La minería de datos

Conceptos fundamentales

Competencias a lograr

- Comprender los fundamentos de la Minería de Datos y sus aplicaciones.
- Identificar las etapas del proceso de Minería de Datos.
- Realizar preprocesamiento básico en un conjunto de datos.
- Diferenciar entre diferentes tipos de datos y atributos.

Definiciones iniciales

Datos

Hechos o medidas que describen características de objetos, eventos o personas.

Información

Datos analizados y presentados en forma adecuada, de interés para un observador en un momento determinado.

Ejemplo:

- **Datos**: Temperaturas registradas a lo largo de una semana: 23°C, 25°C, 24°C, 22°C, 26° C, 21°C, 20°C.
- **Información**: Las temperaturas en la última semana mostraron fluctuaciones, siendo la más alta 26°C y la más baja 20°C. Parece haber ocurrido un descenso de la temperatura al final de la semana.

Definiciones iniciales

Conocimiento

Información procesada para emitir juicios que llevan a conclusiones.

Ejemplo:

- **Datos**: Temperaturas registradas a lo largo de una semana: 23°C, 25°C, 24°C, 22°C, 26°C, 21°C, 20°C.

- **Información**: Las temperaturas en la última semana mostraron fluctuaciones, siendo la más alta 26°C y la más baja 20°C. Parece haber ocurrido un descenso de la temperatura al final de la semana.
- **Conocimiento**: Basándonos en las fluctuaciones de las temperaturas registradas durante la semana y el descenso observado al final, podemos inferir que podría haber un cambio de estación o patrón climático que influye en las temperaturas. Esto podría ser debido a una entrada de aire frío o a otros factores atmosféricos. Con base en esto, podríamos considerar **tomar medidas** para enfrentar cambios de temperatura similares en el futuro, como vestirse adecuadamente o preparar el hogar para la variabilidad climática.

Definiciones iniciales

Metaconocimiento

Conocimiento sobre cómo se aplican las técnicas y algoritmos de Minería de Datos a conjuntos de datos de contextos específicos, así como sobre los resultados y el rendimiento de esos procesos.

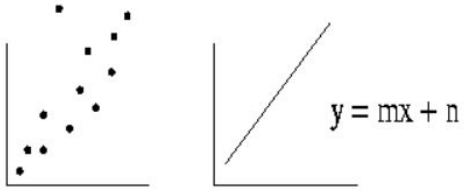
Ejemplo:

Las Máquinas de Vectores de Soporte (SVM) proporcionan buenos resultados en la industria petrolera para la clasificación de formaciones geológicas y para la detección de fallas de equipos o máquinas.

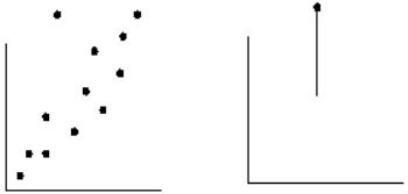
Definiciones iniciales

Modelo vs Patrón

Modelo: Habla de todo el conjunto de datos



Patrón: Habla de una región particular de datos.



Definiciones iniciales

Modelo

Se refiere a una representación simplificada o abstracta (matemática o estadística) de un fenómeno, proceso o sistema real, que aprende de los datos para hacer predicciones o tomar decisiones en nuevas situaciones.

Es una representación o aproximación de la realidad.

Esta representación tiene que ser de calidad.

