# PREDICCIÓN DE PRECIOS DE VIVIENDAS

**PROYECTO DE MACHINE LEARNING**

Sebastián Pomi Rodríguez

Octubre 2023

# INDICES

# INTRODUCCIÓN

En el actual mercado inmobiliario, el análisis de datos se ha convertido en una herramienta indispensable para la predicción de precios de viviendas. Los especialistas en datos inmobiliarios son profesionales altamente buscados, quienes, mediante su experticia técnica y conocimientos específicos, extraen información crucial de conjuntos de datos relacionados con propiedades. Comprender cómo las distintas variables, como la ubicación, tamaño o año de construcción, influyen en el precio de las viviendas en Madrid puede ser esencial tanto para las agencias inmobiliarias como para los potenciales compradores.

En este Análisis Exploratorio de Datos (EDA), nos centraremos en el mercado de viviendas de Madrid, explorando cómo los precios de las propiedades varían en función de diferentes factores. Examinaremos las fluctuaciones de precios entre viviendas de nueva construcción frente a las antiguas, la influencia de las distintas zonas de la ciudad, y cómo características específicas, como la presencia de terrazas o piscinas, pueden afectar el valor de una propiedad.

A través de este EDA, aspiramos a identificar tendencias, correlaciones y patrones que ofrezcan una visión clara sobre la dinámica de precios en el sector inmobiliario madrileño. Estos descubrimientos pueden ser de gran utilidad para las agencias inmobiliarias, que pueden afinar sus estrategias de valoración y marketing, y para los compradores, que obtendrán una visión más informada sobre el valor real de una propiedad.

Mediante visualizaciones detalladas, análisis estadísticos y una profunda exploración de datos, buscamos arrojar luz sobre las variaciones de precios en el mercado de viviendas de Madrid en función de diversos criterios. Esta investigación no sólo será valiosa para entender el panorama actual, sino que también puede sentar las bases para futuros análisis y pronósticos en el dinámico mercado inmobiliario de Madrid.

## Hipótesis:

Las comodidades en una vivienda, incluyendo características como ascensor, aire acondicionado, calefacción, estacionamiento, balcón, piscina o terraza, tienen un impacto considerable en el precio de venta.

## Otras consultas

- Que distrito tiene la media de precios más alta

Descripción de las variables.

1. Address-> Dirección de la vivienda
2. Zipcode-> Código Postal
3. Latitude -> Latitud
4. Longitude-> Longitud
5. Price -> Precio de la casa
6. Date -> Fecha de publicación del aviso de venta
7. Rooms -> Número de habitaciones
8. Bathrooms -> Número de baños
9. Surface-> Superficie de la vivienda en metros cuadrados
10. Floor-> Numero de piso
11. Elevator-> Si tiene ascensor
12. Air_Conditioner-> Si tiene aire acondicionado
13. Heater -> Si tiene calefacción
14. Parking -> Si tiene parking
15. Balcony -> Si tiene balcón
16. Terrace -> Si tiene terraza
17. Swimming_Pool -> Si tiene piscina

# DESARROLLO DEL CONTENIDO

## OBTENCIÓN DE DATOS:

Para este análisis hemos obtenido los datos a partir de un Dataset que contiene información sobre viviendas a la venta en la ciudad de Madrid publicadas desde 2017 hasta 2023. El Dataset cuenta originalmente con 17 columnas y 14130 viviendas

El Dataset lo obtuvimos de la página Kaggle, compartido por el usuario AFERNANDEZ (https://www.kaggle.com/datasets/alefernandezarmas/madrid-real-state-prices)

Información del Dataset

- Tiene 17 columnas
- Tiene 14.130 filas
- Tiene 3 tipos de datos. Objet, Int y Float

Obtenemos los primeros datos:

Como primera visualización del Dataset realizamos estudios para ver el tamaño del Dataset, los nombres de las columnas, los tipos de datos, comprobamos si hay valores nulos o vacíos. También realizamos el primer resumen estadístico de las columnas numéricas para tener una visión general y rápida de las variables.

## TRATAMIENTO DE COLUMNAS.

En este paso eliminamos las columnas que no nos sirven, las traducimos al español y tratamos sus valores nulos y ceros.

Eliminación de columnas

Decidimos eliminar la columna “Date” ya que la fecha de publicación de la vivienda no nos representaría información util

Traducción de columnas:

Renombramos las columnas con su nombre en español para unificar el idioma del estudio.

Creación de columnas:

Agregamos la columna ‘Precio_Medio_cp” que representa el valor medio de las viviendas por código postal.

