**Checkpoint 2 - Grupo 10**

**Análisis Exploratorio**

El conjunto de datos consta inicialmente de 20 columnas y 460154 entradas. Tras filtrar los datos que nos interesan para este TP, nos quedan 94249 entradas de interés.

Los datos que

Hay varios tipos de datos que nos interesan:

- tenemos datos cuantitativos, como las coordenadas o las propiedades, que permiten especificar cada dato y disponer de datos variados y de un conjunto de datos más completo,

- tenemos también datos cualitativos que permiten categorizar y especificar cada propiedad con diversas etiquetas.

**Preprocesamiento de Datos**

* Eliminación de columnas

Decidimos eliminar las columnas "place\_l2", "operation" y "property\_currency", que no son útiles porque eran nuestros filtros iniciales y, por lo tanto, sólo tienen un valor en nuestro conjunto de datos. También eliminamos "place\_l5" y "place\_l6" porque estaban vacías.

Por último, decidimos eliminar "id", "property\_title", "start\_date", "end\_date" y "created\_on" porque sólo identifican una única venta y no un grupo de propiedades. No obstante, no estamos cerrados a reintegrar las columnas "start\_date", "end\_date" y/o "created\_on", que podrían ayudar a identificar tendencias mensuales en las ventas de propiedades.

* Correlaciones

Hemos podido detectar las siguientes correlaciones :

- "property\_rooms" y "property\_bedrooms" : 0.86

- "property\_price" y "property\_rooms" : 0.48

- "property\_price" y "property\_bedrooms" : 0.42

- "property\_surface\_total" y "property\_surface\_covered" : 0.34

- "latitud" y "longitud" : -0.28

* Nuevos features

Aún no hemos generado ninguna feature, pero está previsto crear una nueva variable, el precio del metro cuadrado por venta, que permitiría realizar nuevos análisis, sobre todo en términos de localización.

* Valores atípicos

Por el momento, mediante un análisis unidimensional, hemos identificado algunos valores atípicos (superiores a 3 veces el rango intercuartílico) para las variables "property\_surface\_total", "property\_surface\_covered", "property\_rooms", "property\_bedrooms" y "property\_price". En vista de su escaso número, hemos decidido simplemente eliminarlos, pero un análisis multidimensional podría cambiar nuestra forma de proceder.

* Datos faltantes

Utilizamos tres técnicas diferentes para completar los datos que faltaban.

En primer lugar, para las variables "latitud", "longitud" y "place\_l3" (respectivamente 4,21%, 4,21% y 0,48% de datos faltantes), utilizamos un CSV que delimita geográficamente todos los barrios de Buenos Aires (<https://data.buenosaires.gob.ar/dataset/barrios>) para completar la variable "place\_l3" a partir de las coordenadas y las coordenadas a partir del centro del barrio según el valor de "place\_l3". Este método no es perfecto ni preciso, pero nos permite aumentar nuestros datos de manera coherente. Las entradas que se quedaron sin una de estas variables se suprimieron porque su número era reducido (en torno al 0,6%).

En segundo lugar, suprimimos la columna "place\_l4" (95,97% de datos ausentes) y sustituimos los de "place\_l3" por los de "place\_l4" si existían, ya que sólo contenían divisiones dentro de Palermo.

Por último, hemos utilizado el principio MICE para completar los datos que faltaban en las columnas "property\_rooms" (1,14%), "property\_bedrooms" (11,74%), "property\_surface\_total" (5,54%) y "property\_surface\_covered" (3,8%) a partir de las demás columnas utilizando el principio de regresión e intentando acercarnos lo más posible a la realidad.

**Visualizaciones**



Este gráfico de barras representa la diversidad de los barrios presentes en el conjunto de datos. Sin embargo, podemos ver que algunos barrios, como Palermo y Belgrano, están sobrerrepresentados con relación a los demás, lo que complicará la predicción para algunos barrios menos representados, como Liniers.



Este diagrama de dispersión pone de manifiesto la correlación positiva entre el número de habitaciones y el número de dormitorios. También podemos ver que la distribución está bastante centrada en valores inferiores a 20, lo que parece coherente dadas las propiedades de Buenos Aires.

Puede que no necesitemos transformar los datos, pero sí eliminar aquellos que parezcan incoherentes (¡como los que tienen 70 dormitorios y sólo 4 habitaciones!). Por tanto, un criterio podría ser que el número de habitaciones sea siempre mayor o igual que el número de dormitorios.

**Clustering**

Estimacion de la cantidad de clusters



Consideramos que el mejor numero de cluster es 3.

Analisis de Silhouette.



