Modelo de Machine Learning Supervisado de Regresión para la predicción de precio de propiedades de la Ciudad Autónoma de Buenos Aires .

***Abstract*—**

La Ciudad Autónoma de Buenos Aires se encuentra en el hemisferio sur de América Sur, es la capital de la República Argentina. Conformada por 48 barrios. Se pueden encontrar diferentes tipos de inmuebles residenciales como departamentos, casas, PH, Pétit Hôtel, entre otras. Muchas de las opciones para vivir en la ciudad son por medio de un alquiler o compra de la vivienda. El precio de la vivienda puede variar dependiendo en que zona deseemos vivir.

El principal objetivo de este trabajo es entender cómo afectan el precio de la propiedad, las diferentes variables, tales como factores habitacionales, barrios, entre otros. Para ello, se tratará de predecir el precio acorde a dichos factores, empleando herramientas de Inteligencia Artificial y Machine Learning.

The Autonomous City of Buenos Aires is located in the southern hemisphere of South America, it is the capital of the Argentine Republic. Made up of 48 neighborhoods. You can find different types of residential properties such as apartments, houses, PH, Pétit Hôtel, among others. Many of the options for living in the city are by renting or buying a home. The price of housing can vary depending on which area we want to live.

The main objective of this work is to understand how the price of the property is affected by the different variables, such as housing factors, neighborhoods, among others. For this, it will try to predict the price according to these factors, using Artificial Intelligence and Machine Learning tools.

I. Introducción

La Ciudad Autónoma de Buenos Aires se encuentra en el hemisferio sur de América Sur, es la capital y ciudad más poblada de la República Argentina. Está situada en la región centro-este del país, sobre la orilla sur del Río de la Plata, en la región pampeana. La población de la ciudad es de 3 075 646 habitantes.

Esta metrópolis está conformada aproximadamente por 48 barrios. Los barrios del noreste son los de mayor poder adquisitivo, con tiendas exclusivas y varias áreas residenciales de la clase alta como Recoleta, Retiro, Palermo, Belgrano, Núñez, Las Cañitas, Colegiales así como también Puerto Madero, al este de la ciudad. A excepción del barrio de Barracas, en el que emerge una población de clase media y media alta gracias al auge inmobiliario, la zona sur es la que ostenta los menores indicadores socio-económicos de la ciudad

En la ciudad se pueden apreciar diferentes tipos de inmuebles residenciales como departamentos, casas, PH, Pétit Hôtel, entre otras. Muchas de las opciones para vivir en la ciudad son por medio de un alquiler o compra de la vivienda. El precio de la vivienda puede variar dependiendo en que zona deseemos vivir.

Como se mencionó anteriormente, el principal objetivo de este trabajo es entender cómo afectan el precio de la propiedad, diferentes variables, tales como factores habitacionales, barrios, entre otros. Para ello, se tratará de predecir el precio acorde a dichos factores, empleando herramientas de Inteligencia Artificial y Machine Learning.

II. Análisis exploratorio de datos (EDA)

Para comenzar a procesar la información, se partió de un data set. Para dicho procesamiento, primero se debió llevar a cabo una limpieza de los mismos, extrayendo información de menor relevancia o información nula/errónea. Entre las herramientas de limpieza de datos, se han aplicado algoritmos de feature extraction, para intentar que los algoritmos de Machine Learning(ML) puedan llegar a aprender mejor de los datos que se brindan.

Una vez filtrado el set de datos se dispuso de los datos referidos a las características de las propiedades en la Ciudad Autónoma de Buenos Aires y mediante herramientas de visualización de datos, pudieron observarse diferentes variables que pudieron afectar a la evaluación de propiedades.

En la figura a continuación (Figura 1) se puede obtener una primera visión del target del precio. Se puede observar a continuación, segregado por barrios, los rango de precios. Se pueden observar casos como Recoleta, Puerto Madero, Las Cañitas, Retiro y Belgrano donde el promedio de precio es mayor que el resto.

