# Resumen Data Mining

Por: Pablo Federico Martín Luna

## Tema 1: Introducción al Data Mining

### ¿Qué es Data Mining?

La minería de datos es el proceso de extraer patrones y conocimiento de grandes conjuntos de datos, utilizando técnicas automatizadas y algoritmos sofisticados.

### Knowledge Discovery in Databases (KDD)

Proceso de descubrir patrones y relaciones interesantes en grandes conjuntos de datos almacenados en una base de datos. Se utilizan técnicas automatizadas y algoritmos para analizar los datos y extraer información útil e información relevante que puede ser utilizada para tomar decisiones informadas.

#### Aspectos importantes de KDD

- Patrones válidos: conocimiento contrastable con la realidad.
- Potencialmente útiles: relación con el objetivo que nos proponemos.
- Comprensibles: relacionado con el usuario que maneja el conocimiento extraído de los datos.

#### Proceso y fases de KDD

1. Definir qué se va a hacer
    1. Entender el dominio de aplicación.
    2. Entender el problema a resolver.
    3. Fijar los objetivos: asociación, clasificar, agrupar o predecir.
2. Selección de los datos
    1. Tipos de datos que se van a usar y sus fuentes.
    2. Se seleccionan los datos y se extraen.
    3. Costoso en tiempo y esfuerzo.
3. Preprocesamiento

    Preparación y limpieza de los distintos datos de manera que se puedan manejar en el resto de las fases. (70% del esfuerzo aprox.)

    - Datos necesarios.
    - Datos incompletos.
    - Datos redundantes.
    - Datos incorrectos.
    - Errores de transcripción.
    - Datos envejecidos.
    - Variaciones de datos.
4. Transformación

    Tratamiento preliminar de datos, transformación y creación de nuevas variables.

    - Convertir datos categóricos a numéricos y viceversa.
    - Otras transformaciones: simplificar, agrupar, normalizar…
    - Reducción de la dimensionalidad.

5. Técnicas de Data Mining

Elección de las técnicas que se van a usar para encontrar patrones o información que estaba oculta extrayendo nuevo conocimiento que dé valor extra a los datos.

Tipos:

- Predictivo.
- Descriptivo.

Técnicas:

1. Regla de asociación

   Dado un conjunto de registros, encontrar reglas de dependencia que ayuden a predecir la ocurrencia de un elemento basándonos en la presencia de otros.

   $$X \rightarrow Y$$

   Características:

   - Cuando aparecen elementos por asociación aparecen otros.
   - Se trabaja con atributos categóricos.
   - Los atributos pueden expresarse en forma binaria.
   - Los elementos de X no aparecen en Y.

2. Patrones secuenciales

   Es la extracción de patrones frecuentes relacionados con el tiempo u otro tipo de secuencia.

   Características:

   - Son similares a las reglas de asociación, pero los atributos tienen temporalidad o siguen una secuencia.

3. Clasificación

   Dado un conjunto de registros. Cada uno con un conjunto de atributos y donde uno de ellos se le denomina atributos de clase.

   Encontrar un modelo para el atributo de clase en función de los valores del resto de atributos.

   Hay que validar los resultados sobre un conjunto de datos de validación no usados hasta entonces.

   Características:

   - Trabajar con datos etiquetados, supervisado.
   - Utilizar un conjunto de datos de entrenamiento y clasificar nuevos datos
   - Encuentra los atributos que definen mejora una clase

4. Regresión

Predicción del valor de una variable basándose en el valor que reflejan otras y siguiendo modelos de dependencia que pueden ser lineales o no lineales.

Cuidado con hacer hipótesis de causalidad.

Características:

- Predicen valores numéricos.
- La clasificación se puede entender como la predicción de una clase.

5. Clustering

Dados unos datos y unas medidas de similitud entre ellos, encontrar agrupaciones similares.

Las medidas de similitud pueden ser la distancia euclídea para atributos continuos.

Características:

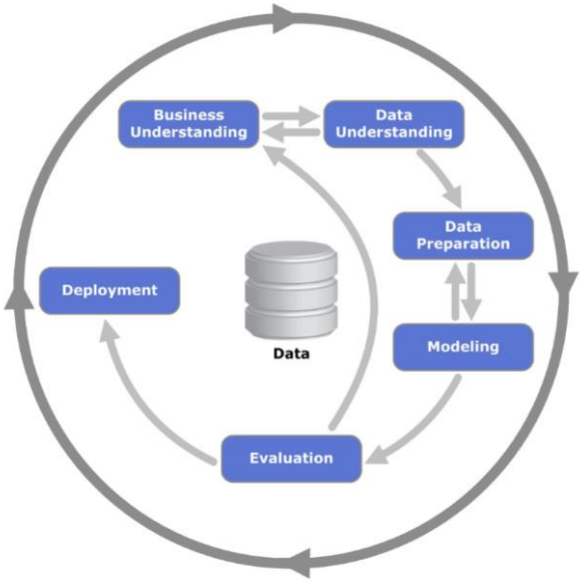
- Trabaja con datos no etiquetados, no supervisados.
- Sugiere nuevos grupos con respecto a los atributos.

6. Interpretación y evaluación

Los modelos o patrones obtenidos deberían aceptar o contradecir la hipótesis inicial. Los patrones deben presentarse de forma que sean entendibles. Es por ello por lo que las técnicas de visualización son muy útiles.

- Integrar los datos:
  - Usar el modelo obtenido en el sistema de información que se está trabajando.
  - Continuar con el proceso de KDD usando los nuevos datos que se obtienen.

## Cross-Industry Standard Process for Data Mining (CRISP-DM)




1. Comprensión del negocio:
   - Establecimiento de los objetivos del negocio.
   - Evaluación de la situación.
   - Establecimiento de los objetivos del Data Mining.
   - Generación del plan del proyecto.
2. Comprensión de los datos:
   - Recopilación inicial de los datos.
   - Descripción de los datos.
   - Exploración de los datos.
   - Verificación de la calidad de los datos.
3. Preparación de los datos:
   - Selección de los datos.
   - Limpieza de los datos.
   - Construcción de los datos.
   - Integración de los datos.
   - Formateo de datos.
4. Modelado:
   - Selección de la técnica de modelado.
   - Diseño de la evaluación.
   - Construcción del modelado.
   - Evaluación del modelado.
5. Evaluación:
   - Evaluación de resultados.
   - Revisar el proceso.
   - Establecimiento de los siguientes pasos o acciones.
6. Despliegue:
   - Planificación de despliegue.
   - Planificación de la monitorización y del mantenimiento.
   - Generación de informe final.
   - Revisión del proyecto.

