Taller PT1

Asignatura: Minería de Datos

Fecha: septiembre 10 de 2025

1. Introducción

- El resultado de este taller debe ser un notebook en un documento en *.pdf.
- Un notebook, con toda la estructura para ejecutarse directamente. Debe contener el código necesario para cargar los datos directamente del repositorio de Kaggle (sin necesidad de descargar los datos a local).

1.1. **Objetivos del taller:**

- **Comprender los diferentes conceptos de Minería de Datos:** Aplicar conceptos fundamentales como tipos de datos, calidad de los datos y preprocesamiento en un escenario real.
- **Análisis Exploratorio de Datos (EDA):** Utilizar Python para explorar y visualizar un conjunto de datos complejo y multimodal.
- **Implementación Notebook Python:** Demostrar la capacidad para estructurar código en un *notebook* de Python, utilizando clases, funciones y bibliotecas especializadas.

1.2. **Descripción del Dataset utilizado en esta actividad:**

- **Dataset de Olist:** Conjunto de datos de E-commerce brasileño, destaca su naturaleza multimodal (datos de ventas, clientes, geolocalización, reseñas, entre otros).
- **Estructura del Dataset:** Múltiples tablas (.csv), requiere implementar fusión para un análisis coherente.
- **Estructura del notebook:**
    - **Encabezado**: utilice ## para nombres integrantes, utilice #### para: correos integrantes (e-mails UV), fecha, título en Markdown.
    - **Sección**: cada sección y explicación del algoritmo implementado debe ir en un Markdown independiente sobre la sección de código implementada.

2. Implementación.

2.1. **Exploración inicial de los datos:**

- **Código:** usted debe incluir el código de importación de bibliotecas (pandas, numpy, matplotlib, seaborn, otros), la carga de cada uno de los archivos ".csv" en DataFrames. Deberá implementar una función "DataProcessor" para el procesamiento de los datos. Todo el código debe realizarse en un "notebook"

- **Descripción:** debe explicar el propósito de cada biblioteca y mostrar un breve resumen de las primeras filas *.head() de los datos, la información del DataFrame .info() y las estadísticas descriptivas *.describe() para cada tabla cargada.
- **Datos:** En esta oportunidad, se debe trabajar con la base de datos “Brazilian_E-Commerce”, disponible en Kaggle: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

2.2. **Preprocesamiento y fusión de datos:**

- **Código de la Clase DataProcessor:** Debe incluir, por lo menos**:**
    - **Método clean_data():** Eliminación de valores nulos y duplicados.
    - **Método merge_tables():** Fusión de las tablas, explicando el tipo de *join* (how='inner' o how='left'), otros.
    - **Método feature_engineering():** Creación de nuevas características, como el cálculo del tiempo de entrega o el valor total del pedido.
- **Descripción:** Detallar cada método de la clase. Justificar las decisiones de preprocesamiento, como la imputación o eliminación de datos. Explicar el flujo de la fusión de tablas, identificando las claves primarias y foráneas.

3. Análisis de la calidad de los datos

- **Valores nulos y duplicados:**
    - **Código:** Verificar la existencia de valores nulos y duplicados en el DataFrame final (master_df).
    - **Resultados:** Presentar la salida del código (isnull().sum(), duplicated().sum()) y confirmar la limpieza de los datos.
- **Tipos de datos:**
    - **Código:** Muestre el resultado de master_df.info() después de la fusión.
    - **Descripción:** Asegurar que las columnas de fecha y hora se hayan convertido correctamente al tipo “datetime” y que los tipos de datos sean adecuados para el análisis.

4. Medidas de similitud y distancia

- **Descripción:**
    - **Propiedades de distancia:** Explicar brevemente cómo la distancia de **Haversine** cumple con las propiedades de una métrica de distancia (no-negatividad, identidad, simetría y desigualdad del triángulo).
    - **Aplicación:** Argumentar por qué esta medida es más adecuada que la distancia euclidiana para datos geográficos.
    - **Comparación:** comparar los resultados al aplicar esta métrica vs la distancia de Manhattan y ;ahalanobis

- **Aporte adicional:** Implemente al menos una medida de similitud y distancia (diferentes a lo solicitado) que usted considere apropiados según su análisis y datos.
- **Visualización (Mapa KDE):**
    - **Código:** Implementar el código para el gráfico de densidad (KDE) de las coordenadas de los clientes o vendedores.
    - **Análisis:** Interpretar el gráfico, identificando las zonas de mayor concentración de actividad de E-commerce.

