

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
Departamento de Ingeniería Comercial

Machine Learning: Comparación de modelos de aprendizaje automático en la predicción de precios de vivienda en la región metropolitana de Santiago

INGENIERÍA COMERCIAL

Mayo 2025

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

Departamento de Ingeniería Comercial

Machine Learning: Comparación de modelos de aprendizaje automático en la predicción de precios de vivienda en la región metropolitana de Santiago

Tesis presentada por

Deivi Yorsy Ramos Almuna

Como requisito para optar al Título de

INGENIERO COMERCIAL

Director de Tesis: **Francesco Pontarelli**

Mayo 2025

TITULO DE TESIS:

“Machine Learning: Comparación de modelos de aprendizaje automático en la predicción de precios de vivienda en la región metropolitana de Santiago”

AUTOR:

Deivi Yorsy Ramos Almuna

TRABAJO DE TESIS, presentando en cumplimiento parcial de los requisitos para el Título de Ingeniero Comercial de la Universidad Técnica Federico Santa María.

Observaciones:

Francesco Pontarelli

Director de Tesis

Macarena Gatica

Correferente

Valparaíso, Mayo 2025

Todo el contenido, análisis, conclusiones
y opiniones vertidas en este estudio son
de mi exclusiva responsabilidad.

Nombre: Deivi Rómulo Almona

Firma: 

Fecha: 26/05/25

AGRADECIMIENTOS

*En primer lugar, quiero expresar mi más profundo agradecimiento a mi familia, quienes han estado presentes en cada paso que he dado, brindándome su apoyo incondicional. Soy el hombre que soy hoy, sin lugar a dudas, gracias a ellos. A mis padres, **Ernestina Almuna Astudillo** y **Carlos Ramos Montoya**, por darme el amor que todo hijo debería recibir; y a mis hermanos, **Jean Carlos Ramos Almuna** y **Yeselya Martell Almuna**, por ser siempre ese apoyo sólido en los momentos difíciles.*

*Extiendo mi gratitud a mis amigos y compañeros de universidad, quienes hicieron de esta travesía una experiencia emocionante y enriquecedora. Gracias por su constante preocupación y por estar presentes en mi vida, especialmente a **Vanessa, Francisca, Alexis, Camilo, Lucas, Camila, Daphne** y **Javiera**. Asimismo, agradezco profundamente a mis amigos psicólogos, quienes siempre tendrán un lugar especial en mi corazón.*

*Quiero expresar también mi sincero agradecimiento a la **Universidad Técnica Federico Santa María** y al **Departamento de Ingeniería Comercial**, por la formación académica que me brindaron y por la oportunidad de desarrollar esta investigación. A mi tutor, el profesor **Francesco Pontarelli**, por su invaluable orientación, paciencia y comentarios que enriquecieron este trabajo. A mi correferente **Macarena Gatica**, por su constante apoyo y valiosas enseñanzas tanto durante mi formación académica como en la elaboración de esta tesis. Además, agradezco a la Dra. **Jocelyn Tapia**, por su apoyo técnico y su disposición al facilitarme la base de datos sobre la cual se construyó esta tesis.*

Índice

1. Introducción.....	7
2. Marco Teórico.....	9
2.1. Aprendizaje Automático: Conceptos y Evolución.....	9
2.2. Predicción de Precios Inmobiliarios.....	9
2.3. Variables Consideradas en el Estudio.....	11
2.4. Modelos de Machine Learning para Predicción de Precios.....	12
2.5. Técnicas de Ensemble: Stacking.....	14
2.6. Optimización de Hiperparámetros.....	14
2.7. Evaluación de Modelos.....	15
2.8 Interpretabilidad del Modelo: SHAP Values.....	17
3. Metodología.....	17
3.1 Enfoque de la investigación.....	17
3.2. Conjunto de datos.....	18
3.3. Análisis Exploratorio de Datos (EDA).....	18
3.4. Preparación de los datos.....	23
3.5. Modelado y optimización.....	24
4. Resultados.....	25
4.1 Resultados Generales.....	25
4.2 Modelos Clave: Random Forest, Ensemble y Linear Regressor.....	26
4.3 Comparación de Valores Predichos versus Valores Reales.....	27
4.4 Errores Promedios por Rango de Precios.....	28
4.5 Distribución Geográfica de las Predicciones.....	29
4.6 Importancia de las variables.....	31
4.7 Desempeño en comunas minoritarias.....	32
5. Discusiones.....	34
5.1. Modelos Avanzados versus Modelos Lineales.....	34
5.2. Desempeño en Comunas con Baja Representación.....	35
5.3. Importancia de la Interpretabilidad del Modelo.....	35
5.4. Sesgo del Modelo y Representación Desigual de Datos.....	36
6. Recomendaciones.....	38
7. Implicaciones Prácticas.....	40
7. Conclusión.....	50
8. Referencias.....	52
9. Anexos.....	54
9. Código.....	63
Código con valores Paramétricos óptimos.....	63
Código con búsqueda de hiperparametros.....	66

1. Introducción

El mercado inmobiliario constituye un pilar esencial de la economía, influenciado por factores económicos, sociales y geográficos. Dentro de este mercado, la valoración de inmuebles representa un desafío significativo, debido a la interacción de múltiples variables intrínsecas, como el tamaño y las comodidades de una vivienda, y extrínsecas, como la ubicación y las condiciones del entorno. En el contexto de la Región Metropolitana de Santiago, este desafío se intensifica por la diversidad y dinamismo del mercado, donde los precios fluctúan rápidamente, reflejando tanto cambios macroeconómicos como transformaciones urbanas.

En este escenario, los modelos de aprendizaje automático han surgido como herramientas prometedoras para abordar problemas de predicción complejos. Estos modelos no solo permiten analizar grandes volúmenes de datos históricos, sino que también identifican patrones no lineales y relaciones complejas que los métodos tradicionales, como la regresión lineal, no logran captar adecuadamente. Sin embargo, dada la amplia variedad de algoritmos disponibles, resulta necesario realizar un análisis comparativo que permita identificar cuál de ellos se adapta mejor a las características específicas del mercado inmobiliario en Santiago.

El presente trabajo tiene como objetivo comparar el desempeño de diferentes modelos de aprendizaje automático para la predicción de precios de viviendas en la Región Metropolitana. Esta evaluación busca no solo determinar el modelo más adecuado, sino también identificar los factores clave que influyen en las predicciones y comprender cómo cada algoritmo responde a las características del mercado. De esta forma, se espera aportar una herramienta más robusta para la toma de decisiones informadas en el ámbito inmobiliario.

Objetivo general

- Comparar el desempeño de diferentes modelos de aprendizaje automático para predecir precios de viviendas en la Región Metropolitana de Santiago, analizando su precisión, capacidad de generalización y los factores más influyentes en la determinación del precio.

Objetivos específicos

1. Implementar y comparar diferentes algoritmos de machine learning.
2. Identificar y analizar los factores más influyentes en la determinación del precio de las viviendas.
3. Evaluar el desempeño de los modelos mediante Métricas de Error
4. Proponer recomendaciones e implicaciones prácticas para la mejora continua de los modelos y su potencial aplicación en la industria inmobiliaria.

Justificación

El mercado inmobiliario en la Región Metropolitana de Santiago enfrenta retos significativos debido a su alta complejidad y dinamismo. La capacidad de predecir precios de viviendas de manera precisa es crucial para optimizar la toma de decisiones en un mercado competitivo, especialmente para actores clave como compradores, vendedores e inmobiliarias.

Tradicionalmente, los métodos estadísticos tradicionales han sido utilizados para abordar este problema; sin embargo, su incapacidad para capturar relaciones no lineales y patrones complejos limita su eficacia en mercados tan variados como el de Santiago. Los modelos de aprendizaje automático representan una alternativa prometedora, pero su implementación plantea la necesidad de evaluar cuál de ellos ofrece el mejor equilibrio entre precisión, robustez y aplicabilidad.

Esta investigación se justifica por su potencial para llenar este vacío, proporcionando una comparación objetiva de modelos de machine learning aplicados a la predicción de precios de viviendas. Los resultados no solo contribuirán a mejorar la precisión de las estimaciones, sino que también ofrecerán insights prácticos para implementar herramientas predictivas en el sector inmobiliario, maximizando su utilidad en contextos reales.

2. Marco Teórico

2.1. Aprendizaje Automático: Conceptos y Evolución

El aprendizaje automático, o machine learning, se define como un conjunto de técnicas que permite a los sistemas computacionales la capacidad de aprender de los datos y realizar predicciones o decisiones sin ser explícitamente programados para cada tarea. Su evolución ha sido notable, desde algoritmos fundacionales como el perceptrón desarrollado por Rosenblatt (Rosenblatt, 1958), hasta la sofisticación actual de modelos complejos y métodos de ensamble.

Esta disciplina ha ganado una relevancia sin precedentes en las últimas décadas, impulsada por el avance exponencial en la capacidad de procesamiento computacional y la creciente disponibilidad de grandes volúmenes de datos. Estas mejoras han permitido abordar problemas de gran complejidad, como la predicción de precios de viviendas, con una precisión y adaptabilidad que superan a muchos métodos estadísticos tradicionales. Dentro del machine learning, los problemas de regresión, que buscan predecir un valor numérico continuo, como el precio de una propiedad, representan el eje central de esta investigación.

En el ámbito de la regresión, algoritmos como la Regresión Lineal sirven como un modelo base fundamental y punto de referencia, permitiendo comprender las relaciones directas entre variables. A partir de esta base, modelos más avanzados como Random Forest (Breiman, 2001) y las Máquinas de Soporte Vectorial (SVM) para regresión (Cortes & Vapnik, 1995) son capaces de capturar interacciones no lineales y patrones más complejos en los datos. Particularmente efectivos son los métodos de ensamble, los cuales combinan las predicciones de múltiples modelos base para mejorar la robustez y reducir el error, lo que los convierte en candidatos prometedores para lograr predicciones precisas en mercados con alta volatilidad, como el inmobiliario. La aplicación exitosa de estos modelos requiere un proceso metodológico que abarca desde la preparación de datos hasta el entrenamiento, evaluación y ajuste de los algoritmos para asegurar predicciones fiables y aplicables.

2.2. Predicción de Precios Inmobiliarios

La predicción de precios de viviendas es un problema multidimensional que integra análisis económico, estadístico y computacional. Uno de los enfoques tradicionales más influyentes es la teoría del precio hedónico (Rosen, 1974), la cual postula que el valor de un bien o servicio puede descomponerse en función de las características individuales que lo componen. En el caso específico del mercado inmobiliario, esto permite descomponer el valor de un inmueble en sus características intrínsecas, como el tamaño, la cantidad de habitaciones o las comodidades, y sus características extrínsecas, como la ubicación, la accesibilidad y el entorno.

Este modelo, basado generalmente en técnicas como la regresión lineal, asume relaciones lineales entre las variables. Sin embargo, presenta limitaciones al intentar capturar

interacciones complejas, como la influencia conjunta de la ubicación y las comodidades, o patrones no lineales inherentes al mercado inmobiliario.

En el caso de Santiago, el mercado inmobiliario está condicionado por factores clave que afectan significativamente la variación de precios entre comunas. La ubicación geográfica y las dinámicas socioeconómicas generan una fuerte segmentación, con áreas como Las Condes, Providencia y Vitacura concentrando propiedades de alto valor, mientras comunas como La Pintana o Cerro Navia presentan precios significativamente más bajos. Indicadores como el Índice de Calidad de Vida Urbana (ICVU) reflejan estas diferencias, considerando factores como el acceso a transporte, servicios educativos y áreas verdes, que tienden a estabilizar los precios y fomentar alzas en comunas con altos índices de calidad. Asimismo, las transformaciones urbanas y demográficas han contribuido a diversificar la oferta inmobiliaria. En sectores como Estación Central y San Miguel, el crecimiento vertical ha impulsado la construcción de departamentos, mientras que el aumento de inmigrantes ha incrementado la demanda de propiedades más accesibles, afectando la valorización en sectores periféricos.

El impacto de la economía nacional también desempeña un papel crucial. Factores como las tasas de interés hipotecarias, la inflación y el acceso al crédito tienen un efecto directo en la demanda de viviendas. Por ejemplo, en períodos de altas tasas, los compradores potenciales tienden a posponer sus decisiones, reduciendo la actividad en el mercado. Además, políticas gubernamentales como los subsidios habitacionales han influido en el acceso a la vivienda, particularmente en sectores específicos.

Frente a estas complejidades, los métodos hedónicos, basados en regresión lineal, enfrentan desafíos al asumir relaciones lineales entre las variables, lo que no permite capturar adecuadamente las interacciones entre factores como ubicación y comodidades. Además, los precios de viviendas en comunas emergentes, como Maipú o Quilicura, suelen presentar variaciones abruptas debido a su rápida urbanización, dificultando la modelación sin técnicas que consideren relaciones no lineales.

Los modelos de machine learning ofrecen una alternativa robusta frente a estas restricciones. Su capacidad para identificar patrones no evidentes en datos geoespaciales y temporales, adaptarse a cambios dinámicos en factores socioeconómicos y urbanos, e integrar múltiples variables complejas, como coordenadas geográficas, calidad municipal y accesibilidad, los posiciona como herramientas robustas para generar predicciones más precisas y adaptativas. En este contexto, el aprendizaje automático no solo mejora la comprensión del mercado inmobiliario, sino que también potencia la capacidad para abordar problemas complejos, combinando fundamentos económicos con herramientas avanzadas de análisis computacional.

2.3. Variables Consideradas en el Estudio

El presente trabajo utiliza un conjunto de variables cuidadosamente seleccionadas para capturar tanto las características intrínsecas de las propiedades como los factores extrínsecos relacionados con su entorno. Estas variables han sido diseñadas para reflejar las dinámicas del mercado inmobiliario en la Región Metropolitana de Santiago.

Las **variables intrínsecas** describen las características inherentes al inmueble. Entre estas se incluyen: el número de baños (*bathrooms*), un indicador clave de comodidad; el número de dormitorios (*bedrooms*), que refleja la capacidad de la vivienda; la ubicación física específica del inmueble (*latitude* y *longitude*); la superficie total de la propiedad en metros cuadrados (*total_surface*); el número de estacionamientos disponibles (*parking_lots*); y la presencia de características adicionales como piscina (*swimpool*) o terraza (*terrace*). Además, se considera si la propiedad es una construcción nueva (*new_house*).

Por otro lado, las **variables extrínsecas** representan factores externos al inmueble que influyen en su valor. Estas incluyen indicadores educativos, medidos a través de los puntajes SIMCE de distintos niveles educativos (4° básico, 8° básico y 2° medio), que reflejan la calidad de las instituciones escolares cercanas. También se consideran variables de proximidad a servicios e infraestructura, que miden la distancia (en kilómetros) a estaciones de metro, parques, hospitales, colegios, universidades, centros comerciales, industrias y vertederos. Además, se incluye el número de escuelas ubicadas en radios de 1 km, 2 km y 3 km alrededor de la propiedad, el tipo de vendedor (*seller_type*, ya sea propietario o agente inmobiliario), la comuna donde se encuentra el inmueble (*municipality*), y el Índice de Calidad de Vida Urbana (*quality_municipality*), que refleja el nivel socioeconómico del sector.

Finalmente, la **variable dependiente** es el precio de la propiedad (*price*), que se utiliza como variable objetivo en el modelo predictivo. Este valor es el principal resultado que se busca estimar a partir de las variables intrínsecas y extrínsecas previamente descritas.

Para mayor claridad, el detalle completo de las variables consideradas en este estudio, junto con sus respectivas descripciones, se presenta en el Anexo, en la Tabla 1.

Selección y Ajustes de Variables

Durante el proceso de limpieza y preprocesamiento de los datos, se realizaron ajustes importantes en las variables para alinear el enfoque de esta investigación con sus objetivos específicos. El dataset original contenía información diversa sobre propiedades, incluyendo tanto casas como departamentos, variables relacionadas con las fotos de características específicas del inmueble (como cocina, baños y sala de estar) y la presencia de bodegas asociadas a cada propiedad. Para esta investigación, el dataset fue filtrado para considerar únicamente casas, lo que permitió un análisis más focalizado y relevante para el mercado inmobiliario en estudio.

En primer lugar, se descartaron las variables relacionadas con la presencia de fotos de características específicas del inmueble. Esta decisión se tomó para distanciar esta

investigación de los objetivos de la Dra. Jocelyn Tapia, quien estudia el impacto de las variables basadas en imágenes en la predicción de precios inmobiliarios mediante técnicas avanzadas de machine learning, incluyendo redes neuronales convolucionales (CNN). Al excluir estas variables, se busca evitar la superposición de enfoques y garantizar la originalidad de los resultados.

Por otro lado, la variable (`storage_units`) que representa el número de bodegas asociadas a cada propiedad, también fue eliminada. Esto se debe a que el conjunto de datos original incluía tanto casas como departamentos, pero esta investigación se centrará exclusivamente en casas. En este contexto, la presencia de bodegas no resulta representativa ni relevante para los objetivos del modelo predictivo.

Estas decisiones se tomaron con el propósito de reducir la dimensionalidad del dataset, optimizar el desempeño de los modelos y garantizar que el análisis sea específico y representativo de las dinámicas del mercado inmobiliario en relación con las casas de la Región Metropolitana de Santiago, manteniendo un enfoque distinto y complementario al de la investigación de la Dra. Tapia.

2.4. Modelos de Machine Learning para Predicción de Precios

En el contexto de la predicción de precios de viviendas, los modelos de machine learning ofrecen una amplia gama de enfoques para capturar relaciones tanto lineales como no lineales entre las variables. Para este propósito, se seleccionaron los siguientes modelos de machine learning, cada uno con características específicas que los hacen adecuados para abordar diferentes aspectos del problema.

- **Lasso Regression:** Un modelo de regresión lineal que incorpora una penalización $L1$, promoviendo la selección automática de características. Este enfoque es especialmente útil cuando se busca simplificar el modelo al eliminar variables con baja relevancia, optimizando la interpretabilidad sin sacrificar precisión.
- **Ridge Regression:** Otra extensión de la regresión lineal, pero con una penalización $L2$ que reduce el impacto de la multicolinealidad entre las variables. Este modelo es ideal cuando todas las características contribuyen en cierta medida a las predicciones y se requiere un balance entre simplicidad y desempeño.
- **Árboles de Decisión:** Algoritmo basado en divisiones jerárquicas que organiza los datos en subconjuntos más pequeños según criterios de decisión. Los árboles de decisión destacan por su facilidad de interpretación y capacidad para manejar tanto variables categóricas como numéricas, aunque su sensibilidad al sobreajuste puede limitar su capacidad de generalización.
- **Random Forest Regressor:** Un modelo de ensamble que combina múltiples árboles de decisión, generando predicciones robustas al promediar los resultados de cada árbol individual. Este enfoque no solo reduce el riesgo de sobreajuste, sino que

también captura relaciones complejas entre variables intrínsecas y extrínsecas del mercado inmobiliario.

- **K-Nearest Neighbors (KNN):** Algoritmo basado en la proximidad, que predice los valores en función del promedio de los k vecinos más cercanos. Su sencillez lo convierte en una herramienta efectiva para detectar patrones locales en los datos, aunque su sensibilidad a la escala de las variables y la selección óptima de k puede representar un desafío.
- **Support Vector Regressor (SVR):** Modelo basado en máquinas de soporte vectorial, diseñado para problemas de regresión. SVR ajusta un hiperplano dentro de un margen de tolerancia (ϵ), utilizando kernels para transformar los datos en un espacio de mayor dimensionalidad donde las relaciones no lineales puedan ser modeladas.
- **Modelo de Ensamble:** Un enfoque avanzado que combina las predicciones de los modelos anteriores, maximizando la precisión y capacidad de generalización. Este tipo de modelo aprovecha las fortalezas individuales de cada modelo base, creando una solución integrada y robusta para la predicción de precios de viviendas.

En la sección de Metodología se explicarán en detalle los modelos empleados, junto con el proceso de selección, ajuste y evaluación aplicado al conjunto de datos utilizado. Además, se describirán las técnicas de validación implementadas para garantizar la robustez de los resultados.

Comparación de los modelos

A continuación, se presenta una síntesis comparativa de los modelos de machine learning seleccionados, enfatizando sus principales ventajas y limitaciones. Este análisis tiene como objetivo ofrecer una visión clara y concisa sobre las fortalezas y desafíos asociados a cada modelo, facilitando la comprensión de su aplicabilidad en el contexto del mercado inmobiliario de Santiago.