## Casos de uso

- Asociación: Encontrar ítems o características que se dan en común.
- Clasificación: Crear un clasificador con una serie de individuos. Luego usarlo para asignar una clase a los individuos nuevos.
- Predicción: Dados unos datos históricos para saber si un evento ocurrirá o no.
- Clustering: Agrupar individuos de manera que los individuos cercanos se les considere de un mismo grupo.

# Tema 2: Preparación de datos

## Objetivo

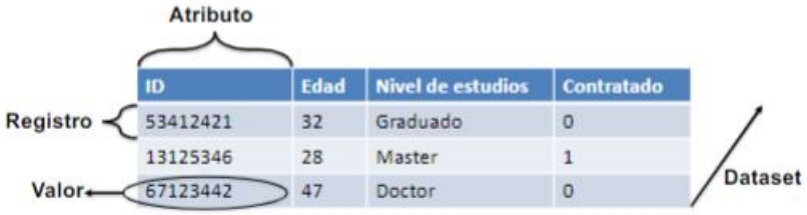
Organizarlos de manera que pueda ser procesados por los programas de construcción de modelos que hayan sido elegidos y, al mismo tiempo, asegurar que los datos se hallan de tal forma que se pueda obtener el mejor modelo posible del conjunto de datos.

## Tipos de datos

### Estructurados

Información, presentada en forma de columnas con cabecera y filas que puede ser ordenada y procesada. Fácil acceso.

- Tablas de datos: consisten en una colección de registros con un conjunto de atributos cada uno.




  - Valor: representación simbólica de un atributo o característica de una entidad.
  - Atributo: representación de una propiedad o característica. Tipos:
    - Numéricos.
    - Booleanos.
    - Categóricos.
  - Registró: una colección de atributos.
  - Colección de registros: tabla.
  - Conjunto de tablas: dataset.
- Transacciones: cada registro o transacción se compone de un conjunto de ítems.

### No estructurados

Aquel que no tiene una estructura interna identificable. Se pueden organizar para hacer búsquedas sencillas. Se almacenan sin que el sistema entienda el formato.

- Texto: un documento consiste en un vector de términos, cada término es una componente o atributo del vector.
- Imágenes: en blanco y negro, matriz de NxM con valores entre 0 y 255. En color, igual, pero con profundidad 3.
- Vídeos: conjunto de X imágenes.
- Redes sociales: la información que se genera en una red social corresponde a un grafo. En él, los nodos serían usuarios y las aristas su manera de relacionarse. Los grafos se representan como matrices.
- Señales: dependiendo de si la señal es analógica/digital, de un canal o multicanal, etc., la complejidad varía. Simplificando, es un vector de N instantes de tiempo con los valores de la frecuencia.

### Semiestructurados

Aquellos que usan etiquetas para identificar distintos datos, permitiendo agruparlos y establece jerarquías (JSON, XML, NoSQL)

**MIRAR TABLA PÁG. 7. 22-23 DATA MINING TEMA 2.PDF**

## Calidad de datos

Es una forma compleja de medir las propiedades de los datos desde diferentes perspectivas. Es un examen exhaustivo de la eficiencia, la fiabilidad y la conveniencia de los datos.

Las 6 dimensiones:

- Completitud: en algunos casos, los datos que no están son irrelevantes, pero cuando se vuelven necesarios para un proceso del negocio, éstos se vuelven críticos.
- Conformidad: los datos que están en los campos de la tabla deben estar en un formato estándar y legible.
- Consistencia: al hacer el cruce de información con los registros, se debe evitar la información contradictoria.
- Precisión/Exactitud: si los datos no son precisos, estos no pueden ser utilizados. Para detectar si lo son, se comparan con una fuente de referencia.
- Duplicación: es importante saber si se tiene la misma información en formatos iguales o similares dentro de la tabla.
- Integridad: otra dimensión de calidad importante radica en el hecho de saber si toda la información relevante de un registro está presente de forma que se pueda utilizar.

## Técnicas de selección

- Registros: sólo se eligen instancias completas representativas del total de los datos disponibles. Aleatoria o por filtros.
- Atributos: se realiza porque haya atributos irrelevantes, redundantes o por exceso de dimensionalidad.

## Técnicas de limpieza

- Datos incompletos: no se ha recogido la información o no se ha proporcionado. Ver si es significante.
- Datos redundantes: pueden ser consecuencias de la mezcla de dos o más conjuntos de datos.
- Datos incorrectos o inconsistentes: se dan porque no hay proceso de control de errores.
- Errores de transcripción: muy típicos con el uso de mayúsculas y minúsculas. Formatos de fechas.
- Variaciones en las referencias a los mismos conceptos: en datos categóricos usar distintas etiquetas para referirse a lo mismo.
- Valores atípicos (outliers): aparece un valor que no se relaciona en nada con el resto, totalmente diferente y aislado, se puede considerar que es una anomalía en la medición.

## Transformación de datos

- Convertir valores continuos a categóricos: asignar una categoría a cada rango de valores que necesitemos, fijando bien una correspondencia entre los valores numéricos y la categoría.
- Datos categóricos a numéricos: proceso inverso al anterior. Con intervalos o valores únicos.

- Simplificación de valores: dividir todos los valores por una constante.
- Agrupación de valores continuos: asignar una “etiqueta” numérica a un evento.
- Expansión de un atributo: cuando el valor de un atributo puede adoptar los valores en un conjunto limitado de categorías.
- Derivación de datos: utilizar los atributos de los datos existentes para derivar nuevos atributos.
- Fusión de datos: combinar dos conjuntos de datos en uno.
- Normalización: consiste en situar los datos sobre una escala de valores equivalente que permita la comparación de atributos que toman valores en dominios o rangos diferentes.
  Tipos:
  - Por máximos: encontrar el valor máximo del atributo a normalizar X y dividir los valores por este.
    $$Z_i = \frac{X_i}{X_{max}}$$
  - Por la diferencia: compensar el defecto de la distancia del valor que tratamos con respecto al máximo de los valores observados.
    $$Z_i = \frac{|X_i - X_{min}|}{|X_{max} - X_{min}|}$$
- Discretización: consiste en establecer un criterio por el medio del cual se pueden dividir los valores de un atributo en dos o más conjuntos disjuntos.
  Características:
  - Se realiza por motivos de coste computacional.
  - El tiempo necesario es menor con datos discretizados.
  - En general necesitan menos espacio de almacenamiento.
  - Los modelos que usan datos continuos pesan más.
  - La comprensión de los modelos es más sencilla usando menos valores para describir un atributo.

