Introducción al DataMining

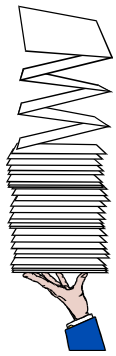
Lluís Garrido
garrido@ecm.ub.es
Universitat de Barcelona

Índice

- ¿Qué es el DataMining?
- ¿Qué puede hacer el DataMining?
- ¿Cómo hacer el DataMining? Técnicas
- Metodología del DataMining
  - Ejemplo 1: venta plan de pensiones
  - Ejemplo 2: predicción de mercado de capitales

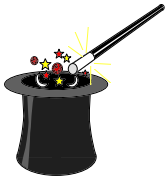
¿Qué es el DataMining?

- análisis y exploración automática
- de grandes bases de datos
- para extraer información útil
- y no evidente



¿Qué no es el DataMining?

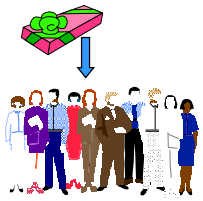
- Queries/ Informes
- Data Warehouse
- Sistemas Expertos
- Estadística
-

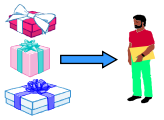


¿Por qué ahora?

- Tenemos grandes bases de datos
- cpu/$ cada vez mejor
- Productos comerciales para DataMining existen
- Competencia cada vez mayor

¿Qué puede hacer el DataMining?





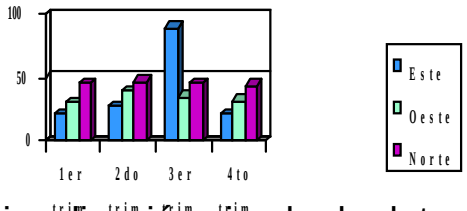
- **Clasificación**
  - Credit scoring
  - Fidelidad: ¿quién se dará de baja?
  - Mejora de campañas: ¿quién de mis clientes es más propenso a comprar mi producto?
  - Próximo producto: ¿qué producto ofrecer al cliente X?
- **Estimación**
  - ¿cuál es la nómina de este cliente?
- **Predicción**
  - ventas en el próximo mes
  - ¿cuál será el IBEX35 de la próxima semana?
- **Clusterización**
  - Como son nuestros clientes
  - por que compran ciertos productos y como evolucionan

En todos los casos utilizaremos la base de datos histórica para crear los modelos

¿Cómo hacer el DataMining? Técnicas

- Estadísticas
- Clusterización
- Árboles
- Redes neuronales
- Algoritmos genéticos
- .... (Market Basket Analysis, Memory Based Reasoning, Link Analysis,...)

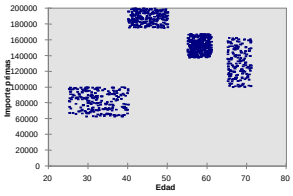
Estadísticas

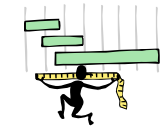


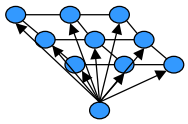
- Útiles en
  - Descripción y visualización simple de datos
  - Test de hipótesis
    - max, min, medias, sigma, frecuencias, ….
    - histogramas 2D
    - búsqueda de correlaciones
    - regresiones lineales
- Dificultades en
  - relaciones no lineales entre variables
  - distribuciones no gaussianas

Técnicas de clusterización

- Todas ellas agrupan los clientes pero con distintos métodos y criterios
- Las técnicas más usadas son
  - clásicas: Se agregan registros hasta llegar al número de grupos deseados o a la distancia mínima
  - redes neuronales
    - Kohonen. Método: el ganador se lo lleva todo
    - Neural-Gas. Simula un gas formado de moléculas
- Es muy recomendable hacer una Clusterización: Podemos descubrir agradable sorpresas







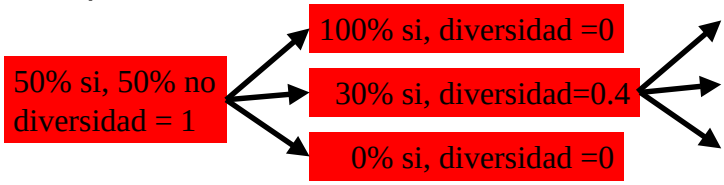
datos

Árboles

- Herramientas muy populares para clasificación
- Atractivo:
  - sus resultados pueden expresarse mediante reglas ⇒ ejecutables directamente en SQL
- Problemas:
  - el número de reglas generalmente es enorme
  - son superadas por las redes (predicción)

Árboles: ¿cómo crecen?

