



# TRABAJO FINAL DATA SCIENCE

MATIAS IGNACIO TOTH

# INDICE

1. Presentación del dataset y objetivos de la investigación
2. Análisis explotario del Dataset
3. Creacion y evaluacion del modelo
4. Conclusiones



## 1 - DATA SET Y OBJETIVOS DEL ANÁLISIS

# DATASET

- Este data set corresponde a Craigslist.org, el conocido sitio de anuncios clasificados.
- En el se encuentran 265 mil datos de viviendas en renta de estados unidos
- Cada publicación incluye la siguiente información (ver imagen).

## Diccionario de datos

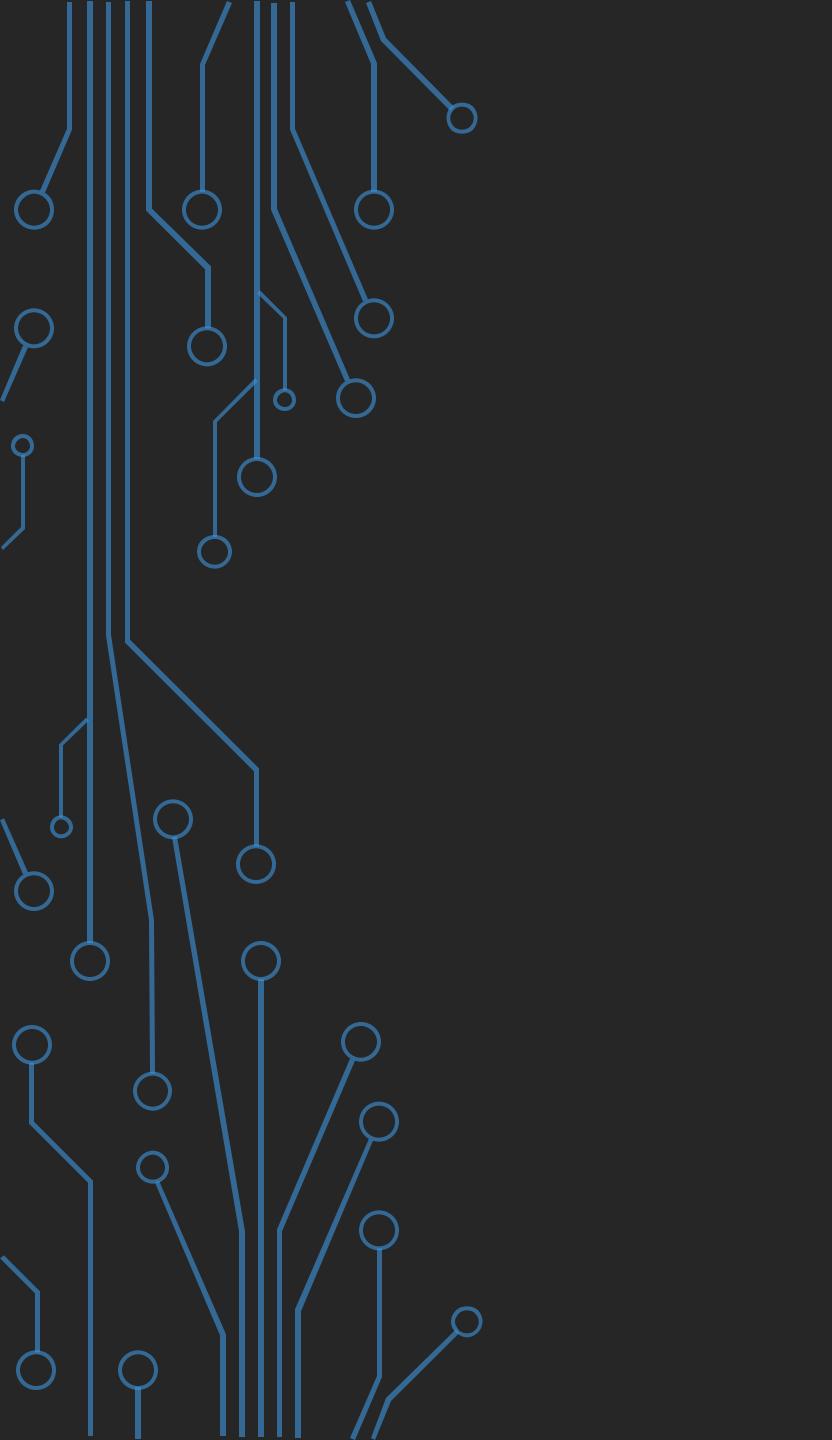
Id: listing id  
url: listing URL  
region: craigslist region  
region\_url: region URL  
price: rent per month (Target Column)  
type: housing type  
sqfeet: total square footage  
beds: number of beds  
baths: number of bathrooms  
cats\_allowed: cats allowed boolean (1 = yes, 0 = no)  
dogs\_allowed: dogs allowed boolean  
smoking\_allowed: smoking allowed boolean  
wheelchair\_access: has wheelchair access boolean  
electric\_vehicle\_charge: has electric vehicle charger boolean  
comes\_furnished: comes with furniture boolean  
laundry\_options: laundry options available  
parking\_options: parking options available  
image\_url: image URL  
description: description by poster  
lat: latitude  
long: longitude  
state: state of listing