  Tipos:
  - Partición en intervalos de la misma amplitud.
  - Partición en intervalos de igual frecuencia.

## Reducción de la dimensionalidad

Asegurar la calidad del modelo resultante:

- Trabajando con menos atributos.

Problemas de la alta dimensionalidad:

- Coste de procesamiento y almacenamiento.
- Atributos relevantes e irrelevantes.
- Cálculo de distancias complejo.
- Data sparsity.
- La maldición de la dimensionalidad.

Tipos:

- Reducción del número de atributos: la reducción del número de atributos consiste en encontrar un subconjunto de los atributos originales que permita obtener modelos de la misma calidad que los que se obtendrían utilizando todos los atributos.
  Fases: Selección de atributos y Extracción de atributos

## Análisis de componentes principales (PCA)

Transformar un conjunto de variables, a las que se denomina originales, en un nuevo conjunto de variables denominadas componentes principales. Estas últimas se caracterizan por estar incómodas entre sí y, además, pueden ordenarse de acuerdo con la información que llevan incorporada.

Fases:

1. Calcular el sumatorio de los valores.
2. Calcular el porcentaje de varianza.
3. Calcular el porcentaje acumulado
4. Ver en los autovectores qué atributos influyen más, mirando los valores más altos, en valor absoluto.
5. Calcular los valores de las componentes para cada caso.
6. Graficar.

**AQUÍ MIRAR EL EJEMPLO QUE HICIMOS DE PCA**

## Reducción de dataset

- Reducción del número de casos: consiste en encontrar una muestra, un subconjunto original de casos, que muestre un comportamiento parecido.

# Tema 3: Exploración de datos

## Visualización de datos

Convertir los datos en formato tabular o visual de modo que las características de los datos y sus relaciones con sus atributos sean más intuitivas.

Se puede usar en dos momentos diferentes:

- Visualización previa: se utiliza para entender mejor los datos y sugerir posibles patrones o que tipo herramienta de KDD utilizar.
- Visualización posterior al proceso de minería de datos: se utiliza para mostrar los patrones y entenderlos.

Dependiendo de la fase, hay dos tipos de usuarios:

- Visualización previa se utiliza frecuentemente por dataset curators[1], para ver la calidad de los datos, tendencias o filones que investigar.
- La visualización posterior se utiliza normalmente para validar y mostrar a los expertos/clientes los resultados.

Diferentes métodos según el tipo de dato:

- Univariante: medida de variable cuantitativa simple. Mide la distribución.
- Bivariante: parejas de individuos de dos variables cuantitativas. Las variables están relacionadas.
- Multivariable: representación multidimensional de datos multivariable.

## Diagrama de barras

Se utiliza para representar la frecuencia de las categorías de una variable del mismo tipo. Si es cuantitativa, hacer una transformación.

## Histograma

Permite la representación de la frecuencia de una variable cuantitativa mediante intervalos.

## Pie chart

Suelen utilizarse para mostrar tamaños relativos de partes de un todo. (ej.: porcentajes)

## Diagrama de dispersión

Son muy útiles para mostrar de forma intuitiva las relaciones entre 2-3 atributos.

## Diagrama de burbujas

Es un tipo especial de diagrama de dispersión al que se le introduce una tercera variable indica el tamaño.

## Polígono de frecuencias → Diagrama líneas

Esta representación se basa en el histograma. Sólo es útil para variables cuantitativas. Los puntos que permiten la unión de las líneas representan el centro de clase y el valor de un punto en el caso de crear un diagrama de líneas.

---

[1] Es la persona que se encarga de que un dataset esté bien recopilado y los formatos se correspondan a las necesidades de la organización.

## Diagrama de tallos y hojas

Permite obtener simultáneamente una distribución y las frecuencias de la variable junto a su representación gráfica.

Como se crea:

1. Ordenar los valores de menor a mayor o viceversa.
2. Elegir el tallo y separarlo de las hojas.
3. Dibujar los tallos de arriba hacia abajo y sus hojas asociadas de izquierda a derecha.

## Diagrama de cajas

Mostrar visualmente grupos de datos numéricos a través de sus cuartiles.

## Mapas coropléticos

Muestran zonas geográficas o regiones divididas en colores, con sombras o dibujos en relación con una variable de datos.

## Diagrama de cuerdas

Visualizar las interrelaciones entre entidades. Las conexiones entre las entidades se utilizan para mostrar el hecho de que comparten algo en común.

## Diagrama de diamantes

Permite mostrar visualmente qué atributos se asocian con un conjunto de individuos.

## Mapas de calor

Visualizar datos a traces de las variaciones de color.

## Grafos

Entender cómo se conectan los diferentes individuos.

## Dendrograma

Entender las jerarquías mediante una representación en forma de árbol.

# Tema 4: Reglas de asociación

## Definición del problema

Dado un conjunto de transacciones, encontrar las reglas que predecirá la aparición o no de un elemento concreto basado en la aparición del resto de los elementos en la transacción.

### Aplicaciones

- Product placement. (donde colocar los productos en el supermercado)
- Promociones y ofertas.
- Gestión de inventarios.

### Definiciones

- Conjunto de elementos (itemset): colección de uno o más elementos.
- K-itemset: conjunto con K elementos.
- Frecuencia de soporte: frecuencia de aparición de un K-itemset.
- Soporte: frecuencia de un K-itemset entre el total de transacciones. La probabilidad condicional de que una transacción que contenga X (antecedente) también contenta Y (consecuente), es decir los elementos de la regla.
- Conjunto de elementos (itemset) es frecuente: un conjunto de elementos cuyo soporte es mayor o igual que un umbral (soporte mínimo) dado o elegido.
- Regla de asociación: una relación X □ Y donde X e Y son conjuntos de elementos. Se tiene que cumplir:
    - Todos los elementos de X e Y pertenecen al conjunto inicial de datos.
    - Los elementos de X no se repiten en Y y viceversa.
- Métricas para la evaluación de reglas:
    - Confianza: Frecuencia con que los elementos en Y aparecen en transacciones que contienen a X. Dicho de otra manera, la probabilidad de que una transacción contenga X e Y dividido por las que contienen sólo a X. Deben ser mayor o igual que un umbral mínimo.