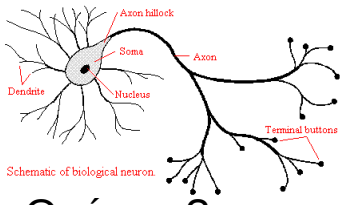
- Cada rama se divide en otras para disminuir la diversidad
  - diversidad: como más baja es, indica predominio de una clase

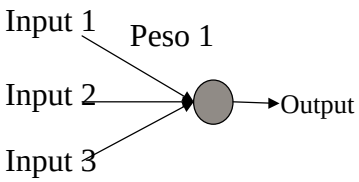


- El proceso termina por criterios de
  - ganancia de información, entropía,...

Redes Neuronales Multicapa

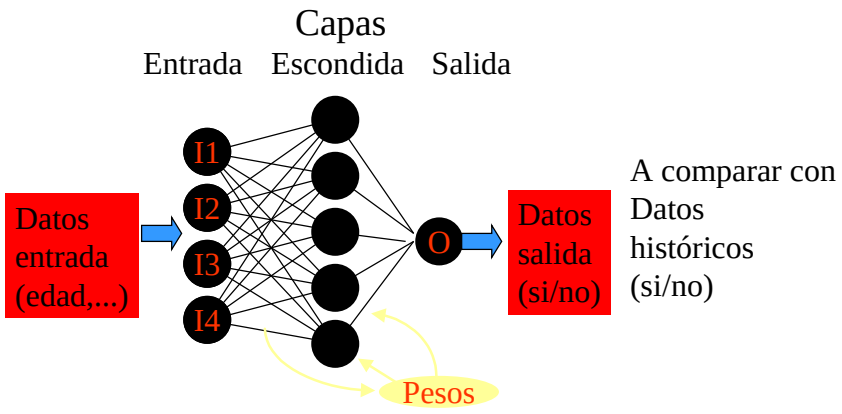
- Inspiradas en la estructura neuronal biológica





- ¿Qué son?
  - Grupo de neuronas (unidad básica de procesamiento) interconectadas con distinta influencia mutua (peso)
- Aprenden" a partir de ejemplos
  - como nosotros!!!!!
- Memoria
  - basada en los pesos

Redes neuronales



Si datos históricos ≠ datos salida ⇒ aprender

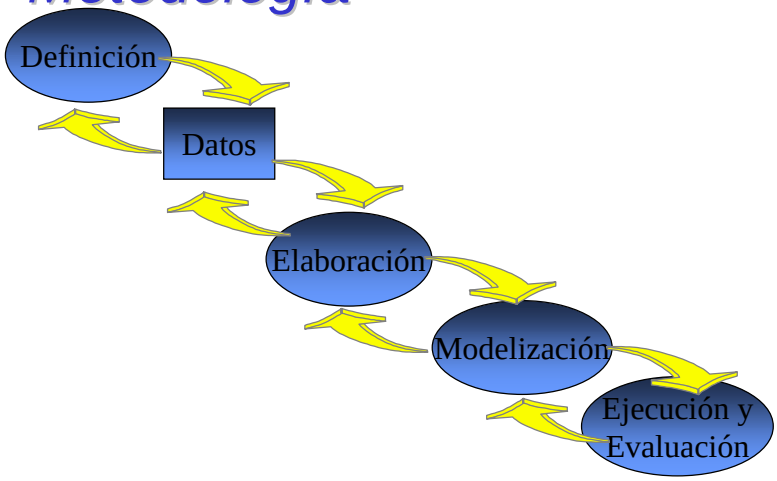
¿Para qué son útiles?