# OBJETIVOS

- Son muchos los casos de inquilino que, se entera luego de alquilar, están pagando más por las características de ese espacio.
- La idea de este proyecto, es poderle permitir a los inquilinos, a la hora de buscar departamentos, tener una referencia de precio para tener como herramienta adicional a la hora de decidir sobre el alquiler de un determinado espacio.
- Me centrar en el análisis de la variable a predecir, PRICE.

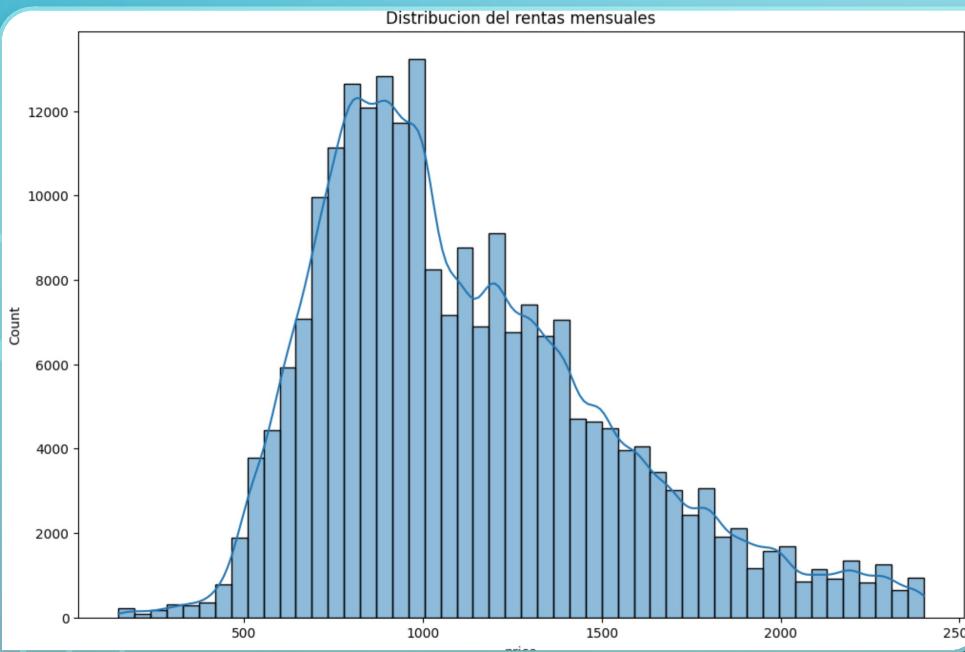
# HIPÓTESIS DE TRABAJO

? Podremos generar un modelo de regresión lo suficientemente confiable como para ayuda a la gente en busca de des paramentos ?



