UNIVERSIDAD NACIONAL DE LA MATANZA

INTELIGENCIA DE NEGOCIOS

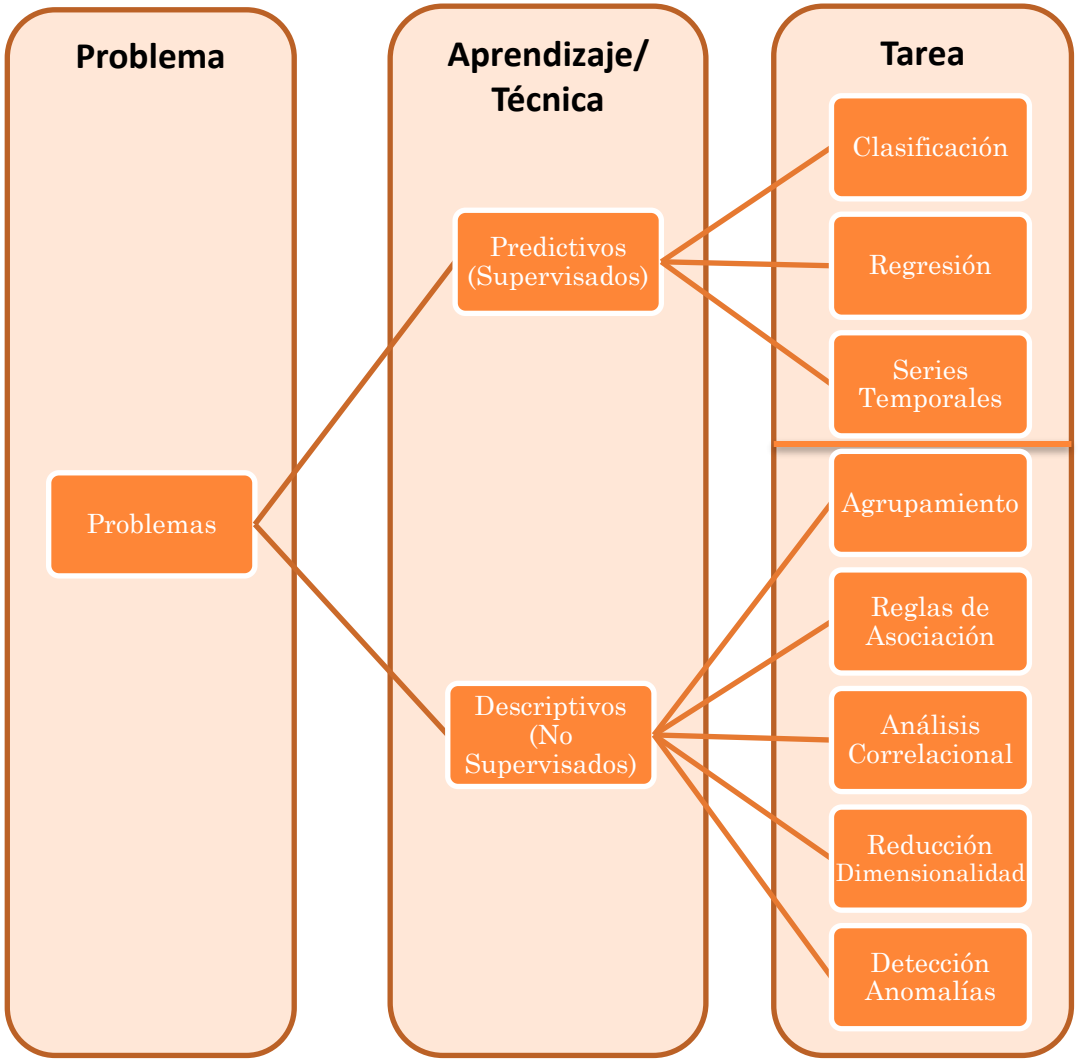
Tecnologías Inteligentes
para Explotación de Información

Docentes: ING. LORENA R. MATTEO

Autores ppt orig.: Lic. HUGO M. CASTRO / MG. DIEGO BASSO

Fecha última actualización.: 10/06/2025

Técnicas de Minería de Datos

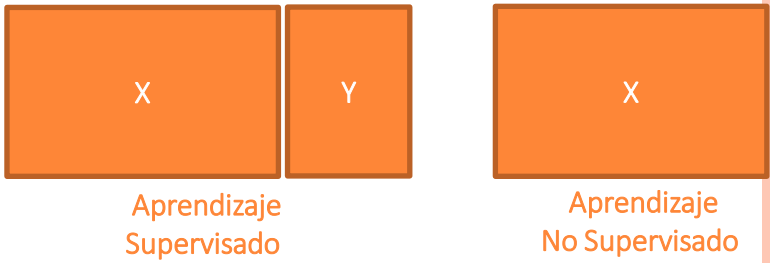


Las técnicas de minería de datos son herramientas que facilitan el descubrimiento de conocimiento.

Modelos Descriptivos

- Tareas de Segmentación (Clustering)
  - Agrupamiento jerárquico o no jerárquico de datos de acuerdo a un determinado criterio.
    - Jerárquico: Puede ser aglomerativo o divisivo.
    - No Jerárquico: N° Grupos determinados de antemano.
- Tareas de Asociación
  - Descubren por medio de reglas de asociación hechos que ocurren en común dentro de un determinado conjunto de datos.
  - Utilizado en análisis de canasta (market basket analysis).
    - {cebollas, vegetales} $\Rightarrow$ {carne}
    - {cerveza} $\Rightarrow$ {leche, pañales}

Modelos Descriptivos



Supervisado vs No Supervisado:

- **Aprendizaje Supervisado:**
  - X (variables de entrada) e Y (variable objetivo): conocidos
  - variable objetivo y sus valores específicos sirven para entrenar modelos y predecir nuevo caso
- **Aprendizaje No Supervisado:**
  - Solo conozco X (variables de entrada), el conjunto de características del caso no existe una variable objetivo predefinida, podría descubrirse dentro de los ejemplos
  - Se busca describir los datos, descubriendo agrupaciones naturales presentes en las variables de entrada.

