Aplicación de redes neuronales artificiales al tratamiento de datos incompletos

José Blas Navarro Pastor

Tesis doctoral dirigida por los Doctores:
Josep Maria Losilla Vidal
Lourdes Ezpeleta Ascaso

Departament de Psicobiologia i Metodologia
de les Ciències de la Salut

Facultat de Psicologia

Universitat Autònoma de Barcelona

1998

6
Simulación estadística

De los métodos de análisis con datos incompletos revisados en la primera parte, se ha seleccionado un amplio subconjunto para ser evaluados mediante simulación estadística. La simulación planteada consiste en obtener matrices de datos incompletas con diferentes tipos de variables y varios niveles de covariación entre ellas, y aplicar las diferentes técnicas de análisis seleccionadas para estimar, finalmente, un posterior modelo de clasificación.

A un primer nivel, los procedimientos estudiados se diferencian en el hecho de que estén o no basados en la imputación de un valor a cada missing, entendiendo por imputación la asignación de un valor obtenido mediante algún sistema que requiera de la información presente en la matriz de datos, y no la mera codificación de los datos missing a un determinado valor elegido arbitrariamente.

6.1. OBJETIVOS

Mediante el estudio de simulación se pretende cubrir los siguientes objetivos:

1. Comparar diferentes métodos de imputación de valor a los datos missing, distinguiendo entre métodos de imputación directa (media, moda, mediana, valor aleatorio de distribución de probabilidad uniforme y valor aleatorio de distribución de probabilidad estimada para la variable) y de imputación por regresión/red a partir de otras variables de la matriz (regresión, EM, *Best estimate* con redes perceptrón multicapa -MLP- y *Best estimate* con redes de función de base radial -RBF-). Estableciendo conclusiones diferenciadas según la naturaleza de la variable que contiene los valores missings a imputar (categórica con dos o más categorías, o cuantitativa), según la forma de su distribución, y según el nivel de covariación entre las variables de la matriz.
2. Estudiar el efecto sobre la imputación por regresión/red a una variable, de la previa imputación directa de los valores faltantes en las variables independientes. Este análisis es importante porque las matrices de datos habitualmente tienen valores desconocidos en más de una variable, y para imputar por regresión/red los de una de ellas es necesario haber imputado previamente los de las variables que actúan como independientes. Además, cuando en una matriz de datos hay una variable de especial relevancia, es interesante conocer cómo se ve afectada la imputación de sus valores perdidos por la presencia de valores perdidos en otras variables.

3. Analizar cómo influye la imputación de los datos faltantes de las variables independientes en la capacidad discriminante de un modelo de clasificación, distinguiendo entre diferentes técnicas de clasificación (regresión logística, redes neuronales MLP, redes neuronales RBF). Comparar estos resultados con los obtenidos mediante métodos de análisis de valores ausentes con los que se realiza directamente la estimación de los parámetros del modelo de clasificación, sin imputar previamente un valor a los datos desconocidos (*listwise* y codificación de missings con inclusión de variables indicadoras). Para poder calcular la aportación neta del uso de variables indicadoras, sin la distorsión provocada por la asociada codificación del valor faltante, se ha realizado también el análisis que incluye únicamente la codificación de los valores desconocidos.
4. Contrastar la hipótesis relativa a la superior capacidad de las redes neuronales artificiales en problemas de clasificación, en comparación con técnicas estadísticas convencionales como la regresión logística, así como evaluar la influencia del método de análisis de datos faltantes en la diferencia en el error de clasificación entre el modelo de red neuronal y el modelo de regresión.
5. Investigar cómo se ve afectada la capacidad de clasificación por el porcentaje de valores faltantes de los datos.
6. A partir de las conclusiones extraídas de los resultados en los apartados previos, elaborar un conjunto de recomendaciones para guiar la elección del procedimiento más óptimo para el tratamiento de los datos missing de las variables independientes en un problema de clasificación, distinguiendo en función de la magnitud de la covariación entre las variables independientes, y entre éstas y la variable dependiente.

6.2. MÉTODO

6.2.1. Diseño de la simulación

6.2.1.1. Matrices con datos completos

Inicialmente se generaron un total de 300 matrices con datos completos (matrices MC1-MC300). Cada matriz tiene 500 registros y un total de 7 variables, obtenidas a partir de una distribución normal multivariante con parámetros $\mu=0$, $\sigma=1$ en cada variable y diferentes niveles de covariación entre ellas. A partir de las 7 variables cuantitativas originales se generaron los 7 tipos de variables que fueron analizados:

a) Una variable binaria con distribución equiprobable (BU): la primera variable de cada matriz se dicotomizó en el valor 0 (que se corresponde al percentil 50

en el mecanismo generador), obteniéndose así una variable binaria procedente de una distribución binomial con π=0.5.

b) Una variable binaria con distribución no equiprobable (BA): la segunda variable de cada matriz se dicotomizó en el valor 0.8416 (que se corresponde al percentil 80 en el mecanismo generador), obteniéndose una variable binaria procedente de una distribución binomial con π=0.8.

c) Una variable ordinal con distribución equiprobable (OU): la tercera variable de cada matriz se recodificó en 3 categorías a partir de los valores -0.4316 (percentil 33.3 en el mecanismo generador) y 0.4316 (percentil 66.6 en el mecanismo generador), consiguiendo una variable procedente de una distribución multinomial con $\pi_1=\pi_2=\pi_3=1/3$.

d) Una variable ordinal con distribución no equiprobable (OA): la cuarta variable de cada matriz se recodificó en 3 categorías a partir de los valores 0 (percentil 50 en el mecanismo generador) y 1.0364 (percentil 85 en el mecanismo generador), consiguiendo una variable procedente de una distribución multinomial con $\pi_1=0.5$, $\pi_2=0.35$, $\pi_3=0.15$. La elección de los parámetros de la distribución multinomial posiblemente influirá en las conclusiones sobre el análisis de valores missing de esta variable, pero puesto que había que tomar una decisión, se optó por elegir una distribución no excesivamente desequilibrada.

e) Una variable cuantitativa con distribución normal (CN): la quinta variable de cada matriz no se transformó, ya que su mecanismo generador ya es una distribución normal con μ=0, σ=1.

f) Una variable cuantitativa con distribución asimétrica (CL): a la sexta variable de cada matriz se le aplicó la función exponencial, dando como resultado una variable cuantitativa con distribución asimétrica, cuyos estadísticos descriptivos básicos, promediados en las 300 matrices, se hallan en la Tabla 18:

Tabla 18. Indices descriptivos de la variable CL promediados en las 300 matrices

	Media	Mediana	Desv.Est.	Asimetría	E.E. Asimetría
CL	1.65	1.00	2.16	4.78	0.109

g) Una variable binaria con distribución equiprobable (DEP): la séptima variable de cada matriz se dicotomizó en el valor 0 (percentil 50 en el mecanismo generador), obteniéndose una variable binaria procedente de una distribución binomial con π=0.5.

Las seis primeras variables (BA, BU, OA, OU, CN, CL) tendrán el rol de variables independientes en los posteriores modelos de clasificación, mientras que la séptima variable (DEP) se utilizará como variable dependiente.

Respecto al nivel de covariación entre las 7 variables, de las 300 matrices con datos completos generadas:

- Las matrices MC1-MC100 se generaron utilizando parámetros de covariación nula entre las siete variables. La media del coeficiente de correlación de Spearman entre cada par de variables en las 100 matrices se halla en la Tabla 19:

Tabla 19. Correlación promedio en las matrices MC1-MC100 entre cada par de variables (correlación de Spearman).

	BA	OU	OA	CN	CL	DEP
BU	0.011	-0.001	-0.004	0.003	-0.008	-0.002
BA		-0.003	0.004	0.003	-0.001	0.005
OU			-0.004	0.002	-0.001	0.007
OA				0.003	-0.002	-0.007
CN					-0.001	-0.008
CL						-0.003

- Las matrices MC101-MC200 se generaron utilizando parámetros de covariación media entre las seis variables independientes, y covariación media-alta entre estas seis variables y la variable dependiente. La media del coeficiente de correlación de Spearman entre cada par de variables en las 100 matrices se halla en la Tabla 20:

Tabla 20. Correlación promedio en las matrices MC101-MC200 entre cada par de variables (correlación de Spearman).

	BA	OU	OA	CN	CL	DEP
BU	0.283	0.371	0.356	0.400	0.399	0.418
BA		0.313	0.286	0.340	0.341	0.338
OU			0.396	0.436	0.442	0.450
OA				0.427	0.431	0.438
CN					0.482	0.486
CL						0.482

- Las matrices MC201-MC300 se generaron utilizando parámetros de covariación baja entre las seis variables independientes, pero covariación media entre cada variable independiente y la variable dependiente. La media del coeficiente de correlación de Spearman entre cada par de variables en las 100 matrices se halla en la Tabla 21:

Tabla 21. Correlación promedio en las matrices MC201-MC300 entre cada par de variables (correlación de Spearman).

	BA	OU	OA	CN	CL	DEP
BU	0.052	0.082	0.074	0.076	0.076	0.314
BA		0.070	0.065	0.074	0.063	0.258
OU			0.082	0.091	0.086	0.352
OA				0.088	0.081	0.336
CN					0.097	0.377
CL						0.370

6.2.1.2. Matrices con datos incompletos

A partir de cada matriz con datos completos se generaron 8 nuevas matrices con datos incompletos. La asignación de los valores faltantes se realizó a partir de un procedimiento aleatorio, consistente en generar números aleatorios entre 1 y 500, que representan el número de registro que contendrá el valor missing. Este mecanismo generador de datos ausentes garantiza que los missing sean completamente aleatorios (*MCAR*). Concretamente, las operaciones realizadas en cada matriz con datos completos para generar las 7 nuevas matrices con datos incompletos son:

1. Eliminación de 15 valores en cada una de las 6 variables independientes, originando una nueva matriz con un 3% de valores faltantes en cada una de sus 6 variables independientes (matrices MM1).
2. Eliminación de 15 valores en la variable BU, dando lugar a una nueva matriz con un 3% de valores faltantes en dicha variable y sin missing en las otras variables independientes (matriz MM2_BU).
3. Eliminación de 15 valores en la variable BA, dando lugar a una nueva matriz con un 3% de valores faltantes en dicha variable y sin missing en las otras variables independientes (matriz MM2_BA).
4. Eliminación de 15 valores en la variable OU, dando lugar a una nueva matriz con un 3% de valores faltantes en dicha variable y sin missing en las otras variables independientes (matriz MM2_OU).
5. Eliminación de 15 valores en la variable OA, dando lugar a una nueva matriz con un 3% de valores faltantes en dicha variable y sin missing en las otras variables independientes (matriz MM2_OA).
6. Eliminación de 15 valores en la variable CN, dando lugar a una nueva matriz con un 3% de valores faltantes en dicha variable y sin missing en las otras variables independientes (matriz MM2_CN).

7. Eliminación de 15 valores en la variable CL, dando lugar a una nueva matriz con un 3% de valores faltantes en dicha variable y sin missing en las otras variables independientes (matriz MM2_CL).
8. Eliminación de 60 valores en cada una de las 6 variables independientes, generando una nueva matriz con un 12% de datos faltantes en cada una de sus 6 variables independientes (matrices MM3).

En ningún caso se generaron datos missing en la variable que tiene el rol de dependiente en el problema de clasificación.

6.2.1.3. Imputación de valor a los datos faltantes

Los valores faltantes de las diferentes variables fueron imputados mediante varios métodos de imputación directa y de imputación por regresión/red, calculándose en cada caso el error de imputación cometido como el error promedio en los registros imputados en cada matriz de datos. Para calcular el error de imputación cometido en cada registro, en las variables binarias y ordinales se calculó la tasa nominal de error, considerando el error igual a 0 si el valor imputado coincidía con el valor real, o igual a 1 en caso de no coincidir. En las variables cuantitativas se calculó el error medio cuadrático, promediando en los registros imputados la suma de las diferencias al cuadrado entre el valor imputado y el valor real.

a. Imputación directa

Puesto que el resultado de los métodos de imputación directa no están afectados por la magnitud de la correlación entre las variables de la matriz de datos, estos métodos sólo fueron evaluados en las matrices MM1_1 a MM1_100. Los métodos de imputación directa evaluados en cada variable son:

a) Los valores missing de la variable BU fueron imputados directamente con la moda, un valor aleatorio de una distribución equiprobable (VADU), y un valor aleatorio de una distribución binomial caracterizada por π=0.5 (VADE).

b) Los valores missing de la variable BA fueron imputados directamente con la moda, un valor aleatorio de una distribución equiprobable (VADU), y un valor aleatorio de una distribución binomial caracterizada por π=0.8 (VADE).

c) Los valores missing de la variable OU fueron imputados directamente con la moda, la mediana, un valor aleatorio de una distribución equiprobable (VADU), y un valor aleatorio de una distribución multinomial caracterizada por $\pi_1=\pi_2=\pi_3=1/3$ (VADE).

d) Los valores missing de la variable OA fueron imputados directamente con la moda, la mediana, un valor aleatorio de una distribución equiprobable

(VADU), y un valor aleatorio de una distribución multinomial caracterizada por π_1=0.5, π_2=0.35, π_3=0.15 (VADE).

e) Los valores missing de la variable CN fueron imputados directamente con la media, la mediana, un valor aleatorio de una distribución uniforme (VADU), y un valor aleatorio de una distribución normal μ=0, σ=1 (VADE).

f) Los valores missing de la variable CL fueron imputados directamente con la media, la mediana, un valor aleatorio de una distribución uniforme (VADU), y un valor aleatorio de una distribución asimétrica, obtenida aplicando la función exponencial a una distribución normal μ=0, σ=1 (VADE).

Además, en las 300 matrices MM1 los valores missing de cada variable fueron imputados con el mejor método de imputación directa de cada variable, obtenido a partir de los resultados de las imputaciones previas, generándose una nueva matriz con las 6 variables independientes imputadas con el mejor método de imputación directa (matriz MID), que será posteriormente utilizada en un modelo de clasificación de la variable dependiente.

b. Imputación por regresión/red

Los valores perdidos de cada variable independiente fueron imputados con la predicción realizada con diferentes modelos establecidos a partir del resto de variables independientes. En concreto, se realizaron imputaciones a partir de modelos de regresión (lineal y logística), redes MLP, redes RBF y mediante el algoritmo EM. Para ello, en primer lugar se estimaron los modelos de regresión y se entrenaron las redes neuronales utilizando los registros con datos completos, posteriormente se aplicaron los modelos estimados a los registros con valores faltantes para obtener un valor predicho, que se considera el valor imputado.

En las matrices MM1 hay que tener en cuenta que, en primer lugar, los valores ausentes en las variables independientes del modelo de imputación fueron reemplazados con el mejor método de imputación directa obtenido en el apartado anterior. En las matrices MM2 no fue necesario realizar este primer paso, ya que en dichas matrices sólo contiene datos faltantes una determinada variable (p.ej. en MM2_BU sólo contiene datos missing la variable BU).

Los métodos de imputación por regresión/red aplicados son:

a) En las matrices MM1, los valores missing de la variable BU fueron imputados con regresión logística, red neuronal MLP y red neuronal RBF. Las correspondientes matrices con las seis variables independientes imputadas, cinco por imputación directa y la sexta con regresión logística, red MLP y red RBF, fueron guardadas con los nombres BU_RE, BU_ML, BU_RB respectivamente, para ser posteriormente utilizadas en un modelo de clasificación de la variable dependiente.

b) En las matrices MM1, los valores missing de la variable BA fueron imputados con regresión logística, red neuronal MLP y red neuronal RBF. Las correspondientes matrices con las seis variables independientes imputadas, cinco por imputación directa y la sexta con regresión logística, red MLP y red RBF, fueron guardadas con los nombres BA_RE, BA_ML, BA_RB respectivamente, para ser posteriormente utilizadas en un modelo de clasificación de la variable dependiente.

c) En las matrices MM1, los valores missing de la variable OU fueron imputados con regresión logística, red neuronal MLP y red neuronal RBF. Las correspondientes matrices con las seis variables independientes imputadas, cinco por imputación directa y la sexta con regresión logística, red MLP y red RBF, fueron guardadas con los nombres OU_RE, OU_ML, OU_RB respectivamente, para ser posteriormente utilizadas en un modelo de clasificación de la variable dependiente.

d) En las matrices MM1, los valores missing de la variable OA fueron imputados con regresión logística, red neuronal MLP y red neuronal RBF. Las correspondientes matrices con las seis variables independientes imputadas, cinco por imputación directa y la sexta con regresión logística, red MLP y red RBF, fueron guardadas con los nombres OA_RE, OA_ML, OA_RB respectivamente, para ser posteriormente utilizadas en un modelo de clasificación de la variable dependiente.

e) En las matrices MM1, los valores missing de la variable CN fueron imputados con regresión lineal, red neuronal MLP, red neuronal RBF y algoritmo EM. Las correspondientes matrices con las seis variables independientes imputadas, cinco por imputación directa y la sexta con regresión logística, red MLP, red RBF y algoritmo EM, fueron guardadas con los nombres CN_RE, CN_ML, CN_RB, CN_EM respectivamente, para ser posteriormente utilizadas en un modelo de clasificación de la variable dependiente.

f) En las matrices MM1, los valores missing de la variable CL fueron imputados con regresión lineal, red neuronal MLP, red neuronal RBF y algoritmo EM. Las correspondientes matrices con las seis variables independientes imputadas, cinco por imputación directa y la sexta con regresión logística, red MLP, red RBF y algoritmo EM, fueron guardadas con los nombres CL_RE, CL_ML, CL_RB, CL_EM respectivamente, para ser posteriormente utilizadas en un modelo de clasificación de la variable dependiente.

g) En las matrices MM2_BU, MM2_BA, MM2_OU, MM2_OA, MM2_CN, MM2_CL se realizaron las mismas imputaciones por regresión/red, midiéndose en cada caso el error de imputación cometido, si bien las correspondientes matrices con datos imputados no fueron almacenadas para ser empleadas posteriormente en el modelo de clasificación de la variable

dependiente, ya que en la práctica no es habitual tener una matriz de datos con datos faltantes únicamente en una de sus variables. Las matrices MM2 sirven exclusivamente para evaluar el objetivo 2, y no para evaluar los objetivos que hacen referencia al efecto de la imputación sobre un posterior modelo de clasificación.

El método de imputación *Network reduction* no fue incluido en el análisis por la imposibilidad de ser ejecutado en las matrices de datos generadas. Efectivamente, con seis variables con valores ausentes, el número de patrones de datos missing diferentes es tan elevado que sería necesario entrenar decenas de redes para imputar todos los missing de una sola matriz, y con el problema añadido de que el número de registros disponibles para el aprendizaje de la red sería, en muchos casos, insuficiente.

6.2.1.4. Clasificación

Para estudiar los objetivos 3 y 4 planteados se estimaron diferentes modelos de clasificación de la variable dependiente (DEP), y se midió en cada uno de ellos el error de clasificación como el porcentaje de clasificaciones incorrectas realizadas en una submuestra aleatoria de 50 registros de la matriz de datos, que actúan como datos de test. Se partió de las matrices con datos imputados, almacenadas en el apartado anterior, y de las matrices con datos incompletos MM1, a las que se aplicaron los métodos de análisis *listwise*, codificación de valores missing y codificación de valores missing con inclusión de variables indicadoras.

Concretamente, las matrices de datos en las que se estimaron los diferentes modelos de clasificación son:

a) En las matrices con todas las variables independientes imputadas con el mejor método de imputación directa (matrices MID) se clasificó la variable dependiente (DEP) mediante regresión logística, red neuronal MLP y red neuronal RBF.

b) En las matrices con cinco variables independientes imputadas con el mejor método de imputación directa, y la sexta imputada por regresión (matrices BU_RE, BA_RE, OU_RE, OA_RE, CN_RE, CL_RE) o algoritmo EM (matrices CN_EM, CL_EM) se clasificó la variable dependiente mediante red neuronal MLP y red neuronal RBF. No se incluye en estas matrices la clasificación mediante regresión logística por la alta colinealidad generada al imputar una variable independiente por regresión o EM a partir de las demás.

c) En las matrices con cinco variables independientes imputadas con el mejor método de imputación directa, y la sexta imputada por red MLP y por red RBF (matrices BU_ML, BA_ML, OU_ML, OA_ML, CN_ML, CL_ML, BU_RB, BA_RB, OU_RB, OA_RB, CN_RB, CL_RB) se clasificó la variable

dependiente mediante regresión logística, red neuronal MLP y red neuronal RBF.

d) En las matrices con valores faltantes en todas las variables independientes (MM1), se clasificó la variable dependiente mediante regresión logística, red neuronal MLP y red neuronal RBF a partir de los casos con datos completos (*listwise*).

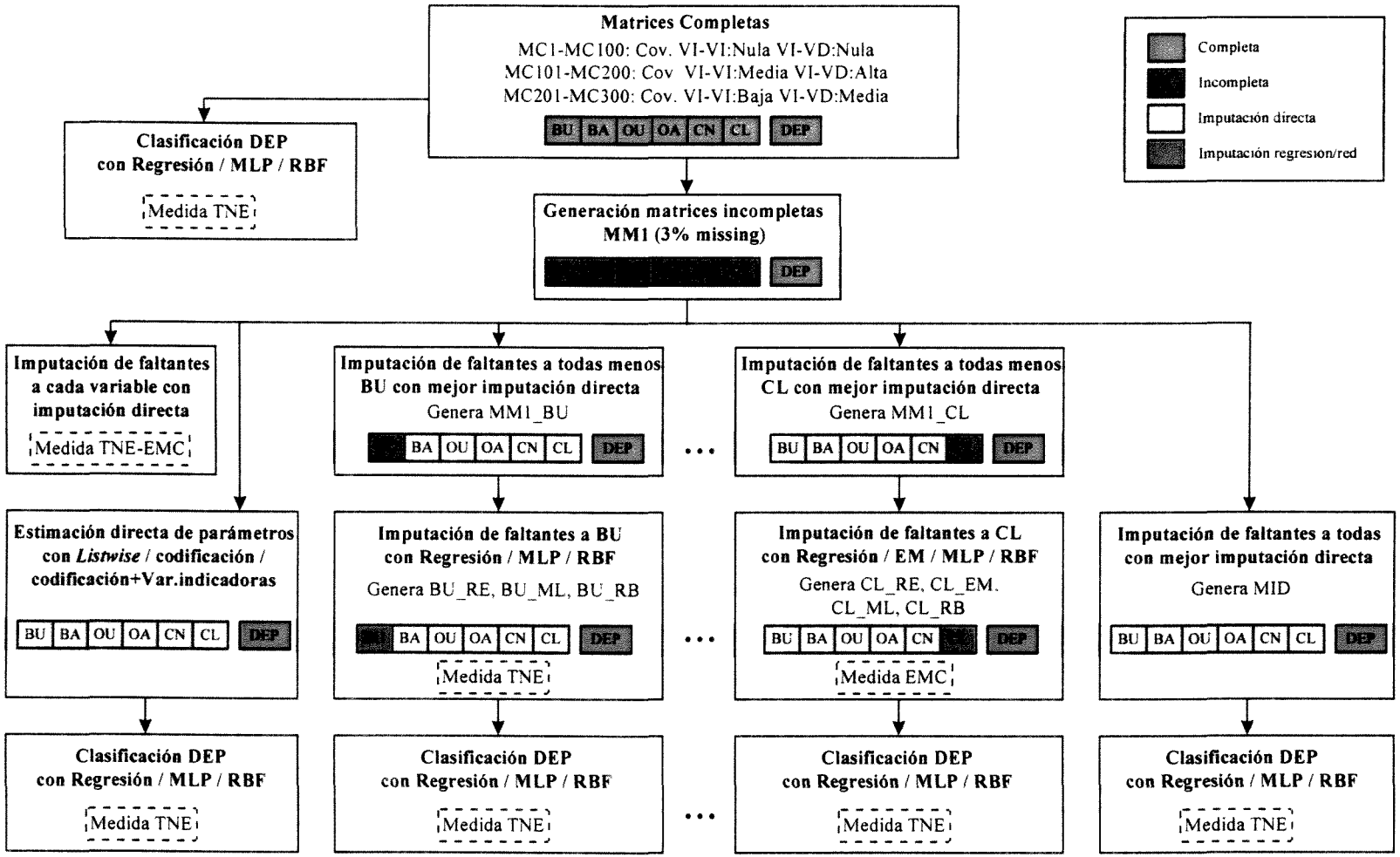
e) En las matrices con valores faltantes en todas las variables independientes (MM1), se codificaron todos los datos missing al valor 99 y se clasificó la variable dependiente mediante regresión logística, red neuronal MLP y red neuronal RBF. Diferentes análisis realizados previamente pusieron de manifiesto que el valor empleado para codificar los datos ausentes (-99, 99, 999) no es relevante en el resultado de la clasificación, siempre y cuando se trate, obviamente, de un valor fuera del rango de valores de la variable.

f) En las matrices con valores faltantes en todas las variables independientes (MM1), se codificaron todos los datos missing al valor 99 y además se generaron seis variables indicadoras, una por cada variable independiente, con el valor uno si el dato es missing y cero en caso contrario. A partir de esta matriz, con un total de doce variables independientes, se clasificó la variable dependiente mediante regresión logística, red neuronal MLP y red neuronal RBF.

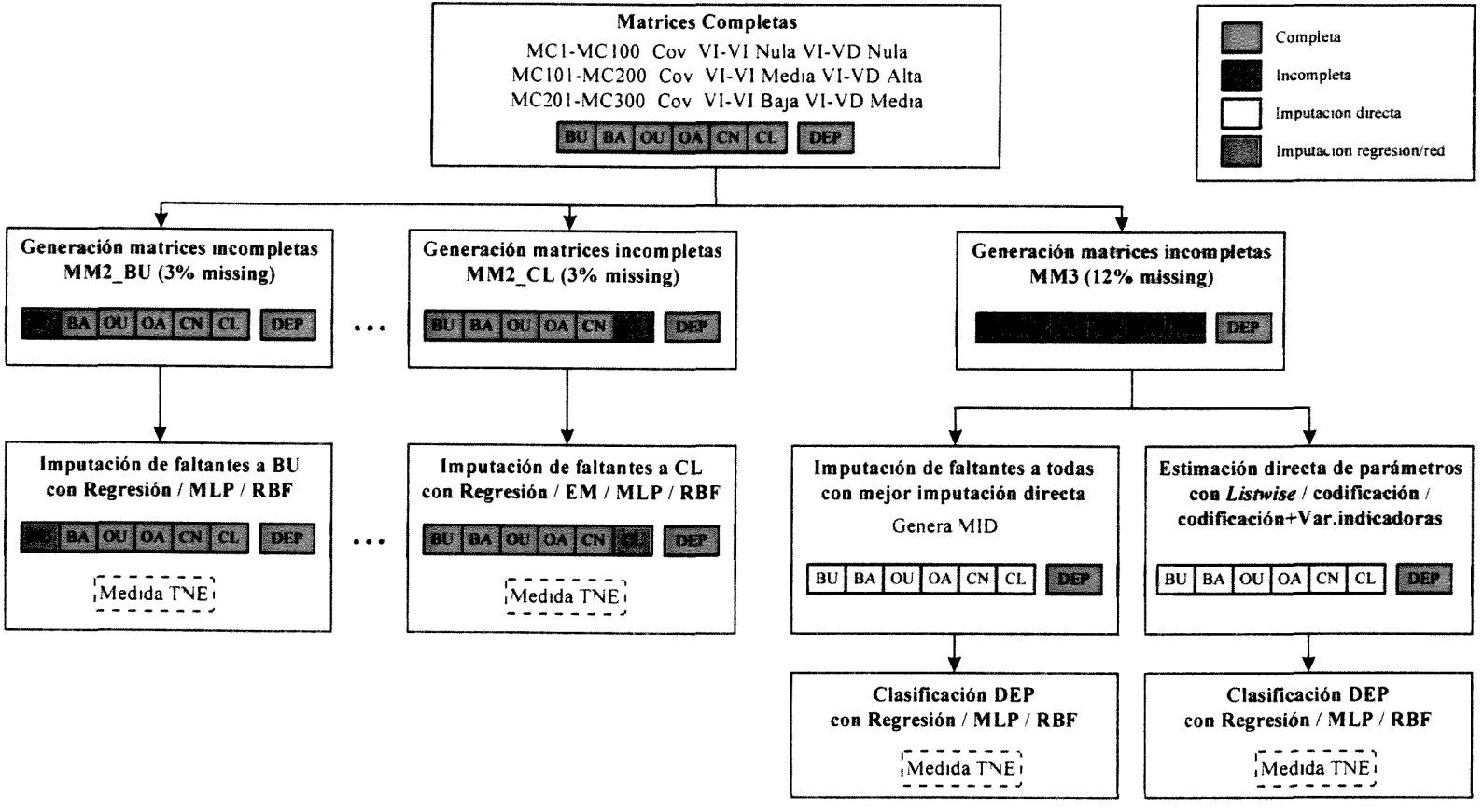
g) En las matrices con datos completos (matrices MC) se clasificó la variable dependiente mediante regresión logística, red neuronal MLP y red neuronal RBF.

Para evaluar el efecto del porcentaje de missings presentes en los datos sobre el resultado de la clasificación (objetivo 5), se estimaron diferentes modelos con las matrices MM3, en las que cada variable independiente contiene un 12% de datos ausentes.

En las siguientes páginas se esquematiza el diseño completo de la simulación realizada.



Diseño de la simulación (1)



Diseño de la simulación (2)

6.2.2. Material

La simulación fue realiza en ordenadores personales equipados con procesadores Pentium II a 233 Mhz. y con 64 Mb de memoria RAM.

La imputación directa de valores faltantes se realizó mediante el módulo de análisis de datos missing de la aplicación SPSS en su versión 7.5.2. para Windows.

La imputación por regresión de los valores faltantes se realizó mediante regresión logística en las variables binarias (BU, BA), regresión logística politómica en las variables ordinales (OU, OA) y regresión lineal en las variables cuantitativas (CN, CL). Para ello se utilizaron las aplicaciones SPSS 7.5.2 (SPSS Inc., 1996) y BMDP 93' (Dixon, 1993).

La imputación mediante el algoritmo EM a las variables cuantitativas (CN, CL) se realizó mediante el módulo de análisis de datos missing del programa SPSS 7.5.2.

La imputación mediante red neuronal MLP y red neuronal RBF se realizó con la aplicación Neural Connection 2.0 (SPSS Inc., 1997). Para decidir la topología de red a emplear en los diferentes apartados de la simulación, se realizaron una serie de pruebas previas con un subconjunto de las matrices disponibles. Partiendo siempre de la topología sugerida por la aplicación Neural Connection, se modificaron tanto el número de unidades ocultas (aumentándolo y disminuyéndolo), como el número de capas de unidades ocultas (una o dos capas). Los resultados demostraron que las diferencias entre las diferentes topologías probadas eran mínimas y no sistemáticas, en el sentido de que en unas matrices el error mejoraba y en otras empeoraba. A raíz de ello, siempre se trabajó con los parámetros para la topología de red sugeridos por la aplicación que se detallan a continuación.

Redes MLP

- *Número de unidades de entrada*: El número total de unidades de entrada de una determinada topología de red es una función del tipo y número de variables independientes. En general, cada variable de entrada cuantitativa se corresponde con una unidad de entrada, y cada variable de entrada categórica se corresponde con tantas unidades de entrada como categorías tenga la variable.
- *Número de unidades y capas ocultas:* El número total de unidades y capas ocultas es una función del número de unidades de entrada y de salida, y de la complejidad de la función a aprender. Las redes neuronales entrenadas tienen una sola capa de unidades ocultas y un número variable de unidades ocultas: desde 4 en las redes más simples hasta 7 en las más complejas, éstas últimas

empleadas sólo en el problema de clasificación mediante la técnica de codificación de los valores missing con inclusión de variables indicadoras.

- *Número de unidades de salida*: El número total de unidades de salida es una función del tipo de variable dependiente. Como en el caso de las unidades de entrada, si la variable dependiente es cuantitativa sólo habrá una unidad de salida, mientras que si la variable dependiente es categórica habrán tantas unidades de salida como categorías tenga la variable.
- *Función de activación*: Para las unidades de entrada la función identidad, para las unidades ocultas la función logística y para las unidades de salida la función identidad.
- *Pesos iniciales*: Valores aleatorios obtenidos de una distribución uniforme en el rango -0.1 a 0.1.
- *Regla de aprendizaje*: El entrenamiento de las redes neuronales se realizó mediante una regla de aprendizaje de segundo orden, denominada "gradiente conjugado". Para acelerar la convergencia, la aplicación Neural Connection realiza el entrenamiento de la red neuronal en 4 fases diferentes. En las fases 2, 3 y 4 se inicia el entrenamiento con la red que ha dado el menor error de generalización en la fase anterior. Los coeficientes de aprendizaje y momento, junto al número de iteraciones (épocas) de cada fase de entrenamiento se hallan en la Tabla 22.

Tabla 22. Características de la regla de aprendizaje de las redes MLP

	Fase 1	Fase 2	Fase 3	Fase 4
Iteraciones	100	100	100	10000
Coef. aprendizaje	0.9	0.7	0.5	0.4
Momento	0.1	0.4	0.5	0.6

Redes RBF:

- *Número de centros*: 5 iniciales posicionados aleatoriamente, con incrementos de 5 en 5 hasta un máximo de 50.
- *Medida de distancia de error*: Distancia euclidea.
- *Función radial*: Función gausiana, con σ^2=0.1

6.3. RESULTADOS

6.3.1. Imputación directa

La media del error cometido con cada método de imputación directa en las seis variables analizadas, y el intervalo de confianza del 95%, se presentan en la Tabla 23.

Tabla 23. Error medio de los métodos de imputación directa en las matrices MM1_1-MM1_100

MM1 ID	BU*	BA*	OU*	OA*	CN**	CL**
Moda	0.477 (0.453-0.501)	0.204 (0.185-0.223)	0.641 (0.616-0.666)	0.497 (0.471-0.523)	N.A.	N.A.
Media	N.A.	N.A.	N.A.	N.A.	1.052 (0.967-1.137)	5.329 (3.191-7.467)
Mediana	N.A.	N.A.	0.673 (0.647-0.699)	0.566 (0.535-0.597)	1.053 (0.968-1.138)	5.720 (3.457-7.983)
VADU	0.475 (0.452-0.499)	0.537 (0.513-0.561)	0.661 (0.635-0.688)	0.661 (0.637-0.686)	4.302 (3.995-4.608)	189.186 (142.3-236.0)
VADE	0.524 (0.500-0.548)	0.291 (0.270-0.312)	0.661 (0.635-0.688)	0.604 (0.580-0.628)	1.926 (1.789-2.064)	11.939 (6.414-17.46)

***: Tasa nominal de error (IC 95%) **: Error medio cuadrático (IC 95%) N.A.: No aplicado**

En la variable BU el menor error se obtiene al imputar la moda y un V.A.D.U. (valor aleatorio de una distribución binomial equiprobable), si bien las diferencias con la imputación de un V.A.D.E. (valor aleatorio de una distribución estimada para la variable, es decir binomial con π=0.5) son debidas al azar, ya que los tres métodos son equivalentes.

Al promediar el error de los tres métodos de imputación de BU se obtiene una tasa nominal de error media de 0.492, muy próxima a 0.5, porcentaje de error esperado, lo que refleja la fiabilidad del procedimiento.

De cara a posteriores análisis, se considera la imputación de la moda como el mejor método de imputación directa de una variable binaria con distribución equiprobable (BU).

Respecto a la variable BA, la imputación de la moda da el menor error ($\bar{X}_{TNE} = 0.204$), que nuevamente coincide con el teórico 0.20 que cabría esperar. La imputación de un V.A.D.E. (distribución binomial caracterizada por π=0.8) tiene asociada una tasa nominal de error de 0.291, cercana a la proporción teórica de error 0.32, que se obtiene como resultado de sumar la probabilidad de imputar un 1 siendo el valor real un 0 (0.8*0.2=0.16), más la probabilidad de imputar un 0 siendo el valor real un 1 (0.2*0.8=0.16).

Finalmente, la imputación de un V.A.D.U. ofrece un error promedio de 0.537, muy superior a los otros métodos evaluados, y próximo al 0.50 previsto (0.5*2+0.5*8).

Los resultados obtenidos en la variable OU siguen el mismo razonamiento presentado para BU, los cuatro métodos de imputación analizados deben dar el mismo error ($0.\hat{6}$), y en general así es. Puesto que se debe seleccionar uno de ellos, la imputación de la moda, con una tasa nominal de error media de 0.641, es el procedimiento de elección para una variable categórica con más de dos categorías con distribución equiprobable.

En lo referente a la variable OA, la imputación de la moda y de la mediana ($\overline{X}_{TNE} = 0.497$ y $\overline{X}_{TNE} = 0.566$ respectivamente) ofrecen los mejores resultados. Las diferencias entre ambos métodos es fácilmente explicable, ya que con una variable como OA, proveniente de una población con π_1=0.5, π_2=0.35, π_3=0.15, la moda prácticamente siempre valdrá 1 (error teórico asociado 0.5), mientras que la mediana fluctuará, siendo aproximadamente en la mitad de matrices 1 y en la otra mitad 2 (error teórico asociado 0.575).

Parece lógico suponer que el mejor método de imputación directa de una variable categórica con distribución no equiprobable depende del número de categorías y del grado de asimetría que tenga: cuanto más prevalente sea una determinada categoría más efectiva será la imputación de la moda, cuantas más categorías tenga la variable mayor será el error de imputación de todos los métodos estudiados.

En la variable CN la imputación de la media y de la mediana dan un error medio cuadrático promedio de 1.052 y 1.053 respectivamente, igualdad esperada, ya que en una variable cuantitativa normal estos dos índices son muy similares.

El método de elección para posteriores análisis será la imputación de la media, ya que, frente a la imputación de la mediana, se trata de un procedimiento ampliamente implementado en los paquetes estadísticos que incorporan el análisis de datos incompletos.

En el caso de la variable CL, la imputación de la media proporciona el menor error medio cuadrático ($\overline{X}_{EMC} = 5.329$). Cabe mencionar el elevado error asociado a la imputación de un V.A.D.U.

6.3.2. Imputación por regresión/red

Matrices 1-100 (Covariación VI-VI:Nula)

En la Tabla 24 se presentan los resultados de la imputación por regresión/red a las 6 variables estudiadas en las matrices MM1_1 a MM1_100 (recuérdese que en este conjunto de matrices la correlación entre las 6 variables es prácticamente nula, y que los valores faltantes en las variables que actúan como independientes han sido previamente imputados con el mejor método de imputación directa).

