UNIVERSIDAD PERUANA UNIÓN

ESCUELA DE POSGRADO

Unidad de Posgrado de Ingeniería y arquitectura

Predicción de precios de viviendas usando Machine Learning en Lima Metropolitana

Trabajo de Investigación para obtener el Grado Académico de Maestro en Ingeniería de Sistemas con Mención en Dirección y Gestión de Tecnologías de Información

Autor:

Mariela Victoria Terán Suárez
Mardeli Beatriz Panduro Del Castillo
Emelson Alex Chire Hernandez

Asesor:

Mg. Johann Alexis Ospina Galindez

Lima, Abril del 2025

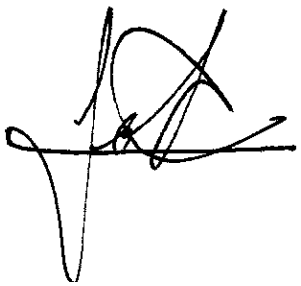
DECLARACIÓN JURADA DE ORIGINALIDAD DE TRABAJO DE INVESTIGACIÓN

Yo Johann Alexis Ospina Galindez, docente de la Unidad de Posgrado de ingeniería y arquitectura, Escuela de Posgrado de la Universidad Peruana Unión.

DECLARO:

Que la presente investigación titulada: **"PREDICCIÓN DE PRECIOS DE VIVIENDAS USANDO MACHINE LEARNING EN LIMA METROPOLITANA"** de los autores Mariela Victoria Terán Suárez, Mardeli Beatriz Panduro, Del Castillo Emelson Alex Chire Hernandez tiene un índice de similitud de 14 % verificable en el informe del programa Turnitin, y fue realizada en la Universidad Peruana Unión bajo mi dirección.

En tal sentido asumo la responsabilidad que corresponde ante cualquier falsedad u omisión de los documentos como de la información aportada, firmo la presente declaración en la ciudad de Lima, a los 03 días del mes de Setiembre del año 2025



_____

Johann Alexis Ospina Galindez

ACTA DE SUSTENTACIÓN DE TRABAJO DE INVESTIGACIÓN

En Lima, Ñaña, Villa unión a 22 días del mes de abril del año 2025, siendo las 09:00 horas, se reunieron de forma online sincrónica, bajo la dirección del presidente del jurado Mg. Geraldine Verónica Alvizuri Llerena, secretario Mg. Junior Israel Pacheco Espinoza; los demás miembros: Mg. Lizeth Geanina Huanca López, PhD. Javier Linkolk López Gonzales y el asesor Mg. Johann Alexis Ospina Galindez, con el propósito de administrar el acto académico de sustentación de Trabajo de Investigación de la Maestría titulada "Predicción de precios de viviendas usando machine learning", conducente a la obtención de la Maestría en Ingeniería de Sistemas con mención en Dirección y Gestión de Tecnologías de Información.

El presidente inició el acto académico de sustentación invitando a los candidatos a hacer uso del tiempo determinado para su exposición. Concluida la exposición, el presidente invitó a los demás miembros del jurado a efectuar las preguntas, cuestionamientos y aclaraciones pertinentes, aquellos que fueron absueltos por los candidatos. Luego, se produjo un receso para las deliberaciones y la emisión del dictaminador del jurado. Posteriormente, el jurado procedió a dejar constancia escrita sobre la evaluación en la presente acta, con el dictamen siguiente:

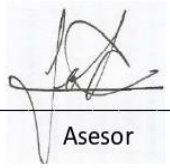
Candidatos: Mardeli Beatriz Panduro Del Castillo, Emelson Alex Chire Hernández y Mariela Victoria Terán Suarez

CALIFICACIÓN	ESCALAS			Mérito
	Vigesimal	Literal	Cualitativa	
Aprobado	19	A	Con nominación de excelente	Excelencia

Finalmente, el presidente del jurado invitó a los candidatos a ponerse de pie, para recibir la evaluación final. Además, el presidente del jurado concluyó el acto académico de sustentación, procediéndose a registrar las firmas respectivas.

Presidente

Secretario



Asesor

Miembro

Miembro

Candidato

Candidato

Candidato

Predicción de precios de viviendas usando machine learning en lima metropolitana

Mardeli Beatriz Panduro Del Castillo, Mariela Victoria Terán Suárez, Alex Chire Fernandez, Johann Alexis Ospina Galindez

Escuela de Posgrado, Universidad Peruana Unión, Lima -Peru.