Table 2: Comparación de modelos de machine learning

Modelo	Ventaja Principal	Limitación Principal
Lasso Regression	Selección automática de características, lo que mejora la interpretabilidad del modelo	Posible pérdida de variables importantes debido a la penalización
Ridge Regression	Reduce el impacto de la multicolinealidad, ofreciendo un balance entre simplicidad y desempeño	Menor capacidad de selección automática de características en comparación con Lasso
Árboles de Decisión	Facilidad de interpretación y manejo de variables categóricas y numéricas	Alta sensibilidad al sobreajuste, lo que puede limitar su generalización
Random Forest	Robustez y manejo de relaciones no lineales	Complejidad computacional y dificultad para interpretar los resultados
KNN (K-Nearest Neighbors)	Sencillez y flexibilidad para detectar patrones locales	Sensibilidad a la escala de las variables y dificultad para elegir el valor óptimo de k
SVR (Support Vector Regressor)	Excelente desempeño en alta dimensionalidad gracias al uso de kernels	Ajuste complejo de hiperparámetros y alta carga computacional
Modelo de Ensamble	Alta precisión y capacidad de generalización combinando fortalezas de otros modelos	Complejidad en la interpretación de los resultados y mayor costo computacional

2.5. Técnicas de Ensamble: Stacking

El stacking es una técnica de aprendizaje automático desarrollada por Wolpert (1992), que se utiliza para mejorar la precisión predictiva al combinar las salidas de varios modelos base mediante un modelo adicional denominado meta-modelo. A diferencia de otros métodos de ensamble, como el bagging o el boosting, el stacking no promedia ni pondera las predicciones, sino que emplea un modelo final que aprende cómo integrar las salidas de los modelos base de manera óptima. Para implementar esta técnica, se entrenan primero diversos modelos base, seleccionados por su capacidad para capturar patrones complementarios en los datos. Las predicciones de estos modelos se utilizan posteriormente como entrada para entrenar el meta-modelo, que se encarga de generar la predicción final. Este enfoque presenta varias ventajas, entre las cuales se destaca su capacidad para integrar modelos de distinta naturaleza, como lineales y no lineales, lo que permite capturar relaciones complejas en los datos. Además, reduce el riesgo de sobreajuste al combinar las fortalezas de los modelos base, mejorando la robustez y precisión de las predicciones. Por estas razones, el stacking es una técnica particularmente útil en problemas donde se requiere maximizar el desempeño predictivo aprovechando la diversidad de los modelos.

2.6. Optimización de Hiperparámetros

Los hiperparámetros son configuraciones específicas de los modelos de machine learning que no se ajustan automáticamente durante el entrenamiento, sino que deben definirse antes de iniciar el proceso. Estos parámetros controlan aspectos fundamentales como la complejidad del modelo, la regularización y su comportamiento general, lo que tiene un

impacto significativo en el desempeño final. Su correcta selección es crucial para equilibrar el sesgo y la varianza del modelo, evitar el sobreajuste y maximizar la capacidad predictiva en datos nuevos. En este estudio, se utilizó GridSearchCV, una técnica que permite evaluar exhaustivamente combinaciones predefinidas de hiperparámetros mediante validación cruzada. Este enfoque asegura una selección óptima al identificar la configuración más adecuada para cada modelo.

En el caso del modelo Random Forest, se ajustaron tres hiperparámetros principales. El número de árboles (*n_estimators*) controla la cantidad de árboles de decisión en el modelo, lo que influye directamente en su capacidad de generalización. La profundidad máxima (*max_depth*) regula el número de niveles en cada árbol, limitando su complejidad para evitar el sobreajuste. Además, el número mínimo de muestras para dividir un nodo (*min_samples_split*) define cómo y cuándo un nodo puede dividirse, impactando la granularidad del modelo.

Para el modelo K-Nearest Neighbors (KNN), se consideraron tres hiperparámetros. El número de vecinos (*n_neighbors*) determina cuántos puntos cercanos se utilizan para realizar una predicción. La métrica de distancia (*p*) establece cómo se mide la proximidad entre los puntos, permitiendo elegir entre la distancia Euclidiana o Manhattan. Finalmente, el peso asignado a los vecinos (*weights*) controla la influencia relativa de los puntos cercanos en las predicciones.

En el caso de Support Vector Regression (SVR), se ajustaron tres hiperparámetros. El parámetro de penalización (*C*) regula el equilibrio entre un modelo más flexible o más estricto, ajustando el margen de tolerancia a los errores. El kernel especifica cómo se transforman los datos para capturar relaciones no lineales, siendo opciones comunes el kernel lineal y RBF (Radial Basis Function). Finalmente, el parámetro *gamma* ajusta la forma en que los datos influyen en el modelo al definir la flexibilidad de la curva de ajuste.

Para los modelos de regresión Ridge y Lasso, se optimizaron tres hiperparámetros. El parámetro de regularización (*alpha*) controla la penalización aplicada a los coeficientes, ayudando a evitar el sobreajuste. El número máximo de iteraciones (*max_iter*) asegura la convergencia del modelo incluso en casos más complejos. Por último, el parámetro relacionado con la resolución del modelo, como el solucionador (*solver* en Ridge o *tol* en Lasso), permite optimizar la precisión en los cálculos de ajuste.

2.7. Evaluación de Modelos

La evaluación de modelos predictivos es esencial para garantizar su precisión y capacidad de generalización. En este trabajo, se han seleccionado las siguientes métricas para analizar, desde diferentes perspectivas, la calidad de las predicciones:

- **Error Absoluto Medio (MAE):** Mide la magnitud promedio de los errores en las predicciones, sin considerar su dirección. Se calcula como:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

Es una métrica intuitiva y útil cuando se busca entender el error promedio.

- **Error Absoluto Porcentual Medio (MAPE):** Calcula el error en términos porcentuales, facilitando la interpretación relativa de los errores respecto a los valores reales. Se define como:

$$MAPE = \frac{1}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right| \cdot 100$$

El MAPE es especialmente útil para comprender el error en términos relativos, lo que permite comparar el desempeño del modelo en diferentes rangos de precios. Sin embargo, es sensible a valores cercanos a cero y puede ser influenciado por outliers.

- **Coeficiente de Determinación (R^2):** El Coeficiente de Determinación mide la proporción de la varianza total en la variable dependiente explicada por el modelo. Se calcula como:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$$

Un valor de R^2 cercano a 1 indica que el modelo explica la mayor parte de la variación en los datos. Sin embargo, R^2 igual a 0 significa que la media sería más útil para predecir los valores.

- **Raíz del Error Cuadrático Medio (RMSE):** Es una de las métricas más comunes para medir la precisión de un modelo de regresión. Se calcula como la raíz del promedio del cuadrado de las diferencias entre los valores predichos ($\hat{y}_i$) y los valores observados (y_i).

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$$

El RMSE penaliza con más fuerza los errores grandes debido a la elevación al cuadrado, lo que lo hace útil para identificar modelos que minimizan errores significativos.

2.8 Interpretabilidad del Modelo: SHAP Values

La interpretabilidad de los modelos de machine learning es un aspecto fundamental, especialmente en problemas complejos como la predicción de precios de viviendas, donde entender el impacto de las variables en las predicciones es crucial. En este contexto, se utilizan los SHAP Values (*SHapley Additive exPlanations*), una técnica basada en la teoría de juegos que permite descomponer las predicciones de un modelo en contribuciones individuales de cada variable, proporcionando una explicación clara y cuantitativa sobre la influencia de cada característica en las predicciones (Lundberg y Lee, 2017).

Los SHAP Values asignan un valor a cada variable de entrada en función de su contribución marginal a la predicción, considerando todas las combinaciones posibles de las otras variables. Este enfoque garantiza una interpretación coherente y consistente, ya que cumple con propiedades fundamentales como la aditividad y la equidad, derivadas de los valores de Shapley en teoría de juegos (Shapley, 1953). Además, los SHAP Values son compatibles con una amplia gama de modelos, desde algoritmos lineales hasta modelos más complejos como Random Forest y Support Vector Regressor.

En este trabajo, se utilizaron los SHAP Values para analizar la importancia de las variables en los modelos de predicción. Esto permitió identificar cuáles características intrínsecas y extrínsecas, como la ubicación geográfica o el número de baños, tienen un mayor impacto en el precio de las viviendas. Las visualizaciones basadas en los SHAP Values, tales como gráficos de resumen y dependencias, ofrecen una visión clara y detallada de cómo los modelos procesan la información, haciendo visibles patrones complejos y relaciones no lineales que, de otro modo, permanecerían ocultos en análisis más tradicionales.

El uso de los SHAP Values no solo mejoró la interpretabilidad del modelo, sino que también proporcionó una base sólida para evaluar la relevancia de las variables, respaldando decisiones fundamentadas y transparentes en el análisis de precios inmobiliarios. En particular, esta técnica fue clave para entender el comportamiento de los modelos, asegurando que los modelos consideraran factores relevantes y reduciendo la posibilidad de interpretaciones erróneas.

3. Metodología

3.1 Enfoque de la investigación

La presente investigación adopta un enfoque cuantitativo y experimental, diseñado para evaluar y comparar el desempeño de diversos modelos de machine learning en la predicción de precios de viviendas en la Región Metropolitana de Santiago. Este estudio se enmarca dentro de un diseño descriptivo - comparativo, ya que analiza datos estructurados con el propósito de identificar las fortalezas y limitaciones de los algoritmos seleccionados en términos de precisión, robustez y capacidad de generalización.

El objetivo principal radica en determinar qué modelos ofrecen un desempeño óptimo al abordar este problema, optimizando no solo su precisión predictiva, sino también su interpretabilidad y aplicabilidad al análisis inmobiliario.

3.2. Conjunto de datos

El conjunto de datos utilizado en este estudio fue proporcionado por la Dra. Jocelyn Tapia y comprende inicialmente 52,039 registros recopilados mediante web scraping de los portales inmobiliarios más visitados en el año 2020. Esta base de datos contiene información sobre propiedades tanto de departamentos como de casas, abarcando características intrínsecas y extrínsecas relacionadas con el mercado inmobiliario. Para alinear el análisis con los objetivos del presente estudio, se realizó un filtrado del conjunto de datos, seleccionando exclusivamente registros correspondientes a viviendas unifamiliares. Este criterio redujo significativamente el tamaño del dataset inicial, dejando un total de 19,971 registros tras la exclusión de entradas asociadas a departamentos.

Las variables del conjunto de datos se clasificaron en dos categorías principales: intrínsecas y extrínsecas. Las variables intrínsecas comprenden características estructurales propias de las propiedades, como el número de baños, dormitorios, superficie total en metros cuadrados y la presencia de comodidades adicionales, tales como piscina o terraza. Por otro lado, las variables extrínsecas abarcan factores contextuales que influyen en el valor de las propiedades, como su ubicación geográfica (latitud y longitud), el Índice de Calidad de Vida Urbana (ICVU) de la comuna y las distancias geodésicas mínimas a servicios clave, como hospitales, colegios, parques, estaciones de transporte público, entre otros.

3.3. Análisis Exploratorio de Datos (EDA)

El análisis exploratorio de datos constituye una etapa fundamental en este estudio, ya que permite comprender el comportamiento inicial de las variables presentes en el dataset. Este proceso tiene como objetivo principal ofrecer una visión detallada de los datos, identificando patrones generales, tendencias, sesgos y posibles valores atípicos que podrían influir en los resultados del análisis.

En esta etapa, se examina la distribución de las variables más relevantes, como el precio de las propiedades (variable objetivo), así como las principales características predictoras que formarán parte del modelo. Es importante señalar que este análisis no implica la exclusión de variables, ya que todas las seleccionadas serán utilizadas en el entrenamiento de los modelos predictivos.

Dado que el precio de las propiedades es la variable objetivo, resulta esencial analizar su distribución antes de aplicar cualquier modelo. Este análisis inicial proporciona una comprensión de su comportamiento, destacando patrones generales, posibles sesgos y la presencia de valores atípicos. A continuación, se presenta el histograma de la variable price, el cual ilustra cómo se distribuyen los valores de las propiedades incluidas en el dataset.

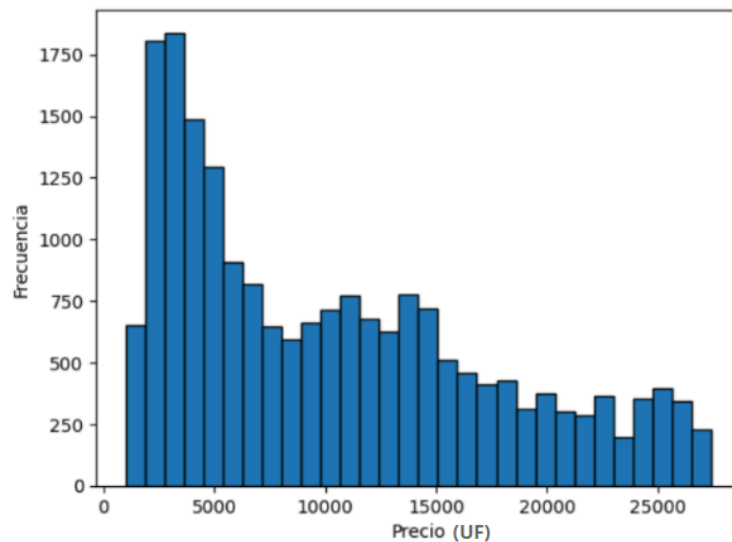


Figura 1. Histograma de precios de propiedades en la Región Metropolitana de Santiago

En la figura 1, se observa que la mayoría de los precios de las propiedades se concentran en el rango de 2.000 UF a 5.000 UF, mostrando una distribución sesgada hacia la derecha. Esto indica la existencia de propiedades con precios significativamente altos que podrían considerarse valores atípicos. Este patrón es común en mercados inmobiliarios, donde una pequeña proporción de propiedades de lujo eleva los valores en los rangos superiores.

Una vez analizada la distribución de precios, es importante entender cómo se comportan las principales variables predictoras. En los histogramas que se presentan a continuación, se analizan características clave como la superficie total, el número de habitaciones y el número de baños, que tienen un impacto significativo en la determinación del precio de las propiedades.

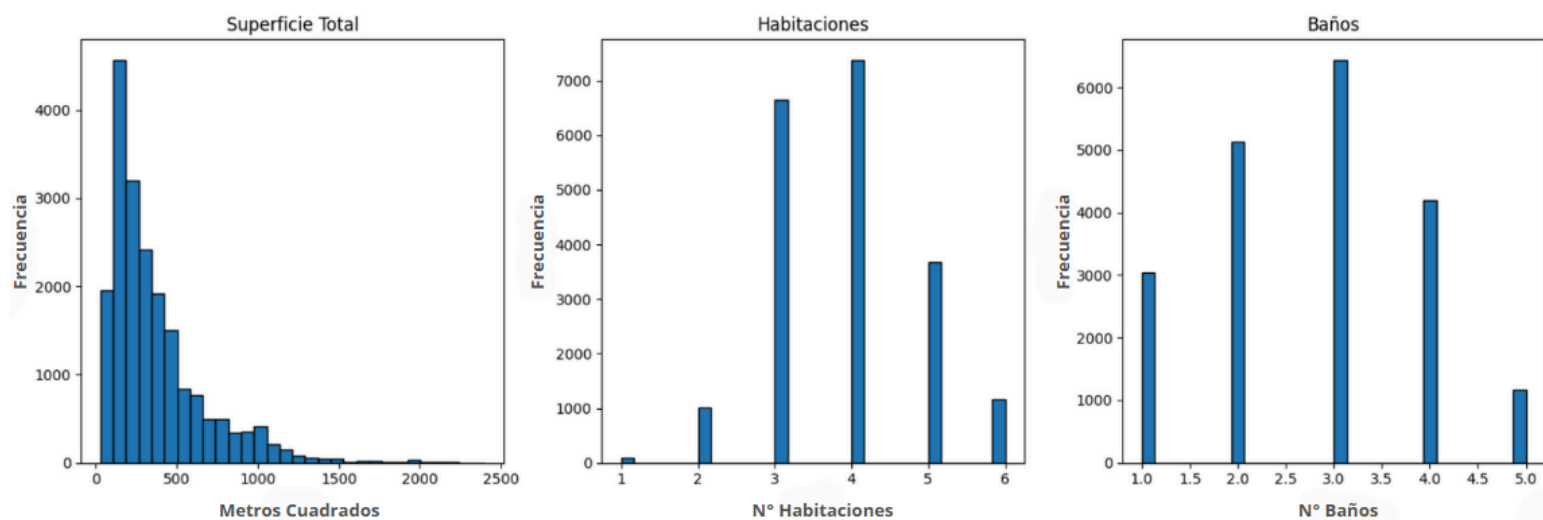


Figura 2. Histogramas de superficie total, número de habitaciones y número de baños.

En la Figura 2, se observa que la superficie total de las propiedades está altamente concentrada en rangos bajos, con una media cercana a 368 m^2 . No obstante, existen propiedades con superficies significativamente mayores, lo que genera una distribución sesgada hacia la derecha. Este patrón también se refleja en el número de habitaciones y baños, donde la mayoría de las propiedades cuentan con 3 o 4 habitaciones y 2 o 3 baños, respectivamente. Estas características evidencian la naturaleza del mercado inmobiliario en la región, predominado por propiedades de tamaño grande, mientras que las configuraciones más pequeñas o extremadamente grandes son menos frecuentes.

Distribución de propiedades en Santiago

Una vez analizadas las características físicas de las propiedades, es relevante contextualizar cómo se distribuyen estas viviendas a lo largo de las distintas comunas de Santiago. Este análisis permite identificar qué comunas concentran la mayor cantidad de datos en el dataset y cómo esto puede influir en el modelo predictivo, dado que ciertas comunas pueden tener mayor peso en el entrenamiento del modelo debido a su representación.

El conteo de propiedades por comuna muestra que las comunas con mayor cantidad de datos en el dataset son Las Condes, Lo Barnechea, y Colina, cada una con más de 1.700 propiedades. Estas comunas, generalmente caracterizadas por valores de propiedades más altos, podrían influir significativamente en la predicción del modelo. En el extremo opuesto, comunas como Curacaví, Calera de Tango, San José de Maipo, Melipilla y Pirque tienen una representación mínima, con menos de 20 registros, es importante destacar que esta distribución no es uniforme, lo que podría limitar el poder predictivo para estas áreas.

En el Anexo 2, se presenta el conteo completo de propiedades por comuna. Este análisis evidencia una clara concentración de datos en comunas específicas, lo que sugiere que los modelos predictivos podrían aprender patrones dominados por estas zonas.

Distribución Geográfica de los Precios en Santiago

Para complementar el análisis de la distribución de propiedades por comuna, se ha generado un mapa geográfico que permite visualizar de manera espacial la variación de precios en el mercado inmobiliario de Santiago. Este mapa utiliza una escala de colores que refleja las diferencias de valor por zona, donde tonos más cálidos (rojos) indican áreas con precios más elevados, mientras que tonos fríos (azules) señalan sectores con valores más accesibles.

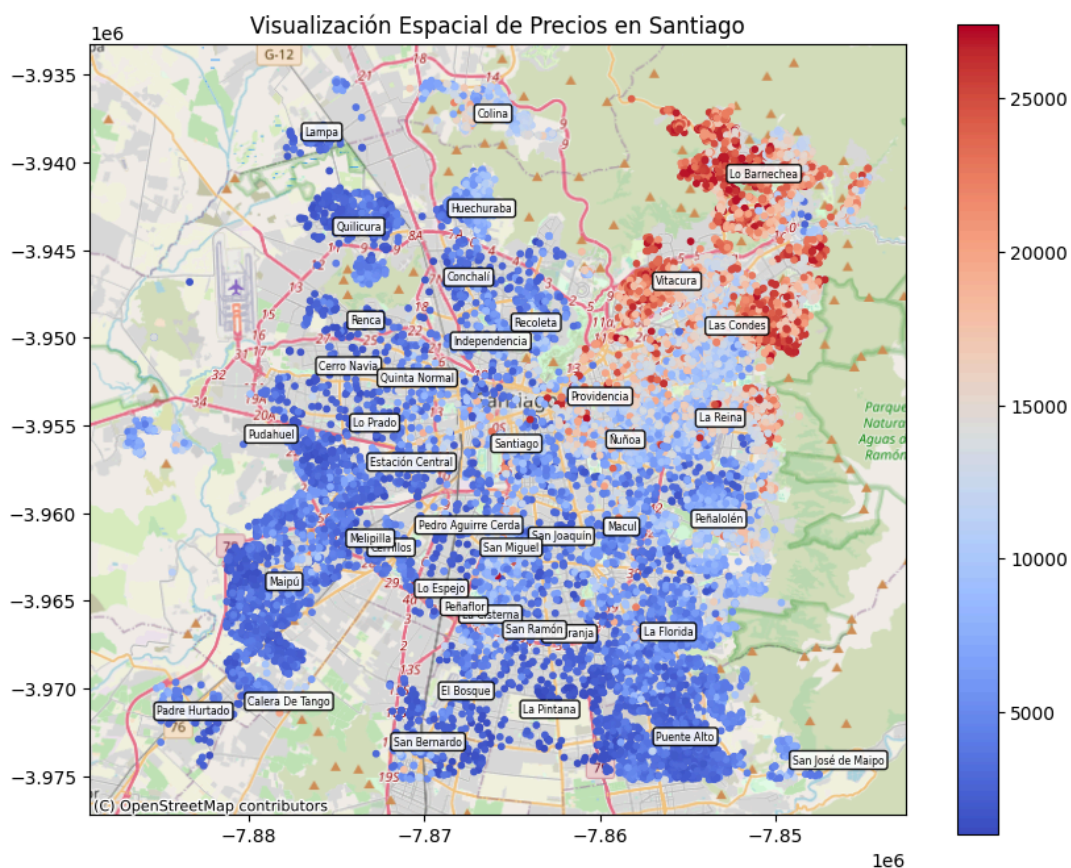


Figura 3. Mapa de distribución de precios de propiedades en Santiago, representado en una escala de colores que varía de azul (precios más bajos) a rojo (precios más altos)

En la Figura 3, se observa que las comunas del sector oriente, como Lo Barnechea, Vitacura, y Las Condes, concentran las propiedades con los precios más altos, representados en tonos rojizos. En contraste, comunas como Maipú, Puente Alto, y San Bernardo, ubicadas principalmente en el sector sur y poniente, predominan en precios más bajos, representados en tonos azulados. Este comportamiento refleja las desigualdades socioeconómicas históricas de la región, con una mayor valorización de las propiedades en comunas con mejores índices de calidad de vida urbana (ICVU) y mejor acceso a servicios.

