modo parecido a como lo hace el método científico J. F. Morales (1994).

Esta nueva propuesta teórica supera a los modelos clásicos de aprendizaje causal, al menos, en varios sentidos que se han hecho notar por diferentes investigadores. En primer lugar, hay aspectos metodológicos que se refieren a la manipulación de las variables que se utilizan en las tareas causales que no son tratados por los modelos clásicos (p. e., Gopnik et al., 2004; Gopnik y Schulz,

2004; Matute, 2002). Por otro lado, la relación entre correlación y causación no se había hecho explícita evitando considerar que existe una relación directa entre ambas (p. e., Cheng, 1997; Gopnik et al., 2004; Perales, Catena, y Maldonado, 2002a). Por último, los modelos clásicos no hacen referencia al conocimiento previo que se tiene sobre la estructura causal de la realidad <sup>12</sup> (Gopnik et al., 2004; Perales et al., 2002a; Perales, Catena, y Maldonado, 2002b).

El paradigma experimental que apoya este nuevo modelo del aprendizaje causal en niños proviene de la tarea de detección de *blickets*<sup>13</sup>. En esta tarea se presentan varios cubos de madera a los niños y se les pregunta sobre su poder causal para hacer funcionar un mecanismo oculto que genera luces y música. Cuando un blicket es colocado sobre la máquina detectora de blickets, esta última emite música y luces de colores. El experimentador determina arbitrariamente qué es un blicket y qué no es un blicket.

En un estudio desarrollado por Gopnik, Sobel, Schulz, y Glymour (2001, experimento 1 y 2) se encontró que niños de cuatro años de edad realizaban juicios causales consistentes con el formalismo de las redes bayesianas en la tarea de detección de blickets. También se ha encontrado que los niños pequeños son capaces de modificar la situación experimental (esto es, intervenir sobre ella) para dar lugar a un efecto que previamente nunca habían observado (p. e., Gopnik et al., 2001, experimento 3; Sobel, Tenenbaum, y Gopnik, 2004, experimento 3). Este último resultado no puede ser explicado por las teorías clásicas del aprendizaje causal, mientras que es comprensible dentro de la teoría causal del aprendizaje desarrollada bajo el formalismo de las redes bayesianas (Gopnik et al., 2004; Spirtes et al., 2000). Otro fenómeno que ha servido para poner en tela de juicio los postulados de las teorías clásicas del aprendizaje ha sido el efecto de bloqueo hacia atrás (p. e., Gopnik et al., 2004; Sobel et al., 2004).

En la investigación en cognición animal se han encontrado resultados similares. En un estudio llevado a cabo por Blaisdell, Sawa, Leising, y Waldmann (2006) se encontró que ratas experimentales fueron capaces de extraer conclusiones causales consistentes con el formalismo normativo de las redes bayesianas. Contrariamente a lo que postulaban los modelos clásicos de aprendizaje, los animales fueron capaces de entender que sus acciones eran la causa de la presentación

---

<sup>12</sup>Excepto en términos generales guiados por supuestos teóricos robustos (Cheng, 1997).

<sup>13</sup>Un *blicket* es un bloque de madera que ha sido cargado arbitrariamente de poder causal en una tarea de detección causal.

de estímulos apetitivos y no que era fruto de una asociación simple como se pensaba anteriormente. Aunque son datos aislados y queda mucho trabajo por hacer a este nivel, estos resultados han llamado la atención de la comunidad científica al sugerir que los animales podrían ser sensibles al razonamiento causal complejo (Clayton y Dickinson, 2006).

Estos avances en teoría del aprendizaje han surgido paralelamente al desarrollo de nuevas teorías estadísticas orientadas a entender la causalidad a partir de datos observacionales pero extrayendo conclusiones equiparables a datos experimentales (p. e., Pearl, 2000, 2002; Spirtes et al., 2000). Esta diferenciación entre el aprendizaje meramente observacional e interventivo<sup>14</sup> está dando lugar a una intensa investigación (p. e., Lagnado y Sloman, 2004; Waldman y Hagmayer, 2005). Por ejemplo, Lagnado y Sloman (2004) han propuesto que la ventaja del aprendizaje basado en la intervención es debido a que conlleva información temporal que es usada eficientemente para determinar la estructura causal. Estos resultados son consistentes con los trabajos clásicos que enfatizaban la mejora del razonamiento y la predicción cuando la información se presentaba en el sentido natural de la realidad (de las causas a los efectos) y no a la inversa (Tversky y Kahneman, 1982).

Por su parte, Krynski y Tenenbaum (2007) y J. López y García (2009) han mostrado recientemente que los juicios de causalidad bajo incertidumbre emitidos por las personas son análogos a los que realiza una red bayesiana. Estos hallazgos son importantes en el contexto del desarrollo de modelos formales en campos aplicados como la intervención psicológica o la ingeniería del control de procesos (J. López, García, y De la Fuente, 2006; J. López y García, 2007c).

### Psicología Forense

El uso de las redes bayesianas en el ámbito forense cobra especial relevancia en la investigación policial (p. e., Garbolino y Taroni, 2002; Oatley y Ewart, 2002, 2003). Por ejemplo, Garbolino y Taroni (2002) proponen que las redes bayesianas podrían ser usadas en el ámbito forense como herramientas que ayudasen a sopesar evidencias (pruebas) de cara a la resolución de los casos.

---

<sup>14</sup>*Interventivo* en este contexto se refiere a un aprendizaje análogo al instrumental u operante. Sin embargo, tiene connotaciones estadísticas que se refieren a ajustar el valor de una variable de un modelo causal en un valor concreto y modificar su estructura cualitativa en algunos casos.

Una ilustración de esta utilidad de las redes bayesianas lo encontramos en el proyecto OVER. Este proyecto fue fruto de la colaboración entre la *West Midlands Police*, el *Centro de Sistemas Adaptativos, Computación y Tecnología*; y la *División de Psicología* de la Universidad de Sunderland. Como resultado de este trabajo conjunto, Oatley y Ewart (2003) han informado de la construcción de un sistema informático que predice la probabilidad de robos en casas de una región metropolitana de Birmingham. Este programa está integrado por varios módulos que funcionan controlados por diferentes herramientas estadísticas. Sin embargo, el motor de inferencia principal es una red bayesiana que estima la probabilidad de asaltos a casas en función de un conjunto de variables (p. e. *modus operandi*, número de asaltos en una zona, etc.). Todo el trabajo está monitorizado por un plano de la ciudad donde se representa el riesgo de asalto con diferentes colores. A su vez, todos los análisis pueden ser exportados a formato Excel para que la policía pueda elaborar informes fácilmente.

Recientemente, (J. López, De la Fuente, et al., 2008) presentaron un modelo de red bayesiana capaz de predecir la tendencia del veredicto en jurados simulados. Estos resultados podrían servir para conseguir que la formación de jurados sea más imparcial y menos sesgada a ciertos veredictos.

### Psicolingüística

Como se ha comentado anteriormente Glymour (2001) defendió la utilización masiva de las redes bayesianas en el quehacer de la Psicología. Su propuesta tocaba varias áreas pero se centraba en el aprendizaje. Sin embargo, hizo referencia a la utilidad que supondrían las redes bayesianas en el estudio de los mecanismos neurocognitivos del lenguaje. Su propuesta pasó por proponer que se aplicase el formalismo de las redes bayesianas sobre los modelos clásicos de la Neuropsicología del lenguaje (Ellis y Young, 1988).

No obstante, fue Daniel Jurafsky (1996) quien propuso una conexión entre la estructura del lenguaje en términos de probabilidades condicionadas y ciertos aspectos léxico-semánticos. Algo parecido ha hecho la *Cambridge University Press* en su *Cambridge Advanced Learner's Dictionary (CALD)* ya que cada una de sus entradas está etiquetada en términos de la probabilidad de uso con base en el *Frequency Information System*. Este sistema de información basado en la frecuencia clasifica las palabras en tres grupos dada una gran base de datos lingüística o

*CORPUS* atendiendo no sólo a su importancia relativa en el discurso, sino también respecto a su significado individual y en unión con otras palabras. El CALD tiene tres grupos principales de palabras: las entradas tipo *E* (de *Essential*), que son palabras de uso muy frecuente en inglés ( $p(E) = 4 \times 10^{-5}$ ) y que son esenciales para desarrollar una comunicación eficiente; las tipo *I* (de *Improver*), que son palabras que aparecen en el idioma inglés con una probabilidad comprendida entre  $2 \times 10^{-5}$  y  $4 \times 10^{-5}$ , las cuales son útiles para definir conceptos importantes; y por último, las palabras tipo *A* (de *Advanced*) son menos frecuentes en inglés ( $1 \times 10^{-5} \leq p(A) \leq 2 \times 10^{-5}$ ) pero son altamente significativas denotando fluidez y naturalidad en la comunicación.

Aunque el trabajo inicial de Jurafsky tan sólo hace una ligera referencia a la utilidad de las redes bayesianas en el estudio del lenguaje a nivel psicológico, en estudios posteriores (Narayan y Jurafsky, 1998, 2002) se han replicado algunos resultados experimentales usando modelos de red bayesiana. Estos hallazgos también han tenido utilidad práctica. Por ejemplo, en un estudio orientado a comparar la habilidad en la recuperación de información electrónica entre Redes Neuronales Artificiales y redes bayesianas, las redes bayesianas ofrecieron mejores resultados (F. Martínez, Díaz, Martín, Rivas, y Ureña, 2003).

Por otro lado, estudios recientes han mostrado que la evolución del lenguaje en términos históricos obedece a un patrón probabilístico basado en el uso relativo de las palabras (Fitch, 2007; Marris, 2007). Por ejemplo, (Lieberman, Michel, Jackson, Tang, y Nowak, 2007) han observado que la vida media ( $VM_v$ ) de los verbos irregulares en el idioma inglés obedece a la regla

$$VM_v \propto \sqrt{fr_v},$$

donde  $fr_v$  se refiere a la frecuencia de uso del verbo. O lo que es lo mismo, las palabras con frecuencia de uso baja evolucionan más rápidamente que las palabras que son usadas frecuentemente. Fitch (2007) ha señalado que esta observación puede favorecer el desarrollo de modelos teóricos basados en *sesgos innatos* que podrían tomar la forma de distribuciones previas bayesianas o teorías destinadas a modelar el proceso de aprendizaje del lenguaje en términos de sistemas estadísticos. A su vez, esta observación abre la posibilidad de estudiar cuantitativamente la evolución del lenguaje en los mismos términos que se supone que evolucionan los seres vivos. Por ejemplo, el concepto de *memes* introducido por

(Dawkins, 1976/1988) que postula que ciertos elementos lingüísticos evolucionan de modo análogo a como lo hacen los genes podría ser abordado empíricamente.

#### **Psicodiagnóstico**

Los sistemas expertos han estado relacionados con las Ciencias de la Salud desde sus inicios; de hecho, el surgimiento de las redes neurales artificiales (que pueden considerarse como un tipo particular de sistema experto) estuvo asociado al diagnóstico médico (Quinlan, 1991).

La red ASIA (Cowell et al., 1999) fue la primera red bayesiana desarrollada para ser utilizada como ayuda en el diagnóstico de enfermedades. Estuvo compuesta por ocho nodos y se dirigió al diagnóstico diferencial de tuberculosis, bronquitis y cáncer de pulmón. Sin embargo, esta red «primigenia» no estaba orientada a resolver problemas reales, sino que, más bien pretendía ser un ejemplo de la potencialidad de las redes bayesianas en el diagnóstico médico (Díez, 1998). La primera red bayesiana con aplicación clínica real se desarrolló en la Universidad de Aalborg (Dinamarca) y tuvo como objetivo el diagnóstico de enfermedades musculares (Hugin Expert, 2004) y la monitorización de insulina en pacientes diabéticos.

En el diagnóstico psicológico, al igual que en el médico, se lleva a cabo un proceso de toma de decisiones donde el profesional, o profesionales, tiene que poner en marcha una búsqueda activa de evidencias y actuar acorde a ellas (Godoy, 1996). Por ello, se han utilizado numerosas herramientas estadísticas para minimizar las posibilidades de error en el diagnóstico. Por ejemplo, la teoría de la utilidad esperada junto a la maximización de la utilidad multiatributo y el uso de la regla de Bayes para calcular probabilidades condicionales han sido usadas con la idea de apoyar a los profesionales en su toma de decisiones (p. e., Esteve, 1996a, 1996b).

Un ejemplo de sistema experto de red bayesiana usado para el diagnóstico psicológico fue MENTOR (Mani et al., 1997), que fue construido en la Universidad de California entre 1959 y 1980 con la colaboración de la Fundación Kaiser y su objetivo era la predicción y el diagnóstico precoz de retraso mental en bebés. Por su parte, el Instituto de Investigación en Salud Mental de Victoria (Melbourne, Australia, [www.mhri.edu.au](http://www.mhri.edu.au)) es una organización que está trabajando con la tecnología de las redes bayesianas para comprender, tratar y prevenir trastor-

nos psicológicos como la esquizofrenia, los trastornos del estado de ánimo o el Alzheimer.

No obstante, como señalan (J. López et al., 2007), todas las iniciativas orientadas a desarrollar modelos, sistemas o herramientas de utilidad en la toma de decisiones clínica han de tener en cuenta los problemas que acarrearán en términos ontológicos relativos a lo *descriptivo* y/o *normativo*. Algunos autores (p. e., Godoy, 1996; Raiffa, 1994) proponen que los mecanismos, como las redes bayesianas, de optimización ante la toma de decisiones clínica tendrían que ser meros *apoyos* y no sustitutos de los profesionales que desarrollan esta actividad.

### Psicología Organizacional

Como herramienta estadística, las redes bayesianas podrían usarse en el estudio de cualquier tópico investigado por la Psicología Social. Sin embargo, los estudios que se han realizado por el momento se han centrado en aplicaciones orientadas al entorno empresarial. El interés por la estadística aplicada a la toma de decisiones en el contexto empresarial se ha acentuado gracias a la evolución de los estilos de dirección, pasando de un estilo de dirección centrado en la experiencia del empresario a un modelo de dirección basado en indicadores estadísticos, acaecida en los últimos años (p. e., Alcazar, 2005; Forester, 1968/1973; Giles, 2005b; León, 1994; A. Martínez, Martínez, y Martínez, 2002).

En el campo de los estudios de mercado las redes bayesianas han mostrado ser herramientas especialmente útiles (Concejero, 2004; Nadkarni y Shenoy, 2001, 2004). Por ejemplo, Baesens et al. (2004) encontraron que una red bayesiana con cuatro nodos y seis arcos (frente a otros modelos más complejos y menos parsimoniosos que implicaban más variables) clasificaba eficientemente a los clientes de una empresa en relación a la perspectiva de compras a largo plazo. Encontraron que la evolución de las compras de los clientes se relacionaba (positivamente) con el porcentaje máximo de productos que adquiriría de una misma categoría; y (negativamente) con la contribución inicial y el número total de artículos comprados.

El vínculo más claro del mundo empresarial y la psicología social aplicada podríamos encontrarlo en la gestión y dirección de recursos humanos. Por ejemplo, como se ha hecho notar por Maciá et al. (1990), la selección de personal es un buen ejemplo de un proceso de toma de decisiones secuencial dentro de una organización. En este sentido, los diagramas de influencia serían apropiados para

### 3.7 El rol de las redes bayesianas en Psicología

ir guiando el proceso de selección de personal en función de ciertas restricciones. HumRRO (*Human Resources Research Organization*, [www.humrro.org](http://www.humrro.org)) es una empresa que se fundó en 1951 y es un ejemplo de la utilización de las redes bayesianas en la consultoría de recursos humanos.

## Fundamentos de las redes bayesianas

# Capítulo 4

## Actitudes emprendedoras

*As long as they vote with their feet and hearts, immigrants are a more powerful engine for social change than armies.*

(Richerson y Boyd, 2008)

### 4.1. Reseña histórica

El espíritu emprendedor probablemente ha sido, junto a la aparición de un sistema lingüístico complejo o las habilidades de aprendizaje, uno de los grandes hitos filogenéticos en la historia de la evolución humana (Boyce, Kral, Allnut y Goetz, 2005). Esta tendencia a la «aventura» podría haber aparecido como rasgo característico de la especie humana hace aproximadamente 60.000 años y podría haber representado el punto arquimédico que empujó al representante prehistórico del hombre moderno a abandonar el continente africano y a conquistar, en cuestión de un par de miles de años, todos los ecosistemas del planeta.

El papel de las personas emprendedoras en el sistema productivo ha sido destacado en contextos de diferente índole (Corman, Lussier, y Nolan, 1996) y es importante por varias razones. Bien porque aportan el dinamismo que las economías necesitan desde un punto de vista microeconómico, porque son una fuente ingente de innovación o porque regulan la tasa de empleo de los países;

## Actitudes emprendedoras

las personas que emprenden la creación de una empresa generan alteraciones beneficiosas en el tejido empresarial en el que se desenvuelven y provocan cambios que son mensurables desde un punto de vista macroeconómico. No obstante, pese a que el interés científico por la tendencia a crear empresas se remonta al siglo XVIII y XIX, son muchas las lagunas y la controversia que hay respecto a su conceptualización (p. e., Cano, García, y Gea, 2003; Rogoff y Lee, 1996).

Podríamos decir que el concepto, a grandes rasgos, de *emprendedor* o *persona emprendedora* tiene su origen en Francia. De hecho, el vocablo inglés acuñado para hacer referencia a este fenómeno (*entrepreneur*) procede del francés. Por otro lado, como señalan Hébert y Link (1989) la taxonomía de las teorías referentes al fenómeno emprendedor tienen su origen en el economista irlandés Richard Cantillon que pasó gran parte de su vida en Francia.

Richard Cantillon ha conseguido la mayor parte de su reputación gracias a su obra *Essai sur la nature du commerce en général*, que permaneció en la oscuridad alrededor de un siglo y que fue descubierta accidentalmente por William Stanley Jevons (Hayek, 1985). El modelo económico propuesto por Cantillon identifica a tres agentes principales que forman parte de una jerarquía de relevancia: a) los terratenientes (*landowners*), que son económicamente independientes; b) los emprendedores, que se embaucan en intercambios de mercado bajo riesgo con la intención de conseguir beneficio; y los empleados (*hirelings*), que son persuadidos por dotaciones económicas para realizar trabajos poco atractivos o desagradables y que evitan la toma de decisiones activa con el fin de asegurarse garantías contractuales de ingresos estables. Aunque Cantillon propone que la cúspide de la jerarquía pertenece a los terratenientes, los emprendedores tienen un papel esencial en el desarrollo económico. Para él los emprendedores son agentes económicos que consiguen beneficios a partir de juicios empresariales bajo incertidumbre y no por medio de rentas o salarios. Con posterioridad otros autores han secundado esta idea de considera al emprendedor como un ente social que desafía al riesgo y a la incertidumbre (p. e., Hayek, 1985; Hébert y Link, 1989; Polopolus y Emerson, 1991; Samuelson, 1970).

Varios aspectos de los destacados por Cantillon cobran especial relevancia en el campo de las actitudes emprendedoras. En primer lugar, destaca claramente que los emprendedores son agentes económicos que tienden a gestionar eficientemente la incertidumbre del mercado. A su vez, acentúa que es la *función* y no la

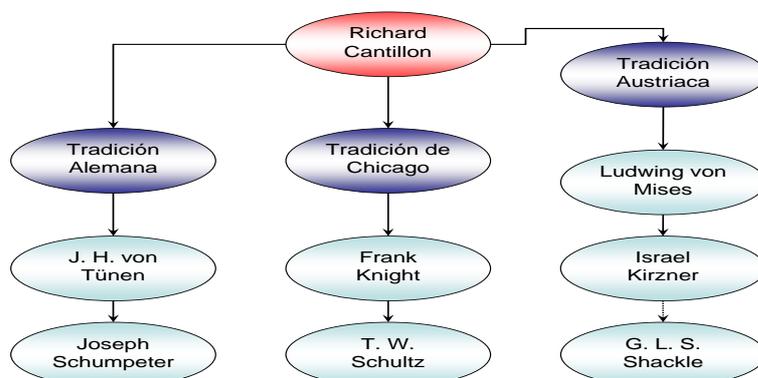
personalidad del emprendedor lo que los diferencia de otros elementos productivos. Por último, desliga la noción de emprendedor de la clase social. Los ejemplos que Cantillon (1755/1931) relata son ilustrativos de su perspectiva:

[...] *El granjero es un emprendedor que promete pagar al propietario, por su granja o su tierra, una suma fija de dinero [...] sin tener la certeza del beneficio que obtendría de esta empresa. [...] El precio de los artículos producidos por el granjero depende naturalmente de estos [...] acontecimientos imprevisibles para él, lo cual significa que conduce la empresa de su granja con incertidumbre. [...] El artesano que ha comprado la lana del comerciante, o directamente del productor, no puede saber qué beneficio obtendrá al vender sus paños y telas al sastre. El lencero es un empresario que compra telas al fabricante, a un determinado precio, para revenderlas a un precio incierto [...] Los tenderos y detallistas de toda especie son empresarios que compran a un precio cierto, y revenden en sus tiendas o en las plazas públicas a un precio incierto. [...] Todos los otros empresarios, como los que se benefician de las minas, o los de espectáculos, edificaciones, etc –lo mismo que los empresarios de su propio trabajo, que no necesitan fondos para establecerse como buhoneros, caldereros, zurcidoras, des-hollinadores, aguadores, etc.–, subsisten con incertidumbre [...] pp. 39–45.*

Aunque la propuesta de Cantillon ha sido criticada por no tener en cuenta algunos estratos de la sociedad (la nobleza o cierto sector del proletariado) a la hora de modelar la dinámica económica, ha servido para que otros autores desarrollasen sus teorías en relación al emprendedor sobre la base de sus postulados. Como señalan Hébert y Link (1989), el trabajo de Cantillon ha sido el germen que ha propiciado el desarrollo de tres escuelas de pensamiento sobre la naturaleza, desarrollo y función del emprendedor (Figura 4.1). Aunque algunas de las flechas de este gráfico no son tan rectas como cabría esperar desde una perspectiva histórica y en algunos casos la asociación entre autores es muy tenue, son útiles desde el punto de vista expositivo y la mayoría de las relaciones que se establecen son ineludibles.

La primera línea de pensamiento surgida del trabajo de Cantillon, la Tradición Alemana, desemboca en Joseph Schumpeter cuya aportación proviene de su

## Actitudes emprendedoras



**Figura 4.1:** Corrientes de pensamiento surgidas a partir del trabajo de Richard Cantillon. Basado en Hébert y Link (1989).

análisis sobre el desarrollo económico como proceso dinámico, donde la perturbación del equilibrio es vista como un paso hacia el avance de la economía. En este modelo el emprendedor tiene un papel clave como pieza fundamental del desarrollo económico por medio de la *creación destructiva*, ya que el emprendedor trata de optimizar sus beneficios desarrollando innovaciones (en productos o en procesos de producción) que le hagan tener una ventaja competitiva en detrimento (en cierto modo destruyendo) a sus competidores (McCraw, 2006). Así pues, el emprendedor es aquella persona que innova, que establece nuevas combinaciones de elementos en el sistema productivo con el fin de aumentar sus beneficios a costa de reducir los beneficios de sus competidores.

En lo que respecta a la Tradición de Chicago, el gráfico de la Figura 4.1 desemboca en Theodore William Schultz que compartió el Premio Nobel Conmemorativo en Economía con William Arthur Lewis en el año 1979. Schultz es reconocido por sus aportaciones sobre el desarrollo económico en relación a la educación de donde surgió la denominada *Teoría del Capital Humano*. En el estudio de las actitudes emprendedoras Schultz hizo dos grandes aportaciones según Hébert y Link (1989): a) definir emprendedor como un agente con habilidad para enfrentarse a los desequilibrios económicos y extender la idea de emprendedor a actividades no mercantiles<sup>1</sup> (arte, estudios, hogar, deporte, etc.), y b) destacar la importancia de la educación sobre la habilidad de los agentes emprendedores para percibir y reaccionar con éxito a los desequilibrios económicos. Otro aspecto

<sup>1</sup>McKenzie, Ugbah, y Smothers (2007) también han hecho recientemente esa misma apreciación.

importante de sus aportaciones es que matiza el establecimiento de una relación directa entre la habilidad para gestionar el riesgo o la incertidumbre y la actividad emprendedora. Desde su punto de vista plantea que, aunque el riesgo y la incertidumbre son importantes para la toma de decisiones de los emprendedores; estos elementos de decisión también están presentes en ausencia de la gestión emprendedora.

Para terminar, una nueva y desafiante teoría sobre el papel de los emprendedores en los sistemas económicos es la aportada por Israel Kirzner enmarcada dentro de la Tradición Austriaca. Para Kirzner el emprendedor es caracterizado por un *estado de alerta* especial que le permite aprovechar las oportunidades de consecución de beneficio en el mercado. A su vez, el papel de los emprendedores es conseguir alcanzar el tipo de ajuste necesario para mover los mercados económicos hacia el punto de equilibrio (Foshee, Heath, y Balic, 2003). De este modo, el rasgo que caracteriza a los emprendedores desde esta perspectiva es la capacidad de percepción para reconocer oportunidades de éxito que conduzcan a los mercados en los que se desenvuelven al punto de equilibrio. La flecha discontinua que aparece entre Kirzner y Shackle es debida a que el segundo no comparte la idea de tendencia al equilibrio.

Como se puede observar, la diferencia entre estas tres corrientes de pensamiento se sintetiza en la relación del emprendedor con el equilibrio económico. Mientras que la escuela alemana identifica al emprendedor como un elemento clave que favorece el desequilibrio económico que produce desarrollo económico y la escuela de Chicago destaca que el emprendedor ha de ser hábil para reaccionar a los desequilibrios económicos, la tradición austriaca sugiere que los emprendedores empujan al sistema hacia el equilibrio. En cualquier caso, todas estas perspectivas economicistas de la función del emprendedor han supuesto el punto de partida para el desarrollo de modelos orientados a identificar y definir la predisposición emprendedora.

## 4.2. Conceptualización y tipologías emprendedoras

Como señalan Hébert y Link (1989), una posible explicación a este relativo desacuerdo podría deberse a las divergencias metodológicas que subyacen en la

## Actitudes emprendedoras

investigación en la creación de empresas. Otro elemento que podría haber influido en el poco grado de acuerdo es el abordaje multidisciplinar que se ha hecho del tema ya que, en algunas ocasiones, el trasvase conceptual entre varias disciplinas puede conducir a imprecisiones conceptuales lo que podría haber conducido a malas interpretaciones y, por consiguiente, a una vaga definición del concepto. Por ejemplo, es posible que el uso de la palabra *empresario* como sinónima de *emprendedor* pueda haber hecho mella en interpretaciones conceptuales en campos tan dispares como la Economía y la Psicología (Veciana, 1989).

De cualquier forma, se han propuesto multitud de definiciones orientadas a clarificar el concepto de emprendedor que han arrojado luz sobre este objeto de estudio. No obstante, se puede apreciar una gran heterogeneidad en la finalidad y contenido de las definiciones de emprendedor. Por ejemplo, desde un punto de vista histórico-sintético, Hébert y Link (1989) señalan que el emprendedor es:

*[...] alguien especializado en asumir la responsabilidad ante la toma de decisiones que afecta a la localización, forma, uso de bienes, fuentes o instituciones [...] p. 39.*

Mientras que desde un punto de vista empresarial (Genesca y Veciana, 1984; Veciana, 1989) se ha señalado que el emprendedor es:

*[...] la persona que ha creado una empresa en marcha y esta sobrevive en años [...] p. 147 y 19 respectivamente.*

Por su parte, una de las definiciones más austeras del emprendedor fue acuñada por Gartner (1988) que lo definió como:

*[...] el rol social que subyace en los individuos que crean organizaciones [...] p. 30.*

En otras ocasiones la figura del emprendedor se ha definido desde un punto de vista más práctico que hiciese posible operacionalizar el concepto. Un ejemplo de ello lo encontramos en el trabajo de Huefner, Hunt, y Robinson (1996) que lo delimita como:

*[...] la persona que se autodefine como emprendedor y que ha poseído y dirigido uno o más negocios [...] p. 62.*

## 4.2 Conceptualización y tipologías emprendedoras

Una aproximación metodológica más reciente es la llevada a cabo por J. López, García, Cano, Gea, y De la Fuente (2009) con la que señalan que la persona emprendedora se caracteriza por ciertos rasgos en términos probabilísticos. En cualquier caso, la figura emprendedora corresponde a una persona que busca beneficio, en contraposición al salario que busca un directivo o ejecutivo, como contrapartida a su actividad emprendedora. O más generalmente, el emprendedor es *la persona o grupo que busca explotar una oportunidad económica* (McKenzie et al., 2007). Por otro lado, los emprendedores se embarcan en nuevas actividades, mientras que el objetivo de los directivos o ejecutivos es mantener funcionando un negocio previamente establecido. Como señala Samuelson (1970) los rasgos más característicos del emprendedor son tener una visión, ser original, poseer valentía e introducir más que descubrir nuevos productos o servicios. Por ejemplo, siguiendo el ejemplo que él mismo plantea en torno al teléfono; Maxwell desarrolló formalmente la teoría de las ondas de radio, Hertz observó el fenómeno empíricamente pero fueron Marconi y Sarnoff quienes realmente lo hicieron económicamente rentable. Es más, Marconi y Sarnoff se antepusieron en cierto modo al sentido común e hicieron productiva una idea que *a priori* no parecía fructífera desde un punto de vista financiero. Por ejemplo, el 19<sup>o</sup> presidente de los Estados Unidos, Rutherford Birchard Hayes, dijo después de una demostración de su funcionamiento en 1876 que *era un gran invento pero ¿a quién le gustaría usarlo?*<sup>2</sup>

Dada la variedad de definiciones existentes en la bibliografía, algunos autores han optado por identificar los diferentes tipos de perfiles que se agrupan bajo el espectro emprendedor. Con esta estrategia se conseguiría reducir la ambigüedad conceptual, ya que se profundiza en la definición conceptual del objeto de estudio, y proveer de información útil desde un punto de vista práctico, ya que serviría para diseñar las investigaciones.

Por ejemplo, Rogoff y Lee (1996) tomaron una muestra de 231 propietarios de negocios y evaluaron tanto sus vías de inicio en los negocios como sus actitudes hacia la consecución de objetivos. Encontraron que este grupo de emprendedores se podía dividir en tres perfiles diferentes: los creadores, los herederos y los operadores. Los emprendedores *creadores* son definidos como aquellos que han iniciado una nueva aventura empresarial con el sueño de crear un nuevo servicio o producto. Los emprendedores con perfil de *herederos* incluye a estas personas que han

---

<sup>2</sup>*It's a great invention but who would want to use it anyway?*

## Actitudes emprendedoras

heredado un negocio familiar o quienes entraron en el mundo de los negocios por medio de una conexión de parentesco. Por último, los emprendedores *operadores* son estos que adquieren un negocio o una franquicia. Este tipo de emprendedor está principalmente motivado por objetivos financieros, por la ausencia de opciones, por el deseo de adquirir un negocio o franquicia con una reducida carga de incertidumbre en el mercado.

Sus resultados mostraron un patrón significativo de diferencias entre los tres grupos. En primer lugar, los creadores, conducidos por la visión de crear nuevos productos o servicios, obtuvieron mayores puntuaciones en las dimensiones de creación de negocios-productos, en la sub-escala de respeto-reconocimiento por parte de los semejantes y en la dimensión de la utilización de las propias habilidades en la escala utilizada. Por su parte, los operadores mostraron un patrón de respuestas consistente con su orientación financiera puntuando más alto en el objetivo de proteger su inversión y en el objetivo de construir algo para la propia familia. Por último, los herederos mostraron un perfil muy diferente a los dos grupos anteriores; puntuaron alto en su orientación al beneficio, bajo en la tendencia a crear servicios o productos y obtuvieron los niveles más bajos de utilización de las propias habilidades así como en la dimensión de contribuir a la sociedad.

Por su parte, Veciana (1989) destaca que recientemente han surgido tres nuevas figuras dentro del marco emprendedor: los emprendedores técnicos, la mujer empresaria y las empresas cooperativistas. El denominado *emprendedor técnico* basa su actividad económica en el sector tecnológico comercializando productos o servicios relacionados con las nuevas tecnologías. Esta forma de aventurero empresarial no se diferencia fundamentalmente del perfil general de emprendedor excepto en cuatro características:

- posee un nivel de formación más alto,
- suele ser más joven a la hora de crear su empresa,
- tiende a crear la nueva empresa en equipo,
- y por lo general ha hecho uso y ha sido asesorado o apoyado en su decisión de crear la empresa por las *incubadoras* o *viveros* de empresas.

Ejemplos de este tipo de emprendedores podemos verlos en la génesis de compañías como *Microsoft*, *Adobe Systems*, *Apple Inc.*, *eBay*, *Google*, *Hewlett-*

## 4.2 Conceptualización y tipologías emprendedoras

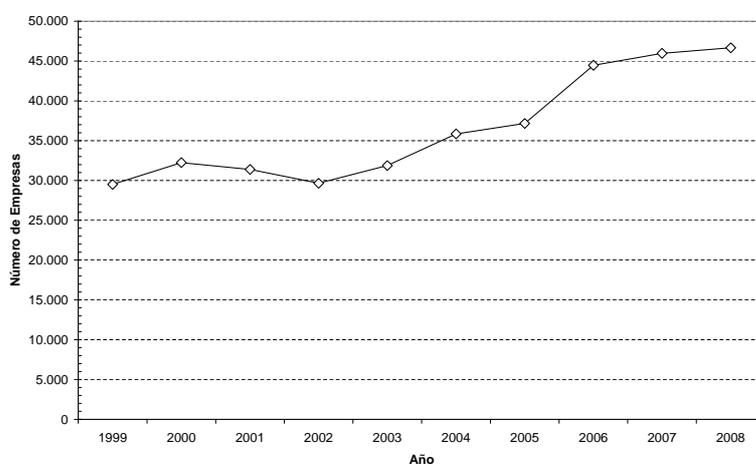
*Packard, Intel, Sun Microsystems, Symantec o Yahoo!* surgidas todas ellas en el *Silicon Valley*. Por lo general todo este tipo de empresas técnicas suelen surgir al amparo de universidades, institutos técnicos o centros de investigación. El caso del nacimiento de *Microsoft* y *Apple* son ejemplos legendarios hasta tal punto que ha sido llevado al cine (Moore y Burke, 1999). Para poner un ejemplo de nombre propio a este perfil emprendedor podemos citar a Torakusu Yamaha fundador de *Yamaha Corporation* en 1887 que inicialmente se dedicó a la fabricación de instrumentos musicales (ello queda reflejado en el logo de la compañía compuesto por tres diapasones cruzados), posteriormente construyó hélices para aviones y más recientemente se ha dedicado al mercado de las motocicletas, entre otros. También podemos citar a Tsutomu Kato y Tadashi Osanai como fundadores en 1962 de *Korg Corporation* y que se ha convertido en un referente mundial en la construcción de dispositivos electrónicos musicales. En la actualidad la idea de emprendedor tecnológico aparece reflejado en lo que ha venido a denominarse *Empresa de Base Tecnológica* o *EBT* (De las Nieves, 2008).

En cuanto a la mujer empresaria, el autor destaca el mayor nivel formativo y la motivación de logro así como la de independencia como los elementos diferenciadores de este nuevo colectivo emprendedor. A su vez, también habría que acentuar las ayudas y subvenciones económicas estatales que promueven o favorecen la entrada de la mujer en el mercado empresarial.

En lo que respecta a los empresarios cooperativistas, el autor ya señalaba en 1989 que se había producido un aumento considerable en el número de cooperativas basadas en la fórmula de la *economía social*, así como en el número de socios que las integraban, durante el periodo 1984-1987. Este formato de asociación muestra una serie de ventajas para los socios cooperativistas y, como puede verse en la gráfica de la Figura 4.2, el aumento progresivo de este tipo de empresas se sigue produciendo en la actualidad.

Por su parte, Huefner et al. (1996) propusieron una definición útil desde el punto de vista práctico basada en dos dimensiones: el número de negocios poseídos y la auto-definición de emprendedor. A partir de una muestra de 335 participantes encontraron que el cruce de estas dos dimensiones daba lugar a cuatro perfiles emprendedores: 1) los propietarios-directivos, que son personas que han poseído uno o más negocios pero que no se autodefinen como emprendedores; 2) los emprendedores potenciales, que son los que se autocalifican como emprendedores

## Actitudes emprendedoras



**Figura 4.2:** Evolución del número de empresas españolas desde 1999 a 2008 que no son trabajadores autónomos, sociedades anónimas o sociedades limitadas. Fuente: Directorio General de Empresas (DIRCE), *Instituto Nacional de Estadística*.

aunque nunca han poseído un negocio; 3) los emprendedores, que han poseído al menos un negocio y se definen como emprendedores; y 4) los no-emprendedores, que son las personas que ni se definen como emprendedores ni nunca han poseído un negocio.

Otra gran tipología de emprendedor, como también han hecho notar los autores clásicos (Hébert y Link, 1989), es la del *emprendedor social*. Los emprendedores sociales son entes dedicados a la producción o prestación de servicios sin la intención de obtener un beneficio económico por ello. Por ejemplo, Light (2005) define al emprendedor social como:

*un individuo, grupo, red, organización o alianza de organizaciones que buscan cambios sostenibles a gran escala a través de rotura de patrones de ideas sobre que y/o como los gobiernos, organizaciones sin ánimo de lucro y empresas han de actuar para gestionar problemas sociales significativos.*

Light (2005) también ha aportado una sub-clasificación dentro de la tipología de emprendedor social. Para ello ha utilizado dos dimensiones: la intensidad de la actividad del emprendedor social (a dos niveles: alta y baja), y el apoyo económico a la actividad emprendedora (también a dos niveles: alto y bajo). De la combinación de estas dos dimensiones con sus dos posibles niveles surgen cuatro tipos

## 4.2 Conceptualización y tipologías emprendedoras

de emprendedores sociales. En primer lugar, estarían los *emprendedores sociales al completo* que serían aquellos entes que reciben un alto volumen de subvenciones y que llevan a cabo una intensa actividad emprendedora. Los emprendedores que ponen en marcha una gran cantidad de actividades emprendedoras con bajos niveles de apoyo económico serían los *emprendedores rebeldes*; mientras que los agentes sociales que reciben una alta cantidad de recursos económicos y que ejecutan pocas acciones emprendedoras serían los denominados *emprendedores difusos*. Para terminar, estarían los *emprendedores falsos* que serían aquellos que reciben poco dinero pero que ponen en marcha pocas iniciativas emprendedoras.

Para por último, también se ha acuñado el concepto de *intraemprendedor*<sup>3</sup> para hacer alusión a las personas que inician o desencadenan el desarrollo de nuevos productos o servicios en el seno de organizaciones ya creadas (Asián, Fernandez, y Otri, 2005; Lumpkin, 2006). Desde el punto de vista de Lumpkin (2006), la figura del intraemprendedor tiene un papel esencial en las organizaciones porque engrana el proceso de innovación dentro de la empresa lo que conduce al éxito corporativo.

Aunque los intraemprendedores tienen un perfil parecido a los emprendedores se pueden establecer diferencias entre ambos entes económicos. Como señalan HayGroup y SAP AG (2003), un alto grado de identificación con la compañía es un rasgo que se encuentra presente en los intraemprendedores, mientras que esto no parece darse en los emprendedores (ellos son la compañía). El nivel educativo es otra variable que funciona diferencialmente en intraemprendedores y emprendedores. Mientras que el nivel educativo parece ser crítico en la permanencia y expansión de los emprendedores, en los intraemprendedores es un aspecto importante en fases primigenias de la carrera profesional. Esto es, la educación es una barrera (un filtro) de entrada a la organización y a los puestos de responsabilidad dentro de ella; pero una vez superada, el nivel educativo parece no marcar la diferencia entre los intraemprendedores de éxito y los profesionales promedio.

---

<sup>3</sup>También denominados como emprendedores corporativos, aventureros corporativos o estrategas corporativos (Baum, Frese, y Baron, 2006).

### 4.3. Teorías explicativas

Existen diferentes modelos que han sido desarrollados para dar una explicación a por qué algunas personas emprenden la creación de empresas. Revisando la bibliografía sobre el tema se puede observar que, salvando las fronteras poco nítidas que los separan, se podrían crear dos grupos de modelos explicativos de la actividad emprendedora (Cano et al., 2003; J. García, Cano, y Gea, 2005). Por un lado, existe un grupo de modelos que enfatizan un origen sociológico, en algunos casos economicista, de la figura del emprendedor. Desde este punto de vista, los emprendedores son el producto de una serie de fuerzas socio-económico-demográficas que condicionan el surgimiento de las actitudes emprendedoras.

Por otro lado, hay un conjunto de modelos que acentúan los condicionantes psicológicos que dan lugar a personas más orientadas hacia la creación de empresas. En este caso, el hecho de que una persona tenga una actitud positiva hacia la creación de empresas suele ser explicado aludiendo a los aspectos psicológicos que la caracterizan. En las siguientes secciones se describen someramente las características de cada uno de estos grupos así como los sub-modelos que se han desarrollado dentro de ellos.

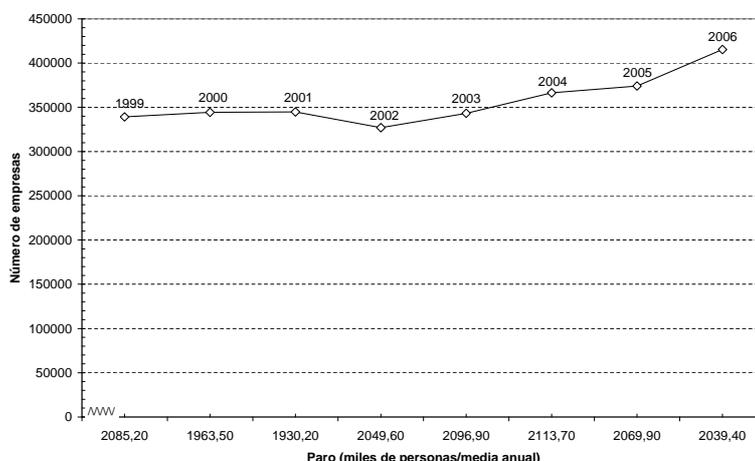
#### 4.3.1. Dimensión socio-económica

Dentro de este grupo de modelos orientados a justificar el surgimiento de la tendencia hacia la creación de empresas podemos encontrar modelos de distinta índole que aluden tanto a la dinámica del mercado laboral como a aspectos sociales relacionados con el estatus socioeconómico.

##### **Modelo de las turbulencias en el mercado laboral**

Bajo esta insignia se recoge uno de los grandes modelos acuñados para explicar el desarrollo de las actitudes hacia la creación de empresas (p. e., Cano et al., 2003; Loufti, 1991; Sánchez, 2003). Este modelo surgió al observar que la creación de empresas correlaciona positivamente con las turbulencias en el mercado laboral, principalmente el desempleo o la amenaza de desempleo (Stanworth, Stanworth, Granger, y Blyth, 1989).

De hecho, si analizamos los datos de paro y de creación de empresas de los años recientes se observa que existe una pequeña correlación positiva entre estas dos



**Figura 4.3:** Relación entre desempleo y creación de nuevas empresas. Fuente: Anuario de Estadísticas Laborales y de Asuntos Sociales. Ministerio de Trabajo y Asuntos Sociales, Demografía de Alta de Empresas (DIRCE), *Instituto Nacional de Estadística*

variables ( $r_{xy} = 0,125$ ), aunque no alcanza a ser estadísticamente diferente de cero ( $p = 0,768$ ) (Figura 4.3). Aunque no exista una correlación significativamente diferente de cero y que los datos analizados aquí corresponden a una serie temporal de ocho años; la relación que se establece entre desempleo y creación de empresas no está clara a nivel empírico. Además, estos datos se limitan a España, por lo que estos resultados no serían generalizables a otros países.

### Modelos *pull* y *push*

Otra forma en la que se manifiesta el modelo socio-económico de la génesis del emprendedor es en términos de las motivaciones que favorecen el entorno social y laboral que rodean a las personas. En este sentido se han acuñado los conceptos de motivaciones *pull* y *push* (Genescá y Capelleras, 2004). El primer enfoque parte de la percepción de una oportunidad de negocio que aportaría gran valor competitivo y beneficio al emprendedor; mientras que el segundo tipo de motivación se deriva de la inexistencia de alternativas laborales deseables, es decir, por la imposibilidad de encontrar un trabajo dependiente satisfactorio. Aunque en la dimensión *push* tiene relación con el modelo de las turbulencias del mercado laboral, la diferencia está en que desde la perspectiva *push* la persona ha encontrado o tiene trabajo, el problema es que no está satisfecho con las recompensas que le proporciona.

## Actitudes emprendedoras

Al parecer, las personas que crean una empresa tienen mayores niveles de motivación *push* que *pull*. Por ejemplo, J. A. García (1997) hizo notar que el 20,4% de las empresas de nueva creación son fundadas por personas que trabajaban por cuenta ajena y que, viendo mermadas sus capacidades de promoción en una organización, intentaron mejorar sus condiciones laborales lanzándose al mundo de los negocios. Por su parte, destaca que tan sólo el 17,4% de las nuevas empresas han sido creadas con base en la materialización de una idea que se ha ido madurando y que podría ser rentable económicamente. Aunque las diferencias porcentuales entre uno y otro tipo de motivación son muy pequeñas, estos datos son consistentes con el modelo de las turbulencias del mercado, ya que la dificultad o imposibilidad de encontrar trabajo por cuenta ajena de calidad parece «empujar» a las personas a crear empresas.

## Modelo de la marginación social

La teoría de la marginación social sugiere que las personas que perciben un fuerte nivel de incongruencia entre sus atributos personales y el papel/es que desempeñan socialmente tenderán a cambiar o reconstruir su realidad social. En algunas ocasiones las personas tenderán a asociarse a movimientos religiosos, políticos o cualquier otra doctrina que les prometa redefinir su mundo en términos más apropiados respecto a su grupo social de referencia. En otros casos, la persona puede considerar el auto-empleo como medio para alcanzar el estatus social que se merecen (Stanworth et al., 1989).

Así pues, como señala (Veciana, 1989), las biografías de los grandes emprendedores se encuentran llenas de temas tales como la «huida de la pobreza, huida de la inseguridad, muerte y muerte repentina, padres que se fueron o padres que fueron abandonados por sus hijos». De este modo, una infancia difícil, marcada por un inicio temprano en la vida laboral y ciertos recuerdos traumáticos podrían estar a la base de una tendencia a la creación de empresas orientada a alcanzar el éxito social. Como señala Allport (1935) recordando los postulados psicoanalíticos, las experiencias dramáticas o los traumas pueden ser vivencias lo suficientemente impactantes como para fijar actitudes en las personas que permanecerán relativamente estables a lo largo de la vida. Ejemplos de este tipo de precariedad social lo podemos encontrar frecuentemente en los inicios de grandes compañías (Hewlett-Packard, Apple, Microsoft, etc.) o de personas con gran éxito

en el mundo de los negocios (p. e., Jesús Echevarría, Ricardo Soriano, Francisco Hernando, etc.).

