SEGUNDA ENTREGA DE PROYECTO

POR:
Carlos Casas Arenas
Sebastian Aristizabal Castañeda

MATERIA:
Introducción a la Inteligencia Artificial

PROFESOR:
Raul Ramos Pollan

UNIVERSIDAD
DE ANTIOQUIA
1 8 0 3

UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA
MEDELLÍN 2023

Descripción del Progreso Alcanzado

1. Preprocesamiento del Dataset:

Inicialmente, se cargaron los datos y se realizó un análisis superficial de las columnas.

Identificación de Redundancia

Se identificaron columnas con datos idénticos, lo que sugiere redundancia en el dataset. Se encontraron las siguientes parejas de columnas con valores casi idénticos:

"antiguedad_original" = "tiempodeconstruido"
"area" = "areaconstruida"
"areabalcon" = "areaterraza"
"valor" = "valorventa"

Limpieza de Columnas

Con el objetivo de reducir la redundancia y el número de columnas, se eliminaron las columnas con más del 60% de datos faltantes. Además, de cada par de columnas idénticas, se mantuvo una y se eliminó la otra. Esto redujo el número de columnas de 45 a 24.

Tratamiento de Datos Faltantes

Se analizaron las columnas restantes con datos faltantes para determinar cómo tratarlos en futuros modelos. Esto permitió un enfoque más claro en las columnas que requerían atención en el análisis de modelos. Notamos que la mayoría de datos faltantes provenían de columnas categóricas, y a su vez la mayor cantidad de columnas categóricas hacen referencia a caracteristicas de casas mayores a estrato 3 o incluso 4, caracteristicas como “conjuntocerrado”, “garajecubierto”, “porteriaovigilancia”, “saloncomunal”, etc

Mejora de la Columna "area"

Se identificaron valores de "area" iguales a cero, lo que no tenía sentido. Se decidió reemplazar estos valores por el promedio de "area" según el estrato de la vivienda. Además, se limitó el valor de "area" a 8000 para mejorar su impacto en el precio de la vivienda.

Límites en Otras Columnas Numéricas

Se impusieron límites en otras columnas numéricas, como "baños", "garajes" y "habitaciones", para reflejar mejor la realidad de los datos y hacerlos más coherentes.

Visualización de Variables Numéricas

Se realizaron gráficos de relación entre variables numéricas para entender mejor su impacto en el precio de las viviendas después del preprocesamiento.

Tratamiento de Variables Categóricas

Se convirtió la variable categórica "antiguedad_original" en una variable numérica que encapsula la antigüedad de cada casa de manera más comprensible. Nos enfocamos más en esta variable categórica ya que consideramos que puede ser una característica importante para el valor de una casa

Visualización de Variables Categóricas

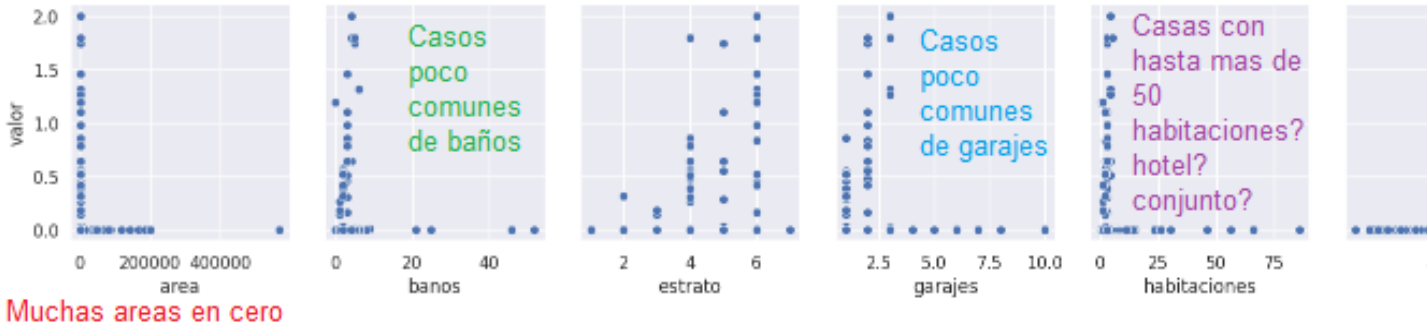
Se realizaron gráficos de relación entre la variable "valor" y otras variables categóricas, destacando las características que parecen influir en un mayor valor de la vivienda.

