



MINERÍA DE DATOS

TÉCNICAS Y HERRAMIENTAS

CÉSAR PÉREZ LÓPEZ

Universidad Complutense de Madrid
Instituto de Estudios Fiscales

Autor colaborador:

DANIEL SANTÍN GONZÁLEZ

Universidad Complutense de Madrid

THOMSON



ÍNDICE

<i>Introducción</i>	<i>XVII</i>
<i>Capítulo 1. Minería de datos: Conceptos, técnicas y sistemas</i>	<i>1</i>
Aproximación al concepto de minería de datos	1
El proceso de extracción del conocimiento	3
Técnicas de minería de datos.....	8
Sistemas de minería de datos	10
<i>Capítulo 2. Entorno de trabajo de SPSS Clementine.....</i>	<i>13</i>
Introducción a Clementine	13
Usando el ratón.....	16
Ayuda en Clementine	16
Panel de control en Clementine.....	18
Ejemplo de trabajo con Clementine	21
Insertar un nodo fuente (origen) de datos en el área de trabajo.....	22
Enlazar un nodo con una fuente de datos	23
Controlar la carga de datos con el <i>nodo Tabla</i>	25
Definir variables predictoras con el <i>nodo Tipo</i>	27
Utilizar un nodo de modelado	29
Ejecutar una ruta.....	29
Interpretar un modelo	32
Predecir con un modelo	34
Guardar un modelo	34
Nodos de orígenes de datos	35
Nodos de operaciones con registros	35

Nodos de operaciones con campos.....	36
Nodos para gráficos	37
Nodos para modelado	38
Nodos de salida	40
Capítulo 3. Entorno de trabajo de SAS Enterprise Miner.....	41
Introducción a SAS Enterprise Miner	41
Comenzando con SAS Enterprise Miner.....	43
Inicio de un proyecto nuevo	47
Menú principal de SAS Enterprise Miner	48
Ejemplo de trabajo con SAS Enterprise Miner	58
Leer ficheros y enlazarlos con Enterprise Miner mediante el <i>nodo</i> <i>Input Data Source</i>	58
Definir tipos de variables con el <i>nodo Input Data Source</i>	63
Enlace de nodos de un diagrama. El <i>nodo Data Partition</i>	65
Utilizar un nodo de modelado	67
Capítulo 4. Fase de selección en minería de datos.....	73
Selección en el proceso de extracción del conocimiento	73
Recopilación e integración de datos: <i>Data Warehouse</i>	74
<i>Data Warehouse</i> y <i>Data Mining</i>	77
Selección de datos mediante muestreo	78
Muestreo aleatorio simple	82
Muestreo estratificado	85
Muestreo sistemático	91
Muestreo unietápico de conglomerados	95
Muestreo bietápico de conglomerados	99
Muestreo polietápico de conglomerados	101
Diseños complejos: Bietápico con estratificación en primera etapa	101
Selección de números aleatorios: Método de Montecarlo.....	102
Selección de características relevantes	104
Análisis de correlaciones.....	105
Capítulo 5. Fase de selección en SAS Enterprise Miner y SPSS Clementine.....	109
La fase de selección en Enterprise Miner	109
El <i>nodo Fuente de Datos</i>	109
El <i>nodo Muestreo</i>	117
El <i>nodo de Partición de Datos</i>	122
El <i>nodo de Selección de Variables</i>	125
El <i>nodo de Series Temporales</i>	129

La fase de selección en SPSS Clementine.....	139
Importación de datos ASCII.....	140
Importación de datos de una fuente ODBC (Access, Excel, etc.)	140
Importación de datos de SPSS.....	143
Importación de datos de SAS	145
Selección de datos	148
Muestreo de datos	149
Capítulo 6. Fase de selección en SPSS Muestras Complejas y SAS Base.....	151
Técnicas de muestreo a través de SPSS	151
Diseños complejos y el asistente de muestreo. Creación de un nuevo plan de muestreo.....	152
Asistente de muestreo: modificar un plan existente	161
Asistente de muestreo: ejecutar un plan de muestreo dado	164
Preparación de una muestra compleja para su análisis: Creación de un nuevo plan de análisis	164
Preparación de una muestra compleja para su análisis.....	168
Cálculos en muestras complejas: Frecuencias, descriptivos, tablas de contingencia y razones	168
Selección de casos en SPSS	174
Selección de casos mediante criterios condicionales	174
Selección de fechas, horas y filas	175
Selección de una muestra aleatoria.....	175
Semilla de aleatorización.....	176
Operadores para la selección en SPSS	176
Operadores aritméticos	176
Operadores relacionales	177
Operadores lógicos.....	177
Funciones de generación de números aleatorios en SPSS.....	177
Selección de la información en SAS Base	180
Declarando valores perdidos con la sentencia MISSING	180
Seleccionando información por grupos: sentencia BY	180
Seleccionando variables de frecuencias: sentencia FREQ	182
Seleccionando variables de pesos: sentencia WEIGHT.....	183
Seleccionando variables de identificación: Sentencia ID.....	184
Operadores para la selección en SAS.....	184
Operadores aritméticos.....	185
Operadores de comparación	185
Operadores lógicos o booleanos.....	186
Operadores MIN, MAX y concatenación.....	187
Orden de evaluación de los operadores en las expresiones.....	188
Funciones de generación de números aleatorios en SAS	189
Cálculos con funciones en SAS.....	191

Capítulo 7. Fase de exploración en minería de datos	193
Exploración en el proceso de extracción del conocimiento	193
Análisis exploratorio	194
Herramientas de exploración visual	194
Histograma de frecuencias	195
Diagrama de tallo y hojas	196
Gráfico de caja y bigotes	198
Gráfico múltiple de caja y bigotes	199
Gráfico de simetría	201
Gráfico de dispersión	203
Gráficos para variables cualitativas	205
Herramientas de exploración formal	207
Contrastes de la bondad de ajuste a una distribución: Test de la Chi-cuadrado	208
Contraste de Kolmogorov-Smirnov Lilliefors de la bondad de ajuste a una distribución	209
Estadísticos robustos de centralización	211
Estadísticos robustos de dispersión	212
Estadísticos robustos de asimetría y curtosis	214
Contrastes de aleatoriedad	216
Transformaciones de las variables	220
Supuestos subyacentes en las técnicas de minería de datos	221
Normalidad	221
Heteroscedasticidad	225
Multicolinealidad	227
Autocorrelación	227
Linealidad	228
Un ejemplo	230
Capítulo 8. Fase de exploración en SAS Enterprise Miner y SPSS Clementine	239
La fase de exploración en Enterprise Miner	239
El nodo Explorador de distribuciones	239
El nodo Multigráficos	243
El nodo de exploración de patrones	250
La fase de exploración en SPSS Clementine	266
El nodo Gráfico	267
El nodo Distribución	270
El nodo Histograma	271
El nodo Malla	273
El nodo Malla Direccional	274
El nodo Gráfico Múltiple	275
El nodo Recolectar	276