Tareas de Segmentación (Clustering)

- ¿Cuándo usar esta tarea?
  - cuando sea necesario dividir los datos en grupos significativos y/o útiles, sin perder la estructura natural de los datos
  - muchas veces es sólo un punto de partida
- ¿Para qué usar esta tarea? Para encontrar….
  - agrupamientos naturales y describir sus propiedades
  - agrupamientos útiles y descubrir posible clase
  - correlación entre las variables
  - representantes para grupos homogéneos, reducción de dim.
  - objetos inusuales (outliers)
  - perturbaciones aleatorias de los datos (noise)

Tareas de Segmentación (Clustering)

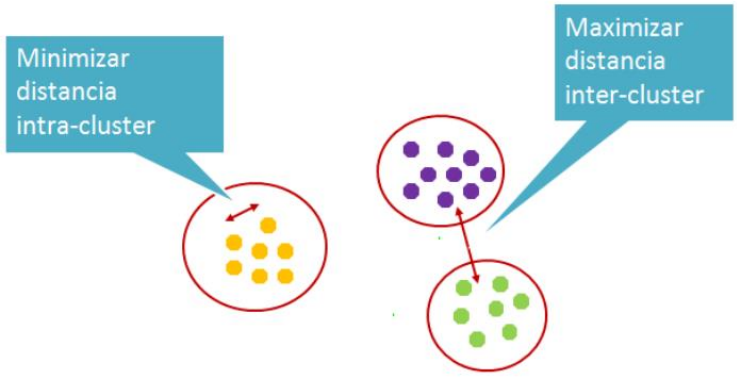
- ¿Qué aplicaciones típicas tiene esta tarea?
  - ***Marketing e Investigación de Mercado:*** descubrimiento de clientes con hábitos de compra similares, esquemas de comportamiento, se agrupan ➔ dirigir campañas: ofertas, publicidades, fidelidad, etc.
  - ***Seguros:*** identificación de grupos de asegurados con características parecidas (siniestros, posesiones, etc.) ➔ ofertar productos que otros clientes de ese grupo ya poseen.
  - ***Fraudes:*** detectar puntos fuera del común de los grupos identificados.
  - ***Planificación urbana:*** identificación de grupos de viviendas de acuerdo a tipo, valor o situación geográfica.
  - ***Web Mining:*** clasificación de documentos, análisis de logs para descubrir patrones de acceso similares.
  - ***Procesamiento de imágenes:*** reconocimiento de patrones, búsqueda de objetos en imágenes, áreas con cierta característica de tierra (GIS), …
  - ***Preprocesamiento****:* para otras técnicas de DM (maldición de la dimensionalidad, correlación de variables, etc.)

Tareas de Segmentación (Clustering)

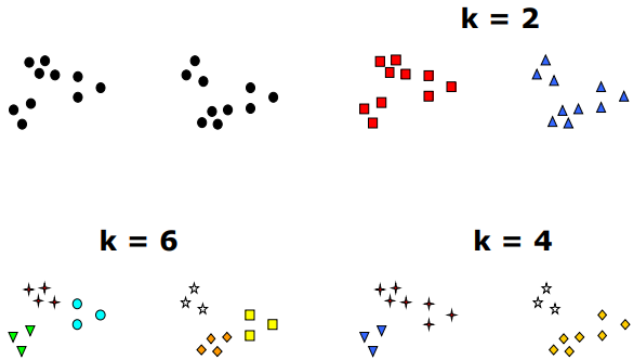
- ¿En qué consiste esta tarea?
  - Dado un conjunto de datos (puntos heterogéneos) se busca encontrar un número (k) de subgrupos homogéneos (clúster) con características similares, donde los miembros sean:
    - similares a cualquier otro incluido en el mismo grupo (están cerca y relacionados)
    - distintos a los incluidos en otros grupos (están alejados)
  - La similitud se define utilizando una medida de distancia:
    - Euclidiana, Minkowski, Jaccard, Coseno, etc.
  - En 2 dimensiones parece una tarea sencilla, pero en la realidad estos puntos están en un espacio de alta dimensionalidad.

Tareas de Segmentación (Clustering)

- ¿Qué debe tenerse en cuenta?
  - Un buen método de clustering debe identificar clústeres que sean tanto compactos somo separados entre sí:
    - Alta similaridad intra-cluster
    - Baja similaridad inter-cluster



  - La decisión del número de clústeres (k) es uno de los retos en agrupamiento, existen distintos métodos para su elección:
    - Método del Codo
    - Conocimiento del campo
    - Decisión de negocios



Tareas de Segmentación (Clustering)

- ¿Qué se espera de este método?
  - Un buen método de agrupamiento debiera:
    - descubrir algunos o todos los patrones ocultos en los datos
    - ser fácil de interpretar
    - ser escalable
    - ser insensible al orden de los registros de entrada
    - ser válido para registros de alta dimensionalidad
    - brindar la capacidad de incorporar restricciones del usuario
    - permitir dar peso a ciertas variables dependiendo de distintos criterios (relativos a su aplicación, …)
  - La calidad del método de clustering depende de la medida de similitud y de su implementación
  - Las funciones de distancia son muy sensibles al tipo de variables.

Tareas de Segmentación (Clustering)

- ¿Cómo medir la similitud entre instancias?
  - Las medidas de distancia/similitud dependen del tipo de variable y son no negativas

Variables Numéricas

Distancia Euclídea
Distancia de Manhattan
Distancia Minkowski

Normalizar/Estandarizar

- z-score
- min-max
- decimal scale

Variables Binarias

Coeficiente de Jaccard

Variables Categóricas

Distancia: 1 si los valores son diferentes, 0 si son iguales

¿Tienen todos los atributos la misma importancia?

- Si no tienen igual importancia, será necesario ponderar los atributos

Tipos de Segmentación (Clustering)

- **No Jerárquico (Particional)**

Divide los datos en subconjuntos (clústeres) sin solapamiento, ⇔ cada dato está en un solo subconjunto

  - nro de grupos (k) determinado de antemano.
  - cada punto pertenece al grupo de "más cercano“
  - *Global Optimal, K-Means, K-Modes, K-Medoids (PAM)*

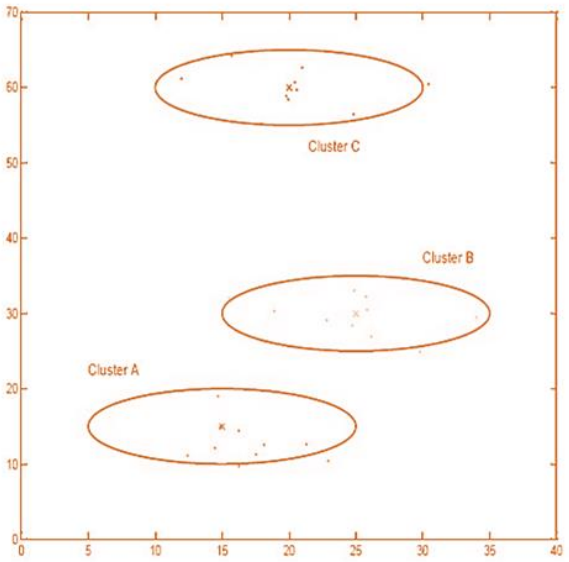
- **Jerárquico**

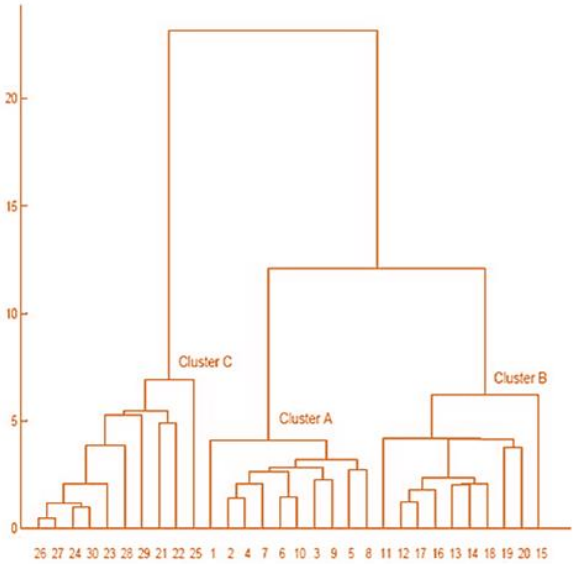
Construye un árbol binario o dendograma a partir de un conjunto de ejemplos

