Proyecto I – Aprendizaje por Reflejo

Objetivo

El objetivo de este proyecto es diseñar, entrenar y evaluar modelos de Machine Learning para un problema de regresión utilizando datos reales. Se espera que los estudiantes apliquen metodologías adecuadas de preprocesamiento, validación, comparación de modelos y análisis crítico de resultados.

El proyecto busca evaluar no solamente la capacidad de implementar algoritmos, sino también el criterio técnico en la toma de decisiones relacionadas con la preparación de datos, selección de modelos y evaluación del desempeño.

Modalidad de trabajo

- El proyecto se realizará en parejas o de manera individual.
- Los datasets proporcionados en el portal del curso son:
  - training_features.csv
  - training_target.csv
- Evaluación de modelo (presencial):s **Lunes 23 de febrero, 2026**
- Reporte escrito: **Viernes 27 de febrero, 2026**

Descripción

A partir de los datos proporcionados, cada grupo deberá desarrollar un modelo de regresión que permita predecir los valores contenidos en *training_target.csv* a partir de las variables en *training_features.csv*.

El trabajo deberá incluir obligatoriamente:

1. Implementación y evaluación de al menos dos algoritmos distintos de regresión.
2. Comparación cuantitativa del desempeño de ambos modelos.
3. Selección justificada del mejor modelo.

Pueden utilizar cualquier algoritmo de regresión que consideren adecuado.

Etapas del proyecto

Exploración y comprensión de los datos

Antes de entrenar modelos, deben realizar un análisis exploratorio que incluya:

- Inspección de dimensiones del dataset.
- Tipos de variables.
- Detección de valores faltantes.
- Identificación de variables categóricas y numéricas.
- Análisis preliminar de distribuciones y posibles outliers.
- Identificación de posibles correlaciones relevantes.

Este análisis debe orientar sus decisiones posteriores.

Preprocesamiento

Durante esta etapa deberán tomar decisiones justificadas sobre:

Selección de variables

- No es obligatorio utilizar todas las variables.
- Pueden descartar variables si lo consideran apropiado, pero deben justificarlo.

Valores faltantes

- Identificar su presencia.
- Decidir si imputarlos o eliminar observaciones.
- Justificar claramente el método utilizado (media, mediana, moda, modelo predictivo, eliminación, etc.).

Variables categóricas

- Determinar si son ordinales o nominales.
- Elegir una estrategia de codificación adecuada (por ejemplo: one-hot encoding, ordinal encoding).
- Justificar la elección.

Escalamiento

- Evaluar si el algoritmo elegido requiere normalización o estandarización.

Separación de datos

El dataset original deberá dividirse en:

- Training set
- Validation set

La evaluación del desempeño del modelo deberá realizarse exclusivamente sobre el validation set. También pueden evaluar el uso de Cross Validation, pero deberán explicar el funcionamiento en el trabajo escrito.

Entrenamiento y comparación de modelos

Cada pareja deberá:

- Entrenar al menos dos algoritmos distintos.
- Ajustar hiperparámetros si lo consideran necesario.
- Comparar el desempeño utilizando métricas adecuadas.

Persistencia del modelo

El mejor modelo desarrollado deberá guardarse utilizando el módulo *pickle* de Python. El archivo del modelo guardado será utilizado el día de la presentación.

Presentación final

El día de la presentación:

- Se les entregará un testing set nuevo. Este tendrá el mismo formato que el *training_set.csv,* en el caso que efectuén transformaciones al training set, asegurense de crear un pipeline que haga estas mismas transformaciones al testing set.
- Cada pareja deberá cargar su modelo previamente guardado.
- Generarán predicciones sobre el testing set.
- Se calculará el **MSE** sobre esas predicciones.

La pareja que obtenga el menor MSE en el testing set obtendrá automáticamente 100 puntos en el proyecto y no deberá entregar reporte escrito.

Reporte escrito

Deberán entregar un informe técnico que incluya:

1. Descripción del análisis exploratorio realizado.
2. Justificación de las decisiones de preprocesamiento.
3. Explicación conceptual de los algoritmos utilizados.
4. Resultados obtenidos en el validation set.
5. Comparación entre modelos.
6. Discusión crítica de resultados.
7. Propuestas de mejora si dispusieran de más tiempo.

El reporte debe demostrar comprensión de los conceptos y no limitarse a mostrar código o resultados numéricos.