El preprocesamiento permitió simplificar y limpiar el dataset, además de identificar las variables clave que influyen en el precio de las viviendas. Estos hallazgos sientan las bases para un análisis y modelos más profundos en etapas posteriores del trabajo.

Dataset antes del Preprocesado

	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



Dataset despues de Preprocesado

No hay columnas repetidas, aun se ven muchos valores nulos, pero son correspondientes a valores categoricos

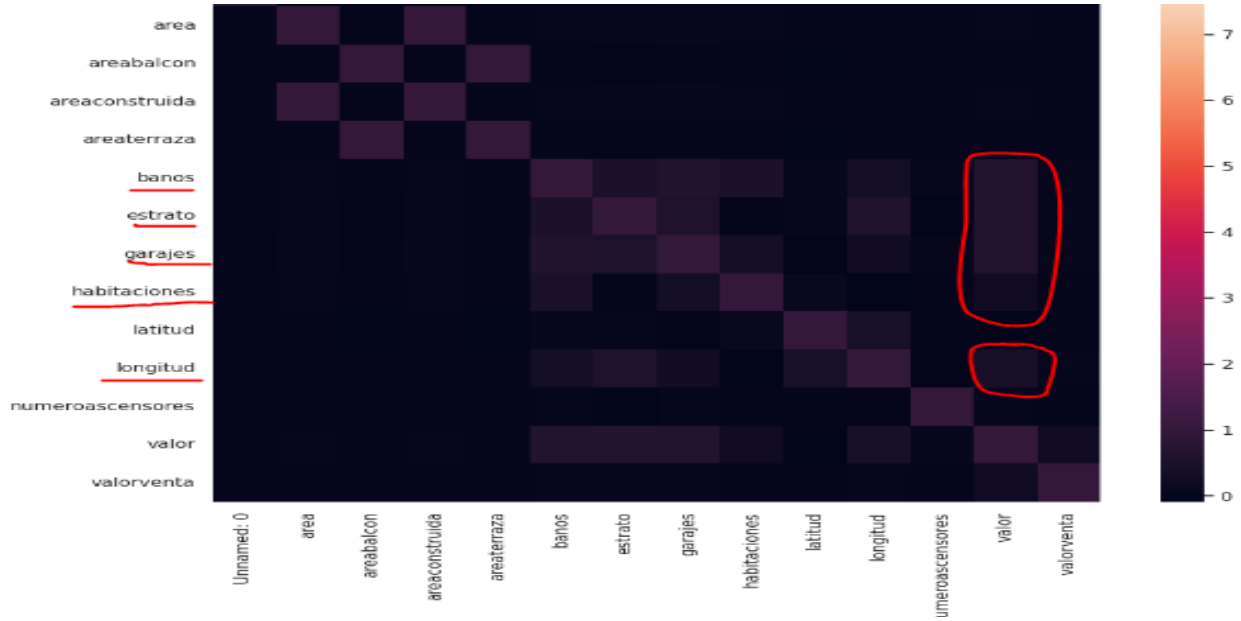
	Unnamed: 0	antiguedad_original	area	balcon	banos	conjuntocerrado	estrato	estudioobiblioteca	garajecubierto	garajes	...	numeroascensores
0	0	10	145.0	Terraza	3.0	NaN	6.0	Si	Si	2.0	...	2.0
1	1	5	114.0	NaN	3.0	NaN	4.0	NaN	NaN	0.0	...	NaN
2	2	10	170.0	Terraza	4.0	Si	6.0	Si	Si	3.0	...	1.0
3	3	5	61.0	Balcón	1.0	NaN	6.0	NaN	Si	1.0	...	1.0
4	4	20	120.5	NaN	3.0	NaN	NaN	Si	NaN	2.0	...	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...
129071	130746	10	260.0	Ninguno	3.0	NaN	6.0	Si	Si	2.0	...	1.0
129072	130747	10	166.0	Terraza	4.0	Si	4.0	Si	Si	2.0	...	2.0
129073	130748	10	87.0	Ninguno	2.0	Si	3.0	Si	NaN	1.0	...	1.0