Definiciones iniciales

Modelo

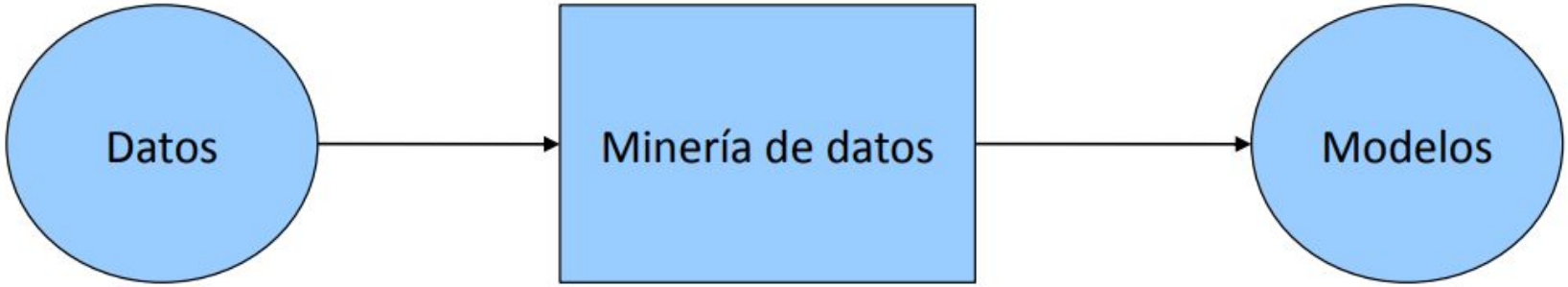
Suponga el siguiente contexto:

Se tiene un modelo que predice el precio de departamentos en la Ciudad Autónoma de Buenos Aires.

Departamento A	Departamento B
Lindo departamento 2 ambientes, 80mts2, frente a la estación de subte “Palermo”, con ambientes climatizados, en un edificio de 5 pisos que tiene 10 años de construido.	Departamento 2 ambientes, 80mts2, a 5 cuadras de la estación de subte “Constitución”, sin climatización, en un edificio de 12 pisos que tiene 80 años de construido.
US$ 130.000	**US$ 375.000**

Predicción del modelo 

Visión simplificada de la minería de datos



- Los “modelos” son el producto de la minería de datos...
- ...y dan soporte a las estrategias de decisión que se tomen

Definiciones iniciales

Datos y Modelos ⇒ Conocimiento

Los datos se obtienen de:

- Bases de datos (relacionales, documentales, de grafos, espaciales, temporales, multimedia, etc).
- APIs
- WWW
- Redes sociales
- …

Los modelos pueden ser:

- **Descriptivos**: identifican patrones que explican o resumen datos.
  - Reglas de asociación
  - Clustering
- **Predictivos**: estiman valores de variables de interés (a predecir) a partir de valores de otras variables.
  - Regresión
  - Clasificación

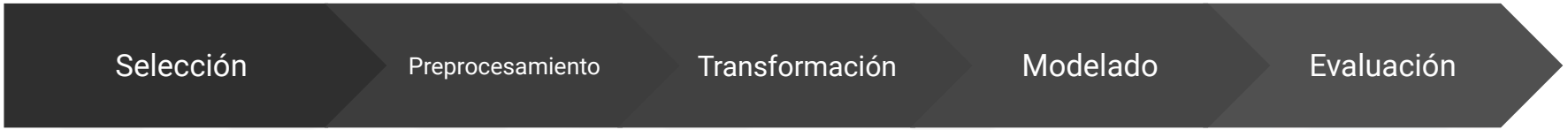
Minería de datos (varias definiciones)

1. La minería de datos tiene como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten hacia la toma de decisión.
2. Minería de datos es la exploración y análisis de grandes cantidades de datos con el objeto de encontrar patrones y reglas significativas (conocimiento).
3. Es un mecanismo de explotación que consiste en la búsqueda de información valiosa en grandes volúmenes de datos.
4. Ligada a las bodegas de datos (información histórica) con la cual los algoritmos de minería de datos obtienen información necesaria para la toma de decisiones.

Minería de datos (varias definiciones)

1. Análisis de grandes volúmenes de datos para encontrar relaciones no triviales, y para resumirlos de manera que sean entendibles y útiles (Hand, Mannila y Smyth).
2. Extracción de patrones y modelos interesantes, potencialmente útiles y datos en base de datos de gran tamaño (Hand).
3. Proceso de extraer conocimiento util y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos (Witten and Frank)
4. Proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles, a partir de los datos (Fayyad y col.)