- Son extremadamente útiles en
  - clasificación
    - Objetivo: aprender la relación
      - datos históricos ⇒ si(=1), no(=0)
    - Para nuevos clientes
      - datos nuevo cliente ⇒ número entre 0 y 1
      - es la probabilidad de que este cliente nos diga sí
      - esta probabilidad junto a una estimación de beneficios nos permitirá decidir como actuar
- Son muy útiles en
  - estimación
  - predicción

Algoritmos genéticos

- Emulan la naturaleza:
  - tenemos una población de individuos (soluciones)
  - estos individuos sobreviven y se reproducen según sus cualidades
- Especializados en problemas de optimización con restricciones
  - búsqueda de la mejor solución entre muchas

Metodología



Ejemplo 1

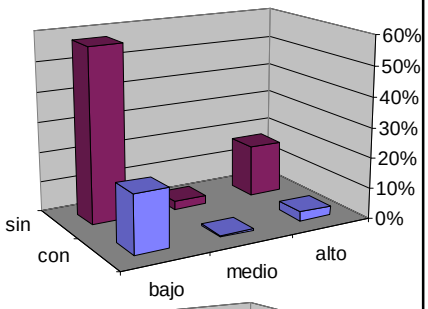
- Problema:
  - Venta de Plan de Pensiones
- Datos:
  - Datos de cliente
  - Datos de cuentas (selección de productos)
  - Muestra aleatoria de todos los clientes del banco
    - Clientes sin el producto
    - Clientes con el producto
    - varios años de datos por falta de estadística

Clustering (Neural Gas)

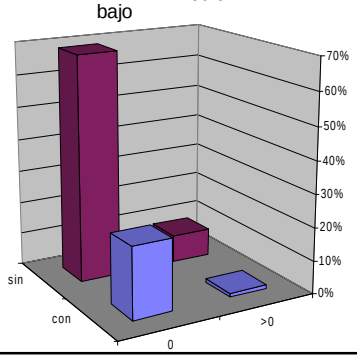
Variable	1	2	3	4	5	6	7
AF	**21**	**368**	81	**17**	**27**	103	**23**
Antiguedad	94	104	81	106	86	**144**	94
Domiciliacions	**65**	**48**	**34**	96	**222**	**50**	**411**
Edad	**64**	**132**	114	90	92	**142**	83
EstadoCivil_C	**0**	**153**	**135**	**153**	**146**	84	85
EstadoCivil_S	**336**	**4**	**10**	**2**	**5**	**55**	**133**
EstadoCivil_V	**0**	**16**	**131**	**10**	**5**	**672**	**32**
EstadoCivil_?	**41**	**26**	**222**	**74**	**220**	**27**	**220**
FInversion	**48**	**28**	**30**	**48**	**191**	**386**	113
Hipoteca	**29**	**6**	**18**	**47**	**39**	**2**	**1255**
IPF	**46**	**67**	**57**	**40**	**132**	**402**	**45**
Jubilacion	86	**15**	**73**	**198**	**182**	**36**	**193**
NivelCultural	120	**67**	82	118	115	77	**126**
NivelSE	102	88	93	110	104	102	107
Nomina	123	**8**	**15**	**64**	**353**	**8**	**240**
Pension	**8**	**303**	**69**	**13**	**4**	**275**	**32**
Pensiones	**9**	**4**	**10**	103	**508**	**50**	**258**
PrestamoP	82	**21**	**37**	**71**	**250**	**13**	**483**
Sexo_H	97	124	**56**	87	**155**	**62**	**142**
Sexo_M	104	**67**	**161**	118	**24**	**153**	**42**
Tarjeta_0	89	**135**	**134**	79	**45**	**135**	**55**
Tarjeta_1	122	**26**	**29**	**145**	**217**	**25**	**196**
Vista	77	105	**51**	**63**	**130**	**210**	124

Visualización simple (correlaciones)

SALDO CUENTA	bajo	medio	alto
con	19,1%	0,6%	3,0%
sin	57,6%	2,7%	17,0%
Factor	1/4	1/6	1/7