En general se obtienen resultados muy similares entre los diferentes procedimientos analizados (regresión, algoritmo EM, red MLP y red RBF), con una ligera tendencia a que la imputación por regresión presente un menor error en las variables categóricas (BU, BA, OU, OA) y el algoritmo EM en las variables cuantitativas (CN, CL), si bien en general se trata de diferencias mínimas.

En cuanto a los valores concretos de error, son muy similares a los obtenidos con el mejor método de imputación directa, como cabe esperar por la ausencia de correlación entre las variables.

Tabla 24. Error medio de los métodos de imputación por regresión/red en las matrices MM1_1-MM1_100

MM1 IR/R	BU*	BA*	OU*	OA*	CN**	CL**
REGR.	0.492 (0.466-0.518)	0.205 (0.186-0.223)	0.665 (0.641-0.689)	0.509 (0.484-0.533)	1.063 (0.977-1.149)	5.352 (3.208-7.496)
EM	N.A.	N.A.	N.A.	N.A.	1.057 (0.972-1.143)	5.350 (3.203-7.496)
MLP	0.511 (0.482-0.539)	0.217 (0.198-0.235)	0.644 (0.619-0.668)	0.531 (0.505-0.556)	1.088 (1.000-1.175)	5.432 (3.297-7.566)
RBF	0.515 (0.488-0.541)	0.210 (0.192-0.228)	0.655 (0.632-0.679)	0.519 (0.493-0.544)	1.086 (0.996-1.177)	5.507 (3.366-7.648)

***: Tasa nominal de error (IC 95%) **: Error medio cuadrático (IC 95%) N.A.: No aplicado**

En la Tabla 25 se hallan los resultados de la imputación por regresión/red a las 6 variables en las matrices MM2_1 a MM2_100 (matrices con correlación nula y con valores faltantes únicamente en una variable).

No se observan diferencias relevantes respecto a la imputación en las matrices MM1, excepto en la variable CL, en la que se obtiene un EMC promedio menor en aproximadamente 0.9, lo que sugiere que la imputación directa de las otras variables ha incrementado levemente la covariación con la variable CL.

Tabla 25. Error medio de los métodos de imputación por regresión/red en las matrices MM2_1-MM2_100

MM2 IR/R	BU*	BA*	OU*	OA*	CN**	CL**
REGR.	0.515 (0.489-0.541)	0.210 (0.191-0.229)	0.661 (0.640-0.682)	0.495 (0.469-0.520)	1.011 (0.937-1.085)	4.444 (3.321-5.568)
EM	N.A.	N.A.	N.A.	N.A.	1.003 (0.930-1.077)	4.437 (3.315-5.559)
MLP	0.477 (0.451-0.502)	0.221 (0.201-0.241)	0.645 (0.623-0.667)	0.535 (0.509-0.562)	1.013 (0.938-1.087)	4.585 (3.479-5.691)
RBF	0.523 (0.497-0.548)	0.211 (0.193-0.230)	0.657 (0.632-0.683)	0.521 (0.494-0.547)	1.028 (0.952-1.104)	4.585 (3.447-5.724)

***: Tasa nominal de error (IC 95%) **: Error medio cuadrático (IC 95%) N.A.: No aplicado**

Matrices 101-200 (Covariación VI-VI:Media)

Los resultados de la imputación por regresión/red en las matrices MM1_101 a MM1_200 (matrices con correlación media entre las 6 variables y con los datos missing en las variables independientes imputados previamente) se hallan en la Tabla 26.

En la variable BU, la imputación por red neuronal MLP o RBF da el menor error ($\overline{X}_{TNE} = 0.273$ y $\overline{X}_{TNE} = 0.279$ respectivamente), aproximadamente un 4% inferior al cometido al imputar mediante regresión logística. Además, las diferencias respecto a los métodos de imputación directa son bastante elevadas (del orden del 23%).

En la variable BA los resultados de los tres procedimientos son similares, posiblemente debido a que el valor imputado en todos los casos es el mismo: el más prevalente. Consecuentemente, apenas se consigue mejoría respecto a la mejor imputación directa (moda).

En la variable OU el menor error promedio ($\overline{X}_{TNE} = 0.438$) se consigue imputando con red RBF. Respecto a los procedimientos de imputación directa el decremento es de nuevo relevante, situándose en un 21.7% en el caso más favorable (moda-RBF).

En lo referente a la variable OA no se puede considerar un procedimiento superior a los demás. Con una tasa nominal de error media en los tres métodos de 0.421, cualquiera de ellos es preferible a la imputación directa de los valores missing.

En la variable CN los mejores resultados se obtienen mediante la imputación con red neuronal RBF ($\overline{X}_{EMC} = 0.595$) y con red neuronal MLP ($\overline{X}_{EMC} = 0.611$). Aunque con el algoritmo EM y la regresión lineal también se reduce el error

medio cuadrático respecto a la imputación directa, el procedimiento de elección en la imputación de una variable cuantitativa normal es mediante red neuronal RBF.

El error cometido al imputar por regresión/red la variable CL es muy similar con los cuatro procedimientos evaluados, y claramente inferior al cometido mediante la imputación directa de esta variable. Unicamente los resultados con red MLP parecen estar ligeramente afectados por la asimetría de la variable CL.

Tabla 26. Error medio de los métodos de imputación por regresión/red en las matrices MM1_101-MM1_200

MM1 IR/R	BU*	BA*	OU*	OA*	CN**	CL**
REGR.	0.313 (0.288-0.337)	0.185 (0.166-0.203)	0.477 (0.447-0.506)	0.414 (0.391-0.437)	0.680 (0.629-0.731)	3.815 (2.632-4.997)
EM	N.A.	N.A.	N.A.	N.A.	0.637 (0.586-0.687)	3.858 (2.658-5.058)
MLP	0.273 (0.252-0.295)	0.194 (0.174-0.214)	0.454 (0.429-0.479)	0.429 (0.407-0.452)	0.611 (0.556-0.666)	4.112 (2.859-5.364)
RBF	0.279 (0.255-0.303)	0.188 (0.169-0.207)	0.438 (0.414-0.462)	0.419 (0.393-0.445)	0.595 (0.542-0.649)	3.836 (2.696-4.976)

***: Tasa nominal de error (IC 95%) **: Error medio cuadrático (IC 95%) N.A.: No aplicado**

En la Tabla 27 se presenta el error de imputación cometido en las matrices MM2_101 a MM2_200 (con correlación media entre las seis variables y con valores faltantes únicamente en una variable).

Las diferencias respecto a la imputación en las matrices MM1_101 a MM1_200 se encuentran principalmente en las variables BU, OU y CL, o dicho de otra manera, la imputación previa de los valores missing en las variables independientes aumenta el error al imputar por regresión/red una variable del tipo BU, OU o CL.

En las variables BU y OU, el error en las matrices MM2 es inferior en aproximadamente un 4% cuando se imputa mediante red neuronal, y en un 3% cuando se imputa por regresión. En la variable CL la disminución se sitúa en alrededor de 1 punto.

Tabla 27. Error medio de los métodos de imputación por regresión/red en las matrices MM2_101-MM2_200

MM2 IR/R	BU*	BA*	OU*	OA*	CN**	CL**
REGR.	0.285 (0.261-0.309)	0.183 (0.161-0.204)	0.446 (0.420-0.473)	0.402 (0.377-0.427)	0.681 (0.633-0.730)	2.909 (1.896-3.921)
EM	N.A.	N.A.	N.A.	N.A.	0.637 (0.589-0.685)	2.903 (1.881-3.924)
MLP	0.237 (0.213-0.261)	0.196 (0.172-0.220)	0.407 (0.378-0.436)	0.422 (0.397-0.448)	0.609 (0.561-0.658)	3.124 (2.088-4.160)
RBF	0.240 (0.215-0.265)	0.192 (0.169-0.215)	0.397 (0.369-0.424)	0.400 (0.374-0.425)	0.627 (0.548-0.707)	3.199 (2.117-4.282)

***: Tasa nominal de error (IC 95%) **: Error medio cuadrático (IC 95%) N.A.: No aplicado**

Matrices 201-300 (Covariación VI-VI:Baja)

El error promedio de la imputación por regresión/red en las matrices MM1_201 a MM1_300 (matrices con correlación baja entre las 6 variables y con los valores missing en las variables independientes imputados previamente) se halla en la Tabla 28.

Tabla 28. Error medio de los métodos de imputación por regresión/red en las matrices MM1_201-MM1_300

MM1 IR/R	BU*	BA*	OU*	OA*	CN**	CL**
REGR.	0.448 (0.424-0.473)	0.185 (0.164-0.206)	0.633 (0.608-0.659)	0.502 (0.478-0.526)	0.978 (0.909-1.047)	5.060 (4.380-5.739)
EM	N.A.	N.A.	N.A.	N.A.	0.967 (0.897-1.037)	5.070 (4.396-5.743)
MLP	0.442 (0.415-0.469)	0.212 (0.191-0.233)	0.603 (0.575-0.630)	0.527 (0.502-0.551)	0.942 (0.872-1.012)	5.153 (4.459-5.846)
RBF	0.424 (0.397-0.451)	0.187 (0.166-0.208)	0.600 (0.575-0.625)	0.519 (0.493-0.545)	0.923 (0.853-0.992)	5.070 (4.369-5.771)

***: Tasa nominal de error (IC 95%) **: Error medio cuadrático (IC 95%) N.A.: No aplicado**

Como es de esperar, en general se obtienen valores medios entre los obtenidos en las matrices 1-100 y en las matrices 101-200. La excepción se halla en las variables categóricas no equiprobables (BA y OA), en las que el error no se reduce respecto a las matrices con ausencia de correlación, y en especial en BA, en la que ni siquiera se incrementa respecto a las matrices con presencia de correlación.

Las diferencias entre los cuatro procedimientos de imputación son mínimas, lo que sugiere que la mayor capacidad de las redes neuronales para detectar patrones de covariación en los datos se pierde cuando dicha covariación es baja.

En la Tabla 29 se presentan los resultados obtenidos al imputar por regresión/red los valores ausentes en las matrices MM2_201 a MM2_300 (con correlación baja entre las seis variables y con valores faltantes únicamente en una variable).

De forma similar a lo observado en las matrices con ausencia total de correlación, apenas se aprecian diferencias respecto a la imputación en las matrices MM1, aunque si parece haber una leve tendencia general a que la imputación previa de los datos perdidos en las variables independientes incremente el error. La excepción nuevamente se halla en la variable CL, en la que estas diferencias se acentúan, y en la variable OA, en la que el error es algo menor en las matrices MM1.

Tabla 29. Error medio de los métodos de imputación por regresión/red en las matrices MM2_201-MM2_300

MM2 IR/R	BU*	BA*	OU*	OA*	CN**	CL**
REGR.	0.440 (0.414-0.467)	0.196 (0.177-0.216)	0.617 (0.590-0.643)	0.529 (0.505-0.554)	0.958 (0.888-1.028)	4.200 (3.281-5.119)
EM	N.A.	N.A.	N.A.	N.A.	0.911 (0.842-0.981)	4.210 (3.276-5.143)
MLP	0.446 (0.422-0.470)	0.214 (0.192-0.235)	0.591 (0.565-0.617)	0.549 (0.525-0.573)	0.928 (0.851-1.005)	4.316 (3.392-5.240)
RBF	0.434 (0.409-0.459)	0.199 (0.179-0.219)	0.567 (0.540-0.593)	0.534 (0.511-0.558)	0.894 (0.817-0.971)	4.202 (3.272-5.131)

***: Tasa nominal de error (IC 95%) **: Error medio cuadrático (IC 95%) N.A.: No aplicado**

Para sintetizar los resultados presentados hasta el momento, se presentan a continuación un conjunto de gráficos con el error de imputación directa y por regresión/red en cada una de las variables estudiadas. Así, la Ilustración 2 presenta el error de imputación (media e intervalo de confianza) en la variable BU para cada uno de los métodos de imputación directa y por regresión/red en las matrices 1 a 100, mientras que en la Ilustración 3 se representa el error en las matrices 101 a 200 y en la Ilustración 4 en las matrices 201 a 300.

Variable BU

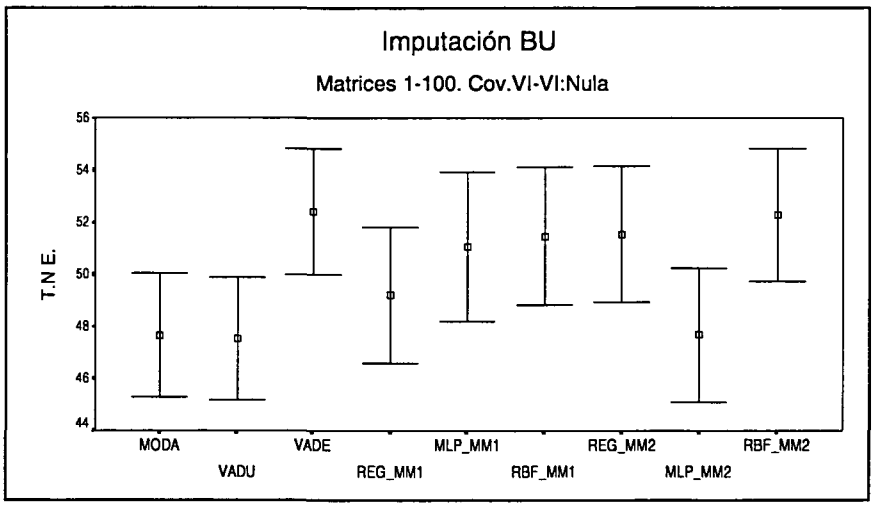


Ilustración 2. Media de la tasa nominal de error (IC 95%) en la imputación de BU, en las matrices 1-100

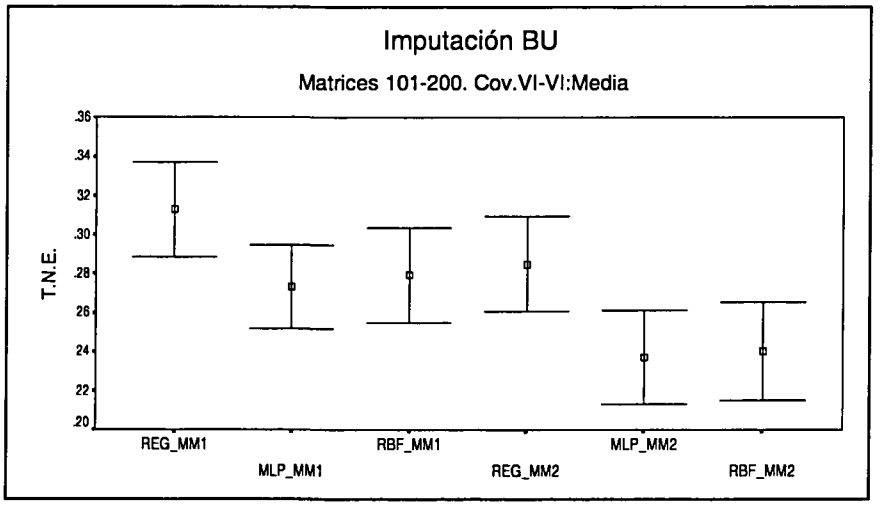


Ilustración 3. Media de la tasa nominal de error (IC 95%) en la imputación de BU, en las matrices 101-200

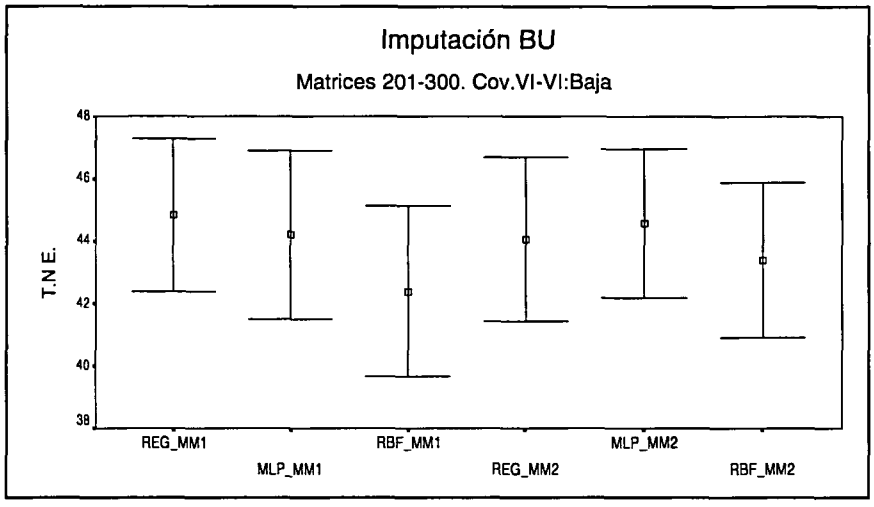


Ilustración 4. Media de la tasa nominal de error (IC 95%) en la imputación de BU, en las matrices 201-300

Variable BA

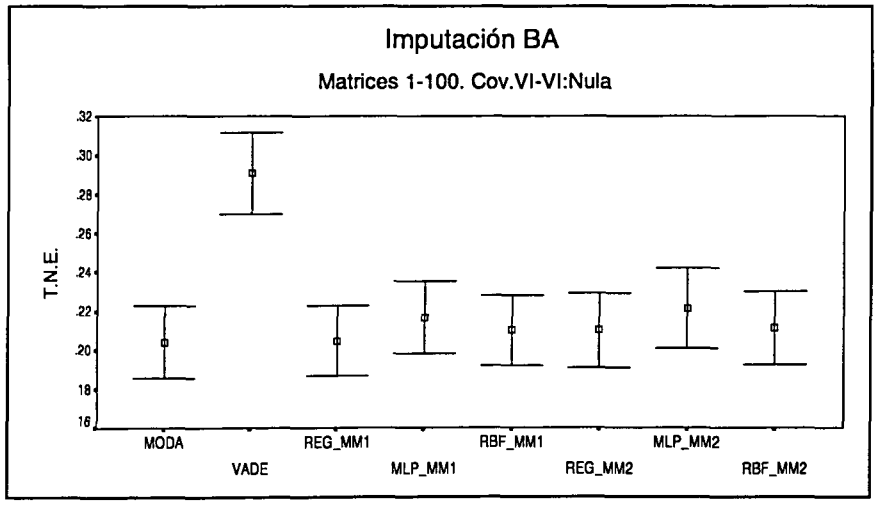


Ilustración 5. Media de la tasa nominal de error (IC 95%) en la imputación de BA, en las matrices 1-100 (VADU no representado)

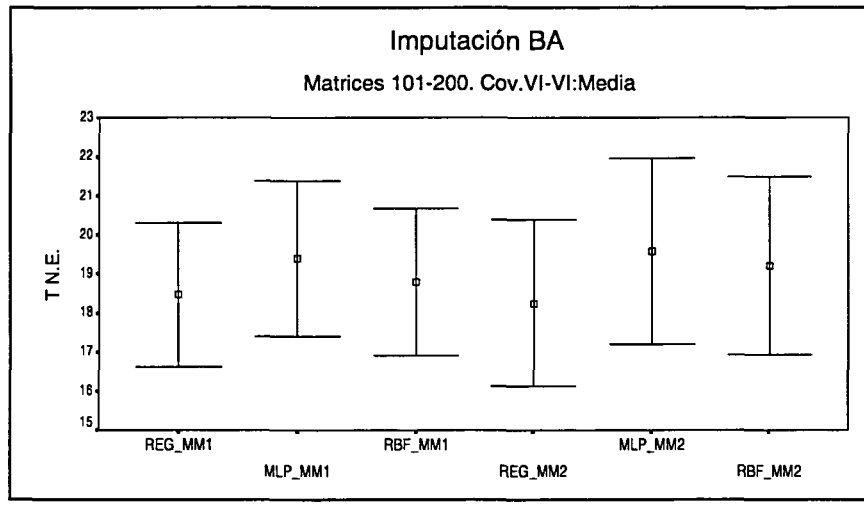


Ilustración 6. Media de la tasa nominal de error (IC 95%) en la imputación de BA, en las matrices 101-200

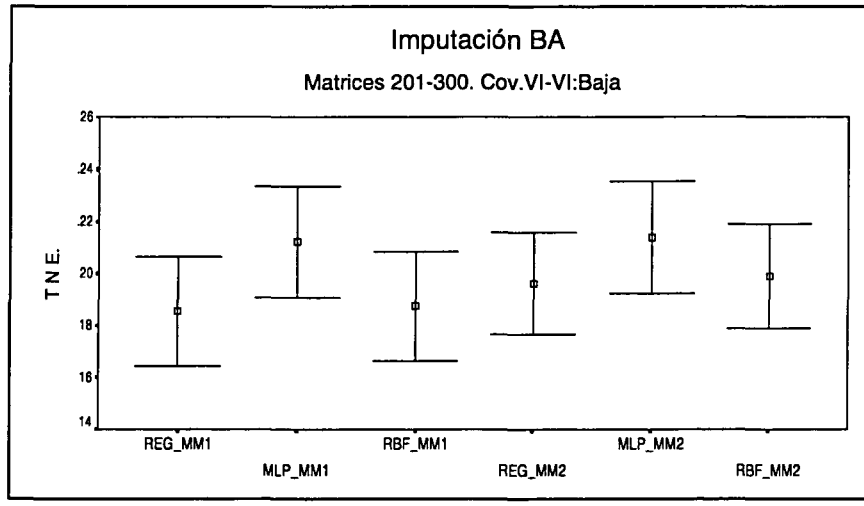


Ilustración 7. Media de la tasa nominal de error (IC 95%) en la imputación de BA, en las matrices 201-300

Variable OU

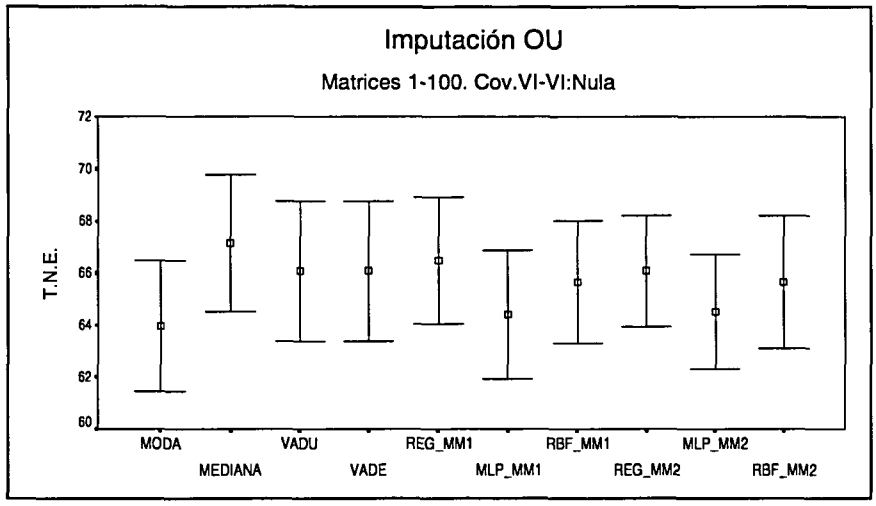


Ilustración 8. Media de la tasa nominal de error (IC 95%) en la imputación de OU, en las matrices 1-100

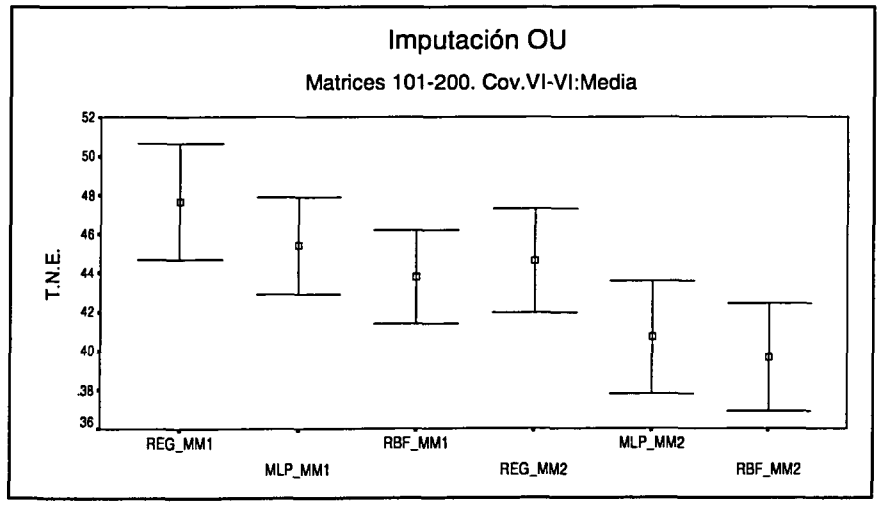


Ilustración 9. Media de la tasa nominal de error (IC 95%) en la imputación de OU, en las matrices 101-200

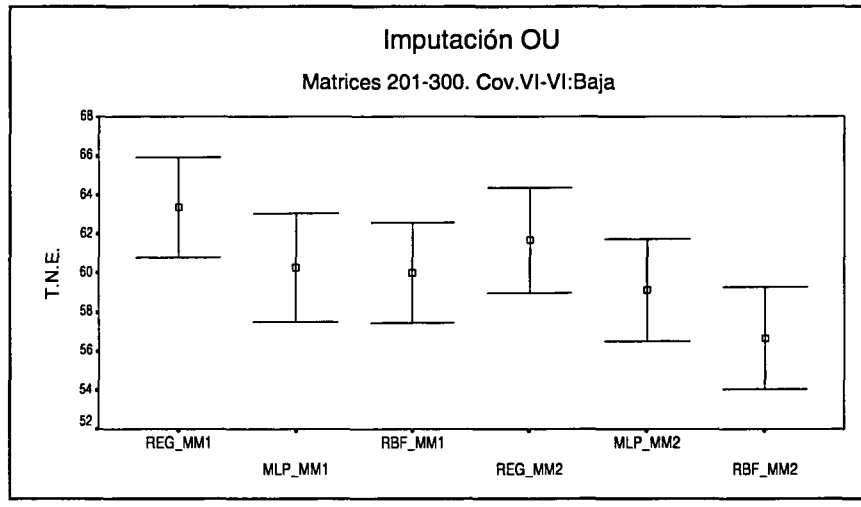


Ilustración 10. Media de la tasa nominal de error (IC 95%) en la imputación de OU, en las matrices 201-300

Variable OA

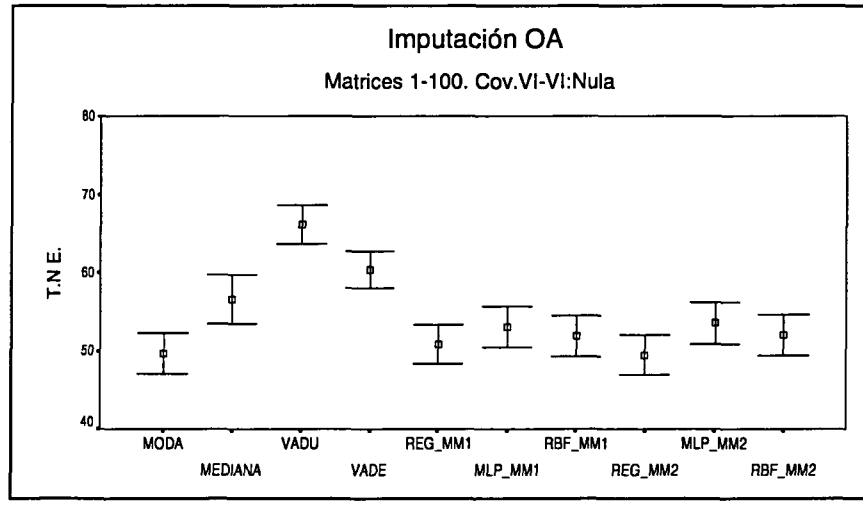


Ilustración 11. Media de la tasa nominal de error (IC 95%) en la imputación de OA, en las matrices 1-100

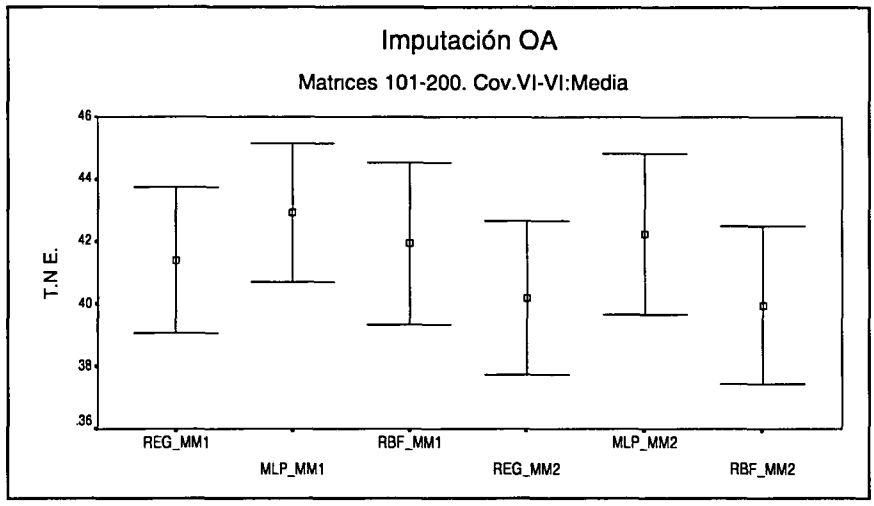


Ilustración 12. Media de la tasa nominal de error (IC 95%) en la imputación de OA, en las matrices 101-200

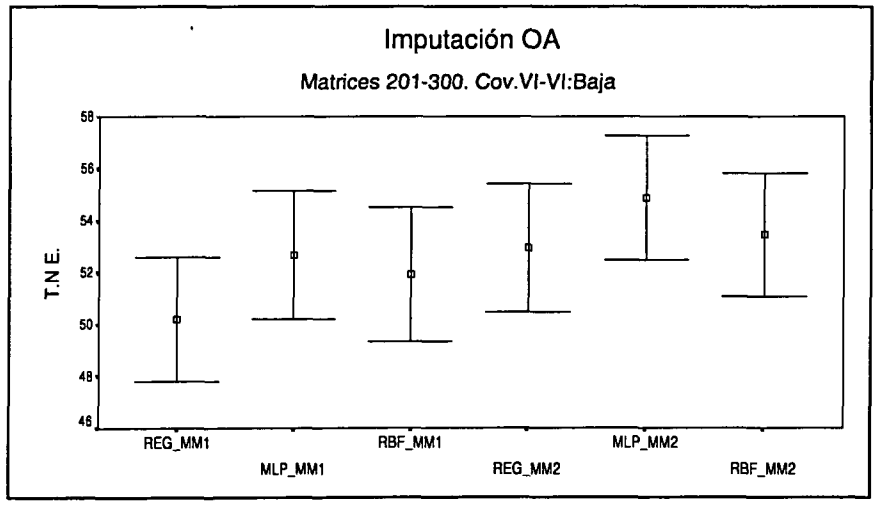


Ilustración 13. Media de la tasa nominal de error (IC 95%) en la imputación de OA, en las matrices 201-300

Variable CN

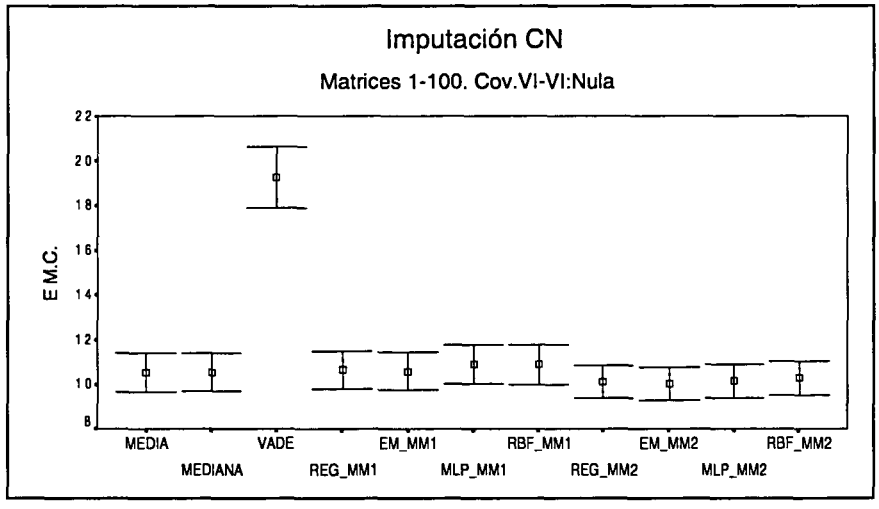


Ilustración 14. Media del error medio cuadrático (IC 95%) en la imputación de CN, en las matrices 1-100 (VADU no representado)

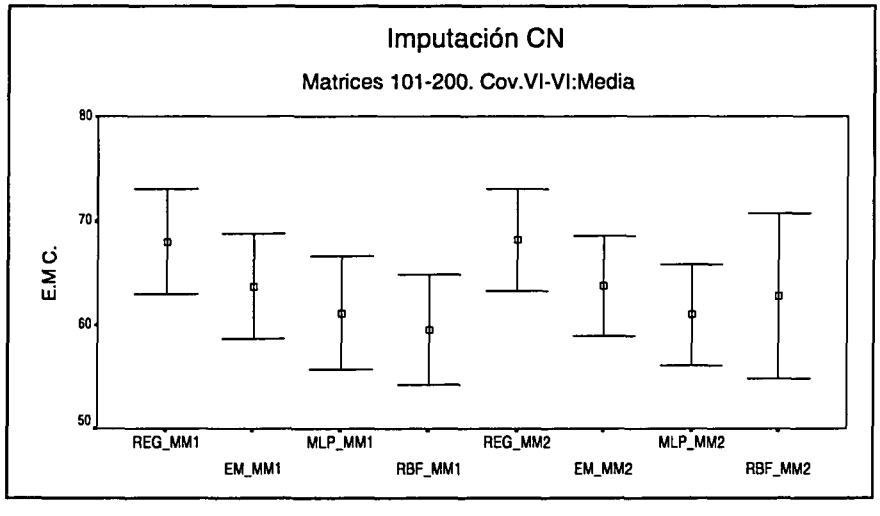


Ilustración 15. Media del error medio cuadrático (IC 95%) en la imputación de CN, en las matrices 101-200

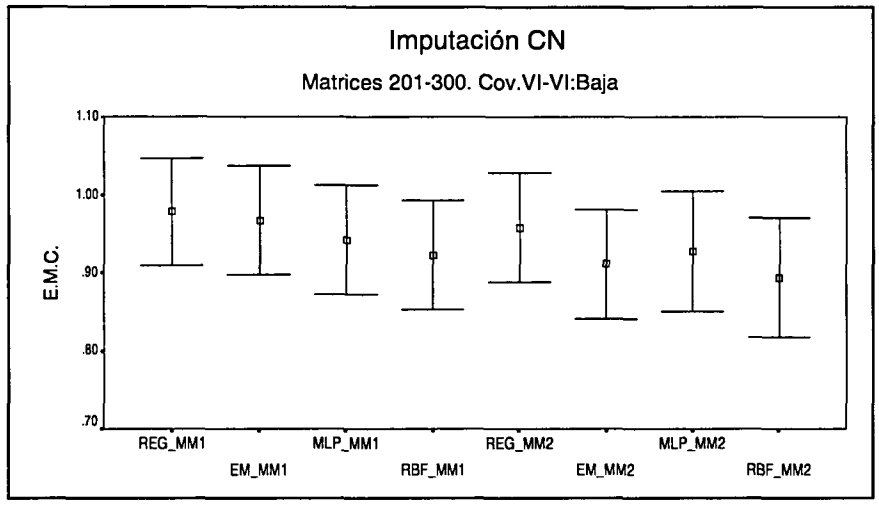


Ilustración 16. Media del error medio cuadrático (IC 95%) en la imputación de CN, en las matrices 201-300

Variable CL

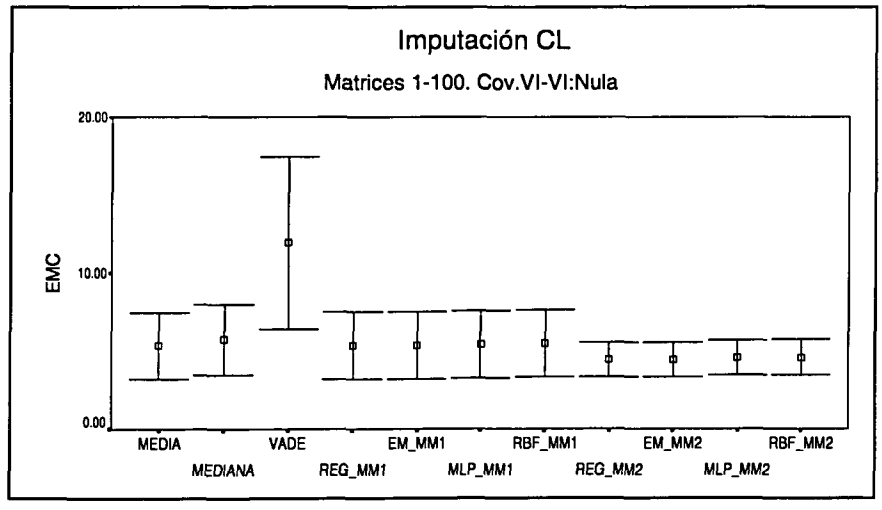


Ilustración 17. Media del error medio cuadrático (IC 95%) en la imputación de CL, en las matrices 1-100 (VADU no representado)

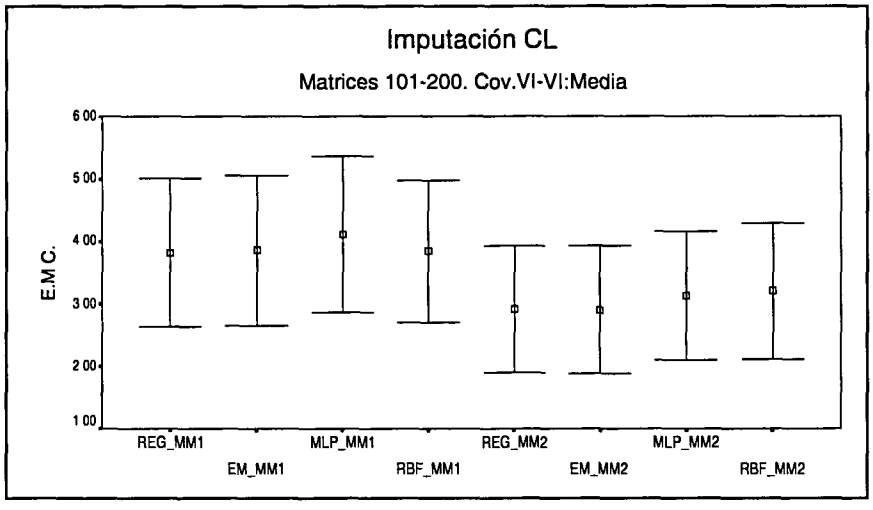


Ilustración 18. Media del error medio cuadrático (IC 95%) en la imputación de CL, en las matrices 101-200

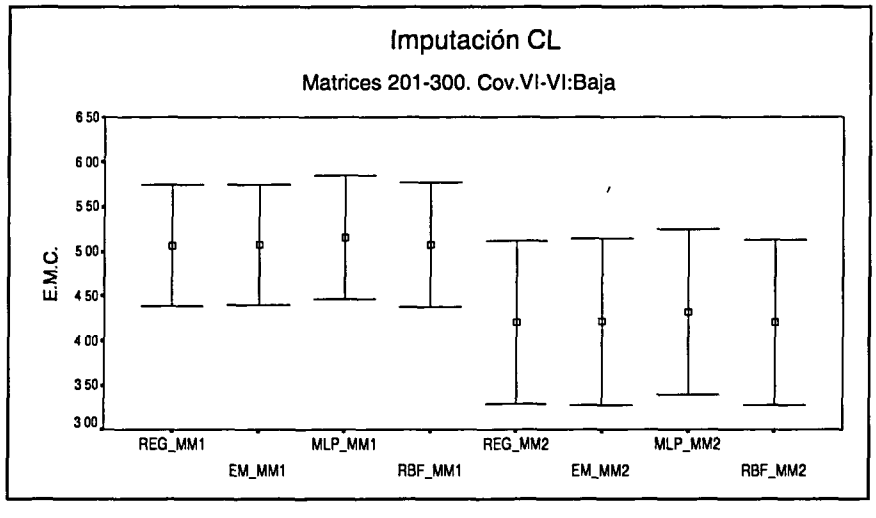


Ilustración 19. Media del error medio cuadrático (IC 95%) en la imputación de CL, en las matrices 201-300

Puesto que cada una de las ilustraciones anteriores representa una única variable, no permiten comparar el error de imputación cometido en variables diferentes, además de no compartir la misma escala en el eje de ordenadas. Las seis gráficas siguientes presentan el error de imputación agrupando las variables en función de su naturaleza, de manera que la Ilustración 20 presenta el error (media de la tasa nominal de error) de la imputación directa y por regresión/red en las variables categóricas para las matrices 1 a 100, y la Ilustración 21 representa el error (media del error medio cuadrático) en las variables cuantitativas para las matrices 1-100. Puesto que los análisis realizados en las matrices MM2 sólo sirven para evaluar el objetivo 2, en las siguientes gráficas se ha obviado su representación. Tenga también presente que en las gráficas que representan las variables cuantitativas no se incluyen los métodos V.A.D.U. y V.A.D.E. aplicados a la variable CL, ya que su error asociado es tan elevado que el incremento en la escala de valores restaría detalle al gráfico.