Proceso general de la minería de datos



Selección	Preprocesamiento	Transformación	Modelado	Evaluación
Elegir el conjunto de datos adecuado.	Limpiar, transformar y reducir los datos.	Convertir datos en una forma adecuada para análisis.	Aplicar algoritmos para descubrir patrones.	Evaluar y seleccionar el mejor modelo.

Datasets

Conceptos fundamentales:

Unidad de datos

Entidad en la población estudiada, como una persona o empresa, sobre la que se recopilan datos y está integrada por diferentes tipos de variables.

Una unidad de datos también se conoce como una de unidad o registro, o simplemente registro.

Elemento de datos

Característica (o atributo) de una unidad de datos que se mide o cuenta, como la altura, el país de nacimiento o los ingresos.

Un elemento de datos también se conoce como variable porque la característica que representa puede variar entre las unidades de datos y puede variar con el tiempo.

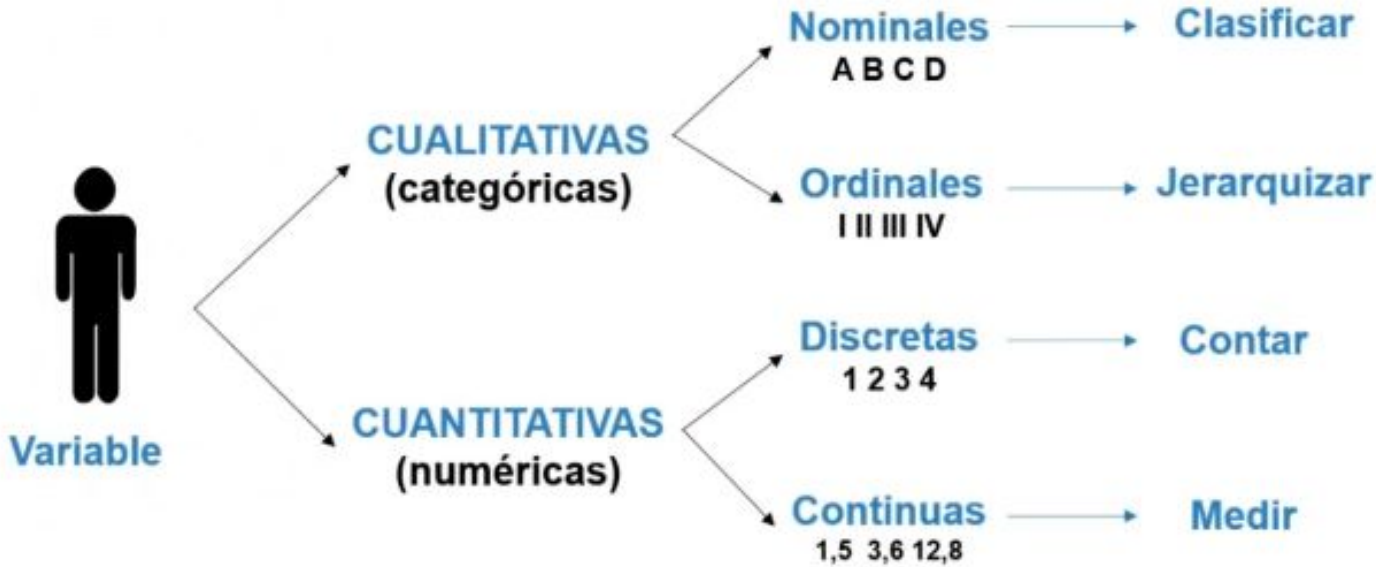
Observación

Ocurrencia de un elemento de datos específico que se registra sobre una unidad de datos.

