PROBLEMA DE NEGOCIO

Una agencia inmobiliaria en California necesita mejorar su estrategia de precios para optimizar la compra y venta de propiedades. Actualmente, no se están considerando adecuadamente las variables espaciales, como la distancia a la costa y la proximidad a ciudades importantes, lo que afecta la rentabilidad. El objetivo del proyecto es desarrollar un modelo predictivo que, usando datos de geolocalización (latitud, longitud, distancia a la costa) y características del inmueble, permita estimar con precisión el precio de las viviendas y mejorar la toma de decisiones en su estrategia de precios.

Proceso de la información:

1. Analizar y comprender el conjunto de datos proporcionado, identificar variables clave y realizar visualizaciones para entender las relaciones entre las variables y seleccionar las características relevantes. Identificar tambien si la posición geográfica es determinante en el precio.

2. Realizar limpieza de datos, limpiar valores atípicos y normalización/escalado de datos.

3. Realizar pruebas con el modelo machine learning XGBoost para predecir el valor medio de las viviendas en el estado de California.

4. Determinar el ajuste optimo del modelo por medio de GridSearchCV para tener predicciones que puedan ser empleadas como herramientas de valorización de los precios.

1. Configuración del Ambiente

```
import pandas as pd
import numpy as np#para funciones matematicas
import matplotlib.pyplot as plt
import seaborn as sns
import folium
```

```
import requests
import branca.colormap as cm
import xgboost as xgb#algoritmo de aprendizaje automatico
para regresion y clasificación
import joblib#exporta el modelo
import geopandas as gpd

from shapely.geometry import Point
from matplotlib.colors import Normalize, to_hex
from folium import Element
from matplotlib import cm
from matplotlib.colors import Normalize, to_hex
from IPython.display import display,HTML
from folium.plugins import HeatMap
from sklearn.model_selection import
train_test_split#metodo que nos ayuda a separar una base
from sklearn.metrics import mean_squared_error#metodo
para calcular RMSE
from sklearn.model_selection import GridSearchCV#para
hiperparametros del modelo
from math import radians, sin, cos, sqrt, atan2
```

2. Exploración de Datos

```
df = pd.read_csv('viviendas_en_california/California_Houses.csv')
df
```

	Median_House_Value	Median_Income	Median_Age	Tot_Rooms	Tot_Bedrooms	Population	Households	Latitude	Longitude	Distance_to_coast	Distance_to_LA	Distance_to_SanDiego	Distance_to_SanJose	Distance_to_SanFrancisco
0	452600.0	8.3252	41	880	129	322	126	37.88	-122.23	9263.040773	556529.158342	735501.806984	67432.517001	21250.213767
1	358500.0	8.3014	21	7099	1106	2401	1138	37.86	-122.22	10225.733072	554279.850069	733236.884360	65049.908574	20880.600400
2	352100.0	7.2574	52	1467	190	496	177	37.85	-122.24	8259.085109	554610.717069	733525.682937	64867.289833	18811.487450
3	341300.0	5.6431	52	1274	235	558	219	37.85	-122.25	7768.086571	555194.266086	734095.290744	65287.138412	18031.047568
4	342200.0	3.8462	52	1627	280	565	259	37.85	-122.25	7768.086571	555194.266086	734095.290744	65287.138412	18031.047568
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
20635	78100.0	1.5603	25	1665	374	845	330	39.48	-121.09	162031.481121	654530.186299	830631.543047	248510.058162	222619.890417
20636	77100.0	2.5568	18	697	150	356	114	39.49	-121.21	160445.433537	659747.068444	836245.915229	246849.888948	218314.424634
20637	92300.0	1.7000	17	2254	485	1007	433	39.43	-121.22	153754.341182	654042.214020	830699.573163	240172.220489	212097.936232
20638	84700.0	1.8672	18	1860	409	741	349	39.43	-121.32	152005.022239	657698.007703	834672.461887	238193.865909	207923.199166
20639	89400.0	2.3886	16	2785	616	1387	530	39.37	-121.24	146866.196892	648723.337126	825569.179028	233282.769063	205473.376575