Este análisis geográfico no solo complementa el análisis numérico, sino que también refuerza la importancia de las variables espaciales, como la latitud, longitud y las distancias a servicios clave, en la formación del precio de las propiedades. Este mapa resalta patrones espaciales significativos que son cruciales para entender cómo la ubicación impacta en el valor inmobiliario.

Relación entre las Variables Predictoras y el Precio

Una vez analizadas las distribuciones y la representación espacial de las propiedades en el dataset, es fundamental explorar las relaciones entre las principales variables predictoras y el precio de las propiedades. Este análisis permite comprender cómo cada característica se

asocia con la variable objetivo y detectar patrones que podrían influir en el comportamiento de los modelos predictivos.

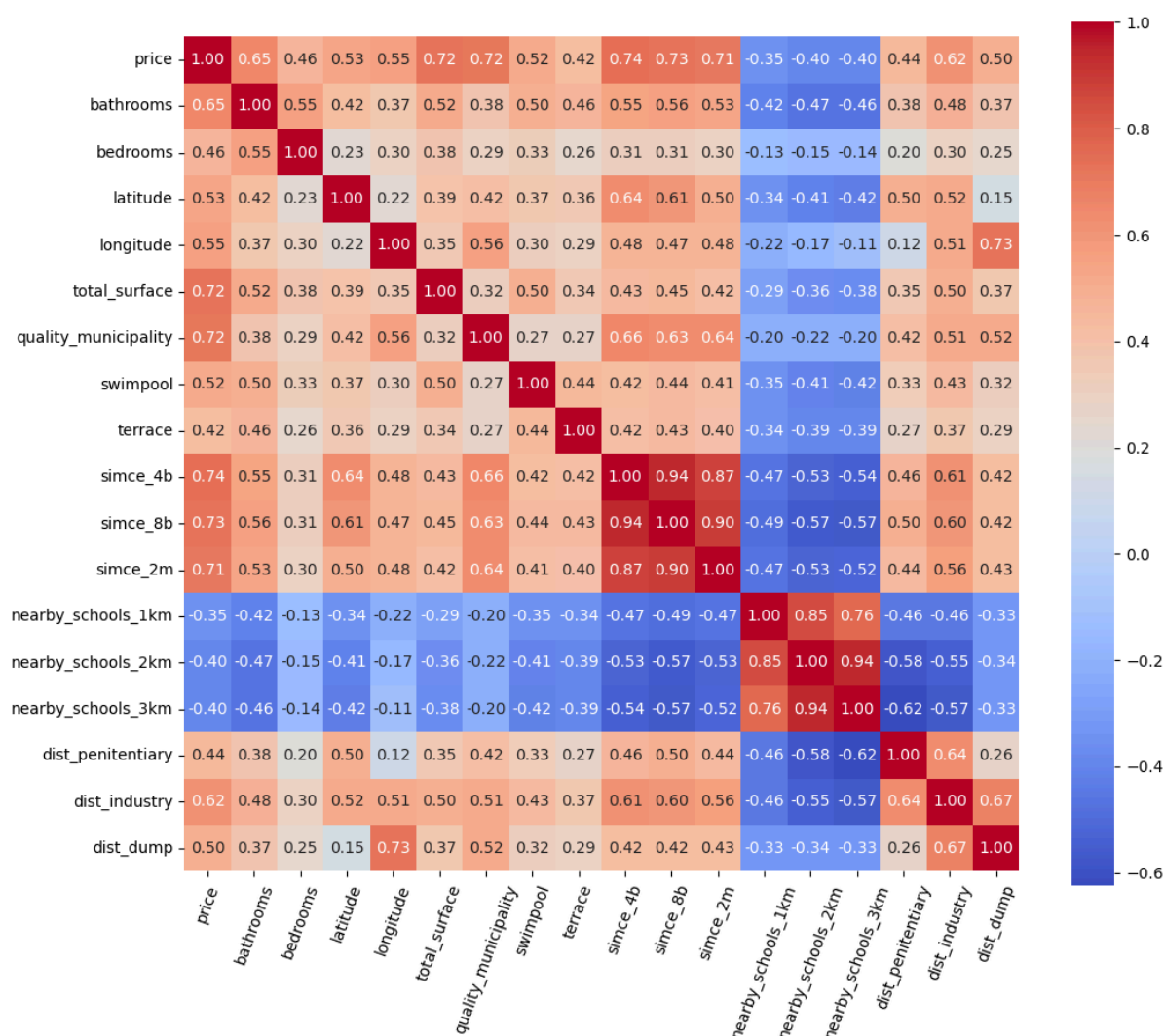


Figura 4. Mapa de correlación entre las variables predictoras y el precio de las propiedades.

En la Figura 4, se observa que variables como la superficie total, la calidad del municipio y el SIMCE de los colegios cercanos presentan una alta correlación positiva con el precio, lo que indica que incrementos en estas características tienden a estar asociados con un aumento en el valor de las propiedades, reflejando también que una mejor calidad educativa en un radio de 2 kilómetros se asocia con valores más altos de las propiedades.

Por otro lado, variables como el número de escuelas cercanas dentro de radios de 1 a 3 kilómetros, presentan correlaciones negativas con el precio. Esto sugiere que la cercanía a escuelas se percibe como una externalidad negativa. Este efecto podría estar asociado a factores como la congestión urbana o la contaminación auditiva generada por el tráfico vehicular y las actividades en el entorno escolar.

Estos resultados están alineados con estudios previos, como el de Sah, Conroy y Narwold (2016), quienes encontraron que la proximidad a escuelas puede generar tanto

externalidades positivas como negativas, dependiendo del contexto. En su trabajo, señalan que, aunque las escuelas suelen incrementar los precios al reducir los costos de transporte y mejorar el acceso educativo, en algunos casos predominan externalidades negativas, como el ruido y la congestión, que disminuyen el atractivo de las propiedades cercanas.

3.4. Preparación de los datos

El proceso de limpieza y selección del conjunto de datos incluyó varias etapas clave. Primero, se filtraron los registros para incluir únicamente propiedades clasificadas como casas, garantizando una mayor especificidad en los resultados. Posteriormente, se eliminaron columnas que no eran relevantes para los objetivos del estudio, como aquellas relacionadas con imágenes de propiedades, dado que su análisis fue objeto de una investigación previa realizada por la Dra. Tapia. También se descartó la variable `storage_units`, que representa el número de bodegas asociadas a cada propiedad, debido a que no es representativa en el contexto de viviendas unifamiliares. Como resultado de este proceso, el número de variables pasó de 47 en el dataset original a 29 en el conjunto de datos final.

Finalmente, el conjunto de datos resultante, compuesto por 19,971 registros, se particionó en tres subconjuntos: entrenamiento, validación y prueba. El conjunto de entrenamiento, destinado a ajustar los modelos, incluyó 12,781 registros, equivalentes al 64 % del total. El conjunto de validación, utilizado para la selección y ajuste de hiperparámetros, abarcó 3,195 registros, representando el 16 %. Por último, el conjunto de prueba, reservado para la evaluación final de los modelos, estuvo compuesto por 3,995 registros, es decir, el 20 % del total. Esta partición garantiza un proceso de modelado robusto, minimizando el riesgo de sobreajuste y permitiendo una evaluación confiable de la capacidad predictiva de los algoritmos seleccionados.

Se realizó un proceso de estandarización en todas las variables numéricas. Este procedimiento ajustó los datos para que tuvieran una media de 0 y una desviación estándar de 1, la estandarización se llevó a cabo debido a que muchos algoritmos, como Support Vector Regressor o K-Nearest Neighbors, son sensibles a las magnitudes relativas de las variables. Al reducir las escalas de las variables a un rango comparable, este procedimiento no solo mejora la capacidad de los modelos para identificar patrones subyacentes, sino que también asegura que las variables con valores más grandes no dominen el proceso de entrenamiento. Los parámetros de escalado fueron calculados exclusivamente a partir del conjunto de entrenamiento y posteriormente aplicados a los conjuntos de validación y prueba, garantizando que los datos futuros fueran transformados de manera consistente.

Además, la variable categórica `municipality` fue transformada utilizando codificación one-hot, un método que descompone las categorías de una variable en múltiples columnas, donde cada columna representa un nivel único de la categoría original. En estas columnas, se asigna un valor de 1 si el registro pertenece a esa categoría específica, y un 0 en caso contrario. Para evitar problemas de colinealidad entre las columnas generadas, se eliminó la primera categoría como referencia. Este procedimiento permitió incorporar la información sobre la comuna de cada propiedad en los modelos de manera numérica, garantizando que

las relaciones entre comunas se analizaran de forma independiente y sin imponer un orden arbitrario, lo que resultó en un análisis más transparente y robusto.

3.5. Modelado y optimización

El ajuste de hiperparámetros se llevó a cabo utilizando una estrategia basada en validación cruzada predefinida mediante el esquema de `PredefinedSplit`. Este método permitió combinar los conjuntos de entrenamiento y validación en un único dataset para realizar una búsqueda exhaustiva de hiperparámetros con `GridSearchCV`. En este esquema, los datos del conjunto de entrenamiento se usaron para entrenar los modelos, mientras que los datos del conjunto de validación se utilizaron para evaluar diferentes combinaciones de hiperparámetros. Esta metodología aseguró una evaluación consistente y evitó la contaminación de los datos durante el proceso de optimización.

Cada modelo ajustado contó con una rejilla específica de hiperparámetros. En el caso de Random Forest, los parámetros ajustados incluyeron la profundidad máxima de los árboles, el número de estimadores y el mínimo de muestras requeridas para dividir un nodo. Para K-Nearest Neighbors, se evaluaron combinaciones del número de vecinos, el tipo de peso asignado y la métrica de distancia. En el modelo Support Vector Regressor, se ajustaron el kernel, el parámetro de penalización y la escala del gamma. En los modelos de regresión Ridge y Lasso, los parámetros optimizados fueron el coeficiente de regularización, el número máximo de iteraciones y la tolerancia para la convergencia.

Los mejores hiperparámetros obtenidos tras este proceso fueron los siguientes: para Random Forest, `max_depth`: None, `min_samples_split`: 2 y `n_estimators`: 200; para K-Nearest Neighbors, `n_neighbors`: 7, `p`: 1 y `weights`: uniform; para SVR, `C`: 10, `gamma`: scale y kernel: linear; para Ridge Regression, `alpha`: 10, `max_iter`: 1000 y `solver`: svd; y finalmente, para Lasso Regression, `alpha`: 1, `max_iter`: 1000 y `tol`: 0.0001. Este proceso de optimización aseguró que cada modelo estuviera configurado de manera óptima para maximizar su capacidad predictiva y minimizar el error en los datos de prueba.

Para complementar los modelos individuales, se construyó un modelo de ensamble utilizando la técnica de Stacking Regressor, que combina las predicciones de varios algoritmos base mediante un modelo final denominado meta-modelo. En este estudio, los modelos base seleccionados para el ensamble fueron Lasso, KNN y SVR, optimizados previamente durante el ajuste de hiperparámetros. Estos modelos base se integraron en el Stacking Regressor, utilizando como meta-modelo una regresión lineal, lo que permitió combinar las fortalezas individuales de cada algoritmo para mejorar la precisión general del sistema predictivo. El proceso de construcción del modelo de ensamble incluyó el ajuste final sobre el conjunto combinado de entrenamiento y validación.

El desarrollo de esta investigación se sustentó en el uso de un conjunto de librerías de Python, ampliamente reconocidas en los campos de la ciencia de datos y el aprendizaje automático. Los nombres de estas librerías, junto con una descripción de sus propósitos específicos en cada etapa del análisis, se detallan en el Anexo 3.

4. Resultados

4.1. Resultados Generales

En esta sección se exponen los resultados obtenidos en la comparación de los modelos evaluados para la predicción de precios de viviendas en la Región Metropolitana. Los modelos considerados incluyen regresión lineal, Ridge, Lasso, Random Forest Regressor (RFS), K-Nearest Neighbors (KNN), Support Vector Regressor (SVR) y un modelo ensamble basado en stacking que combina Lasso, KNN y SVR, la evaluación se realizó utilizando métricas como MAE , $RMSE$, R^2 y $MAPE$.

A Continuación se presentan los resultados de cada modelo en el conjunto de test, ordenados por desempeño general. En el cual, se destacan el modelo de Random Forest Regressor y el Modelo de Ensamble, que lograron los mejores resultados.

Table 4: Resultados de métricas por modelo en la predicción de precios de viviendas

Modelo	MAE (UF)	RMSE (UF)	R^2	MAPE (%)
Random Forest Regressor	1045.08	1676.89	0.95	12.52
Ensemble	1407.50	2117.31	0.91	16.35
K-Nearest Neighbors	1465.05	2239.44	0.90	17.33
Support Vector Regressor	1773.35	2673.39	0.86	20.04
Linear Regressor	1776.25	2534.55	0.88	21.85
Lasso Regression	1778.30	2537.00	0.88	21.88
Ridge Regression	1778.73	2536.83	0.88	21.93

Como se observa en la Tabla 1, el modelo Random Forest Regressor presentó el mejor desempeño general, con un MAE de 1045.08, un RMSE de 1676.89, un R^2 de 0.95 y un MAPE de 12.52%. Estos resultados destacan su capacidad para capturar relaciones complejas entre las variables. Por su parte, el modelo Ensemble, compuesto por la combinación de Lasso, KNN y SVR, también obtuvo métricas destacadas, alcanzando un R^2 de 0.91 y un MAPE de 16.35%.

Por el contrario, los modelos lineales (Ridge y Lasso) junto con el SVR presentaron limitaciones cuando se evaluaron individualmente, mostrando mayores errores y un desempeño inferior frente a los modelos no lineales. Estas diferencias subrayan la relevancia de adoptar enfoques avanzados para abordar la complejidad del mercado inmobiliario.

Para ilustrar de manera más clara las diferencias en el desempeño de los modelos, la Figura 5, presenta un gráfico comparativo de los errores absolutos promedio (MAE y RMSE) para cada modelo, expresados en Unidades de Fomento (UF).

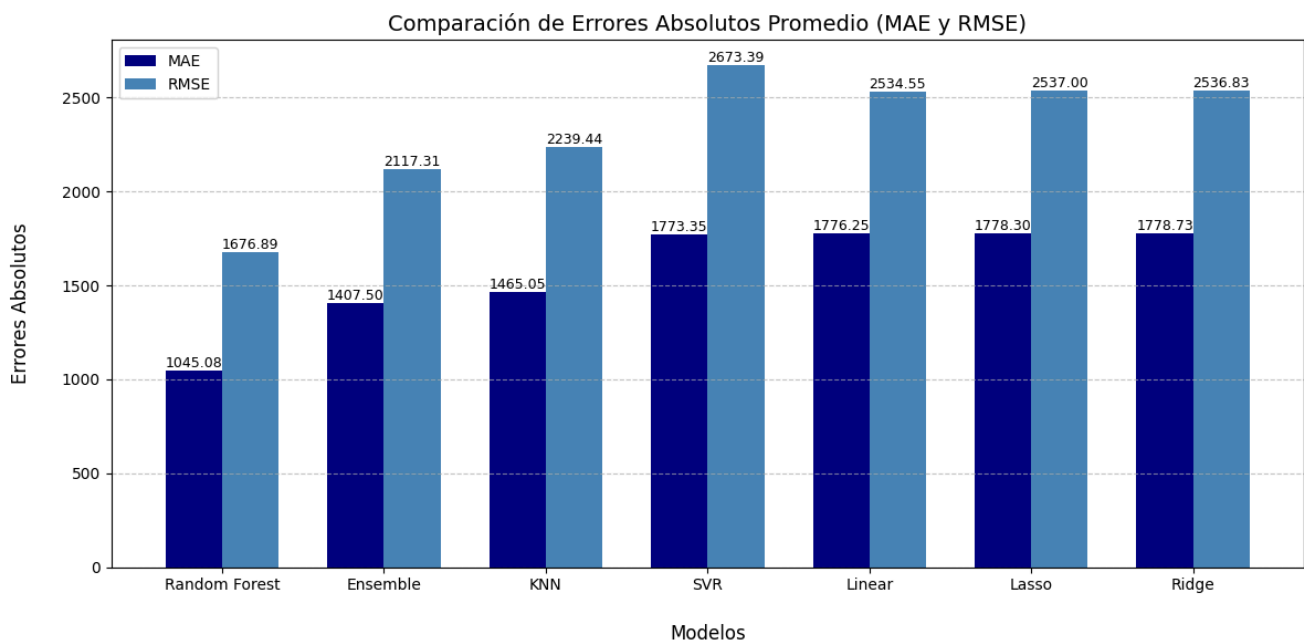


Figura 5. Comparación de errores absolutos promedio (MAE y RMSE) para los modelos evaluados, expresados en Unidades de Fomento Chilena (UF).

El Random Forest exhibe un desempeño notable en comparación con los otros modelos, distanciándose del valor real en un promedio de 1045.08 UF, con un error máximo en promedio de 1676 UF. Estas cifras lo posicionan muy por delante del siguiente modelo en rendimiento, el de Ensemble, el cual se aleja en torno a 1407 UF del valor real (MAE), implicando una diferencia de 362 UF con respecto al Random Forest.

Es fundamental resaltar que el modelo Random Forest es el que logra la mayor reducción del error, medido por su RMSE, aventajando a su competidor más cercano, el Ensemble, en 441 UF. Esta brecha se amplía considerablemente al compararlo con los modelos lineales, mostrando una mejora de unas 858 UF aproximadamente. Por otro lado, el modelo de máquinas de soporte vectorial es el que más dificultades presenta para optimizar la reducción del error en las estimaciones, llegando a un RMSE de 2673.39 UF, lo que supera en aproximadamente 140 UF el error de los modelos lineales.

4.2 Modelos Clave: Random Forest, Ensemble y Linear Regressor

A partir de los resultados obtenidos, se seleccionaron los tres modelos con mejor desempeño general, Random Forest, Ensemble y Linear Regressor. Aunque KNN y SVR obtuvieron métricas superiores a Linear Regressor, la elección de este último se justifica por dos razones principales.

Primero, KNN y SVR ya forman parte del modelo Ensemble, junto con Lasso, lo que asegura que sus contribuciones están representadas en el análisis. Segundo, Linear Regressor sirve como punto de referencia al ser un modelo lineal, lo que permite comparar

su desempeño frente a enfoques más avanzados y evaluar el impacto de las relaciones no lineales en las predicciones.

El objetivo de esta selección es realizar un análisis exhaustivo de las relaciones capturadas por los modelos seleccionados, comprender los mecanismos subyacentes a sus predicciones y determinar las razones detrás de su desempeño relativo. Este análisis se enfoca en evaluar su rendimiento global y desglosar tanto sus fortalezas como sus limitaciones, con el fin de obtener conclusiones más claras y precisas en el contexto de la predicción de precios de viviendas.

4.3 Comparación de Valores Predichos versus Valores Reales

A continuación, se presenta un gráfico de dispersión que compara los valores predichos y reales para los tres modelos seleccionados. Esta visualización permite evaluar la precisión de las predicciones y detectar patrones específicos en el desempeño de cada modelo. Además se incluye la línea de ajuste perfecto ($y = x$) como referencia, que representa el escenario ideal donde las predicciones coinciden exactamente con los valores reales.

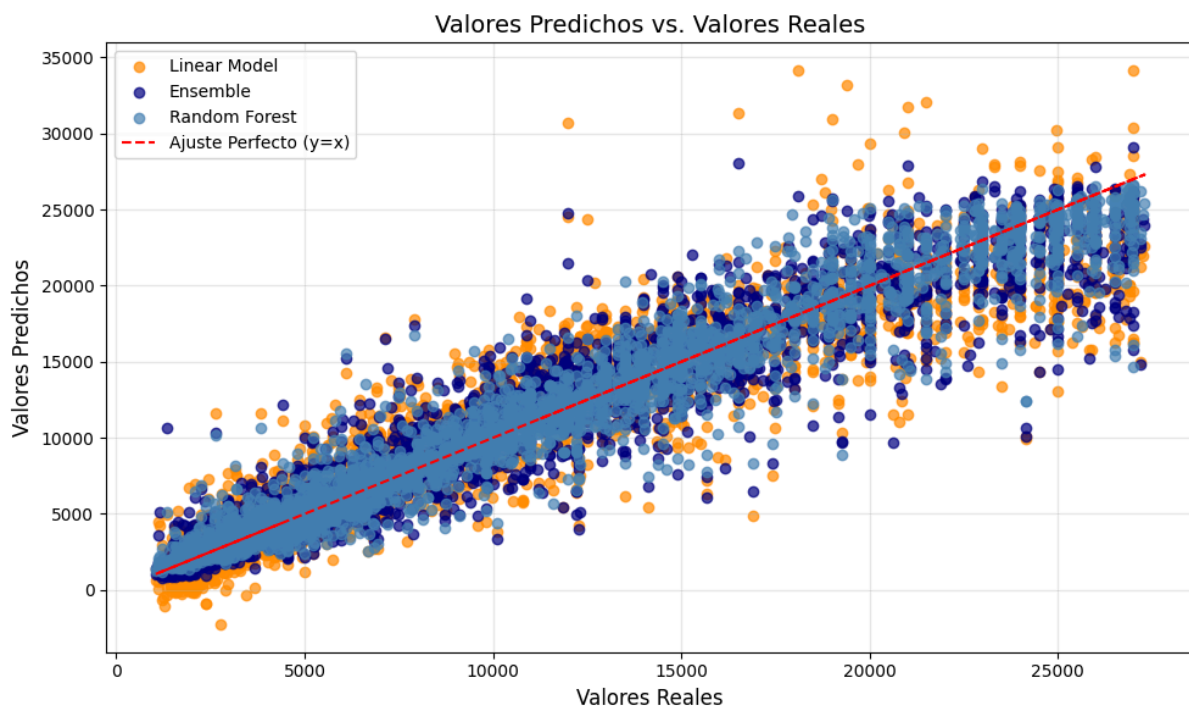


Figura 6. Comparación entre valores predichos y valores reales para los modelos Random Forest, Ensemble y Regresión Lineal. La línea roja representa el ajuste perfecto ($y = x$).