Capítulo 9. Fase de exploración en SPSS y SAS	277
Análisis exploratorio de datos con SPSS. <i>Procedimiento Explorar</i>	277
Gráficos de análisis exploratorio con SPSS	282
Tipos de gráficos	282
Histogramas	283
Gráficos de normalidad	283
Gráficos de caja y bigotes	286
Gráficos de dispersión	288
Gráficos interactivos dinámicos de análisis exploratorio con SPSS	290
Creación interactiva de gráficos a partir de tablas	297
Gráficos interactivos de caja y bigotes	298
Histogramas interactivos	299
Diagramas interactivos de dispersión	301
Análisis exploratorio formal con SPSS	303
Contraste de aleatoriedad. <i>Procedimiento Prueba de rachas</i>	303
Contraste de ajuste a una distribución de frecuencias. <i>Procedimiento Prueba de Kolmogorov-Smirnov</i>	304
Análisis exploratorio de los datos con SAS Base. <i>Procedimiento Univariate</i>	305
Gráficos de análisis exploratorio con SAS	318
Gráficos exploratorios de alta resolución. <i>Procedimiento GCHART</i>	318
Gráficos exploratorios de mapas: <i>Procedimiento GMAP</i>	322
Gráficos exploratorios de caja y bigotes: <i>Procedimiento BOXPLOT</i>	328
Capítulo 10. Fases de limpieza y transformación de datos	333
Limpieza y transformación de datos en el proceso de extracción del conocimiento	337
Valores atípicos (<i>Outliers</i>)	337
Información faltante (Datos <i>missing</i>)	337
Soluciones para los datos ausentes: Supresión de datos e imputación de información faltante	343
Transformación de datos	346
Transponer, fusionar, agregar, segmentar y ordenar archivos	346
Ponderar casos y categorizar y numerizar variables	347
Pareamiento o <i>matching</i>	348
Transformación de datos mediante técnicas de reducción de la dimensión	349
Componentes principales	350
Análisis factorial	357

Capítulo 11. Las fases de limpieza y transformación de datos en SAS Enterprise Miner y SPSS Clementine..... 365

Las fases de limpieza y transformación de datos en Enterprise Miner	365
El nodo Transformación de variables	365
El nodo Asignación de atributos	371
Tratamiento de los datos atípicos con el nodo Filtro de Outliers	378
El nodo Imputación de datos missing.....	384
El nodo Exploración de patrones para Componentes Principales	393
Lasa fases de limpieza y transformación de datos en Clementine	400
El nodo Seleccionar.....	402
El nodo Muestra para procesos de muestreo.....	404
El nodo Combinar para procesos de matching.....	405
El nodo Equilibrar.....	407
El nodo Ordenar.....	408
El nodo Agregar para calcular estadísticos por subgrupos	409
El nodo Distinguir	411
El nodo Añadir para concatenación de archivos	411
El nodo Filtrar.....	412
El nodo Derivar para transformación de variables.....	413
El nodo Tipo para asignar atributos a variables.....	415
El nodo Rellenar para imputación de datos missing	416
El nodo Factor/PCA para Análisis Factorial y Componentes Principales ...	417

Capítulo 12. Fases de limpieza y transformación de datos en SPSS y SAS .. 427

Técnicas de reducción de la dimensión en SPSS Base.....	427
Componentes principales con SPSS.....	428
Análisis factorial con SPSS	439
Transformación de datos en SPSS Base.....	447
Transformación de valores de datos	447
Remodificación de variables	449
Ordenar casos	451
Transponer, fusionar, agregar y segmentar archivos. Matching	451
Ponderar casos	458
Categorizar variables: Categorizador visual.....	459
Asignar rangos a casos y tipificar variables	462
SPSS y el análisis de datos missing. Imputación.....	463
Reemplazar valores perdidos.....	469
Detección de valores atípicos en SPSS	470
Detección de casos atípicos mediante gráficos de control	470
Detección de casos atípicos mediante gráficos de caja y bigotes.....	472
Técnicas de reducción de la dimensión en SAS STAT	475

Componentes principales en SAS. Procedimiento PRINCOMP y Procedimiento FACTOR	475
Análisis factorial en SAS. Procedimiento FACTOR	482
Transformación de datos en SAS Base	487
Operaciones con ficheros: Concatenación y Matching	487
Actualizando ficheros de datos SAS	489
Añadir información. Procedimiento APPEND.....	491
Tipificación de datos: Procedimiento STANDARD.....	494

Capítulo 13. Fase de minería de datos. Técnicas predictivas de modelización ... 497

Técnicas de minería de datos propiamente dichas	497
Técnicas predictivas para la modelización	498
Modelo de regresión múltiple.....	504
Estimación del modelo lineal de regresión múltiple	505
Estimación del modelo, contrastes e intervalos de confianza a través del cálculo matricial	506
Análisis de la varianza en el modelo de regresión múltiple	507
Predicciones.....	510
Análisis de los residuos	511
Técnicas de selección en el modelo de regresión.....	512
Modelos de elección discreta	513
Modelos de elección discreta binaria: Modelo lineal de probabilidad y regresión logística binaria	514
Modelos de elección múltiple: Modelo Logit Multinomial	519
Modelo lineal general de regresión múltiple (GLM)	521
Clasificación ad hoc: Análisis discriminante	521
Hipótesis en el modelo discriminante.....	522
Estimación del modelo discriminante	523
Clasificación mediante el modelo discriminante.....	525

Capítulo 14. Técnicas predictivas de modelización con SAS Enterprise Miner y SPSS Clementine..... 529

Técnicas predictivas de modelización con SAS Enterprise Miner.....	529
El nodo Regression: Modelo de regresión múltiple	530
El nodo Regression: Modelo lineal general GLM.....	538
El nodo Regression: Modelo de elección discreta Logit y Probit	551
Técnicas predictivas de modelización con SPSS Clementine.....	554
El nodo Regresión Lineal: Modelo de regresión múltiple.....	555
El nodo Regresión Logística: Modelos de elección discreta	561