Transformación de columnas:

Transformamos la columna “Codigo Postal” con la librería Geopy para que nos de los valores reales de los códigos postales en base a las columnas latitud y longitud.

## TRATAMIENTO DE DATOS DUPLICADOS

Observamos que el Dataset tiene 4.237 datos duplicados

Eliminamos valores duplicados

## LIMPIEZA DE DATOS POR VARIABLE

### Superficie_m2:

Observamos que existen registros de viviendas con menos de 20 metros cuadrados. Las eliminamos

También eliminamos las viviendas que tienen más de 55 metros cuadrados y ninguna habitación.

Observamos registros de viviendas con menos de 300 metros cuadrados y más de 8 habitaciones. Las eliminamos

### Precio:

Luego de realizar un exhaustivo análisis de la variable, comprobamos que exisitian viviendas cuyo precio era menor a 75.000 euros y con una superficie de más de 60 metros cuadrados. Las eliminamos. También las que tenían precio 0.

Eliminamos las casas que tenían valores mayores a 5.000.000 euros para que no quede desbalanceado el modelo en el futuro.

### Codigo_Postal:

Eliminamos los valores que tengan un código postal que no comience con “280” para así mantener solo las viviendas que se encuentran dentro de la ciudad de Madrid.

### Baños:

Eliminamos las viviendas que no tenían baños.

También eliminamos los estudios que tenían más de 1 baño y las casas que tenían menos de 3 baños al tener más de 6 habitaciones.

Debido a la falta de calidad de los datos en esta variable, hemos tomado la decisión de eliminar la columna en el análisis posterior.