## Algoritmo Apriori

1. Empezando por los ítems individuales, k = 1.
2. Repetir hasta que no haya más itemsets frecuentes.
3. Extraer y evaluar las reglas.

### Ejemplo

Dado el siguiente dataset. Para una frecuencia de soporte mínimo de 2 y una confianza mínima del 60%.

| ID | Transacción |
|---|---|
| **T1** | L1, L2, L5 |
| **T2** | L2, L4 |
| **T3** | L2, L3 |
| **T4** | L1, L2, L4 |
| **T5** | L1, L3 |
| **T6** | L2, L3 |
| **T7** | L1, L3 |
| **T8** | L1, L2, L3, L5 |
| **T9** | L1, L2, L3 |

K = 1

Se crea una tabla con el conteo de soporte de cada ítem en el Dataset (conjunto candidato).

| Itemset | Frecuencia Soporte |
|---|---|
| **L1** | 6 |
| **L2** | 7 |
| **L3** | 6 |
| **L4** | 2 |
| **L5** | 2 |

Comparamos con la frecuencia de soporte mínima. Si la frecuencia de soporte es menor, eliminamos el ítem.

En este caso no se elimina ninguno.

K = 2

Se crea una tabla con la frecuencia de soporte de cada ítem en el dataset (conjunto candidato).

| Itemset | Frecuencia Soporte |
|---|---|
| **L1, L2** | 4 |
| **L1, L3** | 4 |
| **L1, L4** | 1 |
| **L1, L5** | 2 |
| **L2, L3** | 4 |
| **L2, L4** | 2 |
| **L2, L5** | 2 |
| **L3, L4** | 0 |
| **L3, L5** | 1 |
| **L4, L5** | 0 |

Comparamos con la frecuencia de soporte mínima. Si la frecuencia de soporte es menor, eliminamos el ítem.

| Itemset | Frecuencia Soporte |
|---|---|
| **L1, L2** | 4 |
| **L1, L3** | 4 |
| ~~**L1, L4**~~ | ~~1~~ |
| **L1, L5** | 2 |
| **L2, L3** | 4 |
| **L2, L4** | 2 |
| **L2, L5** | 2 |
| ~~**L3, L4**~~ | ~~0~~ |
| ~~**L3, L5**~~ | ~~1~~ |
| ~~**L4, L5**~~ | ~~0~~ |

K = 3

| Itemset | Frecuencia Soporte |
|---|---|
| **L1, L2, L3** | 2 |
| ~~**L1, L2, L4**~~ | |
| **L1, L2, L5** | 2 |
| ~~**L1, L3, L5**~~ | |
| ~~**L2, L3, L4**~~ | |
| ~~**L2, L3, L5**~~ | |
| ~~**L3, L4, L5**~~ | |

En este caso, solo nos quedamos con 2 itemsets porque el resto son subsets de itemsets ya descartados. Y de los que nos quedan nos quedamos con todos.

K = 4

Solo hay un  itemset con 4 elementos, pero al contener un itemset ya eliminado no lo podemos agregar.

EVALUAMOS LAS REGLAS

| Itemset | Frecuencia Soporte |
|---|---|
| **L1, L2, L3** | 2 |
| **L1, L2, L5** | 2 |

Evaluamos para cada ítem de la iteración final.

Para el itemset (L1, L2, L3):

$$L1, L2 \rightarrow L3 = \frac{\sigma(L1, L2, L3)}{\sigma(L1, L2)} = \frac{2}{4} = 0.5 = 50\%$$

$$L1, L3 \rightarrow L2 = \frac{\sigma(L1, L2, L3)}{\sigma(L1, L3)} = \frac{2}{4} = 0.5 = 50\%$$

$$L2, L3 \rightarrow L1 = \frac{\sigma(L1, L2, L3)}{\sigma(L2, L3)} = \frac{2}{4} = 0.5 = 50\%$$

$$L1 \rightarrow L2, L3 = \frac{\sigma(L1, L2, L3)}{\sigma(L1)} = \frac{2}{6} = 0.33 = 33\%$$

$$L2 \rightarrow L1, L3 = \frac{\sigma(L1, L2, L3)}{\sigma(L2)} = \frac{2}{7} = 0.28 = 28\%$$

$$L3 \rightarrow L1, L2 = \frac{\sigma(L1, L2, L3)}{\sigma(L3)} = \frac{2}{6} = 0.33 = 33\%$$

Para el itemset (L1, L2, L5):

$$L1, L2 \rightarrow L5 = \frac{\sigma(L1, L2, L5)}{\sigma(L1, L2)} = \frac{2}{4} = 0.5 = 50\%$$

$$L1, L5 \rightarrow L2 = \frac{\sigma(L1, L2, L5)}{\sigma(L1, L5)} = \frac{2}{2} = 1 = 100\%$$

$$L2, L5 \rightarrow L1 = \frac{\sigma(L1, L2, L5)}{\sigma(L2, L5)} = \frac{2}{2} = 1 = 100\%$$

$$L1 \rightarrow L2, L5 = \frac{\sigma(L1, L2, L5)}{\sigma(L1)} = \frac{2}{6} = 0.33 = 33\%$$

$$L2 \rightarrow L1, L5 = \frac{\sigma(L1, L2, L5)}{\sigma(L2)} = \frac{2}{7} = 0.28 = 28\%$$

$$L5 \rightarrow L1, L2 = \frac{\sigma(L1, L2, L5)}{\sigma(L5)} = \frac{2}{2} = 0.33 = 100\%$$

Ahora tenemos comparamos los porcentajes con la confianza mínima y vemos cuáles nos quedamos.

Para calcular el soporte, esto lo hacemos si el soporte mínimo está entre 0 y 1:

$$Soporte = \frac{Frecuencia\ Soporte}{Número\ transacciones}$$

## Evaluación de patrones de asociación

Interpretación de las reglas:

- El número de reglas obtenidas en casos reales puede ser del orden de miles o millones.
- Muchas de ellas no tienen interés práctico al no revelar nada novedoso.
- Las conclusiones derivadas de una regla pueden ser débiles.