En el gráfico se observa que los tres modelos presentan una alta densidad de puntos cercanos a la línea de ajuste perfecto, lo que evidencia un buen ajuste entre los valores predichos y los valores reales en la mayoría de los casos. Sin embargo, el modelo Random Forest sobresale por su mayor proximidad a la línea de ajuste, mostrando predicciones consistentemente precisas y una menor dispersión, incluso en los valores extremos.

El modelo Ensemble muestra un rendimiento equilibrado, con una combinación efectiva de las fortalezas de los enfoques KNN, SVR y Lasso. Aunque presenta una mayor dispersión que Random Forest, logra capturar relaciones no lineales con mayor eficacia que el modelo lineal, ofreciendo un desempeño intermedio.

Por otro lado, la regresión lineal muestra una dispersión significativamente mayor, particularmente en los rangos de precios extremos. Aunque su desempeño es comparable al de Random Forest y Ensemble en valores medianos y al subestimar precios, sus errores se vuelven mucho más pronunciados al sobreestimar, especialmente en viviendas de mayor valor. Este comportamiento evidencia las limitaciones del modelo lineal para capturar relaciones complejas entre variables, lo que afecta negativamente su desempeño general.

4.4 Errores Promedios por Rango de Precios

Como se evidenció en el análisis anterior, los modelos presentan diferencias en su capacidad para predecir valores extremos. Para profundizar en este comportamiento, resulta fundamental visualizar cómo varían los errores promedio en función de los rangos de precios. La Figura 7, ilustra esta relación, abarcando desde propiedades de menor valor hasta aquellas con precios más altos. Este enfoque permite identificar patrones específicos y evaluar la precisión de los modelos en distintos segmentos del mercado,

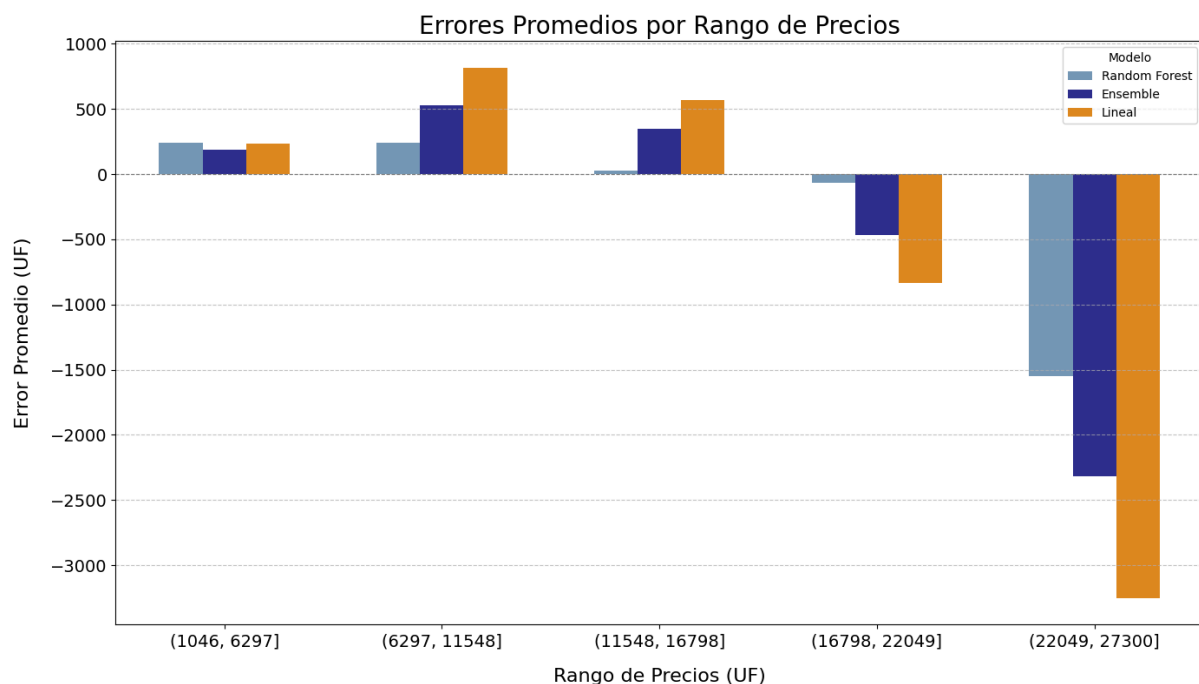


Figura 7. Errores promedio (UF) por rango de precios para los modelos Random Forest, Ensemble y Regresión Lineal.

El análisis de los errores promedio por rango de precios revela diferencias significativas en el desempeño de los modelos seleccionados, particularmente en su capacidad para capturar la verdadera magnitud de los precios. En los rangos iniciales de precios (1.046 a 6.297 UF), todos los modelos exhiben errores promedio cercanos a cero, lo que indica una alta precisión y una baja tendencia a subestimar o sobreestimar. Este comportamiento se observa especialmente en el Ensemble, que muestra un valor más próximo al cero. Esto sugiere que las propiedades de menor valor pueden presentar características más homogéneas o patrones de relación más simples, facilitando su predicción por parte de los modelos.

A medida que los precios aumentan, el comportamiento de los modelos comienza a divergir. En los rangos intermedios, como (6.297 a 11.548 UF), el modelo Random Forest tiende a sobreestimar ligeramente los precios, mientras que el Ensemble muestra una sobrestimación mayor. No obstante, es en los rangos de (11.548 a 16.798 UF) y especialmente en (16.798 a 22.049 UF) es donde el Random Forest demuestra su mejor rendimiento, manteniendo sus errores promedio extremadamente cercanos a cero, lo que indica una notable precisión en estos segmentos de precios medios-altos. El modelo Lineal, por otro lado, evidencia una marcada tendencia a sobreestimar en estos rangos intermedios hasta (16.798 UF)

Sin embargo, en los rangos más altos, esta tendencia se invierte drásticamente para el modelo lineal. En los segmentos de (16.798 a 22.049 UF) y (22.049 a 27.300 UF), el modelo lineal muestra limitaciones significativas al predecir valores, tendiendo a subestimar considerablemente los precios reales. Esto refleja su dependencia de relaciones directas entre las variables, limitando su capacidad para modelar escenarios complejos con interacciones no lineales y valores atípicos de forma adecuada. Por su parte, los modelos Random Forest y Ensemble, aunque también muestran un incremento en la subestimación en los valores más extremos, mantienen un desempeño notablemente superior al modelo lineal, con errores promedio más moderados. Su capacidad para capturar relaciones no lineales y patrones complejos les confiere mayor flexibilidad y precisión en estos segmentos de precios más altos.

4.5 Distribución Geográfica de las Predicciones

Evaluar la distribución geográfica de las predicciones de los modelos, permitirá identificar mejor cómo los modelos capturan las diferencias en los precios de las viviendas en las distintas comunas de la Región Metropolitana, resaltando patrones espaciales específicos y sus limitaciones en determinadas áreas. La Figura 8, muestra los mapas generados con las predicciones de cada modelo.

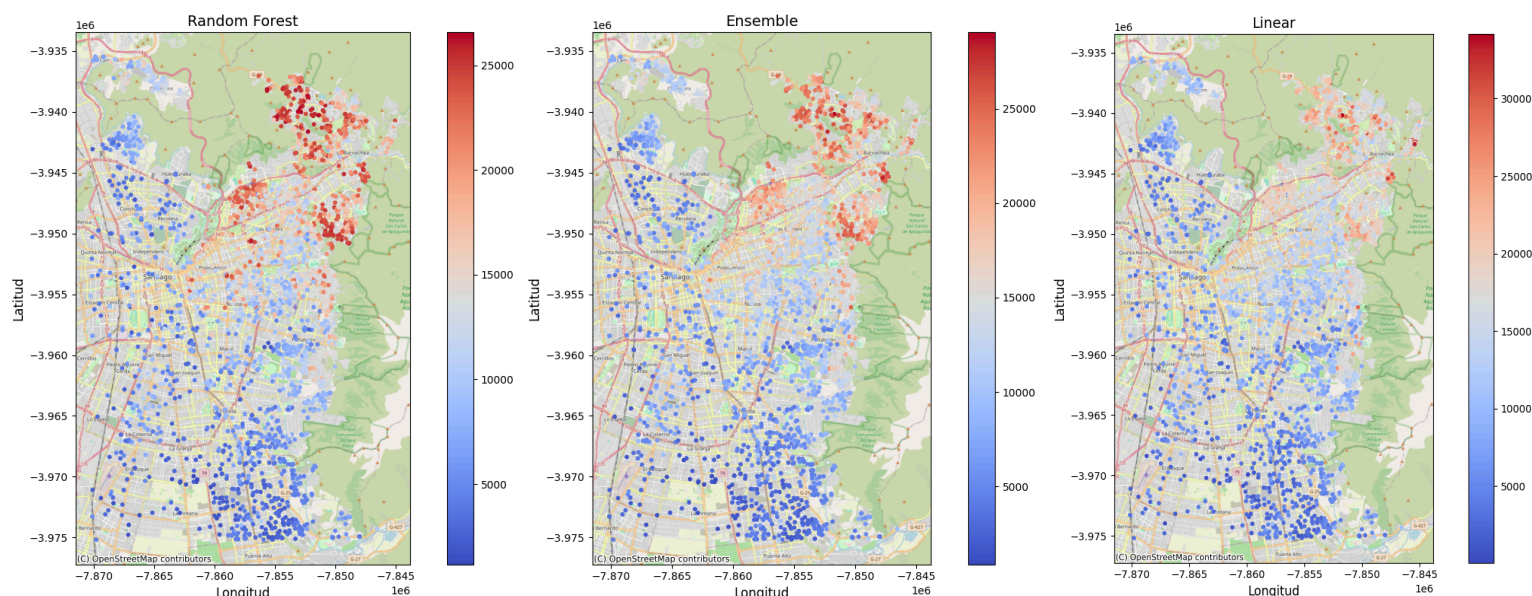


Figura 8. Predicciones geográficas de los modelos Random Forest, Ensemble y Regresión Lineal, representando precios (UF) predichos en la Región Metropolitana, con colores que varían desde tonos azules (precios más bajos) hasta tonos rojos (precios más altos)."

En el caso del modelo Random Forest, se observa un ajuste preciso en las comunas del sector oriente, como Las Condes, Vitacura y Lo Barnechea, donde los altos valores de las propiedades son correctamente capturados, como lo evidencia la concentración de puntos rojos. Asimismo, en las comunas del sector sur y poniente, como Puente Alto y Maipú, el modelo predice consistentemente valores más bajos, lo que refleja su capacidad para adaptarse a las dinámicas del mercado inmobiliario en estas zonas. Este comportamiento reafirma la capacidad del modelo Random Forest para manejar relaciones complejas y no lineales.

El modelo Ensemble también presenta un desempeño sólido en estas áreas, con una distribución de precios similar a la observada en Random Forest. Sin embargo, en comunas intermedias como Vitacura y Providencia, se nota una ligera mayor dispersión en las predicciones, lo que podría atribuirse a la combinación de algoritmos en su estructura, que aporta un enfoque más equilibrado pero algo menos preciso en algunos casos extremos.

Por otro lado, el modelo Lineal muestra limitaciones significativas al predecir valores en comunas con altos precios, como Vitacura y Lo Barnechea, donde tiende a subestimar considerablemente los precios reales. Indicando su dificultad para capturar relaciones complejas y dinámicas en estas áreas. No obstante, es importante destacar que el modelo Lineal, mantiene un rendimiento competente en las comunas con valores intermedios y bajos, donde las relaciones entre las variables son más lineales. Estos resultados reflejan la dependencia del modelo Lineal en relaciones directas entre las variables, limitando su capacidad para modelar escenarios con interacciones no lineales.

En el Anexo 4, se incluyen los gráficos completos que muestran la distribución geográfica de las predicciones generadas por estos modelos.

4.6 Importancia de las variables

El análisis de la importancia de las variables resulta esencial para comprender cómo los modelos seleccionados priorizan las distintas características en la predicción de precios de las viviendas. Este enfoque no solo permite interpretar las relaciones entre las variables y los valores predichos, sino también evaluar la coherencia de los resultados obtenidos con respecto a los factores que tradicionalmente influyen en el mercado inmobiliario. En este contexto, la Figura 9, muestra cómo los valores SHAP (SHapley Additive exPlanations) identifican las variables más relevantes en los modelos, revelando diferencias significativas en la forma en que cada modelo pondera estas variables. Cabe destacar que el orden de las variables en el gráfico sigue la importancia determinada por el modelo Random Forest, dado que este mostró el mejor desempeño general.

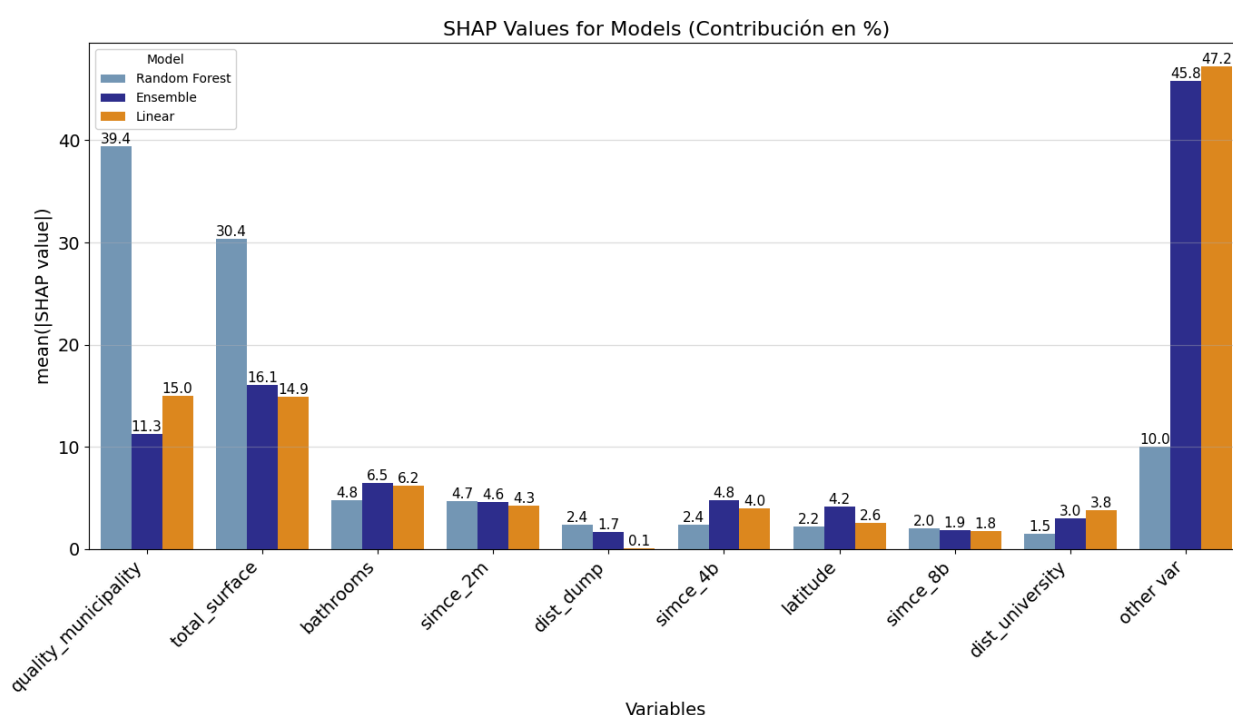


Figura 9. Importancia de las variables según los valores SHAP para los modelos Random Forest, Ensemble y Regresión Lineal, ordenadas según la importancia determinada por Random Forest.

En el modelo Random Forest, el índice de calidad de vida se destaca como la característica más influyente, con una importancia relativa del 39.4%. Esto refleja la prioridad que el modelo asigna a factores extrínsecos relacionados con la calidad del entorno urbano en la predicción del precio de las propiedades, subrayando la relevancia del contexto en la valoración inmobiliaria. Por su parte, la superficie total, con un peso del 30.4%, se posiciona como la característica intrínseca más significativa, evidenciando su estrecha relación con el valor de las viviendas.

El número de baños, con una importancia del 4.8%, aparece también como un factor clave, incluso superando al número de dormitorios. Esto sugiere que el modelo valora más las características funcionales y de comodidad en las viviendas, aspectos que tienen un

impacto notable en la percepción de valor por parte de los compradores. Finalmente, es importante señalar que las características no representadas en el gráfico, o aquellas que no se encuentran entre las nueve más importantes, acumulan un 10% de la importancia total asignada por el modelo. Esto muestra que, aunque el Random Forest concentra gran parte de su peso en un grupo reducido de variables clave, también considera un rango más amplio de características con menor relevancia individual, contribuyendo de manera complementaria al desempeño predictivo del modelo.

En el caso del modelo Ensemble, la superficie total se presenta con un peso relativo del 16.1%, seguida por el índice de calidad de vida municipal, que aporta un 11.3%. Esto evidencia un enfoque más equilibrado del modelo, al combinar tanto características intrínsecas de las propiedades como factores extrínsecos del entorno. El número de baños, aunque relevante, tiene una influencia menor en este modelo, con un aporte del 6.5%, lo que refleja las diferencias en la forma en que las combinaciones de algoritmos del Ensemble capturan las relaciones entre las variables.

Adicionalmente, se destaca que el modelo otorga una importancia significativa a las variables menos influyentes según el modelo Random Forest, acumulando un 45.8% de la contribución total. Este comportamiento sugiere que el Ensemble tiene en cuenta un espectro más amplio de características, integrando factores secundarios que, aunque menos relevantes de forma individual, fortalecen la robustez y capacidad generalizadora del modelo.

Por su parte, el modelo Lineal también muestra una fuerte dependencia de las variables menos consideradas por el Random Forest, que representa el 47.2% de la importancia relativa. En contraste, otras variables clave, como el índice de calidad de vida municipal y la superficie total, tienen una relevancia significativamente menor, con aportes del 15% y 14.9%, respectivamente. Este comportamiento refleja que el modelo lineal prioriza relaciones directas y de gran magnitud, lo cual pone de manifiesto una alta dependencia del modelo hacia un conjunto amplio de características secundarias, lo que, aunque podría enriquecer la representación de datos, también limita su capacidad para captar interacciones más complejas y dinámicas no lineales en el mercado inmobiliario.

4.7 Desempeño en comunas minoritarias.

Este análisis se enfoca en evaluar el desempeño de los modelos en comunas con baja representación de datos en el conjunto de prueba, como Calera de Tango, Curacaví, El Monte, Melipilla, Pirque y San José de Maipo. Esto permite destacar cómo los modelos se comportan en escenarios caracterizados por datos limitados y heterogéneos.

El Anexo 5 presenta un resumen de las métricas de error obtenidas para cada comuna y modelo evaluado. Además, se incluyen gráficos comparativos de las métricas MAE y MAPE, que destacan las diferencias de desempeño entre los modelos en cada comuna.

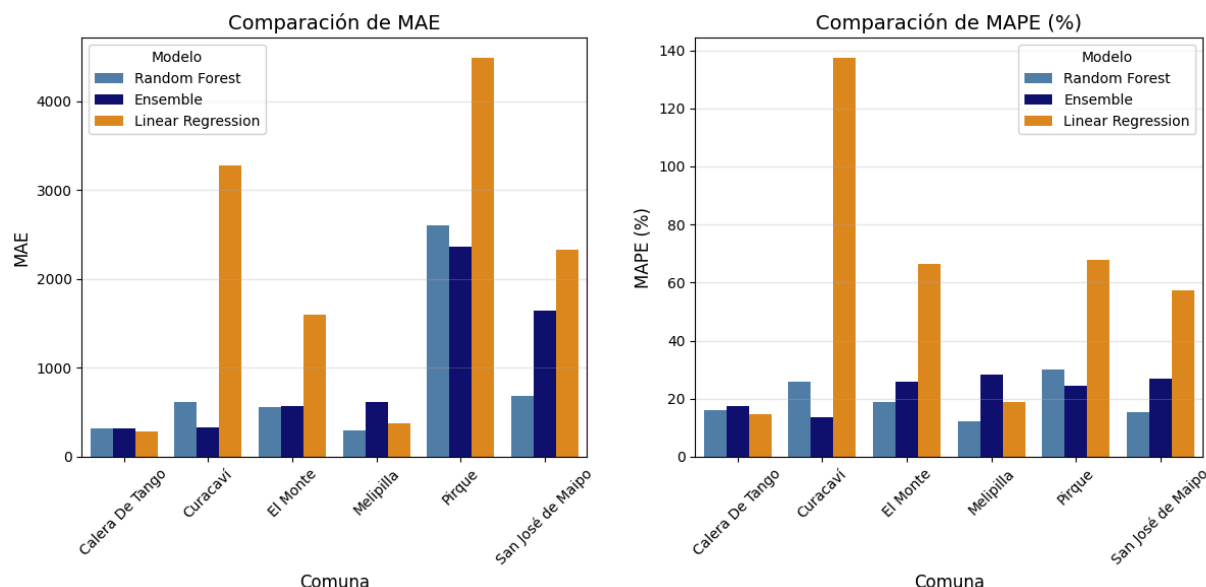


Figura 10. Comparación de las métricas MAE en UF y RMSE, entre los modelos Random Forest, Ensemble y Regresión Lineal para comunas con baja representación, .

