PREDICCIÓN DE PRECIO DE CASAS COLOMBIANAS

CARLOS CASAS ARENAS

SEBASTIAN ARISTIZABAL CASTAÑEDA

DEPARTAMENTO DE INGENIERÍA DE SISTEMAS

INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL

UNIVERSIDAD DE ANTIOQUIA

25 DE NOVIEMBRE DE 2023

2023 - 2

UNIVERSIDAD DE ANTIOQUIA

1 8 0 3

CONTENIDO

1. Planteamiento del problema

1.1. Introducción

A lo largo del desarrollo de los proyectos hemos aprendido sobre la recopilación de información, el análisis de datos, depuración de estos, creación y entendimiento sobre modelos predictivos iniciales.

A medida que se ha avanzado en la comprensión de estos conceptos, hemos podido visionar de mejor manera lo que hicimos y podríamos mejorar. Por ejemplo:

- Elección de dataset: Incluso si un dataset aparenta tener muchos datos, esto no necesariamente significa que la información sea valiosa, nuestro dataset contenía bastantes filas (más de 140 mil) y columnas (40), pero esto no implicó que toda la información fuese útil o relevante para nuestras necesidades, de por si había mucha información faltante y por otro lado poca relación entre las variables de mayor relevancia para nuestro estudio.

- Limpieza de la información: El trato que se le da a los datos faltantes está bastante ligado al objetivo del proyecto, por lo que antes de borrar o cambiar información es recomendable hacer un estudio de lo que implica modificar ciertas variables y su impacto en los modelos.

- Creación de modelos predictivos: Al comienzo de un proyecto es normal hacer muchas pruebas para la creación de modelos que no necesariamente tendrán los mejores resultados, pero puede ir dando una idea del camino que se tomará.

1.2. Dataset

El conjunto de datos seleccionado para este proyecto se obtuvo de Kaggle. Este dataset ha sido utilizado en el contexto de predicción de precios de viviendas en Colombia y ha sido subido a Kaggle hace aproximadamente 4 años.

Número de Filas: 129,076

Número de Columnas: 45

El dataset contiene diversas características que describen propiedades inmobiliarias en Colombia. Algunas de las columnas más relevantes incluyen información sobre la antigüedad, área de la propiedad, número de baños, estrato, presencia de garajes, entre otras. La lista completa de columnas es la siguiente:

['antiguedad_original', 'area', 'areabalcon', 'areaconstruida', 'areaterraza', 'balcon', 'banos', 'banoservicio', 'conjuntocerrado', 'cuarto_de_escoltas', 'cuartodeservicio', 'depositoocuartoutil', 'depositos', 'estrato', 'estudioobiblioteca', 'garajecubierto', 'garajes', 'gimnasio', 'habitaciones', 'halldealcobasoestar', 'instalaciondegas', 'jacuzzi', 'jardin', 'latitud', 'longitud', 'numeroascensores', 'parqueaderovisitantes', 'piscina', 'plantaelectrica', 'porteriaovigilancia', 'remodelado', 'saloncomunal', 'sauna_yo_turco', 'terraza', 'tiempodeconstruido', 'tipodegaraje', 'valor', 'valorventa', 'vigilancia', 'vista', 'zona_de_bbq', 'zonadelavanderia', 'zonaninos', 'zonasverdes']

La falta de descripciones detalladas de las columnas ha generado cierta confusión en la interpretación de los datos, por lo que se requerirá un análisis más profundo durante el proceso.

Este dataset será la base fundamental para el análisis exploratorio, el preprocesamiento de datos y el desarrollo de modelos predictivos y descriptivos en las siguientes fases del proyecto

1.3. Métrica

Como nuestro dataset y predicción tiene un enfoque de regresión, se usa el **RMSE y R2** como métrica de evaluación de desempeño. Este mide la raíz cuadrada del promedio de los errores cuadráticos entre las predicciones del modelo y los valores reales.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

donde

- Y: son los valores reales
- Y sombrero: los valores predichos
- n: número de datos

- Cuanto menor sea el valor de RMSE, mejor será el ajuste del modelo a los datos.
- El RMSE tiene la misma unidad que la variable de respuesta (Valor de la casa), lo que facilita la interpretación.

El mejor valor posible que tenemos con **R2** es 1 y el peor es 0. Una desventaja que tiene es que asume que cada variable ayuda a explicar la variación en la predicción, lo cual no siempre es cierto. Si añadimos otra variable, el valor de R2 se incrementa o permanece

igual, pero nunca disminuye. Esto puede hacernos creer que el modelo está mejorando, pero no necesariamente es así.

$$R^2 = 1 - \frac{\sum (y_i - x_i)^2}{\sum (y_i - \mu_y)}$$

1.4. Variable Objetivo

Nuestra variable objetivo es el precio de la casa, se predicen posibles valores de una casa basado en algunas características de esta, las variables más significativas e influyentes son el número de habitaciones, número de baños, estrato, número de garajes y longitud.

2. Exploración descriptiva del dataset

En esta sección, realizamos una exploración profunda del conjunto de datos para comprender la distribución y las características fundamentales.

2.1. Variables numéricas

```
# Filtra solo las columnas numéricas
numeric_df = df.select_dtypes(include=['number'])
numeric_df.columns
```

```
Index(['Unnamed: 0', 'area', 'areabalcon', 'areaconstruida', 'areaterraza',
       'banos', 'estrato', 'garajes', 'habitaciones', 'latitud', 'longitud',
       'numeroascensores', 'valor', 'valorventa'],
      dtype='object')
```

Se exploraron 13 variables numéricas del conjunto de datos, se analizaron diversas características que proporcionan información valiosa sobre las propiedades en Colombia.

Áreas

"area," "areabalcon," y "areaconstruida": Estas columnas revelan la distribución de tamaños de las propiedades. Se observaron áreas con valores nulos o ceros, lo que sugiere una posible falta de información o la necesidad de corrección. Como acción inicial, se propone sustituir los valores nulos por el promedio de las áreas según el estrato de la vivienda.

Características principales

"baños," "garajes," y "habitaciones": Estas variables ofrecen información sobre las comodidades y la capacidad de las viviendas. Se identificaron rangos inusuales, como más de 40 baños, sugiriendo la necesidad de limitar estos valores para mejorar la coherencia en el análisis.

"estrato": Es una variable importante que refleja la clasificación socioeconómica de la vivienda.

Ubicación Geográfica

"latitud" y "longitud": Estas variables pueden desempeñar un papel crucial en el análisis espacial de las viviendas en Colombia, permitiendo visualizar la distribución geográfica de las propiedades.

