Unidad 2

EL PROCESO DE EXTRACCIÓN DE CONOCIMIENTO (KDD)

Modelizado de Minería de Datos - 2Q2025

Índice

Proceso habitual de recopilación, almacenamiento y análisis

Un término muy utilizado, y el más relacionado con la minería de datos, es la extracción o "descubrimiento de conocimiento en bases de datos" (*Knowledge Discovery in Databases, KDD*)

Propiedades deseables del conocimiento extraído:

- Válidos: precisos para datos nuevos.
- Novedosos: desconocidos para el sistema y el usuario.
- Potencialmente útiles: conducen a beneficios.
- Comprensibles: fáciles de interpretar y usar.

Proceso habitual de recopilación, almacenamiento y análisis

Integración de datos: comprender el negocio, determinar fuentes útiles y obtenerlas.

Transformación a formato común: datawarehouse para unificar y resolver inconsistencias.

Selección, limpieza y transformación: correcciones, tratar nulos y atributos relevantes.

Minería de datos: elegir tarea (clasificación, agrupamiento, etc.) y método.

Evaluación e interpretación: evaluar e iteración si es necesario.

Difusión: usar y compartir el nuevo conocimiento.

Fase de integración y recopilación

El proceso KDD (descubrimiento de conocimiento en bases de datos) requiere:

- Reconocimiento
- Planificación
- Fuentes de datos
- Recopilación de “materia prima”

Surge el concepto de **data warehousing** (almacén de datos) para solucionar este problema.

Fase de transformación a formato común

Un **Datawarehouse** es un repositorio central donde se almacenan grandes volúmenes de datos de diferentes fuentes para análisis y toma de decisiones estratégicas, generalmente, a través de un proceso de ETL.

Su objetivo es integrar datos históricos y permitir la realización de consultas complejas.

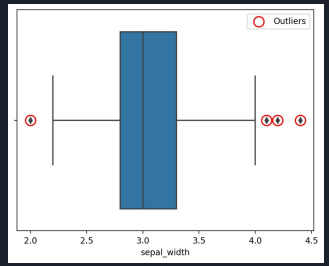
Visita sugerida:
https://aws.amazon.com/what-is/data-warehouse/

Fase de selección, limpieza y transformación

La calidad del conocimiento depende de la calidad de los datos

Problemas clásicos

- Outliers
- Nulos (*NaN*)



	TIME	FUNDS	x**2
0	2020-01	NaN	NaN
1	2019-12	2.0	4.0
2	2019-11	5.0	NaN
3	2019-10	NaN	49.0
4	2019-09	NaN	81.0
5	2019-08	11.0	121.0

Tratamiento

- Eliminarlos
- Imputar moda, media o mediana u otro valor representativo

Fase de selección, limpieza y transformación

La calidad del conocimiento depende de la calidad de los datos

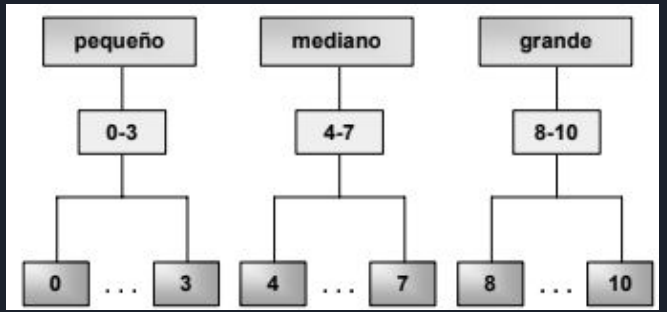
Transformaciones clásicas

- Numerización (variables dummies)
- Discretización

Income	Age	Marital Status
$45,000	23	Single
$48,000	25	Single
$54,000	24	Single
$57,000	29	Single
$65,000	38	Married
$69,000	36	Single
$78,000	40	Married
$83,000	59	Divorced
$98,000	56	Divorced
$104,000	64	Married
$107,000	53	Married