datos numericos

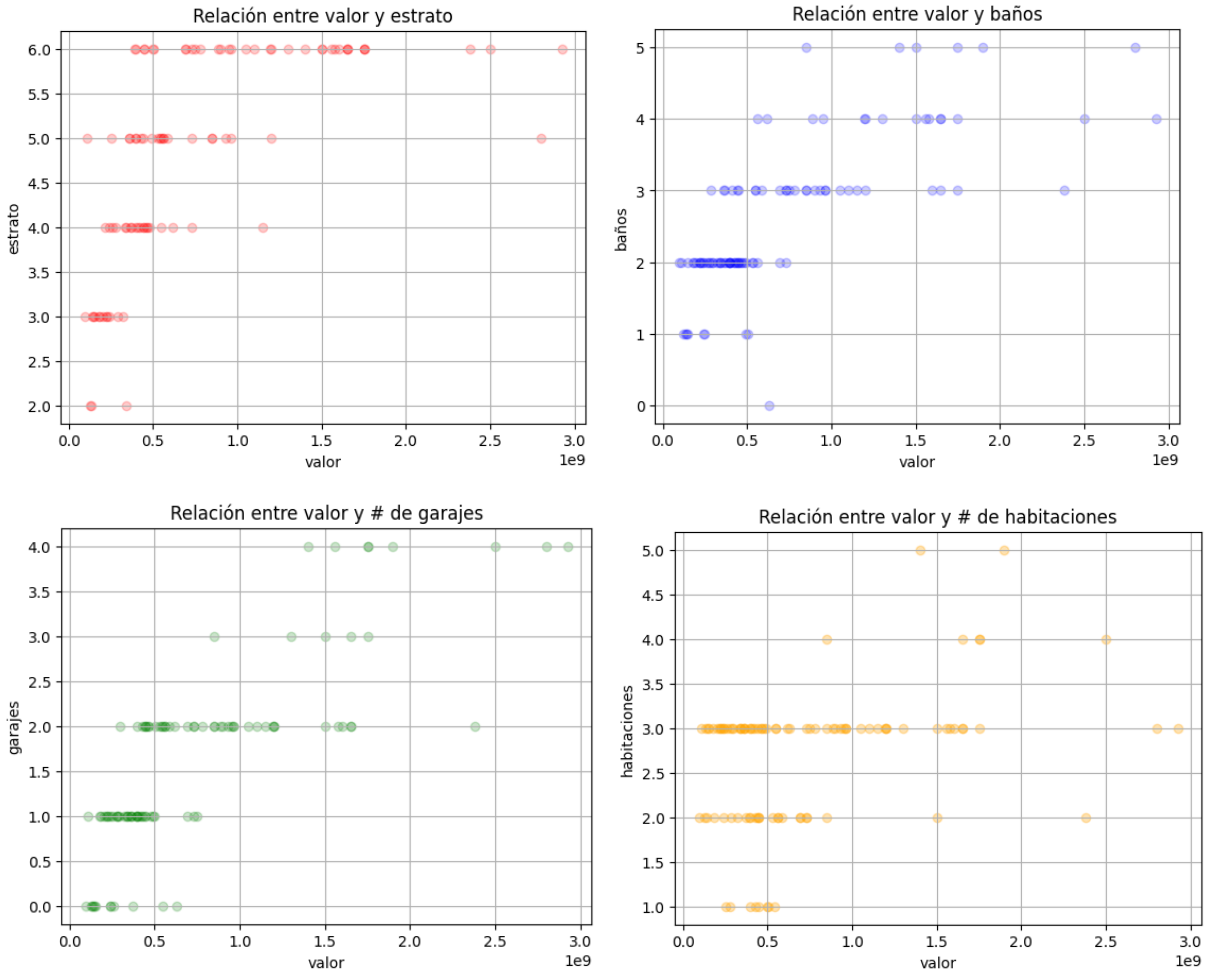
Comenzamos a ver mas claramente una relacion

2. Desarrollo de Modelos Iniciales:

Usando los datos ‘limpios’ tras el preprocesado se comienza a obtener información en busca de validar patrones y tendencias en nuestros futuros modelos. Por lo que se hará un breve análisis de la relación de las variables más relevantes que influyen en los precios de una casa, como se pudo detectar en el preprocesado de datos las cuales fueron estrato, número de baños, número de habitaciones, número de garajes y longitud. Se observa en la matriz de correlación.



