Introducción a la Minería de Datos

DE LOS DATOS AL CONOCIMIENTO...

JAIME MIRANDA

Departamento de Ingeniería Industrial

Universidad de Chile

AGENDA

DEFINICIONES Y NOCIONES BÁSICAS DE DATA-MINING

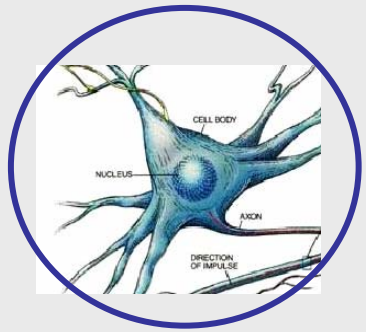
DESCRIPCIÓN DEL CURSO:

→ Parte 1: Análisis multidimensional de datos.

→ Parte 2: Técnicas de Data Mining.

→ Parte 3: Caso aplicado de estudio.

CRITERIOS DE EVALUACIÓN



¿QUÉ ES DATA-MINING?

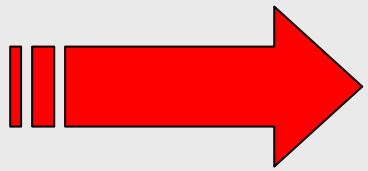
ALGUNAS DEFINICIONES:

→ *"Proceso de extracción de información y patrones de comportamientos que permanecen ocultos entre grandes cantidades de información."*

→ *"Proceso que a través del descubrimiento y cuantificación de relaciones predictivas en los datos, permite transformar la información disponible en conocimiento útil."*

Información

Relaciones



Conocimiento útil

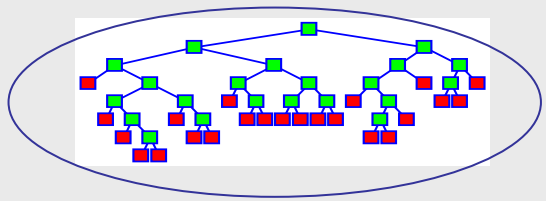
Patrones ocultos

¿POR QUÉ ES NECESARIO?

Las empresas de todos los tamaños necesitan aprender de sus datos para crear una relación “one-to-one” con sus clientes.

Las empresas recogen datos de todos sus procesos.

Los datos recogidos se tienen que analizar, comprender y convertir en información con la que se pueda actuar y aquí es donde Data Mining juega su papel.



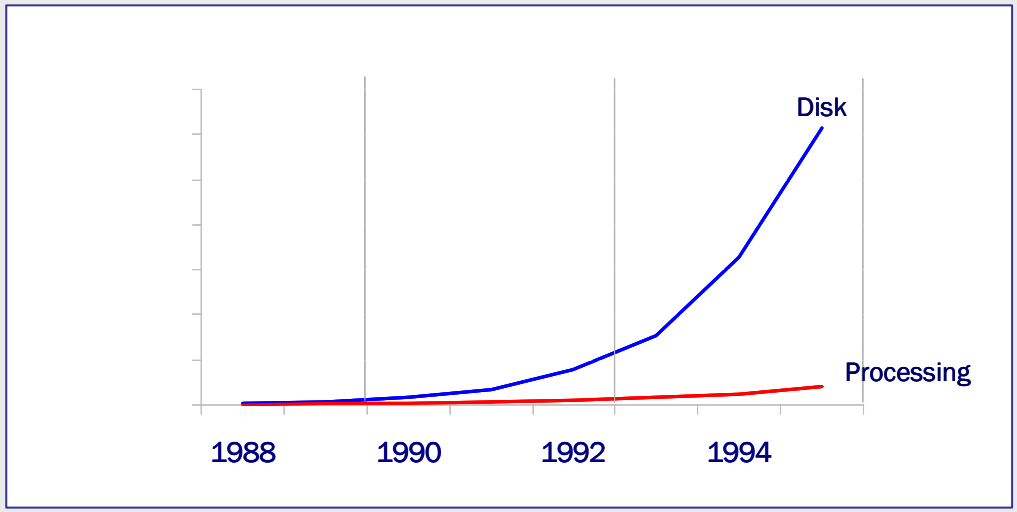
ALMACENAMIENTO Y CAPACIDAD DE PROCESAMIENTO

LEY DE MOORE:

→ “La capacidad de procesamiento se duplica cada 18 meses”

RESPECTO AL ALMACENAMIENTO:

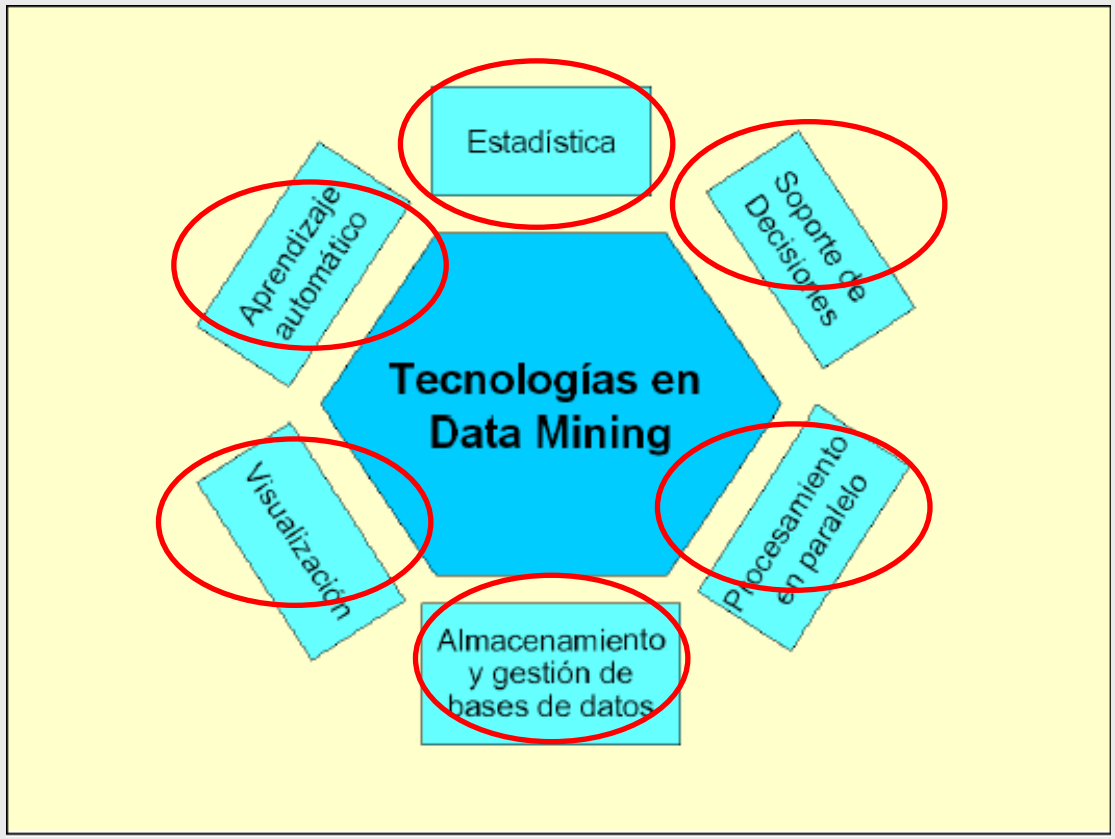
→ “La capacidad de almacenamiento se duplica cada 9 meses”



