Machine Learning en el Mercado Inmobiliario

Benjamín Sáez Antil
Universidad del Bío-Bío, Avda. Collao 1202, Concepción, Chile
patriciosaez0729@gmail.com

1. Introducción

El mercado inmobiliario es fundamental para la economía, especialmente en ciudades como Santiago de Chile, donde la demanda de propiedades sigue en aumento. Valorar inmuebles con precisión es un reto clave para compradores e inversionistas (Cifuentes et al., 2020).

El Machine Learning (en adelante ML) corresponde a una rama de la Inteligencia Artificial, que utiliza datos y algoritmos para simular el aprendizaje humano y mejorar su precisión de forma progresiva (IBM, s.f.). Esta capacidad permite analizar grandes volúmenes de datos, detectar patrones complejos y logra optimizar la toma de decisiones.

Este estudio utiliza una base de datos de mayo del 2020, con un tamaño de 1139 datos sobre propiedades usadas en Chile, los cuales incluyen características como comuna, tipo de vivienda y precios en UF y CLP, entre otros. Dichos datos permiten comparar variados modelos de ML. Donde los datos se obtuvieron de la página web Kaggle (Gonzalez, 2020).

2. Objetivos

Objetivo General
Desarrollar un modelo para predecir precios inmobiliarios en la región metropolitana, por medio de la utilización de técnicas de machine learning, con el fin de mejorar la comprensión del comportamiento del mercado.

Objetivos específicos

1. Realizar un análisis exploratorio del conjunto de datos obtenidos.
2. Usar técnicas de machine learning para evaluar y comparar la efectividad de distintos algorimtos mediante el uso de la herramienta PyCaret.
3. Seleccionar el modelo que más se adapte a los datos para predecir los precios de las propiedades, mientras se utilizan las características de las viviendas como variables principales.
4. Identificar y analizar las variables más relevantes que influyen en el valor de las propiedades, para comprender mejor qué factores tienen mayor impacto en la determinación de los precios.

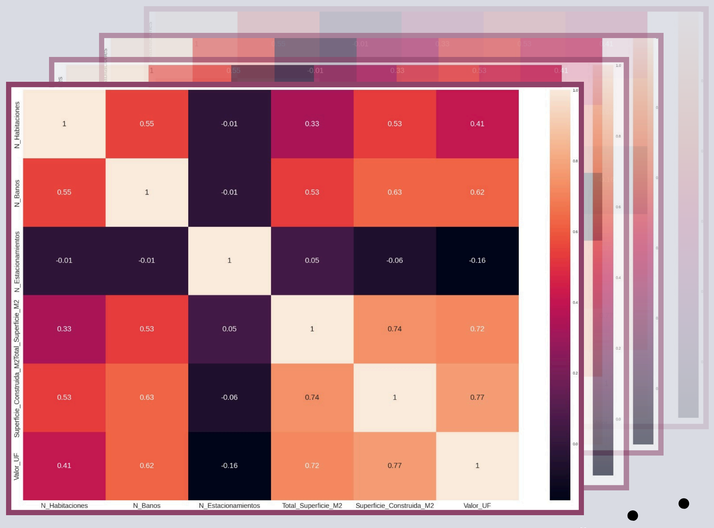
3. Metodología

1.Tratamiento y limpieza de datos:
- a. Recopilación de datos.
- b. Eliminación de columnas irrelevantes.
- c. Adicionar columna relevante llamada "Venta por dueño".

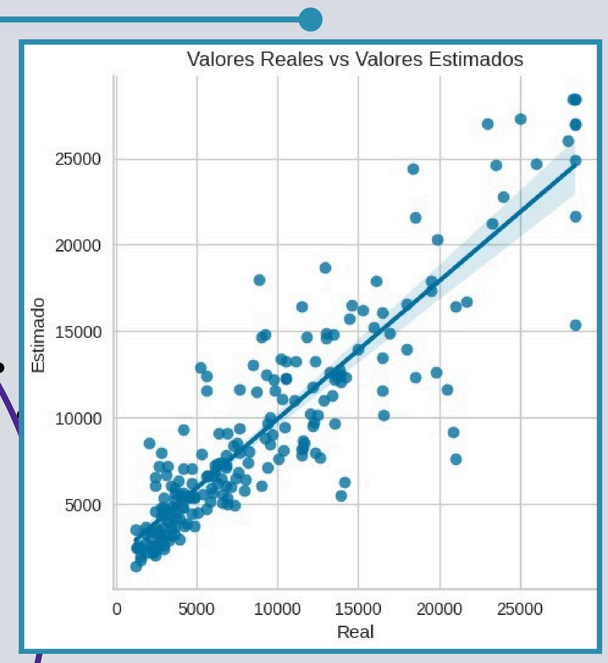
2.Análisis Exploratorio de Datos (EDA):
- a. Análisis descriptivo.
- b. Ajuste de tipos de variables.
- c. Detección y tratamiento de datos ausentes.
- d. Identificación de datos atípicos.
- e. Correlación de variables.

3.Comparación de modelos predictivos:
- a. Comparación de distintos algoritmos, mediante la librería PyCaret de Python.
- b. Utilización de métricas de evaluación para determinar el modelo más óptimo.

4.Construcción modelo predictivo:
Según los resultados del item 3, se selecciona el modelo Random Forest, continuando con los siguientes pasos:
- a. Creación de subconjuntos de datos - con selección de 80% para entrenamiento y 20% para prueba.
- b. Entrenamiento de cada árbol.
- c. Análisis de sobreajuste.
- d. Predicción.