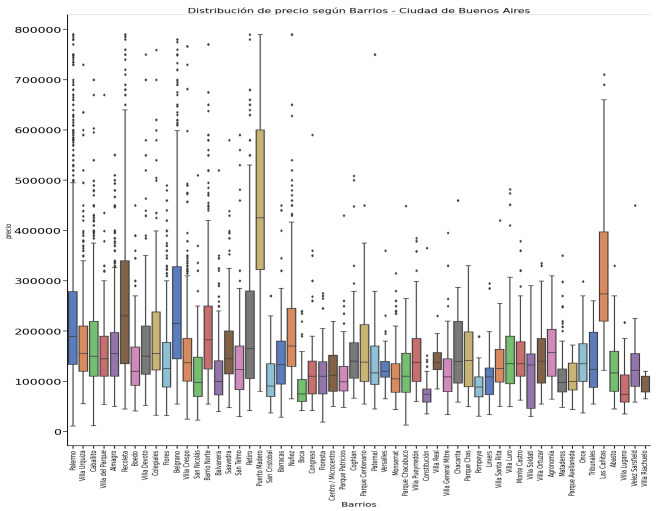


Figura 1 - Distribución de precios de propiedades en la Ciudad de Buenos Aires, segregado por barrios.

Para continuar con el análisis de datos, se procedió a la búsqueda de aglomeraciones de propiedades, por tanto, se llevó a cabo un análisis de la densidad en los distintos barrios de la ciudad.En la siguiente figura se muestra la distribución de las propiedades en venta de CABA (en habitantes por superficie), con el objeto de encontrar los barrios con mayor densidad.

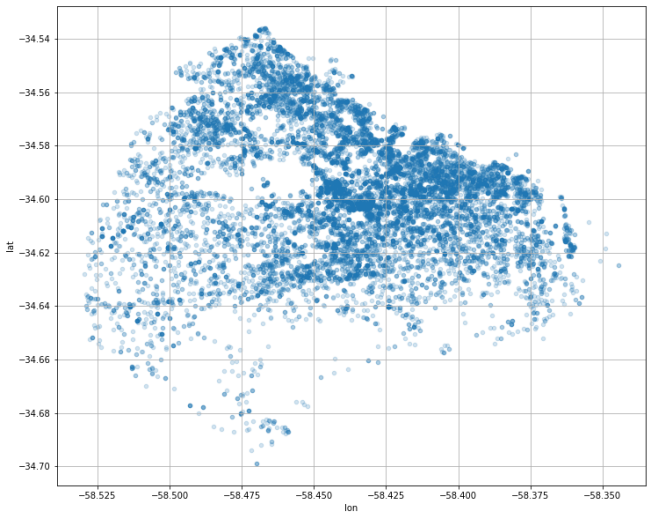


Figura 2 - Densidad propiedades en venta en la Ciudad de Buenos Aires, discriminado por barrios [hab/km2].

Para intentar comparar a través de una visualización geográfica, se realizó un análisis de la cantidad de propiedades a la venta en la ciudad de Buenos Aires también segregando por barrios.

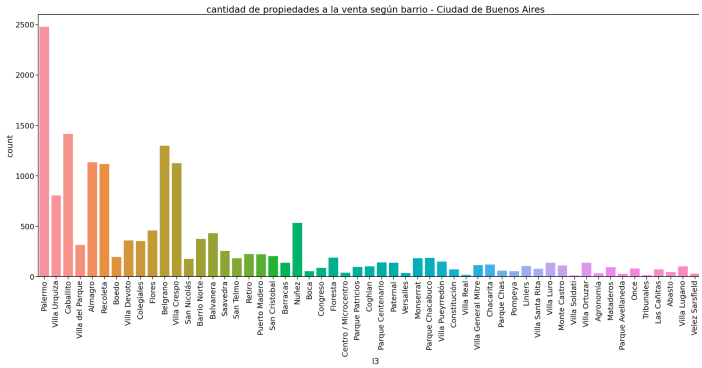


Figura 3 -Cantidad de propiedades a la venta en la Ciudad de Buenos Aires, discriminado por barrios.

En la siguiente figura, se puede observar cierta correlación entre los barrios y su precio.