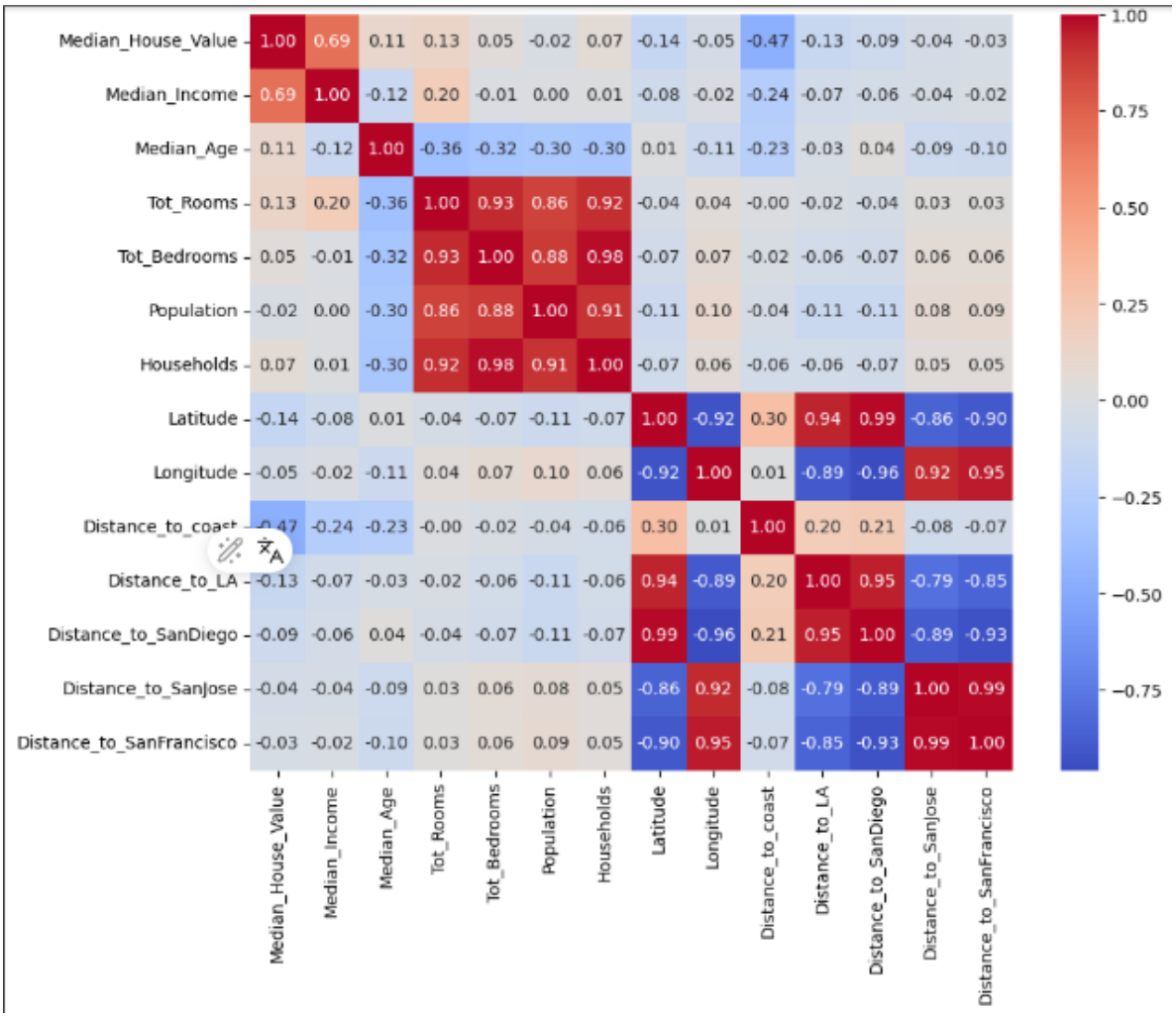
20640 rows × 14 columns

df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 14 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Median_House_Value       20640 non-null  float64
 1   Median_Income            20640 non-null  float64
 2   Median_Age               20640 non-null  int64
 3   Tot_Rooms                20640 non-null  int64
 4   Tot_Bedrooms             20640 non-null  int64
 5   Population               20640 non-null  int64
 6   Households               20640 non-null  int64
 7   Latitude                 20640 non-null  float64
 8   Longitude                20640 non-null  float64
 9   Distance_to_coast        20640 non-null  float64
 10  Distance_to_LA           20640 non-null  float64
 11  Distance_to_SanDiego     20640 non-null  float64
 12  Distance_to_SanJose      20640 non-null  float64
 13  Distance_to_SanFrancisco 20640 non-null  float64
dtypes: float64(9), int64(5)
memory usage: 2.2 MB
```

```
# Calcula la matriz de correlaciones
correlation_matrix = df_original.corr()

