[Unidades del Examen 2 - Inteligencia de Negocios V0.9 (Sin refinar) 3](#_Toc201749233)

[Data mining (Minería de datos) 3](#_Toc201749234)

[Conceptos relacionados 3](#_Toc201749235)

[OLAP vs Minería de datos 3](#_Toc201749236)

[Áreas de aplicación de Minería de datos 4](#_Toc201749237)

[Tareas de Mineria de Datos 4](#_Toc201749238)

[Técnica en la minería de datos 4](#_Toc201749239)

[Trabajo con Algoritmos 4](#_Toc201749240)

[Modelo predictivo – Aprendizaje supervisado 4](#_Toc201749241)

[Metodo descriptivo – Aprendizaje No supervisado 5](#_Toc201749242)

[Tareas de Asociación 5](#_Toc201749243)

[Tareas de Segmentación/Agrupamiento (Clustering) 5](#_Toc201749244)

[Aprendizaje por Refuerzo 5](#_Toc201749245)

[Explotación de Información 5](#_Toc201749246)

[Proceso Descubrimiento de Conocimiento (KDD) 5](#_Toc201749247)

[Metodologías de Explotación de Información 6](#_Toc201749248)

[Metodología CRISP-DM 6](#_Toc201749249)

[SEMMA (Sample, Explore, Modify, Model and Assess) 7](#_Toc201749250)

[Herramienta para la minería de datos 9](#_Toc201749251)

[Modelos predictivos (Aprendizaje Supervisado) 9](#_Toc201749252)

[Tareas de Clasificación 9](#_Toc201749253)

[Árboles de Decisión (Algoritmos de inducción TDIDT - Top Down Induction Trees): 10](#_Toc201749254)

[Bayes: 10](#_Toc201749255)

[Las Redes Bayesianas 10](#_Toc201749256)

[El Clasificador Bayesiano Naïve-Bayes es un tipo específico de red bayesiana. 10](#_Toc201749257)

[Proceso de aprendizaje: 10](#_Toc201749258)

[Evaluación de Modelos de Clasificación: 11](#_Toc201749259)

[La Matriz de Confusión 11](#_Toc201749260)

[Métricas clave de evaluación: 11](#_Toc201749261)

[KNIME 12](#_Toc201749262)

[Características y Funcionamiento de KNIME: 12](#_Toc201749263)

[Funcionalidades de KNIME: 13](#_Toc201749264)

[Creación de Workflows: 13](#_Toc201749265)

[Modelos Descriptivos (Aprendizaje No Supervisado) 14](#_Toc201749266)

[Tareas de Segmentación (Clustering): 14](#_Toc201749267)

[Propósito: 14](#_Toc201749268)

[Aplicaciones Típicas: 14](#_Toc201749269)

[Funcionamiento: 14](#_Toc201749270)

[Tipos de Clustering: 15](#_Toc201749271)

[Mapas Auto-Organizados de Kohonen (SOM): 17](#_Toc201749272)

[Evaluación de Segmentación: 17](#_Toc201749273)

[Métricas Internas: 17](#_Toc201749274)

[Métricas Externas: 17](#_Toc201749275)

[Validación con Expertos: 17](#_Toc201749276)

[Tareas de Asociación: 17](#_Toc201749277)

[Propósito 17](#_Toc201749278)

[Aplicaciones Típicas: 17](#_Toc201749279)

[Reglas de Asociación: 18](#_Toc201749280)

[Métricas de Evaluación: 18](#_Toc201749281)

[Cobertura (Soporte/Support): 18](#_Toc201749282)

[Precisión (Confianza/Confidence): 18](#_Toc201749283)

[Proceso de Extracción de Reglas: 18](#_Toc201749284)

[Calidad de los modelos de minería de datos 19](#_Toc201749285)

[¿Qué hace que un modelo sea bueno? Un buen modelo debe: 19](#_Toc201749286)

[El Problema del Sobreaprendizaje (Overfitting): 19](#_Toc201749287)

[Generalización - Análisis de Errores: 19](#_Toc201749288)

[Selección de Algoritmos de Machine Learning: 20](#_Toc201749289)

[Fundamentos de la IA 20](#_Toc201749290)

[Historia de la IA: 20](#_Toc201749291)

[Inteligencia Artificial Generativa (IAGen): 21](#_Toc201749292)

[Large Language Models (LLMs): 21](#_Toc201749293)

[Small Language Models (SLMs): 21](#_Toc201749294)

[Alucinaciones en IA: 21](#_Toc201749295)

[Ingeniería de Prompts: 22](#_Toc201749296)

[Aplicaciones de IA en Educación: 22](#_Toc201749297)

[Automatización y Agentes IA: 22](#_Toc201749298)

[Explainable AI (XAI) o IA Explicable: 22](#_Toc201749299)

[Objetivo: 22](#_Toc201749300)

[Ejemplos en acción: 23](#_Toc201749301)

[Herramientas prácticas para XAI: 23](#_Toc201749302)

[Ética y Responsabilidad en el Uso de la Minería de Datos y la IA: 23](#_Toc201749303)

[Principios Fundamentales: 23](#_Toc201749304)

