

# Algoritmo ML

## Predicción de precios inmobiliarios

Joaquín Gorostiza  
CoderHouse

# Valor Inmobiliario en Madrid

Una renombrada cadena inmobiliaria se enfrentaba a un dilema común pero desafiante en su búsqueda por brindar un servicio excepcional a sus clientes. Los dueños de viviendas, impulsados por emociones y recuerdos arraigados, a menudo sobrevaloran el precio de sus propiedades al momento de venderlas. Esto conducía a una especulación inmobiliaria que dificulta la transparencia y la equidad en el mercado.

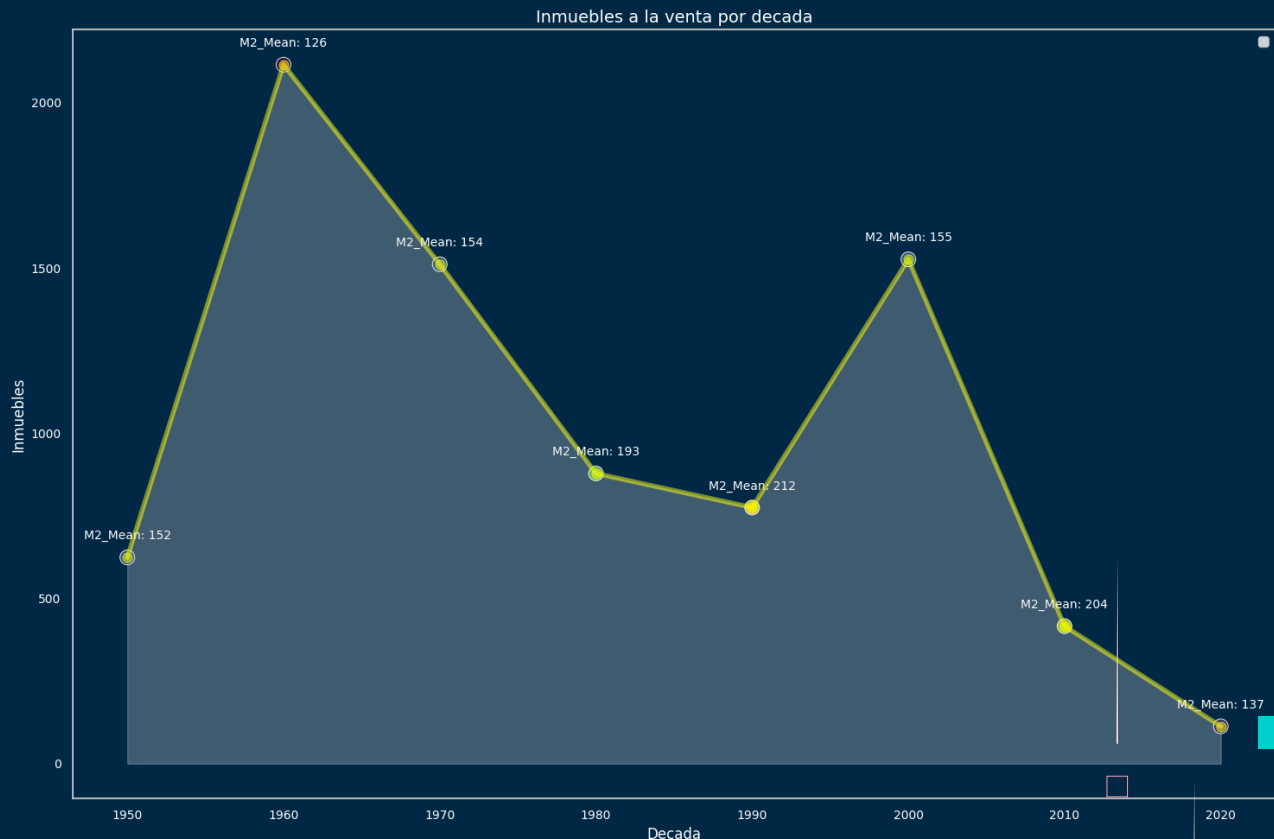
En medio de este panorama, la cadena inmobiliaria se propuso revolucionar la forma en que tasaban las propiedades en venta. Su objetivo era claro: crear una herramienta de predicción que pudiera estimar con precisión el valor de venta de las viviendas, permitiendo a los clientes obtener una visión realista del mercado y facilitando transacciones más justas.