Capítulo 15. Técnicas predictivas de modelización con SAS y SPSS	565
El modelo lineal general con SAS. <i>Procedimiento GLM</i>	565
Modelos del análisis de la varianza y la covarianza con SAS	571
Modelo de elección discreta en SAS	574
Modelo Logit: <i>Procedimiento LOGISTIC</i>	574
Modelo Probit: <i>Procedimiento PROBIT</i>	579
SAS y el análisis discriminante: <i>Procedimiento DISCRIM</i>	581
El modelo lineal general con SPSS. <i>Procedimiento MLG Multivariante</i>	585
Modelo de elección discreta en SPSS	593
Modelo Logit: <i>Procedimiento LOGISTICA MULTINOMIAL</i>	593
Modelo Probit: <i>Procedimiento PROBIT</i>	599
SPSS y el análisis discriminante	601
Capítulo 16. Técnicas descriptivas y predictivas de clasificación. Clusters y árboles de decisión	609
El análisis cluster como técnica descriptiva de clasificación	609
Medidas de similitud	610
Técnicas en el análisis <i>cluster</i>	614
Clusters jerárquicos, secuenciales, aglomerativos y exclusivos (S.A.H.N.)	616
El dendograma en el análisis cluster jerárquico	617
Análisis cluster no jerárquico	617
Los árboles de decisión como técnica predictiva de clasificación	621
Características de los árboles de decisión	622
Herramientas para el trabajo con árboles de decisión	626
Árboles CHAID	627
Árboles CART	628
Árboles QUEST	630
Análisis de conglomerados y árboles de decisión como métodos de segmentación	631
Capítulo 17. Clusters y árboles de decisión con SAS Enterprise Miner y SPSS Clementine	633
Análisis <i>cluster</i> con Enterprise Miner. El <i>nodo Clustering</i>	633
Árboles de decisión con Enterprise Miner. El <i>nodo Tree</i>	641
Entrenamiento interactivo (<i>Interactive Training</i>)	652
Análisis <i>cluster</i> con SPSS Clementine	656
El <i>nodo Entrenar K-medias: Cluster</i> no jerárquico	656
El <i>nodo Cluster Bietápico: Cluster</i> jerárquico	661
Árboles de decisión con SPSS Clementine	662
El <i>nodo Crear C5.0</i>	662
El <i>nodo Árbol C&R</i>	664

Capítulo 18. Clusters y árboles de decisión con SAS y SPSS	665
SPSS y el análisis <i>cluster</i> jerárquico	665
SPSS y el análisis <i>cluster</i> no jerárquico	671
SAS y el análisis <i>cluster</i> jerárquico	675
<i>Procedimiento ACECLUS</i>	675
<i>Procedimiento CLUSTER</i>	677
<i>Procedimiento TREE</i>	678
SAS y el análisis cluster no jerárquico	681
Árboles de decisión (o clasificación) con SPSS	687
Creación de un árbol de decisión: Método CHAID	689
Métodos CRT y QUEST. Poda de árboles	695
Capítulo 19. Redes neuronales	699
Descripción de una red neuronal	699
Definición	699
Función de salida y funciones de transferencia o activación	701
Redes neuronales y ajuste de modelos de regresión	703
Aprendizaje en las redes neuronales	704
Funcionamiento de una red neuronal	707
El algoritmo de aprendizaje Retropropagación (<i>Back-Propagation</i>)	708
Análisis discriminante a través del Perceptrón	709
Análisis de series temporales mediante redes neuronales	713
Análisis de componentes principales con redes neuronales	715
<i>Clustering</i> mediante redes neuronales	717
Capítulo 20. Redes neuronales con SAS Enterprise Miner y SPSS Clementine	721
Redes neuronales con SAS Enterprise Miner	721
Optimización y ajuste de modelos con redes: <i>Nodo Neural Network</i>	722
Análisis en componentes principales a través de redes neuronales: <i>Nodo Princomp/Dmneural</i>	745
Predicción y análisis discriminante a través de redes neuronales: <i>Nodo Two Stage Model</i>	751
Análisis <i>cluster</i> con redes neuronales: <i>Nodo SOM/Kohonen</i>	756
Redes neuronales con SPSS Clementine	765
<i>Nodo Entrenar red</i>	765
<i>Nodo Entrenar Kohonen</i>	769
<i>Nodo Entrenar K-medias</i>	771
Índice alfabético	775

INTRODUCCIÓN

Este libro presenta las técnicas más habituales utilizadas en minería de datos de una forma sencilla y fácil de entender a través de las soluciones de software más comunes de entre las existentes en el mercado. Se persigue como finalidad inicial clarificar las aplicaciones relativas a métodos tradicionalmente calificados como difíciles u opacos. Se busca presentar las aplicaciones en la minería de datos sin necesidad de manejar desarrollos matemáticos elevados ni algoritmos teóricos complicados, que es la razón más común de las dificultades en la comprensión y aplicación de esta materia.

Hoy en día se utiliza la minería de datos en diferentes campos de la ciencia. Cabe destacar las aplicaciones financieras y en banca, en análisis de mercados y comercio, en seguros y salud privada, en educación, en procesos industriales, en medicina, en biología y bioingeniería, en telecomunicaciones y en muchas otras áreas. Lo esencial para empezar a trabajar en minería de datos, sea cual sea el campo en que se aplique, es la comprensión de los propios conceptos, tarea que no exige ni mucho menos el dominio de aparato científico que conlleva la materia. Posteriormente, cuando ya sea necesaria la operatoria avanzada, los programas de ordenador permiten obtener los resultados sin necesidad de descifrar el desarrollo matemático de los algoritmos que están debajo de los procedimientos.

En este libro se describen los conceptos de minería de datos de la forma más sencilla posible, de modo que sean inteligibles por lectores con formación diversa. Los capítulos comienzan describiendo las técnicas en lenguaje asequible y presentando a continuación la forma de tratarlas mediante aplicaciones prácticas. Una parte importante de cada capítulo son casos prácticos totalmente resueltos, incluyendo la interpretación de los resultados, que precisamente es lo más importante en cualquier materia con la que se trabaje.

MINERÍA DE DATOS: CONCEPTOS, TÉCNICAS Y SISTEMAS

APROXIMACIÓN AL CONCEPTO DE MINERÍA DE DATOS

La minería de datos puede definirse inicialmente como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos.

La disponibilidad de grandes volúmenes de información y el uso generalizado de herramientas informáticas ha transformado el análisis de datos orientándolo hacia determinadas técnicas especializadas englobadas bajo el nombre de minería de datos *Data Mining*.

Las técnicas de minería de datos persiguen el descubrimiento automático de conocimiento contenido en la información almacenada de modo ordenado en grandes bases de datos. Estas técnicas tienen como objetivo descubrir patrones, perfiles y tendencias a través del análisis de los datos utilizando tecnologías de reconocimiento de patrones, redes neuronales, lógica difusa, algoritmos genéticos y otras técnicas avanzadas de análisis de datos.

No obstante, la minería de datos es ya un concepto muy evolucionado que necesita ser aproximado conceptualmente por etapas. Inicialmente la finalidad de los sistemas de información era recopilar información sobre una parcela determinada para ayudar en la toma de decisiones. Con la informatización de las organizaciones y la aparición de aplicaciones software operacionales sobre el sistema de información, la finalidad principal de los sistemas de información es dar soporte a los procesos básicos de la organización (ventas, producción, personal...). Una vez satisfecha la necesidad de tener un soporte informático para los procesos básicos de la organización (*sistemas de información para la gestión*), las organizaciones exigen nuevas prestaciones de los sistemas de información (*sistemas de información para la toma de decisiones*).

De esta forma han aparecido diferentes herramientas de negocio para la toma de decisiones (DSS o *Decision Support Systems*) que *coexisten*: EIS, OLAP, consultas e informes, y las propias herramientas de minería de datos.