[Peligros si no se controla: 23](#_Toc201749305)

# Unidades del Examen 2 - Inteligencia de Negocios V0.9 (Sin refinar)

[Recomiendo centrarse más en la práctica]

## Data mining (Minería de datos)

**Proceso de encontrar y analizar automáticamente información útil para tomar decisiones en grandes volúmenes de datos**

Una cantidad de datos y variedad que desborda la capacidad humana, con esto van **aprendiendo** por experiencias, comienzan a tomar decisiones basados en la **experiencia** de extraer esa información de fuentes muy diversas. Encontrando así información histórica útil para predecir

### Conceptos relacionados

* **IA**: Capacidad de máquinas para realizar tareas que requieren inteligencia humana, como el Aprendizaje Automático (ML), Procesamiento de lenguaje natural (PLN) y Visión Artificial
* **Ciencia de Datos (Data Science):** Transforma datos en información relevante aplicando Método Científico y generando conocimiento. Evoluciona de la DM y se usa como herramienta de Ciencia de computación, Probabilidad, Estadísticas y ML
* **Aprendizaje Automático (Machine Learnig - ML):** Técnica de IA que permite a sistemas aprender de datos sin programación explicita, basada en algoritmos que analizan datos, predicen y toman decisiones
* **Aprendizaje Profundo (Deep Learning):** Técnica dentro de ML, se usan **redes neuronales** artificiales con varias capas interconectadas para tareas complejas
* **Datos Masivos (Big Data):** Datos cuyo volumen y complejidad necesitan nuevas arquitecturas, técnicas y algoritmos para gestionar y extraer valor. Características: Volumen, velocidad, variedad, veracidad y valor

### OLAP vs Minería de datos

* **OLAP**: **Produce** **información**, permite al usuario analizar datos desde diferentes vistas y responder preguntas. (¿Subieron ventas en abril? / Venden más sucursales del GBA?)
* **Minería de datos**: **Extrae** **patrones**, hace **modelos**, encuentra relaciones y tendencias, todo esto a partir de **reglas** **que va creando** por los patrones. Ayuda al usuario a tomar decisiones a partir de lo descubierto (¿Qué factores influyen en la venta del producto X? / Cuando un cliente compra este producto?)

### Áreas de aplicación de Minería de datos

* **Comercios/Marketing**: Identificar patrones de compra, asociación de clientes con demografía, predecir respuestas a campañas de mailing, analiza la canasta de compra
* **Bancos**: Detecta patrones de uso fraudulento, identifica clientes leales, predice clientes con probabilidad de darse de baja, encuentra relaciones de correlaciones entre indicadores financieros
* **Salud privada**: Patrones de pacientes con alto riesgo
* **Medicina**: Estudio de factores en distintas patologías, identificar terapias medicas satisfactorias para diferentes enfermedades

### Tareas de Mineria de Datos

* **Problemas Predictivos (Supervisados)**: Clasificación, Regresión y Series Temporales
* **Problemas Descriptivos (No Supervisados)**: Agrupamiento (Clustering), **Reglas de Asociación**, Análisis Correlacional, Reducción de Dimensionalidad, y Detección de Anomalías. Las técnicas de minería de datos son herramientas que facilitan el descubrimiento de conocimiento

### Técnica en la minería de datos

Se busca generalizar, aprender conceptos por sus conjuntos, suelen ser robustos, ajustables a dimensionalidades muy altas complejidad, teniendo aprendizaje inductivo

Convierte los datos en modelos, usando algoritmos de aprendizaje (Entrenamiento) y resolución (Calculo predictivo)

### Trabajo con Algoritmos

El proceso de trabajo con algoritmos de minería de datos incluye 5 pasos:

1. **Selección del Algoritmo**: Elegir y testear por diferentes criterios.
2. **Entrenamiento**: Verificar resultados conforme al algoritmo y datos.
3. **Evaluación de Calidad**: Usar métricas para decidir si el algoritmo es adecuado o necesita ajuste de hiperparámetros.
4. **Ajuste de Hiperparámetros**: Modificar según situación, datos y métricas, y volver al paso 2.
5. **Objetivos y Métricas**: Si se está satisfecho, el modelo entrenado está listo; si no, volver al paso 1

### Modelo predictivo – Aprendizaje supervisado

Aprende a **predecir** la clase de una nueva instancia usando ejemplos pre-etiquetados. Encuentra cosas en común en grupos automáticamente

### Metodo descriptivo – Aprendizaje No supervisado

No hay conocimiento a priori ni instancias etiquetadas. **Agrupa** extrayendo nuevas variables que expliquen los datos. Explora y describe los comportamientos de datos

#### Tareas de Asociación

Descubren hechos que ocurren en común dentro de un conjunto de datos mediante reglas de asociación, usadas en análisis de canasta de compra. Ejemplo: {cebollas, vegetales} ⇒ {carne}

#### Tareas de Segmentación/Agrupamiento (Clustering)

**Agrupamiento** jerárquico o no jerárquico de datos según un criterio. Jerárquico puede ser aglomerativo o divisivo; no jerárquico requiere un número de grupos predeterminado

**Problema**: En el aprendizaje no supervisado, solo se conocen las variables de entrada (X), y no existe una variable objetivo predefinida. El objetivo es describir el comportamiento de los datos, descubriendo agrupaciones naturales o patrones.

