

Laboratorio de Machine Learning y Deep Learning

UNIVERSIDAD EAFIT

Objetivo:

El propósito de este laboratorio es aplicar conceptos fundamentales de Machine Learning (ML) y Deep Learning (DL) para el análisis, preprocesamiento, entrenamiento y evaluación de modelos predictivos. El estudiante deberá realizar un proceso completo de análisis de datos y modelado supervisado, tomando decisiones fundamentadas en la exploración y resultados obtenidos.

Dataset

Cada grupo puede escoger alguno de los siguientes datasets de Kaggle:

<https://www.kaggle.com/datasets/muhammetvar1/laptop-price>

<https://www.kaggle.com/datasets/hellbuoy/car-price-prediction>

<https://www.kaggle.com/datasets/poushal02/student-academic-stress-real-world-dataset>

<https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset>

Instrucciones

1. Análisis preliminar del problema

Para el dataset seleccionado:

- Determine si se trata de un problema de clasificación o regresión. Justifique su respuesta e indique claramente el target (variable objetivo).
- Clasifique las características en tipos de variables (numéricas, categóricas, binarias, ordinales, etc.).
- Investigue y explique el protocolo de adquisición y/o generación de datos que siguieron los investigadores.

2. Análisis exploratorio de datos (EDA)

Realice un EDA completo sobre el dataset:

- Distribuciones de las variables.
- Estadísticos descriptivos.
- Correlaciones entre variables.

- Relación entre variables predictoras y el target.

Cada gráfico o estadística debe ir acompañado de una interpretación detallada, en la cual explique qué información aporta al problema de ML.

Utilice librerías como pandas, numpy, matplotlib y seaborn.

3. Procesamiento de datos

Aplique buenas prácticas de procesamiento y limpieza de datos:

- Manejo de valores nulos.
- Codificación de variables categóricas.
- Normalización o estandarización si aplica.
- Reducción de dimensionalidad si se justifica.

Implemente un pipeline de procesamiento con scikit-learn.

Divida los datos en X_{train} , X_{val} y X_{test} con proporciones justificadas (ej. 70/15/15).

4. Entrenamiento de modelos

a. Entrene y evalúe al menos 3 modelos distintos sobre su dataset:

- **k-Nearest Neighbors (kNN).**
- **Modelo de ensamble** (Random Forest o Gradient Boosting).
- **Deep Neural Network (DNN)** (mínimo 3 capas ocultas, con funciones de activación y regularización).

Muestre los resultados en una tabla comparativa generada en Python, donde se evidencie el desempeño de cada modelo en X_{train} , X_{val} y X_{test} .

b. Responda:

- ¿Cuál modelo tuvo mejor desempeño?
- ¿Alguno presentó overfitting o underfitting? ¿Cómo lo detectó?
- ¿Cuál seleccionaría para producción y por qué?

5. Prueba con muestra artificial

Genere una muestra artificial (nueva) con características inventadas, ingrésela al modelo seleccionado y analice la predicción.

Explique:

¿El resultado tiene sentido?

¿Qué pasaría si modificara una o más variables de la muestra?

6. Investigue y explique las siguientes estrategias:

- K-Fold Cross Validation.
- Leave-One-Out Cross Validation (LOOCV).

Responda:

¿Son aplicables estas estrategias al dataset elegido?

¿Qué beneficios tendrían frente al esquema de validación tradicional (train/val/test)?

ENTREGABLES

Se entregará un proyecto completo con el desarrollo completo del laboratorio, incluyendo código, análisis e interpretaciones. (Usar la Wiki para el desarrollo de los numerales y poner una introducción en el Readme)

Incluir un Diagrama de flujo (en cualquier formato) que represente el pipeline de procesamiento y modelado.

Tabla comparativa final de desempeño de modelos.

Respuestas a las preguntas de análisis planteadas.

Criterios de evaluación

- Claridad en la justificación teórica **(20%)**.
- Calidad del EDA e interpretaciones **(20%)**.
- Correcto preprocesamiento de datos y pipeline **(20%)**.
- Implementación y comparación de modelos **(25%)**.
- Prueba con muestra artificial y análisis de validación cruzada **(15%)**.