Otro aspecto que refuerza este modelo es el hecho de que un considerable número de inmigrantes tienden a emplearse a sí mismos con la intención de salir de la marginación social (Veciana, 1989). Como señalan Arjona y Checa (2006) utilizando el modelo interactivo de desarrollo de negocios étnicos, tanto las características de los grupos de inmigrantes (situación de desventaja en el mercado de trabajo) como los factores de predisposición (puestos de trabajo precarios), convierten al colectivo inmigrante en un foco apropiado para el desarrollo de empresas. Así pues, el problema de acceso a determinados puestos de trabajo, unido a un contexto laboral hostil con pocas oportunidades que raya en la precariedad son las condiciones que generan las ideas emprendedoras de los inmigrantes; las cuales, presumiblemente, sirven para reducir la marginación social en la que se hayan. En palabras de Richerson y Boyd (2008), la inmigración produce un efecto de, entre otros, optimización económica en los países de acogida gracias a ese potencial emprendedor. Por su parte, la experiencia emprendedora también es valiosa para los países de origen cuando los emigrantes retornan ya que el conocimiento que han adquirido durante su migración sirve para engranar el tejido productivo de sus regiones o comarcas (Cassarino, 2004; Ruiz-Ruano y López, 2008a, 2008b).

#### **Modelo intergeneracional**

Para terminar, hay que destacar que existen muchas personas que acceden al mundo de los negocios por medio de familiares o parientes próximos. Esto es, muchas personas, principalmente en lo que se refiere a las pequeñas empresas, acceden al mundo empresarial porque sus padres trabajaron por cuenta propia en estos micronegocios. De este modo, además de haber sido fuertemente animados a integrarse en el negocio familiar, los potenciales emprendedores tienen una relativa facilidad para entrar en el mercado empresarial (Stanworth et al., 1989). De hecho, como señala Allport (1935) uno de los cuatro aspectos transversales de la formación de las actitudes es la imitación de los padres. En el caso de las actitudes hacia la creación de empresas, los hijos podrían ir asimilando una actitud favorable hacia la creación de empresas aunque no tengan experiencia directa con el negocio a edades tempranas de la vida.

El trabajo de Rogoff y Lee (1996) es un apoyo a la teoría intergeneracional ya que pudieron tipificar un perfil concreto de emprendedor cuya función era la de heredar las empresas familiares. El emprendedor heredero tiene un conjunto concreto de características que teóricamente vienen definidas por bajos niveles de creatividad en nuevos productos o servicios, bajos niveles de tendencia a contribuir a la sociedad y bajos niveles de orientación al crecimiento o expansión de la empresa. El resto de rasgos emprendedores identificado por los autores (interés financiero, ganancia de respeto o reconocimiento social, conciencia del riesgo, etc.) permanecen intermedios.

### 4.3.2. Dimensión psicológica

En lo que respecta a los modelos explicativos de las actitudes hacia la creación de empresas desde una óptica psicológica, se pueden encontrar modelos que van desde la perspectiva psicodinámica hasta los avances de la moderna neurociencia.

En cualquier caso, como apuntábamos anteriormente, la subdivisión que hacemos en diferentes parcelas que tratan de explicar el comportamiento emprendedor obedece a motivos expositivos ya que en algunos casos la separación de unos modelos y otros (ya sea dentro de esta misma dimensión psicológica o entre algunos modelos psicológicos y otros socio-económicos) es más bien difusa.

#### Modelo psicodinámico

Esta forma de ver al emprendedor tiene muchos puntos comunes con el modelo de la marginación social presentado anteriormente. Como señalan Stanworth et al. (1989), desde este modelo se considera que la actividad emprendedora es fruto de experiencias tempranas focalizadas en un entorno familiar infeliz que deja a la persona en una situación problemática entorno a la autoestima, inseguridad y auto-confianza. Como resultado, la persona desarrolla un marcado rechazo hacia la autoridad que pone en serio peligro el trabajo con los demás y como consecuencia se desarrolla la tendencia a trabajar como una unidad independiente.

Un elemento clave que ha sido ligado al desarrollo de actitudes emprendedoras como ha sido la necesidad de independencia (p.e., Genesca y Veciana, 1984; Veciana, 1989), podría estar a la base de un conflicto psicodinámico no resuelto. Desde esta perspectiva, la persona que tiene problemas para identificar el rol

autoritario en la figura materna o paterna tiende a rebelarse, en fases posteriores de su desarrollo psicológico, contra la autoridad. Siendo este el desencadenante de la tendencia a crear empresas como medio para materializar sus deseos de instaurar un sistema autoritario que no vivenció en su infancia.

No obstante, esta perspectiva ha sido criticada duramente por reducir la figura del emprendedor a una tipología estereotipada caracterizada por una personalidad incapaz de integrarse en la vida organizacional.

#### **Modelo de los rasgos**

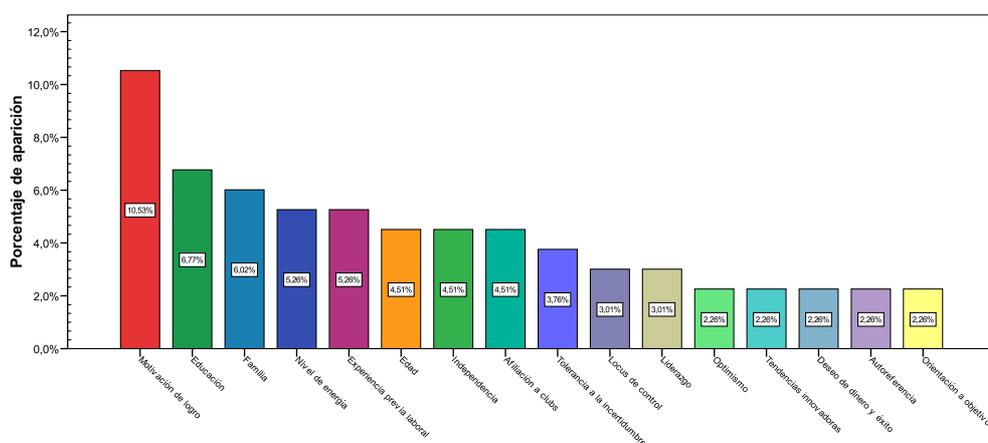
El modelo de los rasgos es probablemente el que más éxito haya tenido dentro del estudio de las actitudes hacia la creación de empresas y el que más volumen de trabajos ha suscitado. Desde esta perspectiva se defiende que las personas que llevan a cabo la creación de una empresa se caracterizan por un conjunto de rasgos de personalidad o un estado fijo de existencia concreto (Gartner, 1988).

Uno de los rasgos más legendarios que se han vinculado a la actividad emprendedora es el de la inteligencia. Por ejemplo, Gottfredson (1998) señala que los directivos empresariales que dependen de sí mismos forman parte del grupo de personas que muestran mayores puntuaciones en cocientes de inteligencia (CI). La característica que mejor definiría a estas personas en términos de su inteligencia sería una marcada tendencia al auto-aprendizaje. No obstante, dentro de los rasgos más investigados en el campo de las actitudes emprendedoras están el de la motivación de logro, nivel de energía, independencia, tolerancia a la incertidumbre, locus de control, liderazgo, optimismo, tendencias innovadoras, deseo de dinero-éxito, auto-referencia y orientación a objetivos (Figura 4.4).

La *motivación o necesidad de logro* es el constructo que más atención ha recibido (p.e., HayGroup y SAP AG, 2003; Huefner et al., 1996; Light, 2005; Stanworth et al., 1989; Veciana, 1989) y se refiere al «deseo de hacerlo bien», a la necesidad interna de conseguir una meta o un objetivo. El hallazgo más frecuente es que las personas caracterizadas por una mayor tendencia hacia la creación de empresas tienen niveles más altos de motivación de logro (McClelland, 1961); esto es, son personas que intentan conseguir sus objetivos por todos los medios. De hecho, la utilización de este constructo en el campo de los recursos humanos sirve para predecir la competencia y la productividad de las personas (Boyatzis, 1982).

Por su parte, el constructo *nivel de energía o energía personal (power)* hace

## Actitudes emprendedoras



**Figura 4.4:** Frecuencia de alusión a los elementos más importantes en la investigación sobre emprendedores. Basado en la revisión de Gartner (1988).

referencia a una fuerza interna que pueda ser utilizada con maestría para vencer contratiempos y reveses, superar obstáculos aparentemente insalvables y un sin fin de dificultades que tiene que superar cualquier persona que realiza un acto verdaderamente creativo (Veciana, 1989). La *independencia* es otro constructo que tiene que ver, como veíamos anteriormente, con la incomodidad por trabajar por cuenta ajena y una marcada antipatía hacia la autoridad.

Dos constructos muy relacionados entre sí y que han estado presentes en la investigación sobre los rasgos que caracterizan a las personas emprendedoras son la *tolerancia a la incertidumbre* y el *locus de control*. Como se vio al principio de este capítulo, ya los autores clásicos que comenzaron a plantear la importancia de los emprendedores dentro de los sistemas económicos hacían hincapié en el papel clave que juega la gestión de la incertidumbre en la creación de empresas (Cantillon, 1755/1931). Aunque la habilidad para gestionar situaciones inciertas puede no ser una característica determinante de la actividad emprendedora, si que es cierto que cierta finura para detectar modos orientados a minimizar los riesgos sería de gran ayuda al emprendedor. Por su parte, el *locus de control* o el nivel de *expectativas generalizadas de control* se refiere al modo en que una persona atribuye el origen de los refuerzos y castigos que recibe (Rotter, 1966). En lo que respecta las investigaciones sobre los niveles de locus de control en los emprendedores, los resultados parecen indicar que las personas que emprenden la creación de una empresa tienen mayores niveles de locus de control interno. Esto es, tienden a atribuir el éxito y el fracaso a ellos mismos. Por ejemplo, Sánchez

(2003) destaca que dentro de la naturaleza multifacética del autoempleado, el locus de control interno tiene una gran relevancia.

Existen muchos otros constructos que han sido relacionados con la actividad emprendedora. De hecho existen tantos que esta perspectiva ha sido criticada por ser tan amplia y por no haber dado una respuesta a las preguntas típicas que se plantean las ciencias sociales en relación a la naturaleza y función del emprendedor. En los años recientes, y con el objetivo de arrojar más luz sobre esta cuestión, se han introducido los últimos avances en el modelado estadístico estocástico al estudio de las actitudes emprendedoras desde el punto de vista de los rasgos psicológicos (J. García et al., 2005, 2006, 2007).

#### **Modelo conductual**

Como consecuencia de las dificultades que ha experimentado la perspectiva de los rasgos para explicar la génesis y la funcionalidad del emprendedor, Gartner (1988) introdujo la idea del *modelo conductual* en el estudio de la tendencia hacia la creación de empresas. En este caso, el énfasis no se pone en *qué* es la persona (esto es, sus rasgos de personalidad), sino en lo que *hace*. De este modo se retomaba la idea planteada por los autores clásicos en relación al papel que desempeña el emprendedor en la sociedad, volcando el interés en la funcionalidad de su trabajo (Hébert y Link, 1989).

No obstante, la perspectiva conductual parece no haber tenido una gran repercusión en la investigación de las actitudes emprendedoras. Ello puede ser debido a los problemas metodológicos que se derivan del intento de estudiar al emprendedor desde una perspectiva como esta o a aspectos sustantivos. Por ejemplo, McKenzie et al. (2007) han retomado el modelo de los rasgos y han criticado la perspectiva de Gartner por ser demasiado restrictiva respecto a lo que es un emprendedor y por descontextualizada.

#### **Modelo de los sesgos**

Otra observación consistente que señalan los trabajos orientados a estudiar la actividad emprendedora es que las personas que desencadenan la creación de una empresa parecen tener una visión sesgada de la realidad (Light, 2005). Ya sea en su habilidad para percibir y gestionar la incertidumbre o para estimar las

## Actitudes emprendedoras

tasas de riesgo en sus decisiones, numerosos trabajos han señalado las desviaciones sistemáticas que comenten los empresarios de éxito cuando razonan (p. e., HayGroup y SAP AG, 2003; Hébert y Link, 1989; Veciana, 1989).

Por ello, se ha acuñado el modelo de los sesgos sobre los emprendedores haciendo referencia al modelo de los heurísticos y los sesgos surgido en la psicología cognitiva de los años setenta del pasado siglo (p. e., Kahneman, 2003; Kahneman et al., 1982; Kahneman y Tversky, 1973; Tversky y Kahneman, 1974, 1982). Más concretamente, este modelo podría girar entorno al concepto de intuición. De esta manera se soslayaría el concepto peyorativo asociado a la idea de sesgo y se exaltaría la habilidad de razonamiento «agudo» por parte de los emprendedores.

Como señala Veciana (1989), el empresario es altamente intuitivo. Aunque durante mucho tiempo la intuición tuvo mala prensa dentro del mundo de los negocios, este concepto se va tildando de un halo de sofisticación entre los directivos de empresas. Como señala Gigerenzer (2007), la intuición es una forma de inteligencia sin pensamiento consciente sin la que no podríamos tomar decisiones. En el mundo de los negocios la intuición es omnipresente ya que, en la mayoría de los casos, una decisión racional al 100 % implicaría calibrar todos los factores, sopesar todas las consecuencias, valorar todos los costes y beneficios, cruzar fríamente miles de variables; en definitiva, supondría realizar una serie de cálculos de dimensiones astronómicas. Este tipo de tarea es un ejemplo para lo que no estamos preparadas las personas desde un punto de vista biológico (Jastrow, 1981/1993). Así pues, los emprendedores se basarían en su intuición para tomar las decisiones y utilizan los informes racionales para apoyar sus decisiones. Las investigaciones más recientes en este sentido se centran en estudiar los patrones neurales que se ponen en marcha cuando tomamos decisiones en condiciones de riesgo (p.e., Hsu, Bhatt, Adolphs, Tranel, y Camerer, 2005; D. Lee, 2005; McCoy y Platt, 2005; Rustichini, 2005).

Lawrence, Clark, Labuzetta, Sahakian, y Vyakarnum (2008) proponen, basándose en la diferenciación clásica que se establece entre ejecutivos o directivos y emprendedores (Samuelson, 1970), que los emprendedores son personas que llevan a cabo comportamientos altamente arriesgados de manera adaptativa y que las consecuencias que se derivan de su conducta arriesgada son de gran valor en contextos económicos de toma de decisiones. Lawrence et al. (2008) sometieron a un grupo de directivos empresariales y a un grupo de emprendedores a

dos tipos de tareas de toma de decisiones que implicaban procesos de decisión «razonados» (decisiones *frías*) o arriesgados (decisiones *calientes*)<sup>4</sup>. Observaron que los emprendedores arriesgaron más en tareas calientes mientras que no apreciaron diferencias estadísticamente significativas entre ambos grupos en tareas que implicaban toma de decisiones frías. Por su parte, Hauser (2008) ha señalado que este tipo de tendencia que caracteriza a las personas emprendedoras puede moldearse en contextos donde se fomente la actitud emprendedora.

## 4.4. Actitudes emprendedoras en la universidad española

Fishbein y Ajzen (1975) hicieron notar a mediados de los años setenta del pasado siglo que el estado-del-arte en la investigación sobre actitudes se encontraba en una situación similar a como lo estuvo cuarenta años antes cuando Gordon Allport (1935) reclamaba la perentoriedad de las actitudes en el quehacer de la psicología. En la última década del siglo XX el estudio de las actitudes seguía siendo un elemento clave dentro del campo de la psicología (Deaux, Dane, y Wrightsman, 1993), aunque no sin pocas vicisitudes en el área de la definición conceptual y metodológica de la actitud.

Como señalan (Deaux et al., 1993), la idea de *actitud* se utilizó en el pasado para referirse a una postura física y a la connotación simbólica que esta tenía. Posteriormente, la actitud se utilizó para referirse a la predisposición de respuesta ante la presentación de estímulos en las tareas de tiempos de reacción. La idea de entender la actitud como un concepto abstracto que implica procesos y estructuras mentales internas o latentes es la más reciente. En cualquier caso, podemos considerar la actitud como *una predisposición aprendida de responder favorable, mixta o desfavorablemente frente a un objeto, persona o idea en particular* (Brehm, Kassin, y Fein, 2005; Feldman, 1998).

A parte del problema que supone la definición de actitud surge el problema de las dimensiones que la integran. Algunos autores clásicos como Allport (1935)

---

<sup>4</sup>La dimensión *fría* se evaluó usando la *Tarea de la Torre de Londres* y se refiere a decisiones emocionalmente neutrales donde no hay que sopesar las recompensas y los castigos que supone una decisión. Por su parte, la dimensión *caliente* (o arriesgada) de la toma de decisiones se valoró con la *Tarea de la Apuesta de Cambridge* y se refiere decisiones donde las recompensas y los castigos están asociadas a resultados emotivamente significativos.

## Actitudes emprendedoras

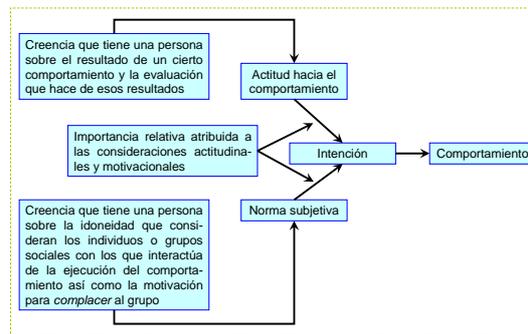
consideran que la actitud es un fenómeno psicológico multidimensional. Por ello, más recientemente se ha acuñado la idea de (basándose en la concepción platónica de la trilogía pensamiento-sentimiento-acción) que la actitud se define como un constructo complejo y multidimensional integrado por tres componentes: *cognitivo*, *afectivo* o *emocional* y el *conductual* o *conativo*. Esta concepción multifacética de la actitud ha recibido numerosas críticas. Por ejemplo, el componente afectivo de la actitud ha sido criticado porque depende del contexto y del objeto evaluado. Por ejemplo, al evaluar la actitud hacia el sabor de las anchoas el componente emotivo puede resultar trivial, aunque si evaluamos la actitud hacia los padres el componente emocional puede tener gran relevancia. No obstante, nosotros vamos a seguir utilizando la perspectiva teórica tridimensional de la actitud ya que sigue siendo una concepción vigente en la actualidad (p.e., Ajzen y Fishbein, 2005; Feldman, 1998; Franzoi, 2005; J. F. Morales, Reboloso, y Moya, 1994). De hecho, las tres variables de respuesta que se van a incluir en los modelos estimados van a representar, en cierto modo, las tres dimensiones clásicas de la actitud (ver la sub-sección de materiales en el capítulo de metodología general).

Por su parte, otro aspecto central del estudio de las actitudes es su relación con el comportamiento. En nuestro caso, vamos a considerar la *Teoría del Comportamiento Planeado* (TCP) como argumento de fondo, en términos de la relación que se establece entre actitud y comportamiento, sobre el que se basará nuestro análisis (Ajzen y Fishbein, 2005). El antecedente de la TCP se conoce como *Teoría de la Acción Razonada* (TAR) (Ajzen y Fishbein, 1980); no obstante, ambas teorías suponen que el comportamiento ha sido razonado, que las personas piensan sobre las consecuencias de sus actos y que toman decisiones orientadas a alcanzar unos resultados y a evitar otros.

Según la concepción inicial de la TAR, el comportamiento está determinado directamente por la intención. A su vez, la intención es el producto de la combinación de la actitud hacia dicho comportamiento y el conjunto de normas subjetivas que ha interiorizado el individuo en relación a los grupos de personas con los que se relaciona. Estos dos determinantes de la intención están modulados por la consideración que da la persona a los aspectos actitudinales y normativos (Figura 4.5).

Por su parte, la TCP se concibió primeramente como un modelo que establecía que la intención de conducta dependía de la actitud y del control percibido de la

#### 4.4 Actitudes emprendedoras en la universidad española



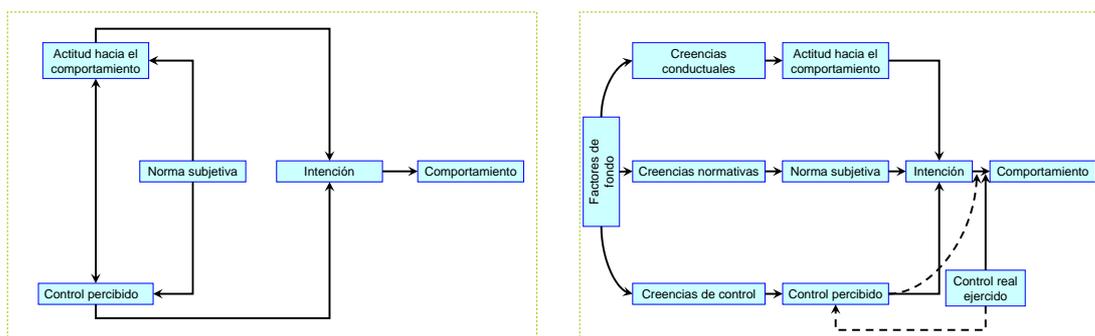
**Figura 4.5:** Modelo de la teoría de la acción razonada que relaciona la actitud y el comportamiento. Adaptada de Ajzen y Fishbein (1980).

situación. Por su parte, el control percibido y la actitud dependían de las normas subjetivas. Estas normas subjetivas se referían, por un lado, a las creencias que posee el individuo sobre lo que la sociedad demanda; y, por otro, a la motivación de la persona para acomodarse a las normas sociales. Por su parte, la actitud hacia el comportamiento se consideraba el fruto de las creencias que poseía la persona respecto al comportamiento y a la previsión de los resultados que se obtendrían al ejecutar la conducta (Figura 4.6 izquierda).

El modelo más reciente de la TCP (Ajzen y Fishbein, 2005) señala que la intención de conducta (como determinante del comportamiento) depende de la actitud hacia el comportamiento, de las normas subjetivas y del control percibido sobre el conjunto de reforzadores contextuales (Figura 4.6 derecha). No obstante, este último modelo también añade un factor modulador de la relación que se establece entre la intención y la conducta: el control real sobre las consecuencias del comportamiento. Por otro lado, este nuevo modelo incorpora las creencias conductuales, normativas y de control como antecedentes de la actitud, norma subjetiva y control percibido respectivamente. Por último, el modelo propone que las creencias dependen de factores de fondo individuales (personalidad, inteligencia, experiencia, etc.), sociales (educación, edad, sexo, religión, cultura, etc.) e informativos (conocimiento, medios de información, etc.).

En el contexto de la investigación en actitudes emprendedoras, Genescá y Cappel­leras (2004) han observado recientemente que el nivel educativo es un predictor importante del éxito de las empresas de nueva creación en España. Es más, hicieron notar que las empresas españolas creadas por titulados universitarios tienen una mayor probabilidad de éxito y que, de éstas, son las que fueron fundadas por

## Actitudes emprendedoras



**Figura 4.6:** Modelos de la teoría del comportamiento planeado que relaciona la actitud y el comportamiento. A la izquierda aparece el modelo más clásico y a la derecha el más actual. Adaptado de Deaux et al. (1993) y Ajzen y Fishbein (2005) respectivamente.

personas con estudios de tercer ciclo las que muestran mayor estabilidad temporal y de negocio. Sánchez (2003) también ha destacado la importancia de la formación en el perfil emprendedor y ha señalado que «el autoempleado del siglo XXI tiene que ser una persona que se encuentre con una formación sólida» (p. 243).

Por su parte, como señalan Gómez, Mira, y Martínez (2007), uno de los condicionantes más importantes de la actitud emprendedora es el nivel formativo o educativo. Este rasgo de la actividad emprendedora ha impulsado el desarrollo de iniciativas, por parte de las instituciones educativas de nivel superior, que catalicen la creación de empresas desde el ámbito universitario (Díaz, 2003).

Así pues, la enseñanza universitaria se perfila como un elemento potente dentro de la formación de personas potencialmente emprendedoras. La formación superior facilitada por las universidades tiene como uno de sus objetivos la estimulación de una visión emprendedora en relación a los conocimientos que proporciona y por ello es útil aproximarse al estudio de las actitudes emprendedoras en la universidad (Cano et al., 2003). De este modo, como señalan Peñas y Quijano (2008), la universidad, junto a los demás agentes sociales implicados en el proceso, tiene la responsabilidad de fomentar el espíritu emprendedor para hacer productivo el conocimiento que allí se maneja. Este objetivo cobra más importancia cuando se detecta que ciertos conocimientos adquiridos en el contexto educativo superior no repercuten directamente en una mayor eficiencia en el puesto de trabajo (p. e., Wong, 2008). En una investigación reciente, por ejemplo, Grundling

#### 4.4 Actitudes emprendedoras en la universidad española

y Steynberg (2008) han desarrollado un procedimiento para utilizar la universidad como una plataforma de lanzamiento para las ideas emprendedoras de base tecnológica. En concreto, desarrollaron un modelo que consta de tres fases en las que establece un sistema de información que permite gestionar de modo eficiente la información y las actividades orientadas al desarrollo de productos o servicios.

Cano et al. (2003) y J. García et al. (2005) han propuesto que la actitud hacia la creación de empresas en estudiantes universitarios consta de varias facetas: creatividad, perseverancia, capacidad de organización, independencia, confianza en sí mismo, riesgos calculados, tolerancia a la incertidumbre, actitud competitiva, capacidad de negociación y locus de control. Estas dimensiones de la actitud han sido validadas con posterioridad en otros estudios utilizando modelos de ecuaciones estructurales y redes bayesianas (J. García et al., 2007). En un trabajo de corte diferente Domingo, Bajo, Chiloeches, y García (2008) observaron que la actividad virtual de crear una empresa ficticia en estudiantes de primer curso universitario produjo un aumento en la confianza en el futuro laboral de los participantes.

En un trabajo reciente con estudiantes universitarios, J. García et al. (2006) observaron que la actitud hacia la creación de empresas dependía del grado de facilidad percibida para crear una empresa, del grado en que se ha considerado seriamente la creación de una empresa y de la deseabilidad subjetiva de crear una empresa. En este trabajo también se puso de manifiesto que hay otras variables relacionadas indirectamente con la actitud general hacia la creación de empresas como la titulación cursada, el sexo o el tipo de trabajo que a una persona le gustaría desempeñar.

Actitudes emprendedoras

**Parte II**

**Trabajo Empírico**



# Capítulo 5

## Metodología general

*"[...] (he) was not interested at all in the changes. It existed, that was all that mattered to him".*

(Highsmith, 1955/1999)

### 5.1. Contextualización de la investigación

En el caso particular de este trabajo, investigar la idoneidad de un modelo estadístico u otro para predecir componentes de la actitud hacia la creación de empresas tiene su justificación sustantiva en la importancia que tienen estos aspectos psicológicos como predictores del comportamiento. Por su parte y en última instancia, el comportamiento emprendedor tiene consecuencias en el desarrollo económico. Como así lo hizo notar McClelland (1961) hace casi medio siglo, los expertos actuales en economía hacen hincapié en que los factores psicológicos juegan un papel crucial para entender el crecimiento, las crisis y el declive en el desarrollo económico (Buchanan, 2009; Farmer y Foley, 2009; Wald, 2008). Concretamente, se utilizarán dos técnicas estadísticas diferentes para predecir las tres dimensiones de la actitud emprendedora (emocional, cognitiva y conductual) y se pondrán de relieve las condiciones en las cuales una u otra técnica genera mejores resultados.

Por lo tanto, la virtud de este trabajo reside en su perspectiva sustantiva, dado que explorará la idoneidad de diferentes modelos estadísticos para predecir los componentes de la actitud emprendedora; y en su perspectiva metodológica, ya que comparará dos técnicas estadísticas que se sustentan en filosofías relativamente diferentes para poner de manifiesto los puntos fuertes y débiles de cada una de ellas bajo diferentes circunstancias.

## 5.2. Objetivos

El objetivo central de esta tesis es comparar la ejecución de la regresión logística y las redes bayesianas para predecir los tres componentes básicos de la actitud emprendedora en términos de eficiencia predictiva. Este trabajo está legitimado porque varios trabajos sugieren que los modelos basados en estructura de red podrían tener ventajas frente a la regresión logística en términos predictivos (p. e., Ankarali, Canan, Akkus, Bugdayci, y Ali, 2007; Bartfay, Mackillop, y Pater, 2006; Eftekhar, Mohammad, Ardebili, Ghodsi, y Ketabchi, 2005; Finch y Schneider, 2007; J. García et al., 2007; Jaimes, Farbiarz, Alvarez, y Martínez, 2005; Kumar, Rao, y Soni, 1995; Terrin, Schmid, Griffith, D'Agostino, y Selker, 2003). No obstante, no hay evidencias robustas que muestren una mejor ejecución de las técnicas basadas en modelos de red frente a la regresión logística.

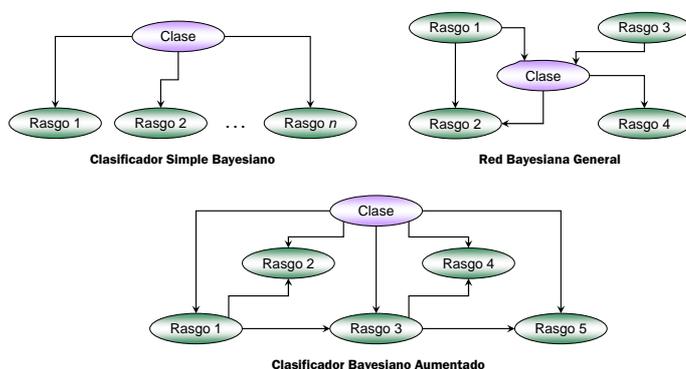
Por su parte, los trabajos que comparan las redes bayesianas y la regresión logística no son numerosos según la bibliografía revisada. Por ejemplo, S. M. Lee et al. (2005) hicieron notar que el uso de las redes bayesianas podrían representar ciertas ventajas frente al uso de la regresión logística. Entre las ventajas que suponen las redes bayesianas frente a la regresión logística cabe destacar la superación de ciertos supuestos estadísticos como el de aditividad, la facilidad para el manejo de una gran cantidad de predictores, la facilidad para identificar e interpretar los efectos de interacción y la facilidad para modelar relaciones no lineales entre variables (S. M. Lee et al., 2005). En la Tabla 5.1 aparecen algunas de las diferencias que existen entre redes bayesianas y regresión logística. Por otro lado, en un trabajo reciente (J. López, Ruiz-Ruano, y García, 2008b) hemos observado que las redes bayesianas son más eficientes en la predicción que la regresión logística. Sin embargo, no fue el objetivo de éste trabajo comparar ambas técnicas formalmente, sino analizar un conjunto de datos con técnicas de análisis

**Tabla 5.1:** Regresión logística vs red bayesiana. Adaptado de S. M. Lee et al. (2005).

	REGRESIÓN LOGÍSTICA	RED BAYESIANA
<i>Aproximación</i>	Basada en la verificación de hipótesis	No basada en la hipotetización
<i>Tipos de estudios</i>	Experimentales / observacionales con un reducido número de variables	Toleran un gran número de variables
<i>Términos usados</i>	Variable independiente (predicador), variable dependiente (respuesta)	Nodo, nodo aleatorio, nodo de información, nodo objetivo
<i>Asunciones</i>	Linealidad para los <i>logits</i> y aditividad	Ninguna
<i>Eficiencia</i>	Necesita que se especifiquen las interacciones y las variables confusoras	La detección de interacciones y variables confusoras puede simplificarse gracias al método gráfico
<i>Selección de variables</i>	La necesidad de especificar las variables puede crear el riesgo de introducir sesgos en la selección de variables relevantes	Los algoritmos de estimación estructural pueden ayudar en la selección de variables importantes
<i>Error</i>	Puede ser influenciado por los errores de los investigadores debido a la ausencia de conocimiento sustantivo	Puede ser reducido utilizando bases de datos y algoritmos de estimación
<i>Datos perdidos</i>	Problemas para predecir con datos incompletos	Predicción aceptable con datos incompletos
<i>Interpretación</i>	Coefficientes, ventajas relativas	Relaciones gráficas, probabilidades, probabilidades posteriores

diferentes.

En concreto, en este trabajo se utilizará el modelo de regresión logística binaria y el clasificador ingenuo de Bayes (*Bayes Naïve Classifier*) (I. Martínez y Rodríguez, 2003), clasificador simple de bayes (Domingos y Pazzni, 1996) o, como la hemos llamado antes, red divergente. Aunque recientemente se han desarrollado técnicas orientadas a incorporar la filosofía de la estadística bayesiana en el análisis de regresión logística binaria (p. e., Genkin, Lewis, y Madigan, 2005; A. J. Ortiz, Martín, Ureña, y García, 2005), este trabajo se centrará en comparar la regresión logística binaria clásica y el clasificador simple de Bayes por dos razones principales. En primer lugar, ambas técnicas surgieron en la misma época, mientras que la versión bayesiana de la regresión logística es un fenómeno



**Figura 5.1:** Taxonomía de clasificadores de red bayesiana. Adaptado de Shen et al. (2003).

más reciente. En segundo lugar, la regresión logística binaria clásica está más extendida que su homóloga bayesiana, con lo cual las comparaciones pueden ser más útiles. En lo que respecta al clasificador ingenuo de Bayes, se ha seleccionado este modelo por ser el que más se parece desde un punto de vista formal a la regresión logística (Greiner et al., 2005; Greiner y Zhou, 2002; Shen et al., 2003). No obstante, hay que hacer notar que esta no es la única forma en la que se puede presentar un clasificador de red bayesiana. Por ejemplo, como se puede apreciar en la Figura 5.1, el clasificador simple de Bayes consiste en un nodo o variable de clasificación y dos o más nodos o variables que representan rasgos; mientras que el clasificador bayesiano aumentado (*Tree Augmented Naïve-Bayes*) tolera que los rasgos estén relacionados entre sí.

Aquí se considera que tanto la regresión logística como la red bayesiana son herramientas de clasificación; esto es, funciones que asignan una etiqueta de clase a ejemplos, típicamente descritas o caracterizadas por un conjunto de atributos (Shen et al., 2003). Por eficiencia predictiva se entenderá, en términos generales, el grado en que una herramienta de clasificación proporciona respuestas correctas de manera frecuente (Greiner et al., 2005). Más concretamente, la validez predictiva se entenderá como el grado en que cada modelo tiene niveles óptimos de sensibilidad, especificidad, falsos negativos, falsos positivos, valores predictivos positivos, valores predictivos negativos o tasas globales clasificaciones correctas.

No obstante, las dos técnicas comparadas en este trabajo producen diferentes tasas de clasificaciones correctas dependiendo del umbral de corte establecido para la clasificación. Así pues, en una primera fase describiremos el comportamiento

de estas técnicas utilizando un valor de corte arbitrario establecido en 0,5 como se suele hacer comúnmente (p. e., Domingos y Pazzni, 1996). Posteriormente, utilizaremos curvas ROC para comparar los niveles de especificidad y sensibilidad para diferentes puntos de corte en las probabilidades proporcionadas por ambos modelos (p. e., DeMaris, 2002; Hanley y McNeil, 1982, 1983).

En primer lugar, presentaremos un estudio inicial donde se construyen los modelos más apropiados de red bayesiana y regresión logística tanto desde un punto de vista estadístico como sustantivo. A continuación, evaluaremos el efecto que tiene el tamaño de la muestra sobre la bondad predictiva en ambas técnicas. Seguidamente, se tratará la influencia del nivel de categorización de las variables independientes sobre la validez predictiva. En tercer lugar, se investigará el efecto que produce el número de casos perdidos en términos de bondad predictiva en ambas técnicas. Todos los estudios comparativos entre técnicas se llevarán a cabo utilizando curvas ROC.

## 5.3. Método

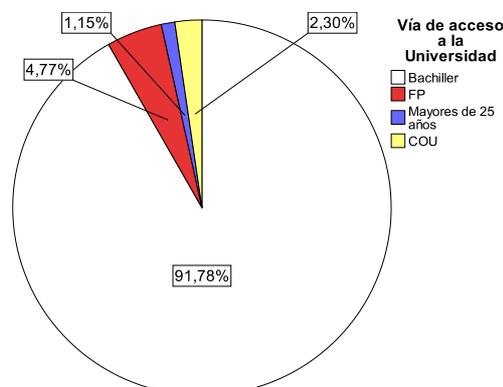
### 5.3.1. Participantes

Se realizó un muestreo por bloques con afijación proporcional por sexo y titulación. La titulación se consideró una variable agregada definida por las tres familias genéricas: a) ciencias humanas y jurídicas, b) ciencias empresariales y c) titulaciones técnicas. El muestreo se realizó sobre los alumnos de la Universidad de Almería en los primeros y últimos cursos de cada titulación<sup>1</sup> durante el curso académico 2005-2006. El error muestral fue del  $\pm 3\%$  con un intervalo de confianza del 95 %,  $z = 1,96$  y  $p = q = 0,5$  para el total de la muestra.

La muestra estuvo compuesta por 1230 participantes de los cuales 426 fueron hombres (34,6 %) y 797 (64,8 %) mujeres, el 0,6 % (7) restante de los participantes no indicaron su sexo. Las edades de los participantes estuvieron comprendidas entre los 17 y los 56 años con una media de 22,45 y una desviación típica de 4,46. La mayor parte de los participantes residían en la ciudad de Almería (65,7 %).

---

<sup>1</sup>Se eligieron estos cursos con el objetivo de estudiar las diferencias entre estudiantes «novatos» y «veteranos» en relación a las actitudes emprendedoras dentro de un proyecto más general que no será tratado aquí.



**Figura 5.2:** Vía de acceso a la universidad de los participantes.

### Perfil académico y laboral

La mayor parte de los participantes ingresaron en la universidad por medio del Bachillerato o a través de algún Ciclo Formativo Profesional (Figura 5.2).

Respecto a las titulaciones que cursaban los participantes, en la Tabla 5.2 aparecen el porcentaje de estudiantes por titulación así como las frecuencias en términos de las familias genéricas definidas al diseñar el muestreo<sup>2</sup>. El 8,9% de los participantes indicaron que poseían otra titulación universitaria además de la que estaban cursando.

La mayor parte de los encuestados se dedican únicamente a estudiar (78,3%, 963). Un 1,5% (18) de los entrevistados manifestó compaginar los estudios con un trabajo por cuenta ajena para una empresa y un 1,8% (22) afirmó trabajar para la función pública. No obstante, el 14,1% (173) de los entrevistados indicó que tenía experiencia laboral por cuenta ajena. Este colectivo muestra un promedio de 4,34 ( $DT = 5,74$ ) años de experiencia y lo más frecuente es que las personas hayan trabajado para una (5,4%), dos (4%) o tres (2%) empresas. Como se puede apreciar en la Figura 5.3 la mayor parte de las personas que trabajan por cuenta ajena desempeñan un trabajo de media o baja cualificación y mayormente lo hacen en el sector privado. Por último, un 1,2% (14) de los participantes manifestaron ser empresarios o profesionales independientes que tienen a su cargo entre una y 20 personas ( $M = 7,43$ ,  $DT = 7,68$ ).

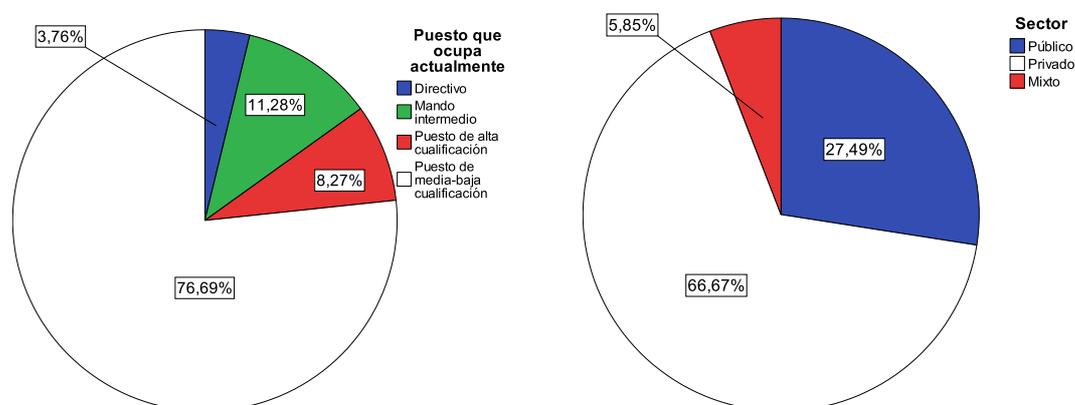
<sup>2</sup>Tres personas indicaron ser estudiantes del Programa Erasmus (un 0,24%) aunque no aparecen en la Tabla.

Tabla 5.2: Estudios cursados por los participantes.

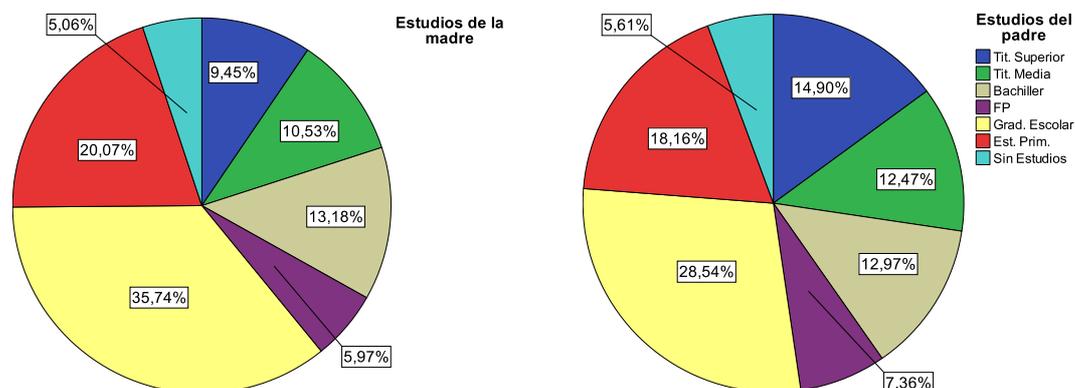
CLUSTER	ESTUDIOS	<i>fr</i>	%
HJ	Magisterio: Educación Infantil		8,37
	Licenciatura en Derecho		7,32
	Licenciatura en Psicología		7,24
	Diplomatura en Relaciones Laborales		4,72
	Magisterio: Lengua Extranjera		4,63
	Diplomatura en Enfermería	557	4,15
	Licenciatura en Filología Inglesa		3,33
	Licenciatura en Psicopedagogía		1,95
	Licenciatura en Ciencias del Trabajo		1,22
	Magisterio: Educación Física		1,06
	Licenciatura en Filología Hispánica		0,81
Magisterio: Educación Musical		0,49	
CE	Diplomatura en Ciencias Empresariales		13,74
	Licenciatura en Administración y Dirección de Empresas	323	8,29
	Diplomatura en Gestión y Administración Pública		2,11
	Diplomatura en Turismo		2,11
TT	Ingeniería Técnica Agrícola: Hortofruticultura y Jardinería		5,85
	Licenciatura en Ciencias Ambientales		3,41
	Ingeniería Técnica Informática de Gestión		3,09
	Ingeniería Técnica Informática de Sistemas		2,76
	Licenciatura en Química		2,20
	Licenciatura en Matemáticas		2,03
	Ingeniería Química	344	1,95
	Ingeniería Informática		1,79
	Ingeniería Agrónoma		1,63
	Ingeniería Técnica Agrícola: Mecanización y Construcción		1,22
	Ingeniería Técnica Agrícola: Industrias Agroalimentarias		1,14
Ingeniería Técnica Agrícola: Explotaciones Agrícolas y Ganaderas		0,81	
Ingeniería de Materiales		0,08	
	Total válido	1227	99,76
	Datos perdidos	3	0,24

HJ: Ciencias Humanas y Jurídicas, CE: Ciencias Empresariales y TT: Titulaciones Técnicas

## Metodología general



**Figura 5.3:** Puesto ocupado y sector de actividad donde trabajan los participantes.



**Figura 5.4:** Nivel de estudios familiar. Tit.: Titulación, FP: Formación Profesional, Grad.: Graduado, Est. Prim.: Estudios Primarios.

## Perfil familiar

Como se puede observar en la Figura 5.4, la proporción de madres que tienen estudios primarios o han conseguido el Graduado Escolar es mayor frente a la de los padres. Por el contrario, es en el grupo de los padres donde se encuentra mayor proporción de personas que han cursado programas de formación profesional. En lo que respecta a la formación universitaria los padres también muestran mayores proporciones. Mientras que los padres de los participantes superan a la proporción de madres que tienen titulaciones medias en alrededor de un dos por ciento, cuando hablamos de titulación superior los padres superan en más de un 5% a las madres con este nivel de estudios.

Casi una cuarta parte de los padres de los participantes eran empleados de una

empresa. Otras profesiones como la de empresario agrícola, pesquero o forestal y empresario de industria o servicios suman juntas casi otra cuarta parte de la muestra. Casi un 20 % de los participantes manifestó que sus padres trabajaban en la función pública y entorno al 10 % indicaron que sus padres eran profesionales de oficio (Tabla 5.5).

La mayoría de las madres de los participantes se dedicaban a labores domésticas mientras que casi un 20 % de los encuestados indicaron que sus madres eran empleadas de una empresa. Casi un 12 % de los participantes manifestaron que sus madres eran empresarias bien en el sector primario o bien en el sector industrial o de servicios. Por último, poco más del 14 % de los participantes señalaron que sus madres trabajaban para la función pública.

En los casos en los que los padres son empresarios, éstos suelen tener más empleados a su cargo ( $M = 12,80$ ,  $DT = 46,34$ , rango 0-500) que cuando lo son las madres ( $M = 5,06$ ,  $DT = 13,57$ , rango 0-100).