# Muestra el mapa de calor
plt.figure(figsize=(10,8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title("Matriz de correlación entre variables")
plt.show()
```



Variable objetivo(target)

Median_House_Value: Es la variable que queremos predecir.

1. Median_Income(Ingreso medio por hogar en la zona): Su correlación positiva es de 0.69 las zonas con más ingresos suelen tenen viviendas más caras.

2. Median_Age(Edad promedio habitantes en la zona): su correlación es baja 0.11, su relación con las otras variables es muy intensa, nos puede ayudar con zonas nuevas o que estan deterioradas.

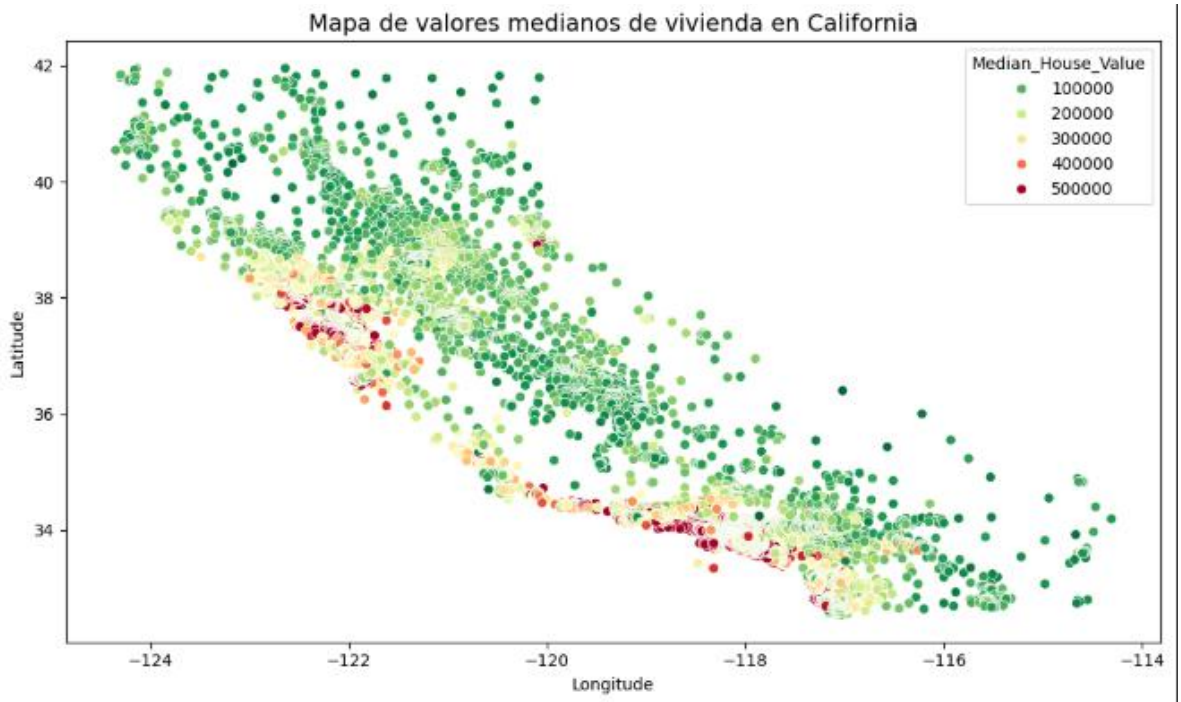
3. Tot_Rooms(Total de habitaciones en las viviendas en la zona): Nos indica el tamaño de la vivienda.