Un EIS (*Executive Information System*) es un sistema de información y un conjunto de herramientas asociadas que proporciona a los directivos acceso a la información de estado y sus actividades de gestión. Está especializado en analizar el estado diario de la organización (mediante indicadores clave) para informar rápidamente sobre cambios a los directivos. La información solicitada suele ser, en gran medida, numérica (ventas semanales, nivel de stocks, balances parciales, etc.) y representada de forma gráfica al estilo de las hojas de cálculo.

Las herramientas OLAP (*On-Line Analytical Processing*) son más genéricas, funcionan sobre un sistema de información (transaccional o almacén de datos) y permiten realizar agregaciones y combinaciones de los datos de maneras mucho más complejas y ambiciosas, con objetivos de análisis más estratégicos. Las herramientas OLAP están basadas, generalmente, en sistemas o interfaces multidimensionales, que presentan la información de una manera matricial. Las herramientas OLAP proporcionan facilidades para “manejar” y “transformar” los datos, producen otros “datos” (más agregados, combinados) y son una gran ayuda para analizar los datos porque producen diferentes vistas de los mismos.

Los *sistemas de informes* o *consultas avanzadas* están basados, generalmente, en sistemas relacionales u objeto-relacionales y el resultado se presenta de forma tabular. Generalmente están implementados en bases de datos relacionales.

Las *herramientas de minería de datos* permiten extraer patrones, tendencias y regularidades para describir y comprender mejor los datos y para predecir comportamientos futuros. La Minería de Datos analiza los datos y el resto de herramientas citadas anteriormente facilitan el acceso a la información para que el análisis sea más efectivo, es decir, son instrumentos de apoyo a la minería de datos.

No obstante las herramientas anteriormente citadas suelen necesitar de la existencia previa de un *almacén de datos* (*Data Warehouse*). El almacén de datos es el *sistema de información central* en todo este proceso. Un almacén de datos es una colección de datos orientada a un dominio, integrada, no volátil y variante en el tiempo para ayudar en la toma de decisiones. Un almacén de datos es un conjunto de datos históricos, internos o externos y descriptivos de un contexto o área de estudio, que están integrados y organizados de tal forma que permiten aplicar eficientemente herramientas para resumir, describir y analizar los datos con el fin de ayudar en la toma de decisiones estratégicas.

Las fuentes internas y externas de datos están separadas. Gran parte de los datos que se incorporan en un almacén de datos provienen de una base de datos transaccional que es el origen de datos interno y cuya información es fruto de las transacciones derivadas de la actividad diaria, pero también existen otras fuentes externas de información.

Existe un sistema especializado para realizar la carga y mantenimiento de un almacén de datos, denominado *sistema ETL* (*Extraction, Transformation, Load*). Este sistema se encarga de la lectura de datos transaccionales, de la incorporación de datos externos, creación de claves, integración de datos, agregaciones, limpieza y transformación de datos, creación y mantenimiento de metadatos, planificación de carga y mantenimiento, indización, pruebas de calidad, etc.

La Figura 1-1, cuya fuente es Orallo, Quintana y Ramírez (*Introducción a la Minería de datos*) ordena los conceptos expuestos en los párrafos anteriores.

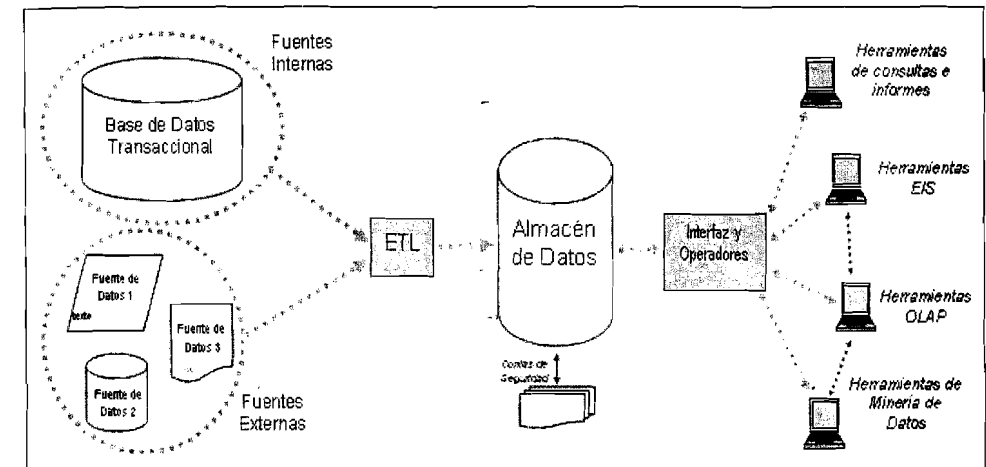


Figura 1-1

No obstante hay que tener claro que los almacenes de datos no son imprescindibles para hacer extracción de conocimiento a partir de los datos. Se puede hacer minería de datos sobre un simple fichero de datos. Pero las ventajas de organizar un almacén de datos para realizar minería de datos se amortizan sobradamente a medio y largo plazo cuando tenemos grandes volúmenes de datos, o éstos aumentan con el tiempo, o provienen de fuentes heterogéneas o se van a combinar de maneras arbitrarias y no predefinidas.

EL PROCESO DE EXTRACCIÓN DEL CONOCIMIENTO

Pero la minería de datos es sólo una etapa del *proceso de extracción de conocimiento a partir de datos* (KDD). Este proceso consta de varias fases como la preparación de datos (selección, limpieza, y transformación), su exploración y auditoría, minería de datos propiamente dicha (desarrollo de modelos y análisis de datos), evaluación, difusión y utilización de modelos (*output*). Además, el proceso de extracción del conocimiento incorpora muy diferentes técnicas (árboles de decisión, regresión lineal, redes neuronales artificiales, técnicas bayesianas, máquinas de soporte vectorial, etc.) de campos diversos (aprendizaje automático e inteligencia artificial), estadística, bases de datos, etc.) y aborda una tipología variada de problemas (clasificación, categorización, estimación/regresión, agrupamiento, etc.). La Figura 1-2 muestra las etapas del KDD.