**Cuándo usarlo**: Es útil cuando se necesita dividir los datos en grupos significativos o útiles, sin perder la estructura natural de los datos. A menudo es un punto de partida para análisis posteriores.

**Para qué usarlo**: Se utiliza para encontrar agrupamientos naturales, describir sus propiedades, descubrir posibles clases, identificar correlaciones entre variables, reducir la dimensionalidad o detectar objetos inusuales (outliers)

#### Aprendizaje por Refuerzo

Los algoritmos aprenden por sí mismos basándose en un esquema de "premios y castigos" en un entorno que cambia con el tiempo. Se busca maximizar la recompensa. Ejemplos incluyen juegos, robótica, medicina y sistemas de navegación de vehículos autónomos

#### Explotación de Información

Es la sub-disciplina informática que aporta a la Inteligencia de Negocios las herramientas (procesos y tecnologías) para la transformación de información en conocimiento. Utiliza la Minería de Datos y aborda problemas de predicción, clasificación y segmentación

## Proceso Descubrimiento de Conocimiento (KDD)

Databases, es un conjunto de actividades organizadas. Sus fases son:

Escala de tiempo

El contenido generado por IA puede ser incorrecto.

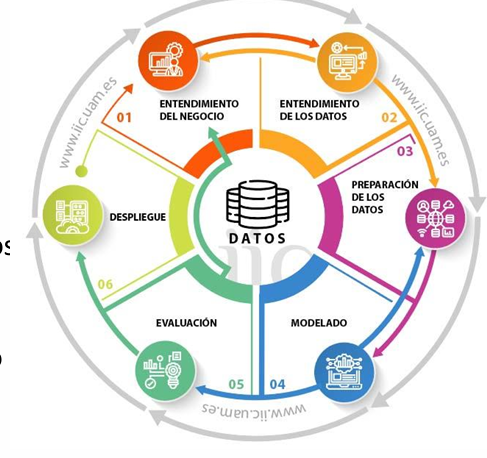
* **Selección de datos**: Datos sobre los que se trabajará.
* **Preprocesamiento**: Preparación y limpieza de datos (manejo de faltantes, inconsistencias, anomalías).
* **Transformación**: Tratamiento preliminar, agregación, normalización y generación de nuevas variables.
* **Minería de Datos**: Construcción de modelos con técnicas predictivas y descriptivas.
* **Evaluación e interpretación**: Análisis del modelo y conocimiento obtenido, validando si los resultados son satisfactorios

### Metodologías de Explotación de Información

Se mencionan metodologías probadas por la comunidad científica como CRISP-DM, SEMMA, P3TQ y DATLAS

### Metodología CRISP-DM

Cross-Industry Standard Process for Data Mining es un **proceso estándar**, no patentado y de libre acceso, para adaptar la minería de datos a la resolución de problemas de negocio o investigación



* **Comprensión de los Datos (Data Understanding)**: Recolectar datos, analizar sus características y formular hipótesis iniciales. Se distinguen atributos nominales (categóricos, discretos) y ordinales (numéricos, continuos)
* **Preparación de los Datos (Data Preparation)**: Tratamiento del conjunto de datos final, incluyendo selección, limpieza y transformación. Se abordan problemas de calidad (anómalos, faltantes, duplicados) y se realizan transformaciones como agregación, selección/creación de atributos, discretización y normalización
* **Modelado**: Aplicar procesos de explotación de información y algoritmos de minería sobre los datos para obtener información oculta y patrones de conocimiento
* **Evaluación**: Analizar los patrones obtenidos según los objetivos organizacionales y determinar si el nuevo conocimiento será implementado
* **Implementación**: Comunicar e implementar el nuevo conocimiento de forma comprensible para el usuario

### SEMMA (Sample, Explore, Modify, Model and Assess)

Desarrollado por SAS Institute, esta metodología incluye las fases

Diagrama

El contenido generado por IA puede ser incorrecto.

* **Sample (Muestro)**: Muestrear la base de datos principal
* **Explore (Exploración)**: Entender los datos, buscar anomalías, patrones y tendencias
* **Modify (Modificación)**: Crear, seleccionar y transformar variables, buscar anomalías y reducir variables
* **Model (Modelado)**: Aplicar métodos estadísticos y evaluar fortalezas.
* **Assess (Evaluar)**: Evaluar la confiabilidad y utilidad de los hallazgos. Similar a KDD y CRISP-DM, si los objetivos no se logran, se puede iterar

Interfaz de usuario gráfica

El contenido generado por IA puede ser incorrecto.

## Herramienta para la minería de datos

Tabla

El contenido generado por IA puede ser incorrecto.

