Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación
Magíster en Ciencia de Datos

Estudio de Propiedades en Chile, RM

Informe Trabajo Final

Grupo 1

Integrantes:
Diego Bascuñán, Matías Espinoza, Matías Gatica, Ignacio Niño De Zepeda, Francisco Villaseca

Fecha de entrega: 01 de Octubre de 2024

Índice de contenido

Índice de Tablas

Índice de figuras

1. Introducción

El presente informe estudia las propiedades, en particular casas en venta en la Región Metropolitana de Chile. Este dataset fue rescatado de kaggle, y contiene cerca de 9.300 registros, que describen una propiedad en variables tales como; el precio de la propiedad, comuna que pertenece, dormitorios, baños, área de construcción, área total del terreno y si tiene o no estacionamiento.

El enfoque de este documento está orientado entre etapas, en primer lugar, se realiza un análisis exploratorio del dataset, para luego, entrenar una serie de algoritmos de regresión que buscan generalizar el precio de una propiedad en función de las variables del dataset. Finalmente, se busca mediante aprendizaje no supervisado estudiar relaciones relevantes a nivel grupos para las distintas propiedades de la región metropolitana.

2. Objetivos

- Realizar un análisis exploratorio de los datos de propiedades de la Región Metropolitana, con el objeto de tener una comprensión profunda de la estructura, calidad, y relación subyacente entre los datos del dataset.
- Evaluar y comprar el desempeño de diferentes modelos de regresión de modo de seleccionar el mejor modelo que generaliza el precio de las propiedades, aplicando una metodología de validación cruzada y métricas como R^2-ajustado para evaluar el desempeño de los modelos.
- Aplicar un algoritmo de clusterización KMeans para segmentar las propiedades de la Región Metropolitana, con el fin de estudiar la composición de cada clusters a nivel de comunas que lo componen.

3. Resultados

3.1. Análisis Exploratorio

Al cargar la data de propiedades se aprecia que se tienen variables de tipos numéricas y categóricas que poseen datos nulos. Antes de realizar el análisis estadístico propiamente tal se descartarán algunas variables como id y Realtor debido a que no representan características intrínsecas de una propiedad. Así mismo, se descarta utilizar la Ubicación debido a que el formato en que se encuentra no representa una ubicación precisa que permita georreferenciar la propiedad. Por último, se utiliza el precio en UF para valorizar las propiedades, ya que, en general, es la unidad monetaria en la que se transan estos bienes. Luego, a priori se tendrán las siguientes variables a estudiar Price UF, Comuna, Dorms, Baths, Built Area, Total Area y Parking.

En una primera instancia se utilizará una estrategia de eliminación de datos faltantes para luego comparar como cambian los estadísticos y distribuciones de los datos. Para esto, la variable base será Parking, ya que, es la que contiene una mayor cantidad de datos nulos. En particular, para este caso, no se puede asumir

con certeza que los datos faltantes sean propiedades sin estacionamiento, por tanto, no se realiza alguna imputación. De las tablas 1 y 2, se puede apreciar que el cambio en las distribuciones no es considerable, teniendo repercusiones marcadas en la media, desviación estándar y máximo de las variables Built Area y Total Area. Luego, se cree que es una buena opción eliminar los datos faltantes, ya que, se eliminan valores extremos que no afectan a la distribución original de las variables sino que afectan a los estadísticos que son susceptible a los valores enunciados.

Tabla 1
Estadísticas básicas para las características sin eliminación de datos faltantes