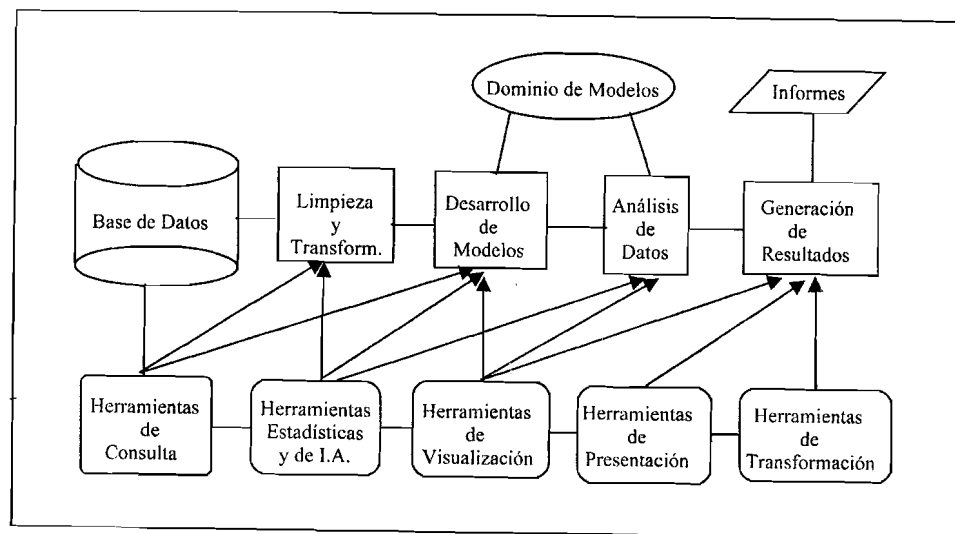


Figura 1-2

El KDD comienza con la *recopilación e integración de la información* a partir de unos datos iniciales de que se dispone (fase de *selección de datos*). Las primeras fases del KDD determinan que las fases sucesivas sean capaces de extraer conocimiento válido y útil a partir de la información original. Generalmente, la información que se quiere investigar sobre un cierto dominio de la organización se encuentra en bases de datos (*Database*) y otras fuentes muy diversas, tanto internas como externas (en general la información se encuentra ordenada en almacenes de datos). Muchas de estas fuentes son las que se utilizan para el trabajo transaccional. El análisis posterior será mucho más sencillo si la fuente es unificada, accesible (interna) y desconectada del trabajo transaccional. Aparte de información interna de la organización, los almacenes de datos pueden recoger información externa, como demografías (censo), páginas amarillas, psicografías (perfiles por zonas), uso de Internet, información de otras organizaciones y bases de datos externas compradas a otras compañías. La disponibilidad de grandes volúmenes de información en esta fase nos lleva a la necesidad de usar técnicas de muestreo para la selección de datos.

La fase siguiente del KDD integra la exploración, la limpieza o criba de datos (*Data Cleaning*) y la transformación de datos. Se deben eliminar el mayor número posible de datos erróneos o inconsistentes (limpieza) e irrelevantes (criba). En esta fase se utilizan herramientas de consulta (*Query tools*) y herramientas estadísticas (*Statistics tools*) casi exclusivamente. En la exploración se usan técnicas de análisis exploratorio de datos como los histogramas y los diagramas de caja, tallo y hojas, que ayudan a detectar datos anómalos o atípicos (*outliers*). La presencia de datos atípicos y valores desaparecidos (datos *missing*) puede llevarnos a usar algoritmos robustos a datos atípicos y desaparecidos (p.ej. árboles de decisión), a filtrar la información, a reemplazar valores mediante *técnicas de imputación* y a transformar datos continuos en discretos mediante *técnicas de discretización*. Entre las técnicas avanzadas de transformación tenemos las de reducción y aumento de la dimensión.

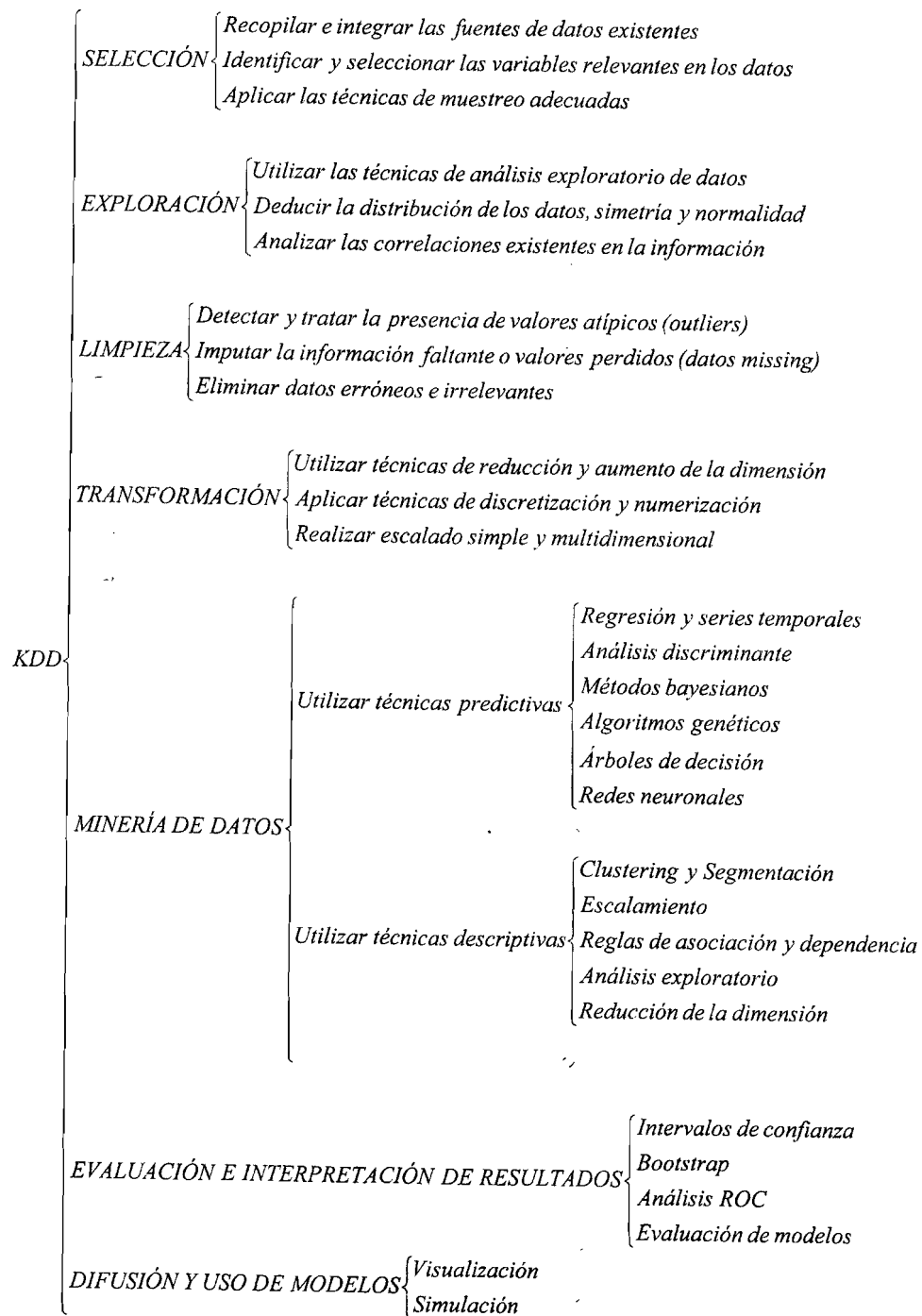
La fase siguiente en el KDD es la propia minería de datos que se llevará a cabo a partir del desarrollo de modelos predictivos y descriptivos (*Model Development*) y mediante el análisis de datos (*Data Analysis*). Una vez recogidos los datos de interés, un explorador puede decidir qué tipo de patrón quiere descubrir. El tipo de conocimiento que se desea extraer va a marcar claramente la técnica de minería de datos a utilizar.