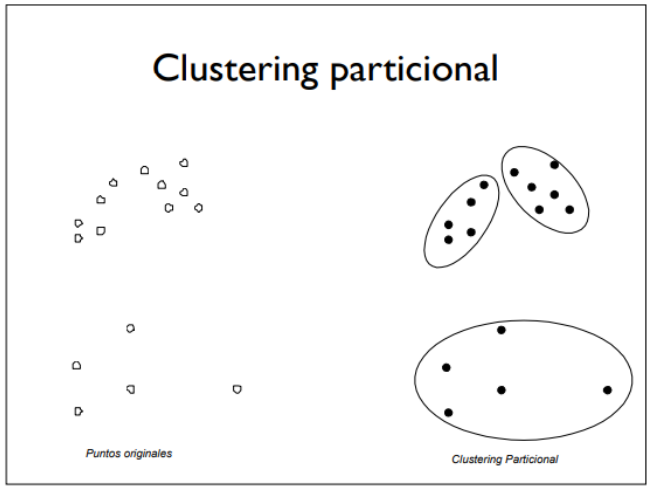
  - **Aglomerativo** (bottom-up - de abajo hacia arriba)
    - Inicialmente, cada punto es un clúster
    - Combina repetidamente los dos grupos "más cercanos" en uno, para formar clústeres mayores
    - *AGNES*
  - **Divisivo** (top-down – de arriba hacia abajo)
    - Comienza con un clúster y de forma recursiva lo va dividiendo.
    - *DIANA*

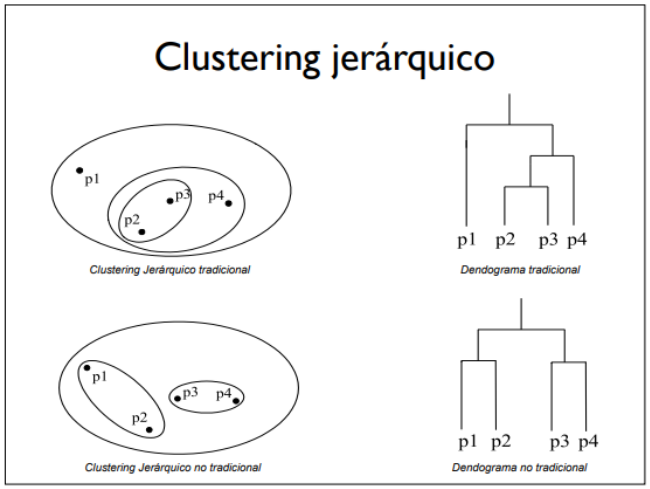
Tipos de Segmentación (Clustering)

- **No Jerárquico (Particional) vs Jerárquico**









Tipos de Segmentación (Clustering)

- **Basados en densidad**

Utilizan funciones de conectividad y densidad

  - *DBSCAN, OPTICS, DENCLUE, CLIQUE*

- **Basados en rejillas (Grilla)**

Utilizan una estructura de granularidad de múltiples niveles

  - *STING, BAN-Clustering/GRIDCLUS, Wave-Clustering, CLIQUE*

- **Basados en datos categóricos**
  - *ROCK*
- **Basados en modelos**

Suponen un modelo para cada uno de los clústeres con el propósito de encontrar el modelo que mejor ajuste.

  - *Redes Neuronales:* SOM: Self Organizing Maps
  - *Machine Learning:* COBWEB (clustering conceptual)
  - *Estadísticos:* Gaussian Mixture Model, Autoclass (Bayesiano)
  - *Lógica Difusa:* Fuzzy C-Means

Modelos Descriptivos - En resumen…

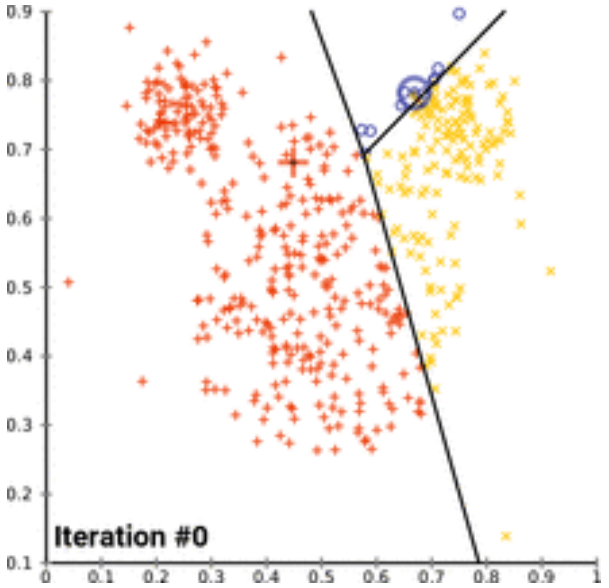
- No realizan predicciones.
- No hay una variable objetivo identificada en el conjunto de datos. En su lugar, se buscan patrones y estructuras entre todas las variables
- Analizan otros aspectos de los datos.
- Problemas tratados que veremos:
  - Segmentación o Agrupamiento (Clustering)
  - Reglas de Asociación
- Tecnologías utilizadas que veremos:
  - K-Means
  - Redes Neuronales SOM
  - Algoritmo Apriori (descubrimiento de ítems frecuentes)

Tareas de Segmentación K-Means

¿De qué se trata?

- K-Means o K-Medias es un algoritmo de clustering no supervisado que agrupa datos similares en K clústeres para descubrir patrones ocultos.

Algoritmo K-Means básico
1: **Select k** puntos/objetos como el centroide inicial
2: **repeat**
3: **From k** clústeres asignando todos los puntos a centroide más cercano.
4: Recalcular el centroide de cada clúster
5: **until** los centroides no cambien.



Centroides: Puntos centrales de cada clúster.

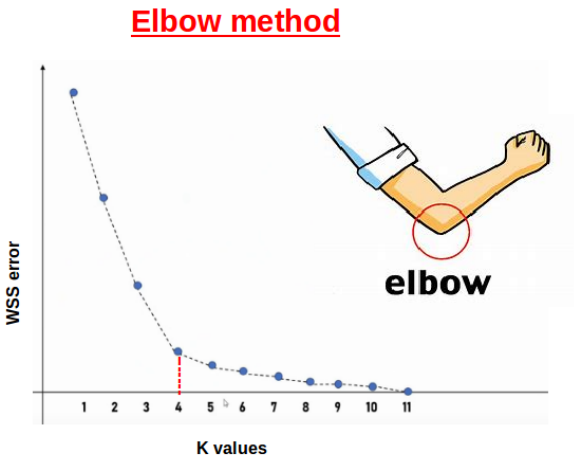
Clústeres: Grupos de datos alrededor de los centroides.

