Trabajo Práctico 2

Exploración y Curación de Datos

Diplomatura en Ciencia de Datos-FAMAF

Integrantes:
-Adrián Zelaya
-Joaquín Velasco
-Javier Kondratiuk
-Mariana Pereyra

-2022-

Informe detallado de las actividades realizadas para la curación de las variables.

##Columnas seleccionadas para la predicción del valor de las propiedades y tratamiento en cada una de ellas

Categóricas:

1. **Suburb:** suburbio. 314 valores posibles
2. **Address:** dirección. 13378 valores posibles
3. **Type:** tipo de propiedad. 3 valores posibles
4. **Method:** método de venta. 5 valores posibles
5. **Regionname:** región general (este, oeste, noreste, etc.)
6. **SellerG:** vendedor
7. **CouncilArea:** concejo gobernante del área

Todas las características categóricas fueron codificadas con un método OneHotEncoding. Excluimos la característica Address del análisis dado que casi el 99% de sus valores son únicos.

Numéricas:

Price : Indica el precio de las propiedades; removimos los outliers del primer y último cuartil.

Landsize: Área del terreno; removimos los outliers del primer y último cuartil.

Lattitude: Latitud

Longtitude: Longitud

YearBuilt: Año de construcción

BuildingArea: Área de la construcción

Rooms: Número de habitaciones/ambientes

Propertycount: Número de propiedades que existen en el suburbio.

Transformaciones:

1. Todas las características numéricas fueron estandarizadas.

2. Se removieron los outliers de las columnas `YearBuilt` y `BuildingArea`. Todo valor que sea mayor a 1.5 veces el rango intercuartil fue considerado outlier.
3. Las columnas `YearBuilt` y `BuildingArea` fueron imputadas utilizando el algoritmo K-Nearest-Neighbor a partir de todas las características numéricas con K=5 (hiper parámetro del algoritmo)

Datos aumentados

Realizamos un PCA solo con las variables numéricas del Dataset y con las variables previamente estandarizadas.

Se agregaron las 2 primeras columnas (las que explicaban la mayor variabilidad de los datos) obtenidas a través del método de PCA, aplicado sobre el conjunto de datos totalmente procesado.