# Inmuebles por década

La necesidad de una herramienta como esta se volvía aún más evidente al observar la variación de inmuebles en venta por década.

Este gráfico revela cómo la cantidad de inmuebles disponibles para la venta ha evolucionado y cómo esta variación influye en la cantidad de tasaciones que los agentes inmobiliarios deben realizar. A medida que los ciclos de ventas y las demandas del mercado cambiaban, el trabajo de tasación se volvía más complejo.



# Metadata de entrada

## Estadísticas

Number of variables 58

Number of observations 21736

Missing cells 530410

Missing cells (%) 42.1%

Duplicate rows (%) 0.0%

## Tipos de variables

Numeric 13

Text 6

Unsupported 10

Categorical 29

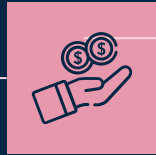


# Cuales son los factores más influyentes en el precio de los inmuebles?

Variable  
objetivo



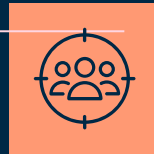
Precio de  
venta



Variables  
predictors?

Hipotéticamente  
podríamos pensar  
que las variables  
más influyentes en el  
precio de un  
inmueble son:

- ubicación geográfica
- amenities
- orientación solar
- metraje construido
- antigüedad del inmueble



A continuación  
lo analizaremos  
con algunos  
gráficos



# Univariado

## Resumen de las variables

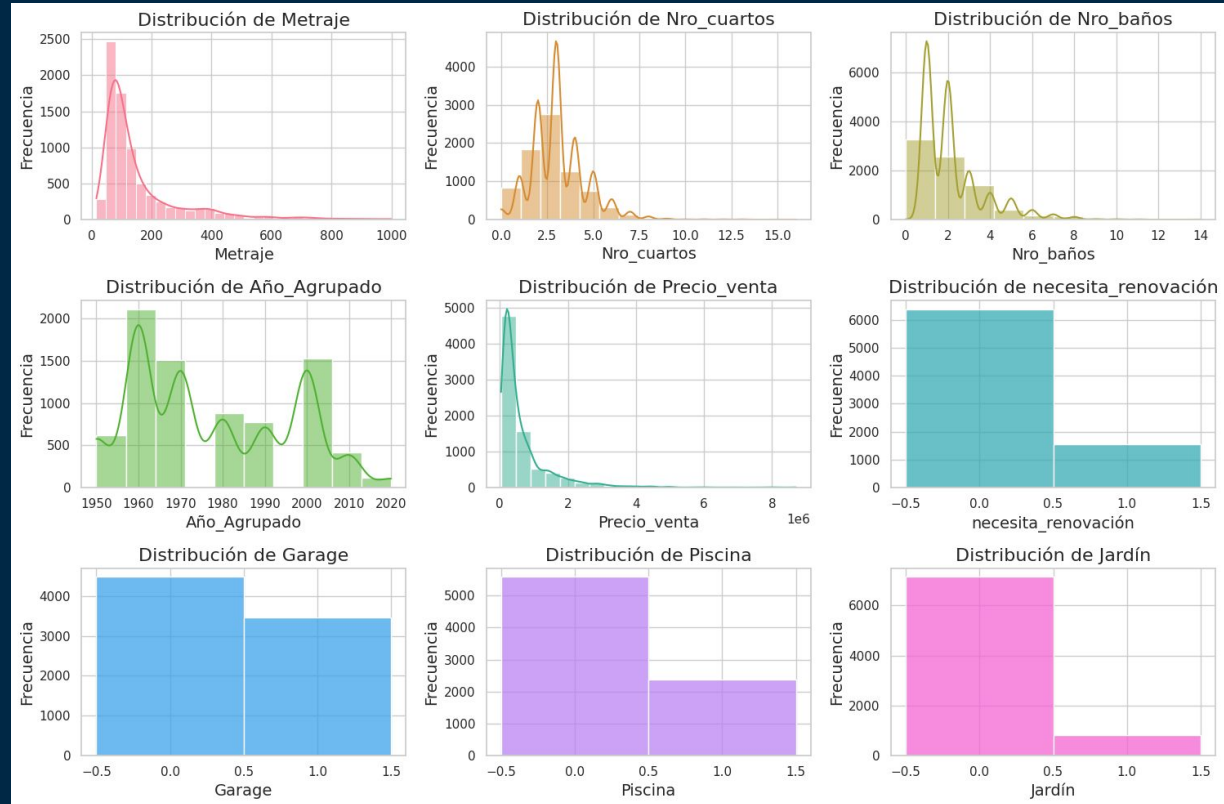
De la lista de variables, se puede observar que la mayoría de ellas son de tipo booleana (Desde Jardín hasta el final de la lista.)