Resumen

El presente estudio abordó el desarrollo de un modelo de Machine Learning para la estimación de precios de alquiler de viviendas en Lima Metropolitana, Perú, utilizando datos obtenidos mediante Web Scraping. Se integraron variables clave como ubicación, metraje, número de habitaciones y baños, y precio, aplicando técnicas de procesamiento, limpieza y estructuración de datos para garantizar su calidad antes del modelado predictivo. Se implementaron y compararon distintos algoritmos de Machine Learning, incluyendo Random Forest, Gradient Boosting, Support Vector Regression (SVR) y modelos de ensamble como Voting y Stacking Regressor. La evaluación de los modelos se llevó a cabo utilizando métricas como el Error Absoluto Medio (MAE), la Raíz del Error Cuadrático Medio (RMSE) y el Coeficiente de Determinación (R^2). Los resultados indicaron que el Voting Regressor obtuvo el mejor desempeño predictivo, superando a los modelos individuales y al Stacking Regressor. El análisis de los residuos reveló que, aunque el modelo capturó la tendencia general de los precios de alquiler, presentó mayores errores en valores elevados, sugiriendo la presencia de outliers y la necesidad de transformaciones en los datos. Además, la curva de aprendizaje evidenció que el modelo mejoró su precisión con más datos, aunque aún existía margen de optimización. En conclusión, el Voting Regressor se consolidó como la mejor alternativa predictiva, destacando la influencia de la ubicación y el metraje como variables determinantes en los precios de alquiler. Futuras investigaciones podrían enfocarse en el uso de XGBoost, redes neuronales y estrategias avanzadas de ajuste de hiperparámetros para mejorar la precisión del modelo y su aplicabilidad en el mercado inmobiliario.

***Palabras clave**: Machine Learning; Predicción de precios; Alquiler de viviendas*

Abstract

This study addressed the development of a Machine Learning model for the estimation of housing rental prices in Metropolitan Lima, Peru, using data obtained through Web Scraping. Key variables such as location, footage, number of bedrooms and bathrooms, and price were integrated, applying data processing, cleaning and structuring techniques to ensure data quality prior to predictive modeling. Different Machine Learning algorithms were implemented and compared, including Random Forest, Gradient Boosting, Support Vector Regression (SVR) and ensemble models such as Voting and Stacking Regressor. The evaluation of the models was carried out using metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Coefficient of Determination (R^2). The results indicated that the Voting Regressor obtained the best predictive performance, outperforming the individual models and the Stacking Regressor. The analysis of the residuals revealed that, although the model captured the general trend in rental prices, it presented greater errors at higher values, suggesting the presence of outliers and the need for data transformations. In addition, the learning curve showed that the model improved its accuracy with more data, although there was still room for optimization. In conclusion, the Voting Regressor was consolidated as the best predictive alternative, highlighting the influence of location and footage as determining variables in rental prices. Future research could focus on the use of XGBoost, neural networks and advanced hyperparameter fitting strategies to improve the accuracy of the model and its applicability in the real estate market.

***Keywords:** Machine Learning; Price prediction; Housing rentals*

1. INTRODUCCIÓN

Un avance que merece destacarse en Lima Metropolitana es la creciente necesidad de abordar los desafíos del mercado inmobiliario, específicamente en la evaluación precisa de precios de vivienda [1]. Esta problemática, exacerbada en los últimos años por el acelerado crecimiento urbano, la desigualdad socioeconómica y la carencia de herramientas modernas para predecir de manera confiable el valor de las propiedades, presenta un área crítica para la innovación [2].

Si bien aún existen grandes retos, la introducción de soluciones tecnológicas avanzadas tiene el potencial de transformar la dinámica del sector, reduciendo la incertidumbre que enfrentan tanto compradores como vendedores [3]. Este impacto no solo podría facilitar decisiones más informadas, sino también fortalecer uno de los pilares clave de la economía nacional, demostrando cómo las herramientas predictivas pueden revolucionar industrias enteras [4].

El mercado inmobiliario en Lima Metropolitana presenta una gran heterogeneidad y complejidad debido a factores como el crecimiento urbano desordenado, las brechas socioeconómicas y las fluctuaciones en la demanda [5]. Estas condiciones hacen que la estimación precisa de precios sea un desafío crítico, especialmente con enfoques tradicionales basados en regresión lineal o evaluaciones intuitivas, los cuales son incapaces de capturar relaciones no lineales y manejar grandes volúmenes de datos [6].

Desde la perspectiva del conocimiento, este desafío representa una valiosa oportunidad para validar enfoques avanzados basados en aprendizaje automático (Machine Learning).

Estas técnicas, ampliamente adoptadas en sectores como la salud, las finanzas y la logística, poseen un potencial transformador para abordar las limitaciones inherentes a los modelos tradicionales del mercado inmobiliario [7]. Algoritmos como los árboles de decisión, redes neuronales y máquinas de soporte vectorial tienen la capacidad de procesar grandes volúmenes de datos estructurados y no estructurados, integrando variables complejas como ubicación, características de las propiedades y tendencias económicas [8].