IPF	0	>0
con	21,5%	1,1%
sin	68,8%	8,6%
Factor	1/4	1/9



Competencia entre IPF y Plan de Pensiones

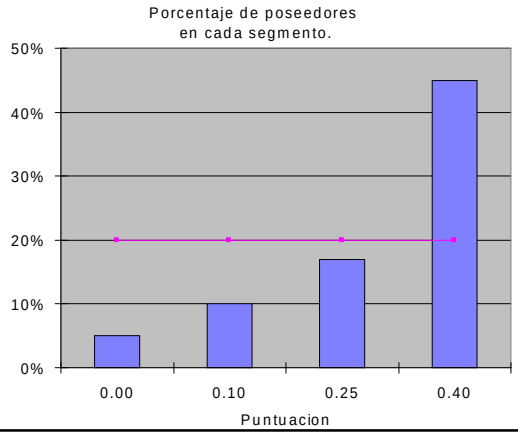
Resultados modelo neuronal

Objetivo: aprender la relación

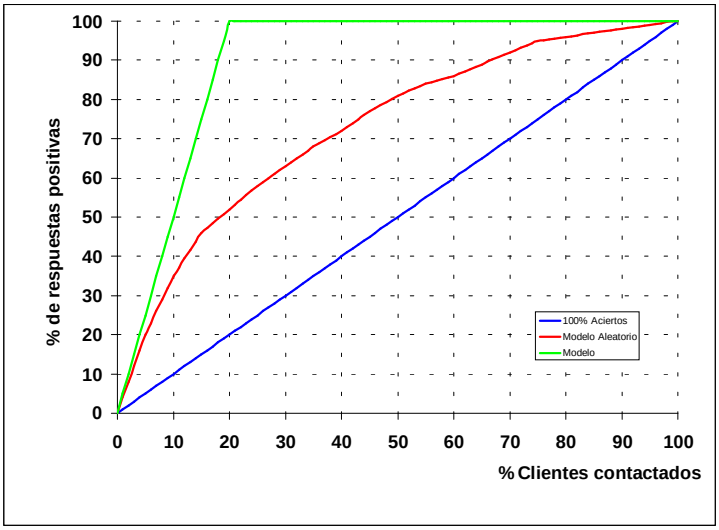
datos cliente ⇒ posee Plan Pensiones (=1), no posee (=0)

Finalizado el aprendizaje la red da una puntuación (de 0 a 1) para cada cliente que puede interpretarse como la propensión de que dicho cliente tenga un plan de pensiones.

Resultados sobre clientes no utilizados durante el aprendizaje:



Lift Chart



Sensibilidad

Variable	Sens.	Dirección	Desv. Stan.	Rango Bajo	Rango Alto
Cuentas	6,32	Negativa	2,16	-5,52	2,95
FIAMM	4,41	Negativa	1,87	-4,62	2,72
Fim1	3,06	Negativa	1,52	-3,85	2,13
Hipoteca	1,85	Positiva	1,10	-1,35	2,95
IPF	1,69	Negativa	1,12	-2,86	1,52
CuentaDep	0,95	Negativa	0,90	-2,14	1,39
CtaAhorro	0,77	Negativa	0,76	-1,92	1,06
SaldoCuentas	0,50	Negativa	0,62	-1,55	0,86
Tarjeta	0,48	Positiva	0,42	-0,27	1,37
DepFinanc	0,42	Negativa	0,59	-1,43	0,88
Fim2	0,40	Negativa	0,57	-1,39	0,85

Resultados

- Para desarrollar la campaña,
  - clientes con score máximo
  - clientes con score mínimo

Posibilidad de medición

- Resultados finales mostraban claramente mayor respuesta en segmento de alto score

  (factor 2)

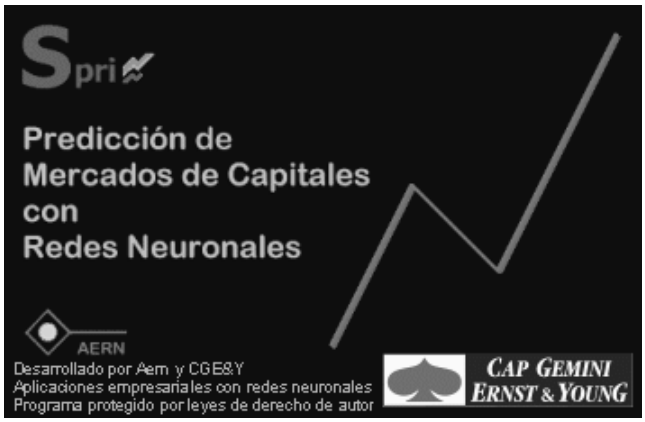
herramientas

Nombre	***Compañía***
Model1	Group1
Marksman	HNC
Enterprise Miner	SAS
Intelligent Miner	IBM
Clementine	SPSS
Sprinn	AERN & CAP GEMINI

La diferencia se centra en sus objetivos, la automatización, la cantidad de datos que puede utilizar, su integración en la base de datos,...