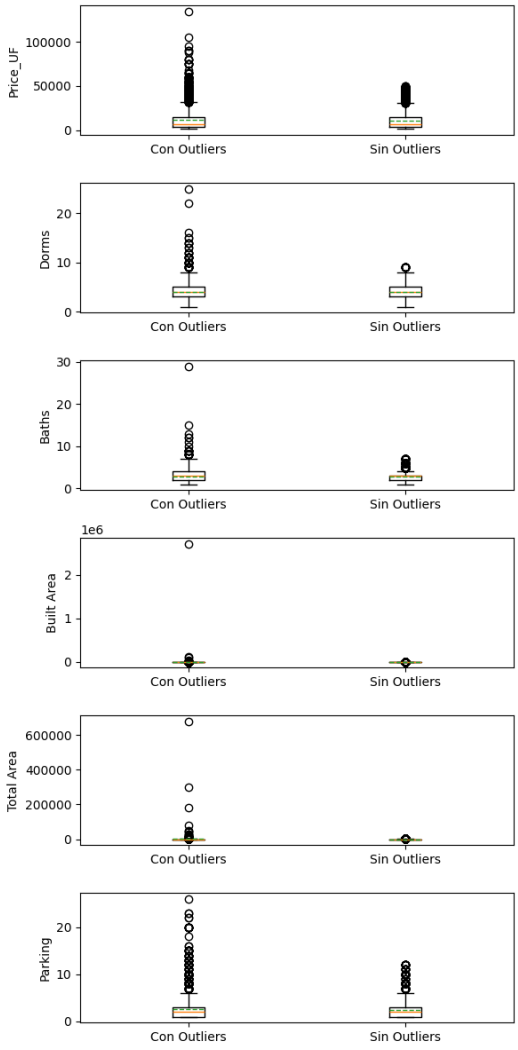
Variables	count	mean	std	min	1 %	2.50 %	10 %	25 %	50 %	75 %	90 %	97.50 %
Price UF	9291	10879.3	11188.0	1026	1469	1663	2300	3553.5	6500	14600	25000	39990
Dorms	9202	4.0	1.7	1	2	2	3	3	4	5	6	8
Baths	9138	2.7	1.5	1	1	1	1	2	3	3	5	6
Built Area	9013	6091.6	527436.7	1	43	50	66	90	131	209	348	600
Total Area	9057	891.0	11291.4	1	56	64	90	134	230	480	1093.2	5000
Parking	6371	2.7	2.2	1	1	1	1	1	2	3	5	9

Tabla 2
Estadísticas básicas para las características sin datos faltantes.

Variables	count	mean	std	min	1 %	2.50 %	10 %	25 %	50 %	75 %	90 %	97.50 %	9
Price UF	6143	11142.4	10897.0	1026	1524	1802	2384	3680	7000	15000	25000	39840.5	49
Dorms	6143	4.0	1.4	1	2	2	3	3	4	5	6	7	
Baths	6143	2.8	1.4	1	1	1	1	2	3	4	5	6	
Built Area	6143	675.2	34636.5	1	45	50	68	90	132	204	340	580	78
Total Area	6143	824.2	9927.8	1	60	68	96	137	240	480	1059.8	5000	5
Parking	6143	2.7	2.2	1	1	1	1	1	2	3	5	9	

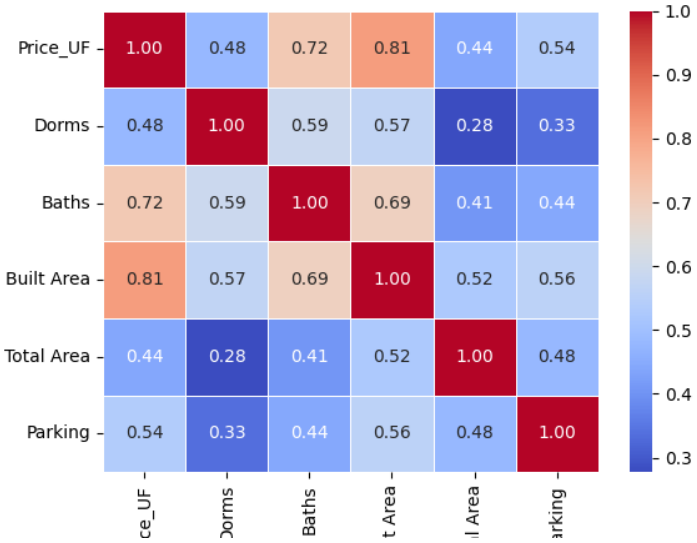
Tomando en cuenta los valores de los percentiles, se toma en cuenta que los valores que superen al percentil 99 % se tomarán como outliers y se eliminarán. En la Figura 1 se pueden observar las distribuciones con y sin outliers.