Esto permite generar predicciones significativamente más precisas, adaptadas a las particularidades del mercado local, y aporta una nueva dimensión al análisis inmobiliario, haciendo posible decisiones informadas y estrategias más efectivas para todos los actores del sector [9].

Estudios recientes han demostrado que algoritmos de Machine Learning como Gradient Boosting y Random Forest pueden abordar estas limitaciones al integrar datos heterogéneos y modelar interacciones complejas [10].

La necesidad de avanzar en los modelos de predicción de precios de viviendas es cada vez más evidente, especialmente ante las limitaciones de las metodologías tradicionales. Estas suelen depender de análisis estadísticos lineales y la experiencia intuitiva de expertos del mercado [11].

Aunque estas aproximaciones han sido útiles históricamente, no cuentan con la capacidad de procesar e integrar grandes volúmenes de datos provenientes de fuentes diversas. Factores como la ubicación geográfica, las características específicas de las propiedades, las fluctuaciones económicas y otros elementos contextuales que afectan de manera crucial los precios quedan subutilizados en estos enfoques [12].

La implementación de herramientas modernas y más sofisticadas tiene el potencial de superar estas restricciones, proporcionando un panorama más detallado y revelador que pueda transformar las dinámicas del mercado inmobiliario [13].

En Lima Metropolitana, las disparidades en los precios son significativas, con valores que varían desde S/. 500 en zonas periféricas hasta S/. 20,000 en distritos como San Isidro o Miraflores. Esta disparidad refleja no solo diferencias en las características físicas de las propiedades, sino también factores contextuales como la accesibilidad, los servicios disponibles y las condiciones socioeconómicas locales [14].

En este contexto, la aplicación de algoritmos de Machine Learning ofrece una oportunidad única

para proporcionar estimaciones más precisas y accionables, beneficiando a compradores, vendedores y entidades financieras [15].

Este estudio tiene como objetivo principal desarrollar un modelo predictivo basado en datos recolectados mediante Web Scraping, aplicando algoritmos de Machine Learning avanzados para optimizar la precisión de las estimaciones. Además, se busca identificar las variables con mayor impacto en los precios de las viviendas, contribuyendo al entendimiento del mercado inmobiliario en Lima Metropolitana.

2. Metodología

El enfoque metodológico de este estudio es cuantitativo, exploratorio y aplicado. La investigación se desarrolló en un entorno virtual utilizando herramientas de programación avanzadas para la extracción, limpieza y análisis de datos. La metodología se estructuró en tres fases principales:

2.1 Recopilación de datos

El proceso automático de obtener datos de sitios web, convertirlos a un formato estructurado y guardarlos para su posterior procesamiento o análisis se conoce como Web Scraping [16]. Los datos se recolectaron mediante esta técnica aplicadas a páginas web de anuncios inmobiliarios que publican precios de alquiler en la ciudad de Lima. Para este proceso, se utilizaron herramientas como Python con las librerías Beautiful Soup, Scrapy o Selenium. Los datos recolectados incluyeron las siguientes variables:

a) Precio
b) Ubicación geográfica
c) Área
d) Número de habitaciones
e) Baños

La cobertura geográfica incluyó distritos de diferentes niveles socioeconómicos para garantizar representatividad.

Se realizó la programación del código de Web Scraping en Python para extraer las características de las viviendas en alquiler, en función de características como la ubicación, tamaño, número de habitaciones, número de baños y precios.

Figura 1. Proceso del Web Scarping

2.2 Preprocesamiento y modelado

Una vez recolectados, los datos fueron procesados para eliminar registros duplicados, valores nulos y errores. Se emplearon herramientas como Pandas y Numpy para garantizar que los datos estuvieran en un formato adecuado para el análisis. También se aplicaron técnicas de normalización y codificación para variables categóricas, como el distrito o el estado del departamento. Se realizaron las siguientes operaciones:

a) Se eliminaron distritos que no pertenecen a Lima Metropolitana.
b) Se eliminaron departamentos con precios menores de 200 soles.
c) Se eliminaron departamentos con precios mayores de 20000 soles.
d) Se eliminaron departamentos con cuartos mayores a 10.
e) Se eliminaron departamentos con áreas mayores a 500 mt2.