La brecha entre capacidad de procesar lo que almacenamos, aumenta con el tiempo

¿DE DONDE SURGE EL DATA-MINING ?

De la integración múltiple...



APLICACIONES DE NEGOCIO

DETECCIÓN DE FRAUDES:

→ Identificar transacciones fraudulentas

MARKETING Y VENTAS:

→ Identificar potenciales clientes; establecer la efectividad de las campañas de marketing

ANÁLISIS DE PROCESOS DE MANUFACTURA:

→ Identificar las causas de fallas en máquinas

ENTENDIENDO COMPORTAMIENTO DE CONSUMIDORES:

→ modelos de retención de clientes, afinidades, clustering

APLICACIONES DE NEGOCIO (2)

APROBAR CRÉDITOS:

→ Establecer Credit Scoring para un cliente a la hora de pedir un préstamo

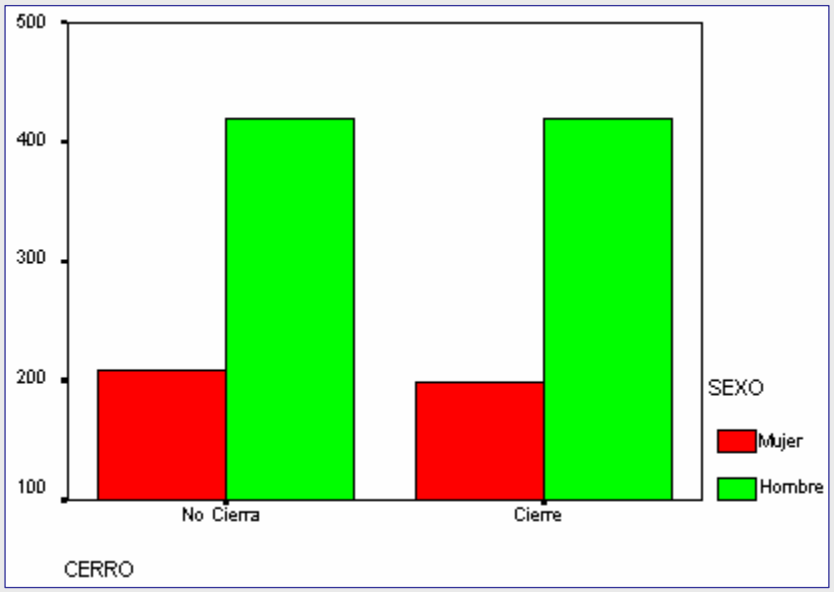
GESTIÓN DE PORTAFOLIO:

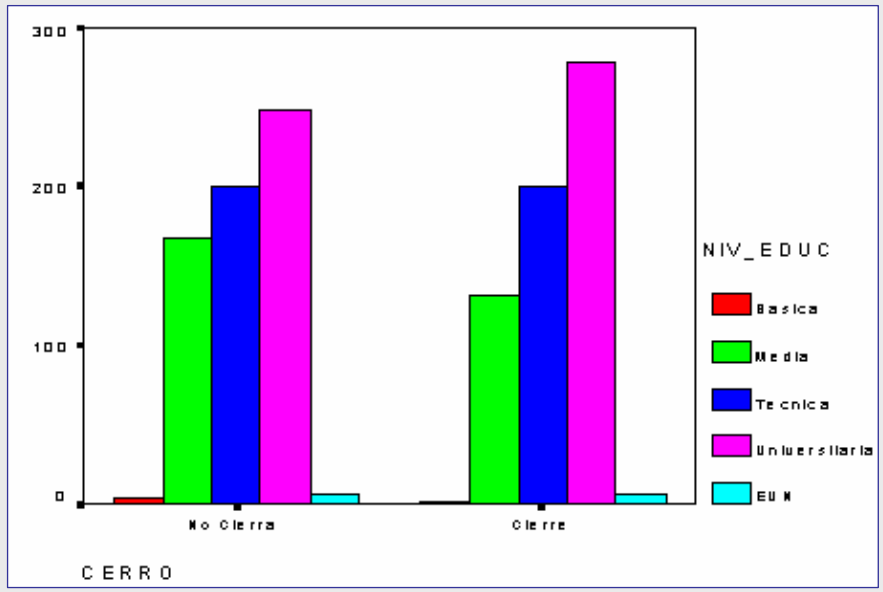
→ optimizar un portafolio de instrumentos financieros maximizando el retorno o minimizando el riesgo

ANÁLISIS DE WEBSITES:

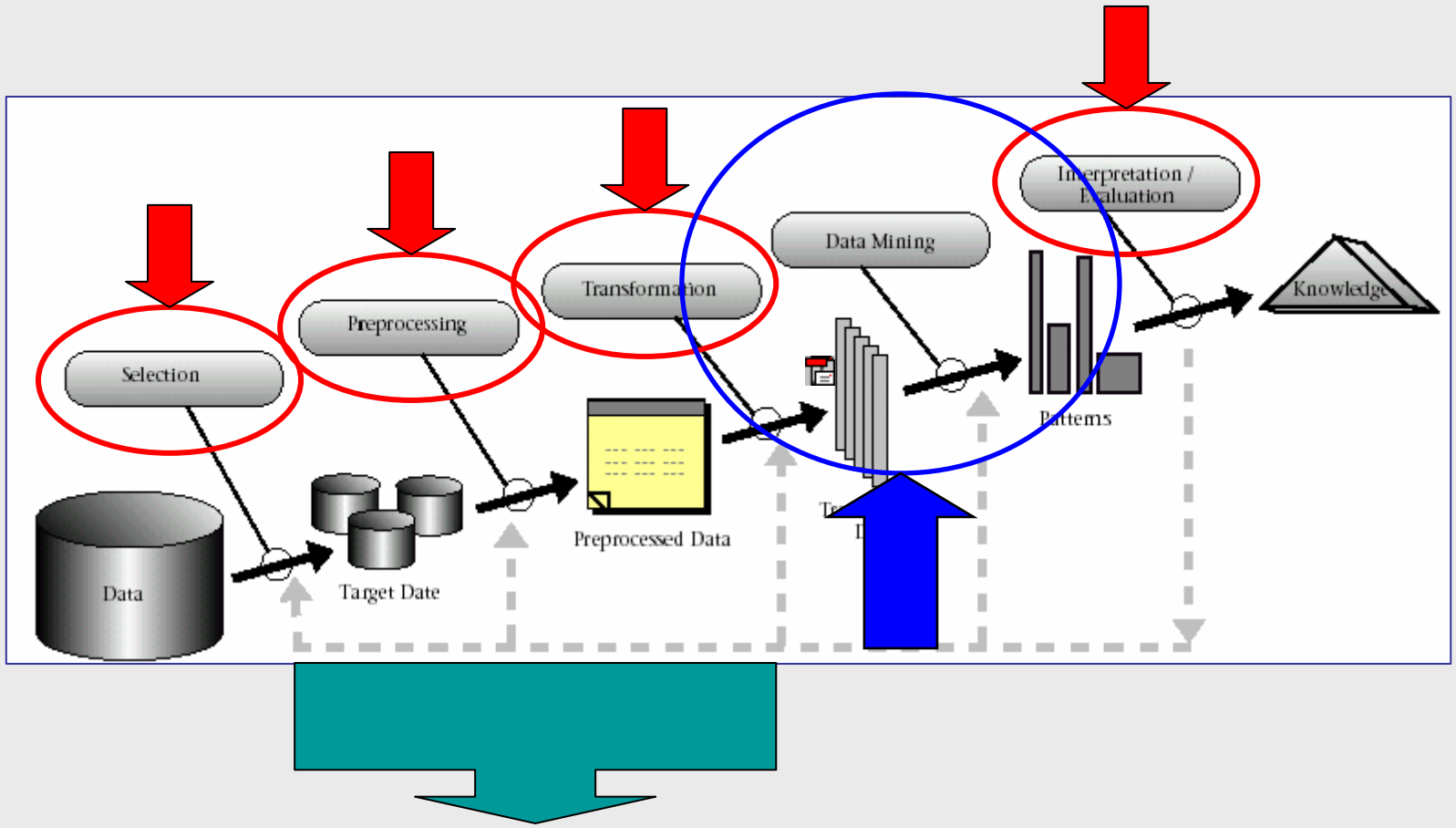
→ modelar preferencias de usuarios desde logs, filtros colaborativos, caminos preferidos, etc.