Imagen 1
Boxplots para cada variable con y sin Outliers



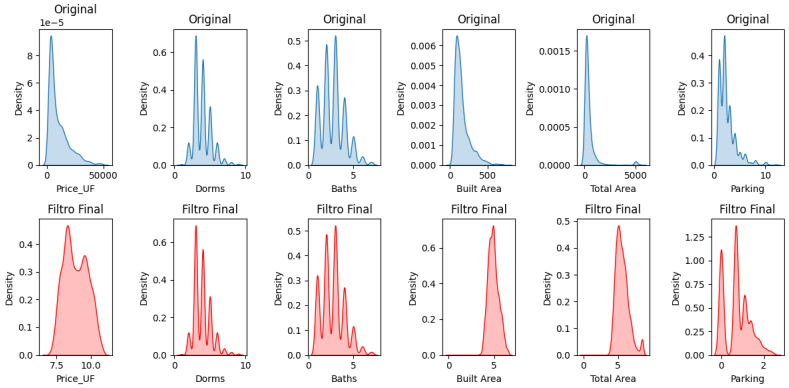
Analizando el valor de correlación para cada par de variables, se aprecia que los pares de variables que se encuentran más correlacionados son Price UF - Baths, Price UF - Built Area y Built Area - Baths, sin embargo, a pesar de que esto es un resultado esperable se cree que los valores de correlación no son lo suficientemente altos para asumir una redundancia lineal entre estas variables. Luego, no se elimina ninguna variable por correlación.

Imagen 2
Mapa de calor para correlaciones



Dadas las distribuciones obtenidas, se aprecia que las variables Price UF, Built Area, Total Area y Parking tienen funciones de densidad de kernel similares a una distribución log-normal, por esto, se aplica una transformación de este tipo a estas variables para estudiar su comportamiento. Como se aprecia en la Figura 3, si bien existen distribuciones irregulares se aprecia que en general tienen tendencia a centralizarse alrededor de un valor, por lo tanto, se utilizará un escalamiento de tipo StandardScaler.

Imagen 3
Distribuciones variables numéricas para datos sin outliers con y sin transformación logarítmica



3.2. Modelos de Regresión

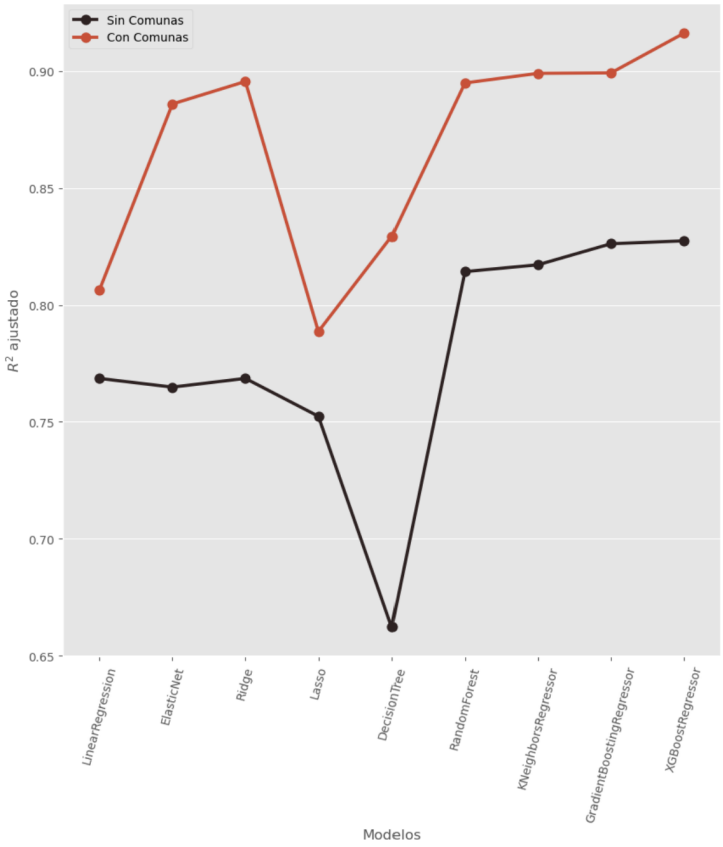
Se realiza una validación cruzada de GridSearchCV, con CV = 5 a los modelos seleccionados con sus respectivos hiperparámetros del anexo(5.1). Adicional a esto, se evalúan 2 opciones en cuanto a las varia-

bles explicativas del precio: (i) el cual no incluye como variable independiente la comuna que pertenece la propiedad, (ii) el que si la incluye. De esta manera, el modelo a estimar a nivel de variables será lo siguiente:

$$Ln(price_{uf}) = \begin{cases} Dorms + Baths + BuiltArea + TotalArea + Parking & (i) \quad sin\,comunas \\ Dorms + Baths + BuiltArea + TotalArea + Parking + Comuna & (ii) \quad con\,comunas \end{cases} \tag{1}$$