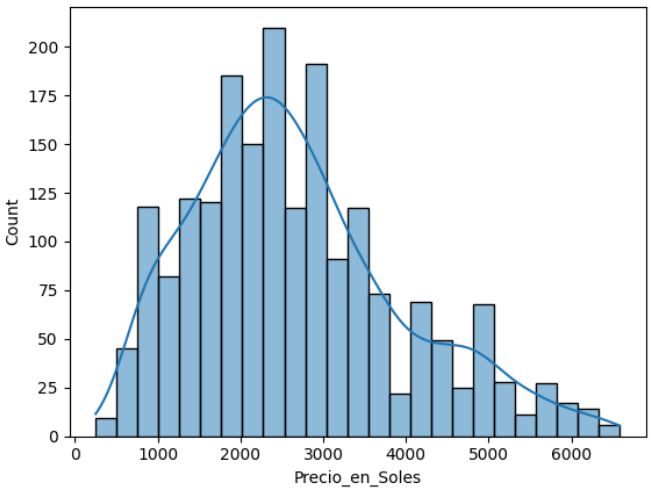


Figura 2. Datos normalizados cantidad - precio

2.2 Modelo de desarrollo

Se dividió el dataset en conjuntos de entrenamiento (80%) y prueba (20%). Por otro lado, se implementaron modelos como:

- Árbol de Decisión
- Bagging
- Random Forest
- AdaBoost
- SVR
- Extra Trees
- XGBoost
- Regresión Lineal
- Gradient Boosting

Además, se probaron modelos de ensamble, como:

- Voting Regressor: Combinación de Random Forest, Gradient Boosting y SVR.
- Stacking Regressor: Combinación jerárquica con un modelo final basado en Ridge Regression.

3. Resultados y discusión

Se evaluaron los modelos utilizando Error Cuadrático Medio (MSE) mide el error promedio de las predicciones y el Coeficiente de Determinación (R^2) Indica qué tan bien se ajusta el modelo a los datos (más cercano a 1 es mejor), todos los modelos se entrenaron con los mismos datos. Se obtuvieron los siguientes resultados:

Tabla 1. Resultados de los Modelos Iniciales

Modelo	MSE	R^2
Árbol de Decisión	1,031,106	0.3616
Bagging	821,414	0.4914
Random Forest	802,936	0.5029
AdaBoost	728,104	0.5492
SVR **(peor modelo)**	1,484,780	0.0807
Extra Trees	893,808	0.4466
XGBoost	906,739	0.4386
Regresión Lineal	755,519	0.5322
Gradient Boosting **(Mejor modelo)**	675,453	0.5818

Estos resultados reflejan que Gradient Boosting destacó como el mejor modelo en términos de MSE (Error Cuadrático Medio) mas bajo y R^2 mas alto, lo que significa que tuvo la mejor capacidad de predicción.

Una de las razones por la cual Gradiente Boosting fue mejor es que es un método de ensamble, lo que significa que combina múltiples modelos (árboles de decisión débiles) para crear un modelo fuerte y preciso [17]. A diferencia de Random Forest, Gradient Boosting construye cada árbol corrigiendo los errores del anterior, mejorando la precisión con cada iteración.

Los precios de renta pueden depender de factores no lineales como interacción entre variables, tendencias ocultas en los datos y/o impacto de valores extremos o atípicos. Gradient Boosting es muy bueno manejando relaciones complejas y no lineales en los datos, lo que le dio ventaja sobre modelos más simples como la Regresión Lineal o Árboles de Decisión [18].

Aunque los árboles de decisión pueden aprender muy bien los datos de entrenamiento, muchas veces sobrefit (aprenden demasiado los detalles de los datos y fallan en datos nuevos) [19]. Gradient Boosting controla el overfitting con:

a) Aprendizaje lento (learning rate): Cada nuevo árbol solo corrige un poco el error del anterior.

b) Número de árboles: Se puede ajustar para encontrar el punto óptimo entre precisión y generalización.
c) Profundidad de los árboles: Limita cuánto puede aprender cada árbol individual.

Por eso, Gradient Boosting es más robusto y generaliza mejor en nuevos datos.

3.1 Evaluación de Modelos

Se mejoraron los parámetros de los modelos de Random Forest, Gradient Boosting y SVR con GridSearchCV (los valores posibles para cada hiperparámetro fueron número de árboles en el bosque (100, 200 o 300), profundidad máxima del árbol (10, 20 o ilimitado None), número mínimo de muestras requeridas para dividir un nodo (2 o 5 o 10) y número mínimo de muestras en cada hoja del árbol (1, 2 o 4).).