	count	mean	std	min	25%	50%	75%	max
Metraje	7904.0	158.308072	148.094898	16.0	72.0	101.0	175.00	999.0
Nro_cuartos	7958.0	3.129178	1.510851	0.0	2.0	3.0	4.00	16.0
Nro_baños	7958.0	2.183840	1.511883	0.0	1.0	2.0	3.00	14.0
Precio_venta	7958.0	647234.741895	784437.135372	36000.0	180000.0	350000.0	775000.00	8700000.0
PrecioxM2	7958.0	3629.537949	1672.160561	706.0	2354.0	3354.5	4479.75	14542.0
Año_Agrupado	7958.0	1977.378738	18.580790	1950.0	1960.0	1970.0	2000.00	2020.0
Jardín	7958.0	0.101910	0.302549	0.0	0.0	0.0	0.00	1.0
Piscina	7958.0	0.297688	0.457270	0.0	0.0	0.0	1.00	1.0
Garage	7958.0	0.436291	0.495956	0.0	0.0	0.0	1.00	1.0
is_floor_under	7699.0	0.108196	0.310648	0.0	0.0	0.0	0.00	1.0
necesita_renovación	7958.0	0.196406	0.397304	0.0	0.0	0.0	0.00	1.0
has_lift	7958.0	0.676301	0.467916	0.0	0.0	1.0	1.00	1.0
is_exterior	6689.0	0.939901	0.237687	0.0	1.0	1.0	1.00	1.0
terrazza	7958.0	0.517090	0.499739	0.0	0.0	1.0	1.00	1.0
balcón	7958.0	0.147022	0.354150	0.0	0.0	0.0	0.00	1.0
zonas_verdes	7958.0	0.241644	0.428106	0.0	0.0	0.0	0.00	1.0
vestidor	7958.0	0.410028	0.491869	0.0	0.0	0.0	1.00	1.0
armarios equipados	7958.0	0.710857	0.453393	0.0	0.0	1.0	1.00	1.0

