UADER

Facultad de Ciencia Y Tenología

Base de Datos Avanzadas

DATA MINING *Conceptos Teóricos*

"Lo esencial es invisible a los ojos " Saint-Exupery

Docentes:

Ing Sato Fernando

A.S Trossero Sebastián

Versión 20191027

Resumen

- Definición de Data Mining y KDD
- Proceso de Data Mining
- Comparación de DM y consultas de DB
- Modelos de Data Mining
- Inconvenientes del Data Mining
- Potenciales aplicaciones
- Funciones de Data Mining
- Técnicas y Herramientas de uso frecuente en Data Mining

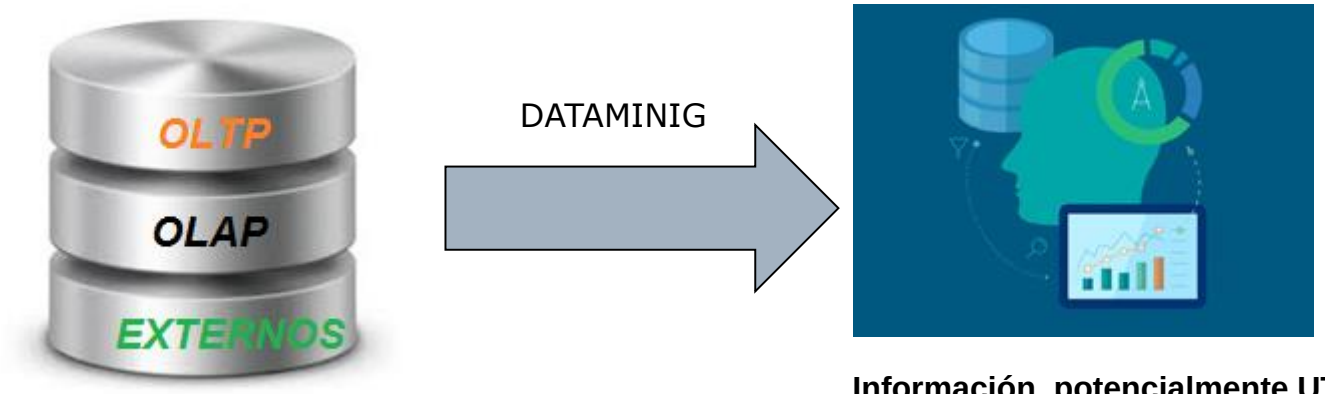
¿ Que es Data Mining?

Transformar datos en conocimiento

"Datos con mucha riquesa, Información muy pobre" → *Conocimiento (Patrones Interesantes)*

Data Mining

Es un proceso automático que permite extraer tendencias interesantes, no triviales y previamente desconocidas, de los datos y descubrir relaciones entre variables



Información potencialmente UTIL

Data Mining

¿Cómo funciona?

Data mining es un proceso analítico que utiliza como materia prima las bases de datos para encontrar patrones y las relaciones que se encuentran ocultos dentro de los datos, para de esa manera crear modelos, representaciones abstractas de la realidad y representación de datos obtenidos

Nota: Además de encontrar información, genera información nueva que posteriormente servirá para apoyar la toma de decisiones.

Data Mining

Actualmente se considera a Data Mining el elemento clave de un proceso mucho más elaborado llamado **KDD** (Knowledge Discovery and Data Mining), el cual está estrechamente ligado a otro importante tema ya aboradado, el Data Warehousing.

Data Mining

Si bien los términos Minería de Datos (Data Mining) y Descubrimiento de Conocimiento en Bases de Datos (KDD) son usados como sinónimos.

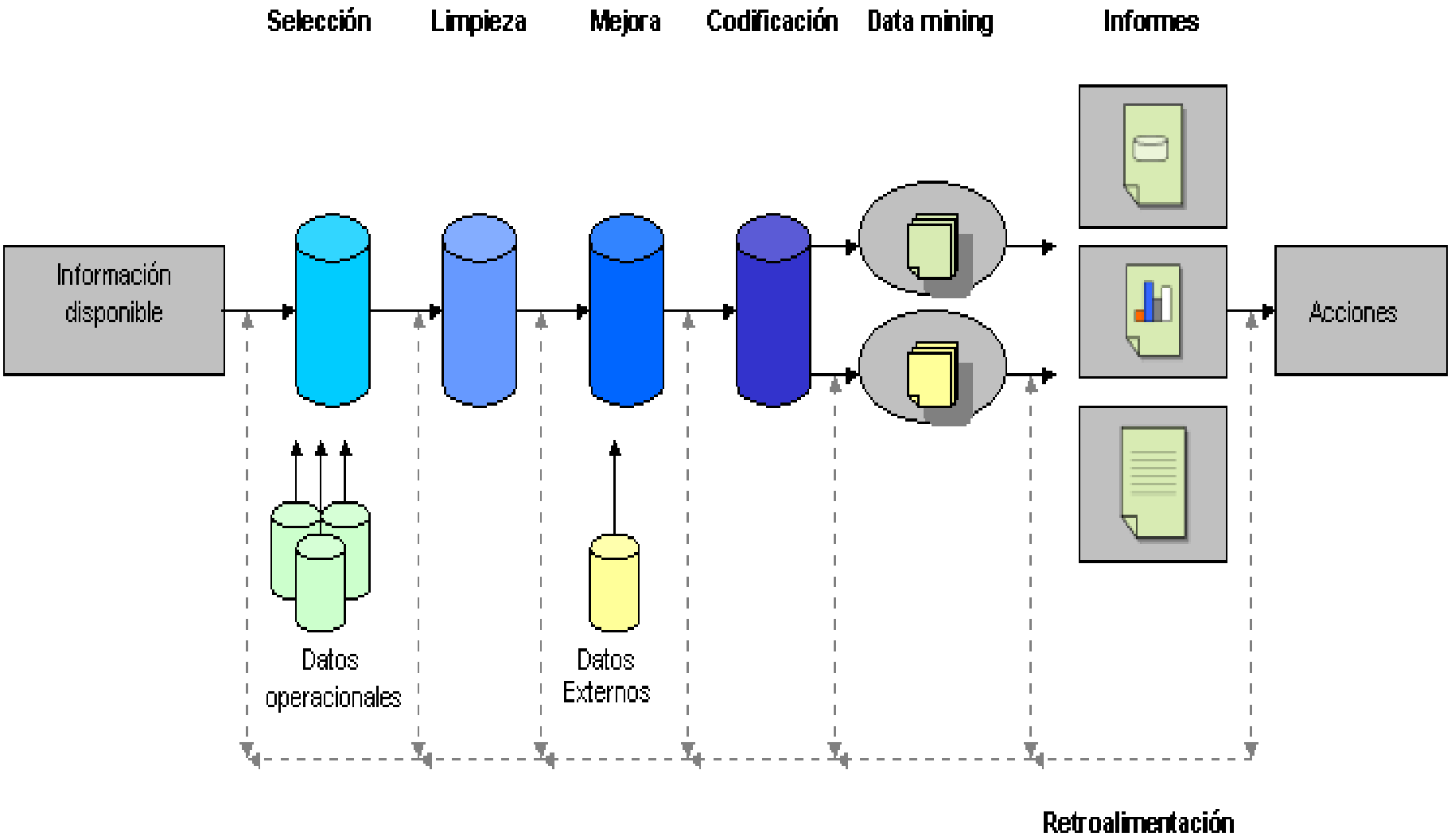
Para Algunos autores, el término KDD describe el proceso completo de extracción de conocimiento a partir de los datos,

Mientras que minería es un conjunto de técnicas.

KDD - Definición

- KDD lo podemos definir como “la extracción no trivial de conocimiento previamente desconocido y potencialmente útil a partir de un gran volumen de datos en el cual la información está implícita”. [Fayyad, 1996b].

KDD - Definición



¿Como nos ayudan DM y KDD?

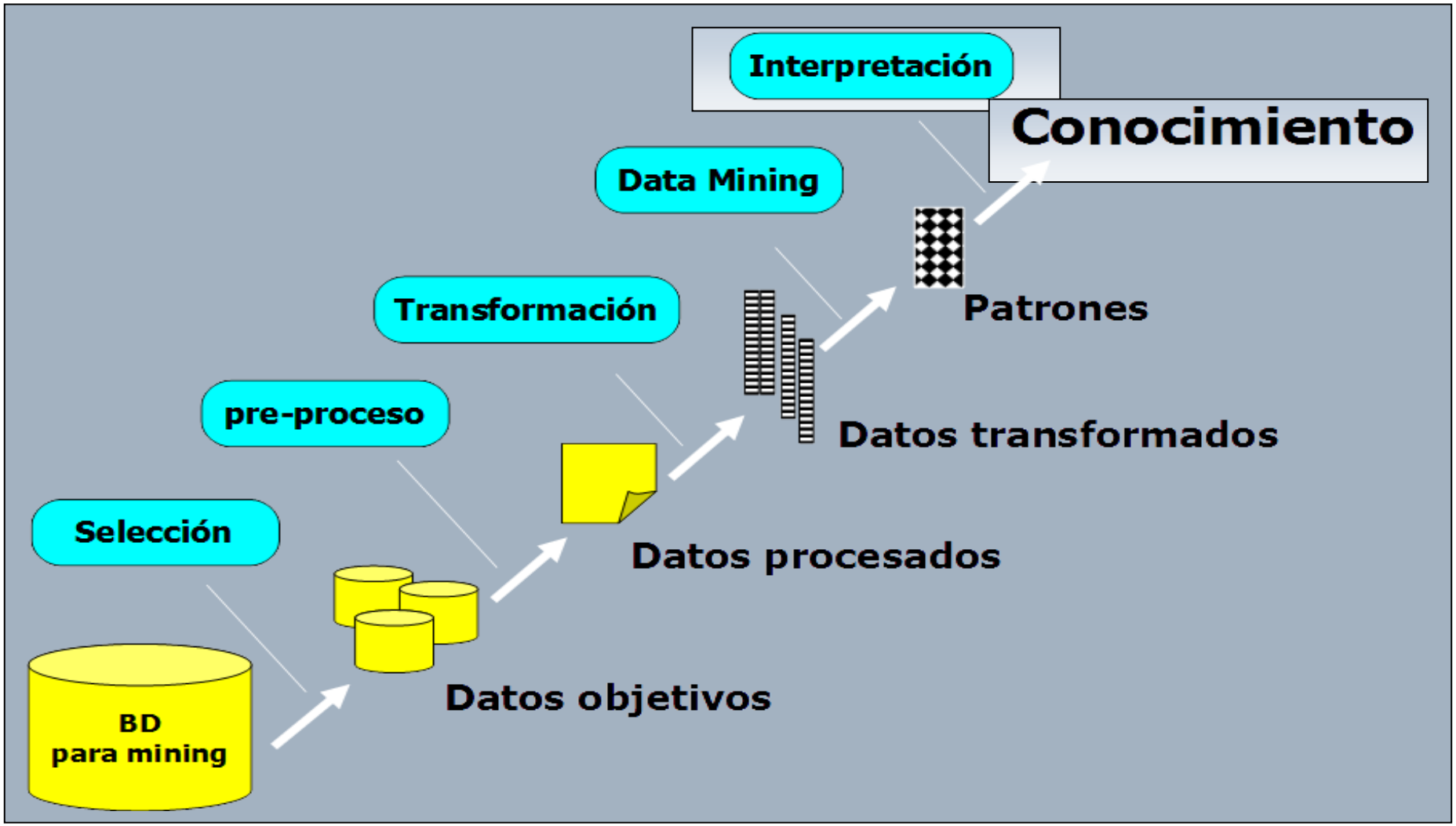
- ¿Qué clientes tienen mas probabilidad de permanecer fieles a nuestra organización?
- ¿Qué clientes están a punto de abandonarnos?
- ¿Dónde debemos localizar la próxima sucursal?
- ¿Qué productos se deben promocionar a qué prospectos?
- ...

- Las respuestas a estas preguntas están enterradas en los datos y se necesitan las técnicas de Data Mining para buscarlas.

Resumen

- Definición de Data Mining y KDD
- Proceso de Data Mining
- Comparación de DM y consultas de DB
- Modelos de Data Mining
- Potenciales aplicaciones
- Funciones de Data Mining
- Técnicas y Herramientas de uso frecuente en Data Mining

Proceso de Data Mining



Selección de los datos

- En este paso se seleccionan los datos necesarios para el análisis. En la mayoría de los casos, estos datos se encuentran almacenados en bases de datos operacionales usadas por los sistemas de información de la organización.
- A las bases de datos operaciones se las suele complementar con información externa proveniente de la competencia y entes regulatorios.

Limpieza

- Una tarea importante en la operación de limpieza es la de eliminar registros duplicados (de-duplicación), por ejemplo un cliente puede aparecer cargado 2 veces como consecuencia de una mala carga y/o un bajo control de integridad.
- Otro problema es la falta de consistencia de los dominios, por ejemplo una transacción listada en una tabla finalizada en 2000 pero la compañía se estableció después de 2001.

Preproceso (Limpieza)

Los datos deben ser analizados a efectos de determinar valores erróneos, valores incorrectos o datos faltantes

Estructura Errónea

Errores en los atributos o en la definición de los mismos

Valores incorrectos: La estrategia a seguir puede ser:

- descartarlos,
- omitir el registro,
- inferir valores a partir de valores conocidos,
- tratarlo como un valor especial
- usar un valor promedio

Transformaciones

Los datos pueden requerir transformaciones por diferentes razones:

1. Debido a las diferencias de escalas de los mismos, a efectos de evitar problemas numérico durante el procesamiento.

   - Estrategia: realizar transformación definiendo una escala apropiada.

2. Debido a la presencia de valores cualitativos (no numéricos), que requieren ser cuantificados mediante escalas apropiadas.

Transformaciones

Codificación: Otras veces la información seleccionada no tiene el formato requerido por los algoritmos de reconocimiento de patrones.

Algunos ejemplos de codificaciones incluyen:

- pasar de dirección a región,
- fecha de nacimiento a edad, o franja etaria.
- dividir ingresos por 1000, etc.