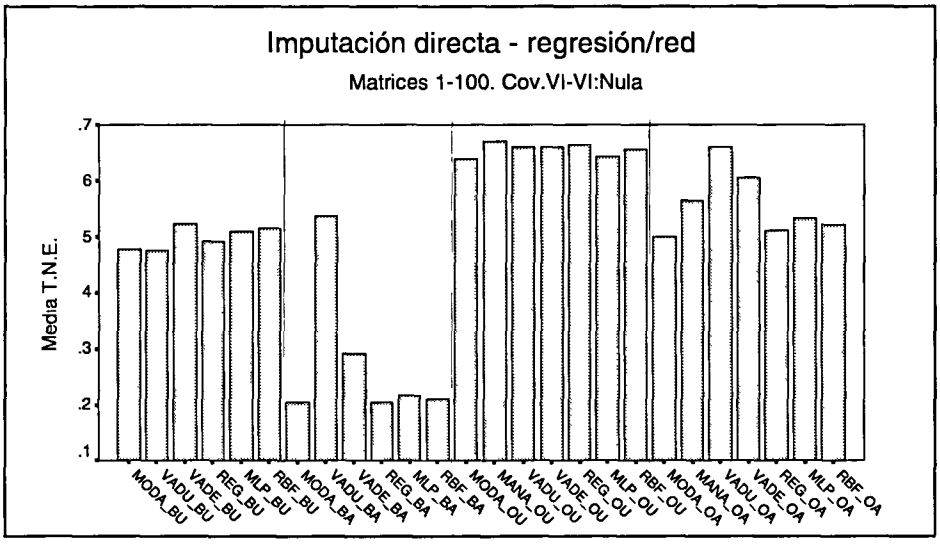


Ilustración 20. Media de la tasa nominal de error en la imputación directa y por regresión/red a las variables categóricas, en las matrices 1-100 (MM1)

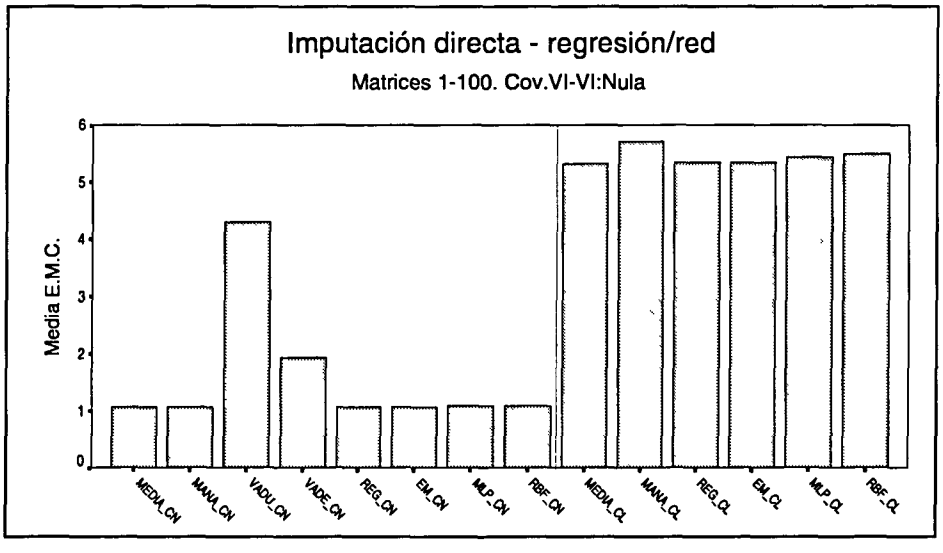


Ilustración 21. Media del error medio cuadrático en la imputación directa y por regresión/red a las variables cuantitativas, en las matrices 1-100 (MM1) (VADU_CL y VADE_CL no representados)

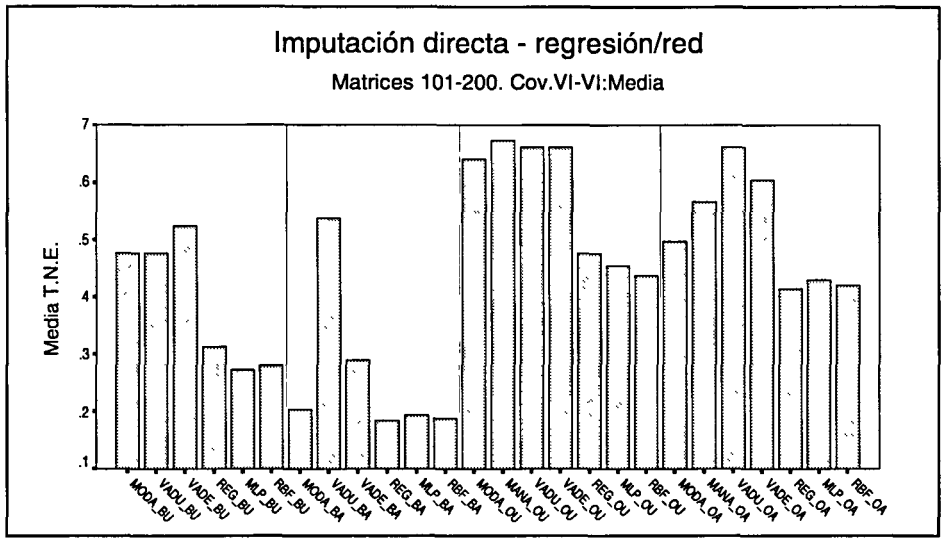


Ilustración 22. Media de la tasa nominal de error en la imputación directa y por regresión/red a las variables categóricas, en las matrices 101-200 (MM1)

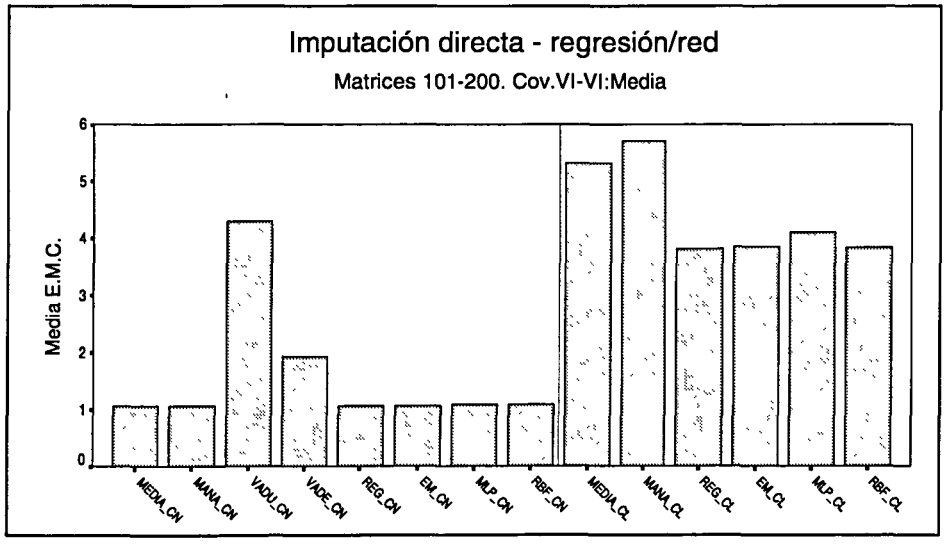


Ilustración 23. Media del error medio cuadrático en la imputación directa y por regresión/red a las variables cuantitativas, en las matrices 101-200 (MM1) (VADU_CL y VADE_CL no representados)

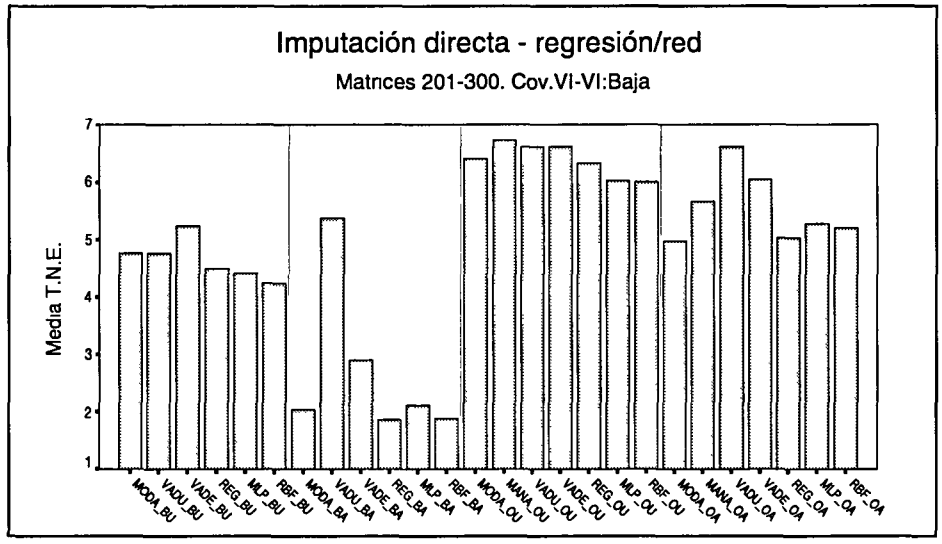


Ilustración 24. Media de la tasa nominal de error en la imputación directa y por regresión/red a las variables categóricas, en las matrices 201-300 (MM1)

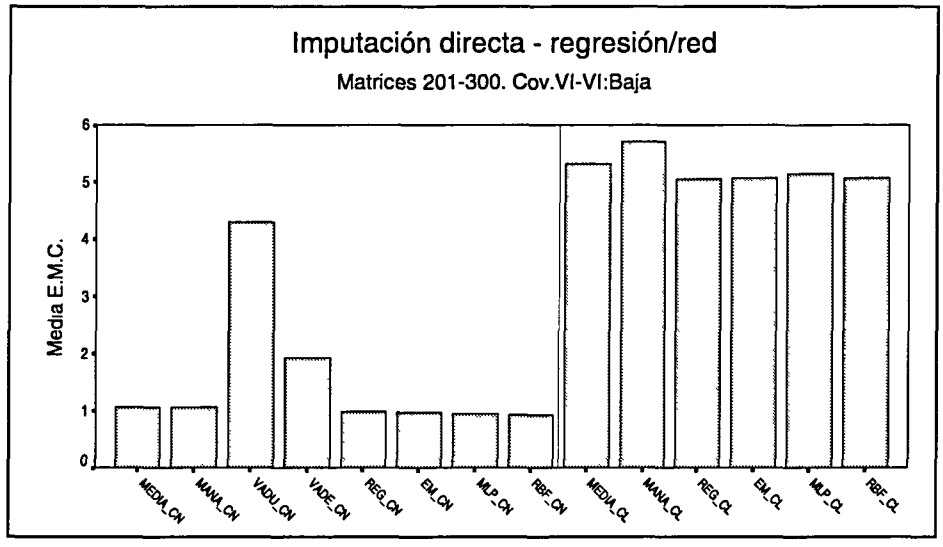


Ilustración 25. Media del error medio cuadrático en la imputación directa y por regresión/red a las variables cuantitativas, en las matrices 201-300 (MM1) (VADU_CL y VADE_CL no representados)

6.3.3. Clasificación

A continuación se presentan los resultados correspondientes a los modelos de clasificación estimados para clasificar la variable dependiente binaria (DEP). En cada tabla se presenta el promedio de la tasa nominal de error de la clasificación, calculada como el porcentaje de registros de test incorrectamente clasificados, junto a su intervalo de confianza. Los resultados se presentan diferenciando la operación realizada con los valores faltantes antes de estimar el modelo de clasificación (imputación directa, *listwise*, imputación por regresión, etc.), el conjunto de matrices analizadas (1-100, 101-200, 201-300) y la técnica de clasificación empleada (regresión logística, red neuronal MLP y red neuronal RBF). A continuación de cada tabla se incluye un gráfico con los resultados obtenidos.

Matrices 1-100 (Covariación VI-VI:Nula, VI-VD:Nula)

En la Tabla 30 se hallan los resultados correspondientes a la clasificación en las matrices 1-100 (con covariación nula entre todas las variables), en las que los valores missing en las seis variables que actúan como independientes han sido tratados de diferente manera:

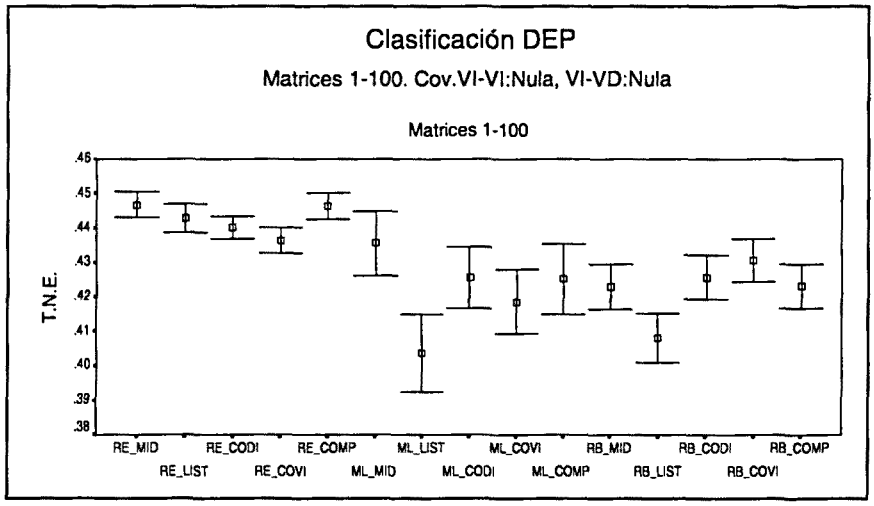
- Imputados con el mejor método de imputación directa (MID).
- Eliminados los registros que tienen datos missing (LIST).
- Codificados al valor 99 (CODI).
- Codificados al valor 99 e incluidas variables indicadoras (COVI).

También se incluye el resultado de la clasificación en las matrices completas (COMP) generadas inicialmente.

El error de clasificación es muy similar en todas las condiciones planteadas, si bien cabe mencionar que la clasificación mediante red neuronal MLP o RBF, habiendo eliminado los registros con datos incompletos (*listwise*), proporciona el menor error ($\overline{X}_{TNE} = 0.404$ y $\overline{X}_{TNE} = 0.408$ respectivamente). Los porcentajes de error son muy elevados si se tiene en cuenta que por azar se clasificarían correctamente un 50% de registros, aunque eran de esperar dada la casi nula correlación existente en las matrices 1-100 entre las variables independientes y la dependiente.

Tabla 30. Media de la tasa nominal de error (IC 95%) en la clasificación de la variable dependiente en las matrices MID, LIST, CODI, COVI, COMP (1-100)

CLAS.	MID	LIST	CODI	COVI	COMP
REGR.	0.447 (0.443-0.451)	0.443 (0.439-0.447)	0.440 (0.437-0.443)	0.436 (0.433-0.440)	0.446 (0.443-0.450)
MLP	0.436 (0.426-0.445)	0.404 (0.392-0.415)	0.426 (0.417-0.434)	0.419 (0.409-0.428)	0.425 (0.415-0.435)
RBF	0.423 (0.416-0.429)	0.408 (0.401-0.415)	0.426 (0.419-0.432)	0.431 (0.424-0.437)	0.423 (0.417-0.429)



La Tabla 31 contiene los resultados correspondientes a la clasificación a partir de las matrices BU_RE, BU_ML y BU_RB (matrices 1-100), en las que los valores missing en la variable BU han sido imputados mediante regresión, red MLP y red RBF respectivamente, y los missing en las otras variables han sido imputados con el mejor método de imputación directa. De la misma manera, en la Tabla 32 se hallan los resultados obtenidos en las matrices BA_RE, BA_ML y BA_RB (1-100), y así sucesivamente para el resto de variables independientes imputadas por regresión/red, hasta la Tabla 36, en la que se presentan los errores de clasificación obtenidos en las matrices CL_RE, CL_EM, CL_ML y CL_RB (1-100).

En todos estos modelos, el error de clasificación promedio oscila entre el 42% y el 44.8%, correspondiendo los porcentajes más bajos a la clasificación mediante red neuronal MLP o RBF, y los más altos a la clasificación mediante regresión logística. Ni el tipo de variable previamente imputado mediante regresión/red ni el tipo de imputación realizada parecen incidir en el resultado de la clasificación.

Tabla 31. Media de la tasa nominal de error (IC 95%) en la clasificación de la variable dependiente en las matrices BU_RE, BU_ML, BU_RB (1-100)

CLAS.	BU_RE	BU_ML	BU_RB
REGR.	N.A.	0.447 (0.443-0.451)	0.448 (0.444-0.452)
MLP	0.425 (0.415-0.435)	0.434 (0.424-0.443)	0.427 (0.417-0.436)
RBF	0.421 (0.415-0.427)	0.422 (0.415-0.428)	0.420 (0.413-0.427)

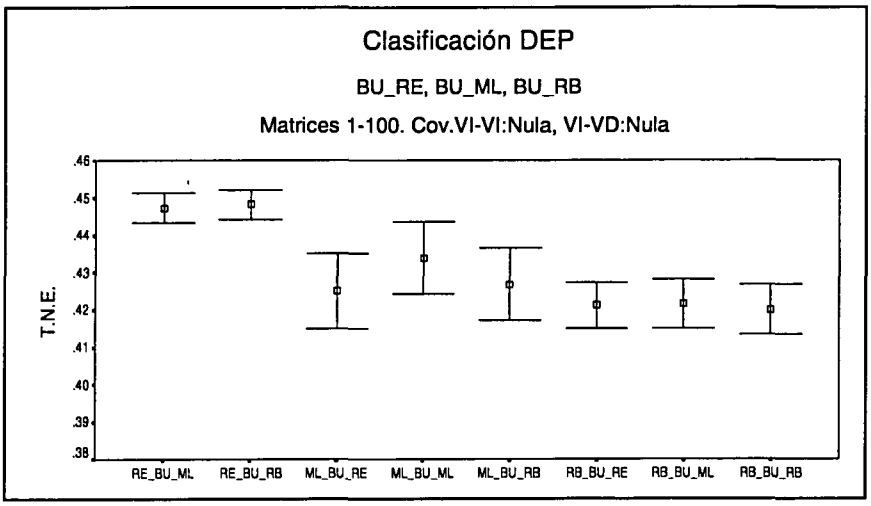


Tabla 32. Media de la tasa nominal de error (IC 95%) en la clasificación de la variable dependiente en las matrices BA_RE, BA_ML, BA_RB (1-100)

CLAS.	BA_RE	BA_ML	BA_RB
REGR.	N.A.	0.447 (0.443-0.450)	0.447 (0.444-0.451)
MLP	0.427 (0.417-0.436)	0.427 (0.417-0.437)	0.429 (0.419-0.438)
RBF	0.422 (0.416-0.428)	0.422 (0.416-0.429)	0.422 (0.415-0.428)

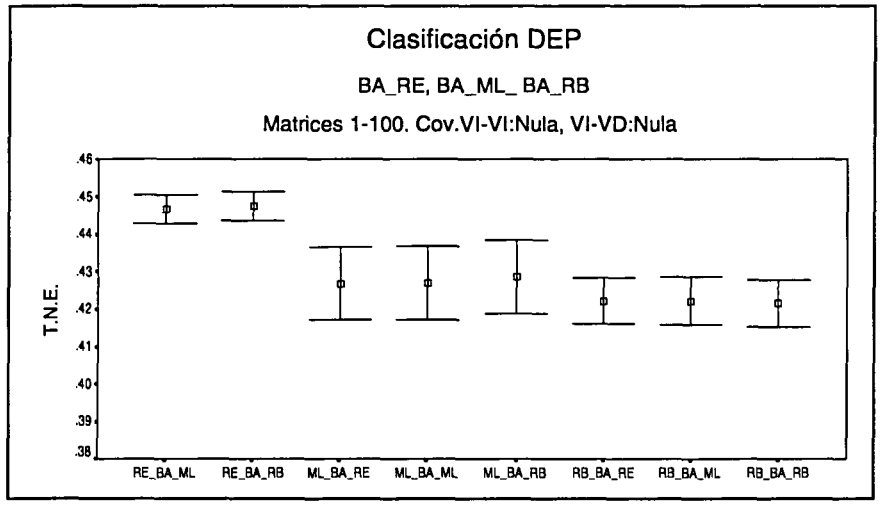


Tabla 33. Media de la tasa nominal de error (IC 95%) en la clasificación de la variable dependiente en las matrices OU_RE, OU_ML, OU_RB (1-100)

CLAS.	OU_RE	OU_ML	OU_RB
REGR.	N.A.	0.448 (0.445-0.452)	0.448 (0.444-0.451)
MLP	0.426 (0.417-0.435)	0.422 (0.413-0.431)	0.423 (0.413-0.432)
RBF	0.422 (0.415-0.428)	0.423 (0.417-0.429)	0.421 (0.415-0.428)

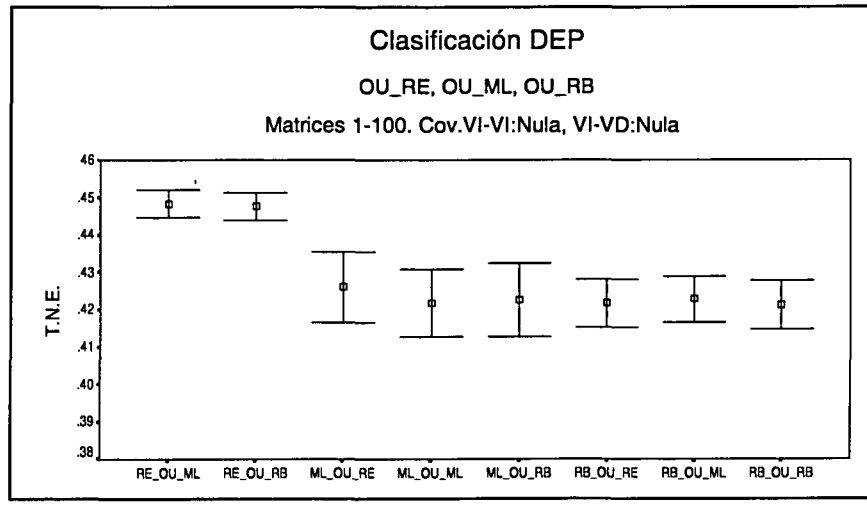


Tabla 34. Media de la tasa nominal de error (IC 95%) en la clasificación de la variable dependiente en las matrices OA_RE, OA_ML, OA_RB (1-100)

CLAS.	OA_RE	OA_ML	OA_RB
REGR.	N.A.	0.446 (0.443-0.450)	0.446 (0.443-0.450)
MLP	0.426 (0.416-0.436)	0.425 (0.416-0.434)	0.423 (0.413-0.433)
RBF	0.421 (0.415-0.428)	0.422 (0.416-0.429)	0.424 (0.417-0.430)

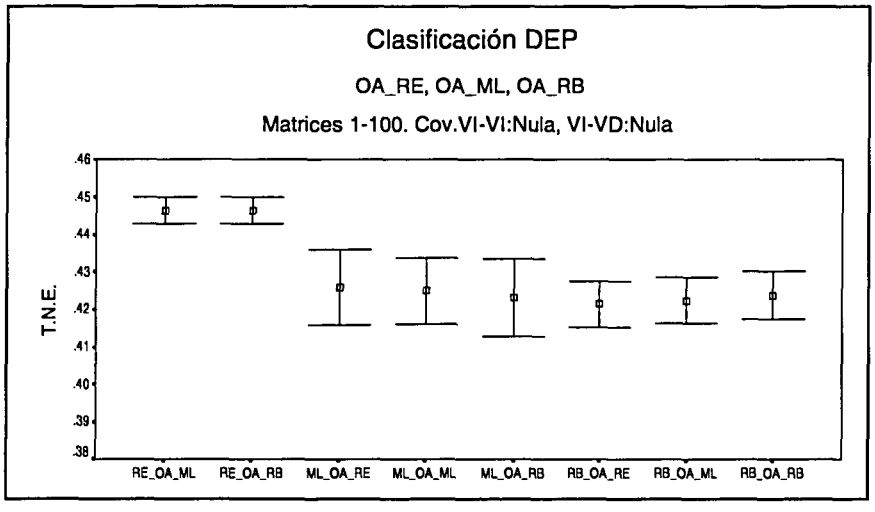


Tabla 35. Media de la tasa nominal de error (IC 95%) en la clasificación de la variable dependiente en las matrices CN_RE, CN_EM, CN_ML, CN_RB (1-100)

CLAS.	CN_RE	CN_EM	CN_ML	CN_RB
REGR.	N.A.	N.A.	0.447 (0.443-0.451)	0.446 (0.442-0.450)
MLP	0.424 (0.415-0.433)	0.423 (0.414-0.432)	0.428 (0.419-0.436)	0.423 (0.414-0.432)
RBF	0.422 (0.415-0.428)	0.422 (0.416-0.428)	0.421 (0.415-0.427)	0.421 (0.415-0.427)

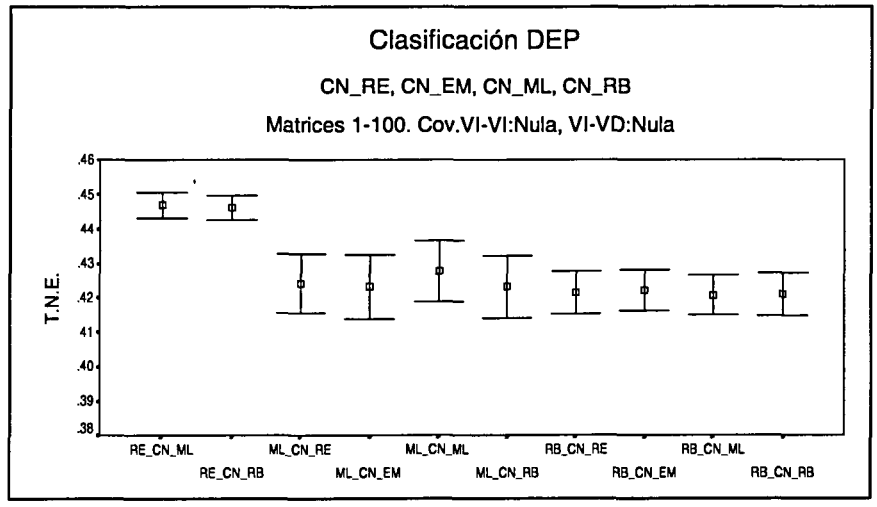
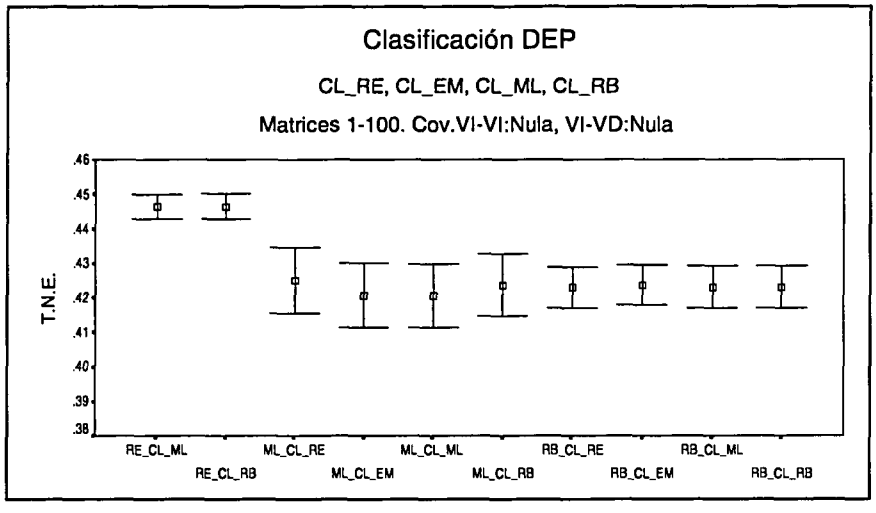


Tabla 36. Media de la tasa nominal de error (IC 95%) en la clasificación de la variable dependiente en las matrices CL_RE, CL_EM, CL_ML, CL_RB (1-100)

CLAS.	CL_RE	CL_EM	CL_ML	CL_RB
REGR.	N.A.	N.A.	0.446 (0.443-0.450)	0.446 (0.443-0.450)
MLP	0.425 (0.416-0.435)	0.421 (0.411-0.430)	0.421 (0.412-0.430)	0.424 (0.415-0.433)
RBF	0.423 (0.417-0.429)	0.424 (0.418-0.430)	0.423 (0.417-0.429)	0.423 (0.417-0.429)



Matrices 101-200 (Covariación VI-VI:Media, VI-VD:Media-alta)

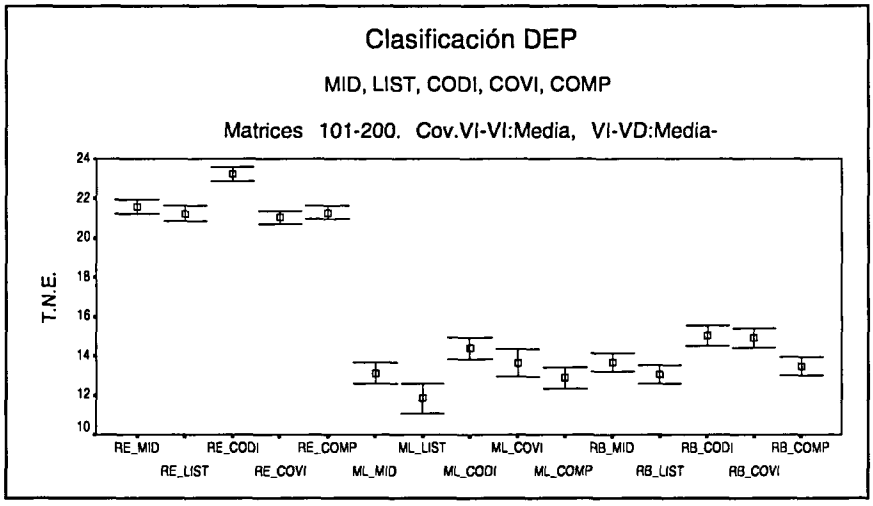
La Tabla 37 contiene los resultados de la clasificación efectuada con los métodos de análisis que directamente estiman el modelo de clasificación, sin imputar previamente un valor a partir de la información observada, en las matrices 101-200, caracterizadas por la correlación de nivel medio entre las variables independientes (VI-VI) y medio-alto con la variable dependiente (VI-VD).

Un primer análisis refleja que la clasificación mediante red MLP reduce en alrededor de un 9% el porcentaje de malas clasificaciones respecto a la regresión logística, mientras que en la clasificación con red RBF la reducción se sitúa en torno al 8%. La diferencia entre las técnicas de clasificación es casi constante en todos los métodos de análisis de datos incompletos, siendo máxima cuando se eliminan los registros incompletos (*listwise*) y mínima cuando los valores missing se codifican y además se introducen variables indicadoras.

Los porcentajes de error conseguidos mediante el empleo de redes neuronales son sustancialmente inferiores en este conjunto de matrices (101-200) respecto a los obtenidos en las matrices 1-100, con un mínimo de 11.8%, asociado nuevamente al método *listwise*, y un máximo de 15%, asociado a la codificación al valor 99. Sin considerar el procedimiento *listwise*, la imputación directa de los valores faltantes ofrece el menor error ($\overline{X}_{TNE} = 0.408$), si bien las diferencias con los otros procedimientos no son en ningún caso superiores al 2%.

Tabla 37. Media de la tasa nominal de error (IC 95%) en la clasificación de la variable dependiente en las matrices MID, LIST, CODI, COVI, COMP (101-200)

CLAS.	MID	LIST	CODI	COVI	COMP
REGR.	0.215 (0.212-0.219)	0.212 (0.208-0.216)	0.232 (0.229-0.236)	0.210 (0.207-0.214)	0.213 (0.209-0.216)
MLP	0.131 (0.126-0.137)	0.118 (0.111-0.126)	0.144 (0.138-0.149)	0.136 (0.130-0.143)	0.129 (0.123-0.134)
RBF	0.137 (0.132-0.142)	0.130 (0.126-0.135)	0.150 (0.145-0.155)	0.149 (0.144-0.154)	0.135 (0.130-0.139)



En las 6 tablas (Tabla 38 a Tabla 43) y gráficos siguientes se hallan los resultados de la clasificación a partir de las matrices BU_RE, BU_ML, BU_RB a CL_RE, CL_EM, CL_ML, CL_RB (matrices 101-200).

Nuevamente la clasificación mediante red neuronal, especialmente MLP, es claramente superior a la clasificación mediante regresión logística, con diferencias que oscilan entre el 5% y el 10%. Dichas diferencias se hacen máximas cuando se analizan las matrices con la variable BU imputada por regresión/red, y mínimas en el análisis de las matrices con CL como variable imputada.

Sin considerar la clasificación mediante regresión logística, las diferencias en función del tipo de variable imputada y del tipo de imputación realizada no son en ningún caso elevadas. Sin embargo, hay una clara tendencia a que la imputación mediante regresión/red de las variables cuantitativas (respecto a las categóricas) empeore el resultado del posterior modelo de clasificación.

Por otra parte, en las variables categóricas con distribución equiprobable (BU, OU) la previa imputación de los valores missing mediante red neuronal hace disminuir el posterior error de clasificación, y contrariamente, en las variables categóricas con distribución no equiprobable (BA, OA), es la imputación por regresión la que a posteriori produce un menor error de clasificación, aunque con diferencias menores respecto a los modelos de red neuronal. Con las variables cuantitativas sucede algo similar, en CN la mejor imputación previa es con red MLP, mientras que en CL lo es mediante regresión o algoritmo EM. En este último aspecto, aunque las diferencias son siempre mínimas también lo son sistemáticas.