### Modelos predictivos (Aprendizaje Supervisado)

Predice el resultado de una variable de interés (variable clase o atributo objetivo) a partir de otras variables (atributos predictores). El entrenamiento busca descubrir las relaciones entre las variables de entrada y la variable objetivo para predecir nuevos casos

## Tareas de Clasificación

Implican la asignación de nuevos casos a una clase con la máxima exactitud y precisión posible

#### Árboles de Decisión (Algoritmos de inducción TDIDT - Top Down Induction Trees):

Aprenden de ejemplos preclasificados, identificando y ubicando en la parte superior del árbol los atributos que mejor separan los ejemplos. Utilizan la teoría de la información para encontrar los atributos con mayor ganancia de información (ej. ID3, C4.5, C5). Generan reglas del tipo IF-THEN

### Bayes:

### Las Redes Bayesianas

Son un tipo de modelo predictivo que pertenece a las tecnologías para clasificación. Una red bayesiana **es un grafo acíclico dirigido que se compone de nodos y arcos.**

* Los nodos representan las variables aleatorias, también conocidas como atributos.
* Los arcos representan las dependencias probabilísticas de cada variable, indicando una influencia directa de una variable sobre otra.

**Fundamento**: Se basan en la probabilidad condicional, utilizando el Teorema de Bayes.

**Relación causa-efecto:** Estas redes representan la relación de causa y efecto entre los atributos y proporcionan una medida cuantitativa y probabilística de la importancia de los atributos en un problema de clasificación de clases.

### El Clasificador Bayesiano Naïve-Bayes es un tipo específico de red bayesiana.

Suposición clave: Considera que cada atributo predictor (A\_i) y el atributo clase (C) son variables aleatorias, y que las relaciones de dependencia entre los atributos A\_i son condicionalmente independientes entre sí, dado el atributo clase C.

**Objetivo**: Dado un registro con atributos A1, A2,..., An, el objetivo es predecir la clase C. Se busca encontrar el valor de C que maximice la probabilidad p(C/A1, A2,..., An).

#### Proceso de aprendizaje:

**Aprendizaje Estructural**: Identifica las relaciones de dependencia e independencia entre las variables.

**Aprendizaje Paramétrico:** Determina las probabilidades a priori de cada clase y las probabilidades condicionales.

**Proceso de inferencia**: Después de aprender, el modelo puede realizar predicciones a partir de nuevas observaciones. Por ejemplo, en un escenario donde se predice si se jugará al tenis, se calculan las probabilidades de "Juega Si" y "Juega No" basándose en las condiciones dadas (soleado, cálido, húmedo, ventoso). El clasificador predice la clase con la probabilidad más alta, como que "no se juega al tenis con una probabilidad del 64%" en el ejemplo dado

### Evaluación de Modelos de Clasificación:

La evaluación es crucial para determinar la calidad del modelo59. No hay un algoritmo universalmente "mejor", y la elección depende del problema. Las métricas se calculan a menudo usando una **Matriz de Confusión**, que compara los resultados predichos con los reales

#### La Matriz de Confusión

Es una herramienta esencial para la evaluación de modelos de clasificación, especialmente en el aprendizaje supervisado78.

* **Propósito:** Permite comparar el resultado obtenido por el modelo predictivo con los resultados reales de los datos de prueba8.
* **Estructura (para 2 clases):** Muestra los valores clasificados correctamente (Verdaderos Positivos - TP, Verdaderos Negativos - TN) y los errores de clasificación (Falsos Positivos - FP, Falsos Negativos - FN)910.
  + **TP (Verdaderos Positivos / NCVA):** Casos reales de la clase A (positiva) que el modelo clasificó correctamente como A910.
  + **FN (Falsos Negativos / NCFB):** Casos reales de la clase A (positiva) que el modelo clasificó incorrectamente como B (negativa)910.
  + **FP (Falsos Positivos / NCFA):** Casos reales de la clase B (negativa) que el modelo clasificó incorrectamente como A (positiva)910.
  + **TN (Verdaderos Negativos / NCVB):** Casos reales de la clase B (negativa) que el modelo clasificó correctamente como B

### Métricas clave de evaluación:

**Exactitud (Accuracy)**: Proporción de casos clasificados correctamente sobre el total de casos65. Evalúa la capacidad de generalización. Una alta exactitud es >70%. Sin embargo, la exactitud no es una buena métrica para datasets desbalanceados, ya que puede ser engañosa

**Precisión (Precision)**: Proporción de casos reales de una clase respecto al total de casos clasificados por el modelo en esa clase. Evalúa la efectividad para clasificar casos a una clase particular

**Exhaustividad (Recall / Sensibilidad / TPR)**: Tasa de verdaderos positivos; proporción de casos positivos bien clasificados respecto a todos los elementos que son realmente positivos. Indica qué tan bien el modelo detecta la clase positiva

**Especificidad (TNR)**: Tasa de verdaderos negativos; proporción de casos negativos bien clasificados respecto a todos los elementos que son realmente negativos. Útil para identificar verdaderos negativos

**F1-Score**: Combina precisión y exhaustividad en un solo valor, siendo la media armónica de ambos. Es útil para comparar el rendimiento combinado

**Coeficiente Kappa (Cohen)**: Mide el grado de acuerdo entre las predicciones del modelo y la realidad, más allá del azar. Es especialmente útil con clases desbalanceadas