UN PEQUEÑO EJEMPLO …



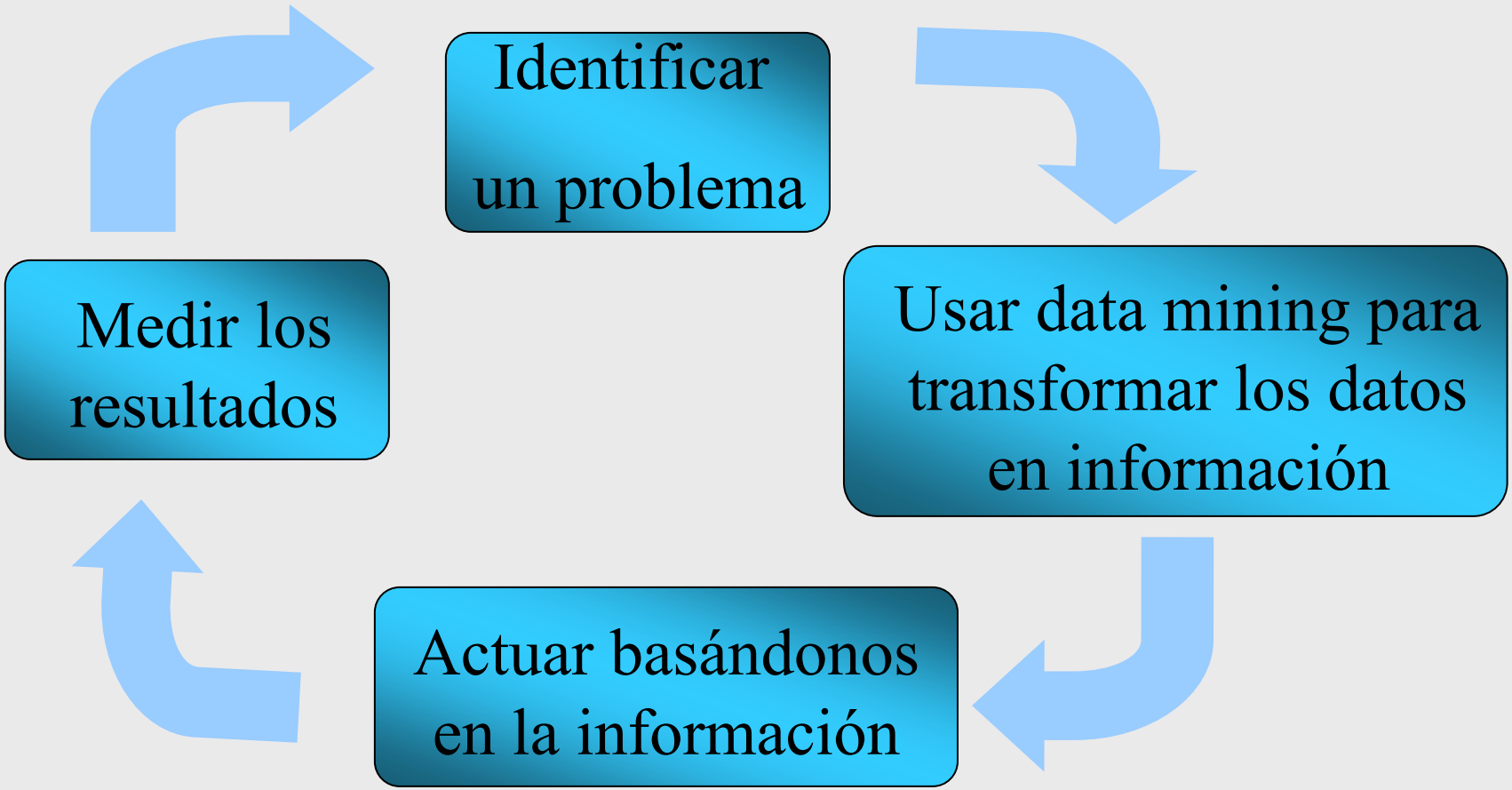


¿DÓNDE ESTAMOS PARADOS?

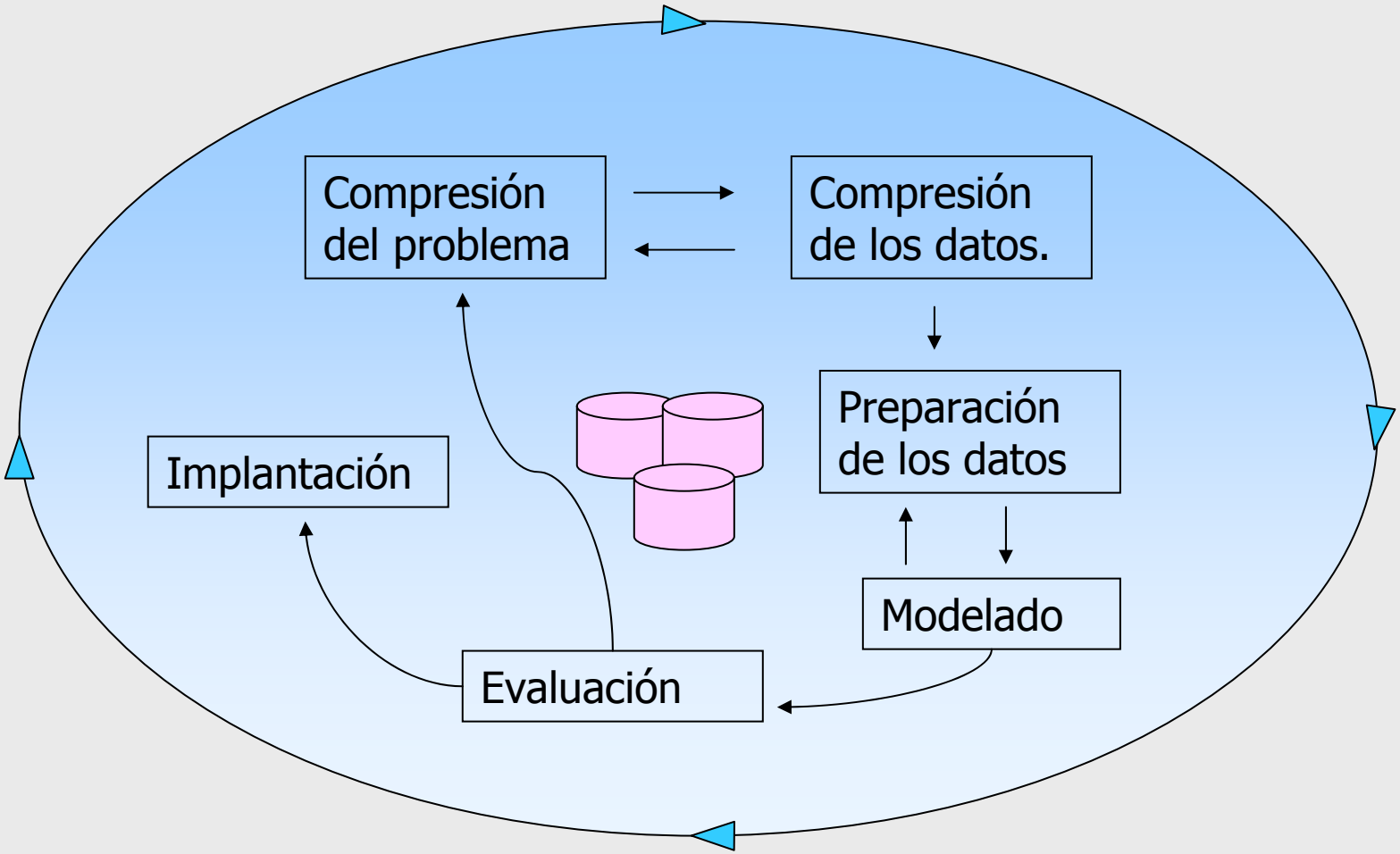


80% tiempo

EL CICLO DE DATA MINING



MÁS EN DETALLE …



ALGUNAS ÁREAS TRATADAS EN DM

CLASIFICACIÓN

→ Consiste en etiquetar los objetos y crear un modelo que los clasifique bajo algún criterio.

ESTIMACIÓN O REGRESIÓN

→ Es la asignación de un valor ausente en un campo, en función de los demás campos presentes en el registro o de los mismos registros existentes.