3.2 StackingRegressor

Es un modelo de ensamble en scikit-learn que combina varios modelos de regresión, pero en lugar de simplemente promediarlos (como VotingRegressor), usa otro modelo adicional (meta-modelo) para aprender la mejor combinación de predicciones [20]. Se obtuvieron los siguientes resultados:

Tabla 2. Resultados del StackingRegressor

Modelo	MSE	R^2
StackingRegressor	675754.2141458034	0.5816039781571894

El modelo de Stacking Regressor, compuesto por Random Forest, Gradient Boosting y SVR, fue entrenado y evaluado utilizando un conjunto de datos de renta de apartamentos. Tras su entrenamiento, el modelo obtuvo un Error Cuadrático Medio (MSE) de 675,754, lo que indicó que, en promedio, las predicciones del modelo presentaron una desviación considerable respecto a los valores reales de renta. Además, el coeficiente de determinación (R^2) fue de 0.5816, lo que significó que el modelo explicó aproximadamente el 58.16% de la variabilidad en los precios de alquiler. Aunque el resultado mostró una capacidad aceptable de predicción, la presencia de un MSE relativamente alto sugirió que el modelo aún podía mejorarse mediante un ajuste más preciso de los hiperparámetros o la incorporación de nuevas características relevantes en el conjunto de datos.

El modelo Stacking Regressor fue optimizado mediante el ajuste del hiperparámetro alpha del meta-modelo Ridge, utilizando GridSearchCV con validación cruzada de cinco pliegues. Tras la optimización, el modelo alcanzó un MSE de 675,754 y un R^2 de 0.5816, valores prácticamente idénticos a los obtenidos antes del ajuste. Esto indicó que la regularización aplicada en el meta-modelo no tuvo un impacto significativo en el desempeño del modelo, lo que sugirió que la combinación de los modelos base ya estaba bien ajustada. Además, al comparar con otros métodos de ensamble, se observó que el Voting Regressor mantuvo el mejor desempeño, al lograr un menor error y una mejor capacidad explicativa. Estos resultados sugirieron que, para mejorar la precisión del modelo de Stacking, podría haber sido más beneficioso probar con un meta-modelo más complejo, como Gradient Boosting, XGBoost o incluso redes neuronales, en lugar de limitarse a una regresión Ridge.

El modelo Stacking Regressor fue ajustado utilizando selección de características con RFE, empleando un Random Forest optimizado para identificar las cinco variables más importantes. Luego, el modelo fue reentrenado y evaluado, obteniendo un MSE de 675,754 y un R^2 de 0.5816, valores que no mostraron mejoras significativas en comparación con el modelo original. Esto indicó que las características eliminadas no contribuían de manera relevante a la predicción del precio de renta. Por lo tanto, la selección de características mediante RFE no tuvo un impacto positivo en el rendimiento del modelo, lo que sugirió que el conjunto de variables original ya contenía la información más relevante para la predicción.

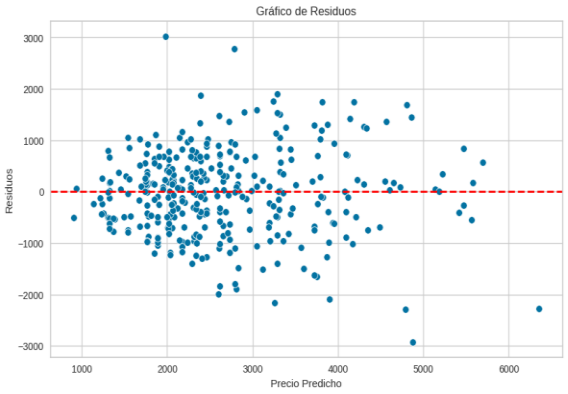


Figura 3. Precio Observado vs. Precio Predicho

El gráfico mostró que el modelo con selección de características (RFE) mantuvo una buena tendencia, pero aún presentó errores en algunos valores atípicos. Esto sugiere que eliminar características con RFE no mejoró significativamente el desempeño del modelo, ya que el error cuadrático medio (MSE) y el coeficiente de determinación (R^2) se mantuvieron prácticamente iguales.

3.2 Selección de variables significativas

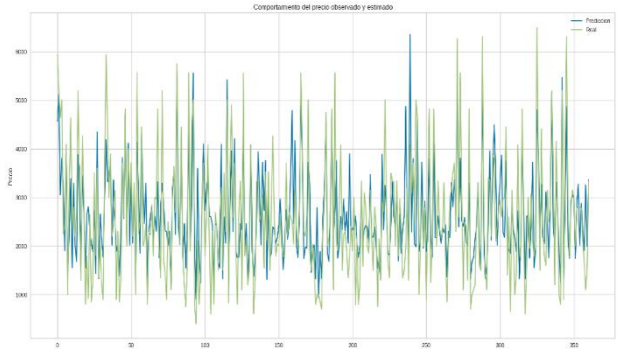


Figura 4. Comportamiento del precio observado y estimado

El gráfico muestra que el modelo con selección de características (RFE) mantiene la tendencia de los precios reales, pero aún presenta errores en algunos valores extremos. Aunque la predicción no es perfecta, el modelo es capaz de captar patrones generales en los datos de renta.