Técnicas de Data Mining

Existen diferentes técnicas que se usan para descubrir conocimiento oculto:

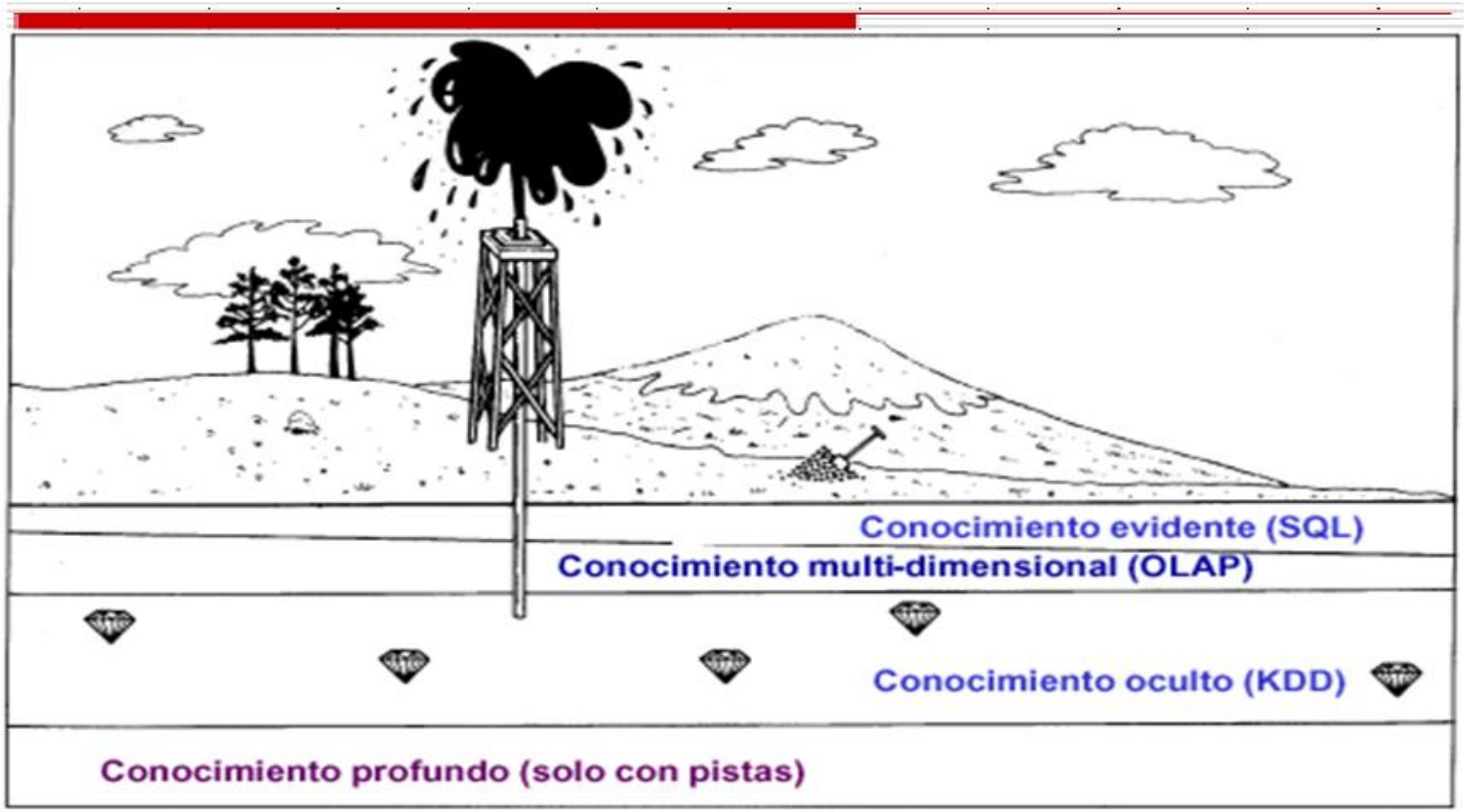
- *Procesamiento analítico en línea (OLAP)*
- *Reglas de asociación*
- *Arboles de decisión*
- *Clustering*

Interpretación

Una vez encontrados los patrones de comportamiento, los mismos deben ser interpretados en el dominio del negocio a efectos de generar información de utilidad (Conocimiento).

Reportes: Los resultados del proceso de Data Mining se pueden mostrar en diferentes formatos. En general, se pueden usar reportes o gráficos para visualizar los resultados.

Descubrimiento de Conocimiento



SQL: Structured Query Language
OLAP: Online Analytical Processing
KDD: Knowledge Discovery on Databases

Conocimiento evidente

- Esta es la información que se puede recuperar fácilmente de bases de datos usando herramientas de consulta tales como SQL.

  Ejemplos:

  - Ventas del último mes de un producto
  - Listado de clientes que no renuevan la póliza de seguro

Conocimiento multidimensional

- Esta es la información que se puede analizar con procesos analíticos en línea OLAP, o también por que no utilizando SQL.
  La ventaja de OLAP es que está optimizada para este tipo de búsqueda y operaciones de análisis.

  Ejemplos:
  - Distribución Geográfica de Ventas por producto (con visión temporal).
  - Distribución de Ventas por producto y Clientes.

Conocimiento oculto

- Estos datos se pueden encontrar fácilmente con KDD y en particular con algoritmos de Data Mining. Una vez más, se podría utilizar SQL para encontrar estos patrones pero se consumiría una enorme cantidad de tiempo. Es decir, utilizando algoritmos de Data Mining se pueden encontrar datos ocultos en minutos, mientras que utilizando SQL se tardarían meses para conseguir los mismos resultados.

Conocimiento oculto

Ejemplos:

- Qué características comparten los clientes que no renovaron la póliza de seguros y cómo difieren de aquellos que la renovaron?
- Por qué la sucursal Concordia es más rentable que Paraná?

Conocimiento profundo

- Esta es la información que está almacenada en la base de datos pero sólo puede ser localizada si se tienen pistas que indiquen donde buscar.

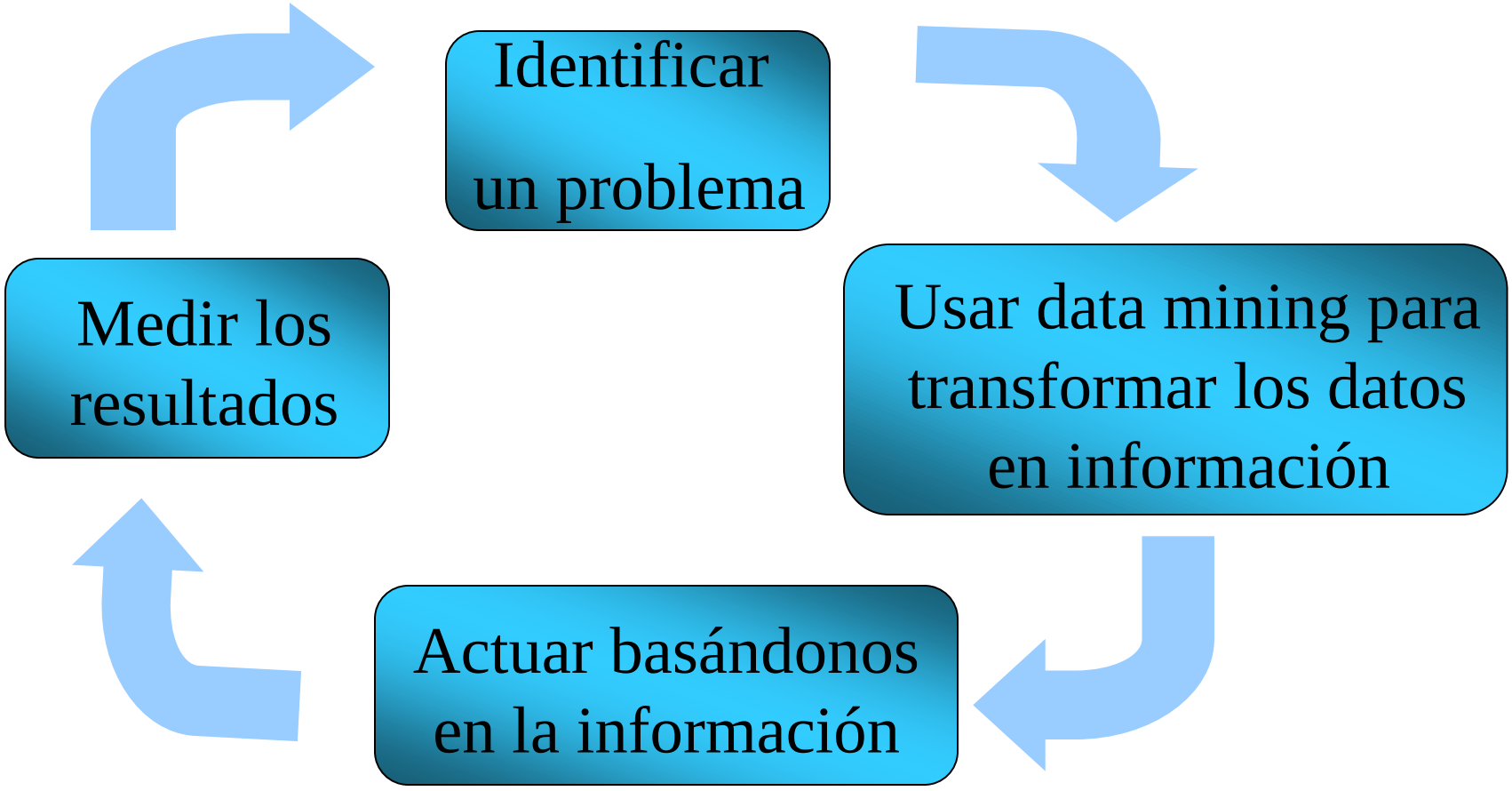
  Ejemplo:

  - Determinar los tipos de productos que incrementan su venta, después de eventos como partidos de futbol de la selección nacional y/o cobro de aguinaldo.

OLAP Vs Data Mining

OLAP	Minería de datos
¿Cuál es la proporción media de accidentes entre fumadores y no fumadores?	¿Cuál es la mejor predicción para accidentes?
¿Cuál es la factura telefónica media de mis clientes y de los que han dejado la compañia?	¿Dejara X la compañia? ¿Qué factores afectan a los abandonados?
¿Cuánto es la compra media diaria de tarjetas robadas y legítimas?	¿Cuáles son los patrones de compra asociados con el fraude de tarjetas?

El ciclo de data mining



Preparación del Data Mining

- Se exploran los datos
  - Distribución
  - Relación
  - Influencia
- Se preparan los datos
  - Se eligen variables
  - Se eligen las tablas y filas que inciden
  - Se crean nuevas variables
  - Se transforman las variables

El Proceso de DM en la Practica

- Se construye el modelo
- Se entrena y ejecuta el modelo
  - Datos para entrenamiento
  - Datos para prueba
- Se prueba el modelo
- Se evalúan los resultados
- Se rehacen corridas si es necesario
- Se guardan los resultados

Data Mining

Dónde se puede utilizar

- Marketing: Segmentación, campañas, rentabilidad, lealtad,...
- Ventas: Esquemas de comportamiento, hábitos de compra
- Finanzas: Inversiones, administración de cartera
- Bancos y Seguros: Aprobación de créditos y pólizas
- Seguridad: Detección de fraudes
- Medicina: Análisis de tratamientos
- Fabricación: control de calidad, adjudicación de recursos
- Internet: Análisis de clicks (Web mining)

Modelos de Data Mining

- Entrenamiento
  - Supervisado
  - No supervisado
- Prueba
- Evaluación

Modelos de Data Mining

PRUEBA

De los casos históricos disponibles se destina una cierta cantidad para *entrenar* el modelo y se reserva una porción de ellos para *probar* el modelo

Se presentan los casos como si fueran nuevos y se coteja la respuesta del modelo con los valores reales

Modelos de Data Mining

Matriz de confusión

Cantidad de casos **900**

Real \ Predicción	Sí	No
Sí	**455**	**29**
No	**32**	**384**

Matriz de confusión

Sobre un total de 900 casos el modelo predijo

455 como sí y en realidad era sí

384 como no y en realidad era no

839 predicciones correctas (93,2%)

El resto (6,8%) los predijo en forma incorrecta

PRECISION

Tecnicas del DATAMINING

- Regresión Lineal
  Método matemático, que crea un modelo entre la relación de las variables dependientes, las variables independiente y un término aleatorio.
- Redes Neuronales
  prototipo de aprendizaje y procesamiento automático, infundido netamente en la forma de trabajar del sistema nervioso animal
- Árbol de Decisión
  modelo de predicción, el cual construye diagramas de construcciones lógicas para representar y categorizar una serie de condiciones que ocurren de manera sucesiva, para la resolución de un problema.
- Algoritmo de Agrupamiento (clustering)
  Consiste en la agrupación de una serie de vectores de acuerdo a un criterio de cercanía, la cual se determina en términos de funciones de distancia o variables discretas.

Resumen

- Definición de Data Mining
- Proceso de Data Mining
- Modelos de Data Mining
- Técnicas y Herramientas de uso frecuente en Data Mining

Modelos de Data Mining

Se dividen en dos grandes grupos:

Descriptivos: Describen el comportamiento de los datos de manera que puedan ser interpretados por un experto.

Como trabajan → Identifican patrones por diversos métodos que explican o resumen los datos.

Modelos de Data Mining

Se dividen en dos grandes grupos:

Descriptivos: Describen el ….

Predictivos: además de describir se usan para predecir el valor de un nuevo caso.

Como trabajan → Estiman valores de variables de interés (a predecir) a partir de otras variables (predictoras).

Modelos de Data Mining

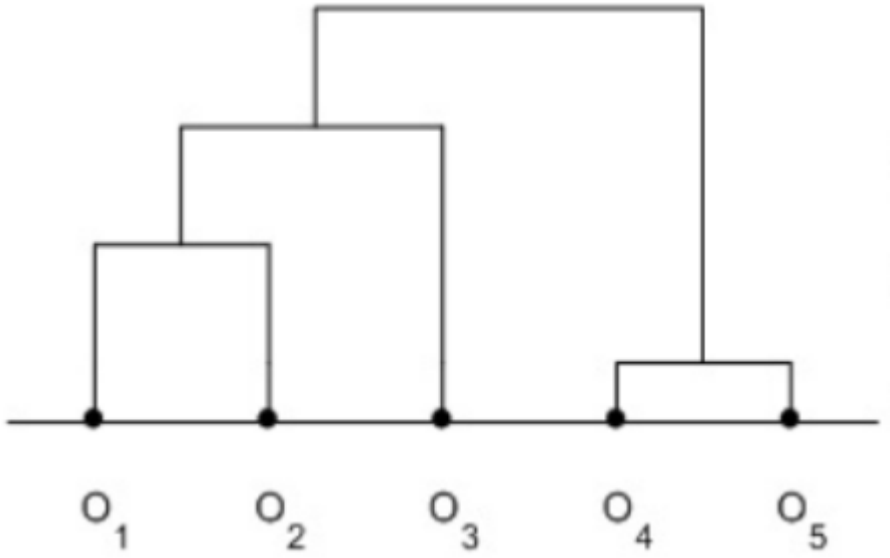
A su vez se subclasifican en:

- **Descriptivos:**
  - **Clustering:** Agrupamiento de casos homogéneos.
  - **Asociación:** Expresan patrones de comportamiento de los datos.
- **Predictivos**:
  - **Regresión:** Variable a predecir continua.
  - **Clasificación:** Variable a predecir discreta.