SEGMENTACIÓN:

→ Consiste en fraccionar el conjunto de los registros (población) en subpoblaciones de comportamiento similar.

PROBLEMAS DE CLASIFICACIÓN

Examinar las características de un nuevo objeto y asignarlo a una clase dentro de un conjunto de clases predefinido.

→ Clasificar personas que piden créditos como alto medio o bajo riesgo
→ Determinar el patrón de las quejas de seguros fraudulentas
→ Patrón de los clientes que nos dejarán en los próximos 6 meses

Se ha de disponer de un conjunto de entrenamiento en el que todos los registros estén clasificados

El problema consiste en construir un modelo que aplicado a un nuevo ejemplo sin clasificar lo clasifique.

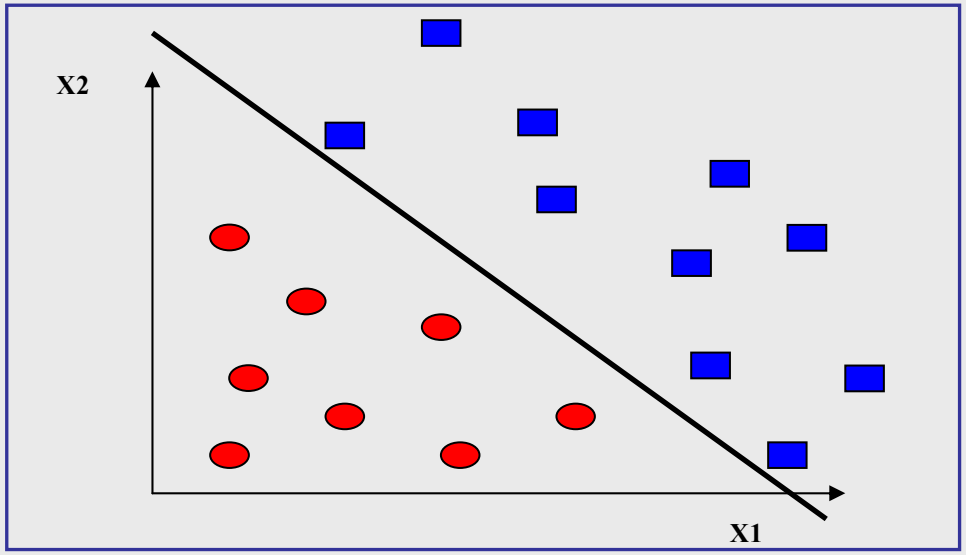
Se tiene siempre un número limitado de clases y se esperar poder asignar cualquier nuevo objeto en una de esas clases.

PROBLEMAS DE CLASIFICACIÓN (2)

Determinación de la pertenencia de un objeto a una cierta clase específica.

Encontrar la mejor función que discrimine este fenómeno.

Aplicar la función encontrada a nuevos objetos.



PROBLEMAS DE REGRESIÓN

La clasificación trata con problemas de salidas discretas (si o no, alto, medio o bajo riesgo, responderá o no responderá...)

La estimación trata con problemas donde el valor a clasificar puede tomar valores en un rango continuo (ingresos, balance de la tarjeta de crédito, probabilidad de que sea jugador)

Ejemplos

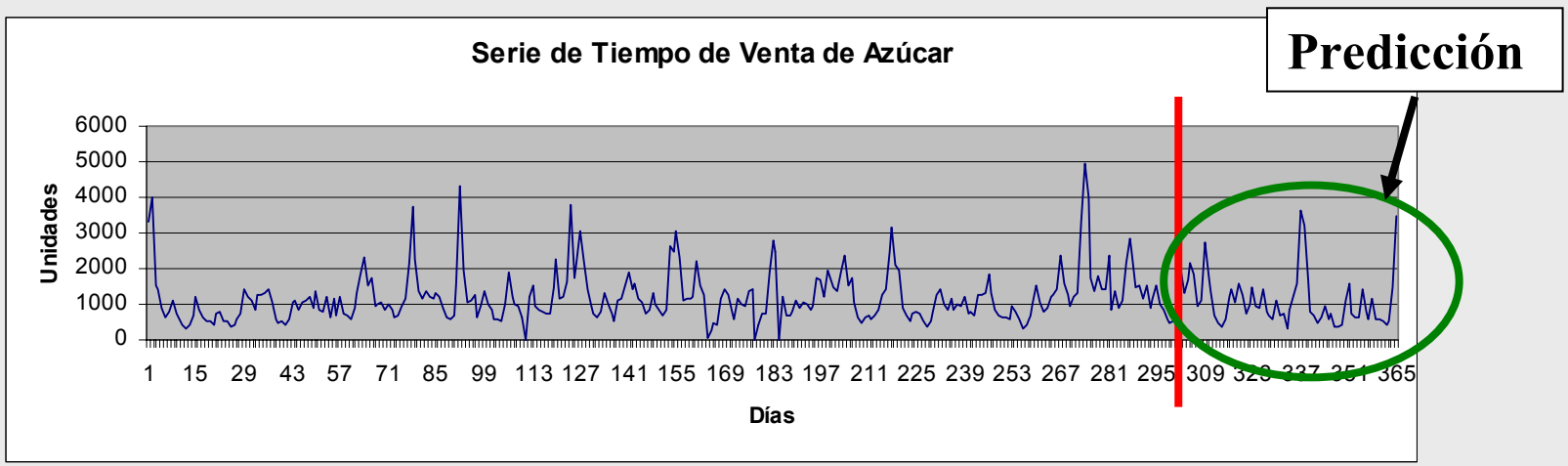
→ Estimar el número de hijos de una familia
→ Estimar la probabilidad de que alguien conteste a un mailing
→ Estimar el tiempo de vida de un cliente
→ Estimar los ingresos totales de una familia

PROBLEMAS DE REGRESIÓN (2)

Estudiar el comportamiento temporal y dinámico de alguna variable.

Encontrar la mejor función que describa este fenómeno.

Aplicar la función encontrada a la predicción de nuevos valores de la serie.



ASOCIACIÓN O CANASTA DE LA COMPRA

IDEA CENTRAL: Determinar que cosas van juntas.

