Proyecto 2 – Introducción a la ciencia de los datos
Héctor Fabio Ocampo Arbeláez
R.A. 1

Objetivo del Proyecto

El propósito de este proyecto es que los estudiantes apliquen los conocimientos adquiridos en análisis exploratorio de datos, normalización y limpieza de datos y modelos de aprendizaje automático.

Grupos de 3 personas (Los mismos grupos del proyecto 1)

Data Set

Data Set:

- https://www.kaggle.com/datasets/neuromusic/avocado-prices/data

El objetivo es predecir el valor de un aguacate según las caracteristicas.

O pueden buscar un dataset a su gusto en el siguiente link:
https://www.kaggle.com/datasets?search=prices
Nota: el DataSet debe tener variables categóricas, variables numéricas, una variable objetivo y más de 1000 registros.

Desarrollo del proyecto

1. Análisis Descriptivo del Dataset (20%)

Los estudiantes deben:

- Cargar el dataset en Python usando **Pandas**.
- Generar **visualizaciones exploratorias** con **Matplotlib y Seaborn**, como:
  - Histogramas de las variables numéricas.
  - Boxplots para detectar outliers.
  - Correlación entre variables mediante heatmaps.
  - Y las que considere necesarias para entender el dataset

Proyecto 2 – Introducción a la ciencia de los datos
Héctor Fabio Ocampo Arbeláez
R.A. 1

Entregables:

- Código Python del análisis exploratorio.
- Gráficos explicativos con descripciones.
- Documento con insights obtenidos.

2. Limpieza y Normalización de Datos (30%)

Los estudiantes deben:

- Manejar los **valores nulos** con diferentes estrategias.
- Detectar y tratar **valores atípicos.**.
- Convertir **variables categóricas** a dummies.
- Estandarizar y normalizar datos..

Entregables:

- Código Python del proceso de limpieza y normalización.
- Documentación de los pasos seguidos y justificación de las decisiones.

3. Implementación de Modelos Predictivos (30%)

Los estudiantes deben entrenar y evaluar al menos **tres modelos de Machine Learning** para una variable objetivo del dataset.
Posibles modelos a utilizar:

1. **Regresión** (para datasets con variables continuas):
    a. Regresión Lineal.
    b. Random Forest Regressor.
    c. Redes Neuronales.
2. **Clasificación** (para datasets con categorías discretas):
    a. K-Nearest Neighbors (KNN).
    b. Árboles de Decisión.
    c. Support Vector Machines (SVM).
3. **Agrupamiento (Clustering)**:
    a. **K-Means** para segmentar los datos.
    b. Evaluación con métricas como **Silhouette Score**.

Se debe comparar la precisión de los modelos mediante métricas como **MSE, R^2, Accuracy o F1-score**.

Entregables:

- Código Python con la implementación de los modelos.
- Comparación de métricas de desempeño.
- Documento con el análisis y justificación de los modelos seleccionados.

4. Conclusiones y Presentación Final (20%)

En esta fase, los estudiantes deben:

- **Analizar los hallazgos obtenidos** en cada etapa del proyecto.
- **Explicar la efectividad de los modelos predictivos** y sugerir posibles mejoras.
- **Elaborar visualizaciones finales** con los resultados de los modelos.

Entregables:

- Informe final con:
  - Análisis del dataset.
  - Técnicas de limpieza y normalización utilizadas.
  - Modelos entrenados y comparación de desempeño.
  - Conclusiones y posibles mejoras.

Evaluación del Proyecto

Fase	Ponderación (%)
Análisis Descriptivo	20%
Limpieza y Normalización de Datos	30%
Implementación de Modelos Predictivos	30%
Conclusiones y Presentación Final	20%
Total	**100%**