### 5.3.2. Materiales

Para la recogida de información se utilizó un cuestionario de tres folios tipo A4 impresos a doble cara y grapados por su esquina superior izquierda. En la primera página del formulario aparecieron los membretes del Consejo Social de la Universidad de Almería, del Servicio Universitario de Empleo y de la Fundación Mediterránea Empresa-Universidad de Almería. A continuación aparecía una breve presentación de finalidad del estudio y un requerimiento formal para colaborar

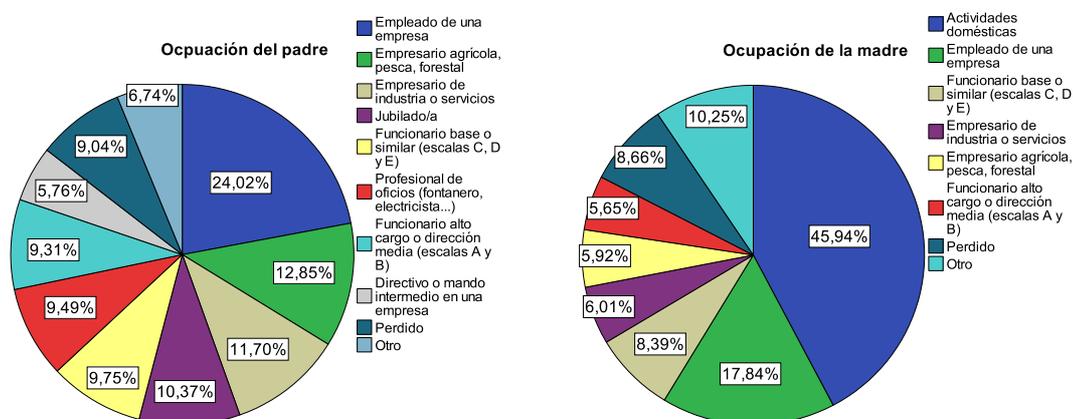


Figura 5.5: Ocupación principal de los padres.

## Metodología general

en el mismo. También se indicó que los datos recogidos con el cuestionario serían tratados confidencialmente y se facilitó información de contacto con los responsables de la investigación por si alguien tenía alguna duda. Para terminar se agradeció la colaboración en la investigación. Una versión completa de la portada del formulario puede encontrarse en la página 263 del Anexo.

En el reverso de la portada aparecieron los espacios dedicados a recoger los datos sociodemográficos y una pregunta referida a la preferencia laboral. En la siguiente página se mostraron varias preguntas relativas a la intención de crear una empresa estándar o sin ánimo de lucro y una pregunta para evaluar el grado en que los participantes consideraban más o menos fácil la creación de una empresa respecto al pasado. Las tres primeras preguntas de esta segunda página son las variables sobre las que se han basado los modelos de regresión logística y de red bayesiana en términos predictivos. Esto es, cada una de estas variables fueron consideradas (una a una) como variables dependientes, o de respuesta, en los modelos de regresión logística y como nodos clase, o divergentes, en los modelos de red bayesiana. Las preguntas fueron:

1. ¿Considera deseable crear una empresa propia?
2. ¿Ha pensado seriamente, como una opción real a corto/medio plazo, montar su propia empresa?
3. ¿Ha iniciado en algún momento acciones encaminadas a montar su propia empresa?

Estas preguntas fueron elegidas porque representan las tres dimensiones de la actitud propuestas por el modelo actitudinal expuesto anteriormente. De este modo, la pregunta número uno hace referencia al componente emotivo de la actitud ya que está orientada a evaluar la deseabilidad o agradabilidad ante la creación de empresas (DES). Por su parte, la segunda pregunta se referiría al componente cognitivo de la actitud (COG) hacia la creación de empresas mientras que la tercera pregunta se referiría a la dimensión conductual (CON). Estas tres preguntas fueron respondidas por los participantes en términos de «sí» o «no». Para responder a cada pregunta los participantes tuvieron que marcar la respuesta deseada marcando en una casilla de verificación que aparecía a la izquierda de cada posible alternativa de respuesta.

Bajo este conjunto de preguntas relativas a la creación de empresas apareció, en primer lugar, una mini-escala de carencias formativas percibidas (C) y a continuación otra escala breve de preparación percibida (P); ambas relevantes para el objetivo de este trabajo ya que se consideraron como variables independientes, o factores, para la regresión logística y como nodos hijo, o de evidencia, en la red bayesiana. La escala C constó de 12 ítems puntuables en una escala tipo Likert de cinco alternativas que indicaban el grado de conocimiento de diferentes aspectos organizacionales relativos a la creación de empresas. La escala P tuvo cuatro ítems orientados a evaluar la preparación percibida por los participantes en relación a los conocimientos experiencia contactos y en términos generales para crear una empresa. Cada ítem se puntuó en una escala tipo Likert de cinco alternativas que osciló desde muy bajo a muy alto. En la página 265 se pueden ver los ítems tal y como fueron presentados en la escala C y en la página 266 aparecen los de la escala P. En ambos casos los ítems estuvieron expresados en sentido directo indicando una mayor puntuación en el ítem una mayor carencia percibida o un mayor nivel de preparación percibida respectivamente.

En el reverso de esta segunda página aparecieron las escalas de motivación para crear un negocio propio y de obstáculos percibidos frente a la creación de una empresa. La escala de motivación para crear un negocio propio (M) constó de 11 ítems puntuables en una escala tipo Likert de cinco alternativas, mientras que la escala de obstáculos percibidos ante la creación de una empresa contuvo 17 ítems puntuables del mismo modo. En ambas escalas todos los ítems estuvieron expresados en sentido directo indicando una mayor puntuación en cada ítem una mayor motivación para crear una empresa o un mayor grado de obstáculos relacionados con la creación de una empresa. En la página 267 del Anexo aparece la escala M y en la página 268 aparece la escala O.

En la tercera página apareció la Escala de Actitudes hacia la Creación de Empresas (ACEM, en adelante A) (Cano et al., 2003; J. García et al., 2005, 2006, 2007). La escala A consta de 29 ítems tipo Likert de cuatro alternativas sobre la frecuencia con que se realizan ciertos comportamientos o pensamientos. La escala consta de 13 ítems inversos (1, 2, 3, 4, 5, 7, 9, 12, 20, 21, 22, 23 y 25) que hay que recodificar para hacer una interpretación positiva de la puntuación, a más cantidad de puntuación mayor cantidad de actitud emprendedora. La escala está integrada por 10 facetas que cargan sobre la actitud general hacia

tareas emprendedoras: Creatividad, Perseverancia, Capacidad de Organización, Independencia, Confianza en si Mismo, Riesgo Calculado, Tolerancia a la Incertidumbre, Competitividad, Negociación y Locus de Control. Aunque el método de corrección de la escala en su modo original fue la suma de la puntuación en cada ítem aquí se ha utilizado otro procedimiento que se comentará más abajo. En la página 271 del Anexo aparece la escala A junto a sus claves para la corección.

Para la obtención de la puntuación de cada una de las escalas se utilizó la ecuación

$$\frac{\sum_{i=1}^n x_i}{n_{x_i}} \quad (5.1)$$

donde  $x_i$  se refiere a la puntuación del ítem  $i$  y  $n_{x_i}$  se refiere al número total de ítems de cada escala. De esta manera la puntuación de cada una de las escalas osciló entre la puntuación mínima y máxima en la que se podía responder a cada ítem individualmente.

En la Tabla 5.3 aparecen estadísticos descriptivos sobre la dispersión y forma de las distribuciones de las puntuaciones de las escalas, mientras que las distribuciones de frecuencias de las escalas pueden verse en la Figura 5.6. Como se puede observar, la variable con menos asimetría respecto a la distribución normal es la Actitud Emprendedora medida con la ACEMP mientras que la que muestra mayor asimetría negativa es la motivación. La variable con la curtosis más parecida a la normal es la de Carencias y las variables de Obstáculos y de Motivación son más leptocúrticas que la distribución normal.

## Propiedades psicométricas de las escalas

### *Fiabilidad*

Cano et al. (2003), J. García et al. (2005), J. García et al. (2007) así como J. López et al. (2009) han aportado evidencias consistentes de la fiabilidad de las puntuaciones de la escala ACEMP. Por ejemplo, la estimación de la fiabilidad entendida como consistencia interna de las puntuaciones en una muestra de estudiantes universitario fue de 0,84 utilizando un Análisis Factorial Confirmatorio considerando la escala como unidimensional (Cano et al., 2003).

En la Tabla 5.4 aparecen los índices de consistencia interna para las escalas

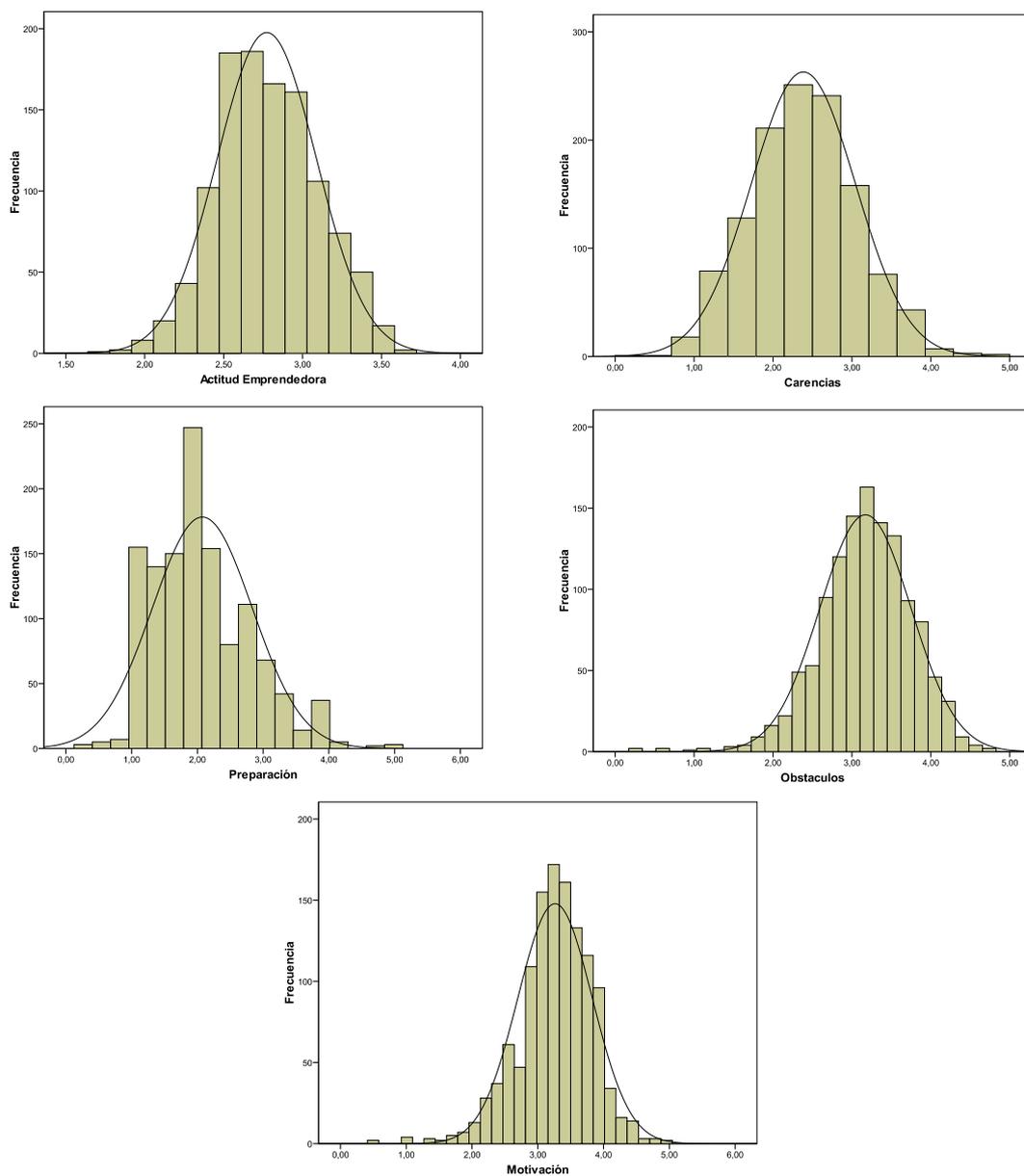


Figura 5.6: Distribución de frecuencias de las puntuaciones de las escalas.

**Tabla 5.3:** Descriptivos de las distribuciones de las puntuaciones de las escalas.

ESCALA	<i>Min</i>	<i>Max</i>	<i>M</i>	<i>DT</i>	<i>A</i>	<i>ET<sub>A</sub></i>	<i>C</i>	<i>ET<sub>C</sub></i>
Actitud Emprendedora	1,69	3,69	2,77	0,31	0,07	0,07	-0,26	0,15
Carencias	0,25	4,83	2,38	0,66	0,18	0,07	-0,05	0,14
Preparación	0,25	5,00	2,08	0,76	0,51	0,07	0,40	0,14
Obstaculos	0,24	4,71	3,17	0,58	-0,50	0,07	1,49	0,14
Motivación	0,45	5,00	3,26	0,57	-0,61	0,07	1,59	0,14

*A*: asimetría y *C*: curtosis.

**Tabla 5.4:** Consistencia interna de las escalas.

ESCALA	<i>n</i>	$\alpha$
Carencias Formativas Percibidas	12	0,892
Preparación Percibida	4	0,852
Motivación para Crear Empresas	11	0,878
Obstáculos Percibidos	17	0,829
Actitud hacia la Creación de Empresas	29	0,749

*n*: número de ítems,  $\alpha$ : índice de consistencia interna de Cronbach.

utilizadas en este estudio. Para cada una de ellas se ha calculado en índice de consistencia interna  $\alpha$  de Cronbach. Como se puede apreciar, todas las escalas muestran índices apropiados de consistencia interna y en ningún caso son menores de 0,7.

### **Validez**

J. García et al. (2005), J. García et al. (2007), Cano et al. (2003) así como J. López et al. (2009) también han aportado evidencias que apoyan la validez de la escala ACEMP, principalmente, desde el punto de vista de la estructura factorial.

Como se puede apreciar en la Tabla 5.5, las escalas utilizadas son capaces de diferenciar entre las personas que responden positiva y negativamente a las preguntas clave relativas a la deseabilidad de crear una nueva empresa, a la in-

**Tabla 5.5:** Validez discriminante para las escalas utilizadas.

PREGUNTA	RESPUESTAS	ESCALA	<i>n</i>	<i>M</i>	<i>DT</i>	<i>ET</i>	<i>t</i>	<i>gl</i>	<i>p</i>	<i>d</i>
<i>Des</i>	Sí	A	834	2,81	0,31	0,01	7,51	1120	**	0,48
	No		288	2,66	0,31	0,02				
	Sí	C	907	2,45	0,65	0,02	5,76	1215	**	0,38
	No		310	2,2	0,66	0,04				
	Sí	P	912	2,15	0,74	0,02	6	1218	**	0,38
	No		308	1,86	0,77	0,04				
	Sí	O	911	3,15	0,57	0,02	-2,81	1220	0,01	-0,18
	No		311	3,25	0,56	0,03				
	Sí	M	911	3,35	0,51	0,02	8,18	446,13	**	0,57
	No		309	3,02	0,64	0,04				
<i>Cog</i>	No	A	691	2,71	0,3	0,01	-9,07	1120	**	-0,56
	Sí		431	2,88	0,31	0,01				
	No	C	759	2,27	0,66	0,02	-7,6	1216	**	-0,46
	Sí		459	2,56	0,61	0,03				
	No	P	758	1,91	0,71	0,03	-10	924,01	**	-0,60
	Sí		462	2,35	0,76	0,04				
	No	O	760	3,23	0,55	0,02	4,35	1221	**	0,26
	Sí		463	3,08	0,59	0,03				
	No	M	758	3,19	0,57	0,02	-5,72	1219	**	-0,35
	Sí		463	3,38	0,53	0,02				
<i>Com</i>	No	A	1021	2,76	0,31	0,01	-5,17	1116	**	-0,52
	Sí		97	2,93	0,34	0,03				
	No	C	1106	2,35	0,65	0,02	-5,86	1211	**	-0,61
	Sí		107	2,74	0,63	0,06				
	No	P	1107	2,01	0,72	0,02	-9,06	124,1	**	-0,96
	Sí		109	2,76	0,84	0,08				
	No	O	1109	3,19	0,56	0,02	2,47	124,3	0,01	0,26
	Sí		109	3,03	0,65	0,06				
	No	M	1108	3,25	0,56	0,02	-2,75	1214	0,01	-0,27
	Sí		108	3,4	0,57	0,05				

*Des*: Deseabilidad: ¿Considera deseable crear una empresa propia?, *Cog*: Cognición: ¿Ha pensado seriamente, como una opción real a corto/medio plazo, montar su propia empresa?, *Com*: Comportamiento: ¿Ha iniciado en algún momento acciones encaminadas a montar su propia empresa?, A: actitud medida con ACEMP, C: carencias formativas percibidas, P: preparación percibida, O: obstáculos percibidos, M: motivación, *n*: número de observaciones, *M*: media, *DT*: desviación típica, *ET*: error típico de la media, \*\*:  $p < 0,01$ , *d*: *d* de Cohen para el tamaño del efecto.

tención de crear una empresa o al hecho de haber iniciado acciones encaminadas a la creación de empresas.

Como se puede observar en la Tabla 5.5, en todas las preguntas clave que hemos utilizado para evaluar la tendencia a crear empresas (deseabilidad, pensamiento y comportamiento) las personas que responden positivamente son las que tienen una mayor puntuación en las escalas A, C, P y M; mientras que son las que obtienen una menor puntuación en la escala O. Esto es, las personas que consideran deseable crear una empresa, han pensado seriamente crear una empresa o han iniciado acciones encaminadas a crear una empresa son las que puntúan alto en actitud, conocimientos, preparación y motivación; mientras que

obtienen menores puntuaciones de obstáculos percibidos. Todas estas diferencias son estadísticamente significativas.

En cuanto a los tamaños del efecto, se puede observar que la variable que sistemáticamente obtiene valores del tamaño del efecto intermedios en todas las dimensiones es la puntuación en la escala A. Sin embargo, el mayor tamaño del efecto observado corresponde a la variable de preparación percibida cuando se refiere al hecho de haber iniciado alguna acción encaminada a la creación de empresas.

### **5.3.3. Procedimiento**

Los cuestionarios fueron administrados entre los meses de diciembre de 2005 y mayo de 2006 a estudiantes de los primeros y últimos cursos de los estudios que se imparten en la Universidad de Almería.

El cuestionario fue administrado de forma grupal en horas lectivas con el consentimiento del profesorado y la participación voluntaria del alumnado. Para solicitar la colaboración por parte de los profesores/as y con el fin de que donaran parte de su horario lectivo para que su alumnado pudiese cumplimentar el cuestionario, se contactó con ellos/as con un correo electrónico que fue el mismo en todos los casos.

Antes de la administración se dieron unas instrucciones generales y se pidió a los participantes que leyesen detenidamente la carta de presentación que aparecía en la primera hoja del formulario. Durante la cumplimentación del mismo el administrador del formulario resolvía las dudas que pudiesen surgir. La cumplimentación del cuestionario tenía una duración de entre 20 y 30 minutos. No se dio ninguna recompensa por la colaboración excepto el agradecimiento verbal.

# Capítulo 6

## Estudio 1. Construcción y depuración de modelos de red bayesiana y regresión logística

*"And if we fight for our rights we become  
children of the sun  
(Children of the sun we are)"*

(Yanou, 2008)

### 6.1. Descripción

En este primer estudio se construirán los modelos de red bayesiana y regresión logística más óptimos tanto desde un punto de vista sustantivo como estadístico. No obstante, para ambas técnicas se tendrán que seleccionar las mismas variables con el fin de poder realizar las comparaciones correspondientes en los siguientes estudios.

El objetivo de este primer estudio es encontrar un modelo común de red bayesiana y regresión logística por cada una de las variables de respuesta que representan las facetas de la actitud (DES, GOG y CON) y que fueron descritas

## Estudio 1. Construcción y depuración de modelos de red bayesiana y regresión logística

en la sección de metodología general. El objetivo principal será seleccionar los modelos que mayor tasa de clasificaciones correctas produzcan teniendo en cuenta el principio de parsimonia del modelo.

## 6.2. Método

### 6.2.1. Muestra

La muestra estuvo compuesta por 1230 estudiantes (426 hombres y 797 mujeres) de los primeros y últimos cursos de las carreras que se cursan en la Universidad de Almería. Las edades de los participantes estuvieron comprendidas entre los 17 y los 56 años, donde la media fue de 22,45 y la desviación típica fue de 4,46. La mayor parte de los participantes residían en la ciudad de Almería e indicaron haber ingresado en la universidad a través del Bachillerato o por medio de algún Ciclo Formativo Profesional.

Una descripción más detallada del procedimiento de muestreo, así como de la muestra en si puede encontrarse en la sub-sección 5.3.1 del capítulo de Metodología General.

### 6.2.2. Materiales

La recogida de la información se llevó a cabo con un cuestionario impreso en tres folios tamaño A4 a doble cara y grapados en su esquina superior izquierda. Para construir los modelos se utilizaron como variables dependientes, o nodos clase, las preguntas consideradas como los componentes de la actitud emprendedora. En concreto, para modelar el componente emocional (DES) se utilizó la pregunta *¿Considera deseable crear una empresa propia?*, para representar el componente cognitivo (COG) se usó la pregunta *¿Ha pensado seriamente, como una opción a corto/medio plazo, montar su propia empresa?*, mientras que para referirse al componente conductual (CON) se analizó la pregunta *¿Ha iniciado en algún momento acciones encaminadas a montar su propia empresa?*.

Por su parte, la puntuación en las escalas de actitud hacia la creación de empresas [ACEMP] (A), carencias formativas percibidas (C), preparación percibida (P), obstáculos percibidos (O), motivación (M) fueron utilizadas como variables

independientes en los modelos de regresión logística y como variables de divergencia en los clasificadores ingenuos bayesianos.

Una descripción más detallada de la herramienta de recogida de datos se puede consultar en la sub-sección 5.3.2 del capítulo de Metodología General.

### 6.2.3. Análisis de datos

Dado que tenemos cinco variables independientes que pueden ser candidatas para formar parte del mejor modelo teórico y estadístico, y teniendo en cuenta el principio de parsimonia respecto al número de variables; llevamos a cabo una comparación de todas las posibles combinaciones de variables independientes que se podrían dar en esta situación.

Para cada variable dependiente se estimaron 31 modelos, de los cuales uno contuvo a las cinco variables independientes, cinco contuvieron todas las posibles combinaciones de cuatro variables, diez modelos recogieron todas las combinaciones posibles entre tres variables, otros diez agruparon las variables por pares y se estimaron otros cinco modelos más incluyendo cada una de las variables por separado. Ya que tenemos tres variables independientes, el proceso se repitió tres veces, con lo que tenemos 93 modelos. Por tanto, se estimaron un total de 93 modelos por cada tipo de técnica estadística, lo que dio lugar a un total de 186 modelos estadísticos.

No fueron introducidos términos de interacción para no tratar la complejidad de este caso. Aunque podríamos haber tratado este tema, estudios posteriores tendrán que dirigir investigaciones a solventar este tópico. En nuestro caso, los resultados que se obtengan aquí, en cualquier caso, se pueden tomar como un indicativo conservador del caso más complejo que incluiría a los términos de interacción (Peduzzi et al., 1995).

La estimación de los modelos de regresión logística se llevó a cabo utilizando el paquete estadístico SPSS<sup>1</sup> en su versión 15.0 en su formato no condicionado, sin incluir ningún término de interacción y configurando el nivel de confianza de las estimaciones al 95 %. No se utilizó ningún método de selección secuencial de variables independientes y por tanto se forzó al programa a introducir en los modelos de regresión logística las variables definidas previamente por el diseño

---

<sup>1</sup>*Statistical Package for the Social Sciences.*

## Estudio 1. Construcción y depuración de modelos de red bayesiana y regresión logística

del estudio.

Los parámetros de las redes bayesianas fueron estimados utilizando el método de máxima verosimilitud corregido con la ecuación de la sucesión de Laplace. Tanto la estimación paramétrica como la posterior actualización de probabilidades, así como la estimación de los parámetros de bondad de ajuste se llevaron a cabo con Netica 4.02 (*Norsys Software*). Las estructuras gráficas fueron creadas manualmente con base al diseño del estudio.

Para cada modelo se estimaron sus correspondientes índices de bondad de ajuste y para todos se obtuvieron los valores de sensibilidad, especificidad, proporción de falsos positivos, proporción de falsos negativos, valor predictivo positivo, valor predictivo negativo y proporción general de clasificaciones correctas. Todos estos parámetros se obtuvieron utilizando una macro diseñada para Microsoft Excel 2003 que incorporaba las salidas de SPSS y Netica.

### 6.3. Resultados

En la Tabla 12.1 que aparece en la página 275 del Anexo aparecen los parámetros de validez predictiva de los 186 modelos estimados. Por su parte, en la Tabla 6.1 aparecen los modelos, para cada técnica y variable dependiente, más predictivos. Un primer detalle que destaca de la Tabla 6.1 es que el componente conductual de la actitud (CON) es el que mejor es predicho por ambas técnicas en términos de la tasa de clasificaciones correctas. No obstante, el modelo más parsimonioso es el 165 que incluye tres variables independientes frente al modelo 63 que incluye a todas las variables. Ambas técnicas muestran niveles similares en todos los índices predictivos, excepto en la proporción de falsos negativos, donde la regresión logística obtiene una mayor puntuación. Así pues, el modelo que incorpora la puntuación en ACEMP, la preparación percibida y los obstáculos percibidos es el elegido para modelar el componente conductual de la actitud.

El componente emocional, o de deseabilidad (DES), es el siguiente elemento mejor predicho por ambas técnicas. Nuevamente aquí, la red bayesiana selecciona un modelo más parsimonioso frente a lo que estima la regresión logística. Por su parte, el modelo seleccionado por la regresión logística produce una tasa más elevada de falsos positivos. Así pues, en este caso, el modelo seleccionado es el que incorpora la puntuación de la escala ACEMP, la preparación percibida, los

obstáculos percibidos y el grado de motivación como factores explicativos del componente emocional de la actitud hacia la creación de empresas.

Por último, el componente cognitivo es el que peor ha sido predicho por ambas técnicas. Sin embargo, es la regresión logística la que obtiene un modelo más parsimonioso con la menor pérdida en la tasa de clasificaciones correctas. Por lo tanto, el modelo seleccionado para explicar el componente cognitivo de la actitud (COG) es el que incluye a la puntuación de la ACEMP, las carencias percibidas, y la preparación percibida.

### 6.3.1. Modelos de regresión logística

En esta subsección se presentarán los parámetros de los modelos de regresión logística y se estudiará la bondad de ajuste de cada uno de ellos.

#### Componente emocional

El conjunto de variables seleccionado para modelar la dimensión emocional de la actitud hacia la creación de empresas produjo los parámetros que aparecen en la Tabla 6.2. Como se puede observar, todos los valores del estadístico de contraste de Wald son estadísticamente diferentes de cero, lo que indica que todas las variables son relevantes para explicar la variable dependiente. A su vez, ningún intervalo de confianza para el exponencial del parámetro beta contiene al valor uno. También se puede observar que todas las variables independientes se asocian positivamente con la tendencia a considerar deseable la creación de una empresa, excepto los obstáculos percibidos que se relacionan inversamente con esta emoción. Esto es, a mayor puntuación en la escala ACEMP, mayor preparación percibida y mayor puntuación en la escala de motivación, la probabilidad de considerar deseable la creación de una empresa aumenta. El mayor parámetro del modelo está asociado a la puntuación de la escala ACEMP.

En la Tabla 6.3 aparecen las frecuencias observadas y estimadas para los diez grupos en los que el procedimiento de Hosmer-Lemeshow divide la muestra. El valor de  $\chi^2$  es de 12,33 con una significación de 0,14, luego el modelo se ajusta adecuadamente a los datos.

Por su parte, cuando analizamos los residuos estudentizados al cuadrado en función de las probabilidades pronosticadas se observa que los mayores desajustes

Estudio 1. Construcción y depuración de modelos de red bayesiana y regresión logística

**Tabla 6.1:** Los modelos más predictivos en función de cada una de las técnicas y de las variables dependientes.

Modelo	165 *	63	98 *	1	125	38 *
VD	CON	CON	DES	DES	COG	COG
Técnica	RB	RL	RB	RL	RB	RL
VI	A P O	A C P O M	A P O M	A C P O M	A C P O M	A C P
S	0,22	0,03	0,94	0,96	0,57	0,41
E	1,00	1,00	0,34	0,21	0,83	0,86
PFP	0,00	0,00	0,66	0,79	0,17	0,14
PFN	0,78	0,97	0,06	0,04	0,43	0,59
VPP	0,83	0,75	0,81	0,78	0,66	0,65
VPN	0,93	0,92	0,66	0,62	0,76	0,70
TCC	0,93	0,92	0,79	0,77	0,73	0,69
$R^2$		0,164		0,179		0,160
$H - L_p$		0,949		0,115		0,137
$-2\log$		559,982		1114,411		1340,970
$\chi^2$		83,074		143,177		139,231
PL	0,233		0,457		0,542	
PC	0,128		0,299		0,363	
CE	0,932		0,833		0,795	

VD: variable dependiente, RB: red bayesiana, RL: regresión logística, VI: variables independientes, A: puntuación ACEMP, C: carencias percibidas, P: preparación percibida, O: obstáculos percibidos, M: motivación, S: sensibilidad, E: especificidad, PFP: proporción de falsos positivos, PFN: proporción de falsos negativos, VPP: valor predictivo positivo, VPN: valor predictivo negativo, TCC: tasa de clasificaciones correctas,  $R^2$ :  $R$  de Nagelkerke,  $H - L_p$ :  $p$ -valor del test de Hosmer-Lemeshow, PL: pérdida logarítmica, PC: pérdida cuadrática, CE: compensación esférica. El \* indica el modelo elegido para la correspondiente combinación de variables independientes y variable dependiente.

**Tabla 6.2:** Modelo de regresión logística para explicar el componente emocional de la actitud.

Variable	$\hat{\beta}$	ET	Wald ( $gl = 1$ )(*)	$e^{\hat{\beta}}$	$IC_{inf}$	$IC_{sup}$
A	0,985	0,264	13,956	2,677	1,597	4,489
P	0,497	0,111	19,954	1,643	1,321	2,043
O	-0,596	0,156	14,663	0,551	0,406	0,747
M	1,084	0,153	50,434	2,956	2,192	3,986
Constante	-4,139	0,913	20,550	0,016		

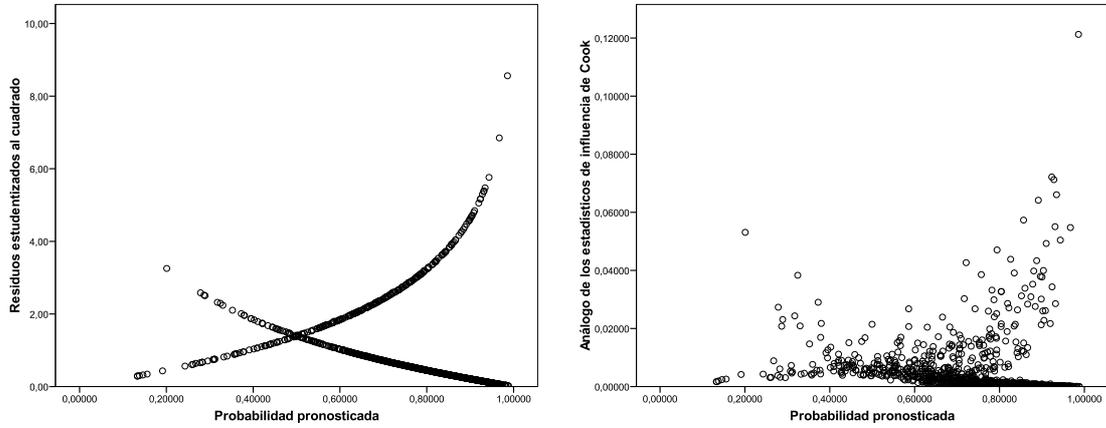
\*:  $p < 0,001$  en todos los casos, A: puntuación ACEMP, P: preparación percibida, O: obstáculos percibidos, M: motivación,  $IC_{inf}$ : límite inferior del intervalo de confianza para la estimación de  $e^{\hat{\beta}}$  al 95 %,  $IC_{sup}$ : límite superior del intervalo de confianza para la estimación de  $e^{\hat{\beta}}$  al 95 %.

**Tabla 6.3:** Test de ajuste Hosmer-Lemeshow para el componente emocional.

	DES=NO		DES=SÍ		Total
	Observado	Esperado	Observado	Esperado	
1	65	65,71	46	45,29	111
2	52	45,41	59	65,59	111
3	42	37,18	69	73,82	111
4	23	31,92	88	79,08	111
5	24	27,17	87	83,83	111
6	27	23,37	84	87,63	111
7	13	19,34	98	91,66	111
8	14	15,74	97	95,26	111
9	14	11,02	97	99,98	111
10	9	6,14	106	108,86	115

DES: ¿Considera deseable crear una empresa propia?

## Estudio 1. Construcción y depuración de modelos de red bayesiana y regresión logística



**Figura 6.1:** Comparación de los residuos estandarizados al cuadrado (izquierda) y de las distancias de Cook (derecha) con la probabilidad pronosticada para el modelo DES.

del modelo se producen cuando la respuesta a la pregunta utilizada como variable dependiente es negativa. A su vez, el gráfico de las distancias de Cook informa de que los casos con mucha influencia sobre el modelo no son muchos aunque están cargados hacia altas probabilidades estimadas (Figura 6.1).

### Componente cognitivo

En la Tabla 6.4 se puede observar que de las tres variables seleccionadas para explicar el componente cognitivo de la actitud (COG) sólo dos son estadísticamente significativas, la puntuación en la escala ACEMP y la preparación percibida. También se observa que tanto la puntuación en ACEMP como la preparación percibida se relacionan positivamente con la probabilidad de considerar seriamente la creación de una empresa a corto/medio plazo. El intervalo de confianza para la estimación de beta contiene a la unidad en la variable carencias percibidas. Nuevamente, el mayor parámetro del modelo recae sobre la puntuación ACEMP indicando que las personas que puntúan alto en esta escala tienen un *riesgo* entre 2,9 y 7 veces más grande de pensar seriamente la creación de una empresa.

Este modelo, sin embargo, muestra un buen ajuste a los datos (Tabla 6.5) cuando lo evaluamos con el estadístico de Hosmer-Lemeshow ( $G_{HL}^2 = 12,32$ ,  $p = 0,137$ ). Por su parte, como se observa en la Figura 6.2, aunque los residuos estandarizados al cuadrado muestran que las mayores desviaciones se producen cuando la respuesta cognitiva es positiva, los residuos de Cook sugieren que es

**Tabla 6.4:** Modelo de regresión logística para explicar el componente cognitivo de la actitud.

Variable	$\hat{\beta}$	ET	Wald ( $gl = 1$ )	$p$	$e^{\hat{\beta}}$	$IC_{inf}$	$IC_{sup}$
A	1,499	0,228	43,357	*	4,476	2,865	6,993
C	0,158	0,118	1,813	0,178	1,172	0,930	1,476
P	0,624	0,102	37,373	*	1,867	1,528	2,281
Constante	-6,363	0,641	98,474	*	0,002		

\*:  $p < 0,001$ , A: puntuación ACEMP, C: carencias percibidas, P: preparación percibida,  $IC_{inf}$ : limite inferior del intervalo de confianza para la estimación de  $e^{\hat{\beta}}$  al 95 %,  $IC_{sup}$ : limite superior del intervalo de confianza para la estimación de  $e^{\hat{\beta}}$  al 95 %.

debido a un conjunto de casos limitado.

### Componente conductual

Como se puede observar en la Tabla 6.6, el modelo de regresión logística estimó el parámetro para la variable obstáculos percibidos en el sentido apropiado aunque este no alcanzó a ser estadísticamente significativo. A su vez, su intervalo de confianza contiene al valor 1. Nuevamente la variable con más fuerza para explicar el componente conductual de la actitud fue la puntuación en la escala ACEMP, que en este caso estimó que las personas que puntúan alto en esta escala tiene 1,5 y 6,6 veces más posibilidades de crear una empresa.

En la Tabla 6.7 se puede apreciar que las frecuencias observadas y las frecuencias predichas por el modelo están muy cercanas y el estadístico de contraste muestra niveles que sugieren un buen ajuste ( $G_{HL}^2 = 4,02$ ,  $p = 0,855$ ). Por su parte, el análisis de los residuos sugiere que los mayores problemas del modelo aparecen cuando pretende pronosticar respuestas positivas en la variable dependiente (Figura 6.3).

#### 6.3.2. Modelos de red bayesiana

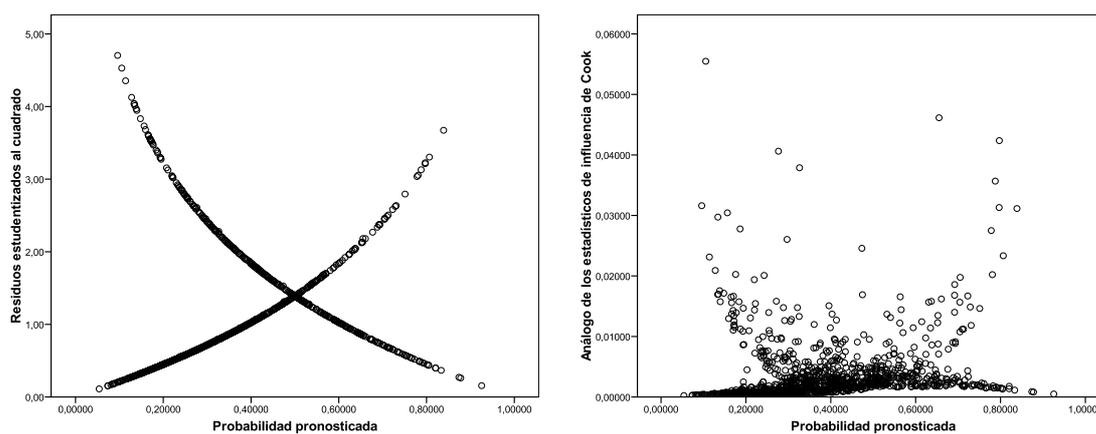
Los modelos de red bayesiana no se basan en ningún supuesto distribucional de las variables luego esta sección se limitará a presentar las probabilidades previas

Estudio 1. Construcción y depuración de modelos de red bayesiana y regresión logística

**Tabla 6.5:** Test de ajuste Hosmer-Lemeshow para el componente cognitivo.

	COG=NO		COG=SÍ		Total
	Observado	Esperado	Observado	Esperado	
1	87	95,447	24	15,553	111
2	97	88,709	14	22,291	111
3	81	83,670	31	28,330	112
4	85	78,083	26	32,917	111
5	73	73,355	38	37,645	111
6	69	68,007	42	42,993	111
7	62	62,941	49	48,059	111
8	53	56,015	58	54,985	111
9	45	47,394	66	63,606	111
10	34	32,378	78	79,622	112

COG: ¿Ha pensado seriamente, como una opción real a corto/medio plazo, montar su propia empresa?



**Figura 6.2:** Comparación de los residuos estudentizados al cuadrado (izquierda) y de las distancias de Cook (derecha) con la probabilidad pronosticada para el modelo COG.

**Tabla 6.6:** Modelo de regresión logística para explicar el componente conductual de la actitud.

Variable	$\hat{\beta}$	ET	Wald ( $gl = 1$ )	$p$	$e^{\hat{\beta}}$	$IC_{inf}$	$IC_{sup}$
A	1,148	0,379	9,153	0,002	3,152	1,498	6,631
P	1,121	0,150	55,781	*	3,069	2,287	4,119
O	-0,083	0,203	0,168	0,682	0,920	0,618	1,370
Constante	-7,994	1,433	31,140	*	0,000		

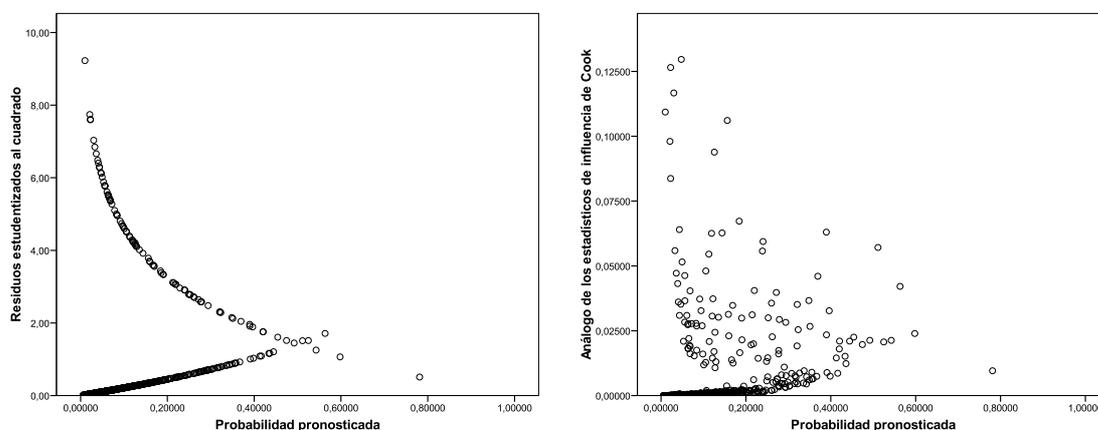
\*:  $p < 0,001$ , A: puntuación ACEMP, P: preparación percibida, O: obstáculos percibidos,  $IC_{inf}$ : limite inferior del intervalo de confianza para la estimación de  $e^{\hat{\beta}}$  al 95 %,  $IC_{sup}$ : limite superior del intervalo de confianza para la estimación de  $e^{\hat{\beta}}$  al 95 %.

**Tabla 6.7:** Test de ajuste Hosmer-Lemeshow para el componente conductual.

	CON=No		CON=Sí		Total
	Observado	Esperado	Observado	Esperado	
1	110	109,55	1	1,45	111
2	108	108,62	3	2,38	111
3	108	107,44	3	3,56	111
4	107	106,44	4	4,56	111
5	104	105,32	7	5,68	111
6	102	104,82	10	7,18	112
7	105	102,18	6	8,82	111
8	97	98,99	14	12,01	111
9	97	93,88	14	17,12	111
10	77	77,76	35	34,24	112

CON: ¿Ha iniciado en algún momento acciones encaminadas a montar su propia empresa?

## Estudio 1. Construcción y depuración de modelos de red bayesiana y regresión logística



**Figura 6.3:** Comparación de los residuos estudentizados al cuadrado (izquierda) y de las distancias de Cook (derecha) con la probabilidad pronosticada para el modelo CON.

estimadas para cada variable. Como se puede observar en las Figuras 6.4, 6.5 y 6.5 las distribuciones de las variables independientes no cambian mucho de un modelo a otro. La única diferencia aparece en la variable de divergencia, o clase, donde varía la proporción de respuestas positivas a cada una de las preguntas.

### 6.3.3. Ajuste predictivo de los modelos seleccionados

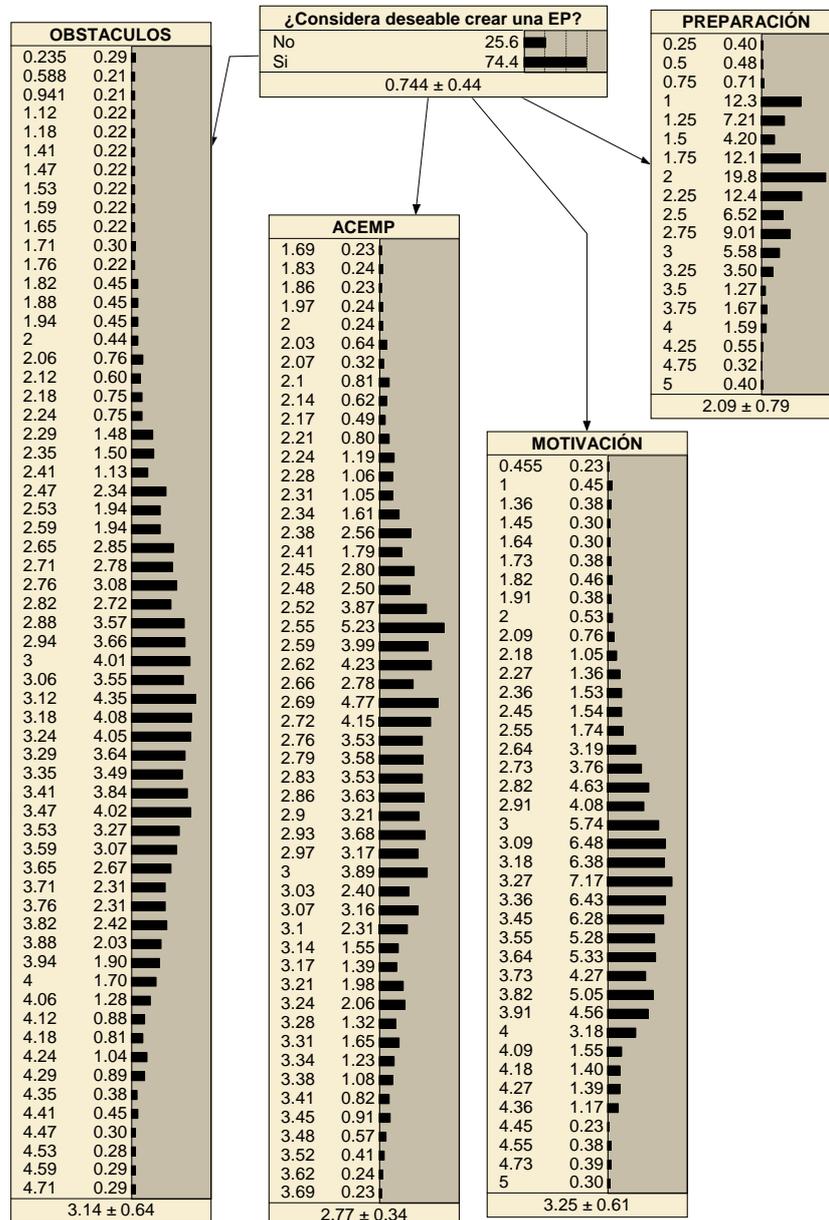
Como se puede apreciar en la Tabla 6.8 no existen grandes diferencias en términos predictivos en los modelos seleccionados en ambas técnicas.

El aspecto en el que ambas técnicas están más parejas es en la tasa de clasificaciones correctas no superándose la diferencia entre una técnica y otra en un 2,2% en ninguno de los casos. Por el contrario, la diferencia más importante se observa en el modelo conductual en el valor predictivo positivo, alcanzándose una diferencia de aproximadamente del 33%.

En lo que respecta a la sensibilidad, ambas técnicas están parejas y las diferencias oscilan entre el 1,9% en el componente emocional y el 18,9% en el componente conductual.

