Proyecto Final de Introducción a Ciencia de Datos

Gustavo Lopez - Josué Arbulú

May 17, 2025

Introducción

El presente proyecto busca aplicar de forma práctica los conocimientos fundamentales de Ciencia de Datos adquiridos a lo largo del curso. Está dividido en dos fases (entregas) que abarcan de forma progresiva el ciclo completo de un proyecto real de análisis de datos: desde la identificación del problema, adquisición y preparación de un conjunto de datos, pasando por su exploración estadística, hasta la construcción de un modelo predictivo usando algoritmos de Machine Learning.

El objetivo principal es que cada estudiante sea capaz de estructurar un proyecto ordenado, justificado y reproducible, tal como se espera en entornos profesionales. Se espera rigurosidad en la limpieza de datos, claridad en los gráficos y análisis, y una explicación detallada de cada decisión tomada.

Lista de Datasets Disponibles

Cada grupo deberá elegir uno de los siguientes conjuntos de datos para desarrollar su proyecto. Los datasets se dividen en dos tipos de tareas: **clasificación** y **regresión**.

Tareas de Clasificación

1. **Detección de Diabetes - Grupo 5 (Líder: Alba)**
   Pima Indians Diabetes Database
2. **Fuga de Clientes Telco - Grupo 6 (Líder: Getulio Marin)**
   Telco Customer Churn
3. **Detección de Fraudes de Seguros - Grupo 2 (Líder: Gianfranco Mauro))**
   Vehicle Claim Fraud Detection
4. **Fallas en Máquinas Industriales - Grupo 3 (Líder: Bricet)**
   Machine Predictive Maintenance
5. **Fuga de Clientes de Banco - Grupo 8 (Líder: Marcelo Salinas)**
   Bank Customer Churn Prediction

Tareas de Regresión

1. **Precio de Autos Usados - Grupo 7 (Líder: Brandon)**
   Craigslist Car Listings
2. **Consumo de Combustible (Auto MPG) - Grupo 9 (Líder: Santiago Quiñones)**
   UCI Auto MPG Dataset
3. **Venta de Videojuegos - Grupo 1 (Líder: Alondra)**
   Video Game Sales
4. **Precio de Propiedades en Melbourne - Grupo 4 (Líder: Leonardo)**
   Melbourne Housing Market

Estructura General del Proyecto

Cada entrega debe incluir dos productos clave:

- Un **Jupyter Notebook (el archivo .ipynb del Google Colab)** documentado, ordenado y comentado.
- Una **presentación visual (en formato .ppt)** que resuma y explique el proceso de análisis, resultados obtenidos y conclusiones.

Importante: Todas las decisiones deben estar justificadas técnica o estadísticamente. Cada gráfico debe ir acompañado de su interpretación y el porqué de su uso.

Entrega P1 – Exploración y Preprocesamiento

Objetivo

Aplicar las etapas iniciales de un proyecto de Ciencia de Datos:

- Comprensión y contextualización del dataset
- Limpieza de datos
- Análisis exploratorio con visualizaciones

Estructura esperada del Notebook (Referencial)

1. **Contexto del Dataset**
   - Descripción del problema.
   - Fuente del dataset.
   - Objetivo del análisis (¿qué se desea predecir o entender?).
2. **Carga y Exploración Inicial**
   - Dimensiones y estructura del dataset.
   - Tipos de variables y primeras estadísticas con `.describe()`, `.info()`.
   - Visualización de la distribución de valores nulos, etc.
3. **Preprocesamiento de los datos (Data Wrangling)**
   - Identificación y tratamiento de valores nulos y duplicados.
   - Transformación de tipo de datos.
   - Cambio de nombre de columnas.
   - Justificación de métodos de imputación: media, eliminación, etc.
   - Enriquecimiento de los datos - Creación de nuevas variables.
     (Ejemplo: Cálculo de la edad a partir de la fecha de nacimiento, etc.)
4. **Análisis Exploratorio de Datos (EDA)**
   - **1. Análisis Univariado**
     - **1.1 Con gráficos**
       - **Boxplots:** Para cada variable numérica, se grafican boxplots que permiten visualizar la dispersión, posibles outliers y simetría. Se deben comentar los hallazgos relevantes por variable.
       - **Histogramas:** Se genera un histograma por cada variable numérica para analizar la distribución de frecuencias. Se debe interpretar la forma de la distribución y su implicancia.