# Univariado

## Histogramas

Análisis de  
algunas  
variables

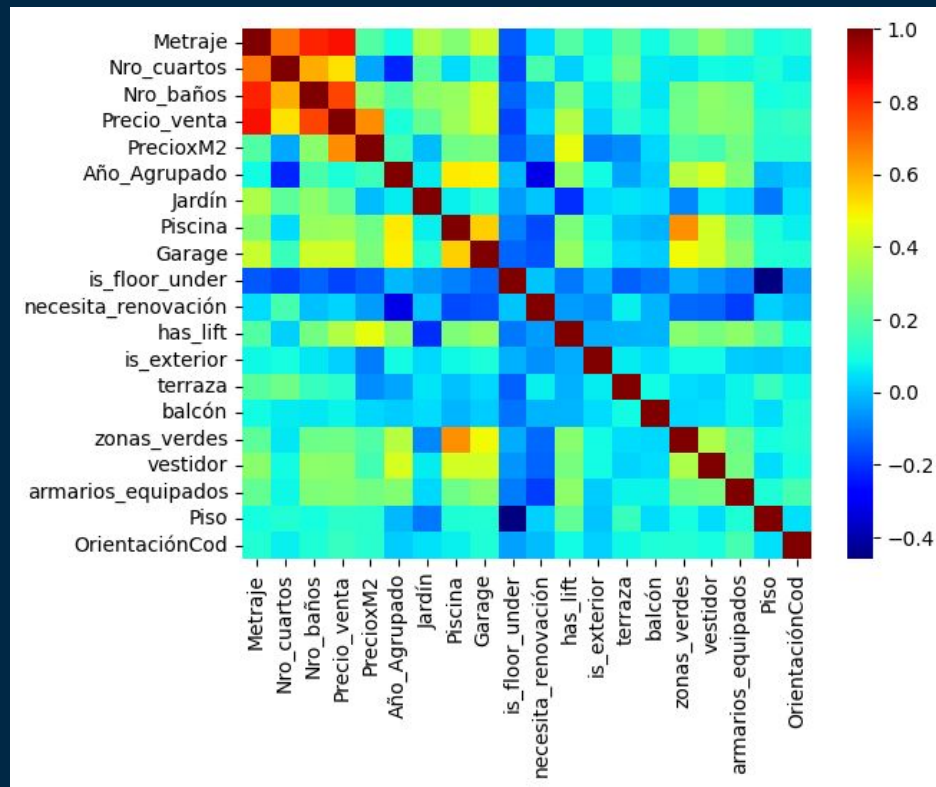


# Bivariado

## Correlación entre variables

Las variables con mayor correlación se da en la esquina superior izquierda entre metraje, nro de cuartos, nro de baños con el precio de venta.

Por otro lado, no se evidencia una alta correlación entre el precioxm2 con el precio de venta, pero los análisis realizados reflejan que la combinación entre metraje y precioxm2 generan una predicción exacta, por lo cual se debe interpretar como una redundancia. Por este motivo se decide eliminar la variable precioxm2.



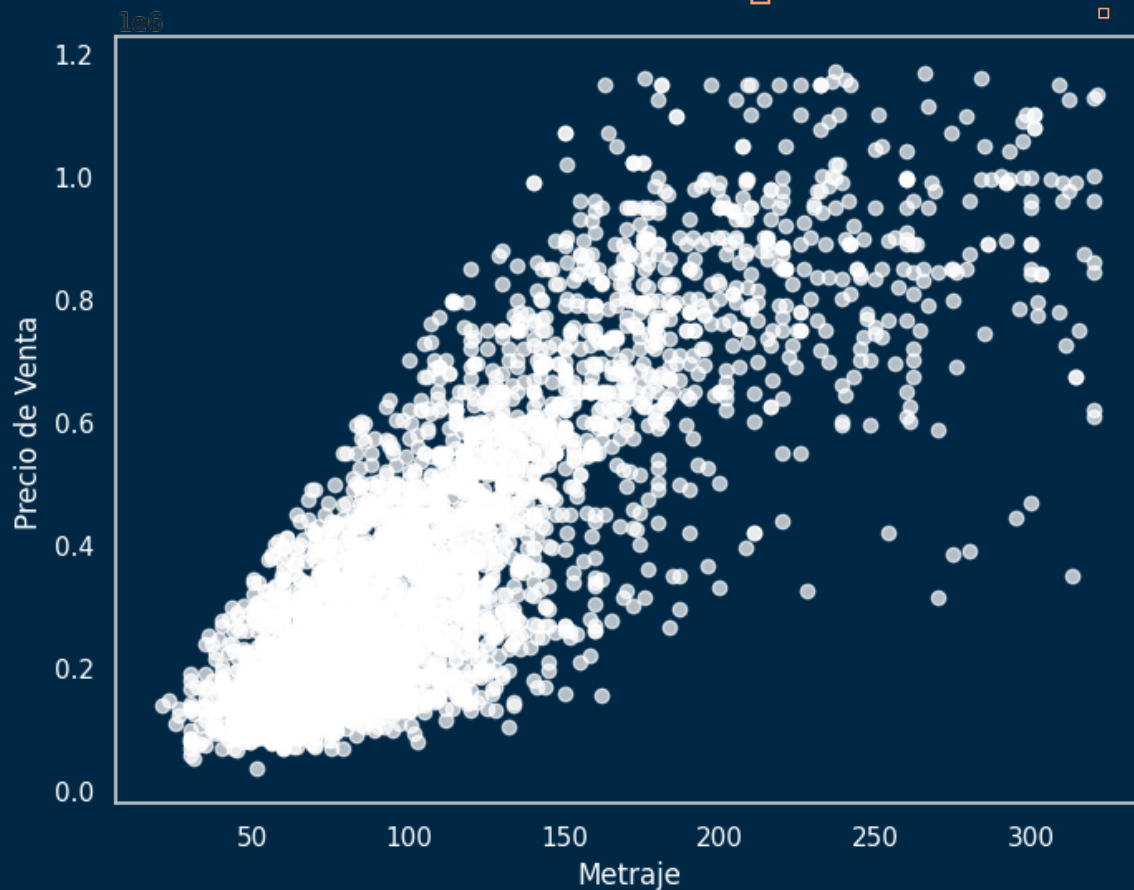


# Bivariado

## Precio venta vs Metraje

Se puede observar una correlación fuerte entre las variables precio de venta y metraje.