Clustering - Agrupamiento

Dados unos datos sin etiquetar, el objetivo es encontrar grupos naturales de instancias.

a) Particional



b) Jerárquico

Tipos de Clustering

Clustering particional
Partición de los objetos en grupos o clusters.
Todos los objetos pertenecen a alguno de los k clusters, los cuales son disjuntos.
Problema => elección de k

Clustering ascendente jerárquico
Crear un dendograma, es decir, crear un conjunto de agrupaciones anidadas hasta construir un árbol jerárquico.

Mas Complejo → Menos usado.

Caso de Estudio: Asociación

Problema

Dado un conjunto de transacciones, encontrar reglas que describen tendencias en los datos:

cuando la ocurrencia
culo esta asociada a la
de otros artículos en
transacción.

Caso Estudio: Asociación

TID	Artículos
1	Pan, leche, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, cerveza
4	Pan, leche, pañales, cerveza
5	Pan, leche, huevos, cerveza

Ejemplo Asociación – Análisis Ticket Compra Supermercado

TID	Artículos
1	Pan, leche, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, cerveza
4	Pan, leche, pañales, cerveza
5	Pan, leche, huevos, cerveza

TID	Artículos
1	A, B, C
2	A, D, E
3	B, D, E
4	A, B, D, E
5	A, B, C, E

ITEMS: A=Pan B=Leche C=Huevos D=Pañales E=Cerveza

Ejemplo de Asociacion

Convertimos a Binario.

TID	Artículos
1	A, B, C
2	A, D, E
3	B, D, E
4	A, B, D, E
5	A, B, C, E

TID	A	B	C	D	E
1	1	1	1	0	0
2	1	0	0	1	1
3	0	1	0	1	1
4	1	1	0	1	1
5	1	1	1	0	1

Se convierten los atributos en identificadores binarios, ej item=leche, --> se convierte como B con valores (1,0) presente o no.

Ej Asociación - **Definiciones**

TID	Artículos
1	Pan, leche, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, cerveza
4	Pan, leche, pañales, cerveza
5	Pan, leche, huevos, cerveza

- ItemSet: conjunto de uno o mas items
  - Por ej: {A,B,E} (orden sin importancia)

- Transacción: (TID,ItemSet) subconjunto de la transacción relevante para el proceso de asociación.
  - TID, es el identificador de la transacción.

Ej Asociación – **Soporte Itemset e ItemSet Frecuente**

- Soporte ItemSet k.
  - supp(K) = Numero de transacciones (t) que soportan (contienen) el conjunto de elementos k sobre total t.

Ejemplo de Estudio:

  - supp({A,B})=3/5
  - supp({B,C})=2/5

TID	Artículos
1	Pan, leche, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, cerveza
4	Pan, leche, pañales, cerveza
5	Pan, leche, huevos, cerveza

- ItemSet frecuente: itemset con Soporte mayor o igual a un umbral establecido por el usuario.
  - supp(I) >= minsup. Ej: supp({leche,pan}>0.10

Ej Asociación
Reglas de Asociación

- Transacciones

TID	Artículos
1	Pan, leche, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, cerveza
4	Pan, leche, pañales, cerveza
5	Pan, leche, huevos, cerveza

- Reglas de Asociación

{panales} → {cerveza}

{leche, pan} → {huevos}

{cerveza, pan}
→ {leche, huevos}

Expresión de la forma **X** → **Y**
Donde **X** e **Y** son itemsets.

!OJO! → implica co-ocurrencia, no causalidad.

Ej Asociación – **Medidas de Evaluación de Reglas**

Medidas de evaluación de las reglas de asociación

- **Soporte de la regla: supp (X → Y)**

  Fracción de las transacciones que contiene tanto a X como a Y.

- **Confianza de la regla: conf (X → Y)**

  Fracción de las transacciones en las que aparece X que también incluyen a Y; esto es, la confianza mide con qué frecuencia aparece Y en las transacciones que incluyen X.

Ej Asociación – **Medidas de Evaluación de Reglas**

TID	Artículos
1	Pan, leche, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, cerveza
4	Pan, leche, pañales, cerveza
5	Pan, leche, huevos, cerveza

Veamos en el ejemplo:

supp({pañales}) = 3/5 = 0.6

supp({cerveza}) = 4/5 = 0.8

supp({cerveza} → {pañales}) = supp({pañales,cerveza})
= 3/5 = 0.6 → **60 %**

conf({cerveza} → {pañales})
= supp({pañales},{cerveza}) / supp({cerveza})
= (3/5) / (4/5) = ¾ = 0.75 → **75 %**

Ej Asociación – **Aplicaciones (I)**

"Product placement": Colocación de productos en las gondolas

- Objetivo: Identificar artículos que muchos clientes compran conjuntamente.
- Solución: Procesar los datos de los tickets proporcionados.
- Ejemplo: Si un cliente compra pañales, es muy probable que compre cerveza

  No nos sorprendamos si vemos las cervezas colocadas al lado de los pañales en el súper.

Ej Asociación – **Aplicaciones (II)**

"Gestión de inventarios" Problema

Una empresa de reparación de lavarropas quiere anticipar la naturaleza de las reparaciones que realiza y mantener a sus vehículos equipados con las piezas que permitan reducir el números de visitas a casa de sus clientes.

Solución

Procesar los datos sobre herramientas y piezas utilizadas en reparaciones previas para descubrir patrones de co-ocurrencia.

Ej Asociación – **Aplicaciones (III)**

"Promociones y Ofertas" Problema

Si se identifica una regla: {impresora} → {toner}

Tóner en el consecuente

- Puede determinarse cómo incrementar venta toner.

Impresora en el antecedente

- Puede determinarse qué productos se verían afectados si dejamos de vender impresoras.

Impresora antecedente y tóner consecuente

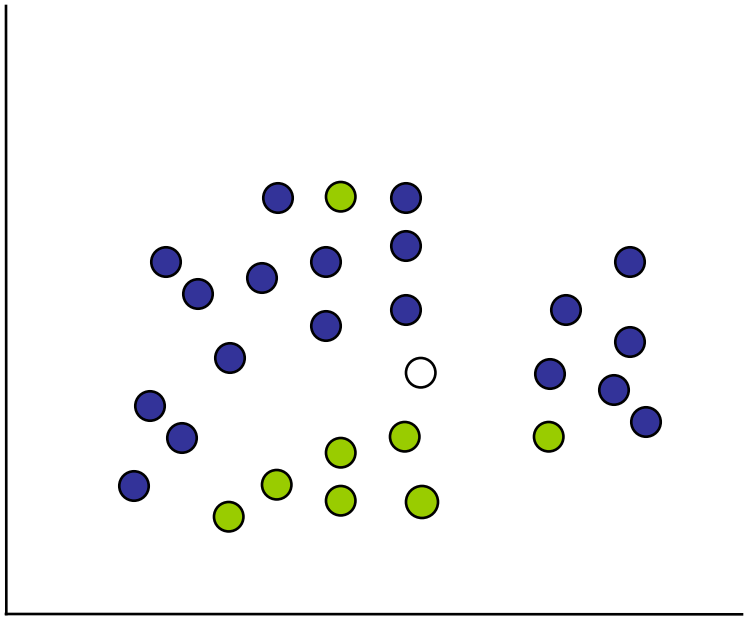
Puede utilizarse para ver qué productos deberían venderse con impresoras para promocionar las ventas de tóner.

Modelos Predictivos

- Clasificación
  - Predice un valor discreto
    - Sí / No
    - Alto / Mediano / Bajo
- Regresión
  - Predice un valor continuo
    - Importes
    - Cantidades

Clasificación

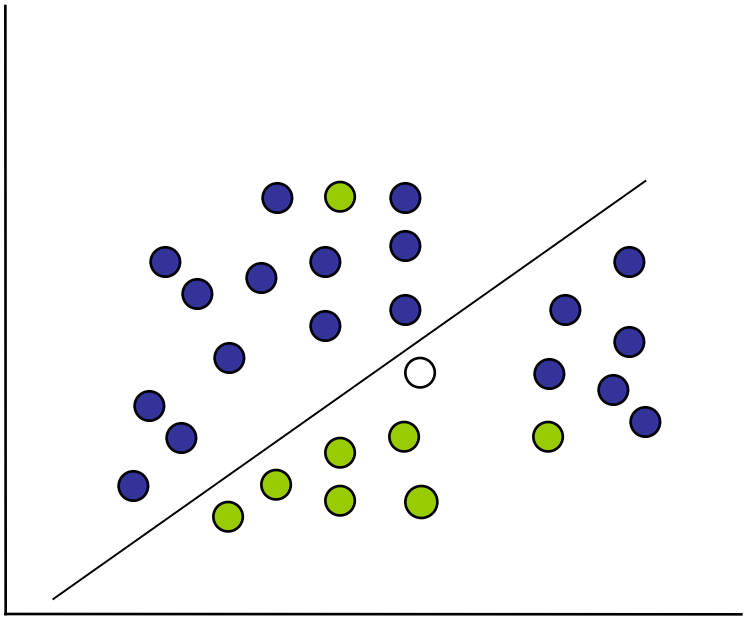
Métodos para clasificar objetos.



<u>Muchos Enfoques:</u>
Regresión.
Arboles de Decisión.
Bayesianos.
Redes Neuronales.
...

Dado un conjunto de puntos,
¿Cuál es la clase de un nuevo punto?

Clasificación: Regresión Lineal

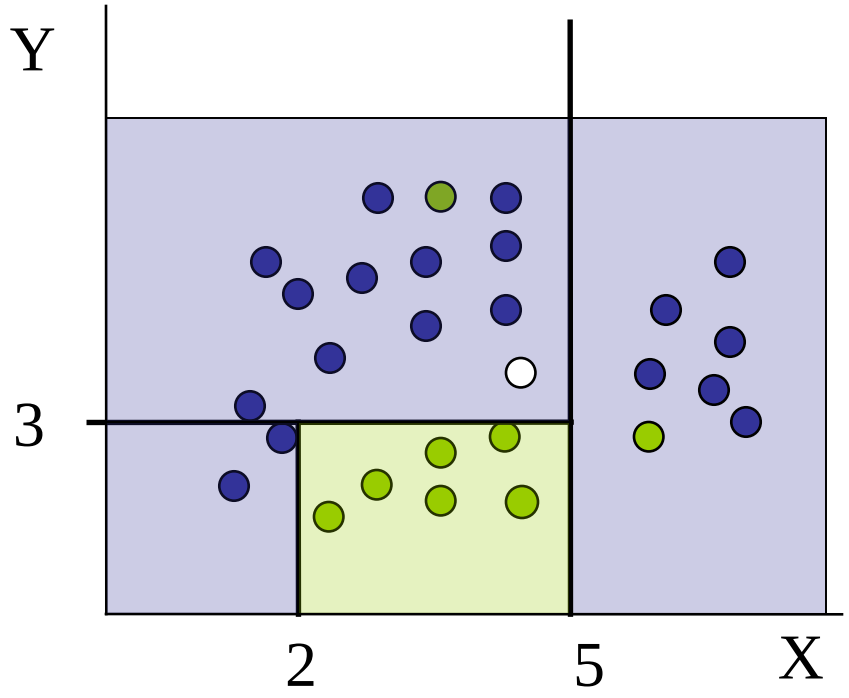


- Regresión lineal.

$$y = \beta_0 + \beta_1 x_1 + u$$

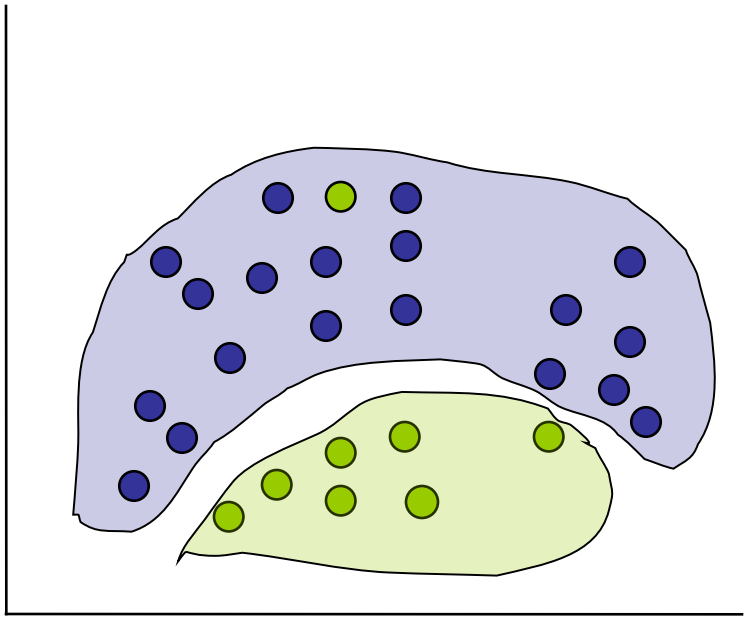
- Regresión
  calcula wi de los datos para minimizar el error cuadrado.
- Poco Flexible

Clasificación: Arboles de Decisión



If(X > 5) **then** blue
else if (Y > 3) **then** blue
else if (X > 2) **then** green
else blue

Clasificación: Redes Neuronales



- Pueden seleccionar regiones mas complejas
- Gralmente son mas precisos.
- Pueden sobre ajustar los datos (aprendizaje).