En datasets **desbalanceados** (donde hay muchas instancias de una clase y muy pocas de otra, como en detección de fraudes o diagnóstico médico), la precisión, recall y F1-Score son más representativas que la exactitud. Los errores en la clasificación pueden tener diferentes costos para un negocio (ej. un falso negativo en un banco donde el cliente se va es más costoso)

## KNIME

Fue desarrollado en 2004 por el Departamento de Bioinformática y Minería de Datos de la Universidad de Konstanz, Alemania, y su desarrollo es continuado por KNIME.com GmbH

### Características y Funcionamiento de KNIME:

* Desarrollado sobre la plataforma Eclipse y programado principalmente en Java.
* Su uso se basa en el **diseño de un flujo de ejecución** (workflow) que representa las distintas etapas de un proyecto de minería de datos.
* El análisis de datos en KNIME se realiza mediante un pipeline de **nodos**, conectados por "**puentes**" que transportan datos o modelos. Cada nodo procesa las entradas y produce resultados.
* Permite **explorar** los datos **visualmente** a través de vistas.
* Es una herramienta para **análisis**, **manipulación**, **visualización** y **generación** de informes de datos, basada en el paradigma de programación gráfica (drag & drop).
* Proporciona una amplia gama de **extensiones**, incluyendo Text Mining, Network Mining, Cheminformatics, e integraciones con lenguajes y herramientas como Java, R, Python, **Weka**, Keras, Plotly, H2O, etc..
* Los **nodos** están **agrupados** en **fichas** para diferentes propósitos: Entrada de datos (IO > Read), Salida de datos (IO > Write), Preprocesamiento (Data Manipulation), Minería de datos (Analytics/Mining para modelos como reglas de asociación, clustering, clasificación), y Salida de resultados (Data Views). También tiene nodos específicos para WEKA (Bayes, Trees, Rules).
* La interfaz de usuario de KNIME presenta varios paneles: KNIME Explorer, Workflow Coach, Node Repository, Workflow Editor, Outline, Console & Node Monitor, Node Description y KNIME Community Hub. Existen tanto una interfaz clásica como una moderna.
* Ofrece más de 4000 rutinas de análisis de datos.

### Funcionalidades de KNIME:

* Software libre bajo licencia **GNU**.
* Combinación de datos y herramientas, analíticas poderosas, más de 4000 módulos y en crecimiento.
* Conectores para todos los formatos de archivos y bases de datos comunes.
* Soporte para gran variedad de tipos de datos, combinación y transformación de datos nativa y en la base de datos.
* Funciones matemáticas y estadísticas, algoritmos de predicción avanzados y Machine Learning.
* **Control de flujo, herramienta de unión para R, Python, SQL, Java, Weka**.
* Vistas de datos y reportes interactivos.
* Se recomienda instalar la versión de 64 bits y algunas extensiones útiles para Data Generation, Textprocessing, Python Integration, Weka Integration, Machine Learning Interpretability (Labs), Plotly, entre otras

### Creación de Workflows:

**Un flujo de trabajo** es una secuencia configurable de nodos que realizan tareas específicas.

* Los datos fluyen de **izquierda a derecha** a través de los nodos.
* Para construir un flujo, se arrastran y sueltan nodos desde el repositorio al Editor Windows y se conectan las salidas de unos nodos con las entradas de otros.
* Un flujo básico podría ser: **Nodo de lectura de datos → Nodo de preprocesamiento → Nodo de modelado → Nodo de salida de resultados**.
* Los nodos tienen estados visuales que indican si están no configurados, configurados, ejecutados o con error.

## Modelos Descriptivos (Aprendizaje No Supervisado)

Gráfico

El contenido generado por IA puede ser incorrecto.

### Tareas de Segmentación (Clustering):

#### Propósito:

**Dividir los datos en grupos significativos y homogéneos** (clústeres) con características similares entre sus miembros, pero distintos de los de otros grupos, sin perder la estructura natural de los datos. Es un punto de partida común para encontrar agrupamientos naturales o útiles, correlaciones, representantes para grupos, objetos inusuales (outliers) o ruido.

#### Aplicaciones Típicas:

**Marketing** (descubrimiento de hábitos de compra, campañas dirigidas), **Seguros** (identificación de asegurados con características similares para ofertar productos), **Fraudes** (detección de puntos inusuales), **Planificación urbana** (identificación de grupos de viviendas), **Web Mining** (clasificación de documentos, análisis de logs), **Procesamiento de imágenes** (reconocimiento de patrones), y como preprocesamiento para otras técnicas de DM.

#### Funcionamiento:

**La similitud entre datos se define usando una medida de distancia** (Euclidiana, Minkowski, Jaccard, Coseno, etc.). Un buen método de clustering debe identificar clústeres compactos (alta similaridad intra-clúster) y bien separados (baja similaridad inter-clúster). La elección del número de clústeres (k) es un reto, para lo cual se usan métodos como el del Codo, el de la Silueta, o el conocimiento del dominio/negocio.

Diagrama

El contenido generado por IA puede ser incorrecto.