3.4 Residual plot

Muestra la diferencia entre los valores reales y los valores predichos por el modelo de Machine Learning.

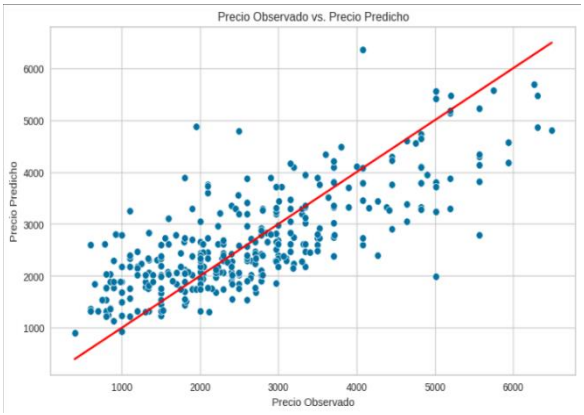


Figura 5. Comportamiento del precio observado y estimado

Este gráfico muestra la distribución de los errores (residuos) del modelo, es decir, la diferencia entre los valores reales y los valores

predichos. Es una herramienta fundamental para evaluar si el modelo tiene algún sesgo sistemático o si los errores están distribuidos de manera aleatoria

Los residuos no están perfectamente distribuidos. Se observa una mayor dispersión en valores altos de precio, lo que sugiere que el modelo tiene más errores en predicciones de alquileres altos. Hay más acumulación de puntos cerca de 0 en valores bajos y medios, lo que sugiere que en precios más bajos el modelo predice mejor. Algunos puntos extremos (outliers) muestran errores muy grandes (residuos mayores a 2000 o menores a -2000), lo que indica errores importantes en algunos casos.

El modelo sigue una tendencia aceptable, pero presenta errores más grandes en predicciones de precios altos. No hay una tendencia fuerte en los residuos, lo que indica que el modelo no está sesgado gravemente.

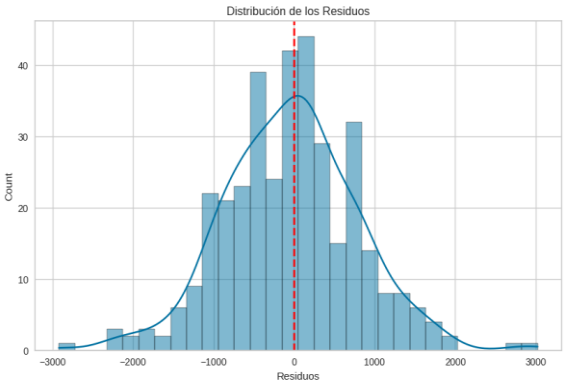


Figura 6. Distribución de residuos

Este gráfico muestra la distribución de los errores (residuos) del modelo, permitiendo evaluar si los errores siguen una distribución normal y si el modelo tiene algún sesgo. El modelo tiene una distribución de errores aceptable, pero no es perfectamente simétrica, lo que sugiere un ligero sesgo. La mayoría de los residuos están cerca de 0, lo que indica que el modelo funciona bien en general.

3.5 Quantile-Quantile

Permite evaluar si los residuos del modelo siguen una distribución normal, lo cual es un supuesto importante en regresión.

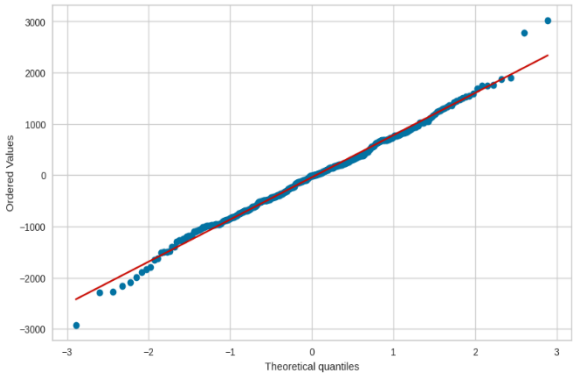


Figura 6. Distribución de residuos

La mayoría de los puntos siguen la línea roja en la parte central, lo que sugiere que los residuos se comportan aproximadamente como una distribución normal en su rango medio, Sin embargo, en los extremos hay desviaciones significativas, lo que indica que hay outliers y valores extremos en los residuos. Este comportamiento sugiere que los errores del modelo no son completamente normales, especialmente en los valores más altos y más bajos de los residuos.