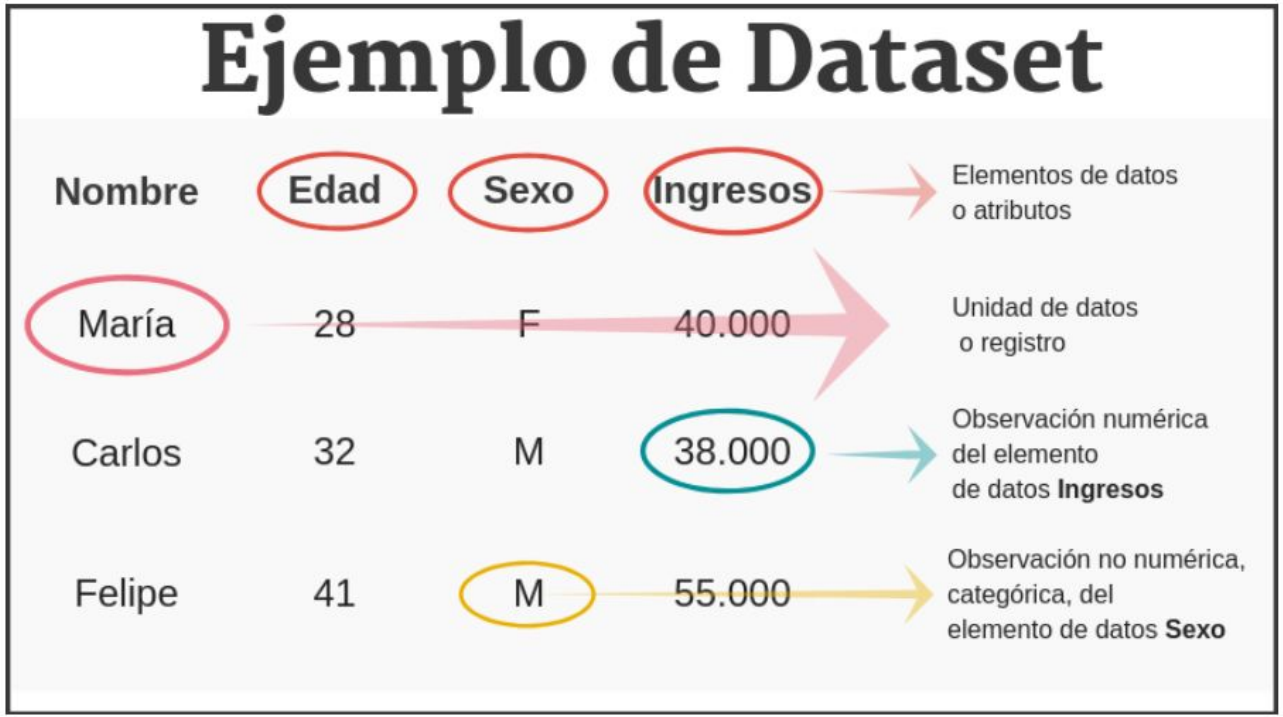
Una observación puede ser numérica o no numérica (categórica).

Datasets

Tipos de variables:



Datasets


Ejemplo de Dataset
Nombre
Edad
Sexo
Ingresos
Elementos de datos o atributos
María
28
F
40.000
Unidad de datos o registro
Carlos
32
M
38.000
Observación numérica del elemento de datos Ingresos
Felipe
41
M
55.000
Observación no numérica, categórica, del elemento de datos Sexo

Ejemplo

Explorar algunos datasets:

- IBM HR Analytics Employee Attrition
- MP Onboarding Metrics
- Twitter Personality OCEAN

Librerías Python a utilizar:

- Pandas (documentación)

Probar con Dataset Search de Google

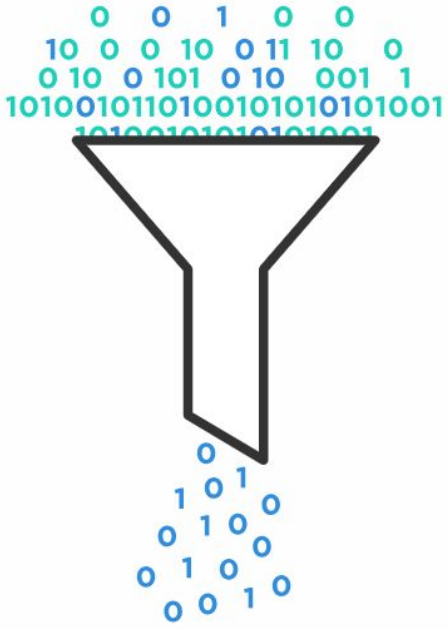
Etapa 2: preprocesamiento

Preprocesamiento básico

Objetivo: Mejorar la calidad de los datos

Datos de mala calidad son datos que pueden:

- Estar incompletos
- Tener valores faltantes
- Tener ruido o errores



Ejemplo

Explorar el siguiente dataset

- MP Onboarding Metrics

Librerías Python a utilizar:

- Pandas (documentación)

Etapas del preprocesamiento

Limpieza de datos

Se eliminan los valores duplicados, faltantes o incorrectos que puedan distorsionar los resultados del análisis.

Esto implica tomar decisiones sobre cómo tratar los valores faltantes y cómo corregir los datos ruidosos o erróneos.

Valores duplicados

Se refieren a la presencia de elementos idénticos o repetidos en un conjunto de datos, como una lista, un array o una base de datos.

Esto puede ser un problema en muchos contextos, ya que los valores duplicados pueden distorsionar análisis, generar resultados incorrectos o causar confusiones en la interpretación de los datos.

Acciones: Identificar y eliminar

Ejemplo

Explorar el siguiente dataset

- MP Onboarding Metrics

Librerías Python a utilizar:

- Pandas (documentación)

Valores faltantes

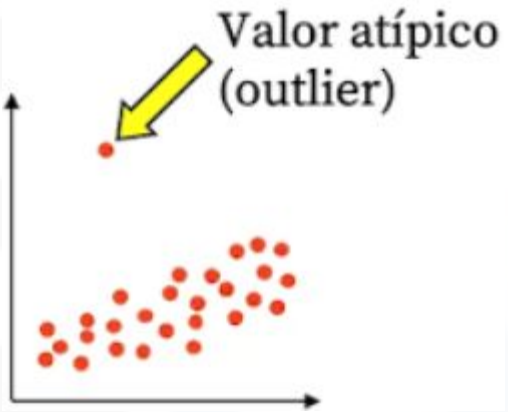
Se refieren a datos que están ausentes o no disponibles en un conjunto de datos.

Esto puede ocurrir por diversas razones, como errores en la recopilación de datos, problemas técnicos, omisión de información o simplemente porque algunos valores no fueron registrados.

Acciones: Analizar y eliminar o imputar

Valores atípicos

Son observaciones que se desvían significativamente del patrón general de un conjunto de datos y pueden tener un impacto importante en el análisis estadístico y en la interpretación de los resultados.
Pueden surgir por errores de medición, errores de entrada de datos, fenómenos inusuales en el mundo real o simplemente variabilidad natural en los datos.



Acciones: Analizar y permitir o transformar o truncar o eliminar

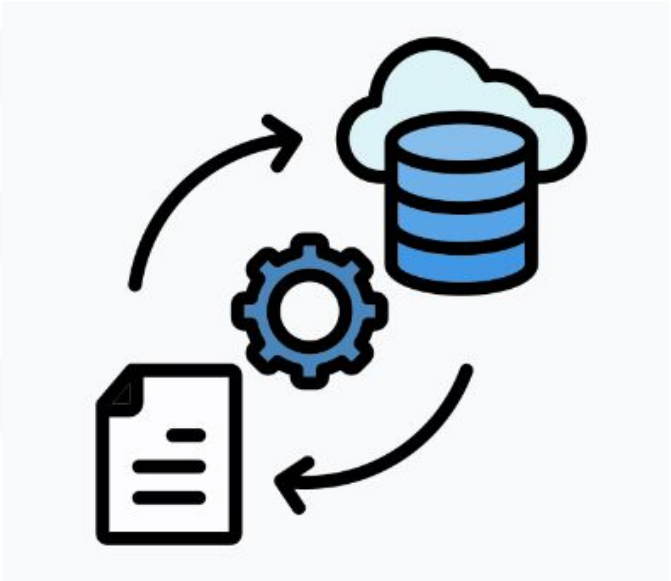
Etapa 3: transformación

Transformación

Objetivo: Preparar los datos para análisis y modelos

En esta etapa, los datos se pueden transformar mediante operaciones matemáticas, estadísticas o lógicas para crear nuevas variables o resaltar patrones.