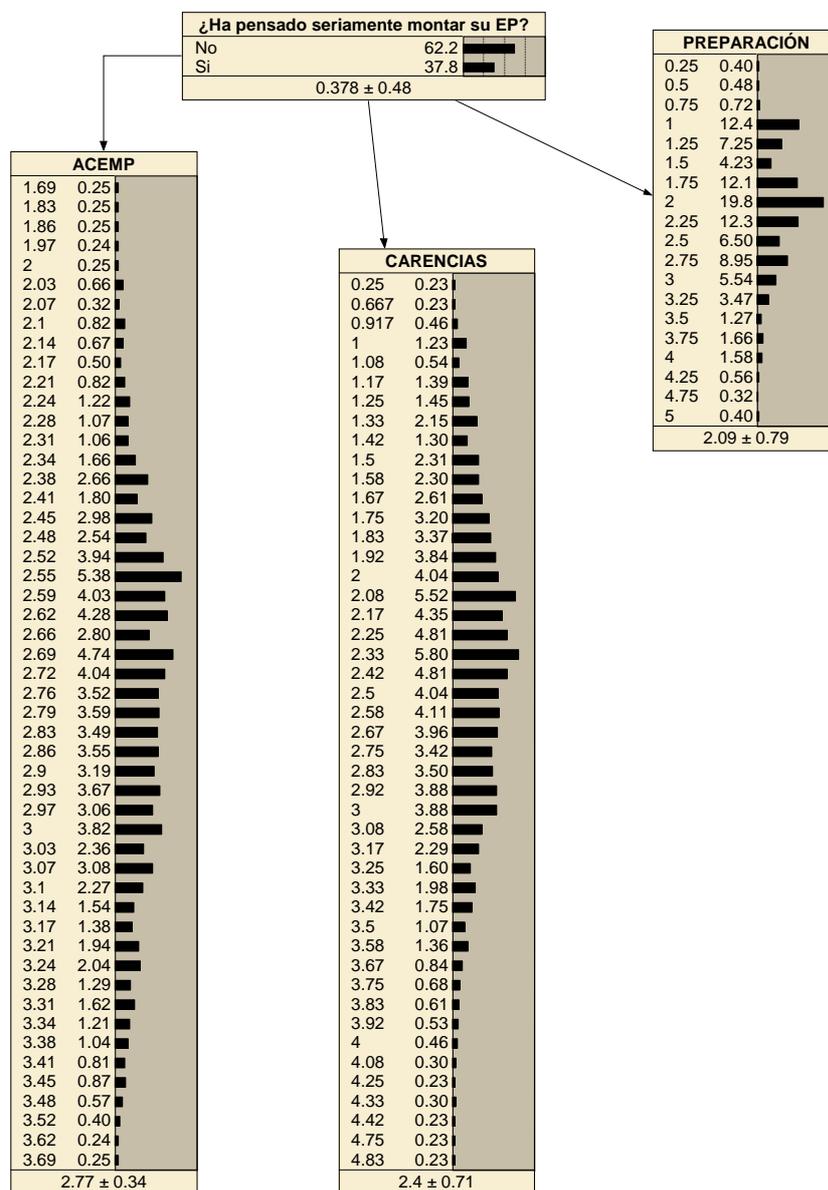
En términos de especificidad, el modelo de red bayesiana se comporta mejor para predecir el componente emocional pero la regresión logística la sobrepasa en los componentes cognitivo y conductual.

Por otro lado, la red bayesiana genera más falsos positivos en los componentes cognitivo y conductual, pero es en esas dos mismas dimensiones donde la regresión



**Figura 6.4:** Probabilidades previas del componente emocional en una red bayesiana. EP: empresa propia. En la base de cada nodo se muestra la media ± la desviación típica.

Estudio 1. Construcción y depuración de modelos de red bayesiana y regresión logística



**Figura 6.5:** Probabilidades previas del componente cognitivo en una red bayesiana. EP: empresa propia. En la base de cada nodo se muestra la media ± la desviación típica.

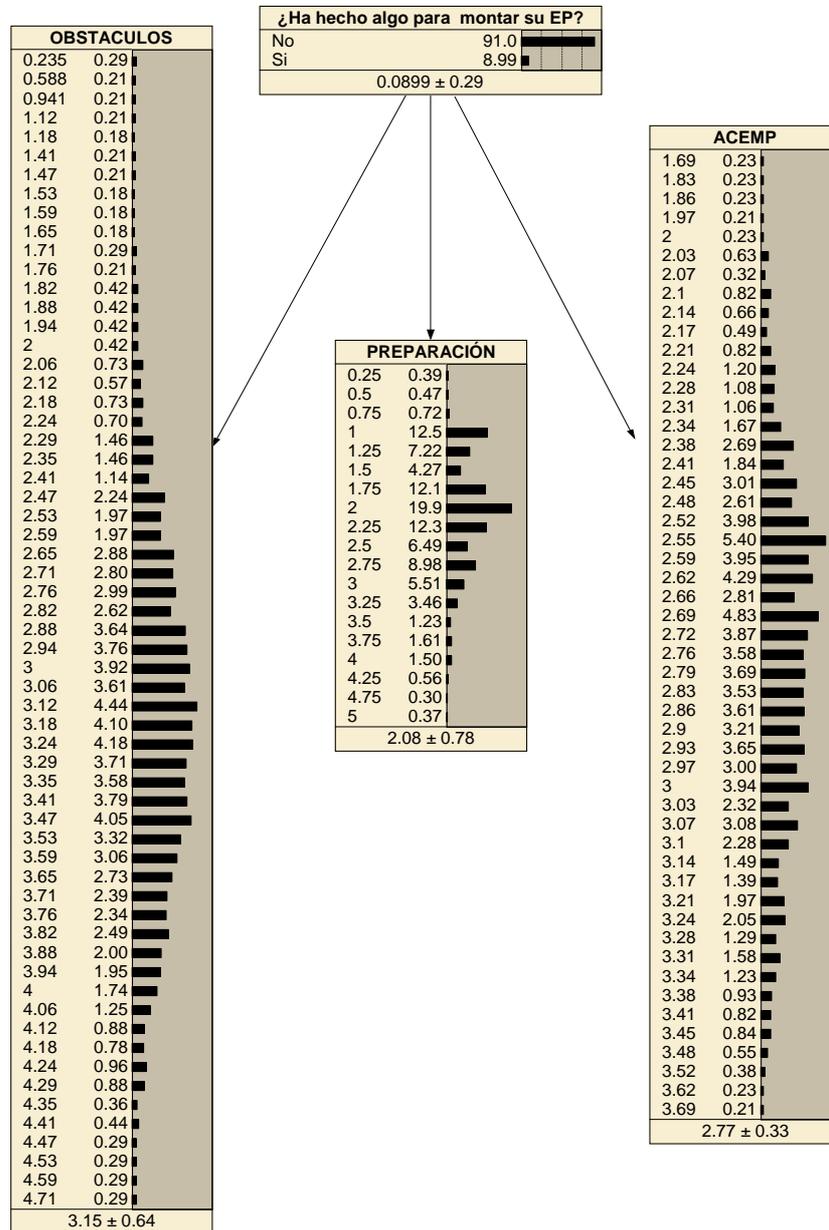


Figura 6.6: Probabilidades previas del componente conductual en una red bayesiana. EP: empresa propia. En la base de cada nodo se muestra la media ± la desviación típica.

**Tabla 6.8:** Ajuste predictivo para los modelos seleccionados.

	RB			RL		
	DES	COG	CON	DES	COG	CON
S	0,939	0,508	0,220	0,958	0,410	0,031
E	0,342	0,815	0,996	0,198	0,864	0,997
PFP	0,658	0,185	0,004	0,802	0,136	0,003
PFN	0,061	0,492	0,780	0,042	0,587	0,969
VPP	0,806	0,625	0,828	0,778	0,678	0,500
VPN	0,656	0,732	0,929	0,615	0,703	0,915
TCC	0,786	0,699	0,926	0,765	0,692	0,913

RB: red bayesiana, RL: regresión logística, DES: componente emocional, COG: componente cognitivo, CON: componente conductual, S: sensibilidad, E: especificidad, PFP: proporción de falsos positivos, PFN: proporción de falsos negativos, VPP: valor predictivo positivo, VPN: valor predictivo negativo y TCC: tasa de clasificaciones correctas.

logística se ve sobrepasada en términos de la proporción de falsos negativos.

#### 6.3.4. Comparación preliminar entre técnicas

En la Tabla 6.9 aparecen las pruebas  $t$  de Student para comparar los estadísticos de ajuste predictivo entre la regresión logística y la red bayesiana. Como se puede apreciar, no existen diferencias estadísticamente significativas entre ambas técnicas en esta primera fase en la que se han construido todos los modelos posibles combinando todas las variables independientes.

No obstante, hay que destacar que se observa que las redes bayesianas tienen valores mayores de sensibilidad y especificidad, así como menores valores en la proporción de falsos positivos y falsos negativos. Las redes bayesianas también tienen los mayores valores de valores positivos predichos, valores negativos predichos y de la tasa de clasificaciones correctas.

**Tabla 6.9:** Comparación preliminar entre técnicas.

		<i>n</i>	<i>M</i>	<i>DT</i>	<i>ET</i>	<i>t</i>	<i>gl</i>	<i>p</i>	<i>Dif. M</i>	<i>ET</i>	<i>IC 95</i>	
S	RL	93	0,44	0,41	0,04	-1,068	184	0,29	-0,06	0,06	-0,17	0,05
	RB	93	0,50	0,36	0,04							
E	RL	93	0,66	0,41	0,04	-0,466	184	0,64	-0,03	0,06	-0,14	0,08
	RB	93	0,68	0,36	0,04							
PFP	RL	93	0,34	0,41	0,04	0,466	184	0,64	0,03	0,06	-0,08	0,14
	RB	93	0,32	0,36	0,04							
PFN	RL	93	0,56	0,41	0,04	1,068	184	0,29	0,06	0,06	-0,05	0,17
	RB	93	0,50	0,36	0,04							
VPP	RL	77	0,67	0,09	0,01	-1,923	168	0,06	-0,03	0,02	-0,06	0,00
	RB	93	0,70	0,11	0,01							
VPN	RL	91	0,71	0,20	0,02	-1,774	181	0,08	-0,04	0,02	-0,09	0,00
	RB	92	0,76	0,13	0,01							
TCC	RL	93	0,78	0,10	0,01	-0,849	184	0,40	-0,01	0,01	-0,04	0,02
	RB	93	0,79	0,10	0,01							

S: sensibilidad, E: especificidad, PFP: proporción de falsos positivos, PFN: proporción de falsos negativos, VPP: valor predictivo positivo, VPN: valor predictivo negativo y TCC: tasa de clasificaciones correctas.

## 6.4. Discusión parcial

En este primer estudio hemos seleccionado tres conjuntos de variables para modelar el componente emocional, cognitivo y conductual de la tendencia hacia la creación de empresas. Se han seleccionado los modelos más parsimoniosos que incluían menos variables y que suponían poca pérdida de validez predictiva para ambas técnicas. No obstante, hay algunos aspectos que destacar sobre la bondad de ajuste de algunos parámetros de los modelos de regresión logística.

En primer lugar, de las variables seleccionadas para modelar el componente cognitivo de la actitud hacia la creación de empresas, la variable carencias percibidas no mostró un buen ajuste en términos del estadístico de Wald. Teóricamente se tendría que haber eliminado esta variable del modelo pero, dado que el modelo funciona razonablemente para predecir la variable dependiente, esta variable se conservó. Además, el test de Hosmer-Lemeshow indicó un buen ajuste en las frecuencias observadas y esperadas del modelo así como el análisis gráfico de los residuos no mostró indicios de una falta de ajuste considerable. En lo que respecta al conjunto de variables seleccionado para modelar el componente conductual, sucedió lo mismo con la variable motivación, pero se decidió mantenerla

## Estudio 1. Construcción y depuración de modelos de red bayesiana y regresión logística

en el modelo debido al óptimo nivel de clasificaciones correctas que generaba. Esta decisión de conservar estas variables en los modelos se ve reforzada por el hecho de que no hay diferencias considerables en los niveles predictivos de ambas técnicas.

Por último, también hemos realizado una comparación preliminar de los modelos y hemos observado que no se aprecian diferencias estadísticamente significativas en el comportamiento de ambas técnicas en esta fase inicial con respecto a los estadísticos de predicción empleados.

Los resultados de este primer estudio tienen relevancia sustantiva en el ámbito aplicado de la tesis ya que sirven para caracterizar el perfil del emprendedor potencial entre los estudiantes universitarios (Huefner et al., 1996). En este sentido, hemos creado tres modelos diferentes que sirven para predecir los tres componentes (emocional, cognitivo y conductual) de la actitud emprendedora en estudiantes universitarios utilizando dos técnicas estadísticas diferentes que han mostrado niveles razonables de ajuste predictivo.

El primer elemento a destacar de los tres modelos de las dimensiones de la actitud emprendedora es que en todos los casos aparece la puntuación en la escala ACEMP como variable relevante. Este hecho tiene repercusiones tanto a nivel práctico como a nivel teórico. Desde un punto de vista práctico, la validez de la escala ACEMP se ve reforzada, así como su potencialidad de uso como herramienta para la detección precoz de emprendedores potenciales (Cano et al., 2003; J. García et al., 2005, 2006, 2007). Por su parte, el hecho de que la puntuación en nuestra escala de actitudes emprendedoras sea un elemento relevante ante la predicción de la tendencia a crear empresas, se puede considerar un elemento útil ante la definición de emprendedor potencial. Por ello, sería deseable explorar en detenimiento las dimensiones que la componen para evaluar cuales son los factores más útiles en este sentido (Cano et al., 2003).

Los resultados obtenidos en este primer estudio también son consistentes con otros trabajos orientados a caracterizar al emprendedor ya que también hemos encontrado que la preparación percibida es un predictor importante de la tendencia a crear empresas (p. e., Genescá y Capelleras, 2004; Genesca y Veciana, 1984; Gómez et al., 2007; HayGroup y SAP AG, 2003; Rogoff y Lee, 1996; Sánchez, 2003; Veciana, 1989). Así pues, los programas de intervención social destinados a potenciar y valorar positivamente a las personas emprendedoras cobran relevancia

en el contexto universitario (p. e., Díaz, 2003; Peñas y Quijano, 2008).

Aparte de los factores comunes a los tres componentes de la actitud hacia la creación de empresas, se ha obtenido una relación importante entre el aspecto motivacional y la dimensión emocional de la actitud emprendedora. Como se ha visto en el modelo de regresión logística para predecir la deseabilidad de crear una empresa, las personas que obtienen mayores valores en la escala de motivación tienden a considerar deseable la creación de una empresa propia entre dos y cuatro veces más que las personas que obtienen puntuaciones bajas en la escala motivacional. Sin embargo, la deseabilidad de crear una empresa propia se ve afectada negativamente por la cantidad de obstáculos percibidos.

## Estudio 1. Construcción y depuración de modelos de red bayesiana y regresión logística

# Capítulo 7

## Estudio 2. Tamaño de muestra y área bajo la curva ROC

*"And sometimes I get nervous,  
when I see an open door.  
Close your eyes, clear your heart,  
cut the cord"*

(Killers, 2008)

### 7.1. Descripción

El tamaño de la muestra es uno de los problemas que se han destacado del análisis de regresión logística porque las estimaciones de los coeficientes del modelo son sesgadas a medida que la muestra se hace más pequeña (p. e., Hsieh, 1989; Hsieh et al., 1998; E. King y Ryan, 2002; G. King y Zeng, 2001a, 2001b; Whittermore, 1981). Así pues, en este capítulo, se estudiará el efecto que tiene el tamaño de la muestra sobre el nivel de clasificación general evaluado con curvas ROC. Previsiblemente, en ambas técnicas se producirá un descenso en el área bajo la curva ROC a medida que disminuye el tamaño de la muestra pero se espera que

## Estudio 2. Tamaño de muestra y área bajo la curva ROC

este efecto sea menos severo en las redes bayesianas. En concreto, se operacionalizará la idea de tamaño muestral en términos de eventos por variable (Concato et al., 1995; Peduzzi et al., 1995, 1996).

## 7.2. Método

### 7.2.1. Muestra

Para realizar este estudio se seleccionaron 21 muestras aleatorias simples de diferentes tamaños a partir del archivo original de datos compuesto por 1230 observaciones provenientes del muestreo de los primeros y últimos cursos de las titulaciones que se imparten en la Universidad de Almería. Un total de 426 participantes fueron hombres y el resto (797) mujeres, mientras que la edad de los participantes estuvo comprendida entre los 17 y los 56 años, donde la media fue de 22,45 y la desviación típica fue de 4,46. Una descripción más detallada del procedimiento de muestreo, así como de la muestra en si puede encontrarse en la sub-sección 5.3.1 del capítulo de Metodología General.

Se calcularon tamaños muestrales diferentes para cada uno de los modelos teniendo en cuenta la tasa de respuestas positivas, con el objetivo de estudiar la influencia relativa del número de eventos por variable en la exactitud global de cada técnica usando el área bajo la curva ROC (Concato et al., 1995; Peduzzi et al., 1995, 1996). Con el fin de conseguir un número de eventos por variable cercano a 5, 10, 15, 20, 25, 30, 100 y 200 se determinaron los tamaños muestrales especificados en la Tabla 7.1 para cada uno de los tres componentes de la actitud.

### 7.2.2. Análisis de datos

Se estimaron un total de 42 modelos (21 de regresión logística y 21 de redes bayesianas) utilizando los tamaños muestrales definidos en la Tabla 7.1. De los 21 modelos que se estimaron con cada técnica, ocho fueron para modelar el componente emocional de la actitud emprendedora, siete para el componente cognitivo y seis para el conductual.

La estimación de los modelos de regresión logística se llevó a cabo utilizando el paquete estadístico SPSS 15.0 en su formato no condicionado, sin incluir ningún término de interacción y configurando el nivel de confianza de las estimaciones al

**Tabla 7.1:** Tamaños de muestra en función de el número deseado de eventos por variable y de la tasa de respuestas positivas.

$VD$	$fr(y = 1)$	$EPV_{obs}$	$EPV_{des}$							
			5	10	15	20	25	30	100	200
DES	913	304	20	40	61	81	101	121	404	808
COG	463	116	53	106	159	213	266	319	1063	
CON	109	36	169	339	508	677	846	1016		

$EPV_{des}$ : eventos por variable deseados,  $fr(y = 1)$ , frecuencia de respuestas positivas,  $EPV_{obs}$ : eventos por variable observados,  $VD$ : variable dependiente, DES: componente emocional, COG: componente cognitivo y CON: componente conductual.

95 %. No se utilizó ningún método de selección secuencial de variables independientes y por tanto se forzó al programa a introducir en los modelos de regresión logística las variables definidas previamente por el diseño del estudio.

Los parámetros de las redes bayesianas fueron estimados utilizando el método de máxima verosimilitud corregido con la ecuación de la sucesión de Laplace. Tanto la estimación paramétrica como la posterior actualización de probabilidades, así como la estimación de los parámetros de bondad de ajuste se llevaron a cabo con Netica 4.02. Las estructuras gráficas fueron creadas manualmente con base al diseño del estudio.

El área bajo la curva ROC ( $\theta$ ) se estimó usando la técnica no paramétrica propuesta por Hanley y McNeil (1982) que se basa en el estadístico  $U$  de Mann-Whitney. La ( $U$ ) de Mann-Whitney se estima usando la ecuación

$$U = \sum_{i=1}^{n_p} \sum_{j=1}^{n_a} S(x_{p_i}, x_{a_j}), \quad (7.1)$$

donde

Estudio 2. Tamaño de muestra y área bajo la curva ROC

$$S(x_{p_i}, x_{a_j}) = \begin{cases} 1 & \text{si } x_p > x_a, \\ \frac{1}{2} & \text{si } x_p = x_a, \\ 0 & \text{si } x_p < x_a, \end{cases} \quad (7.2)$$

lo que, *al menos conceptualmente*, puede entenderse como una comparación entre todas las posibles combinaciones  $n_p \times n_a$  en las sub-muestras de casos donde se presenta el rasgo de interés ( $n_p$ ) y donde no se presenta el rasgo ( $n_a$ ).

De este modo, la estimación del área bajo la curva, o suma de rangos de Wilcoxon ( $W$ ), según Hanley y McNeil (1982) viene dada por

$$\hat{\theta} = \frac{1}{n_p \times n_a} \times U. \quad (7.3)$$

La desviación típica de ésta estimación del área bajo la curva ROC viene dada por

$$DT_{\hat{\theta}} = \sqrt{\frac{\hat{\theta} \times (1 - \hat{\theta}) + (n_p - 1) \times (Q_1 - \hat{\theta}^2) + (n_a - 1) \times (Q_2 - \hat{\theta}^2)}{n_p \times n_a}}, \quad (7.4)$$

donde  $Q_1$  representa la probabilidad de que las respuestas de dos casos seleccionados al azar del grupo que presentan el rasgo sean mayores que la de un caso elegido aleatoriamente del grupo donde no se presenta el rasgo; mientras que  $Q_2$  representa la probabilidad de que la respuesta de la técnica sobre un caso seleccionado aleatoriamente del grupo que presenta el rasgo sea mayor que la probabilidad estimada para dos casos seleccionados al azar del grupo que no presenta el rasgo. Las estimaciones para ambos parámetros vienen dados por las siguientes expresiones aproximadas según la distribución exponencial

$$\hat{Q}_1 = \frac{\hat{\theta}}{2 - \hat{\theta}} \quad \text{y} \quad \hat{Q}_2 = \frac{2 \times \hat{\theta}^2}{1 + \hat{\theta}}. \quad (7.5)$$

Las curvas fueron dibujadas utilizando una suavización *logit* basada en la

ecuación de la regresión lineal y tomando 100 intervalos constantes de 0,01 respecto a la Proporción de Falsos Positivos ( $PFPP$ ). Primero se calcularon los *logit* de cada valor de corte para la  $PFPP$  aplicando

$$\log \widehat{PFPP}^* = \ln \frac{PFPP}{1 - PFPP} . \quad (7.6)$$

Seguidamente, se utilizaron las estimaciones de la pendiente ( $\hat{\alpha}$ ) y del punto de corte en el eje  $y$  ( $\hat{\beta}$ ) con un análisis de regresión lineal sobre las tasas observadas de casos con y sin el rasgo de interés para cada umbral de corte con lo que se obtuvo la estimación logística de la sensibilidad ( $\log \hat{S}$ ):

$$\log \hat{S}^* = \hat{\alpha} + \hat{\beta} \times \log \widehat{PFPP}^* . \quad (7.7)$$

Por último, la sensibilidad estimada ( $\hat{S}^*$ ) para cada punto de corte simulado de la  $PFPP$  se obtuvo como se refleja en la ecuación 7.8.

$$\hat{S}^* = \frac{e^{\log \hat{S}^*}}{1 - e^{\log \hat{S}^*}} . \quad (7.8)$$

Aunque las inferencias y contrastes entre curvas se llevaron a cabo usando el valor de  $\hat{\theta}$  por entenderse más apropiado desde un punto de vista metodológico ya que puede considerarse como el límite inferior del verdadero área bajo la curva ROC, también se obtuvo una estimación del área bajo la curva ROC suavizada basada en el procedimiento gráfico descrito anteriormente y que viene definida como

$$\hat{\theta}^* = \sum_{i=0,01}^{0,99} w_i \times \hat{h}_i, \quad (7.9)$$

donde  $w_i$  es la anchura del intervalo y  $\hat{h}_i$  es la altura interpolada del intervalo.

Para comprobar si existían diferencias estadísticamente significativas entre las curvas ROC generadas por la regresión logística y la red bayesiana se uti-

## Estudio 2. Tamaño de muestra y área bajo la curva ROC

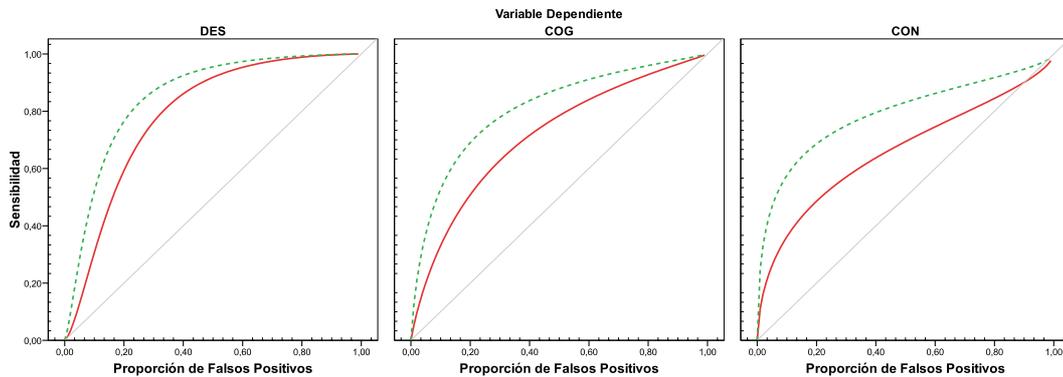
lizó el procedimiento propuesto por Hanley y McNeil (1983). Dado que se trata de muestras pareadas, se tuvo en cuenta la correlación entre ambas áreas bajo la curva.

Las estimaciones del tamaño del área bajo la curva ROC, su desviación típica, su estimación suavizada, los valores de los estadísticos de contraste, las probabilidades para los contrastes de hipótesis sobre la diferencia entre curvas así como los intervalos de confianza fueron estimados con una macro para Microsoft Excel 2003 diseñada específicamente para este trabajo que incorporaba las salidas de SPSS y de Netica. Las coordenadas de las curvas ROC suavizadas también fueron estimadas utilizando esta macro.

### 7.3. Resultados

Como se puede apreciar en la Figura 7.1, la manipulación del tamaño de la muestra beneficia globalmente a las redes bayesianas. Aunque en ambas técnicas se produce una reducción del área bajo la curva ROC a medida que se reduce el porcentaje de respuestas positivas en la variable dependiente, el área bajo la curva ROC que producen las redes bayesianas es mayor que la obtenida con la regresión logística. Así, mientras que el área bajo la curva ROC para el componente emocional de la actitud emprendedora estimada en la regresión logística es de 0,7761 ( $z = 6,86, p < 0,001$ ), el de la red bayesiana es 0,8344 ( $z = 9,15, p < 0,001$ ); el área para el componente cognitivo estimada usando la regresión logística es 0,6932 ( $z = 5,96, p < 0,001$ ) mientras que el de la red bayesiana es 0,7945 ( $z = 11,12, p < 0,001$ ); y, por último, la estimación del área bajo la curva ROC para la regresión logística es de 0,6337 ( $z = 2,81, p = 0,002$ ) cuando la estimada para la red bayesiana es 0,7235 ( $z = 4,91, p < 0,001$ ).

En términos diferenciales globales, la mayor diferencia estadísticamente significativa entre las curvas producidas por la regresión logística y la red bayesiana se observa con la variable dependiente del componente cognitivo de la actitud emprendedora, donde las redes bayesianas mejoran en un 10,14% respecto a la regresión logística ( $z = 5,76, p < 0,001$ ); o lo que es lo mismo, cuando la tasa de respuestas positivas y negativas está más equilibrada. Por otro lado, cuando la tasa de respuestas positivas de la variable dependiente es baja (como sucede con el componente conductual de la actitud emprendedora) las redes bayesia-



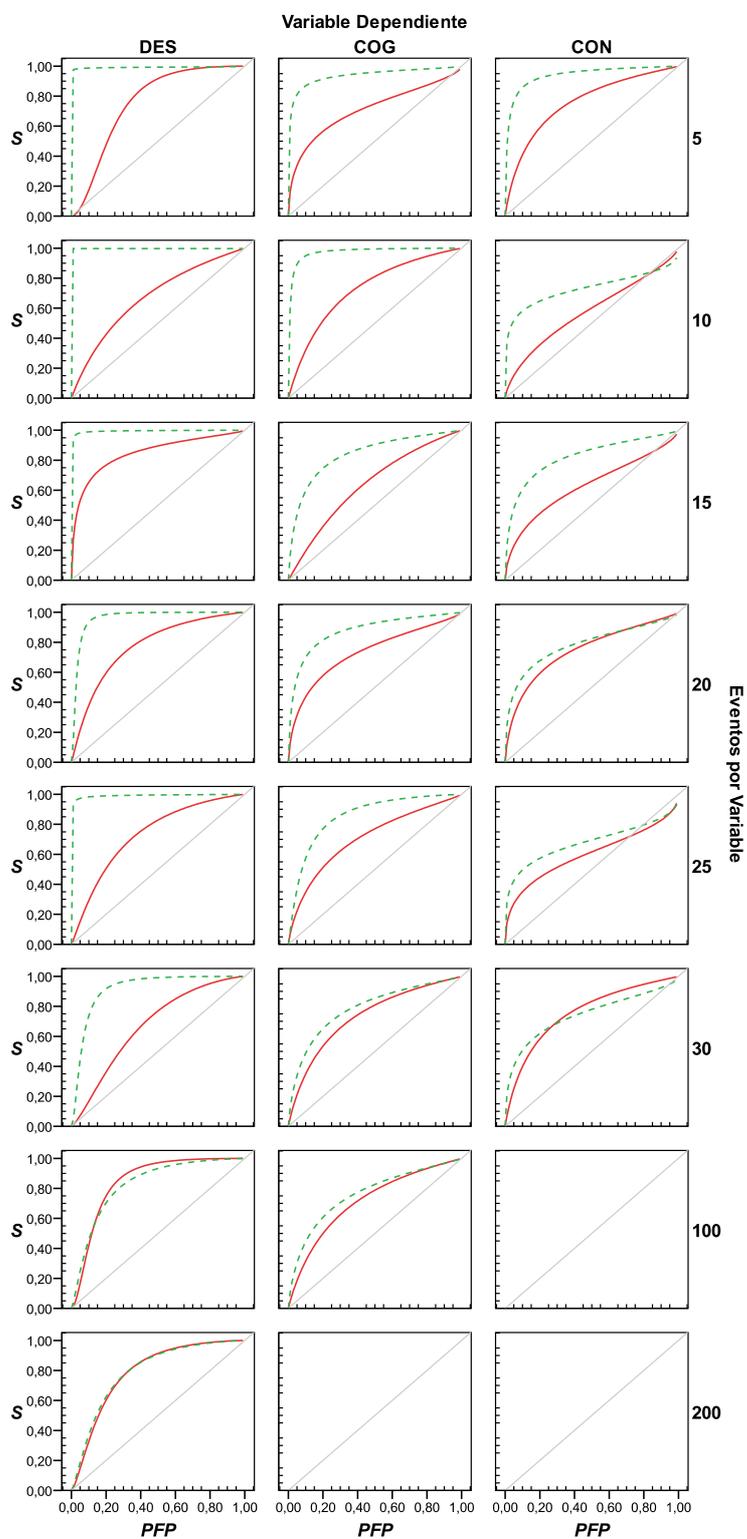
**Figura 7.1:** Curvas ROC globales en función de la variable dependiente. La línea continua representa a la regresión logística y la discontinua a la red bayesiana.

nas también muestran una mejora estadísticamente significativa del 8,68 % en la predicción en comparación con la regresión logística ( $z = 2,41, p = 0,008$ ). Sin embargo, no se observan diferencias estadísticamente significativas en términos globales entre las redes bayesianas y la regresión logística cuando la tasa de respuestas positivas es muy elevada en la variable dependiente como pasa en el caso del componente emocional de la actitud emprendedora ( $z = 1,31, p = 0,095$ ).

Si hacemos un análisis detallado teniendo en cuenta el tamaño de la muestra en términos de eventos por variable independiente y nos fijamos en las áreas bajo las curvas ROC podemos extraer conclusiones interesantes. Cuando la tasa de respuestas positivas en la variable dependiente es muy alto, como pasa con el componente emocional de la actitud emprendedora, las redes bayesianas muestran una ejecución casi perfecta cuando el número de eventos por variable es relativamente bajo, con tamaños muestrales inferiores a 100 casos. Sin embargo, cuando el número de eventos por variable es relativamente grande (100 o 200) que suponen muestras entre 400 y 800 casos, la ejecución de las redes bayesianas y la regresión logística es prácticamente la misma (Tabla 7.2) y no existen diferencias estadísticamente significativas entre las áreas ROC generadas por ambas técnicas (Tabla 7.3 y Figura 7.2). Otra observación importante es, como se puede apreciar en la Tabla 7.2, que las curvas ROC generadas por la regresión logística cuando el tamaño de muestra es de 20 y 40 (con 5 y 10 eventos por variable respectivamente) no alcanzan a ser estadísticamente diferentes de 0,5 con un nivel de confianza del 95 %.

Cuando analizamos las curvas ROC generadas por los modelos de red bayesia-

Estudio 2. Tamaño de muestra y área bajo la curva ROC



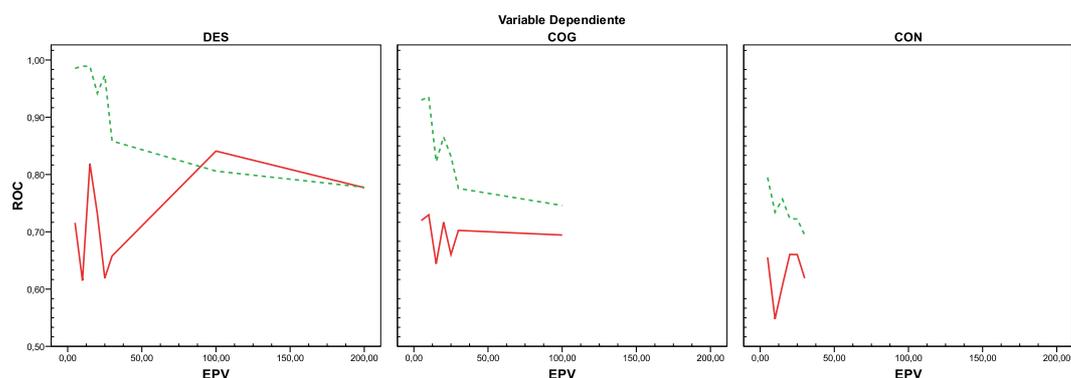
**Figura 7.2:** Curvas ROC en función del número de eventos por variable. La línea continua representa a la regresión logística y la discontinua a la red bayesiana.

**Tabla 7.2:** Áreas bajo la curva ROC en función del tamaño de la muestra ( $n$ ), de la variable dependiente ( $VD$ ) y del tipo de técnica.

Técnica	$VD$	$n$	$\theta$	$DT_\theta$	$\theta^*$	$z_{\theta=0}$	$p_{\theta=0}$	$IC_{inf}$	$IC_{sup}$
RL	DES	20	0,7157	0,1343	0,7444	1,61	0,054	0,4525	0,9789
	DES	40	0,6149	0,1280	0,6599	0,90	0,185	0,3640	0,8657
	DES	61	0,8191	0,0624	0,8364	5,11	*	0,6968	0,9414
	DES	81	0,7316	0,0588	0,7621	3,94	*	0,6164	0,8468
	DES	101	0,6192	0,0707	0,7205	1,69	0,046	0,4807	0,7577
	DES	121	0,6579	0,0615	0,6595	2,57	0,005	0,5373	0,7785
	DES	404	0,8409	0,0248	0,8354	13,75	*	0,7923	0,8894
	DES	804	0,7770	0,0202	0,7784	13,70	*	0,7374	0,8166
	COG	53	0,7197	0,0897	0,7003	2,45	0,007	0,5438	0,8956
	COG	106	0,7297	0,0542	0,7162	4,24	*	0,6235	0,8360
	COG	159	0,6442	0,0463	0,6257	3,12	0,001	0,5535	0,7350
	COG	213	0,7167	0,0386	0,7222	5,61	*	0,6411	0,7924
	COG	266	0,6602	0,0356	0,6969	4,50	*	0,5905	0,7299
	COG	319	0,7025	0,0315	0,7207	6,42	*	0,6407	0,7644
	COG	1063	0,6944	0,0175	0,7016	11,10	*	0,6601	0,7288
	CON	169	0,6555	0,0908	0,7475	1,71	0,043	0,4775	0,8335
	CON	339	0,5478	0,0677	0,5685	0,71	0,240	0,4152	0,6804
	CON	508	0,6065	0,0548	0,6202	1,94	0,026	0,4990	0,7139
	CON	677	0,6605	0,0452	0,7368	3,55	*	0,5719	0,7491
	CON	846	0,6602	0,0405	0,5906	3,95	*	0,5808	0,7396
CON	1016	0,6191	0,0365	0,7454	3,27	0,001	0,5477	0,6906	
RB	DES	20	0,9854	0,0365	0,9773	13,29	*	0,9138	1,0000
	DES	40	0,9891	0,0308	0,9821	15,88	*	0,9287	1,0000
	DES	61	0,9889	0,0134	0,9804	36,57	*	0,9627	1,0000
	DES	81	0,9417	0,0375	0,9505	11,79	*	0,8683	1,0000
	DES	101	0,9732	0,0146	0,9775	32,45	*	0,9447	1,0000
	DES	121	0,8584	0,0529	0,9052	6,78	*	0,7547	0,9621
	DES	404	0,8060	0,0274	0,8166	11,17	*	0,7523	0,8597
	DES	804	0,7778	0,0203	0,7834	13,71	*	0,7381	0,8175
	COG	53	0,9299	0,0397	0,9260	10,82	*	0,8520	1,0000
	COG	106	0,9356	0,0218	0,9635	19,94	*	0,8927	0,9784
	COG	159	0,8225	0,0352	0,8292	9,18	*	0,7536	0,8914
	COG	213	0,8656	0,0271	0,8662	13,51	*	0,8126	0,9187
	COG	266	0,8315	0,0254	0,8451	13,05	*	0,7817	0,8814
	COG	319	0,7758	0,0278	0,7764	9,92	*	0,7213	0,8303
	COG	1063	0,7460	0,0156	0,7507	15,75	*	0,7154	0,7766
	CON	169	0,7950	0,0793	0,9289	3,72	*	0,6395	0,9504
	CON	339	0,7343	0,0663	0,7211	3,53	*	0,6043	0,8643
	CON	508	0,7570	0,0490	0,8132	5,24	*	0,6609	0,8530
	CON	677	0,7230	0,0426	0,7739	5,24	*	0,6395	0,8065
	CON	846	0,7222	0,0373	0,6694	5,97	*	0,6492	0,7953
CON	1016	0,6943	0,0344	0,7324	5,65	*	0,6269	0,7617	

RL: regresión logística, RB: red bayesiana, DT: desviación típica, IC: intervalo de confianza al 95% y \*:  $p < 0,001$ .

## Estudio 2. Tamaño de muestra y área bajo la curva ROC



**Figura 7.3:** Evolución del área bajo la curva ROC en función de la variable dependiente y del número de eventos por variable. La línea discontinua representa a la red bayesiana y la sólida a la regresión logística.

na y regresión logística que incluyen como variables dependientes al componente cognitivo y conductual de la actitud hacia la creación de empresas se observa que las diferencias son menos grandes. Sin embargo, en todos los casos la curva ROC generada por la red bayesiana deja bajo sí un área mayor que la que produce la regresión logística (Tabla 7.2).

Otra observación importante es que el área bajo la curva ROC se va reduciendo sensiblemente en la curvas generadas por las redes bayesianas a medida que se aumenta en número de eventos por variables. Este cambio no es tan evidente en el caso de la regresión logística (Tabla 7.2 y Figura 7.2).

Así pues, como se puede observar en la Figura 7.3, se pueden hacer dos apreciaciones básicas en relación a la evolución del área bajo la curva ROC en ambas técnicas a medida que aumenta el número de eventos por variable. En primer lugar, el área bajo la curva ROC tiende a descender a medida que el número de eventos por variable va aumentando cuando se utiliza la red bayesiana mientras que el área para la regresión logística tiende a mantenerse más estable. Además, el área bajo la curva ROC producida por la regresión logística tiende a ser inferior a la producida por la red bayesiana excepto en el caso de variables dependientes con una alta tasa de respuestas positivas donde ambas técnicas funcionan similarmente cuando la tasa de eventos por variable está en torno al 100.

Por último, el tamaño de las diferencias entre las áreas bajo la curva ROC para la regresión logística y la red bayesiana aparecen en la Tabla 7.4. Como se puede observar, las mayores diferencias se observan cuando el número de eventos por

**Tabla 7.3:** Comparación de las áreas bajo la curva ROC en función de la variable dependiente ( $VD$ ) y del tamaño de muestra ( $n$ ).

$VD$	$n$	$r_p$	$r_a$	$r$	$DT_{\theta_{RL}-\theta_{RB}}$	$z_{\theta_{RL}-\theta_{RB}}$	$p_{\theta_{RL}\neq\theta_{RB}}$
DES	20	0,5014	0,4961	0,41	0,12	2,177	0,015
DES	40	0,2000	0,0000	0,08	0,13	2,895	0,002
DES	61	0,4559	0,1150	0,21	0,06	2,783	0,003
DES	81	0,3348	0,5503	0,38	0,06	3,724	*
DES	101	0,0516	0,2135	0,07	0,07	4,977	*
DES	121	0,2481	0,2051	0,19	0,07	2,742	0,003
DES	404	0,5904	0,7952	0,63	0,02	1,546	0,061
DES	804	0,7144	0,7915	0,7	0,02	0,049	0,481
COG	53	0,4639	0,1993	0,28	0,09	2,405	0,008
COG	106	0,6847	0,3076	0,42	0,05	4,183	*
COG	159	0,7420	0,4566	0,54	0,04	4,429	*
COG	213	0,6669	0,4715	0,51	0,03	4,376	*
COG	266	0,5373	0,6258	0,54	0,03	5,605	*
COG	319	0,7082	0,7001	0,66	0,02	2,966	0,002
COG	1063	0,8841	0,8434	0,84	0,01	5,405	*
CON	169	0,7359	0,7000	0,67	0,07	1,995	0,023
CON	339	0,6669	0,5309	0,55	0,06	2,934	0,002
CON	508	0,6894	0,6638	0,63	0,04	3,347	*
CON	677	0,8774	0,6671	0,73	0,03	1,933	0,027
CON	846	0,8258	0,6890	0,71	0,03	2,083	0,019
CON	1016	0,7952	0,7025	0,71	0,03	2,778	0,003

$DT$ : desviación típica,  $r_p$ : correlación entre las respuestas de las técnicas para los casos con respuesta positiva,  $r_a$ : correlación entre las respuestas de las técnicas para los casos con respuesta negativa,  $r$ : correlación entre las áreas bajo las curvas, \*:  $p < 0,001$ .

## Estudio 2. Tamaño de muestra y área bajo la curva ROC

**Tabla 7.4:** Tamaño de las diferencias entre áreas bajo la curva ROC en función del número de eventos por variable y la variable dependiente (VD).

		VD		
		DES	COG	CON
Eventos por variable	5	26,96	21,01	13,95
	10	37,42	20,58	18,65
	15	16,98	17,83	15,05
	20	21,02	14,89	6,25
	25	35,41	17,14	6,20
	30	20,05	7,33	7,52
	100	3,49	5,16	
	200	0,08		
		<i>M</i>	20,17	14,85
	<i>DT</i>	13,46	6,26	5,31

Expresado en porcentaje y valor absoluto.

variable es inferior a 100 para el caso de la dimensión emocional de la actitud hacia la creación de empresas, cuando el número de eventos por variable es inferior a 30 en el componente cognitivo de la actitud y cuando el número de eventos positivos es igual o inferior a 25 en el componente conductual. En relación a la variable dependiente, se puede apreciar que el componente de deseabilidad en la creación de empresas es el que produce una mayor ejecución diferencial, mientras que el componente cognitivo es el que muestra un porcentaje de diferencia más bajo.

## 7.4. Discusión parcial

El hallazgo más relevante de este estudio, como pensábamos inicialmente, es que la manipulación del tamaño de la muestra, o más concretamente del número de eventos por variable, tiene un efecto sobre la habilidad predictiva más pronunciado sobre la regresión logística que sobre las redes bayesianas. En términos generales, la habilidad predictiva de la regresión logística se ve más amenazada, en comparación con las redes bayesianas, a medida que la muestra se hace más pequeña. El hecho de que ambas técnicas converjan en su tasa predictiva en muestras grandes pone de manifiesto la naturaleza asintótica de la teoría clásica (p.

e., Firth, 1993; Freedman y Pee, 1989) frente a la perspectiva bayesiana (p. e., Alonso y Tubau, 2002; Cowell et al., 1999; De la Fuente et al., 2002; Heckerman, 1995; Serrano, 2003).

También se ha puesto de manifiesto que el área bajo la curva ROC se reduce a medida que disminuye la tasa de respuestas positivas en la variable dependiente aunque, en todos los casos, la red bayesiana sigue comportándose mejor, en términos predictivos, que la regresión logística. Estos resultados son consistentes con la idea de que en muestras balanceadas se necesita menor muestra para que funcione bien la regresión logística mientras que se necesita un mayor tamaño muestral cuando hay una baja prevalencia en la variable resultado (Hsieh et al., 1998; Whittermore, 1981).

Estos resultados son consistentes con los encontrados por Concato et al. (1995) quienes observaron que los coeficientes de modelo de regresión logística tienden a ser sesgados cuando la variable dependiente tiene valores reducidos de eventos por variable en el modelo. Así pues, además de producirse un aumento en la varianza de las estimaciones del modelo de regresión logística, de producirse ajustes paradójicos y de reducirse la potencia estadística (Peduzzi et al., 1995, 1996); la reducción del número de eventos por variable tiende a ser más crítico para la regresión logística que para las redes bayesianas en términos de validez predictiva.

Los resultados aquí obtenidos pueden ser de utilidad como criterio de toma de decisiones ante la elección de una u otra técnica teniendo en cuenta las condiciones de los datos y del modelo deseado. Dado que una de las desventajas más llamativas de la regresión logística es el trabajar con muestras pequeñas (p. e., Bull et al., 2002; Long, 1997; Ortega y Cayuela, 2002; Silva y Barroso, 2004), el uso de las redes bayesianas puede considerarse como una alternativa más a las opciones ya disponibles para subsanar tal debilidad. Por tanto, el uso de una red bayesiana para la construcción de modelos probabilísticos puede tener relevancia como alternativa a la regresión logística clásica junto a la regresión logística exacta (p. e., Hirji et al., 1987; Mehta y Patel, 1995) o a las funciones de minimización del error asintótico en la estimación (p. e., Bull et al., 2002).

## Estudio 2. Tamaño de muestra y área bajo la curva ROC

# Capítulo 8

## Estudio 3. Nivel de categorización y área bajo la curva ROC

*"Everything is dangerous, [...]. If it wasn't so, life wouldn't be worth living."*

(Wilde, 1895/1998)

### 8.1. Descripción

Reade y Kupper (1995) señalan que la recategorización de las variables predictoras puede ser una amenaza para los resultados de la regresión logística en los estudios sociales. Pese a que la recategorización de variables puede tener consecuencias positivas desde el punto de vista estadístico y sustantivo (Cusmille y Bangdiwala, 2000; Zhao y Kolonel, 1992), también pueden aparecer efectos indeseables que es necesario tratar adecuadamente (p. e., Altman et al., 1994; Flegal et al., 1991). Por tanto, en este estudio se evaluará el efecto diferencial que produce la recodificación de todas las variables independientes de un modelo de regresión logística y de red bayesiana en términos del área ROC que generan. Se espera que la categorización de variables tenga un menor impacto sobre el área bajo la curva ROC generada por las redes bayesianas en contraposición a la generada por la regresión logística.