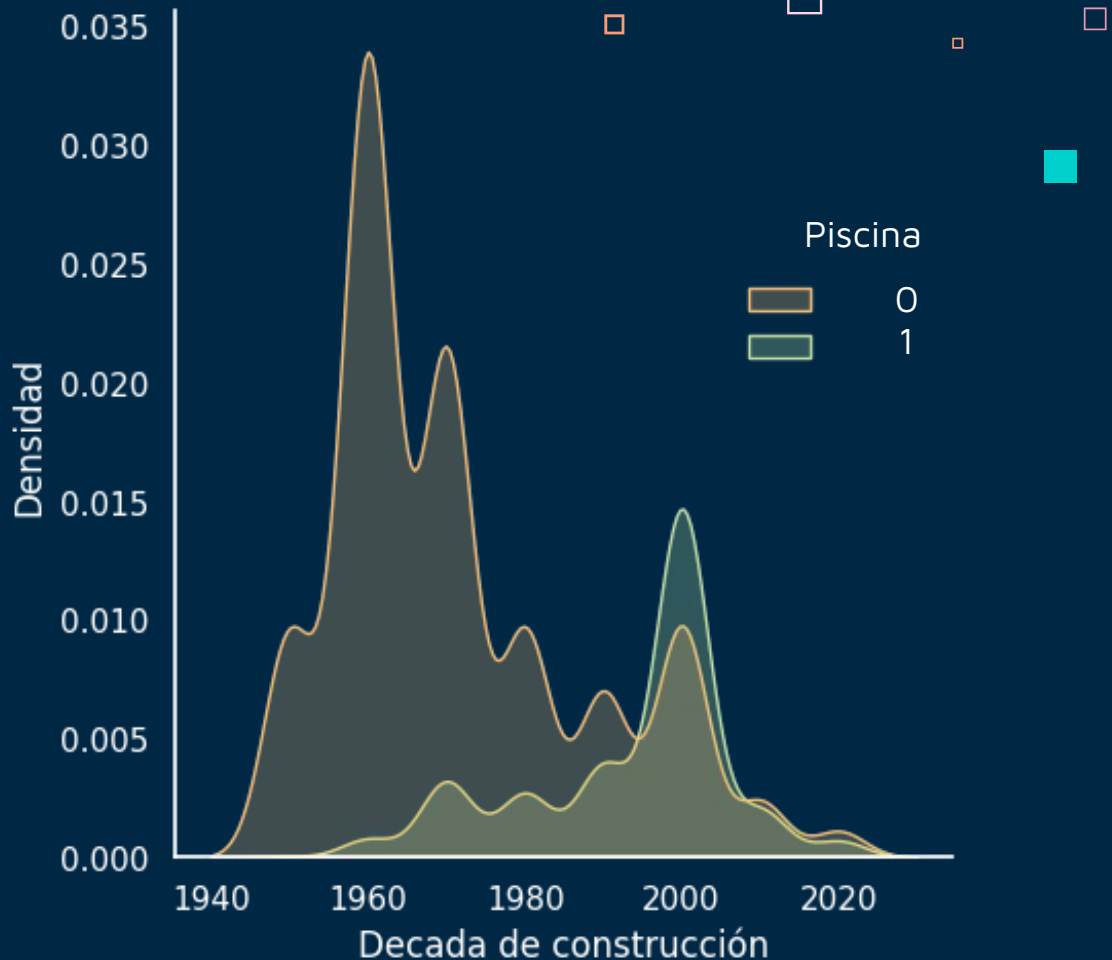
Son directamente proporcionales, ya que cuanto mayor es el valor del metraje, mayor es el valor del precio de venta.