4. Latitude: Su correlación es moderada -0.14 nos permite analizar patrones espaciales importantes como que el sur tiende a tener viviendas más costosas que el norte.

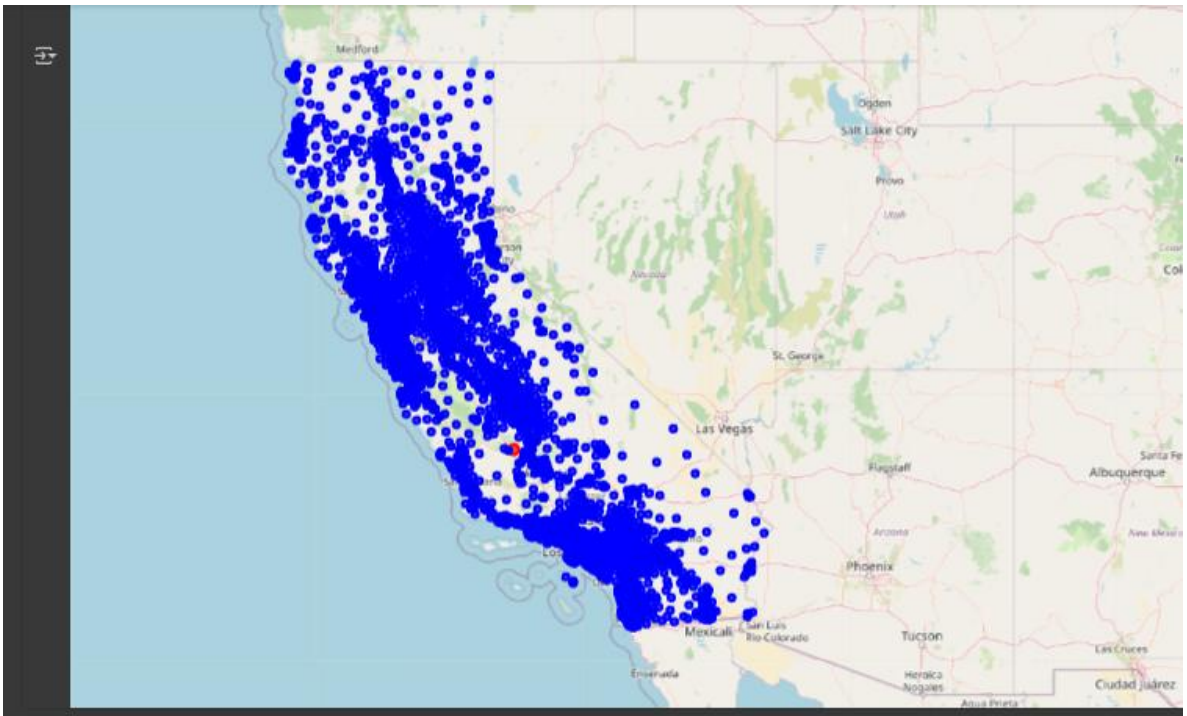
5. Longitude: Su correlación es baja -0.046 nos puede ayudar junto a latitude si hay diferencias de valor en la costa y el interior del estado.

6. Distance_to_coast: Tiene una correlación negativa -0.47 esto quiere decir que mientras más cerca del mar, más cara es la vivienda.

7. Distance_to_LA(distancia de la zona a los Angeeles): De las ciudades Los Angeles es la que mayor tiene correlacción -0.13 y afecta directamenete al valor inmobiliario.



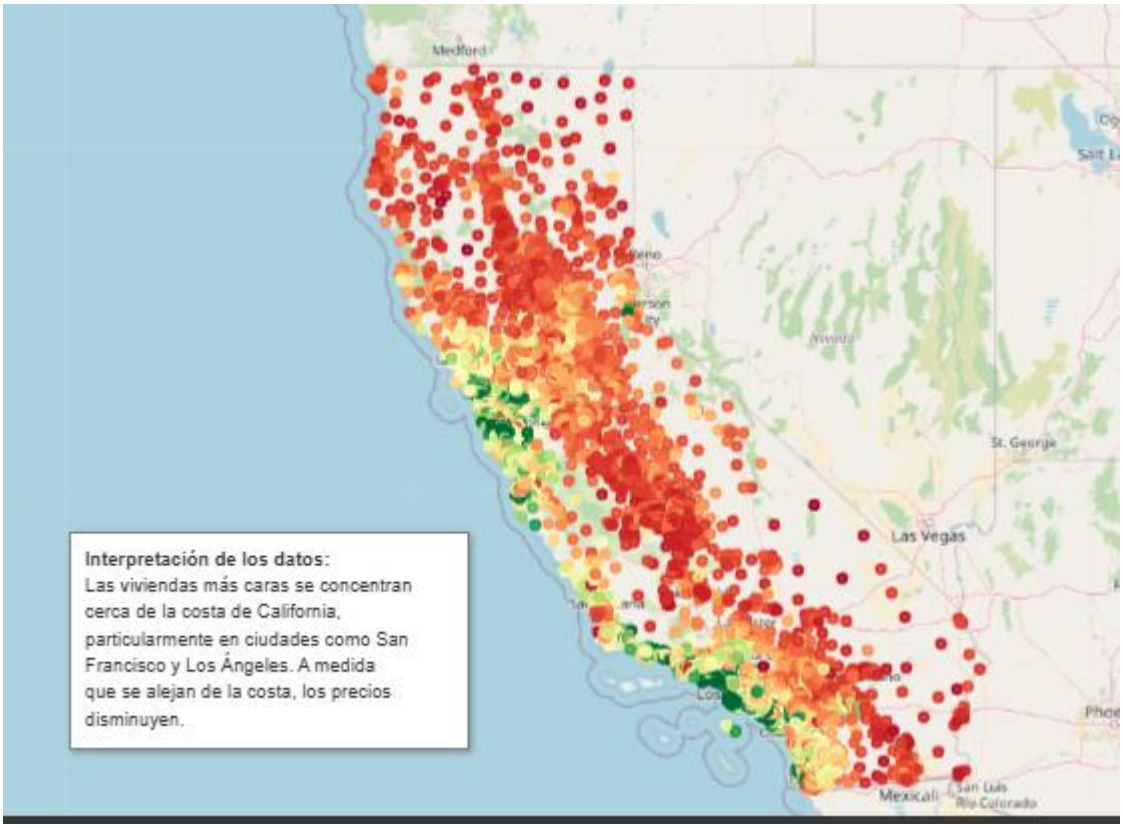
