Ingeniería de Sistemas

Minería de Datos

SISTEMAS DE INTELIGENCIA EMPRESARIAL

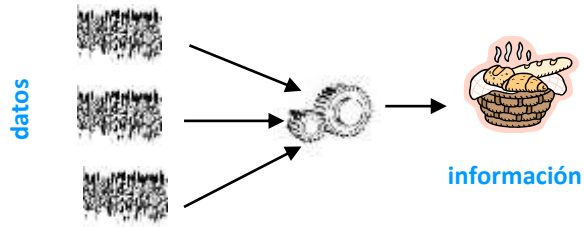
2020-1 | 10 MAYO

UNIVERSIDAD DE LIMA

Business Intelligence

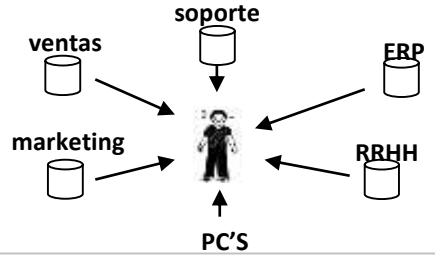
MUCHOS DATOS PERO POCA INFORMACIÓN

- **Hay que procesar los datos para que se conviertan en información útil**



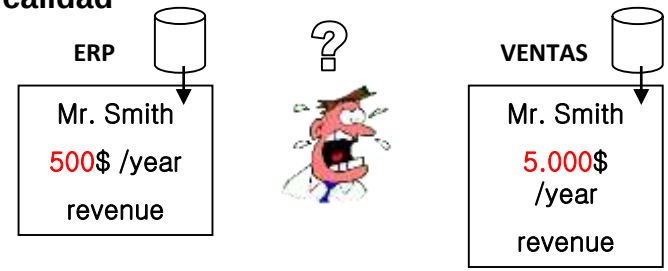
MÚLTIPLES FUENTES DE DATOS

- **La agregación de las diferentes fuentes facilita el acceso estructurado a la información.**

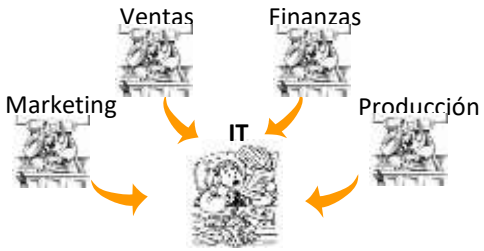


BAJA CALIDAD DE LOS DATOS

- **La fiabilidad de la información depende de su calidad**



SATURACIÓN DEL DEPARTAMENTO DE I.T.



- **Es importante descentralizar la gestión de la información, facilitando el acceso a los usuarios.**

Las empresas se enfrentan al reto de transformar los datos almacenados en una valiosa fuente de información para la toma de decisiones.

Business Intelligence: etapas

01

ANÁLISIS DE NEGOCIO Y ANÁLISIS TECNOLÓGICO

- Definición y análisis de los objetivos de negocio del proyecto.
- Definición de las necesidades tecnológicas asociadas a los objetivos de negocio.

02

ANÁLISIS DE NEGOCIO Y ANÁLISIS TECNOLÓGICO

- Integración y normalización de las diferentes fuentes de datos involucrados en el proyecto.
- Datawarehouse estructurado para el acceso óptimo a la información.

03

IMPLANTACIÓN DE PROCESO "OLAP" DE ANÁLISIS

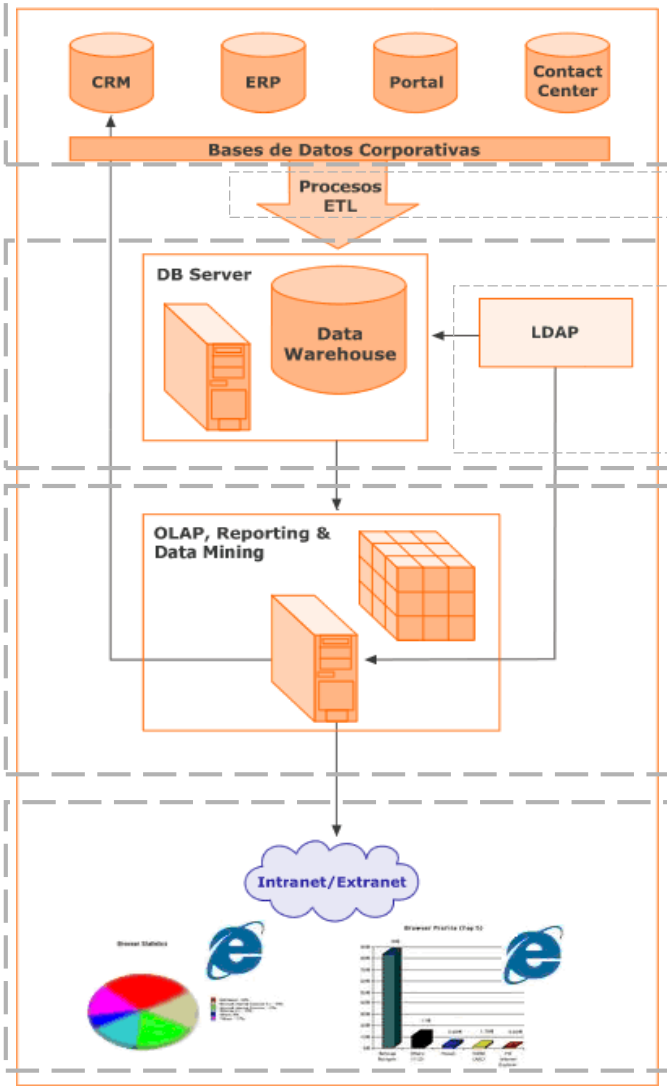
- Creación de una estructura multidimensional para el acceso a la información.
- Posibilidad de realizar consultas de datos e informes a diferentes niveles de detalle.