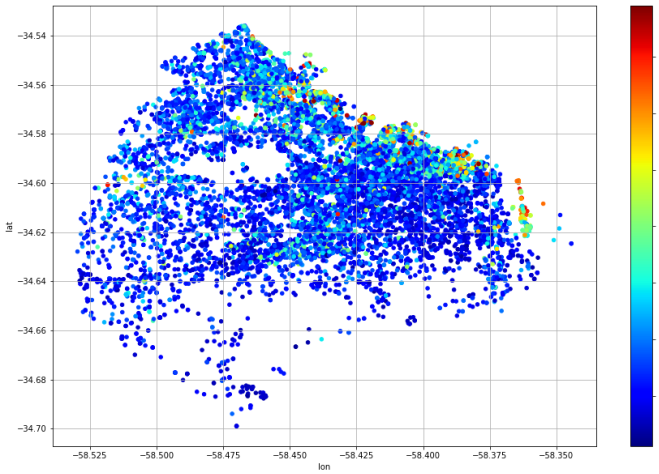


Figura 4 -Cantidad de propiedades a la venta en la Ciudad de Buenos Aires, relación con el precio

Con los mapas se puede llegar a intuir dónde se encuentran los lugares con mayor precios.

III.DATASET

El DataSet empleado para el armado del modelo que verifique la hipótesis son los que se detallan a continuación:

- Dataset con publicaciones de propiedad de Capital Federal. Cuenta con un total de 38.656 registros y 26 variables con distintas características de las propiedades.

IV.MÉTODOS

Con el objetivo de predecir los precios de las propiedades, se decidió utilizar un modelo de aprendizaje supervisado con algoritmos de regresión. Un modelo de regresión es un modelo matemático que busca determinar la relación entre una variable dependiente (Y), con respecto a otras variables

de entrada (features), llamadas explicativas o independientes (X). La regresión lineal múltiple permite generar un modelo lineal en el que el valor de la variable dependiente o respuesta (Y) se determina a partir de un conjunto de variables independientes llamadas predictores (X1, X2, X3…). Es una extensión de la regresión lineal simple, por lo que es fundamental comprender esta última. Los modelos de regresión múltiple pueden emplearse para predecir el valor de la variable dependiente o para evaluar la influencia que tienen los predictores sobre ella (esto último se debe analizar con cautela para no malinterpretar causa-efecto).

Los modelos lineales múltiples siguen la siguiente ecuación:

$$Yi=(\beta 0+\beta 1X1i+\beta 2X2i+\cdots+\beta nXni)+ei$$

$f(x) = y \quad y \in \Re$ Pilar de los algoritmo de regresión

- β0: es la ordenada en el origen, el valor de la variable dependiente Y cuando todos los predictores son cero.
- βi: es el efecto promedio que tiene el incremento en una unidad de la variable predictora Xi sobre la variable dependiente Y, manteniéndose constantes el resto de variables. Se conocen como coeficientes parciales de regresión.
- ei: es el residuo o error, la diferencia entre el valor observado y el estimado por el modelo.

El residuo es un elemento fundamental para la medición de error del sistema, o dicho en otras palabras para la eficiencia del modelo. El objetivo es intentar minimizar en un proceso iterativo las métricas de error para poder realizar predicciones más precisas. Algunas de esas métricas son el error cuadrático medio(MSE), raíz cuadrada del error cuadrático medio(RMSE) , media del error(MAE) y el coeficiente de determinación R^2.

$$R^2 = \frac{TSS - RSS}{TSS} \qquad MSE = \frac{\Sigma(\bar{Y}-Y)^2}{n} \qquad RMSE = \sqrt{\frac{\Sigma(\bar{Y}-Y)^2}{n}}$$

$$MAE = \frac{|\Sigma(\bar{Y}-Y)^2|}{n}$$

Para obtener la función de regresión se determinaron como parámetros.

y(target) = Precio (predicción)

En cuanto a las features de entrada (X) se emplearon las siguientes dentro del modelo de regresión.

- Latitud
- Longitud
- Rooms
- Bathrooms
- Surface_total.
- Surface_covered
- L3.
- Property_type.

Todas las features mencionadas anteriormente son numéricas con excepción de L3 y property_type, por lo cual, mediante herramientas de Feature Engineering, se transformaron las variables categóricas en variables binarias.