El modelo Random Forest demostró ser el más robusto en general, destacando por presentar los menores errores en la mayoría de las comunas analizadas. En Melipilla, obtuvo un MAE de 294.90 y un MAPE de 12.35%, lo que refleja una alta precisión tanto en valores absolutos como relativos al rango de precios observados. Sin embargo, en comunas como Pirque, donde el MAE aumentó a 2609.27 y el MAPE alcanzó un 29.93%, evidencia dificultades en escenarios con menor representación de los datos, la cantidad por comuna se puede observar en el Anexo 2.

El modelo Ensemble presentó un desempeño competitivo, logrando métricas cercanas a las de Random Forest en comunas como Calera de Tango y Curacaví, con MAE de 322.42 y 326.87, respectivamente. En estas comunas, el MAPE fue de 17.33% y 13.71%, lo que indica una adecuada captura de relaciones complejas en contextos de baja representación. Sin embargo, su desempeño en comunas como Melipilla y Pirque mostró mayor variabilidad, con un MAPE de 28.22% y 24.37%, respectivamente, lo que sugiere inconsistencias en su capacidad de generalización.

Por otro lado, el modelo de Regresión Lineal presentó el desempeño más limitado. En comunas como Curacaví, donde solo hubo un registro en el conjunto de prueba, el modelo alcanzó un MAPE de 137.39%, mientras que en Pirque obtuvo un MAPE de 67.83%. Estos resultados reflejan que este enfoque no logra capturar adecuadamente las relaciones complejas y no lineales en los datos minoritarios, resultando en errores significativamente altos tanto en valores absolutos como relativos.

En resumen, los modelos avanzados como Random Forest y Ensemble han mostrado un desempeño significativamente superior en la mayoría de los escenarios, mientras que los enfoques lineales presentan limitaciones evidentes en la captura de relaciones complejas. A continuación, se analiza cómo estos hallazgos se relacionan con las características del mercado inmobiliario y la literatura previa

5. Discusiones

Los resultados obtenidos en esta investigación destacan la relevancia de emplear modelos avanzados para la predicción de precios inmobiliarios, especialmente en un contexto tan diverso y complejo como el de la Región Metropolitana. En esta sección, se interpretan los hallazgos principales, relacionándolos con el marco teórico y los antecedentes revisados, y se analizan las implicancias prácticas, limitaciones y oportunidades para futuras investigaciones.

5.1. Modelos Avanzados versus Modelos Lineales

El modelo Random Forest demostró un desempeño notablemente superior al de la regresión lineal y, en la mayoría de los casos, al del modelo Ensemble, gracias a su capacidad para capturar relaciones complejas y manejar la heterogeneidad presente en los datos. En el conjunto general, alcanzó un MAE promedio de 1045.08 y un R^2 de 0.95, lo que resalta su eficacia como herramienta predictiva robusta en comparación con los otros modelos. Este modelo sobresale incluso en comunas con alta variabilidad, como San José de Maipo, donde logró un MAE de 690.51 y un MAPE de 15.31%, destacando su capacidad para adaptarse a escenarios con una mayor dispersión de precios. Estos hallazgos están alineados con estudios previos, James et al. (2013) y El Mouna et al. (2023), que enfatizan la eficacia de los modelos no lineales en contextos de alta dimensionalidad y baja multicolinealidad.

El análisis de los errores promedios por rangos de precios, presentado en la Figura 7, revela que Random Forest mantiene consistentemente los menores errores en la mayoría de los rangos, lo que evidencia su capacidad para manejar la distribución desigual de precios en el mercado inmobiliario. En comparación, la regresión lineal mostró limitaciones significativas, especialmente en propiedades de alto valor, donde los errores promedios superaron las 3.000 UF. Esto se observa claramente en los gráficos de dispersión de la Figura 6, donde las predicciones del modelo lineal se alejan consistentemente de la línea de ajuste. Este comportamiento pone de manifiesto la incapacidad del modelo lineal para capturar interacciones complejas y patrones no lineales.

En comunas como Melipilla, Random Forest logró un MAPE de 12.35%, destacando su capacidad para ofrecer predicciones precisas incluso en áreas con representación limitada. Sin embargo, en comunas más heterogéneas como Pirque, el MAPE ascendió a 29.93%, lo que refleja el impacto de una mayor diversidad de precios y características en la precisión del modelo. A pesar de esto, el desempeño sigue siendo superior al de la regresión lineal, que en Pirque alcanzó un MAPE superior al 67%, mostrando su vulnerabilidad ante escenarios con alta variabilidad.

El modelo Ensemble, aunque en general competitivo, mostró resultados mixtos en términos de precisión. Por ejemplo, en Melipilla obtuvo un MAPE ligeramente más alto, de 28.22%, en comparación con Random Forest, pero logró mantener un desempeño notablemente mejor que la regresión lineal en comunas heterogéneas. Esto sugiere que el Ensemble

puede ser una alternativa viable en ciertos contextos, aunque su desempeño tiende a ser más sensible a la variabilidad de los datos.

5.2. Desempeño en Comunas con Baja Representación

En comunas con menos de 10 registros, los modelos avanzados enfrentan desafíos particulares debido a la limitada cantidad de datos disponibles. Curacaví, con un solo registro en el conjunto de prueba, representó un caso extremo. En este contexto, el modelo Random Forest mostró un desempeño más estable, mientras que la regresión lineal registró un MAPE alarmante de 137.39%, evidenciando su incapacidad para modelar relaciones no lineales y su dependencia de una mayor cantidad de datos para obtener predicciones más precisas. Este resultado pone de manifiesto la superioridad de los modelos avanzados en escenarios con datos escasos y alta incertidumbre.

En Calera de Tango, una comuna caracterizada por menor heterogeneidad, la regresión lineal obtuvo métricas competitivas, con un MAE de 287.10 y un RMSE de 293.28, lo que sugiere que este modelo puede ser adecuado en contextos con baja variabilidad. Sin embargo, en comunas más heterogéneas, como Pirque. Random Forest logró un MAE de 2609.27 y un MAPE de 29.93%, superando a la regresión lineal, que alcanzó un MAPE superior al 67%, reflejando su incapacidad para adaptarse a la diversidad de características y precios en esta comuna. El modelo Ensemble, por su parte, obtuvo un MAPE competitivo de 17.33% en Calera de Tango.

En Melipilla, Random Forest volvió a destacar con un MAPE de 12.35%, reafirmando su capacidad para modelar patrones complejos incluso en áreas con baja representación de datos. Este desempeño superior demuestra que los modelos no lineales son más aptos para capturar interacciones y relaciones intrincadas en los datos, independientemente de la cantidad de registros disponibles. Por otro lado, el Ensemble, aunque competitivo en algunos contextos, mostró una mayor dispersión con un MAPE de 28.22%, mientras que la regresión lineal amplificó significativamente los errores debido a su sensibilidad ante interacciones no lineales y la ausencia de una relación directa y sencilla entre las variables predictoras y el precio.

Estos resultados subrayan la importancia de utilizar modelos avanzados como Random Forest en comunas con baja representación de datos. Aunque la regresión lineal puede ser útil en áreas con baja variabilidad, su desempeño es significativamente limitado en contextos más complejos. Por el contrario, el Ensemble, aunque menos consistente que Random Forest, puede ser una opción viable cuando se busca un enfoque intermedio que integre múltiples perspectivas, especialmente en comunas con características más homogéneas.

5.3. Importancia de la Interpretabilidad del Modelo

La interpretabilidad de los modelos de machine learning se ha convertido en un aspecto crucial en muchas aplicaciones, y la predicción de precios de viviendas no es una excepción. Si bien los modelos avanzados como Random Forest y Ensemble pueden ofrecer una mayor precisión predictiva, a menudo se les critica por su naturaleza de "caja negra", lo que dificulta la comprensión de cómo llegan a sus predicciones.

En el contexto del mercado inmobiliario, la interpretabilidad es fundamental para generar confianza en los modelos y facilitar su adopción por parte de los diferentes actores involucrados. Los agentes inmobiliarios, los compradores, los vendedores y las instituciones financieras necesitan comprender qué factores están impulsando las predicciones de precios para tomar decisiones informadas y justas.

La falta de interpretabilidad puede generar desconfianza y resistencia al uso de los modelos, especialmente en situaciones donde se requiere transparencia y rendición de cuentas. Por ejemplo, si un modelo predice un valor significativamente diferente al esperado para una propiedad, los interesados querrán saber qué variables están contribuyendo a esta predicción y si estas variables son relevantes y justificables.

Afortunadamente, existen técnicas que pueden ayudar a mejorar la interpretabilidad de los modelos de machine learning. En esta investigación, se utilizó la metodología SHAP (SHapley Additive exPlanations) para analizar la importancia de las variables en las predicciones. Los SHAP values permiten cuantificar la contribución de cada variable a la predicción individual de cada vivienda, lo que proporciona una visión más clara de cómo el modelo está funcionando.

El análisis de los SHAP values reveló que variables como la calidad del municipio, la superficie total y el número de baños tienen una influencia significativa en las predicciones de precios. Esta información puede ser valiosa para los agentes inmobiliarios, ya que les permite comprender qué aspectos de una propiedad son más valorados por el mercado y ajustar sus estrategias de venta en consecuencia.

En resumen, la interpretabilidad de los modelos de machine learning es un factor clave para su adopción y uso responsable en el mercado inmobiliario. Técnicas como SHAP permiten abrir la "caja negra" de los modelos y proporcionar una mayor transparencia en el proceso de predicción, lo que a su vez fomenta la confianza, la equidad y la toma de decisiones informadas.

5.4. Sesgo del Modelo y Representación Desigual de Datos

Un aspecto crítico a considerar en el desarrollo y evaluación de modelos predictivos, especialmente en datasets con distribuciones de datos heterogéneas como el presentado en el Anexo 2, es el potencial sesgo inducido por la baja representación de ciertas comunas. Si bien los modelos avanzados como Random Forest demostraron superioridad en escenarios de datos escasos, la disparidad en la cantidad de registros por comuna introduce desafíos específicos en la generalización y equidad de las predicciones.

En el dataset empleado, la cantidad de registros varía drásticamente entre comunas, desde Las Condes con 2300 registros hasta Curacaví con solo 7. Esta desproporción puede llevar a que los modelos, a pesar de su sofisticación, tiendan a desempeñarse mejor en comunas con una alta densidad de datos, donde han tenido más oportunidades de aprender patrones complejos y representativos. Por el contrario, en comunas con muy pocos registros, el modelo puede no haber aprendido lo suficiente sobre las características y dinámicas de precios específicas de esas áreas. Esto se traduce en una mayor incertidumbre y errores para estas comunas, a pesar de los valores promedio reportados en el conjunto general, ya que el modelo podría estar generalizando patrones aprendidos de comunas con más datos, que no necesariamente aplican a las características únicas de las zonas subrepresentadas.

Además, esta distribución desigual puede generar un sesgo implícito en las predicciones. El modelo podría sesgar sus resultados hacia las características predominantes de las comunas mejor representadas. Por ejemplo, si la mayoría de los datos provienen de comunas de alto valor, el modelo podría subestimar o sobrestimar los precios en comunas con características socioeconómicas distintas y con menos registros, incluso si estas diferencias no están explícitamente capturadas por las variables predictoras. Este sesgo puede llevar a injusticias o a decisiones erróneas en el mercado inmobiliario, afectando a compradores y vendedores de estas áreas. La interpretabilidad, si bien se abordó con SHAP en la sección 5.3, también puede verse afectada en este contexto. Si el modelo no ha aprendido adecuadamente las interacciones en una comuna con pocos datos, la explicación SHAP para esas predicciones podría ser menos fiable o reflejar patrones más globales del dataset que patrones locales.

Para mitigar este sesgo en futuras investigaciones, se podrían explorar diversas estrategias. La solución ideal, aunque no siempre factible, sería la recopilación de más datos en las comunas subrepresentadas. Otra alternativa es asignar mayor peso a los errores de predicción en comunas con menor cantidad de registros durante el entrenamiento, lo que permitiría que el algoritmo preste más atención a su rendimiento en estas áreas. También se podría considerar el desarrollo de modelos específicos para grupos de comunas con características similares y suficiente representación, o incluso enfoques de transfer learning, donde un modelo pre-entrenado en un dataset grande se afina con los datos escasos de las comunas subrepresentadas.

6. Recomendaciones

Si bien los resultados obtenidos en esta investigación son prometedores, es fundamental reconocer que el estudio no está exento de limitaciones. Estas limitaciones, lejos de disminuir el valor de los hallazgos, abren un abanico de oportunidades para futuras investigaciones que busquen fortalecer las metodologías empleadas, ampliar el alcance de los resultados y explorar mejoras adicionales en el campo de la predicción de precios de viviendas.

Una de las principales limitaciones identificadas se relaciona con la baja representación de ciertas comunas dentro del conjunto de datos analizado. Esta disparidad en la cantidad de información disponible para cada comuna puede afectar la estabilidad y la capacidad de generalización de los modelos predictivos, especialmente en aquellas áreas con menor cantidad de datos. Por otro lado, la ausencia de variables temporales en el análisis constituye otra limitación importante. La dinámica del mercado inmobiliario está influenciada por una variedad de factores temporales, como las fluctuaciones estacionales, los ciclos económicos y las variaciones en las tasas de interés, que pueden tener un impacto significativo en los precios de las viviendas. Al no considerar estas variables temporales, los modelos pueden perder la capacidad de capturar tendencias dinámicas importantes, lo que limita su precisión y aplicabilidad en escenarios reales.

En este contexto, las siguientes recomendaciones se proponen como posibles vías para optimizar el desempeño de los modelos y mejorar su capacidad predictiva en escenarios con desafíos específicos, como la baja representatividad de los datos o la alta heterogeneidad del mercado inmobiliario:

1. **Incorporar técnicas de generación de muestras sintéticas:** : Para abordar el problema de la baja representación de ciertas comunas, se sugiere explorar la utilización de técnicas de generación de muestras sintéticas. Métodos como SMOTE (Synthetic Minority Oversampling Technique) o la generación de datos sintéticos con modelos probabilísticos pueden ser empleados para aumentar la cantidad de datos disponibles para las comunas con menor representación. Esto permitiría mejorar la estabilidad de los modelos predictivos en áreas periféricas y heterogéneas (Chawla et al., 2002), donde la escasez de datos puede ser un factor limitante.
2. **Evaluar el impacto de variables temporales:** Se recomienda integrar variables relacionadas con la temporalidad en los modelos predictivos. Factores como los ciclos económicos, la estacionalidad y las variaciones en las tasas de interés pueden tener una influencia significativa en los precios inmobiliarios, y su inclusión en los modelos permitiría capturar de manera más precisa las dinámicas del mercado. Esto podría lograrse mediante la utilización de modelos basados en series temporales o enfoques híbridos.

3. **Explorar técnicas avanzadas de modelado:** Si bien los modelos Random Forest y Ensemble demostraron un buen desempeño en esta investigación, se sugiere explorar la utilización de técnicas de modelado aún más avanzadas. La comparación del desempeño de Random Forest y Ensemble con redes neuronales profundas y arquitecturas como Long Short-Term Memory (LSTM), que son especialmente efectivas para manejar datos con dependencias temporales, podría arrojar resultados interesantes. Asimismo, la incorporación de enfoques como Gradient Boosting Machines (GBM) y CatBoost, conocidos por su capacidad para manejar grandes volúmenes de datos y capturar interacciones complejas, podría ser beneficiosa para evaluar su desempeño en comparación con los modelos utilizados en este estudio.
4. **Validación con datos de otras regiones:** Para evaluar la capacidad de generalización de los modelos en contextos diversos, se recomienda realizar una validación con conjuntos de datos independientes o provenientes de otras regiones geográficas. Esto permitiría determinar si los modelos son capaces de mantener su precisión y robustez al ser aplicados a datos diferentes de aquellos con los que fueron entrenados.
5. **Implementar Stratified Cross-Validation:** En escenarios con datos desbalanceados, como las comunas con baja representación, se sugiere implementar la técnica de Stratified Cross-Validation para evaluar los modelos. Esta técnica asegura que las proporciones de las clases o grupos de interés se mantengan constantes en cada partición de entrenamiento y prueba, lo que mejora la representatividad de las métricas de evaluación y reduce el sesgo en la comparación entre modelos. Esto es especialmente relevante para fortalecer las predicciones en contextos heterogéneos y garantizar una evaluación más precisa del desempeño de los modelos.
6. **Realizar ingeniería de características:** Se propone aplicar técnicas de ingeniería de características para diseñar y seleccionar nuevas variables que puedan mejorar la capacidad predictiva de los modelos. Este proceso implica transformar, combinar y derivar nuevas variables a partir de las existentes, con el objetivo de capturar relaciones complejas e interacciones entre variables que no son evidentes en el conjunto de datos original. La ingeniería de características puede ser una herramienta poderosa para optimizar el desempeño de los modelos y revelar patrones ocultos en los datos, lo que podría conducir a predicciones más precisas y robustas.

7. Implicaciones Prácticas

Los resultados de esta investigación tienen implicaciones significativas para el sector inmobiliario, el mercado financiero, el desarrollo tecnológico y la planificación urbana. La capacidad de los modelos de machine learning para procesar grandes volúmenes de datos y detectar patrones ocultos proporciona una herramienta poderosa para optimizar procesos de evaluación, mejorar la toma de decisiones y diseñar estrategias comerciales más precisas. A continuación, se detallan diversas aplicaciones y oportunidades que surgen a partir de estos modelos predictivos.

Valuación de Propiedades y Estrategias Comerciales

Los modelos de machine learning han demostrado ser herramientas clave en la optimización de la valuación inmobiliaria, permitiendo establecer precios más precisos y reduciendo la incertidumbre en las transacciones. La estimación automatizada de precios facilita la toma de decisiones tanto para vendedores como compradores, proporcionando una evaluación basada en datos históricos y tendencias del mercado en tiempo real. Además, al integrar variables económicas y geográficas en la predicción, estos modelos pueden identificar variaciones específicas según la ubicación y características de cada propiedad.

En el sector inmobiliario, el uso de estos modelos puede mejorar significativamente la segmentación del mercado. En comunas como Vitacura y Lo Barnechea, donde los valores de las propiedades son más estables, las predicciones pueden proporcionar información confiable sobre la evolución de los precios. Sin embargo, en zonas con mayor volatilidad, como Santiago Centro o Independencia, la capacidad de estos modelos para detectar fluctuaciones y tendencias resulta clave para ajustar estrategias de comercialización y establecer precios más competitivos. Además, estos modelos pueden facilitar estrategias de venta dinámicas, donde los precios se ajusten en función de la demanda en tiempo real, similar a la fijación de precios en la industria del turismo y el comercio electrónico.

Otra aplicación relevante es la estimación del valor de una propiedad en base a su potencial de renovación o remodelación. Utilizando datos históricos sobre la apreciación de viviendas tras mejoras estructurales, los modelos pueden sugerir qué tipos de remodelaciones generarían mayor retorno de inversión, brindando información valiosa tanto a propietarios como a desarrolladores inmobiliarios.

Identificación de Oportunidades Comerciales y Gestión de Activos

El mercado inmobiliario está en constante evolución, lo que abre la posibilidad de identificar oportunidades estratégicas a través del análisis avanzado de datos. Los modelos de machine learning permiten detectar patrones en la fluctuación de precios, la oferta y demanda de viviendas, y la valorización de determinadas zonas geográficas, facilitando la toma de decisiones informadas para inversionistas y administradores de activos.

Uno de los principales usos de estos modelos es la identificación de propiedades subvaloradas o con alto potencial de apreciación. Los algoritmos pueden predecir qué áreas tienen mayores probabilidades de experimentar incrementos de valor en el corto o mediano plazo. Factores como el crecimiento de la infraestructura, el desarrollo de nuevos polos comerciales o mejoras en la conectividad pueden influir en la plusvalía de una zona, y los modelos de machine learning pueden detectar estos patrones antes de que se reflejen en los precios del mercado.

Otra aplicación clave es la optimización de carteras de inversión inmobiliaria. Los inversionistas pueden emplear modelos predictivos para evaluar el desempeño de distintos tipos de propiedades y definir estrategias de diversificación de activos. Por ejemplo, mediante análisis de clustering, los modelos pueden segmentar mercados en función de su rentabilidad y estabilidad, permitiendo una mejor distribución de inversiones entre zonas consolidadas y mercados emergentes. Asimismo, el uso de algoritmos de refuerzo permite ajustar estrategias de compra y venta en tiempo real, maximizando la rentabilidad de cada transacción.

En el sector corporativo, las empresas inmobiliarias pueden utilizar machine learning para mejorar la administración de activos y la gestión de portafolios. Modelos avanzados pueden predecir la depreciación de propiedades y recomendar intervenciones estratégicas para mantener su competitividad en el mercado. Además, la integración de datos de mantenimiento, ocupación y rentabilidad de los inmuebles en los modelos permite optimizar la toma de decisiones sobre cuándo vender, remodelar o reposicionar un activo dentro del portafolio de la empresa.

Asimismo, la detección de tendencias emergentes en la demanda permite a las empresas ajustar su oferta en función de las expectativas del mercado. Por ejemplo, si los modelos detectan un aumento en la preferencia por viviendas con espacios de coworking o áreas verdes, las inmobiliarias pueden adaptar sus proyectos para alinearse con estas tendencias y mejorar su competitividad.

