UNIVERSIDAD ESAN

Uso de Machine Learning para la predicción de precios de departamentos en Lima

Trabajo de investigación presentado en satisfacción parcial de los requerimientos para obtener el grado de Maestro en Finanzas por:

Código 2208740 Barrientos Villegas, Renzo David

Código 2208767 Delgado Luque, Renzo Augusto

Código 2208766 Escalante Carty, Laura Patricia

Código 2201792 Febres Bustamante, Gonzalo

Código 2207462 Hisbes Malca, Estefany Brizet

Programa de la Maestría en Finanzas

MAF/22-2

Lima, 21 de marzo de 2025

Uso de Machine learning para la predicción de precios de departamentos de Lima.docx

INFORME DE ORIGINALIDAD

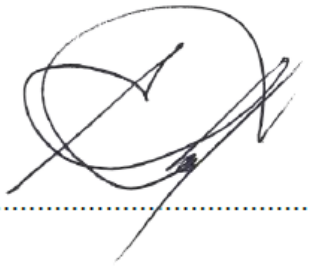
0%	0%	0%	0%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

Excluir citas	Activo	Excluir coincidencias	< 2%
Excluir bibliografía	Activo		

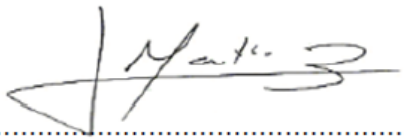
Este trabajo de investigación

Uso de Machine Learning para la predicción de precios de departamentos en Lima ha sido aprobado.



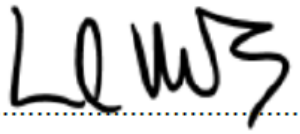
.....

Ernesto Fernando Cuadros Tenorio (Jurado)



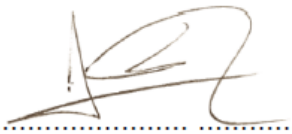
.....

César Armando Martínez La Rosa (Jurado)



.....

Luis Carlos Chávez-Bedoya Mercado (Co-asesor)



....

Luis Francisco Rosales Marticorena (Co-asesor)

Universidad ESAN

2025

A mí familia por su apoyo incondicional y creer en mi desde siempre, a mi grupo de la maestría, que son ahora mis amigos, por su apoyo constante y consejos para sacar adelante cada uno de los procesos que vivimos juntos y a mi novia por impulsarme a seguir creciendo profesionalmente.

Renzo Barrientos

A mis padres, y a mi futura esposa a quienes me motivaron para emprender este camino de desafíos y retos. A mis compañeros de la maestría con quienes compartimos noches de estudios, risas y desafíos. Por el apoyo mutuo, la motivación constante y por demostrar que el camino es más llevadero cuando se recorre en equipo.

Renzo Delgado

A mi madre, a mi futuro esposo, a Sh. y Ql., por su amor y su constante ayuda. Pero, sobre todo, al grupo 8, unos grandes amigos.

Laura Escalante

A mi esposa, Diana, y a mis hijos, José Eduardo y Joaquín Ignacio.

Gonzalo Febres

A mis padres, Manuel y María, por ser mi mayor inspiración. A mis hermanos mayores, y a los más pequeños S y K por su amor incondicional. Y a mi novio, mi compañero de vida, por su apoyo constante.

Estefany Hisbes

Renzo David Barrientos Villegas

Candidato a Maestro en Finanzas de ESAN Graduate School of Business. Contador de la UNMSM. Experiencia de 4 años en preparación de estados financieros, análisis de cuentas y registros contables, Experiencia de 3 años en Finanzas con el control de gastos presupuestados, seguimiento de indicadores financieros, proyección de resultados, presupuesto anual de empresa y evaluación de nuevos proyectos.

FORMACIÓN

2022 – 2024 ESAN Graduate School of Business
Maestría en Finanzas, Finanzas Corporativas.

2023 - UPF Barcelona School of Management
Maestría de Gerencia Bancaria y Financiera.

2014 – 2018 - Universidad Nacional Mayor de San Marcos
Bachiller en Contabilidad

EXPERIENCIA

2020 - Actualidad	**SANNA División Ambulatoria – Servicios médicos** - Empresa con 30 años en el sector salud y venta de medicinas.
	Jefe de Planeamiento y Finanzas. •Responsable de mantener la confiabilidad de los resultados y cumplimiento de los objetivos de la empresa. •Manejo de proyección de ventas y resultados mensuales y anuales. •Control de indicadores financieros y control del presupuesto anual. •Evaluación de nuevos proyectos o negocios para el sector salud.