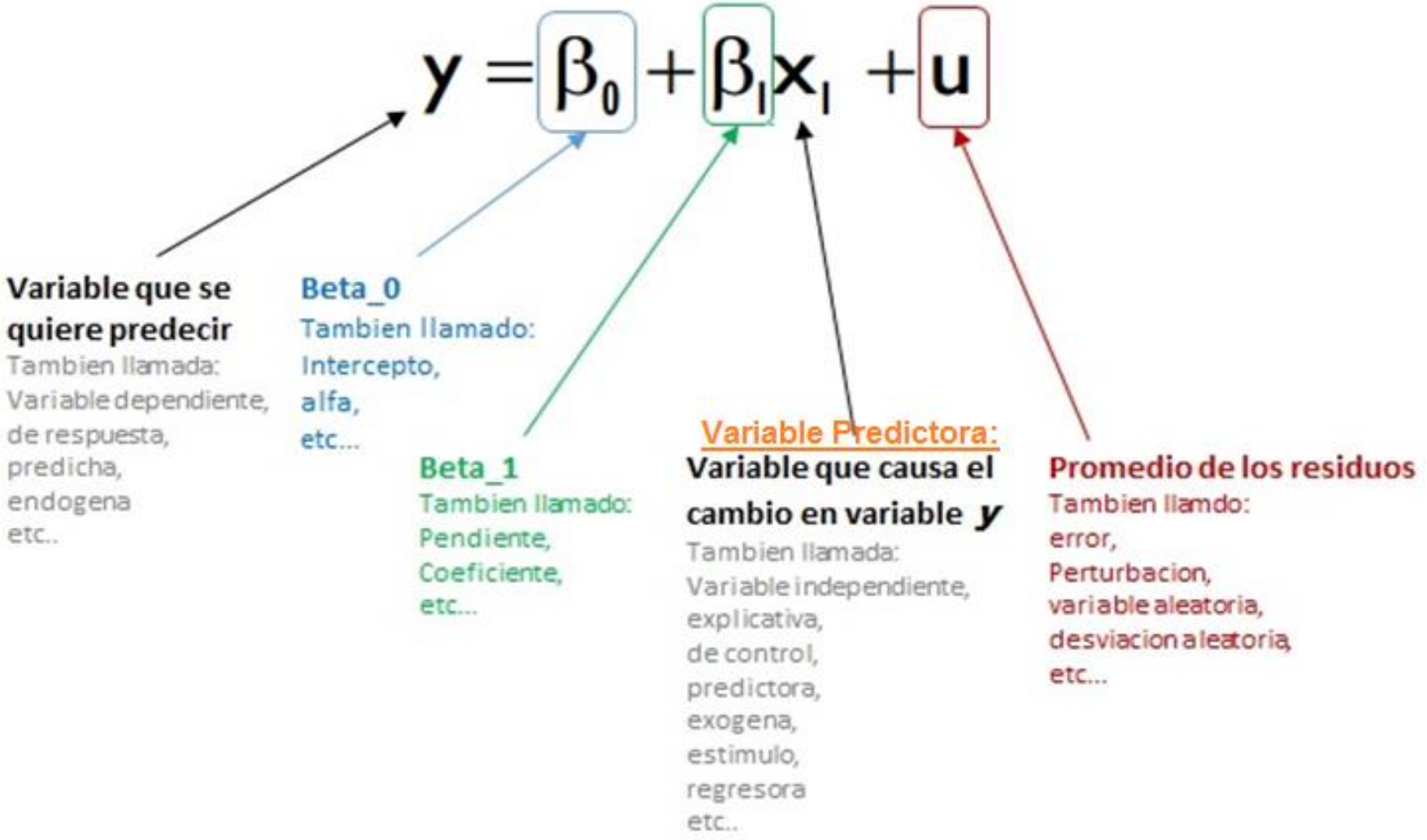
Resumen

- Definición de Data Mining
- Proceso de Data Mining
- Modelos de Data Mining
- Técnicas, otra mirada.

Regresión lineal

- El análisis de regresión es una técnica estadística para investigar la relación funcional entre dos o más variables, ajustando algún modelo matemático.
- La regresión lineal simple utiliza una sola variable de regresión y el caso más sencillo es el modelo de línea recta. Supóngase que se tiene un conjunto de n pares de observaciones (xi,yi), se busca encontrar una recta que describa de la mejor manera cada uno de esos pares observados.

Regresión lineal



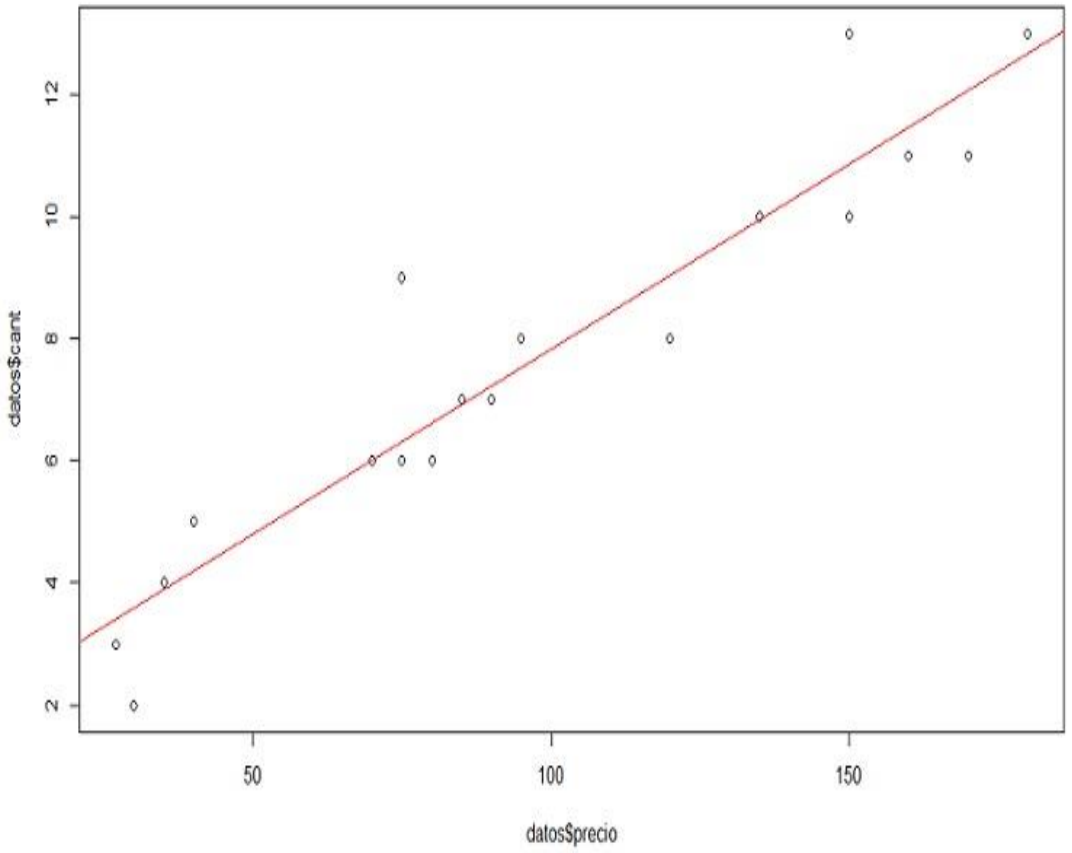
Regresión lineal

	DATOS		Calculos para REGRESION				
Nro	(cant)	(precio)	z1	z1 * (cant)	z1 * z1	predic_ŷ	((cant)-predic_ŷ)^2)
1	9	75	-18,60	-167	346	6,32	7,17
2	11	170	76,40	840	5837	12,08	1,16
3	6	70	-23,60	-142	557	6,02	0,00
4	4	35	-58,60	-234	3434	3,90	0,01
5	7	85	-8,60	-60	74	6,93	0,01
6	8	120	26,40	211	697	9,05	1,10
7	11	160	66,40	730	4409	11,47	0,22
8	6	70	-23,60	-142	557	6,02	0,00
9	5	40	-53,60	-268	2873	4,20	0,64
10	6	80	-13,60	-82	185	6,63	0,39
11	10	150	56,40	564	3181	10,87	0,75
12	8	95	1,40	11	2	7,53	0,22
13	2	30	-63,60	-127	4045	3,60	2,55
14	3	27	-66,60	-200	4436	3,41	0,17
15	7	90	-3,60	-25	13	7,23	0,05
16	13	180	86,40	1123	7465	12,68	0,10
17	10	135	41,40	414	1714	9,96	0,00
18	4	35	-58,60	-234	3434	3,90	0,01
19	6	75	-18,60	-112	346	6,32	0,10
20	13	150	56,40	733	3181	10,87	4,55
	7,45	**93,60**		**2835**	**46785**		**19,21**

error estandar del residuo	**1,03**
Beta_1	**0,06**
Intercepto	**1,78**
R (coef. Correl)	0,95
R cuadrado	0,90
Total Observaciones	20
R cuadrado ajustado	0,89

El sig ej usa la regresión lineal para predecir la cantidad de compra(o demanda) según el precio.

Regresión lineal



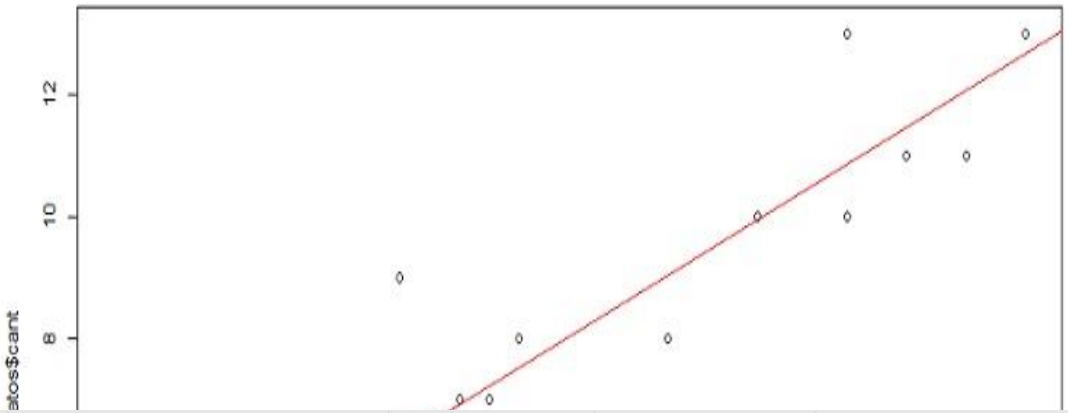
(cant) = 1.78 + 0.06 (precio)+1.03

error estandar del residuo	**1,03**
Beta_1	**0,06**
Intercepto	**1,78**
R (coef. Correl)	0,95
R cuadrado	0,90
Total Observaciones	20
R cuadrado ajustado	0,89

Nota: Modelo desarrollado con sofware r

El sig ej usa la regresión lineal para predecir la cantidad de compra(o demanda) según el precio.

Con Excel - Regresión lineal



(cant) = 1.78 + 0.06 (precio)+1.03

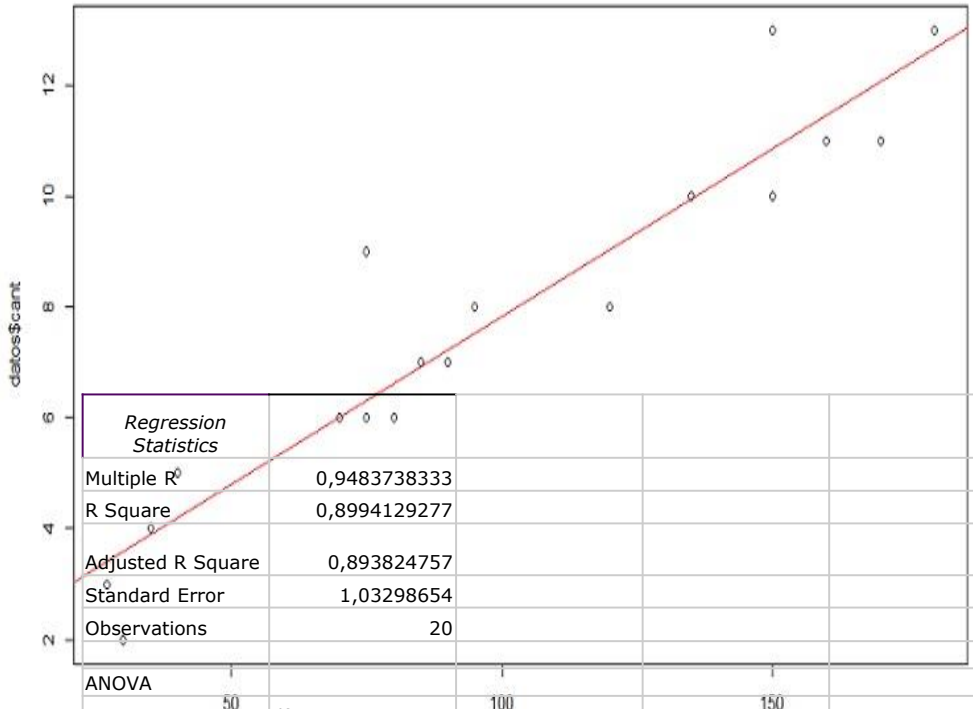
De "R"

error estandar del residuo	**1,03**
Beta_1	**0,06**
Intercepto	**1,78**
R (coef. Correl)	0,95
R cuadrado	0,90
Total Observaciones	20
R cuadrado ajustado	0,89

Con Excel: Datos --> Analisis de Datos --> Regresión

	A	B	C	D	E	F	G	H	I
1	Resumen								
2									
3	*Estadísticas de la regresión*								
4	Coeficiente de correlación múltiple	0.948373833							
5	Coeficiente de determinación R^2	0.899412928							
6	R^2 ajustado	0.893824757							
7	Error típico	1.03298654							
8	Observaciones	20							
9									
10	ANÁLISIS DE VARIANZA								
11		*Grados de libertad*	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*			
12	Regresión	1	171.7428985	171.7428985	160.949437	0.00000000021			
13	Residuos	18	19.20710145	1.067061192					
14	Total	19	190.95						
15									
16		*Coeficientes*	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*nferior 95.0%*	*uperior 95.0%*
17	Intercepción	1.778958123	0.503161816	3.535558674	0.00236256	0.721854374	2.83606187	0.72185437	2.83606187
18	Variable X 1	0.060588054	0.004775758	12.68658494	2.0501E-10	0.05055456	0.07062155	0.05055456	0.07062155

Con Google - Regresión lineal



(cant) = 1.78 + 0.06 (precio)+1.03

De "R"

error estandar del residuo	**1,03**
Beta_1	**0,06**
Intercepto	**1,78**
R (coef. Correl)	0,95
R cuadrado	0,90
Total Observaciones	20
R cuadrado ajustado	0,89

Regression Statistics	
Multiple R	0,9483738333
R Square	0,8994129277
Adjusted R Square	0,893824757
Standard Error	1,03298654
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	171,7428985	171,7428985	160,9494375	0,0000000002050104531
Residual	18	19,20710145	1,067061192		
Total	19	190,95			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95%*	*Upper 95%*
Intercept	1,778958123	0,503161816	3,535558674	0,002362558276	0,7218543869	2,836061859	0,7218543869	2,836061859
X Variable 1	0,06058805424	0,004775757584	12,68658494	0,0000000002050104821	0,05055456	0,07062154848	0,05055456	0,07062154848

Regresión Lineal

- Se considera que la variable X es la variable independiente o regresiva y se mide sin error, mientras que Y es la variable respuesta para cada valor específico xi de X; y además Y es una variable aleatoria con alguna función de densidad para cada nivel de X.
- Se obtiene el diagrama de dispersión

Regresión lineal

- El objetivo del análisis de la regresión lineal es analizar un modelo que pretende explicar el comportamiento de una variable (Variable endógena, explicada o dependiente), que denotaremos por **Y,** utilizando la información proporcionada por los valores tomados por un conjunto de variables (explicativas, exógenas o independientes), que denotaremos por **X1 , X2 , ….., X n**

Regresión Lineal - Conclusión

El análisis de regresión sirve tanto para EXPLORAR datos como para CONFIRMAR teorías.