Otro aspecto relevante es el impacto de factores macroeconómicos en la gestión de activos inmobiliarios. La capacidad de machine learning para integrar datos de indicadores económicos, tasas de interés y políticas gubernamentales permite anticipar posibles cambios en el mercado y ajustar estrategias de inversión en consecuencia. Esto es especialmente útil en períodos de incertidumbre económica, donde la toma de decisiones basada en datos puede mitigar riesgos y mejorar la resiliencia de los portafolios de inversión.

Apoyo a Políticas Públicas y Desarrollo Urbano

El uso de modelos de machine learning en la planificación urbana y en la formulación de políticas públicas está transformando la manera en que los gobiernos gestionan el crecimiento de las ciudades y la distribución de recursos. A partir del análisis avanzado de datos, estos modelos permiten tomar decisiones basadas en evidencia, optimizar el desarrollo de infraestructura y mejorar la equidad en el acceso a la vivienda.

Uno de los principales aportes de estas tecnologías es la identificación de zonas en crecimiento y la predicción de la evolución del mercado inmobiliario en diferentes sectores urbanos. Mediante el análisis de datos históricos de precios, patrones de migración, accesibilidad a servicios y proyección del desarrollo de infraestructuras, los modelos pueden ayudar a los gobiernos a planificar el uso del suelo de manera más eficiente, asegurando un desarrollo equilibrado de las ciudades y evitando la sobreconcentración de inversiones en ciertas áreas en detrimento de otras.

Asimismo, los modelos de machine learning permiten la detección temprana de zonas con potencial riesgo de gentrificación. A través del análisis de tendencias en el incremento de precios, cambios en la composición demográfica y aumento en la inversión inmobiliaria, es posible anticipar desplazamientos de poblaciones vulnerables y diseñar políticas que mitiguen estos efectos, como la regulación de arriendos o la implementación de incentivos para la preservación de viviendas accesibles.

Otra aplicación clave es la optimización en la asignación de infraestructura y servicios públicos. A partir de modelos predictivos, los gobiernos pueden determinar en qué zonas es más urgente la construcción de hospitales, escuelas, transporte público o espacios recreativos. Esto garantiza que los recursos se distribuyan de manera más equitativa y que las inversiones tengan un impacto significativo en la calidad de vida de los ciudadanos. Por ejemplo, al cruzar datos de densidad poblacional, acceso a transporte y costos de vivienda, es posible identificar barrios con déficit de conectividad y planificar la construcción de nuevas estaciones de metro o buses.

Además, la capacidad de estos modelos para integrar múltiples fuentes de información, como imágenes satelitales, sensores urbanos y registros administrativos, permite un monitoreo en tiempo real del estado de la ciudad. Esto facilita la implementación de políticas dinámicas que se ajusten a los cambios en las necesidades de la población y al impacto de eventos externos, como crisis económicas, desastres naturales o pandemias.

En el ámbito de la regulación del mercado inmobiliario, los modelos de machine learning pueden ser utilizados para prevenir la especulación excesiva y el uso indebido de terrenos urbanos. A través del análisis de transacciones inmobiliarias, es posible detectar compras masivas de propiedades con fines especulativos y diseñar estrategias para desincentivar estas prácticas, promoviendo una mayor estabilidad en los precios y asegurando que el acceso a la vivienda no se vea limitado por intereses especulativos.

Por último, el uso de machine learning en la planificación urbana no solo permite un desarrollo más ordenado de las ciudades, sino que también facilita la transición hacia modelos de urbanismo sostenible. Al evaluar variables ambientales, como el impacto de la contaminación, el consumo energético y la accesibilidad a espacios verdes, estos modelos pueden contribuir a la formulación de estrategias para reducir la huella ecológica de las ciudades y mejorar la resiliencia urbana frente al cambio climático.

Mercado Emergente y Expansión Comercial

Los mercados emergentes presentan desafíos únicos en términos de acceso a información confiable, estabilidad económica y regulación del sector inmobiliario. En estos contextos, la aplicación de modelos de machine learning puede marcar una diferencia significativa al mejorar la transparencia, reducir la incertidumbre y facilitar la toma de decisiones informadas.

Uno de los principales beneficios del uso de machine learning en mercados emergentes es la capacidad de construir bases de datos más completas y precisas. En muchos países en desarrollo, la información sobre transacciones inmobiliarias es fragmentada y carece de estandarización, lo que dificulta la evaluación de precios y el análisis de tendencias. Mediante el uso de técnicas de aprendizaje automático, es posible recopilar, limpiar y estructurar estos datos para generar modelos de predicción de precios más confiables y detallados. Esto permite que los compradores y vendedores accedan a información más precisa y reduzcan el riesgo de sobrevaloración o subvaloración de propiedades.

Además, los modelos de machine learning pueden ser utilizados para detectar patrones de crecimiento en mercados con alta volatilidad. En economías emergentes, factores como cambios políticos, inversión extranjera y desarrollo de infraestructura pueden influir rápidamente en la valorización de ciertas áreas. Al analizar grandes volúmenes de datos económicos y demográficos, los modelos pueden prever la expansión de nuevos polos comerciales, identificar zonas con alto potencial de revalorización y recomendar estrategias de inversión basadas en datos cuantificables. Esto es particularmente valioso para desarrolladores inmobiliarios y fondos de inversión que buscan maximizar su rentabilidad en mercados con alto potencial de crecimiento.

Otro aspecto clave es la mitigación de riesgos en mercados con fluctuaciones económicas pronunciadas. En muchas regiones, las crisis financieras y las variaciones en las tasas de interés afectan directamente la estabilidad del sector inmobiliario. Los modelos predictivos pueden analizar datos históricos y factores macroeconómicos para anticipar ciclos de recesión y expansión, permitiendo a los inversionistas y empresas inmobiliarias ajustar sus estrategias en función de escenarios futuros. Esta capacidad de análisis prospectivo ayuda a minimizar pérdidas y optimizar el tiempo de entrada y salida de inversiones en estos mercados.

En términos de expansión comercial, el machine learning también juega un papel clave en la identificación de oportunidades para la internacionalización de empresas del sector inmobiliario. A través del análisis comparativo de distintos mercados, los modelos pueden evaluar qué regiones ofrecen condiciones más favorables para la inversión, considerando factores como estabilidad política, demanda habitacional, costo de la tierra y potencial de crecimiento. Esto permite a las empresas planificar su expansión de manera más estratégica, optimizando la asignación de recursos y reduciendo la exposición a mercados con alta incertidumbre.

Impacto en el Sector Financiero y Bancario

La implementación de modelos de machine learning en el sector financiero y bancario ha generado avances significativos en la evaluación de riesgos, la optimización de productos financieros y la mejora en la estabilidad del crédito hipotecario. La capacidad de estos modelos para procesar grandes volúmenes de datos y detectar patrones ocultos ha permitido a las instituciones financieras tomar decisiones más informadas, minimizar riesgos y diseñar productos más adaptados a las necesidades del mercado inmobiliario.

Uno de los principales beneficios de la aplicación de machine learning en el sector bancario es la mejora en la evaluación crediticia. Tradicionalmente, los bancos han utilizado modelos estadísticos para determinar la solvencia de los solicitantes de crédito, basándose en información como el historial financiero, la relación deuda-ingreso y el puntaje de crédito. Sin embargo, estos enfoques presentan limitaciones al no considerar un análisis más profundo de las dinámicas del mercado y las condiciones económicas. Con machine learning, es posible construir modelos que integren múltiples fuentes de información, como el comportamiento de pago en otros sectores financieros, patrones de gasto y tendencias macroeconómicas, permitiendo una evaluación más precisa del riesgo crediticio.

Además, el uso de algoritmos predictivos en la concesión de hipotecas permite anticipar el comportamiento de los clientes y reducir la probabilidad de impago. Al analizar datos históricos de morosidad y la evolución del mercado inmobiliario, los bancos pueden ajustar sus políticas de préstamo, estableciendo tasas de interés personalizadas y condiciones de pago que reflejen el riesgo real de cada solicitante. Esto no solo optimiza la asignación de créditos, sino que también mejora la estabilidad financiera al reducir la exposición a crisis hipotecarias.

El impacto del machine learning en el sector financiero también se extiende a la detección de fraudes en transacciones inmobiliarias. La automatización de la detección de patrones sospechosos en la compra y venta de propiedades, así como en la transferencia de activos, ha permitido a los bancos y organismos reguladores mejorar sus mecanismos de prevención de lavado de dinero y fraude hipotecario. A través del análisis de comportamientos anómalos y la identificación de redes de transacciones irregulares, las instituciones financieras pueden reforzar la seguridad del sistema y garantizar una mayor transparencia en las operaciones inmobiliarias.

Asimismo, la personalización de productos financieros se ha visto beneficiada por la implementación de machine learning en el análisis de clientes. A través del procesamiento de datos sobre las preferencias y necesidades de los consumidores, los bancos pueden ofrecer productos hipotecarios más ajustados a cada perfil, mejorando la experiencia del cliente y aumentando la accesibilidad al crédito. Por ejemplo, modelos de segmentación pueden identificar clientes con alta probabilidad de adquirir una propiedad en los próximos años y ofrecerles opciones de financiamiento con condiciones adaptadas a su capacidad de pago y perfil de riesgo.

Desarrollo de Herramientas Tecnológicas

El avance de los modelos de machine learning ha impulsado el desarrollo de herramientas tecnológicas que transforman la manera en que se gestionan, analizan y comercializan los bienes raíces. Estas herramientas han permitido automatizar procesos complejos, mejorar la precisión en la toma de decisiones y facilitar el acceso a información en tiempo real tanto para compradores como para inversionistas y desarrolladores inmobiliarios.

Uno de los desarrollos más destacados es la integración de machine learning en plataformas digitales de valuación y comercialización de propiedades. A través de algoritmos que analizan datos históricos de precios, características del inmueble y condiciones del mercado, estas plataformas pueden ofrecer estimaciones precisas en cuestión de segundos. Además, permiten a los usuarios comparar propiedades similares, predecir tendencias de precios y recibir recomendaciones personalizadas sobre oportunidades de inversión, optimizando así el proceso de compra y venta de bienes raíces.

Otro avance clave es la implementación de sistemas de análisis geoespacial, que utilizan machine learning para evaluar patrones de crecimiento urbano y detectar zonas con potencial de revalorización. Estos sistemas combinan datos de ubicación, infraestructura, transporte, accesibilidad y actividad económica para ofrecer mapas dinámicos que ayudan a inversionistas y urbanistas a tomar decisiones fundamentadas sobre el desarrollo inmobiliario. Gracias a estas herramientas, es posible predecir el impacto de nuevos proyectos en la valorización de una zona, facilitando estrategias de inversión y planificación a largo plazo.

En el ámbito de la gestión inmobiliaria, los modelos de machine learning han permitido la creación de sistemas de mantenimiento predictivo para edificios y complejos habitacionales. A través del análisis de datos en tiempo real provenientes de sensores IoT (Internet of Things), estas herramientas pueden identificar fallas en la infraestructura antes de que ocurran problemas graves, optimizando los costos de mantenimiento y prolongando la vida útil de los activos inmobiliarios. Este enfoque no solo reduce gastos operativos, sino que también mejora la seguridad y eficiencia energética de los inmuebles.

Además, el uso de visión por computadora en el sector inmobiliario ha revolucionado la forma en que se evalúan y promocionan las propiedades. Mediante el análisis automatizado de imágenes y videos, los modelos de machine learning pueden extraer información detallada sobre el estado de una vivienda, identificar características clave y detectar daños estructurales. Esta tecnología es particularmente útil en tasaciones remotas y en la optimización de anuncios inmobiliarios, ya que permite generar descripciones más precisas y atractivas basadas en un análisis objetivo de las imágenes.

En el ámbito de las inversiones inmobiliarias, han surgido plataformas que emplean machine learning para facilitar la tokenización de activos inmobiliarios, permitiendo la fragmentación de propiedades en unidades más accesibles para inversionistas individuales. Estas plataformas analizan datos de mercado y establecen modelos de rentabilidad basados en predicciones de demanda y apreciación del inmueble, democratizando el acceso a la inversión inmobiliaria y fomentando la liquidez del mercado.

Adicionalmente, la combinación de machine learning con realidad aumentada (AR) y realidad virtual (VR) ha mejorado la experiencia de compra y venta de propiedades. Mediante simulaciones interactivas, los usuarios pueden explorar virtualmente inmuebles en diferentes ubicaciones, personalizar diseños y evaluar detalles arquitectónicos sin necesidad de visitas presenciales. Estas herramientas han demostrado ser especialmente valiosas en el contexto del comercio internacional y en la promoción de desarrollos inmobiliarios en fases tempranas de construcción.

Desafíos y Consideraciones Éticas

Si bien el uso de machine learning en el sector inmobiliario ha generado avances significativos en la precisión de las predicciones y la optimización de estrategias comerciales, también plantea desafíos importantes que deben ser abordados para garantizar un desarrollo ético y equitativo de estas tecnologías. La transparencia en la toma de decisiones algorítmicas, la protección de datos personales y la equidad en el acceso a la vivienda son algunas de las cuestiones clave que requieren atención.

Uno de los principales desafíos es la transparencia y explicabilidad de los modelos. Muchos algoritmos de machine learning, especialmente aquellos basados en redes neuronales profundas, funcionan como “cajas negras”, donde los resultados son difíciles de interpretar. Esto puede generar desconfianza entre compradores, vendedores y reguladores, ya que no siempre es claro cómo se han calculado los valores de las propiedades o qué factores han influido en las predicciones. Para mitigar este problema, es necesario desarrollar métodos de interpretabilidad que permitan comprender las decisiones del modelo y justificar sus predicciones de manera clara y accesible.

Otro aspecto crítico es la privacidad y seguridad de los datos. Los modelos de machine learning requieren grandes volúmenes de información para entrenarse, incluyendo datos sensibles sobre transacciones inmobiliarias, ubicaciones de propiedades y patrones de comportamiento de los usuarios. Sin un marco adecuado de protección, existe el riesgo de filtraciones de información o de un uso indebido de los datos recopilados. La implementación de técnicas como el aprendizaje federado y el encriptado de datos es fundamental para garantizar la confidencialidad y el cumplimiento de normativas de privacidad como la Ley de Protección de Datos Personales (BCN, 2024).

Además, el uso de machine learning en la valuación de propiedades y en la concesión de créditos hipotecarios puede amplificar sesgos históricos y desigualdades preexistentes. Si los modelos se entrenan con datos que reflejan patrones de discriminación en la asignación de préstamos o en la valorización de inmuebles en ciertas zonas, las predicciones resultantes podrían perpetuar estas injusticias. Para evitarlo, es crucial realizar auditorías constantes de los modelos, implementar técnicas de mitigación de sesgos y garantizar que los datos utilizados sean representativos de la diversidad del mercado inmobiliario.

Un desafío adicional es la regulación y control del uso de estas tecnologías. A medida que el machine learning se vuelve más influyente en la determinación de precios y en la gestión de activos inmobiliarios, los organismos reguladores deben establecer normativas que aseguren un uso responsable de los modelos. Esto incluye la definición de estándares para

la transparencia de los algoritmos, la protección de los derechos de los consumidores y la prevención de la manipulación del mercado a través de estrategias automatizadas.

Asimismo, la accesibilidad a estas herramientas es un tema que debe ser abordado con especial atención. Aunque la implementación de machine learning en el sector inmobiliario ha beneficiado a grandes desarrolladores e inversionistas, existe el riesgo de que pequeñas empresas y compradores individuales no tengan el mismo acceso a estas tecnologías. Esto podría generar una concentración de ventajas competitivas en pocos actores, aumentando la brecha en el acceso a oportunidades de inversión. Para contrarrestar este problema, es necesario promover el desarrollo de plataformas abiertas y accesibles, que permitan a una mayor cantidad de usuarios beneficiarse de estas innovaciones.

De igual forma, la interacción entre machine learning y la sostenibilidad urbana representa un reto emergente. Si bien los modelos pueden contribuir a una mejor planificación del desarrollo urbano, también podrían incentivar la sobreexplotación de ciertas zonas o la especulación inmobiliaria en áreas con potencial de revalorización. Es fundamental que las predicciones generadas por estos modelos se utilicen para fomentar un crecimiento equilibrado y sostenible, evitando impactos negativos en el medio ambiente y en las comunidades locales.

A continuación, se examinarán dos casos emblemáticos, Redfin y Zillow, que ilustran vívidamente cómo la integración de machine learning ha redefinido el panorama inmobiliario. La implementación de estas tecnologías ha transformado no solo la valuación de propiedades y la gestión de activos, sino también la dinámica del mercado, ofreciendo lecciones valiosas sobre los éxitos y complejidades inherentes a su aplicación en un entorno real.

Caso 1: Redfin y la Transformación del Mercado Inmobiliario a Través de Machine Learning

Redfin, una agencia inmobiliaria de tecnología, se ha posicionado como un ejemplo destacado de la aplicación exitosa de machine learning para revolucionar el sector inmobiliario. Fundada por ingenieros de software, la compañía ha optimizado el proceso de compraventa de viviendas y se ha convertido en el sitio web de agencias inmobiliarias más grande de la web. Redfin ha integrado la inteligencia artificial en varios aspectos de su plataforma, siendo sus algoritmos de estimación y recomendación dos de las funciones basadas en machine learning más destacadas.

Redfin Estimate: Valuación de Propiedades Impulsada por Datos

Redfin Estimate es un producto central que calcula el valor de mercado de cualquier vivienda, utilizando un algoritmo de machine learning que analiza exhaustivamente diversas variables. El modelo considera más de 500 puntos de datos para cada cálculo, abarcando información extraída de los servicios de listado múltiple (MLS), datos geográficos y la participación de los usuarios en Redfin.com.

Inicialmente, Redfin Estimate se concibió como una herramienta para atraer a los usuarios, permitiéndoles obtener estimaciones de valor de sus viviendas de forma gratuita. Sin embargo, su utilidad se ha expandido significativamente, convirtiéndose en un activo esencial.

Redfin ha demostrado un compromiso continuo con la mejora de la precisión de Redfin Estimate. En 2019, la compañía anunció una optimización en la exactitud de su estimación para viviendas fuera del mercado, reduciendo el error medio del 6.3% al 5.6%. Esta mayor precisión no solo beneficia a los usuarios al proporcionarles información más confiable.

Redfin Recommendations: Personalización Inteligente de la Búsqueda

Redfin Recommendations es otro componente clave de la plataforma, impulsado por machine learning y responsable de una parte importante del tráfico del sitio web. Este sistema inteligente ofrece listas de viviendas recomendadas a los clientes, tanto en el sitio web como a través de correos electrónicos personalizados y notificaciones push, alertándolos sobre nuevas propiedades que se ajustan a sus preferencias.

Un hallazgo interesante, destacado por Akshat Kaul (2021), es la capacidad del algoritmo de recomendaciones para superar las propias búsquedas guardadas de los usuarios en la predicción de las viviendas que realmente les interesarán. Esto ilustra el poder del machine learning para descubrir patrones de comportamiento y preferencias latentes, proporcionando recomendaciones más precisas y relevantes que mejoran la experiencia del usuario.

Infraestructura Tecnológica y Enfoque en los Datos

Se ha distinguido por su enfoque en la tecnología y los datos, lo que le permite desarrollar soluciones innovadoras y eficientes. La empresa utiliza una combinación de modelos clásicos y de aprendizaje profundo para sus diversas aplicaciones, incluyendo el modelo de Redfin Estimate, que emplea técnicas como el Random Forest, el aumento de gradiente, modelos de ensemble y modelos jerárquicos.

La compañía se nutre de una amplia variedad de fuentes de datos, que abarcan desde los MLS y datos geográficos hasta la información sobre la participación de los usuarios en su plataforma. Redfin también ha realizado inversiones significativas en el desarrollo de capacidades para extraer información valiosa de datos no estructurados, como documentos de transacciones inmobiliarias, con el objetivo de enriquecer aún más sus modelos y servicios.

Ética, Responsabilidad e Impacto en el Ecosistema Inmobiliario

Redfin reconoce la importancia de la IA responsable y se esfuerza por garantizar la equidad y transparencia en sus algoritmos. La empresa excluye explícitamente los datos sobre delincuencia de sus análisis para evitar perpetuar sesgos y desigualdades sociales.

Al mismo tiempo, ha declarado que comprenden el impacto potencial de sus algoritmos en el mercado inmobiliario, ya que las decisiones de compra de los usuarios pueden estar

influenciadas por las recomendaciones de la plataforma. Por lo tanto, la compañía se compromete a mantener la transparencia en el funcionamiento de sus modelos y a trabajar proactivamente para mitigar cualquier efecto negativo no deseado.

Redfin representa un caso ejemplar de cómo el machine learning puede transformar la experiencia de búsqueda de vivienda, optimizar la valuación de propiedades y mejorar la eficiencia operativa en el sector inmobiliario, al tiempo que se abordan consideraciones éticas y se reconoce el impacto en el mercado.

Caso 2: Zillow y los Desafíos de la Implementación de Machine Learning en un Mercado Volátil

Zillow, un destacado mercado inmobiliario en línea, sufrió una significativa debacle con el cierre de su negocio Zillow Offers, impulsado por algoritmos de iBuying. Un algoritmo defectuoso en la valoración de propiedades obligó a la compañía a reducir el valor estimado de las viviendas adquiridas en el tercer y cuarto trimestre de 2021 en más de 500 millones de dólares, lo que, para mitigar el impacto, llevó a la previsión de una reducción del 25% de su plantilla.

Sobreestimación de Valores y Expansión Agresiva

El problema central radicó en que los algoritmos de Zillow sobreestimaron el valor de las viviendas que adquirieron. Simultáneamente, Zillow expandió agresivamente su programa de compras, adquiriendo más viviendas en los dos últimos trimestres que en los dos años anteriores. Esta combinación de sobrevaloración y expansión acelerada resultó en que la compañía se viera obligada a vender grandes cantidades de viviendas a precios inferiores al de compra, debido al alto costo de mantener las casas vacías con la esperanza de una recuperación del mercado.