Ahora bien, en cuanto a los resultado se puede observar en el gráfico de a continuación, la inclusión de la comuna es significativa a nivel de R^2 independiente del modelo. A esto adicionalmente, se desprende que el mejor modelo para estimar el precio, es el **XGBoostRegressor** con un valor $R^2 = 0.916$. En el anexo (5.2) se puede observar las métricas de los modelos.

Imagen 4
Gráfico resultado validación cruzada modelos



Finalmente, se concluye que el modelo que mejor generaliza al precio de las propiedades contiene los siguientes hiperpámetros. Los cálculos se pueden observar en la sesión (4) del notebook adjunto.

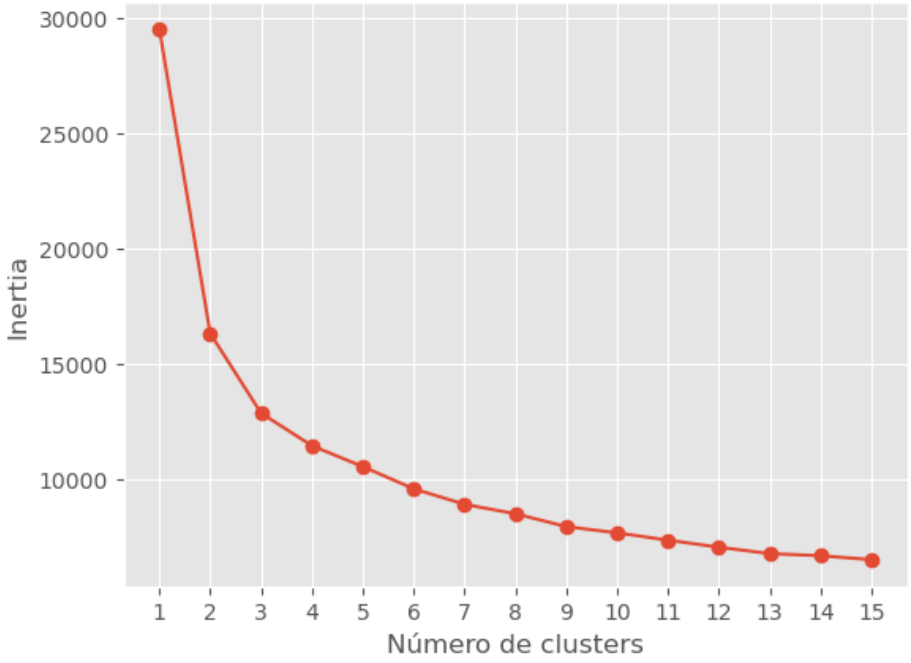
```
{'colsample_bytree': 0.8, 'learning_rate': 0.1, 'max_depth': 5, 'min_child_weight': 3, '
    n_estimators': 300, 'reg_alpha': 0, 'reg_lambda': 0.1, 'subsample': 0.7}
```

3.3. Clusterización

Con el objetivo de identificar grupos de propiedades con características similares y evaluar si estos grupos se relacionan con las comunas de Santiago, se realiza un análisis de clúster utilizando el algoritmo K-Means. Este análisis permite segmentar los datos, sin considerar la variable respuesta (comuna), para descubrir si existen patrones y/o relaciones en los parámetros de las propiedades.

Para seleccionar el número apropiado de clústeres, se utilizó Elbow Method. Se entrenaron modelos K-Means variando el número de clústeres desde 1 hasta 15, se registró la inercia para cada uno y se procedió a generar el gráfico correspondiente.