# Bivariado

## Decada constructiva vs Piscina

Al analizar las variables década de construcción vs viviendas con piscina se puede observar que existe un desfase entre ambas variables desde 1940 hasta la década del 1990, en donde a partir de aquí sí se correlacionan ambas variables. Por tal motivo, ambas variables tienen una correlación fuerte parcialmente, pero una correlación débil si evaluamos ambas variables a nivel general.

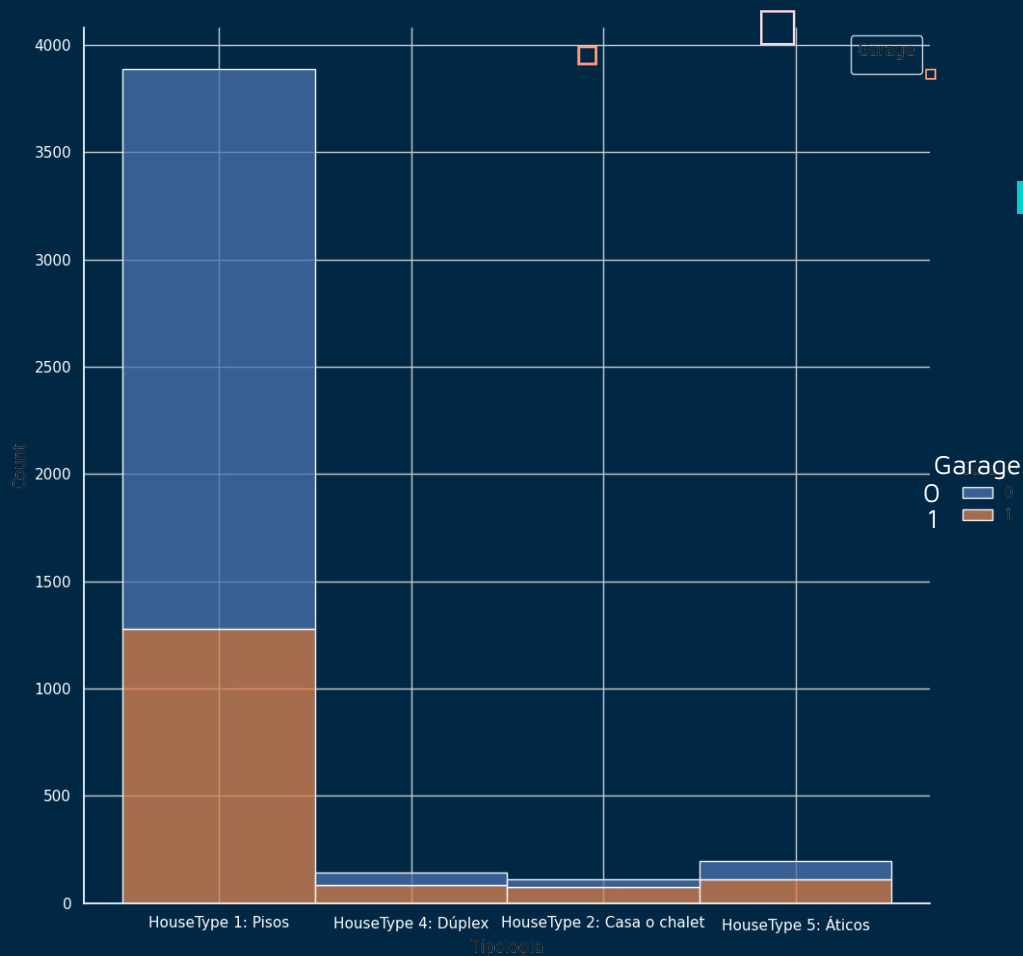


# Bivariado

## Tipología vs garage

A nivel general, podemos interpretar que la proporción de viviendas con o sin garage se reparte casi uniformemente para todas las tipologías.