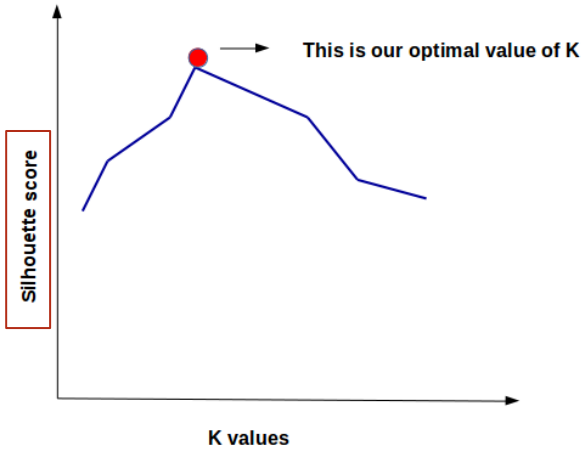
TAREAS DE SEGMENTACIÓN K-MEANS

¿Cómo seleccionar el número correcto de clústeres k?

El algoritmo de K-Means es muy sensible a la posición inicial de los centroides de los grupos, su elección requiere un delicado equilibrio. **Métodos para determinar el número óptimo de clústeres:**

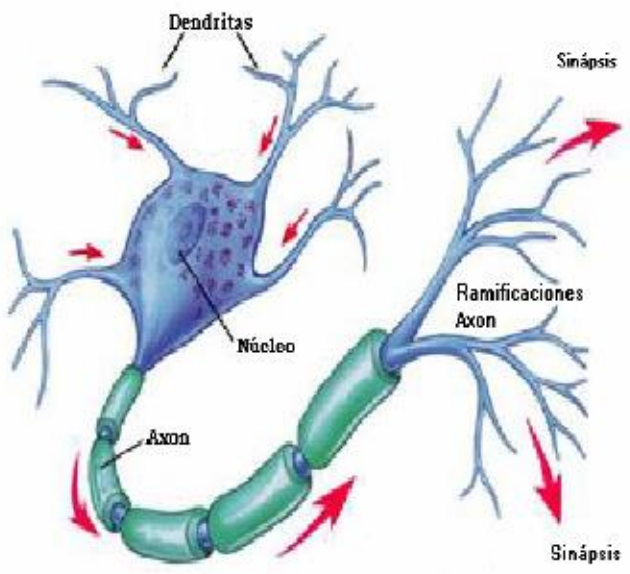
- **Método del Codo:** Se calcula la suma de las distancias al cuadrado dentro de los clústeres (WCSS) para k=x, se grafica y busca donde se forma el "codo".
- **Método de la Silueta:** Se la calcula la medida de la silueta para evaluar la calidad de la agrupación y seleccionar el K que maximice esta medida.
- **Conocimiento del Dominio:** Utilizar el conocimiento previo del problema para establecer un rango probable para K y combinarlo con métodos cuantitativos.





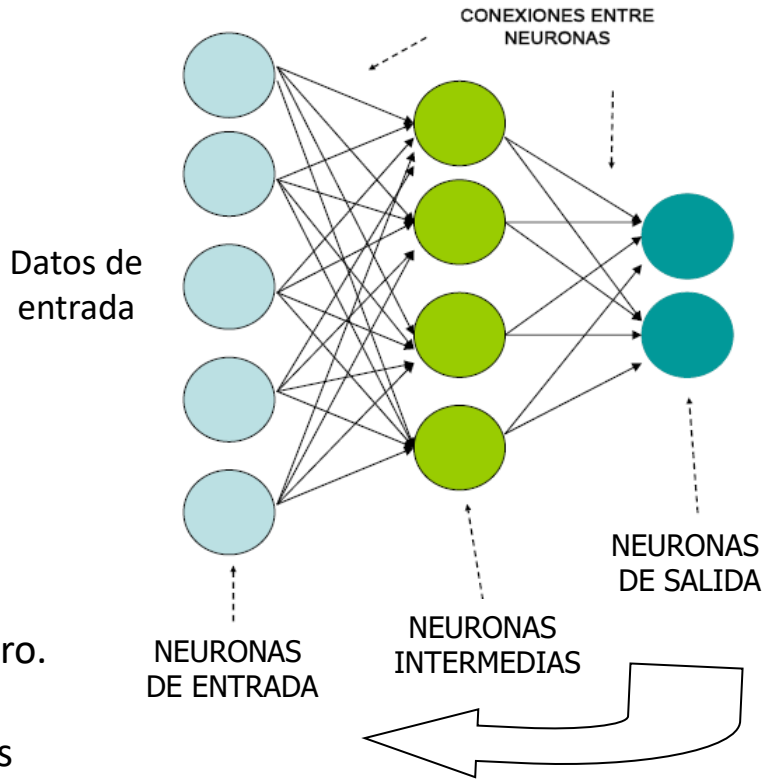
SEGMENTACIÓN REDES NEURONALES

Neurona Biológica



La neurona es la unidad fundamental del sistema nervioso y en particular del cerebro. Cada neurona es una simple unidad procesadora que recibe y combina señales desde y hacia otras neuronas. Si la combinación de entradas es suficientemente fuerte la salida de la neurona se activa

Neurona Artificial



Casos de entrenamiento con sus respuestas
Casos de Prueba

Redes Neuronales

Neuronas Biológicas vs. Artificiales

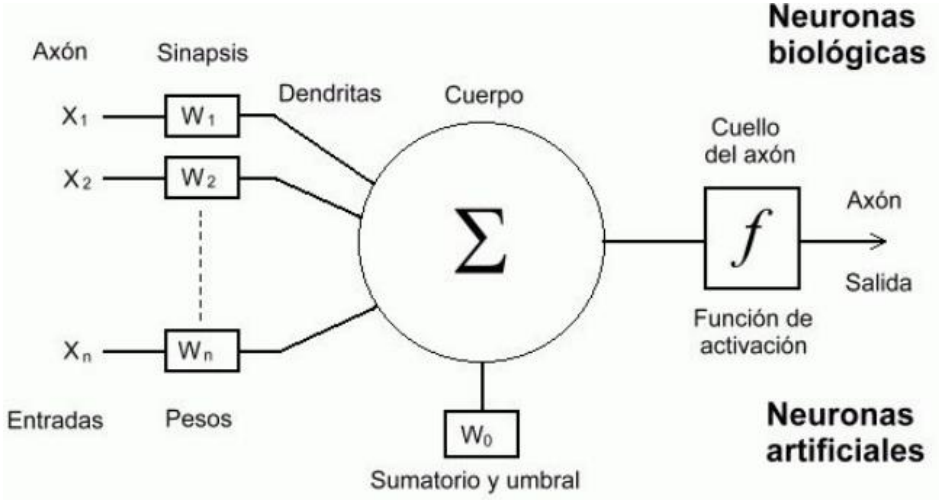
En 1956, se acuñaría el propio término “Inteligencia Artificial” por John McCarthy, Marvin Minsky y Claude Shannon en una conferencia en Darthmounth.

