Prueba de Conocimiento Científico de Datos

Hebert Gomez

Proceso de Limpieza y Modelado de Avalúos de Vivienda

1. Carga y Limpieza de Datos

- **Carga de Datos:** Se cargó el conjunto de datos desde el archivo CSV train_precios_vivienda.csv.
- **Proceso de Limpieza:**
  - **Eliminación de Columnas:** Se eliminó la primera columna y la columna 'fecha_aprobación' del conjunto de datos, debido a que no es relevante para el cálculo del avalúo.
  - **Corrección de Codificación y Minúsculas:** Se implementó una función para corregir texto mal codificado y convertirlo a minúsculas, eliminando espacios adicionales.
  - **Eliminación de Filas:** Se eliminaron las filas donde el valor de 'valor_total_avaluo' era nulo o igual a cero.
  - **Normalización de Formato:** Se eliminaron los puntos de la columna 'valor_total_avaluo' y se guardó el DataFrame limpio en train_precios_vivienda_clean.csv.

Preprocesamiento y Transformación de Datos

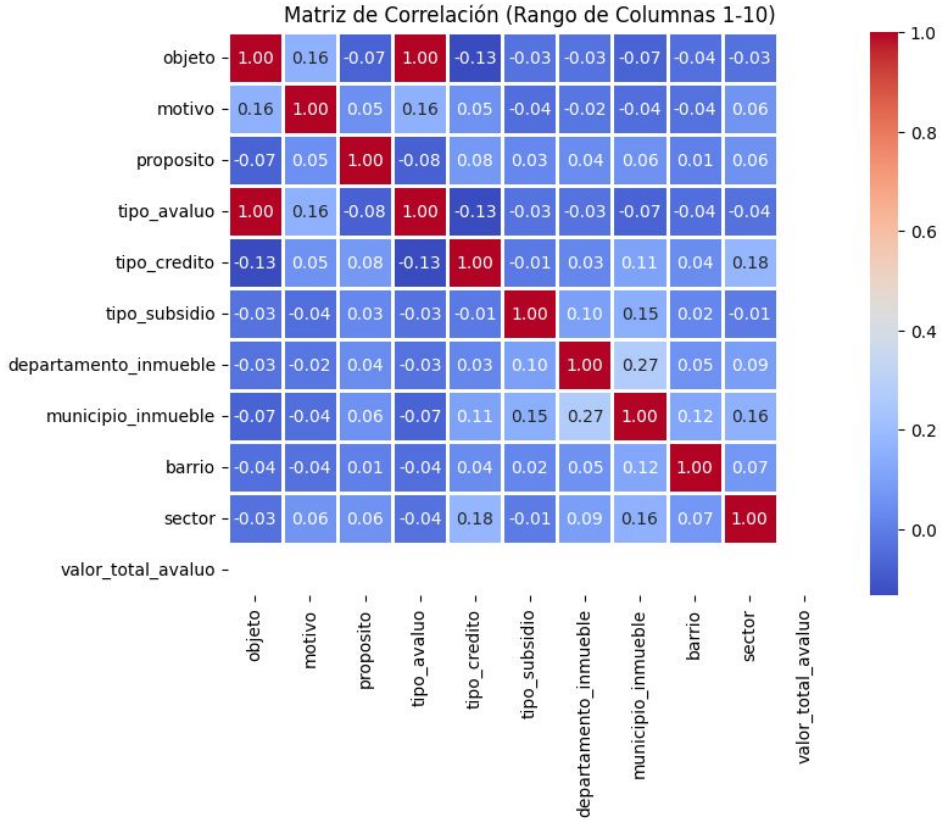
- **Codificación de Variables Categóricas:**
  - Se asignó un valor numérico único a cada cadena de texto en las columnas categóricas para su procesamiento posterior.
  - Se guardó el DataFrame codificado en `train_precios_vivienda_encoded_manual.csv`.

Este proceso se realizó debido a que se encontraron columnas con valores categóricos y que podrían influir en el resultado del cálculo del avalúo.