### Tabla de contingencia

Ejemplo:

| | Café | No café | |
|---|---|---|---|
| **Té** | 150 | 50 | 200 |
| **No té** | 750 | 50 | 800 |
| | 900 | 100 | 1000 |

Hipótesis: “La gente que bebe té tienden a beber café”

Según la tabla de contingencia:

- Los bebedores de café son el 90% independientemente si toman té o no.
- Los bebedores de café en el caso de comprar té disminuyen a un 15%.

Conclusión: la hipótesis no cuadra.

### Medidas de evaluación

- Lift o Interest: cuantifica la relación existente entre X e Y.

$$Lift = \frac{conf(X,Y)}{sop(Y)}$$

  - Interpretación:
    - Lift = 1, si X e Y son independientes.
    - Lift < 1, si X e Y están negativamente correlados.
    - Lift > 1, si X e Y están positivamente correlados.
  - Características:
    - Cuando es igual a 1 quiere decir que hay independencia estadística. No nos dice nada.
    - Es simétrica, Lift(X □ Y) = Lift(Y □ X).

## Algoritmo GSP

1. Obtener todas las secuencias frecuentes de 1 elemento.
2. Para k ≥ 2. Mientras se encuentren nuevas secuencias frecuentes.
   - Generar k-secuencias candidatas a partir de las (k-1)-secuencias frecuentes.
   - Podar k-secuencias candidatas que contengan alguna (k-1)-secuencia no frecuente.
   - Recorrer nuevamente el conjunto de datos para obtener el soporte de las candidatas.
   - Eliminar las k-secuencias candidatas cuyo soporte real esté por debajo del MinSap.

## Ejemplo

Dadas las siguientes secuencias y una frecuencia de soporte de 2.

| Transaction Date | Customer ID | Items Purchased | Customer Sequence |
|---|---|---|---|
| **1** | 1 | A | AB(FG)CD |
| **3** | 1 | B | |
| **7** | 1 | FG | |
| **9** | 1 | C | |
| **10** | 1 | D | |
| **1** | 2 | B | BGD |
| **4** | 2 | G | |
| **6** | 2 | D | |
| **1** | 3 | B | BFG(AB) |
| **5** | 3 | F | |
| **8** | 3 | G | |
| **9** | 3 | AB | |
| **2** | 4 | F | F(AB)CD |
| **6** | 4 | AB | |
| **8** | 4 | C | |
| **10** | 4 | D | |
| **3** | 5 | A | A(BC)GF(DE) |
| **4** | 5 | BC | |
| **7** | 5 | G | |
| **9** | 5 | F | |
| **10** | 5 | DE | |

K = 1

| Item | Support |
|---|---|
| **A** | 4 |
| **B** | 5 |
| **C** | 3 |
| **D** | 4 |
| **E** | 1 |
| **F** | 4 |
| **G** | 4 |

K = 2

Items comprados por separado (Secuencia)

| | A | B | C | D | F | G |
|---|---|---|---|---|---|---|
| A | AA | AB | AC | AD | AF | AG |
| B | BA | BB | BC | BD | BF | BG |
| C | CA | CB | CC | CD | CF | CG |
| D | DA | DB | DC | DD | DF | DG |
| F | FA | FB | FC | FD | FF | FG |
| G | GA | GB | GC | GD | GF | GG |

Items comprados a la vez (Transacción

| | A | B | C | D | F | G |
|---|---|---|---|---|---|---|
| A | | (AB) | (AC) | (AD) | (AF) | (AG) |
| B | | | (BC) | (BD) | (BF) | (BG) |
| C | | | | (CD) | (CF) | (CG) |
| D | | | | | (DF) | (DG) |
| F | | | | | | (FG) |
| G | | | | | | |

| Customer Sequence | AA | AB | AC | AD | AF | AG |
|---|---|---|---|---|---|---|
| AB(FG)CD | x | 1 | 1 | 1 | 1 | 1 |
| BGD | x | x | x | x | x | x |
| BFG(AB) | x | x | x | x | x | x |
| F(AB)CD | x | x | 1 | 1 | x | x |
| A(BC)GF(DE) | x | 1 | 1 | 1 | 1 | 1 |
| Total | x | 2 | 3 | 3 | 2 | 2 |

| Customer Sequence | BA | BB | BC | BD | BF | BG |
|---|---|---|---|---|---|---|
| AB(FG)CD | x | x | 1 | 1 | 1 | 1 |
| BGD | X | X | X | 1 | X | 1 |
| BFG(AB) | 1 | 1 | X | X | 1 | 1 |
| F(AB)CD | X | X | 1 | 1 | X | X |
| A(BC)GF(DE) | X | X | X | 1 | 1 | 1 |
| Total | 1 | 1 | 2 | 4 | 3 | 4 |

| Customer Sequence | CA | CB | CC | CD | CF | CG |
|---|---|---|---|---|---|---|
| AB(FG)CD | x | x | x | 1 | x | X |
| BGD | x | x | x | x | x | X |
| BFG(AB) | x | x | x | x | x | X |
| F(AB)CD | x | x | x | 1 | x | x |
| A(BC)GF(DE) | X | X | X | 1 | 1 | 1 |
| Total | X | X | X | 3 | 1 | 1 |

| Customer Sequence | DA | DB | DC | DD | DF | DG |
|---|---|---|---|---|---|---|
| AB(FG)CD | X | X | X | X | X | X |
| BGD | X | X | X | X | X | X |
| BFG(AB) | X | X | X | X | X | X |
| F(AB)CD | X | X | X | X | X | X |
| A(BC)GF(DE) | X | X | X | X | X | X |
| Total | X | X | X | X | X | X |

| Customer Sequence | FA | FB | FC | FD | FF | FG |
|---|---|---|---|---|---|---|
| AB(FG)CD | X | X | 1 | 1 | X | X |
| BGD | X | X | X | X | X | X |
| BFG(AB) | 1 | 1 | X | X | X | 1 |
| F(AB)CD | 1 | 1 | 1 | 1 | X | X |
| A(BC)GF(DE) | X | X | X | 1 | X | X |
| Total | 2 | 2 | 2 | 3 | X | 1 |