5. Medidas de correlación

- **Correlación de Pearson:** Implementar el código para calcular la correlación entre variables numéricas, según su mejor decisión.
- **Correlación de Spearman:** Implementar el código para el análisis con Spearman en el caso de relaciones no lineales, según su mejor decisión.
- **V de Cramer:** Implementación y cálculo de la V de Cramer para la asociación entre variables categóricas.
- **Análisis:**
    - **Resultados numéricos:** Presentar y explicar los valores obtenidos para cada tipo de correlación.
    - **Mapa de calor:** Adjuntar el mapa de calor generado y discutir las relaciones entre las variables, por ejemplo, price vs. review_score, entre otras combinaciones.
- **Aporte adicional:** Implemente al menos una métrica más (diferente a lo solicitado) que usted considere apropiado según su análisis y requerimientos de los datos.

6. Visualizaciones avanzadas con Seaborn y Matplotlib

- **Códigos y gráficos:**
    - **Boxplot:** Distribución de precios por estado del cliente.
    - **Gráfico de Violín:** Distribución del tiempo de entrega por estado.
    - **Gráfico de Pastel:** Distribución de las 10 categorías de productos más vendidas.
    - **Pairplot:** Exploración de relaciones entre múltiples variables numéricas.
- **Aporte adicional:** Implemente al menos dos gráficos más (diferentes a los solicitados) que usted considere apropiados según su análisis.
- **Análisis:** Para cada gráfico, proporcionar una breve interpretación de lo que revelan los datos. No se limite solo a características visibles u “obvias” en los gráficos.

7. Conclusiones y desafíos

- **Hallazgos clave:** Resumir los “insights” más importantes obtenidos del análisis exploratorio, como la relación entre la distancia y el tiempo de entrega o las categorías de productos más populares, entre otros aspectos relevantes.
- **Limitaciones:** Identificar las limitaciones del análisis realizado (datos geográficos aproximados, falta de información sobre devoluciones, otros).

Rúbrica de evaluación del taller (PT1)

Criterio	Descripción	Puntaje máximo
Formato y estructura del notebook	Encabezado completo (nombres, correos, fecha, título), uso correcto de Markdown para secciones y explicaciones, según las indicaciones. Notebook ejecutable sin errores. Exportado correctamente a PDF.	10
Objetivos y comprensión del problema	Claridad en la presentación de los objetivos del taller. Demuestra comprensión del propósito del análisis, del dataset multimodal y su desarrollo.	5
Carga y exploración inicial de datos	Importación correcta de bibliotecas, carga de archivos desde Kaggle, uso de funciones y argumentos. Explicación clara del propósito de cada biblioteca y función en el código.	10
Implementación del algoritmo	Implementación funcional de los métodos “clean_data(), merge_tables(), feature_engineering(), otros” de la función solicitada. Justificación de decisiones de preprocesamiento y fusión de los datos.	15
Análisis de calidad de los datos	Verificación de nulos y duplicados, conversión correcta de tipos de datos, especialmente fechas. Explicación clara de resultados y gráficos de los resultados obtenidos.	10
Medidas de similitud y distancia	Explicación teórica de la distancia de (Haversine y demás solicitadas) y comparación con las otras métricas. Implementación correcta de cálculo y visualización “KDE” e interpretación del gráfico/resultados obtenidos.	10
Medidas de correlación	Implementación de Pearson, Spearman y V de Cramer. Presentación de resultados numéricos. Mapa de calor diseñado, otros gráficos y su análisis.	10
Visualizaciones avanzadas	Implementación de los gráficos solicitados (boxplot, violín, pastel, pairplot, otros solicitados). Interpretación profunda de cada gráfico en relación con los tipos de datos.	10

Conclusiones y desafíos	Identificación de hallazgos clave. Reflexión sobre limitaciones del análisis. Redacción clara y crítica de los resultados y análisis.	10
Creatividad, aporte y profundidad del análisis	Aporte de ideas propias, métodos, gráficos, visualizaciones adicionales relevantes y análisis más allá de lo solicitado en este documento del taller.	10