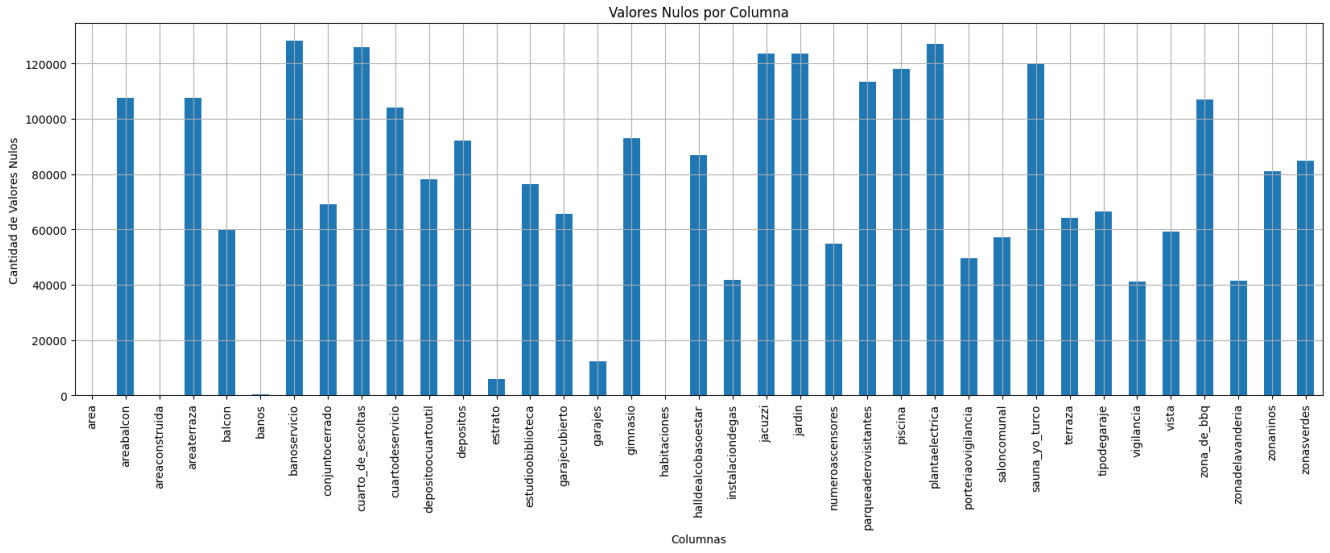
2.1.1. Datos faltantes

Durante la exploración, se identificaron 5244 valores nulos distribuidos entre las variables numéricas. Este hallazgo resalta la importancia de abordar la integridad de los datos antes de realizar análisis más avanzados. Para mejorar la calidad de los datos, se propusieron acciones correctivas:

Sustitución de Datos en Áreas de Vivienda: Se sugiere sustituir los valores nulos en las columnas de área por el promedio de las áreas según el estrato de la vivienda.

Limitación de Valores Inusuales: Se propone limitar los valores inusuales, como en el caso de baños, para garantizar la coherencia en el conjunto de datos.

Estas acciones buscan asegurar la fiabilidad de los datos numéricos, sentando las bases para análisis más detallados y la creación de modelos predictivos.



2.2. Variables categoricas

```
[ ]  1  # Filtra solo las columnas categoricas
     2  cate_df = df.select_dtypes(include=['object'])
     3  cate_df.columns
```

```
Index(['antiguedad_original', 'balcon', 'banoservicio', 'conjuntocerrado',
       'cuarto_de_escoltas', 'cuartodeservicio', 'depositoocuartoutil',
       'depositos', 'estudioobiblioteca', 'garajecubierto', 'gimnasio',
       'halldealcobasoestar', 'instalaciondegas', 'jacuzzi', 'jardin',
       'parqueaderovisitantes', 'piscina', 'plantaelectrica',
       'porteriaovigilancia', 'remodelado', 'saloncomunal', 'sauna_yo_turco',
       'terraza', 'tiempodeconstruido', 'tipodegaraje', 'vigilancia', 'vista',
       'zona_de_bbq', 'zonadelavanderia', 'zonaninos', 'zonasverdes'],
      dtype='object')
```

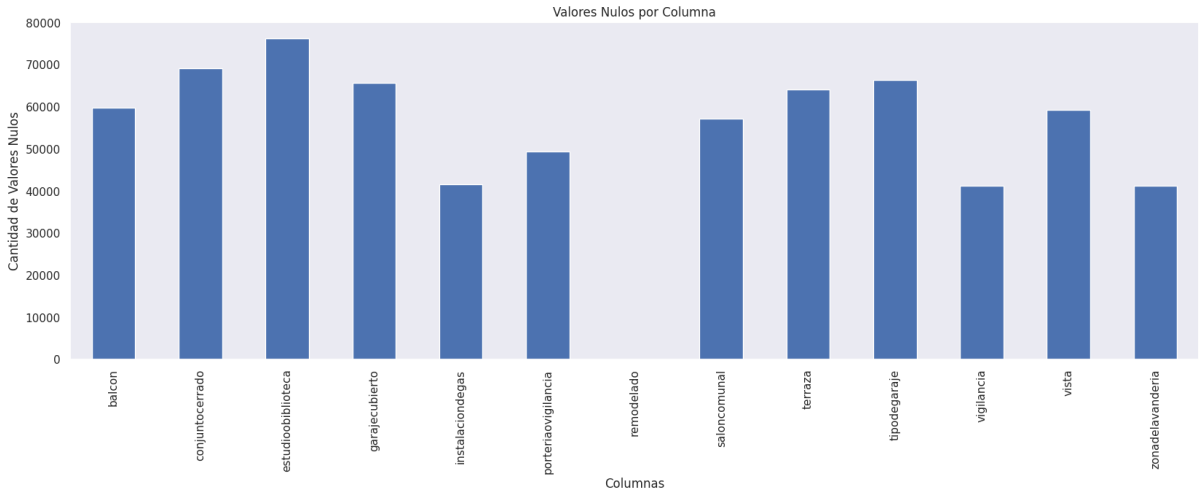
Se hallaron 34 variables categóricas en este dataset. Las variables categóricas, a diferencia de las numéricas, representan atributos que pueden clasificarse en categorías discretas, proporcionando información valiosa sobre aspectos específicos de las viviendas.

Las variables categóricas abarcan una amplia gama de características, desde la presencia de balcones hasta la existencia de vigilancia, pasando por detalles como la instalación de gas y la presencia de zonas verdes.

Algunas variables presentan una única opción de valor, lo que significa que todas las entradas para esa característica son idénticas. Estas columnas fueron objeto de atención especial para evaluar su relevancia en el análisis general.