| Customer Sequence | GA | GB | GC | GD | GF | GG |
|---|---|---|---|---|---|---|
| AB(FG)CD | X | X | 1 | 1 | X | X |
| BGD | X | X | X | 1 | X | X |
| BFG(AB) | 1 | 1 | X | X | X | X |
| F(AB)CD | X | X | X | X | X | X |
| A(BC)GF(DE) | X | X | X | 1 | 1 | X |
| Total | 1 | 1 | 1 | 3 | 1 | X |

| Customer Sequence | (AB) | (AC) | (AD) | (AF) | (AG) |
|---|---|---|---|---|---|
| AB(FG)CD | X | X | X | X | X |
| BGD | X | X | X | X | X |
| BFG(AB) | 1 | X | X | X | X |
| F(AB)CD | 1 | X | X | X | X |
| A(BC)GF(DE) | X | X | X | X | X |
| Total | 2 | X | X | X | X |

| Customer Sequence | (BC) | (BD) | (BF) | (BG) |
|---|---|---|---|---|
| AB(FG)CD | X | X | X | X |
| BGD | X | X | X | X |
| BFG(AB) | X | X | X | X |
| F(AB)CD | X | X | X | X |
| A(BC)GF(DE) | 1 | X | X | X |
| Total | 1 | X | X | X |

| Customer Sequence | (CD) | (CF) | (CG) |
|---|---|---|---|
| AB(FG)CD | X | X | X |
| BGD | X | X | X |
| BFG(AB) | X | X | X |
| F(AB)CD | X | X | X |
| A(BC)GF(DE) | X | X | X |
| Total | X | X | X |

| Customer Sequence | (DF) | (DG) |
|---|---|---|
| AB(FG)CD | X | X |
| BGD | X | X |
| BFG(AB) | X | X |
| F(AB)CD | X | X |
| A(BC)GF(DE) | X | X |
| Total | X | X |

| Customer Sequence | (FG) |
|---|---|
| AB(FG)CD | 1 |
| BGD | X |
| BFG(AB) | X |
| F(AB)CD | X |
| A(BC)GF(DE) | X |
| Total | 1 |

K = 3

Primero vemos que secuencias podemos combinar, siendo estas las que compartan tengan el mismo valor en -Las que el -1st qué queremos combinar

| 2-seq | -1st | -Last |
|---|---|---|
| **AB** | B | A |
| **AC** | C | A |
| **AD** | D | A |
| **AF** | F | A |
| **AG** | G | A |
| **BC** | C | B |
| **BD** | D | B |
| **BF** | F | B |
| **BG** | G | B |
| **CD** | D | C |
| **FA** | A | F |
| **FB** | B | F |
| **FC** | C | F |
| **FD** | D | F |
| **GD** | D | G |
| **(AB)** | A | B |
| | B | A |

| 2-seq (1) | 2-sep - 1st | 2-seq (2) | 2-seq Last | 3-seq after join | 3-seq after prune | Support Count | 3- seq supported |
|---|---|---|---|---|---|---|---|
| **AB** | B | BC | B | ABC | ABC | 1 | |
| **AB** | B | BD | B | ABD | ABD | 2 | ABD |
| **AB** | B | BF | B | ABF | ABF | 2 | ABF |
| **AB** | B | BG | B | ABG | ABG | 2 | ABG |
| **AB** | B | (AB) | B | A(AB) | | | |

Esto se repetiría para todos los casos de la tabla anterior y comprobaremos las secuencias por si hubiese alguna que no apareció en K = 2, como sucede en A(AB) que podemos encontrar AA que no cumplió con el soporte mínimo.

Final:

| 1-item | 2-items | 3-items | 4-items |
|---|---|---|---|
| **A** | AB | ABD | ABFD |
| **B** | AC | ABF | ABGD |
| **C** | AD | ABG | |
| **D** | AF | ACD | |
| **F** | AG | AFD | |
| **G** | BC | AGD | |
| | BD | BCD | |
| | BF | BFD | |
| | BG | BGD | |
| | CD | F(AB) | |
| | FA | FCD | |
| | FB | | |
| | FC | | |
| | FD | | |
| | GD | | |
| | (AB) | | |

# Tema 5: Clasificación

## Métodos:

- Modelos basados en reglas.
- Árboles de decisión.
- Algoritmo k-nearest-neighbors.
- Redes bayesianas.
- Support vector machines.

## Modelo basado en reglas

### Ejemplo

| Cliente | Edad | Diagnóstico | Lágrima | Recomendación |
|---|---|---|---|---|
| **1** | Joven | Miope | Normal | Duras |
| **2** | Joven | Hipermétrope | Normal | Duras |
| **3** | Pre Presbicia | Miope | Normal | Duras |
| **4** | Presbicia | Miope | Normal | Duras |
| **5** | Joven | Miope | Reducida | Nada |
| **6** | Joven | Miope | Reducida | Nada |
| **7** | Joven | Hipermétrope | Reducida | Nada |
| **8** | Joven | Hipermétrope | Reducida | Nada |

Recomendación: Duras

- Edad:
  - Joven: 2/6
  - Pre Presbicia: 1/1
  - Presbicia: 1/1
- Diagnóstico:
  - Miope: 3/5
  - Hipermétrope 1/3
- Lágrima:
  - Normal: 4/4
  - Reducida: 0/4

Escogemos para sacar la norma que tenga mayor valor y abarque más cantidad de casos, por lo tanto:

Si Lágrima = Normal, entonces Recomendación = Duras

Y ahora buscamos la regla para el siguiente valor del atributo Recomendación.

Recomendación: Nada

- Edad:
  - Joven: 4/4
- Diagnóstico:
  - Miope: 2/2
  - Hipermétrope: 2/2
- Lágrima:
  - Reducida: 4/4

En este caso, como hay dos valores iguales y que abarcan la misma cantidad de datos, se escoge la que se quiera, en mi caso escojo la siguiente:

Si Lágrima = Reducida, entonces Recomendación = Nada

# Árboles de decisión

## Fórmulas

- Entropía (Hunt):

$$E(t) = -\sum_{j=1}^{k} p(j\ t) \cdot p(j\ t)$$

- Efectividad:

$$Ef(t) = -\sum_{i=1}^{n} \frac{|E_i|}{|E|} \cdot E(t)$$

- Ganancia:

$$G(t) = E(t) - Ef(t, X)$$

- GINI:

$$GINI(t) = 1 - \sum_{j=1}^{k} p(j\ t)^2$$

- Error de clasificación:

$$Error(t) = 1 - (j\ t)$$

Ejemplo

| Cliente | Horario | Sexo | Nivel físico | Clave |
|---|---|---|---|---|
| **1** | Mañana | Hombre | Alto | 1 |
| **2** | Mañana | Mujer | Normal | 1 |
| **3** | Mediodía | Mujer | Normal | 2 |
| **4** | Tarde | Mujer | Normal | 2 |
| **5** | Tarde | Mujer | Alto | 1 |
| **6** | Mediodía | Mujer | Bajo | 2 |
| **7** | Tarde | Hombre | Bajo | 2 |
| **8** | Mañana | Mujer | Normal | 1 |