Imagen 5
Gráfico: Elbow Method - Clusterización Propiedades RM



Como se puede evidenciar en el gráfico 4, hay una fuerte caída en los primeros 3 clusters, lo que sugiere que los primeros clusters agregan mucho valor en términos de explicar la variabilidad de los datos. Luego, alrededor de los clusters 3-4, la curva comienza a aplanarse lo que indica que a partir de este punto los clusters comienzan a tener rendimientos decrecientes en términos de la reducción de inercia. Finalmente, en el cluster 6-7, la pendiente de la curva se aplana aún más indicando que elegir el número de cluster más allá del número 7 no reducirá de manera importante la variabilidad entre los grupos y aumentaremos el riesgo de sobreajuste del modelo.

3.3.1. Análisis de resultados

El gráfico 5, muestra la relación del precio de la propiedad y cantidad de dormitorios, en donde se observa que no se presenta una fuerte correlación, ya que se observa una gran dispersión entre los precios de cada

propiedad a nivel de dormitorios. No obstante, se observa que en los clústers 1 y 4, tienden a concentrarse las propiedades con los precios más altos, mientras que los clústers 2 se agrupa mayormente los precios más bajos. Lo anterior desprende, existen otros factores (ie. variables de interés) que suben el precio de propiedad, como lo podría ser la ubicación (comuna).

Imagen 6
Gráfico: Dispersión entre precio propiedad y Dormitorios

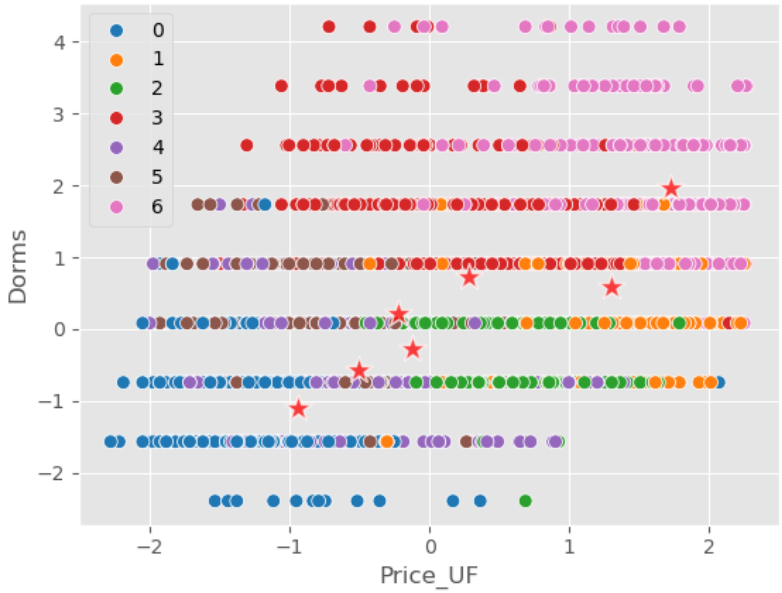
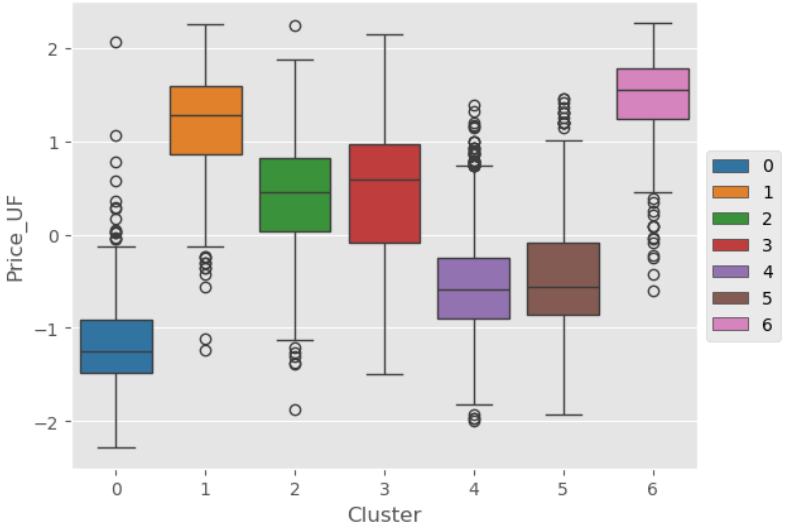


Imagen 7
Gráfico: Boxplot precios propiedades según clúster



El gráfico anterior, confirma la observación que el confirma la observación de que el clúster número 1 agrupa las propiedades de mayor precio, mientras que el clúster número 2 contiene las propiedades de