Tabla 38. Media de la tasa nominal de error (IC 95%) en la clasificación de la variable dependiente en las matrices BU_RE, BU_ML, BU_RB (101-200)

CLAS.	BU_RE	BU_ML	BU_RB
REGR.	N.A.	0.215 (0.212-0.219)	0.216 (0.212-0.219)
MLP	0.131 (0.125-0.136)	0.117 (0.112-0.122)	0.129 (0.123-0.135)
RBF	0.135 (0.130-0.139)	0.127 (0.122-0.131)	0.136 (0.132-0.141)

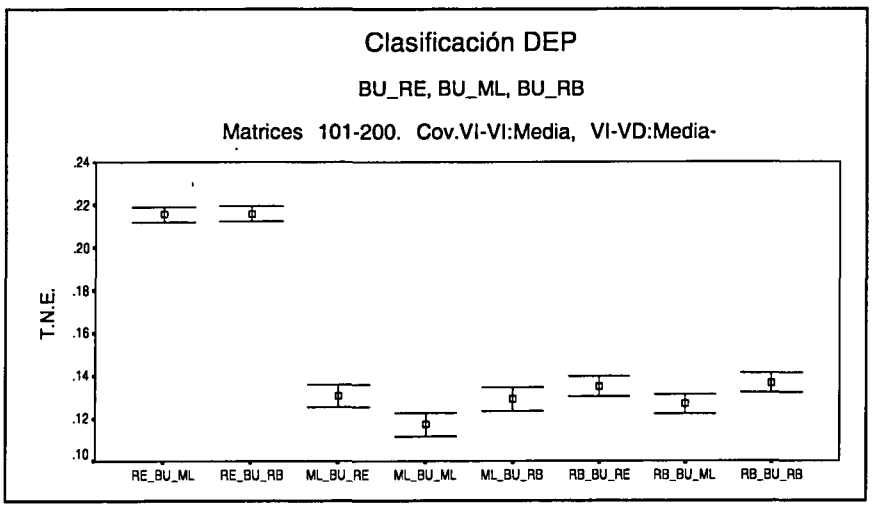


Tabla 39. Media de la tasa nominal de error (IC 95%) en la clasificación de la variable dependiente en las matrices BA_RE, BA_ML, BA_RB (101-200)

CLAS.	BA_RE	BA_ML	BA_RB
REGR.	N.A.	0.226 (0.222-0.229)	0.226 (0.222-0.229)
MLP	0.121 (0.115-0.126)	0.150 (0.145-0.155)	0.149 (0.143-0.154)
RBF	0.127 (0.122-0.132)	0.156 (0.151-0.161)	0.157 (0.153-0.162)

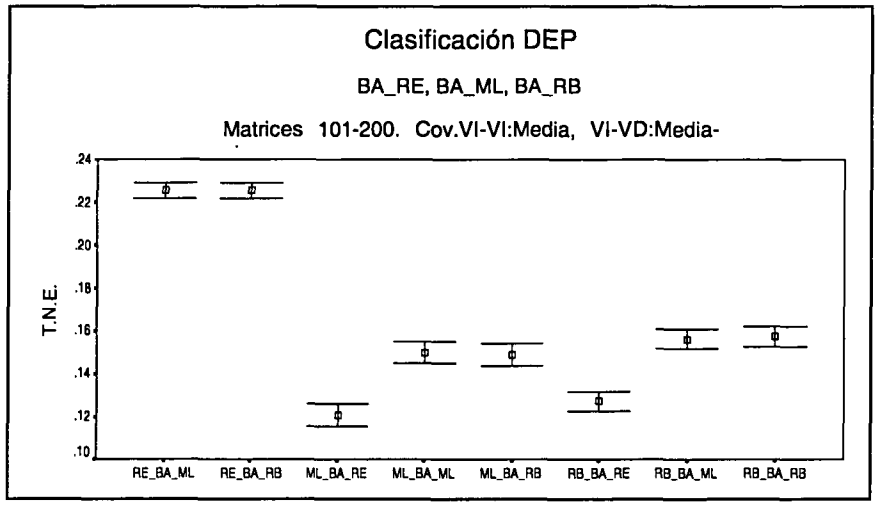


Tabla 40. Media de la tasa nominal de error (IC 95%) en la clasificación de la variable dependiente en las matrices OU_RE, OU_ML, OU_RB (101-200)

CLAS.	OU_RE	OU_ML	OU_RB
REGR.	N.A.	0.215 (0.212-0.219)	0.215 (0.212-0.219)
MLP	0.137 (0.132-0.142)	0.132 (0.126-0.137)	0.132 (0.126-0.137)
RBF	0.146 (0.141-0.150)	0.137 (0.133-0.142)	0.137 (0.132-0.142)

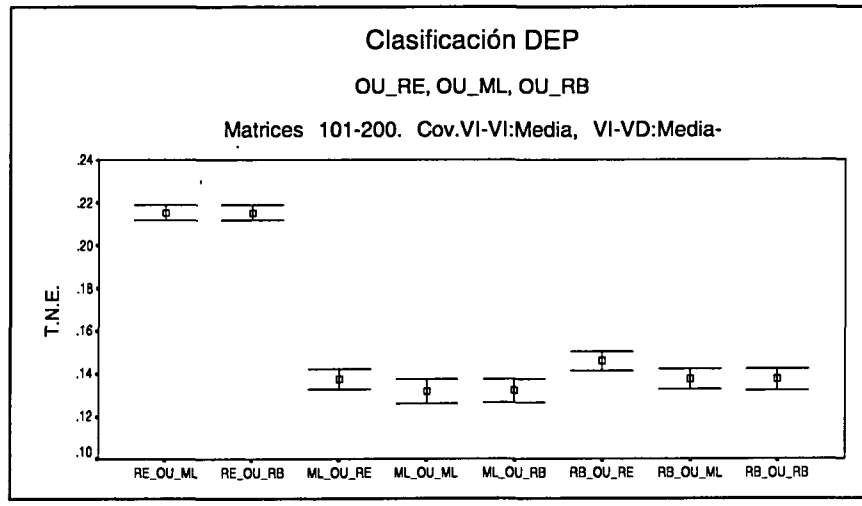


Tabla 41. Media de la tasa nominal de error (IC 95%) en la clasificación de la variable dependiente en las matrices OA_RE, OA_ML, OA_RB (101-200)

CLAS.	OA_RE	OA_ML	OA_RB
REGR.	N.A.	0.215 (0.212-0.219)	0.215 (0.212-0.219)
MLP	0.126 (0.121-0.131)	0.129 (0.124-0.134)	0.131 (0.126-0.136)
RBF	0.137 (0.132-0.141)	0.137 (0.132-0.141)	0.137 (0.133-0.142)

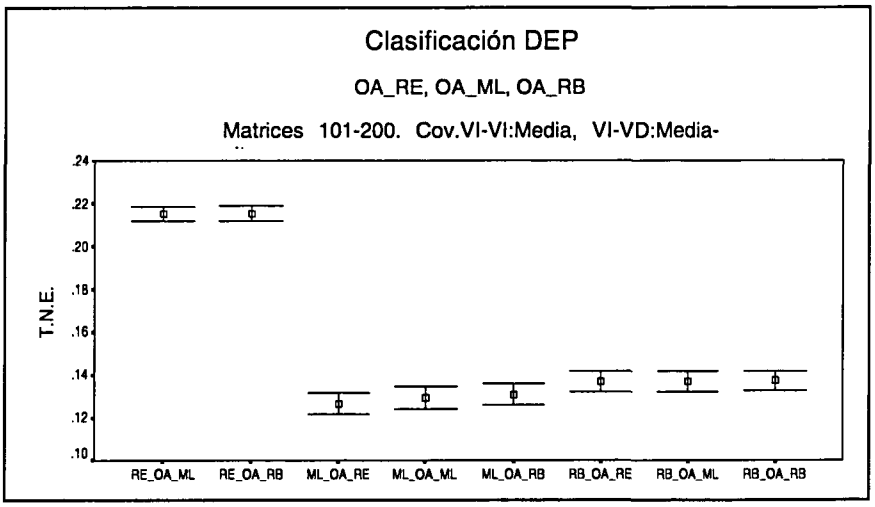


Tabla 42. Media de la tasa nominal de error (IC 95%) en la clasificación de la variable dependiente en las matrices CN_RE, CN_EM, CN_ML, CN_RB (101-200)

CLAS.	CN_RE	CN_EM	CN_ML	CN_RB
REGR.	N.A.	N.A.	0.215 (0.212-0.219)	0.215 (0.212-0.219)
MLP	0.151 (0.146-0.156)	0.150 (0.145-0.155)	0.138 (0.133-0.143)	0.148 (0.143-0.153)
RBF	0.157 (0.152-0.161)	0.157 (0.152-0.162)	0.146 (0.141-0.151)	0.157 (0.152-0.152)

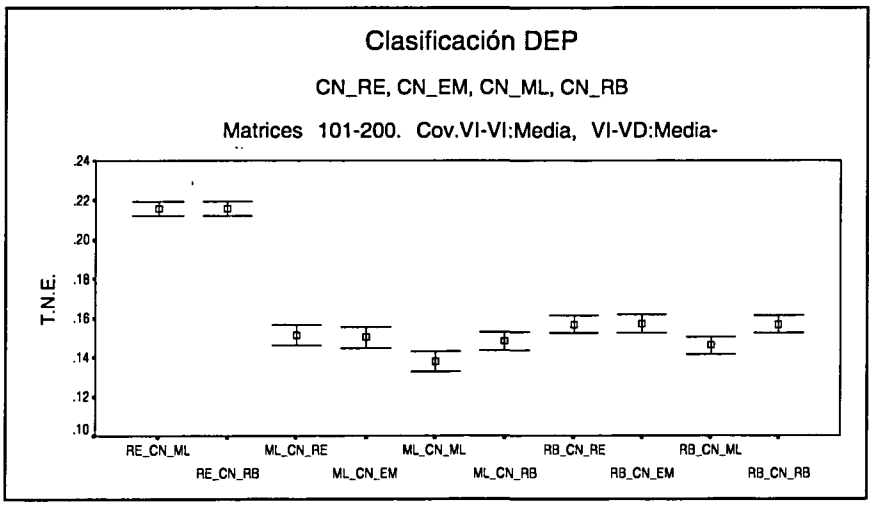
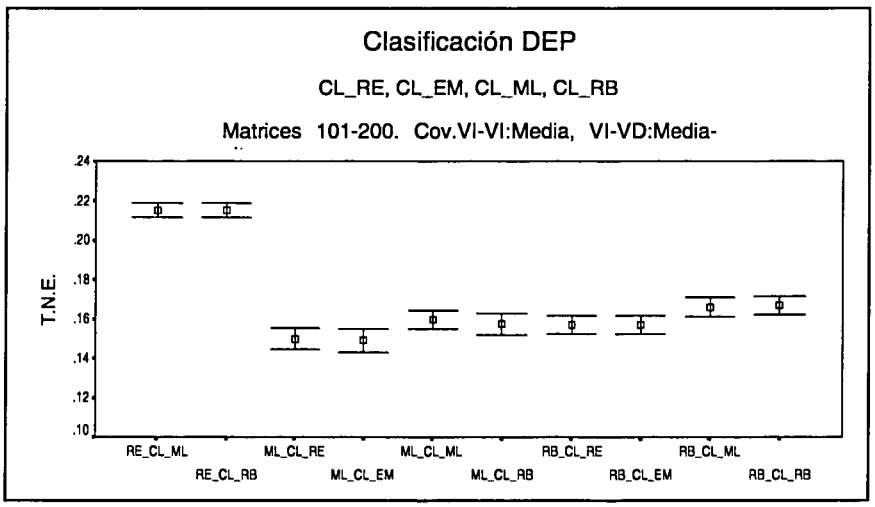


Tabla 43. Media de la tasa nominal de error (IC 95%) en la clasificación de la variable dependiente en las matrices CL_RE, CL_EM, CL_ML, CL_RB (101-200)

CLAS.	CL_RE	CL_EM	CL_ML	CL_RB
REGR.	N.A.	N.A.	0.215 (0.211-0.219)	0.215 (0.211-0.219)
MLP	0.150 (0.145-0.155)	0.149 (0.143-0.155)	0.160 (0.155-0.164)	0.157 (0.152-0.163)
RBF	0.157 (0.153-0.162)	0.157 (0.153-0.162)	0.166 (0.161-0.171)	0.167 (0.162-0.171)



Matrices 201-300 (Covariación VI-VI:Baja, VI-VD:Media)

En la Tabla 44 se resumen los resultados correspondientes a la clasificación mediante los métodos directos de análisis con datos missing en el conjunto de las matrices 201-300, con correlación baja entre las variables independientes (VI-VI) y media con la variable dependiente (VI-VD).

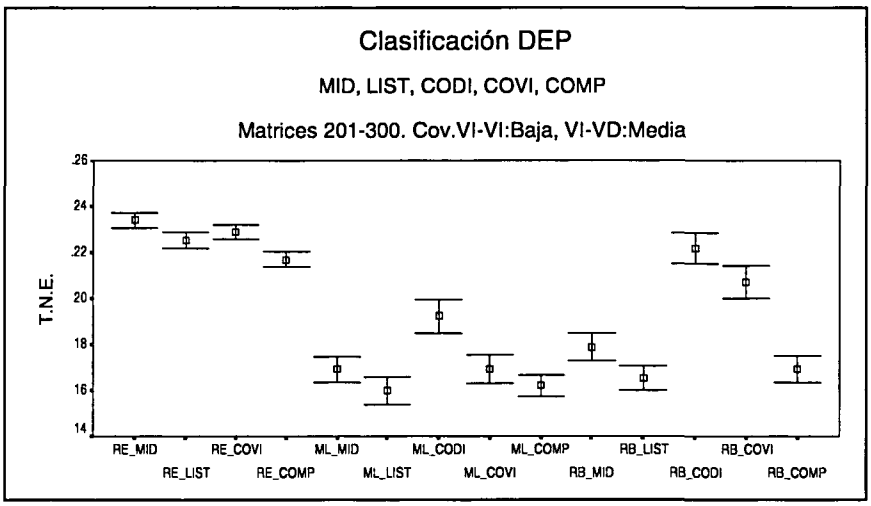
De nuevo se constata la superioridad de las redes neuronales respecto a la regresión logística, que se traduce en porcentajes de error inferiores entre el 2.2% y el 9%. Sin embargo, y a diferencia de las conclusiones derivadas para los conjuntos de matrices previamente analizados, los dos tipos de red estudiados no se comportan siempre de manera similar. En concreto, cuando los valores ausentes son codificados (CODI) o codificados e incluidas variables indicadoras (COVI), las redes MLP se muestran bastante más eficaces que las redes RBF, con diferencias entre ambas de hasta un 3.8%. Con el resto de procedimientos, aunque las diferencias no son tan extremas, sí son siempre favorables a las redes MLP.

Centrándonos en los datos obtenidos mediante la clasificación con red MLP, el error de clasificación es claramente superior con el procedimiento de codificación, en tanto que la eliminación de registros *(listwise)* ofrece el menor error. En un punto intermedio, la imputación directa (MID) y la codificación con inclusión de variables indicadoras, son los métodos más efectivos si no se quiere (o puede) eliminar registros.

En lo referente al análisis de los valores específicos de error, con redes MLP éstos se hallan entre el 16% y el 17%, superiores en aproximadamente un 4% a los obtenidos en las matrices 101-200, como cabe esperar dado que la correlación con la variable dependiente es inferior. Un análisis más exhaustivo pone de manifiesto que, en la clasificación con regresión, las diferencias entre ambos conjuntos de matrices es inferior al mencionado 4% comentado para redes MLP, siendo de aproximadamente un 2%, lo que sugiere que las redes neuronales MLP son mejores que la regresión logística para captar un pequeño incremento de la correlación con la variable dependiente.

Tabla 44. Media de la tasa nominal de error (IC 95%) en la clasificación de la variable dependiente en las matrices MID, LIST, CODI, COVI, COMP (201-300)

CLAS.	MID	LIST	CODI	COVI	COMP
REGR.	0.234 (0.231-0.237)	0.225 (0.222-0.229)	0.282 (0.278-0.287)	0.229 (0.226-0.232)	0.217 (0.214-0.220)
MLP	0.169 (0.164-0.175)	0.160 (0.154-0.166)	0.192 (0.185-0.200)	0.169 (0.163-0.176)	0.162 (0.158-0.167)
RBF	0.179 (0.173-0.185)	0.165 (0.160-0.171)	0.222 (0.215-0.228)	0.207 (0.200-0.214)	0.169 (0.164-0.175)



Las tablas que se hallan a continuación (Tabla 45 a Tabla 50) presentan la tasa nominal de error de la clasificación en las matrices BU_RE, BU_ML, BU_RB hasta CL_RE, CL_EM, CL_ML, CL_RB (matrices 201-300).

Se mantiene la mayor capacidad de las redes neuronales respecto a la regresión, especialmente de las redes MLP, que de manera estable se muestran algo superiores a las redes RBF.

Como sucede en los dos conjuntos de matrices anteriormente analizados, e incluso más acentuadamente, apenas hay diferencias en función del tipo de variable imputada y del método de imputación aplicado. Unicamente la previa imputación de la variable CL incrementa invariablemente el error de clasificación, aunque tan solo en aproximadamente un 1%.

Tabla 45. Media de la tasa nominal de error (IC 95%) en la clasificación de la variable dependiente en las matrices BU_RE, BU_ML, BU_RB (201-300)

CLAS.	BU_RE	BU_ML	BU_RB
REGR.	N.A.	0.234 (0.231-0.237)	0.234 (0.231-0.237)
MLP	0.171 (0.166-0.176)	0.169 (0.163-0.175)	0.169 (0.164-0.174)
RBF	0.179 (0.173-0.184)	0.177 (0.171-0.183)	0.179 (0.173-0.185)

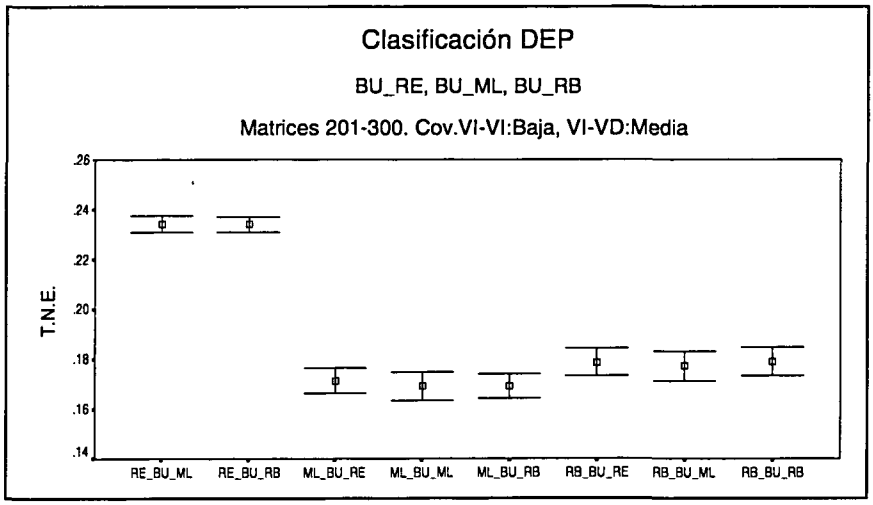


Tabla 46. Media de la tasa nominal de error (IC 95%) en la clasificación de la variable dependiente en las matrices BA_RE, BA_ML, BA_RB (201-300)

CLAS.	BA_RE	BA_ML	BA_RB
REGR.	N.A.	0.225 (0.222-0.229)	0.226 (0.223-0.229)
MLP	0.169 (0.164-0.175)	0.170 (0.165-0.175)	0.169 (0.164-0.175)
RBF	0.179 (0.173-0.185)	0.178 (0.172-0.183)	0.178 (0.173-0.184)

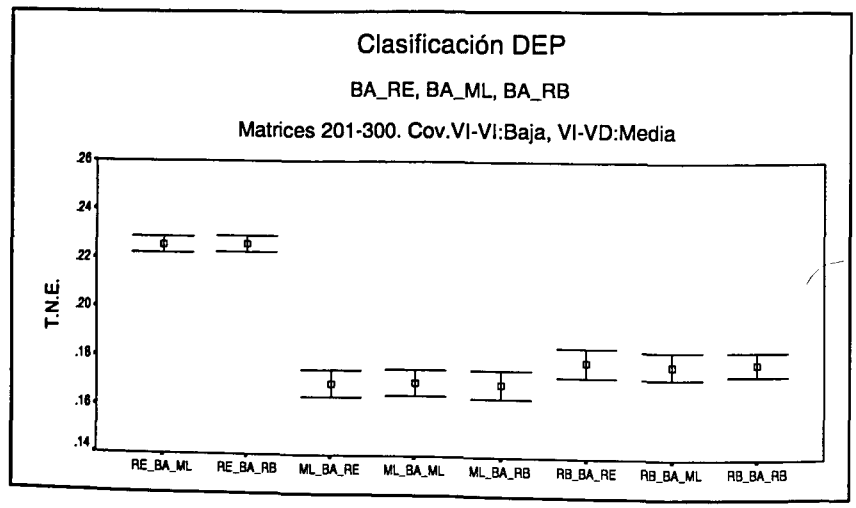


Tabla 47. Media de la tasa nominal de error (IC 95%) en la clasificación de la variable dependiente en las matrices OU_RE, OU_ML, OU_RB (201-300)

CLAS.	OU_RE	OU_ML	OU_RB
REGR.	N.A.	0.235 (0.232-0.238)	0.234 (0.231-0.238)
MLP	0.171 (0.166-0.176)	0.171 (0.166-0.176)	0.174 (0.169-0.179)
RBF	0.177 (0.172-0.182)	0.178 (0.173-0.183)	0.179 (0.173-0.184)

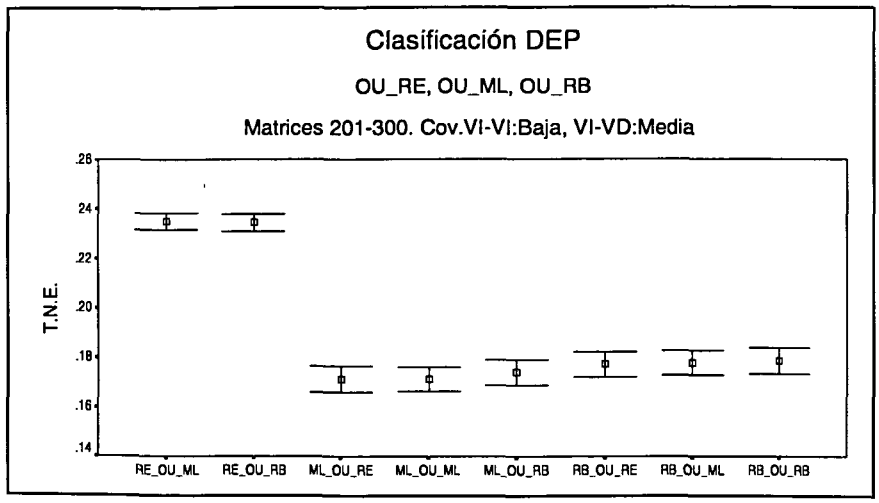


Tabla 48. Media de la tasa nominal de error (IC 95%) en la clasificación de la variable dependiente en las matrices OA_RE, OA_ML, OA_RB (201-300)

CLAS.	OA_RE	OA_ML	OA_RB
REGR.	N.A.	0.224 (0.220-0.227)	0.224 (0.220-0.227)
MLP	0.172 (0.166-0.178)	0.171 (0.165-0.176)	0.170 (0.164-0.175)
RBF	0.178 (0.172-0.184)	0.179 (0.173-0.185)	0.177 (0.172-0.183)

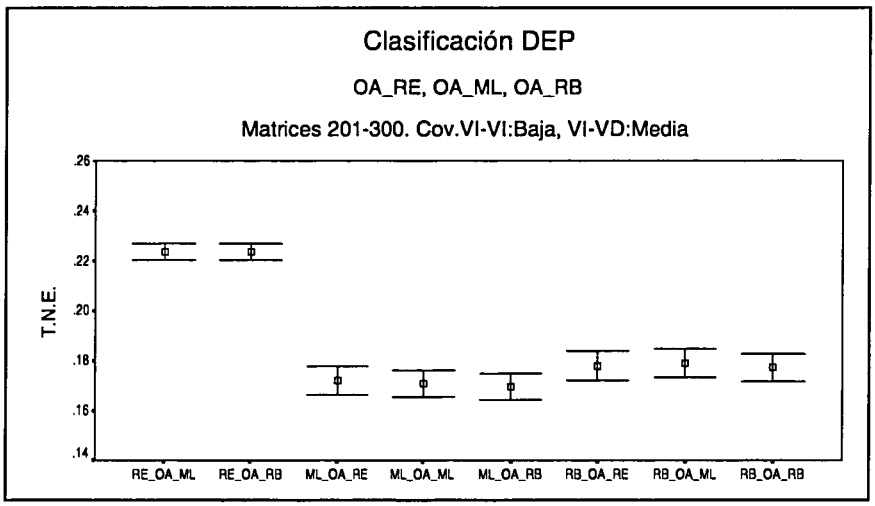


Tabla 49. Media de la tasa nominal de error (IC 95%) en la clasificación de la variable dependiente en las matrices CN_RE, CN_EM, CN_ML, CN_RB (201-300)

CLAS.	CN_RE	CN_EM	CN_ML	CN_RB
REGR.	N.A.	N.A.	0.234 (0.231-0.237)	0.225 (0.221-0.228)
MLP	0.173 (0.167-0.178)	0.171 (0.165-0.176)	0.172 (0.167-0.178)	0.172 (0.167-0.177)
RBF	0.180 (0.174-0.186)	0.180 (0.174-0.185)	0.178 (0.172-0.184)	0.179 (0.173-0.185)

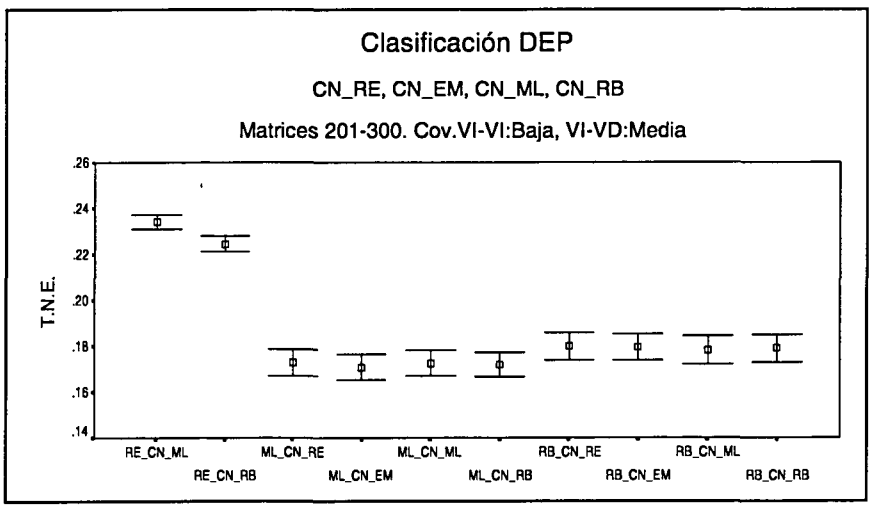
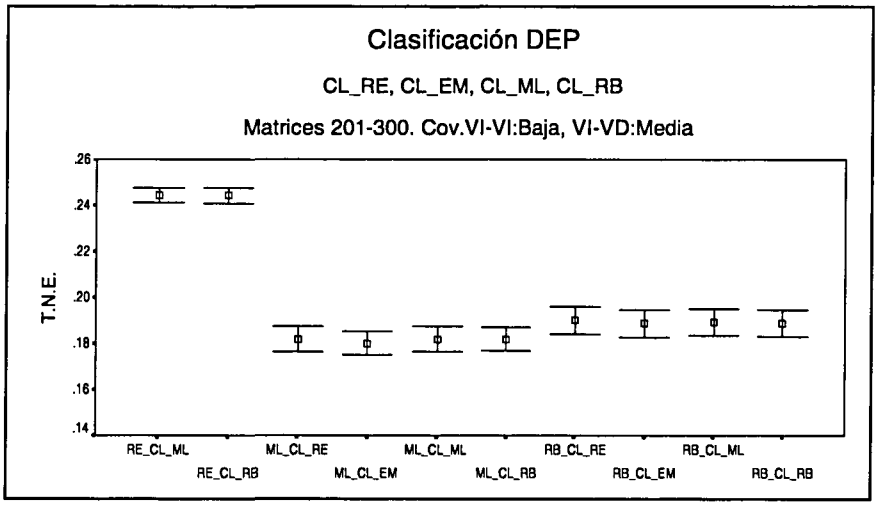


Tabla 50. Media de la tasa nominal de error (IC 95%) en la clasificación de la variable dependiente en las matrices CL_RE, CL_EM, CL_ML, CL_RB (201-300)

CLAS.	CL_RE	CL_EM	CL_ML	CL_RB
REGR.	N.A.	N.A.	0.244 (0.241-0.248)	0.244 (0.241-0.248)
MLP	0.182 (0.176-0.187)	0.180 (0.175-0.185)	0.182 (0.177-0.187)	0.182 (0.177-0.187)
RBF	0.190 (0.184-0.196)	0.189 (0.183-0.195)	0.189 (0.183-0.195)	0.189 (0.183-0.195)



6.3.4. Influencia del porcentaje de valores faltantes

Puesto que, en general, no se han hallado diferencias relevantes en los resultados de la clasificación ni en función del tipo de variable imputada ni del tipo de imputación por regresión/red efectuada, y teniendo en cuenta además que, ante un problema de clasificación, lo más sencillo en la práctica es utilizar alguna de las técnicas de estimación directa de los parámetros del modelo de clasificación, sólo se estimó un subconjunto de modelos de clasificación en las matrices con un 12% de datos missing. En concreto, se analizaron las matrices en que los valores faltantes fueron:

- Imputados con el mejor método de imputación directa (MID).
- Eliminados los registros incompletos (LIST).
- Codificados al valor 99 (CODI).
- Codificados al valor 99 e incluidas variables indicadoras (COVI).

Para facilitar la comparación se incluye también el resultado de la clasificación en las matrices completas (COMP), aunque ya se ha presentado en los apartados anteriores.

Las matrices empleadas en estos análisis son las que presentan correlación no nula entre las variables independientes y la dependiente, es decir las matrices MM3_101 a MM3_200 y MM3_201 a MM3_300, ya que, si un incremento del porcentaje de valores missing repercute sobre el resultado de la clasificación, sólo se manifestará cuando los datos estén relacionados. En la Tabla 51 se presentan los resultados para las matrices 101-200 y en la Tabla 52 para las matrices 201-300.

Tabla 51. Media de la tasa nominal de error (IC 95%) en la clasificación de la variable dependiente en las matrices MM3_101-MM3_200 MID, LIST, CODI, COVI, COMP

CLAS.	MID	LIST	CODI	COVI	COMP
REGR.	0.233 (0.229-0.236)	0.207 (0.202-0.213)	0.263 (0.259-0.267)	0.246 (0.243-0.250)	0.213 (0.209-0.216)
MLP	0.147 (0.143-0.152)	0.111 (0.102-0.119)	0.153 (0.147-0.159)	0.148 (0.142-0.153)	0.129 (0.123-0.134)
RBF	0.150 (0.145-0.155)	0.134 (0.127-0.141)	0.159 (0.154-0.164)	0.150 (0.145-0.155)	0.135 (0.130-0.139)

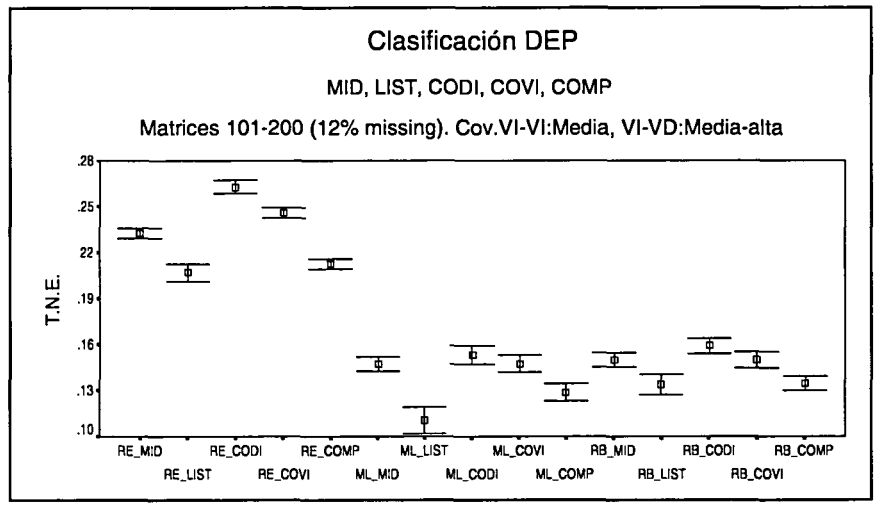
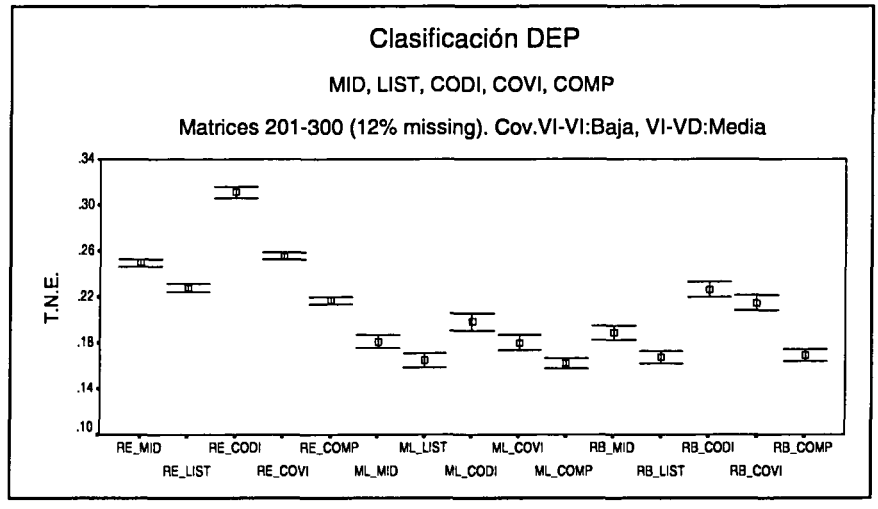


Tabla 52. Media de la tasa nominal de error (IC 95%) en la clasificación de la variable dependiente en las matrices MM3_201-MM3_300 MID, LIST, CODI, COVI, COMP

CLAS.	MID	LIST	CODI	COVI	COMP
REGR.	0.250 (0.246-0.254)	0.228 (0.222-0.234)	0.311 (0.308-0.314)	0.256 (0.253-0.259)	0.217 (0.214-0.220)
MLP	0.181 (0.175-0.187)	0.165 (0.158-0.172)	0.198 (0.193-0.202)	0.180 (0.176-0.184)	0.162 (0.158-0.167)
RBF	0.189 (0.183-0.196)	0.167 (0.160-0.174)	0.227 (0.223-0.231)	0.215 (0.210-0.220)	0.169 (0.164-0.175)



La técnica de eliminación de registros ofrece de nuevo el menor error de clasificación, que prácticamente coincide con el obtenido en las matrices de datos con un 3% de valores faltantes. Dejando de lado el método *listwise*, todos los porcentajes de error obtenidos en las matrices MM3 son superiores a los correspondientes calculados en las matrices MM1. Un estudio más detallado pone de manifiesto que con la imputación directa (MID) y la codificación de missings con inclusión de variables indicadoras (COVI) se consiguen los mejores resultados, que aumentan respecto a los obtenidos en las matrices MM1 de forma diferente según la técnica de clasificación empleada. En las matrices 101-200, el aumento medio al clasificar mediante regresión logística es del 2.8%, mediante red MLP del 1.2% y mediante red RBF del 0.8%, mientras que en las matrices 201-300, el incremento del porcentaje de valores missing aumenta el error en un 2.4%, 0.9% y 0.7%, cuando las técnicas de clasificación son regresión logística, red MLP y red BRF respectivamente.

6.4. CONCLUSIONES Y DISCUSIÓN

En los objetivos planteados al inicio del experimento de simulación se distinguen dos líneas principales de estudio: (1) el análisis de la imputación de valor a los datos faltantes y (2) el estudio de un problema de clasificación en una matriz de datos incompletos. En ambas cuestiones, las conclusiones presentadas son válidas cuando los valores missing son aleatorios (*MAR*) y, en el problema de clasificación, sólo se hallan valores faltantes en las variables independientes.

En lo referente a la evaluación del error cometido al imputar valor a los datos missing, nuestros resultados sugieren que con la imputación por regresión/red se consigue el menor error para cualquier tipo de variable y nivel de covariación, si bien es cierto que el costo computacional y la dedicación requeridas por este tipo de imputación son mucho más elevados que los necesarios al realizar una imputación directa. Por ello, y teniendo presente también que en matrices de datos en que las variables no están correlacionadas la imputación directa ofrece porcentajes de error similares a la imputación por regresión/red, para seleccionar el mejor método, definido como una combinación del error de imputación y de su coste de aplicación, hay que tener en consideración el nivel de correlación entre las variables de la matriz, lo que se traduce en la realización, en primer lugar, de un estudio de los patrones de covariación entre los datos, como mínimo a nivel bivariante. Cuando el nivel de correlación entre las variables es prácticamente nulo, la imputación directa de la moda en las variables categóricas, y de la media en las cuantitativas, es el procedimiento de elección. Por contra, si las variables presentan un cierto nivel de relación entre sí, como suele ocurrir en la práctica, la imputación mediante regresión/red es aconsejable. En esta última situación, si se trata de variables procedentes de una distribución no equiprobable (categóricas) o asimétrica (cuantitativas), tanto la imputación por regresión como por red

neuronal artificial ofrecen resultados similares, mientras que si se imputan variables procedentes de una distribución equiprobable (categóricas) o normal (cuantitativas), es recomendable la imputación mediante red neuronal. En la elección final se debe tener presente que las redes neuronales no establecen ningún tipo de supuesto sobre la distribución de los datos, mientras que el modelo de regresión y el algoritmo EM requieren la comprobación empírica de que la distribución poblacional de las variables se ajusta a un determinado modelo probabilístico, aspecto que, como afirman Graham, Hofer y Mackinnon (1996) difícilmente puede ser corroborado a partir de los datos disponibles en una muestra.

Respecto a la arquitectura de red óptima, las redes MLP y RBF tienen un comportamiento similar en varias situaciones. Sin embargo, en otras condiciones la imputación mediante red RBF es preferible, tanto por el hecho de que ofrece mejores resultados como porque tiene un menor coste computacional. La Tabla 53 presenta el método de imputación recomendado en función del nivel de covariación entre los datos y del tipo de variable imputada. Recuerde que para elegir el mejor método se ha considerado, en primer lugar, su error de imputación asociado, en segundo lugar, el coste computacional y humano que requiere su ejecución y, en tercer lugar, las restricciones que realiza sobre la especificación de la distribución poblacional de los datos. Si en una celda de la tabla hay más de un método todos ellos se consideran similares, a pesar de lo cual, el orden en que aparecen refleja nuestra preferencia.

Las conclusiones referentes a las variables categóricas no equiprobables pueden verse modificadas por el nivel de no equiprobabilidad. En nuestros datos, la variable binaria no equiprobable procede de una población caracterizada por π=0.8, y la variable ordinal no equiprobable se generó a partir de una población con π_1=0.5, π_2=0.35, π_3=0.15. A medida que estas proporciones tiendan a valores equiprobables, las correspondientes conclusiones se aproximarán a las presentadas para variables con distribución equiprobable. Algo similar sucede con la variable cuantitativa asimétrica y su nivel de asimetría.