Redes neuronales

- 1) Una nueva forma de computación, inspirada en modelos biológicos.
- 2) Un modelo matemático compuesto por un gran número de elementos procesales organizados en niveles.
- 3) ...un sistema de computación compuesto por un gran número de elementos simples, elementos de procesos muy interconectados, los cuales procesan información por medio de su estado dinámico como respuesta a entradas externas.
- 4) Redes neuronales artificiales son redes interconectadas masivamente en paralelo de elementos simples (usualmente adaptativos) y con organización jerárquica, las cuales intentan interactuar con los objetos del mundo real del mismo modo que lo hace el sistema nervioso biológico.

Redes Neuronales

Son capaces de:

- aprender de la experiencia,
- generalización de casos anteriores a nuevos casos,
- abstraer características esenciales a partir de entradas que representan información irrelevante

Método

- Arboles de Decisión

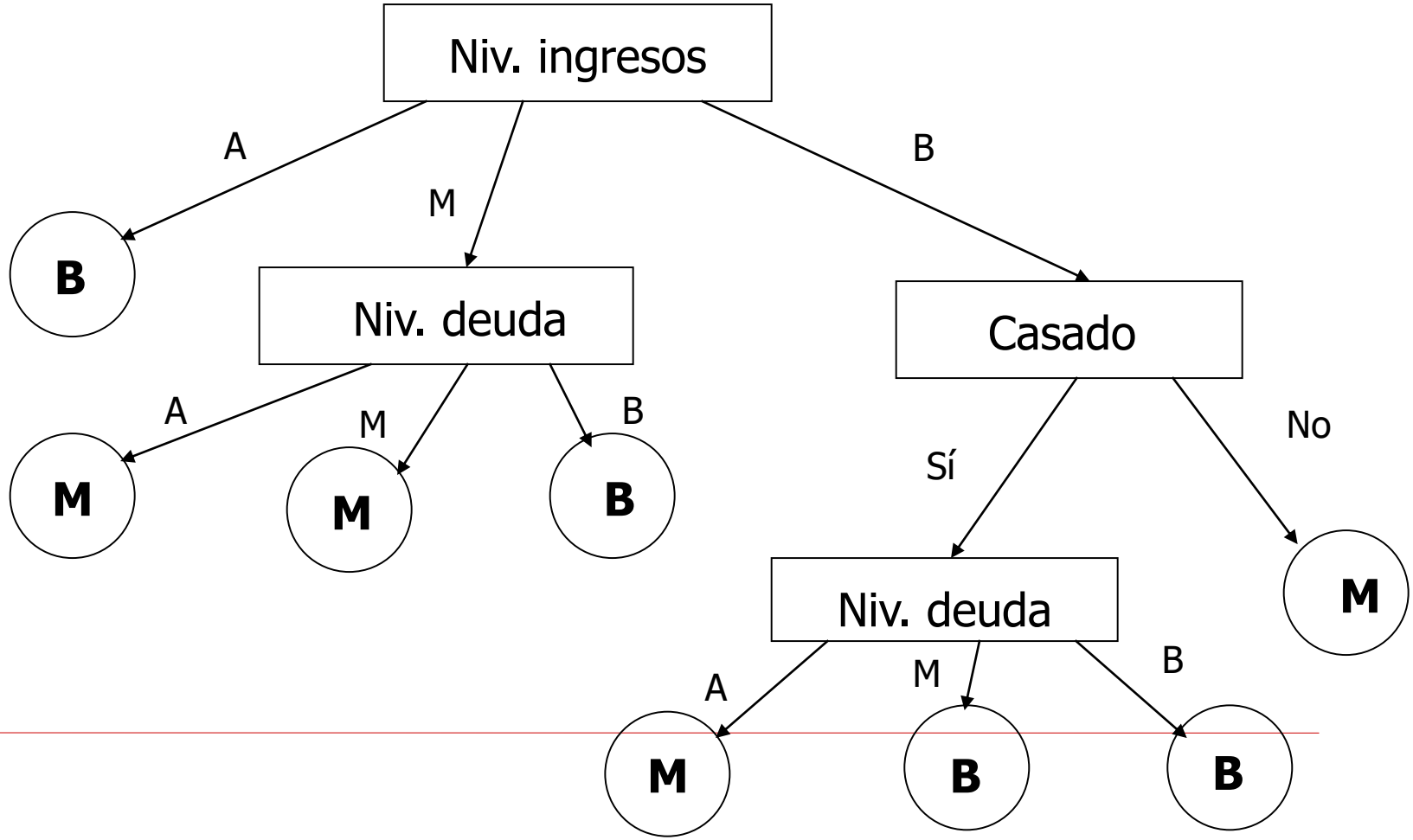
Arbol de decisión

- Un árbol de decisión es una forma gráfica y analítica de representar todos los eventos (sucesos) que pueden surgir a partir de una decisión asumida en cierto momento.
- Nos ayudan a tomar la decisión más "acertada", desde un punto de vista probabilístico, ante un abanico de posibles decisiones. Estos árboles permiten examinar los resultados y determinar visualmente cómo fluye el modelo. Los resultados visuales ayudan a buscar subgrupos específicos y relaciones que tal vez no encontraríamos con estadísticos más tradicionales.

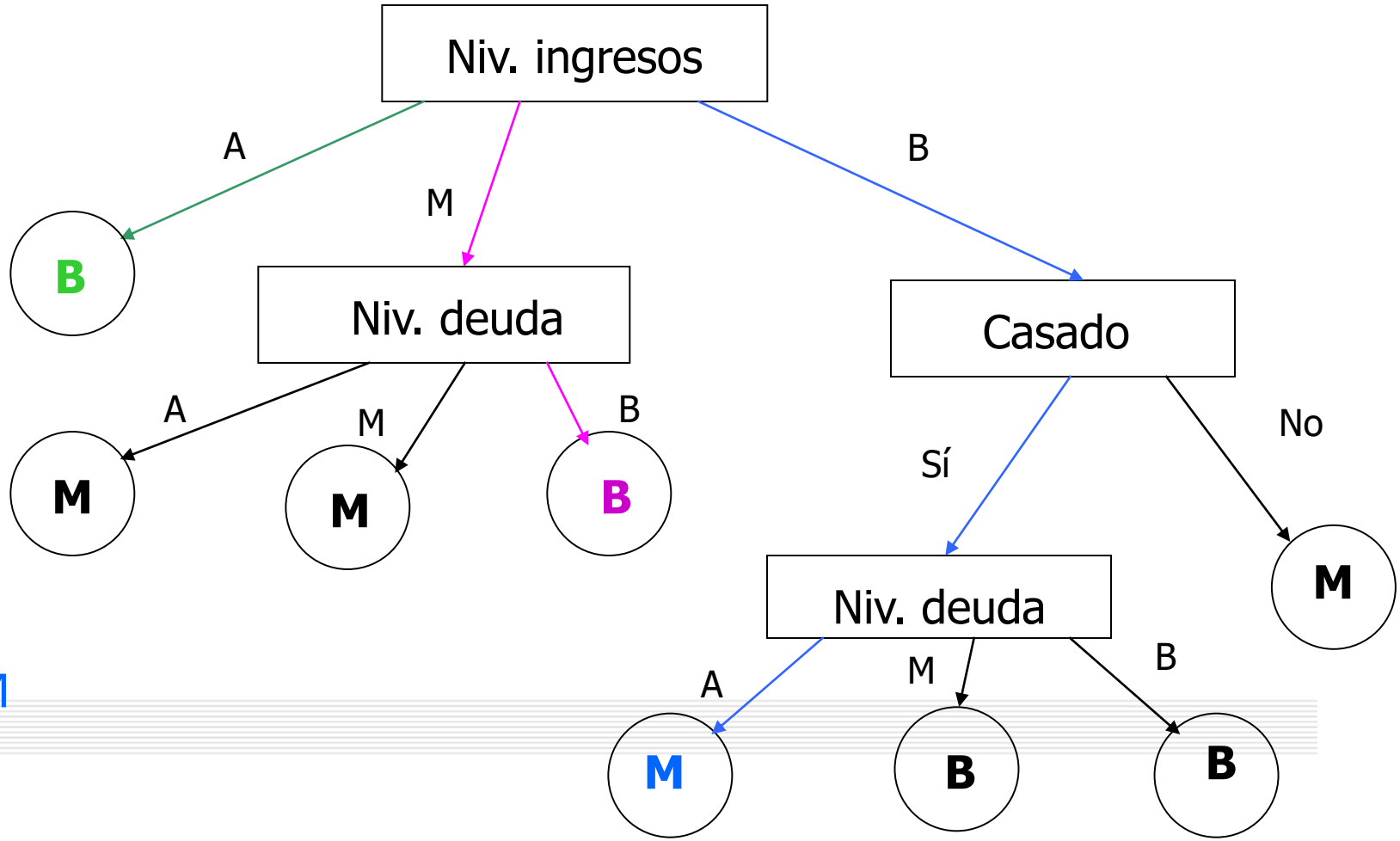
Arbol de decisión

- Los árboles de decisión son una técnica estadística para la segmentación, la estratificación, la predicción, la reducción de datos y el filtrado de variables, la identificación de interacciones, la fusión de categorías y la discretización de variables continuas.
- La función árboles de decisión permite crear árboles de clasificación y de decisión para identificar grupos y predecir evento futuros

Árbol de decisión
Ejemplo



Árbol de decisión
Ejemplo Clasif Tomas,Lucia ...



Tomás B

Lucía B

Horacio M

Árboles de Clasificación

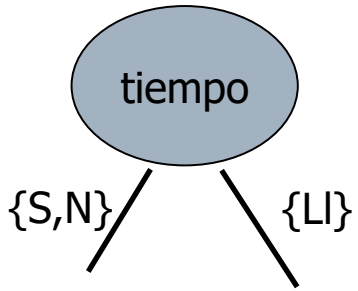
- **Nodo raíz**
  - no tiene entradas, y tiene una o más salidas
- **Nodo interno**
  - tiene una sola entrada, y dos o más salidas
- **Nodo terminal**
  - tiene una sola entrada, y no tiene salidas
- Cada nodo terminal es asignado a una clase
- Los nodos no terminales contienen **condiciones de test de atributos** (**CTA**) que permiten dividir el conjunto de registros en sub-conjuntos con diferentes características

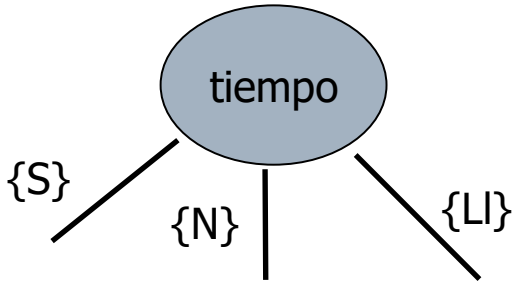
Construcción de un árbol de Clasificación (Inducción)

- Características
  - Número exponencial de alternativas
  - Algunos son más exactos que otros
  - No es posible asegurar el óptimo

Características de las CTA

- Atributo binario:
  - genera dos posibles resultados
- Atributo nominal:
  - puede tener varios valores:
    - División múltiple
    - División binaria sucesiva

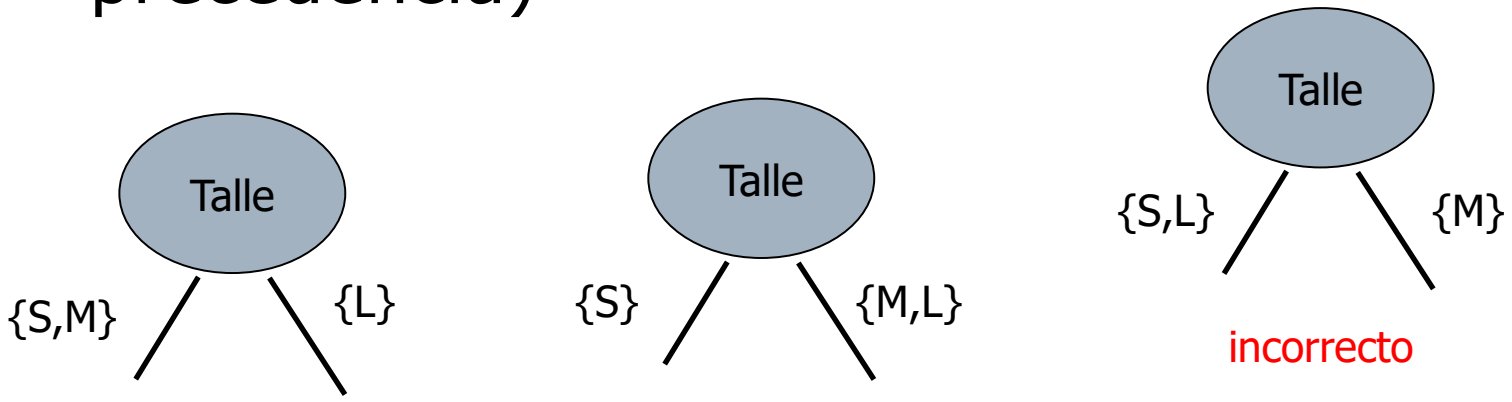




Clases: **S**oleado, **N**ublado, **Ll**uvia

Características de las CTA

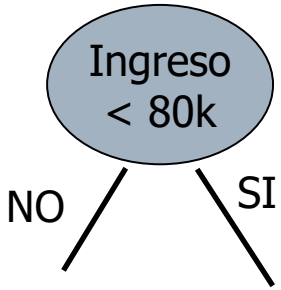
- Atributo ordinal:
  - puede usarse división múltiple o binaria (en este caso mantener la relación de precedencia)