04

DESARROLLO DE UN MODELO DE DATAMINING

- Extracción de tendencias ocultas en grandes volúmenes de datos.
- Explicaciones de las tendencias descubiertas en función de parámetros accionables de negocio.

Business Intelligence: arquitectura



Fuentes de datos corporativos.

Extracción, Transformación y carga de los datos

Integración del sistema de B.I. con el entorno de seguridad corporativo

Almacenamiento masivo de datos estructurado para soportar consultas multidimensionales ágilmente.

Generación y publicación de análisis para los informes.

Acceso global vía web a la información según definición de perfiles de usuario.

Minería de datos

Proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos (Witten and Frank, 2000).

Aproximación

Una visión simplificada de la minería de datos

Se obtienen de:

- Bases de datos (relacionales, espaciales, temporales, documentales, multimedia, etc)
- World Wide Web

Son el producto de la minería de datos...

...y dan soporte a las estrategias de decisión que se tomen

Tipos de aprendizaje

Modelos descriptivos o aprendizaje no supervisado

Modelos predictivos o aprendizaje supervisado

Identifican patrones que explican o resumen datos

- Reglas de asociación: expresan patrones de comportamiento en los datos.
- Clustering: agrupación de casos homogéneos.

Estiman valores de variables de interés a partir de valores de otras variables.

- Regresión: Variable a predecir continua
- Clasificación supervisada: Variable a predecir discreta.

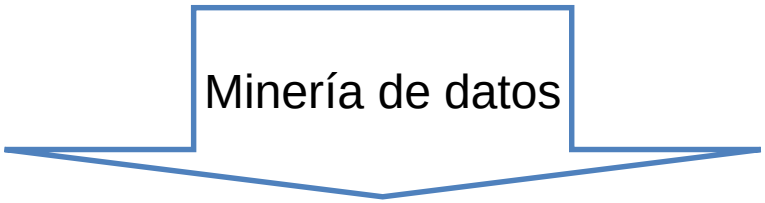
Ejemplo

Agente comercial: ¿Debo conceder una hipoteca a un cliente?

Datos:

cId	Credit-p (years)	Credit-a (euros)	Salary (euros)	Own House	Defaulter accounts	...	Returns-credit
101	15	60.000	2.200	yes	2	...	no
102	2	30.000	3.500	yes	0	...	yes
103	9	9.000	1.700	yes	1	...	no
104	15	18.000	1.900	no	0	...	yes
105	10	24.000	2.100	no	0	...	no
...	...	...	...	...	...	...	...

Modelo generado:

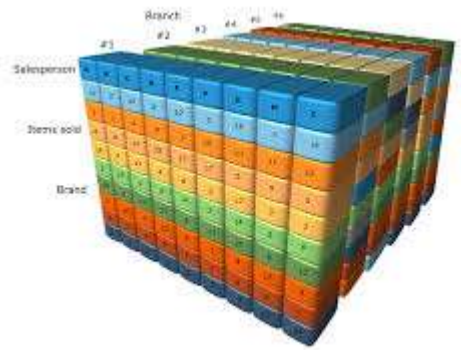


If Defaulter-accounts > 0 **then** Returns-credit = no

If Defaulter-accounts = 0 **and** [(Salary > 2500) **or** (Credit-p > 10)] **then** Returns-credit = yes

OLAP vs Minería de datos

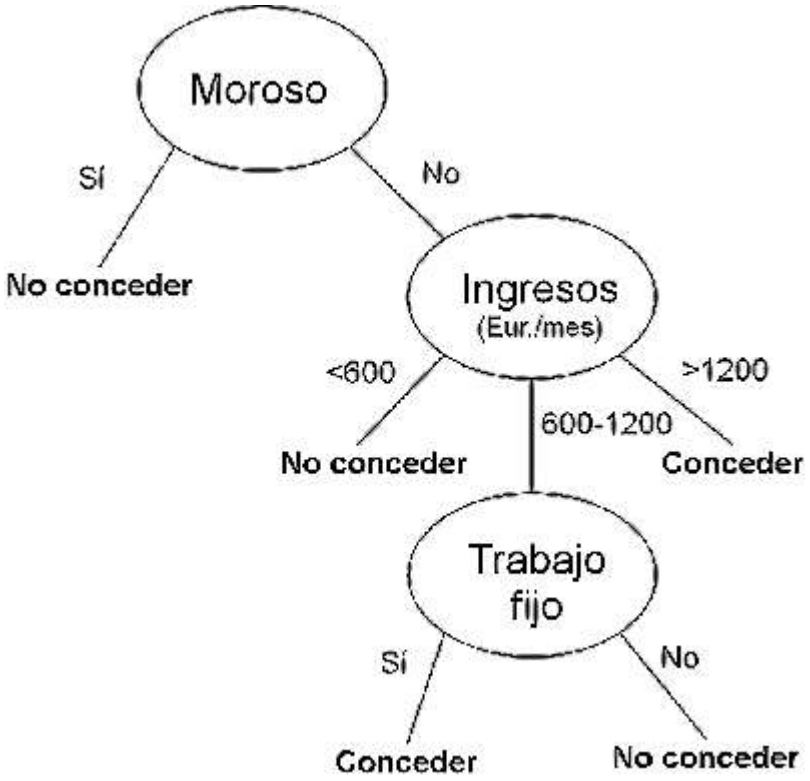
OLAP	Minería de datos
¿Cuál es la proporción media de accidentes entre fumadores y no fumadores?	¿Cuál es la mejor predicción para accidentes?
¿Cuál es la factura telefónica media de mis clientes y de los que han dejado la compañia?	¿Dejara X la compañia? ¿Qué factores afectan a los abandonados?
¿Cuánto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



Técnicas mas Usadas en la Minería de Datos

Árboles de decisiones

Conceder o no créditos a un cliente.