→ Pañales y cerveza se compran juntos los fines de semana

El ejemplo típico es observar qué productos suelen ir juntos en la cesta de la compra

Se puede utilizar para establecer los almacenes, escaparates y estrategias de Cross-selling.

PROBLEMAS DE SEGMENTACIÓN

Segmentar una población heterogénea en un número de subgrupos homogéneos o clusters.

No hay clases predefinidas

Registros agrupados en base a su similitud.

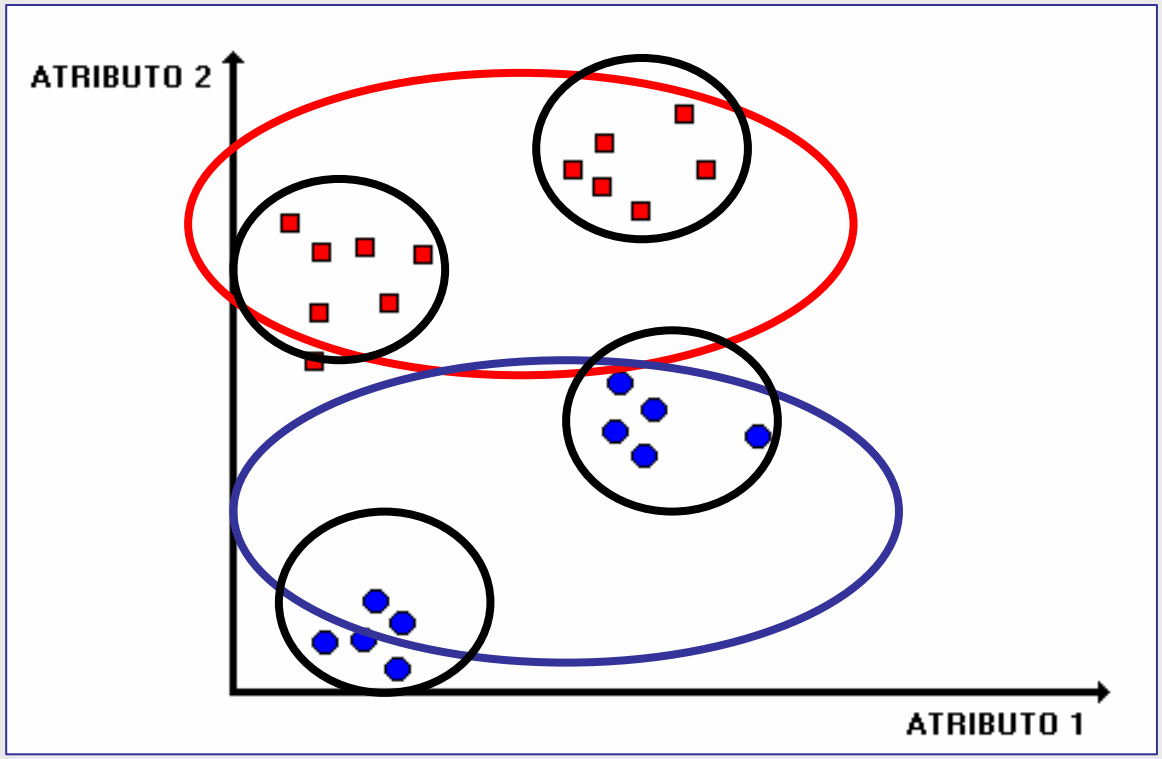
Se realiza a menudo antes de otras tareas de descubrimiento.

→ Encontrar clientes con hábitos de compra similares

PROBLEMAS DE SEGMENTACIÓN (2)

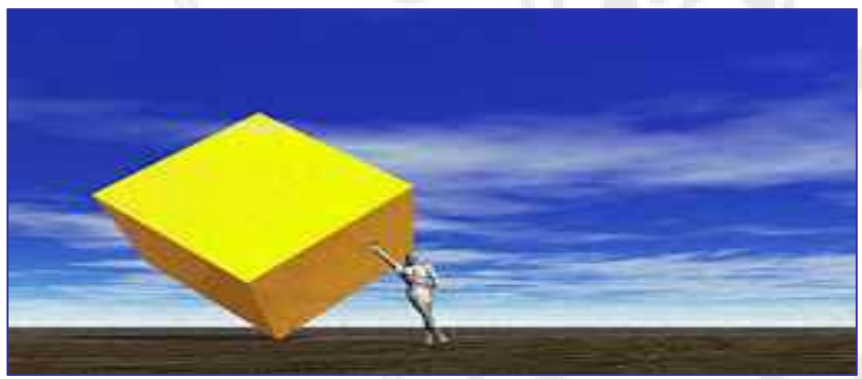
Encontrar patrones característicos no visibles a simple vista.

Encontrar soluciones entre subconjuntos o subpoblaciones.



PARTE Nº1

ANÁLISIS MULTIDIMENSIONAL DE DATOS



ANÁLISIS MULTIDIMENCIONAL

→ Trabaja con un conjunto reducido de datos.

→ Genera reportes mas claros y más específicos.

