Análisis y Modelado de Precios de Propiedades

CONTEXTO

Este examen tiene como objetivo evaluar la capacidad de los estudiantes para analizar, preprocesar y modelar un conjunto de datos utilizando *Random Forest Regressor*, además de visualizar y explicar los resultados. Se les proporcionará un dataset con características de propiedades en Lima, y deberán realizar cada una de las etapas mencionadas

1. **EXPLORACIÓN DE DATOS (5 pts.)**
   - 1.1. Cargar el dataset proporcionado
   - 1.2. Mostrar los tipos de datos de cada columna y la cantidad de null (una sola función)
   - 1.3. Mostrar un resumen estadístico de tanto variables numéricas y categóricas
   - 1.4. ¿Qué tipos de datos predominan en el dataset?
   - 1.5. Crear histogramas para visualizar la distribución de las características numéricas principales (elige al menos 3)
   - 1.6. ¿Observas algún patrón interesante o posible valor atípico en alguna característica?
2. **PREPROCESAMIENTO DE LOS DATOS (6 pts.)**
   - 2.1. Seleccionar las siguientes características para trabajar en el modelado: 'Distrito', 'Vista', 'Proximidad a transporte', 'Estado de Inmueble', 'Área', 'Precio'
   - 2.2. ¿Por qué crees que es importante utilizar 'Área' y 'Estado de Inmueble' en el modelo?
   - 2.3. Convertir las columnas categóricas seleccionadas ('Distrito', 'Vista', 'Proximidad a transporte', 'Estado de Inmueble') a tipo *category*
   - 2.4. Aplicar One-Hot Encoding a 'Distrito' y 'Vista'
   - 2.5. Aplicar Ordinal Encoding a 'Proximidad a transporte' y 'Estado de Inmueble'
   - 2.6. ¿Qué diferencias existen entre One-Hot Encoding y Ordinal Encoding? ¿Por qué usamos uno u otro en cada caso?
   - 2.7. Eliminar las filas que contengan valores nulos
   - 2.8. ¿Qué problemas podrían surgir si se eliminan demasiados registros con valores nulos?
3. **MODELADO (6 pts.)**
   - 3.1. Dividir el dataset en un conjunto de entrenamiento y un conjunto de prueba, utilizando una proporción del 80% para entrenamiento y 20% para prueba
   - 3.2. ¿Por qué es importante dividir el dataset antes de entrenar un modelo?
   - 3.3. Crear un modelo de *Random Forest Regressor* con 100 árboles (*n_estimators*=100) y entrenarlo con el conjunto de entrenamiento (aplicar paralelismo)
   - 3.4. Realizar el entrenamiento
   - 3.5. ¿Cuál es la ventaja de utilizar programación paralela en Random Forest?
   - 3.6. Utilizar el conjunto de prueba para predecir el precio de las propiedades
4. **VISUALIZACIÓN DE RESULTADOS (3 pts.)**
   - 4.1. Visualizar la importancia de cada característica utilizando un gráfico de barras
   - 4.2. ¿Qué características parecen ser las más importantes para predecir el precio de las propiedades?
   - 4.3. Crear un gráfico de dispersión para comparar las predicciones del modelo frente a los valores reales de 'Precio'
   - 4.4. ¿Qué tan bien se alinean las predicciones con los valores reales? ¿Qué podrías hacer para mejorar el modelo?