## 8.2. Método

### 8.2.1. Muestra

Una muestra de respuestas al cuestionario que incluía las escalas de actitud hacia la creación de empresas, motivación, preparación percibida, obstáculos percibidos y carencias percibidas proveniente de 1230 estudiantes (426 hombres y 797 mujeres) de la Universidad de Almería fue usada en este estudio. Las edades de los participantes oscilaron entre los 17 y los 56 años, siendo la media 22,45 y la desviación típica 4,46. Una descripción más detallada del procedimiento de muestreo así como de las características de la muestra puede encontrarse en la sub-sección 5.3.1.

### 8.2.2. Análisis de datos

Todas las variables fueron recategorizadas en dos, cinco y diez clases a intervalos de igual anchura (Domingos y Pazzni, 1996). En primer lugar se calculó la anchura del intervalo para cada tipo de recodificación con  $\frac{max_n - min_n}{nc_i}$ , donde  $max_n - min_n$  se refiere al rango de la variable  $n$  y  $nc_i$  es el número de categorías de la nueva variable. En la Tabla 8.1 aparecen los valores de corte usados para construir las nuevas variables.

Se construyeron 24 modelos, de los cuales 12 fueron redes bayesianas y otros 12 modelos de regresión logística. Para cada variable dependiente (componente emocional, componente cognitivo y componente conductual de la actitud emprendedora) se estimaron cuatro modelos: uno donde no se recodificaron las variables independientes y los otros tres donde se categorizaron en dos, tres, cinco y diez categorías.

Los modelos de regresión logística se estimaron con el paquete estadístico SPSS 15.0. No se incluyó ningún término de interacción y el nivel de confianza de las estimaciones se ajustó al 95%. Se forzó al programa a introducir en los modelos de regresión logística las variables definidas previamente por el diseño del estudio.

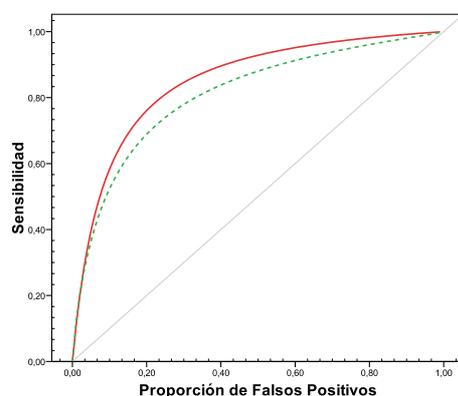
La estimación paramétrica de las redes bayesianas se hizo con el método de máxima verosimilitud usando la corrección de Laplace, mientras que las estructuras cualitativas se construyeron manualmente con base al diseño del estudio.

**Tabla 8.1:** Valores de corte para las variables independientes.

Cateogρίας	A	C	P	O	M
2	1,69	0,25	0,25	0,24	0,45
	2,69	2,54	2,63	2,47	2,73
	3,69	4,83	5,00	4,71	5,00
5	1,69	0,25	0,25	0,24	0,45
	2,09	1,17	1,20	1,13	1,36
	2,49	2,08	2,15	2,02	2,27
	2,89	3,00	3,10	2,92	3,18
	3,29	3,92	4,05	3,81	4,09
	3,69	4,83	5,00	4,71	5,00
10	1,69	0,25	0,25	0,24	0,45
	1,89	0,71	0,73	0,68	0,91
	2,09	1,17	1,20	1,13	1,36
	2,29	1,63	1,68	1,58	1,82
	2,49	2,08	2,15	2,02	2,27
	2,69	2,54	2,63	2,47	2,73
	2,89	3,00	3,10	2,92	3,18
	3,09	3,46	3,58	3,36	3,64
	3,29	3,92	4,05	3,81	4,09
	3,49	4,38	4,53	4,26	4,55
3,69	4,83	5,00	4,71	5,00	

A: nivel de actitud emprendedora medida con la escala ACEMP, C: carencias percibidas, P: preparación percibida, O: obstáculos percibidos, M: motivación.

### Estudio 3. Nivel de categorización y área bajo la curva ROC



**Figura 8.1:** Efecto global de la recategorización sobre la curva ROC. La línea continua representa a la regresión logística y la discontinua a la red bayesiana.

Para estimar los parámetros y para obtener las probabilidades posteriores se usó el programa Netica 4.02.

Las estimaciones del tamaño del área bajo la curva ROC (Hanley y McNeil, 1982), su desviación típica, su estimación suavizada, los valores de los estadísticos de contraste, las probabilidades para los contrastes de hipótesis sobre la diferencia entre curvas (Hanley y McNeil, 1983) así como los intervalos de confianza fueron estimados con una macro para Microsoft Excel 2003 diseñada específicamente para este trabajo que incorporaba las salidas de SPSS y de Netica. Las coordenadas de las curvas ROC suavizadas también fueron estimadas utilizando esta macro.

Una descripción más detallada sobre los análisis de los datos puede encontrarse en la sub-sección 7.2.2.

## 8.3. Resultados

El resultado más llamativo es, contrariamente a lo que pensábamos en un principio y como se puede apreciar en la Figura 8.1, que la recategorización benefició globalmente a la regresión logística. Aunque tanto la regresión logística ( $\theta = 0,8365, z = 28,19, p < 0,001$ ) como las redes bayesianas ( $\theta = 0,7992, z = 23,18, p < 0,001$ ) generaron áreas bajo la curva ROC aceptables, se observa una diferencia estadísticamente significativa del 3,72 % favorable a la regresión logística ( $z = 4,06, p < 0,001$ ).

Como se puede observar en la Tabla 8.2 y en la Figura 8.2 las redes bayesianas experimentan problemas cuando se recodifican las variables predictoras para

explicar la dimensión cognitiva y conductual de la actitud hacia la creación de empresas. En estos casos, las redes bayesianas no llegan a mostrar una ejecución que genere una curva ROC estadísticamente diferente de 0,5 (Tabla 8.2) mientras que en el caso de la regresión logística las áreas bajo la curva ROC son siempre estadísticamente diferentes de este valor de referencia.

No obstante, cuando la variable dependiente es la deseabilidad de crear una empresa, las redes bayesianas se comportan igual o mejor que la regresión logística. Así, cuando la recodificación de las variables independientes fue en cinco o en diez categorías las redes bayesianas generan áreas bajo la curva ROC estadísticamente superiores a las generadas por la regresión logística ( $z = 1,93, p < 0,05$  y  $z = 2,00, p < 0,05$ , respectivamente). En el caso de la dicotomización, la regresión logística obtiene un 1,3% de ventaja predictiva respecto a las redes bayesianas pero esta diferencia no es estadísticamente significativa ( $z = 1,36, p = 0,09$ ).

También se puede observar (Tabla 8.2) que las redes bayesianas son mejores técnicas de predicción cuanto las puntuaciones de las variables independientes no están categorizadas independientemente de cual sea la variable dependiente y que esta superioridad es estadísticamente significativa (Tabla 8.3).

Por último, hay que señalar que la mayor superioridad de la regresión logística sobre las redes bayesianas se observa en el caso del modelo diseñado para predecir la dimensión cognitiva de la actitud emprendedora donde las diferencias entre ambas técnicas en términos de áreas bajo la curva ROC oscila entre el 13,19% y el 18,06% cuando se recodifican las variables independientes (Tabla 8.4).

## 8.4. Discusión parcial

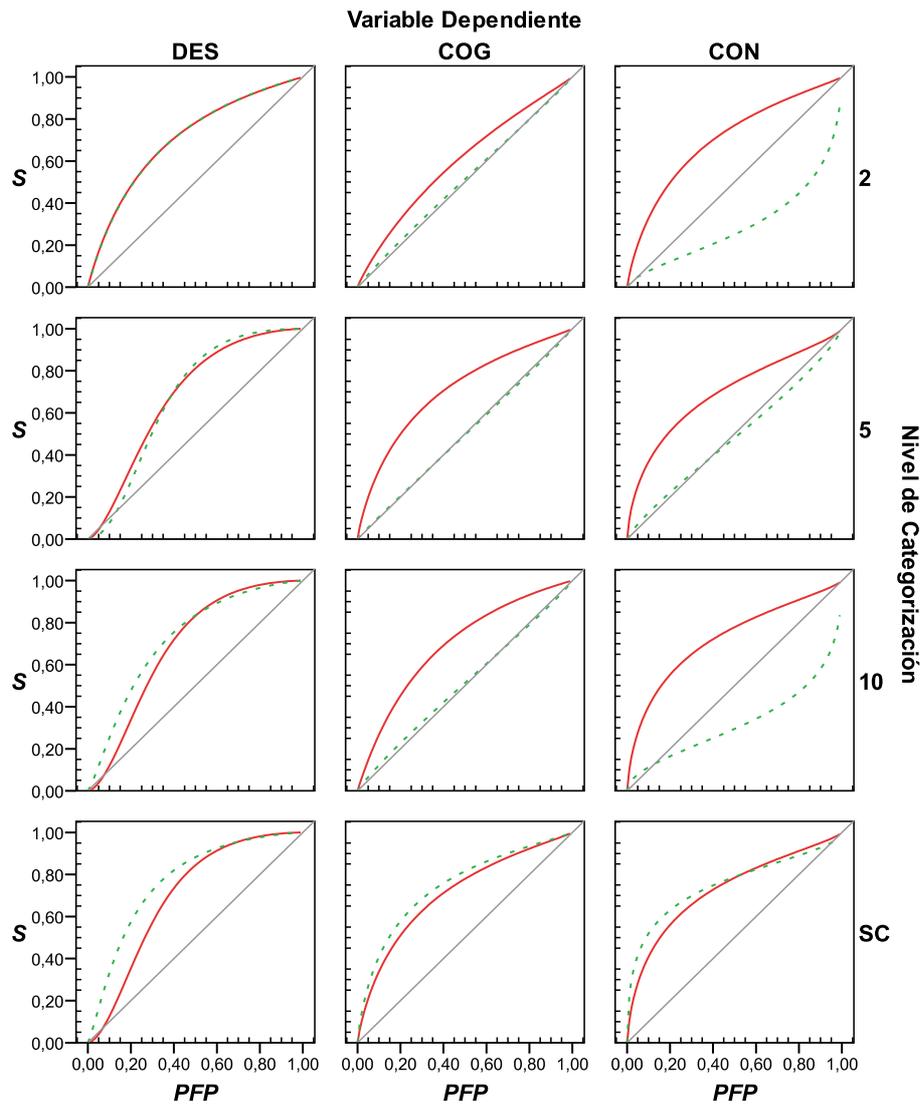
La observación más llamativa de este estudio es que la regresión logística funciona mejor ante la recodificación en dos, cinco y diez categorías cuando las respuestas positivas y negativas están presentes en proporciones similares en la variable dependiente (como es el caso de la dimensión cognitiva de la actitud hacia la creación de empresas) o cuando la tasa de respuestas negativas es muy elevada como sucede en el caso de la dimensión conductual de la actitud. No obstante, cuando la proporción de respuestas positivas es muy elevada, como sucede con el aspecto emocional de la actitud ante la creación de empresas, las redes bayesianas muestran una mejor ejecución en la clasificación cuando las variables independientes

Estudio 3. Nivel de categorización y área bajo la curva ROC

**Tabla 8.2:** Áreas bajo la curva ROC en función del Nivel de categorización (NC), de la variable dependiente ( $VD$ ) y del tipo de técnica.

Técnica	$VD$	NC	$\theta$	$DT_\theta$	$\theta^*$	$z_{\theta=0}$	$p_{\theta=0}$	$IC_{inf}$	$IC_{sup}$
RL	DES	2	0,6673	0,0189	0,6943	8,87	*	0,6303	0,7042
	DES	5	0,6752	0,0186	0,6716	9,42	*	0,6388	0,7117
	DES	10	0,6972	0,0183	0,6795	10,80	*	0,6614	0,7330
	DES	SC	0,7030	0,0182	0,6878	11,14	*	0,6673	0,7387
	COG	2	0,6385	0,0175	0,5849	7,90	*	0,6041	0,6729
	COG	5	0,6753	0,0165	0,6955	10,63	*	0,6430	0,7076
	COG	10	0,6772	0,0166	0,6822	10,66	*	0,6446	0,7097
	COG	SC	0,6880	0,0164	0,7006	11,43	*	0,6558	0,7202
	CON	2	0,6320	0,0326	0,6926	4,05	*	0,5680	0,6959
	CON	5	0,6366	0,0335	0,6837	4,08	*	0,5710	0,7022
	CON	10	0,6492	0,0335	0,7121	4,46	*	0,5837	0,7148
	CON	SC	0,6412	0,0336	0,7171	4,21	*	0,5754	0,7070
RB	DES	2	0,6549	0,0182	0,6941	8,53	*	0,6193	0,6905
	DES	5	0,6924	0,0175	0,6654	11,01	*	0,6582	0,7267
	DES	10	0,7161	0,0174	0,7155	12,43	*	0,6821	0,7502
	DES	SC	0,7549	0,0167	0,7604	15,28	*	0,7222	0,7876
	COG	2	0,5066	0,0170	0,4994	0,39	0,349	0,4732	0,5400
	COG	5	0,4947	0,0170	0,4866	-0,31	0,622	0,4614	0,5281
	COG	10	0,5050	0,0170	0,4991	0,30	0,383	0,4717	0,5384
	COG	SC	0,7290	0,0148	0,7369	15,47	*	0,7000	0,7580
	CON	2	0,5136	0,0293	0,2834	0,46	0,322	0,4562	0,5709
	CON	5	0,5247	0,0298	0,4761	0,83	0,204	0,4662	0,5831
	CON	10	0,5334	0,0299	0,3150	1,12	0,132	0,4749	0,5919
	CON	SC	0,7015	0,0314	0,7406	6,41	*	0,6398	0,7631

RL: regresión logística, RB: red bayesiana,  $DT$ : desviación típica,  $IC$ : intervalo de confianza al 95 %, SC: sin categorizar, \*:  $p < 0,001$ .



**Figura 8.2:** Curvas ROC en función del nivel de categorización. La línea continua representa a la regresión logística y la discontinua a la red bayesiana. SC: sin categorizar.

Estudio 3. Nivel de categorización y área bajo la curva ROC

**Tabla 8.3:** Comparación de áreas bajo la curva ROC en función del nivel de categorización (NC), de la variable dependiente (VD) y del tipo de técnica.

VD	NC	$r_p$	$r_a$	$r$	$DT_{\theta_{RL}-\theta_{RB}}$	$z_{\theta_{RL}-\theta_{RB}}$	$p_{\theta_{RL}\neq\theta_{RB}}$
DES	2	0,9973	0,9985	0,88	0,01	1,358	0,087
DES	5	0,9336	0,9588	0,88	0,01	1,931	0,027
DES	10	0,8688	0,9147	0,86	0,01	1,997	0,023
DES	SC	0,7064	0,7753	0,71	0,01	3,883	*
COG	2	0,0495	0,0053	0,02	0,02	5,449	*
COG	5	0,0787	0,0397	0,04	0,02	7,777	*
COG	10	0,0810	0,0548	0,05	0,02	7,424	*
COG	SC	0,8789	0,8571	0,84	0,01	4,569	*
CON	2	0,0651	0,0168	0,04	0,04	2,757	0,003
CON	5	0,2000	0,0197	0,09	0,04	2,618	0,004
CON	10	0,0525	0,0025	0,02	0,04	2,610	0,005
CON	SC	0,8094	0,7212	0,73	0,02	2,513	0,006

*DT*: desviación típica, *SC*: sin categorizar,  $r_p$ : correlación entre las respuestas de las técnicas para los casos con respuesta positiva,  $r_a$ : correlación entre las respuestas de las técnicas para los casos con respuesta negativa,  $r$ : correlación entre las áreas bajo las curvas, \*:  $p < 0,001$ .

**Tabla 8.4:** Tamaño de las diferencias entre áreas bajo la curva ROC en función del número de categorías en las variables independientes y la variable dependiente (VD). *SC*: sin categorizar.

	VD			
	DES	COG	CON	
	2	1,23	13,19	11,84
Número de categorías	5	1,72	18,06	11,19
	10	1,89	17,21	11,59
	SC	5,19	4,10	6,02
	<i>M</i>	2,51	13,14	10,16
	<i>DT</i>	1,81	6,39	2,77

Expresado en porcentaje y valor absoluto.

son recodificadas en cinco o diez categorías.

En cualquier caso, estas conclusiones preliminares sobre la comparación del efecto que tiene la recodificación sobre el área bajo la curva ROC han de tomarse con cautela porque se pueden identificar ciertos aspectos que pueden estar sesgando los resultados. En primer lugar, el formato de recodificación elegido puede haber influido, principalmente, sobre la bondad predictiva de las redes bayesianas. Por ejemplo, J. García et al. (2007) compararon las redes bayesianas con los modelos de ecuaciones estructurales recodificando los valores de la red usando los cuartiles y observaron índices de validez predictiva similares entre ambas técnicas. Estos resultados pueden considerarse, a falta de más investigaciones, una prueba indirecta, en contraste con el trabajo de J. García et al. (2007), para intuir cuáles son los métodos de recodificación más recomendables cuando se utiliza una u otra técnica. Es más, como señalan Zhao y Kolonel (1992) el uso de una categorización basada en los cuartiles de las variables tiene ciertas ventajas sobre aspectos como los supuestos del modelo o sobre la potencialidad de comparación entre diferentes estudios.

En segundo lugar, la naturaleza distribucional de cada variable dependiente también tendría que haberse tenido en cuenta a la hora de hacer la recodificación (Ragland, 1992). Por tanto, aunque habría que testar específicamente esta hipótesis, los resultados obtenidos apuntan a que la regresión logística parece ser menos sensible al modo de recodificación de las variables independientes sin tener en cuenta los supuestos distribucionales en comparación con las redes bayesianas.

Otra hipótesis tentativa para explicar la falta de superioridad de las redes bayesianas frente a la regresión logística podría pasar por considerar que la verdadera estructura de dependencias no es la que describe el clasificador simple bayesiano. En este caso, sería deseable diseñar otras estructuras de red bayesiana (clasificadores bayesianos enriquecidos, por ejemplo) que contengan a las variables actitudinales y testar las hipótesis de diferencia en la ejecución predictiva nuevamente.

### Estudio 3. Nivel de categorización y área bajo la curva ROC

# Capítulo 9

## Estudio 4. Tipo de categorización y área bajo la curva ROC

*"In every live we have some troubles  
but when you worry you make it double."*

(McFerrin, 1988)

### 9.1. Descripción

Este estudio surge como respuesta a los resultados obtenidos en el Estudio 3. Dado que, contrariamente a las hipótesis planteadas, la regresión logística toleró mejor la recodificación de las variables independientes frente a las redes bayesianas; en este estudio se realizarán otro tipo de recodificaciones sobre estas variables para ponderar su efecto sobre las curvas ROC generadas por una y otra técnica. En concreto, se utilizarán parámetros muestrales (percentiles y puntuaciones típicas normalizadas) para generar nuevas variables recodificadas. En consonancia con nuestras hipótesis previas y con los estudios revisados (p. e., J. García et al., 2007; Zhao y Kolonel, 1992), creemos que la validez predictiva de las redes bayesianas igualará o mejorará a la de la regresión logística en estas circunstancias.

## 9.2. Método

### 9.2.1. Muestra

Las 1230 observaciones provenientes del muestreo de los primeros y últimos cursos de las titulaciones que se imparten en la Universidad de Almería fueron utilizadas en este estudio. Un total de 426 participantes fueron hombres y el resto (797) mujeres; mientras que la edad de los participantes estuvo comprendida entre los 17 y los 56 años, donde la media fue de 22,45 y la desviación típica fue de 4,46. Una descripción más detallada del procedimiento de muestreo, así como de la muestra en si puede encontrarse en la sub-sección 5.3.1 del capítulo de Metodología General.

### 9.2.2. Análisis de datos

Todas las variables independientes fueron recodificadas en esta ocasión utilizando dos procedimientos basados en parámetros muestrales. Hubo dos tipos de recodificaciones: basadas en los percentiles, y basadas en la normalización de las puntuaciones típicas de cada variable. Todas las variables se recodificaron en dos, cinco y diez categorías.

Los puntos de corte para el formato de la recodificación en formato de percentiles fueron: a) el percentil 50 en el caso de la dicotomización, b) los percentiles 20, 40, 60 y 80 en el caso de cinco categorías, y c) los deciles. En la Tabla 9.1 aparecen los puntos de corte para cada variable independiente utilizada en el estudio.

La estandarización de las variables se realizó estimando las puntuaciones típicas  $z$  esperadas que corresponderían a la curva normal acumulada tipificada utilizando la expresión genérica

$$Y_i^t = \Phi^{-1} \left( \frac{r_i - c}{n - 2c + 1} \right), \quad (9.1)$$

donde  $\Phi^{-1}(\cdot)$  se refiere a la función de distribución normal acumulativa,  $r_i$  es el rango asignado y  $n$  representa el número de casos válido. Dado que se utilizó la transformada de Tukey, el parámetro  $c$  valió  $\frac{1}{3}$ . Las categorías se formaron usando intervalos de igual anchura utilizando el rango comprendido entre  $-3,5$  y  $3,5$ .

Se estimaron 36 modelos estadísticos de los cuales una mitad fueron modelos

**Tabla 9.1:** Valores de corte para las variables independientes para la recodificación basada en los percentiles.

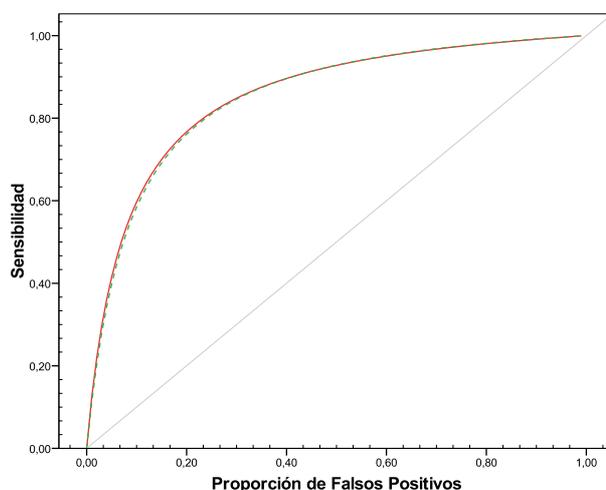
$P_n$	Act. Empr.	Preparación	Obstaculos	Motivación	Carencias
10	2,379	1,000	2,471	2,582	1,500
20	2,517	1,250	2,706	2,818	1,833
30	2,586	1,750	2,882	3,000	2,000
40	2,690	2,000	3,059	3,182	2,167
50	2,759	2,000	3,177	3,273	2,333
60	2,862	2,250	3,353	3,455	2,500
70	2,931	2,250	3,471	3,546	2,750
80	3,035	2,750	3,647	3,727	2,917
90	3,207	3,000	3,882	3,909	3,250

$P_n$ : percentil enésimo, Act. Empr.: actitud emprendedora.

de regresión logística y la otra de red bayesiana. De los 18 modelos que se estimaron para cada técnica un tercio fueron para predecir el componente emocional, otro para el componente cognitivo y el resto para el componente conductual de la actitud emprendedora. Por último, para cada componente se estimaron modelos con variables recodificadas en dos, cinco y diez categorías basados en los percentiles, y modelos con variables recodificadas en dos, cinco y diez categorías basadas en la normalización de las variables predictoras.

Los modelos de regresión logística se estimaron con el paquete estadístico SPSS 15.0, sin incluir ningún término de interacción y el nivel de confianza de las estimaciones se ajustó al 95%. El programa fue forzado a introducir en los modelos de regresión logística las variables definidas previamente por el diseño del estudio. La estimación de los parámetros de las redes bayesianas fue hecha con el método de máxima verosimilitud usando la corrección de Laplace y las estructuras cualitativas se construyeron manualmente con base al diseño del estudio. Para estimar los parámetros y para obtener las probabilidades posteriores se usó el programa Netica 4.02. Las estimaciones del tamaño del área bajo la curva ROC (Hanley y McNeil, 1982), su desviación típica, su estimación suavizada, los valores de los estadísticos de contraste, las probabilidades para los contrastes de hipótesis sobre la diferencia entre curvas, los intervalos de confianza (Hanley y McNeil, 1983) así como las coordenadas de las curvas ROC suavizadas fueron estimadas

#### Estudio 4. Tipo de categorización y área bajo la curva ROC



**Figura 9.1:** Efecto global de la categorización sobre la curva ROC en función de la técnica. La línea continua representa a la regresión logística y la discontinua a la red bayesiana.

con una macro para Microsoft Excel 2003 diseñada específicamente para este trabajo que incorporaba las salidas de SPSS y de Netica.

Una descripción más detallada sobre los análisis de los datos puede encontrarse en la sub-sección 7.2.2.

### 9.3. Resultados

Como se puede apreciar en la Figura 9.1 aunque el área bajo la curva ROC global de este estudio generada por la regresión logística ( $0,8406$ ,  $z = 48,28$ ,  $p < 0,001$ ) es mayor que la generada por la red bayesiana ( $0,8377$ ,  $z = 49,73$ ,  $p < 0,001$ ), esta diferencia en validez predictiva del  $0,29\%$  no es estadísticamente significativa ( $z = 0,82$ ,  $p = 0,207$ ).

Sin embargo, cuando se estiman las curvas segmentando las clasificaciones hechas por cada técnica en función de la variable dependiente, se observan diferencias estadísticamente significativas ( $z = 7,353$ ,  $p < 0,001$ ) a favor de la red bayesiana ( $0,6398$ ,  $z = 11,01$ ,  $p < 0,001$ ) frente a la regresión logística ( $0,5923$ ,  $z = 6,83$ ,  $p < 0,001$ ) para predecir el componente conductual de la actitud emprendedora. Por su parte, tanto las redes bayesianas como la regresión logística obtienen áreas bajo la curva ROC estadísticamente similares cuando se trata de predecir el componente emocional de la actitud (red bayesiana:  $0,6910$ ,

$z = 26,58, p < 0,001$ ; regresión logística:  $0,6931, z = 25,98, p < 0,001$ ; diferencia:  $z = 0,601, p = 0,274$ ) y el componente cognitivo (red bayesiana:  $0,6847, z = 28,97, p < 0,001$ ; regresión logística:  $0,6816, z = 27,06, p < 0,001$ ; diferencia:  $z = 0,957, p = 0,169$ ).

En lo que respecta al análisis pormenorizado que aparece en la Tabla 9.2 y la Figura 9.2 cabe destacar, en primer lugar, que hay dos condiciones (en el modelo para predecir el componente conductual de la actitud emprendedora cuando se dicotomizan las variables tanto con los percentiles como con las variables tipificadas) donde la regresión logística no alcanza a generar áreas bajo la curva ROC estadísticamente diferentes de cero.

También se observa que las redes bayesianas generan curvas ROC más grandes en todas las condiciones planificadas para predecir el componente conductual de la actitud emprendedora (Tabla 9.2) y que todas estas diferencias son estadísticamente significativas excepto para el caso en que se recodificaron las variables atendiendo a los cinco grupos definidos por los percentiles y a los diez grupos generados a partir de la normalización de las variables independientes (Tabla 9.3).

Por su parte, la regresión logística obtiene áreas bajo la curva ROC mayores a las generadas por la red bayesiana en todos los modelos diseñados para predecir el componente emocional de la actitud emprendedora. No obstante, como se puede apreciar en la Tabla 9.3 sólo se observan tres diferencias estadísticamente significativas en las condiciones donde se dicotomizaron las variables (tanto con los percentiles como con las puntuaciones típicas) y cuando se usaron los deciles para formar los grupos.

En el caso del componente cognitivo, se puede apreciar que en la mitad de los casos las redes bayesianas funcionan mejor mientras que en la mitad restante la regresión logística genera curvas ROC más grandes (Tabla 9.2). Sin embargo, la única diferencia estadísticamente significativa que se observa está a favor de las redes bayesianas para el caso en que se recodificaron las variables en cinco categorías utilizando la tipificación de las variables independientes (Tabla 9.3).

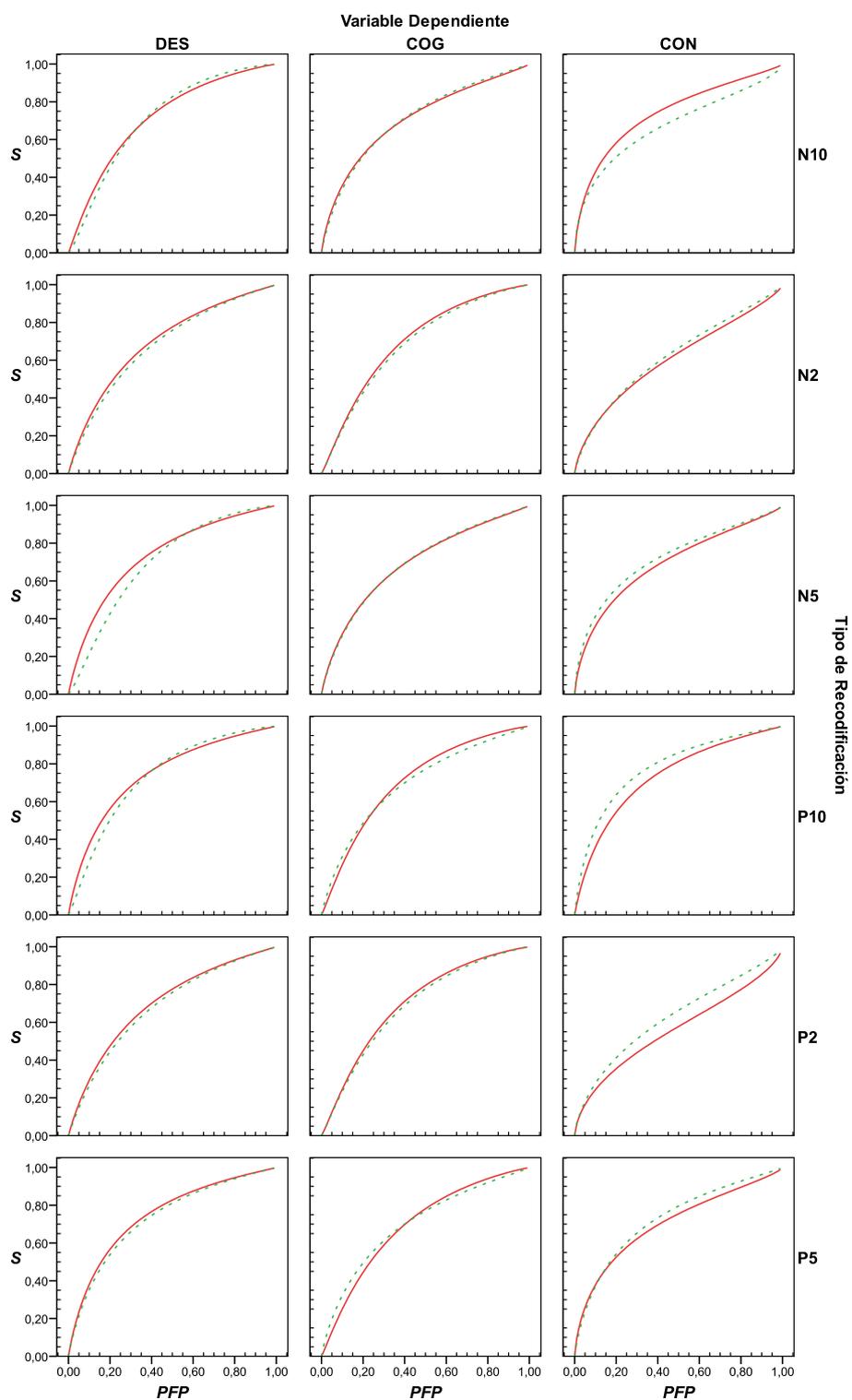
Para terminar, como se puede observar en la Tabla 9.4 las mayores diferencias entre ambas técnicas se encuentran a la hora de predecir el componente conductual de la actitud emprendedora, que como se comentó más arriba es la situación donde mejor funcionan las redes bayesianas. En concreto, la mayor diferencia a

Estudio 4. Tipo de categorización y área bajo la curva ROC

**Tabla 9.2:** Áreas bajo la curva ROC en función del tipo de recodificación (TR), de la variable dependiente (VD) y de la técnica.

Técnica	VD	TR	$\theta$	$DT_{\theta}$	$\theta^*$	$z_{\theta=0}$	$p_{\theta=0}$	$IC_{inf}$	$IC_{sup}$
RL	DES	N2	0,6773	0,0184	0,6897	9,66	*	0,6414	0,7133
	DES	N5	0,6941	0,0183	0,7257	10,59	*	0,6582	0,7301
	DES	N10	0,6892	0,0183	0,7053	10,31	*	0,6532	0,7252
	DES	P2	0,6773	0,0184	0,6897	9,66	*	0,6414	0,7133
	DES	P5	0,7126	0,0179	0,7381	11,89	*	0,6775	0,7476
	DES	P10	0,7066	0,0180	0,7365	11,51	*	0,6715	0,7418
	COG	N2	0,6839	0,0162	0,6917	11,32	*	0,6520	0,7157
	COG	N5	0,6604	0,0169	0,6865	9,50	*	0,6273	0,6935
	COG	N10	0,6854	0,0165	0,7002	11,23	*	0,6530	0,7178
	COG	P2	0,6839	0,0162	0,6917	11,32	*	0,6520	0,7157
	COG	P5	0,6862	0,0165	0,6845	11,26	*	0,6538	0,7186
	COG	P10	0,6937	0,0162	0,7016	11,94	*	0,6619	0,7255
	CON	N2	0,5130	0,0309	0,5981	0,42	0,336	0,4525	0,5736
	CON	N5	0,6185	0,0337	0,6846	3,51	*	0,5524	0,6846
	CON	N10	0,6374	0,0334	0,7310	4,11	*	0,5719	0,7030
	CON	P2	0,5135	0,0309	0,5522	0,44	0,331	0,4530	0,5741
	CON	P5	0,6432	0,0326	0,6929	4,39	*	0,5793	0,7072
	CON	P10	0,6325	0,0331	0,7253	4,01	*	0,5677	0,6973
RB	DES	N2	0,6627	0,0177	0,6735	9,17	*	0,6279	0,6974
	DES	N5	0,6853	0,0179	0,6896	10,33	*	0,6501	0,7204
	DES	N10	0,7048	0,0175	0,7051	11,71	*	0,6705	0,7390
	DES	P2	0,6627	0,0177	0,6735	9,17	*	0,6279	0,6974
	DES	P5	0,7029	0,0176	0,7230	11,51	*	0,6684	0,7375
	DES	P10	0,7229	0,0170	0,7240	13,15	*	0,6897	0,7561
	COG	N2	0,6761	0,0156	0,6785	11,25	*	0,6454	0,7068
	COG	N5	0,6783	0,0158	0,6871	11,27	*	0,6473	0,7093
	COG	N10	0,6981	0,0155	0,7021	12,80	*	0,6678	0,7284
	COG	P2	0,6761	0,0156	0,6785	11,25	*	0,6454	0,7068
	COG	P5	0,6889	0,0156	0,6922	12,11	*	0,6583	0,7195
	COG	P10	0,6879	0,0156	0,6904	12,06	*	0,6574	0,7185
	CON	N2	0,5912	0,0306	0,6139	2,98	0,001	0,5312	0,6513
	CON	N5	0,6584	0,0313	0,7126	5,07	*	0,5971	0,7198
	CON	N10	0,6486	0,0316	0,6663	4,70	*	0,5867	0,7105
	CON	P2	0,5949	0,0306	0,6165	3,10	0,001	0,5350	0,6548
	CON	P5	0,6602	0,0307	0,7173	5,23	*	0,6001	0,7203
	CON	P10	0,6827	0,0309	0,7726	5,91	*	0,6221	0,7432

RL: regresión logística, RB: red bayesiana,  $DT$ : desviación típica,  $IC$ : intervalo de confianza al 95%. N2, N5 y N10: recodificación de las puntuaciones típicas normalizadas a dos, cinco y diez categorías respectivamente. P2, P5 y P10: recodificación utilizando los percentiles en dos, cinco y diez categorías respectivamente. \*:  $p < 0,001$ .



**Figura 9.2:** Curvas ROC en función del tipo de categorización. La línea continua representa a la regresión logística y la discontinua a la red bayesiana.

Estudio 4. Tipo de categorización y área bajo la curva ROC

**Tabla 9.3:** Comparación de las áreas bajo la curva ROC en función de la variable dependiente (VD) y del tipo de recodificación (TR).

<i>VD</i>	TR	$r_p$	$r_a$	$r$	$DT_{\theta_{RL}-\theta_{RB}}$	$z_{\theta_{RL}-\theta_{RB}}$	$p_{\theta_{RL}\neq\theta_{RB}}$
DES	N2	0,9835	0,9809	0,88	0,01	1,655	0,049
DES	N5	0,9213	0,9442	0,88	0,01	0,999	0,159
DES	N10	0,8788	0,9055	0,86	0,01	1,638	0,051
DES	P2	0,9835	0,9809	0,88	0,01	1,655	0,049
DES	P5	0,9353	0,9388	0,88	0,01	1,108	0,134
DES	P10	0,8773	0,8941	0,86	0,01	1,751	0,040
COG	N2	0,9779	0,9787	0,88	0,01	0,992	0,161
COG	N5	0,9673	0,9578	0,88	0,01	2,211	0,014
COG	N10	0,9443	0,9309	0,88	0,01	1,606	0,054
COG	P2	0,9779	0,9787	0,88	0,01	0,992	0,161
COG	P5	0,9596	0,9582	0,88	0,01	0,337	0,368
COG	P10	0,9558	0,9514	0,88	0,01	0,737	0,230
CON	N2	0,9780	0,9853	0,88	0,02	5,190	*
CON	N5	0,9343	0,9172	0,88	0,02	2,480	0,007
CON	N10	0,8859	0,8643	0,84	0,02	0,604	0,273
CON	P2	0,9780	0,9853	0,88	0,02	5,401	*
CON	P5	0,9389	0,9416	0,88	0,02	1,087	0,138
CON	P10	0,8715	0,8462	0,82	0,02	2,602	0,005

*DT*: desviación típica,  $r_p$ : correlación entre las respuestas de las técnicas para los casos con respuesta positiva,  $r_a$ : correlación entre las respuestas de las técnicas para los casos con respuesta negativa,  $r$ : correlación entre las áreas bajo las curvas. N2, N5 y N10: recodificación de las puntuaciones típicas normalizadas a dos, cinco y diez categorías respectivamente. P2, P5 y P10: recodificación utilizando la los percentiles en dos, cinco y diez categorías respectivamente. \*:  $p < 0,001$ .

**Tabla 9.4:** Tamaño de las diferencias entre áreas bajo la curva ROC en función del tipo de recategorización en las variables independientes y de la variable dependiente (VD).

		VD		
		DES	COG	CON
Tipo de recodificación	N2	1,47	0,78	7,82
	N5	0,89	1,79	3,99
	N10	1,56	1,27	1,12
	P2	1,47	0,78	8,13
	P5	0,96	0,27	1,70
	P10	1,63	0,58	5,02
	<i>M</i>	1,33	0,91	4,63
<i>DT</i>	0,32	0,54	2,97	

Expresado en porcentaje y valor absoluto.

favor de la red bayesiana se observa en el modelo que utiliza la dicotomización basada en la mediana, donde la regresión logística no genera una curva estadísticamente diferente de los niveles de azar. La dimensión emocional de la actitud emprendedora genera diferencias entorno al 1,63% y el 0,89% que suponen una ventaja media global de la regresión logística sobre la red bayesiana del 1,33%. Por último, el componente cognitivo de la actitud hacia la creación de empresas es predicho de manera similar por ambas técnicas donde las diferencias entre las curvas generadas por la regresión logística y la red bayesiana difieren globalmente en menos de un 1%.

## 9.4. Discusión parcial

El hallazgo más significativo de este estudio, en consonancia con las hipótesis iniciales y con un estudio anterior donde se utilizaron modelos de ecuaciones estructurales (J. García et al., 2007), ha sido que no existen diferencias estadísticamente significativas en términos globales entre las redes bayesianas y la regresión logística cuando se categorizan las variables independientes de los modelos para predecir los componentes de la actitud emprendedora. Este resultado combinado con el del estudio anterior puede servir en el contexto aplicado ante la toma de

#### Estudio 4. Tipo de categorización y área bajo la curva ROC

decisiones para la construcción de modelos. Mientras que la regresión logística parece funcionar relativamente mejor cuando la recodificación se hace en términos absolutos, la red bayesiana se iguala a esta primera cuando se utilizan recodificaciones relativas basadas en la tipificación de las variables o en los cuantiles. En cualquier caso, la elección de los puntos de corte ha de hacerse con todas las garantías posibles que indiquen que se ha alterado lo menos posible la naturaleza de los datos (Altman et al., 1994).

No obstante, se ha observado que las redes bayesianas generan áreas bajo la curva ROC estadísticamente más grandes que las generadas por la regresión logística cuando se pretende predecir el componente conductual de la actitud hacia la creación de empresas. Por el contrario, la regresión logística genera áreas bajo la curva ROC mayores cuando se está intentando predecir el componente emocional de la actitud. Como se vio en el estudio anterior el patrón de resultados es inverso cuando se recodifican las variables atendiendo a intervalos arbitrarios de la misma anchura; esto es, la red bayesiana funciona mejor cuando hay una alta prevalencia mientras que la regresión logística lo hace mejor cuando la tasa de prevalencia es baja. Dado que la diferencia entre una variable dependiente y otra estiva en la tasa de respuestas positivas, esta disociación podría estar indicando una mayor vulnerabilidad predictiva de la regresión logística, como sucede en relación a otros parámetros (p. e., Hsieh et al., 1998; Whittermore, 1981), cuando la recodificación se basa en puntos de corte relativos y cuando la tasa de respuestas positiva es muy baja.

En cualquier caso, sería conveniente profundizar en este asunto utilizando datos reales o simulados donde la tasa de respuestas positiva en la variable dependiente varíe sensiblemente. Aspectos como la naturaleza distribucional de las variables predictoras también podrían ser parámetros a tener en cuenta en la evaluación de la bondad predictiva de una técnica y otra tras la recodificación (Ragland, 1992).

# Capítulo 10

## Estudio 5. Porcentaje de casos perdidos y área bajo la curva ROC

*"[...] even the details of the conversation were less clear. They were like the missing letters on a Roman tomb and I the archaeologist filling in the gaps according to the bias of my scholarship".*

(Greene, 1955/2004)

### 10.1. Descripción

Es sabido que la regresión logística presenta problemas prácticos cuando es sometida a la estimación de parámetros a partir de bases de datos con casos perdidos (p. e., Eftekhar et al., 2005; Ng y Jordan, 2002). Dado que este tema no ha sido muy estudiado en el contexto de la comparación entre la regresión logística y las redes bayesianas en términos de validez predictiva (p. e., Greiner et al., 2005; Greiner y Zhou, 2002; Shen et al., 2003), el objetivo de este estudio será estudiar la relación entre el porcentaje de datos perdidos en una muestra y el área generada bajo la curva ROC por la regresión logística y las redes bayesianas. En este estudio se espera observar un buen comportamiento de las redes bayesianas ante

## Estudio 5. Porcentaje de casos perdidos y área bajo la curva ROC

la presencia de datos perdidos, mientras que se espera que la regresión logística sea más vulnerable en términos predictivos en esta situación.

## 10.2. Método

### 10.2.1. Muestra

De la muestra general descrita en la sección 5.3.1 se eliminaron todos los casos que contenían algún valor perdido en cualquiera de las variables implicadas en los modelos considerados. En total se eliminaron un 10,24% (126) de los casos de la base de datos original y la muestra se quedó en 1114 casos.

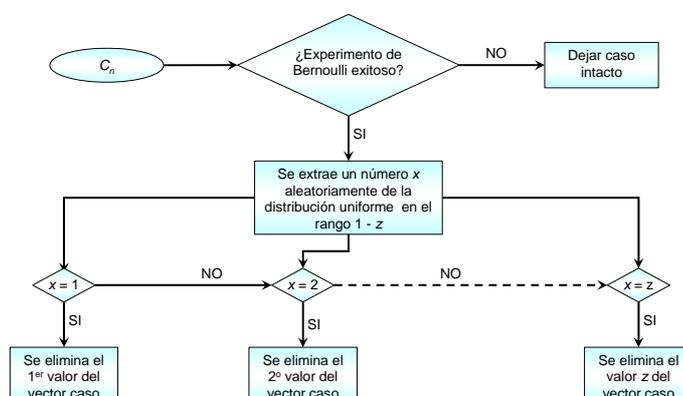
En esta nueva sub-muestra sin datos perdidos el 34,9% (389) de los participantes fueron hombres mientras que el 64,7% (721) fueron mujeres. Por su parte, en un 0,4% (4) de la muestra los participantes no indicaron su sexo. Las edades de los participantes estuvieron comprendidas entre los 17 y los 56 años, donde la media fue de 22,35 y la desviación típica de 4,25. Por otro lado, el 24,3% (271) fueron estudiantes de las carreras de Ciencias Empresariales, el 47,2% (526) fueron estudiantes de Ciencias Humanas y Jurídicas, mientras que el 28,1% (313) fueron estudiantes de Titulaciones Técnicas.