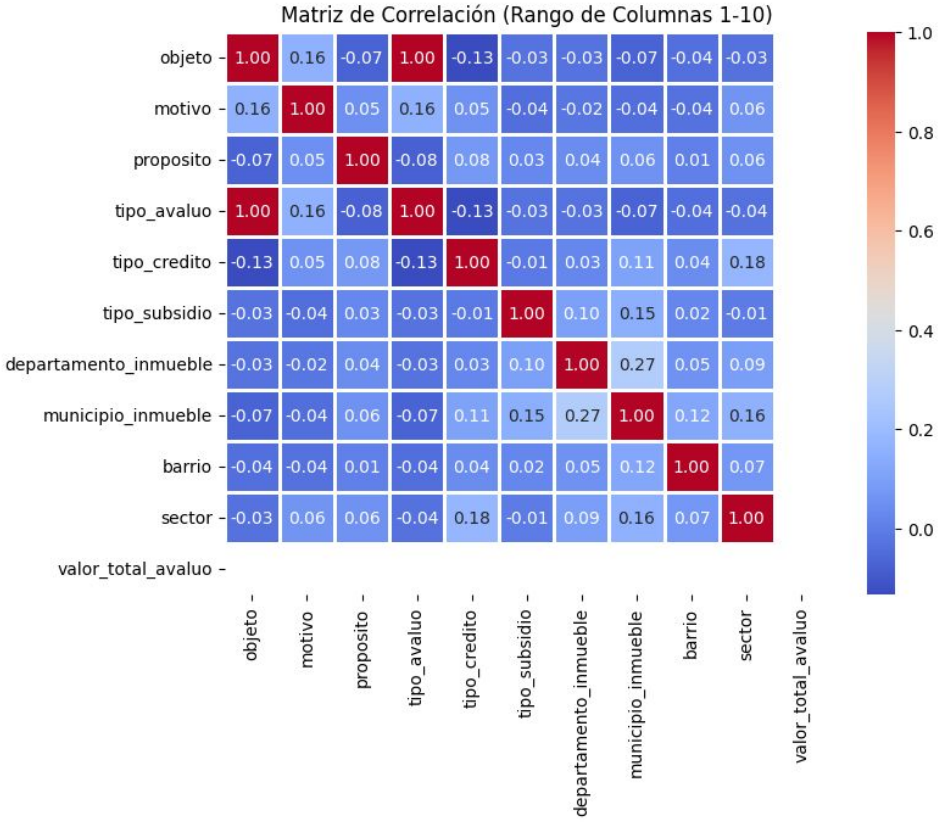
Análisis de Correlación

- **Visualización de Correlaciones:**
  - Se realizaron análisis de correlación por segmentos de columnas para explorar las relaciones entre las variables.
  - Se utilizó un mapa de calor para visualizar las matrices de correlación y entender las interacciones entre las variables.

Partiendo de los datos limpios y teniendo en cuenta que se tienen más de 200 parámetros, se realizan las gráficas de correlación en grupos de a 10, teniendo presente, la relación con el valor_total_avaluo en todas.



Al Iterar sobre las columnas de la matriz de correlación, se encuentra que las correlaciones son muy bajas, por debajo del 0.6. Esto puede deberse tanto a factores de toma de datos y calculo del valor del avalúo, como factores de manejo y limpieza de los datos.



Entrenamiento y Evaluación del Modelo

Modelado y Evaluación

- **Modelo de Regresión Random Forest:**
  - Se utilizó un modelo Random Forest para predecir los precios de vivienda.
  - Se implementó una búsqueda aleatoria de hiperparámetros para optimizar el modelo.
  - Se evaluó el modelo utilizando métricas como el Error Cuadrático Medio (MSE) y el Coeficiente de Determinación (R^2).

Basándonos en los resultados obtenidos del modelo de Regresión Random Forest y la búsqueda aleatoria de hiperparámetros, podemos realizar las siguientes conclusiones y análisis:

Entrenamiento y Evaluación del Modelo

Desempeño del Modelo de Regresión Random Forest:

- **Error Cuadrático Medio (MSE):** El MSE obtenido es extremadamente alto (aproximadamente 2.30e+21), lo cual indica que el modelo tiene una gran discrepancia entre los valores predichos y los valores reales. Esto sugiere que el modelo no está capturando de manera efectiva las relaciones entre las características de entrada y la variable objetivo.
- **Coeficiente de Determinación (R^2):** El valor de R^2 es negativo (-0.593), lo cual es muy bajo y sugiere que el modelo no es mejor que un modelo que simplemente predice la media de la variable objetivo. Esto indica un mal ajuste del modelo a los datos observados.