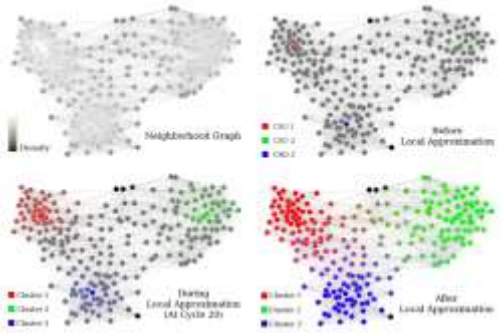


Clustering

- ***El problema de la clasificación***
  - Identificar grupos de individuos/objetos de características similares
  - Tipologías
    - Economía
    - Biología
    - Medicina

- ***Definición de análisis cluster***

  Conjunto de técnicas multivariantes cuyo principal propósito es la agrupación de individuos en conglomerados (cluster) basándose en las características de los mismos



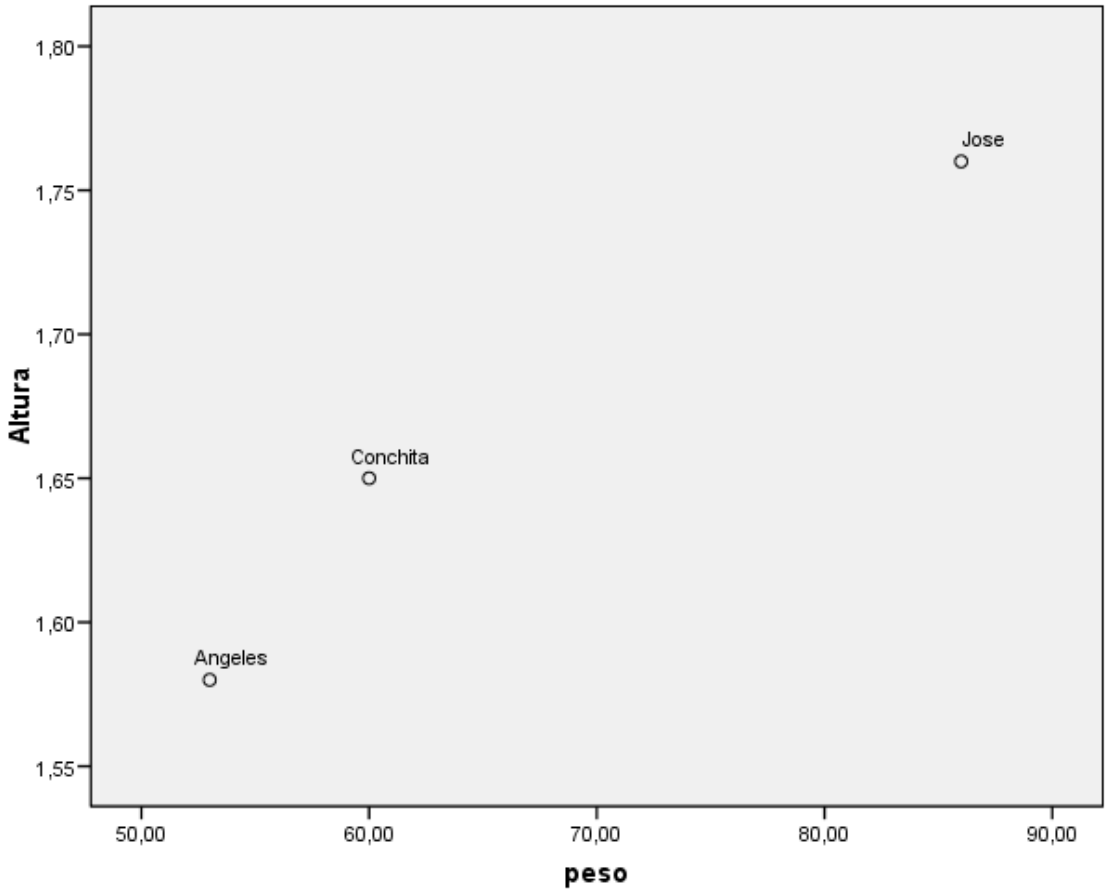
Conceptos básicos

- Matriz de datos

$$\begin{matrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{matrix}$$

Peso	Altura
86	1,76
53	1,58
60	1,65

Conceptos básicos



Representación gráfica de la matriz de datos

Conceptos básicos

- Distancia
  - La distancia es un índice de disimilaridad que verifica las siguientes propiedades:

$$D(a,b) \geq 0$$

$$D(a,b) = D(b,a)$$

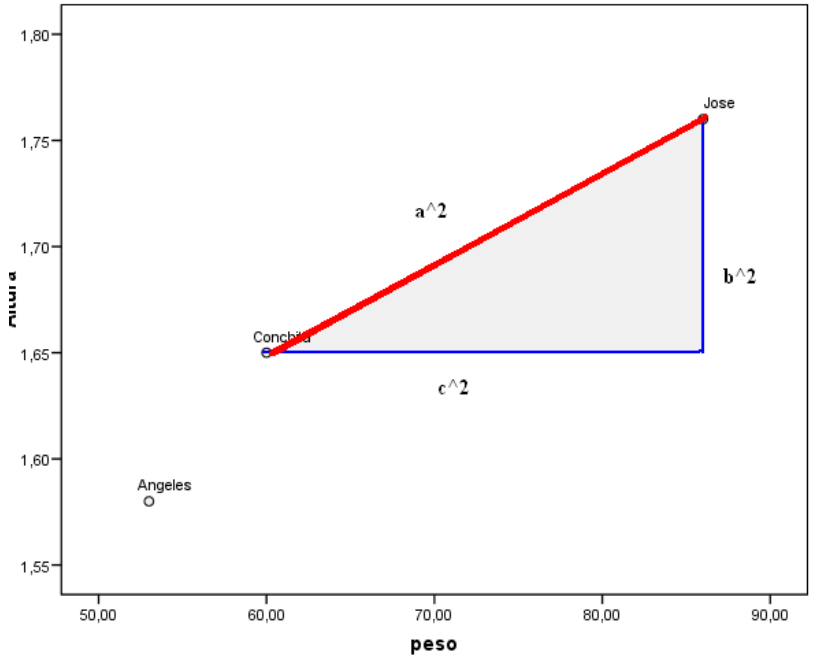
$$D(a,a) = 0$$

$$D(a,c) \leq D(a,b) + D(b,c)$$

Clustering

Conceptos básicos

- Distancia euclídea



Matriz de distancias

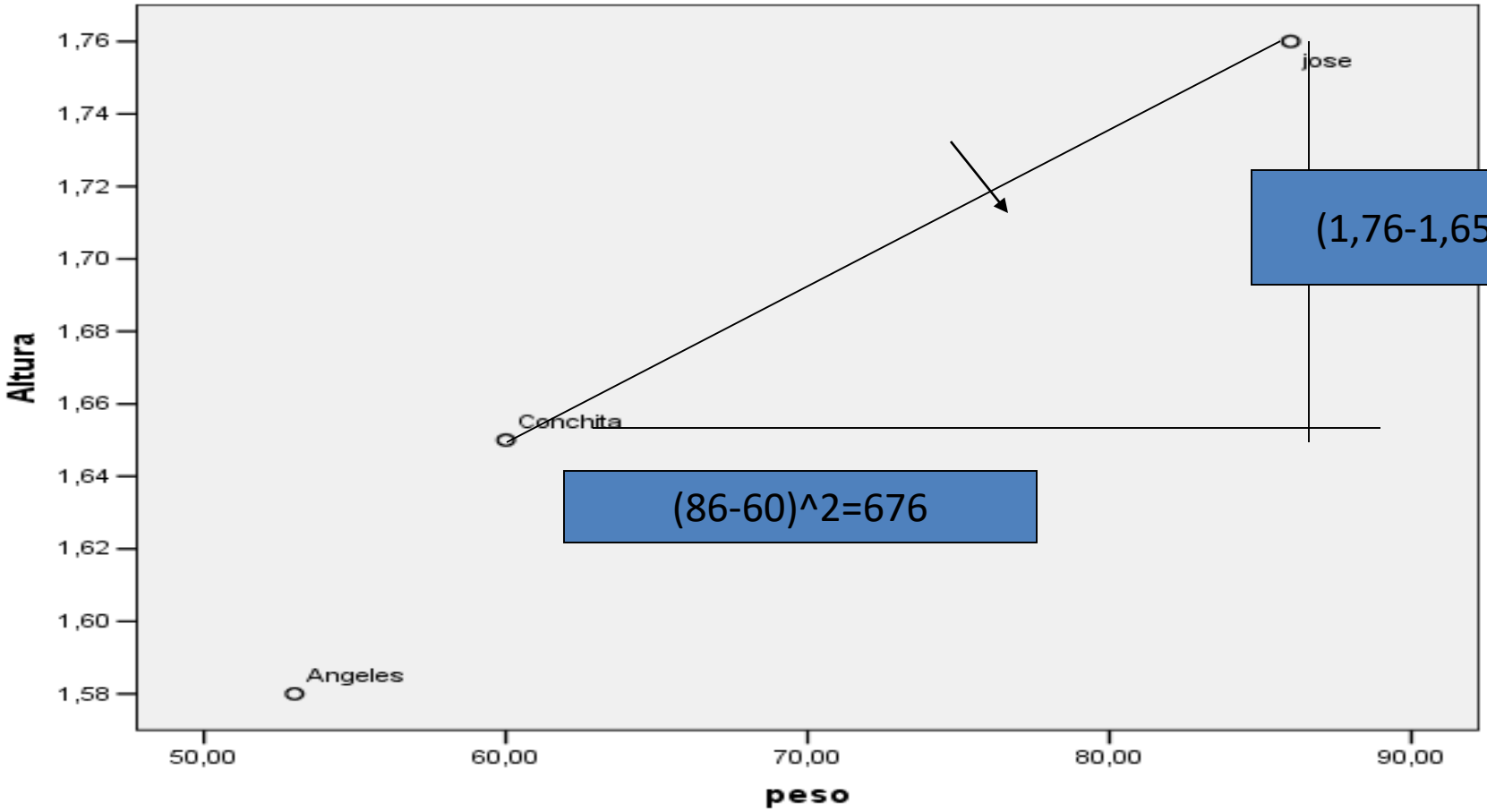
Caso	distancia euclídea		
	1:Jose	2:Angeles	3:Conchita
1:Jose	,000	33,000	26,000
2:Angeles	33,000	,000	7,000
3:Conchita	26,000	7,000	,000

Esta es una matriz de disimilaridades

$$D(a_1, a_2) = \sqrt{(a_{11} - a_{21})^2 + (a_{12} - a_{22})^2}$$

Clustering

Conceptos básicos



(1,76-1,65)^2=0,01

(86-60)^2=676

Estandarización de variables

- El problema de utilizar variables con distinto recorrido.