### 10.2.2. Análisis de datos

Para simular un patrón de datos perdidos completamente aleatorio (Hair et al., 1998) se llevaron a cabo, en primer lugar, experimentos Bernoulli para cada caso con probabilidades de éxito de 0,05, 0,10, 0,15, 0,20, 0,25, 0,50, y 0,70 para cada dimensión de la actitud hacia la creación de empresas. Para cada vector caso ( $c_n$ ) del archivo original

$$c_n = [vd, vi_1, vi_2, \dots, vi_p]$$

y para cada una de las variables dependientes o predichas ( $vd$ ), donde  $vi_p$  se refiere al conjunto de las variables predictoras o independientes, se realizaron experimentos de Bernoulli con las probabilidades de éxito indicadas anteriormente. Cuando el experimento resultaba exitoso se etiquetaba el caso con el valor 1. A continuación, para los casos que resultaron exitosos en el experimento Bernoulli se



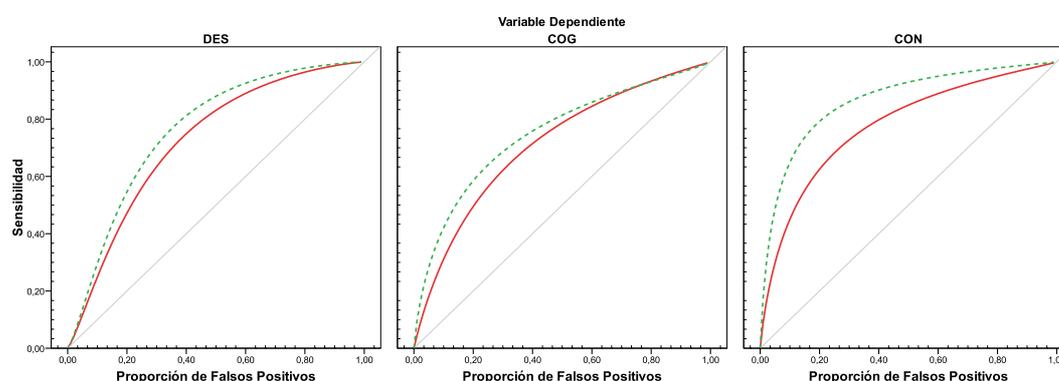
**Figura 10.1:** Esquema del proceso de simulación de caos perdidos.

utilizó la distribución uniforme aleatoria comprendida entre uno y  $z$  (indicando  $z$  el número de variables implicadas en el modelo) para eliminar un valor del vector caso en la base de datos. Todas las simulaciones de casos perdidos se realizaron con una macro para Microsoft Excel 2003 utilizando el módulo de Análisis de Datos. En la Figura 10.1 aparece representado gráficamente el proceso de simulación de casos perdidos.

En total, se generaron 24 bases de datos con patrones aleatorios de datos perdidos caracterizados por diferentes porcentajes de casos perdidos (0 %, 5 %, 10 %, 15 %, 20 %, 25 %, 50 % y 70 %); de las cuales un tercio fueron destinadas a estimar los modelos del componente emocional, otro tercio a la estimación de los modelos del componente cognitivo y el resto se destinó a la estimación de los modelos del componente conductual de la actitud emprendedora.

Los modelos de regresión logística se estimaron con el paquete estadístico SPSS 15.0 sin incluir términos de interacción, se forzó al programa a introducir en los modelos de regresión las variables definidas previamente por el diseño del estudio y el nivel de confianza de las estimaciones se ajustó al 95 %. La estimación paramétrica de las redes bayesianas se hizo con el método de máxima verosimilitud usando la corrección de Laplace, mientras que las estructuras cualitativas se construyeron manualmente con base al diseño del estudio. Todos los parámetros y las probabilidades posteriores se calcularon con el programa Netica 4.02. Las estimaciones del tamaño del área bajo la curva ROC (Hanley y McNeil, 1982), su desviación típica, su estimación suavizada, los valores de los estadísticos de contraste, las probabilidades para los contrastes de hipótesis sobre la diferencia

## Estudio 5. Porcentaje de casos perdidos y área bajo la curva ROC



**Figura 10.2:** Curvas ROC globales en función de la variable dependiente. La línea continua representa a la regresión logística y la discontinua a la red bayesiana.

entre curvas (Hanley y McNeil, 1983) así como los intervalos de confianza fueron estimados con una macro para Microsoft Excel 2003 diseñada específicamente para este trabajo que incorporaba las salidas de SPSS y de Netica. Las coordenadas de las curvas ROC suavizadas también fueron estimadas utilizando esta macro.

Una descripción más detallada de los análisis de datos puede encontrarse en la sub-sección 7.2.2.

### 10.3. Resultados

Como se puede observar en la Figura 10.2 las redes bayesianas funcionan globalmente mejor ante la presencia de datos perdidos al predecir los tres componentes de la actitud emprendedora en comparación con la regresión logística. La mayor diferencia entre áreas bajo la curva ROC se observa en el componente emocional ( $z = 2,93, p = 0,002$ ) donde la regresión logística obtiene un área de 0,6993 ( $z = 9,47, p < 0,001$ ) mientras que en el caso de la red bayesiana es de 0,7468 ( $z = 13,35, p < 0,001$ ).

En el modelo de la dimensión cognitiva de la actitud hacia la creación de empresas la diferencia entre curvas es del 4,64% ( $z = 5,18, p < 0,001$ ) y la regresión logística obtiene un área bajo la curva ROC de 0,6881 ( $z = 11,44, p < 0,001$ ) mientras que la red bayesiana genera una curva con un área de 0,7351 ( $z = 15,30, p < 0,001$ ). Por último, aunque la red bayesiana tiene un área global bajo la curva mayor (0,6597,  $z = 4,68, p < 0,001$ ) que la regresión logística (0,6440,  $z = 4,31, p < 0,001$ ) esta diferencia no es estadísticamente significativa

( $z = 0,59, p = 0,28$ ).

En la Tabla 10.1 aparecen las áreas bajo la curva ROC estimadas para cada dimensión de la actitud hacia la creación de empresas en función del porcentaje de casos perdidos en la base de datos y en la Figura 10.3 aparece la representación gráfica de las curvas suavizadas. Como aparece en la Tabla 10.1, las áreas bajo la curva ROC generadas por ambas técnicas en todas las condiciones de casos perdidos son estadísticamente diferentes de 0,5. No obstante, se pueden hacer algunas apreciaciones en relación a la diferencia entre técnicas.

En primer lugar, en consonancia con lo anteriormente expuesto, aunque en todos los casos el área bajo la curva generada por la regresión logística es inferior a la generada por la red bayesiana, no se observan diferencias estadísticamente significativas entre las curvas generadas por una y otra técnica en el modelo diseñado para predecir el componente conductual de la actitud hacia la creación de empresas (Tabla 10.2). En el resto de los casos si que se observan diferencias estadísticamente significativas entre las curvas generadas por la regresión logística y la red bayesiana excepto en los casos en que se trabaja con el modelo del componente emocional de la actitud al 20 y al 70 % de casos perdidos, donde las diferencias pueden considerarse marginales.

Por otro lado, se observa que cuando se está trabajando con el modelo del componente cognitivo de la actitud emprendedora, donde el porcentaje de respuestas positivas y negativas está más balanceado, el área bajo la curva ROC estimada para la regresión logística es mayor cuando hay un porcentaje moderado de casos perdidos (entre el 5 y el 15 %). Cuando el porcentaje de casos perdidos sobrepasa el 15 % el área bajo la curva ROC se va reduciendo ligeramente. Las redes bayesianas parecen verse menos afectadas en su habilidad predictiva por la presencia de casos perdidos y en todas las condiciones simuladas de casos perdidos producen áreas bajo la curva ROC en torno al 0,73.

Por último, no se aprecia ningún patrón destacado en la comparación de las curvas ROC generadas por el modelo del componente emocional de la actitud emprendedora aunque en todos los casos el área bajo la curva ROC generada por la red bayesiana es mayor a la generada por la regresión logística. Como se puede ver en la Tabla 10.3 los mayores tamaños de la diferencia entre las curvas se observan en la dimensión emocional donde se supera el 5 % en tres ocasiones, mientras que en el modelo del componente cognitivo las variaciones son más

Estudio 5. Porcentaje de casos perdidos y área bajo la curva ROC

**Tabla 10.1:** Áreas bajo la curva ROC en función del porcentaje de casos perdidos (CP), de la variable dependiente (VD) y del tipo de técnica.

Técnica	VD	CP	$\theta$	$DT_{\theta}$	$\theta^*$	$z_{\theta=0}$	$p_{\theta=0}$	$IC_{inf}$	$IC_{sup}$
RL	DES	0	0,7016	0,0182	0,7099	11,05	*	0,6659	0,7374
	DES	5	0,6913	0,0190	0,7005	10,08	*	0,6541	0,7285
	DES	10	0,7081	0,0192	0,7198	10,83	*	0,6705	0,7458
	DES	15	0,6968	0,0197	0,7083	10,00	*	0,6582	0,7353
	DES	20	0,7025	0,0204	0,7142	9,93	*	0,6625	0,7425
	DES	25	0,7034	0,0206	0,7183	9,88	*	0,6630	0,7437
	DES	50	0,6798	0,0274	0,7114	6,57	*	0,6262	0,7335
	DES	70	0,7025	0,0336	0,7272	6,02	*	0,6365	0,7684
	COG	0	0,6881	0,0165	0,7010	11,44	*	0,6559	0,7204
	COG	5	0,6915	0,0167	0,7013	11,49	*	0,6588	0,7241
	COG	10	0,7021	0,0171	0,7095	11,85	*	0,6686	0,7355
	COG	15	0,7056	0,0174	0,7098	11,83	*	0,6715	0,7396
	COG	20	0,6856	0,0186	0,7031	9,97	*	0,6491	0,7221
	COG	25	0,6811	0,0194	0,6877	9,35	*	0,6432	0,7191
	COG	50	0,6708	0,0241	0,6829	7,08	*	0,6235	0,7181
	COG	70	0,6812	0,0292	0,6900	6,20	*	0,6239	0,7384
	CON	0	0,6440	0,0334	0,7497	4,31	*	0,5784	0,7095
	CON	5	0,6386	0,0340	0,7404	4,08	*	0,5720	0,7051
	CON	10	0,6485	0,0353	0,7473	4,20	*	0,5792	0,7177
	CON	15	0,6517	0,0369	0,7786	4,11	*	0,5794	0,7241
CON	20	0,6655	0,0367	0,7783	4,51	*	0,5936	0,7374	
CON	25	0,6248	0,0376	0,7297	3,32	*	0,5511	0,6986	
CON	50	0,6817	0,0446	0,8048	4,07	*	0,5943	0,7692	
CON	70	0,6604	0,0498	0,7225	3,22	0,001	0,5627	0,7581	
RB	DES	0	0,7592	0,0178	0,7756	14,56	*	0,7243	0,7940
	DES	5	0,7387	0,0184	0,7413	12,94	*	0,7025	0,7748
	DES	10	0,7618	0,0179	0,7742	14,64	*	0,7268	0,7969
	DES	15	0,7477	0,0183	0,7345	13,51	*	0,7118	0,7836
	DES	20	0,7245	0,0189	0,7264	11,89	*	0,6875	0,7615
	DES	25	0,7550	0,0180	0,7590	14,17	*	0,7197	0,7902
	DES	50	0,7460	0,0191	0,7377	12,89	*	0,7086	0,7834
	DES	70	0,7406	0,0196	0,7407	12,29	*	0,7022	0,7790
	COG	0	0,7351	0,0154	0,7419	15,30	*	0,7050	0,7652
	COG	5	0,7307	0,0157	0,7359	14,73	*	0,7000	0,7614
	COG	10	0,7316	0,0158	0,7355	14,66	*	0,7007	0,7626
	COG	15	0,7370	0,0157	0,7441	15,12	*	0,7062	0,7677
	COG	20	0,7355	0,0159	0,7411	14,77	*	0,7042	0,7667
	COG	25	0,7259	0,0162	0,7327	13,97	*	0,6943	0,7576
	COG	50	0,7212	0,0168	0,7334	13,15	*	0,6882	0,7541
	COG	70	0,7301	0,0170	0,7402	13,51	*	0,6967	0,7635
	CON	0	0,6597	0,0341	0,8419	4,68	*	0,5929	0,7266
	CON	5	0,6634	0,0340	0,8844	4,80	*	0,5967	0,7302
	CON	10	0,6711	0,0346	0,8116	4,95	*	0,6033	0,7389
	CON	15	0,6611	0,0356	0,8610	4,53	*	0,5914	0,7309
CON	20	0,6624	0,0349	0,8717	4,65	*	0,5940	0,7309	
CON	25	0,6365	0,0348	0,8328	3,92	*	0,5682	0,7048	
CON	50	0,6387	0,0363	0,8391	3,82	*	0,5675	0,7098	
CON	70	0,6462	0,0369	0,8622	3,96	*	0,5739	0,7185	

RL: regresión logística, RB: red bayesiana, DT: desviación típica, IC: intervalo de confianza al 95%, \*:  $p < 0,001$ .

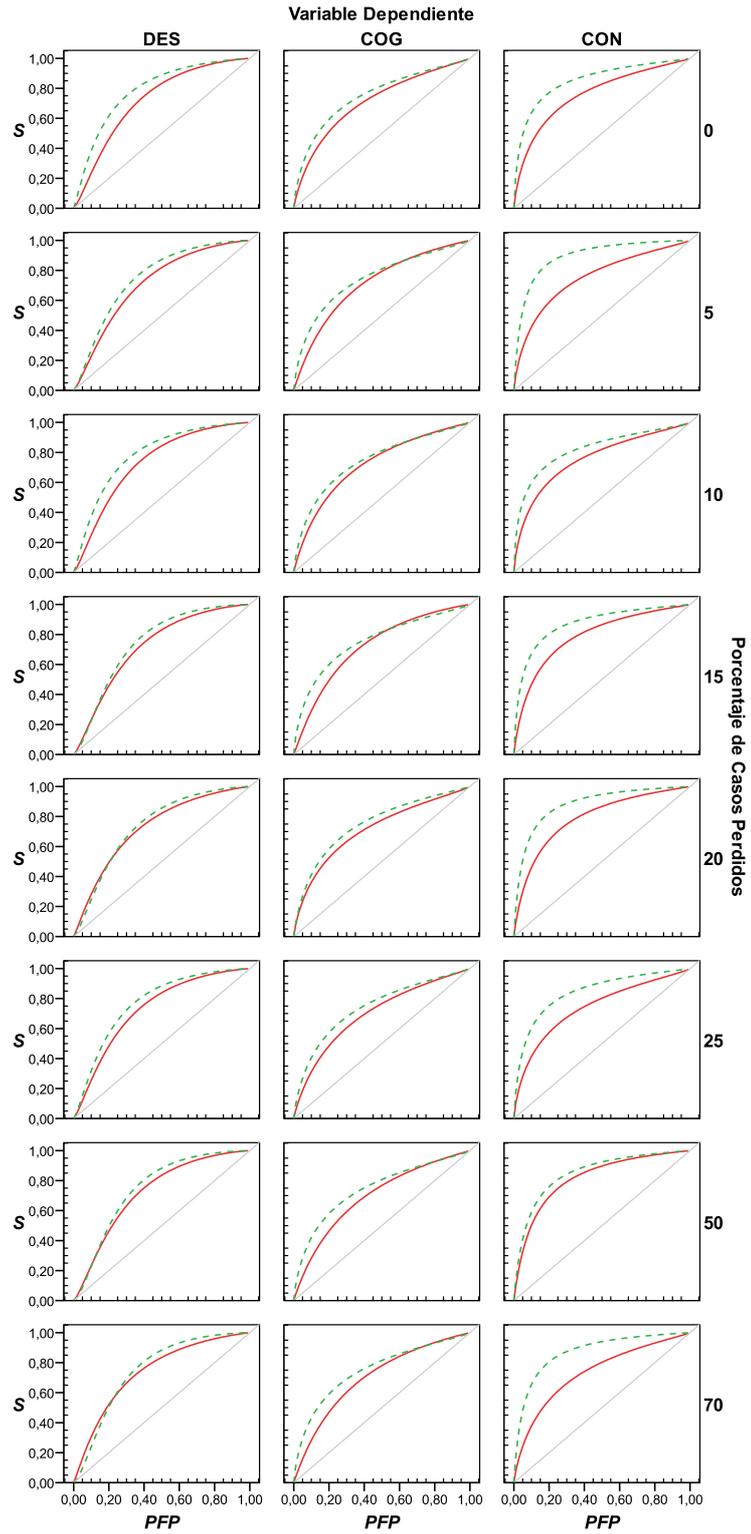


Figura 10.3: Curvas ROC en función del porcentaje de casos perdidos. La línea continua representa a la regresión logística y la discontinua a la red bayesiana.

Estudio 5. Porcentaje de casos perdidos y área bajo la curva ROC

**Tabla 10.2:** Comparación de áreas bajo la curva ROC en función del porcentaje de casos perdidos (CP), de la variable dependiente (VD) y del tipo de técnica.

VD	CP	$r_p$	$r_a$	$r$	$DT_{\theta_{RL}-\theta_{RB}}$	$z_{\theta_{RL}-\theta_{RB}}$	$p_{\theta_{RL}\neq\theta_{RB}}$
DES	0	0,7017	0,7661	0,69	0,01	4,056	*
DES	5	0,6521	0,7538	0,67	0,02	3,115	0,001
DES	10	0,7034	0,7742	0,69	0,01	3,664	*
DES	15	0,6425	0,7555	0,65	0,02	3,193	0,001
DES	20	0,6724	0,7807	0,69	0,02	1,413	0,079
DES	25	0,6939	0,7394	0,66	0,02	3,210	0,001
DES	50	0,6670	0,6888	0,63	0,02	3,098	0,001
DES	70	0,6255	0,7671	0,65	0,03	1,488	0,068
COG	0	0,8745	0,8517	0,84	0,01	5,182	*
COG	5	0,8818	0,8453	0,84	0,01	4,266	*
COG	10	0,8812	0,8590	0,84	0,01	3,154	0,001
COG	15	0,8608	0,8338	0,82	0,01	3,127	0,001
COG	20	0,8478	0,8206	0,79	0,01	4,345	*
COG	25	0,8883	0,8472	0,84	0,01	4,265	*
COG	50	0,8413	0,8340	0,79	0,01	3,368	*
COG	70	0,8316	0,7917	0,77	0,02	2,519	0,006
CON	0	0,7727	0,6850	0,69	0,03	0,594	0,276
CON	5	0,7942	0,6405	0,67	0,03	0,900	0,184
CON	10	0,7796	0,6743	0,69	0,03	0,823	0,205
CON	15	0,7490	0,6760	0,67	0,03	0,318	0,375
CON	20	0,7424	0,6908	0,67	0,03	0,106	0,458
CON	25	0,8376	0,6391	0,69	0,03	0,407	0,342
CON	50	0,6482	0,6652	0,61	0,04	1,179	0,119
CON	70	0,6861	0,6990	0,65	0,04	0,372	0,355

RL: regresión logística, RB: red bayesiana, DT: desviación típica,  $r_p$ : correlación entre las respuestas de las técnicas para los casos con respuesta positiva,  $r_a$ : correlación entre las respuestas de las técnicas para los casos con respuesta negativa,  $r$ : correlación entre las áreas bajo las curvas, \*:  $p < 0,001$ .

**Tabla 10.3:** Tamaño de las diferencias entre áreas bajo la curva ROC en función del porcentaje de casos perdidos y la variable dependiente.

		VD		
		DES	COG	CON
Porcentaje de casos perdidos	0	5,76	4,69	1,58
	5	4,74	3,92	2,49
	10	5,37	2,96	2,27
	15	5,09	3,14	0,94
	20	2,19	4,99	0,31
	25	5,16	4,48	1,17
	50	6,61	5,04	4,31
	70	3,82	4,89	1,42
	<i>M</i>	4,84	4,17	1,46
	<i>DT</i>	1,34	0,83	1,23

Expresado en porcentaje y valor absoluto.

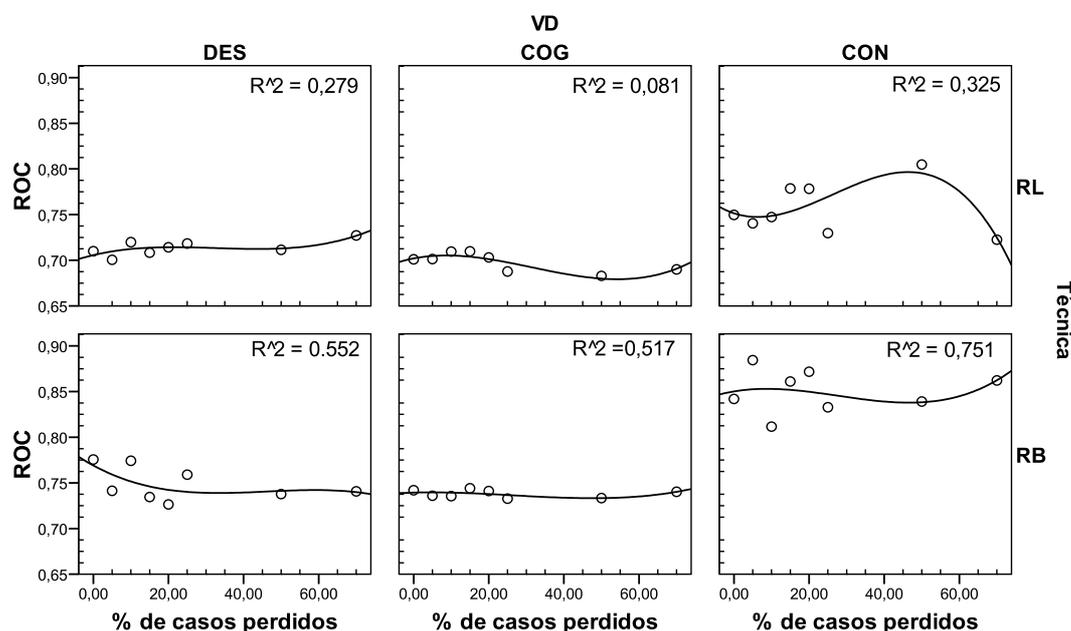
estables.

Para terminar, en la Figura 10.4 se han representado el tamaño de las áreas suavizadas bajo las curvas ROC en función de la variable dependiente, de la técnica y del porcentaje de casos perdidos para ver como evolucionan en cada caso. El aspecto más llamativo es que las redes bayesianas obtienen áreas más grandes en todos los casos. El caso más significativo se encuentra en el modelo conductual donde la ejecución de la regresión logística se reduce drásticamente cuando el porcentaje de casos perdidos pasa del 50% mientras que el de la red bayesiana mejora ligeramente.

## 10.4. Discusión parcial

El resultado más destacable es este estudio es, en consonancia con nuestras hipótesis iniciales, que la técnica de la regresión logística es más vulnerable ante la presencia de casos perdidos cuando la proporción de respuestas positivas y negativas a la variable dependiente tiende a estar equilibrada o está sesgada al valor positivo (componentes cognitivo y emocional de la actitud emprendedora respectivamente). Sin embargo, cuando la tasa de respuestas negativas es muy elevada

Estudio 5. Porcentaje de casos perdidos y área bajo la curva ROC



**Figura 10.4:** Evolución del área ROC en función del porcentaje de casos perdidos, de la técnica (RL: regresión logística, RB: red bayesiana) y de la variable dependiente (VD). La  $R^2$  ha sido ajustada para un modelo de regresión cúbica sólo con fines descriptivos.

en comparación con las positivas (caso del componente conductual de la actitud hacia la creación de empresas) las diferencias entre técnicas desaparecen aunque la red bayesiana sigue siendo ligeramente superior.

Estos resultados son consistentes con anteriores trabajos que destacaban la idoneidad de las redes bayesianas para trabajar en condiciones caracterizadas por la presencia de datos perdidos (p. e., Heckerman, 1995; Nadkarni y Shenoy, 2004; Jansen et al., 2003). Resultados similares se han encontrado en otros estudios donde se ha comparado la regresión logística con herramientas estadísticas basadas en una filosofía de red similar a la utilizada por las redes bayesianas. Por ejemplo, Eftekhar et al. (2005) observaron que las Redes Neuronales Artificiales podían ser una alternativa plausible en términos predictivos a la regresión logística cuando los datos a analizar contenían un volumen considerable de datos perdidos.

Sin embargo, es destacable el hecho de que la regresión logística se comporte de igual manera ante la presencia de datos perdidos cuando la tasa de respuestas negativas es muy elevada en la variable dependiente. Futuros desarrollos que

conjuguen las propiedades estructurales de la regresión logística en términos de modelos gráficos podrían aportar técnicas estadísticas más fiables y válidas en estas situaciones (p. e., Greiner et al., 2005; Greiner y Zhou, 2002).

## Estudio 5. Porcentaje de casos perdidos y área bajo la curva ROC

**Parte III**

**Conclusiones**



# Capítulo 11

## Discusión general

*"We are just a moment in time,  
a blink of an eye,  
a dream for the blind,  
visions from a dying brain [...]"*

(Anathema, 1998)

En este capítulo se hará una discusión general de la tesis evaluando los resultados obtenidos así como las implicaciones que conllevan en el contexto de la investigación. En primer lugar se tratará el aspecto metodológico de la tesis y se evaluarán los resultados obtenidos en términos diferenciales respecto a los niveles de validez predictiva alcanzados por la regresión logística y las redes bayesianas bajo las condiciones analizadas anteriormente. Seguidamente se dedicará un apartado a tratar las implicaciones que este trabajo tiene en el campo de estudio de las actitudes emprendedoras.

### 11.1. Redes bayesianas *versus* regresión logística

En términos generales se ha observado que tanto la manipulación del tamaño de la muestra como del porcentaje de casos perdidos no afecta igualmente a la regresión

## Discusión general

logística y a las redes bayesianas, y que en ambos casos las redes bayesianas parecen ser más resistentes en términos predictivos ante la manipulación de estos parámetros. No obstante, se ha mostrado que en ambos casos la distribución no balanceada de la variable predicha tiene efectos sobre el área bajo la curva ROC.

Como se indicó en la sección teórica, el tamaño de la muestra es uno de los aspectos que más preocupan en la aplicación práctica de la regresión logística (p. e., Jovel, 1995; Long, 1997; Silva y Barroso, 2004). Por ello, se han desarrollado métodos como la regresión logística exacta (p. e., Hirji et al., 1987; Mehta y Patel, 1995) o las funciones de penalización del error asintótico (p. e., Bull et al., 2002; Firth, 1993) para tratar de minimizar el efecto que tienen las muestras pequeñas sobre la estimación de la regresión logística. En este sentido, el uso de las redes bayesianas en condiciones de tamaños muestrales reducidos puede considerarse como otra alternativa a las ya disponibles para estas situaciones dado que son capaces de tratar eficientemente los datos en estas situaciones sin poner en peligro la validez predictiva de los modelos.

El grado en que una muestra presenta datos perdidos es otra debilidad para el procedimiento de estimación paramétrica de la regresión logística ya que pueden producirse problemas numéricos que sesguen los modelos (Irala et al., 1997). En este estudio se ha observado que el porcentaje de casos perdidos es un parámetro que tiene efectos más drásticos sobre la validez predictiva de la regresión logística en comparación con las redes bayesianas. No obstante, esta ventaja de las redes bayesianas para tratar eficientemente con datos perdidos desaparece cuando la tasa de respuestas positivas en la variable criterio es muy baja. La idoneidad de trabajar con redes bayesianas cuando las muestras presentan datos perdidos ya era conocida (p. e., Heckerman, 1995) pero, como se ha puesto de manifiesto en este estudio, es necesario llevar a cabo más estudios comparativos con diferentes técnicas para determinar cuales son las condiciones óptimas de aplicación de una u otra técnica. Un avance prometedor es el que implica el desarrollo de procedimientos de estimación paramétrica híbridos a partir de las propiedades estructurales de las redes bayesianas y de la regresión logística (Greiner et al., 2005; Greiner y Zhou, 2002; Ross, Wettig, Grünwald, Myllymäki, y Tirri, 2005).

En el único caso donde se han encontrado problemas para establecer claramente la superioridad de una técnica sobre otra ha sido al manipular el nivel de categorización de las variables dependientes. Pese a que uno de los parámetros

que influyen negativamente en el análisis de regresión logística es la dicotomización de la variable resultado (p. e., Altman et al., 1994; Ragland, 1992; Reade y Kupper, 1995), se han observado áreas bajo la curva ROC mayores para la regresión logística en términos generales y específicos cuando se categorizaron las variables independientes utilizando recodificaciones arbitrarias con intervalos de igual amplitud. No obstante, estas diferencias, aunque estadísticamente significativas, no son muy grandes y podrían deberse a la influencia de otros factores. Dado que el método de categorización lo consideramos determinante para obtener estos resultados, se realizó un segundo estudio sobre este mismo tópico. Estas diferencias iniciales al recodificar las variables predictoras desaparecieron cuando se utilizaron recodificaciones relativas basadas en las puntuaciones típicas normalizadas y en los cuantiles. No obstante se observó una disociación influenciada por la tasa de respuestas positivas en la variable dependiente. Así, cuando la tasa de respuestas positivas es muy elevada la red bayesiana funcionó mejor al ser recodificada en términos absolutos; mientras que cuando la tasa de prevalencia fue menor funcionó mejor la recodificación relativa para la red bayesiana. El patrón inverso fue observado para la regresión logística. Futuras investigaciones en este tópico permitirán conocer más específicamente cuales son las condiciones bajo las cuales ambas técnicas se desenvuelven mejor. Por ejemplo, un aspecto que podría estudiarse en más detenimiento sería el efecto de la clasificación no diferencial y diferencial sobre el área bajo la curva ROC en ambas técnicas (p. e., Flegal et al., 1991; Fung y Howe, 1984).

Aparte de los resultados obtenidos en los estudios presentados, es conveniente hacer algunas apreciaciones sobre las debilidades metodológicas de este trabajo. En primer lugar, los modelos trabajados podrían criticarse por adolecer de una buena especificación ya que en algunos casos, principalmente en algunos modelos de regresión logística, los parámetros del modelo no alcanzaban la significación estadística deseable. Sin embargo, la selección llevada a cabo de los modelos queda justificada por varias razones. Primeramente, el principal criterio, aunque no el único, de selección de los modelos fue la validez predictiva; esto es, se seleccionaron modelos que fueron adecuados a la hora de predecir correctamente las variables criterio (Greiner et al., 2005). Esta perspectiva pragmática tiene la ventaja de igualar a ambas técnicas independientemente de los supuestos que cada una de ellas requieren, además de enfrentarlas en la misma tarea de clasificación.

## Discusión general

En segundo lugar, dada la naturaleza del estudio basada en el análisis de casos naturales o reales, no se ha manipulado el balance de las respuestas en la variable dependiente de los modelos. Esto puede haber dado lugar a ciertos sesgos y sería recomendable que en el futuro se llevasen a cabo estudios de simulación donde este parámetro esté controlado. Por otro lado, la naturaleza distribucional de las variables dependientes puede ser otro factor determinante en la validez predictiva de ambas técnicas por lo que sería deseable que también se tuviese en cuenta en los mismos términos. Estos aspectos podrían haberse controlado realizando simulaciones de datos, aunque nuestro interés se centró en estudiar lo que tiende a pasar en la práctica donde suelen presentarse casos en que la distribución de las variables no sigue condiciones ideales.

Por último, el hecho de no haber seguido un procedimiento de validación cruzada puede atentar contra la generalización externa de los resultados (Harrel et al., 1985). No obstante, cabría esperar un menor sobreajuste y una mayor potencialidad de generalización externa en las redes bayesianas (Heckerman, 1995) como hemos observado en estudios anteriores (J. García et al., 2006, 2007; J. López, De la Fuente, et al., 2008). En cualquier caso, estos resultados pueden considerarse como un ejemplo óptimo de lo que sucedería en el mejor de los casos suponiendo que los datos de estimación y de validación se ajustasen bien al modelo.

Para terminar esta sección se señalarán algunos aspectos generales a tener en cuenta respecto a la idoneidad de utilizar las redes bayesianas frente a la regresión logística aparte de lo tratado en este estudio. Retomemos para ello la Tabla 5.1 donde podemos apreciar algunas de las ventajas y diferencias que aportan las redes bayesianas en comparación a la regresión logística.

Primeramente, las redes bayesianas pueden considerarse más apropiadas en el modelado estadístico porque permiten representar al unísono y «elegantemente» la dimensión cualitativa y cuantitativa de un problema, fenómeno o situación (Edwards, 1998; Heckerman, 1995; J. López et al., 2005, 2006, 2007). Esta operacionalización que hacen las redes bayesianas de las relaciones que se establecen entre las variables de un modelo en términos de dependencia e independiencia condicional con base a la teoría de grafos (Harary, 1969; Gould, 1988; Spirtes et al., 2000; Tutte, 1984; Xiang, 2002), podría considerarse como un avance en la metodología cualitativa aplicada en términos analíticos. Dado que la polémi-

## 11.1 Redes bayesianas *versus* regresión logística

ca que suscita en términos generales por la dicotomía cualitativo-cuantitativo es actual y de importancia práctica en la investigación de las ciencias sociales (G. King y Zeng, 2001a), podría ser interesante utilizar esta técnica estadística para superar estas vicisitudes.

Por otro lado, el hecho de que las redes bayesianas no tengan que estar limitadas a un modelo subyacente en los datos las hace más flexibles y potencialmente más aplicables a un amplio espectro de situaciones. Adicionalmente, la regresión logística podría no ser apropiada para modelar sistemas complejos con dinámicas no lineales ya que su modelo estadístico es una combinación lineal de variables. Por el contrario, las redes bayesianas serían más útiles en situaciones de este tipo ya que, análogamente a otras técnicas basadas en filosofía de red como las Redes Neurales Artificiales, permiten representar relaciones estadísticas no lineales y hacen más amigable la detección e interpretación de interacciones entre variables (S. M. Lee et al., 2005; Eftekhar et al., 2005).

Otra ventaja adicional es que las redes bayesianas representan todo el conocimiento en términos de probabilidades, lo que puede convertirse en una ventaja cuando se usan para realizar inferencias (Huete, 1998), frente a la utilización de *odds ratio* en la regresión logística que pueden ser difíciles de interpretar o no directamente obtenidas dependiendo del software estadístico utilizado (Irala et al., 1997).

Por último, S. M. Lee et al. (2005) opinan que la regresión logística es una técnica estadística más centrada, o limitada, por el proceso de verificación de hipótesis. Esto es, en un modelo de regresión logística las relaciones que se esperan encontrar entre las variables del modelo han de ser establecidas a priori y, como consecuencia, los predictores considerados como no importantes para el modelo no son incluidos. Esto puede tener consecuencias indeseables cuando se analizan grandes bases de datos con un gran número de variables ya que esta técnica es insensible a la detección de relaciones desconocidas o inesperadas entre las variables. Por su parte, las redes bayesianas pueden facilitar la detección de patrones y tendencias en los datos que no han sido hipotetizadas previamente.

## 11.2. Actitudes emprendedoras

El resultado que más podríamos destacar a este nivel ha sido que la dimensión conductual de la actitud hacia la creación de empresas es la más difícil de predecir correctamente a partir de las variables utilizadas en el estudio. En todos los casos, el área bajo la curva ROC generada tanto con regresión logística como con las redes bayesianas ha sido la menor comparada con la dimensión emocional y cognitiva. Este hallazgo es consistente con los hallazgos derivados del estudio de las actitudes en términos generales dado que no es fácil identificar la relación entre los componentes actitudinales y la conducta (p. e., Brehm et al., 2005; Deaux et al., 1993; Feldman, 1998; Franzoi, 2005; J. F. Morales et al., 1994). De este modo, aunque aquí se está haciendo referencia a la dificultad de predecir el componente conductual de la actitud emprendedora a partir de un conjunto de variables, se pone de manifiesto que la dimensión conductual de la actitud hacia la creación de empresas es difícil de predecir análogamente a como lo sería el comportamiento en sí.

Sin desligarme de lo anteriormente expuesto, y aprovechando la idea de los usos potenciales de las redes bayesianas en Psicología, una opción tentativa para mejorar la predicción conductual a partir de las actitudes podría ser el uso de las redes bayesianas como modelo estadístico subyacente. Por ejemplo, siguiendo la línea teórica de las actitudes que se ha usado en esta tesis, se podrían desarrollar modelos de red bayesiana para predecir el comportamiento sobre la base de la Teoría de la Acción Razonada o propuestas afines (p. e., Ajzen y Fishbein, 2005, 1980; Fishbein y Ajzen, 1975). En el caso concreto de las actitudes hacia la creación de empresas esta estrategia sería deseable tanto desde un punto de vista teórico como aplicado ya que, por un lado, ahondaría en el entendimiento del fenómeno emprendedor y, por otro, serviría para diseñar estrategias dirigidas a optimizar la gestión del potencial emprendedor en las personas.

En términos de las variables que se han apreciado relacionadas con la actitud hacia la creación de empresas, se ha observado que la puntuación en la escala de actitudes emprendedoras ACEMP ha sido un buen predictor de todas las dimensiones de la actitud hacia la creación de empresas. Así pues, este estudio podría verse como otra prueba añadida sobre la validez de la escala (Cano et al., 2003; J. García et al., 2005, 2006, 2007; J. López et al., 2009). Adicionalmente,

estudios posteriores tendrán que profundizar en las relaciones que se establecen entre la actitud hacia la creación de empresas y las dimensiones de la ACEMP. De este modo, dado que la crítica de Gartner (1988) hacia el modelo de los rasgos en el estudio de la creación de empresas ha sido rebatida recientemente (McKenzie et al., 2007), nuestra escala podría ser de utilidad en diferentes contextos, tanto teóricos como aplicados, para identificar y caracterizar el perfil de la persona que emprende la creación de una aventura empresarial.

Otra variable que hemos encontrado relacionada con todas las dimensiones de la actitud emprendedora ha sido la preparación. Además, se ha observado que, aunque en todos los casos la relación entre el nivel de preparación y la tendencia a crear empresas es positiva, la dimensión conductual es la que más relación tiene con la preparación percibida mientras que el aspecto emocional y el cognitivo están al mismo nivel. Este resultado es consistente con los trabajos anteriores que mostraban la importancia de la formación en el perfil de la persona emprendedora (p. e., Genescá y Capelleras, 2004; Genesca y Veciana, 1984; Gómez et al., 2007; HayGroup y SAP AG, 2003; Rogoff y Lee, 1996; Sánchez, 2003; Veciana, 1989).

Por consiguiente, dado que la formación se perfila como un elemento esencial en la caracterización de la persona emprendedora así como un rasgo determinante para predecir el éxito competitivo de los emprendedores, sería deseable llevar a cabo programas de intervención destinados a recomendar y crear itinerarios formativos de corte emprendedor en las personas potencialmente emprendedoras. En el ámbito universitario, por ejemplo, ya ha sido destacada la perentoriedad de propiciar un contexto favorable al desarrollo profesional e intelectual vinculado al desarrollo de nuevas empresas (p. e., Cano et al., 2003; Díaz, 2003; Peñas y Quijano, 2008).

Aparte de la puntuación en la escala ACEMP y del grado de preparación percibida, que son factores comunes a las tres dimensiones actitudinales, se ha observado<sup>1</sup> que la dimensión emocional de la creación de empresas está relacionada con un aspecto motivacional y con la cantidad de obstáculos percibidos por la persona. Así, las personas más motivadas a crear su propia empresa tendrán un mayor componente emocional de la actitud emprendedora, mientras que las personas que perciben más obstáculos tenderán a tener una reducida afinidad

---

<sup>1</sup>Todos los comentarios de este párrafo están referidos a los parámetros *beta* obtenidos en los modelos de regresión logística.

## Discusión general

emocional hacia la creación de empresas. Por su parte, el componente emocional está relacionado inversamente con el número de obstáculos percibidos ante la creación de empresas. Esto es, cuantos más obstáculos percibe una persona potencialmente emprendedora, más pequeña será su actitud hacia la creación de empresas en su dimensión conductual. Por último, se ha observado que las personas con más carencias percibidas son aquellas que mayor tendencia cognitiva tienen hacia la creación de empresas. Aunque este aspecto ha sido el que menos fiabilidad reporta, sería consistente con el estudio de Rogoff y Lee (1996) donde encontraron que las personas que más capacidades tenían para llevar a cabo la gestión de una empresa eran, de hecho, las que menos conocimientos tenían al respecto.

En términos globales, estos resultados pueden servir para caracterizar al emprendedor o, más concretamente, al emprendedor potencial dado que se ha trabajado con una muestra de estudiantes universitarios (p. e., Huefner et al., 1996; J. López et al., 2009). Esto tiene utilidad práctica porque los emprendedores potenciales son valiosos económicamente y por ello hay que protegerlos o, al menos, orientarles vocacionalmente (p. e., Valls, 1996). En este sentido, los resultados aquí presentados pueden utilizarse institucionalmente para clasificar e identificar a los emprendedores dentro de la universidad con el objetivo de optimizar sus probabilidades de éxito en sus empresas (Díaz, 2003) ya que, como ha señalado recientemente Morillas (2009) la potenciación de los emprendedores estimula los programas de desarrollo local y es un caldo de cultivo para el empleo.

## Parte IV

### Epílogo



# Capítulo 12

## Reflexiones finales

*"You and I are now in confrontation, but I see no violence."*

(Briley, 1989/2008)

A parte de lo expuesto en este trabajo, considero que las redes bayesianas pueden tener una aceptable utilidad en el quehacer de la disciplina psicológica. Como hemos señalado en diferentes lugares (J. López et al., 2005, 2006, 2007; J. López y García, 2007c), opino que la naturaleza bayesiana y probabilística de este tipo de redes puede favorecer el avance de la disciplina en áreas tan dispares como el conocimiento del funcionamiento neural y la gestión de recursos humanos. En lo que resta de este capítulo dedicaré algunas líneas a exponer someramente mi opinión al respecto. En términos generales la mayoría de las perspectivas potenciales de las redes bayesianas que aquí expongo están relacionadas con la percepción, el aprendizaje y razonamiento causal; bien sea desde un punto de vista básico, biológico o terapéutico. Es posible, aunque no estoy seguro, que mi interés por el área de la percepción, el aprendizaje y el razonamiento causal se puedan remontar a cuando cursé la asignatura *Aprendizaje y cognición* de la que hablé en el Prefacio. Sin embargo, lo que es cierto es que el estudio de la causalidad desde un punto de vista psicológico fue, es y seguirá siendo un asunto de interés profundo para Psicología (Notterman, 2004).

## Reflexiones finales

Desde un punto de vista básico, las redes bayesianas pueden aportar un doble beneficio al estudio empírico de la percepción, aprendizaje y razonamiento causal. Por un lado, dado que son herramientas estadísticas concebidas para modelar estructuras y procesos causales, pueden servir para plantear hipótesis y diseñar experimentos sobre estos aspectos psicológicos en humanos y animales. Por ejemplo, ya se ha destacado la habilidad de las redes bayesianas para modelar el aprendizaje causal en niños pequeños (Gopnik et al., 2004; Gopnik y Schulz, 2004) y en términos generales (Glymour, 2001, 2003); pero, aunque sería viable, no se han llevado a cabo trabajos orientados a modelar el aprendizaje causal en términos neurales basándose en las propiedades cuantitativas y cualitativas de las redes bayesianas. La introducción de la *hipótesis de la codificación bayesiana* es un claro ejemplo de la potencialidad que tendrían estas herramientas para modelar procesos sensoriales y de aquellos derivados de ellos (Knill y Pouget, 2004; Ma, Beck, Latham, y Pouget, 2006; Salinas, 2006).

En cuanto al razonamiento o inferencia causal, tanto Krynski y Tenenbaum (2007) como J. López y García (2009) han mostrado recientemente que el razonamiento causal humano es análogo al llevado a cabo por una red bayesiana. En cuanto a la percepción de la causalidad, considero que sería apropiado retomar los estudios clásicos de Michotte (1946) utilizando el paradigma de la *Teoría del Marco Relacional* centrándose en los conceptos de equivalencia estimular, transferencia de funciones y marcos temporales (p. e., Barnes, 1994; Barnes y Hampson, 1993; Barnes y Holmes, 1991; O'Hora, Barnes, Roche, y Smeets, 2004). Todo este volumen de conocimiento básico podría servir, al menos, para avanzar en la comprensión y tratamiento de ciertos trastornos psicológicos donde subyace un componente causal (Matute, 2002) como los del estado del ánimo o la esquizofrenia.

En el caso de los trastornos del estado del ánimo, una de las teorías más fuertes es la de la indefensión aprendida en la que subyace en un trasfondo atribucional causal (Overmier y LoLordo, 1998). Esto es, se entiende que ciertos trastornos del estado del ánimo que sufren algunas personas han sido provocados por un patrón atribucional no adaptativo donde, por ejemplo, atribuyen sus fracasos a ellas mismas mientras que los éxitos son atribuidos al azar o a factores externos. En este sentido, creo que sería interesante desarrollar, en primer lugar, modelos probabilísticos de la atribución causal frente a tareas formalizadas bajo el entorno

de la filosofía de las redes bayesianas. Estos modelos podrían servir para profundizar en el conocimiento de los procesos de aprendizaje percepción, aprendizaje e inferencia causal y, a su vez, podrían ser el germen del desarrollo de test de diagnóstico orientados al tratamiento cognitivo, conductual y emocional de este tipo de trastornos.

En lo que respecta a la esquizofrenia, pienso que los síntomas positivos que la caracterizan podrían tener también una interpretación causal además de la atencional como se defiende actualmente (p. e., Fuentes, 2001; Rodríguez-Arias, Aguilar, Espert, y Miñarro, 1996). Así, por ejemplo, el modelo de inhibición latente como modelo de los síntomas positivos de la esquizofrenia (Weiner, 2000) podría reinterpretarse en términos causales y ser estudiado con modelos estadísticos de redes bayesianas.

Por otro lado, hemos propuesto que el estudio de la percepción, aprendizaje y razonamiento causal podrían servir para desarrollar mejoras en las redes bayesianas (J. López et al., 2006). Así pues, el estudio y la comprensión de los procesos de percepción, aprendizaje y razonamiento causal podrían dar lugar al desarrollo de nuevos algoritmos de aprendizaje estructural y/o paramétrico implementables en los programas desarrollados para trabajar con redes bayesianas. Nosotros proponemos un proyecto de investigación parecido al que llevan a cabo el *Brain Mind Institute* (en la *Ecole Polytechnique de Lausanne*) y en *International Business Machines* (IBM) que va desde la observación del sistema natural hasta el modelado y a la inversa de modo iterativo (Markram, 2006). En este sentido, pienso que el aspecto estructural sería más delicado en términos de modelado porque posiblemente implicaría problemas *np-duros* siguiendo la nomenclatura de la Teoría de la Complejidad Computacional. Así mismo, creo que el aspecto paramétrico tendría que retomarse, al menos en su fase inicial, partiendo de los postulados de Pavlov (1927) y Rescorla y Wagner (1972) para adaptarlos a la realidad de las redes bayesianas. También creo que un aspecto crucial para los algoritmos futuros de aprendizaje en este contexto, independientemente de si se trabaja con la dimensión estructural o paramétrica de las redes, ha de ser la filosofía del *ensayo-por-ensayo*. Esto es, creo que los algoritmos de aprendizaje causal artificiales tendrían que funcionar de manera análoga a como hacemos las personas y los animales: caso-por-caso o ensayo-por-ensayo. Por ejemplo, como se ha comentado más arriba, el algoritmo PC implementado en TETRAD tiene

## Reflexiones finales

gran importancia para la Psicología porque es capaz de modelar, hasta cierto punto, el razonamiento en humanos. No obstante, este procedimiento de modelado no es plausible desde un punto de vista natural ya que sus estimaciones se realizan tomando una muestra de datos en su conjunto y no ensayo-por-ensayo o caso-por-caso que es como, presumiblemente, aprendemos los humanos.