Ejemplos incluyen la creación de variables derivadas (como calcular el índice de masa corporal a partir de peso y altura) o la conversión de variables categóricas en variables numéricas.



Tratamiento de variables categóricas

Variables ordinales

El tratamiento de variables ordinales implica manejar datos que tienen un orden o jerarquía predefinida, pero donde las diferencias entre los valores no necesariamente son iguales o bien definidas.

Estas variables son comunes en encuestas de opinión, escalas de calificación, niveles de educación, calificaciones de rendimiento, entre otros.

Acciones: asignar números a las categorías ordinales

Tratamiento de variables categóricas

Variables nominales

Son aquellas que representan categorías o etiquetas sin un orden inherente.

Algunos ejemplos comunes de variables nominales son el género, el estado civil, la ciudad de residencia, el color favorito, etc.

Acciones: codificar las variables

Normalización y estandarización

Normalización

Es útil cuando las características tienen diferentes rangos y se requiere que todas estén en un rango común.

Puede ser especialmente útil en algoritmos que utilizan distancias euclidianas, como el k-means o algoritmos basados en gradientes.

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \in [0\ ,\ 1]$$

Normalización y estandarización

Estandarización

Es útil cuando las características tienen diferentes escalas y distribuciones, y deseas que todas tengan una media de 0 y una desviación estándar de 1.

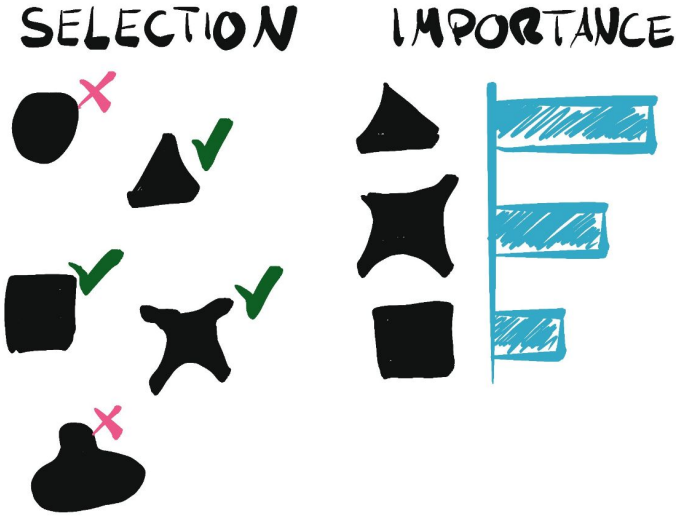
Algunos algoritmos, como SVM y regresión logística, pueden beneficiarse de características estandarizadas.

$$X_{standard} = \frac{X - \mu}{\sigma} \quad \in [0\ ,\ 1]$$

Selección de atributos

Implica identificar y elegir las variables más relevantes y significativas para un análisis o modelo en particular.

El objetivo principal de la selección de variables es mejorar la precisión y la eficiencia de los modelos, reduciendo el ruido y la complejidad innecesaria en los datos.



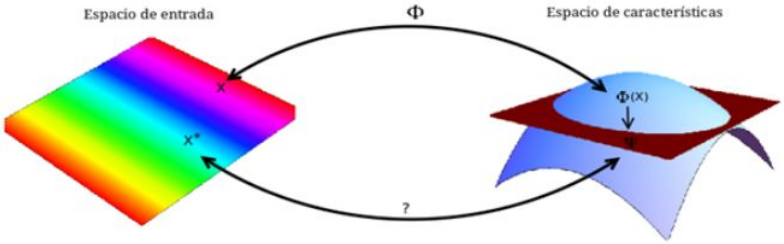
Reducción de la dimensionalidad

Extracción de características

Se crean nuevas características que son combinaciones lineales de las características originales.