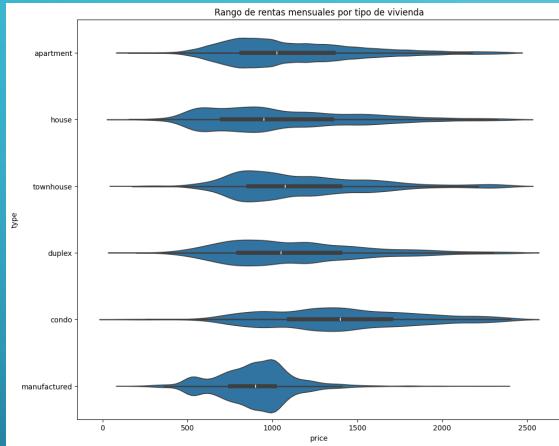
## 2 - ANÁLISIS EXPLORATORIO DEL DATA SET

# DISTRIBUCIÓN DE LA VARIABLE OBJETIVO

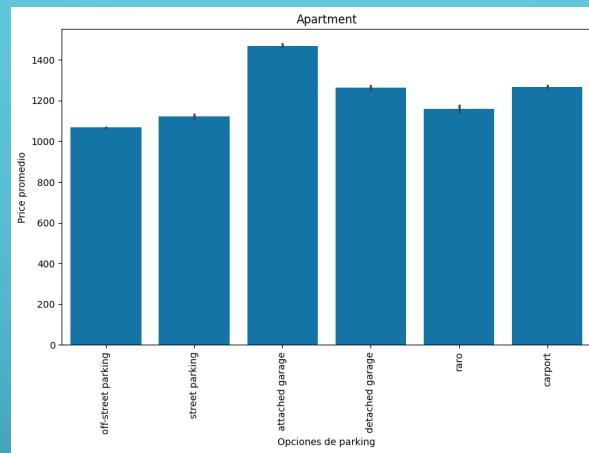


- Precio de las vivienda en alquiler oscilan desde los 150 USD a los 2400 USD
- El precio promedio de una vivienda en alquiler es de 1122 USD

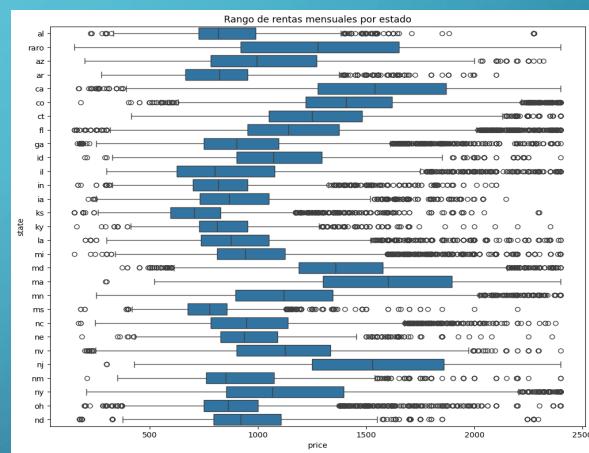
# RELACIÓN DE LA VARIABLE TARGET CON VARIABLES CATEGÓRICAS



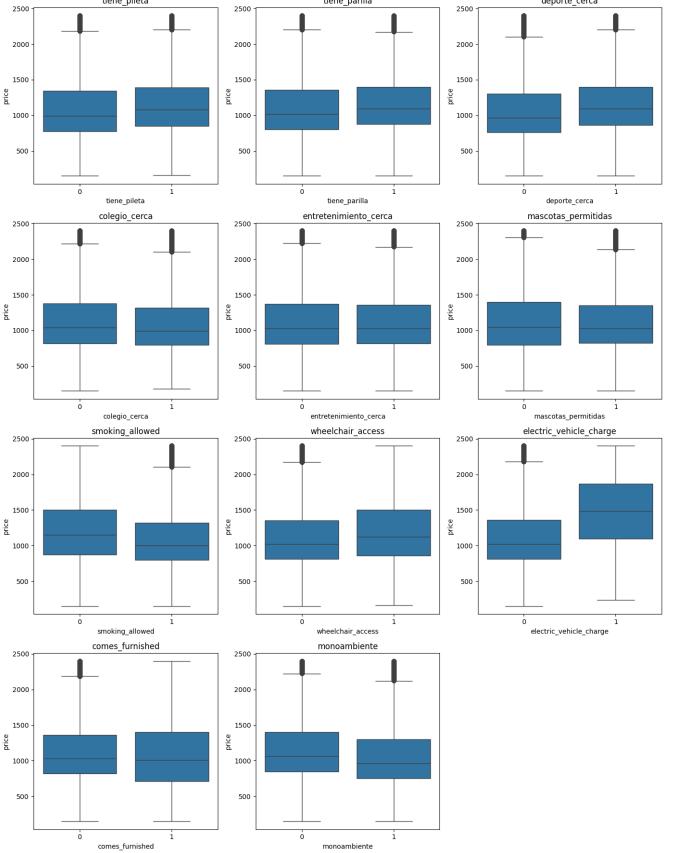
Por la forma del violín de 'manufactured', podemos ver que los precios suelen ser mas baratos que el resto