  -Homogeneizar las escalas en el intervalo 0-1.

$$a'_{i,j} = \frac{a_{i,j} - \min(a_{*,j})}{\max(a_{*,j}) - \min(a_{*,j})}$$

86	1,76
53	1,58
60	1,65

1,00	1,00
0,00	0,00
0,21	0,39

Clustering -Método de agrupación K means

01 Se determina a priori el número de clusters que se desea construir (k).

02 Se establece una configuración aleatoria de los centros de estos clusters, estos centros se denominan centroides.

03 Los elementos se asignan al cluster cuyo centroide esté más cerca.

04 Se recalculan (actualizan) nuevamente los centroides en función de los elementos que les han sido asignados

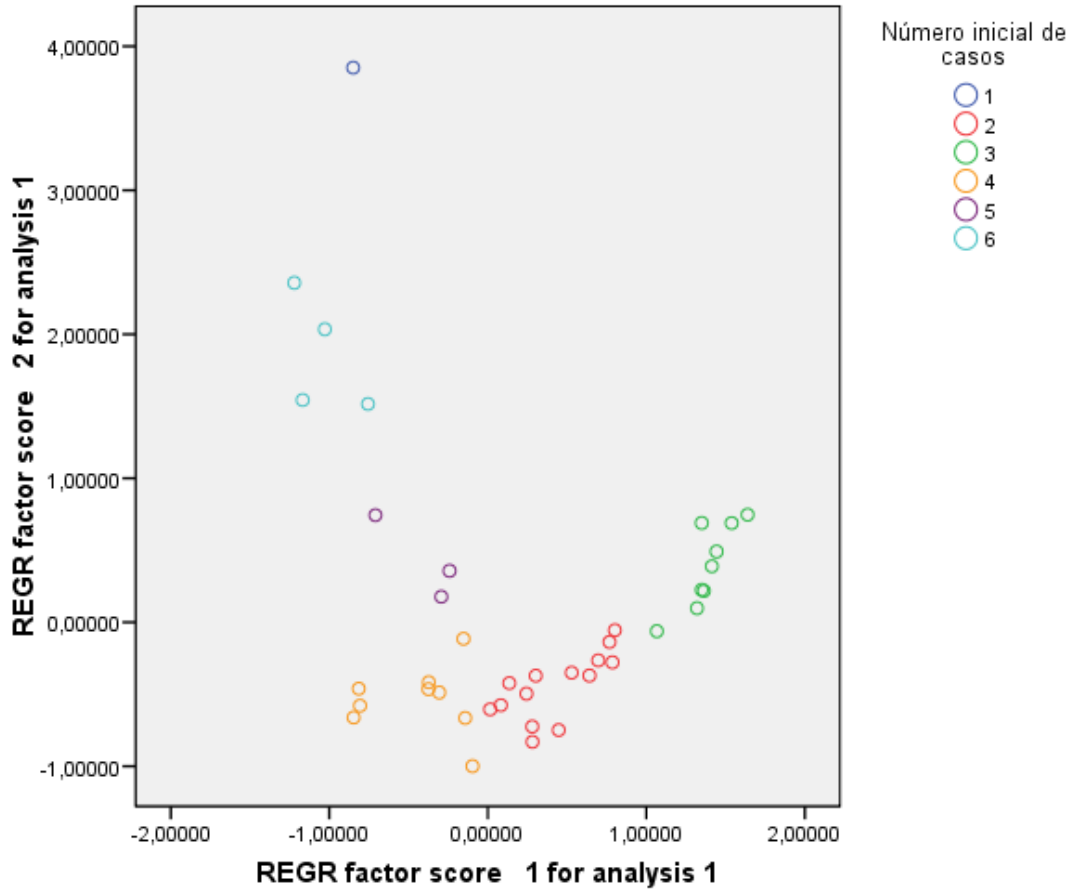
05 Se repite el algoritmo desde el paso 3, hasta que los centroides dejan de cambiar.

Clustering -Método de agrupación K means

- https://www.youtube.com/watch?v=_aWzGGNrcic
- Minuto 4:22 al 7:34

Métodos de agrupación K means

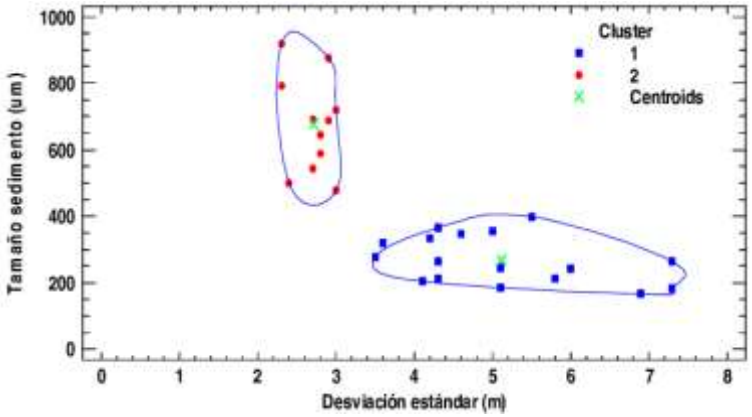
Ejemplo agrupación con 6 clusters



Otros métodos de clustering

Métodos de conglomeración

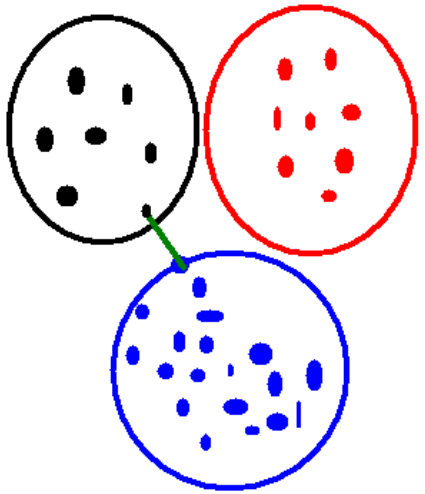
- ✓ **Vecino más cercano.**
- ✓ **Vecino más lejano.**
- ✓ **Centroide**



Otros métodos de clustering

Vecino más cercano

- ✓ La distancia entre dos conglomerados se define como la distancia (en la métrica considerada) de los dos elementos más cercanos.
- ✓ Este método tiende a maximizar lo conexo.