2.2.1. Datos faltantes

En las variables categóricas se hallaron bastantes datos faltantes, también algunas columnas solo tenían un valor discreto, es decir, solo existía una opción, y si una casa no tenía esa opción estaba vacía.Algunas de las columnas relevantes con datos faltantes fueron:



2.3. Análisis de variable objetivo

En este análisis, la variable objetivo de interés es "venta", que representa el valor de venta de las propiedades. Se identificaron dos columnas idénticas que reflejaban nuestra variable objetivo: "venta" y "valorventa". La decisión fue mantener la columna "venta" para representar de manera unificada el valor de venta de cada casa en pesos colombianos.

La columna "venta" se revela como un indicador clave para entender la variación en los precios de las propiedades.

valor	valorventa
9.000000e+08	9.000000e+08
5.481475e+08	5.481475e+08
1.500000e+09	1.500000e+09
4.950000e+08	4.950000e+08
8.500000e+08	8.500000e+08
...	...
1.450000e+09	1.450000e+09
6.000000e+08	6.000000e+08
3.400000e+08	3.400000e+08
3.712500e+08	3.712500e+08
4.500000e+08	4.500000e+08

2.4. Correlación de variables

Durante la exploración, se observó una relación sustancial entre el precio de venta de una casa y diversas características, entre las cuales destacan:

Área: El tamaño de la propiedad, medida en metros cuadrados, muestra una influencia significativa en el precio de venta.

Baños: El número de baños en una casa se correlaciona con los precios, sugiriendo que propiedades con más baños tienden a tener un valor de venta más alto.

Estrato: La clasificación del estrato de la propiedad se identifica como un factor relevante, indicando una relación entre la ubicación y la categoría socioeconómica, y el precio de venta.

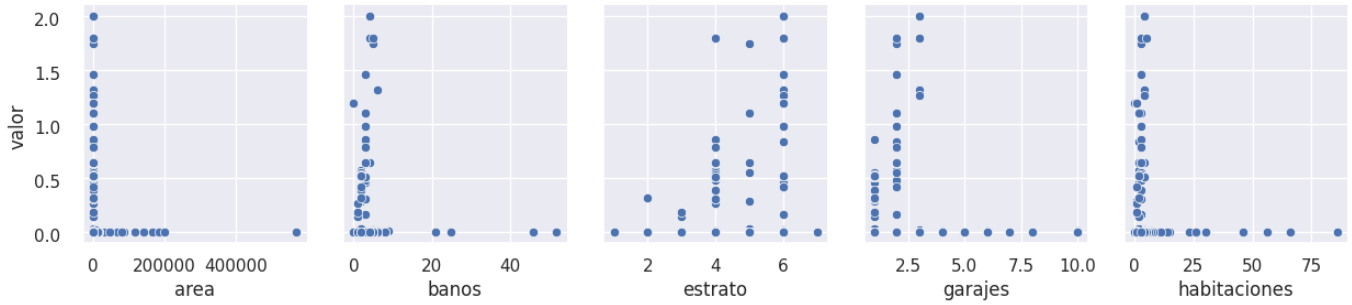
Garajes: La cantidad de garajes disponibles influye en el valor de venta, sugiriendo que propiedades con más espacios de estacionamiento pueden tener precios más altos.

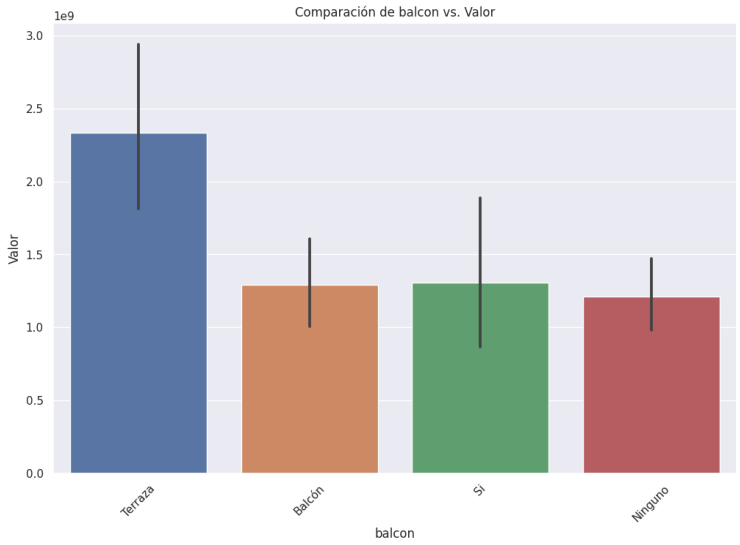
Habitaciones: El número de habitaciones en una casa se asocia con el precio, indicando que propiedades con más habitaciones pueden tener un valor de venta más alto.

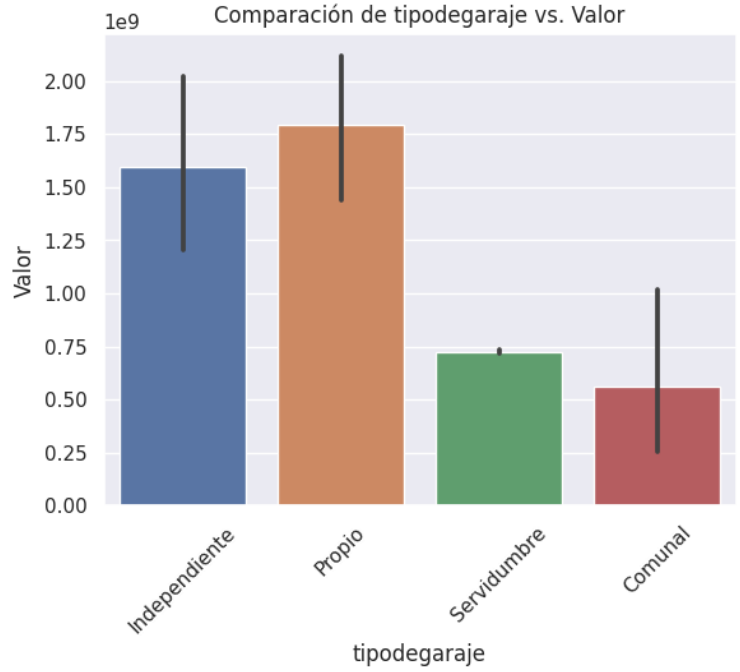
Vista Exterior: La presencia de una vista exterior se correlaciona con precios más elevados, sugiriendo que propiedades con características escénicas pueden tener un mayor atractivo y, por lo tanto, un valor de venta superior.