Entrenamiento y Evaluación del Modelo

Búsqueda Aleatoria de Hiperparámetros:

- Se realizó una búsqueda aleatoria para optimizar los hiperparámetros del modelo. Los mejores hiperparámetros encontrados fueron:
  - `n_estimators`: 100
  - `min_samples_split`: 2
  - `min_samples_leaf`: 2
  - `max_depth`: None
- A pesar de la optimización de hiperparámetros, el desempeño del modelo no mejoró significativamente en términos de MSE y R^2.

Entrenamiento y Evaluación del Modelo

Los resultados indican que el modelo actual no es adecuado para predecir los valores de `valor_total_avaluo` en este conjunto de datos. Es posible que se requiera un enfoque diferente.

La alta magnitud del MSE y el bajo valor de R^2 sugieren que podría haber problemas fundamentales en la calidad de los datos, la representación de las características o la elección del modelo.

Se recomienda explorar otras técnicas de modelado, como modelos de ensamble más avanzados, redes neuronales u otros modelos de regresión que puedan capturar mejor las complejidades de los datos.

Funciones para Limpiar Coordenadas Geográficas

Para lograr un análisis complementario del comportamiento del avalúo de un inmueble, con los parámetros del mismo, se debe considerar su geolocalización y su cercanía con puntos de interés.

Para esto, se tiene el documento “**PuntosInteres.csv**”, el cual contiene sitios de interés categorizados y relacionados con un punto cardinal.

Con el fin de lograr un acercamiento a la relación entre el valor del avalúo de un inmueble y su ubicación con respecto a los puntos de interés, se identificaron las coordenadas en un radio de aproximadamente 1 Km, hallando los puntos de interés asociados a cada inmueble.

Finalmente, con los puntos encontrados, se genera una columna con los resultados vectorizados, esto para que sea compatible con el input del modelo.

Funciones para Limpiar Coordenadas Geográficas

Limpieza y Conversión de Coordenadas Geográficas:

- Las funciones `clean_longitude` y `clean_latitude` están diseñadas para limpiar y convertir valores de longitud y latitud a un formato numérico adecuado. Esto incluye eliminar puntos y manejar el formato de cadena para asegurar que se puedan convertir correctamente a tipo `float`.

Lectura y Manipulación de Datos desde CSV:

- Se carga el archivo CSV `PuntosInteres.csv` utilizando `pd.read_csv()`, especificando el separador como `;` y nombres de columna personalizados.
- Se realizan operaciones para ajustar el formato de las columnas de longitud y latitud para asegurar que estén en un formato numérico adecuado.

Funciones para Limpiar Coordenadas Geográficas

Asignación de Categorías basada en Coordenadas:

- Un diccionario `dic` se utiliza para almacenar las categorías correspondientes a cada punto de interés identificado por `id`, basado en la proximidad de las coordenadas (longitud y latitud) a un conjunto de ubicaciones definidas (`df_loc`).
- Se itera sobre los datos del DataFrame `df` y `df_loc` para asignar las categorías correspondientes a cada punto de interés según su `id` y ubicación geográfica.

Funciones para Limpiar Coordenadas Geográficas

Entrenamiento de Modelo Word2Vec:

- Se tokenizan las categorías asignadas para cada punto de interés.
- Se entrena un modelo Word2Vec sobre estos tokens para obtener representaciones vectoriales de las palabras.
- Se define una función para obtener el vector de frase promedio de las categorías asignadas a cada punto de interés utilizando el modelo Word2Vec entrenado.

Creación de Vectores de Frases:

- Se aplica la función definida para generar vectores de frases para cada conjunto de categorías tokenizadas asignadas a los puntos de interés.
- Los resultados, que incluyen las categorías originales y sus representaciones vectoriales, se imprimen para su visualización.