Clustering – Vecino mas cercano

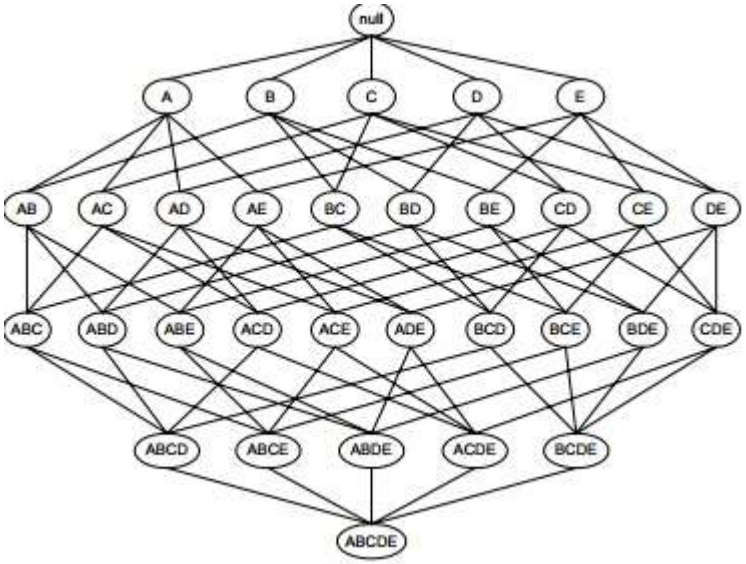
- https://www.youtube.com/watch?v=UqYde-LULfs
- Minuto 00 al 1:46

Reglas de asociación

- Las reglas de asociación se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos.

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

  - ***Detectar cuándo la ocurrencia de un artículo está asociada a la ocurrencia de otros artículos en la misma transacción.***



Reglas de asociación

Expresión de la forma X → Y

→ donde X e Y son itemsets.

$$\{ pan, jamon \} \Rightarrow \{ queso \}$$

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.

Medidas de evaluación de las reglas de asociación

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

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) = 3/4 = 0.75 = 75%

Análisis de la canasta del mercado

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

Reglas de asociación

{pañales} → {cerveza}

{leche, pan} → {cerveza}

Implica co-ocurrencia, no causalidad

Análisis de la canasta del mercado. Aplicaciones

Colocación de producto

- Objetivo: Identificar artículos que muchos clientes compran conjuntamente.
- Solución: Procesar los datos de los terminales de punto de venta proporcionados por los escáneres de códigos de barras.
- Ejemplo: Si un cliente compra pañales, es muy probable que compre cerveza (¡no se sorprenda si ve las cervezas colocadas al lado de los pañales en el súper!)

Redes Neuronales

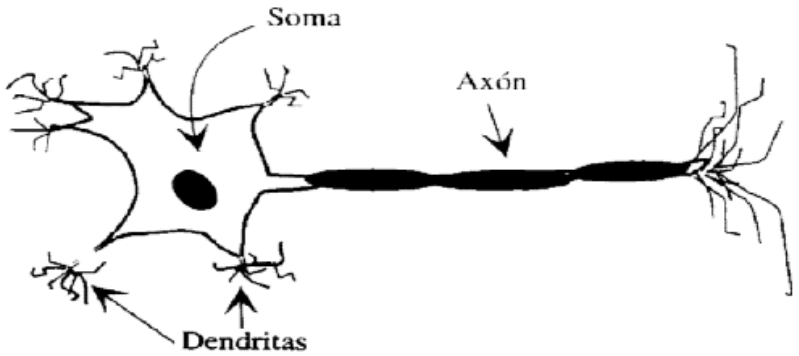
- Una red neuronal es "un sistema para el tratamiento de la información, cuya unidad básica de procesamiento está inspirada en la célula fundamental del sistema nervioso humano: **la neurona**".
- Las neuronas son un componente relativamente simple pero conectadas de a miles forman un poderoso sistema.
- Se utilizan para reconocer patrones, incluyendo imágenes, manuscritos, tendencias financieras, etc.
- Tienen la capacidad de aprender y mejorar su funcionamiento.

Redes Neuronales

Neurona

- El cerebro humano contiene más de cien mil millones de neuronas.
- La clave para el procesamiento de la información son las conecciones entre ellas llamadas **sinápsis**.