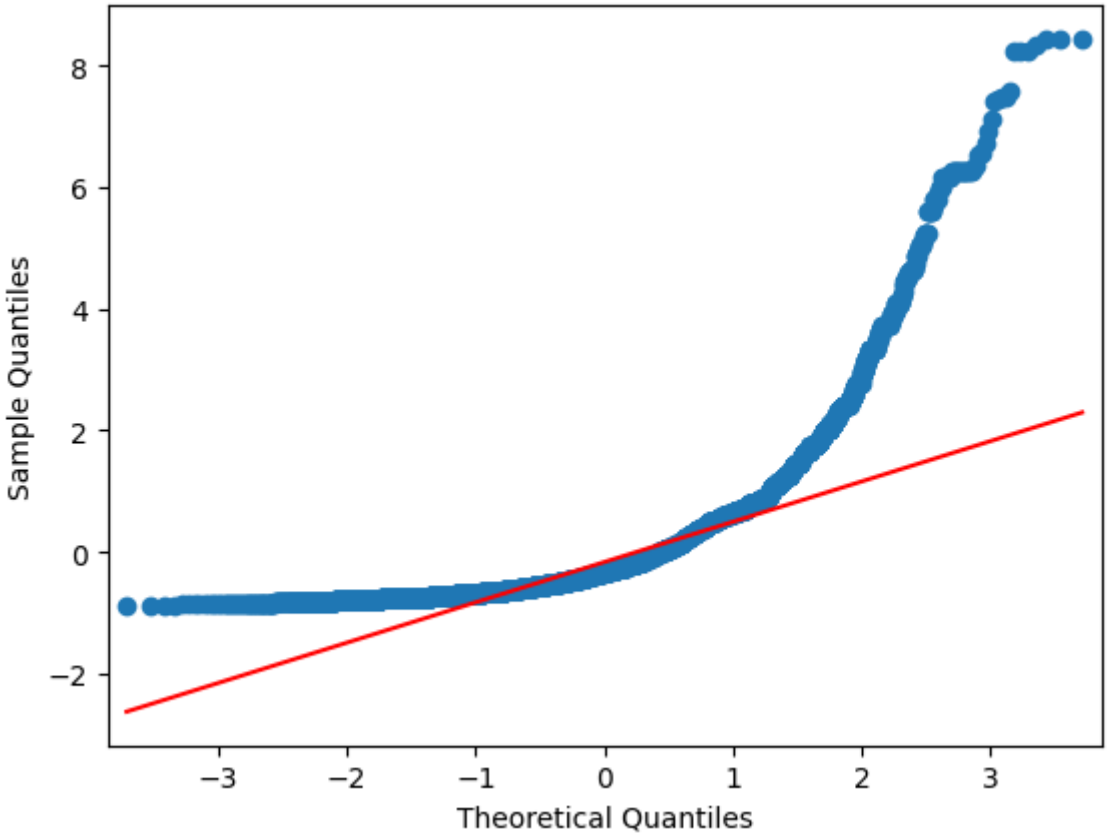
# ANÁLISIS DE DATOS

## ANÁLISIS UNIDIMENSIONAL

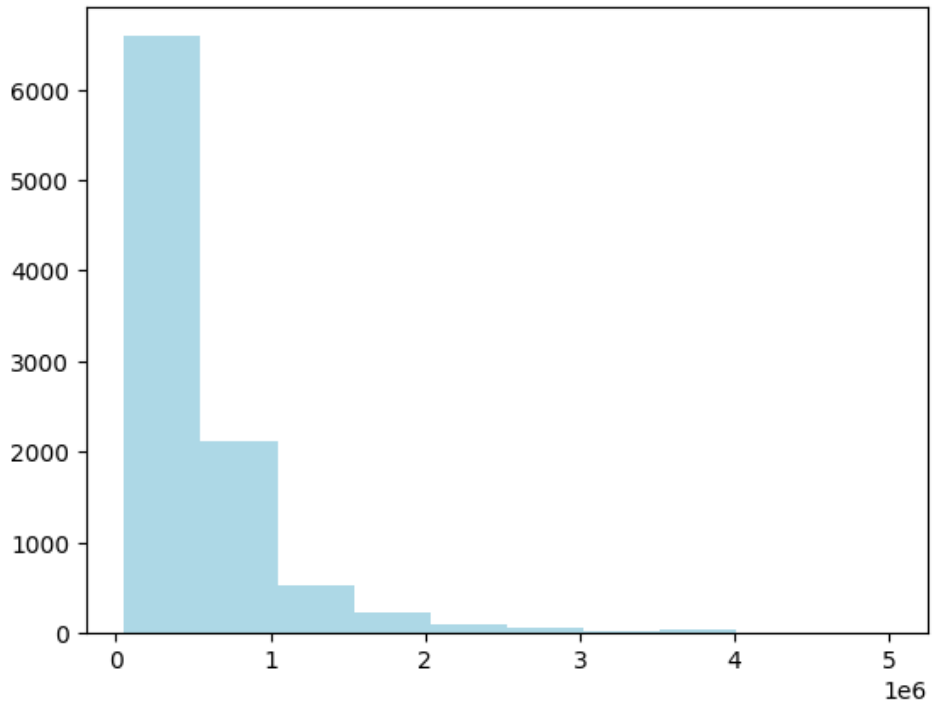
Análisis unidimensional de variables numéricas:

- Resumen estadístico para observar los principales datos estadísticos.
- Histograma para visualizar la distribución y comportamiento de las variables.
- Comprobamos que la variable dependiente “Precio” no sigue una distribución normal (fig.1)
- Gráfico para ver distribución y simetría de la variable “Precio”(fig.2)
- Cálculo del precio medio por habitación
- Gráficos de torta para ver porcentajes de viviendas con Ascensor, Aire Acondicionado, Calefacción, Balcón, Parking, Terraza y PiscinaAnálisis unidimensional de variables no numéricas:

*1 Distribución de la variable Precio con respecto a la Normal*



*2 Distribución y Simetria del Precio*

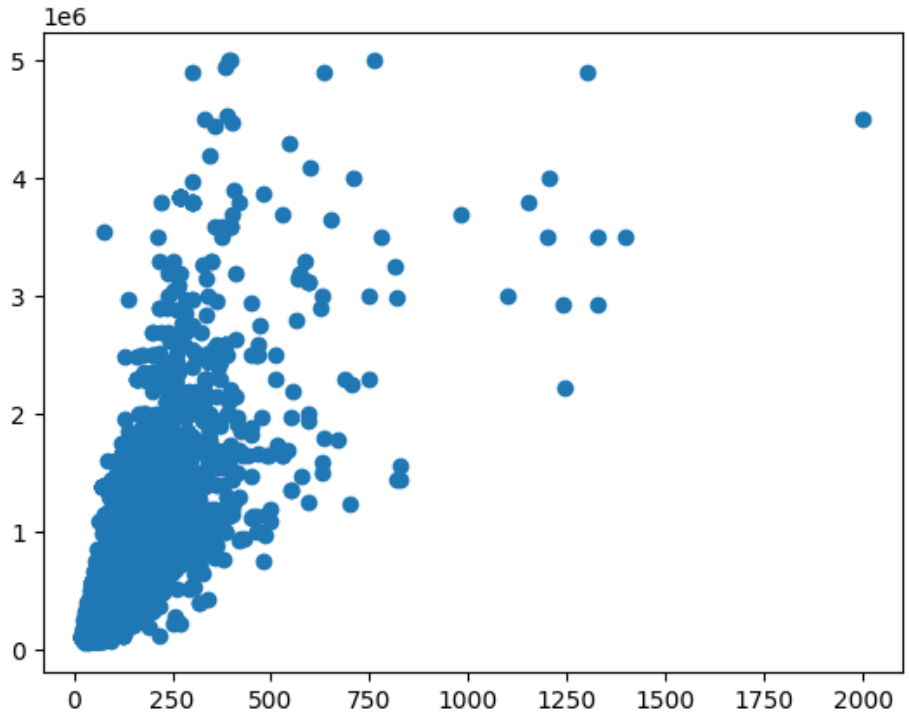



# ANÁLISIS MULTIDIMENSIONAL

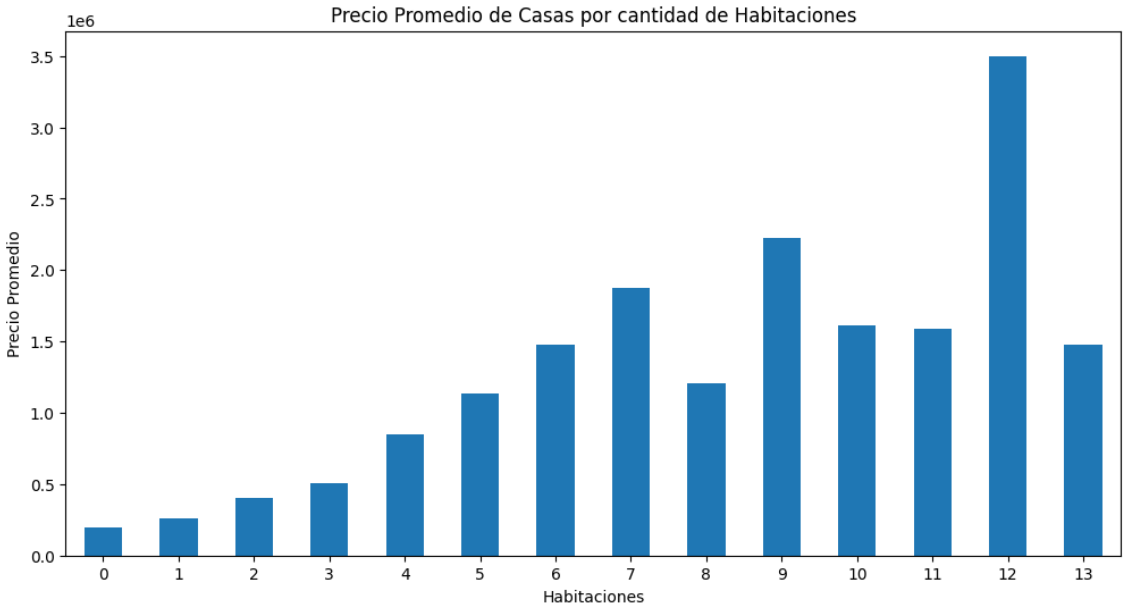
## Relación Precio - Superficie

Vemos como a medida que aumenta la superficie, aumenta el precio.

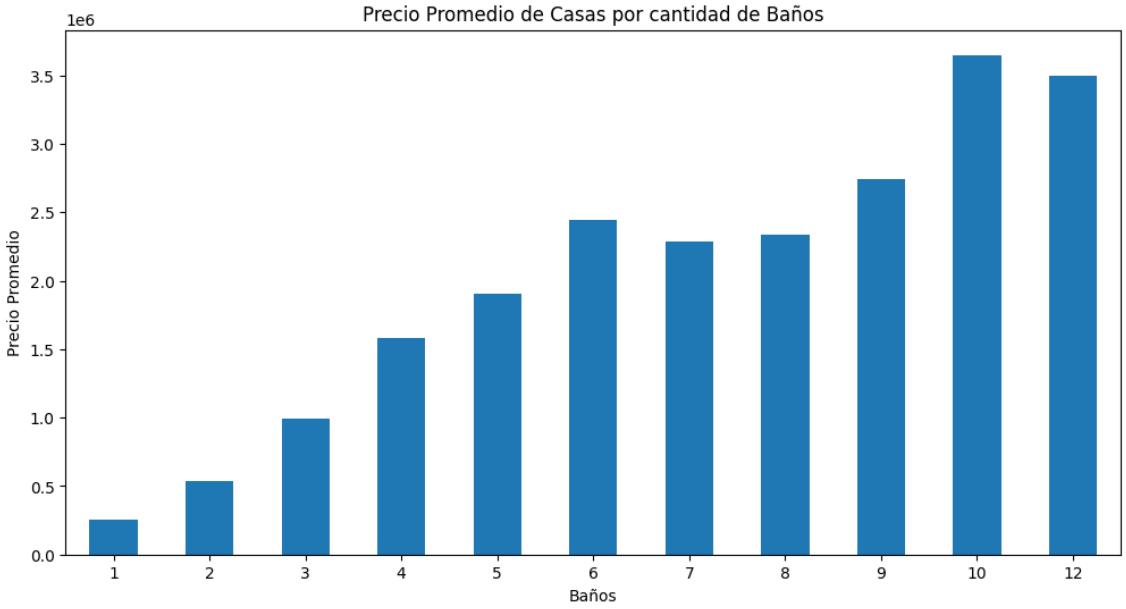
*3Relación Superficie (en metros cuadrados) y Precio (€)*