En 1958, Frank Rosenblatt diseña la primera red neuronal artificial, el **Perceptrón**.

Modelo neuronal con n entradas, que consta de:

- Un conjunto de **entradas** x1,…xn
- Los **pesos sinápticos** w1,…wn, correspondientes a cada entrada
- Una **función de agregación**, Σ
- **Una función de activación**, f
- Una **salida**:

$$Y = f\left(\sum_{i=0}^{n} w_i\, x_i\right)$$



Segmentación Redes Neuronales SOM

- **Mapas Auto-Organizados de Kohonen (SOM)**

¿De qué se trata?

- Los SOM son un tipo de red neuronal **no supervisada** desarrollada por Teuvo Kohonen, utilizada para reducción de dimensionalidad, visualización de datos y segmentación.

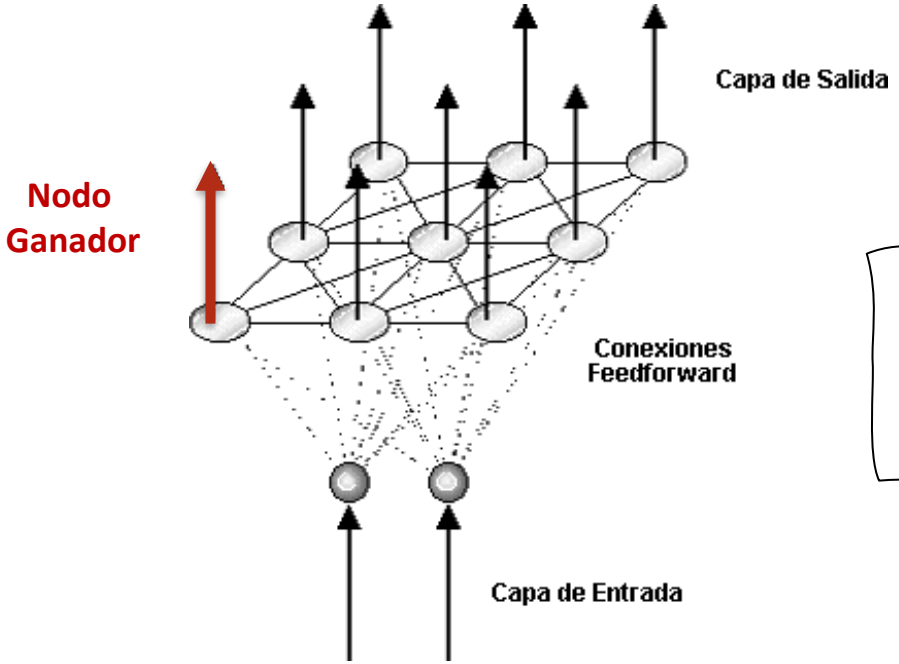
¿Cuál es su propósito?

- Transformar datos de alta dimensión en un mapa de baja dimensión (generalmente 2D) que preserva la topología de los datos originales. Frecuente los humanos aprendemos de manera no supervisada.

1. Visualización y Exploración de Datos:
•**Reducción de Dimensionalidad:** Los SOM transforman datos de alta dimensionalidad en un mapa bidimensional, lo que facilita la visualización y el análisis.
•**Preservación Topológica:** Mantienen las relaciones topológicas entre los datos, lo que significa que los datos similares en el espacio original permanecen juntos en el mapa autoorganizado resultante.

2. Agrupación Natural:
•**Clústeres Naturales:** Durante el entrenamiento, las neuronas de la red SOM se organizan de manera que datos similares se mapean cerca unos de otros en el mapa, formando clústeres naturales.
•**Identificación de Patrones:** Esto permite identificar patrones y segmentos en los datos de manera intuitiva y visual.

Redes Neuronales SOM



Neurona: Cada nodo en el SOM es una neurona con un vector de pesos.

Vecindad: Neuronas cercanas en la cuadrícula tienen pesos similares.

- Cada vez que se presenta un registro de entrada, las neuronas *"compiten"* y una se define como la ganadora.
- Si se presenta un registro de entrada parecido al anterior, es muy posible que el ganador sea el mismo nodo de salida.

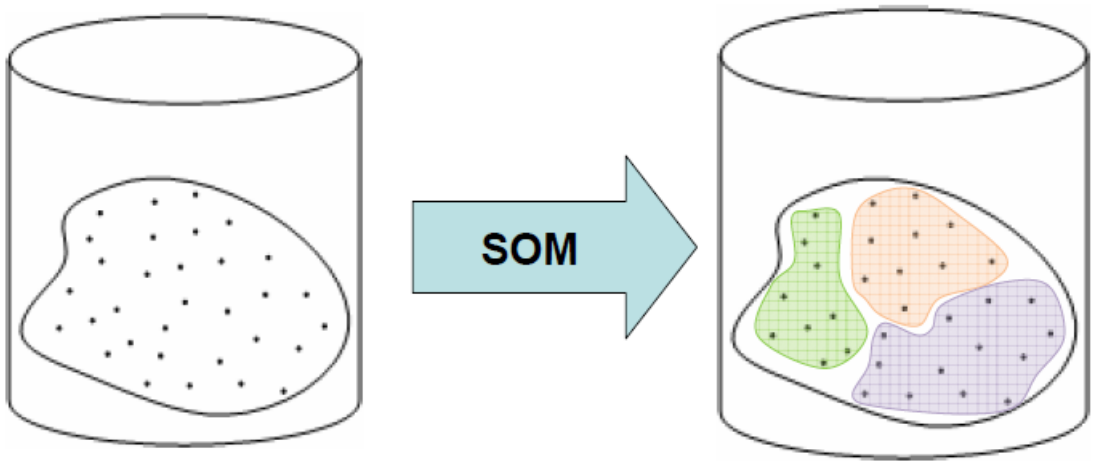
Redes Neuronales SOM

Proceso de Entrenamiento No Supervisado:

1. **Inicialización:** Los pesos de las neuronas se configuran aleatoriamente.
2. **Selección de Muestra:** Se elige una instancia del conjunto de datos de entrada.
3. **Competencia:** Las neuronas "compiten" para ver cuál se parece más a la instancia. La neurona con la distancia (en gral, euclidiana) más pequeña (la más cercana) es la "neurona ganadora" o "Best Matching Unit (BMU)".
4. **Cooperación y Actualización de Pesos:** La neurona ganadora y sus vecinas cercanas ajustan sus pesos para parecerse más al dato de entrada (instancia).
5. **Reducción de Vecindad:** A medida que avanza el entrenamiento, el tamaño de esta vecindad se reduce, lo que permite un ajuste más preciso y localizado.