Fisiología de una neurona elemental

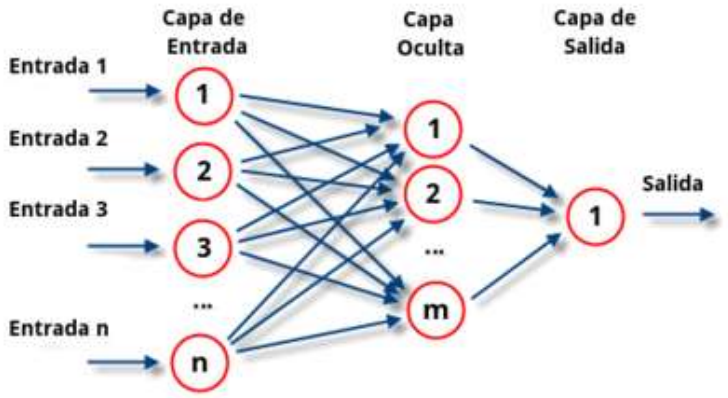


Redes Neuronales

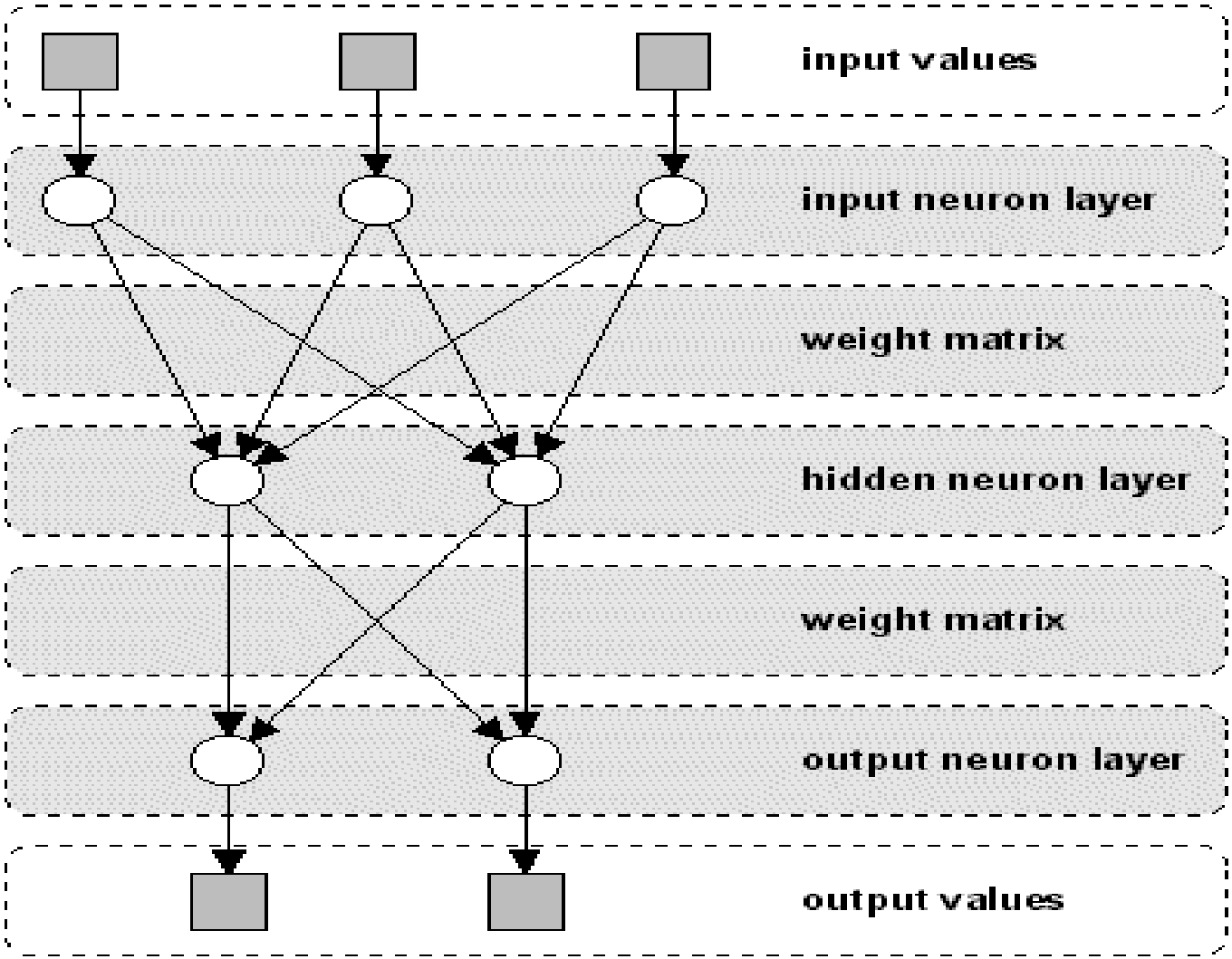
Elementos de una red neuronal

Se interconectan neuronas en tres tipos de capas:

- De entrada: reciben estímulos externos.
- Oculta: elementos internos de procesamiento (se pueden estructurar en varias capas).
- De salida: reciben la información procesada y retornan la respuesta del sistema al exterior.



Redes Neuronales



Redes Neuronales

Ventajas:

- Aprendizaje adaptativo: lo necesario es aplicar un buen algoritmo y disponer de patrones (pares) de entrenamiento.
- Auto-organización => conduce a la generalización
- Tolerancia a fallos: las redes pueden aprender patrones que contienen ruido, distorsión o que están incompletos.
- Operación en tiempo real: procesan gran cantidad de datos en poco tiempo.
- Facilidad de inserción en tecnología ya existente.

Redes Neuronales

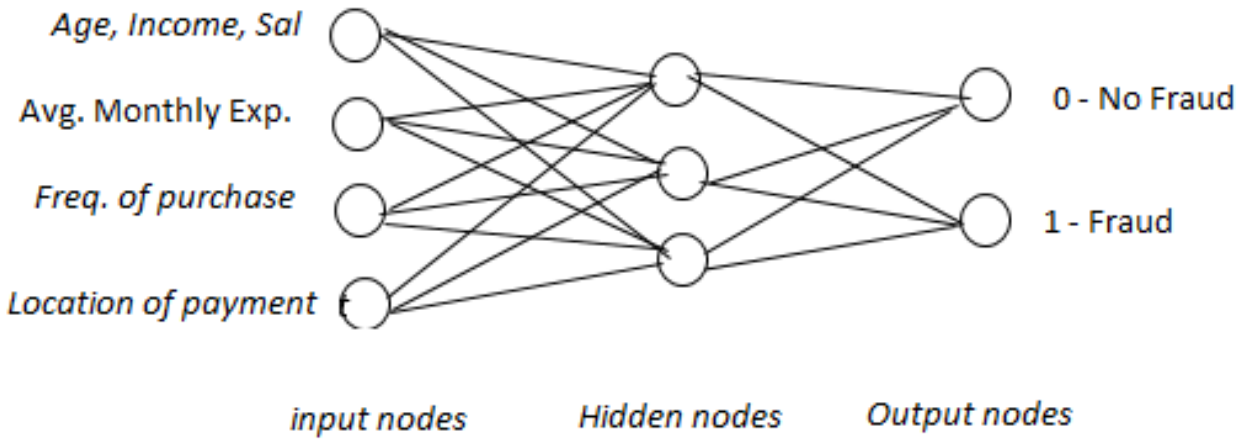
Inconvenientes:

- La base de conocimientos no es transparente (caja negra) (Parcialmente resuelto)
- Aprendizaje algunas veces dificil / lento
- Capacidad de almacenamiento limitado
- Demasiados parametros
  - Numero de capas
  - Numero de neuronas
    - Demasiadas neuronas requieren mas tiempo de entrenamiento
  - Velocidad de aprendizaje

Redes Neuronales

Aplicaciones: Detección de fraude en pago con tarjeta de crédito

- En 2010, 33% de los propietarios de tarjetas de crédito reportaron fraudes en sus cuentas en las últimos 5 anos.