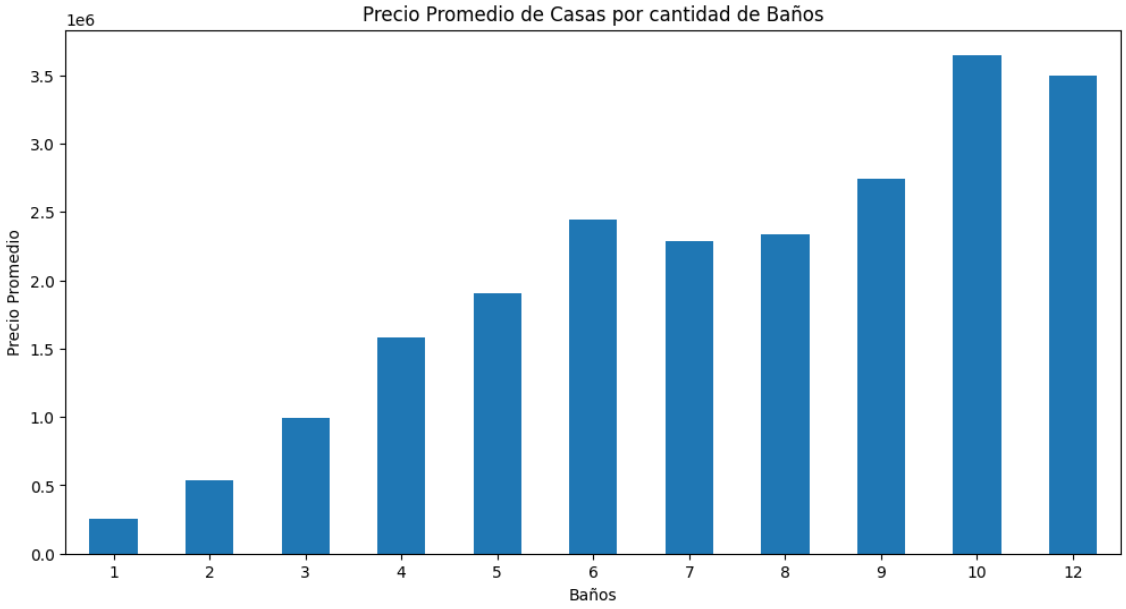
Precio Promedio por cantidad de Habitaciones