Para seleccionar y validar los modelos anteriores es necesaria una nueva fase consistente en el uso de criterios de evaluación de hipótesis. El despliegue del modelo a veces es trivial pero otras veces requiere un proceso de implementación o interpretación. En esta fase se utilizan adicionalmente herramientas estadísticas y de visualización (*Visualization tools*).

Una fase posterior del KDD es la relativa a la difusión y uso del conocimiento derivado de las técnicas de minería de datos a través de los modelos correspondientes que habitualmente desembocan en la generación de resultados (*Output Generation*). El modelo puede tener muchos usuarios y necesitar difusión, con lo que puede requerir ser expresado de una manera comprensible para ser distribuido en la organización. En esta fase se utilizan herramientas de visualización (*Visualization tools*), presentación (*Presentation tools*) y transformación de datos (*Data transformation tools*).

Por lo tanto, observamos en el proceso de extracción del conocimiento KDD la secuencia de fases siguiente: SELECCIÓN → EXPLORACIÓN → LIMPIEZA → TRANSFORMACIÓN → MINERÍA DE DATOS → EVALUACIÓN → DIFUSIÓN

En la fase de *selección* se integran y recopilan los datos, se determinan las fuentes de información que pueden ser útiles y dónde conseguirlas, se identifican y seleccionan las variables relevantes en los datos y se aplican las técnicas de muestreo adecuadas. Todo ello se facilita disponiendo de un almacén de datos con la información en formato común y sin inconsistencias. Dado que los datos provienen de diferentes fuentes, es necesaria su *exploración* mediante técnicas de análisis exploratorio de datos, buscando entre otras cosas la distribución de los datos, su simetría y normalidad y las correlaciones existentes en la información. A continuación es necesaria la *limpieza* de los datos, ya que pueden contener valores atípicos, valores faltantes y valores erróneos. En esta fase se analiza la influencia de los datos atípicos, se imputan los valores faltantes y se eliminan o corrigen los datos incorrectos. A continuación, si es necesario, se lleva a cabo la *transformación* de los datos generalmente mediante técnicas de reducción o aumento de la dimensión y escalado simple y multidimensional, entre otras. Las cuatro primeras fases se suelen englobar bajo el nombre de *preparación de datos*. En la fase de *minería de datos*, se decide cuál es la tarea a realizar (clasificar, agrupar, etc.) y se elige la técnica descriptiva o predictiva que se va a utilizar. En la fase de *evaluación e interpretación* se evalúan los patrones y se analizan por los expertos, y si es necesario se vuelve a las fases anteriores para una nueva iteración. Finalmente, en la fase de *difusión* se hace uso del nuevo conocimiento y se hace partícipe de él a todos los posibles usuarios. Entonces, la clasificación de las fases del proceso de extracción del conocimiento podría resumirse en el siguiente esquema:



No obstante, la clasificación anterior no es la única que aparece en la literatura de esta materia. Existen otras interpretaciones del concepto de minería de datos, en la línea de considerar las fases del proceso de extracción del conocimiento expresadas previamente como técnicas de minería de datos. Por ejemplo, SAS Institute define el concepto de *Data Mining* como el proceso de Seleccionar (*Selecting*), Explorar (*Exploring*), Modificar (*Modifying*), Modelizar (*Modeling*) y Valorar (*Assessment*) grandes cantidades de datos con el objetivo de descubrir patrones desconocidos que puedan ser utilizados como ventaja comparativa respecto a los competidores. Este proceso es resumido con las siglas SEMMA. La Figura 1-3 ilustra las fases del proceso de minería de datos según SAS Institute.

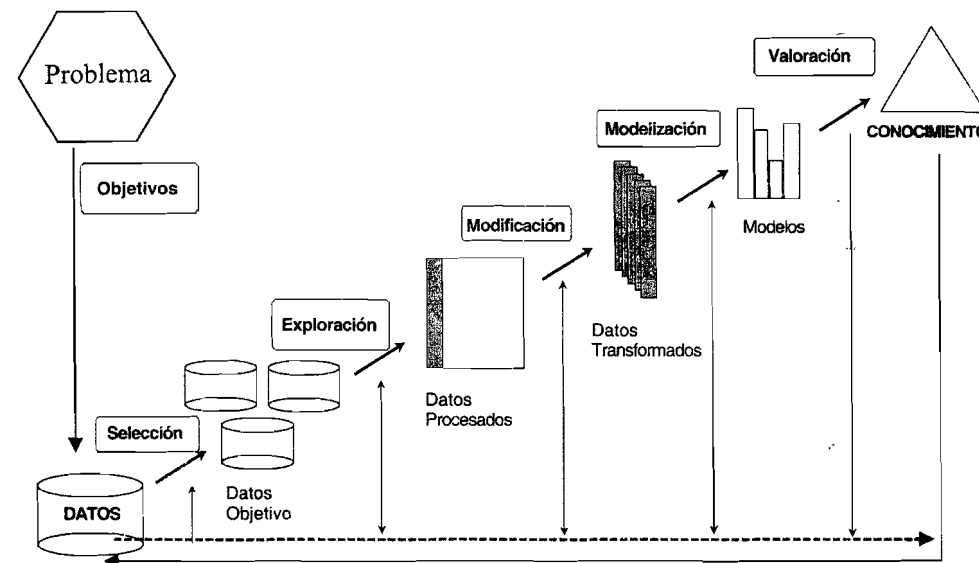


Figura 1-3

Se observa la equivalencia entre las componentes del concepto de minería de datos de SAS Institute y las fases del KDD expuestas anteriormente. Las fases de Limpieza y Transformación en KDD equivalen a la fase de Modificación en SAS, y la fase de Minería de Datos equivale a Modelización. Evaluación y Valoración pueden considerarse sinónimos. SAS Institute implementa la minería de datos en el software *Enterprise Miner*, que será utilizado en este libro, y en otros procedimientos y módulos (STAT, ETS,...).

Por su parte SPSS considera que las seis fases que forman el proceso de la minería de datos son: la comprensión del negocio, la comprensión de los datos, la preparación de los datos, el modelado, la evaluación y el uso del modelo. SPSS implementa esta filosofía de la minería de datos en el software *Clementine*, que será utilizado en este libro, y en otros procedimientos y módulos (*Answer Tree*, *Neural Connection*,...).

TÉCNICAS DE MINERÍA DE DATOS

La clasificación inicial de las técnicas de minería de datos distingue entre técnicas predictivas, en las que las variables pueden clasificarse inicialmente en dependientes e independientes (similares a las técnicas del análisis de la dependencia o métodos explicativos del análisis multivariante), técnicas descriptivas, en las que todas las variables tienen inicialmente el mismo estatus (similares a las técnicas del análisis de la interdependencia o métodos descriptivos del análisis multivariante) y técnicas auxiliares.