Tabla 53. Mejor método de imputación de los valores faltantes en función del nivel de correlación entre variables y del tipo de variable a imputar

		Nivel de correlación		
		Nulo	Bajo	Medio-alto
Tipo variable	Binaria equiprobable	MODA	RBF	MLP / RBF
	Binaria no equiprobable	MODA	RBF / REG	RBF / REG
	Ordinal equiprobable	MODA	MLP / RBF	RBF
	Ordinal no equiprobable	MODA	REG	RBF / MLP / REG
	Cuantitativa normal	MEDIA	RBF / MLP	RBF / MLP
	Cuantitativa asimétrica	MEDIA	RBF / MLP / REG	RBF / REG / EM

REG:Regresión ; MLP:Red Perceptrón multicapa
EM:Algoritmo EM ; RBF:Red de función base radial

En la literatura revisada no hemos hallado ningún trabajo que evalúe los diferentes métodos de imputación en función del tipo de variable y del nivel de correlación de los datos. Sólo Prechelt (1994b) hace referencia a la naturaleza de la variable en el momento de codificar los valores missing en las variables independientes de una red neuronal, si bien, al tratarse de una primera comunicación a una lista de correo, plantea más preguntas que respuestas ofrece.

En una matriz con datos faltantes en más de una variable, la imputación por regresión/red de los valores missing de una determinada variable requiere de la anterior imputación de los missing de las variables que actúan como independientes, que se acostumbra a realizar mediante algún procedimiento de imputación directa, normalmente la media o la moda. Las consecuencias de esta imputación anterior se manifiestan, en forma de un incremento del error, principalmente si la correlación entre las variables es media-alta y la variable imputada es binaria u ordinal equiprobable, o cuantitativa asimétrica. En estas situaciones, el error asociado a la imputación de la variable de interés se ve incrementado por el error cometido en la imputación de las variables que actúan como independientes.

En lo referente al estudio de la clasificación de una variable binaria a partir de un conjunto de variables independientes con datos incompletos, a un primer nivel se deben distinguir dos situaciones:

En primer lugar, cuando en el conjunto de variables independientes hay una determinada variable con una especial relevancia en el diseño de la investigación, y dicha variable presenta un nivel de correlación medio-alto con otras variables, la imputación de sus valores faltantes se debe realizar mediante un procedimiento de imputación por regresión/red, ya que así se consigue disminuir el sesgo en el efecto que dicha variable tiene sobre la dependiente, reduciéndose además de forma colateral el error de la clasificación. Para ello, los valores missing en el resto de variables han de ser imputados mediante el mejor procedimiento de imputación directa, que como se comenta en las conclusiones anteriores, es la moda en variables categóricas y la media en variables cuantitativas. La elección del tipo de imputación por regresión/red en un problema de clasificación de tipo explicativo, con un nivel de correlación medio o alto entre las variables independientes, depende del tipo de variable cuyos valores missing se imputen (ver Tabla 54). Cuando la variable en estudio no se halle correlacionada con otras variables registradas, la imputación directa de sus valores faltantes es el procedimiento recomendado.

Tabla 54. Mejor método de imputación por regresión/red en un problema de clasificación con correlación media o *alta entre las variables independientes.*

Tipo de variable					
Binaria equiprobable	Binaria no equiprobable	Ordinal equiprobable	Ordinal no equiprobable	Cuantitativa normal	Cuantitativa asimétrica
MLP	REG	RBF / MLP	MLP / REG	MLP	EM / REG

Por otra parte, cuando el modelo de clasificación es eminente predictivo y, por tanto, no hay una variable independiente de particular interés, se deben distinguir dos situaciones: si los valores ausentes se dan predominantemente en unas determinadas variables que se hallan correlacionadas con otras variables predictoras, éstas deben ser imputadas por regresión/red (ver Tabla 54) y el resto por imputación directa. Si, por contra, los valores perdidos se reparten más o menos equilibradamente entre todas las variables, los métodos de estimación directa de los parámetros del modelo de clasificación son aconsejables por su simplicidad de uso. En este último caso, el menor error de clasificación se obtiene siempre con el método de eliminación de los registros que tienen algún valor missing (*listwise*). Una explicación a la superioridad de la técnica *listwise* se puede hallar en la estricta aleatoriedad otorgada a la asignación de los datos

missing en nuestro experimento, ya que en estas circunstancias, como señalan Little y Rubin (1987), la eliminación de los registros con datos incompletos no implica un sesgo de las estimaciones. Además, parece lógico que un método que se caracteriza por aniquilar el problema en lugar de solucionarlo ofrezca un error de clasificación global bajo. A mi juicio, los inconvenientes del uso del método *listwise*, comentados en el primer capítulo, son la reducción en la precisión de las estimaciones de los parámetros poblacionales y la pérdida de potencia en las pruebas estadísticas de significación. En este sentido, se puede considerar un primer indicio la mayor amplitud de los intervalos de confianza de la tasa nominal de error calculados con el método *listwise* respecto a los otros procedimientos. A primera vista, nuestros resultados son diametralmente opuestos a los obtenidos por Pitarque y Ruiz (1996), quienes concluyen que con la técnica *listwise* se obtiene el peor error de clasificación. Sin embargo, las condiciones de la simulación realizada permiten explicar las discrepancias. Pitarque y Ruiz (1996) trabajan con matrices de datos con 15 variables independientes y 100 registros, de los cuales la mitad tienen dos valores ausentes, de manera que al eliminar los registros incompletos disponen de 50 casos para estimar 103 parámetros (implicados en la red MLP 15-6-1 que emplean). Por contra, nuestras matrices de datos contienen 500 registros y 6 variables independientes, con un 3% de valores faltantes en cada variable. En el peor de los casos (ningún registro con 2 valores missing), ello se traduce en la existencia de 90 registros con datos incompletos, lo que supone que quedan 410 casos para estimar 57 parámetros (implicados en la red MLP 12-4-1 que empleamos en la mayoría de modelos).

Dejando de lado el método *listwise*, la imputación directa de los datos ausentes de cada variable y la codificación de los valores missing con la inclusión de variables indicadoras, ofrecen los mejores resultados en matrices de datos con correlación media o media-alta con la variable dependiente. Nuestros resultados coinciden con los hallados por Ding, Denoeux y Helloco (1993), quienes exponen la superioridad del uso de variables indicadoras frente a la imputación de valor, y con los obtenidos por Vamplew y Adams (1992), quienes, al clasificar semillas en uno de diez grupos a partir de siete variables independientes, obtienen los mejores porcentajes de clasificación con la inclusión de variables indicadoras y mediante la imputación de los valores faltantes con una red neuronal. En un contexto psicopatológico, con el objetivo de clasificar sujetos como patológicos o no a partir de la presencia/ausencia de determinados síntomas, Taylor y Amir (1994) deducen que el empleo de variables indicadoras es una solución óptima, aunque tanto como una solución más sencilla, consistente en imputar a cada valor missing un código intermedio entre los que representan la presencia y la ausencia del síntoma.

La simple codificación de los datos faltantes a un valor fuera del rango de valores válidos, sin la inclusión de variables indicadoras, no es por si sola efectiva. La incorporación de variables indicadoras incrementa la eficacia de este método, dependiendo dicho incremento de la técnica de clasificación empleada: cuando la clasificación se realiza mediante regresión logística, la inclusión de variables indicadoras es más efectiva que cuando se clasifica a través de una red neuronal.

En muchos estudios se ha constatado la superioridad de las redes neuronales artificiales frente a otras técnicas de clasificación: Huang y Lippman (1987), Sethi y Otten (1990), Bonelli y Parodi (1991) comparan modelos tradicionales y redes neuronales, obteniendo resultados ligeramente favorables a estas últimas. Furlanello, Giuliani y Trentin (1995) corroboran la mejor capacidad de discriminación de las redes RBF frente a la de modelos lineales multivariables. Ripley (1993) concluye que las redes MLP son equivalentes a complejas técnicas estadísticas como la *projection pursuit regression*, Garson (1991) confirma la superioridad de las redes frente a los modelos de regresión, *path analysis* y análisis discriminante, y Schrodt (1993), trabajando con variables fuertemente modales, determina que las redes neuronales son más eficaces que los algoritmos de inducción de reglas y el análisis discriminante. En otros trabajos, por contra, no se ha podido establecer la superioridad de una técnica en concreto (Michie, Spiegelhalter y Taylor, 1994; Thrun, Mitchell y Cheng, 1994). Nuestras conclusiones a este respecto son claramente favorables al uso de las redes neuronales, en detrimento de la regresión logística, en un problema de clasificación. Incluso cuando el nivel de correlación con la variable dependiente es prácticamente nulo, el error asociado a las redes es levemente inferior al de la regresión. La diferencia entre ambos modelos se acrecienta a medida que aumenta la correlación con la variable dependiente, indicando la mayor capacidad de las redes neuronales para captar un incremento del nivel de relación en los datos.

Las diferencias entre las dos topologías de red analizadas son mínimas en la mayoría de situaciones, pero siempre a favor de las redes MLP. Considerando la potencia actual de los ordenadores y el crecimiento exponencial que están experimentando, creemos que el superior coste computacional de las redes MLP frente a las RBF no puede ser un factor relevante a la hora de decidirse por una u otra arquitectura, lo que nos conduce a recomendar el empleo de las redes perceptrón multicapa en un problema de clasificación como el planteado.

Las conclusiones establecidas hasta el momento son válidas para matrices de datos con un porcentaje bajo de datos faltantes (alrededor del 3%). Independientemente de la técnica de análisis de datos incompletos, los resultados respecto a la influencia del porcentaje de valores missing presentes en los datos sugieren que, manteniendo constante el nivel de correlación con la variable

dependiente en un valor medio o medio-alto, la capacidad discriminante de un modelo de clasificación se reduce a medida que aumenta el porcentaje de datos desconocidos, como demuestra el mayor error cometido en las matrices con un 12% de valores missing en las variables independientes. También se aprecia una tendencia a que cuanto más alto es el nivel de correlación con la variable dependiente, más se ve afectado el error de clasificación por un incremento del porcentaje de datos missing. Por último, la disminución de la capacidad predictiva es mayor cuando se clasifica mediante regresión logística respecto a cuando se hace con redes neuronales, y dentro de éstas, las redes MLP se ven algo más afectadas que las redes RBF. Todo ello hace sospechar que las redes neuronales, y especialmente las redes RBF, son más resistentes al incremento del porcentaje de valores faltantes que la regresión logística.

Respecto a la selección de la mejor estrategia de afrontamiento de los valores missing, ésta no parece estar afectada por su cantidad. Así, con un 12% de valores desconocidos, nuevamente con la eliminación de registros se obtienen los mejores resultados, seguido de la imputación directa y de la codificación con inclusión de variables indicadoras. Obviamente, si el método *listwise* se descartó en las conclusiones anteriores por las consecuencias derivadas de la reducción del tamaño muestral que comporta, también se debe descartar cuando el número de sujetos que se deberían eliminar se incrementa (como acostumbra a ocurrir cuando el número de datos perdidos se eleva). Por todo ello, ante un problema de clasificación de tipo predictivo, con un elevado porcentaje de valores faltantes distribuidos en diversas variables, la estrategia óptima es imputar los valores missing con el mejor método de imputación directa y, posteriormente, establecer el modelo de clasificación con una red neuronal MLP o RBF.

La estimación de diferentes modelos de clasificación con las matrices de datos completas permite estudiar cómo se ve afectada la capacidad discriminante por la presencia de datos incompletos. Con un porcentaje de valores faltantes bajo (3% en nuestro experimento), las mejores estrategias de análisis de datos incompletos dan resultados similares a los obtenidos con datos completos, mientras que cuando el porcentaje de valores missing es elevado (12% en nuestro experimento), cualquier procedimiento de análisis sobre las matrices incompletas da peores resultados que el análisis de la matriz completa. De ello se deduce que las técnicas presentadas para afrontar el problema de los datos incompletos son efectivas cuando hay un número reducido de valores perdidos, mientras que, en caso contrario, cualquier procedimiento se muestra incapaz de compensar totalmente la ausencia de información provocada por la pérdida de datos.

7
Aplicación a un estudio de psicopatología infantil

7.1. PRESENTACIÓN DEL ESTUDIO

Las conclusiones obtenidas en el experimento de simulación se han aplicado a una matriz de datos reales, obtenida a partir de un amplio estudio para evaluar diferentes entrevistas y cuestionarios de uso habitual en psicopatología infantil.

Nuestro objetivo es establecer un modelo predictivo de la opinión de los padres sobre la necesidad de ayuda psicológica de su hijo. Para ello hemos seleccionado un amplio conjunto de variables predictoras, que en los posteriores listados se identificarán con el nombre indicado entre paréntesis:

- Sexo (*Sexo*): Codificado como 0 (Masculino) y 1 (Femenino).
- Edad del niño (*Edad*). Edad en años cumplidos.
- Nivel socioeconómico (*SES*). Fue determinado con el Índice de Posición Social de Hollingshead (1975), y codificado como 1 (Medio a alto), 2 (Medio-bajo) y 3 (bajo).
- Antecedentes familiares de psicopatología (*AP*). Se obtuvo a partir de un cuestionario sociodemográfico aplicado a los padres. Fue codificado como 0 (No) y 1 (Sí).
- Problemas conductuales tempranos (*PCT*): La entrevista *Diagnostic Interview for Children and Adolescent* (DICA-R; Reich, Shayka y Taibleson, 1991), aplicada a los padres, recoge información sobre los problemas conductuales del niño en la infancia temprana. Esta variable se codificó como 1 (sin problemas), 2 (uno o dos problemas) y 3 (más de dos problemas).
- Estilo educativo. El *Egna Minnen av Barndoms Uppfostran* (EMBU; Perris, Jacobson, Lindström, Knorring y Perris, 1980) es un cuestionario originalmente desarrollado para evaluar los recuerdos del adulto sobre el estilo educativo de sus padres. Castro, Toro, Arrindell, Van Der Ende y Puig (1990) presentan una modificación del cuestionario para registrar el estilo educativo de los padres en el momento presente. Estos autores desarrollaron tres versiones diferentes: para niños, para adolescentes y para los padres del niño/adolescente. Posteriormente, Castro, Toro, Van Der Ende y Arrindell (1993) realizaron un estudio psicométrico en la población española, poniendo

de manifiesto la existencia de cuatro dimensiones referentes al estilo educativo del padre y de la madre: rechazo, calor emocional, sobreprotección y favorecedor del sujeto. En nuestro estudio, cada una de estas escalas es obtenida de tres informantes diferentes: el padre, la madre y el hijo, evaluando este último por separado el estilo educativo del padre y el de la madre. Así, hay un total de 16 variables que miden estilo educativo:

Tabla 55. Variables que miden estilo educativo.

	Informante			
Escala	Padre	Madre	Niño (sobre padre)	Niño (sobre madre)
Rechazo	*Recha_p*	*Recha_m*	*Recha_np*	*Recha_nm*
Calor emocional	*Calor_p*	*Calor_m*	*Calor_np*	*Calor_nm*
Sobreprotección	*Prote_p*	*Prote_m*	*Prote_np*	*Prote_nm*
Favorecedor	*Favor_p*	*Favor_m*	*Favor_np*	*Favor_nm*

Respecto a la variable dependiente (Dep), refleja la opinión de los padres sobre la necesidad de ayuda del hijo. Esta pregunta, que se obtiene mediante la entrevista DICA-R, se ha codificado con los valores 0 (no necesita ayuda) y 1 (sí necesita ayuda). Al estudiar la distribución de frecuencias de la variable *Dep* (ver Fig. 26) se llega a la conclusión de que puede provenir de un modelo de equiprobabilidad.

¿Hijo necesita ayuda?

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	No	73	36 5	36.5	36.5
	Sí	127	63 5	63 5	100.0
	Total	200	100.0	100.0	
Total		200	100.0		

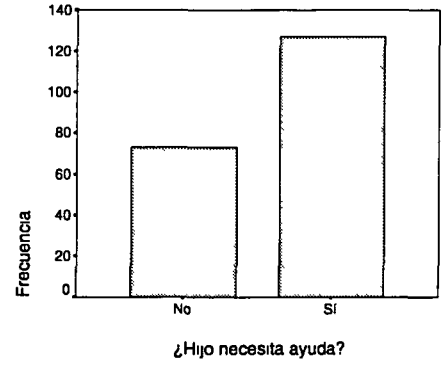


Fig. 26. Distribución de frecuencias de la variable dependiente

7.2. ANÁLISIS DE LOS VALORES FALTANTES

El análisis de los datos comienza con un exhaustivo estudio de los valores missing. En primer lugar se ha evaluado la aleatoriedad de las observaciones (OAR), para lo cual se ha generado una variable indicadora de la presencia/ausencia de valor asociada a cada variable con valores ausentes (más de un 5%), y se han comparado las medias de cada variable cuantitativa en los grupos formados por las citadas variables indicadoras. Los resultados (ver Anexo) avalan la hipótesis de que los valores registrados son observaciones aleatorias. También se ha estudiado la distribución de los valores ausentes en las categorías de cada variable categórica, no hallándose tampoco diferencias relevantes. Por otra parte, se asume que los valores missing son aleatorios, aspecto que, como se comenta en el capítulo 1, no puede ser empíricamente evaluado.

En la Tabla 56 se presentan los porcentajes de valores perdidos en cada variable (diagonal principal), junto al porcentaje de casos en que una de las variables es desconocida y la otra no. Se observa que el mayor porcentaje de valores faltantes se da en las variables que miden el estilo educativo desde el punto de vista del padre, y especialmente en la escala *Favor_p*, que es desconocida en un 29% de la muestra. Además, estos valores perdidos se concentran en los mismos sujetos, como se deduce del hecho de que, en las variables *Prote_p*, *Calor_p* y *Recha_p* no hay ningún registro en que una de las variables sea missing y alguna de las otras dos no. Por otra parte, las variables *Favor_m* y *PCT* también destacan al tener un 11% y 6% de valores perdidos respectivamente.

Tabla 56. Porcentaje de valores missing por variable

Porcentaje de valores missing [a,b]

	PCT	FAVOR_M	FAVOR_P	PROTE_P	CALOR_P	RECHA_P
PCT	6 00					
FAVOR_M	16 00	11 00				
FAVOR_P	30 00	19 00	29 00			
PROTE_P	21 50	26 50	9 50	19 50		
CALOR_P	21 50	26 50	9 50	00	19 50	
RECHA_P	21 50	26 50	9 50	00	00	19 50

No se incluyen las variables con menos de un 5% de valores missing

a En la diagonal se halla el porcentaje de valores missing de la variable

b Fuera de la diagonal se halla el porcentaje de casos en que una de las variables es missing y la otra no

Los patrones de valores missing presentes en los datos se hallan en la Tabla 57. Sólo un 62.5% (125) de sujetos tienen valor en las 21 variables independientes registradas, lo que indica que en un análisis convencional con eliminación de registros incompletos el tamaño muestral se vería sustancialmente reducido. La combinación de valores ausentes más frecuente corresponde a la falta de valor en las 4 dimensiones que evalúan el estilo educativo según el padre, combinación presente en un 14.5% (29) de registros. También se observa un elevado porcentaje de casos (7.5%) con valor desconocido en la combinación *Favor_p-Favor_m*. Los restantes patrones de datos missing se dan en muy pocos sujetos.

Tabla 57. Patrones de valores missing

Patrones de valores missing

Nº de casos	Patrones de missing[a]																				Completo si...[b]
	EDAD	FAVOR_NP	SEXO	FAVOR_NM	PROTE_M	CALOR_M	RECHA_M	CALOR_NP	PROTE_NP	RECHA_NP	RECHA_NM	CALOR_NM	PROTE_NM	AP	SES	PCT	FAVOR_M	FAVOR_P	RECHA_P	PROTE_P	CALOR_P	
125																						125
2																		X				127
15																	X	X				142
1															X		X	X				146
1															X	X	X	X				151
4																X						129
1														X		X						131
1														X								126
1														X	X							130
3															X							128
1														X	X		X					131
1					X	X										X						126
1							X									X						130
4																X		X	X	X	X	164
29																		X	X	X	X	156
2														X				X	X	X	X	159
1																	X	X	X	X	X	172
1										X	X						X	X	X	X	X	173
1									X	X							X	X	X	X	X	173
1						X	X										X	X	X	X	X	173
1				X							X	X	X		X							129
2								X	X	X	X	X	X									127
1								X	X	X	X	X	X			X						132

a. Las variables están ordenadas según los patrones de missing

b. Número de casos completos si las variables missing en ese patrón (marcadas con X) son eliminadas

Para estudiar la distribución de los valores faltantes por sujeto se ha generado una nueva variable que contiene el número de datos missing en cada registro. Los resultados se hallan en la Tabla 58. Ningún caso tiene más de un 33.3% (7) de valores perdidos, lo cual garantiza que la imputación por regresión/red sea fiable en todos los casos en que se efectúe. Cabe mencionar que un 15% de sujetos tienen 4 valores perdidos, que corresponden en casi su totalidad a las ya comentadas variables obtenidas con el padre como informante. A partir de los datos de la Tabla 58, y teniendo en cuenta que se han registrado 22 variables (21 variables independiente y 1 dependiente) se puede obtener fácilmente el porcentaje total de valores desconocidos, que en nuestra muestra es del 5.77%.

Tabla 58. Número de valores faltantes por sujeto.

		Frecuencia	Porcentaje	Porcentaje acumulado
Válidos	0	125	62.5	62.5
	1	10	5.0	67.5
	2	19	9.5	77.0
	3	2	1.0	78.0
	4	30	15.0	93.0
	5	8	4.0	97.0
	6	2	1.0	98.0
	7	4	2.0	100.0
	Total	200	100.0	
Total		200	100.0	

7.3. ANÁLISIS DESCRIPTIVO

7.3.1. Univariante

En la Tabla 59 se hallan los estadísticos descriptivos de las variables independientes cuantitativas, y en la Tabla 60 las distribuciones de frecuencias de las variables independientes categóricas. Los correspondientes histogramas, gráficos de caja y diagramas de barras se presentan en el Anexo. En función de su naturaleza y la forma de su distribución empírica, las variables registradas se han clasificado en:

- Binaria equiprobable: Variable *Sexo*.
- Binaria no equiprobable: Variable *AP*.
- Ordinal equiprobable: Variable *SES*.
- Ordinal no equiprobable: Variable *PCT*.
- Cuantitativa normal: Variables *Prote_p, Prote_m, Calor_nm, Calor_np, Edad.*

- Cuantitativa asimétrica: *Recha_p, Recha_m, Recha_np, Recha_nm, Calor_p, Calor_m, Prote_np, Prote_nm, Favor_p, Favor_m, Favor_np, Favor_nm.*

Clasificación que determina el método de imputación recomendado que posteriormente será aplicado.

Tabla 59. Estadísticos descriptivos de las variables independientes cuantitativas.

Estadísticos descriptivos

	N	Mínimo	Máximo	Media		Desv. típ
	Estadístico	Estadístico	Estadístico	Estadístico	Error típico	Estadístico
Edad	200	6	17	11.96	.22	3.08
Rechazo (niño-padre)	195	11.00	83 00	25.4959	1.0806	15.0903
Calor emocional (niño-padre)	197	21.00	69.23	46.7900	.6599	9.2623
Sobreprotección (niño-padre)	196	10.00	51.00	24.9808	.6332	8 8647
Favorecedor (niño-padre)	200	5.00	14 00	6.7738	.1399	1 9788
Rechazo (niño-madre)	195	11.00	76.19	25.8493	1.0724	14 9750
Calor emocional (niño-madre)	196	20 00	67.00	48 0525	.5678	7.9499
Sobreprotección (niño-madre)	196	10 00	52 00	26.2705	.6549	9.1684
Favorecedor (niño-madre)	199	5 00	14.00	6 8982	.1458	2.0570
Rechazo (padre)	161	13.00	26.00	17.6612	.2543	3 2266
Calor emocional (padre)	161	32.00	68.00	53.7658	.6326	8 0263
Sobreprotección (padre)	161	22.80	67.56	38 8160	.5071	6 4344
Favorecedor (padre)	142	3.00	7.00	3.5775	8.441E-02	1 0058
Rechazo (madre)	198	13.00	29 00	17.4893	.2344	3 2979
Calor emocional (madre)	198	34.00	68 00	56.5834	.4975	7.0005
Sobreprotección (madre)	199	21.00	60.17	40.3852	.4878	6.8816
Favorecedor (madre)	178	3 00	12 00	3.7753	.1105	1 4749

Estadísticos descriptivos

	Asimetría		Curtosis	
	Estadístico	Error típico	Estadístico	Error típico
Edad	-.121	.172	-1.171	.342
Rechazo (niño-padre)	1.274	.174	1.652	.346
Calor emocional (niño-padre)	-.627	.173	.348	.345
Sobreprotección (niño-padre)	.672	.174	-.134	.346
Favorecedor (niño-padre)	1.083	.172	.804	.342
Rechazo (niño-madre)	1.099	.174	.873	.346
Calor emocional (niño-madre)	-.647	.174	.968	.346
Sobreprotección (niño-madre)	.532	.174	-.409	.346
Favorecedor (niño-madre)	1.084	.172	.765	.343
Rechazo (padre)	.498	.191	-.491	.380
Calor emocional (padre)	-.413	.191	-.358	.380
Sobreprotección (padre)	.645	.191	1.887	.380
Favorecedor (padre)	1.863	.203	2.836	.404
Rechazo (madre)	.900	.173	.895	.344
Calor emocional (madre)	-.668	.173	.158	.344
Sobreprotección (madre)	.217	.172	-.129	.343
Favorecedor (madre)	2 543	.182	7.784	.362

Tabla 60. Distribución de frecuencias de las variables independientes categóricas.

Sexo

		Frecuencia	Porcentaje	Porcentaje válido
Válidos	Masculino	97	48 5	48 5
	Femenino	103	51.5	51.5
	Total	200	100.0	100 0
Total		200	100 0	

Nivel socieconómico

		Frecuencia	Porcentaje	Porcentaje válido
Válidos	Medio a alto	62	31.0	32.3
	Medio-bajo	56	28.0	29.2
	Bajo	74	37.0	38.5
	Total	192	96.0	100.0
Perdidos	Perdidos del sistema	8	4.0	
	Total	8	4.0	
Total		200	100 0	

Problemas conductuales tempranos

		Frecuencia	Porcentaje	Porcentaje válido
Válidos	0	104	52.0	55.3
	1-2	57	28.5	30.3
	>2	27	13.5	14.4
	Total	188	94.0	100.0
Perdidos	Perdidos del sistema	12	6.0	
	Total	12	6.0	
Total		200	100 0	

Antecedentes psicopatología

		Frecuencia	Porcentaje	Porcentaje válido
Válidos	No	67	33.5	34.5
	Sí	127	63.5	65 5
	Total	194	97.0	100.0
Perdidos	Perdidos del sistema	6	3.0	
	Total	6	3.0	
Total		200	100 0	

7.3.2. Bivariante

Para estudiar el nivel de covariación entre las variables registradas en la matriz de datos se ha calculado el coeficiente de correlación de Spearman entre cada par de variables. Dado el elevado número de medidas registradas, y con el objetivo de resumir los resultados, en la Tabla 61 se presenta la media de la correlación de cada variable independiente con todas las demás, así como la media de la correlación de la variable dependiente con el conjunto de independientes. En el cálculo de dicha media se ha empleado el valor absoluto del coeficiente de correlación. En el Anexo del capítulo actual se hallan los resultados completos.

En general se han obtenido niveles de correlación de tipo medio-bajo entre las variables independientes, con valores comprendidos entre $\overline{\rho} = 0.056$, en la escala *Favorecedor (madre)* y $\overline{\rho} = 0.325$ en *Rechazo (niño-padre).* Las variables que miden el estilo educativo según el padre presentan valores comprendidos entre $\overline{\rho} = 0.119$ y $\overline{\rho} = 0.207$. Respecto a la variable dependiente, la correlación promedio con el conjunto de variables independientes es de $\overline{\rho} = 0.188$.

Tabla 61. Correlación media entre variables (Coeficiente de correlación de Spearman)

VARIABLE	Correlación Media
Edad	0.258
Sexo	0.109
Nivel socieconómico	0.085
Problemas conduct. Tempranos	0.126
Antecedentes psicopatología	0.120
Rechazo (niño-padre)	0.325
Calor emocional (niño-padre)	0.196
Sobreprotección (niño-padre)	0.266
Favorecedor (niño-padre)	0.139
Rechazo (niño-madre)	0.315
Calor emocional (niño-madre)	0.190
Sobreprotección (niño-madre)	0.274
Favorecedor (niño-madre)	0.111
Rechazo (padre)	0.199
Calor emocional (padre)	0.207
Sobreprotección (padre)	0.162
Favorecedor (padre)	0.119
Rechazo (madre)	0.174
Calor emocional (madre)	0.197
Sobreprotección (madre)	0.206
Favorecedor (madre)	0.056
¿Hijo necesita ayuda?	0.188

7.4. IMPUTACION DE LOS VALORES FALTANTES

Como se ha comentado anteriormente, en la matriz de datos analizada el mayor porcentaje de valores perdidos se halla en las variables provenientes de la entrevista con el padre, que presentan un nivel de correlación moderado con el resto de variables independientes registradas. Ante esta situación, la estrategia de análisis que hemos seleccionado consiste en la imputación mediante red RBF de los valores faltantes en dichas variables, previa imputación directa del resto de variables independientes. Así, en la Tabla 62 se especifica la técnica de imputación aplicada a cada variable independiente que tiene valores faltantes.

Tabla 62. Método de imputación de cada variable independiente que presenta valores faltantes

VARIABLE	MÉTODO DE IMPUTACIÓN
Nivel socieconómico	Moda
Problemas conductuales tempranos	Moda
Antecedentes psicopatología	Moda
Rechazo (niño-padre)	Media
Calor emocional (niño-padre)	Media
Sobreprotección (niño-padre)	Media
Favorecedor (niño-padre)	Media
Rechazo (niño-madre)	Media
Calor emocional (niño-madre)	Media
Sobreprotección (niño-madre)	Media
Favorecedor (niño-madre)	Media
Rechazo (padre)	Red RBF
Calor emocional (padre)	Red RBF
Sobreprotección (padre)	Red RBF
Favorecedor (padre)	Red RBF
Rechazo (madre)	Media
Calor emocional (madre)	Media
Sobreprotección (madre)	Media
Favorecedor (madre)	Media

La red neuronal RBF empleada hace uso de los mismos parámetros presentados en el experimento de simulación del capítulo anterior: función gausiana como función de base radial, distancia euclídea y 5 centros iniciales con incrementos de 5 en 5 hasta un máximo de 100.

En la matriz de datos completa resultante de la imputación (que denominaremos matriz completa RBF) se han calculado diversos estadísticos descriptivos que se hallan en la Tabla 63. En general, el incremento del tamaño muestral conlleva una reducción del error estándar de los estadísticos calculados, lo que se traduce en un aumento de la precisión de las estimaciones y en un incremento de la

potencia de las pruebas de significación. La media de las variables imputadas mediante red neuronal RBF apenas cambia respecto a la obtenida en la matriz *listwise*. Las distribuciones de frecuencias de las variables independientes categóricas no se presentan porque, al haber imputado sus valores missing con la moda, se pueden obtener fácilmente a partir de los datos de la Tabla 60.

Tabla 63. Estadísticos descriptivos de las variables independientes cuantitativas calculados en la matriz completa

Estadísticos descriptivos

	N	Mínimo	Máximo	Media		Desv. típ
	Estadístico	Estadístico	Estadístico	Estadístico	Error típico	Estadístico
Edad	200	6	17	11.96	.22	3 08
Rechazo (niño-padre)	200	11.00	83 00	25.4959	1.0536	14.8995
Calor emocional (niño-padre)	200	21.00	69.23	46.7900	.6500	9.1923
Sobreprotección (niño-padre)	200	10.00	51.00	24.9808	.6205	8.7751
Favorecedor (niño-padre)	200	5 00	14 00	6.7738	.1399	1.9788
Rechazo (niño-madre)	200	11.00	76.19	25.8493	1.0455	14.7856
Calor emocional (niño-madre)	200	20.00	67.00	48.0525	.5565	7.8696
Sobreprotección (niño-madre)	200	10.00	52.00	26.2705	.6418	9.0758
Favorecedor (niño-madre)	200	5.00	14.00	6 8982	.1451	2.0518
Rechazo (padre)	200	13.00	26 00	17.5903	.2118	2.9955
Calor emocional (padre)	200	32.00	68 00	53.6381	.5327	7.5340
Sobreprotección (padre)	200	22.80	67.56	38.6062	.4239	5.9945
Favorecedor (padre)	200	3.00	7.00	3.5742	6.107E-02	.8637
Rechazo (madre)	200	13 00	29.00	17.4893	.2320	3.2813
Calor emocional (madre)	200	34.00	68.00	56.5834	.4925	6.9652
Sobreprotección (madre)	200	21.00	60.17	40.3852	.4854	6 8643
Favorecedor (madre)	200	3 00	12 00	3.7753	9 835E-02	1 3909

Estadísticos descriptivos

	Asimetría		Curtosis	
	Estadístico	Error típico	Estadístico	Error típico
Edad	-.121	.172	-1.171	.342
Rechazo (niño-padre)	1.290	.172	1.771	.342
Calor emocional (niño-padre)	-.631	.172	.399	.342
Sobreprotección (niño-padre)	.679	.172	-.075	.342
Favorecedor (niño-padre)	1.083	.172	.804	.342
Rechazo (niño-madre)	1.113	.172	.972	.342
Calor emocional (niño-madre)	-.654	.172	1.049	.342
Sobreprotección (niño-madre)	.537	.172	-.356	.342
Favorecedor (niño-madre)	1.087	.172	.784	.342
Rechazo (padre)	.559	.172	-.167	.342
Calor emocional (padre)	-.377	.172	-.171	.342
Sobreprotección (padre)	.708	.172	2.382	.342
Favorecedor (padre)	2.096	.172	4.597	.342
Rechazo (madre)	.905	.172	.934	.342
Calor emocional (madre)	-.671	.172	.190	.342
Sobreprotección (madre)	.217	.172	-.115	.342
Favorecedor (madre)	2 693	.172	9 091	.342

Como se ha comentado en el capítulo 2, la mayoría de investigadores que incluyen en sus análisis el tratamiento de los valores perdidos se limitan a imputar el valor medio o la moda a todas las variables incompletas. Ello ha motivado que también generemos una matriz de datos completa (que denominaremos matriz completa MID) imputando a todas las variables independientes cuantitativas (incluidas las procedentes de la entrevista con el padre) el valor medio, y a las categóricas la moda. Esta nueva matriz de datos completos será empleada en el problema de clasificación que se presenta en el siguiente apartado.

7.5. CLASIFICACIÓN

En las matrices de datos completos generadas se han estimado diferentes modelos de clasificación de la variable dependiente, utilizando tanto el modelo de regresión logística como la red neuronal MLP. El porcentaje de clasificaciones correctas con cada técnica de clasificación se midió en un subconjunto de casos que actúan como datos de test (10% de la muestra disponible). Dado el reducido tamaño muestral, para compensar la incidencia del azar en la selección de los registros de test se recurrió a la técnica de validación cruzada. Así, el porcentaje de clasificaciones correctas se midió en un total de 10 conjuntos de datos de test, seleccionados de forma aleatoria, y el resultado final se obtuvo como el promedio en dichas 10 submuestras.

7.5.1. Regresión logística

En la Tabla 64 se hallan las estimaciones del modelo de regresión logística obtenido a partir de la matrices completas RBF y MID. Con fines comparativos también se presenta el modelo de regresión estimado aplicando el método *listwise*. Los porcentajes de clasificaciones correctas obtenidos en las tres matrices analizadas son muy similares. Sin embargo, estamos ante un claro ejemplo en el que es inviable eliminar los registros incompletos, ya que se trata de sujetos con nombre y apellidos. En la práctica clínica no se puede dejar de asignar un paciente a un grupo porque su padre, por ejemplo, no rellene el cuestionario que recoge su estilo educativo.

Respecto a las estimaciones de los parámetros del modelo de regresión, el método *listwise* conduce a intervalos de confianza menos precisos que los obtenidos sobre las matrices completas, empobreciendo la interpretación del efecto que cada variable independiente tiene sobre la dependiente. Un ejemplo evidente lo constituye la variable *PCT*, que en el análisis sobre la matriz *listwise* resulta irrelevante en la clasificación de *Dep* (ya que los intervalos de confianza de las dos variables ficticias generadas incluyen el valor 1), mientras que al ser analizada en las matrices completas se constituye en un factor de riesgo (en el

sentido de que la presencia de problemas conductuales tempranos incrementa la creencia de que el hijo necesita ayuda).