Predicción de Mercados de Capitales

AERN conjuntamente con **Cap Gemini Ernst & Young** y en colaboración de 10 personas del sector, ha desarrollado un aplicativo orientado a la predicción en el ámbito de Mercados Financieros



¿Qué es SpriNN Intitucional?

- Herramienta de predicción de series financieras.
- Se basa en redes neuronales
- Esta herramienta otorga flexibilidad al analista financiero
  - permite crear y evaluar sus propios modelos de predicción
  - permite ajustar el nivel de riego
  - permite utilizar series no financieras (opinión,...)
  - permite evaluar influencias
- Spri*NN* genera evaluaciones de riesgo y muestra oportunidades de inversión.

Funcionalidades

(I) Definición de la estructura de datos de cada serie

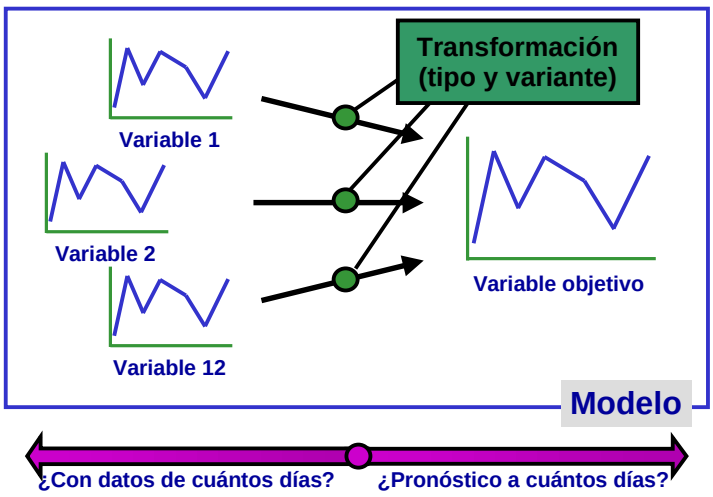
(II) Carga de datos según estructura

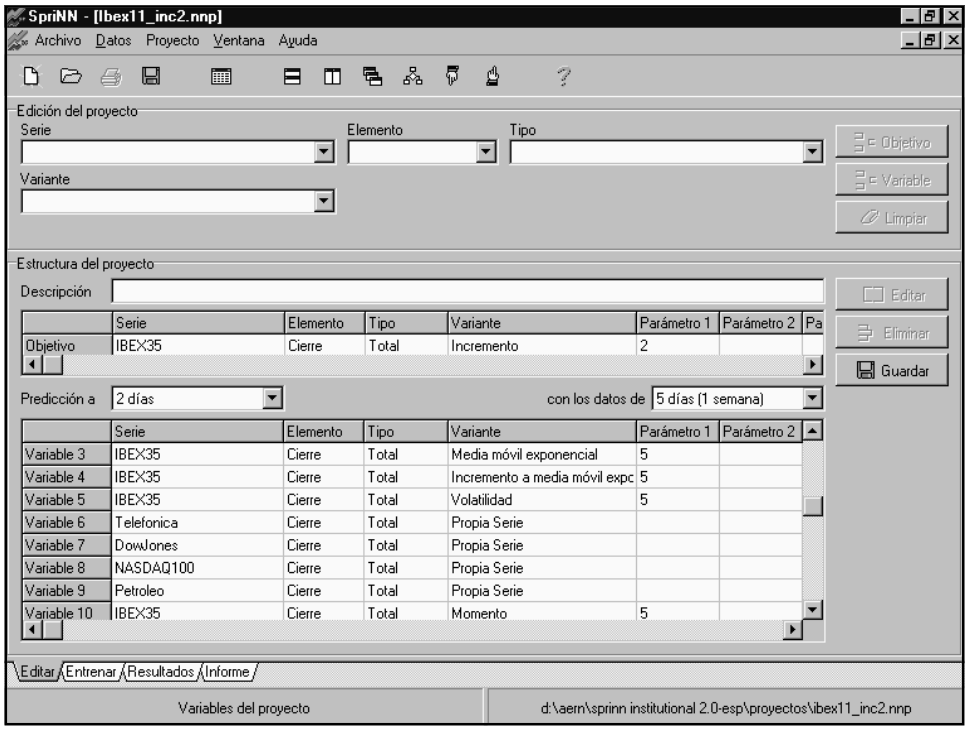
(III) Definición del modelo predictivo

(IV) Entrenamiento del modelo con Redes Neuronales

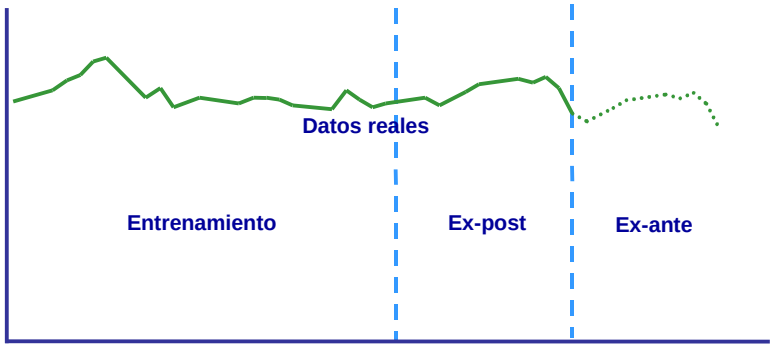
(V) Visualización de los resultados y recomendaciones