La Deriva de Concepto y la Falta de Adaptación al Mercado

El análisis de la cronología de los eventos sugiere que los algoritmos de Zillow no lograron ajustarse a tiempo a los cambios en las condiciones del mercado inmobiliario. Los modelos continuaron asumiendo un mercado activo y sobreestimaron los precios de las viviendas, un fenómeno conocido en machine learning como "deriva de concepto".

La deriva de concepto ocurre cuando las relaciones entre las variables de entrada y la variable objetivo cambian con el tiempo, lo que puede llevar a los modelos de machine learning a realizar predicciones inexactas. En el caso de Zillow, los modelos no capturaron con precisión cómo la relación entre las características de las viviendas (número de dormitorios, baños, superficie, etc.) y los precios de las viviendas se modificó a medida que el mercado se estancó, siendo afectado por la pandemia.

Factores Contribuyentes a la Deriva del Modelo

Varios factores pudieron haber contribuido a esta deriva del modelo. En primer lugar, Zillow experimentó dificultades para renovar y revender las viviendas con la suficiente rapidez

debido a la escasez de contratistas durante la pandemia de COVID-19. No está claro si los modelos de Zillow incorporaron este factor adecuadamente en sus cálculos. En segundo lugar, Zillow pudo haber estado comprando en áreas que experimentaron aumentos de precios impulsados por la demanda de viviendas suburbanas y rurales durante la pandemia. Sin embargo, la mayor disponibilidad de vacunas pudo haber reducido el atractivo de estas áreas, llevando a una estabilización o caída de los precios, mientras que los algoritmos seguían esperando aumentos.

Lecciones Aprendidas y la Importancia del Monitoreo del Modelo

La debacle de Zillow subraya la importancia crítica de monitorear continuamente el rendimiento de los modelos de machine learning y de contar con herramientas para detectar y mitigar la deriva del modelo. En el caso de Zillow Offers, habría sido beneficioso medir continuamente la precisión del modelo, los resultados del modelo y las entradas del modelo para identificar posibles problemas y activar actualizaciones oportunas del modelo.

Este caso destaca que, si bien el machine learning ofrece un gran potencial para la valoración de propiedades, es esencial reconocer sus limitaciones y la necesidad de una supervisión humana y una adaptación continua a las dinámicas del mercado.

8. Conclusión

Esta investigación demuestra el significativo potencial de los modelos de machine learning para predecir precios de vivienda en la Región Metropolitana de Santiago. Donde se realizó una exhaustiva comparación de modelos lineales, no lineales y de ensamble, utilizando un conjunto de datos robusto y representativo del mercado inmobiliario. Los resultados fueron claros, el modelo Random Forest exhibió el mejor desempeño en precisión y capacidad de generalización, incluso en comunas con baja representación. Sin embargo, no podemos ignorar que el modelo de Ensamble también arrojó resultados muy sólidos, superando al Random Forest en ciertos escenarios. Esto subraya la eficacia de estas herramientas para obtener valoraciones inmobiliarias más precisas.

El análisis de la importancia de las variables reveló que factores como la superficie construida, la ubicación (comuna) y el número de dormitorios tienen una influencia significativa en el precio de las viviendas. Además, la aplicación de técnicas de interpretabilidad, como los SHAP values, permitió comprender mejor la contribución de cada variable a las predicciones individuales, lo cual tiene implicaciones valiosas para la toma de decisiones en el sector inmobiliario.

Sin embargo, es importante reconocer las limitaciones de este estudio y las áreas donde se podría profundizar en futuras investigaciones. Se ha identificado la necesidad de realizar un análisis más detallado de los errores cometidos por los modelos, con el fin de comprender mejor sus fortalezas y debilidades en diferentes segmentos del mercado. Asimismo, se propone explorar las interacciones entre variables y la incorporación de la dimensión temporal en los modelos, para capturar de manera más precisa las dinámicas del mercado inmobiliario.

Además, se destaca la importancia de abordar las implicaciones éticas del uso de machine learning en la valoración de propiedades, incluyendo la identificación y mitigación de posibles sesgos en los datos y los modelos, para garantizar la equidad y la transparencia en el acceso a la vivienda.

En resumen, esta tesis contribuye al conocimiento existente sobre la aplicación de machine learning en el sector inmobiliario y proporciona una base para el desarrollo de herramientas prácticas que puedan mejorar la eficiencia y la transparencia del mercado. Se espera que los hallazgos y las recomendaciones de esta investigación sean de utilidad para agentes inmobiliarios, compradores, vendedores, instituciones financieras y formuladores de políticas.

9. Referencias

- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
<https://doi.org/10.1023/A:1010933404324>
- Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3), 273–297.
- T. Cover and P. Hart, Nearest neighbor pattern classification, in IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, January 1967, doi: 10.1109/TIT.1967.1053964.
- Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1), 34–55. <https://doi.org/10.1086/260169>
- Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2), 241–259.
[https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1)
- Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*.
- Shapley, L. S. (1953). A Value for n-Person Games. In *Contributions to the Theory of Games* (pp. 307–317). Princeton University Press.
- Sah, V., Conroy, S. J., & Narwold, A. (2016). Estimating school proximity effects on housing prices: The importance of robust spatial controls in hedonic estimations. *The Journal of Real Estate Finance and Economics*, 53(1), 50–76. <https://doi.org/10.1007/s11146-015-9490-3>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer.
- El Mouna, M., et al. (2023). A comparative study of urban house price prediction using machine learning algorithms. *E3S Web of Conferences*, 418, 03001.
<https://doi.org/10.1051/e3sconf/202341803001>
- Biblioteca del Congreso Nacional de Chile. (2024). *Ley N° 21.719: Regula la protección y el tratamiento de los datos personales y crea la Agencia de Protección de Datos Personales*.
<https://www.bcn.cl/leychile/navegar?idNorma=1209272>
- Kaul, A. (2021, 16 de noviembre). House Hunters: How ML at Redfin is influencing the housing market with Akshat Kaul. TWIML AI Podcast.
<https://twimlai.com/article/house-hunters-how-ml-at-redfin-is-influencing-the-housing-market-with-akshat-kaul/>

Redfin. (2019, 19 de diciembre). Redfin Improves Accuracy of the Redfin Estimate and Expands to an Additional 11 Million Properties. Redfin.
<https://investors.redfin.com/news-events/press-releases/detail/345/redfin-improves-accuracy-of-the-redfin-estimate-and-expands>

Inside AI News. (2021). The \$500mm+ Debacle at Zillow Offers – What Went Wrong with the AI Models?,
<https://insideainews.com/2021/12/13/the-500mm-debacle-at-zillow-offers-what-went-wrong-with-the-ai-models/>

MarketWatch. (2021). Zillow to stop flipping homes for good as it stands to lose more than \$550 million, will lay off a quarter of staff,
<https://www.marketwatch.com/story/zillow-to-stop-flipping-homes-for-good-as-it-stands-to-lose-more-than-550-million-will-lay-off-a-quarter-of-staff-11635885027>

A continuación, se citan la documentación de las bibliotecas de los modelos utilizados:

Scikit-learn developers. (n.d.-a). sklearn.ensemble.RandomForestRegressor. scikit-learn. Recuperado el [10 de septiembre de 2024] de
<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>

Scikit-learn developers. (n.d.-b). sklearn.neighbors.KNeighborsRegressor. scikit-learn. Recuperado el [10 de septiembre de 2024] de
<https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html>

Scikit-learn developers. (n.d.-c). sklearn.svm.SVR. scikit-learn. Recuperado el [10 de septiembre de 2024] de
<https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html>

Scikit-learn developers. (n.d.-d). sklearn.linear_model.Lasso. scikit-learn. Recuperado el [10 de septiembre de 2024] de
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html

Scikit-learn developers. (n.d.-e). sklearn.linear_model.Ridge. scikit-learn. Recuperado el [10 de septiembre de 2024] de
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html

Scikit-learn developers. (n.d.-f). sklearn.linear_model.LinearRegression. scikit-learn. Recuperado el [10 de septiembre de 2024] de
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

10. Anexos

Anexo 1. Tabla de descripción de las variables utilizadas del dataset.

Table 1: Descripción de las variables del dataset con identificación numérica

No.	Variable	Descripción
1	price	Precio publicado de la propiedad
2	bathrooms	Número de baños
3	bedrooms	Número de dormitorios
4	latitude	Latitud de la ubicación de la propiedad
5	longitude	Longitud de la ubicación de la propiedad
6	seller_type	Indica si el vendedor es el propietario (1) o un agente inmobiliario (0)
7	total_surface	Superficie total de la propiedad
8	municipality	Municipalidad a la que pertenece
9	quality_municipality	Índice de Calidad de Vida Urbana (ICVU, 2019) de la municipalidad
10	parking_lots	Número de estacionamientos
11	swimpool	Indica si la propiedad tiene piscina (1) o no (0)
12	terrace	Indica si la propiedad tiene terraza (1) o no (0)
13	new_house	Indica si la propiedad es una construcción nueva (1) o no (0)
14-16	Simce_x	Puntaje promedio SIMCE de escuelas dentro de un radio de 2 km del nivel x
17-19	nearby_schools_x	Número de escuelas cercanas dentro de un radio de x kilómetros
20-29	dist_x	Distancia geodésica mínima (en kilómetros) a un servicio o lugar x

Nota:

- La variable (simce_x) incluye los puntajes SIMCE correspondientes a 4° básico, 8° básico y 2° medio.
- La variable (nearby_schools_x) considera el número de escuelas cercanas en radios de 1 km, 2 km y 3 km.
- La variable (dist_x) incluye distancias geodésicas mínimas en kilómetros a servicios específicos: estaciones de metro, parques, hospitales, comisarías, colegios, universidades, centros comerciales, penitenciaría, industrias y vertederos.

Anexo 2. Tabla de distribución de datos por comuna en el dataset

Table 3: Distribución de datos por comuna en el dataset

Comuna	Cantidad	Comuna	Cantidad
Las Condes	2300	El Bosque	131
Lo Barnechea	2182	Independencia	130
Colina	1703	Padre Hurtado	126
Puente Alto	1603	Cerrillos	123
Maipú	1463	Conchalí	113
Peñalolén	1229	San Joaquín	96
Vitacura	1130	Renca	94
La Florida	1093	Quinta Normal	93
La Reina	1013	La Granja	76
Ñuñoa	573	La Pintana	71
San Bernardo	488	Pedro Aguirre Cerda	62
Huechuraba	487	Talagante	62
Providencia	421	San Ramón	52
Quilicura	380	Lo Prado	46
Pudahuel	350	Isla De Maipo	41
Santiago	308	Paine	41
Macul	272	Cerro Navia	41
Lampa	243	Tiltil	33
Buín	234	Lo Espejo	30
San Miguel	219	El Monte	28
Peñaflor	206	Melipilla	19
Recoleta	194	San José de Maipo	17
La Cisterna	185	Calera De Tango	17
Estación Central	132	Pirque	17
		Curacaví	7

Anexo 3. Librerías utilizadas para la implementación de los modelos.

A continuación, se presenta un resumen de las principales librerías utilizadas y su propósito:

1. Manipulación y análisis de datos

- Pandas: Utilizado para la carga, manipulación y análisis de datos tabulares. Proporcionó las herramientas necesarias para filtrar, agregar y transformar los datos en un formato adecuado para el análisis.
- NumPy: Empleado para cálculos numéricos de alto rendimiento, incluyendo operaciones con matrices y arreglos necesarios para el preprocesamiento de datos.

2. Visualización de datos

- Matplotlib: Permitió la creación de gráficos como histogramas y gráficos de dispersión, esenciales para explorar la distribución y relaciones entre variables.
- Seaborn: Utilizado también para la creación de visualizaciones, como el mapa de correlación, que facilitó la identificación de relaciones entre las variables predictoras y la variable objetivo.

3. Preprocesamiento de datos

- Scikit-learn: Proporcionó herramientas para la estandarización de variables numéricas mediante StandardScaler y la codificación de variables categóricas con OneHotEncoder. Además, incluyo funciones para particionar los datos en conjuntos de entrenamiento, validación y prueba.

4. Modelado y optimización

- Scikit-learn: fue la librería que se utilizó para la implementación de modelos de machine learning como Random Forest, K-Nearest Neighbors, Support Vector Regressor, Ridge y Lasso Regression. Además, se utilizó para la búsqueda de hiperparámetros con GridSearchCV y la construcción del modelo de ensamble con Stacking Regressor.

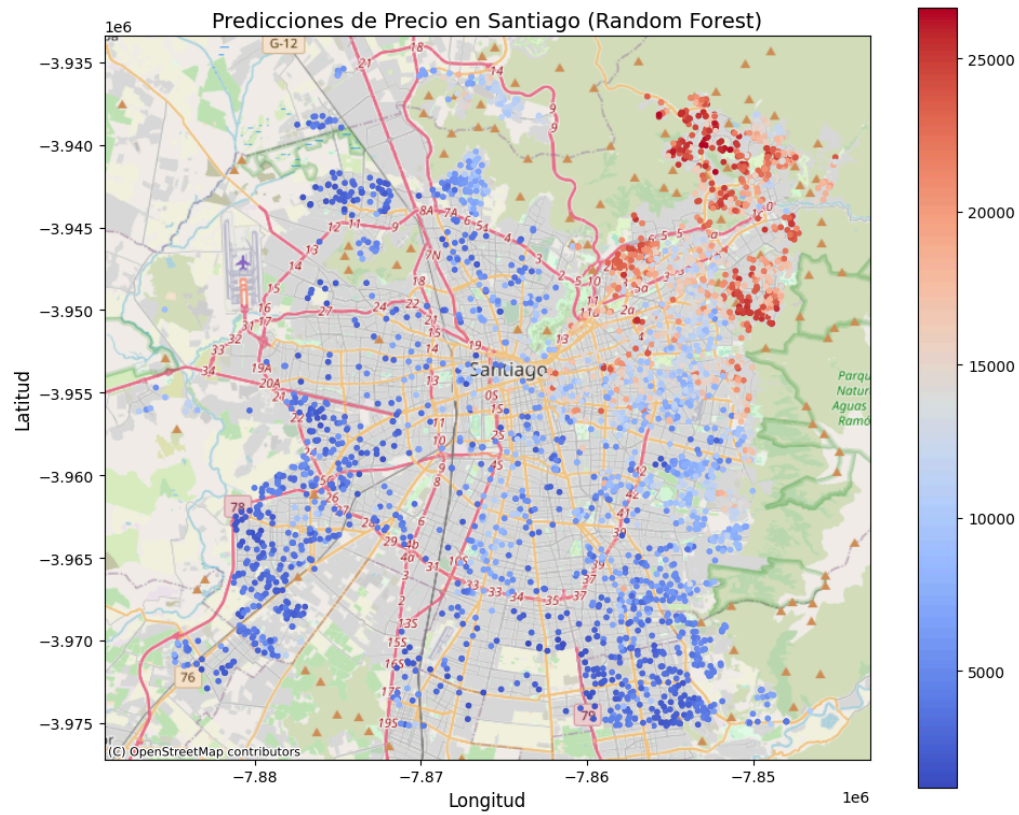
5. Métricas de error:

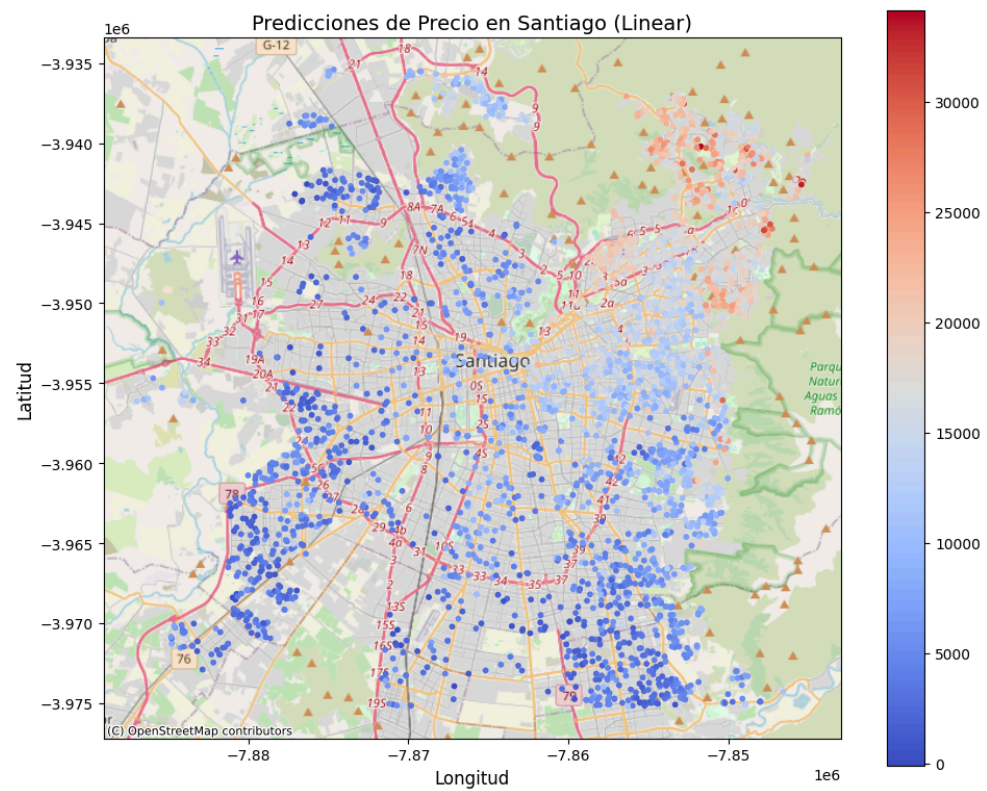
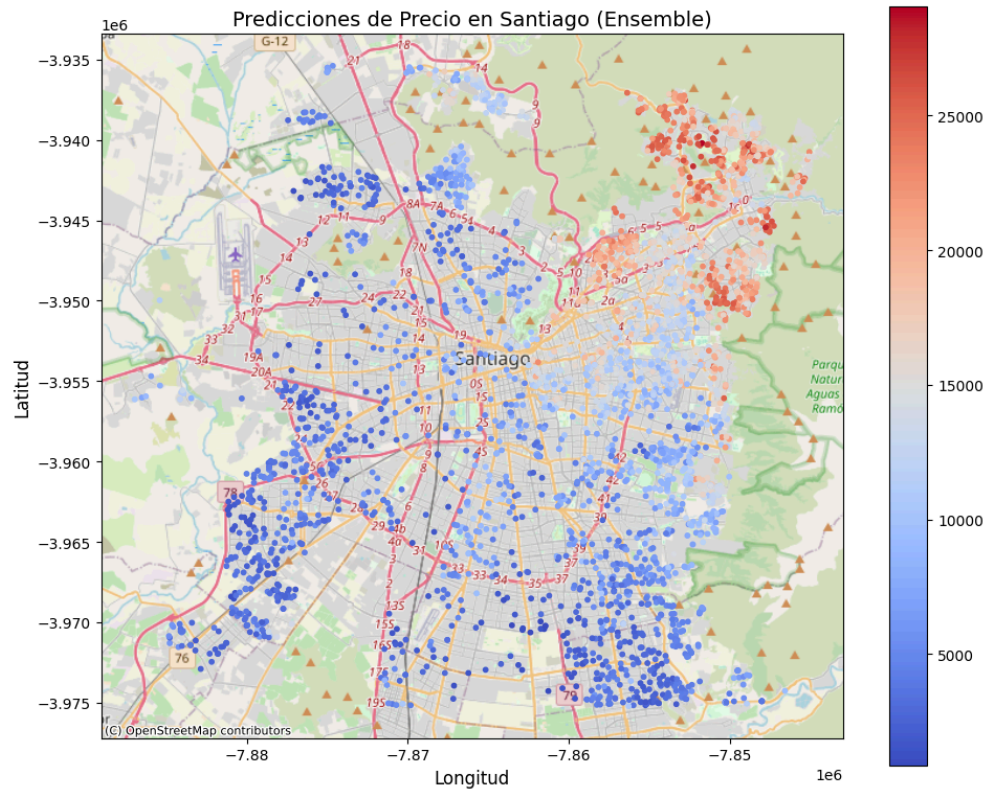
- Scikit-learn: se utilizó para la evaluación de modelos mediante métricas como mean_absolute_error (MAE), que mide la magnitud promedio del error, mean_squared_error (MSE), que cuantifica el error promedio al cuadrado, y r2_score, que evalúa el ajuste general del modelo.

6. Análisis geoespacial

- GeoPandas: Herramienta clave para la manipulación de datos geográficos, como la creación y manejo de geometrías (puntos, polígonos) y la integración de datos espaciales.
- Shapely: Utilizado para la creación de geometrías y operaciones espaciales, como la transformación de coordenadas en puntos.
- Contextily: Facilitó la superposición de mapas base interactivos, como mapas de OpenStreetMap, en visualizaciones geoespaciales

Anexo 4. Gráficos de distribución geográfica de las distribuciones de los modelos Random Forest, Ensemble y Linear.