1) Calculamos la entropía para el conjunto inicial de datos.

$$E_{inicial} = -\left(\frac{4}{8}\cdot\frac{4}{8} + \frac{4}{8}\cdot\frac{4}{8}\right) = 1$$

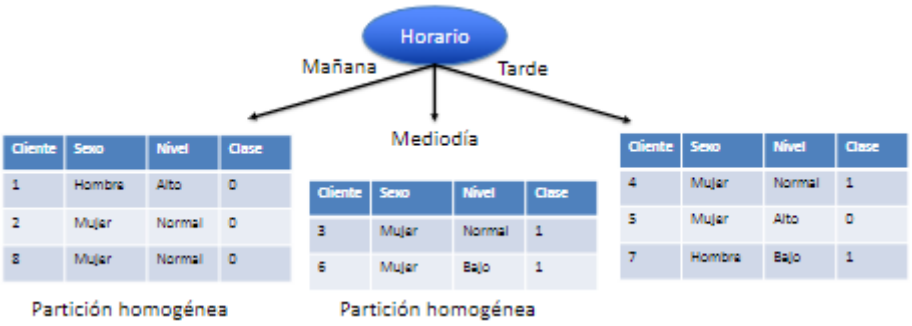
2) Medimos la efectividad para cada atributo.

$$Ef(X, Sexo) = \frac{2}{8}\cdot\left(-\left(\frac{1}{2}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{1}{2}\right)\right) + \frac{6}{8}\cdot\left(-\left(\frac{3}{6}\cdot\frac{3}{6} + \frac{3}{6}\cdot\frac{3}{6}\right)\right) = 1$$

3) Calculamos la ganancia y elegimos la mayor.

$$Sexo = 1 - 1 = 0$$

4) Dividir el nodo elegido y generar subtablas.




5) Repetimos hasta que no haya más subtablas no homogéneas.
6) Obtenemos las reglas.
   a. R1: Horario = “Mañana” □ Clase 1
   b. R2: Horario = “Mediodía” □ Clase 2
   c. R3: Horario = “Tarde” A Nivel Físico = “Alto” □ Clase 1
   d. R4: Horario = “Tarde” A Nivel Físico = “Normal” □ Clase 2
   e. R5: Horario = “Tarde” A Nivel Físico = “Bajo” □ Clase 1
7) Evaluamos las reglas.
8) Calculamos el error para cada regla.
9) Por último, se calcula el error global usando el error y la proporción de registros que tiene esa regla.

$$E_{global} = \sum_{i=1}^{n} w_i \cdot E_i$$

## Algoritmo k-nearest-neighbors

### Métricas para la distancia (para valores continuos)

- Distancia Euclídea:

$$Distancia = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

- Distancia de Manhattan:

$$Distancia = \sum_{i=1}^{k} |x_i - y_i|$$

### Métricas para la distancia (para valores categóricos)

- Distancia de Hamming:

$$Distancia = \sum_{i=1}^{k} |x_i - y_i|$$

## Ejemplo

| Largo sépalo | Ancho sépalo | Largo pétalo | Ancho pétalo | Clase |
|---|---|---|---|---|
| **5.1** | 3.5 | 1.4 | 0.2 | Setosa |
| **4.0** | 3.0 | 1.4 | 0.2 | Setosa |
| **4.7** | 3.0 | 1.3 | 0.2 | Setosa |
| **7.0** | 3.2 | 4.7 | 1.4 | Versicolor |
| **3.4** | 3.2 | 4.5 | 1.5 | Versicolor |
| **6.9** | 3.1 | 4.9 | 1.5 | Versicolor |

1) Clasificar un nuevo individuo (5.0 3.3 1.4 0.2), se calcula la distancia Euclídea a cada individuo del set de entrenamiento.

D(X1,X) = 0.22 D(X4,X) = 4.04

D(X2,X) = 1.04 D(X5,X) = 3.72

D(X3,X) = 0.43 D(X6,X) = 4.19

2) Para K = 1

D(X1,X) = 0.22 → Setosa

3) Para K = 2

D(X1,X) = 0.22 → Setosa

D(X3,X) = 0.43 → Setosa

NOTA:

En caso de haber valores repetidos haríamos un conjunto con los N valores repetidos.

## Redes bayesianas

| Age | Spectacle prescription | Astigmatism | Tear production rate | Recommended Lenses |
|---|---|---|---|---|
| young | myope | no | reduced | none |
| young | myope | no | normal | soft |
| young | myope | yes | reduced | none |
| young | myope | yes | normal | hard |
| young | hypermetrope | no | reduced | none |
| young | hypermetrope | no | normal | soft |
| young | hypermetrope | yes | reduced | none |
| young | hypermetrope | yes | normal | hard |
| pre-presbyopic | myope | no | reduced | none |
| pre-presbyopic | myope | no | normal | soft |
| pre-presbyopic | myope | yes | reduced | none |
| pre-presbyopic | myope | yes | normal | hard |
| pre-presbyopic | hypermetrope | no | reduced | none |
| pre-presbyopic | hypermetrope | no | normal | soft |
| pre-presbyopic | hypermetrope | yes | reduced | none |
| pre-presbyopic | hypermetrope | yes | normal | none |
| presbyopic | myope | no | reduced | none |
| presbyopic | myope | no | normal | none |
| presbyopic | myope | yes | reduced | none |
| presbyopic | myope | yes | normal | hard |
| presbyopic | hypermetrope | no | reduced | none |
| presbyopic | hypermetrope | no | normal | soft |
| presbyopic | hypermetrope | yes | reduced | none |
| presbyopic | hypermetrope | yes | normal | none |

### Edad

| | Ninguna | Blandas | Duras |
|---|---|---|---|
| Joven | 4/15 | 2/5 | 2/4 |
| Pre Presbicia | 5/15 | 2/5 | 1/4 |
| Presbicia | 6/15 | 1/5 | 1 |