Repetición: Este proceso se repite n veces con todos los datos de entrada hasta que los pesos se estabilizan y el mapa se organiza de manera útil.

Proceso Intuitivo de Agrupamiento



- Los registros semejantes van a parar a la misma categoría (clúster).
- Una vez entrenado, el mapa de Kohonen se puede usar para categorizar nuevos registros.
  - El resultado final es la creación del llamado mapa autoorganizado donde se representan los rasgos más sobresalientes del espacio de entrada.

Medidas de Evaluación de Segmentación

¿Es necesario validar los clústeres?

- En Clasificación, la validación es parte integral del proceso y tenemos medidas claras para ello: exactitud, precisión, etc.
- No así en Clustering….

¿Cómo saber si nuestros clústeres son buenos?

- No hay una repuesta absoluta
- Depende de la aplicación

Entonces, ¿para qué evaluar? Para…

- evitar encontrar patrones en el ruido
- comparar algoritmos de clustering diferentes
- comparar conjuntos de clústeres diferentes
- comprar dos clústeres

Medidas de Evaluación de Segmentación

¿Qué aspectos se consideran en la validación?

- Determinar la tendencia de agrupamiento (clustering tendency), *por ej. si existe una estructura no-aleatoria en los datos*
- Encontrar el número correcto de clústeres (k)
- Evaluar qué tan bien los resultados se ajustan a los datos (sin consultar datos externos)
- Comparar resultados con resultados externos, *por ej. clases asignadas manualmente*
- Comparar dos conjuntos de clústeres para saber cuál es mejor

¿Qué tipos de validación se puede emplear?

- Métricas de Validación Interna (Internal Index)
- Métricas de Validación Externa (External Index)
- Métricas de Validación Relativa (Relative Index)
- Otros: Validación usando correlación, Validación con Expertos, Enfoque visual.

Medidas de Evaluación de Segmentación

Métricas de Validación Interna

Evalúan que tan buena es la estructura del clustering sin necesidad de información ajena al propio algoritmo y su resultado, únicamente basadas en información de los datos.

- **Cohesión**: mide qué tan cercanos son los objetos en un clúster. *Por ej.: Sum of Squared Within (SSW/SSE)*
- **Separación**: mide qué tan diferente o bien separado es un clúster de otros. Hay varios enfoques para medir esta distancia entre clústeres: distancia entre el miembro más cercano, distancia entre los miembros más distantes o la distancia entre los centroides. *Por ej.: Sum of Squared Between (SSB)*
- **Coeficiente de Silhouette:** es una métrica para evaluar la calidad del agrupamiento obtenido con algoritmos de clustering, su objetivo es identificar cuál es el número óptimo de agrupamientos.

Medidas de Evaluación de Segmentación

Métricas de Validación Externa

Una vez finalizado el algoritmo de agrupación, se compara el clúster en el que fue asignado cada elemento, con la etiqueta de clase que traía de antemano (información externa, no siempre disponible).

- **Pureza**: nivel en que un clúster contiene elementos de una sola clase (se usa la clase predominante)
- **Entropía**: cantidad de clases diferentes que contiene un clúster

Métricas de Validación Relativa

Se utiliza para comparar dos agrupaciones diferentes o clúster. A menudo se utiliza un índice externo o interno para esta función, *por ejemplo, SSW/SSE o Entropía*

Validación con Expertos

Se pueden evaluar los clústeres para ver si producen el resultado esperado y comparar con otras soluciones • Se puede generar una clasificación de validación.

Segmentación (Clustering) - Conclusiones

- Aprendizaje No Supervisado
- No realizan predicciones
- Admite muchos enfoques
- La calidad del agrupamiento depende de la medida de similitud utilizada por el método y de su implementación.
- La calidad de un método de agrupamiento también se mide por su capacidad para descubrir algunos o todos los patrones ocultos.
- La evaluación objetiva es problemática: generalmente realizado por inspección humana / experta.

Tareas de **Asociación**

¿En qué consiste esta tarea?

- Descubre por medio de reglas de asociación hechos que ocurren en común dentro de un determinado conjunto de datos:
  - patrones frecuentes, asociaciones, correlaciones o estructuras causales entre conjuntos de artículos u objetos
  - secuencias o patrones temporales

¿Qué aplicaciones típicas tiene esta tarea?

- Análisis de Canasta (Market Basket Analysis)
  - *Análisis de clientes:* Se utiliza información de las compras de un cliente para ofrecer una aproximación sobre quién es y porqué hace ciertas compras (comportamiento)
  - *Análisis de productos:* Aporta información sobre qué productos tienden a ser comprados juntos.
- Diseño de Catálogos

Tareas de **Asociación**

Reglas de Asociación

- Dado un conjunto de transacciones se quiere encontrar reglas que puedan predecir la ocurrencia de un ítem a partir de otros ítems de la transacción , con un mínimo de confianza y soporte..
- Análisis de Canasta (Market Basket Analysis)

Ticket ID	Ítems
1	Pan, Leche
2	Pan, Pañales, Cerveza, Huevos
3	Leche, Pañales, Cerveza, Gaseosa
4	Pan, Leche, Pañales, Cerveza
5	Pan, Leche, Pañales, Gaseosa

Ejemplos de Reglas

{Pañales} → {Cerveza}

{Cerveza, Pan} → {Leche}

Soporte ≥ minSupp
Confianza ≥ minConf

- Reglas de Asociación X → Y representan implicancias.
  - Antecedente → Consecuente

Reglas de Asociación

Métricas de Evaluación

- Cobertura de la Regla (Soporte/Support): Proporción de casos a los que se le puede aplicar cada regla. Es decir, el número de instancia que la regla predice correctamente.

$$\text{COBER}(\text{R}) = \frac{\text{N°casos que satisfacen la aplicación de la regla R}}{\text{N°casos totales de la clase}}$$

  - N.º casos R: transacciones que contienen todos los ítems del antecedente o del consecuente. Toma valores entre 0 y 1.
- Reglas con mayor cobertura:
  - Representativas y útiles para obtener características que definen el comportamiento de una clase.
  - Credibilidad e interés del modelo para clasificar nuevos casos a una clase.
- Aplicable también a modelos de clasificación.