### Tipos de Clustering:

* **No Jerárquico (Particional**): Divide los datos en subconjuntos sin solapamiento, con un número de grupos (k) predeterminado. Cada punto pertenece al grupo "más cercano". Ejemplos: K-Means, K-Modes, K-Medoids (PAM).
* **K-Means**: Algoritmo que agrupa datos similares en K clústeres. Selecciona k centroides iniciales, asigna puntos al centroide más cercano y recalcula los centroides hasta que no cambien. Es sensible a la posición inicial de los centroides.

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.

* **Jerárquico**: Construye un árbol binario (dendrograma). Puede ser Aglomerativo (bottom-up, AGNES), donde cada punto es un clúster y se combinan los más cercanos; o Divisivo (top-down, DIANA), donde se parte de un clúster y se divide recursivamente.
* **Otros tipos incluyen basados en densidad** (DBSCAN, OPTICS), basados en rejillas (STING, CLIQUE), basados en datos categóricos (ROCK), y basados en modelos (**Redes Neuronales SOM**, COBWEB, Gaussian Mixture Model, Fuzzy C-Means).

Diagrama

El contenido generado por IA puede ser incorrecto.

Diagrama

El contenido generado por IA puede ser incorrecto.

## Mapas Auto-Organizados de Kohonen (SOM):

Tipo de red neuronal no supervisada usada para reducción de dimensionalidad, visualización y segmentación. Transforma datos de alta dimensión en un mapa de baja dimensión (generalmente 2D) que preserva la topología original. Las neuronas "compiten" para ser la ganadora (Best Matching Unit), y ella y sus vecinas ajustan sus pesos.

### Evaluación de Segmentación:

Es más compleja que en clasificación, ya que no hay una "verdad" conocida. Se evalúa para evitar patrones en el ruido, comparar algoritmos o conjuntos de clústeres. Aspectos a considerar son la tendencia de agrupamiento, el número correcto de clústeres (k), el ajuste de los resultados a los datos (interna), la comparación con datos externos (externa) o entre diferentes conjuntos de clústeres (relativa).

### Métricas Internas:

**Evalúan la estructura del clustering sin información externa**. Incluyen **Cohesión** (qué tan cercanos son los objetos en un clúster, ej. SSW/SSE) y **Separación** (qué tan diferentes son los clústeres entre sí, ej. SSB). El Coeficiente de Silhouette evalúa la calidad del agrupamiento e identifica el número óptimo de clústeres.

### Métricas Externas:

**Comparan el clúster asignado con una etiqueta de clase preexistente** (si disponible). Incluyen **Pureza** (nivel en que un clúster contiene elementos de una sola clase) y **Entropía** (cantidad de clases diferentes que contiene un clúster).

### Validación con Expertos:

**Evaluación** **subjetiva** para ver si los clústeres producen el resultado esperado y comparar con otras soluciones.

### Tareas de Asociación:

Propósito:

**Descubrir reglas que puedan predecir la ocurrencia de un ítem a partir de otros en una transacción**. Buscan patrones frecuentes, asociaciones, correlaciones o estructuras causales entre conjuntos de artículos.

#### Aplicaciones Típicas:

**Análisis** **de Canasta** (Market Basket Analysis) para entender el comportamiento del cliente y qué productos se compran juntos. También en Diseño de Catálogos.

### Reglas de Asociación:

**Concepto:** Se trata de encontrar reglas que puedan predecir la ocurrencia de un ítem a partir de otros ítems en una transacción, con un mínimo de confianza y soporte

Se representan como implicaciones **X → Y (Antecedente → Consecuente).**

**Tabla

El contenido generado por IA puede ser incorrecto.**

## Métricas de Evaluación:

### Cobertura (Soporte/Support):

**Proporción de transacciones que contienen todos los ítems del antecedente y el consecuente**. Reglas con mayor cobertura son más representativas y útiles. Un soporte alto significa que la regla es frecuente. Un soporte bajo puede indicar una ocurrencia casual y poca utilidad.

### Precisión (Confianza/Confidence):

**Proporción de transacciones que cumplen con la regla (**contienen antecedente y consecuente) respecto al total de transacciones que contienen el antecedente. **Una confianza alta indica mayor fiabilidad e interés para asociar ítems y descubrir conocimiento**. Una confianza baja sugiere que es probable que no haya relación entre antecedente y consecuente.

### Proceso de Extracción de Reglas:

Para grandes volúmenes de datos, se descompone en dos pasos: **encontrar conjuntos de artículos frecuentes (con soporte mínimo) y generar reglas de asociación "fuertes" (que satisfagan mínimos de soporte y confianza).** El algoritmo "**A priori**" utiliza la precisión y luego el soporte como criterios de selección de reglas. Valores típicos para reglas de asociación son Soporte = 2-10% y Confianza = 70-90%.

## Calidad de los modelos de minería de datos

### ¿Qué hace que un modelo sea bueno? Un buen modelo debe:

* **Generalizar bien con datos nuevos**, es decir, aprender de manera efectiva con muchos tipos de datos.
* **Evaluarse con datos no usados** **durante el entrenamiento.**
* **Compararse con otros modelos usando métricas** de desempeño como Precisión y Recall. La clave del éxito reside en la capacidad predictiva y la simplicidad del modelo, más allá de la rapidez en su construcción o clasificación. Las métricas se calculan contrastando valores predichos con valores reales de la variable objetivo.