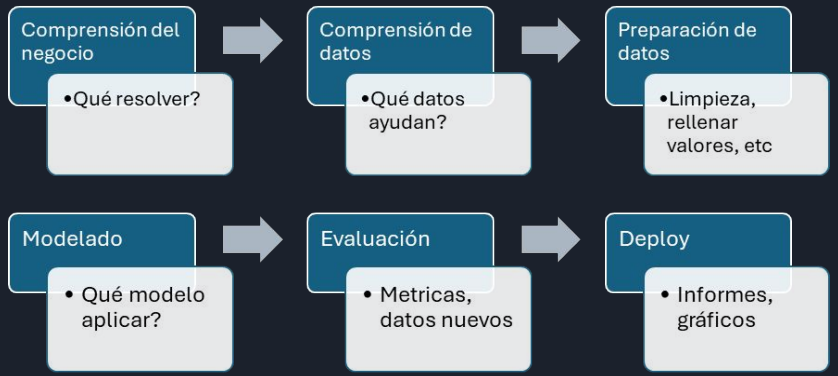
→

Income	Age	Married	Divorced
$45,000	23	0	0
$48,000	25	0	0
$54,000	24	0	0
$57,000	29	0	0
$65,000	38	1	0
$69,000	36	0	0
$78,000	40	1	0
$83,000	59	0	1
$98,000	56	0	1
$104,000	64	1	0
$107,000	53	1	0



Fase de minería de datos

Su objetivo es generar conocimiento a través de modelos que describen patrones y relaciones



Fase de minería de datos: Modelado

Seleccionar el modelo adecuado. Las tareas de minería de datos pueden ser predictivas (clasificación, regresión) o descriptivas (agrupamiento, reglas de asociación, correlaciones)

Modelos predictivos:

- Predicen valores futuros o clasificaciones basadas en datos históricos.
- Se utilizan para tomar decisiones informadas y anticipar resultados.
- Ejemplo: un modelo de regresión que predice las ventas futuras de un producto.

Modelos descriptivos:

- Describen los patrones y relaciones existentes en los datos.
- Ayudan a comprender la estructura y las características de los datos.
- Ejemplo: un modelo de agrupamiento que identifica grupos de clientes con comportamientos similares.

Tipos de modelos - Predictivos

Clasificación:

- Asigna una categoría o etiqueta a un dato basado en sus características.
- Ejemplo: un modelo que clasifica los correos electrónicos como spam o no spam.

Regresión:

- Predice un valor numérico continuo basado en las relaciones entre las variables.
- Ejemplo: un modelo que predice el precio de una casa basado en su tamaño y ubicación.

Tipos de modelos - Predictivos

Predicción:

- Identifica el valor de los datos en función de la descripción de otro correspondiente.
- Por ejemplo, en la detección de fraudes con tarjetas de crédito.

Time Series Analysis

- Sirve como variable independiente para predecir la variable dependiente en el tiempo.
- Ejemplo: tendencias, demanda, estacionalidad.

Tipos de modelos - Descriptivos

Secuencia:

- Busca relaciones entre eventos que ocurren en un orden específico.
- Ejemplo: un modelo que agrupa a los clientes en segmentos de mercado basados en sus comportamientos de compra.

Summarization (resumen):

- Condensa información de un texto extenso en una versión más corta y concisa.
- Ejemplo: Generar un resumen breve y preciso de un texto más largo.

Tipos de modelos - Descriptivos

Agrupamiento:

- Agrupa datos similares en clústeres o grupos basados en sus características.
- Ejemplo: un modelo que agrupa a los clientes en segmentos de mercado basados en sus comportamientos de compra.

Reglas de asociación:

- Descubre relaciones y dependencias entre variables en un conjunto de datos.
- Ejemplo: un modelo que identifica que los clientes que compran pan también tienden a comprar leche.

Fase de evaluación e interpretación

Objetivo: Evaluar la validez y utilidad de los patrones descubiertos y transformarlos en conocimiento comprensible.

Evaluación del Modelo: medir el rendimiento del modelo utilizando métricas relevantes (precisión, exactitud, etc.)

Comparar el modelo con otros modelos o estándares existentes: validar la robustez del modelo con nuevos datos.

Fase de evaluación e interpretación

Matriz de confusión: es una tabla que muestra el número de predicciones correctas e incorrectas realizadas por el modelo, en comparación con los valores reales. Permite evaluar el rendimiento del modelo en detalle, identificando errores específicos. Esencial para entender dónde falla el modelo y qué tipos de errores comete.

- Verdaderos positivos (VP): Casos en los que el modelo predijo correctamente la clase positiva.
- Verdaderos negativos (VN): Casos en los que el modelo predijo correctamente la clase negativa.
- Falsos positivos (FP): Casos en los que el modelo predijo incorrectamente la clase positiva (error tipo I)
- Falsos negativos (FN): Casos en los que el modelo predijo incorrectamente la clase negativa (error tipo II).