El clastering parece ser bueno porque los outliers son pequeños relativamente al resto de las predicciones. vemos que hay dos grupos en los cuales estan los outliers, uno con pocos outliers y el otro con aproximamente un cuarto de las observaciones.

**Clasificación**

Luego de crear el Target.

mostramos la distribucion del precio por metro cuadrado.



probamos las tres altenativas, dividimos la variable pxm2 en 3 intervalos con igual cantidad de observaciones, otra alternativa,dividir la variable pxm2 en 3 intervalos, el primero con el 25% de las observaciones, el siguiente con el 50% y el último con el 25% de las observaciones restantes y la otra alternativa, trabajamos la variable pxm2 relativa a cada tipo de propiedad y luego dividirla como en el punto anterior.

* **Construcción del modelo**

Arbol de Decision **SIN EMPEZAR**

* ¿Optimizaron hiperparámetros? ¿Cuáles?
* ¿Utilizaron K-fold Cross Validation?¿Cuántos folds utilizaron?
* ¿Qué métrica utilizaron para buscar los hiperparámetros?
* Añadir imagen del árbol generado e incluir descripciones que consideren adecuadas para entender el mismo. Si es muy extenso mostrar una porción representativa.

Random Forest **SIN EMPEZAR**

* ¿Optimizaron hiperparámetros? ¿Cuáles?
* ¿Utilizaron K-fold Cross Validation?¿Cuántos folds utilizaron?
* ¿Qué métrica utilizaron para buscar los hiperparámetros?
* Mostrar la conformación final de uno de los árboles generados. Si es muy extenso mostrar una porción representativa y explicar las primeras reglas.

Modelo a Elección **SIN EMPEZAR**

* ¿Optimizaron hiperparámetros? ¿Cuáles?
* ¿Utilizaron K-fold Cross Validation?¿Cuántos folds utilizaron?
* ¿Qué métrica utilizaron para buscar los hiperparámetros?
* Mostrar la conformación final de uno de los árboles generados. Si es muy extenso mostrar una porción representativa y explicar las primeras reglas.
* **Cuadro de Resultados SIN EMPEZAR**
* Realizar un cuadro de resultados comparando los modelos que entrenaron (entre ellos debe figurar cuál es el que seleccionaron como mejor predictor).

Medidas de rendimiento en el conjunto de TEST:

* F1
* Precision
* Recall
* Accuracy
* XXX: si seleccionaron alguna métrica adicional…

Confeccionar el siguiente cuadro con esta información:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Modelo | F1-Test | Precision Test | Recall Test | Accuracy Test | XXX |
| Arbol de Decision |  |  |  |  |  |
| Random Forest |  |  |  |  |  |
| … |  |  |  |  |  |

En cada caso ¿Cómo resultó la performance respecto al set de entrenamiento?

**Nota: indicar brevemente en qué consiste cada modelo de la tabla, por ejemplo**

Arbol de decision: {'max\_depth': 7, 'criterion': 'entropy', 'ccp\_alpha': 0.02}

Random Forest: {'criterion': 'entropy', 'min\_samples\_leaf': 1, 'min\_samples\_split': 16, 'n\_estimators': 50}

**Estado de Avance**

* **Análisis Exploratorio y Preprocesamiento de Datos**

**Porcentaje de Avance:** 100%/100%

**Tareas en curso:** /

**Tareas planificadas:** /

**2. Agrupamiento**

**Porcentaje de Avance:** 70%/100%

**Tareas en curso:** Graficar sobre un mapa de CABA los avisos coloreados según el grupo al que pertenecen.

**Tareas planificadas:** . Repetir el análisis anterior, utilizando sólo 3 grupos..

**3. Clasificación**

**Porcentaje de Avance:** 40%/100%

**Tareas en curso:** Mostrar en un mapa de CABA los avisos coloreados por tipo\_precio.

**Tareas planificadas:** Eleccion de altenativa del target, Entrenamiento y Predicción.

**4. Regresión**

**Porcentaje de Avance:** 0%/100%

**Tareas en curso:** sin empezar.

**Tareas planificadas:** sin empezar.

**Impedimentos:** sin empezar.

**Tiempo dedicado**.

|  |  |  |
| --- | --- | --- |
| Integrante | Tarea | Prom. Hs Semana |
| Alan Richmond | Armado de Reporte  Agrupamiento  Visualización de graficos | 5 |
| Alan Mejia | Agrupamiento  Armado de Reporte  Revisión y corrección | 4 |
| Flavian Ferré | Agrupamiento  Clasificacion  Armado de Reporte  Visualización de graficos | 10 |
|  |  |  |