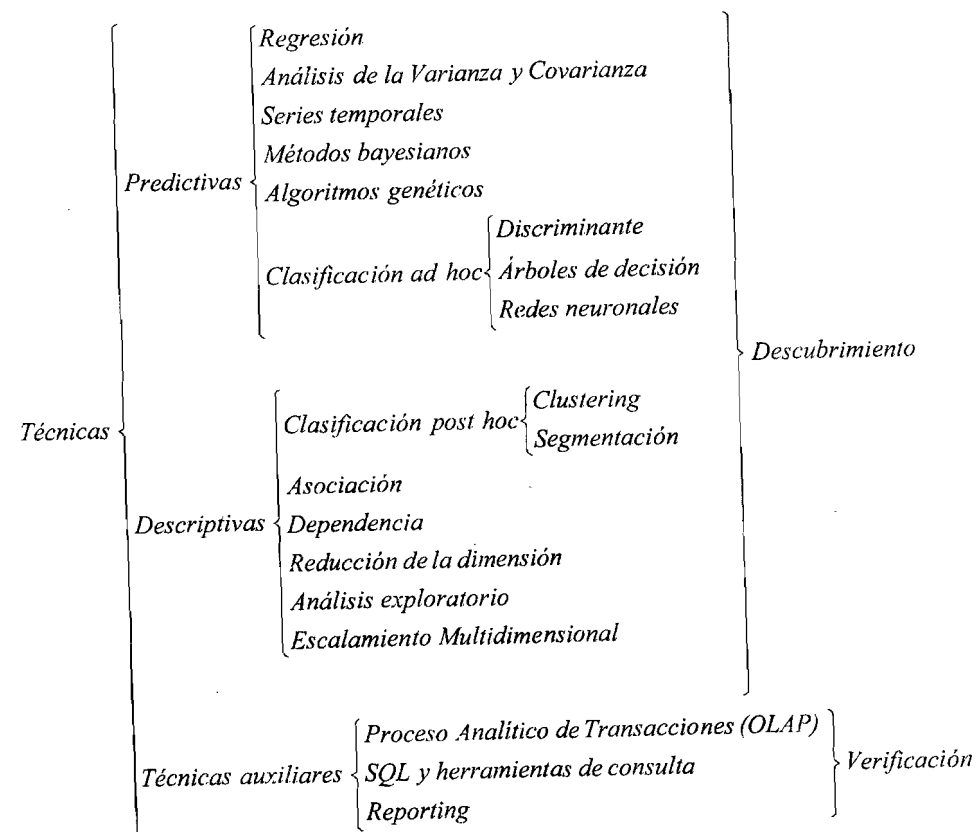
Las *técnicas predictivas* especifican el modelo para los datos en base a un conocimiento teórico previo. El modelo supuesto para los datos debe contrastarse después del proceso de minería de datos antes de aceptarlo como válido. Formalmente, la aplicación de todo modelo debe superar las fases de *identificación objetiva* (a partir de los datos se aplican reglas que permitan identificar el mejor modelo posible que ajuste los datos), *estimación* (proceso de cálculo de los parámetros del modelo elegido para los datos en la fase de identificación), *diagnosis* (proceso de contraste de la validez del modelo estimado) y *predicción* (proceso de utilización del modelo identificado, estimado y validado para predecir valores futuros de las variables dependientes). En algunos casos, el modelo se obtiene como mezcla del conocimiento obtenido antes y después del *Data Mining* y también debe contrastarse antes de aceptarse como válido. Por ejemplo, las *redes neuronales* permiten descubrir modelos complejos y afinarlos a medida que progresa la exploración de los datos. Gracias a su capacidad de aprendizaje, permiten descubrir relaciones complejas entre variables sin ninguna intervención externa. Podemos incluir entre estas técnicas todos los tipos de regresión, series temporales, análisis de la varianza y covarianza, análisis discriminante, árboles de decisión, redes neuronales, algoritmos genéticos y técnicas bayesianas. Tanto los árboles de decisión, como las redes neuronales y el análisis discriminante son a su vez *técnicas de clasificación* que pueden extraer perfiles de comportamiento o clases, siendo el objetivo construir un modelo que permita clasificar cualquier nuevo dato. Los árboles de decisión permiten clasificar los datos en grupos basados en los valores de las variables. El mecanismo de base consiste en elegir un atributo como raíz y desarrollar el árbol según las variables más significativas.

En las *técnicas descriptivas* no se asigna ningún papel predeterminado a las variables. No se supone la existencia de variables dependientes ni independientes y tampoco se supone la existencia de un modelo previo para los datos. Los modelos se crean automáticamente partiendo del reconocimiento de patrones. En este grupo se incluyen las técnicas de *clustering* y segmentación (que también son técnicas de clasificación en cierto modo), las técnicas de asociación y dependencia, las técnicas de análisis exploratorio de datos y las técnicas de reducción de la dimensión (factorial, componentes principales, correspondencias, etc.) y de escalamiento multidimensional.

Tanto las técnicas predictivas como las técnicas descriptivas están enfocadas al *descubrimiento del conocimiento* embebido en los datos.

Las *técnicas auxiliares* son herramientas de apoyo más superficiales y limitadas. Se trata de nuevos métodos basados en técnicas estadísticas descriptivas, consultas e informes y enfocados en general hacia la *verificación*.

A continuación se muestra una *clasificación de las técnicas de Data Mining*.



Se observa que las *técnicas de clasificación* pueden pertenecer tanto al grupo de técnicas predictivas (discriminante, árboles de decisión y redes neuronales) como a las descriptivas (*clustering* y segmentación). Las técnicas de clasificación predictivas suelen denominarse *técnicas de clasificación ad hoc* ya que clasificar individuos u observaciones dentro de grupos previamente definidos. Las técnicas de clasificación descriptivas se denominan *técnicas de clasificación post hoc* porque realizan clasificación sin especificación previa de los grupos.

En la Figura 1-4 se muestra un diagrama con la clasificación de las técnicas de minería de datos, que es clásico en la literatura de esta materia.