Podemos ver que las opciones de departamentos con garage son mas caras, que el resto de opciones



Podemos ver que los precios medio de una unida varí entre los distintos estados



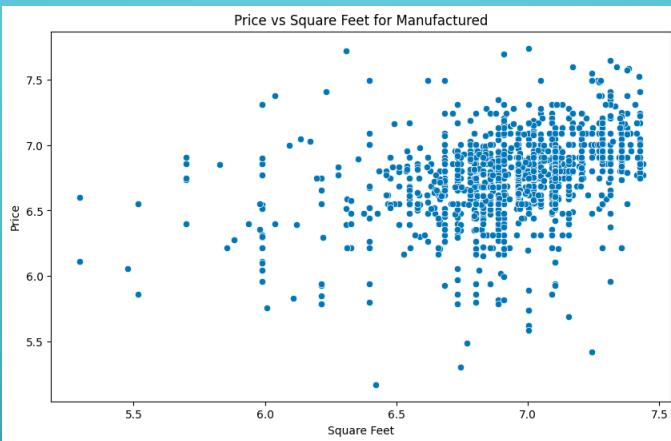
- Las viviendas con comodidades como pileta, parrilla y gym\_cerca, suelen ser mas caras en promedio
- Los espacios con accesos de silla de rueda, también tienden a ser mas caros
- Lo mismo sucede con las viviendas que cuentan con cargador de auto eléctricos
- Los departamento que tiene permitido fumar suelen ser mas baratos
- Los monoambiente son en promedio mas baratos

Cual es la relación entre el precio y los pies cuadrados?

Y como se comporta entre los distintos tipos de viviendas?