	•Cumplimiento de objetivos por tipo de negocio de salud dentro de la empresa, tanto en costos como en resultados y niveles de ventas.
	<u>Analista de Finanzas</u> •Responsable del seguimiento de indicadores financieros, proyecciones mensuales por negocio y apoyo en elaboración de presupuesto. •Elaboraciones de indicadores por negocio y el tipo de pacientes asegurados en centros clínicos. •Apoyo en la proyección y elaboración de presupuesto.
	<u>Analista de Contabilidad</u> •Análisis de cuentas contables, elaboración de estados financieros y control de provisiones por negocio. •Registro de provisiones mensuales por negocio •Elaboración de anexos para la explicación de cuentas contables •Control de las conciliaciones bancarias y flujos de caja.

Renzo Augusto Delgado Luque

Candidato a Maestro en Finanzas de ESAN Graduate School of Business. Administrador de Empresas en la Universidad Peruana de Ciencias Aplicadas (UPC) con más de 5 años de experiencia en el Área de Finanzas a nivel corporativo en puestos de jefatura, coordinador y analista de finanzas en empresas del sector construcción.

FORMACIÓN ACADÉMICA

2022 – 2024 ESAN Graduate School of Business
Maestría en Finanzas, Finanzas Corporativas.

2023 - UPF Barcelona School of Management
Maestría de Gerencia Bancaria y Financiera.

2014 - 2019 Universidad Peruana de Ciencias Aplicadas (UPC) – Bachiller en Administración de Empresas.

EXPERIENCIA PROFESIONAL

2018 - Actualidad	**FLESAN DEL PERU -** Empresa constructora chilena con participación en Perú por más de 15 años teniendo en su cartera más de 1,000 obras a nivel nacional.
	Jefe de Finanzas • Gestión Financiera y Bancaria: Administración de líneas de financiamiento, emisión de garantías, gestión de deuda y optimización de costos financieros. • Análisis y Presentación de Estados Financieros: Evaluación de ratios financieros, cierre contable mensual y elaboración de informes de gestión.

	• Planificación y Estructuración de Proyectos: Proyección financiera, estructuración de proyectos inmobiliarios y energéticos, y adquisición de financiamiento. • Sustentación y Reportes Estratégicos: Presentación de estados financieros ante entidades de financiamiento y elaboración de reportes para la alta dirección.
2017 - 2018	**BANCO INTERNACIONAL DEL PERU (INTERBANK) -** Entidad financiera que pertenece al Grupo INTERCORP
	<u>Analista de Prevención de Fraudes</u> •Análisis de eventos de fraude, elaborando informes de prevención y control de diversas unidades. •Realizar seguimiento de las posibles modalidades de fraude en el ámbito local e internacional •Diseñar y proponer controles que reduzcan la probabilidad de ocurrencia de fraude interno y/o externo de los procesos y procedimientos del Banco. •Proponer actividades para fomentar una cultura de prevención contra riesgos de fraude, destinada a salvaguardar el patrimonio del Banco. •Proponer en coordinación con las unidades orgánicas responsables las modificaciones en los sistemas, procedimientos y controles para minimizar el riesgo de fraude.

	Jefe de Operaciones y Servicios •Administrador de una oficina bancaria supervisando las actividades operativas, de control y seguridad. Impulsar la venta de productos y servicios que ofrece la entidad financiera. •Definir los lineamientos generales del modelo operativo de la Banca Comercial. •Recabar la información que sustente los requerimientos canalizados al Área de Procesos. •Cumplir con las metas indicadas por parte del área comercial del banco (venta de tarjetas de crédito, cuentas de ahorro, depósitos a plazo, etc).

Laura Patricia Escalante Carty

Candidata a Maestro en Finanzas de ESAN. Administradora de la UPC. Experiencia en el sector minero, especializada en administración y tesorería. Especialista en elaboración de flujos de caja. Alta capacidad de adaptación, actitud proactiva, excelentes relaciones interpersonales con influencia positiva y comunicación efectiva a todo nivel.

FORMACIÓN

2022 – 2024 ESAN Graduate School of Business
Maestría en Finanzas, Finanzas Corporativas.

2023 - UPF Barcelona School of Management
Maestría de Gerencia Bancaria y Financiera.

2017 – 2021 Universidad Peruana de Ciencias Aplicadas UPC
Bachiller en Administración de Empresas. Quinto superior.

EXPERIENCIA

2016 - Actualidad	**Silver X Mining Corporation-** Empresa minera canadiense con más de 10 años de experiencia en la exploración y producción de metales preciosos, especializada en la extracción de plata y oro en América Latina.
	<u>Jefa de Administración</u> • Gestión de las áreas Administrativa, Financiera y Tesorería. • Elaboración y supervisión de flujos de caja para la optimización de recursos. • Coordinación de estrategias financieras para la operación minera en el Perú.