En términos biológicos, uno de los campos donde la potencialidad de las redes bayesianas no ha sido explotada es en el estudio de los sustratos neuroquímicos de la incertidumbre. El uso de este tipo de herramientas estadísticas podría encuadrarse dentro de uno de los grandes flancos que constituyen a la Neurociencia Cognitiva y que ha venido a denominarse como *simulación computacional* (Kandel et al., 1995/1997; Parasuraman, 2000). En este contexto, un aspecto básico del comportamiento animal es la habilidad para decidir ante alternativas sobre las que tenemos información incierta y uno de los objetivos de la Neurociencia Cognitiva es la comprensión de los mecanismos subyacentes en la transformación de las señales sensoriales en una decisión (Ditterich, Mazurek, y Shadlen, 2003). Los trabajos de Shadlen y Newsome (1996) pueden considerarse como unos de los primeros orientados a investigar la relación que se establece entre los procesos sensoriales y los mecanismos que subyacen a la acción en términos de toma de decisiones. Sin embargo, el proceso neural que sustenta la toma de decisiones no está basado únicamente en lo sensorial (Ditterich et al., 2003; Lafuente y Romo, 2003), sino que implica procesos emocionales (Bechara, 2005; Hsu et al., 2005; Rustichini, 2005; Sugrue, Corrado, y Newsome, 2005) e implica tanto el trabajo de áreas prefrontales (p. e., Barraclough, Conroy, y Lee, 2004; Bechara, 2005; Hsu et al., 2005; Platt, 2004; Romo y Salinas, 2003) y otras áreas distribuidas por el cerebro (p. e., D. Lee, 2005; McCoy y Platt, 2005; Platt y Glimcher, 1999).

El estudio del sustrato neuroquímico de la incertidumbre ha comenzado a usar la estadística bayesiana y el uso de redes bayesianas podría servir para aumentar el conocimiento que tenemos de estos fenómenos (Cohen y Aston-Jones, 2005). Por ejemplo, Yu y Dayan (2005b) han propuesto un modelo basado en la estadística bayesiana capaz de modelar el comportamiento de nuestra atención ante diferentes tipos de incertidumbre en relación con dos neurotransmisores: la noradrenalina y la acetilcolina. A grandes rasgos, el modelo propone que existen dos sistemas de incertidumbre que interactúan entre sí: la *incertidumbre esperada*, mediada por la acetilcolina (ACh), y la *incertidumbre no esperada*, donde el

neurotransmisor implicado es la noradrenalina (NA) (Cohen y Aston-Jones, 2005; Yu y Dayan, 2003, 2005a, 2005b). La incertidumbre esperada se refiere al conocimiento relativo que tenemos sobre la ocurrencia de eventos en nuestro entorno natural; esto es, a la probabilidad que asociamos a la ocurrencia de los eventos ambientales. Por su parte, la incertidumbre no esperada se refiere a cambios bruscos, repentinos e inesperados que se producen sobre nuestra percepción incierta del entorno. El modelo de Yu y Dayan es consistente con los datos fisiológicos, farmacológicos y conductuales reportados hasta el momento. A su vez, tiene la virtud de proponer dos sistemas de incertidumbre mediados por diferentes neurotransmisores. En un estudio reciente que no usa la estadística bayesiana y donde las redes bayesianas podrían ser de utilidad, por otra parte, Bromberg-Martin y Hikosaka (2009) han mostrado que la actividad de las neuronas dopaminérgicas en macacos correlacionan con las claves ambientales que indican una reducción de incertidumbre en las decisiones.

Así mismo, las redes bayesianas también podrían ser de utilidad como herramientas estadísticas para modelar procesos cognitivos básicos como la memoria<sup>1</sup>. En el caso del estudio de la memoria es sabido que este proceso básico puede descomponerse en diferentes módulos de procesamiento (Baddeley, 1990; Squire, Knowlton, y Musen, 1993) y que estos fenómenos son susceptibles de ser simulados computacionalmente (p. e., Hasselmo y McClelland, 1999). Por tanto, las redes bayesianas podrían utilizarse como complemento a las actuales técnicas computacionales para modelar las bases biológicas del recuerdo y la consolidación (p. e., Abel y Lattal, 2001; Dudai, 2002; Sara, 2000).

En un sentido más general, las redes bayesianas podrían considerarse apropiadas ante el modelado de los circuitos neurales propiamente dichos desde que se sabe que la proporción de neuronas y conexiones sinápticas sigue un patrón predecible que tiende a optimizar los recursos de cómputo del cerebro (Chklovskii, Schikorski, y Stevens, 2002; Weese y Zador, 2006). En este sentido las redes bayesianas siguen una filosofía parecida en su funcionamiento ya que tienden a optimizar los recursos de cálculo en la actualización de probabilidades consumiendo el mínimo de energía necesaria. De ahí que el modelado de los circuitos neurales que sustentan los procesos psicológicos podrían analizarse bajo la óptica de los

---

<sup>1</sup>Valga el ejemplo presentado en el Capítulo 3 para ejemplificar el funcionamiento de una red bayesiana.

## Reflexiones finales

principios de dependencia e independencia condicional interneuronal y contrastar si el fenómeno de computación local tiene cabida en el modelado neural. No obstante, cabe destacar que las redes bayesianas no han sido concebidas como modelos del sistema neural (como sucede con las *Redes Neuronales Artificiales*) y existen aspectos, como la propagación de probabilidades y la representación del conocimiento, que habría que depurar (Díez, 1994). A su vez, también sería deseable que el desarrollo de sistemas híbridos *Red Neural Artificial-Red Bayesiana* se gestasen al amparo de los conceptos propuestos por las teorías de la complejidad como la *borrosidad*, *fractalidad* o la *no linealidad* (p. e., Díez, 1994; Widjaja, 2000).

Por último, creo que la investigación básica destacada más arriba podría dar lugar al desarrollo de redes bayesianas orientadas a fines prácticos. En concreto, podrían desarrollarse sistemas probabilísticos expertos que sirviesen de ayuda ante el diagnóstico y tratamiento de los trastornos psicológicos. Ya se ha comentado como las redes bayesianas serían útiles para modelar el lenguaje y los trastornos asociados a él (Glymour, 2001) en pacientes neuropsicológicos. El diagnóstico de trastornos como la esquizofrenia puede optimizarse usando estas herramientas en conjunción con las técnicas de resonancia magnética (Davatzikos et al., 2005; Wild, 2005). A su vez, gracias al conocimiento de los mecanismos neuroquímicos responsables de la percepción de la incertidumbre y a la comprensión de las redes neurales implicadas en la toma de decisiones podríamos desarrollar tratamientos eficientes ante las adicciones (Bechara, 2005). Por otro lado, el desarrollo de modelos neurales con redes bayesianas podría ayudar en el desarrollo de terapias frente a enfermedades neurodegenerativas donde se presupone una alteración funcional de los circuitos neurales (Palop, Chin, y Mucke, 2006).

Para terminar, quería hacer un par de apreciaciones sobre la potencialidad de las redes bayesianas en la investigación psicosocial. Primeramente, dado que las redes bayesianas pueden formar parte del «corazón» de un agente inteligente (p. e., Xiang, 2002), podrían utilizarse para modelar procesos sociales complejos como los las dinámicas grupales en diferentes contextos. Un posible uso actual, dada la situación económica global, podría ser el desarrollo de modelos económicos donde los factores psicológicos individuales y sociales estuviesen presentes (p. e., Buchanan, 2009; Farmer y Foley, 2009; Wald, 2008). Por otro lado, como se señaló previamente, las redes bayesianas tienen un uso potencial en la selección

de recursos humanos para las organizaciones y, como se indicó, hay empresas que explotan esta utilidad económicamente. Una aplicación interesante de las redes bayesianas pasaría por utilizarlas como técnica de análisis dentro de los cuadros de mando de recursos humanos en las empresas. Así, teniendo en cuenta que el cuadro de mando se ha entendido metafóricamente como el *panel de control desde donde un equipo de pilotos maneja un avión* (o compañía), y dado que las redes bayesianas son herramientas que estiman probabilidades tras mostrarles evidencias; esta simbiosis podría ser muy interesante en términos prácticos ya que proporcionaría información relevante al equipo directivo de cara la toma de decisiones empresarialmente relevante.

Reflexiones finales

# Parte V

## Anexo



# Portada del formulario

ENCUESTA SOBRE ACTITUDES EMPRENDEDORAS EN LA UNIVERSIDAD

Almería 2006

Estimado/a Sr./Sra.

Desde el *Consejo Social*, la *Fundación Mediterránea* y el *Servicio Universitario de Empleo de la Universidad de Almería* se está desarroyando una investigación sobre **Actitudes Emprendedoras en el Ámbito Universitario**, con la finalidad de Identificar las barreras que limitan la puesta en marcha de empresas u otras alternativas emprendedoras, y ofrecer acciones y alternativas que puedan fomentar el espíritu emprendedor de los estudiantes de la Universidad de Almería.

Consideramos que su colaboración es una vía fundamental para intercambiar conocimientos y experiencias en torno a la actitud emprendedora de los estudiantes universitarios. Por ello, le rogamos dedique unos minutos a contestar este cuestionario.

Le garantizamos la total confidencialidad de los datos, dado que el tratamiento y posterior presentación de los resultados es global y anónima. Si se le presenta cualquier duda o consulta sobre la naturaleza de la investigación, el cuestionario o cualquier otro aspecto, por favor, contacte con los responsables de la investigación - - - en el teléfono - - -, correos electrónicos - - -, o por correo postal a la dirección - - -.

Agradeciéndole de antemano el tiempo dedicado y su contribución a esta investigación, reciba un cordial saludo.



# Escala de carencias formativas percibidas y escala de preparación percibida

## **Escala de carencias formativas percibidas**

**¿Cuáles crees que son tus principales carencias formativas para la puesta en marcha de una empresa propia?** (Por favor, evalúe en una escala de 1 a 5 (1 conocimiento muy bajo o nulo, 2 conocimiento bajo, 3 conocimiento medio, 4 conocimiento bueno, 5 conocimiento muy bueno)<sup>2</sup>.

1. Trámites administrativos
2. Estrategia y organización
3. Contabilidad y finanzas
4. Marketing (comercialización)
5. Producción y operaciones
6. Recursos humanos
7. Aspectos legales (formas jurídicas...)
8. Ayudas y subvenciones
9. Habilidades sociales (comunicación...)

---

<sup>2</sup>Esta escala apareció a la derecha de cada ítem y los participantes tuvieron que hacer una marca sobre el número que correspondía a la respuesta elegida.

10. Idiomas
11. Informática
12. Otras

## Escala de preparación percibida

¿Qué nivel de preparación considera que posee Vd. para crear su propia empresa, en conocimientos, experiencia, contactos, y con carácter general? <sup>3</sup>

1. Conocimientos
2. Experiencia
3. Contactos
4. GENERAL

---

<sup>3</sup>Cada ítem tuvo cinco alternativas de respuesta (1 muy bajo, 2 bajo, 3 aceptable, 4 alto y 5 muy alto) que aparecieron a la derecha donde los participantes tuvieron que hacer una marca para la respuesta elegida dentro de una casilla de verificación.

# Escalas de motivación y obstáculos percibidos ante la creación de una empresa

## **Escala de motivaciones ante la creación de empresas**

**¿Cuáles son sus principales motivaciones para crear un negocio propio?** Por favor, evalúe en una escala de 1 a 5 (1 nada importante, 2 poco importante, 3 importancia media, 4 bastante importante, 5 muy importante).

1. Puesta en marcha de ideas propias
2. Encontrar empleo
3. Independencia económica
4. Independencia personal (ser el propio jefe)
5. Posibilidad de obtener ingresos elevados
6. Dirigir a un grupo de personas
7. Frustración trabajando por cuenta ajena
8. Experiencia enriquecedora en lo profesional y/o personal (aprendizaje)

9. Probarse a uno mismo (desafío, reto)
10. Ayudar a los demás
11. Otra

## **Escala de obstáculos percibidos ante la creación de empresas**

**¿Qué importancia concede a los siguientes obstáculos para poner en marcha una empresa propia?** Por favor, evalúe en una escala de 1 a 5 (1 nada importante, 2 poco importante, 3 importancia media, 4 bastante importante, 5 muy importante).

1. Tener que trabajar demasiadas horas
2. Falta de formación teórica
3. Dificultades para reunir el capital inicial
4. Falta de creatividad para encontrar ideas originales/innovadoras
5. Gran responsabilidad
6. Falta de experiencia
7. Problemas para dirigir personas
8. Cargas fiscales excesivas
9. Riesgo elevado
10. Temor al fracaso
11. Ingresos irregulares
12. Imagen negativa del empresario
13. Estar trabajando en la actualidad
14. Aspectos familiares (falta de apoyo familiar...)

15. Desconocimiento del sector de actividad
16. Trámites burocráticos para la puesta en marcha de la empresa
17. Otras



# Escala de actitudes emprendedoras

A continuación le mostramos una serie de cuestiones. Por favor evalúe las frases que se le indican en función de lo que usted opina con respecto a sus pensamientos y acciones habituales (1 casi nunca me ocurre, 2 algunas veces, 3 a menudo, 4 casi siempre)<sup>4</sup>

1. Prefiero realizar un trabajo lo más estructurado posible, sin margen a la improvisación \*
2. Considero que mi futuro depende fundamentalmente de otras circunstancias externas más que de mis propias decisiones \*
3. Todos mis logros los consigo al primer intento; si no es así, abandono \*
4. Para la realización de mis tareas suelo tener problemas en distribuir el tiempo \*
5. Negociar algo me resulta muy complicado \*
6. Me resulta atractivo dirigir a un grupo de personas en el desempeño de un trabajo
7. Me pongo muy nervioso/a cuando realizo algo públicamente \*
8. Ante los problemas, analizo la situación y planteo distintas soluciones alternativas
9. Las situaciones de tensión son muy desagradables para mi \*

---

<sup>4</sup>Los ítems negativos están marcados con un asterisco.

10. Cuando tengo que competir, me crezco y obtengo mejores resultados
11. Siento que me falta tiempo para llevar a cabo todas mis ideas
12. Suelo tener problemas al hacer planes para emplear mi tiempo libre \*
13. Suelo conseguir lo que me propongo
14. Cuando realizo algún trabajo, me considero que estoy capacitado/a para su desempeño
15. Soy una persona a la que le gusta competir. No importa el lugar, ni con quién
16. Me gusta la aventura, pero freno si veo que el riesgo es elevado
17. Conseguir un buen empleo depende más de mi habilidad que de otros factores
18. Soporto bien las situaciones confusas, poco claras o de indefinición
19. Me gusta plantear y discutir oportunidades de negocio que respondan a necesidades de mi entorno
20. Las situaciones de incertidumbre me generan ansiedad \*
21. Para realizar mis tareas, prefiero que me digan lo que debo hacer\*
22. Participo más y me gustan los juegos de colaboración, donde la responsabilidad de las decisiones es compartida \*
23. Si encuentro dificultades en el trabajo suelo «escurrir el bulto» \*
24. Mi grupo de amigos me busca para organizar fiestas y viajes
25. Suelo tomar mis decisiones de forma rápida, sin valorar los pros y los contras \*
26. Soy capaz de buscar mil razones para conseguir lo que me propongo
27. Cuando tomo una decisión me mantengo tranquilo/a

28. Me agrada poner en práctica una actividad distinta a la que realizo habitualmente
29. Cuando se obtiene éxito en el trabajo es porque se merece



# Ajuste predictivo para los modelos estimados

**Tabla 12.1:** Ajuste predictivo de los modelos estimados

MODELO	VARIABLE DEPENDIENTE	TÉCNICA	VARIABLES INDEPENDIENTES	S	E	FPF	PFN	VPP	VPN	TCC
1	DES	RL	A C P O M	0,9577	0,2057	0,7943	0,0423	0,7795	0,6237	0,7665
2	DES	RL	A C P O	0,9722	0,0742	0,9258	0,0278	0,7545	0,4773	0,7435
3	DES	RL	A C P M	0,9589	0,1738	0,8262	0,0411	0,7731	0,5904	0,7595
4	DES	RL	A C O M	0,9674	0,1649	0,8351	0,0326	0,7709	0,6351	0,7619
5	DES	RL	A P O M	0,9579	0,1979	0,8021	0,0421	0,7781	0,6154	0,7648
6	DES	RL	C P O M	0,9668	0,1414	0,8586	0,0332	0,7700	0,5890	0,7591
7	DES	RL	A C P	0,9831	0,0707	0,9293	0,0169	0,7560	0,5882	0,7509
8	DES	RL	A C O	0,9819	0,0490	0,9510	0,0181	0,7495	0,4828	0,7426
9	DES	RL	A C M	0,9626	0,1754	0,8246	0,0374	0,7725	0,6173	0,7612
10	DES	RL	A P O	0,9784	0,0599	0,9401	0,0216	0,7530	0,4857	0,7446
11	DES	RL	A P M	0,9579	0,1802	0,8198	0,0421	0,7745	0,5930	0,7605
12	DES	RL	A O M	0,9675	0,1638	0,8362	0,0325	0,7703	0,6351	0,7614
13	DES	RL	C P O	0,9989	0,0131	0,9869	0,0011	0,7496	0,8000	0,7498
14	DES	RL	C P M	0,9657	0,1349	0,8651	0,0343	0,7687	0,5694	0,7568
15	DES	RL	C O M	0,9669	0,1270	0,8730	0,0331	0,7655	0,5652	0,7541
16	DES	RL	P O M	0,9659	0,1410	0,8590	0,0341	0,7702	0,5811	0,7586
17	DES	RL	A C	0,9843	0,0455	0,9545	0,0157	0,7495	0,5000	0,7437
18	DES	RL	A P	0,9808	0,0599	0,9401	0,0192	0,7537	0,5152	0,7466
19	DES	RL	A O	0,9892	0,0347	0,9653	0,0108	0,7477	0,5263	0,7440
20	DES	RL	A M	0,9748	0,1568	0,8432	0,0252	0,7704	0,6818	0,7652
21	DES	RL	C P	0,9989	0,0000	1,0000	0,0011	0,7467	0,0000	0,7461
22	DES	RL	C O	1,0000	0,0129	0,9871	0,0000	0,7481	1,0000	0,7490
23	DES	RL	C M	0,9823	0,0945	0,9055	0,0177	0,7620	0,6444	0,7576
24	DES	RL	P O	0,9967	0,0000	1,0000	0,0033	0,7471	0,0000	0,7453
25	DES	RL	P M	0,9692	0,1246	0,8754	0,0308	0,7676	0,5758	0,7572
26	DES	RL	O M	0,9747	0,1197	0,8803	0,0253	0,7653	0,6167	0,7580

DES: Deseabilidad de crear una empresa, COG: componente cognitivo de la creación de empresas, CON: componente conductual de la creación de empresas, RL: regresión logística, RB: red bayesiana, A: actitud medida con ACEMP, C: carencias formativas percibidas, P: preparación percibida, O: obstáculos percibidos, M: motivación, S: sensibilidad, E: especificidad, FPF: proporción de falsos positivos, PFN: proporción de falsos negativos, VPP: valor predictivo positivo, VPN: valor predictivo negativo, TCC: tasa de clasificaciones correctas.

*continúa en la página siguiente*

continúa de la página anterior

MODELO	VARIABLE DEPENDIENTE	TÉCNICA	VARIABLES INDEPENDIENTES	S	E	PFP	PFN	VPP	VPN	TCC
27	DES	RL	A	0,9856	0,0313	0,9688	0,0144	0,7466	0,4286	0,7406
28	DES	RL	C	0,9989	0,0000	1,0000	0,0011	0,7451	0,0000	0,7445
29	DES	RL	P	1,0000	0,0000	1,0000	0,0000	0,7475	-	0,7475
30	DES	RL	O	1,0000	0,0000	1,0000	0,0000	0,7455	-	0,7455
31	DES	RL	M	0,9857	0,0712	0,9288	0,0143	0,7578	0,6286	0,7541
32	COC	RL	A C P O M	0,4085	0,8594	0,1406	0,5915	0,6444	0,6996	0,6862
33	COC	RL	A C P O	0,3967	0,8642	0,1358	0,6033	0,6450	0,6973	0,6850
34	COC	RL	A C P M	0,4085	0,8582	0,1418	0,5915	0,6421	0,6996	0,6856
35	COC	RL	A C O M	0,3700	0,8586	0,1414	0,6300	0,6196	0,6865	0,6712
36	COC	RL	A P O M	0,4070	0,8582	0,1418	0,5930	0,6434	0,6971	0,6840
37	COC	RL	C P O M	0,3581	0,8695	0,1305	0,6419	0,6260	0,6895	0,6758
38	COC	RL	A C P	0,4131	0,8644	0,1356	0,5869	0,6543	0,7034	0,6915
39	COC	RL	A C O	0,3443	0,8677	0,1323	0,6557	0,6176	0,6807	0,6673
40	COC	RL	A C M	0,3560	0,8574	0,1426	0,6440	0,6080	0,6817	0,6652
41	COC	RL	A P O	0,4093	0,8659	0,1341	0,5907	0,6567	0,7005	0,6900
42	COC	RL	A P M	0,4047	0,8555	0,1445	0,5953	0,6374	0,6960	0,6816
43	COC	RL	A O M	0,3318	0,8605	0,1395	0,6682	0,5983	0,6727	0,6568
44	COC	RL	C P O	0,3362	0,8886	0,1114	0,6638	0,6471	0,6879	0,6799
45	COC	RL	C P M	0,3341	0,8723	0,1277	0,6659	0,6145	0,6826	0,6686
46	COC	RL	C O M	0,2789	0,8767	0,1233	0,7211	0,5792	0,6663	0,6505
47	COC	RL	P O M	0,3680	0,8765	0,1235	0,6320	0,6464	0,6933	0,6831
48	COC	RL	A C	0,3232	0,8563	0,1437	0,6768	0,5823	0,6712	0,6523
49	COC	RL	A P	0,4023	0,8690	0,1310	0,5977	0,6578	0,6991	0,6893
50	COC	RL	A O	0,2877	0,8580	0,1420	0,7123	0,5586	0,6585	0,6387
51	COC	RL	A M	0,3271	0,8621	0,1379	0,6729	0,5975	0,6719	0,6563
52	COC	RL	C P	0,3275	0,8796	0,1204	0,6725	0,6224	0,6835	0,6713
53	COC	RL	C O	0,2200	0,8851	0,1149	0,7800	0,5372	0,6518	0,6340
54	COC	RL	C M	0,2309	0,8821	0,1179	0,7691	0,5436	0,6536	0,6359
55	COC	RL	P O	0,3247	0,8849	0,1151	0,6753	0,6329	0,6820	0,6724
56	COC	RL	P M	0,3268	0,8700	0,1300	0,6732	0,6064	0,6784	0,6637
57	COC	RL	O M	0,2160	0,9075	0,0925	0,7840	0,5882	0,6543	0,6451
58	COC	RL	A	0,2900	0,8596	0,1404	0,7100	0,5631	0,6600	0,6408
59	COC	RL	C	0,1656	0,8906	0,1094	0,8344	0,4780	0,6383	0,6174
60	COC	RL	P	0,3636	0,8496	0,1504	0,6364	0,5957	0,6866	0,6656
61	COC	RL	O	0,0454	0,9882	0,0118	0,9546	0,7000	0,6295	0,6312
62	COC	RL	M	0,0821	0,9551	0,0449	0,9179	0,5278	0,6301	0,6241
63	CON	RL	A C P O M	0,0319	0,9990	0,0010	0,9681	0,7500	0,9173	0,9167
64	CON	RL	A C P O	0,0316	0,9990	0,0010	0,9684	0,7500	0,9166	0,9160
65	CON	RL	A C P M	0,0319	0,9980	0,0020	0,9681	0,6000	0,9173	0,9159
66	CON	RL	A C O M	0,0000	1,0000	0,0000	1,0000	-	0,9152	0,9152
67	CON	RL	A P O M	0,0417	0,9980	0,0020	0,9583	0,6667	0,9167	0,9153
68	CON	RL	C P O M	0,0660	0,9973	0,0027	0,9340	0,7000	0,9171	0,9153
69	CON	RL	A C P	0,0316	0,9990	0,0010	0,9684	0,7500	0,9167	0,9161
70	CON	RL	A C O	0,0000	1,0000	0,0000	1,0000	-	0,9145	0,9145
71	CON	RL	A C M	0,0000	1,0000	0,0000	1,0000	-	0,9153	0,9153
72	CON	RL	A P O	0,0309	0,9970	0,0030	0,9691	0,5000	0,9150	0,9128

DES: Deseabilidad de crear una empresa, COG: componente cognitivo de la creación de empresas, CON: componente conductual de la creación de empresas, RL: regresión logística, RB: red bayesiana, A: actitud medida con ACEMP, C: carencias formativas percibidas, P: preparación percibida, O: obstáculos percibidos, M: motivación, S: sensibilidad, E: especificidad, PFP: proporción de falsos positivos, PFN: proporción de falsos negativos, VPP: valor predictivo positivo, VPN: valor predictivo negativo, TCC: tasa de clasificaciones correctas.

continúa en la página siguiente

continúa de la página anterior

MODELO	VARIABLE DEPENDIENTE	TÉCNICA	VARIABLES INDEPENDIENTES	S	E	FPF	PFN	VPP	VPN	TCC
73	CON	RL	A P M	0,0313	0,9980	0,0020	0,9688	0,6000	0,9159	0,9145
74	CON	RL	A O M	0,0000	1,0000	0,0000	1,0000	-	0,9139	0,9139
75	CON	RL	C P O	0,0374	0,9955	0,0045	0,9626	0,4444	0,9140	0,9105
76	CON	RL	C P M	0,0472	0,9964	0,0036	0,9528	0,5556	0,9156	0,9129
77	CON	RL	C O M	0,0000	1,0000	0,0000	1,0000	-	0,9123	0,9123
78	CON	RL	P O M	0,0556	0,9982	0,0018	0,9444	0,7500	0,9151	0,9140
79	CON	RL	A C	0,0000	1,0000	0,0000	1,0000	-	0,9146	0,9146
80	CON	RL	A P	0,0412	0,9970	0,0030	0,9588	0,5714	0,9159	0,9137
81	CON	RL	A O	0,0000	1,0000	0,0000	1,0000	-	0,9132	0,9132
82	CON	RL	A M	0,0000	1,0000	0,0000	1,0000	-	0,9140	0,9140
83	CON	RL	C P	0,0374	0,9964	0,0036	0,9626	0,5000	0,9142	0,9115
84	CON	RL	C O	0,0000	1,0000	0,0000	1,0000	-	0,9116	0,9116
85	CON	RL	C M	0,0000	1,0000	0,0000	1,0000	-	0,9123	0,9123
86	CON	RL	P O	0,0275	0,9973	0,0027	0,9725	0,5000	0,9122	0,9101
87	CON	RL	P M	0,0275	0,9973	0,0027	0,9725	0,5000	0,9122	0,9101
88	CON	RL	O M	0,0000	1,0000	0,0000	1,0000	-	0,9111	0,9111
89	CON	RL	A	0,0000	1,0000	0,0000	1,0000	-	0,9132	0,9132
90	CON	RL	C	0,0000	1,0000	0,0000	1,0000	-	0,9118	0,9118
91	CON	RL	P	0,0000	1,0000	0,0000	1,0000	-	0,9118	0,9118
92	CON	RL	O	0,0000	1,0000	0,0000	1,0000	-	0,9105	0,9105
93	CON	RL	M	0,0000	1,0000	0,0000	1,0000	-	0,9112	0,9112
94	DES	RB	A C P O M	0,9244	0,3674	0,6326	0,0756	0,8100	0,6250	0,7822
95	DES	RB	A C P O	0,9200	0,2843	0,7157	0,0800	0,7895	0,5494	0,7577
96	DES	RB	A C P M	0,9233	0,3419	0,6581	0,0767	0,8036	0,6045	0,7749
97	DES	RB	A C O M	0,9354	0,2939	0,7061	0,0646	0,7944	0,6093	0,7716
98	DES	RB	A P O M	0,9387	0,3419	0,6581	0,0613	0,8062	0,6564	0,7863
99	DES	RB	C P O M	0,9441	0,2843	0,7157	0,0559	0,7937	0,6357	0,7757
100	DES	RB	A C P	0,9211	0,2332	0,7668	0,0789	0,7780	0,5034	0,7455
101	DES	RB	A C O	0,9617	0,1821	0,8179	0,0383	0,7743	0,6196	0,7626
102	DES	RB	A C M	0,9463	0,2556	0,7444	0,0537	0,7876	0,6202	0,7700
103	DES	RB	A P O	0,9474	0,2268	0,7732	0,0526	0,7814	0,5966	0,7635
104	DES	RB	A P M	0,9409	0,3195	0,6805	0,0591	0,8013	0,6494	0,7822
105	DES	RB	A O M	0,9540	0,2588	0,7412	0,0460	0,7897	0,6585	0,7765
106	DES	RB	C P O	0,9441	0,1917	0,8083	0,0559	0,7731	0,5405	0,7520
107	DES	RB	C P M	0,9354	0,2236	0,7764	0,0646	0,7785	0,5426	0,7537
108	DES	RB	C O M	0,9595	0,2013	0,7987	0,0405	0,7780	0,6300	0,7659
109	DES	RB	P O M	0,9606	0,2460	0,7540	0,0394	0,7880	0,6814	0,7781
110	DES	RB	A C	0,9715	0,1438	0,8562	0,0285	0,7680	0,6338	0,7602
111	DES	RB	A P	0,9595	0,1757	0,8243	0,0405	0,7725	0,5978	0,7594
112	DES	RB	A O	0,9770	0,1214	0,8786	0,0230	0,7644	0,6441	0,7586
113	DES	RB	A M	0,9628	0,2268	0,7732	0,0372	0,7841	0,6762	0,7749
114	DES	RB	C P	0,9474	0,1629	0,8371	0,0526	0,7675	0,5152	0,7471
115	DES	RB	C O	0,9869	0,0767	0,9233	0,0131	0,7571	0,6667	0,7545
116	DES	RB	C M	0,9628	0,1757	0,8243	0,0372	0,7731	0,6180	0,7618
117	DES	RB	P O	0,9726	0,1182	0,8818	0,0274	0,7629	0,5968	0,7545
118	DES	RB	P M	0,9584	0,1853	0,8147	0,0416	0,7743	0,6042	0,7610

DES: Deseabilidad de crear una empresa, COG: componente cognitivo de la creación de empresas, CON: componente conductual de la creación de empresas, RL: regresión logística, RB: red bayesiana, A: actitud medida con ACEMP, C: carencias formativas percibidas, P: preparación percibida, O: obstáculos percibidos, M: motivación, S: sensibilidad, E: especificidad, FPF: proporción de falsos positivos, PFN: proporción de falsos negativos, VPP: valor predictivo positivo, VPN: valor predictivo negativo, TCC: tasa de clasificaciones correctas.

continúa en la página siguiente

continúa de la página anterior

MODELO	VARIABLE DEPENDIENTE	TÉCNICA	VARIABLES INDEPENDIENTES	S	E	PFP	PFN	VPP	VPN	TCC
119	DES	RB	O M	0,9682	0,1693	0,8307	0,0318	0,7727	0,6463	0,7643
120	DES	RB	A	0,9923	0,0479	0,9521	0,0077	0,7525	0,6818	0,7512
121	DES	RB	C	0,9989	0,0288	0,9712	0,0011	0,7500	0,9000	0,7512
122	DES	RB	P	1,0000	0,0000	1,0000	0,0000	0,7447	-	0,7447
123	DES	RB	O	1,0000	0,0128	0,9872	0,0000	0,7471	1,0000	0,7480
124	DES	RB	M	0,9748	0,1374	0,8626	0,0252	0,7672	0,6515	0,7610
125	COC	RB	A C P O M	0,5659	0,8270	0,1730	0,4341	0,6650	0,7584	0,7284
126	COC	RB	A C P O	0,5400	0,8204	0,1796	0,4600	0,6460	0,7461	0,7145
127	COC	RB	A C P M	0,5572	0,8008	0,1992	0,4428	0,6293	0,7488	0,7088
128	COC	RB	A C O M	0,5054	0,8440	0,1560	0,4946	0,6629	0,7377	0,7162
129	COC	RB	A P O M	0,5054	0,8480	0,1520	0,4946	0,6686	0,7386	0,7186
130	COC	RB	C P O M	0,5076	0,8440	0,1560	0,4924	0,6638	0,7385	0,7170
131	COC	RB	A C P	0,5076	0,8152	0,1848	0,4924	0,6250	0,7318	0,6990
132	COC	RB	A C O	0,4557	0,8493	0,1507	0,5443	0,6472	0,7200	0,7007
133	COC	RB	A C M	0,4730	0,8218	0,1782	0,5270	0,6169	0,7199	0,6900
134	COC	RB	A P O	0,4816	0,8558	0,1442	0,5184	0,6697	0,7312	0,7145
135	COC	RB	A P M	0,4795	0,8401	0,1599	0,5205	0,6453	0,7268	0,7039
136	COC	RB	A O M	0,4341	0,8611	0,1389	0,5659	0,6547	0,7149	0,6998
137	COC	RB	C P O	0,4816	0,8218	0,1782	0,5184	0,6212	0,7232	0,6933
138	COC	RB	C P M	0,4795	0,8100	0,1900	0,5205	0,6049	0,7194	0,6852
139	COC	RB	C O M	0,3952	0,8742	0,1258	0,6048	0,6559	0,7043	0,6933
140	COC	RB	P O M	0,4320	0,8689	0,1311	0,5680	0,6667	0,7160	0,7039
141	COC	RB	A C	0,4060	0,8388	0,1612	0,5940	0,6045	0,6995	0,6754
142	COC	RB	A P	0,4795	0,8453	0,1547	0,5205	0,6529	0,7280	0,7072
143	COC	RB	A O	0,3888	0,8807	0,1193	0,6112	0,6642	0,7037	0,6949
144	COC	RB	A M	0,3715	0,8506	0,1494	0,6285	0,6014	0,6904	0,6697
145	COC	RB	C P	0,4492	0,8152	0,1848	0,5508	0,5960	0,7092	0,6770
146	COC	RB	C O	0,3283	0,8820	0,1180	0,6717	0,6281	0,6839	0,6729
147	COC	RB	C M	0,3240	0,8755	0,1245	0,6760	0,6122	0,6809	0,6672
148	COC	RB	P O	0,3974	0,8702	0,1298	0,6026	0,6502	0,7041	0,6917
149	COC	RB	P M	0,3952	0,8598	0,1402	0,6048	0,6310	0,7009	0,6843
150	COC	RB	O M	0,2570	0,9148	0,0852	0,7430	0,6467	0,6699	0,6664
151	COC	RB	A	0,3067	0,8834	0,1166	0,6933	0,6147	0,6774	0,6656
152	COC	RB	C	0,1512	0,9332	0,0668	0,8488	0,5785	0,6443	0,6378
153	COC	RB	P	0,4471	0,8060	0,1940	0,5529	0,5831	0,7061	0,6705
154	COC	RB	O	0,1685	0,9450	0,0550	0,8315	0,6500	0,6519	0,6517
155	COC	RB	M	0,0994	0,9554	0,0446	0,9006	0,5750	0,6361	0,6321
156	CON	RB	A C P O M	0,2844	0,9829	0,0171	0,7156	0,6200	0,9334	0,9206
157	CON	RB	A C P O	0,2661	0,9847	0,0153	0,7339	0,6304	0,9320	0,9206
158	CON	RB	A C P M	0,2202	0,9820	0,0180	0,7798	0,5455	0,9278	0,9141
159	CON	RB	A C O M	0,1376	0,9919	0,0081	0,8624	0,6250	0,9215	0,9157
160	CON	RB	A P O M	0,2202	0,9946	0,0054	0,7798	0,8000	0,9287	0,9255
161	CON	RB	C P O M	0,2110	0,9811	0,0189	0,7890	0,5227	0,9270	0,9124
162	CON	RB	A C P	0,2110	0,9820	0,0180	0,7890	0,5349	0,9271	0,9133
163	CON	RB	A C O	0,1009	0,9919	0,0081	0,8991	0,5500	0,9185	0,9124
164	CON	RB	A C M	0,0826	0,9928	0,0072	0,9174	0,5294	0,9170	0,9116

DES: Deseabilidad de crear una empresa, COG: componente cognitivo de la creación de empresas, CON: componente conductual de la creación de empresas, RL: regresión logística, RB: red bayesiana, A: actitud medida con ACEMP, C: carencias formativas percibidas, P: preparación percibida, O: obstáculos percibidos, M: motivación, S: sensibilidad, E: especificidad, PFP: proporción de falsos positivos, PFN: proporción de falsos negativos, VPP: valor predictivo positivo, VPN: valor predictivo negativo, TCC: tasa de clasificaciones correctas.

continúa en la página siguiente

continúa de la página anterior

MODELO	VARIABLE DEPENDIENTE	TÉCNICA	VARIABLES INDEPENDIENTES	S	E	PPF	PFN	VPP	VPN	TCC
165	CON	RB	A P O	0,2202	0,9955	0,0045	0,7798	0,8276	0,9288	0,9264
166	CON	RB	A P M	0,1651	0,9937	0,0063	0,8349	0,7200	0,9240	0,9198
167	CON	RB	A O M	0,0917	0,9964	0,0036	0,9083	0,7143	0,9180	0,9157
168	CON	RB	C P O	0,1835	0,9856	0,0144	0,8165	0,5556	0,9250	0,9141
169	CON	RB	C P M	0,1743	0,9829	0,0171	0,8257	0,5000	0,9240	0,9108
170	CON	RB	C O M	0,0734	0,9973	0,0027	0,9266	0,7273	0,9166	0,9149
171	CON	RB	P O M	0,1651	0,9955	0,0045	0,8349	0,7826	0,9241	0,9214
172	CON	RB	A C	0,0734	0,9946	0,0054	0,9266	0,5714	0,9164	0,9124
173	CON	RB	A P	0,1376	0,9955	0,0045	0,8624	0,7500	0,9218	0,9190
174	CON	RB	A O	0,0550	0,9982	0,0018	0,9450	0,7500	0,9152	0,9141
175	CON	RB	A M	0,0550	0,9973	0,0027	0,9450	0,6667	0,9151	0,9133
176	CON	RB	C P	0,1651	0,9829	0,0171	0,8349	0,4865	0,9232	0,9100
177	CON	RB	C O	0,0550	0,9982	0,0018	0,9450	0,7500	0,9152	0,9141
178	CON	RB	C M	0,0459	0,9973	0,0027	0,9541	0,6250	0,9143	0,9124
179	CON	RB	P O	0,1193	0,9973	0,0027	0,8807	0,8125	0,9204	0,9190
180	CON	RB	P M	0,1009	0,9919	0,0081	0,8991	0,5500	0,9185	0,9124
181	CON	RB	O M	0,0550	0,9991	0,0009	0,9450	0,8571	0,9152	0,9149
182	CON	RB	A	0,0183	1,0000	0,0000	0,9817	1,0000	0,9123	0,9124
183	CON	RB	C	0,0092	1,0000	0,0000	0,9908	1,0000	0,9115	0,9116
184	CON	RB	P	0,1101	0,9919	0,0081	0,8899	0,5714	0,9192	0,9133
185	CON	RB	O	0,0367	1,0000	0,0000	0,9633	1,0000	0,9138	0,9141
186	CON	RB	M	0,0275	0,9991	0,0009	0,9725	0,7500	0,9130	0,9124

DES: Deseabilidad de crear una empresa, COG: componente cognitivo de la creación de empresas, CON: componente conductual de la creación de empresas, RL: regresión logística, RB: red bayesiana, A: actitud medida con ACEMP, C: carencias formativas percibidas, P: preparación percibida, O: obstáculos percibidos, M: motivación, S: sensibilidad, E: especificidad, PFP: proporción de falsos positivos, PFN: proporción de falsos negativos, VPP: valor predictivo positivo, VPN: valor predictivo negativo, TCC: tasa de clasificaciones correctas.



## Referencias

- Abel, T., y Lattal, K. M. (2001). Molecular mechanisms of memory acquisition, consolidation and retrieval. *Current Opinion in Neurobiology*, *11*, 180–187.
- Adarraga, P., y Zaccagnini, J. L. (1988). Sistemas expertos y psicología cognitiva: una visión general. *Estudios de Psicología*, *36*, 97–127.
- Adarraga, P., y Zaccagnini, J. L. (1992). DAI: a knowledge-based system for diagnosing autism. A case study on the application of artificial intelligence to psychology. *European Journal of Psychology Assessment*, *8*, 25–46.
- Ajzen, I., y Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood, NJ: Prentice-Hall.
- Ajzen, I., y Fishbein, M. (2005). The influence of attitudes on behavior. En D. Albarracín, B. T. Hohnson, y M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 173–221). Mahwah, NJ: Erlbaum.
- Alcazar, P. (2005, Mayo). Métodos para decidir mejor. *Emprendedores*, *92*, 87–92.
- Alderete, A. M. (2006). Fundamentos del análisis de regresión logística en la investigación psicológica. *Evaluar*, *6*, 52–67.
- Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: SAGE University Papers.
- Allport, G. W. (1935). Attitudes. En C. Murchison (Ed.), *Handbook of social psychology* (pp. 798–844). Worcester, MA: Clark University Press.
- Alonso, D., y Tubau, E. (2002). Inferencias bayesianas: una revisión. *Anuario de Psicología*, *33*, 25–47.
- Altman, D. G., Lausen, B., Sauerbrei, W., y Schumacher, M. (1994). Dangers of using “optimal” cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute*, *86*, 829–835.

## Referencias

- American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5 ed.). Washington, DC: Autor.
- Anathema. (1998). Shroud of false. En *Alternative 4* [CD]. Londres: Peaceville Records.
- Ankarali, H., Canan, A., Akkus, Z., Bugdayci, R., y Ali, M. (2007). Comparison of logistic regression model and classification tree: An application to post-partum depression data. *Expert Systems with Applications*, 32, 987–994.
- Arjona, A., y Checa, J. C. (2006). Empresariado extracomunitario en Almería: estructuras de oportunidad, características del grupo y estrategias étnicas. *Revista Española de Investigaciones Sociológicas*, 115, 297–317.
- Asián, R., Fernandez, J. C., y Otri, A. M. (2005). Medición de la capacidad emprendedora de los empresarios de Coria del Río: la autoeficacia percibida. En *Cities in competition: Situation of the entrepreneurship, business creation, human resource management and family business perspectives* (pp. 83–98). Sevilla: Universidad de Sevilla / Mergablum.
- Ato, M., y López, J. J. (1996). *Análisis estadístico para datos categóricos*. Madrid: Síntesis.
- Baddeley, A. (1990). *Human memory. Theory and practise*. Hove: Lawrence Erlbaum.
- Baesens, B., Verstraeten, G., Van den Poel, D., Egmont-Petersen, M., Van Kenhove, P., y Vanthienen, J. (2004). Bayesian classifiers for identifying the slope of the customer life cycle of long-life customers. *European Journal of Operational Research*, 156, 508–523.
- Barnes, D. (1994). Stimulus equivalence and relational frame theory. *The Psychological Record*, 44, 91–124.
- Barnes, D., y Hampson, P. J. (1993). Stimulus equivalence and connectionism: implications for behavior analysis and cognitive science. *The Psychological Record*, 43, 617–638.
- Barnes, D., y Holmes, Y. (1991). Radical behaviorism, stimulus equivalence, and human cognition. *The Psychological Record*, 41, 19–31.
- Barón, F. J., y Téllez, F. (2005). *Manual básico de bioestadística*. Málaga: Téllez Montiel, F.
- Barraclough, D. J., Conroy, M. L., y Lee, D. (2004). Prefrontal cortex and decision making in a mixed strategy game. *Nature Neuroscience*, 7, 404–410.

- Bartfay, E., Mackillop, W. J., y Pater, J. L. (2006). Comparing the predictive value of neural network models to logistic regression models on the risk of death for small-cell lung cancer patients. *European Journal of Cancer Care*, *15*, 115–124.
- Batista, J. M., y Coenders, G. (2000). *Modelos de ecuaciones estructurales*. Madrid: La Muralla / Hespérides.
- Baum, J. R., Frese, M., y Baron, R. (2006). *The psychology of entrepreneurship*. Mahwah, NJ: Lawrence Erlbaum.
- Bautista, L. E. (1995). “Razón relativa” y “tasa relativa” como traducciones de *odds ratio* y *hazard ratio*. *Boletín de la Oficina Sanitaria Panamericana*, *119*, 278–279.
- Bechara, A. (2005). Decision making, impulse control and loss of willpower to resist drugs: a neurocognitive perspective. *Nature Neuroscience*, *8*, 1458–1463.
- Blaisdell, A. P., Sawa, K., Leising, K. J., y Waldmann, M. R. (2006, Febrero 17). Causal reasoning in rats. *Science*, *311*, 1020–1022.
- Borsboom, D. (2002). [Revisión del libro *The mind's arrows. Bayes nets and graphical causal models in psychology*]. *Acta Psychologica*, *111*, 355–357.
- Boyatzis, R. E. (1982). Competence at work. En A. J. Stewart (Ed.), *Motivation an society* (pp. 221–243). San Francisco, CA: Jossey-Bass.
- Boyce (Productor), S., Kral (Editor), K., Allnut (Editor), C., y Goetz (Editor), A. (2005). *DNA mysteries: the search for Adam* [Documental]. Estados Unidos: National Geographic Television and Film Productions / National Geographic Society.
- Brehm, S. S., Kassin, S., y Fein, S. (2005). *Social psychology* (6 ed.). New York: Houghton Mifflin.
- Briley, J. (2008). *Cry freedom*. Hong Kong: Oxford University Press. (Trabajo original publicado en 1989)
- Bromberg-Martin, E. S., y Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron*, *63*, 119–126.
- Bucay, J. (1999). *Déjame que te cuente...: los cuentos que me enseñaron a vivir*. Barcelona: Círculo de Lectores.
- Buchanan, M. (2009, Agosto 6). Meltdown modeling. *Nature*, *460*, 680–682.