Esto se hace mediante técnicas como el Análisis de Componentes Principales (PCA), que busca proyectar los datos en un nuevo espacio dimensional donde las primeras componentes capturen la mayor varianza de los datos.

Otros métodos de extracción de características incluyen el Autoencoder y el t-SNE (t-Distributed Stochastic Neighbor Embedding).



Etapa 4: modelado

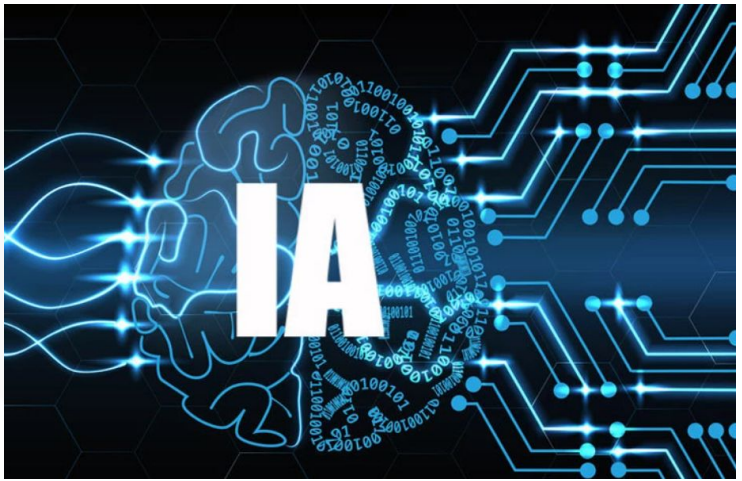
Análisis y modelado en la minería de datos

- En esta etapa, se aplican algoritmos de Minería de Datos para **descubrir patrones y relaciones en los datos**.
- Se utilizan técnicas de Aprendizaje Automático para **entrenar modelos que capturan la estructura subyacente de los datos**.

Nota: Los términos "algoritmos de minería de datos" y "técnicas de aprendizaje automático" a menudo se utilizan de manera intercambiable debido a su relación cercana, pero en realidad se refieren a conceptos ligeramente diferentes dentro del campo de la ciencia de datos.

Área de la minería de datos

Razonamiento inductivo en la Inteligencia Artificial



Algoritmos de minería de datos

Algoritmo	Descripción
Apriori	Utilizado en la minería de asociación para descubrir **patrones de co-ocurrencia** en conjuntos de transacciones, como compras de clientes.
Kmeans	Un algoritmo de agrupamiento que divide un conjunto de datos en clústeres o grupos basados en similitudes
DBSCAN	Otra técnica de agrupamiento que identifica clústeres en función de la densidad de puntos en el espacio de características
Jerárquico	Agrupa los datos de manera jerárquica, formando clústeres anidados.
PCA	Una técnica de reducción de dimensionalidad que proyecta los datos en un espacio de menor dimensión manteniendo la mayor varianza posible.

Algoritmos de minería de datos

Algoritmo / Análisis	Descripción
Redes sociales	Se utiliza para descubrir relaciones y patrones en redes sociales, como identificar comunidades o nodos influyentes.
Texto	Utilizado para extraer información y patrones de documentos de texto, como identificar temas, entidades y opiniones.
Secuencias	Utilizado para descubrir patrones en secuencias temporales, como en análisis de series de tiempo o secuencias genéticas.
Grafos	Explora las relaciones y conexiones entre entidades en forma de un grafo, revelando patrones y estructuras complejas.
Anomalías	Identifica puntos de datos atípicos o anómalos que difieren significativamente del comportamiento normal.

Técnicas de aprendizaje automático - supervisado

Técnica	Descripción
Regresión lineal	Predice valores numéricos continuos basados en relaciones lineales entre variables.
Regresión logística	Clasifica objetos en categorías discretas utilizando una función logística.
SVM	Clasifica objetos mediante la búsqueda de un hiperplano que mejor separe las clases.
Árboles de decisión	Utiliza una estructura de árbol para hacer decisiones clasificatorias o de regresión.
Bosques aleatorios.	Conjunto de árboles de decisión que combinan sus predicciones.