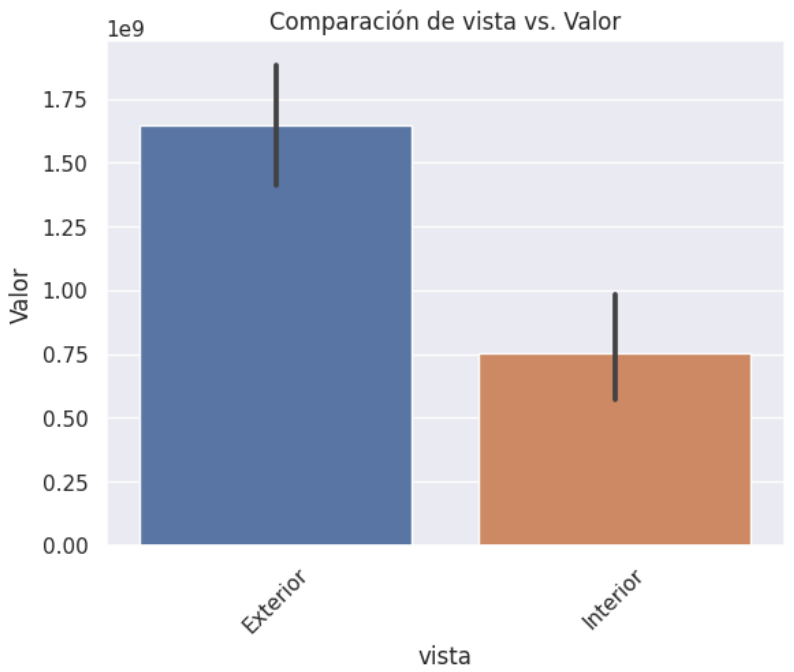
Garaje: La disponibilidad de garaje se identifica como un factor importante, sugiriendo que propiedades con garaje pueden tener precios más altos.

Terraza: La presencia de una terraza en la propiedad muestra una relación positiva con el valor de venta.


valor
area
banos
estrato
garajes
habitaciones


Comparación de balcon vs. Valor
1e9
Valor
balcon
Terraza
Balcón
Si
Ninguno


Comparación de tipodegaraje vs. Valor
1e9
Valor
tipodegaraje
Independiente
Propio
Servidumbre
Comunal



3. Iteraciones de desarrollo

En esta fase, se llevaron a cabo varias iteraciones de preprocesamiento de datos para optimizar la calidad y la relevancia de la información. A continuación, se describen las principales etapas del preprocesamiento:

3.1. Preprocesado de datos

Identificación de Columnas Redundantes
Se identificaron columnas con valores idénticos o altamente similares, como "antiguedad_original" y "tiempodeconstruido", "area" y "areaconstruida", "areabalcon" y "areaterraza", y "valor" y "valorventa". Se decidió conservar una única columna representativa y eliminar las demás.

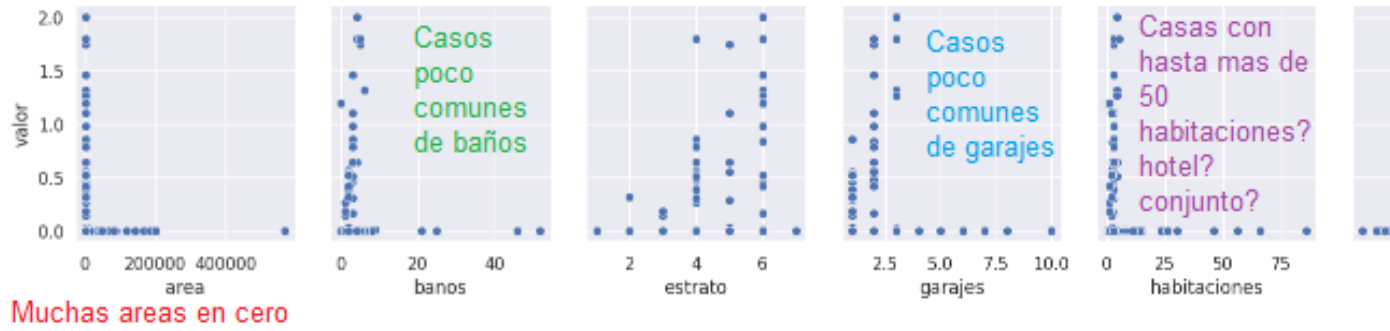
	Unnamed: 0	antiguedad_original	area	areabalcon	areaconstruida	areaterraza	balcon	banos	banoservicio	conjuntocerrado	...	tiempodeconstruido	tipo
0	0	Entre 5 y 10 años	145.00	10.0	145.00	10.0	Terraza	3.0	NaN	NaN	...	Entre 5 y 10 años	Inde
1	1	Entre 0 y 5 años	114.00	NaN	114.00	NaN	NaN	3.0	NaN	NaN	...	Entre 0 y 5 años	
2	2	Entre 5 y 10 años	170.00	30.0	170.00	30.0	Terraza	4.0	NaN	Si	...	Entre 5 y 10 años	
3	3	Entre 0 y 5 años	61.00	NaN	61.00	NaN	Balcón	1.0	NaN	NaN	...	Entre 0 y 5 años	Inde
4	4	Más de 20 años	120.50	NaN	120.50	NaN	NaN	3.0	NaN	NaN	...	Más de 20 años	Se
5	5	Más de 20 años	56.00	NaN	56.00	NaN	NaN	1.0	NaN	NaN	...	Más de 20 años	
6	6	Entre 0 y 5 años	58.00	NaN	58.00	NaN	NaN	2.0	NaN	Si	...	Entre 0 y 5 años	
7	7	Entre 10 y 20 años	211.00	NaN	211.00	NaN	NaN	3.0	NaN	NaN	...	Entre 10 y 20 años	
8	8	Entre 0 y 5 años	57.76	NaN	57.76	NaN	NaN	2.0	NaN	NaN	...	Entre 0 y 5 años	Se

datos categoricos con rangos ineficientes

Eliminación de Columnas con Datos Faltantes

Las columnas con más del 60% de datos faltantes fueron eliminadas. Este umbral se estableció para mantener la integridad de la información restante.

Ajuste de Valores Atípicos
Se observaron valores atípicos en algunas variables, como el número de baños, garajes y habitaciones. Se aplicaron límites razonables para mejorar la coherencia de los datos y evitar distorsiones en el análisis.



Imputación de Datos en Cero
Se corrigieron valores nulos en la variable "area" que originalmente estaban registrados como cero. Se imputaron valores promedio según el estrato de la propiedad.