Análisis relaciones entre variables:



Relación entre el valor de una casa y su estrato:

Esta relación muestra que hasta precios de 500 millones lo más seguro es encontrar casas de estrato 4 y con algunas otras influencias tal vez como el sector, casas de estrato 5, de igual manera tras superar los 500 millones de pesos se pueden encontrar casas con mayor influencia en el estrato 5 y 6, ya a partir de los 1000 millones es seguro conseguir casas de estrato 6, cabe resaltar que esta información está siendo influenciada por otras variables como el número de habitaciones con las que cuenta, el número de baños, su localización etc…

Relación entre el valor de una casa y el número de baños:

Parece ser que en un rango entre los 200 millones y 500 millones de pesos aproximadamente es altamente posible conseguir casas con hasta 2 baños, por otra parte con precios superiores a 500 millones y menores a 1000 millones se encontrarán en su mayoría casas de 3 baños. Mientras que casas con valores de 1000 millones en adelante, exclusivamente serán hogares con 3 ,4 y hasta 5 baños.

Relación entre el valor de una casa y el número de garajes:

Se observa que la gran mayoría de casas con precios menores a 500 millones de pesos llegan a tener un solo garaje, prácticamente precios menores a 100 millones no tienen. Por otro lado, precios mayores a 500 millones son casi una garantía de tener al menos 2 garajes. Si se superan los 1000 millones de pesos se pueden encontrar casas de hasta 3 y 4 garajes pero parece no ser un dato de relevancia para el comprador puesto que, aun con estos altos precios, se adquieren bastantes casas con al menos 2 garajes.

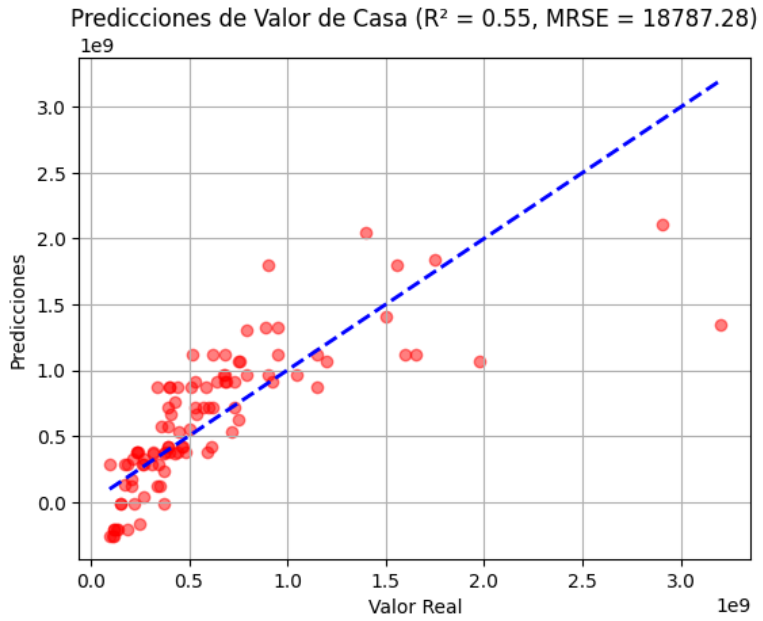
Relación entre el valor de una casa y el número de habitaciones:

La mayor tendencia observada es que casas con precios superiores a 200 millones y menores a 1250 millones de pesos aproximadamente es bastante alta la posibilidad de tener 2 o 3 habitaciones. Mientras que casas que cuestan 1500 millones de pesos en adelante se encuentran con más de 3 habitaciones (hasta 5) pero al igual que sucede con los garajes, parece que 3 habitaciones suele ser suficiente en la mayoría de los casos, solo algunos compradores en específico querrán 4 o 5 habitaciones.