3.7 VotingRegressor

Utilizando VotingRegressor, que es un modelo de ensamble en scikit-learn que combina múltiples modelos de regresión para mejorar la precisión de las predicciones. En lugar de confiar en un solo modelo, "vota" entre varios modelos y devuelve un valor promedio ponderado de sus predicciones [21]. Se obtuvieron los siguientes resultados:

Tabla 3. Resultados del VotingRegressor

Modelo	MSE	R^2
VotingRegressor	671887.3591633316	0.5839981574428552

3.8 Curva de aprendizaje

Es una herramienta clave para evaluar el desempeño del modelo en función del tamaño del conjunto de entrenamiento. Permite identificar si el modelo está sufriendo de sobreajuste (overfitting) o subajuste (underfitting).

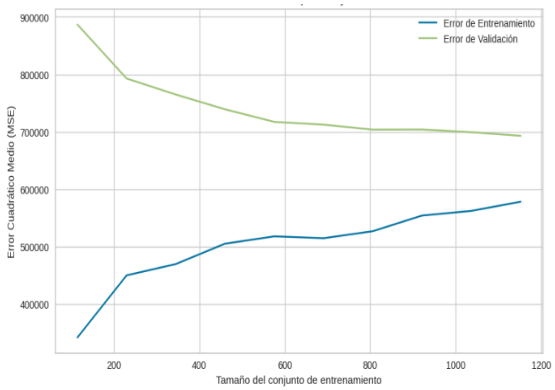


Figura 7. Curva de aprendizaje

El análisis de la curva de aprendizaje del modelo Voting Regressor mostró una mejora progresiva en el desempeño a medida que se incrementó el tamaño del conjunto de entrenamiento. Inicialmente, el error en el conjunto de entrenamiento fue bajo, lo que indicó que el modelo tenía una alta capacidad de ajuste cuando se trabajaba con pocos datos. Sin embargo, conforme se añadieron más ejemplos, el error de entrenamiento aumentó ligeramente, mientras que el error de validación disminuyó de manera significativa hasta estabilizarse.

Se observó que el modelo no sufrió de un sobreajuste extremo, ya que la diferencia entre el error de entrenamiento y el de validación se redujo con más datos, lo que sugirió que el modelo pudo generalizar de manera razonable. A pesar de ello, el error de validación se estabilizó en un valor relativamente alto (~700,000 MSE), lo que indicó que el modelo aún tenía margen de mejora. Esto sugirió que, aunque el Voting Regressor logró aprender patrones generales de los datos, su capacidad predictiva podría mejorarse mediante una optimización más profunda de hiperparámetros, la incorporación de modelos más complejos como XGBoost o redes neuronales, o el aumento del conjunto de datos para mejorar su capacidad de generalización.

4. CONCLUSIÓN

El estudio evaluó la aplicación de Machine Learning en la predicción del precio de renta de apartamentos, probando modelos individuales, ensambles y técnicas de optimización. Se determinó que el Voting Regressor fue el modelo más eficiente, logrando el menor Error Cuadrático Medio (MSE) y el mayor R^2, mientras que el Stacking Regressor y la selección de características con RFE no mejoraron el rendimiento. La optimización con GridSearchCV tampoco generó mejoras significativas, lo que sugirió que el modelo ya estaba bien ajustado.

El análisis de residuos evidenció una tendencia general bien capturada, pero con errores mayores en precios elevados y la presencia de outliers. La curva de aprendizaje mostró que el modelo mejoró con más datos, aunque aún tenía margen de mejora en su capacidad de generalización.

En conclusión, el Voting Regressor se consolidó como la mejor opción predictiva, aunque la precisión pudo haberse optimizado con más datos, transformaciones de variables o modelos más complejos como XGBoost o redes neuronales. Estos resultados destacaron la importancia de evaluar múltiples enfoques antes de seleccionar un modelo óptimo en problemas de regresión.

REFERENCIAS

[1] Fontalvo-Herrera, Tomás J, De la Hoz, Enrique, & Efraín. (2020). DETERMINANTES DEL MERCADO INMOBILIARIO QUE AFECTA LA VOLATILIDAD DEL PRECIO

FUNDAMENTAL POR METRO CUADRADO DE LOS INMUEBLES MULTIFAMILIARES EN LIMA METROPOLITANA DURANTE EL PERIODO 2002-2014. *Dimensión Empresarial*, *18*(2).

[2] Borja, J. (2019). Ciudadanía, derecho a la ciudad y clases sociales. O la democracia vs el Derecho. En *Derecho a la ciudad:una evocación de las transformaciones urbanas en América latina* (Vol. 1).