Transformación de Variables Categóricas
Las variables categóricas fueron examinadas y se tomó la decisión de eliminar aquellas con solo una opción de valor, como "conjuntocerrado", "estudioobiblioteca", "garajecubierto", "saloncomunal" y "zonadelavanderia". Además, se realizaron transformaciones específicas en las variables "balcon", "instalaciondegas", "remodelado", "tipodegaraje" y "vigilancia" para abordar datos faltantes y mejorar la interpretación.

```
[ ]  1  # inspeccionando las variables categoricas
     2  for c in ccols:
     3      print ("%10s"%c, np.unique(dfPre[c].dropna()))
```

```
antiguedad_original ['1 a 8 años' '16 a 30 años' '9 a 15 años' 'Entre 0 y 5 años'
 'Entre 10 y 20 años' 'Entre 5 y 10 años' 'Menos de 1 año'
 'Más de 20 años' 'Más de 30 años' 'Remodelado']
    balcon ['Balcón' 'Ninguno' 'Si' 'Terraza']
conjuntocerrado ['Si']
estudioobiblioteca ['Si']
garajecubierto ['Si']
instalaciondegas ['Natural' 'Ninguno' 'Propano' 'Si']
porteriaovigilancia ['12hrs' '24hrs' 'Si']
remodelado ['No' 'Si']
saloncomunal ['Si']
   terraza ['Balcón' 'Ninguno' 'Si' 'Terraza']
tipodegaraje ['Comunal' 'Independiente' 'Propio' 'Servidumbre']
vigilancia ['12hrs' '24hrs' 'Si']
     vista ['Exterior' 'Interior']
zonadelavanderia ['Si']
```

Análisis de Relaciones con Variables Categóricas
Se exploraron las relaciones entre las variables categóricas transformadas y el precio de venta. Esto permitió identificar características, como la presencia de terraza, garaje y vista exterior, que mostraron una asociación significativa con precios más altos.

Estas iteraciones de preprocesamiento forman la base para el análisis exploratorio y la construcción de modelos predictivos. La optimización continua de este proceso es esencial para garantizar la calidad y la validez de los resultados obtenidos.

Dataset antes del Preprocesado
filas: 129076
columnas: 45

	Unnamed: 0	antiguedad_original	area	areabalcon	areaconstruida	areaterraza	balcon	banos	banoservicio	conjuntocerrado	...
0	0	Entre 5 y 10 años	145.0	10.0	145.0	10.0	Terraza	3.0	NaN	NaN	..
1	1	Entre 0 y 5 años	114.0	NaN	114.0	NaN	NaN	3.0	NaN	NaN	..
2	2	Entre 5 y 10 años	170.0	30.0	170.0	30.0	Terraza	4.0	NaN	Si	..
3	3	Entre 0 y 5 años	61.0	NaN	61.0	NaN	Balcón	1.0	NaN	NaN	..
4	4	Más de 20 años	120.5	NaN	120.5	NaN	NaN	3.0	NaN	NaN	..
...	...	...	...	...	...	...	...	...	...	...	..
129071	130746	Entre 5 y 10 años	260.0	NaN	260.0	NaN	Ninguno	3.0	NaN	NaN	..
129072	130747	Entre 5 y 10 años	166.0	40.0	166.0	40.0	Terraza	4.0	NaN	Si	..
129073	130748	Entre 5 y 10 años	87.0	NaN	87.0	NaN	Ninguno	2.0	NaN	Si	..
129074	130749	Entre 0 y 5 años	55.0	1.0	55.0	1.0	Balcón	2.0	NaN	NaN	..
129075	145551	Más de 30 años	106.0	NaN	106.0	NaN	Si	2.0	NaN	NaN	..

129076 rows × 45 columns

Dataset después del Preprocesado
filas: 129076
columnas: 17

	Unnamed: 0	antiguedad_original	area	balcon	banos	estrato	garajes	habitaciones	instalaciondegas	latitud	lon
0	0	10	145.0	Terraza	3.0	6.0	2.0	3.0	Natural	4.697760	-74.0
1	1	5	114.0	No	3.0	4.0	0.0	3.0	No	4.734622	-74.0
2	2	10	170.0	Terraza	4.0	6.0	3.0	2.0	Natural	4.653789	-74.0
3	3	5	61.0	Balcón	1.0	6.0	1.0	1.0	Natural	4.679389	-74.0
4	4	20	120.5	No	3.0	NaN	2.0	2.0	No	4.705831	-74.0
...	...	...	...	...	...	...	...	...	...	...	
129071	130746	10	260.0	Ninguno	3.0	6.0	2.0	3.0	Natural	4.653242	-74.0
129072	130747	10	166.0	Terraza	4.0	4.0	2.0	3.0	Natural	4.734498	-74.0
129073	130748	10	87.0	Ninguno	2.0	3.0	1.0	3.0	Natural	4.721652	-74.0
129074	130749	5	55.0	Balcón	2.0	4.0	1.0	1.0	Natural	4.611292	-74.0
129075	145551	30	106.0	Balcón	2.0	3.0	1.0	3.0	No	4.640000	-74.0

129076 rows × 17 columns

3.2. Modelos supervisados

Para la realización de modelos supervisados se divide un porcentaje (70%-30%) del dataset entre datos de entrenamiento y datos de prueba

```
# Dividir el conjunto de datos en características (X) y la variable objetivo (y)
X = df_null[['banos', 'estrato', 'garajes', 'habitaciones']].head(500)
y = df_null['valor'].head(500)

# Dividir el conjunto de datos en conjuntos de entrenamiento y prueba 30%
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
```

Luego se hacen los ajustes basados en el modelo seleccionado, se obtiene predicciones y se obtienen las métricas deseadas.

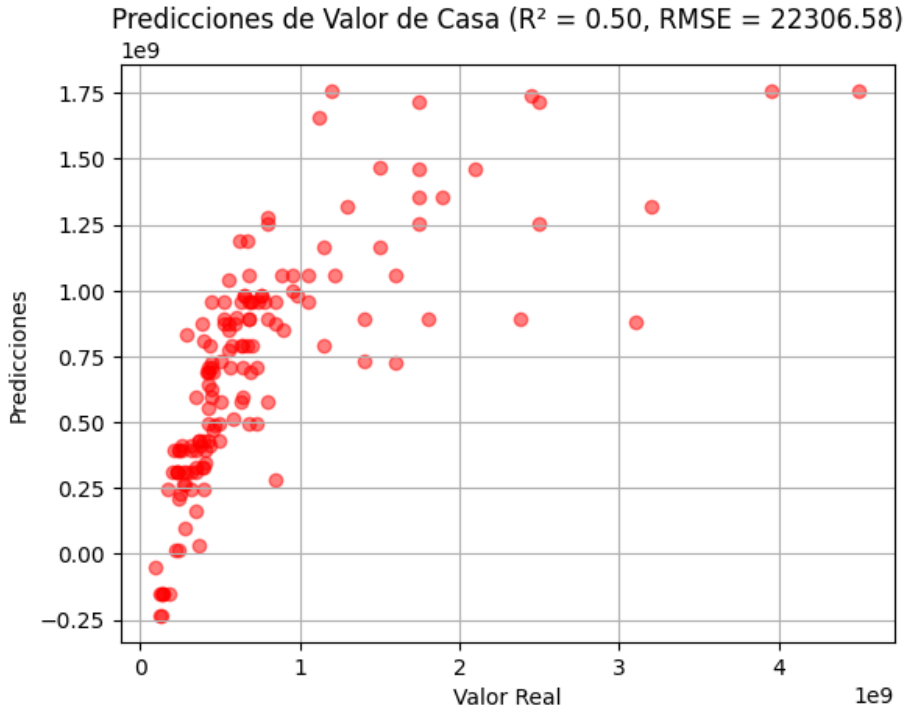
```
# Crear un modelo de Regresión Lineal Múltiple
mlr_reg = LinearRegression()

# Ajustar el modelo a los datos de entrenamiento
mlr_reg.fit(X_train, y_train)

# Realizar predicciones en el conjunto de prueba
y_preds = mlr_reg.predict(X_test)

# Calcular el coeficiente de determinación R² y rmse
r2 = r2_score(y_test, y_preds)
rmse = np.sqrt(mean_squared_error(y_test, y_preds))
```