Precio Promedio por cantidad de Baños



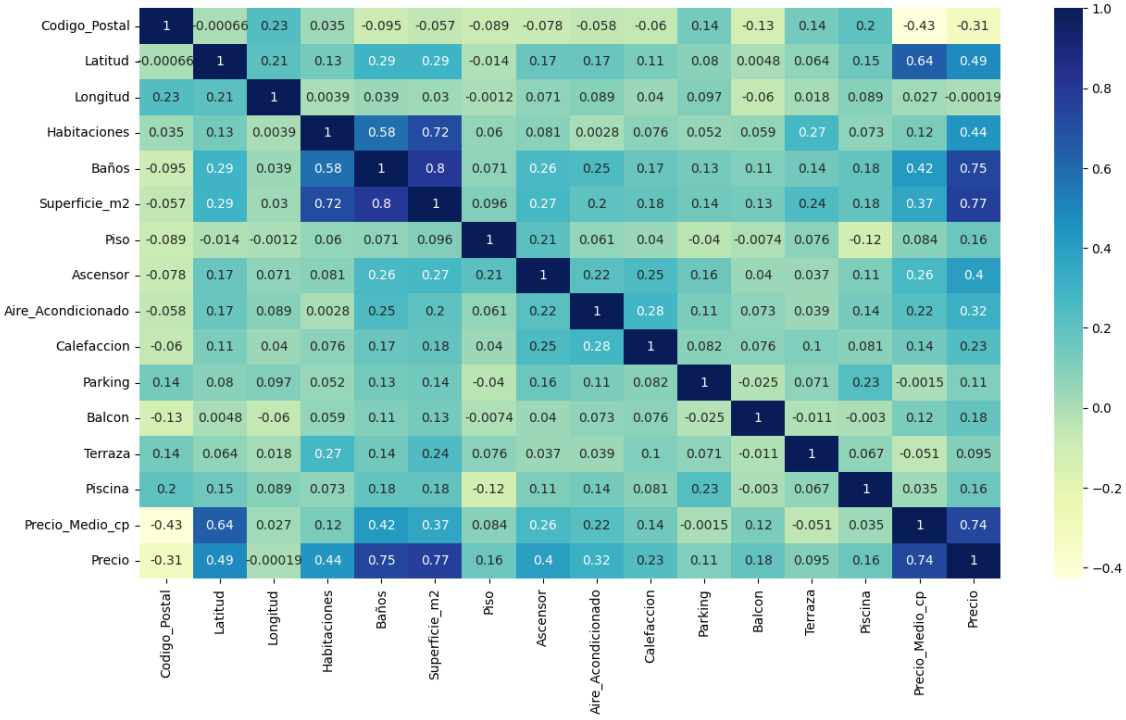
Precio promedio de venta por código postal



Evaluación de la correlación con la variable objetivo

Análisis de correlación

Analizamos y estudiamos como se correlacionan las variables independientes con nuestra variable dependiente Precio. Para ello realizamos una prueba de correlación con el método de Spearman ya que la variable dependiente no sigue una distribución normal




```
Correlación Spearman:
Codigo_Postal        -0.307985
Latitud               0.488509
Longitud             -0.000185
Habitaciones          0.441141
Baños                 0.746738
Superficie_m2         0.773335
Piso                  0.158238
Ascensor              0.399028
Aire_Acondicionado    0.316816
Calefaccion           0.231332
Parking               0.106026
Balcon                0.182708
Terraza               0.094742
Piscina               0.159311
Precio_Medio_cp       0.738680
Precio                1.000000
```

- **Superficie_m2** -> Tamaño de la casa está fuertemente relacionado con su precio

- **Baños** -> Las casas con más baños suelen ser más caras. Es una caracteristica relevante

- **Latitud** -> Si dividimos a Madrid de Norte a Sur, podemos ver como los barrios más caros se encuentran en el Norte

- **Precio_Medio_cp** -> Representa el precio medio por código postal y tiene una fuerte correlación con el precio. <br> Sin embargo, tendremos cuidado al incluir esta característica ya que puede generar un problema de <br> "data leakage", es decir, podrías estar incluyendo una característica que ya tiene la información que estás tratando de predecir. Probaremos modelos con y sin esta caracterisitca.

- **Habitaciones** -> Muestra una correlación relativamente alta

-**Ascensor, Aire Acondicionado, calefacción, Parking, Balcón, Piscina, Terraza** -> correlación entre 0.1 y 0.4. Probaremos modelos con y sin estas variables.

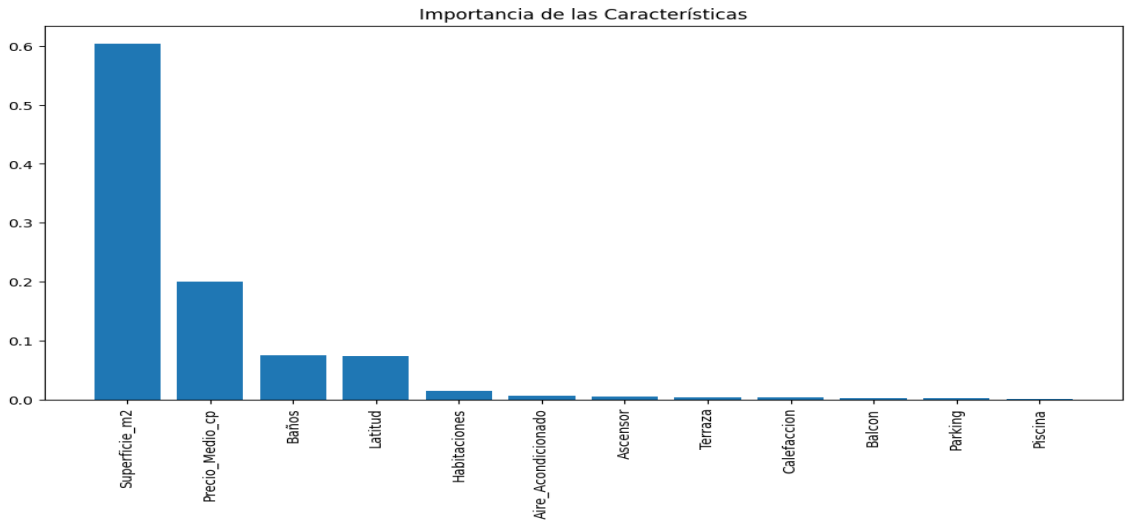
- **Longitud**-> Correlación muy baja, no la tendremos en cuenta

