Trabajo Final Inteligencia de Negocios 2025

Maestría en Economía Aplicada, Facultad de Ciencias Económicas de la Universidad de Buenos Aires

Condiciones de Entrega

a) La fecha límite de entrega es el domingo 20 de julio a las 23:59:59.
b) Deben enviar una un archivo .ipynb (notebook de jupyter o colab) con el formato tp_final_bi_2025_{apellido}.ipynb. Ejemplo: si el alumno se llamase Juan Pérez, debe enviar el archivo tp_final_bi_2025_perez.ipynb.
c) Este archivo debe ser adjuntado en un correo electrónico a fmastelli@gmail.com con el asunto "TP final bi 2025 {apellido}". Ejemplo: si la alumna se llamara Marta Calvo, debe enviar el email con el asunto "TP final bi 2025 Calvo".
d) El archivo que adjuntan en el mail debe estar totalmente ejecutado sin errores. Debe contener tanto el código como las explicaciones.

Conjunto de datos

a) Contarán con 2 datasets de publicaciones de venta de inmuebles:
   i) Entrenamiento
   ii) Test

Consignas

1) Lea el archivo "df_train.csv". ¿Qué puede decir acerca de la estructura del dataset? Mencione cantidad y tipos de columnas, datos faltantes, de qué va el conjunto de datos en términos generales.
2) Análisis exploratorio
   a) Obtener la matriz de correlaciones para las variables numéricas (ignore lat y lon). ¿Qué puede decir acerca de la correlación entre surface_total y surface_covered? ¿y entre rooms y bathrooms?
   b) ¿Cómo es la correlación de la variable a explicar, price, con el resto de las variables?
   c) Obtener estadísticas descriptivas para la variable target (price) y realizar un histograma de la misma. Comente los resultados obtenidos
   d) Obtener las mismas estadísticas descriptivas del target para cada tipo de propiedad y realizar boxplots paralelos de la variable según tipo de propiedad. ¿Qué diferencias encuentran entre los tipos de propiedad?
   e) Graficar un scatterplot de la variable price y surface_total. ¿Detecta alguna anomalía?
   f) Eliminar los outliers univariados de las variables price, rooms y surface_total. Utilizar y fundamentar el o los criterio/s y métodos que consideren adecuados. Deberá trabajar con este dataset filtrado en lo que resta del Trabajo Práctico
   g) Vuelva a realizar la matriz de correlaciones y el histograma de la variable price. Comente los cambios observados si los hubiera.
3) Modelado tradicional

Durante esta consigna, deberá trabajar **sin la columna de descripciones. Con el resto de columnas puede generar las transformaciones que considere conveniente.**

a) Ajuste un modelo lineal sobre el dataset filtrado.
   i) ¿Qué puede concluir e interpretar acerca del signo y tamaño de los coeficientes?
   ii) ¿Qué puede decir acerca de la significatividad estadística de los mismos?

b) Ahora ajuste un modelo lineal LASSO (previa estandarización/normalización de variables), optimizando el parámetro de penalización (lambda en nuestra clase teórica, alpha en sklearn). ¿Alguna variable quedó eliminada? Justifique.

4) Modelos de Aprendizaje

Durante esta consigna, deberá trabajar **sin la columna de descripciones. Con el resto de columnas puede generar las transformaciones que considere conveniente.**

a) Random Forest: Realice búsqueda de hiperparámetros (puede usar conjunto de validación o validación cruzada) sobre el dataset y obtenga un mejor modelo. Reporte métricas de validación (rmse, mae).
b) Boosting: Realice búsqueda de hiperparámetros (puede usar conjunto de validación o validación cruzada) sobre el dataset y obtenga un mejor modelo. Reporte métricas de validación (rmse, mae).
c) Redes Neuronales: ídem. Puede probar al menos 3 arquitecturas diferentes variando la cantidad de capas y neuronas de la red densamente conectada (Dense en keras).

5) Modelos de Aprendizaje + Procesamiento de Lenguaje Natural

Ahora sí vamos a **usar la columna de descripciones**

a) Representar vectorialmente la columna de descripciones. Puede recurrir a BoW, TF-IDF, o embeddings pre-entrenados (cualquiera de los modelos de huggingface vistos en clase). Elija la que considere más conveniente.
b) Incorporar esta representación en el dataset con el que venía trabajando.
c) Repita el paso 4, ahora con el nuevo dataset completo.

6) Performance

Tras las consignas 4 y 5 el alumno deberá tener a su alcance 6 modelos optimizados: Random Forest con y sin descripciones, algún modelo de Boosting con y sin descripciones, y redes neuronales con y sin descripciones.

a) Evalúe la performance de los 6 modelos sobre el conjunto de datos df_test.csv. Reporte RMSE y MAE. ¿Puede concluir que había información relevante en las descripciones que estaba omitida en el resto de los atributos?