La suma de valores medianos del norte: $1,552,110,040$, *Sur* : 2,717,394,021
La viviendas del sur tienen un mayor precio medio que las del norte.



Punto Rojo: Centro geografico calculado apartir del promedio de latitudes y longitudes de df_original.

Criculos azules: Cada punto corresponde a la posición de cada vivienda.

Al hacer clic en cada punto, se despliega una ventana emergente que muestra el valor medio de cada vivienda.



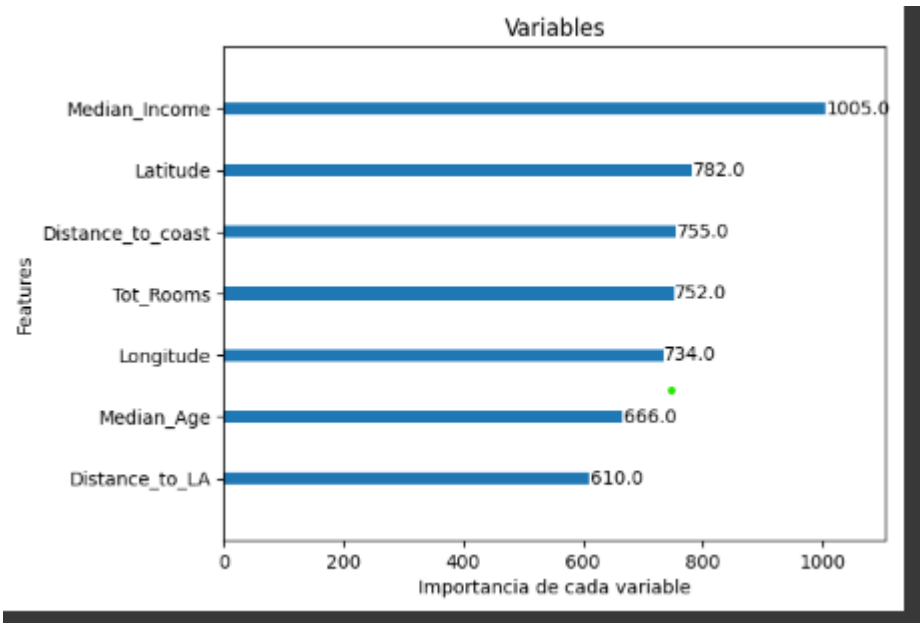
3. Entrenando el Modelo XGBoost

	Median_Income	Median_Age	Tot_Rooms	Latitude	Longitude	Distance_to_coast	Distance_to_LA
1	8.3014	21	7099	37.86	-122.22	10225.733072	554279.850069
2	7.2574	52	1467	37.85	-122.24	8259.085109	554610.717069
3	5.6431	52	1274	37.85	-122.25	7768.086571	555194.266086