Definición del modelo predictivo

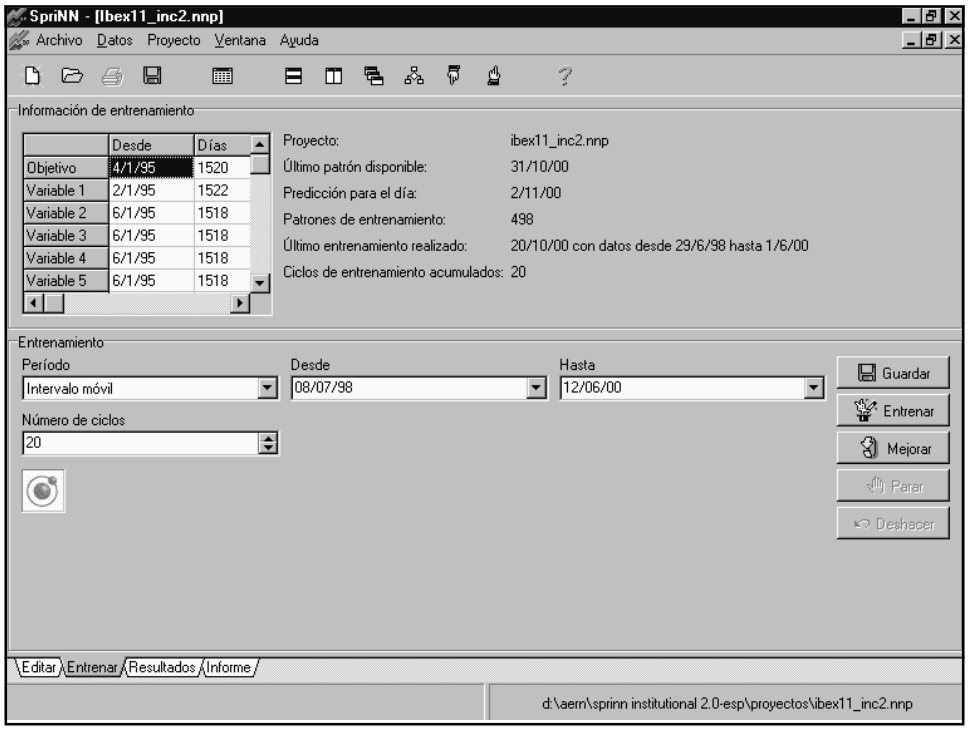


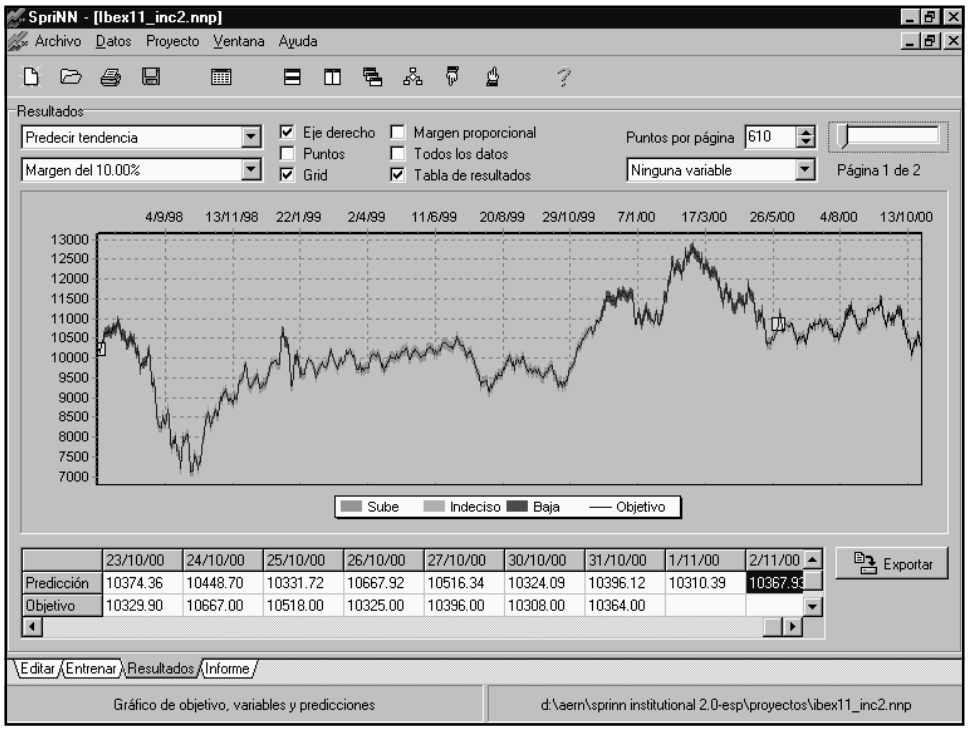


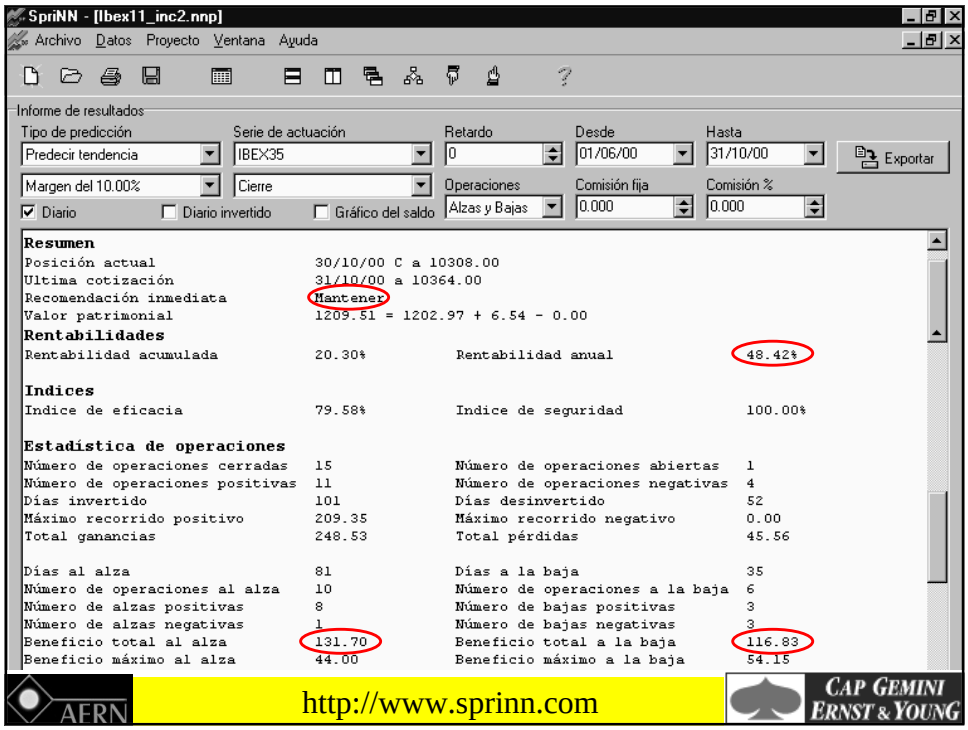
Entrenamiento, test y predicción



Una vez realizado el entrenamiento el análisis gráfico nos permite validar el modelo según su capacidad de predicción: 1) en el periodo de entrenamiento, y 2) en el intervalo ex-post. Todo esto con el objetivo de poder predecir ex-ante.