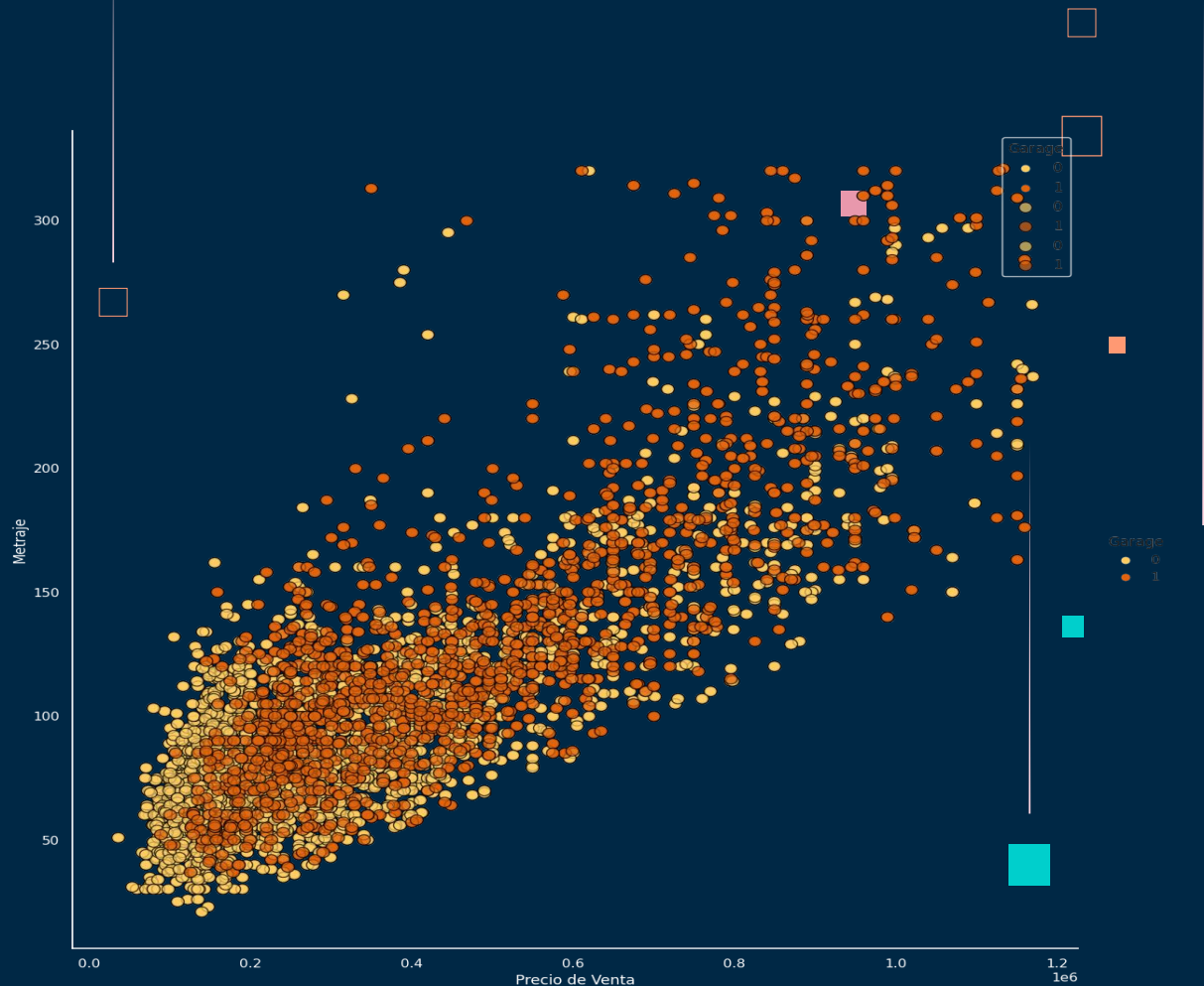
Aunque, si miramos más en detalle se ven pequeñas diferencias en las proporciones para la tipología pisos(monoambiente) en el cual predominan las viviendas sin garage, dado que presentan menores metrajes edificados; o por el contrario, la tipología casas o chalet, el cual significa mayor metraje construido, se aprecia una leve mayoría de viviendas construidas con garages.



# Multivariado

## Precio venta vs Metraje + C/S Garage

Podemos observar que existe un límite en 100 k en donde viviendas que cuestan menos de 100k y tienen menos de 100 mt2 no presentan garage, y en caso contrario, mayores a 100k y con mayor metraje que 100 mt2 se visualiza una mancha uniforme de viviendas con presencia de garage.



# INSIGHTS

Se han seleccionado un muestreo de análisis gráfico, de tantos otros que podemos realizar.