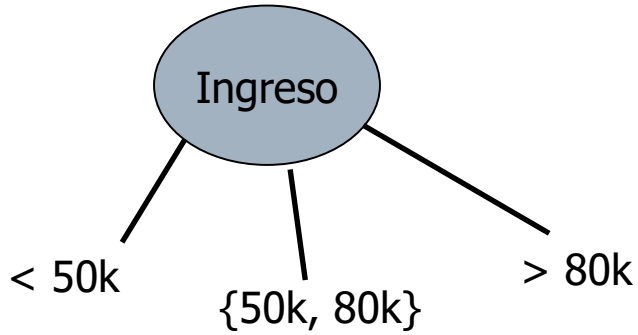


Clases: **S**mall, **M**edio, **L**argo

Características de las CTA

- Atributo continuo:
  - Puede ser expresado como un test de comparación con:
    - División binaria: $(A < v) \vee (A \geq v)$
    - División múltiple $v_i \leq A < v_{i+1} \ \forall \ i=1,\ldots,k$



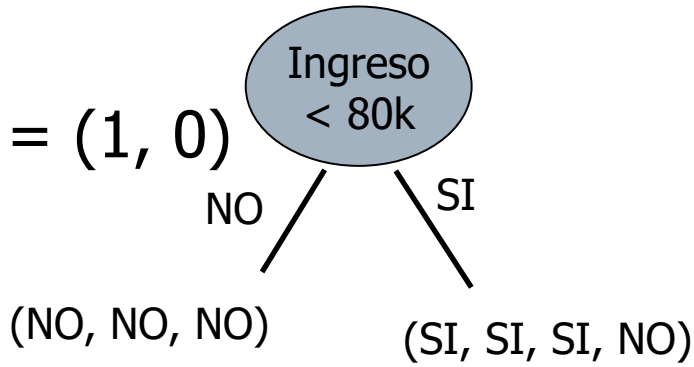


Métricas para elegir la división

- Sea p(i/t) fracción de registros de la clase i en un dado nodo t (también pi)
- Ejemplo: en un caso de 2 clases, la distribución de clases en un nodo puede ser expresada como (p_0, p_1) donde $p_1 = 1 - p_0$

- (1,0) o (0,1) máxima pureza
- (0.5, 0.5) mínima pureza

- Nodo hijo 1: (p_{NO}, p_{SI}) = (3/3, 0/3) = (1, 0)
- Nodo hijo 2 (p_{NO}, p_{SI}) = (1/4, 3/4)



Métricas para elegir la división

Entropía(t) = $-\sum p(i/t) * \text{Log}_2\ p(i/t)\ \forall\ i=0,\ldots,C-1$

Error de clasificación(t)= $1- \max_i [p(i/t)]$

Gini(t)=$1-\sum [p(i/t)]^{**}2\ \forall\ i=0,\ldots,C-1$

C: número de casos

Criterios para medir la performance de una división

Ganancia G

Es usable con la métrica **Error de Clasificación** o para el caso de división binaria

Razón de Ganancia

RG = G / Métrica

Es usable con todas las métricas y divisiones binarias o múltiples.

Ganancia G

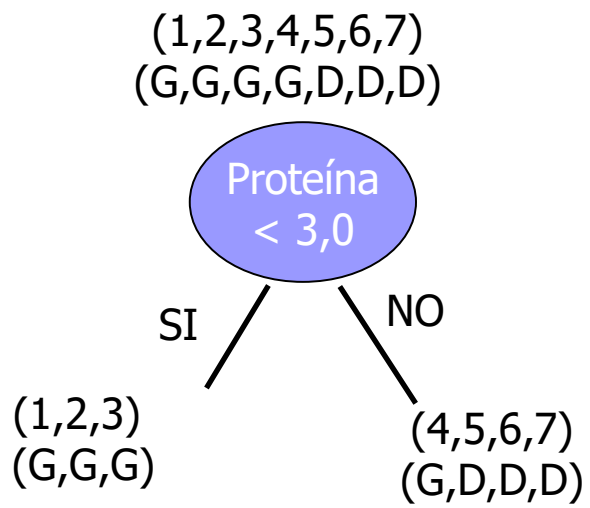
- Criterio para medir la performance de una división

$$\mathbf{G = I(padre) - \sum N(v_j)/N * I(v_j)} \ \forall \ \mathbf{j=1,\ldots,k}$$

- I() = medida de impureza
- N= número de registros del nodo padre
- k = número de valores del atributo
- $N(v_j)$ = número de registros del nodo hijo v_j

Ejemplo uso métrica:
Error de clasificación (EC)

Alimento	Proteína	Lípidos	Categoría
1	1,3	3,1	G
2	1,9	4	G
3	1,6	2,9	G
4	3,1	3,2	G
5	3	1,6	D
6	4,0	1,8	D
7	4,1	2,0	D



Atributo: proteína <3

Nodo padre:

N=7

I(p) = 1-max(3/7,4/7) = 3/7

Nodo Hijo 1:

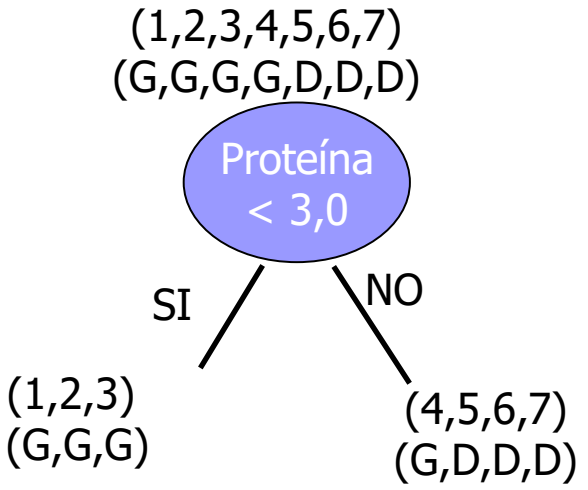
N(1)=3

I(1) = 1-max(0/3,3/3) = 0

Nodo Hijo 2:

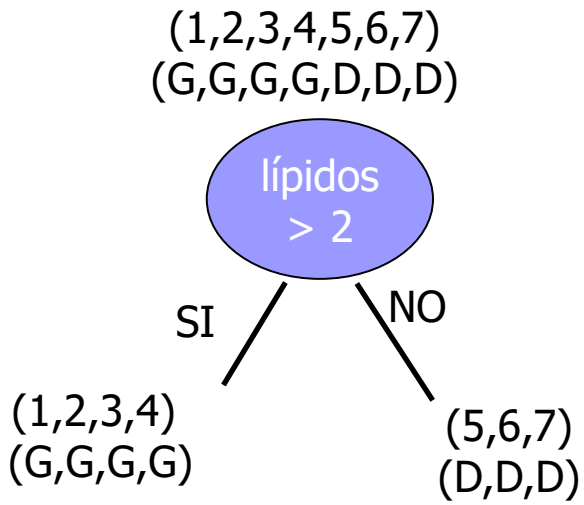
N(2)=4

I(2) = 1-max(3/4,1/4) = 1/4



- G = I(padre) - $\sum$ N(vj)/N*I(vj) $\forall$ j=1,…,k
- G = 3/7 – 3/7*0 – 4/7*1/4 = 3/7-1/7 = 2/7 =

Alimento	Proteína	Lípidos	Categoría
1	1,3	3,1	G
2	1,9	4	G
3	1,6	2,9	G
4	3,1	3,2	G
5	3	1,6	D
6	4,0	1,8	D
7	4,1	2,0	D



Atributo: lípidos >2

Nodo padre:

N=7

I(p) = 1-max(3/7,4/7) = 3/7

Nodo Hijo 1:

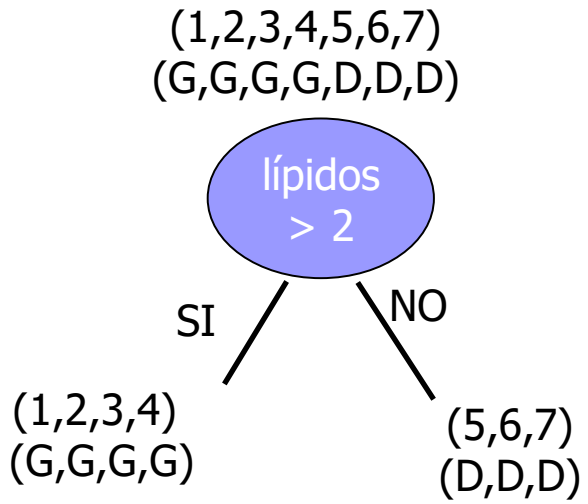
N(1)=4

I(1) = 1-max(0/4,4/4) = 0

Nodo Hijo 2:

N(2)=3

I(2) = 1-max(3/3,0/3) = 0



- G = I(padre) - $\sum$ N(vj)/N*I(vj) $\forall$ j=1,…,k
- G = 3/7 – 4/7*0 – 3/7*0 = 3/7 = **0.43**
- Luego, es conveniente usar como CTA= lípido >2

Agenda

- Árboles de Clasificación
- Algoritmo de Hunt
- Métodos para la CTA
  - Métricas para elegir la división
  - Criterios para medir la performance de una división
- Errores de un Modelo

Errores de un Modelo

Error de entrenamiento

Es producido por los registros del conjunto de entrenamiento que resultan mal clasificados

Error entrenamiento =

$$=[\sum e(ti)\ \forall\ i=1,.k]\ /\ [\sum n(ti)\ \forall\ i=1,.k]$$

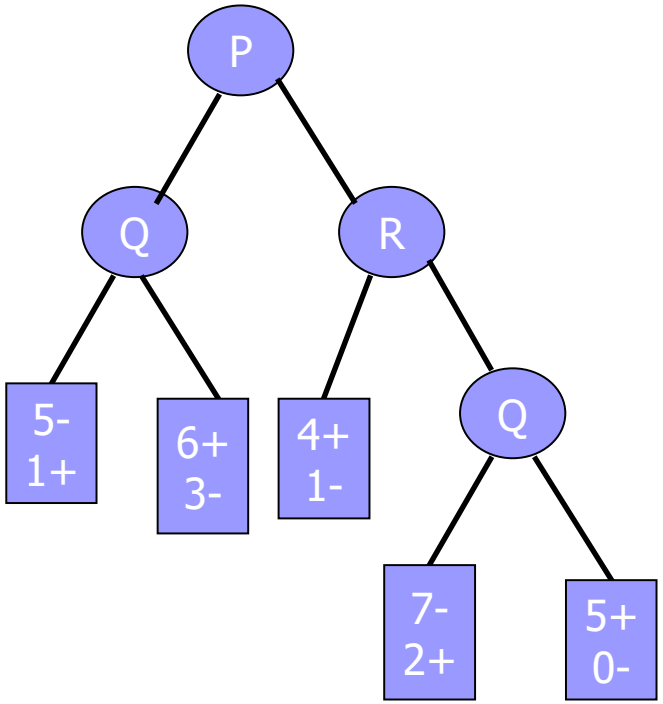
ti: nodo terminal

k: cantidad de nodos terminales

e(ti): registros de entrenamiento mal clasificados en ti

n(ti): total de registros de entrenamiento clasificado en ti

Ejemplo



k: 5
Nodo terminal 1:
e(t_1):1
n(t_1):6
Nodo terminal 2:
e(t2):3
n(t2):9
Nodo terminal 3:
e(t3):1
n(t3):5
Nodo terminal 4:
e(t4):2
n(t4):9
Nodo terminal 5:
e(t5):0
n(t5):5

EE=(1+3+1+2+0)/(6+9+5+9+5)=0,21

Errores de un Modelo

Error de generalización

- Es producido por los registros del conjunto de validación que resultan mal clasificados

- Un buen modelo debe tener bajo error de entrenamiento y de generalización

Errores de un Modelo

Sobre-entrenamiento

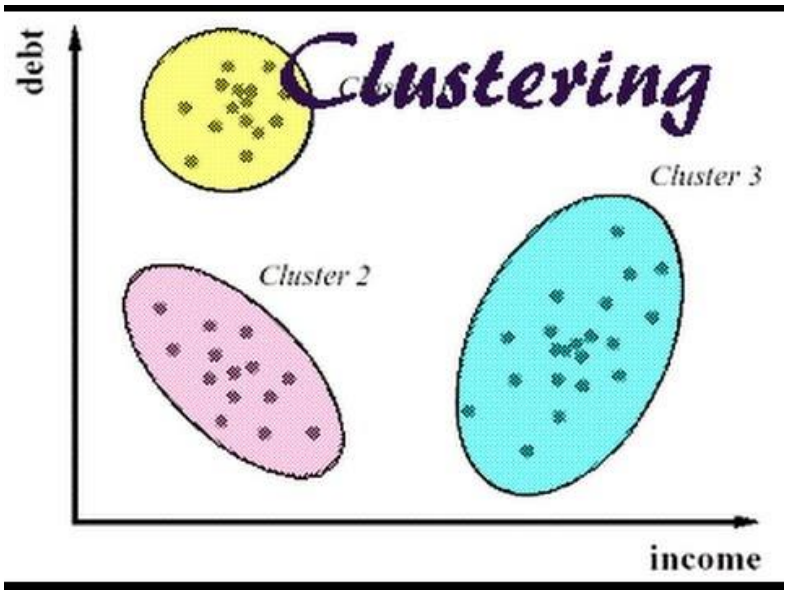
Cuando un modelo tiene un error de entrenamiento muy bajo y un alto error de generalización