## Referencias

- Bull, S. B., Mak, C., y Greenwood, C. M. T. (2002). A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics and Data Analysis*, 39, 57–74.
- Burgeño, M. J., García-Bastos, J. L., y González-Buitrago, J. M. (1995). Las curvas ROC en la evaluación de las pruebas diagnósticas. *Medicina Clínica*, 104, 661–670.
- Cano, C. J., García, J., y Gea, A. B. (2003). *Actitudes emprendedoras y creación de empresas en los estudiantes universitarios*. Almería: Servicio de Publicaciones de la Universidad de Almería / Consejo Social de la Universidad de Almería.
- Cantillon, R. (1931). *Ensayo sobre la naturaleza del comercio en general*. México: Fondo de Cultura Económica. (Trabajo original publicado en 1755)
- Carlson, N. R. (2000). *Fisiología de la conducta* (3ª ed.). Barcelona: Ariel. (Trabajo original publicado en 1993)
- Cassarino, J. P. (2004). Theorising return migration: the conceptual approach to return migrants revisited. *International Journal on Multicultural Societies*, 6, 253–279.
- Castillo, E., Gutiérrez, J. M., y Hadi, A. S. (1998). *Sistemas expertos y modelos de redes probabilísticas*. Madrid: Academia Española de Ingeniería.
- Céspedes, A. J., Rumí, R., Salmerón, A., y Soler, F. J. (2003, Abril). *Análisis del sector agrario del poniente almeriense mediante redes bayesianas*. Artículo presentado en el 27 Congreso Nacional de Estadística e Investigación Operativa. Lleida.
- Cheng, P. W. (1997). From covariation to causation: a causal power theory. *Psychological Review*, 104, 367–450.
- Chevalier, T. (1999). *A girl with a pearl earring*. Londres: Harper Collins.
- Chklovskii, D. B., Schikorski, T., y Stevens, C. F. (2002). Wiring optimization in cortical circuits. *Neuron*, 34, 341–347.
- Chuang, H. L. (1997). High school youths' dropout and re-enrollment behavior. *Economics of Education Review*, 16, 171–186.
- Clayton, N., y Dickinson, A. (2006). Rational rats. *Nature Neuroscience*, 9, 472–474.
- Cohen, J. D., y Aston-Jones, G. (2005, Julio 28). Decision amid uncertainty. *Nature*, 436, 471–472.

- Conati, C., Gertner, A., y VanLehn, K. (2002). Using bayesian networks to manage uncertainty in student modeling. *Modeling and User-Adapted Interaction*, 12, 371–417.
- Concato, J., Peduzzi, P., Holford, T. R., y Feinstein, A. R. (1995). Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy. *Journal of Clinical Epidemiology*, 48, 1495–1501.
- Concejero, P. (2004). *Comparación de los modelos de curvas ROC para la evaluación de procedimientos estadísticos de predicción en investigación de mercados*. Tesis doctoral no publicada, Facultad de Psicología, Universidad Complutense de Madrid.
- Conejo, R., Guzmán, E., Millán, E., Trella, M., Perez-de-la-Cruz, J. L., y Ríos, A. (2004). Siette: A web-based tool for adaptive testing. *International Journal of Artificial Intelligence in Education*, 14, 1–33.
- Conejo, R., Millán, E., Perez de la Cruz, J. L., y Trella, M. (2001). Modelado del alumno: un enfoque bayesiano. *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*, 12, 50–58.
- Corman, J., Lussier, R., y Nolan, K. G. (1996). Factors that encourage entrepreneurial start-ups and existing firm expansion: a longitudinal study comparing recession and expansion periods. *Academy of Entrepreneurship Journal*, 1, 43–55.
- Cowell, R. G., Dawid, A. P., Lauritzen, S. L., y Spiegelhalter, D. J. (1999). *Probabilistic networks and expert systems*. Harrisonburg, VA: Springer.
- Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics*, 14, 1–13.
- Cusmille, F., y Bangdiwala, S. I. (2000). Categorización de variables en el análisis estadístico de datos: consecuencias sobre la interpretación de resultados. *Revista Panamericana de Salud Pública*, 8, 348–354.
- Das, B. (2004). Generating conditional probabilities for bayesian networks: easing the knowledge acquisition problem. Descargado el 10 de Septiembre, 2005, desde <http://arxiv.org/abs/cs.AI/0411034>.
- Davatzikos, C., Shen, D., Gur, R. C., Wu, X., Liu, D., Fan, Y., et al. (2005). Whole-brain morphometric study of schizophrenia revealing a spatially complex set of focal abnormalities [Abstract]. *Archives of General Psychiatry*,

## Referencias

- 62, 1218–1227.
- Dawkins, R. (1988). *El gen egoísta*. Barcelona: Salvat. (Trabajo original publicado en 1976)
- Díaz, J. C. (2003). *La creación de empresas en extremadura. Un análisis institucional*. Tesis doctoral no publicada, Departamento de Economía Financiera y Contabilidad, Universidad de Extremadura.
- De la Fuente, E. I., García, J., y De la Fuente, L. (2002). Estadística bayesiana en la investigación psicológica. *Metodología de las Ciencias del Comportamiento*, 4, 185–200.
- De las Nieves, F. J. (2008, Mayo 15). Prólogo. *Campus ETB*, 1, 3.
- Deaux, K., Dane, C. F., y Wrightsman, L. S. (1993). *Social psychology in the 90s* (6 ed.). Pacific Grove, CA: Brooks/Cole.
- De Cervantes, M. (2004). *El ingenioso hidalgo Don Quixote de la Mancha*. Castilla-La-Mancha: Biblioteca IV Centenario. (Trabajo original publicado en 1605)
- Delgado, M. L. (2003). *Aplicación de las redes neurales artificiales a la estadística*. Madrid: Muralla / Hespérides.
- DeMaris, A. (2002). Explained variance in logistic regression. A Monte Carlo study of proposed measures. *Sociological Methods & Research*, 31, 27–74.
- Díez, F. J. (1994). *Sistema experto para la ecocardiografía*. Tesis doctoral no publicada, Facultad de Ciencias, Universidad Nacional de Educación a Distancia.
- Díez, F. J. (1998). Aplicaciones de los modelos gráficos probabilistas en medicina. En J. A. Gámez y J. M. Puerta (Eds.), *Sistemas expertos probabilísticos* (pp. 239–263). Cuenca: Servicio de Publicaciones de la Universidad de Castilla-La Mancha.
- Ditterich, J., Mazurek, M. E., y Shadlen, M. (2003). Microestimulation of visual cortex affects the speed of perceptual decisions. *Nature Neuroscience*, 6, 891–898.
- Dixon, J. R. (1970). *Introducción a la probabilidad. Texto programado*. México: Limusa-Wiley. (Trabajo original publicado en 1964)
- Domingo, A., Bajo, A. M., Chiloeches, A., y García, V. (2008, Noviembre). «PROFESIONAL.ES - EMPRENDEDOR.ES». *discovery of motivation, skills and the value of knowledge looking at the professional future. a co-*

- llective experience with the students of first year of chemistry*. Comunicación presentada en la International Conference of Education, Research and Innovation (ICERI 2008). Madrid.
- Domingos, P., y Pazzani, M. (1996). Beyond independence: conditions for the optimality of the simple bayesian classifier. En L. Saitta (Ed.), *Proceedings of the 13th international conference on machine learning* (pp. 105–112). Bari, Italia: Morgan Kaufman.
- Dudai, Y. (2002). Molecular bases of long-term memories: a question of persistence. *Current Opinion in Neurobiology*, 12, 211–216.
- Duncan Dhu. (1989). El ritmo de la calle. En *Autobiografía* [LP]. Inglaterra: Dro East West.
- Edwards, W. (1998). Hailfinder. Tools for and experiences with bayesian normative modeling. *American Psychologist*, 53, 416–428.
- Edwards, W., y Fasolo, B. (2001). Decision technology. *Annual Review of Psychology*, 52, 581–606.
- Eftekhari, B., Mohammad, K., Ardebili, H. E., Ghodsi, M., y Ketabchi, E. (2005). Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. *BMC Medical Informatics and Decision Making*, 5, –.
- Ellis, A., y Young, A. W. (1988). *Human cognitive neuropsychology*. Hove: Lawrence Erlbaum.
- Esteve, M. R. (1996a). Teoría de la decisión (II): el teorema de Bayes. En A. Godoy (Ed.), *Toma de decisiones y juicio clínico. Una aproximación psicológica* (pp. 191–247). Madrid: Pirámide.
- Esteve, M. R. (1996b). Teoría de la decisión (I): teoría de la utilidad esperada. En A. Godoy (Ed.), *Toma de decisiones y juicio clínico. Una aproximación psicológica* (pp. 175–190). Madrid: Pirámide.
- Farmer, J. D., y Foley, D. (2009, Agosto 6). The economy needs agent-based modelling. *Nature*, 460, 685–686.
- Feldman, R. S. (1998). *Social psychology* (2 ed.). Upper Saddle River, NJ: Prentice Hall.
- Fernández, J. L., y Pinillos, J. L. (1992). *Atención y percepción*. Madrid: Alhambra.
- Finch, H., y Schneider, M. K. (2007). Classification accuracy of neural net-

## Referencias

- works vs. discriminant analysis, logistic regression, and classification and regression trees. *Methodology*, 3, 47–57.
- Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80, 27–38.
- Fishbein, M., y Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.
- Fitch, W. T. (2007, Octubre 11). An invisible hand. *Nature*, 449, 665–667.
- Flegal, K. M., Keyl, P. M., y Nieto, F. J. (1991). Differential misclassification arising from nondifferential errors in exposure measurement. *American Journal of Epidemiology*, 134, 1233–1244.
- Forester, J. (1973). *Estadística en la toma de decisiones*. Bilbao: Deusto. (Trabajo original publicado en 1968)
- Foshee, A., Heath, W. C., y Balic, M. (2003). Price adjustment and the market process: dealing with disequilibrium. *Journal of Economic and Economic Education Research*, 4(3), 33–47.
- Franco, M., y Vivo, J. M. (2007). *Análisis de curvas ROC. Principios y aplicaciones*. Madrid: La Muralla.
- Franzoi, S. L. (2005). *Social psychology* (4 ed.). New York: Mc Graw Hill.
- Freedman, L. S., y Pee, D. (1989). Return to a note on screening regression equations. *The American Statistician*, 43, 279–282.
- Fuentes, L. J. (2001). Déficit de atención selectiva en la esquizofrenia. *Revista de Neurología*, 32, 387–391.
- Fung, K. Y., y Howe, G. R. (1984). Methodological issues in case-control studies III: The effect of joint misclassification of risk factors and confounding factors upon estimation and power. *International Journal of Epidemiology*, 13, 366–370.
- Gámez, J. A. (1998). Abducción en modelos gráficos. En J. A. Gámez y J. M. Puerta (Eds.), *Sistemas expertos probabilísticos* (pp. 79–111). Cuenca: Servicio de Publicaciones de la Universidad de Castilla-La Mancha.
- Garbolino, P., y Taroni, F. (2002). Evaluation of scientific evidence using bayesian networks. *Forensic Science International*, 125, 149–155.
- García, J., Cano, C. J., y Gea, A. B. (2005). Actitudes emprendedoras en estudiantes universitarios y empresarios. Evidencias de validez de un instrumento. *Iberpsicología*, 10(8), art. 12.

- García, J., López, J., Cano, C. J., Gea, A. B., y De la Fuente, E. I. (2006, Septiembre). *Aplicación de las redes bayesianas al modelado de las actitudes emprendedoras*. Comunicación presentada en el IV Congreso de Metodología de Encuestas. Pamplona.
- García, J., López, J., De la Fuente, L., Cano, C. J., y Gea, A. B. (2007, Febrero). *Modelos de ecuaciones estructurales y redes bayesianas. Una perspectiva confirmatoria aplicada a las actitudes emprendedoras*. Comunicación presentada en el X Congreso de Metodología de las Ciencias Sociales y de la Salud. Barcelona.
- García, J. A. (1997, Abril). *Estudio de los factores que condicionan el éxito o fracaso de las PYMES en Andalucía*. Comunicación presentada en el I Congreso de Ciencia Regional de Andalucía: Andalucía en el Umbral del Siglo XXI. Jerez de la Frontera.
- Gartner, W. B. (1988). "Who is an entrepreneur?" Is the wrong question. *American Journal of Small Business*, 12(4), 11–32.
- Genescá, E., y Capelleras, J. L. (2004). Un análisis comparativo de las características de las microempresas en España. *Universia Business Review*, 2, 72–93.
- Genesca, E., y Veciana, J. M. (1984). Actitudes hacia la creación de empresas. *Información Comercial Española*, 611, 147–155.
- Genkin, A., Lewis, D. D., y Madigan, D. (2005). BBR: Bayesian logistic regression software. Descargado el 7 de Marzo de 2009, desde <http://www.stat.rutgers.edu/~madigan/BBR/>.
- Gigerenzer, G. (1996). On narrow norms and vague heuristic: a reply to Kahneman and Tversky. *Psychological Review*, 103, 592–596.
- Gigerenzer, G. (2007). *Gut feelings: the intelligence of unconscious*. Londres: Penguin Books.
- Giles, J. (2005a, Diciembre 15). Internet encyclopaedias go head to head. *Nature*, 438, 900–901.
- Giles, J. (2005b, Noviembre 17). Wisdom of the crowd. *Nature*, 438, 281.
- Glymour, C. (2001). *The mind's arrows. Bayes nets and graphical causal models in psychology*. Cambridge, MA: MIT Press.
- Glymour, C. (2003). Learning, prediction and causal Bayes nets. *Trends in Cognitive Sciences*, 7, 43–48.

## Referencias

- Gómez, J. M., Mira, I., y Martínez, J. (2007). Condicionantes de la actividad emprendedora e instituciones de apoyo desde el ámbito local: el caso de la provincia de Alicante. *Revista de Empresa*, 20, 20–31.
- Godoy, A. (1996). *Toma de decisiones y juicio clínico. Una aproximación psicológica*. Madrid: Pirámide.
- Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., y Danks, D. (2004). A theory of causal learning in children: causal and bayes nets. *Psychological Review*, 111, 3–32.
- Gopnik, A., y Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends in Cognitives Sciences*, 8, 371–377.
- Gopnik, A., Sobel, D. M., Schulz, L., y Glymour, C. (2001). Causal learning mechanisms in very young children: two, three, and four-years-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, 37, 620–629.
- Gottfredson, L. S. (1998). The general intelligence factor. *Scientific American Presents*, 9(4), 24–29.
- Gould, R. (1988). *Graph theory*. Menlo Park, CA: Benjamin/Cummings.
- Greene, G. (2004). *The quiet American*. Londres: Vintage. (Trabajo original publicado en 1955)
- Greiner, R., Su, X., Shen, B., y Zhou, W. (2005). Structural extension to logistic regression: discriminative parameter learning of belief net classifiers. *Machine Learning*, 59, 297–322.
- Greiner, R., y Zhou, W. (2002). Structural extension to logistic regression: discriminative parameter learning of belief net classifiers. *Proceedings of the Eighteenth Annual National Conference on Artificial Intelligence, Aug, 2002*, 167–173.
- Grundling, J., y Steynberg, L. (2008, Noviembre). *Fostering entrepreneurship through management of innovation research: the case of Tshwane University of Technology, South Africa*. Comunicación presentada en la International Conference of Education, Research and Innovation (ICERI 2008). Madrid.
- Hair, J. F., Anderson, R. E., Tatham, R. L., y Black, W. C. (1998). *Multivariate data analysis*. Englewood Cliffs, NY: Prentice Hall.
- Hanley, J. A., y McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143, 29–36.

- Hanley, J. A., y McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148, 839–843.
- Harary, F. (1969). *Graph theory*. Philippines: Westview Press.
- Harrel, F. E., Lee, K. E., Matchar, D. B., y Reichert, T. A. (1985). Regression models for prognostic prediction: advantages, problems, and suggested solutions. *Cancer Treatment Reports*, 69, 1071–1077.
- Hasselmo, M. E., y McClelland, J. L. (1999). Neural models of memory. *Current Opinion in Neurobiology*, 9, 184–188.
- Hauser, H. (2008, Diciembre 11). Right environment can enhance ‘innate’ entrepreneurial skills. *Nature*, 456, 700.
- Hayek, F. A. (1985). Robert cantillon. *The Journal of Libertarian Studies*, 7, 217–247.
- HayGroup, y SAP AG. (2003). *Factbook, recursos humanos*. Navarra: Aranzadi.
- Hébert, R. F., y Link, A. (1989). In search of the meaning of entrepreneurship. *Small Business Economics*, 1, 39–49.
- Heckerman, D. (1995). *A tutorial on learning with bayesian networks* (Rep. Téc. MS-TR-95-06). Redmon, WA: Microsoft Research.
- Heckerman, D., Mamdani, A., y Wellman, M. P. (1995). Real-world applications of bayesian networks. *Communications of the Association for Computing Machinery*, 38(3), 24–26.
- Herskovits, E. H., y Dagher, A. P. (1997). *Applications of bayesian networks to health care* (Rep. Téc. NSI-TR-1997-02). Baltimore, MD: Noetic Systems.
- Highsmith, P. (1999). *The talented Mr. Ripley*. Londres: Vintage. (Trabajo original publicado en 1955)
- Hirji, K. F., Mehta, C. R., y Patel, N. R. (1987). Computing distributions for exact logistic regression. *Journal of the American Statistical Association*, 82, 1110–1117.
- Hoeting, J. A., Madigan, D., y Raftery, A. E. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 14, 382–401.
- Hopley, L., y van Schalkwyk, J. (2001). The magnificent ROC (Receiver Operating Characteristic curve). Descargado el 10 de Enero de 2006, desde <http://www.anaesthetist.com/mnm/stats/roc/Findex.htm>.
- Hosmer, D. W., Taber, S., y Lemeshow, S. (1991). The importance of assessing

## Referencias

- the fit of logistic regression models: a case study. *American Journal of Public Health*, 81, 1630–1635.
- Hothersall, D. (1997). *Historia de la psicología* (3<sup>a</sup> ed.). México: McGraw-Hill. (Trabajo original publicado en 1995)
- Hsieh, F. Y. (1989). Sample size tables for logistic regression. *Statistics in Medicine*, 8, 795–802.
- Hsieh, F. Y., Bloch, D. A., y Larsen, M. D. (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine*, 17, 1623–1634.
- Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., y Camerer, C. F. (2005, Diciembre 9). Neural systems responding of uncertainty in human decision-making. *Science*, 310, 1680–1683.
- Huefner, J. C., Hunt, H. K., y Robinson, P. B. (1996). A comparison of four scales predicting entrepreneurshp. *Academy of Entrepreneurship Journal*, 1, 56–80.
- Huete, J. F. (1998). Sistemas expertos probabilísticos: modelos gráficos. En J. A. Gámez y J. M. Puerta (Eds.), *Sistemas expertos probabilísticos* (pp. 1–40). Cuenca: Servicio de Publicaciones de la Universidad de Castilla-La Mancha.
- Hugin Expert. (2004). Tutorials. Using learning facilities. Descargado el 21 de Septiembre de 2005, desde <http://www.hugin.com/tutorials/>.
- Indeep. (1982). Last night a DJ saved my life. En *Last night a DJ saved my life* [Single]. New York: Sonud of New York.
- Irala, J., Fernández-Crehuet, R., y Serrano, A. (1997). Intervalos de confianza anormalmente amplios en regresión logística: interpretación de resultados de programas estadísticos. *Revista Panamericana de Salud Pública*, 1, 230–234.
- Jaimes, F., Farbiarz, J., Alvarez, D., y Martínez, C. (2005). Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room. *Critical Care*, 9, 150–156.
- Jansen, R., Yu, H., Greenbaum, D., Kluger, N., Krogan, N. J., Chung, S., et al. (2003, Octubre 17). A bayesian network approach for predicting protein-protein interactions from genomic data. *Science*, 302, 449–453.
- Jastrow, R. (1993). *El telar mágico. El cerebro humano y la computadora*. Bar-

- celona: Salvat. (Trabajo original publicado en 1981)
- Johnson, V. E., y Albert, J. H. (1999). *Ordinal data modelling*. New York: Springer-Verlag.
- Jovel, A. J. (1995). *Análisis de regresión logística*. Madrid: Centro de Investigaciones Sociológicas.
- Juan, A. M. (2003). Regresión lineal múltiple. En Y. del Águila et al. (Eds.), *Técnicas estadísticas aplicadas al análisis de datos* (pp. 144–172). Almería: Servicio de Publicaciones de la Universidad de Almería.
- Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, 20, 137–194.
- Kadie, C. M., Hovel, D., y Horvitz, E. (2001). *MSBNx: a component-centric toolkit for modeling and inference with bayesian networks* (Rep. Téc. MST-TR-2001-67). Redmon, WA: Microsoft Research.
- Kahneman, D. (2003). Experiences of collaborative research. *American Psychologist*, 58, 723–730.
- Kahneman, D., Slovic, P., y Tversky, A. (1982). *Judgement under uncertainty: heuristic and biases*. New York: Cambridge University Press.
- Kahneman, D., y Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237–251.
- Kandel, E. R., Schwartz, J. H., y Jessell, T. M. (1997). *Neurociencia y conducta*. Madrid: Prentice Hall. (Trabajo original publicado en 1995)
- Kansas. (1977). Dust in the wind. En *Point of know return* [LP]. Bogalusa, LA: Kirshner Records.
- Killers, T. (2008). Human. En *Day & Age* [CD]. New York: Island, Vertigo.
- King, E., y Ryan, T. P. (2002). A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression. *The American Statistician*, 56, 163–170.
- King, G., y Zeng, L. (2001a). Explaining rare events in international relations. *International Organization*, 55, 693–715.
- King, G., y Zeng, L. (2001b). Logistic regression in rare events data. *Political Analysis*, 9, 137–163.
- Kleinbaum, D. G. (1994). *Logistic regression*. New York: Springer.
- Knill, D. C., y Pouget, A. (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27, 712–719.

## Referencias

- Korb, K. B., y Nicholson, A. E. (2004). *Bayesian artificial intelligence*. Boca Raton, FL: Chapman & Hall/CRC.
- Krynski, T. R., y Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, *136*, 430–450.
- Kumar, A., Rao, V. R., y Soni, H. (1995). An empirical comparison of neural network and logistic regression models. *Marketing Letters*, *6*, 251–263.
- Lafuente, V., y Romo, R. (2003). Decisions arising from opposing views. *Nature Neuroscience*, *6*, 792–793.
- Lagnado, D. A., y Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 856–876.
- Lawrence, A., Clark, L., Labuzetta, J., Sahakian, B., y Vyakarnum, S. (2008, Noviembre 13). The innovative brain. *Nature*, *456*, 168–169.
- Lee, D. (2005). Neuroeconomics: making risky choices in the brain. *Nature Neuroscience*, *8*, 1129–1130.
- Lee, S. M., Abbott, P., y Johantgen, M. (2005). Logistic regression and bayesian networks to study outcomes using large data sets. *Nursing Research*, *2*, 133–138.
- León, O. G. (1994). *Análisis de decisiones. Técnicas situacionales aplicables a directivos y profesionales*. Madrid: McGraw-Hill.
- Lieberman, E., Michel, J. B., Jackson, J., Tang, T., y Nowak, M. A. (2007, Octubre 11). Quantifying the evolutionary dynamics of language. *Nature*, *449*, 713–716.
- Light, P. C. (2005, Noviembre). *Searching for social entrepreneur: who they might be, where they might be found, what they do*. Trabajo presentado en la Annual Meetings of the Association for Research on Nonprofit and Voluntary Associations. Washington.
- Lilienfeld, D. E., y Pyne, D. A. (1984). The logistic analysis of epidemiologic prospective studies: investigation by simulation. *Statistics in Medicine*, *3*, 15–26.
- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: SAGE Publications.
- López, I., y Píta, S. (1998). Curvas ROC. *Cuadernos de Atención Primaria*,

- 5(4), 229–235.
- López, J., De la Fuente, M. L., García, J., y Ortega, E. M. (2008, Julio). *Bayesian networks applied to prediction of verdict tendency in mock juries*. Comunicación presentada en el III European Congress of Methodology. Oviedo.
- López, J., y García, J. (2007a, Marzo). *A bayes net inside an intelligent tutoring system. How does it work?* Comunicación presentada en la International Technology, Education and Development Conference (INTED 2007). Valencia.
- López, J., y García, J. (2007b, Marzo). *Intelligent tutoring systems with bayesian nets*. Comunicación presentada en la International Technology, Education and Development Conference (INTED 2007). Valencia.
- López, J., y García, J. (2007c). Modelado causal predictivo basado en redes bayesianas. *Actas de las XXVIII Jornadas de Automática*, 1381(1)–1381(6).
- López, J., y García, J. (2007d). Valores, actitudes y comportamiento ecológico modelados con una red bayesiana. *Medio Ambiente y Comportamiento Humano*, 8, 159–175.
- López, J., y García, J. (2008). Sistemas de tutorización inteligente basados en redes bayesianas. *Revista Electrónica de Metodología Aplicada*, 13, 13–25.
- López, J., y García, J. (2009). Asimetría en el razonamiento causal bayesiano bajo incertidumbre. *Boletín de Psicología*, 95, 43–58.
- López, J., García, J., Cano, C. J., Gea, A. B., y De la Fuente, L. (2009, Septiembre). *A definition of potential entrepreneur from a probabilistic point of view*. Comunicación presentada en el XI Congreso de Metodología de las Ciencias Sociales y de la Salud. Málaga.
- López, J., García, J., De la Fuente, L., y De la Fuente, E. I. (2007). Las redes bayesianas como herramienta de modelado en psicología. *Anales de Psicología*, 23, 307–316.
- López, J., García, J., y De la Fuente, E. I. (2005, Septiembre). *Las redes bayesianas en psicología*. Comunicación presentada en el IX Congreso de Metodología de las Ciencias Sociales y de la Salud. Granada.
- López, J., García, J., y De la Fuente, L. (2006). Modelado causal con redes bayesianas. *Actas de las XXVII Jornadas de Automática*, 198–202.
- López, J., Ruiz-Ruano, A. M., y García, J. (2008a, Noviembre). *Competence sequential assessment methodology in higher education modelled with a naïve*

## Referencias

- bayesian classifier*. Comunicación presentada en la International Conference of Education, Research and Innovation (ICERI 2008). Madrid.
- López, J., Ruiz-Ruano, A. M., y García, J. (2008b, Noviembre). *Relationship between self-assessment and marks in higher education: linear, logistic and bayesian analysis*. Comunicación presentada en la International Conference of Education, Research and Innovation (ICERI 2008). Madrid.
- Loufti, M. F. (1991). Configuración y postulados básicos del trabajo por cuenta propia en Europa. *Revista Internacional del Trabajo*, 110(2), 153–174.
- Lumpkin, G. T. (2006). Intrapreneurship and innovation. En J. R. Baum, M. Frese, y R. Baron (Eds.), *The psychology of entrepreneurship* (pp. 237–263). Mahwah, NJ: Lawrence Erlbaum.
- Ma, W. J., Beck, J. M., Latham, P. E., y Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9, 1432–1438.
- Maciá, M. A., Barbero, M. I., Pérez-Llantada, M. C., y Vila, E. (1990). *Psicología y teoría de la decisión: aplicaciones*. Madrid: Universidad Nacional de Educación a Distancia.
- Mani, S., McDermott, S., y Valtorta, M. (1997). MENTOR: a bayesian model for prediction of mental retardation in newborns. *Research in Developmental Disabilities*, 18, 303–318.
- Markram, H. (2006). The blue brain project. *Nature Reviews Neuroscience*, 7, 153–160.
- Marris, E. (2007). How 'holp' became 'helped'. *Nature News*. Descargado el 23 de Marzo, 2009, desde <http://www.nature.com/news/2007/071011/full/news.2007.152.html>.
- Martin, J., y VanLehn, K. (1995). Student assessment using bayesian nets. *International Journal of Human-Computer Studies*, 42, 575–591.
- Martínez, A., Martínez, S., y Martínez, H. (2002). *Estadística empresarial*. Almería: Servicio de Publicaciones de la Universidad de Almería.
- Martínez, F., Díaz, M. C., Martín, M. T., Rivas, V. M., y Ureña, L. A. (2003, Septiembre). *Aplicación de redes neuronales y redes bayesianas en la detección de multipalabras para tareas IR*. Artículo presentado en las II Jornadas de Tratamiento y Recuperación de la Información. Madrid.
- Martínez, I., y Rodríguez, C. (2003). Modelos gráficos. En Y. del Águila et al. (Eds.), *Técnicas estadísticas aplicadas al análisis de datos* (pp. 217–257).

- Almería: Servicio de Publicaciones de la Universidad de Almería.
- Matute, H. (2002). Introducción: aprendizaje de relaciones causales. *Cognitiva*, 14, 7–14.
- McClelland, D. (1961). *The achieving society*. New York: Free Press.
- McCoy, A., y Platt, M. L. (2005). Risk-sensitive neurons in macaque posterior cingulate cortex. *Nature Neuroscience*, 8, 1220–1227.
- McCraw, T. K. (2006). Schumpeter’s business cycles as business history. *Business History Review*, 80, 231–261.
- McFerrin, B. (1988). Don’t worry, be happy. En *Don’t worry, be happy* [LP]. Londres: EMI.
- McKenzie, B., Ugbah, S., y Smothers, N. (2007). “Who is an entrepreneur” is still the wrong question? *Academy of Entrepreneurship Journal*, 13, 23–43.
- Mehta, C. R., y Patel, N. R. (1995). Exact logistic regression: theory and examples. *Statistic in Medicine*, 14, 2143–2160.
- Michotte, A. E. (1946). *La perception de la causalité*. Louvain: L’Institut Supérieur de Philosophie.
- Millán, E., Perez-de-la-Cruz, J. L., y Suárez, E. (2002). Adaptive bayesian networks for multilevel student modelling. En G. Gauthier, C. Frasson, y K. VanLehn (Eds.), *Lecture notes in computer science 1839* (pp. 534–543). Berlin: Springer-Verlag.
- Mislevy, R. J., y Gitomer, D. H. (1996). The role of probability-based inference in an intelligent tutoring system. *User-Mediated and User-Adapted Interaction*, 128, 253–282.
- Monserrat, J. (1998). *La percepción visual: la arquitectura del psiquismo desde el enfoque de la percepción visual*. Madrid: Biblioteca Nueva.
- Moore (Productor), L., y Burke (Escritor/Director), M. (1999). *Pirates of Silicon Valley* [Película de vídeo]. Estados Unidos: Warner Home Video / TNT Original.
- Morales, J. F. (1994). Procesos de atribución. En J. F. Morales (Ed.), *Psicología social* (pp. 239–252). Madrid: McGraw-Hill.
- Morales, J. F., Reboloso, E., y Moya, M. (1994). Actitudes. En J. F. Morales (Ed.), *Psicología social* (pp. 495–524). Madrid: McGraw-Hill.
- Morales, M. E. (2006). *Modelización y predicción en estadística universitaria*. Tesis doctoral no publicada, Facultad de Ciencias Experimentales, Univer-

## Referencias

- alidad de Almería.
- Morillas, J. J. (2009). *Los nuevos yacimientos de empleo entre las estrategias para el apoyo a emprendedores en el marco de los programas de desarrollo local: situación actual en Andalucía*. Tesis doctoral no publicada, Departamento de Economía Aplicada, Universidad de Almería.
- Nadkarni, S., y Shenoy, P. P. (2001). A bayesian network approach to making inferences in causal maps. *European Journal of Operational Research*, 128, 479–498.
- Nadkarni, S., y Shenoy, P. P. (2004). A causal mapping approach to constructing bayesian networks. *Decision Support Systems*, 38, 259–281.
- Narayan, S., y Jurafsky, D. (1998, Agosto). *Bayesian models of human sentence processing*. Comunicación presentada en la XX Annual Meeting of the Cognitive Science Society. Madison.
- Narayan, S., y Jurafsky, D. (2002). A bayesian model predicts human parse preference and reading times in sentence processing. *Advances in Neural Information Processing*, 14, 59–65.
- Neapolitan, E. R., y Morris, S. (2004). Probabilistic modelling with bayesian networks. En D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 371–390). Thousand Oaks, CA: SAGE Publications.
- Ng, A. Y., y Jordan, M. I. (2002). On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. *Advances in Neural Information Processing Systems*, 14, 841–848.
- Norsys Software. (2004). Netica Tutorials. Avanced Topics. Descargado el 4 de Junio de 2004, desde <http://www.norsys.com/>.
- Notterman, J. M. (2004). Persistent conceptual issues in psychology. A selective update. *Theory and Psychology*, 14, 239–260.
- Oatley, G. C., y Ewart, B. W. (2002, Septiembre). *Constructing a bayesian belief network to determine the likelihood of burglary*. Poster presentado en la V International Conference on Forensic Statistic. Venecia.
- Oatley, G. C., y Ewart, B. W. (2003). Crimes analysis software: ‘pins in maps’, clustering and bayes net prediction. *Expert Systems with Applications*, 25, 569–588.
- O’Hora, D., Barnes, D., Roche, B., y Smeet, P. (2004). Derived relational networks

- and control by novel instructions: a possible model of generative verbal responding. *The Psychological Record*, 54, 437–460.
- Olea, J., y Ponsoda, V. (1996). Tests adaptativos informatizados. En J. Muñiz (Ed.), *Psicometría* (pp. 729–783). Madrid: Universitas.
- Ortega, M., y Cayuela, A. (2002). Regresión logística no condicionada y tamaño de muestra: una revisión bibliográfica. *Revista Española de Salud Pública*, 76, 85–93.
- Ortiz, A. J., Martín, M. T., Ureña, L. A., y García, M. A. (2005). Detección automática de SPAM usando regresión logística bayesiana. *Procesamiento del Lenguaje Natural*, 35, 127–133.
- Ortiz, I. M., y Oña, I. (2003). El modelo de regresión lineal simple. En Y. del Águila et al. (Eds.), *Técnicas estadísticas aplicadas al análisis de datos* (pp. 119–142). Almería: Servicio de Publicaciones de la Universidad de Almería.
- Overmier, J. B., y LoLordo, V. M. (1998). Learned helplessness. En W. O'Donohue (Ed.), *Learning and behavior therapy* (pp. 352–373). Needham Heights, MA: Allyn & Bacon.
- Padilla, J. H., Lara, A., y Márquez, P. R. (2001, Octubre). *Sistema inteligente para la enseñanza de las matemáticas*. Artículo presentado en el XXIII Congreso Internacional de Ingeniería Electrónica ELECTRO. México.
- Palop, J. J., Chin, J., y Mucke, L. (2006, Noviembre 19). A network dysfunction perspective on neurodegenerative diseases. *Nature*, 443, 468–773.
- Parasuraman, R. (2000). The attentive brain. Issues and prospects. En R. Parasuraman (Ed.), *The attentive brain* (pp. 3–15). Cambridge, MA: MIT Press.
- Pavlov, I. P. (1927). *Conditioned reflexes*. Londres: Oxford University Press.
- Pearl, J. (2000). *Causality. Models, reasoning and inference*. New York: Cambridge University Press.
- Pearl, J. (2001). *Bayesian networks, causal inference and knowledge discovery* (Rep. Téc. R-281). Los Angeles, CA: University of California.
- Pearl, J. (2002). Causal inference in the health sciences: a conceptual introduction. *Health Services and Outcomes Research Methodology*, 2, 189–220.
- Peñas, M. J., y Quijano, J. (2008, Abril). *¿Es posible fomentar el carácter emprendedor desde la universidad? Un diseño de la asignatura «Empresa Familiar»*. Comunicación presentada en el Congreso Internacional de Em-

## Referencias

- prendedores Ciudad de Salamanca. Salamanca.
- Peduzzi, P., Concato, J., Feinstein, A. R., y Holford, T. R. (1995). Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *Journal of Clinical Epidemiology*, *48*, 1503–1510.
- Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., y Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, *49*, 1373–1379.
- Peng, C. Y. J., y So, T. S. H. (2002). Logistic regression analysis and reporting: a primer. *Understanding Statistics*, *1*, 31–70.
- Perales, J. C., Catena, A., y Maldonado, A. (2002a). Aprendizaje de relaciones de contingencia y causalidad: hacia una análisis integral del aprendizaje causal desde una perspectiva computacional. *Cognitiva*, *14*, 15–41.
- Perales, J. C., Catena, A., y Maldonado, A. (2002b). Tesis, antítesis y síntesis del aprendizaje causal (o defensa de un modelo integrador). *Cognitiva*, *14*, 75–93.
- Platt, M. L. (2004). Unpredictable primates and prefrontal cortex. *Nature Neuroscience*, *7*, 233–238.
- Platt, M. L., y Glimcher, P. W. (1999, Julio 15). Neural correlates of decision variables in parietal cortex. *Nature*, *400*, 233–238.
- Polopolus, L. C., y Emerson, R. D. (1991). Entrepreneurship, sanctions, and labor contracting. *Southern Journal of Agricultural Economics*, *12*, 57–68.
- Pérez, C. (2005). *Métodos estadísticos avanzados con SPSS*. Madrid: Thomson.
- Quinlan, P. (1991). *Connectionism and psychology: a psychological perspective on new connectionist research*. Hertfordshire: Cambridge University Press.
- Ragland, D. R. (1992). Dichotomizing continuous outcome variables: dependence of the magnitude of association and statistical power of the cutpoint. *Epidemiology*, *3*, 434–440.
- Raiffa, H. (1994). The prescriptive orientation of decision making: a synthesis of decision analysis, behavioral decision making, and game theory. En S. Ríos (Ed.), *Decision theory and decision analysis: trends and challenges* (pp. 3–13). Norwell, MA: Kluwer Academic Press.
- Reade, S., y Kupper, L. L. (1995). On the effects of predictor misclassification in multiple linear regression analysis. *Communications in Statistics: Theory*

- and Methods*, 24, 13–37.
- Rescorla, R. A., y Wagner, A. R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. En A. H. Black y W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.
- Richerson, P. J., y Boyd. (2008, Diciembre 18). Migration: an engine for social change. *Nature*, 456, 877.
- Ríos, S. (1995). *Modelización*. Madrid: Alianza Universidad.
- Rodríguez-Arias, M., Aguilar, M., Espert, R., y Miñarro, J. (1996). La esquizofrenia: una perspectiva psicobiológica. *Revista de Neurología*, 24, 1210–1220.
- Rogoff, E. G., y Lee, M. S. (1996). Does firm origin matter? An empirical examination of types of small business owners and entrepreneurs. *Academy of Entrepreneurship Journal*, 1, 1–17.
- Romo, R., y Salinas, E. (2003). Flutter discrimination: neural codes, perception, memory and decision making. *Nature Reviews Neuroscience*, 4, 203–217.
- Ross, T., Wettig, H., Grünwald, P., Myllymäki, P., y Tirri, H. (2005). On discriminative bayesian networks classifiers and logistic regression. *Machine Learning*, 59, 267–296.
- Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcements. *Psychological Monographs*, 80, 1–28.
- Ruiz, F., García, M. E., y Pérez, A. (2005). *Estilos de vida en ciudad de la Habana. Hábitos físico-deportivos y de salud*. Madrid: Gymnos.
- Ruiz, J. M. (1994). *La memoria humana : función y estructura*. Madrid: Alianza.
- Ruiz-Ruano, A. M., y López, J. (2008a). El reto de la educación y la formación intercultural en la migración de retorno. En E. Soriano y A. J. González (Eds.), *Ciudadanía, democracia y participación* (pp. 313–324). Almería: Editorial Universidad de Almería.
- Ruiz-Ruano, A. M., y López, J. (2008b). *Retorno fácil. Análisis de la percepción de los emigrantes y retornados españoles de las barreras psicosociales y socio-políticas del retorno*. Almería: Artes Gráficas SEMAR.
- Rustichini, A. (2005, Diciembre 9). Emotion and reason in making decisions. *Science*, 310, 1624–1625.
- Salinas, E. (2006). Noisy neurons can certainly compute. *Nature Neuroscience*, 9, 1349–1350.

## Referencias

- Samuelson, P. A. (1970). *Economics* (8<sup>a</sup> ed.). New York: McGraw-Hill.
- Sara, S. J. (2000). Retrieval and reconsolidation: toward a neurobiology of remembering. *Learning and Memory*, 7, 73–84.
- Scheines, R., Spirtes, P., Glymour, C., Meek, C., y Richardson, T. (2005). TETRAD 3: tools for causal modeling. User's manual. Descargado el 14 de Febrero de 2005, desde <http://www.phil.cmu.edu/projects/tetrad/>.
- Scooter. (1995). Move your ass! En *...and the beat goes on!* [CD]. Hamburgo: Club Tools.
- Seldmeier, P., y Gigerenzer, G. (2001). Teaching bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, 130, 380–400.
- Serrano, J. (2003). *Iniciación a la estadística bayesiana*. Madrid: Muralla / Hespérides.
- Shadlen, M., y Newsome, W. T. (1996). Motion perception: seeing and deciding. *Proceedings of Natural Academic Sciences*, 93, 628–633.
- Sharkey, N. (2008, Diciembre 19). The ethical frontiers of robotics. *Science*, 322, 1800–1801.
- Shen, B., Su, X., Greiner, R., Musilek, P., y Cheng, C. (2003, Noviembre). *Discriminative parameter learning of general bayesian network classifiers*. Comunicación presentada en la 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI-03). Sacramento, California.
- Silva, L. C., y Barroso, I. M. (2004). *Regresión logística*. Madrid: La Muralla / Hespérides.
- Sánchez, M. L. (2003). *El perfil psicológico del autoempleado*. Tesis doctoral publicada en edición electrónica, Facultad de Psicología, Universidad Complutense de Madrid.
- Sobel, D. M., Tenenbaum, J. B., y Gopnik, A. (2004). Children's causal inferences from indirect evidence: backwards blocking and bayesian reasoning in preschoolers. *Cognitive Science*, 28, 303–333.
- Sofroniou, N., y Hutcheson, G. D. (2002). Confidence intervals for the predictions of logistic regression in the presence and absence of a variance-covariance matrix. *Understanding Statistics*, 1, 3–18.
- Spirtes, P., Glymour, C., y Scheines, R. (2000). *Causation, prediction and search* (2<sup>a</sup> ed.). Cambridge, MA: MIT Press.
- SPSS. (2006). *Manual del usuario de SPSS base 15.0*. Chicago, IL: SPSS Incor-

- porated.
- SPSS, y Recognition Systems. (1997). *Neural connection 2. 0. User's guide*. Chicago, IL: SPSS y Recognition Systems.
- Squire, L. R., Knowlton, B., y Musen, G. (1993). The structure and organization of memory. *Annual Review of Psychology*, *44*, 553–495.
- Stanworth, J., Stanworth, C., Granger, B., y Blyth, S. (1989). Who becomes an entrepreneur? *International Small Business Journal*, *8*(1), 11–22.
- Sugrue, L. P., Corrado, G. S., y Newsome, W. T. (2005). Choosing the greater of two goods: neural currencies for valuation and decision making. *Nature Reviews Neuroscience*, *6*, 363–375.
- Tapia, J. A. (1995). Respuesta a: “Razón relativa” y “tasa relativa” como traducciones de *odds ratio* y *hazard ratio*. *Boletín de la Oficina Sanitaria Panamericana*, *119*, 280–282.
- Terrin, N., Schmid, C. H., Griffith, J. L., D’Agostino, R., y Selker, H. P. (2003). External validity of predictive models: A comparison of logistic regression, classification trees, and neural networks. *Journal of Clinical Epidemiology*, *56*, 721–729.
- Tutte, W. T. (1984). *Graph theory*. Cambridge: Cambridge University Press.
- Tversky, A., y Kahneman, D. (1974, Septiembre 27). Judgement under uncertainty: heuristic and biases. *Science*, *185*, 380–400.
- Tversky, A., y Kahneman, D. (1982). Causal schemas in judgements under uncertainty. En D. Kahneman, P. Slovic, y A. Tversky (Eds.), *Judgement under uncertainty: heuristic and biases* (pp. 117–128). New York: Cambridge University Press.
- Uña, I., Tomeo, V., y San Martín, J. (2003). *Lecciones de cálculo de probabilidades*. Madrid: Thomson.
- Urretavizcaya, M. (2001). Monografía: sistemas inteligentes en el ámbito de la educación. *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*, *12*, 2–4.
- Valencia, J. F. (1994). Racionalidad y comportamiento. En J. F. Morales (Ed.), *Psicología social* (pp. 609–621). Madrid: McGraw-Hill.
- Valls, F. (1996). *Programa autoaplicable de asesoriamento vocacional*. Almería: Editorial Universidad de Almería.
- Veciana, J. M. (1989). Características del empresario en España. *Papeles de*

## Referencias

- Economía Española*, 39, 19–36.
- Wald, C. (2008, Diciembre 18). Crazy money. *Science*, 322, 1624–1626.
- Waldman, M. R., y Hagmayer, Y. (2005). Seeing versus doing: two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 216–227.
- Walker, S. H., y Duncan, D. B. (1967). Estimation of the probability of an event as function of several independent variables. *Biometrika*, 54, 167–179.
- Weese, M. R., y Zador, A. (2006, Febrero 23). Efficiency measures. *Nature*, 439, 920–921.
- Weiner, I. (2000). The latent inhibition model of schizophrenia. En M. Myslobodsky y I. Weiner (Eds.), *Contemporary issues in modeling psychopathology* (pp. 197–230). Boston, MA: Kluwer Academic Press.
- Whittermore, A. S. (1981). Sample size for logistic regression with small response probability. *Journal of American Statistical Association*, 76, 27–32.
- Widjaja, I. (2000). Fractal neural network. Descargado el 7 de Agosto de 2003, desde <http://www.comp.nus.edu.sg/inns/IvoWidjaja/report.htm>.
- Wild, J. (2005, Noviembre 24). Software shakes up schizophrenia diagnosis. *Nature*, 438, 407.
- Wilde, O. (1998). *An ideal husband*. Genova: Black Cat. (Trabajo original publicado en 1895)
- Wong, R. (2008, Noviembre). *An evaluation of generic skill outcomes in engineering education*. Comunicación presentada en la International Conference of Education, Research and Innovation (ICERI 2008). Madrid.
- Xiang, Y. (2002). *Probabilistic reasoning in multiagent systems. A graphical models approach*. Cambridge: Cambridge University Press.
- Yanou. (2008). Children of the sun. En *Children of the sun* [CD]. Köln: Zoo-groove.
- Yarandi, H., y Simpson, S. H. (1991). The logistic regression model and the odds of testing HIV positive. *Nursing Research*, 40, 372–373.
- Yu, A., y Dayan, P. (2003). Expected and unexpected uncertainty: ACh and NE in the neocortex. *Advances in Neural Information Processing Systems*, 15, 157–164.
- Yu, A., y Dayan, P. (2005a). Inference, attention, and decision in a bayesian neural architecture. *Advances in Neural Information Processing Systems*,

- 17, 1577–1584.
- Yu, A., y Dayan, P. (2005b). Uncertainty, neuromodulation, and attention. *Neuron*, 46, 681–692.
- Zhao, L. P., y Kolonel, L. (1992). Efficiency loss from categorizing quantitative exposures into qualitative exposures in case-control studies. *American Journal of Epidemiology*, 136, 464–474.