### Diagnóstico

| | Ninguna | Blandas | Duras |
|---|---|---|---|
| Miope | 7/15 | 2/5 | 3/4 |
| Hipermétrope | 8/15 | 3/5 | 1/4 |

### Astigmatismo

| | Ninguna | Blandas | Duras |
|---|---|---|---|
| Sí | 8/15 | 0/5 | 4/4 |
| No | 7/15 | 5/5 | 0/4 |

### Lágrimas

| | Ninguna | Blandas | Duras |
|---|---|---|---|
| Normal | 3/15 | 5/5 | 4/4 |
| Reducida | 12/15 | 0/5 | 0/4 |

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$$

{Joven, Hipermétrope, No astigmatismo, Lágrima normal}

$$P(Blandas|E) = \frac{P(\{Joven, Hipermétrope, No\ astigmatismo, Lágrima\ normal\}|Blandas) \cdot P(Blandas)}{P(E)}$$

$$P(Ninguna|E) = \frac{P(\{Joven, Hipermétrope, No\ astigmatismo, Lágrima\ normal\}|Ninguna) \cdot P(Ninguna)}{P(E)}$$

$$P(Duras|E) = \frac{P(\{Joven, Hipermétrope, No\ astigmatismo, Lágrima\ normal\}|Duras) \cdot P(Duras)}{P(E)}$$

## Support Vector Machines

**(Mirar en las diapositivas)**

# Tema 6: Regresión

## Objetivo

Obtener estimaciones razonables de Y para distintos valores de X a partir de una muestra de n pares de valores $(x_1, y_1) \dots (x_n, y_n)$.

## Definiciones

Variable independiente (X): también llamada explicativa o exógena. La que se usa para predecir.

Variable dependiente (Y): también llamada respuesta o endógena. La que se predice.

Análisis de correlación: fuerza y dirección de la relación lineal.

Análisis de regresión: predice o estima una variable en función del valor de otra variable.

Correlación: existe una dependencia entre las variables. Determina cuál es el grado de dependencia y tipo de relación.

- Determinista: Conociendo el valor de X, el valor de Y queda perfectamente establecido.
- No determinista: Conociendo el valor de X, el valor de Y no queda perfectamente establecido.
- Lineal positiva: Cuando la función es lineal y ascendente.
- Lineal negativa: Cuando la función es lineal y descendente.
- No correlados: La dispersión entre individuos es tan grande que es imposible encontrar una relación.

Regresión: permite estimar los valores de Y a partir de X.

Coeficiente de correlación: cuantifica la intensidad de la relación lineal entre dos variables. El valor es independiente de las unidades utilizadas por las variables y es sensible a las anomalías.

- Formulas:
  - Muestra: $r\ Pearson = \frac{cov\ muestra\ (X,Y)}{\sqrt{s_x{}^2 \cdot sy^2}} = \frac{\sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}}{\sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} \cdot \sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{n-1}}}$
  - Población: $p = \frac{cov(X,Y)}{\sqrt{var(X) \cdot var(Y)}} = \frac{E(X-\mu_x)(Y-\mu_y)}{\sqrt{E[(X-\mu_x)^2] \cdot E[(Y-\mu_y)^2]}}$
- Interpretación:
  - r Pearson:
    - = 0: no hay correlación lineal.
    - > 0 and ≥1: variables directamente proporcionales.
    - < 0 and ≤-1: variables inversamente proporcionales.

## Regresión lineal simple

### Formulas

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$S_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

$$S_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}$$

$$S_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n}$$

### Ejemplo

| Id | Edad | Peso |
|---|---|---|
| 1 | 2 | 14 |
| 2 | 3 | 20 |
| 3 | 5 | 32 |
| 4 | 7 | 42 |
| 5 | 8 | 44 |

- Sumatorios

| Xi | Yi | Xi^2 | Yi^2 | Xi*Yi |
|---|---|---|---|---|
| 2 | 14 | 4 | 196 | 28 |
| 3 | 20 | 9 | 400 | 60 |
| 5 | 32 | 25 | 1024 | 160 |
| 7 | 42 | 49 | 1764 | 294 |
| 8 | 44 | 64 | 1936 | 352 |
| **25** | **152** | **151** | **5320** | **894** |

1. Tipo de correlación. (Aplicamos las fórmulas)

$$\bar{x} = 5$$

$$\bar{y} = 30.4$$

$$S_x^2 = \frac{151}{5} - 5^2 = 5.2$$

$$S_y^2 = \frac{5320}{5} - 30.4^2 = 139.86$$

$$S_{xy} = \frac{894}{5} - 5 * 30.4 = 26.8$$

$$r\ Pearson = \frac{26.8}{\sqrt{5.2 * 139.84}} = 0.99$$

Como r Pearson > 0, correlación lineal muy alta positiva.

2. Modelos de regresión.

$$x - 5 = \frac{S_{xy}}{S_y^2} * (y - 30.4) \rightarrow x = 0.192y - 0.76$$

$$y - 30.4 = \frac{S_{xy}}{S_x^2} * (x - 5) \rightarrow y = 5.15x + 4.56$$

## Regresión lineal múltiple

### Ejemplo

$$\begin{bmatrix} 5 & 14 & 193 \\ 14 & 46 & 549 \\ 193 & 549 & 7477 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 1625.5 \\ 4862.9 \\ 63196.9 \end{bmatrix}$$

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 5 & 14 & 193 \\ 14 & 46 & 549 \\ 193 & 549 & 7477 \end{bmatrix}^{-1} \begin{bmatrix} 1625.5 \\ 4862.9 \\ 63196.9 \end{bmatrix}$$

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 76.650 & 2.3045 & -2.1477 \\ 2.3045 & 0.2450 & -0.0775 \\ -2.1477 & -0.0775 & 0.0613 \end{bmatrix} \begin{bmatrix} 1625.5 \\ 4862.9 \\ 63196.9 \end{bmatrix}$$

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \frac{228581}{3125} \\ \frac{79231}{2000} \\ \frac{600887}{100000} \end{bmatrix} \rightarrow y = \frac{228581}{3125} + \frac{79231}{2000} x_1 + \frac{600887}{100000} x_2$$

**El resto del tema mirarlo directamente de las diapositivas.**

# Tema 7: Clustering