→ Responden con mayor rapidez a las consultas

Mejorar el rendimiento en línea y
el rendimiento de las consultas a las bases

CUBOS MULTIDIMENSIONALES

COMPONENTES PRINCIPALES

→ Dimensiones

→ Medidas

PRINCIPALES ARQUITECTURAS

→ Estrella

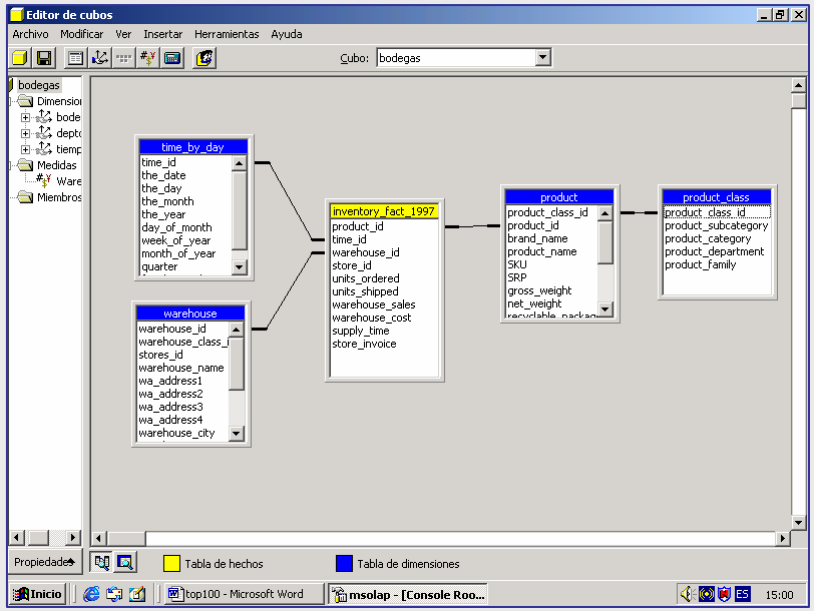
→ Copo de nieve

PRINCIPALES TIPOS

→ OLAP

→ MOLAP

→ ROLAP



SQL Server 7.0

TAREA Nº1: ANÁLISIS MULTIDIMENCIONAL

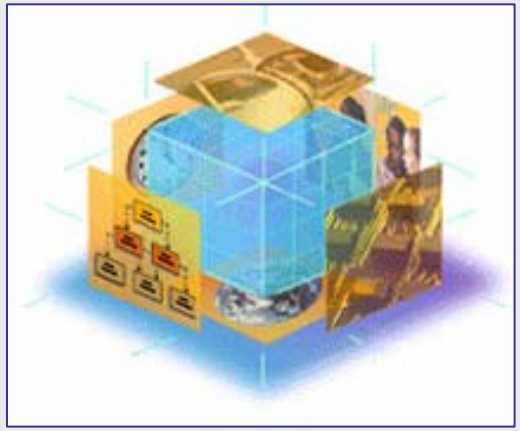
OBJETIVO GENERAL

→ "Entender y usar las herramientas de análisis multidimensional con el fin de comprender y describir el comportamiento de una variable dada"

SE TENDRÁ:

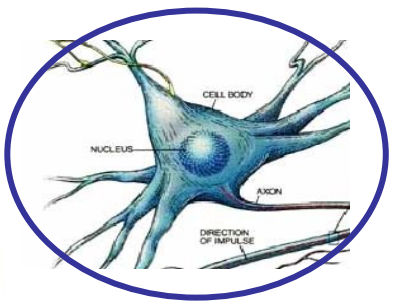
→ Una base de datos relacional.

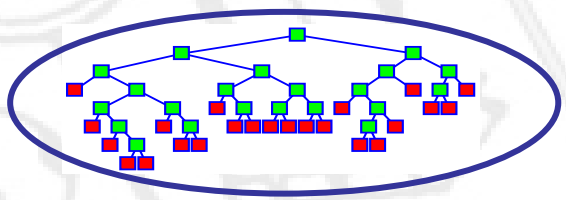
→ Una serie de consultas y preguntas sobre el comportamiento de una o más variables.



PARTE Nº2

TÉCNICAS DE MINERÍA DE DATOS



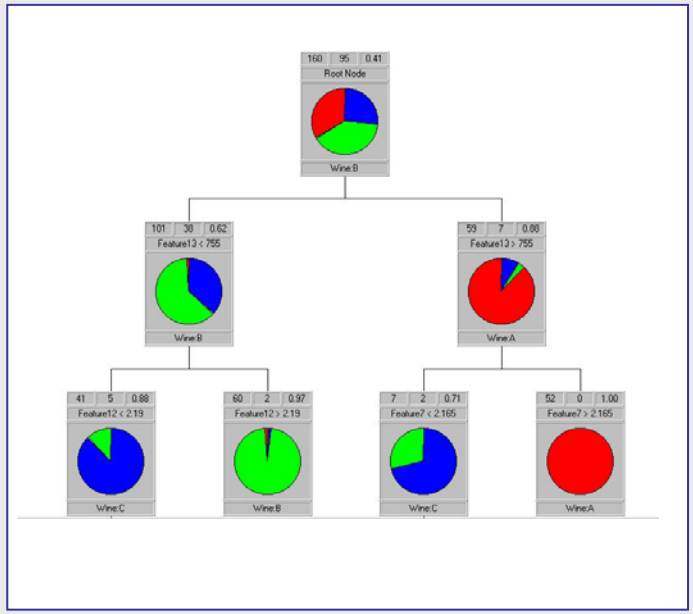


TÉCNICAS EMPLEADAS

MÉTODOS SUPERVISADOS

→ Redes neuronales

→ Árboles de decisión



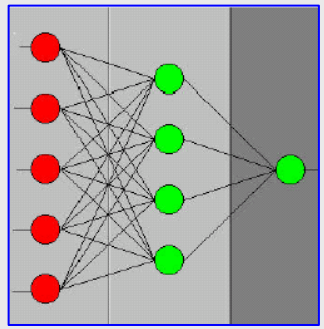
MÉTODOS NO SUPERVISADOS

→ Fuzzy C-means (Cluster)

→ Mapas Kohonen

REDES NEURONALES MLP

APLICACIONES



→ Retención o fuga de clientes
→ Detección de fraudes
→ Scoring (varios)

FORTALEZAS

→ Fuerte en lo referente a la modelación no lineal
→ Trabaja tanto con variables categóricas como continuas
→ Alta aplicabilidad (variadas áreas de estudio)

DEBILIDADES

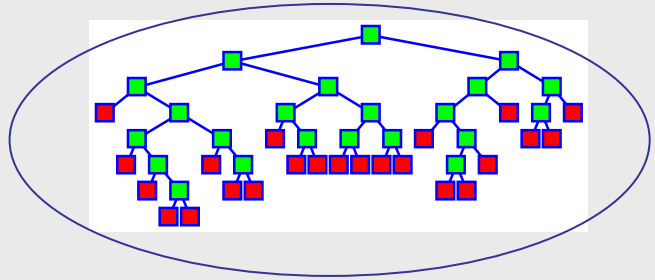
→ Difícil interpretación de las relaciones entre las variables (Heurísticas)
→ Sobreajuste

ÁRBOLES DE DECISIÓN

APLICACIONES

→ Segmentación de clientes

→ Generación de reglas de clasificación en general



FORTALEZAS

→ Fácil interpretación y entendimiento

→ Genera un ranking automático de variables

→ Rápida convergencia del algoritmo

DEBILIDADES

→ Si poseen mucha “profundidad” son difíciles de interpretar

→ Posibilidades discretas: relacionado a variables con muchas categorías

ALGORITMO FUZZY C-MEANS

APLICACIONES

→ Marketing
- Segmentación de clientes
- Ofertas focalizadas

FORTALEZAS

→ No asume ninguna distribución estadística entre los datos.