Causas posible

- Presencia de ruidos (errores en los registros)
- Falta de registros representativos en el conjunto de entrenamiento
- Otros

Método

- Clustering



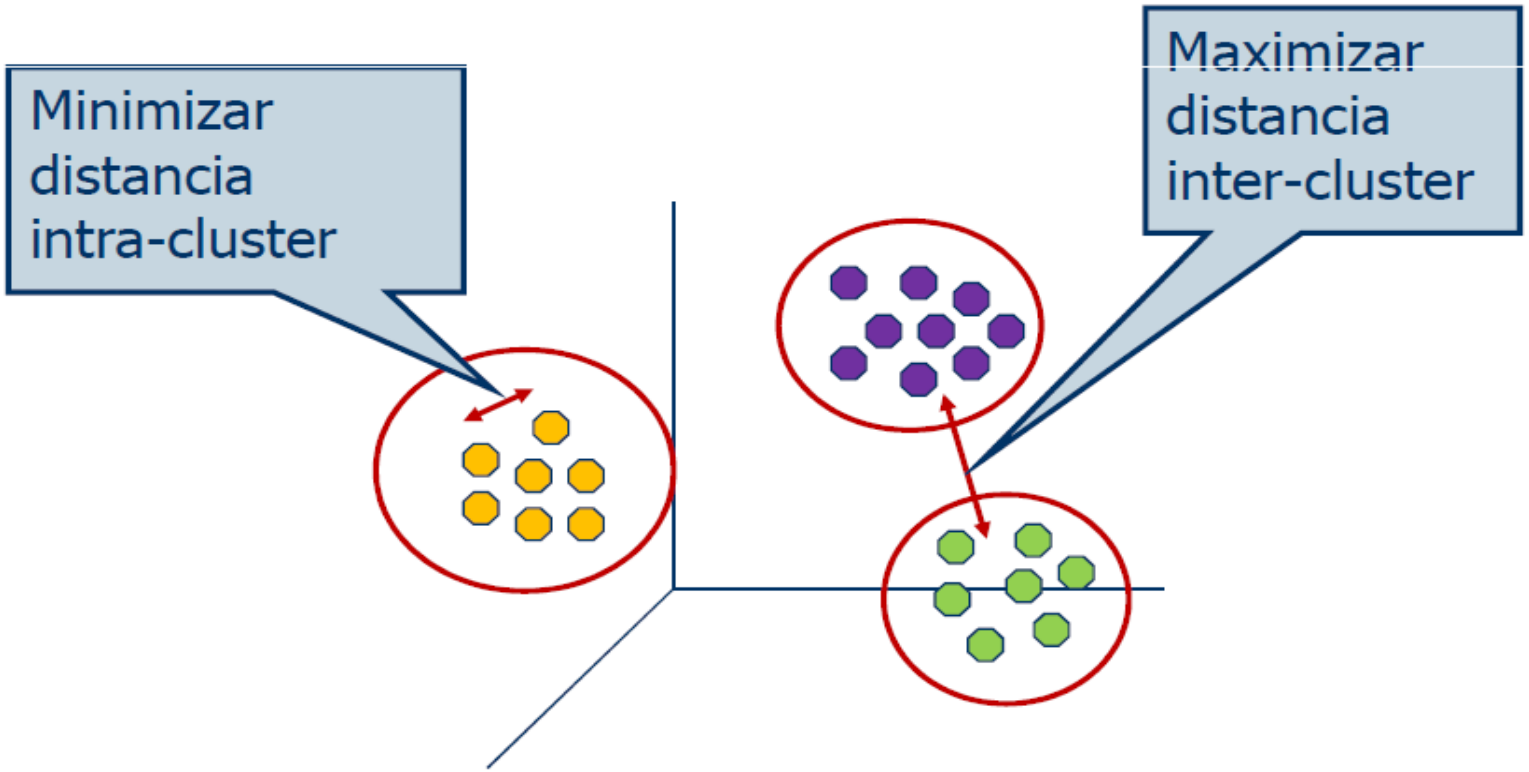
Clustering

- Se basa en intentar responder como es que ciertos **Objetos** (casos) pertenecen o **"caen"** naturalmente en cierto número de **clases** o grupos, de tal manera que estos objetos comparten ciertas características.
- Esta definición asume que los objetos pueden dividirse, razonablemente, en grupos que contienen objetos similares. Si tal división existe, ésta puede estar oculta y debe ser descubierta.

  Este es el **objetivo principal** de las técnicas o estudios de clustering.

Clustering -**Objetivo**

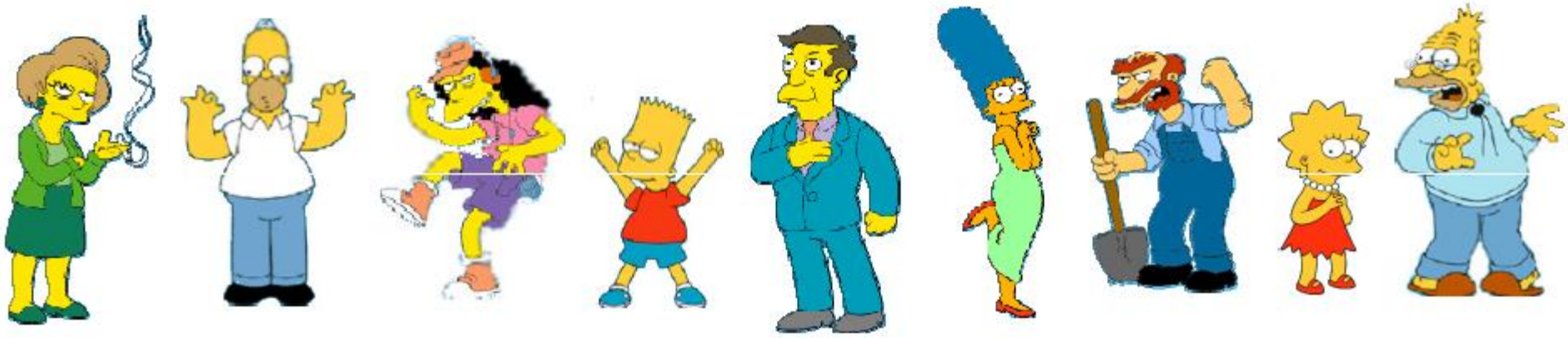
- Encontrar agrupamientos de tal forma que los objetos de un grupo sean similares entre sí y diferentes de los objetos de otros [clusters].



Clustering -**Caracteristica**

- **Aprendizaje no Supervisado:**
  No existen clases (cluster) Predefinidas.

¿Que grupos o clases se nos ocurren?



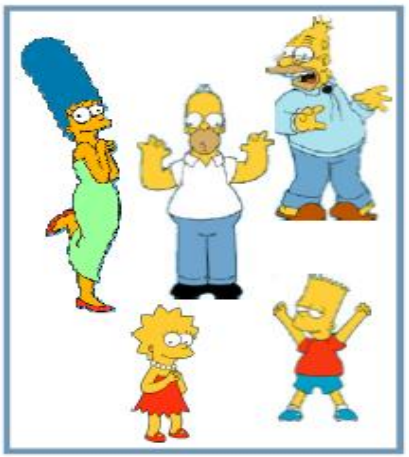
Clustering -**Caracteristica**

Aprendizaje no Supervisado:

No existen clases (cluster) Predefinidas.

¿Mujeres Vs Hombres?

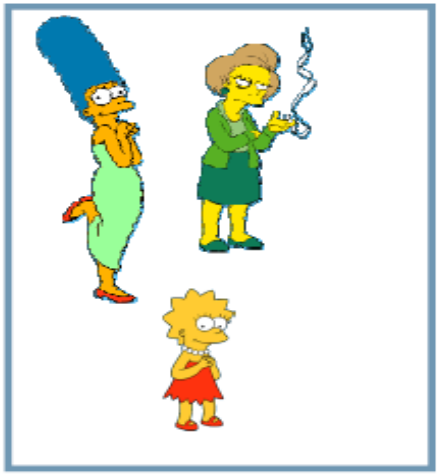
¿Familia Simpson Vs Empleados Escuela?



Clustering -**Caracteristicas**

- **Los Resultados del Algoritmo dependerán:**
  - Del algoritmo de agrupamiento seleccionado.
  - Del conjunto de datos disponible (variables elegidas).
  - La medida de similitud utilizada para comparar objetos (usualmente, definida como medida de distancia).

¿Mujeres Vs Hombres?



Clustering
Metodos Calculo Distancia

Existen diversos métodos para medir la distancia:

- Dist. Minkowski

$$d_r(x,y) = \left(\sum_{j=1}^{J} |x_j - y_j|^r\right)^{\frac{1}{r}}, \quad r \geq 1$$

- Dist Manhattan (city block)

$$d_1(x,y) = \sum_{j=1}^{J} |x_j - y_j|$$

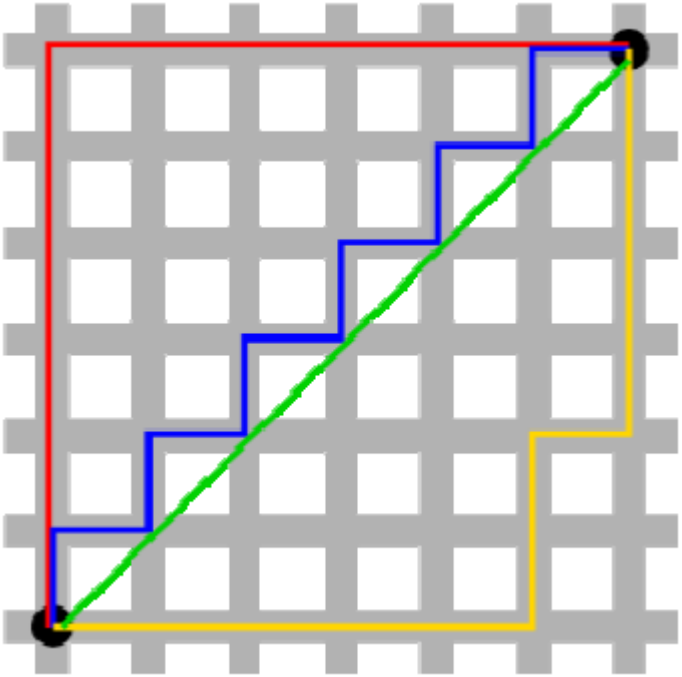
- Dist Euclídea (Pitágoras)

$$d_2(x,y) = \sqrt{\sum_{j=1}^{J} (x_j - y_j)^2}$$

- Distancia Chebyshev (rey ajedrez)

$$d_\infty(x,y) = \text{máx}_{j=1..J} |x_j - y_j|$$

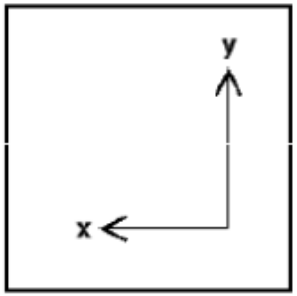
Clustering
Distancias



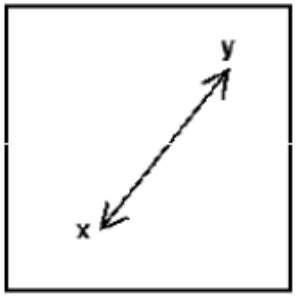
Representación de solo 2 Variables independientes, Una por eje.

Cada método dará un resultado distinto.:

- Dist Manhattan =12 (roja, azul o amarilla)
- Dist Euclídea = 8.5 (verde)
- Distancia Chebyshev = ¿6?



Manhattan



Euclidean

Clustering

Aplicaciones

- Reconocimiento de formas. (OCR)
- Mapas temáticos (GIS)
- Marketing: (Segmentación de clientes)
- Clasificación de documentos (Text Mining)
- Análisis de web logs, patrones de acceso similares (Análisis de Vulnerabilidades)

Nota: Muy usado también como paso previo a otra técnica de Minería.

Ej: (Clustering y luego k-vecinos)

Clustering

- Un Objeto es un dato, el cual esta formado por un conjunto finito de variables.
- Variables:
  - Numéricas: son números reales en general
  - Nominales : Son variables discretas pero que no tienen un orden especificado (estado civil)
  - Ordinales: Son variables discretas con una relación de orden (temp. Alta, Media, baja)
  - Binarias: solo pueden tomar dos estados posibles (dicotómicas)

Método

- K-vecinos

K-vecinos

- También llamada **CBR** (Case Based Reasoning, **Razonamiento basado en casos)**
  - Resuelve un problema tomando en cuenta casos parecidos
  - Función de vecindad o de distancia
  - Función de combinación

K-vecinos

- El modelo de los K-vecinos no tiene fase de entrenamiento
- Entra directamente en la fase de producción
- K indica la cantidad de casos parecidos (vecinos) que se van a considerar
- En este caso vamos a tomar K = 3

K-vecinos

Función de vecindad

Analizamos un caso de 3 variables:

- para nivel de **ingresos** y nivel de **deudas**
  - 0 si son iguales
  - 1 si uno tiene Alto y el otro Medio
  - 1 si uno tiene Medio y el otro Bajo
  - 2 si uno tiene Alto y el otro Bajo
- para **casado**
  - 0 si son iguales
  - 1 si son distintos

K-vecinos

Función de vecindad

- Para cada caso a resolver se confronta con todos los casos testigo
- Se suman los 3 valores
- Se eligen los 3 (K) casos testigo que tienen el menor valor de esta función

K-vecinos

Lote para determinación de vecinos

	Resultado	N_Ingresos	N_Deuda	Casado
Paola	Bueno	Alto	Bajo	Si
Andrea	Bueno	Alto	Medio	No
Jorge	Bueno	Alto	Alto	Si
Débora	Bueno	Medio	Bajo	No
Román	Bueno	Medio	Medio	Si
Carlos	Malo	Medio	Alto	No
Gala	Malo	Bajo	Bajo	No
Vanesa	Bueno	Bajo	Medio	Si
Sergio	Malo	Bajo	Alto	No
Mario	Malo	Bajo	Alto	Si

K-vecinos

	Tomás	Lucía	Horacio
Jorge	0 2 1 = 3	1 2 0 = 3	2 0 0 = 2
Carlos	1 2 0 = 3	0 2 1 = 3	1 0 1 = 2
Andrea	**0 1 0 = 1**	1 1 1 = 3	2 1 1 = 4
Débora	**1 0 0 = 1**	**0 0 1 = 1**	1 2 1 = 4
Sergio	2 2 0 = 4	1 2 1 = 4	**0 0 1 = 1**
Vanesa	2 1 1 = 4	1 1 0 = 2	**0 1 0 = 1**
Mario	2 2 1 = 5	1 2 0 = 3	**0 0 0 = 0**
Gala	2 0 0 = 2	1 0 1 = 2	0 2 1 = 3
Paola	**0 0 1 = 1**	**1 0 0 = 1**	2 2 0 = 4
Román	1 1 1 = 3	**0 1 0 = 1**	1 1 0 = 2

Casos Nuevos a Predecir

K-Vecinos

- Tomás tiene como *vecinos* a
  - Andrea (1), Débora (1), Paola (1)
- Lucía tiene a
  - Débora (1), Paola (1), Román (1)
- Y Horacio a
  - Sergio (1), Vanesa (1), Mario (0)

K-Vecinos

Función de combinación

- Vamos a tomar como valor de la predicción sobre la calificación aquella que corresponda a la mayoría de los vecinos.
- Ejemplos
  - Vecinos: B B M - Predicción: B
  - Vecinos: M B M - Predicción: M

K-Vecinos

PREDICCIONES

- Andrea B, Débora B, Paola B
  - Predicción para **Tomás**→ **Buena**
- Débora B, Paola B, Román B
  - Predicción para **Lucía** → **Buena**
- Sergio M, Vanesa B, Mario M
  - Predicción para **Horacio**→ **Mala**

K-Vecinos

Actividad: Prediga los casos para k=5.