Los algoritmos de regresión que se utilizaron son del tipo supervisado y se mencionan a continuación:

- Support Vector Regression
- KNN Regression
- Random Forest Regression

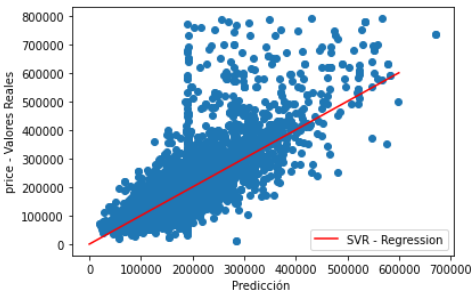
Una vez definidos los modelos, se escalaron los datos empleando un StandardScaler, para que los datos del dataset a emplear respeten una distribución standard. Una vez definidos los modelos que se van a utilizar para entrenar y testear el dataset, se realizó un split entre datos de entrenamiento y datos de testeo determinando:

- Datos de entrenamientos = 70%
- Datos de testeo= 30%

Mediante herramientas de ScikitLearn, tales como GridSearch y CrossValidation, se determinó la mejor combinación de hiperparámetros para cada modelo para así obtener un mayor rendimiento del modelo.

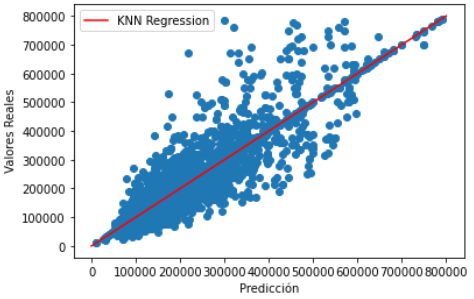
V. Resultados y conclusiones

VI. Support Vector Regression



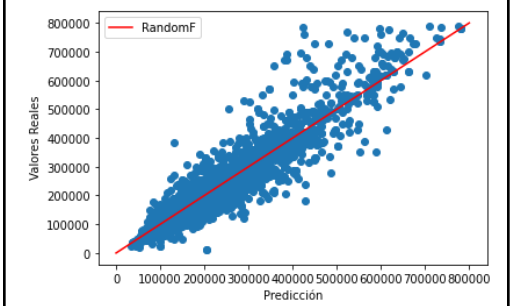

```
R2 score: 0.611596
MSE: 5.690335e+09
MAE: 47084.580814
```

VII. KNN Regression




```
R2 score: 0.836001
MAE: 2.402671e+09
MSE: 24786.633604
```

VIII. Random Forest Regression




```
R2 score: 0.900804
MAE: 1.453273e+09
MSE: 21667.800567
```

	Model	R2	MSE	MAE
1	SVR	0.611	5.690335e+09	47084.58
2	KNN	0.836	2.402671e+09	24786.63
3	RF	0.900	1.453273e+09	21667.80

Figura 6 - Resultados obtenidos por los diferentes modelos de Regresión.

Los modelos de KNN y Random Forest Regression presentaron un rendimiento aceptable, alcanzando una precisión de más del 83%. El modelo Support Vector Regression no presentó resultados aceptables hasta el momento. Empleando estos modelos, podría predecir con una considerable precisión el precio de una propiedad en la ciudad de Buenos Aires.

Referencias

[1] Material extraído de la asignatura “Ciencia de Datos”, correspondiente al plan de carrera de Ingeniería Industrial en Universidad Tecnológica Nacional (UTN) Facultad Regional Buenos Aires, año 2022.

[2] Material oficial de la Ciudad de Buenos Aires. https://www.buenosaires.gob.ar/laciudad/

https://www.buenosaires.gob.ar/jefaturadegabinete/desarrollo-urbano/primera-parte-el-analisis

[3] Liu, H., & Motoda, H. (Eds.). (1998). Feature extraction, construction and selection: A data mining perspective (Vol. 453). Springer Science & Business Media., pp.206–207.

[4] Geron, Aurélien (2017). Hands-on Machine Learning with Scikit Learn, Keras and tensorflow. Pp 110-130.