Aprendizaje de máquina

- https://www.youtube.com/watch?v=TnUYcTuZJpM&t=212s

- Minuto 0:00 al 3:30
- Minuto 7:18 al 9:35

Otros usos de la Minería de Datos

- **Empresas de telecomunicaciones, tarjetas de crédito y compañías de seguros** para la detección de fraudes, optimización de campañas de marketing, descripción y segmentación de clientes, predicción de fidelidad de clientes.
- **La industria del comercio** para diseñar y evaluar campañas de marketing, definir ofertas más apropiadas o recomendaciones de productos a clientes, y predecir riesgo en asignación de créditos a clientes.
- **La industria de la medicina** para predecir la efectividad de procedimientos quirúrgicos, exámenes médicos y medicamentos
- **Bancos e Instituciones Financieras**...

Restricciones Iniciales en la Minería de Datos

- **Sistemas parcialmente desconocidos**: Si el modelo del sistema que produce los datos es bien conocido, entonces no necesitamos de la minería de datos ya que todas las variables son de alguna manera predecibles.
- **Enorme cantidad de datos:** Bases de datos muy grandes compensan la limitaciones de un modelo incompleto. Esto es particularmente cierto cuando las redes neuronales y otros técnicas adaptativas son utilizadas. En estos casos, se requieren suficiente cantidad de datos para el entrenamiento y la verificación.
- **Potente hardware y software:** Muchas de las herramientas presentes en la minería de datos están basadas en el uso intensivo de la computación, en consecuencia convenientes equipos y software eficientes aumentarán el desempeño del proceso, el cual a veces debe vérselas con producciones de datos del orden de los Gbytes/hora.

Un caso famoso acerca del comportamiento de los consumidores

Una situación muy popular sucedió en una cadena de víveres en USA. Utilizando un software de minería de datos para estudiar el comportamiento de sus clientes, encontraron relaciones interesantes entre **pañales, cervezas, hombres, y día de la semana.**

Encontraron que los días jueves y sábado, los hombres que compraban pañales también compraban cerveza.

CRISP-DM

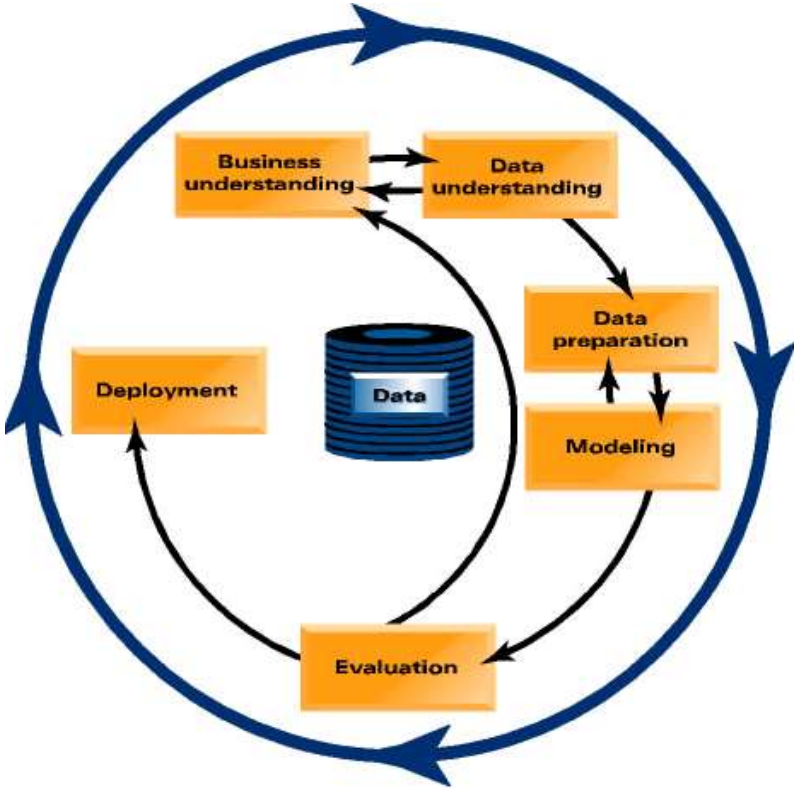
CRoss-Industry Standard Process for Data Mining

- No tiene propietario
- Neutral para Applicaciones/Industrias
- Neutral para herramientas
- Se enfoca en las necesidades del negocio
- Ofrece un framework guía.

¿Porque se debe usar un proceso estándar?

- El proceso de minería de datos debe ser confiable y repetible por personas con poco conocimiento de minería de datos.
- Permite que los proyectos sean replicables
- Ayuda al planeamiento y administración del Proyecto.
- Factor de comfort para personas que recién lo usan
  - Demuestra madurez en Minería de Datos.
  - Reduce la dependencia de “estrellas”

CRISP-DM: Introducción



- Metología de minería de datos
- Modelo de Proceso
- **Puede ser usado por cualquiera**
- **Provee un mapa completo.**
- **Ciclo de Vida: 6 fases**

CRISP – DM: Fases y tareas