	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lr	Linear Regression	2085.9277	9662837.7248	3069.6552	0.8390	0.4226	0.2980	0.0690
rf	Random Forest Regressor	2070.8564	11216423.0486	3326.0907	0.8103	0.3250	0.2635	0.4450
mlp	MLP Regressor	3246.8087	20069614.9539	4459.7171	0.6596	0.5310	0.4570	1.6640
dt	Decision Tree Regressor	2721.9054	20933501.6554	4529.2992	0.6424	0.4394	0.3413	0.7030
knn	K Neighbors Regressor	3241.6141	21555227.8408	4611.0718	0.6342	0.4809	0.4424	0.0600

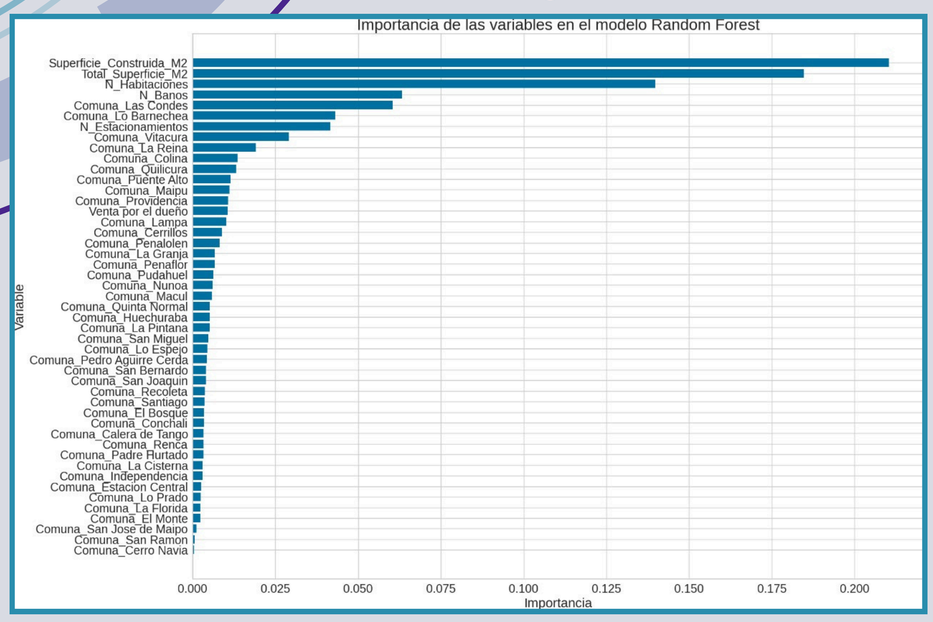


4. Resultados y Discusión

El análisis de correlación muestra que las variables más importantes para predecir el valor de las propiedades son la superficie construida, la superficie total, el número de baños y el número de habitaciones. Estas características tienen una relación directa con el precio en UF, lo que justifica su relevancia en los modelos de predicción, y se considera que el modelo Random Forest uno de los más efectivo (Bozanic, 2020). En cambio, factores como los estacionamientos o si la propiedad es vendida directamente por el dueño tienen una influencia menor.

Este patrón también se refleja en el gráfico de dispersión, donde los valores estimados por el modelo Random Forest se ajustan bastante bien a los valores reales, con la mayoría de los puntos cercanos a la línea de identidad. Sin embargo, el modelo tiende a desviarse un poco en propiedades más caras, lo que podría estar relacionado con las variables que tienen menor peso en la predicción. A pesar de estos casos puntuales, el modelo tiene un buen desempeño, con un R^2 de 80.4%, lo que indica que predice correctamente la mayoría de los precios.

Por último, el gráfico de barras que muestra la importancia de las variables confirma que las características más influyentes son la superficie construida y la superficie total, seguidas por la comuna de la propiedad, donde lugares como Las Condes y Lo Barnechea destacan por su mayor demanda. Aunque otras variables como el número de baños y habitaciones también son relevantes, factores como los estacionamientos o comunas menos cotizadas tienen un impacto mucho menor en el valor final de las propiedades.



5. Conclusión

- El modelo Random Forest demostró su capacidad para capturar la complejidad de los datos y reflejar relaciones no lineales entre las variables.
- La diferencia entre el R^2 ajustado (74.6%) y el valor de R^2, junto con el 19.6% de variabilidad no explicada, sugieren limitaciones en la generalización del modelo, posiblemente por sobreajuste o la falta de factores considerados.
- Este estudio proporcionó conocimientos valiosos sobre la dinámica del mercado inmobiliario y la eficacia de distintas técnicas de modelado.
- La preferencia por Random Forest frente a la Regresión Lineal destaca la importancia de usar modelos que manejen relaciones complejas entre variables.
- Se recomienda, para estudios futuros, ampliar el conjunto de datos mediante Web Scraping y considerar el uso de datos más actuales, además de evaluar el modelo en datos independientes y explorar nuevas variables para mejorar la precisión en la predicción del valor de las propiedades.

Referencias

- Cifuentes, A., Cid, B., & Gallardo, F. (2020). Determinantes de los precios en el mercado inmobiliario sobre la base de un índice de cualidades de la vivienda. Revista Chilena de Economía y Sociedad, 22
- Ibm. (s.f.). ¿Qué es Machine learning (ML)? Explicaciones. Recuperado 25 de septiembre de 2024, de https://www.ibm.com/mx-es/topics/machine-learning
- Gonzalez, G. (2020). Valor casas usadas Región Metropolitana Chile. Kaggle. Recuperado 17 de agosto de 2024, de https://www.kaggle.com/datasets/gorkigonzalez/casas-usadas-rm-chile-mayo-2020
- Bozanic Leal, M. S. (2020). Sistema de predicción de precios-venta de inmuebles en el mercado del sector inmobiliario de la Región Metropolitana de Chile con el uso de algoritmos de Machine Learning (Memoria de título, Universidad de Chile). Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas, Departamento de Ingeniería Industrial