        * **Gráficos de barras:** Aplicables a variables categóricas, muestran la frecuencia de cada categoría. Se debe explicar qué representa cada gráfico y resaltar categorías dominantes o atípicas.
    - **1.2 Sin gráficos**
        * **Medidas de tendencia central:** Media, mediana y moda por variable.
        * **Medidas de dispersión:** Desviación estándar, varianza, rangos, cuartiles.
        * Se debe interpretar cada estadístico en el contexto de los datos, destacando casos de alta dispersión o valores extremos.
- **2. Análisis Multivariado**
    - **2.1 Con gráficos**
        * **Mapa de calor de correlaciones:** Aplica a variables numéricas. Permite identificar relaciones lineales fuertes y posibles problemas de colinealidad.
        * **Boxplots comparativos:** Comparación de una variable numérica entre grupos definidos por una variable categórica.
        * **Histogramas y gráficos de barras agrupadas:** Comparaciones de distribuciones y frecuencias entre subgrupos.
        * **Diagramas de dispersión (scatterplots):** Se deben incluir al menos 5 combinaciones de pares de variables numéricas. Permiten observar tendencias, patrones lineales/no lineales y agrupamientos.
        * **Se debe incluir la interpretación de cada uno de los gráficos**
    - **2.2 Sin gráficos**
        * **Covarianza:** Medida de relación lineal entre pares de variables numéricas. Se interpreta su signo e intensidad.
        * **Correlación:** Se analiza el coeficiente (Pearson, Spearman, etc.), destacando relaciones fuertes o débiles, positivas o negativas.
- **3. Conclusiones Generales del EDA**
    - Resumen de hallazgos relevantes a nivel univariado y multivariado.
    - Discusión sobre la presencia de colinealidad y posibles relaciones lineales.
    - Identificación de variables clave que aportan valor al análisis.

Entregables de P1

- Notebook .ipynb bien documentado.
- Presentación en PPT o Canva con:
    - Introducción del dataset.
    - Proceso de limpieza y decisiones.
    - Principales insights obtenidos del análisis.

Entrega P2 – Modelamiento y Evaluación

Objetivo

Construir un modelo de aprendizaje supervisado para predecir o clasificar según el dataset elegido.

Estructura esperada del Notebook (Referencial)

1. **Preparación Final del Dataset**
    - Selección de features relevantes - Eliminación de variables irrelevantes.
    - Medición de la fuerza predictora de las variables
    - Codificación de variables categóricas.
    - Normalización o estandarización si es necesario, etc.
2. **División en Conjuntos de Entrenamiento y Prueba**
    - Uso de `train_test_split`.
    - Justificación del porcentaje elegido (ej. 80/20 o 70/30).
3. **Entrenamiento de Modelos de Machine Learning**
    - **1. Selección de Modelos**
        - Selección de al menos dos modelos adecuados al tipo de problema (regresión o clasificación). En caso escoja un modelo fuera de la lista proporcionada, justifique su elección.
        - **Modelos para problemas de regresión:**
            * Regresión Lineal Múltiple
            * Regresión Polinomial
            * Regresión No Lineal Múltiple
            * K-Nearest Neighbors (KNN) para regresión
        - **Modelos para problemas de clasificación:**
            * Regresión Logística
            * Support Vector Machines (SVM)
            * K-Nearest Neighbors (KNN) para clasificación
        - Justificación técnica y conceptual para cada modelo seleccionado en función de los datos y del objetivo del análisis.
    - **2. Entrenamiento y Ajuste de Modelos**
        - Entrenamiento de los modelos seleccionados.
        - Ajuste de hiperparámetros utilizando técnicas como grid search o random search.

- **3. Evaluación de Modelos**
  - **Métricas para regresión:**
    - RMSE
    - MAE
    - R^2
  - **Métricas para clasificación:**
    - Accuracy
    - F1-Score (Precision y Recall)
    - Matriz de Confusión

4. **Comparación de Modelos y Conclusiones**
   - Comparación del rendimiento de los modelos según las métricas aplicadas.
   - Discusión de ventajas, limitaciones y comportamiento de cada modelo.
   - **Conclusiones generales:**
     - Evaluación del impacto de los datos en el rendimiento (balanceo, calidad, outliers).
     - Recomendaciones sobre mejoras de datos (nuevas variables, transformación, ingeniería de features).
     - Consideraciones para futuras implementaciones (modelos más complejos, reducción de dimensionalidad, etc.).

Entregables de P2

- Notebook .ipynb con código comentado y estructura clara.
- Presentación (PPT o Canva) que incluya:
  - Elección del modelo y razonamiento.
  - Resultados y métricas.
  - Conclusiones técnicas.

Notas Finales

- Toda visualización debe incluir su **razón de ser** y una **interpretación clara**.
- Toda transformación debe estar respaldada por una **justificación técnica**.
- La organización del notebook, calidad del código y claridad explicativa son claves para una buena evaluación.
- Puedes usar librerías como `pandas`, `numpy`, `matplotlib`, `seaborn`, `scikit-learn`, entre otras.