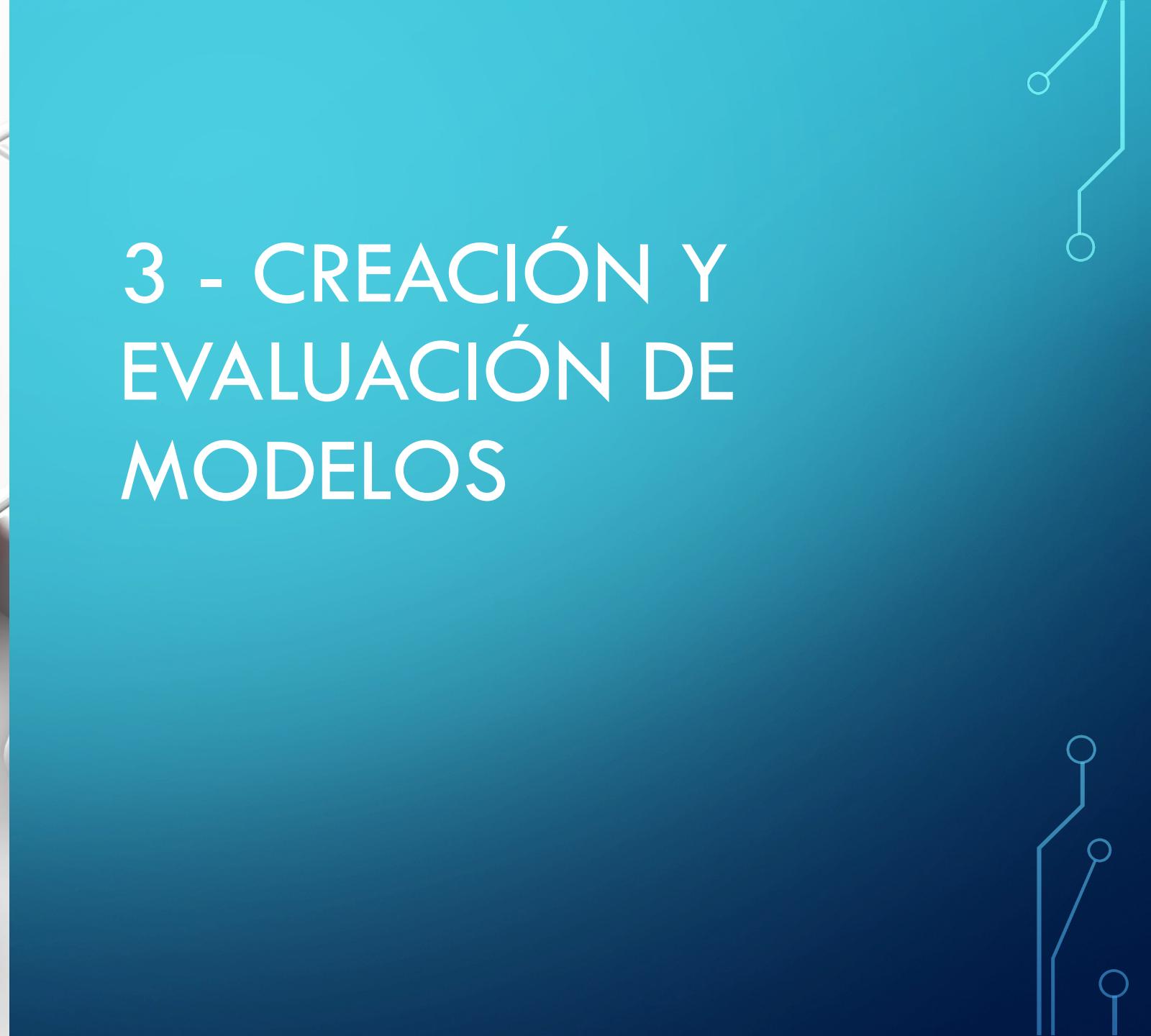
En este caso, no parece haber una relación muy lineal entre pies cuadrados y el precio de una vivienda



Se puede ver que para cada tipo de vivienda llega un nivel de pies, donde aparecen todo tipo de precios



# 3 - CREACIÓN Y EVALUACIÓN DE MODELOS



# RESUMEN

Seleccioné 3 algoritmos para la creación de mis modelos:

- Regresión Lineal
- Random Forest Regressor
- Xgboost Regressor

El data set fue separado en 80/20 para la creación de los sets de entrenamiento y testeo de los algoritmos

Las métricas de evaluación que utilice para medir el desempeño de estos fueron:

- MAE – es el error medio que tenemos por predicción
- RSME – es la diferencia promedio esperada entre nuestra predicción y el valor real
- R2 – mide la correlación entre nuestras predicciones y los valores reales

Selección y optimización del mejor algoritmo, utilizando Grid search y cross validation