SpriNN - [Ibex11_inc2.nnp]
Archivo Datos Proyecto Ventana Ayuda
Información de entrenamiento
Desde Días
Objetivo 4/1/95 1520
Variable 1 2/1/95 1522
Variable 2 6/1/95 1518
Variable 3 6/1/95 1518
Variable 4 6/1/95 1518
Variable 5 6/1/95 1518
Proyecto: ibex11_inc2.nnp
Último patrón disponible: 31/10/00
Predicción para el día: 2/11/00
Patrones de entrenamiento: 498
Último entrenamiento realizado: 20/10/00 con datos desde 29/6/98 hasta 1/6/00
Ciclos de entrenamiento acumulados: 20
Entrenamiento
Período
Intervalo móvil
Desde
08/07/98
Hasta
12/06/00
Número de ciclos
20
Guardar
Entrenar
Mejorar
Parar
Deshacer
Editar Entrenar Resultados Informe
d:\aern\sprinn institutional 2.0-esp\proyectos\ibex11_inc2.nnp


SpriNN - [Ibex11_inc2.nnp]
Archivo Datos Proyecto Ventana Ayuda
Resultados
Predecir tendencia
Margen del 10.00%
Eje derecho
Puntos
Grid
Margen proporcional
Todos los datos
Tabla de resultados
Puntos por página 610
Ninguna variable
Página 1 de 2
4/9/98 13/11/98 22/1/99 2/4/99 11/6/99 20/8/99 29/10/99 7/1/00 17/3/00 26/5/00 4/8/00 13/10/00
13000 12500 12000 11500 11000 10500 10000 9500 9000 8500 8000 7500 7000
Sube Indeciso Baja Objetivo
23/10/00 24/10/00 25/10/00 26/10/00 27/10/00 30/10/00 31/10/00 1/11/00 2/11/00
Predicción 10374.36 10448.70 10331.72 10667.92 10516.34 10324.09 10396.12 10310.39 10367.9
Objetivo 10329.90 10667.00 10518.00 10325.00 10396.00 10308.00 10364.00
Exportar
Editar Entrenar Resultados Informe
Gráfico de objetivo, variables y predicciones
d:\aern\sprinn institutional 2.0-esp\proyectos\ibex11_inc2.nnp



Análisis de las recomendaciones de C/V

Diario de operaciones

C = Abrir Alza, V = Cerrar Alza, W = Abrir Baja, Z = Cerrar Baja

Tipo	Operación	Días	Beneficio
	Capital inicial el 1/6/00		
Baja	2/6/00 W a 11142.00 y 7/6/00 Z a 10688.90	6	40.67
Alza	13/6/00 C a 10806.00 y 14/6/00 V a 10881.00	2	7.22
Alza	16/6/00 C a 10724.00 y 5/7/00 V a 10758.00	20	3.32
Baja	10/7/00 W a 10829.00 y 11/7/00 Z a 10875.00	2	-4.47
Alza	13/7/00 C a 10849.00 y 17/7/00 V a 10950.00	5	9.74
Alza	18/7/00 C a 10792.00 y 21/7/00 V a 10880.00	4	8.61
Alza	25/7/00 C a 10715.00 y 8/8/00 V a 10868.00	15	15.21
Baja	10/8/00 W a 11111.00 y 17/8/00 Z a 11156.00	8	-4.38
Alza	21/8/00 C a 11094.00 y 22/8/00 V a 10935.00	2	-15.42
Alza	23/8/00 C a 10799.00 y 4/9/00 V a 11247.00	13	44.00
Baja	15/9/00 W a 11464.00 y 20/9/00 Z a 10902.00	6	54.15
Alza	21/9/00 C a 10804.00 y 25/9/00 V a 11160.00	5	38.18
Baja	27/9/00 W a 10958.00 y 2/10/00 Z a 11153.00	6	-21.30
Baja	4/10/00 W a 11160.00 y 10/10/00 Z a 10951.00	7	22.02
Alza	12/10/00 C a 10619.00 y 24/10/00 V a 10667.00	13	5.41
Alza	30/10/00 C a 10308.00 abierta		
	Ultima cotización el 31/10/00 a 10364.00	**2**	**6.54**

Resultados reales (1 año)

Estrategia	5 sesiones	20 sesiones	40 sesiones	60 sesiones
Agresiva	***63%(219)***	***62%(229)***	***58%(228)***	***48%(236)***
Media	***63%(145)***	***66%(61)***	***74%(131)***	***56%(143)***
Conservadora	***60%(72)***		***89%(80)***	***100%(44)***