Reglas de Asociación – Ejemplo Cobertura

- Ítems de compras:

Ticket ID	Ítems
1	Pan, Leche
2	Pan, Pañales, Cerveza, Huevos
3	Leche, Pañales, Cerveza, Gaseosa
4	Pan, Leche, Pañales, Cerveza
5	Pan, Leche, Pañales, Gaseosa

- *Numerador* = Cobertura (Cant. Casos R)
- *Denominador* = Cant. Casos Totales

Min. soporte 30%
Min. confianza 50%

- Consideremos las siguientes reglas:
  - {Leche, Pañales} → {Cerveza} Cober(R) = 2/5 = 0.4 (2=Tickets 3y4)
  - {Leche, Cerveza} → {Pañales} Cober(R) = 2/5 = 0.4 (2=Tickets 3y4)
  - {Pañales, Cerveza} → {Leche} Cober(R) = 2/5 = 0.4 (2=Tickets 3y4)
  - {Cerveza} → {Leche, Pañales} Cober(R) = 2/5 = 0.4 (2=Tickets 3y4)
  - {Pan} → {Pañales, Gaseosa} Cober(R) = 1/5 = 0.2 (1=Ticket 5)

Reglas de Asociación

- Precisión de la Regla (Confianza/Confidence): Proporción de casos que cumplen con la regla respecto del total de casos considerados en la precondición de la misma.

$$\text{Prec(R)} = \frac{\text{N°casos que satisfacen la aplicación de la regla R}}{\text{N°casos que satisfacen la precondición}}$$

  - una transacción que contiene el antecedente y también el consecuente. Toma valores entre 0 y 1.

- Cuanto mayor sea la precisión de una regla más confiable e interesante resulta para asociar ítems y descubrir nuevo conocimiento.
- Aplicable también a modelos de clasificación.

Reglas de Asociación – Ejemplo Precisión

- Utilizando los mismos ítems de compra:

Ticket ID	Ítems
1	Pan, Leche
2	Pan, Pañales, Cerveza, Huevos
3	Leche, Pañales, Cerveza, Gaseosa
4	Pan, Leche, Pañales, Cerveza
5	Pan, Leche, Pañales, Gaseosa

- *Numerador* =Idem Cober (Cant. Casos R)
- *Denominador* =Casos PreCondición

- Considerando las reglas anteriores:
  - {Leche, Pañales} → {Cerveza} Prec(R) = 2/3 = 0.67 (3=Tickets 3,4,5)
  - {Leche, Cerveza} → {Pañales} Prec(R) = 2/2 = 1 (2=Tickets 3,4)
  - {Pañales, Cerveza} → {Leche} Prec(R) = 2/3 = 0.67 (3=Tickets 2,3,4)
  - {Cerveza} → {Leche, Pañales} Prec(R) = 2/3 = 0.67 (3=Tickets 2,3,4)
  - {Pan} → {Pañales, Gaseosa} Prec(R) = 1/4 = 0.25 (4=Tickets 1,2,4,5)

Reglas de Asociación

- Conclusiones obtenidas:
  - {Leche, Pañales} → {Cerveza} Cober(R) = 0.4 Prec(R) = 0.67
  - {Leche, Cerveza} → {Pañales} Cober(R) = 0.4 Prec(R) = 1
  - {Pañales, Cerveza} → {Leche} Cober(R) = 0.4 Prec(R) = 0.67
  - {Cerveza} → {Leche, Pañales} Cober(R) = 0.4 Prec(R) = 0.67
  - {Pan} → {Pañales, Gaseosa} Cober(R) = 0.2 Prec(R) = 0.25
- La regla {Leche, Cerveza} → {Pañales} es la que mejor describe características de consumo, y la que genera mayor confianza e interés para descubrir patrones de compra frecuente.
- Reglas con alta precisión, pero baja cobertura son irrelevantes y de poco interés en un modelo de explotación de información.

Reglas de Asociación – Proceso de Extracción de Reglas

Encontrar todas las reglas X & Y ➔ Z con un mínimo de confianza y soporte.

- Cobertura (Soporte) (s): probabilidad de que una transacción contenga {X & Y & Z}
- Precisión (Confianza) (c): probabilidad condicional P(Z|X&Y)

# transacción	artículos
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Sea el valor mínimo para confianza y soporte 50%,

$A \Rightarrow C$ (50%, 66.6%)

$C \Rightarrow A$ (50%, 100%)

Reglas de Asociación – Proceso de Extracción de Reglas

Al tratar con grandes volúmenes de datos, el proceso se descompone en dos pasos:

- Encontrar conjuntos de artículos frecuentes
  - Mayor ocurrencia que el soporte mínimo fijado.
- Generar reglas de asociación “fuerte” a partir de los conjuntos de artículos frecuentes
  - Deben satisfacer el mínimo fijado tanto para soporte como para confianza.

Reglas de Asociación – Proceso de Extracción de Reglas – Ejemplo 1

Encontrar todas las reglas X & Y ➔ Z con un mínimo de confianza y soporte.

- Cobertura (Soporte) (s): probabilidad de que una transacción contenga {X & Y & Z}
- Precisión (Confianza) (c): probabilidad condicional P(Z|X&Y)

EJEMPLO 1

# transacción	artículos
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Frequent Itemset	Cobertura (Soporte)	Cálculo
{A}	75%	=3/4
{B}	50%	=2/4
{C}	50%	=2/4
{A,C}	50%	=2/4

Min. soporte 50%
Min. confianza 50%

	Cobertura ({ A &C })	Cálculo	Confianza (Precisión) = Cobertura({ A &C })/Cobertura({ A })	Cálculo
Para la regla A ➔ C	**50%**	=2/4	**66,67%**	=(2/4) / (3/4) = 0,5/0,75 = 0,666666667 **o bien =2/3**
Para la regla C ➔ A	**50%**	=2/4	**100%**	=(2/4) / (2/4) = 0,5/0,5= 1 **o bien =2/2**

Reglas de Asociación – Proceso de Extracción de Reglas – Ejemplo 2

Encontrar todas las reglas X & Y ➔ Z con un mínimo de confianza y soporte.

- Cobertura (Soporte) (s): probabilidad de que una transacción contenga {X & Y & Z}
- Precisión (Confianza) (c): probabilidad condicional P(Z|X&Y)

EJEMPLO 2

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

I = {Beer, Bread, Jelly, Milk, PeanutButter}

Min. soporte 60%

Cobertura (Soporte) {Bread,PeanutButter}	Cálculo
60%	=3/5

$X \Rightarrow Y$	s	α
Bread ⇒ PeanutButter	60%	75%
PeanutButter ⇒ Bread	60%	100%
Beer ⇒ Bread	20%	50%
PeanutButter ⇒ Jelly	20%	33.3%
Jelly ⇒ PeanutButter	20%	100%
Jelly ⇒ Milk	0%	0%

Cobertura (Soporte)	Cálculo	Confianza(X ➔ Y) = Soporte (X,Y) / Soporte (X)	Cálculo
60%	=3/5	**75%**	=(3/5) / (4/5) = 0,6/0,8 = 0,75 **o bien =3/4**
60%	=3/5	**100%**	=(3/5) / (3/5) = 0,6/0,6 = 1 **o bien =3/3**
20%	=1/5	**50%**	=(1/5) / (2/5) = 0,2/0,4 = 0,5 **o bien =1/2**
20%	=1/5	**33,3%**	=(1/5) / (3/5) = 0,2/0,6 = 0,333 **o bien =1/3**
20%	=1/5	**100%**	=(1/5) / (1/5) = 0,2/0,2 = 1 **o bien =1/1**
0%	=0/5	**0%**	=(0/5) / (1/5) = 0/0 = 0 **o bien =0/1**

Interpretación de las métricas de Evaluación

- **Regla con Soporte (Cobertura)**
  - Alto ➔ Representativas y útiles para obtener características que definen el comportamiento de una clase. | Credibilidad e interés del modelo para clasificar nuevos casos a una clase.
  - Bajo ➔ puede haber aparecido por casualidad. | Poco interesante desde el punto de vista del negocio. | Sirve para eliminar reglas poco interesantes.
- **Regla con Confianza (Precisión)**
  - Alta ➔ Cuanto mayor sea la precisión de una regla más confiable e interesante resulta para asociar ítems y descubrir nuevo conocimiento. | Mayor probabilidad de observar Y en transacciones que tengan X.
  - Baja ➔ Es probable que no exista relación entre antecedente y consecuente.