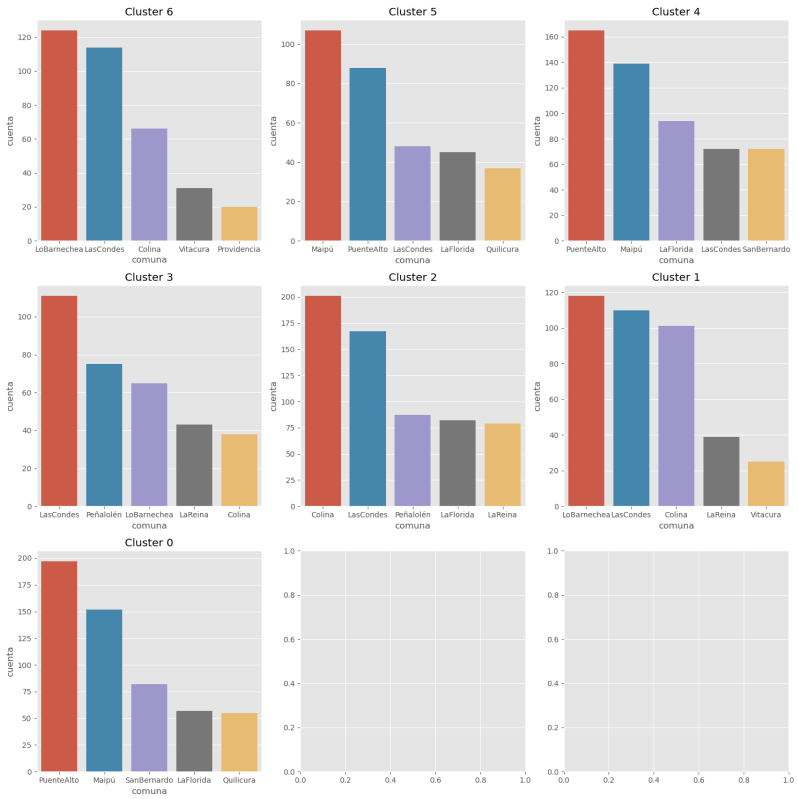
menor precio. Por otro lado, considerando que los datos están normalizados, las desviaciones estándar son relativamente altas. Esto refuerza la conclusión de que existe una dispersión significativa en los precios y una superposición considerable en términos de precios entre los diferentes clústeres.

3.3.2. Relación Clústers con las comunas

Para explorar si existe una relación entre los clústeres y las comunas de Santiago, se incorporó la variable 'Comuna' al conjunto de datos y se creó un dataframe agrupado por clúster y comuna, contando el número de propiedades. Adicionalmente, se estudiará la cantidad de veces que aparece una comuna en un clúster correspondiente.

Los gráficos 8 de barras muestran que el clúster 1, identificado previamente como el que agrupa las propiedades de mayor valor, incluye comunas del sector oriente como Lo Barnechea, Las Condes, Vitacura y La Reina, además de Colina, lo cual concuerda con la idea general de que estas comunas son las más costosas del país, confirmando que la comuna tiene una gran influencia en el precio de las propiedades. Por otro lado, se observa que las propiedades del clúster 2 corresponden a comunas percibidas como más asequibles (Puente Alto, Maipú, San Bernardo, Quilicura, La Florida), con valores más bajos. En resumen, la ubicación geográfica (comuna) pareciera ser un factor clave en la segmentación de precios en el mercado inmobiliario de Santiago.

Imagen 8
Gráfico: Top comunas a nivel clústers



4. Conclusiones

El análisis exploratorio muestra que existe una correlación fuerte el siguiente par de variables Price UF - Baths, Price UF - Built Area y Built Area – Baths. Así mismo, se concluye que estos valores son esperables, no significando que exista redundancia lineal entre la variable.

En lo que respecta a las variables, su distribución (ver gráfico 3) sigue una curva muy parecida a una log-normal, motivo por el cual se decide aplicar un escalamiento de tipo StandarScarler.

Ahora bien, en cuanto al análisis de datos se destaca que la inclusión de la variable comuna es significativo para estimar el precio de la propiedad, lo cual se observa en el gráfico 4. En cuanto al modelo que mejor estima, en ambos grupos de modelos (con comuna y sin comuna) destaca el modelo XGBoostRegressor. Este estudio también arroja que los mejor hiperpárametros seleccionados con validación cruzada.