Se grafican las predicciones:

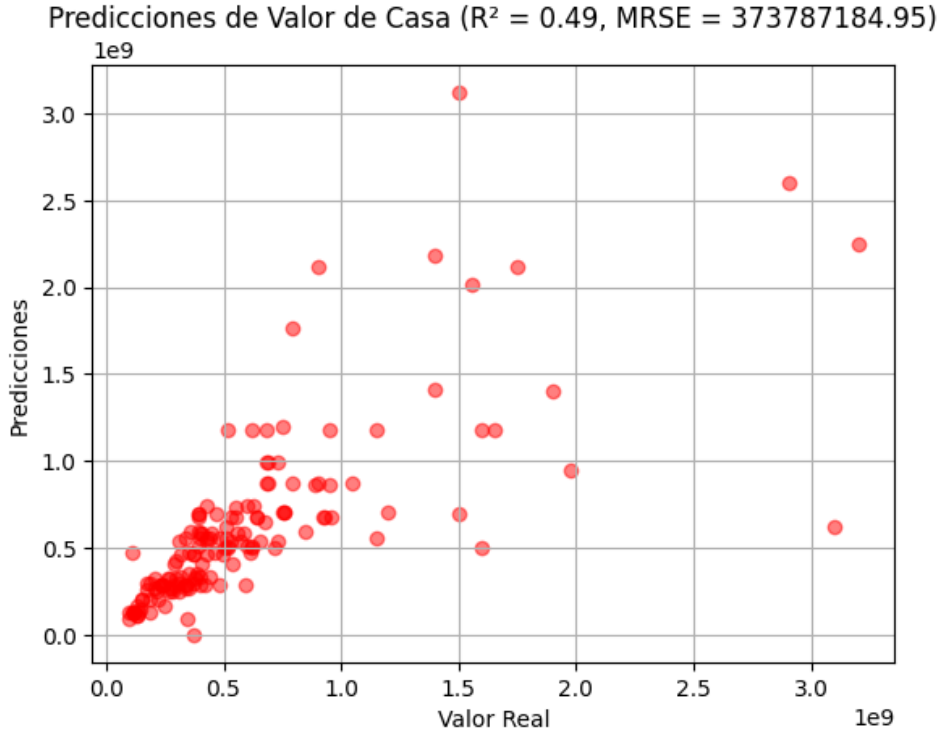
```
# Visualizar los datos originales y las predicciones
plt.scatter(y_test, y_preds, color='red', alpha=0.5)
plt.xlabel('Valor Real')
plt.ylabel('Predicciones')
```

3.2.1. Modelos seleccionados

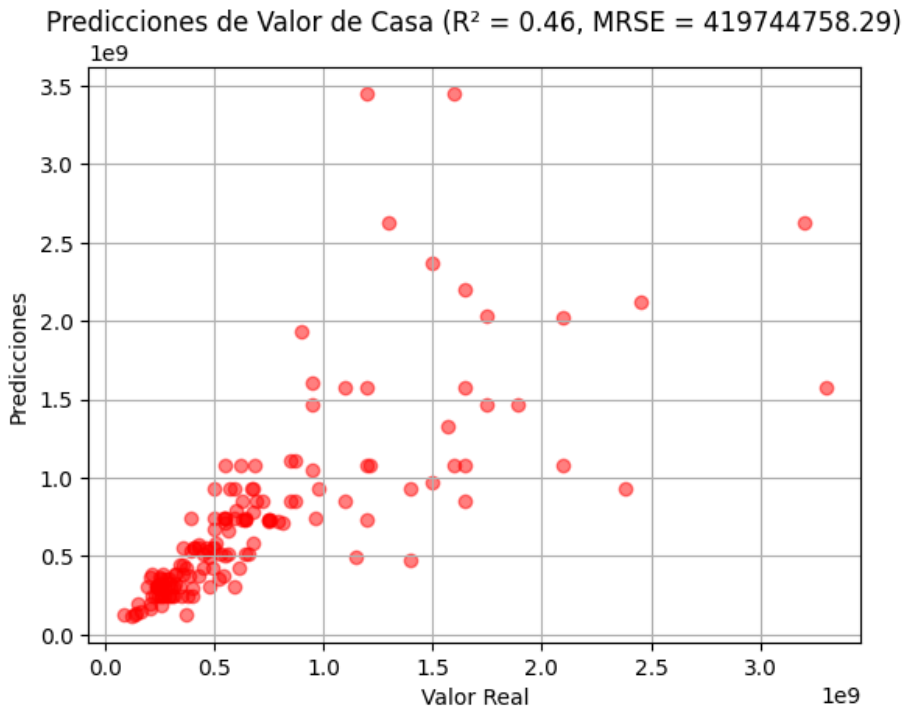
Regresión lineal:



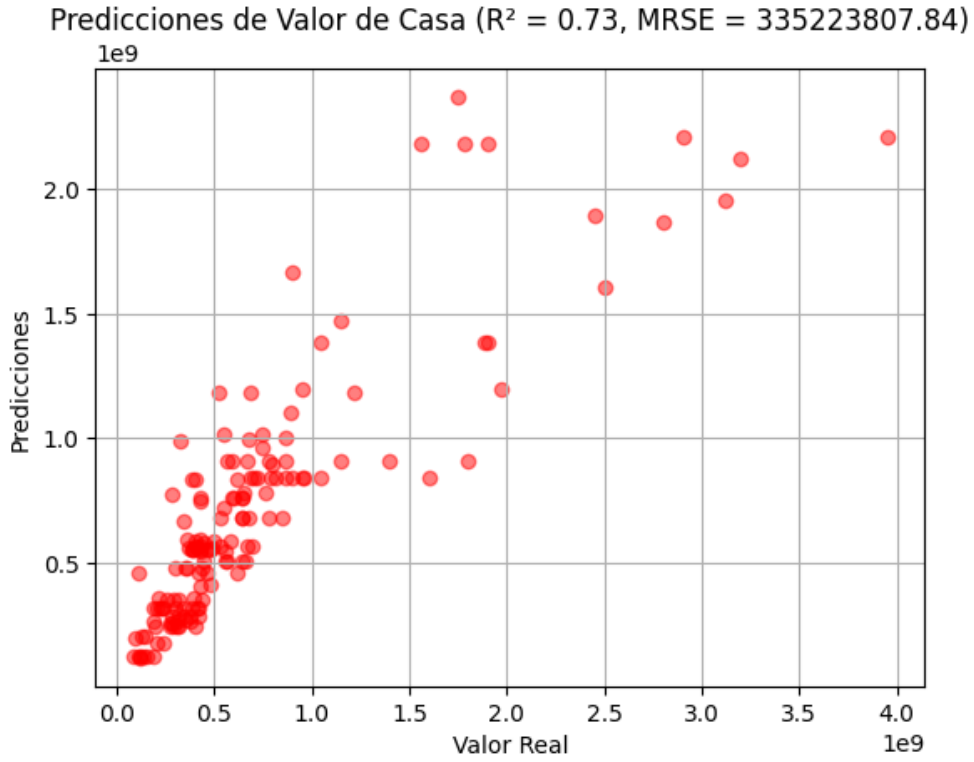
Decision Tree regressor:



Random forest regressor:



Aunque cambiando el hiper parámetro del random forest, a veces se obtenían mejores coeficientes de determinación:



Siendo regresión lineal y random forest regressor los modelos que tiene el mejor R2 y rmse.

3.2.2. Ajuste de métrica

Como no es lo mismo tener un error de un 10% en una casa de bajo precio con una que tiene un precio mucho más elevado, se hacen ciertos ajustes en las métricas

Por lo que se hace uso del **error relativo absoluto promedio**, el cual ajusta sus resultados al ponderar la diferencia entre valores predichos y los reales:

```
def rel_mrae(estimador, X, y):
  preds = estimador.predict(X)
  return np.mean(np.abs(preds-y)/y)
```

Cuando se aplican estos ajustes a cada modelo, se obtiene:

Regresión lineal: mrae = 0.8066954477138085
Decision tree regressor: mrae = 0.23085402790406237
Random forest regressor: mrae = 1.7168703572458808 (al parecer hubo un sobreajuste)

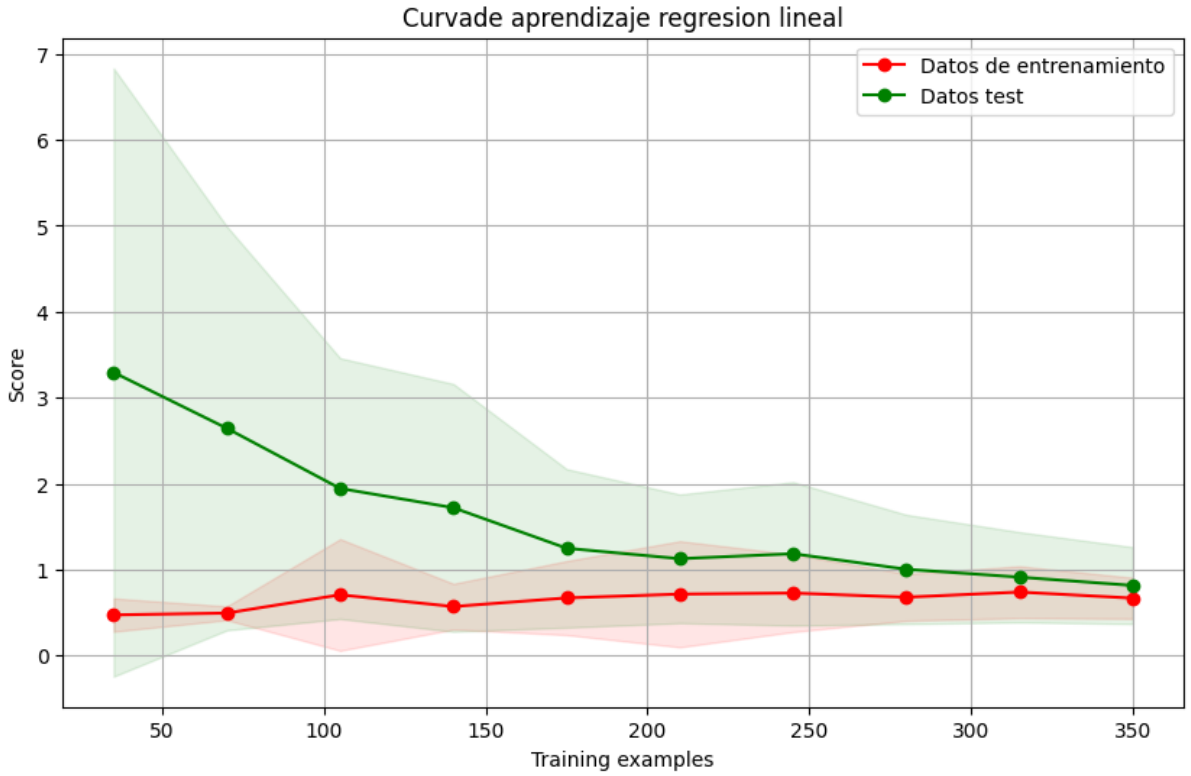
Siendo regresión Lineal y random forest regressor los mejores de los posibles modelos.