Tabla 64. Clasificación mediante regresión logística en las matrices *listwise*, completa RBF y completa MID

	EXP(B) (IC 95%)		
	COMPLETA RBF N=200	COMPLETA MID N=200	*LISTWISE* N=125
SEXO(1)[a]	0.784 (0.315 a 1.949)	0.606 (0.247 a 1.483)	0.402 (0.109 a 1.480)
EDAD	1.165 (0.926 a 1.466)	1.258 (0.999 a 1.585)	1.561 (1.087 a 2.243)
PCT(1)[a]	0.196 (0.040 a 0.953)	0.069 (0.009 a 0.495)	0.194 (0.023 a 1.631)
PCT(2)[a]	0.108 (0.020 a 0.580)	0.056 (0.007 a 0.430)	0.359 (0.036 a 3.575)
AP(1)[a]	0.497 (0.197 a 1.252)	0.644 (0.263 a 1.576)	0.726 (0.215 a 2.442)
SES(1)[a]	0.389 (0.121 a 1.252)	0.210 (0.066 a 0.667)	0.385 (0.077 a 1.930)
SES(2)[a]	0.460 (0.151 a 1.401)	0.427 (0.145 a 1.256)	0.254 (0.056 a 1.151)
RECHA_NP	1.086 (0.910 a 1.295)	1.104 (0.929 a 1.312)	0.905 (0.715 a 1.145)
CALOR_NP	1.018 (0.878 a 1.180)	1.005 (0.868 a 1.164)	0.968 (0.713 a 1.315)
PROTE_NP	1.069 (0.882 a 1.294)	1.023 (0.839 a 1.248)	0.855 (0.621 a 1.177)
FAVOR_NP	1.516 (0.786 a 2.922)	1.399 (0.704 a 2.781)	4.824 (0.980 a 23.74)
RECHA_NM	0.978 (0.835 a 1.146)	0.965 (0.827 a 1.126)	1.104 (0.863 a 1.412)
CALOR_NM	0.902 (0.760 a 1.072)	0.923 (0.777 a 1.097)	0.872 (0.614 a 1.238)
PROTE_NM	0.864 (0.722 a 1.035)	0.892 (0.740 a 1.076)	1.054 (0.783 a 1.419)
FAVOR_NM	0.661 (0.365 a 1.196)	0.766 (0.411 a 1.427)	0.226 (0.046 a 1.106)
RECHA_P	0.913 (0.738 a 1.130)	0.998 (0.818 a 1.216)	1.196 (0.903 a 1.584)
CALOR_P	0.947 (0.871 a 1.028)	0.984 (0.912 a 1.062)	1.002 (0.902 a 1.113)
PROTE_P	1.035 (0.947 a 1.132)	1.010 (0.919 a 1.110)	1.008 (0.891 a 1.140)
FAVOR_P	1.023 (0.554 a 1.888)	1.005 (0.533 a 1.895)	0.705 (0.372 a 1.338)
RECHA_M	1.021 (0.844 a 1.235)	1.050 (0.880 a 1.251)	0.925 (0.726 a 1.178)
CALOR_M	0.919 (0.850 a 0.995)	0.955 (0.887 a 1.028)	0.855 (0.757 a 0.966)
PROTE_M	1.082 (0.995 a 1.176)	1.083 (1.000 a 1.172)	1.110 (0.961 a 1.282)
FAVOR_M	1.553 (0.973 a 2.476)	1.529 (0.990 a 2.363)	2.146 (1.034 a 4.452)
% Clasificac. Correctas	69.06%	68.12%	69.23%

[a] **Última categoría como referencia**

7.5.2. Red neuronal artificial MLP

Respecto a la clasificación mediante red neuronal artificial, la red MLP empleada es similar a la presentada en el capítulo anterior, con la diferencia de que, dada la mayor complejidad del problema, el número de unidades ocultas se ha incrementado a cinco. Como se aprecia en la Tabla 65, el porcentaje de sujetos correctamente clasificados es ligeramente superior cuando se calcula sobre la matriz eliminando los registros incompletos (93.23%). Respecto a las matrices completas, en la matriz RBF se obtiene un 3.44% más de clasificaciones correctas que en la matriz MID, poniendo de manifiesto, tal y como sugieren nuestras conclusiones del experimento de simulación, que la imputación

mediante red RBF de las variables obtenidas en la entrevista con el padre es mejor que la imputación de la media.

La comparación con los resultados conseguidos mediante regresión pone de manifiesto que el porcentaje de clasificaciones correctas se incrementa de forma considerable (20.63% en la matriz completa RBF), como es de esperar dada la constatada superioridad de los modelos de red neuronal en problemas de clasificación.

Tabla 65. Clasificación mediante red neuronal MLP en las matrices *listwise*, completa RBF y completa MID

	COMPLETA RBF N=200	COMPLETA MID N=200	*LISTWISE* N=125
% Clasificac. Correctas	89.69%	86.25%	93.23%

La medida del efecto que una variable independiente tiene sobre la dependiente en un modelo de red neuronal es un tema en continuo estudio. Existe la creencia generalizada de que las redes neuronales son una especie de "caja negra" en la que no se sabe como circula la información. A pesar de que los modelos de red neuronal artificial son eminentemente predictivos y no explicativos, cada día son más los avances en este terreno; así, a nivel teórico, Garson (1991) y Hashem (1992) presentan sendos métodos basados en la descomposición de los pesos de las unidades ocultas a las de salida en los componentes asociados a cada unidad de entrada, y Bremner, Gotts y Denham (1994) aplican los diagramas de Hinton (Hinton y Shallice, 1991) para visualizar la sensibilidad de las unidades de salida a determinados cambios en las unidades de entrada. A un nivel más aplicado, el programa *Neural Connection 2.0* ofrece dos potentes herramientas que permiten evaluar el cambio provocado en la variable dependiente por un determinado incremento en una variable independiente. La primera es una representación gráfica de la función estimada que relaciona dos variables independientes con la dependiente, manteniendo el resto de variables predictoras fijadas a su valor medio. Así, a modo de ejemplo, a partir de la red neuronal entrenada con la matriz de datos completa RBF, en la Ilustración 26 se representa la función estimada que relaciona las variables *Edad*, *Recha_np* y *Calor_np* con la variable *Dep*.

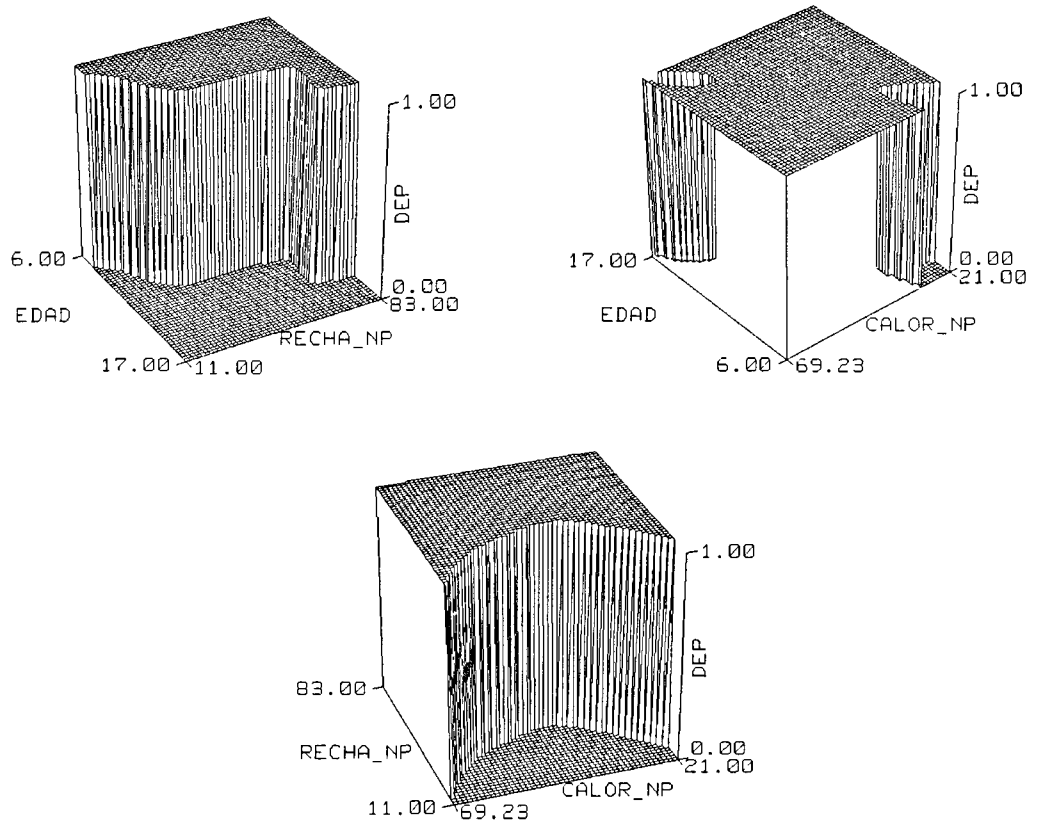


Ilustración 26. Función estimada por la red MLP que relaciona las variables *Edad, Recha_np* y *Calor_np* con la variable *Dep*

Analizando las gráficas de la Ilustración 26 se puede llegar a la conclusión de que el efecto de una variable independiente no puede ser estudiado de forma aislada. Efectivamente, en el caso de la variable *Recha_np*, por ejemplo, la primera gráfica sugiere que un valor bajo en *Recha_np* disminuye la probabilidad del valor 1 en la variable *Dep* (hijo necesita ayuda). Sin embargo, al analizar la tercera gráfica comprobamos que, aunque el valor en *Recha_np* sea bajo, si en *Calor_np* se tiene una puntuación extrema (alta o baja) la probabilidad del valor 1 en *Dep* (hijo necesita ayuda) es elevada.

No obstante, no todas las funciones estimadas por la red neuronal tienen la complejidad del ejemplo anterior. En la Ilustración 27 se presentan las gráficas que relacionan las variables *Recha_m*, *Calor_m* y *Prote_m* con *Dep*. De ellas se deduce que, considerando la interacción entre dichas variables, la probabilidad del valor 1 en *Dep* (hijo necesita ayuda) se incrementa a medida que aumentan las puntuaciones en las tres variables predictoras, o, en otros términos, que un elevado calor emocional, rechazo familiar o sobreprotección, informados por la madre, son factores de riesgo de la necesidad de ayuda por parte del hijo.

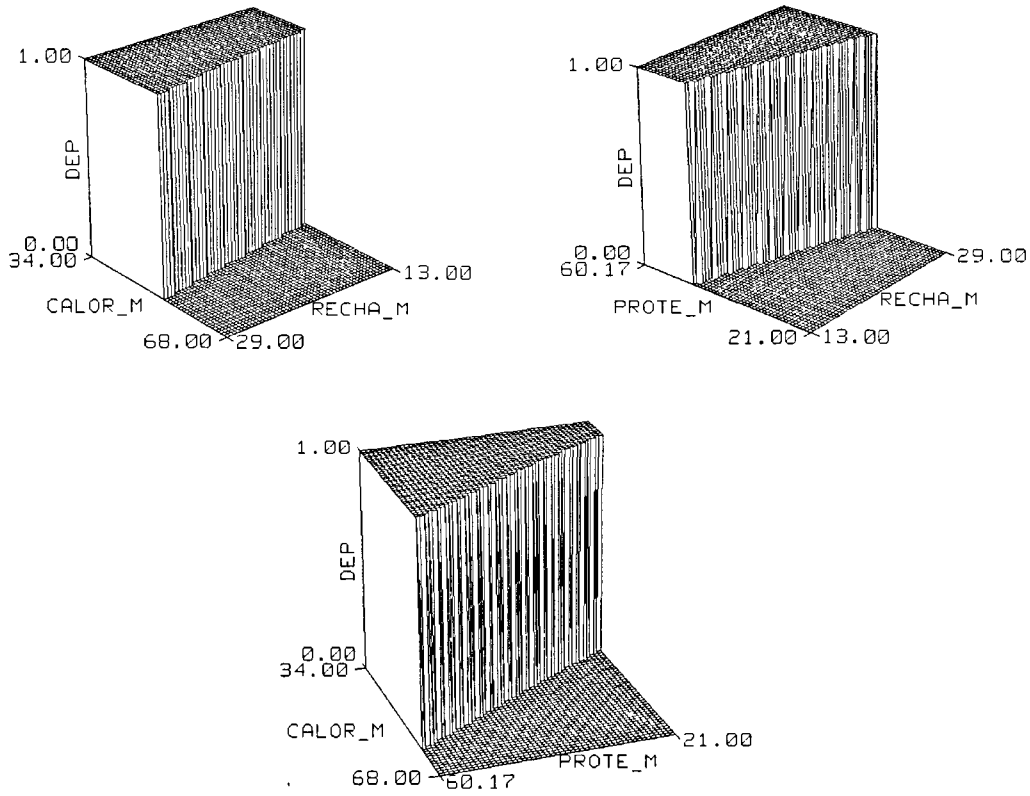


Ilustración 27. Función estimada por la red MLP que relaciona las variables *Recha_m*, *Calor_m* y *Prote_m* con la variable *Dep*

Complementando las representaciones gráficas, los análisis de sensibilidad (*sensitivity analysis*) permiten cuantificar el efecto de una variable independiente sobre la dependiente. En modelos no lineales, como el obtenido mediante una red neuronal MLP, el efecto sobre la variable dependiente de incrementar en una unidad una variable independiente no es constante para todo el rango de valores de la variable independiente. A modo de ejemplo, no es lo mismo valorar la influencia en la necesidad de ayuda del hijo de un incremento de la edad de 7 a 8 años que de 10 a 11 años. Por ello, los análisis de sensibilidad se acostumbrar a realizar a través de una interfaz dinámica, como la que ofrece *Neural Connection 2.0*. En la Ilustración 28 se ejemplifica cómo, fijado *Recha_np* en 45 puntos, y el resto de variables predictoras a su valor medio, un incremento en la edad de 10 a 11.22 años modifica la predicción de la variable dependiente de 1 (hijo necesita ayuda) a 0 (hijo no necesita ayuda).

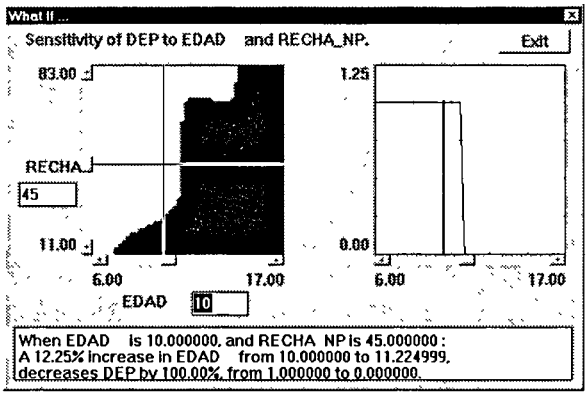


Ilustración 28. Análisis de sensibilidad del efecto de *Edad* sobre *Dep* fijado *Recha_np*

Debido a la gran cantidad de representaciones gráficas y análisis de sensibilidad que serían necesarios, y a que nuestro trabajo se centra en la evaluación de una perspectiva diferente de análisis de datos incompletos, y no tanto en el examen exhaustivo del problema de clasificación planteado, hemos obviado el estudio de la influencia específica que cada variable independiente tiene sobre la dependiente. Sí queda ilustrada, en los párrafos precedentes, la dinámica que se debe seguir para conseguir dicho objetivo en un modelo de clasificación establecido mediante red neuronal artificial.

Nos gustaría citar, a modo de conclusión del capítulo actual, tres aspectos que adquieren relevancia en el ámbito de la psicopatología infantil:

- El tradicional método de eliminación de registros es del todo inaplicable, ya que la eliminación de un sujeto que presente algún dato faltante resulta inviable en la psicopatología aplicada, que no podemos olvidar, debe ser la beneficiaria de nuestras conclusiones.
- En matrices de datos provenientes de investigaciones de carácter psicológico es habitual que los valores faltantes se concentren en unas determinadas variables, como consecuencia, por ejemplo, del cansancio del entrevistado, que deja de responder las últimas preguntas, por la incomparecencia de un determinado informante (padre, profesor, etc.), o, en algunas ocasiones, por un deficiente diseño de la investigación. En esta situación, la imputación de los valores missing mediante red neuronal es mejor que la imputación directa del valor medio.
- El estudio del efecto de una determinada variable independiente sobre la dependiente, aspecto fundamental desde una perspectiva preventiva, debe ser realizado mediante los denominados análisis de sensibilidad en el marco de clasificación con modelos no lineales como las redes neuronales artificiales.

Los factores que rigen la conducta humana son tan intrincados que en muchos casos invalidan conclusiones generales del tipo *"un incremento de una unidad en la variable x provoca un cambio z en la variable y"*. Las complejas interacciones entre factores de riesgo y factores protectores requieren ser estudiadas en detalle, para establecer conclusiones parciales circunscritas a un rango de valores determinado.

ANEXO

Aleatoriedad de las observaciones

t tests (varancias separadas)[a]

		EDAD	RECHA_NP	CALOR_NP	PROTE_NP	FAVOR_NP	RECHA_NM	CALOR_NM	PROTE_NM	FAVOR_NM
RECHA_P	t	2	1 0	- 2	9	6	1 3	-1 3	8	- 2
	gl	62 9	63 7	52 1	60 6	61 0	69 1	57 2	66 2	52 6
	P(Bilat)	871	343	853	395	564	196	205	426	877
	# Presente	161	158	158	158	161	157	157	157	160
	# Missing	39	37	39	38	39	38	39	39	39
	Media(Presente)	11 98	25 9377	46 7222	25 2314	6 8121	26 4446	47 6842	26 5089	6 8859
	Media(Missing)	11 90	23 6090	47 0645	23 9386	6 6154	23 3897	49 5353	25 3105	6 9487
CALOR_P	t	2	1 0	- 2	9	6	1 3	-1 3	8	- 2
	gl	62 9	63 7	52 1	60 6	61 0	69 1	57 2	66 2	52 6
	P(Bilat)	871	343	853	395	564	196	205	426	877
	# Presente	161	158	158	158	161	157	157	157	160
	# Missing	39	37	39	38	39	38	39	39	39
	Media(Presente)	11 98	25 9377	46 7222	25 2314	6 8121	26 4446	47 6842	26 5089	6 8859
	Media(Missing)	11 90	23 6090	47 0645	23 9386	6 6154	23 3897	49 5353	25 3105	6 9487
PROTE_P	t	2	1 0	- 2	9	6	1 3	-1 3	8	- 2
	gl	62 9	63 7	52 1	60 6	61 0	69 1	57 2	66 2	52 6
	P(Bilat)	871	343	853	395	564	196	205	426	877
	# Presente	161	158	158	158	161	157	157	157	160
	# Missing	39	37	39	38	39	38	39	39	39
	Media(Presente)	11 98	25 9377	46 7222	25 2314	6 8121	26 4446	47 6842	26 5089	6 8859
	Media(Missing)	11 90	23 6090	47 0645	23 9386	6 6154	23 3897	49 5353	25 3105	6 9487
FAVOR_P	t	1 4	1 9	- 5	6	1	2 1	-1 2	1 2	- 4
	gl	101 2	124 6	99 4	109 2	125 3	127 6	106 7	128 8	106 6
	P(Bilat)	168	058	647	527	941	034	222	250	711
	# Presente	142	139	139	139	142	138	138	138	141
	# Missing	58	56	58	57	58	57	58	58	58
	Media(Presente)	12 16	26 6923	46 5871	25 2328	6 7799	27 1933	47 6004	26 7238	6 8635
	Media(Missing)	11 48	22 5262	47 2762	24 3661	6 7586	22 5952	49 1284	25 1920	6 9828
FAVOR_M	t	1 9	2 3	-1 5	- 2	- 9	1 6	- 9	3	- 5
	gl	24 8	25 9	27 0	24 9	32 6	25 9	25 9	28 9	33 8
	P(Bilat)	070	032	155	830	377	112	359	780	634
	# Presente	178	175	175	175	178	174	174	174	177
	# Missing	22	20	22	21	22	21	22	22	22
	Media(Presente)	12 13	26 1979	46 4592	24 9327	6 7402	26 4284	47 8539	26 3283	6 8799
	Media(Missing)	10 64	19 3535	49 4213	25 3810	7 0455	21 0505	49 6239	25 8131	7 0455
PCT	t	- 1	- 2	8	- 2	5	- 3	1 1	- 3	1
	gl	12 3	11 3	11 0	11 1	13 9	11 1	11 8	11 6	12 5
	P(Bilat)	899	830	443	813	654	790	314	784	911
	# Presente	188	184	186	185	188	184	185	185	187
	# Missing	12	11	11	11	12	11	11	11	12
	Media(Presente)	11 96	25 4386	46 9316	24 9413	6 7859	25 7750	48 1774	26 2302	6 9024
	Media(Missing)	12 08	26 4545	44 3957	25 6444	6 5833	27 0909	45 9519	26 9475	6 8333

Para cada variable cuantitativa, los dos grupos se forman mediante las correspondientes variables indicadoras

a No se incluyen las variables indicadoras con menos de un 5% de valores missing

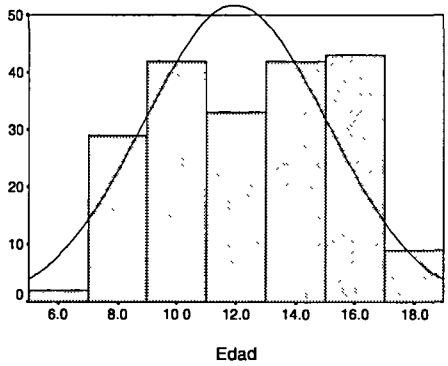
t tests (varianzas separadas)[a]

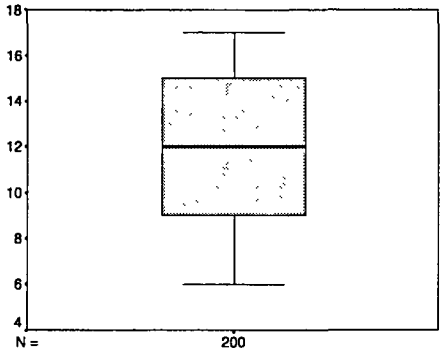
		RECHA_P	CALOR_P	PROTE_P	FAVOR_P	RECHA_M	CALOR_M	PROTE_M	FAVOR_M
RECHA_P	t					1 4	8	1 0	- 5
	gl					60 7	53 0	53 7	53 5
	P(Bilat)					181	404	315	615
	# Presente	161	161	161	142	160	160	160	143
	# Missing	0	0	0	0	38	38	39	35
	Media(Presente)	17 6612	53 7658	38 8160	3 5775	17 6340	56 7977	40 6470	3 7483
	Media(Missing)					16 8802	55 6810	39 3111	3 8857
CALOR_P	t					1 4	8	1 0	- 5
	gl					60 7	53 0	53 7	53 5
	P(Bilat)					181	404	315	615
	# Presente	161	161	161	142	160	160	160	143
	# Missing	0	0	0	0	38	38	39	35
	Media(Presente)	17 6612	53 7658	38 8160	3 5775	17 6340	56 7977	40 6470	3 7483
	Media(Missing)					16 8802	55 6810	39 3111	3 8857
PROTE_P	t					1 4	8	1 0	- 5
	gl					60 7	53 0	53 7	53 5
	P(Bilat)					181	404	315	615
	# Presente	161	161	161	142	160	160	160	143
	# Missing	0	0	0	0	38	38	39	35
	Media(Presente)	17 6612	53 7658	38 8160	3 5775	17 6340	56 7977	40 6470	3 7483
	Media(Missing)					16 8802	55 6810	39 3111	3 8857
FAVOR_P	t	1 3	-2 8	1 4		9	3	2	-1 0
	gl	23 0	26 9	24 8		107 6	99 9	95 6	51 4
	P(Bilat)	208	010	179		381	758	811	340
	# Presente	142	142	142	142	141	141	141	141
	# Missing	19	19	19	0	57	57	58	37
	Media(Presente)	17 7817	53 2436	39 0467	3 5775	17 6182	56 6832	40 4645	3 7163
	Media(Missing)	16 7611	57 6687	37 0918		17 1705	56 3367	40 1924	4 0000
FAVOR_M	t	1 0	-3 3	2		-1 2	- 1	-1 8	
	gl	21 3	26 7	22 0		22 7	24 1	25 4	
	P(Bilat)	325	003	819		224	952	091	
	# Presente	143	143	143	141	177	177	177	178
	# Missing	18	18	18	1	21	21	22	0
	Media(Presente)	17 7536	53 2080	38 8564	3 5816	17 3621	56 5722	40 0596	3 7753
	Media(Missing)	16 9275	58 1979	38 4951		18 5620	56 6784	43 0044	
PCT	t	-1 3	- 6	-1 5	-1 4	- 8	4	-1 6	6
	gl	7 3	7 7	7 7	6 3	10 8	12 1	12 2	12 1
	P(Bilat)	242	587	182	206	455	711	134	535
	# Presente	153	153	153	135	187	186	187	167
	# Missing	8	8	8	7	11	12	12	11
	Media(Presente)	17 5520	53 6817	38 6430	3 5407	17 4375	56 6372	40 1695	3 7904
	Media(Missing)	19 7500	55 3750	42 1250	4 2857	18 3712	55 7500	43 7454	3 5455

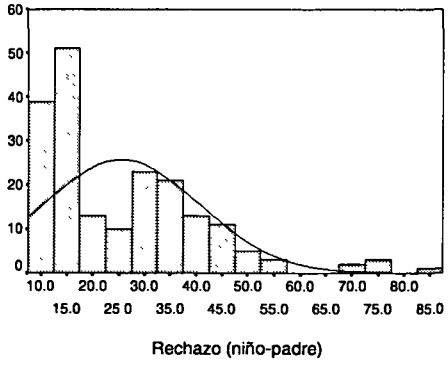
Para cada variable cuantitativa, los dos grupos se forman mediante las correspondientes variables indicadoras

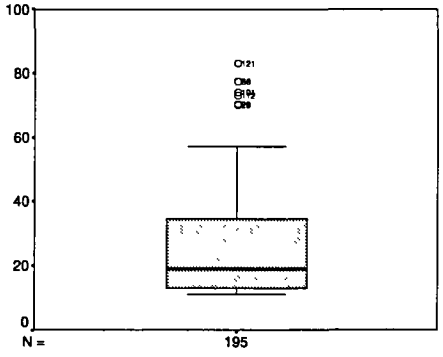
a No se incluyen las variables indicadoras con menos de un 5% de valores missing

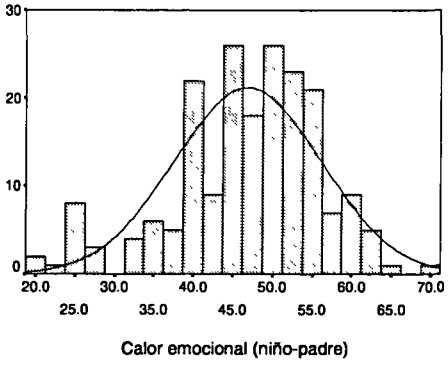
Histogramas y gráficos de caja de las variables independientes cuantitativas

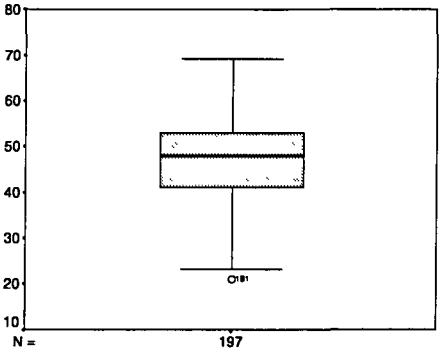


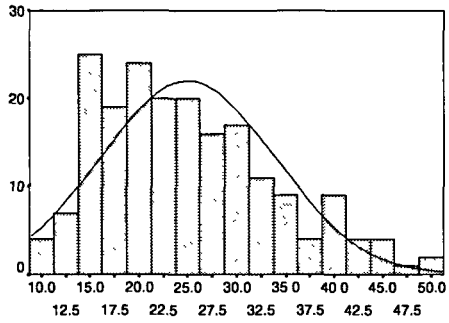




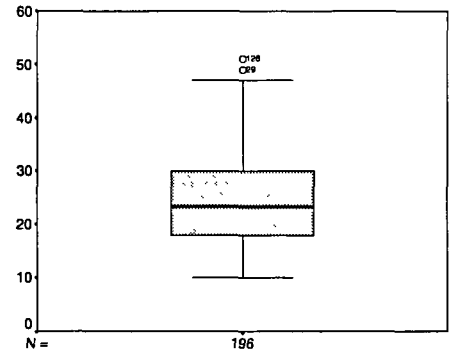


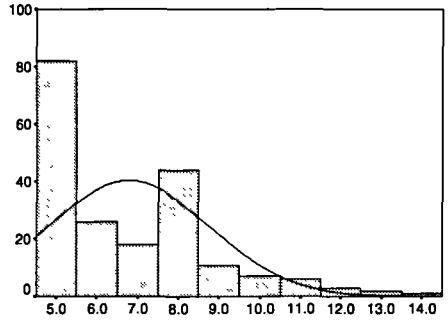




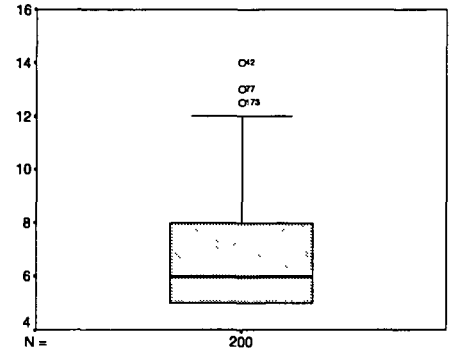


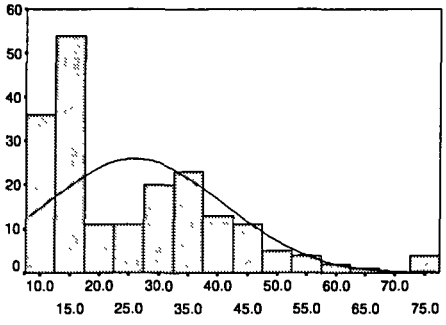
Sobreprotección (niño-padre)



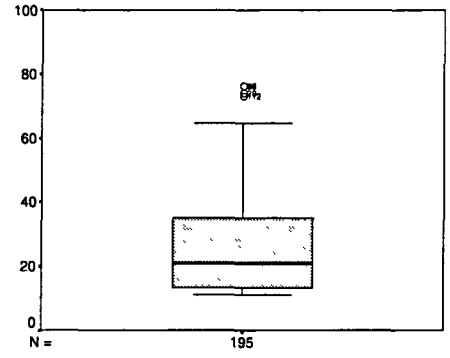


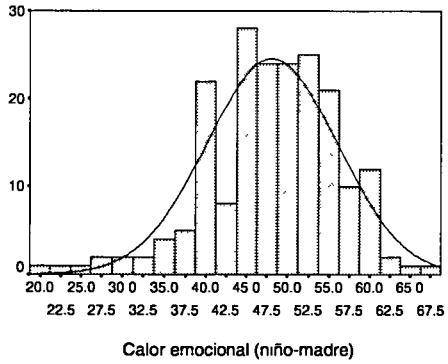
Favorecedor (niño-padre)



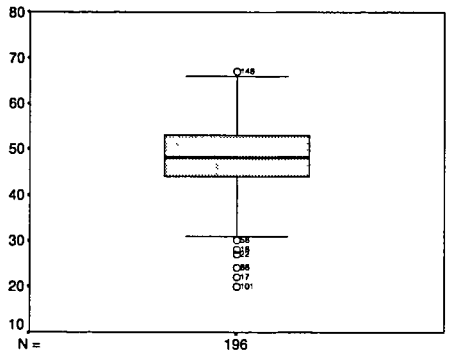


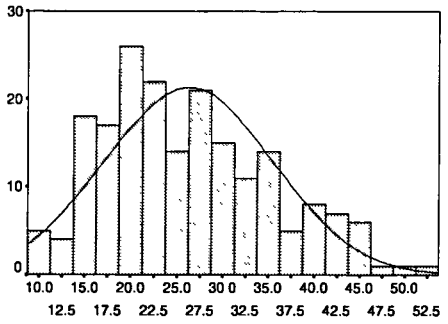
Rechazo (niño-madre)



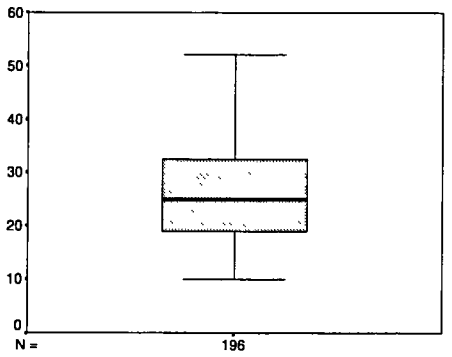


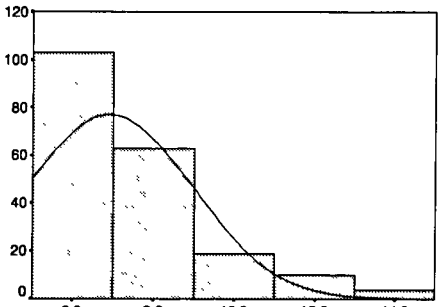
Calor emocional (niño-madre)



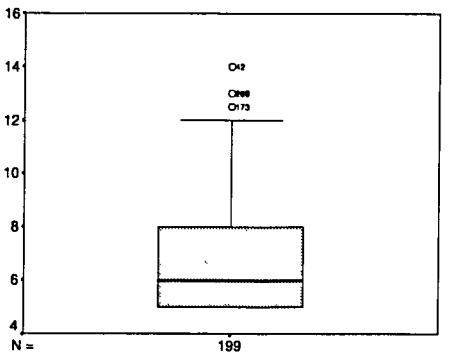


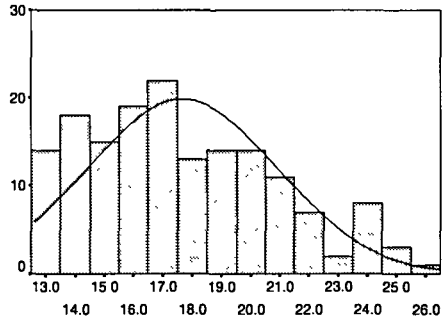
Sobreprotección (niño-madre)



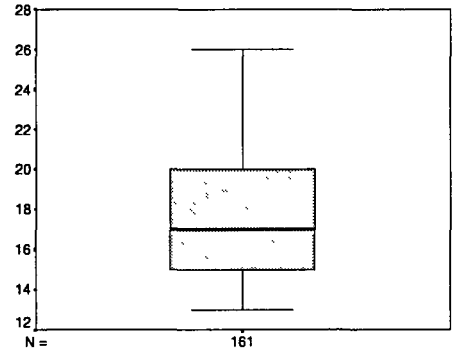


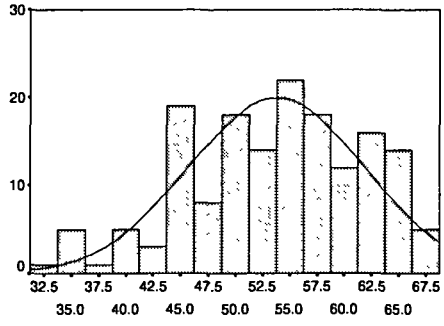
Favorecedor (niño-madre)



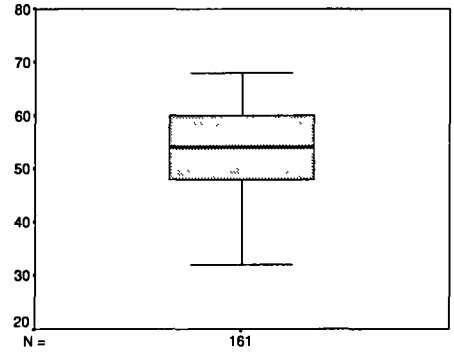


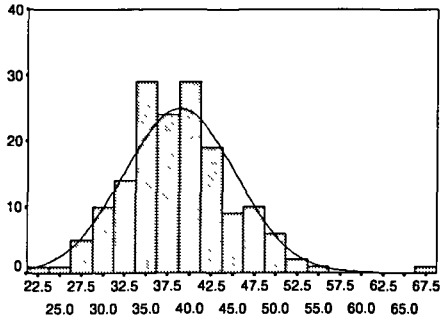
Rechazo (padre)



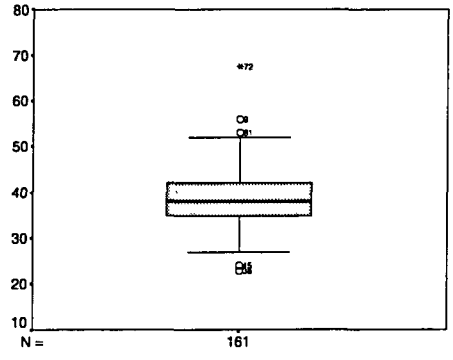


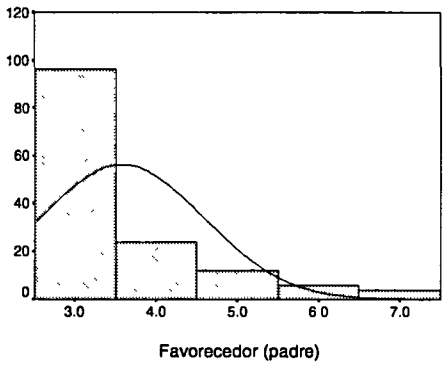
Calor emocional (padre)



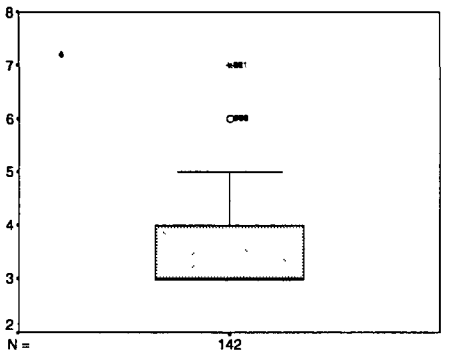


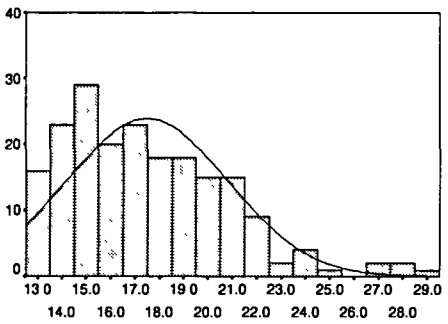
Sobreprotección (padre)



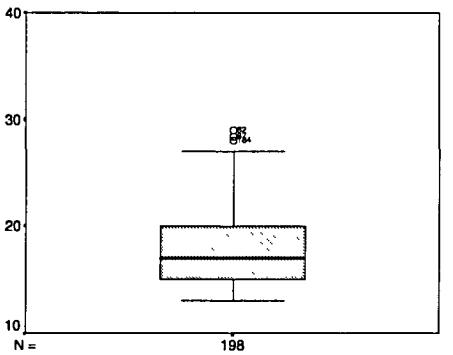


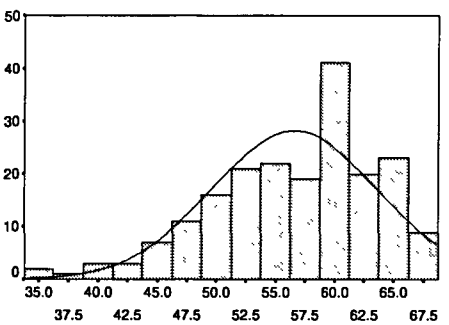
Favorecedor (padre)



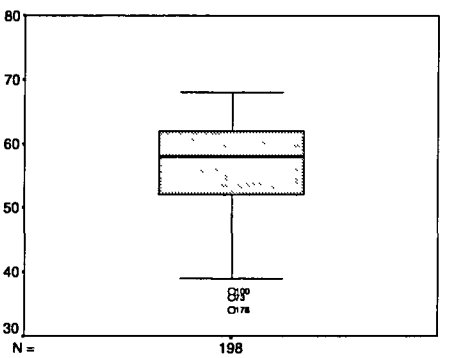


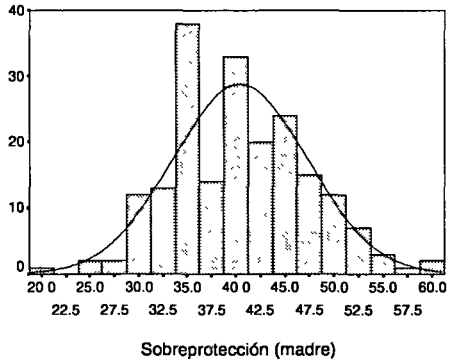
Rechazo (madre)