4	3.8462	52	1627	37.85	-122.25	7768.086571	555194.266086
5	4.0368	52	919	37.85	-122.25	7768.086571	555194.266086
...	...	...	...	...	...	...	...
20635	1.5603	25	1665	39.48	-121.09	162031.481121	654530.186299
20636	2.5568	18	697	39.49	-121.21	160445.433537	659747.068444
20637	1.7000	17	2254	39.43	-121.22	153754.341182	654042.214020
20638	1.8672	18	1860	39.43	-121.32	152005.022239	657698.007703
20639	2.3886	16	2785	39.37	-121.24	146866.196892	648723.337126

18986 rows × 7 columns

	Median_House_Value
1	358500.0
2	352100.0
3	341300.0
4	342200.0
5	269700.0
...	...
20635	78100.0
20636	77100.0
20637	92300.0
20638	84700.0
20639	89400.0

18986 rows × 1 columns

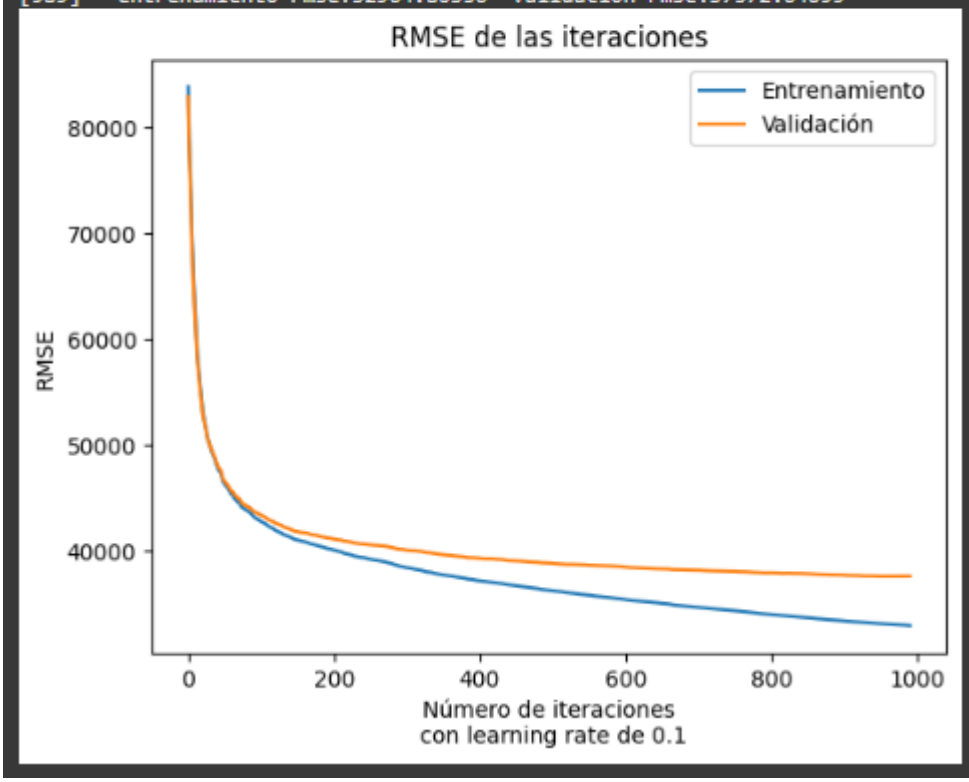


4. Hiperpametros para mejorar el modelo