→ Los resultados son mas intuitivos y fáciles de entender e interpretar

DEBILIDADES

→ Son muy sensitivos a los valores fuera de rango (outliers)

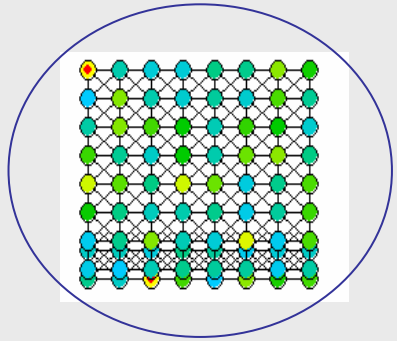
→ No trabajan bien con variables categóricas

MAPAS DE KOHONEN

APLICACIONES

→ Marketing

- Segmentación de clientes
- Ofertas focalizadas



FORTALEZAS

→ Reduce la dimensionalidad del espacio.

→ Los resultados son mas intuitivos y fáciles de entender e interpretar

DEBILIDADES

→ Sólo nos da una visón espacial de los resultados

→ No trabajan bien con variables categóricas

TAREA Nº2: CONSTRUCCIÓN DE MODELOS DE DM

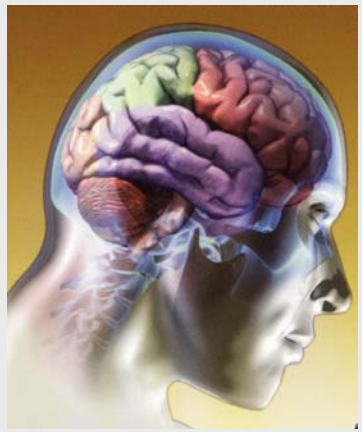
UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
INGENIERÍA INDUSTRIAL

OBJETIVO GENERAL

→ Comprender y estudiar los distintos métodos de minería de datos
→ Hacer sensibilidad de parámetros y obtener configuraciones optimas
→ Seleccionar de atributos relevantes para el análisis

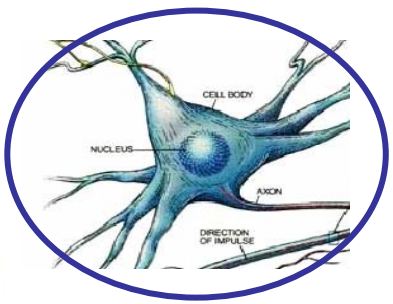
SE TENDRÁ:

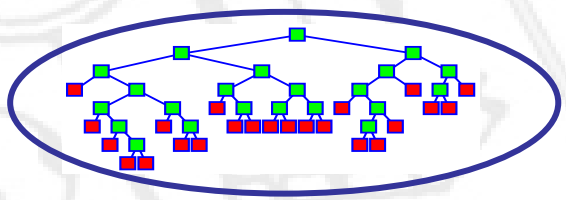
→ Una base de datos de un problema de clasificación con la clase especifica
→ Una base sin la clase y el problema será predecir dicha clase



PARTE Nº3

CASOS APLICADOS A LA INDUSTRIA





OBJETIVOS DE LOS CASOS APLICADOS

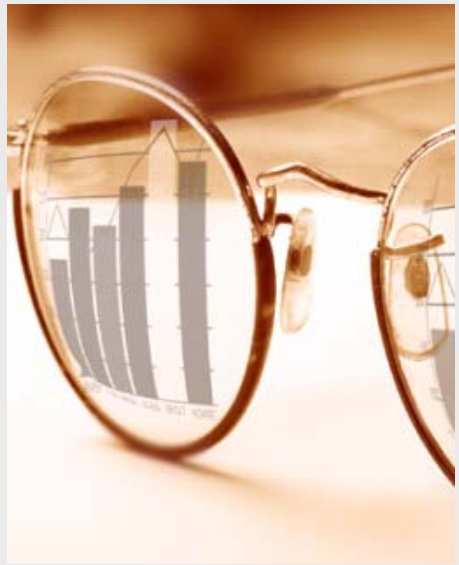
OBJETIVOS

→ Lograr un a primera aproximación a un problema real en Data-Mining.

→ Desarrollar de habilidades en el uso practico de algunas de las técnicas usadas en la minería de datos.

→ Encontrar el mejor modelo para la resolución del problema.

→ Generar políticas comerciales asociadas a cada problema en particular.



CASOS PRELIMINARES DE ESTUDIO

UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
INGENIERÍA INDUSTRIAL

CASO Nº1: Modelo Predictivo de Ofertas Focalizadas

CASO Nº2 : Modelo Predictivo para Créditos de Consumo

CASO Nº3 : Modelo Predictivo de Fugas de Cliente

CRITERIOS DE EVALUACIÓN

TAREAS

Nota Tarea = 0.85* Nota Informe + 0.15* Nota Interrogación.
Nota Tarea Final = Promedio de tareas.

CTP´S

6 * (número de CTPs aprobados/número total de CTPs) +1

NOTA FINAL

Nota examen 40%
Nota promedio de tareas 40%
Nota CTP 20%

Introducción a la Minería de Datos

DE LOS DATOS AL CONOCIMIENTO...

JAIME MIRANDA

Departamento de Ingeniería Industrial

Universidad de Chile