Valores típicos: Soporte = 2-10 % | Confianza = 70-90 %

El **primer criterio de selección de reglas del algoritmo "A priori"** es la **precisión o confianz**a, dada por el porcentaje de veces que instancias que cumplen el antecedente cumplen el consecuente, pero el **segundo es el soporte**, dado por el número de instancias sobre las que es aplicable la regla. Si hay reglas de muy baja precisión (muy cerca del umbral mínimo definido) habría que considerarlas simplemente como ciertas tendencias.

Referencias y Material Complementario

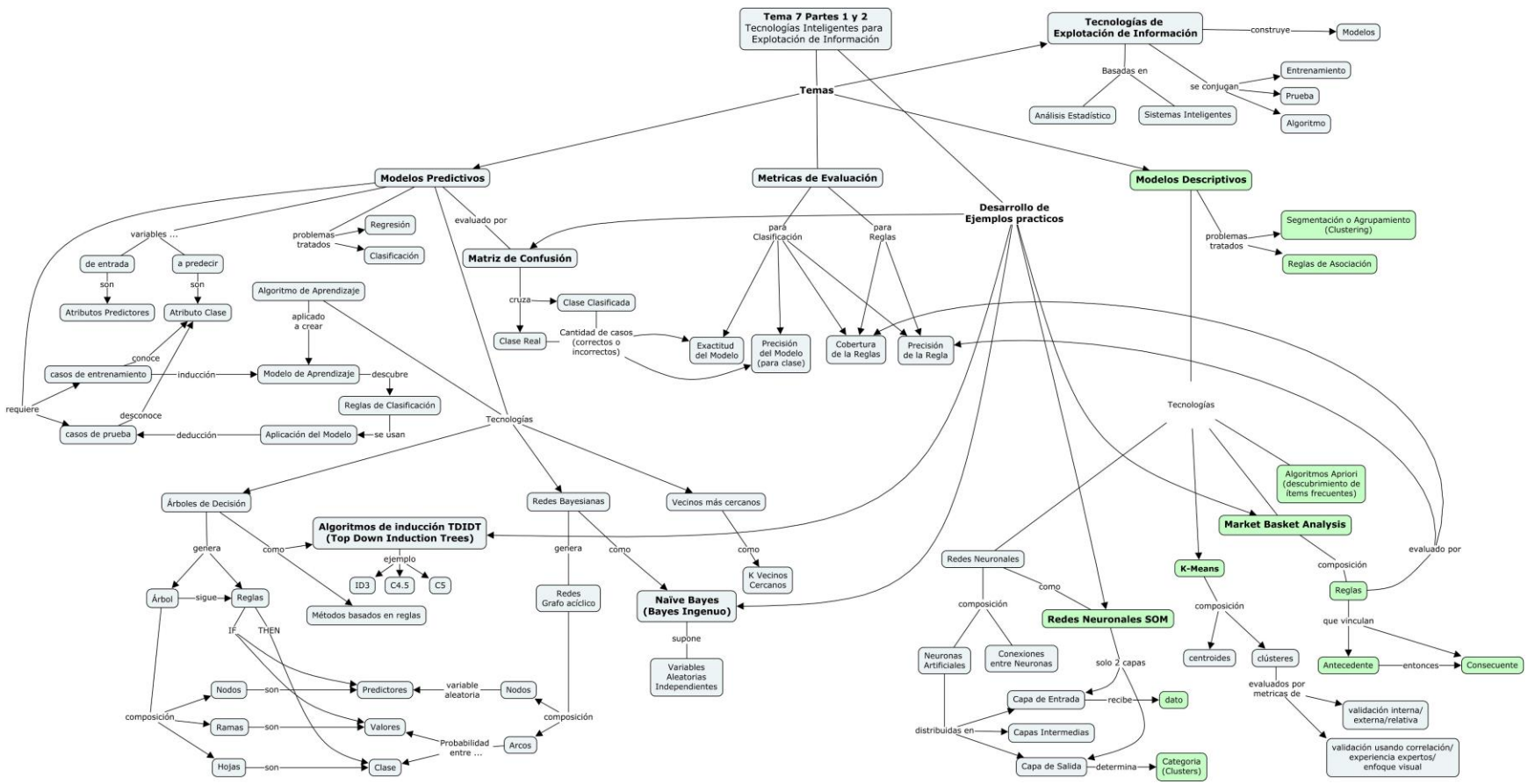
Referencias:

- Curso Inteligencia de Negocios, Universidad de Granada, España
- Curso DM Datos, Prof. F.Bravo/ H. Sarmiento / B. Poblete, Universidad de Chile
- Prof. Dra. S. Schiaffino (Univ.Centro BA-IA , Argentina
- Curso Data Mining, Kent State University, USA
- Curso Minería de Datos, Universidad de Valencia, España
- Curso Mining Massive Datasets |Stanford University, USA
- Curso Advanced Machine Learning | Educación IT , Argentina

Material Complementario:

- https://www.cs.us.es/~fsancho/Blog/posts/Aprendizaje_Supervisado_No_Supervisado.md.html
- https://www.cs.us.es/~fsancho/Blog/posts/Clustering/
- https://dcain.etsin.upm.es/~carlos/bookAA/03.1_Clustering-K-Means.html
- K-MEANS Segmentación con KNIME : https://www.youtube.com/@fbombab547 https://www.youtube.com/watch?v=Uu7t90GX_u4
- KNIME TV - Training Clustering Algorithm: https://www.youtube.com/watch?v=7luMauX0KWM

Resumen Clase Tema 7 – Parte 2



Trabajo Práctico de Minería de Datos
Caso: Créditos Bancarios
Entrega 18/06/2025

- Utilice la herramienta KNIMNE para desarrollar los procesos de explotación de información identificados en el Caso de Estudio; incluyendo tareas de Preprocesamiento, Modelos Predictivos, Descriptivos y Evaluación.

- Entregue un informe que contenga resultados, conclusiones obtenidas, gráficos, una tabla comparativa de métodos aplicados, y las recomendaciones que daría, de acuerdo con lo requerido en el enunciado del trabajo práctico.