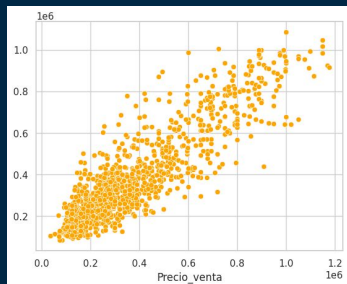
De las distintas variables analizadas, la que evidencia de mayor manera una correlación fuerte con el precio de venta de los inmuebles es el metraje construido.

# APLICACIÓN de ALGORITMOS DE ML

Se desarrollan 5 modelos de predicción con 3 algoritmos diferentes, para comparar y buscar el mayor desempeño:

1. Random Forest Regressor (variables sin transformación)
2. XG- BOOST Regressor (Variables sin transformación|PCA|Robust Scaler)
3. Red neuronal Keras (variables sin transformación)

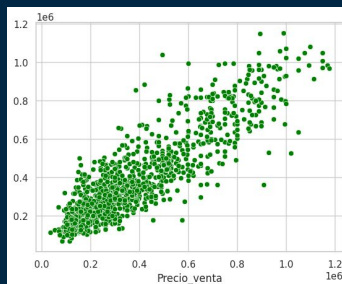
# Comparación de resultados



RANDOM FOREST  
REGRESSOR

Variables  
origen

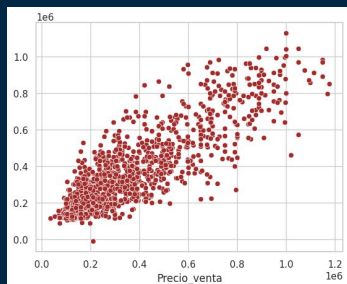
MSE: 11428889589.062  
R2: 0.79725106640864  
MAE 77879.7968530193



XG - BOOST

Variables  
origen

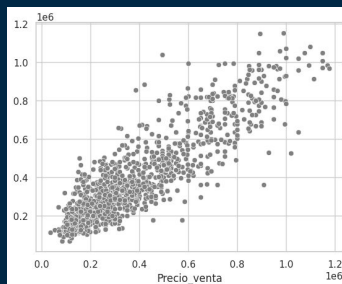
MSE: 11623555839.3  
r2: 0.793797679765  
MAE 77611.48008125



XG - BOOST

PCA

MSE: 14891970431.65  
r2: 0.7358158812748  
MAE 88447.440907812



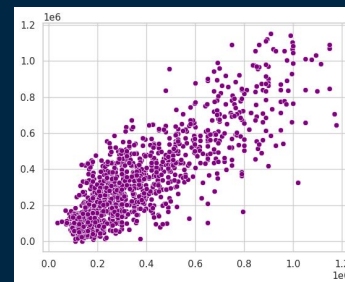
XG - BOOST

ROBUST SCALER

MSE: 11623555839.3  
r2: 0.793797679765  
MAE 77611.48008125

RED NEURONAL KERAS

Variables  
origen



MSE: 19111274057.1347  
R^2: 0.6610  
MAE: 105705.9517

# Mejorando el Algoritmo



Tomo los 2 mejores desempeños para tuneo de hiper parámetros para medir el  $r^2$ :

## 1.Random Forest Regressor

### GRID SEARCH

Mejor  $r^2$  en el conjunto de prueba

0.8045962691857891

### HALVING RANDOMIZED SEARCH

Mejor  $r^2$  en el conjunto de prueba

0.7979057995307111

## 2.XG Boost

### HALVING GRID SEARCH

Mejor  $r^2$  en el conjunto de prueba

$R^2$ : 0.8033

**Al modificar los hiper parámetros conseguimos una leve mejora en el desempeño del modelo.**



# INSIGHT FINAL

Dada la información inicial, con el objetivo de desarrollar un algoritmo eficaz para predecir el precio de nuevos inmuebles en el mercado, podemos concluir que el rendimiento del algoritmo obtenido es de nivel medio a medio-alto, aunque no alcanza una precisión excelente. Este algoritmo podría funcionar como una herramienta de apoyo, permitiendo evitar la necesidad de realizar un porcentaje significativo de tasaciones, dejando el restante para ser evaluado por los agentes inmobiliarios.



# FIN

Alumno Joaquín Gorostiza - Profesor Arturo Tapias - Tutor Estefania Susanj - CoderHouse 2023