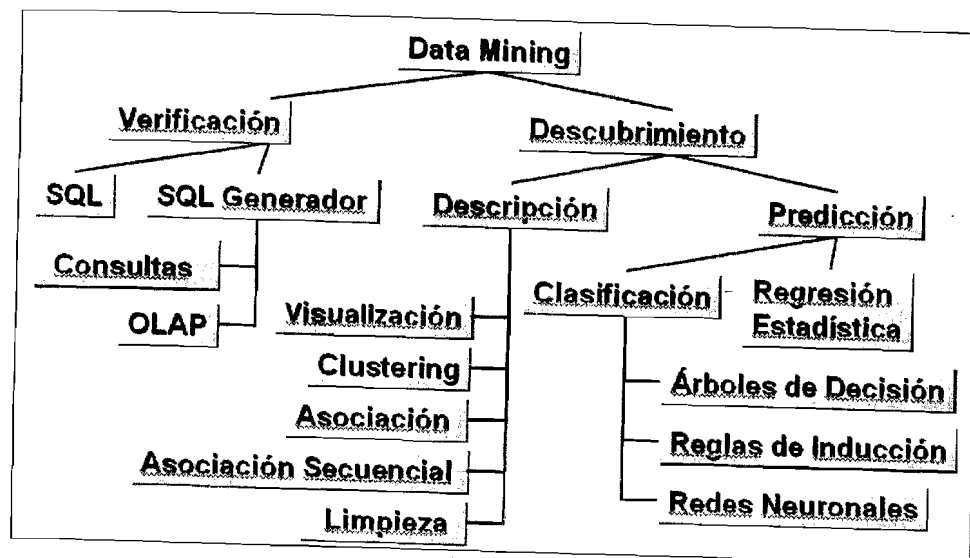


Figura 1-4

SISTEMAS DE MINERÍA DE DATOS

La Figura 1-5 muestra los sistemas de minería de datos más utilizados en el mercado junto con las técnicas que tratan cada uno de ellos, las plataformas sobre las que trabajan y los interfaces de lectura de datos.

Producto	Compañía	Técnicas	Plataforma	Interfaz
Knowledge Seeker	Angoss	Árboles de Decisión	Win	ODBC
CART	Salford Systems	Árboles de Decisión	Win/UNIX	
Clementine	SPSS	Amplio abanico	Win/UNIX	ODBC
Data Surveyor	Data Distilleries	Amplio abanico	UNIX	ODBC
Gain Smarts	Urban Science	Gráficos-Ganancias	Win/UNIX	
Intelligent Miner	IBM	Amplio abanico	UNIX (AIX)	IBM, DB2
Micostrategy	Micostrategy	Datawarehouse	Win	Oracle
Polyanalyst	Megaputer	Simbólicas	Win	Oracle, ODBC
Darwin	Oracle	Amplio abanico	Win/UNIX	Oracle
Enterprise Miner	SAS Institute	Amplio abanico	Win/UNIX/Mac	
SGI MineSet	Silicon Graphics	Asociación y Clasificación	UNIX	Oracle, Sybase, Informix
Wizsoft/Wizwhy	Wizsoft			

Figura 1-5

Los sistemas de minería de datos que utilizaremos en este libro son *SPSS Clementine* y *SAS Enterprise Miner*.

SPSS Clementine es un sistema de minería de datos que contempla diferentes fuentes de datos (ASCII, Oracle, Informix, Sybase, Ingres, etc.), una interfaz visual sencilla y distintas herramientas de minería de datos (redes neuronales, árboles de decisión, regresión, series temporales, *cluster*, etc.). Trabaja bajo los sistemas operativos UNIX y Windows.

SAS Enterprise Miner es una herramienta completa que incluye conexión a bases de datos (a través de ODBC y SAS datasets), muestreo e inclusión de variables derivadas, partición de la evaluación del modelo respecto a conjuntos de entrenamiento, validación y chequeo, distintas herramientas de minería de datos (algoritmos y tipos de árboles de decisión, redes neuronales, regresión y *clustering*, etc.), comparación de modelos y conversión de los modelos en código SAS. Dispone de un interfaz gráfico muy sencillo e incluye herramientas para flujo de proceso, tratando el proceso KDD como un proceso y las fases se pueden repetir, modificar y grabar.

Existen en el mercado otros sistemas que permiten realizar *Data Mining* a través de bases de datos. Concretamente, las bases de datos Oracle y SQL Server disponen de sistemas de minería de datos asociados.

Oracle dispone de herramientas de "Business Intelligence" y "Data Mining" (http://www.oracle.com/ip/analyze/warehouse/bus_intell/index.html) que tienen una orientación más empresarial y de sistemas de información. También dispone de herramientas de OLAP, Datawarehouse e Informes Avanzados. Asimismo, presenta herramientas propias de Minería de Datos a través del producto Oracle Darwin (<http://www.oracle.com/ip/analyze/warehouse/datamining/index.html>).

Microsoft SQL Server dispone del producto *Analysis Services* que implementa la minería de datos. Se fundamenta en el "OLE DB for Data Mining" e implementa una extensión del SQL que trabaja con DMM (*Data Mining Model*) que permite crear el modelo, entrenarlo y realizar predicciones. La versión SQL Server 2005, en su módulo *Analysis Services* cuenta con los algoritmos de minería de datos más avanzado entre los que se incluyen árboles de decisión y regresión, series temporales, agrupación en clústeres, reglas de asociación, algoritmo Naive Bayes y minería de textos. Dispone de un asistente y diseñador para minería de datos que permite construir modelos sofisticados a través de una interfaz fácil de usar. Además, se proporcionan gráficos de elevación y beneficios, por lo que podrá comparar y contrastar la calidad de los modelos antes de dedicarse a la distribución.

Existe una representación clásica de los sistemas de minería de datos cuya fuente es Elder Research (www.datamininglab.com) y que se presenta en la Figura 1-6.

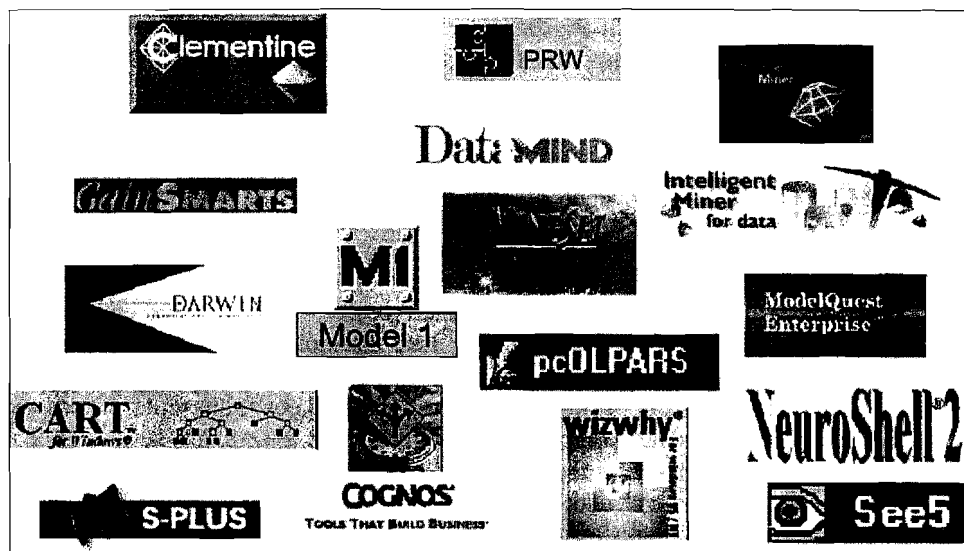


Figura 1-6