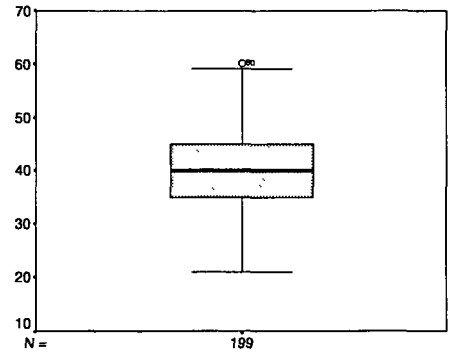


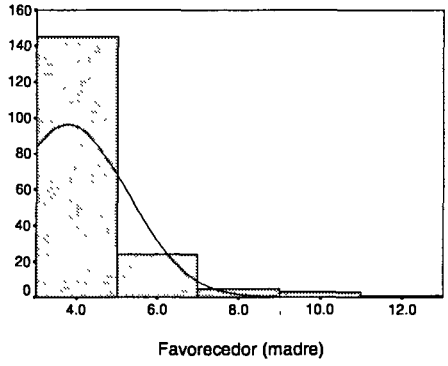


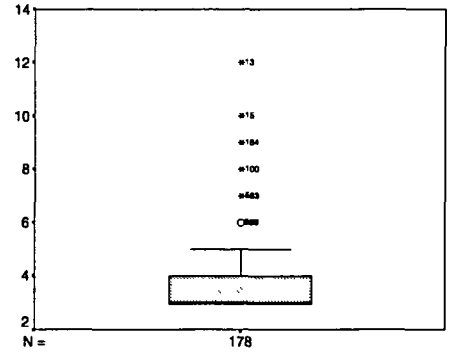
Calor emocional (madre)



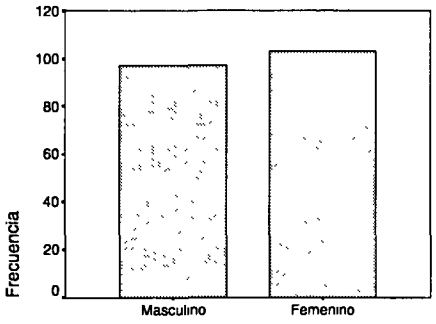

40
30
20
10
0
20 0 25.0 30.0 35.0 40 0 45.0 50.0 55.0 60.0
22.5 27.5 32.5 37.5 42.5 47.5 52.5 57.5
Sobreprotección (madre)

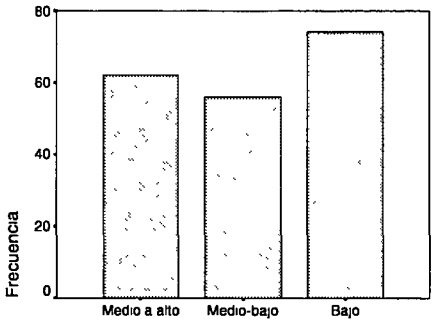

70
60
50
40
30
20
10
N =
199

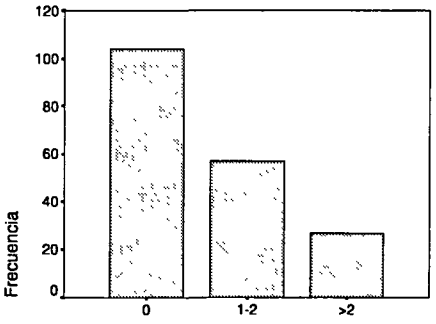

160
140
120
100
80
60
40
20
0
4.0 6.0 8.0 10.0 12.0
Favorecedor (madre)

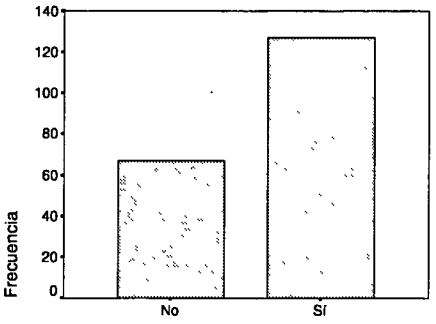

14
12
10
8
6
4
2
N =
178

Diagramas de barras de las variables independientes categóricas









Correlación

Coeficientes de correlación de Spearman

	Rechazo (niño-padre)	Calor emocional (niño-padre)	Sobreprotección (niño-padre)	Favorecedor (niño-padre)
Rechazo (niño-padre)	1.000	-.339	.798	-.126
Calor emocional (niño-padre)	-.339	1.000	.009	.091
Sobreprotección (niño-padre)	.798	.009	1.000	-.100
Favorecedor (niño-padre)	-.126	.091	-.100	1 000
Rechazo (niño-madre)	.980	-.308	.794	-.116
Calor emocional (niño-madre)	-.310	.930	.004	.106
Sobreprotección (niño-madre)	.804	-.048	.956	- 085
Favorecedor (niño-madre)	-.049	.028	-.041	.913
Rechazo (padre)	.236	-.180	.174	.011
Calor emocional (padre)	-.353	.425	-.227	.168
Sobreprotección (padre)	.233	-.131	.272	.114
Favorecedor (padre)	.206	-.035	.189	-.023
Rechazo (madre)	.184	-.152	.135	-.024
Calor emocional (madre)	-.318	.318	-.218	.118
Sobreprotección (madre)	.339	-.130	.332	.046
Favorecedor (madre)	.031	-.094	.031	-.048
Sexo	.161	.049	.127	-.094
Edad	.736	-.206	.673	-.333
Probl. conduct. tempr.	.090	-.196	.043	.066
Anteced. psicopat.	.165	-.191	.114	.039
Nivel socieconómico	-.047	-.055	-.081	.164
¿Hijo necesita ayuda?	.324	-.327	.186	- 084

Coeficientes de correlación de Spearman (continuación)

	Rechazo (niño-madre)	Calor emocional (niño-madre)	Sobreprotección (niño-madre)	Favorecedor (niño-madre)
Rechazo (niño-padre)	.980	-.310	.804	-.049
Calor emocional (niño-padre)	-.308	.930	-.048	.028
Sobreprotección (niño-padre)	.794	.004	.956	-.041
Favorecedor (niño-padre)	-.116	.106	-.085	.913
Rechazo (niño-madre)	1.000	-.295	.813	-.061
Calor emocional (niño-madre)	-.295	1.000	-.020	.060
Sobreprotección (niño-madre)	.813	-.020	1.000	-.018
Favorecedor (niño-madre)	-.061	.060	-.018	1.000
Rechazo (padre)	.216	-.165	.182	.036
Calor emocional (padre)	-.355	.411	-.288	.091
Sobreprotección (padre)	.220	-.153	.236	.116
Favorecedor (padre)	.198	.019	.212	-.029
Rechazo (madre)	.182	-.200	.134	.007
Calor emocional (madre)	-.299	.344	-.212	.028
Sobreprotección (madre)	.329	-.131	.340	.061
Favorecedor (madre)	-.002	-.030	.054	.036
Sexo	.162	.016	.179	-.085
Edad	.726	-.180	.666	-.271
Probl. conduct. tempr.	.091	-.169	.035	.095
Anteced. psicopat.	.138	-.167	.128	.034
Nivel socieconómico	-.022	-.084	-.064	.165
¿Hijo necesita ayuda?	.311	-.352	.168	-.053

Coeficientes de correlación de Spearman (continuación)

	Rechazo (padre)	Calor emocional (padre)	Sobreprotección (padre)	Favorecedor (padre)
Rechazo (niño-padre)	.236	-.353	.233	.206
Calor emocional (niño-padre)	-.180	.425	-.131	-.035
Sobreprotección (niño-padre)	.174	-.227	.272	.189
Favorecedor (niño-padre)	.011	.168	.114	-.023
Rechazo (niño-madre)	.216	-.355	.220	.198
Calor emocional (niño-madre)	-.165	.411	-.153	.019
Sobreprotección (niño-madre)	.182	-.288	.236	.212
Favorecedor (niño-madre)	036	.091	.116	-.029
Rechazo (padre)	1.000	-.363	.358	.304
Calor emocional (padre)	-.363	1 000	.000	-.068
Sobreprotección (padre)	.358	.000	1.000	.130
Favorecedor (padre)	.304	-.068	.130	1 000
Rechazo (madre)	.490	-.166	.208	.174
Calor emocional (madre)	-.164	.395	-.147	-.079
Sobreprotección (madre)	.279	-.083	.434	.174
Favorecedor (madre)	.026	-.104	.023	.275
Sexo	-.170	-.016	-.143	.002
Edad	.003	-.245	.091	.103
Probl. conduct. tempr.	.212	-.131	.149	.021
Anteced. psicopat.	.230	-.134	.036	.089
Nivel socieconómico	-.184	.114	.043	-.056
¿Hijo necesita ayuda?	087	- 217	.148	.022

Coeficientes de correlación de Spearman (continuación)

	Rechazo (madre)	Calor emocional (madre)	Sobreprotección (madre)	Favorecedor (madre)
Rechazo (niño-padre)	.184	-.318	.339	.031
Calor emocional (niño-padre)	-.152	.318	-.130	-.094
Sobreprotección (niño-padre)	.135	-.218	.332	.031
Favorecedor (niño-padre)	-.024	.118	.046	-.048
Rechazo (niño-madre)	.182	-.299	.329	-.002
Calor emocional (niño-madre)	-.200	.344	-.131	-.030
Sobreprotección (niño-madre)	.134	-.212	.340	.054
Favorecedor (niño-madre)	.007	.028	.061	.036
Rechazo (padre)	.490	-.164	.279	.026
Calor emocional (padre)	-.166	.395	-.083	-.104
Sobreprotección (padre)	.208	-.147	.434	.023
Favorecedor (padre)	.174	-.079	.174	.275
Rechazo (madre)	1.000	-.366	.422	-.006
Calor emocional (madre)	-.366	1.000	-.245	-.071
Sobreprotección (madre)	.422	-.245	1.000	-.020
Favorecedor (madre)	-.006	-.071	-.020	1.000
Sexo	-.154	-.032	-.131	.043
Edad	.042	-.248	.165	-.005
Probl. conduct. tempr.	.154	-.169	.299	.050
Anteced. psicopat.	.123	-.080	.137	.123
Nivel socieconómico	-.152	-.088	.026	-.042
¿Hijo necesita ayuda?	.131	-.337	271	045

Coeficientes de correlación de Spearman (continuación)

	Sexo	Edad	Probl conduct tempr	Anteced psicopat	Nivel socieconómico	¿Hijo necesita ayuda?
Rechazo (niño-padre)	161	736	090	165	- 047	324
Calor emocional (niño-padre)	049	- 206	- 196	- 191	- 055	- 327
Sobreprotección (niño-padre)	127	673	043	114	- 081	186
Favorecedor (niño-padre)	- 094	- 333	066	039	164	- 084
Rechazo (niño-madre)	162	726	091	138	- 022	311
Calor emocional (niño-madre)	016	- 180	- 169	- 167	- 084	- 352
Sobreprotección (niño-madre)	179	666	035	128	- 064	168
Favorecedor (niño-madre)	- 085	-.271	095	034	165	- 053
Rechazo (padre)	- 170	003	212	230	- 184	087
Calor emocional (padre)	- 016	-.245	- 131	- 134	114	-.217
Sobreprotección (padre)	- 143	091	149	036	043	148
Favorecedor (padre)	002	103	021	089	- 056	022
Rechazo (madre)	- 154	042	154	123	- 152	131
Calor emocional (madre)	- 032	- 248	- 169	- 080	- 088	- 337
Sobreprotección (madre)	- 131	165	299	137	026	271
Favorecedor (madre)	043	- 005	050	123	- 042	045
Sexo	1 000	219	- 266	- 032	- 102	033
Edad	219	1 000	- 054	143	- 045	312
Probl conduct tempr	-.266	- 054	1 000	173	- 055	172
Anteced psicopat	- 032	143	173	1 000	- 120	177
Nivel socieconómico	- 102	- 045	- 055	- 120	1 000	148
¿Hijo necesita ayuda?	033	312	172	177	148	1 000

8
Consideraciones finales

Una vez presentadas detalladamente las conclusiones y discusión de los resultados de nuestro estudio en el capítulo 6, y tras aplicar dichas conclusiones a un problema real en el capítulo 7, a continuación se presentan un conjunto de reflexiones de carácter general sobre la relevancia de nuestra investigación, sus limitaciones, las líneas de estudio futuras, etc., con las que pretendemos ubicar nuestro trabajo desde la perspectiva más amplia del análisis de la calidad de los datos.

Hace ya bastantes meses, cuando nos planteamos la posibilidad de trabajar sobre el problema de datos incompletos, decidimos que nuestra meta debía ser aportar una solución sencilla a un problema engorroso. Sin duda, el segundo calificativo empleado no es exagerado; ¿quién no ha topado con el escollo de qué hacer ante una matriz en la que faltan datos? Nuestro dilema era cómo conseguir que el calificativo "sencillo" lo fuera, no tanto para nosotros, hasta cierto punto acostumbrados a especificar modelos, comprobar supuestos y descubrir relaciones que a otros podrían pasar inadvertidas, sino para el colectivo de profesionales de la psicología que, al no disponer de una suficiente formación estadística, necesitan soluciones simples que puedan aplicar sin incertidumbre sobre las consecuencias de sus operaciones sobre los datos.

El tradicional método *listwise* de eliminación de registros con datos incompletos, descalificado por diferentes autores, y a pesar de ello empleado masivamente en la investigación actual, sólo puede ser relegado al olvido si su heredero goza de la única virtud que, indiscutiblemente, se le puede atribuir: simplicidad. Las redes neuronales artificiales, que constituyen la implementación de los modelos conexionistas de procesamiento en paralelo, no requieren un profundo conocimiento matemático previo a su uso aplicado. La base de ello se halla en una particular virtud de este tipo de modelos, consistente en que la función que relaciona variables independientes con dependientes se obtiene empíricamente a partir de los datos disponibles, dando respuesta al clásico interrogante de si el modelo se debe ajustar a los datos o viceversa (Sprent, 1998).

Otros procedimientos habituales, particularmente la imputación del valor medio o la moda, también poseen la cualidad de ser aplicados con suma simplicidad, pero en determinadas circunstancias conllevan sesgos tan importantes que su uso es más perjudicial que beneficioso. En este sentido, creemos que un aspecto relevante de nuestro trabajo consiste en que la selección de la mejor estrategia para imputar los valores faltantes está en función del tipo de variable a imputar y

del nivel de correlación en los datos. Si se considera además la distinción entre modelos explicativos y modelos predictivos, presentada en el capítulo 6, las propuestas generales que se derivan de nuestros resultados se pueden resumir en:

- En análisis de tipo univariado el mejor método de imputación cuando las variables están, como mínimo, moderadamente correlacionadas, es mediante red neuronal RBF. En caso de ausencia de correlación se debe imputar el valor medio o la moda, según se trate de una variable cuantitativa o categórica.

- En un problema de clasificación a uno de dos grupos que presentan una frecuencia similar, si existen unas variables predictoras de especial relevancia, o los valores perdidos se agrupan en unas determinadas medidas, éstas han de ser imputadas con red neuronal RBF, previa imputación del resto de variables con la media o la moda. En la situación complementaria, cuando ninguna variable independiente tiene especial relevancia en el diseño y, además, los valores perdidos se distribuyen entre todas ellas, el procedimiento óptimo consiste en utilizar un valor fuera del rango válido para codificar los valores faltantes, acompañando esta operación con la generación de una variable indicadora de presencia/ausencia de valor por cada variable incompleta.

- Independientemente de la técnica de imputación seleccionada, el procedimiento más óptimo para estimar los parámetros del modelo de clasificación es emplear una red neuronal del tipo perceptrón multicapa (MLP). En este caso, para examinar el efecto de una variable independiente sobre la dependiente se debe recurrir a los análisis de sensibilidad.

- Cuando el porcentaje de valores desconocidos es elevado (aproximadamente a partir del 10%), cualquier estrategia de análisis es incapaz de compensar la pérdida de información, si bien las técnicas recomendadas minimizan dicha pérdida.

Todo estudio responde menos preguntas de las que plantea, y en este sentido el nuestro no es una excepción. A continuación presentamos una propuesta sobre los aspectos, no incluidos en el presente trabajo, que a nuestro juicio deben orientar la investigación futura sobre el análisis de datos incompletos en general, y sobre la aplicación de redes neuronales artificiales en particular.

La literatura sobre redes neuronales artificiales sugiere dedicar una especial atención al preprocesamiento de la información, fase previa al análisis con redes neuronales, que incluye principalmente transformaciones de variables continuas con la finalidad de normalizarlas, y ponderaciones sobre el tamaño muestral para equilibrar la distribución de frecuencias de las variables categóricas (Sarle, 1998a; Smith, 1993; SPSS Inc., 1997a). Sin embargo, en el uso aplicado de los modelos de red neuronal, pocos son los investigadores que siguen tales recomendaciones. Así, con el objetivo de acercarnos lo más fielmente posible a

la realidad en el uso aplicado de los modelos de red, los datos empleados en nuestro estudio no han sido sometidos a transformaciones previas a su modelización. Es de esperar que el preprocesamiento de los datos mejore sensiblemente los resultados obtenidos, si bien este aspecto debería ser empíricamente confirmado en futuras investigaciones.

Para comparar los resultados de la clasificación conseguidos mediante redes neuronales hemos tomado como referencia los obtenidos con el modelo de regresión logística. Actualmente existen decenas de procedimientos de clasificación, y si bien es cierto que muchos de ellos apenas son utilizados dada su escasa implementación en los paquetes de análisis estadístico de uso habitual, no lo es menos que los avances en la tecnología de las comunicaciones permitirán en un futuro próximo disponer masivamente de tales herramientas. Posteriores estudios deberían comparar otras técnicas de clasificación como, por ejemplo, métodos no paramétricos para estimar funciones de densidad de probabilidad (*k* vecinos más cercanos, regresión con funciones *kernel*, etc.), métodos de clasificación basados en pruebas exactas (pruebas de permutación, pruebas de aleatorización, etc.), remuestreo *bootstrap*, algoritmos genéticos, etc., además de otras topologías de redes supervisadas diferentes a las MLP y RBF (por ejemplo *Learning Vector Quantization* -LVQ-, redes Bayesianas, etc.). Mención especial requieren los algoritmos de inducción de reglas, ya que, al tratar los valores perdidos como un valor más de los datos, permiten incorporarlos al análisis de forma natural, sin ningún tipo de operación previa.

Las conclusiones de nuestro estudio son aplicables a problemas de clasificación en que la variable dependiente es binaria y, aproximadamente, equiprobable. Se deberían realizar experimentos de simulación para estudiar la generalización de nuestros resultados cuando una categoría de la variable dependiente es mucho más frecuente que la otra, así como cuando se trata de una variable categórica con más de dos categorías. Del mismo modo, tiene un especial interés, por el amplio conjunto de problemas reales que abarca, la modelización de datos incompletos en que la variable dependiente es de naturaleza continua.

Desde la perspectiva más general de la calidad de datos, posteriores investigaciones deberían indagar la robustez de las redes neuronales frente a otros modelos de clasificación cuando existen valores aberrantes en los datos, resultantes, por ejemplo, de un error de medida, de una respuesta incoherente o, sencillamente, de un error durante el registro de los datos.

El tema abordado en la presente tesis es muy amplio y en consecuencia, como además se deduce de las numerosas líneas de estudio futuro que hemos sugerido, queda mucho trabajo por hacer al respecto. No obstante, consideramos que las conclusiones específicas que se desprenden de nuestros resultados son de gran utilidad, especialmente por la elevada cantidad de problemas reales a los que son

aplicables. Por otra parte, la exhaustiva revisión sobre los aspectos teóricos del análisis estadístico de datos incompletos y de los modelos de redes neuronales artificiales, realizada en la primera parte de la tesis, puede ser de gran utilidad para aquellos investigadores que se inicien en esta línea de estudio.

Con nuestro trabajo también hemos pretendido establecer, con la máxima rigurosidad posible, la metodología de trabajo que debe guiar la investigación futura; en este sentido, apostamos por los experimentos de simulación estadística como herramienta básica de investigación, ya que permite manipular, y por tanto controlar, diferentes aspectos que pueden incidir en las conclusiones finales.

Sin embargo, desde mi punto de vista, la principal contribución del presente trabajo se dará en la medida en que sensibilice a sus lectores sobre la necesidad de dedicar, al menos una parte de su tiempo, tanto al análisis de los valores faltantes como a la aplicación de nuevas tecnologías, como las redes neuronales, en dichos análisis.

REFERENCIAS BIBLIOGRÁFICAS

Afifi, A.A. y Elashoff, R.M. (1966). Missing observations in multivariate statistics I: Review of the literature. *Journal of the American Statistical Association, 61*, 595-604.

Allison, P.D. (1982*). Maximum likelihood estimation of linear models when data are missing* (Manuscrito no publicado). Philadelphia: University of Pennsylvania, .

Almeida, L.B. (1987). *A learning rule for asynchronus perceptrons with feedback in a combinatorial enviroment.* IEEE first international conference on neural networks, San Diego.

Amari, S.I. (1977). Neural theory of association and concept formation. *Biological Cibernetics, 26*, 175-185.

Anderson, J.A. (1972). A simple neural network generating an interactive memory. *Mathematical Biosciences, 14*, 197-220.

Anderson, J.A. (1977). Neuronal models with cognitive implicatiosn. En D. Laberge y (Eds.). *Basic processes in reading perception and comprehension* (pp. 27-90). Hillsdale: Erlbaum.

Anderson, J.A. y Rosenfeld, E. (Eds.). (1988). *Neurocomputing: foundations of research.* Cambridge: MIT Press.

Anderson, J.A., Silverstein, J.W., Ritz, S.A. y Jones, R.S. (1977). Distinctive features, categorical perception and probability learning: some applications of a neural model. *Psychological Review, 84*, 413-451.

Anderson, T.W. (1957). Maximum likelihood estimates for the multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association, 52*, 200-203.

Ash, T. (1989). *Dynamic node creation in backpropagation networks* (Informe técnico N° 8901). San Diego: Institute for Cognitive Science, Universidad de California San Diego.

Ato, M. y López, J.A. (Eds.). (1997). *IV Simposio de metodología de las ciencias del comportamiento.* Murcia: Servicio de publicaciones. Universidad de Murcia.

Azorín, F. y . (1986). *Métodos y aplicaciones del muestreo.* Madrid: Alianza Editorial.

Barndorff-Nielsen, O.E., Jensen, J.L. y Kendall, W.S. (Eds.). (1993). *Networks and chaos: statistical and probabilistic aspects.* Londres: Chapman and Hall.

Barron, A.R. (1991). Complexity regularization with application to artificial neural networks. En G. Roussas (Ed.). *Nonparametric functional estimation and related topics* (pp. 561-576). Berlín: Kluwer.

Bartlett, P.L. (1997). For valid generalization, the size of the weights is more important than the size of the network. En M.C. Mozer, M.I. Jordan y T. Petsche (Eds.). *Advances in neural information processing systems (NIPS)* (pp. 134-140). Cambridge: MIT Press.

Barto, A.G. y Sutton, R.S. (1981). *Goal seeking components for adaptative intelligence: an initial assessment* (Informe técnico Nº AFWAL-TR-81-1070). Dayton: Air Force Wright aeronautical laboratories/Avionics laboratory.

Battiti, R. (1992). First and second order methods for learning: between steepest descent and Newton's method. *Neural Computation, 4*(2), 141-166.

Baum, E.B. y Haussler, D. (1989). What size net gives valid generalization?. *Neural Computation, 1*(1), 151-160.

Baum, L.E., Petrie, T., Soules, G. y Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist., 41*, 164-171.

Beale, E.M.L. y Little, R.J.A. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society, B37*, 129-145.

Becker, S. y Le Cun, Y. (1988). Improving the convergence of backpropagation learning with second order methods. En D. Touretsky, G.E. Hinton y T.J. Sejnowski (Eds.). *Proceedings of the 1988 Connectionist Models Summer School* (pp. 29-37). San Mateo: Morgan Kaufmann.

Bertsekas, D.P. (1995). *Nonlinear programming*. Belmont: Athena Scientific.

Bienenstock, E., Fogelman-Souli, F. y Weisbuch, G. (Eds.). (1986). *Disordered systems and biological organization*. Berlín: Springer-Verlag.

Bishop, C.M. (1995). *Neural networks for pattern recognition*. Oxford: Oxford University Press.

Bloomhead, D.S. y Lowe, D. (1988). Multi-variable functional interpolation and adaptative networks. *Complex Systems, 2*, 749-749.

Bonelli, P. y Parodi, A. (1991). An efficient classifier system and its experimental comparisons with two representative learning methods on three medical domains. *Proceedings of the International Conference on Genetic Algorithms*, 288-295.

Boswell, R.A. (1992). *HyperNewID and NewID* (Informe técnico Nº TI/P2154/RAB/4/9.2). Glasgow: Turing Institute.

Box, G.E.P. y Tiao, G.C. (1973). *Bayesian inference in statistical analysis*. Massachusetts: Addison-Wesley.

Breiman, L. y Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation (con discusión). *Journal of the American Statistical Association, 80*, 580-619.

Breiman, L., Friedman, J.H., Olshen, R.A. y Stone, C.J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth and Brooks.

Bremner, F.J., Gotts, S.J. y Denham, D.L. (1994). Hinton diagrams: Viewing connection strengths in neural networks. *Behavior Research Methods, Instruments, & Computers. Scientific Visualization, 26*, 215-218.

Buck, S.F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society, B22*, 302-306.

Burke, H., Rosen, D. y Goodman, P. (1994). Comparing the prediction accuracy of statistical models and artificial neural networks in breast cancer. En J.D. Cowan, G. Tesauro y J. Alspector (Eds.). *Advances in neural information processing systems (NIPS)*. San Mateo: Morgan Kaufmann.

Carpenter, G.A. y Grossberg, S. (1987a). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics and Image Processing, 37*, 54-115.

Carpenter, G.A. y Grossberg, S. (1987b). ART2: self-organization of stable category recognition codes for analog input patterns. *Applied Optics, 26*, 4919-4930.

Carpenter, G.A. y Grossberg, S. (1990). ART3: hierarchical search using chemical transmitters in self-organizing pattern recognition architectures. *Neural Networks, 3*(4), 129-152.

Casdagli, M. y Eubank, S. (Eds.). (1992). *Nonlinear modeling and forecasting*. Massachusetts: Addison-Wesley.

Castro, J., Toro, J., Arrindell, W.A., Van Der Ende, J. y Puig, J. (1990). Perceived parental rearing style in Spanish adolescent, children and their parents: three new forms of the EMBU. En C.S. Stefanis, C.R. Soldatos y A.D. Rabavilas (Eds.)*Psychiatry: a world perspective. Social psychiatry: ethics and law, history of psychiatry: psychiatric education* (pp. 340-344). Amsterdam: Elsevier Science Publishers B.V.

Castro, J., Toro, J., Van Der Ende, J. y Arrindell, W.A. (1993). Exploring the feasibility of assessing perceived parental rearing styles in Spanish children with the EMBU. *International Journal of Social Psychiatry, 39*, 47-57.

Catalina, A. (1996). *Introducción a las redes neuronales artificiales*. Universidad de Valladolid, Facultad de informática. Acceso HTTP: http://www.gui.uva.es.

Cater, J.P. (1987). Succesfully using peak learning rates of 10 (and greater) in back-propagation networks with the heuristic learning algorithm. *Proceedings of the IEEE International conference on neural networks*, II, 645-652.

Chatterjee, S. y Laudato, M. (1995). *Statistical applications of neural networks* (Informe). Massachusetts: Northeastern University Boston, .

Cheng, B. y Titterington, D.M. (1994). Neural networks: a review from a statistical perspective. *Statistical Science, 9*(1), 2-54.

Clark, P. y Niblett, T. (1988). The CN2 induction algorithm. *Machine Learning, 3*, 261-283.

Cobos, A. (1995). El síndrome GIGO. *JANO. Notas de Metodología y Estadística, XLIX*, 481-482.

Cochran, W.G. (1977). *Sampling techniques* (3ª ed.). Nueva York: John Wiley & Sons.

Cohen, J. y Cohen, P. (1975). *Applied multiple regression/correlation analisys for the behavioral sciences*. Hillsdale: Lawrence Erlbaum.

Colledge, M.J., Johnson, J.H., Paré, R. y Sande, I.G. (1978*). Large scale imputation of survey data* (Proceedings). Washington, DC: American Statistical Association, Business and economics section.

Collins, L.M. y Seitz, L.A. (Eds.). (1994). *Advances in data analysis for prevention intervention research*. Washington, DC: National Institute on Drug Abuse. (NIDA Research Monograph 142).

Cowan, J.D., Tesauro, G. y Alspector, J. (Eds.). (1994). *Advances in neural information processing systems (NIPS)* (Vol. 6). San Mateo: Morgan Kaufmann.

Cox, D.R. y Hinkley, D.V. (1974). *Theoretical statistics*. Nueva York: John Wiley & Sons.

Craik, K.J.W. (1943). *The nature of explanation*. Cambridge: Cambridge University Press.

Crawford, S.L. (1989). Extensions to the cart algorithm. *International Journal of Man-Machine Studies, 31*, 197-217.

Crick, F.H. y Asanuma, C. (1986). Certain aspects of the anatomy and physiology of the cerebral cortex. En J.L. McClelland y D.E. Rumelhart (Eds.). *Parallel distributed processing: explorations in the microstructure of cognition. Psychological and biological models*. Cambridge: MIT Press.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematical Control, Signal and Systems, 2*, 303-314.

Dagli, C.H., Kumara, S.R.T. y Shin, Y.C. (Eds.). (1991). *Intelligent engineering systems through artificial neural networks*. Nueva York: ASME Press.

Dempster, A.P., Laird, N.M. y Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B39*, 1-38.

Dempster, A.P. y Rubin, D.B. (1983). Overview. En W.G. Madow, I. Olkin y D.B. Rubin (Eds.). *Incomplete data in sample surveys: Theory and annotated bibliography* (pp. 3-10). Nueva York: Academic Press.

DeRouin, E., Brown, J., Beck, H., Fausset, L. y Schneider, M. (1991). Neural network training on unequally represented classes. En C.H. Dagli, S.R.T. Kumara y Y.C. Shin (Eds.). *Intelligent engineering systems through artificial neural networks* (pp. 135-141). Nueva York: ASME Press.

Ding, X., Denoeux, T. y Helloco, F. (1993). Tracking rain cells in radar images using multilayer neural networks. *Proceedings of ICANN'93*, 962-967. Springer-Verlag.

Dixon, W. (1993). *Biomedical computer programs (BMDP)* [Programa para ordenador]. SPSS Inc. (Productor). Chicago: SPSS Inc. (Distribuidor).

Dixon, W.J. (1983). *BMDP Statistical software*. Berkeley: University of California Press.

Dixon, W.J. (1990). *BMDP Statistical software*. Berkeley: University of California Press.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvements on cross-validation. *Journal of the American Statistical Association, 78*, 316-331.

Fahlman, S.E. (1988). Faster-learning variations on back-propagation: an empirical study. En D. Touretsky, G.E. Hinton y T.J. Sejnowski (Eds.). *Proceedings of the 1988 Connectionist Models Summer School* (pp. 38-51). San Mateo: Morgan Kaufmann.

Fahlman, S.E. y Lebiere, C. (1990). The cascade-correlation learning architecture. En D. Touretsky (Ed.). *Advances in neural information processing systems (NIPS)* (pp. 524-532). San Mateo: Morgan Kaufmann.

Fausset, L. (1994). *Fundamentals of neural networks*. New Jersey: Prentice-Hall.

Feldman, J.A. (1985). Connectionist models and their applications: intoduction. *Cognitive Science, 9*, 1-2.

Feldman, J.A. y Ballard, D.H. (1982). Connectionist models and their properties. *Cognitive Science, 6*, 205-254.

Fernández, E.J. (1993). Modelos conexionistas: conceptos generales, origen y evolución. En I. Olmeda y S. Barba-Romero (Eds.). *Redes neuronales artificiales: fundamentos y aplicaciones* (pp. 3-21). Madrid: Universidad Alcalá de Henares. Servicio de publicaciones.

Fix, E. y Hodges, J.L. (1951). *Discriminatory analysis, nonparametric estimation: consistency properties* (Informe Nº 4). Texas: School of Aviation Medicine.

Flexer, A. (1995). *Connectionist and statisticians, friends or foes?*. The Austrian Research Institute for Artificial Intelligence. Acceso FTP: Nombre del servidor: ai.univie.ac.at Archivo: oefai-tr-95-06_ps(1).ps.

Frean, M. (1990). The upstart algorithm: a method for constructing and training feedforward neural networks. *Neural Computation, 2*, 198-209.

Freedman, V.A. y Wolf, D.A. (1995). A case study on the use of multiple imputation. *Demography, 32*(3), 459-470.

Friedman, J.H. (1991). Multivariate adaptive regression splines (con discusión). *Annals of Statistics, 19*, 1-141.

Friedman, J.H. y Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association, 76*, 817-823.

Fukushima, K. (1975). Cognitron: a self-organizing multilayered neural network. *Biological Cibernetics, 20*, 121-136.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cibernetics, 36*, 193-202.

Fukushima, K. (1988). Neocognitron: a hierarchical neural network model capable of visual pattern recognition. *Neural Networks, 1*(2), 119-130.

Fukushima, K., Miyake, S. y Ito, T. (1983). Neocognitron: a neural network model for a mechanism of visual pattern recognition. *IEEE Transactions on Systems, Man and Cybernetics, 13*, 826-834.

Funahashi, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks, 2*, 183-192.

Furlanello, C., Giuliani, D. y Trentin, E. (1995). Connectionist speaker normalization with generalized resource allocating networks. En G. Tesauro, D. Touretsky y T.K. Leen (Eds.). *Advances in neural information processing systems (NIPS)* (pp. 867-874). Cambridge: MIT Press.

Garson, G.D. (1991). A comparison of neural network and expert systems algorithms with common multivariate procedures for analysis of social science data. *Social Science Computer Review, 9*, 399-434.

Gassman, J.J., Owen, W.W., Kuntz, T.E., Martin, J.P. y Amoroso, W.P. (1995). Data quality assurance, monitoring, and reporting. *Controlled Clinical Trials, 16*, 104S-136S.

Geman, S., Bienenstock, E. y Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation, 4*(1), 1-58.

Geman, S. y Geman, D. (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transactions on Pattern Ananlysis and Machine Learning*, 6, 721-741.

Ghahramani, Z. y Jordan, M.I. (1994). *Learning from incomplete data* (Informe técnico). Massachusetts: Massachusetts Institute of Technology, Artificial Intelligence laboratory.

Gill, P.E., Murray, W. y Wright, M.H. (1981). *Practical optimization*. Londres: Academic Press.

Goodman, R.M. y Smyth, P. (1989). The induction of probabilistic rule sets - the Itrule algorithm. En B. Spatz (Ed.)*Proceedings of the sixth international workshop on machine learning* (pp. 129-132). San Mateo, CA: Morgan Kaufmann.

Graham, J.W., Hofer, S.M. y Mackinnon, D.P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: an application of maximum likelihood procedures. *Multivariate Behavioral Research, 31*(2), 197-218.

Graham, J.W., Hofer, S.M. y Piccinin, A.M. (1994). Analisys with missing data in drug prevention research. En L.M. Collins y L.A. Seitz (Eds.). *Advances in data analysis for prevention intervention research* (pp. 13-63). Washington, DC: National Institute on Drug Abuse.

Grossberg, S. (1976). Adaptative pattern classification and universal recoding, I:parallel development and coding of neural feature detectors. *Biological Cibernetics, 23*, 121-134.

Grossberg, S. (1978). A theory of visual coding, memory and development. En E.L.J. Leeuwenberg y H.F. Buffart (Eds.). *Formal theories of visual perception*. Nueva York: Wiley.

Grossberg, S. (1980). How does a brain build a cognitive code?. *Psychological Review, 87*, 1-51.

Grossberg, S. (1982). *Studies of mind and brain*. Dordrecht (Holland): Reidel.

Grossberg, S. (1987). *The adaptative brain: cognition, learning, reinforcement and rhythm* (Vol. I). Amsterdam: North-Holland.

Grossberg, S. (1987a). *The adaptative brain: vision, speech, language and motor control* (Vol. II). Amsterdam: North-Holland.

Groves, R.M. (1989). *Survey errors and survey costs*. New York: John Wiley & Sons.

Gurney, K. (en prensa). *Neural nets*. Londres: UCL Press Limited.

Haitovsky, Y. (1968). Missing data in regression analysis. *Journal of the Royal Statistical Society B, 30*, 67-81.

Hartley, H.O. (1958). Maximum likelihood estimation from incomplete data. *Biometrics, 14*, 174-194.

Hartley, H.O. y Hocking, R.R. (1971). The analysis of incomplete data. *Biometrics, 27*, 783-808.

Hashem, S. (1992). Sensitivity analysis for feedforward artificial neural networks with differentiable activation functions. *Proceedings of the International Joint Conference on Neural Networks*, I, 419-424.

Haykin, S. (1994). *Neural networks: a comprehensive foundation*. Nueva York: Macmillan.

Healy, M.J.R. y Westmacott, M. (1956). Missing values in experiments analyzed on automatic computers. *Applied Statistics, 5*, 203-206.

Hebb, D. (1985). *The organization of behavior*Trad.). Madrid: Debate. (Traducción del original La organización de la conducta, 1949).

Hecht-Nielsen, R. (1989). Theory on the back-propagation neural network. *Proceedings of the International Joint Conference on Neural Networks*, I, 593-606.

Hecht-Nielsen, R. (1990). *Neurocomputing*. Massachusetts: Addison-Wesley.

Henery, R.J. (1994). Classification. En D. Michie, D.J. Spiegelhalter y C.C. Taylor (Eds.). *Machine learning, neural and statistical classification* (pp. 6-16). Londres: Ellis Horwodd.

Henery, R.J. (1994b). Methods of comparison. En D. Michie, D.J. Spiegelhalter y C.C. Taylor (Eds.). *Machine learning, neural and statistical classification* (pp. 107-124). Londres: Ellis Horwodd.