[3] Yuan, D., Yau, Y., Hou, H., & Liu, Y. (2021). Factors influencing the project duration of urban village redevelopment in contemporary China. Land, 10(7). https://doi.org/10.3390/land10070707

[4] Díaz Rodríguez, G., & Navarro Fernández, C. (2023). El derecho de acceso a la vivienda en Lima Metropolitana a propósito de la Ley 31313, Ley de Desarrollo Urbano Sostenible y el Proyecto de Reglamento de Vivienda de Interés Social. THEMIS Revista de Derecho, 83. https://doi.org/10.18800/themis.202301.007

[5] Sánchez Barrera, J. C., & Valdivia Loro, A. (2022). Calidad de la vivienda en Lima metropolitana. Índice, satisfacción y propuesta de un instrumento. Cuadernos de Vivienda y Urbanismo, 15(1). https://doi.org/10.11144/javeriana.cvu15.cvlm

[6] Lozano Bazán, H. A., & Luna Durand, D. (2017). Rentabilidad de los bienes raíces residenciales en el Perú : ¿existe burbuja intrínseca? Pontificia Universidad Católica Del Perú.

[7] International Monetary Fund. (2010). Are House Prices Rising too Fast in China? IMF Working Papers, 10(274). https://doi.org/10.5089/9781455210817.001

[8] Wang, W. C. V., Lung, S. C. C., & Liu, C. H. (2020). Application of machine learning for the in-field correction of a PM2.5 low-cost sensor network. Sensors (Switzerland), 20(17). https://doi.org/10.3390/s20175002

[9] Xie, X., Zhang, X., Shen, J., & Du, K. (2022). Poplar's Waterlogging Resistance Modeling and Evaluating: Exploring and Perfecting the Feasibility of Machine Learning Methods in Plant Science. Frontiers in Plant Science, 13. https://doi.org/10.3389/fpls.2022.821365

[10] Aziz, R. M., Sharma, P., & Hussain, A. (2024). Machine Learning Algorithms for Crime Prediction under Indian Penal Code. Annals of Data Science, 11(1). https://doi.org/10.1007/s40745-022-00424-6

[11] Ding, Y., Zhu, H., Chen, R., & Li, R. (2022). An Efficient AdaBoost Algorithm with the Multiple Thresholds Classification. Applied Sciences (Switzerland), 12(12). https://doi.org/10.3390/app12125872

[12] Zhang, P., Shi, X., Khan, S. U., Ferreira, B., Portela, B., Oliveira, T., Borges, G., Domingos, H., Leitão, J., Mohottige, I. P., Gharakheili, H. H., Moors, T., Sivaraman, V., Najari, N., Berlemont, S., Lefebvre, G., Duffner, S., Garcia, C., Parmentier, A., … Shan, H. (2019). IEEE Draft Standard for Spectrum Characterization and Occupancy Sensing. IEEE Access, 9(2).

[13] Kadavi, P. R., Lee, C. W., & Lee, S. (2018). Application of ensemble-based machine learning models to landslide susceptibility mapping. Remote Sensing, 10(8). https://doi.org/10.3390/rs10081252

[14] Calderón, J. (2015). Hacia una vivienda pública de alquiler en el Perú. Mercado de alquileres y Estado. WASI, 2(3).

[15] Kadavi, P. R., Lee, C. W., & Lee, S. (2018). Application of ensemble-based machine learning models to landslide susceptibility mapping. Remote Sensing, 10(8). https://doi.org/10.3390/rs10081252

[16] Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the internet for use in psychological research. Psychological Methods, 21(4). https://doi.org/10.1037/met0000081

[17] Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in Neurorobotics, 7(DEC). https://doi.org/10.3389/fnbot.2013.00021

[18] Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54(3). https://doi.org/10.1007/s10462-020-09896-5

[19] Kadavi, P. R., Lee, C. W., & Lee, S. (2018). Application of ensemble-based machine learning models to landslide susceptibility mapping. Remote Sensing, 10(8). https://doi.org/10.3390/rs10081252

[20] Ruchay, A., Gritsenko, S., Ermolova, E., Bochkarev, A., Ermolov, S., Guo, H., & Pezzuolo, A. (2022). A Comparative Study of Machine Learning Methods for Predicting Live Weight of Duroc, Landrace, and Yorkshire Pigs. Animals, 12(9). https://doi.org/10.3390/ani12091152

[21] Zhou, M., Zhong, X., Sun, Y., & Gan, L. (2023). Prediction Model of Coal Consumption Based on Random Forest Variable Selection and Random-Grid Hyperparametric Optimization Algorithm. Proceedings - 2023 International Conference on Power System Technology: Technological Advancements for the Construction of New Power System, PowerCon 2023. https://doi.org/10.1109/PowerCon58120.2023.10331350