Ejercicio 5: Documentación

El presente trabajo se ha basado en el conjunto de datos de la competencia Kaggle sobre estimación de precios de ventas de propiedades en la ciudad de Melbourne, Australia:
(https://www.kaggle.com/dansbecker/melbourne-housing-snapshot)

Se ha utilizado el conjunto de datos reducido producido por DanB:
(https://www.kaggle.com/dansbecker).

Se ha subido una copia a un servidor de la Universidad Nacional de Córdoba para facilitar su acceso remoto:
(https://cs.famaf.unc.edu.ar/~mteruel/datasets/diplodatos/melb_data.csv)

melb_df : Es el dataset con los datos que surgen de la URL anterior.

airbnb_df : Es el dataset con información adicional respectiva al entorno de una propiedad, que se puede encontrar en la siguiente URL:
(https://www.kaggle.com/tylerx/melbourne-airbnb-open-data?select=cleansed_listings_dec18.csv)

Criterios de exclusión de ejemplos y características

1. Se ha explorado el dataset melb_df para determinar las variables que contienen datos NAN, identificando un faltante de datos significativos en las columnas BuildingArea y YearBuilt, los que fueron imputados a posterior con el método KNN.
2. Se definió una función y clases para automatizar la ejecución de gráficos de Boxplots, Histplots y Displots. En base a la cantidad de columnas a graficar, la función display_plots y las clases CustomBoxPlot, CustomHistPlot y CustomDistPlot deciden cómo trabajar los axes, cómo mostrar los labels y si mostrar escalado o no los datos.
3. Se definió una función y clase para tratar los outliers, estableciendo dos estrategias para su tratamiento:
   - IqrStrategy: Todas los datos que no caen en el intervalo (Q1 – (Q3 – Q1) *1.5 , Q3 + (Q3 – Q1)* 1.5) son eliminados.
   - UpperLowerStrategy: Todos los datos que no caen en (Límite Inferior, Límite Superior) son eliminados. Los límites son expresados en el rango de 0 – 1, donde por ejemplo UpperLowerStrategy (0.1, 0.9) filtra todos los datos que están en el cuantil 10% o inferior y los que están en el cuantil 90% o superior.
4. Se han eliminado los valores extremos de las variables numéricas indicadas en la lista referenciada como *delete_outlayiers_columns*, usando la estrategia IqrStrategy.
5. Se han eliminado las siguientes variables / columnas:
   - PropertyCount: Por tener una baja correlación (cercana a cero).
   - PostCode: Por su baja correlación debería ser eliminada, pero se conserva debido a que se usa para hacer el join con el dataset de Airbnb.
   - Date: Por considerar que es una variable que no aporta información relevante al estudio, y porque su permanencia generaría un gran crecimiento de las columnas del dataset en el encoding del mismo.
   - Regionname: Debido a que se cuenta con información más detalla en las variables Suburb y CouncilArea.

- Address: Porque contiene información ya contenida en la columna Suburb.
- Method: Por juicio experto(Se supone que se cuenta con asesoramiento).
- SellerG: Por juicio experto.
- Distance: Por juicio experto.

6. Se han eliminado ejemplos (filas) cuando se descartaron los outliers en el dataset melb_df. En el dataset airbnb_df, se agruparon filas por zipcode para agregar variables (columnas) útiles para la predicción del precio de los inmuebles.
7. Se ha observado en el dataset de airbnb que, para el mismo zipcode y en algunas ocasiones, no todos los valores suburb son iguales. Por ello, se han filtrado y devuelto por cada zipcode de Airbnb_df, el suburb que más veces aparece (moda). En caso que el zipcode que se busca relacionar no esté en el data-frame de airbnb, se devuelve como suburb, *NOT_VALID*.

Características seleccionadas

Características categóricas

1. Suburb: Barrio donde se ubica el inmueble. 314 valores posibles.
2. Type: Tipo de propiedad. 3 valores posibles.
3. CouncilArea: Municipio donde se encuentra el inmueble. En esta variable, se agruparon las categorías que tenían menos de 100 filas, con el objetivo de balancear la cantidad de datos; pasando de 33 valores posibles a 21.
4. Regionname: Región donde se ubica el inmueble. 8 valores posibles.

Características numéricas

1. Rooms: Cantidad de habitaciones.9 valores posibles.
2. Price: Precio de venta del inmueble.2204 valores posibles.
3. Bedroom2: Cantidad de dormitorios.12 valores posibles.
4. Bathroom: Cantidad de baños. 9 valores posibles.
5. Car: Cantidad de cocheras del inmueble. 11 valores posibles.
6. Postcode: Código postal de la zona donde se ubica el inmueble. 198 valores posibles.
7. Landsize: Superficie del terreno de la propiedad. 1448 valores posibles.
8. Lattitude: Grados de latitud de la ubicación del inmueble. 6503 valores posibles.
9. Longtitude: Grados de longitud de la ubicación del inmueble. 7063 valores posibles.
10. BuildingArea: Superficie cubierta del inmueble. 602 valores posibles.
11. YearBuilt: Año de construcción del inmueble. 144 valores posibles.
12. Zipcode: Código postal de la zona donde se ubica el inmueble. Esta característica ha sido tomada de publicaciones de la plataforma AirBnB, previa agrupación, y agregada al dataset melb_df en los valores de intersección con la variable "Postcode". Es la característica elegida para realizar la unión de ambos dataset.
13. airbnb_price_mean: Se agrega el precio promedio diario de publicaciones de la plataforma AirBnB en el mismo código postal.
14. airbnb_record_count: Es el número de ejemplos (filas) que se repiten en cada código postal (zipcode) agrupado.
15. airbnb_weekly_price_mean: Se agrega el precio promedio semanal de publicaciones de la plataforma AirBnB en el mismo código postal.
16. airbnb_monthly_price_mean: Se agrega el precio promedio mensual de publicaciones de la plataforma AirBnB en el mismo código postal.

Transformaciones realizadas

1. **Encoding**: Todas las variables categóricas tipo float y tipo object fueron transformadas a columnas binarias utilizando OneHotEncoder.
   La matriz esparsa resultante se ha convertido a matriz densa, previo cálculo del size de la misma.

2. **Imputación por KNN**: Se obtuvieron todas las columnas que son numéricas, sacando del dataset la columna "Price".
   Se escalaron todos los valores entre 0 y 1 mediante MinMaxScaler, imputando los valores faltantes de las columnas `Year_built` y "Building_area" utilizando KNN. Luego, las columnas fueron transformadas de manera inversa para llevarlas a su escala original.
   Se tomaron todas las características de tipo numérico para hacer la estimación de "Year_built" y "Bulding_area".

Datos aumentados

Se han estandarizado los valores de las variables del dataset mediante StandarScaler, por motivo de que PCA es sensible a la varianza de las mismas, es decir, que las variables con mayor varianza dominarían a las de menor varianza.
Finalmente, se agregaron las 5 primeras columnas obtenidas a través del método de Análisis de Componentes Principales, aplicado sobre el conjunto de datos totalmente procesado.

Composición del resultado

Se creó un string compuesto de una función y métodos para facilitar el acceso al análisis de los datos mencionados anteriormente.