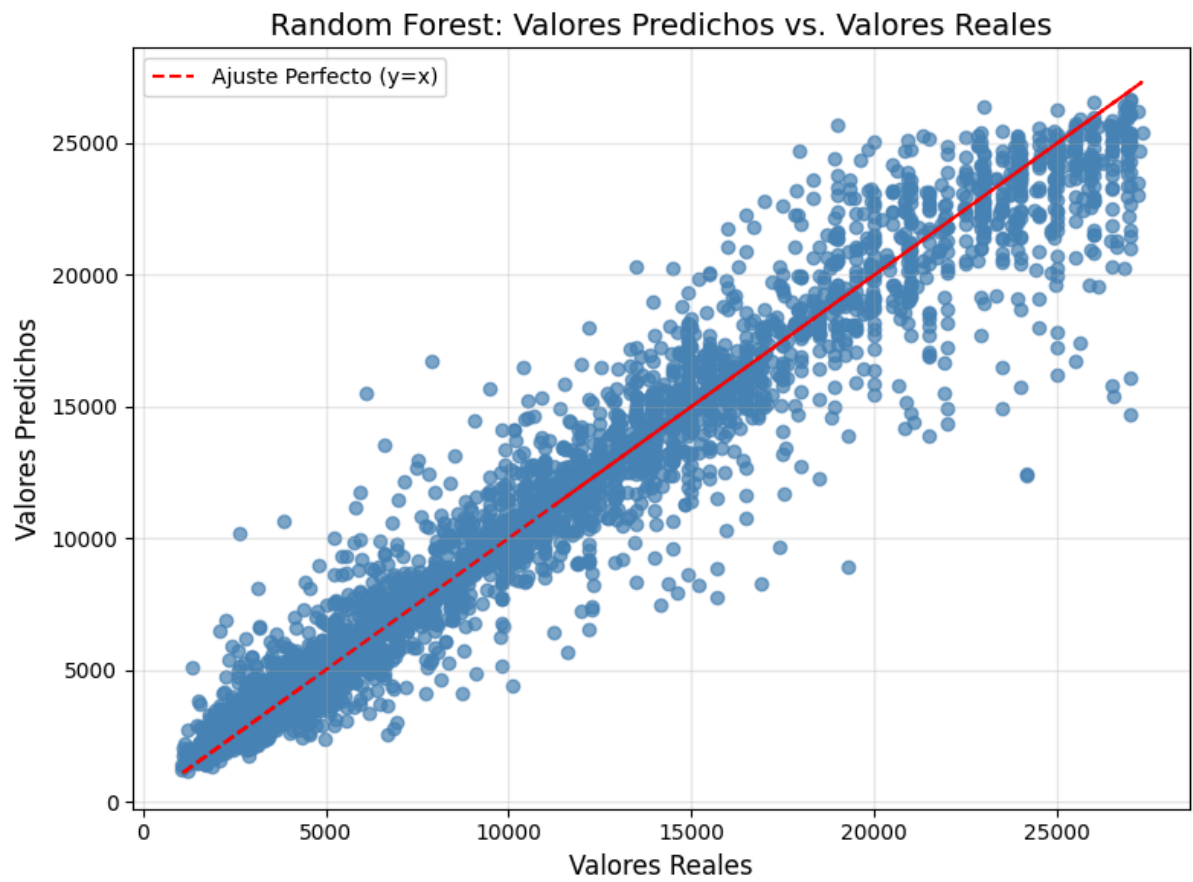


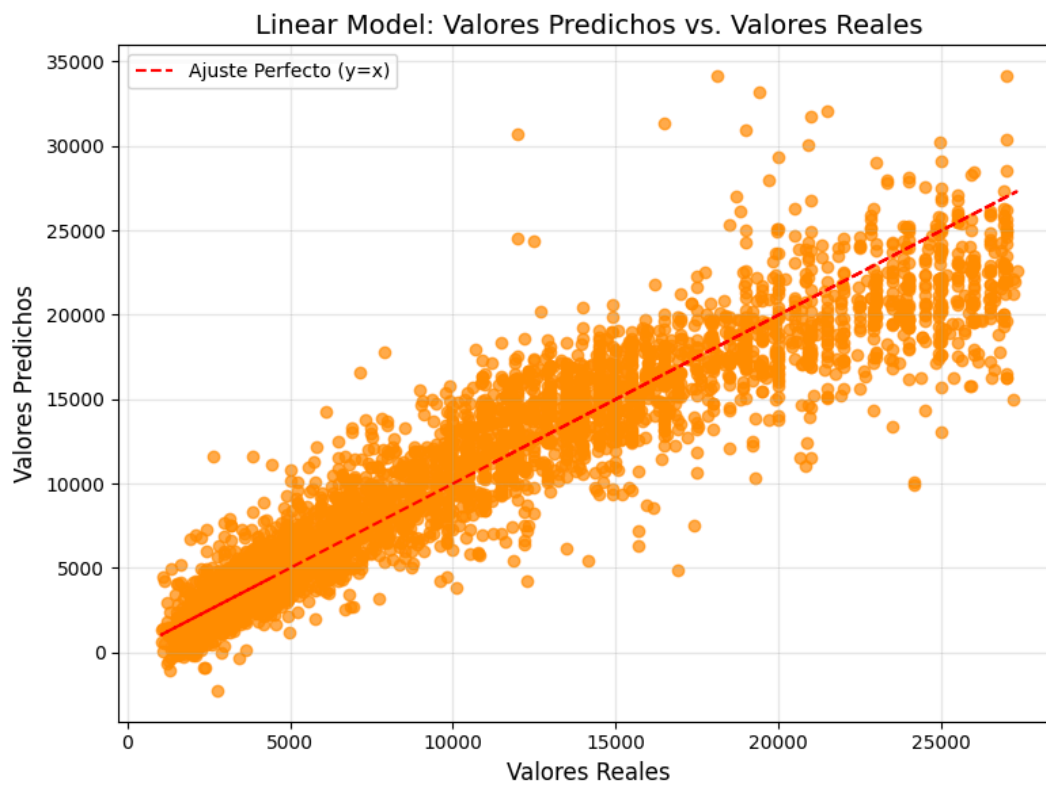
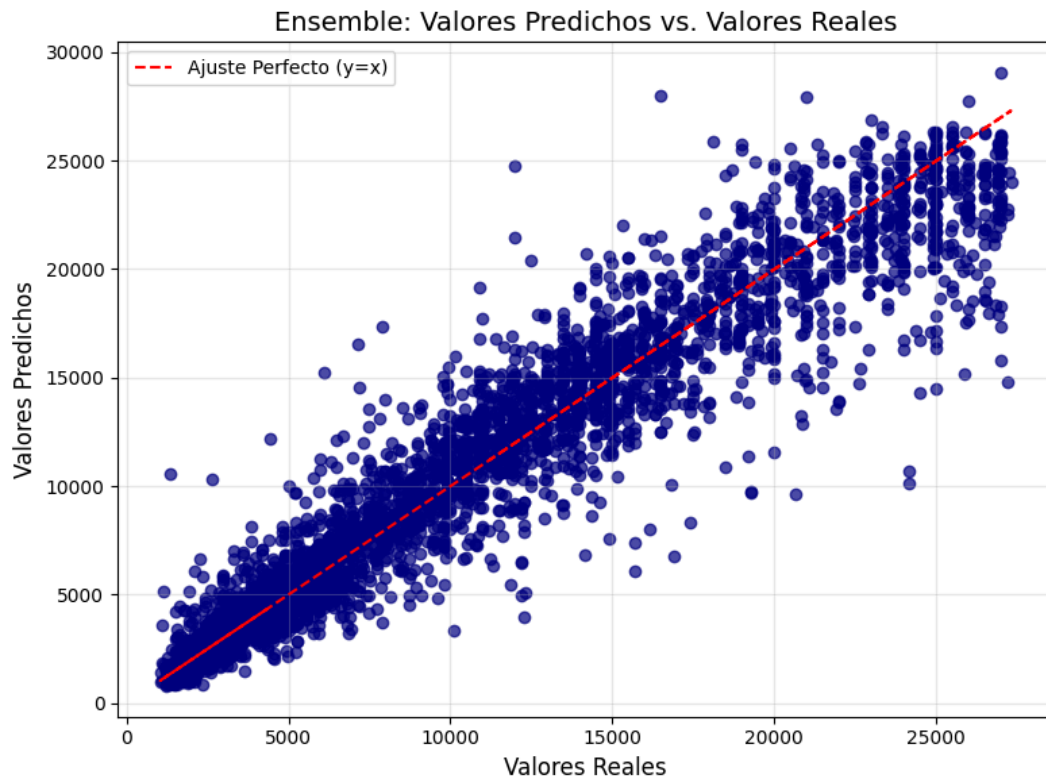
Anexo 5. Tabla de Resultados de métricas por comuna minoritaria y modelo

Table 1: Resultados de métricas por comuna y modelo

Comuna	MAE	RMSE	R ²	MAPE (%)
Random Forest				
Calera De Tango	324.21	326.07	0.134	16.14
Curacaví	615.38	615.38	-	25.82
El Monte	560.22	879.15	0.113	18.92
Melipilla	294.90	325.46	0.626	12.35
Pirque	2609.27	4035.24	0.452	29.93
San José de Maipo	690.51	946.87	0.885	15.31
Ensemble				
Calera De Tango	322.42	375.45	-0.147	17.33
Curacaví	326.87	326.87	-	13.71
El Monte	569.89	894.79	0.081	25.82
Melipilla	616.63	643.68	-0.461	28.22
Pirque	2365.10	4570.84	0.297	24.37
San José de Maipo	1645.12	2019.72	0.476	26.93
Linear Regression				
Calera De Tango	287.10	293.28	0.299	14.53
Curacaví	3274.15	3274.15	-	137.39
El Monte	1596.11	2198.45	-4.55	66.33
Melipilla	379.55	506.25	0.096	18.79
Pirque	4491.72	5905.66	-0.174	67.83
San José de Maipo	2332.99	2786.05	0.002	57.18

Anexo 7. Gráfico de dispersión por modelo





11. Código

Código con valores Paramétricos óptimos

```
# Importación de librerías
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split, GridSearchCV,
cross_val_score
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.ensemble import RandomForestRegressor
from sklearn.neighbors import KNeighborsRegressor
from sklearn.svm import SVR
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import PredefinedSplit
from sklearn.metrics import mean_absolute_error, mean_squared_error,
r2_score
from sklearn.linear_model import Ridge, Lasso, ElasticNet
from sklearn.ensemble import StackingRegressor
import shap
import joblib
import os

# Carga de datos y eliminación de columnas no necesarias
df = pd.read_csv("data.csv")
df.drop(['storage_units', 'photo_kitchen', 'photo_bathroom',
'photo_livingroom', 'photo_bedroom', 'photo_interior',
'photo_diningroom', 'photo_patio', 'photo_exterior', 'photo_gym',
'photo_balcony', 'photo_floorplans', 'photo_laundryroom',
'photo_street', 'photo_gameroom', 'photo_parkinglot',
'photo_swimpool'], axis=1, inplace=True)
df = df[df['building_type'] == 1]
df = pd.get_dummies(df, columns=['municipality'], drop_first=True)

#Quitamos la variable objetivo
x = df.drop('price', axis=1)
y = df['price']

#Conjuntos
```

```

x_temp, x_test, y_temp, y_test = train_test_split(x, y, test_size=0.2,
random_state=42)
x_train, x_val, y_train, y_val = train_test_split(x_temp, y_temp,
test_size=0.2, random_state=42)

#Escalamiento de los datos

#Identificamos las columnas numéricas
col_numeric = x_train.select_dtypes(include=['int64',
'float64']).columns

#Creamos una copia de los conjuntos
x_train_scaled = x_train.copy()
x_val_scaled = x_val.copy()
x_test_scaled = x_test.copy()

#Escalamos con el conjunto de entrenamiento
scaler = StandardScaler()
x_train_scaled[col_numeric] =
scaler.fit_transform(x_train[col_numeric])

#Escalamos el conjunto de validación y test
x_val_scaled[col_numeric] = scaler.transform(x_val[col_numeric])
x_test_scaled[col_numeric] = scaler.transform(x_test[col_numeric])

# Concatenar X_train_scaled y X_val_scaled
x_train_val_scaled = pd.concat([x_train_scaled, x_val_scaled])
y_train_val = pd.concat([y_train, y_val])

# Función para evaluar los modelos
def evaluar_modelo(modelo, x_test, y_test):
    y_pred = modelo.predict(x_test)
    mae = mean_absolute_error(y_test, y_pred)
    rmse = np.sqrt(mean_squared_error(y_test, y_pred))
    r2 = r2_score(y_test, y_pred)
    mape = np.mean(np.abs((y_test - y_pred) / y_test)) * 100
    print(f"MAE: {mae:.2f}")
    print(f"RMSE: {rmse:.2f}")
    print(f"R2: {r2:.2f}")

```

```

# Modelo Random Forest con sus mejores hiperparametros encontrados
best_params_rf = {'max_depth': None, 'min_samples_split': 2,
                  'n_estimators': 200}
best_rf = RandomForestRegressor(**best_params_rf, random_state=42,
                                n_jobs=-1)
best_rf.fit(x_train_val_scaled, y_train_val)
evaluar_modelo(best_rf, x_test_scaled, y_test)

# Modelo knn con sus mejores hiperparametros encontrados
best_params_knn = {'n_neighbors': 7, 'p': 1, 'weights': 'uniform'}
best_knn = KNeighborsRegressor(**best_params_knn, n_jobs=-1)
best_knn.fit(x_train_val_scaled, y_train_val)
evaluar_modelo(best_knn, x_test_scaled, y_test)

# Modelo svr con sus mejores hiperparametros encontrados
best_params_svr = {'C': 10, 'gamma': 'scale', 'kernel': 'linear'}
best_svr = SVR(**best_params_svr)
best_svr.fit(x_train_val_scaled, y_train_val)
evaluar_modelo(best_svr, x_test_scaled, y_test)

# Modelo ridge con sus mejores hiperparametros encontrados
best_params_ridge = {'alpha': 10, 'max_iter': 1000, 'solver': 'svd'}
best_ridge = Ridge(**best_params_ridge, random_state=42)
best_ridge.fit(x_train_val_scaled, y_train_val)
evaluar_modelo(best_ridge, x_test_scaled, y_test)

# Modelo Lasso con sus mejores hiperparametros encontrados
best_params_lasso = {'alpha': 1, 'max_iter': 1000, 'tol': 0.0001}
best_lasso = Lasso(**best_params_lasso, random_state=42)
best_lasso.fit(x_train_val_scaled, y_train_val)
evaluar_modelo(best_lasso, x_test_scaled, y_test)

# Modelo Lineal
linear_model = LinearRegression(n_jobs=-1)
linear_model.fit(x_train_val_scaled, y_train_val)

# Modelo Ensemble

estimators = [
    ('Lasso', best_lasso),
    ('knn', best_knn),
    ('svr', best_svr),
]

```

```

sr = StackingRegressor(
    estimators = estimators,
    final_estimator = linear_model,
    n_jobs = -1
)

sr.fit(x_train_val_scaled, y_train_val)

```

Código con búsqueda de hiperparametros

```

# Importación de librerías
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, GridSearchCV,
PredefinedSplit
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestRegressor, StackingRegressor
from sklearn.neighbors import KNeighborsRegressor
from sklearn.svm import SVR
from sklearn.linear_model import LinearRegression, Ridge, Lasso
from sklearn.metrics import mean_absolute_error, mean_squared_error,
r2_score

# Carga de datos
df = pd.read_csv('data.csv')

# Eliminación de columnas no necesarias
df.drop([
    'storage_units', 'photo_kitchen', 'photo_bathroom',
    'photo_livingroom', 'photo_bedroom',
    'photo_interior', 'photo_diningroom', 'photo_patio',
    'photo_exterior', 'photo_gym',
    'photo_balcony', 'photo_floorplans', 'photo_laundryroom',
    'photo_street', 'photo_gameroom',
    'photo_parkinglot', 'photo_swimpool'
], axis=1, inplace=True)

# Filtrado de datos para casas y eliminación de la columna
`building_type`
df = df[df['building_type'] == 1]
df.drop(['building_type'], axis=1, inplace=True)

```

```

# Creación de variables dummy para la columna `municipality`
df = pd.get_dummies(df, columns=['municipality'], drop_first=True)

# Separación de variables independientes (X) y dependientes (y)
x = df.drop('price', axis=1)
y = df['price']

# División del conjunto de datos
x_temp, x_test, y_temp, y_test = train_test_split(x, y, test_size=0.2,
random_state=42)

x_train, x_val, y_train, y_val = train_test_split(x_temp, y_temp,
test_size=0.2, random_state=42)

# Identificación de columnas numéricas para el escalado
col_numeric = x_train.select_dtypes(include=['int64',
'float64']).columns

# Escalado de datos
scaler = StandardScaler()

x_train_scaled = x_train.copy()
x_val_scaled = x_val.copy()
x_test_scaled = x_test.copy()

x_train_scaled[col_numeric] =
scaler.fit_transform(x_train[col_numeric])
x_val_scaled[col_numeric] = scaler.transform(x_val[col_numeric])
x_test_scaled[col_numeric] = scaler.transform(x_test[col_numeric])

# Concatenación de conjuntos de entrenamiento y validación
x_train_val_scaled = pd.concat([x_train_scaled, x_val_scaled])
y_train_val = pd.concat([y_train, y_val])

# Creación de índices para PredefinedSplit
test_fold = np.concatenate((
    np.full(x_train_scaled.shape[0], -1), # -1 para entrenamiento
    np.zeros(x_val_scaled.shape[0])      # 0 para validación
))
ps = PredefinedSplit(test_fold)

# Función para evaluar los modelos

```

```

def evaluar_modelo(modelo, x_test, y_test):
    y_pred = modelo.predict(x_test)
    mae = mean_absolute_error(y_test, y_pred)
    rmse = np.sqrt(mean_squared_error(y_test, y_pred))
    r2 = r2_score(y_test, y_pred)
    mape = np.mean(np.abs((y_test - y_pred) / y_test)) * 100
    print(f"MAE: {mae:.2f}")
    print(f"RMSE: {rmse:.2f}")
    print(f"R2: {r2:.2f}")
    print(f"MAPE: {mape:.2f}%")

# Random Forest
rf = RandomForestRegressor(random_state=42, n_jobs=-1)
# Cuadrícula con Hiperparametros a probar para RF
param_grid_rf = {
    'n_estimators': [100, 200],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5]
}

grid_rf = GridSearchCV(
    estimator=rf,
    param_grid=param_grid_rf,
    cv=ps,
    scoring='neg_mean_squared_error',
    n_jobs=-1
)

grid_rf.fit(x_train_val_scaled, y_train_val)
best_rf = grid_rf.best_estimator_

# KNeighborsRegressor
knn = KNeighborsRegressor()
# Cuadrícula con Hiperparametros a probar para Knn
param_grid_knn = {
    'n_neighbors': [3, 5, 7, 10, 20, 30, 40],
    'weights': ['uniform'],
    'p': [1, 2]
}

grid_knn = GridSearchCV(
    estimator=knn,
    param_grid=param_grid_knn,

```

```

        cv=ps,
        scoring='neg_mean_squared_error',
        n_jobs=-1
    )

    grid_knn.fit(x_train_val_scaled, y_train_val)
    best_knn = grid_knn.best_estimator_

# SVR
svr = SVR()
# Cuadrícula con Hiperparametros a probar para svr
param_grid_svr = {
    'kernel': ['linear', 'rbf'],
    'C': [1, 10],
    'gamma': ['scale', 'auto']
}

grid_svr = GridSearchCV(
    estimator=svr,
    param_grid=param_grid_svr,
    cv=ps,
    scoring='neg_mean_squared_error',
    n_jobs=-1
)

grid_svr.fit(x_train_val_scaled, y_train_val)
best_svr = grid_svr.best_estimator_

# Ridge Regression
ridge = Ridge()
# Cuadrícula con Hiperparametros a probar para ridge
param_grid_ridge = {
    'alpha': [0.01, 0.1, 1, 10, 100],
    'solver': ['auto', 'svd', 'cholesky'],
    'max_iter': [1000, 5000, 10000]
}

grid_ridge = GridSearchCV(
    estimator=ridge,
    param_grid=param_grid_ridge,
    cv=ps,
    scoring='neg_mean_squared_error',
    n_jobs=-1

```

```

)

grid_ridge.fit(x_train_val_scaled, y_train_val)
best_ridge = grid_ridge.best_estimator_

# Lasso Regression
lasso = Lasso()
# Cuadrícula con Hiperparametros a probar para lasso
param_grid_lasso = {
    'alpha': [0.0001, 0.001, 0.01, 0.1, 1, 10, 100],
    'max_iter': [1000, 5000, 10000]
    'tol': [1e-4, 1e-3, 1e-2]
}

grid_lasso = GridSearchCV(
    estimator=lasso,
    param_grid=param_grid_lasso,
    cv=ps,
    scoring='neg_mean_squared_error',
    n_jobs=-1
)

grid_lasso.fit(x_train_val_scaled, y_train_val)
best_lasso = grid_lasso.best_estimator_

# Liner Regression
linear = LinearRegression()
linear.fit(x_train_val_scaled, y_train_val)

# Stacking Regressor (Ensemble)
estimadores = [
    ('lasso', best_lasso),
    ('knn', best_knn),
    ('svr', best_svr)
]

ensemble = StackingRegressor(
    estimators=estimadores,
    final_estimator=LinearRegression(),
    n_jobs=-1
)

```

```
ensemble.fit(x_train_val_scaled, y_train_val)

# Evaluación de modelos
evaluar_modelo(best_rf, x_test_scaled, y_test)
evaluar_modelo(best_knn, x_test_scaled, y_test)
evaluar_modelo(best_svr, x_test_scaled, y_test)
evaluar_modelo(best_ridge, x_test_scaled, y_test)
evaluar_modelo(best_lasso, x_test_scaled, y_test)
evaluar_modelo(ensemble, x_test_scaled, y_test)
```