- **Codigo_Postal**-> Es una característica categórica, no lo tendremos en cuenta, pero si la guardaremos como diccionario para relacionar las zonas más adelante

Análisis de importancia (RandomForest)

Entrenamos un modelo de RandomForestRegressor para calcular las variables más importantes con respecto a la variable dependiente

```
Superficie_m2: 0.6019
Precio_Medio_cp: 0.2063
Latitud: 0.0744
Baños: 0.0711
Habitaciones: 0.0157
Aire_Acondicionado: 0.0076
Ascensor: 0.0060
Terraza: 0.0042
Balcon: 0.0037
Calefaccion: 0.0035
Piscina: 0.0026
Parking: 0.0023
```



# ENTRENAMIENTO Y VALIDACION DE LOS MODELOS

## Entrenamiento:

Primero separamos el dataset en X (variables independientes) e Y (variable dependiente "Precio).

Luego realizamos el train_test_split para dividir el dataset en los datos de entrenamiento y testeo de los modelos

Los valores de entrenamiento representan el 80% del dataset total (7.732 datos) mientras que los de testeo el otro 20% (1.935 datos)

## StandatdScaler:

Escalamos los datos con StandardScaler para normalizar las variables y que tengan una media de 0 y una desviación estándar de 1. Lo hacemos para convertir todas las medidas estandarizadas en el mismo rango. Para mantener los valores de entrenamiento y testeo, los guardamos en una variable diferente. Luego se lo aplicaremos a todos los modelos para compararlo con los valores originales

## Regresión Lineal:

La regresión lineal encuentra la mejor línea recta que predice una variable a partir de otra(s). Ajusta esta línea usando datos para minimizar el error entre las predicciones y los valores reales.

Lo entrenamos con valores originales y escalados

## Ridge:

La regresión Ridge es una versión de la regresión lineal que incluye un término de regularización. Su objetivo es no solo ajustarse a los datos sino también mantener los pesos del modelo (coeficientes) lo más pequeños posible. Es especialmente útil para prevenir el sobreajuste cuando hay muchas características. Lo entrenamos con valores originales y escalados Alpha = 1

## Lasso:

La regresión Lasso es una variante de la regresión lineal que también incorpora un término de regularización. A diferencia de Ridge, Lasso tiende a hacer que algunos coeficientes sean exactamente cero, lo que efectivamente selecciona características importantes y excluye las no esenciales. Es útil cuando sospechamos que muchos rasgos son irrelevantes o redundantes. Lo entrenamos con valores originales y escalados. Alpha = 1

## Random Forest Regresor:

El "Random Forest Regressor" es un algoritmo basado en un conjunto de árboles de decisión. En lugar de depender de un solo árbol, construye múltiples árboles usando subconjuntos aleatorios del dataset y características seleccionadas al azar. Cada árbol produce una predicción, y el resultado del Random Forest es el promedio de estas predicciones. Al agrupar múltiples árboles, el modelo es más robusto y menos propenso a errores de un solo árbol, reduciendo el sobreajuste y mejorando la precisión general.

Lo entrenamos con valores originales y escalados

También realizamos la técnica de Grid Search para conseguir los mejores hiperparámetros para el modelo.

## Gradient Boosting Regressor:

El "Gradient Boosting Regressor" es un algoritmo que construye una serie de árboles de decisión pequeños y débiles de manera secuencial. Cada árbol trata de corregir los errores de los árboles anteriores. En lugar de dar igual importancia a cada error, pone más énfasis en los datos que fueron predichos erróneamente por los árboles previos (siguiendo el "gradiente" del error). Al sumar las predicciones de todos los árboles, se obtiene un modelo final más preciso y robusto para regresión

## Pipeline:

Es una secuencia ordenada de pasos donde cada paso realiza una tarea específica, como limpiar datos, transformarlos o entrenar un modelo. El "Pipeline" asegura que estos pasos se realicen en el orden correcto y de manera consistente, simplificando el proceso y evitando errores comunes.