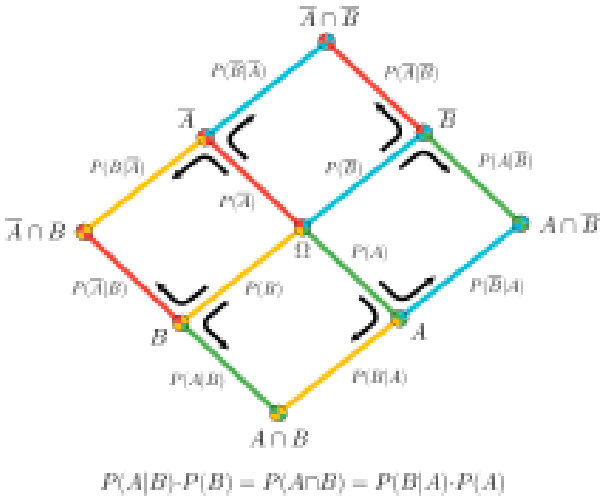
Anto NI=A, ND=A, C=N

Manu NI=B, ND=M, C=N

Xabier NI=M, ND=M C=NC (optar por posición conservadora).

Método

- Modelo Bayes



Clasificadores Bayes

Es un modelo probabilístico de clasificación

Constituyen una muy buena opción alternativa a los Arboles de Decisión.

Como Construir un Clasificador Bayes

- Suponga que desea predecir la salida Y que tiene grado o aridad n y valores v_1, v_2, ... v_n.

  Por ej: Y=clasificación crediticia → v1=B, v2=M

- Supongamos que hay m atributos de entrada llamados X_1, X_2, ... X_m.
- Divida los datos en n_Y conjuntos de datos (**D**ata **S**et) más pequeños llamados DS_1, DS_2,... DS_{n_Y}.
- Definir DS_i = Registros en los que Y = v_i (concreto)
- Para cada DSi, aprende el Estimador de Densidad Mi para modelar la distribución de entrada entre los registros Y=vi. Mi estima P (X1, X2, ... Xm | Y = vi)

Como Construir un Clasificador Bayes

- Suponga que desea predecir la salida Y que tiene grado o aridad n y valores v_1, v_2, ... v_n.

  Por ej: Y=clasificación crediticia → v1=B, v2=M

Datos del problema

- De los 10 casos hay 6 con calificación **B**uena y 4 con calificación **M**ala.

- Sin saber nada más, la probabilidad *a priori* de que la calificación sea B es 0,6 y de que sea M es 0,4

Nota: La información sobre:

Nivel de Ingresos Nivel de Deudas y casado apunta a calcular la probabilidad *a posteriori* de que sea calificado crediticiamente como B o M.

Como Construir un Clasificador Bayes

- Suponga que desea predecir la salida Y con aridad 2.

  Por ej: Y=clasificación crediticia → v1=B, v2=M

- Supongamos que hay m atributos de entrada llamados X_1, X_2, ... X_m.

  Por ej: X1=Nivel Ingreso, X2=Nivel Deuda, X3=EstCivil
- Divida los datos en n_y conjuntos de datos (**D**ata **S**et) más pequeños llamados DS_1, DS_2,... DS_{n_y}.
- Definir DS_i = Registros en los que Y = v_i
- Para cada DSi, aprenda el Estimador de Densidad Mi para modelar la distribución de entrada entre los registros Y=vi. Mi estima P (X1, X2, ... Xm | Y = vi)

Ver próxima pagina.

Como Construir un Clasificador Bayes

Lote de Entrenamiento

	Resultado	N_Ingresos	N_Deuda	Casado
Paola	Bueno	Alto	Bajo	Si
Andrea	Bueno	Alto	Medio	No
Jorge	Bueno	Alto	Alto	Si
Débora	Bueno	Medio	Bajo	No
Román	Bueno	Medio	Medio	Si
Carlos	Malo	Medio	Alto	No
Gala	Malo	Bajo	Bajo	No
Vanesa	Bueno	Bajo	Medio	Si
Sergio	Malo	Bajo	Alto	No
Mario	Malo	Bajo	Alto	Si

Como Construir un Clasificador Bayes

- **Nivel de Ingresos**
  - De los que tienen Nivel de Ingresos **A**lto:
    - hay 3 **B**uenos
    - 0 **M**alos.
  - De los que tienen Nivel de Ingresos **M**edio:
    - hay 2 **B**uenos
    - 1 **M**alo.
  - De los que tienen Nivel de Ingresos **B**ajo:
    - hay 1 **B**ueno
    - 3 **M**alo.

Como Construir un Clasificador Bayes

- De los 6 Buenos hay 3 que tienen Nivel del Ingresos Alto. Luego

  P(B/Ingresos A) = 3/6 = 0.5

- De los 6 Buenos hay 2 que tienen Nivel del Ingresos Medio.

  P(B/Ingresos M) = 2/6 = 0.33

- De los 6 Buenos hay 1 que tiene Nivel del Ingresos Bueno.

  P(B/Ingresos B) = 1/6 = 0.17

Como Construir un Clasificador Bayes

De la misma manera, de los 4 Malos y Nivel de Ingreso

- P(M/Ingresos A) = 0
- P(M/Ingresos M) = 0.25
- P(M/Ingresos B) = 0.75

Como Construir un Clasificador Bayes

Nivel de Deudas

- De los que tienen Nivel de Deudas A hay 1 B y 3 M.
- De los que tienen Nivel de Deudas M hay 3 B y 0 M.
- De los que tienen Nivel de Deudas B hay 2 B y 1 M.

Como Construir un Clasificador Bayes

Calculamos las probabilidades de la misma forma que en el caso anterior

- P(B/Deudas A) = 0.17
- P(B/Deudas M) = 0.50
- P(B/Deudas B) = 0.33
- P(M/Deudas A) = 0.75
- P(M/Deudas M) = 0
- P(M/Deudas B) = 0.25

Como Construir un Clasificador Bayes

Casado

- De los que tienen Casado sí hay 4 B y 1 M
- De los que tienen Casado no hay 2 B y 3 M

Con lo que

- P(B/Casado sí) = 0.67
- P(B/Casado no) = 0.33
- P(M(Casado sí) = 0.25
- P(M/Casado no) = 0.75

Como Construir un Clasificador Bayes

		Frecuencias		Probabilidades condicionales	
		Bueno	Malo	Bueno	Malo
		6	4	0,60	0,40
Nivel de ingresos	Alto	3	0	0,50	0
	Mediano	2	1	0,33	0,25
	Bajo	1	3	0,17	0,75

Como Construir un Clasificador Bayes

		Frecuencias		Probabilidades condicionales	
		Bueno	Malo	Bueno	Malo
		6	4	0,60	0,40
Nivel de Deudas	Alto	1	3	0,17	0,75
	Mediano	3	0	0,50	0
	Bajo	2	1	0,33	0,25
Casado	Sí	4	1	0,67	0,25
	No	2	3	0,33	0,75

Clasificación Caso

Se desea clasificar a **Tomas** quien posee:

- Nivel de ingresos **Alto**,
- Nivel de deuda **Bajo**
- **Soltero**.

¿Bueno o Malo?

Producción: Usamos el Clasificador B.

Caso Tomás tiene

- Nivel de Ingresos **A**
- Nivel de Deudas **B**
- Casado **no**

La probabilidad *a posteriori* de que Tomás tenga una calificación **B**uena sale del producto de

- P(B) = 0.6
- P(B/Ingresos A) = 0.5
- P(B/Deudas B) = 0.33
- P(B/Casado no) = 0.33

Esta probabilidad resulta 0.6x0.5x0.33x0.33= **0.03267**

Producción: Usamos el Clasificador B.

La probabilidad a posteriori de que Tomás tenga una calificación **M**ala sale del producto de

- P(M) = 0.4
- P(M/Ingresos A) = 0
- P(M/Deudas B) = 0.25
- P(M/Casado no) = 0.75

Esta probabilidad resulta 0.4 x 0 x 0.25 x 0.75= **0**

Como la probabilidad de tener calificación Buena es mayor que la de tener Mala, resulta que

El modelo predice que Tomás califica como Buena

Clasificación de Otro Caso

Se desea clasificar a **Horacio** quien posee:

- Nivel de ingresos **Bajo**
- Nivel de deuda **Alto**
- **Casado**.

¿Bueno o Malo?

Producción: Usamos el Clasificador B.

Caso Horacio tiene

- Nivel de Ingresos **B**
- Nivel de Deudas **A**
- Casado **sí**

La probabilidad *a posteriori* de que Horacio tenga una calificación **B**uena sale del producto de

- P(B) = 0.6
- P(B/Ingresos B) = 0.17
- P(B/Deudas A) = 0.17
- P(B/Casado si) = 0.67

Esta probabilidad resulta 0.6x0.17x0.17x0.67= **0.0116**

Producción: Usamos el Clasificador B.

La probabilidad a posteriori de que Horacio tenga una calificación **M**ala sale del producto de

- P(M) = 0.4
- P(M/Ingresos B) = 0.75
- P(M/Deudas A) = 0.75
- P(M/Casado si) = 0.25

Esta probabilidad resulta 0.4 x 0.75 x 0.75 x 0.25 = **0.0562**

Como la probabilidad de tener calificación Buena es menor que la de tener Mala, resulta que

El modelo predice que Horacio califica como Mala

Clasificación de Otro Caso

Se desea clasificar a **Lucia** quien posee un nivel de ingresos **Medio**, un nivel de deuda **Baja** y es **Soltera**.

¿Bueno o Malo?

Bayes

En base a estos datos, Lucía tiene

- probabilidad de calificación B igual a 0,6 x 0,33 x 0,33 x 0,67 = 0,04378
- probabilidad de calificación M igual a 0,4 x 0,25 x 0,25 x 0,25 = 0,00625

Bayes

Como la probabilidad de tener calificación B es mayor que la de tener M, resulta que

El modelo predice que Lucía va a tener calificación Bueno

Tablas de Métodos

Nombre	DESCRIPTIVO		
	Agrupamiento	Reglas de asociación	Correlaciones/Factorizaciones
Redes neuronales	X		
Árboles de decisión ID3, C4.5, C5.0			
Árboles de decisión CART			
Otros árboles de decisión	X	X	
Redes de Kohonen	X		
Regresión lineal y logarítmica			X
Regresión logística		X	
Kmeans	X		
Apriori		X	
Naive Bayes			
Vecimos más próximos	X		
Análisis factorial y de comp. principales			X
Twostep, Cobweb	X		
Algoritmos genéticos y evolutivos	X	X	X
Máquinas de vectores soporte	X		
CN2 rules (cobertura)		X	
Análisis discriminante multivariante			

Resumen

- Definición de Data Mining
- Proceso de Data Mining
- Modelos de Data Mining
- Inconvenientes del Data Mining
- Potenciales aplicaciones
- Funciones de Data Mining
- Relación DM y KDD con la BDA

Importante

- La promesa de Data Mining es encontrar los patrones
- Simplemente el hallazgo de los patrones no es suficiente
- Debemos ser capaces de entender los patrones, responder a ellos, actuar sobre ellos, para finalmente convertir los datos en información, la información en acción y la acción en valor para la empresa.

Esto no es un tema trivial.

Data Mining y KDD

- Trabajo conjunto
  - Negocios (Expertos del Dominio)
  - Especialistas en Estadísticas
  - Especialistas de Sistemas

Incumbencia de BD

Problemática

- Encontrar Problemas puntuales
- Analizar datos al nivel más detallado
- No aplicar un solo enfoque
- No hay una sola solución

Aclaraciones respecto materia BDA

- El OLTP y las copias históricas proporcionan los datos. En menor medida el Datawarehouse.
- La inteligencia permitirá buscar en OLTP y OLAP tratando de encontrar patrones, descubrir reglas, nuevas ideas que probar, y hacer predicciones acerca del futuro.
- EL estudio profundo de las técnicas y herramientas que añaden la "inteligencia" para explotar los datos de los clientes y sacar el máximo rendimiento escapan a los objetivos de esta materia.
- Vemos solo los conceptos de esta disciplina.

Software de Data Mining

- **KNIME** http://www.knime.org/
- **RapidMiner** http://rapidminer.com/
- **Weka** www.cs.waikato.ac.nz/ml/weka/
- **R** http://www.r-project.org/
- **SPSS Modeler**
  http://www.spss.com/software/modeler/
- **SAS Enterprise Min** http://www.sas.com/
- **Power BI y Google Document** tb aportan.

Bibliografía

Libro: Introduction to Data Mining
Autor: Pang-Ning Tan /M. Steinbach / Vipin Kumar:
Editorial: Addison-Wesley

Libro: Data Mining - Concepts and Techniques
Autor: Jiawei Han / Micheline Kamber
Editorial: Morgan Kaufmann Publishers