```
Fitting 5 folds for each of 27 candidates, totalling 135 fits
El mejor RMSE de todas las pruebas:  53163.228279748975
Mejores parametros:  {'colsample_bytree': 0.6, 'max_depth': 3, 'subsample': 0.8}
```

5. Tasa de aprendizaje del modelo XGBoost

```
[100]   entrenamiento-rmse:42738.59817  validación-rmse:43340.43150
[200]   entrenamiento-rmse:40036.81906  validación-rmse:41085.56612
[300]   entrenamiento-rmse:38378.86972  validación-rmse:40034.16051
[400]   entrenamiento-rmse:37137.88485  validación-rmse:39260.92954
[500]   entrenamiento-rmse:36189.53499  validación-rmse:38778.83430
[600]   entrenamiento-rmse:35348.57449  validación-rmse:38418.21900
[700]   entrenamiento-rmse:34619.59597  validación-rmse:38143.95604
[800]   entrenamiento-rmse:33953.11162  validación-rmse:37892.05892
[900]   entrenamiento-rmse:33324.37697  validación-rmse:37663.69838
[989]   entrenamiento-rmse:32904.86350  validación-rmse:37572.04099
```



6. Evaluando el desempeño del modelo

	train-rmse-mean	train-rmse-std	test-rmse-mean	test-rmse-std
0	83857.279905	173.762424	83880.284266	815.571942
1	79613.269918	154.962366	79660.253443	751.315203
2	77595.187044	154.237054	77672.087635	861.798829
3	75511.223161	153.112999	75606.537573	793.599376
4	72600.039792	149.313486	72727.824663	743.380951
...	...	...	...	...
1195	31030.852364	111.439813	38632.471234	525.601697
1196	31025.867184	110.951465	38629.657028	526.527760
1197	31018.919906	110.479176	38627.981715	527.346114
1198	31013.505348	109.767630	38627.386899	527.597618
1199	31008.937408	108.372153	38626.375244	526.415820

1200 rows × 4 columns

7. Valores Pronosticados por el Modelo.

	Latitude	Longitude	Median_Income	Median_Age	Tot_Rooms	Distance_to_LA	Distance_to_coast	Valor_Pronosticado
0	39.762771	-123.744643	4.20	29.1	4	169.44	96.68	293511.03125
1	38.474072	-120.018943	10.38	57.2	3	680.89	109.25	190590.62500
2	36.211936	-119.576525	6.27	48.7	4	341.92	71.71	305110.34375
3	38.601087	-120.547908	2.73	50.5	12	769.58	75.78	172211.93750
4	34.152634	-115.423463	13.61	28.1	4	578.32	127.68	222400.43750

8. Posición de la Viviendas en California.