Hemos implementado un Pipeline para estructurar y automatizar el proceso de modelado. Dentro de este pipeline, hemos considerado dos tipos de modelos: LinearRegression y RandomForestRegressor.

Posteriormente, utilizamos GridSearchCV para optimizar y seleccionar el mejor conjunto de hiperparámetros de los modelos mencionados anteriormente.

Este enfoque garantiza que el modelo seleccionado esté bien ajustado y optimizado para nuestros datos, proporcionando una base sólida para predicciones futuras.

# VALIDACIÓN DE LOS MODELOS

Para evaluar los modelos tomamos en cuenta las siguientes métricas de validación:

- **$R^2$ Score (Coeficiente de Determinación):** Indica qué porcentaje de la variabilidad total de la variable dependiente (Precio) es explicada por el modelo. Los valores oscilan entre 0 y 1, siendo 1 un ajuste perfecto.
- **MAE (Mean Absolute Error):** Es el promedio de los errores absolutos entre las predicciones del modelo y los valores verdaderos. Un MAE de 0 indica predicciones perfectas.
- **MSE (Mean Squared Error):** Es el promedio de los errores al cuadrado entre las predicciones y los valores reales. Es una métrica que penaliza fuertemente los errores grandes, es decir, si hay predicciones que se desvían mucho del valor real, el MSE será significativamente mayor.
- **MAPE (Mean Absolute Percentage Error):** Es el promedio de los errores absolutos como un porcentaje de los valores reales.

Regresión Lineal

```
R2_score linear_model 0.7491
MAE linear_model 148695.3066
MSE linear_model 72982603813.42
MAPE linear_model 33.1827
```

Regresión Lineal Escalado

```
R2_score linear_model_scal 0.7491
MAE linear_model_scal 148695.3066
MSE linear_model_scal 72982603813.42
MAPE linear_model_scal 33.1827
```

Lasso

```
R2_score Lasso 0.7491570829491363
MAE Lasso 148690.47391514675
MSE Lasso 72982239730.74428
MAPE Lasso 33.1805921772023
```

Lasso Escalado

```
R2_score Lasso_scal 0.7491
MAE Lasso_scal 148694.39
MSE Lasso_scal 72982382822.09
MAPE Lasso_scal 33.1822
```

Ridge:

```
R2_score Ridge 0.7491
MAE Ridge 148299.72
MSE Ridge 72986084145.87
MAPE Ridge 33.0142
```

Ridge Escalado

```
R2_score Ridge_scal 0.7491
MAE Ridge_scal 148688.038
MSE Ridge_scal 72978899962.593
MAPE Ridge_scal 33.1789
```

Random Forest

```
R2_score RandomForest 0.8886
MAE RandomForest 82489.58
MSE RandomForest 32387450573.0439
MAPE RandomForest 15.3352
```

Random Forest Normalizado

```
R2_score RandomForest_Scal 0.8876
MAE RandomForest_Scal 83080.309
MSE RandomForest_Scal 32691094521.42
MAPE RandomForest_Scal 15.4111
```

Random Forest Grid Search

```
R2_score RandomForest_GS 0.8619
MAE RandomForest_GS 95200.38
MSE RandomForest_GS 40167535573.23
MAPE RandomForest_GS 17.6107
```

GardientBoosting

```
R2_score GB 0.8713
MAE GB 101765.36
MSE GB 37419352640.86
MAPE GB 19.89
```

Pipeline

```
R2_score pipeline 0.8736
MAE pipeline 89545.35
MSE pipeline 36768057863.41
MAPE pipeline 16.640
```

# CONCLUCIONES

## CONCLUSIONES EVALUACION

Basado en las métricas presentadas, **el modelo Random Forest** es el que proporciona los mejores resultados en términos de precisión y error. Este modelo no solo explica una mayor proporción de la variabilidad de los datos, sino que también tiene errores más bajos en comparación con los demás modelos.

Por lo tanto, **recomendamos utilizar el modelo Random Forest** para las predicciones futuras en este proyecto. Es importante destacar que, si bien este modelo proporciona resultados superiores, es esencial tener en cuenta el costo computacional y la interpretabilidad del modelo dependiendo del contexto del proyecto y las necesidades de los stakeholders.

## CONCLUCIONES GENERALES

Nuestro análisis sugiere que, al determinar el precio de una vivienda, los factores primordiales son su ubicación, superficie en metros cuadrados y número de habitaciones. Aunque las comodidades influyen, no resultan ser determinantes para el precio dentro del contexto estudiado.