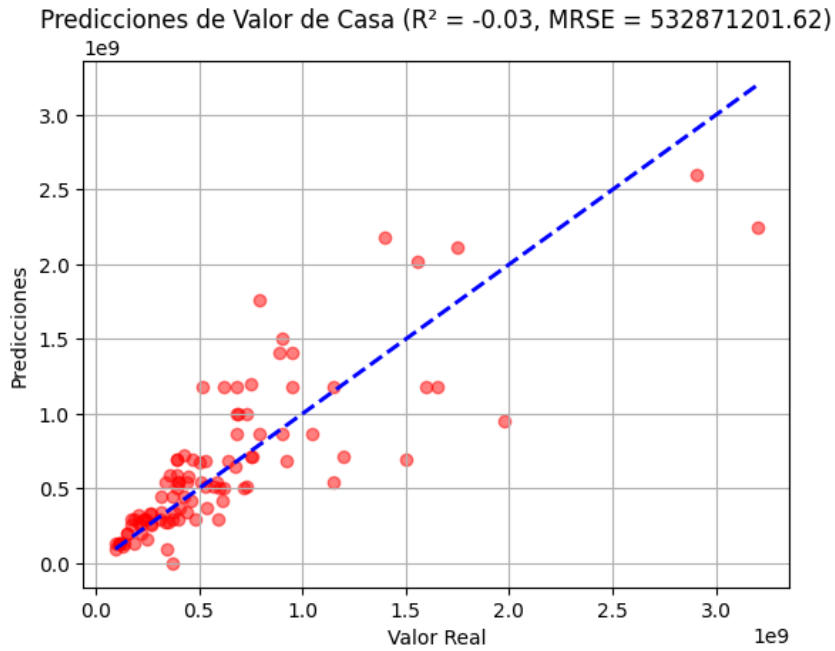
Primeros modelos predictivos:

Se usarán una cantidad limitada de datos 1000 aproximadamente para no generar un overfitting, se hacen pruebas con diferentes hiperparametros en TreeRegressor y Random Forest Regressor para encontrar el mejor modelo, también se usaran 20% datos de prueba y 80% datos de entrenamiento.

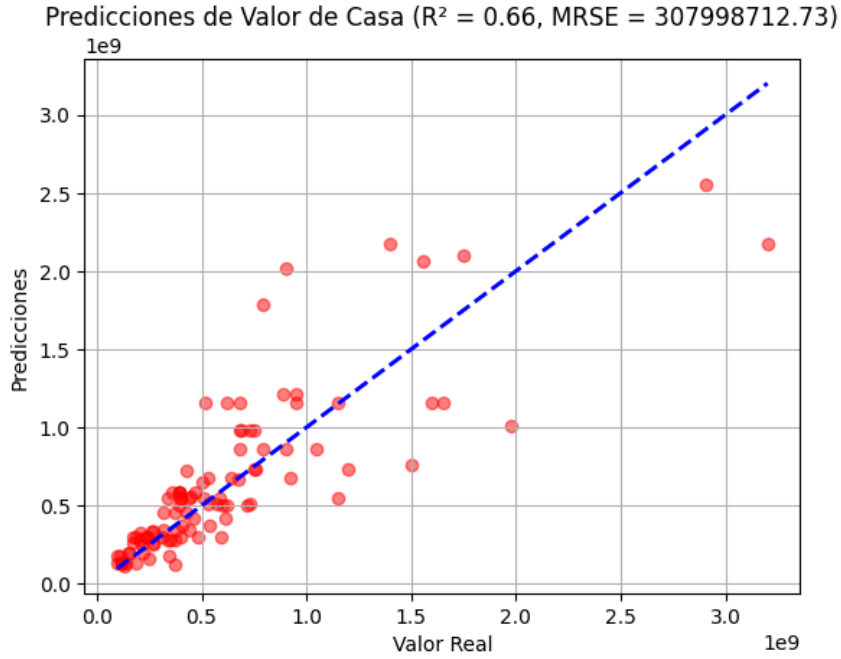
Regresión lineal: contiene un R2 de 0.55 lo cual indica un modelo no muy bueno, el rmse es muy grande lo cual señala mucha variabilidad entre los datos y este tipo de modelo puede no ser el mejor



TreeRegressor: contiene un R2 de -0.03 lo cual indica un modelo peor que el de regresión lineal, el mrse es muy grande lo cual señala mucha variabilidad entre los datos y este tipo de modelo no es bueno



RandomForest Regressor: contiene un R2 de 0.66 lo cual indica ser el mejor tipo de modelo hasta ahora, el mrse es muy grande lo cual señala mucha variabilidad entre los datos y este tipo de modelo puede no estar en las mejores condiciones



Aunque los modelos no son los mejores, hemos notado discrepancia en las gráficas, puesto que no debería ser líneas rectas para un tree Regressor, o un RandomForest Regressor, para ajustar la precisión de los datos se harán futuras pruebas eliminando algunas variables como números de garajes o agregando otras como longitud de la casa, e incluso usando un support vector machine etc… Además se verifican con mayor exactitud el comportamiento de las gráficas y sus datos para comprobar una correcta respuesta de las predicciones.