Técnicas de aprendizaje automático - supervisado

Técnica	Descripción
Redes Neuronales	Modelos inspirados en la biología que pueden manejar tareas complejas.
KNN	Clasifica objetos basados en la mayoría de las clases vecinas más cercanas.
SVM	Clasifica objetos mediante la búsqueda de un hiperplano que mejor separe las clases.
Naive Bayes	Usa el teorema de Bayes para calcular probabilidades de pertenencia a una clase.
XLM	(Máquinas de Aprendizaje Extremo) Extensión de SVM para manejar conjuntos de datos grandes y dimensionalidad alta.

Técnicas de aprendizaje automático - no supervisado

Algoritmos de minería de datos

Etapa 5: evaluación

Evaluación de modelos

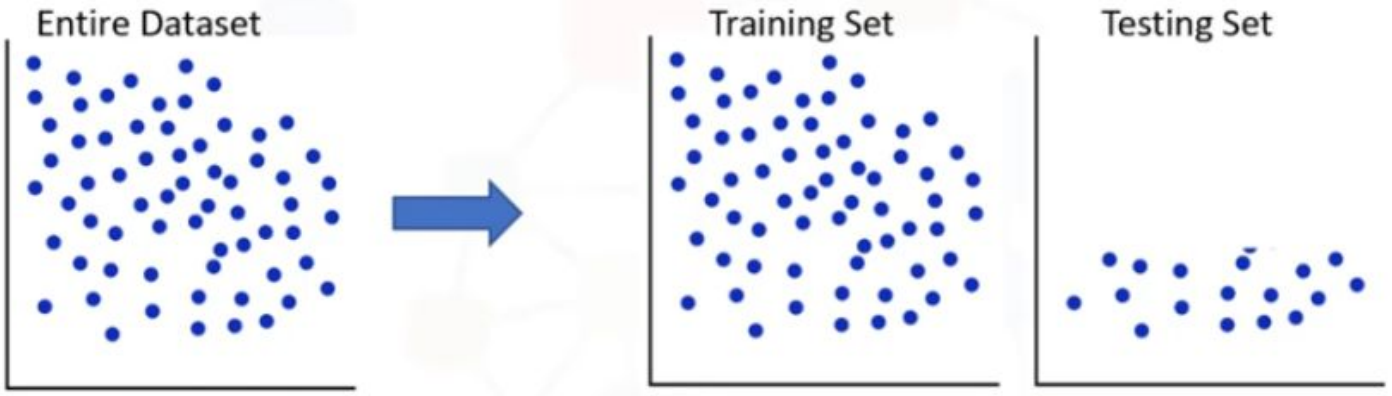
- Los modelos generados deben evaluarse para medir su rendimiento y su capacidad para generalizar a nuevos datos.
- Se utilizan métricas de evaluación como precisión, recall, F1-score y matriz de confusión, dependiendo del tipo de problema.

Importancia de la evaluación

- La evaluación permite determinar si un modelo es **adecuado para su implementación en situaciones reales** o si necesita ajustes adicionales.
- Ayuda a **comparar diferentes modelos** y técnicas para seleccionar la mejor opción para un problema específico.

Conjuntos de datos de evaluación

- Los conjuntos de datos utilizados para evaluar modelos se llaman conjuntos de prueba o conjuntos de validación.
- Estos conjuntos deben ser distintos de los datos de entrenamiento y representativos del mundo real.



Métricas de evaluación

- Las métricas se utilizan para medir la calidad del rendimiento del modelo. La elección de las métricas depende del tipo de problema (clasificación, regresión, etc.).
- Para problemas de clasificación, se utilizan métricas como precisión, recall, F1-score, matriz de confusión y curva ROC.
- Para problemas de regresión, se utilizan métricas como el error cuadrático medio (MSE) y el coeficiente de determinación (R^2).

Resumen

- Fundamentos de la Minería de Datos y sus aplicaciones.
- Etapas del proceso de Minería de Datos.
- Preprocesamiento básico en un conjunto de datos.
- La minería de datos en la IA

Preguntas…