4. Mejores hiperparametros

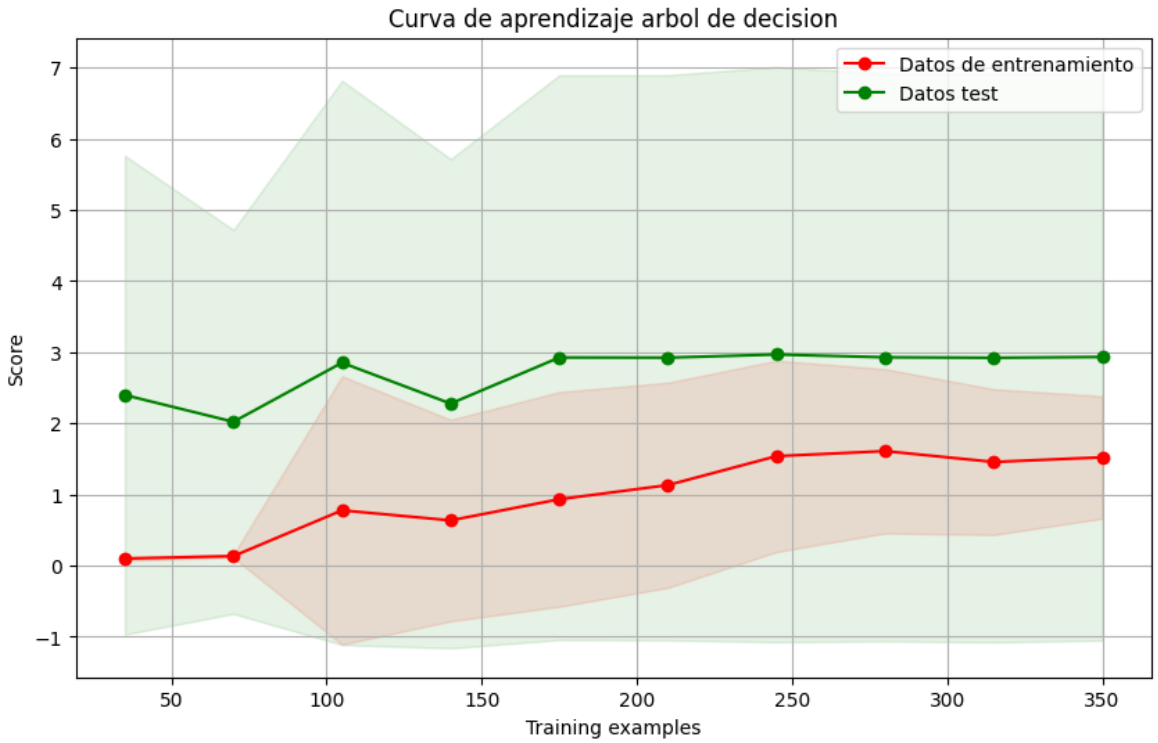
4.1. Algoritmos predictivos

4.1.1. Curvas de aprendizaje

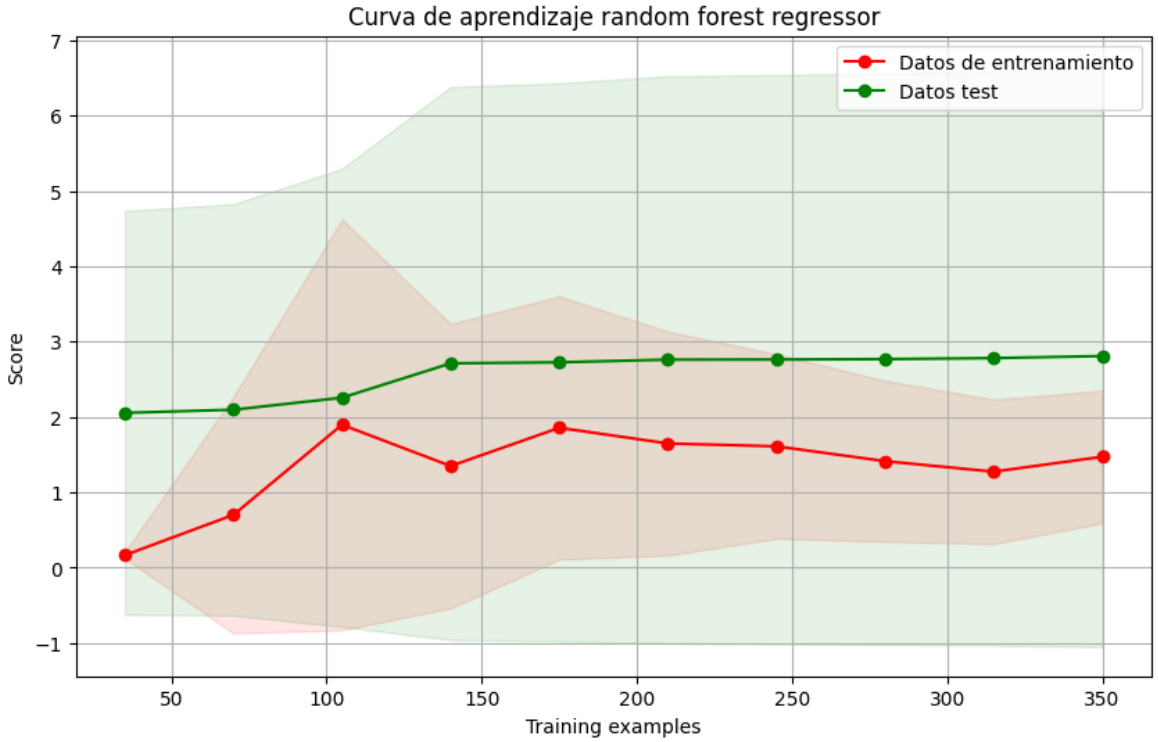
Una curva de aprendizaje es una representación gráfica de cómo el rendimiento de un modelo cambia con respecto al tamaño del conjunto de entrenamiento. Muestra cómo mejora o se estabiliza el rendimiento del modelo a medida que se incrementa la cantidad de datos utilizados para entrenarlo.



A medida que se diagnostican más datos, hay menos variabilidad por lo que es más estable, pero esto se evidencia hasta un límite de aproximadamente unos 2000 datos, cuando se hacen pruebas con más de esta cantidad, el modelo tiende a perder estabilidad. (aumenta su error)



Con un árbol de decisión, ni siquiera probando diferentes hiperparametros el modelo no logra disminuir su error.



Dependiendo de los hiperparametros como la máxima profundidad en el bosque, forest tree regressor, suele ser muy inestable o un poco mejor. Con una profundidad de 40, solía tener más error, mientras que con profundidades pequeñas (alrededor de 10) como hiper parámetro, solía mejorar las métricas.

5. Conclusiones

El proyecto de predicción de precios de casas en Colombia se destacó por abordar desafíos significativos en el procesamiento de datos y la interpretación de resultados. La identificación y manejo de variables redundantes y la imputación de datos faltantes en áreas específicas, como la antigüedad de las casas, fueron aspectos cruciales del preprocesamiento. La transformación de variables categóricas y la eliminación de columnas con datos poco informativos contribuyeron a simplificar el conjunto de datos y mejorar la capacidad predictiva del modelo.

Durante el análisis exploratorio, se identificaron variables clave, como el área, los baños, el estrato y la presencia de balcones, terrazas y garajes, que demostraron tener una relación significativa con el precio de venta. El enfoque en estos aspectos específicos proporcionó insights valiosos para el desarrollo del modelo predictivo.

La presencia de variables sin una descripción clara dificulta la interpretación precisa de ciertos aspectos y plantea obstáculos en la toma de decisiones sobre cómo manejar esos datos. Este aspecto subraya la importancia de contar con conjuntos de datos más completos y detallados para mejorar la calidad del análisis y la modelización.

Bibliografía

https://www.kaggle.com/datasets/danieleduardofajardo/colombia-house-prediction/data