Fase de evaluación e interpretación

		Predicted Values	
		0	1
Actual Values	0	True Negative y_true : 0 y_pred : 0	False Positive y_true : 0 y_pred : 1
	1	False Negative y_true : 1 y_pred : 0	True Positive y_true : 1 y_pred : 1

Fase de evaluación e interpretación

Métricas: Supervisado - Clasificación

Accuracy (Precisión):

- Mide la proporción de predicciones correctas del modelo. `(VP+VN) / (VP+VN+FP+FN)`
- Es útil en problemas de clasificación con clases balanceadas.

Precission

- Del total de casos positivos que el modelo predice, cuántos son realmente positivos `VP/(VP+FP)`

Recall

- Del total de casos positivos reales, cuantos predijo correctamente el modelo `VP/(VP+FN)`

Fase de evaluación e interpretación

Métricas: Supervisado - Clasificación

Curva ROC (Receiver Operating Characteristic):

- Gráfico que muestra el rendimiento de un modelo de clasificación binaria a diferentes umbrales de clasificación.
- El eje Y representa la tasa de verdaderos positivos (TPR), y el eje X representa la tasa de falsos positivos (FPR).
- El área bajo la curva ROC (AUC) indica la capacidad del modelo para distinguir entre clases.

Fase de evaluación e interpretación

Métricas: Supervisado - Regresión

MSE (Error Cuadrático Medio):

- Calcula el promedio de los errores al cuadrado

RMSE (Raíz del Error Cuadrático Medio):

- Es simplemente la raíz cuadrada del MSE.

MAE (Error Medio Absoluto):

- Calcula el promedio de la diferencia absoluta entre las predicciones y los valores reales.

Fase de evaluación e interpretación

Los modelos no supervisados, como los de agrupamiento, no tienen un objetivo de predicción predefinido y trabajan sin etiquetas. Su objetivo es encontrar patrones y estructuras ocultas en los datos. Por lo tanto, su evaluación es más compleja y se centra en la calidad de los clústeres que se formaron.

- Índice de Silueta (Silhouette Score): Esta métrica mide qué tan bien cada objeto se agrupa dentro de su propio clúster en comparación con otros clústeres.
- Índice de Davies-Bouldin: Evalúa la calidad de un agrupamiento midiendo la relación entre la dispersión dentro del clúster y la distancia entre clústeres.

Fase de Difusión, Uso y Monitorización

Implementación del Modelo:

- Recomendar acciones a analistas.
- Aplicar el modelo a nuevos conjuntos de datos.
- Integrarlo en aplicaciones (ej., sistemas de análisis de crédito, filtros de spam).

Difusión del Conocimiento:

- Comunicar y distribuir el modelo a los usuarios a través de canales organizacionales (reuniones, intranet, etc.).
- Integrar el nuevo conocimiento en el know-how de la organización.

Fase de Difusión, Uso y Monitorización

Monitorización Continua:

- Evaluar el rendimiento del modelo a lo largo del tiempo.
- Reevaluar, re-entrenar o reconstruir el modelo debido a posibles cambios en los patrones de datos.
- Los factores externos pueden afectar el modelo, por lo cual la monitorización es muy importante.

Puntos Clave:

- El modelo debe ser utilizado y difundido para generar valor.
- La monitorización asegura que el modelo siga siendo preciso y relevante.

Proceso KDD

Conclusiones

- El proceso de KDD es iterativo: la construcción del modelo implica explorar alternativas, ajustar parámetros y posiblemente regresar a fases previas para optimizar el resultado.
- La elección de la tarea y el algoritmo influye en la preparación de los datos.
- Es crucial dividir los datos en conjuntos de entrenamiento y validación para asegurar la robustez y precisión del modelo.
- El objetivo final es encontrar el modelo que mejor resuelva el problema planteado.

Anexo: OLTP, OLAP y Datawarehouse

OLTP y **OLAP** son dos enfoques para gestionar bases de datos. OLTP se centra en transacciones en tiempo real, mientras que OLAP se centra en analizar datos

	OLTP	**OLAP** *Visita sugerida: https://aws.amazon.com/es/what-is/olap/*
Objetivo	Procesar transacciones comerciales	Analizar datos para tomas de decisiones
Operaciones	Insertar, actualizar, eliminar	Consultas analíticas complejas
Actualizaciones	Real time	Programadas
Copias de seguridad	Muy frecuentes	Menos frecuentes
Ejemplos	MySQL, MariaDB, PostgreSQL	AWS RedShift, Google BigQuery