### El Problema del Sobreaprendizaje (Overfitting):

* **Ocurre cuando un modelo aprende demasiado bien los datos de entrenamiento,** hasta el punto de memorizarlos, **pero falla con datos nuevos.** Es como un estudiante que memoriza respuestas sin entender el problema.
* Un modelo sobreajustado tiene un **error bajo en el entrenamiento**, pero un **error alto en la prueba**
* El propósito del modelo no es clasificar el conjunto de entrenamiento, sino los datos cuya clase no se conoce. Por lo tanto, el conjunto de prueba debe ser siempre independiente del conjunto de entrenamiento. La complejidad excesiva del modelo tiende a provocar sobreaprendizaje.

## Generalización - Análisis de Errores:

Se identifican dos tipos de errores principales:

* **Bias** (Sesgo del Modelo): **Ocurre cuando el modelo es demasiado simple y no captura la complejidad subyacente de los datos**. Resulta en errores altos tanto en entrenamiento como en prueba (Underfitting). Para corregirlo, se necesita un modelo más complejo, más muestras de datos o añadir funciones/hiperparámetros.
* **Varianza**: **Ocurre cuando el modelo es demasiado complejo y es muy sensible a las fluctuaciones en los datos de entrenamiento**. Resulta en un error bajo en entrenamiento, pero alto en prueba (Overfitting). Se reduce la varianza añadiendo más datos y simplificando el modelo.

**El mejor modelo logra un equilibrio entre Sesgo y Varianza.**

* **Train Error LOW + Test Error HIGH**: Indica un modelo demasiado complejo o necesidad de más datos (sobreajuste).
* **Both Errors HIGH**: Indica un modelo demasiado simple o falta de datos (subajuste).
* **Train Error HIGH + Test Error LOW**: Es inusual, podría significar que los datos de prueba son demasiado similares a los de entrenamiento, requiriendo más datos de prueba.
* **Both Errors LOW**: Indica un modelo ideal.

## Selección de Algoritmos de Machine Learning:

La elección del algoritmo depende principalmente de dos aspectos:

1. **Qué se desea hacer con los datos**: Cuál es la pregunta de negocio a responder.

2. **Requisitos del caso de estudio**: Precisión, tiempo de entrenamiento, linealidad, cantidad de parámetros y características que admite la solución.

Algunos algoritmos hacen suposiciones específicas sobre la estructura de los datos o los resultados deseados, y encontrar uno que se adapte puede llevar a mejores resultados

## Fundamentos de la IA

### Historia de la IA:

* **Pre-1950s**: Ideas fundacionales y lógica simbólica, trabajo de Alan Turing.
* **1950s**: Nacimiento de la IA, con el Taller de Dartmouth (1956) como evento fundacional. Primeros programas como "Logic Theorist" y "General Problem Solver".
* **1960s**: Optimismo y primeros éxitos (resolución de problemas de álgebra, ajedrez, comprensión limitada del lenguaje natural como ELIZA).
* **1970s**: Primer "Invierno" de la IA, desilusión por falta de progreso en problemas complejos y reducción de financiación.
* **1980s**: Renacimiento con los Sistemas Expertos, que emulaban conocimiento humano en dominios específicos (diagnóstico médico).
* **Finales 80s y 1990s**: Segundo "Invierno" de la IA debido a limitaciones de los sistemas expertos y falta de aplicaciones generalizadas.
* **Finales 90s y Principios 2000s**: Auge del Machine Learning, impulsado por disponibilidad de datos y potencia computacional (filtrado de spam, sistemas de recomendación, reconocimiento de voz).
* **2010s**: Revolución del Deep Learning, gracias a redes neuronales con múltiples capas, avances en visión por computadora, NLP y reconocimiento de voz.
* **Presente (2020s):** Era de los Large Language Models (LLMs) y la IA Generativa, con desarrollo de modelos como GPT-3, Gemini y aplicaciones para imágenes, audio y video, generando debates éticos.

### Inteligencia Artificial Generativa (IAGen):

* Es una categoría de IA que crea contenido original, tomando entradas de **lenguaje natural** (**Prompts**) y devolviendo respuestas en diversos formatos (lenguaje natural, imágenes, código).
* Se basa en la arquitectura **Transformer** y los usuarios interactúan con ella a través de aplicaciones de chat.

#### Large Language Models (LLMs):

* **Forma avanzada de IA entrenada con grandes volúmenes de texto** para identificar patrones lingüísticos y generar lenguaje coherente.
* **Descomponen el texto en tokens** y predicen el siguiente token para formar respuestas relevantes.
* **"Gran escala"** se refiere a la enorme cantidad de parámetros (neuronas artificiales) y el vasto volumen de texto usado para el entrenamiento, lo que les permite realizar tareas variadas y creativas**. Ejemplos: ChatGPT, Gemini, Copilot.**

#### Small Language Models (SLMs):

* **Versiones reducidas y especializadas de los LLMs,** optimizadas para tareas o dominios específicos.
* **Son más rápidos de ajustar y desplegar,** requieren menos recursos computacionales y consumen menos energía, lo que contribuye a una IA más sostenible y permite su ejecución en dispositivos limitados.