Herzog, T.N. y Rubin, D.B. (1983). Using multiple imputation to handle nonresponse in sample surveys. En W.G. Madow, I. Olkin y D.B. Rubin (Eds.). *Incomplete data in sample surveys: Theory and annotated bibliography*. Nueva York: Academic Press.

Hilera, J.R. y Martínez, V.J. (1995). *Redes neuronales artificiales: fundamentos, modelos y aplicaciones*. Madrid: RA-MA.

Hinton, G.E., Sejnowski, T.J. y Ackley, D.H. (1985). A Learning algorithm for Boltzman machines. *Cognitive Science, 9*, 147-149.

Hinton, G.E. y Shallice, T. (1994). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review, 98*, 74-92.

Hirose, Y., Yamashita, K. y Hijiya, S. (1991). Backpropagation algorithm which varies the number of hidden units. *Neural Networks, 4*, 61-66.

Hollingshead, A.B. (1975). *Four factor index of social status* (Manuscrito no publicado)Universidad de Yale, Department of Sociology.

Hopfield, J.J. (1982). Neural networks and psysical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, 79*, 2554-2558.

Hopfield, J.J. y Tank, D.W. (1985). Neural computation of decisions in optimization problems. *Biological Cibernetics, 52*, 141-152.

Hopfield, J.J. y Tank, D.W. (1986). Computing with neural circuits. *Science, 233*, 625-633.

Hornik, K. (1993). Some new results on neural network approximation. *Neural Networks, 6*, 1069-1072.

Hornik, K., Stinchcombe, M. y White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*(5), 359-366.

Hornik, K., Stinchcombe, M. y White, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks, 3*, 551-560.

Howell, D.C. (1996). *Treatment of missing data*. Universidad de Vermont. Acceso HTTP: http://moose.uvm.edu/~dhowell/StatPages/StatHomePage.html.

Huang, W.Y. y Lippmann, R.P. (1987). Comparisons between neural net and conventional classifiers. *Proceedings of the IEEE International conference on neural networks*, I, 485-494.

Huberty, C.J. y Julian, M.W. (1995). An ad hoc analysis strategy with missing data. *Journal of Experimental Education, 63*, 333-342.

Irie, B. y Miyake, S. (1988). Capabilities of three-layered perceptrons. *Proceedings of the IEEE International conference on neural networks*, I, 161-172.

Jacobs, R.A. (1988). Increased rates of convergence through learning rate adaptation. *Neural Networks, 1*(4), 295-308.

Jarrett, R.G. (1978). The analysis of designed experiments with missing observations. *Applied Statistics, 27*, 38-46.

Kim, J.O. y Curry, J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods and Research, 6*, 215-240.

Klimasauskas, C.C. (Ed.). (1989). *The 1989 neuro-computing bibliography*. Cambridge: MIT Press.

Kohonen, T. (1972). Correlation matrix memories. *IEEE Transactions on Computers*, C21, 353-359.

Kohonen, T. (1977). *Associative memory*. Berlín: Springer-Verlag.

Kohonen, T. (1984). *Self-organization and associative memory*. Berlín: Springer-Verlag.

Kohonen, T. (1995). *Self-organizing maps*. Berlín: Springer-Verlag.

Koistinen, P. y Holmstrom, L. (1992). Kernel regression and backpropagation training with noise. En J.E. Moody, S.J. Hanson y R.P. Lippmann (Eds.). *Advances in neural information processing systems (NIPS)* (pp. 1033-1039). San Mateo: Morgan Kaufmann.

Kolmogorov, A.N. (1957). On the representation of continuous functions of several variables by means of superpositions of continuous functions of one variable. *Doklady Akademii Nauk SSSR, 114*, 953-956.

Kung, S.Y. y Hwang, J.N. (1988). Al algebraic projection technique for optimal hidden units size and learning rates in back-propagation networks. *Proceedings of the IEEE International conference on neural networks*, I, 363-370.

Laberge, D. y (Eds.). (1977). *Basic processes in reading perception and comprehension*. Hillsdale: Erlbaum.

Lachenbruch, P.A. y Mickey, M.R. (1975). *Discriminant analysis*. Nueva York: Hafner Press.

Lashley, K.S. (1950). *In search of the engram*. Cambridge: Cambridge University Press.

Le Cun, Y. (1986). Learning processes in an assimetric treshold network. En E. Bienenstock, F. Fogelman-Souli y G. Weisbuch (Eds.). *Disordered systems and biological organization*. Berlín: Springer-Verlag.

Le Cun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.A. et al . (1990). Handwritten digit recognition with a backpropagation neural network. En D. Touretsky (Ed.). *Advances in neural information processing systems (NIPS)* (pp. 396-404). San Mateo: Morgan Kaufmann.

Le Cun, Y., Denker, J.S. y Solla, S.A. (1990). Optimal brain damage. En D. Touretsky (Ed.). *Advances in neural information processing systems (NIPS)* (pp. 598-605). San Mateo: Morgan Kaufmann.

Lee, S. y Kil, R.M. (1988). Multilayer feedforward potential function network. *Proceedings of the IEEE International conference on neural networks*, II, 161-172.

Leeuwenberg, E.L.J. y Buffart, H.F. (Eds.). (1978). *Formal theories of visual perception.* Nueva York: Wiley.

Li, K.H. (1985). *Hypothesis testing in multiple imputation - with emphasis on mixed-up frequencies in contingency tables.* University of Chicago. Tesis Doctoral.

Lindley, D.V. (1965). *Introduction to probability and statistics from a Bayesian viewpoint* (Vol. II). Cambridge: Cambridge University Press.

Lippmann, R.P., Moody, J.E. y Touretsky, D. (Eds.). (1991). *Advances in neural information processing systems (NIPS)* (Vol. 3). San Mateo: Morgan Kaufmann.

Little, R.J.A. y Rubin, D.B. (1987). *Statistical analysis with missing data.* Nueva York: John Wiley & Sons.

Luria, A.R. (1973). *The working brain.* Londres: Penguin.

MacKay, D.J.C. (1992). Bayesian model comparison and backprop nets. En J.E. Moody, S.J. Hanson y R.P. Lippmann (Eds.). *Advances in neural information processing systems (NIPS)* (pp. 839-846). San Mateo: Morgan Kaufmann.

MacKay, D.J.C. (1992a). Bayesian interpolation. *Neural Computation, 4*, 415-447.

MacKay, D.J.C. (1992b). A practical Bayesian framework for backpropagation. *Neural Computation, 4*, 448-472.

Madow, W.G. y Olkin, I. (Eds.). (1983). *Incomplete data in sample surveys: Symposium on incomplete data, Proceedings* (Vol. III). Nueva York: Academic Press.

Madow, W.G., Olkin, I. y Rubin, D.B. (Eds.). (1983). *Incomplete data in sample surveys: Theory and annotated bibliography* (Vol. II). Nueva York: Academic Press.

Martín , M. (1993). Redes de propagación hacia adelante. En I. Olmeda y S. Barba-Romero (Eds.). *Redes neuronales artificiales: fundamentos y aplicaciones* (pp. 67-82). Madrid: Universidad Alcalá de Henares. Servicio de publicaciones.

Masters, T. (1995). *Advanced algorithms for neural networks: a C++ sourcebook.* Nueva York: John Wiley & Sons.

Matthai, A. (1951). Estimation of parameters from incomplete data with application to design of sample surveys. *Sankhya, 2*, 145-152.

McClelland, J.L. y Rumelhart, D.E. (Eds.). (1986). *Parallel distributed processing: explorations in the microstructure of cognition. Psychological and biological models* (Vol. II). Cambridge: MIT Press.

McClelland, J.L., Rumelhart, D.E. y Hinton, G.E. (1986). El atractivo del procesamiento distribuido en paralelo. En D.E. Rumelhart y J.L. McClelland (Eds.). *Introducción al procesamiento distribuido en paralelo* (pp. 39-80). Madrid: Alianza Editorial.

McClelland, J.L. y Rumelhart, D.E. (1988). *Explorations in parallel distributed processing*. Cambridge: MIT Press.

McCulloch, W.S. y Pitts, W. (1943). A logical calculus of the ideas immanent en nervous acivity. *Bulletin of Mathematical Biophysics, 5*, 115-133.

McKendrick, A.G. (1926). Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematics Society, 44*, 98-130.

Michalski, R.S., Carbonell, J.G. y Mitchell, R.M. (Eds.). (1983). *Machine learning: an artificial intelligence approach*. Palo Alto, CA: Tioga.

Michie, D., Spiegelhalter, D.J. y Taylor, C.C. (1994). Introduction. En D. Michie, D.J. Spiegelhalter y C.C. Taylor (Eds.). *Machine learning, neural and statistical classification* (pp. 1-5). Londres: Ellis Horwodd.

Michie, D., Spiegelhalter, D.J. y Taylor, C.C. (Eds.). (1994). *Machine learning, neural and statistical classification*. Londres: Ellis Horwodd.

Minsky, M. y Papert, S. (1969). *Perceptrons*. Cambridge: MIT Press.

Mira, J. (1993). Fundamentos biológicos de las redes de neuronas artificiales. En I. Olmeda y S. Barba-Romero (Eds.). *Redes neuronales artificiales: fundamentos y aplicaciones* (pp. 23-41). Madrid: Universidad Alcalá de Henares. Servicio de publicaciones.

Moody, J.E. (1992). The effective number of parameters: an anlysis of generalization and regularization in nonlinear learning systems. En J.E. Moody, S.J. Hanson y R.P. Lippmann (Eds.). *Advances in neural information processing systems (NIPS)* (pp. 847-854). San Mateo: Morgan Kaufmann.

Moody, J.E. y Darken, C. (1988). Learning with localized receptive fields. En D. Touretsky, G.E. Hinton y T.J. Sejnowski (Eds.). *Proceedings of the 1988 Connectionist Models Summer School* (pp. 133-143). San Mateo: Morgan Kaufmann.

Moody, J.E. y Darken, C. (1989). Fast learning in networks on locally-tuned processing units. *Neural Computation, 1*, 281-294.

Moody, J.E., Hanson, S.J. y Lippmann, R.P. (Eds.). (1992). *Advances in neural information processing systems (NIPS)* (Vol. 4). San Mateo: Morgan Kaufmann.

Mozer, M.C., Jordan, M.I. y Petsche, T. (Eds.). (1997). *Advances in neural information processing systems (NIPS)* (Vol. 9). Cambridge: MIT Press.

Mueller, P. y Insua, D.R. (1995). *Issues in Bayesian analysis of neural network models* (Informe técnico Nº 95-31). Institute of Statistics and Decision Sciences. Acceso FTP: Nombre del servidor: isds.duke.edu Directorio: pub/WorkingPapers/ Archivo: 95-31.ps.

Murtagh, F. (1996). *Neural networks: historical background.* Acceso HTTP: http://infm.ulst.ac.uk/research/ac460/ip9a/node3.html.

Neal, R.M. (1996). *Bayesian learning for neural networks.* Nueva York: Springer-Verlag.

Nelson, M.M. y Illingworth, W.T. (1991). *A practical guide to neural nets* (3ª ed. rev.). Massachusetts: Addison-Wesley.

NGuyen, D. y Widrow, B. (1990). Improving the learning speed of two-layer neural networks by choosing initial values of the adaptative weights. *Proceedings of the IEEE International conference on neural networks*, III, 21-26.

Nigrin, A. (1993). *Neural networks for pattern recognition.* Cambridge: MIT Press.

Olmeda, I. (1993). Aprendizaje y generalización. En I. Olmeda y S. Barba-Romero (Eds.). *Redes neuronales artificiales: fundamentos y aplicaciones* (pp. 43-63). Madrid: Universidad Alcalá de Henares. Servicio de publicaciones.

Olmeda, I. y Barba-Romero, S. (Eds.). (1993). *Redes neuronales artificiales: fundamentos y aplicaciones.* Madrid: Universidad Alcalá de Henares. Servicio de publicaciones.

Orchard, T. y Woodbury, M.A. (1972). A missing information principle: theory and applications. *Proceedings of the 6th. Berkeley Symposium on Mathematics, Statistics and Probability*, 1, 697-715.

Orme, J.G. y Reis, J. (1991). Multiple regression with missing data. *Journal of Social Service Research, 15*, 61-91.

Orr, M.J.L. (1996). *Introduction to radial basis functions.* Edimburgh: University of Edinburgh, Centre for cognitive science.

Ozturk, A. y Romeu, J.L. (1992). A new method for assessing multivariate normality with graphical applications. *Communications in Statistics-Simulation, 21*, 15-34.

Parker, D. (1982). *Learning logic invention* (Informe técnico Nº S81-64). Stanford: Office of Technology Licensing, Stanford university.

Parker, D. (1985). *Learning logic* (Informe técnico Nº TR-87). Cambridge: Center for computational research in economics and management science.

Parker, D. (1987). Learning algorithms for connectionist networks: applied gradient methods of nonlinear optimization. *Proceedings of the IEEE International conference on neural networks*, II, 593-600.

Pearce, S.C. (1965). *Biological statistics: an introduction.* New York: McGraw-Hill.

Peña, D. (1991). *Estadística, modelos y métodos: Fundamentos* (2ª ed. rev.) (Vol. I). Madrid: Alianza Editorial.

Pineda, F.J. (1987). Generalization of back-propagation to recurrent neural networks. *Physical Review Letters, 59*, 2229-2232.

Pitarque, A. y Ruiz, J.C. (1996). Encoding missing data in back-propagation neural networks. *Psicológica, 17*, 83-91.

Prechelt, L. (1994). *A set of neural network benchmark problems and benchmarking rules* (Informe técnico N° 21/94). Karlsruhe: Universidad de Karlsruhe.

Prechelt, L. (1994b). *Encoding missing values* [Archivo de datos informático]. Acceso e-Mail: ml-connectionists-request@TELNET-1.SRV.CS.CMU.EDU Nombre de la lista: Connectionists. Emisor: prechelt@ira.uka.de.

Preece, D.A. (1971). Iterative procedures for missing values in experiments. *Technometrics, 13*, 743-753.

Prina Ricotti, L., Ragazzini, S. y Martinelli, G. (1988). Learning of word stress in a sub-optimal second order back-propagation neural network. *Proceedings of the IEEE International conference on neural networks*, I, 355-362.

Quinlan, J.R. (1986). Induction of decision trees. *Machine Learning, 1*, 81-106.

Quinlan, J.R. (1983). Learning efficient classification procedures and their application to chess end games. En R.S. Michalski, J.G. Carbonell y R.M. Mitchell (Eds.). *Machine learning: an artificial intelligence approach* (pp. 463-482). Palo Alto, CA: Tioga.

Quinlan, J.R. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

Redman, T.C. (1992). *Data quality: management and technology*. New York: Bantam Books.

Reich, W., Shayka, J. y Taibleson, C. (1991). *Diagnostic Interview for Children and Adolescent-Revised DSM-III-R version 7.2.* (Manuscrito no publicado)Universidad de Washington, Division of Child Psychiatry. St. Louis.

Riedmiller, M. y Braun, H. (1993). A direct adaptative method for faster backpropagation learning: the RPROP algorithm. *Proceedings of the IEEE International conference on neural networks*, I.

Ripley, B.D. (1993). Statistical aspects of neural networks. En O.E. Barndorff-Nielsen, J.L. Jensen y W.S. Kendall (Eds.). *Networks and chaos: statistical and probabilistic aspects*. Londres: Chapman and Hall.

Ripley, B.D. (1995). *Multiple imputation and simulation methods* (Informe) . Acceso HTTP: http://www.stats.ox.ac.uk/~ripley/talks.html.

Ripley, B.D. (1996). *Pattern recognition and neural networks*. Cambridge: Cambridge University Press.

Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics, 14*(3), 1080-1100.

Rohwer, R., Wynne-Jones, M. y Wysotzky, F. (1994). Neural networks. En D. Michie, D.J. Spiegelhalter y C.C. Taylor (Eds.). *Machine learning, neural and statistical classification* (pp. 84-106). Londres: Ellis Horwodd.

Rosenblatt, F. (1958). The Perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review, 65*, 386-408.

Rosenblatt, F. (1960). Perceptron simulation experiments. *Proc. IRE, 48*, 301-309.

Rosenblatt, F. (1962). *Principles of neurodynamics*. Nueva York: Spartan.

Roussas, G. (Ed.). (1991). *Nonparametric functional estimation and related topics*. Berlín: Kluwer.

Rubin, D.B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.

Rubin, D.B. (1978). *Multiple imputation in sample surveys* (Proceedings). Washington, DC: American Statistical Association, Survey research methods section.

Rubin, D.B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business, Economy and Statistics, 4*, 87-94.

Rubin, D.B. y Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association, 81*, 366-374.

Ruiz, J.C., Pitarque, A. y Gómez, M. (1997). Codificación de los valroes faltantes en el entrenamiento de redes neuronales. En M. Ato y J.A. López (Eds.)*IV Simposio de metodología de las ciencias del comportamiento* (pp. 245-251). Murcia: Servicio de publicaciones. Universidad de Murcia.

Rumelhart, D.E., Hinton, G.E. y McClelland, J.L. (1986). Un marco general para el procesamiento distribuido en paralelo. En D.E. Rumelhart y J.L. McClelland (Eds.). *Introducción al procesamiento distribuido en paralelo* (pp. 81-109). Madrid: Alianza Editorial.

Rumelhart, D.E., Hinton, G.E. y Williams, R.J. (1986). El aprendizaje de las representaciones internas por propagación del error. En D.E. Rumelhart y J.L. McClelland (Eds.). *Introducción al procesamiento distribuido en paralelo* (pp. 211-252). Madrid: Alianza Editorial.

Rumelhart, D.E., Hinton, G.E. y Williams, R.J. (1986a). Learning representations by back-propagation. *Nature, 323*, 533-536.

Rumelhart, D.E. y McClelland, J.L. (Eds.). (1992). *Introducción al procesamiento distribuido en paralelo* (García, J.A., Trad.). Madrid: Alianza Editorial. (Traducción del original Paralell distributed processing, 1986).

Rumelhart, D.E. y McClelland, J.L. (1986). Modelos PDP y cuestiones generales de la ciencia cognitiva. En D.E. Rumelhart y J.L. McClelland (Eds.). *Introducción al procesamiento distribuido en paralelo* (pp. 143-176). Madrid: Alianza Editorial.

Sande, I.G. (1983). Hot deck imputation procedures. En W.G. Madow y I. Olkin (Eds.). *Incomplete data in sample surveys: Symposium on incomplete data, Proceedings*. Nueva York: Academic Press.

Sarle, W.S. (1994). *Neural networks and statistical models* Proceedings of the nineteenth anual SAS users group international conference, Massachusetts.

Sarle, W.S. (1995). Stopped training and other remedies for overfitting. *Proceedings of the 27th Symposium on the Interface of Computing Science and statistics*, 1, 352-360.

Sarle, W.S. (1996). *Neural network and statistical jargon*. Acceso FTP: Nombre del servidor: sas.com Directorio: pub/neural Archivo: jargon.

Sarle, W.S. (1998). *Neural network FAQ: Introduction*. Acceso FTP: Nombre del servidor: sas.com Directorio: pub/neural Archivo: faq.html.

Sarle, W.S. (1998a). *Neural network FAQ: Learning*. Acceso FTP: Nombre del servidor: sas.com Directorio: pub/neural Archivo: faq2.html.

Sarle, W.S. (1998b). *Neural network FAQ:Generalization*. Acceso FTP: Nombre del servidor: sas.com Directorio: pub/neural Archivo: faq3.html.

Schafer, J.L. (1994). *Analysis of incomplete multivariate data by simulation*. Londres: Chapman and Hall.

Schafer, J.L. (1997). *Analysis of incomplete multivariate data*. Londres: Chapman and Hall.

Schrodt, P.A. (1991). Prediction of interstate conflict outcomes using a neural network. *Social Science Computer Review, 9*, 359-380.

Segovia, J. (1993). Redes de neuronas recurrentes. En I. Olmeda y S. Barba-Romero (Eds.). *Redes neuronales artificiales: fundamentos y aplicaciones* (pp. 127-138). Madrid: Universidad Alcalá de Henares. Servicio de publicaciones.

Sejnowski, T.J. y Rosenberg, C.R. (1986). Paralell networks that learns to pronounce English text. *Complex Systems, 1*, 145-168.

Sethi, I.K. y Jain, A.K. (Eds.). (1991). *Artificial neural networks and statistical pattern recognition*. Amsterdam: Elsevier Science Publishers B.V.

Sethi, I.K. y Otten, M. (1990). Comparison between entropy net and decision tree classifiers. *Proceedings of the International Joint Conference on Neural Networks*, I, 63-68.

Sharpe, P.K. y Solly, R.J. (1995). Dealing with missing values in neural network-based diagnostic systems. *Neural Computing and Applications, 3*, 73-77.

Sietsma, J. y Dow, R.J.F. (1991). Creating artificial neural networks tha generalize. *Neural Networks, 4*, 67-79.

Silva, F.M. y Almeida, L.B. (1990). Acceleration techniques for the back-propagation algorithm. *Lecture Notes in Computer Science, 412*, 110-119.

Smith, M. (1993). *Neural Networks for Statistical Modeling*. Nueva York: Van Nostrand Reinhold.

Smith, T.W. (1991). *An analysis of missing income information on the General Social Surveys* (Informe metodológico Nº 71), Universidad de Chicago. Acceso HTTP: http://www.icpsr.umich.edu/gss/report/m-report/meth71.htm.

Smolensky, P. (1988). On the proper treatment of connectionism. *The Behavioral and Brain Sciences, 11*, 1-74.

Spatz, B. (Ed.). (1989). *Proceedings of the sixth international workshop on machine learning*. San Mateo, CA: Morgan Kaufmann.

Sprent, P. (1998). *Data driven statistical analysis*. London: Chapman and Hall.

SPSS Inc. (1996). *Statistical Package for Social Sciences 7.5.2* [Programa para ordenador]. SPSS Inc. (Productor). Chicago: SPSS Inc. (Distribuidor).

SPSS Inc. (1997). *Neural Connection 2.0* [Programa para ordenador]. SPSS Inc. (Productor). Chicago: SPSS Inc. (Distribuidor).

SPSS Inc. (1997a). *Neural Connection 2.0 Users Guide* [Manual de programa para ordenador]. Chicago: SPSS Inc.

Stefanis, C.S., Soldatos, C.R. y Rabavilas, A.D. (Eds.). (1990). *Psychiatry: a world perspective. Social psychiatry: ethics and law, history of psychiatry: psychiatric education* (Proceedings of the VIIIth World Congress of Psychiatry, Athens) (Vol. 4). Amsterdam: Elsevier Science Publishers B.V.

Stenberg, S. (1969). Memory scanning: mental processes revealed by reaction-time experiments. *American Scientist, 57*, 421-457.

Stern, C.S. (1997). New tools for analyzing data, making decisions and forecasting trends. *Drug Benefit Trends, 9*(5), 43-49.

Stricker, R. (1998). *Imbalanced classes* [Archivo de datos informático] Nombre de la lista: comp.ai.neural-nets Emisor: rst@tbus.muc.de.

Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics, 1*, 49-58.

Sutton, R.S. (1986). Two problems with backpropagation and other steepest-descent learning procedures for networks. *Proceedings of the eighth annual conference on the Cognitive Science Society*, 823-831.

Tank, D.W. y Hopfield, J.J. (1987). Collective computation in neuronlike circuits. *Scientific American, 257*, 104-114.

Tanner, M.A. y Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association, 82*(398), 528-550.

Taylor, M.A. y Amir, N. (1994). The problem of missing clinical data for research in psychopathology. *The Journal of Nervous and Mental Disease, 182*, 222-229.

Tesauro, G., Touretsky, D. y Leen, T.K. (Eds.). (1995). *Advances in neural information processing systems (NIPS)* (Vol. 7). Cambridge: MIT Press.

Thodberg, H.H. (1996). A review of Bayesian neural networks with an application to near infrared spectroscopy. *IEEE Transactions on Neural Networks, 7*, 56-72.

Thrun, S., Bala, J., Bloedorn, E. y Bratko, I. (Eds.). (1991). *The MONK's problem - a performance comparison of different leraning algorithms.* Pittsburg: Carnegie Mellon University.

Thrun, S., Mitchell, T. y Cheng, J. (1991). The MONK's comparison of learning algorithms - introduction and survey. En S. Thrun, J. Bala, E. Bloedorn y I. Bratko (Eds.). *The MONK's problem - a performance comparison of different leraning algorithms* (pp. 1-6). Pittsburg: Carnegie Mellon University.

Touretsky, D. (Ed.). (1990). *Advances in neural information processing systems (NIPS)* (Vol. 2). San Mateo: Morgan Kaufmann.

Touretsky, D., Hinton, G.E. y Sejnowski, T.J. (Eds.). (1988). *Proceedings of the 1988 Connectionist Models Summer School.* San Mateo: Morgan Kaufmann.

Tresp, V., Ahmad, S. y Neuneier, R. (1994). Training neural networks with deficient data. En J.D. Cowan, G. Tesauro y J. Alspector (Eds.). *Advances in neural information processing systems (NIPS).* San Mateo: Morgan Kaufmann.

Turing, A. (1937). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society, 2*(42), 230-265.

Turing, A. (1950). Computing machinery and intelligence. *Mind, 49*, 433-460.

Vach, W. (1994). *Logistic regression with missing values in the covariates.* New York: Springer-Verlag.

Vach, W. y Blettner, M. (1991). Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *American Journal of Epidemiology, 134*, 895-907.

Vamplew, P. y Adams, A. (1992). Missing values in a backpropagation neural net. *Proceedings of the 3rd. Australian Conference on Neural Networks (ACNN)*, I, 64-66.

Vapnik, V.N. (1992). Principles of risk minimization for learning theory. En J.E. Moody, S.J. Hanson y R.P. Lippmann (Eds.). *Advances in neural information processing systems (NIPS)* (pp. 831-838). San Mateo: Morgan Kaufmann.

Von Der Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striata cortex. *Kybernetik, 14*, 85-100.

Von Eye, A. (1990). *Statistical methods in longitudinal research.* San Diego: Academic Press.

Wachter, K.W. y Trusell, J. (1982). Estimating historical heights. *Journal of the American Statistical Association, 77*, 279-301.

Wasserman, P. (1988). *Combined backpropagation-Cauchy machine* Proceedings of the International Neural Network Society, .

Watrous, R.L. (1987). Learning algorithms for connectionist networks: applied gradient methods of nonlinear optimization. *Proceedings of the IEEE International conference on neural networks*, II, 619-628.

Weigend, A.S., Huberman, B.A. y Rumelhart, D.E. (1990). Predicting the future: a connectionist approach. *International Journal of Neural Systems, 2*, 193-209.

Weigend, A.S., Huberman, B.A. y Rumelhart, D.E. (1992). Predicting sunspots and exchange rates with connectionist networks. En M. Casdagli y S. Eubank (Eds.). *Nonlinear modeling and forecasting.* Massachusetts: Addison-Wesley.

Weigend, A.S., Rumelhart, D.E. y Huberman, B.A. (1991). Generalization by weight-elimination with application to forecasting. En R.P. Lippmann, J.E. Moody y D. Touretsky (Eds.). *Advances in neural information processing systems (NIPS)* (pp. 875-882). San Mateo: Morgan Kaufmann.

Weir, M. (1991). A method for self-determination of adaptative learning rates in backpropagation. *Neural Networks, 4*, 371-379.

Weiss, S.M. y Kulikowski, C.A. (1991). *Computer systems that learn.* San Mateo: Morgan Kaufmann.

Werbos, P.J. (1974*). Beyond regression: new tools for prediction an analysis in behavioral sciences.* Harvard University. Tesis doctoral.

Werbos, P.J. (1991). Links between artificial neural networks (ANN) and statistical pattern recognition. En I.K. Sethi y A.K. Jain (Eds.). *Artificial neural networks and statistical pattern recognition* (pp. 11-33). Amsterdam: Elsevier Science Publishers B.V.

Werbos, P.J. (1994). *The roots of backpropagation.* Nueva York: John Wiley & Sons.

White, H. (1989). Learning in artificial neural networks. *Neural Computation, 1*, 425-464.

White, H. (1992). *Artificial neural networks: approximation and learning theory.* Oxford: Blackwell.

Whitehead, J.C. (1994). Item nonresponse in contingent valuation: Should CV researchers impute values for missing independent variables?. *Journal of Leisure Research, 26*, 296-303.

Widrow, B. y Hoff, M.E. (1960). *Adaptative switching circuits*Institute of radio engineers, Western Electronic Show and Convention, Convention Record, parte 4, 96-104.

Widrow, B. y Lehr, M.A. (1990). 30 years of adaptative neural networks: perceptron, adaline, and backpropagation. *Proceedings of the IEEE International conference on neural networks*, 78(9), 1415-1442.

Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics, 11*, 95-103.

Wynne-Jones, M. (1992). Node splitting: a constructive algorithm for feed-forward neural networks. En J.E. Moody, S.J. Hanson y R.P. Lippmann (Eds.). *Advances in neural information processing systems (NIPS)* (pp. 1072-1079). San Mateo: Morgan Kaufmann.

Wynne-Jones, M. (1991). Constructive algorithms and pruning: improving the multi layer perceptron. *Proceedings of IMACS'91, the 13th World Congress on Computation and Applied Mathematics*, 2, 747-750.

Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Emp. J. Exp. Agric., 1*, 129-142.

Z Solutions. (1997). *An introduction to neural networks*. Z Solutions. Acceso HTTP: http://www.zsolutions.com/index.htm.

Zurada, J.M. (1992). *Introduction to artificial neural systems*. Boston: PWS Publishing Company.

FUENTES DOCUMENTALES SOBRE REDES NEURONALES ARTIFICIALES EN INTERNET

En este apartado presentamos una selección de las fuentes documentales sobre redes neuronales artificiales que se pueden encontrar en Internet. No pretendemos compilar los infinitos recursos sobre redes disponibles, para lo cual ya existen multitud de buscadores como, por ejemplo, Yahoo, Altavista o Infoseek, sino enumerar los que consideramos más valiosos. Para ello hemos evaluado diferentes aspectos: calidad de la información, solidez de la entidad o persona responsable, periodicidad de las actualizaciones, velocidad de acceso al servidor y volumen de participación (en listas de correo y grupos de noticias).

Las fuentes documentales seleccionadas se presentan agrupadas en 5 categorías:

- *Catálogos bibliográficos*. Contienen cientos, y en algunos casos miles, de referencias a libros, artículos, informes técnicos, manuales, etc.
- *Revistas electrónicas*. Listado de las revistas electrónicas sobre redes neuronales artificiales accesibles de forma gratuita. Actualmente, la mayoría de revistas sólo incluyen los títulos y resúmenes de los artículos publicados en los últimos años.
- *Centros e Institutos*. Direcciones de los principales grupos de trabajo sobre redes neuronales artificiales.
- *Software*. Comprende una amplia relación de programas informáticos de redes neuronales artificiales. En muchos casos es posible obtener una copia de evaluación.
- *Listas de correo y Grupos de noticias*. Principales foros de debate sobre redes neuronales artificiales. Este tipo de fuente documental se caracteriza por su dinamismo y por la posibilidad de que participe toda la comunidad investigadora.

a. Catálogos bibliográficos

Austrian Research Institute for Artificial Intelligence.
http://www.ai.univie.ac.at/oefai/nn/conn_biblio.html

Bibliographic Search Tools Pacific Northwest Laboratory.
http://www.emsl.pnl.gov:2080/proj/neuron/neural/search.html

Bibliographic Search Pattern Recognition Group (TUD)
http://www.ph.tn.tudelft.nl/bibliographic.html

Dutch Foundation for Neural Networks
http://www.mbfys.kun.nl/SNN/Pointers/search.html

Neural Information Processing Systems
http://www.cs.cmu.edu/Groups/NIPS/

Neural Network Researchers' Database
http://www.neuronet.ph.kcl.ac.uk/neuronet/researchers.html

Neuroscience Web Search
http://www.acsiom.org/nsr/neuro.html

Sensor Signal and Information Processing (CSSIP)
ftp://ftp.cssip.edu.au/pub

University of Texas at Arlington
http://www-ee.uta.edu/ip/papers/papers.html

The Collection of Computer Science Bibliographies
http://liinwww.ira.uka.de/bibliography/Neural/index.html

b. Revistas electronicas

Connection Science
http://www.eeb.ele.tue.nl/neural/contents/connection_science.html

IEEE Transactions on Neural Networks
http://www.eeb.ele.tue.nl/neural/contents/ieee_trans_on_nn.html

International Journal of Neural Systems
http://www.wspc.co.uk/wspc/Journals/ijns/ijns.html

Network: Computation in Neural Systems
http://www.emsl.pnl.gov:2080/proj/neuron/journals/Network_Computation.html

Neural Computation
http://www.eeb.ele.tue.nl/neural/contents/neural_computation.html

Neural Networks
http://www.eeb.ele.tue.nl/neural/contents/neural_networks.html

Neural Network World
http://www.uivt.cas.cz/~dani/nnw.html

Neural Processing Letters
http://www.dice.ucl.ac.be/neural-nets/NPL/NPL.html

NeuroComputing
http://www.emsl.pnl.gov:2080/proj/neuron/journals/Neurocomputing.html

c. Centros e institutos

Austrian Research Institute for Artificial Intelligence (OFAI)
http://www.ai.univie.ac.at/oefai/nn/

Brigham Young University - Neural Networks and Machine Learning Lab
http://synapse.cs.byu.edu/home.html

Crin-Inria - Cortex Group
http://www.loria.fr/exterieur/equipe/rfia/cortex/cortex.html

Defence Research Agency - Pattern and Information Processing Group
http://www.dra.hmg.gb/cis5pip/Welcome.html

Dutch Foundation for Neural Networks
http://www.mbfys.kun.nl/SNN/

European Neural Network Society
http://www.neuronet.ph.kcl.ac.uk/neuronet/organisations/enns.html

Heudiasyc Lab - Neural Networks Research Group
http://www.hds.univ-compiegne.fr/WEB/scanu/RNA.html

International Neural Network Society
http://sharp.bu.edu/inns/

Italian Neural Network Society
http://mcculloch.ing.unifi.it/neural/siren/sirenEN.html

Japanese Neural Network Society
http://jnns-www.okabe.rcast.u-tokyo.ac.jp/jnns/home.html

Carnegie Mellon University - PDP Group
http://www.cnbc.cmu.edu/PDP++/

Centre for Neural Networks - King's College London
http://physig.ph.kcl.ac.uk/cnn/cnn.html/

Lebedev Physical Institute, Russia - Neural Network Group
http://canopus.lpi.msk.su/project31/

MIT - Center for Biological and Computational Learning
http://www.ai.mit.edu/projects/cbcl/web-homepage/web-homepage.html

Neural Computing Applications Forum
http://www.neuronet.ph.kcl.ac.uk/neuronet/organisations/ncaf.html

NEuroNet - King's College London
http://www.neuronet.ph.kcl.ac.uk/

Stuttgart Neural Network Simulator (SNNS)
http://www.informatik.uni-stuttgart.de/ipvr/bv/projekte/snns/snns.html

UCL-DICE Neural Net Group
http://www.dice.ucl.ac.be/neural-nets/NNgroup.html

Universidad de Buenos Aires
http://galileo.fi.uba.ar/

University of Cincinnati - Artificial Neural Systems Lab
http://www.ece.uc.edu/~ansl/

University of Florence - DSI Neural Networks
http://www-dsi.ing.unifi.it/neural/home.html

University of Illinois at Urbana-Champaign - Artificial Neural Networks and Computational Brain Theory Group
http://anncbt.ai.uiuc.edu/

University of Nijmegen Neural Network Group
http://www.mbfys.kun.nl/SNN/groups/nijmegen/

University of Texas Austin - Laboratory for Artificial Neural Systems
http://www.lans.ece.utexas.edu/

UTCS Neural Nets Research Group
http://www.cs.utexas.edu/users/nn/

d. Software

"AutoSet"
Common Sense Systems
http://www.commonsensesystems.com/

"BackPack Neural Network System"
Z SolutionS
http://www.zsolutions.com/index.htm

"BioNet Simulator"
Elonet Network Systems Ltd
http://www.ncsb-bionet.com/

"BrainMaker"
California Scientific Software
http://wallstreetdirectory.com/

"Clementine"
TAD Sistemas
http://www.tad.es/

"DataEngine"
MIT-Management Intelligenter
http://www.mitgmbh.de/

"LoFlyte"
Accurate Automation Corporation
http://www.accurate-automation.com/

"NetProphet"
International Neural Machines inc.
http://www.ineural.com/

"Neural Connection"
SPSS, Inc.
http://www.spss.com/software/Neuro/

"Neurascript"
Neurodynamics, Inc.
http://www.neurodynamics.com/

"NeuroGenetic Optimiser"
BioComp Systems
http://www.bio-comp.com/

"NeuroShell Predictor, NeuroShell Classifier, NeuroShell Trader"
Ward Systems Group, Inc.
http://www.wardsystems.com/

"NeuroSolutions"
NeuroDimension, Inc.
http://www.nd.com/

"NevProp"
University of Nevada Centre for Biomedical Research
ftp://ftp.scs.unr.edu/pub/goodman/nevpropdir/

"PDP++"
PDP++ Software Home Page - Carnegie Mellon University
http://www.cnbc.cmu.edu/PDP++/

"Pygmalion"
Esprit project
ftp://ftp.funet.fi/pub/sci/neural/sims/pygmalion.tar.Z

"Tcl/Tk"
Partek, Inc.
http://www.partek.com/

"Trajan"
Trajan Software Ltd
http://www.trajan-software.demon.co.uk/

e. Listas de correo y grupos de noticias

Biomedical Applications of Neural Networks Special Interest Group
Lista de correo. Dirección subscripción: Majordomo@mcs.anl.gov

Cellular Neural Networks - CNN
Lista de correo. Dirección subscripción: mb@tce.ing.uniroma1.it

Connectionist Mailing List
Lista de correo. Dirección subscripción: Connectionists-Request@cs.cmu.edu

International Student Society for Neural Networks News Group - ISSNNET
Grupo de noticias: Comp.org.issnnet

Neural Networks News Group
Grupo de noticias: comp.ai.neural-nets

Neuron Digest Mailing List
Lista de correo. Dirección subscripción: neuron-request@psych.upenn.edu

Neuron Emulation Mailing List
Lista de correo. Dirección subscripción: listserv@ucsd.edu

Polish Mailing List on Neural Networks
Lista de correo. Dirección subscripción: listserv@plaern.edu.pl