Finalmente, en lo que respecta al modelo KMeans, se observa que no existe una significativa correlación a nivel de clúster entre las variables cantidad de habitaciones y el precio de la casa, de hecho, existe gran dispersión. De hecho, al estudiar la composición de los clústers que contienen las propiedades más caras, están asignadas fundamentalmente a comunas consolidadas, entre las que destacan Vitacura, Las Condes, Providencia, Lo Barnechea. De esta manera, el modelo no supervisado refuerza la inclusión de la comuna como variable clave para estimar el precio de una propiedad.

5. Anexos

5.1. Modelos de regresión evaluados

```
models = {
'LinearRegression': {
    'model': LinearRegression(),
    'params': {
        'fit_intercept': [True, False]
                        }
},
'ElasticNet': {
    'model': ElasticNet(),
    'params': {
        'alpha': [0.1, 1, 10],
        'l1_ratio': [0.1, 0.5, 0.9],
        'fit_intercept': [True, False]
    }
},
'Ridge': {
    'model': Ridge(),
    'params': {
        'alpha': [0.1, 1, 10],
        'fit_intercept': [True, False]
```

```
        }
    },
    'Lasso': {
        'model': Lasso(),
        'params': {
            'alpha': [0.1, 1, 10],
            'fit_intercept': [True, False]
        }
    },
    'DecisionTree': {
        'model': DecisionTreeRegressor(),
        'params': {
            'criterion': ['mse', 'friedman_mse', 'mae'],
            'splitter': ['best', 'random']
        }
    },
    'RandomForest': {
        'model': RandomForestRegressor(),
        'params': {
            'n_estimators': [10, 50, 100],
            'criterion': ['squared_error', 'absolute_error','friedman_mse']
        }
    },

    'KNeighborsRegressor':{
        'model': KNeighborsRegressor(),
        'params': {
            'n_neighbors':[3,5,6,10],
            'weights':['uniform','distance'],
            'metric':['cosine','euclidean','manhattan']
        }

    },
    'GradientBoostingRegressor':{
        'model': GradientBoostingRegressor(),
        'params': {
            'n_estimators': [10, 50, 100],
            'criterion': ['squared_error', 'friedman_mse']
        }

    },
    'XGBoostRegressor':{
        'model': xgb.XGBRegressor(objective='reg:squarederror'),
        'params': {'n_estimators': [100, 300],
                   'learning_rate': [0.01, 0.1],
                   'max_depth': [3, 5, 7],
                   'subsample': [0.7, 0.8],
                   'colsample_bytree': [0.8],
                   'min_child_weight': [1, 3],
                   'reg_alpha': [0, 0.1],
```

```
                    'reg_lambda': [0.1, 0.5]
        }
    }

}
```

5.2. Resultados Validación Cruzada: Modelos de Regresión

Tabla 3
Resultados indicadores validación cruzada - modelo sin comunas

Modelos	R^2	MAE	MSE	R^2-Ajustado
LinearRegression	0,769	0,367	0,232	0,769
ElasticNet	0,765	0,375	0,235	0,765
Ridge	0,769	0,368	0,232	0,769
Lasso	0,753	0,394	0,248	0,752
DecisionTree	0,663	0,415	0,338	0,662
RandomForest	0,814	0,316	0,186	0,814
KNeighborsRegressor	0,817	0,314	0,183	0,817
GradientBoostingRegressor	0,826	0,309	0,174	0,826
XGBoostRegressor	0,827	0,309	0,173	0,827

Tabla 4
Resultados indicadores validación cruzada - Modelo con comunas

Modelos	R^2	MAE	MSE	R^2-Ajustado
LinearRegression	0,808	0,244	0,189	0,806
ElasticNet	0,887	0,255	0,113	0,886
Ridge	0,896	0,240	0,104	0,895
Lasso	0,791	0,362	0,210	0,789
DecisionTree	0,831	0,289	0,169	0,829
RandomForest	0,896	0,226	0,104	0,895
KNeighborsRegressor	0,900	0,224	0,100	0,899
GradientBoostingRegressor	0,900	0,233	0,100	0,899
XGBoostRegressor	0,917	0,210	0,083	0,916