**Performance Regresion Lineal:**  
MAE: 258.48265019485655  
RMSE: 348.11297576470514  
R2\_Score: 0.2910658721688594

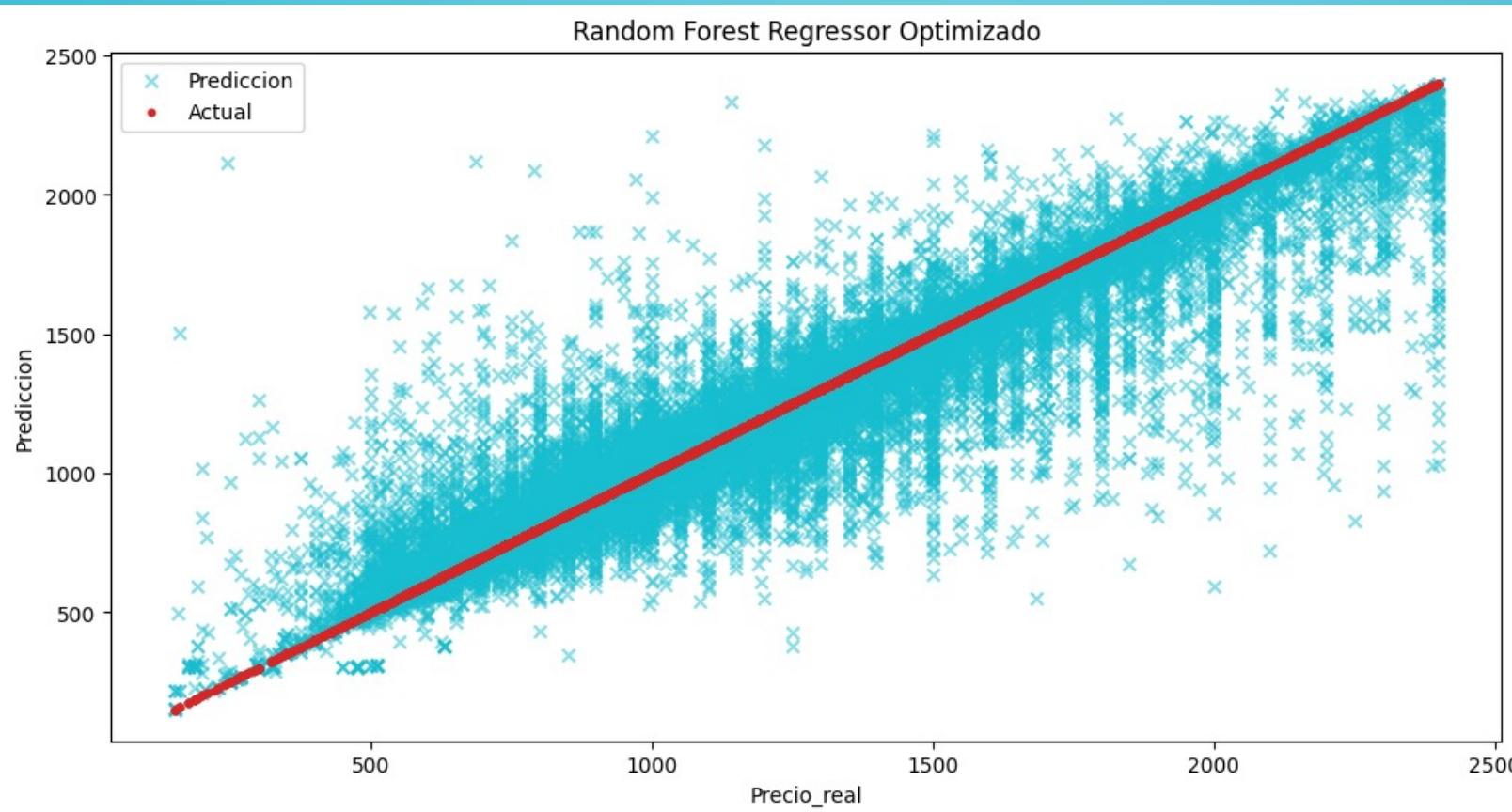
**Performance Xgboost Regressor:**  
MAE: 98.102183563086  
RMSE: 153.4978756491491  
R2\_Score: 0.8621616077180317

**Performance Random Forest Regressor:**  
MAE: 57.841119898520745  
RMSE: 124.9067376863561  
R2\_Score: 0.9087280789661641

TANTO RANDOM FOREST COMO XGBOOST, TUVIERON GRANDES RESULTADOS

# RESULTADOS

# MODELO OPTIMIZADO



Performance Random Forest Regressor optimizado:  
MAE: 57.68189915029307  
RMSE: 124.61401590801148  
R2\_Score: 0.9091553733340575

## 4 - CONCLUSIONES

## CONCLUSIONES

- A través del análisis exploratorio pudimos conocer el dataset para trabajar sobre este
- Pudimos entender e identificar la relación del precio y varía de las variables que tiene un efecto sobre este
- El modelo regresor adoptado de Random Forest, permite ayuda como precio de referencia a la hora de identificar una posible vivienda
- El modelo tiene un MAE de 57,68 un RMSE de 124 y un R2 de 90,92%