### Alucinaciones en IA:

* Aunque los modelos IAGen generan respuestas coherentes, **no siempre son verídicas y pueden producir "alucinaciones**" (información inventada o incorrecta).
* **Ocurren debido a entrenamiento estadístico sin comprensión semántica**, errores de aprendizaje, falta de verificación de hechos o prompts ambiguos.

### Ingeniería de Prompts:

* Un **prompt** es una **instrucción de texto clara y específica** que guía la respuesta de un modelo de IA. Es fundamental para interactuar eficazmente con un **LLM**.
* Una estructura de prompt efectiva **incluye la tarea, contexto, ejemplos, persona, formato y tono deseado.** Los prompts vagos o ambiguos pueden llevar a que el modelo se pierda o use suposiciones.
* **El acrónimo MIEL (Meta + Identidad + Estilo + Lenguaje) ayuda a construir prompts claros y eficaces, útil en educación, comunicación, diseño o investigación.**

### Aplicaciones de IA en Educación:

* Útiles para **resumir textos**, **generar ideas/actividades, simular debates**, corregir textos, traducir, crear materiales, asistencia en redacción, apoyo a investigación y evaluación automática.
* **Limitaciones y Riesgos**: **Pérdida de pensamiento crítico y autoría**, dificultad para evaluar autoría, respuestas erróneas (alucinaciones), reproducción de sesgos y falta de políticas institucionales.
* **Es crucial promover un uso reflexivo y con acompañamiento docente**, evaluando la comprensión más allá del texto final, y enmarcarlo en principios éticos y pedagógicos.

### Automatización y Agentes IA:

* La automatización con IA es clave para enfocarse en pensamiento crítico, **ahorrar tiempo en tareas repetitivas y aprender habilidades demandadas** en el mercado tecnológico.
* Los Agentes IA son programas que p**erciben su entorno y toman decisiones automáticas usando modelos de IA.** Un verdadero agente IA consta de un LLM (para razonar), uso de herramientas (para actuar), memoria (corta y larga), y autorreflexión (para ajustarse). A diferencia de un LLM, los agentes IA agregan funcionalidad orientada a objetivos, pueden usar herramientas, procesar sus salidas y tomar decisiones en múltiples pasos.

### Explainable AI (XAI) o IA Explicable:

**Conjunto de métodos y técnicas que permiten a los humanos comprender y confiar en los resultados de los modelos de IA.**

#### Objetivo:

**Hacer los modelos transparentes, interpretables y auditables, evitando el problema de las "cajas negras".**

Importancia: Genera confianza, asegura ética y cumplimiento normativo (ej. GDPR), ayuda a depurar modelos.

#### Ejemplos en acción:

En crédito bancario (explica por qué se deniega un préstamo), diagnóstico médico (resalta áreas críticas en imágenes), automóviles autónomos (muestra el motivo de una frenada) y detección de fraude (explica patrones inusuales).

#### Herramientas prácticas para XAI:

LIME (explica predicciones individuales), SHAP (explica la contribución de cada variable a la predicción), InterpretML, IBM AI Explainability 360. KNIME ofrece nodos específicos para interpretar modelos de ML con SHAP y LIME sin necesidad de programación.

### Ética y Responsabilidad en el Uso de la Minería de Datos y la IA:

* La IA está creciendo rápidamente, y el acceso eficiente y económico a la información requiere regulación.
* **Motivación**: Las decisiones algorítmicas afectan a las personas.

#### Principios Fundamentales:

* **Imparcialidad**: Evitar sesgos y tratar a todos de manera justa.
* **Confiabilidad y seguridad**: Los sistemas deben funcionar correctamente y sin fallas críticas.
* **Privacidad y protección de datos**: Cuidado de los datos personales durante y después del entrenamiento.
* **Inclusión**: Funcionar para todos sin exclusión por discapacidad, identidad o contexto social.
* **Transparencia**: Ser claro sobre cómo funcionan los sistemas de IA y cómo toman decisiones.
* **Responsabilidad:** Establecer quién es responsable de las acciones de la IA.

### Peligros si no se controla:

**Predicciones erróneas, discriminación por sesgos, pérdida de privacidad y poca transparencia**

* **La responsabilidad recae en los desarrolladores de soluciones de IA,** quienes deben **anticipar riesgos, ser críticos y éticos.**
* **Uso Sostenible de la IA**: Considera el **alto volumen de datos y el consumo de energía de los centros de datos.** Estrategias incluyen eficiencia de datos, uso de energía renovable y eliminación responsable de datos.
* **El futuro de la IA depende de un desarrollo y uso responsable**, con colaboración entre investigadores, legisladores, la industria y la sociedad. Es fundamental tener diversidad de antecedentes en quienes desarrollan estos sistemas e invertir en conjuntos de datos que reflejen la diversidad social para detectar y corregir daños.