	• Manejo de relaciones con entidades regulatorias y financieras. • Implementación de procesos para mejorar la eficiencia en la gestión administrativa.
2015 - 2016	**Mines & Metals Trading Perú-** Empresa minera peruana con proyectos polimetálicos. Adquirió la Unidad de Producción Minera "Recuperada" de la Compañía de Minas Buenaventura S.A.A.
	<u>Administradora</u> • Gestión de recursos humanos: Supervisar la contratación, capacitación y bienestar del personal, asegurando el cumplimiento de las normativas laborales y políticas internas. • Administración financiera**:** Controlar presupuestos, gestionar costos operativos y coordinar pagos a proveedores, garantizando una administración financiera eficiente. • Coordinación de actividades administrativas**:** Gestionar tareas administrativas relacionadas con la operación minera, incluyendo la supervisión de personal administrativo y la implementación de procedimientos internos.

Gonzalo Febres Bustamante

Candidato a Maestro en Finanzas de ESAN. Economista de la Universidad de Piura. Con experiencia en banca y finanzas. Con capacidad para trabajar en equipo, con visión y habilidad de plantear y emprender alternativas para dar soluciones ágiles y acertadas sobre los objetivos señalados.

FORMACIÓN

2022 – 2024 ESAN Graduate School of Business
Maestría en Finanzas, Finanzas Corporativas.

2023 - UPF Barcelona School of Management
Maestría de Gerencia Bancaria y Financiera.

2017 – 2021 Universidad de Piura
Bachiller en Economía

EXPERIENCIA

2024 - Actualidad	**Centros de Salud Peruanos S.A.C. Clínica AVIVA -,** es una empresa peruana del grupo Interbank dedicada a la prestación de servicios médicos integrales. Ofrece atención en diversas especialidades, incluyendo consultas ambulatorias, emergencias, laboratorio, imágenes y procedimientos quirúrgicos.
	<u>Jefe de Planificación Financiera</u> • Elaboración de planes financieros a corto, mediano y largo plazo para optimizar los recursos financieros de la clínica. • Coordinación de la gestión de presupuestos anuales, alineados con los objetivos estratégicos de la clínica.

	• Control de costos operativos y evaluación de la rentabilidad de los servicios y proyectos. • Implementación de indicadores financieros (KPIs) para medir el desempeño económico de la clínica. • Presentación de informes financieros a la alta dirección para la toma de decisiones estratégicas. • Evaluación y proyección de la viabilidad financiera de proyectos e inversiones dentro de la clínica.
2021 - 2024	**Clínica San Felipe -** con más de 60 años de experiencia, ofrece servicios médicos integrales en más de 45 especialidades, contando con más de 350 profesionales.
	<u>Controller Financiero</u> •Contribuir con el planteamiento de la estrategia de las principales actividades propias de Finanzas, Costos, Presupuestos y su viabilidad hacia procesos transversales. •Implementar un adecuado modelo de costos y de rentabilidad, salvaguardando su ejecución. •Responsable de la elaboración de Presupuestos y Control Presupuestal. •Responsable de la evaluación de Proyectos de Inversión: Ampliación de clínica y equipos en específico. •Responsable del Control Logístico de Proyectos de Ampliación: Flujo de pagos, valorizaciones, control contable de capex y opex por centros de costos asociados al Proyecto. •Responsable de la Proyección de los EEFF. •Responsable de la presentación y explicación de EEFF al Directorio. •Responsable de la identificación e implementación de KPIs

	•Controlar los días de cobro y anticuamiento, a través de indicadores de medición. Representante del Proyecto de BI Corporativo de Clínica San Felipe: Proyecto corporativo para la elaboración de visores y repositorios corporativos para análisis de actividad, resultados y costos.
2021 - 2022	**Clínica SANNA – Clínica del Sur -** en Arequipa, forma parte de la red SANNA y ofrece servicios médicos especializados con tecnología avanzada. Brinda atención ambulatoria, hospitalaria y de emergencia en diversas especialidades.
	<u>Jefe de Finanzas</u> •Estrategia y Control Financiero: Definir estrategias para Finanzas, Tesorería, Compras, Liquidación, Cobranzas y Facturación, optimizando el ciclo de efectivo. •Gestión de Facturación y Cobranza: Supervisar la correcta liquidación, emisión y cobranza de expedientes para pacientes particulares, asegurados y financiadores. •Indicadores y Seguimiento: Controlar los días de facturación, cobro y anticuamiento mediante indicadores de medición para servicios ambulatorios, hospitalarios y de urgencias. •Gestión de Seguros y Tarifas: Administrar seguros de salud particulares, tarificación, promoción corporativa y ajuste de tarifas. •Compras y Proveedores: Optimizar la gestión de compras, negociar con proveedores médicos y aseguradoras, y supervisar el control de activos. •Flujo de Caja y Cumplimiento: Mejorar la gestión del flujo de caja, rendimientos sobre excedentes y dar seguimiento a fiscalizaciones.
2020 - 2021	**Clínica SANNA – El Golf -** Clínica privada que forma parte de la red SANNA, ofreciendo servicios médicos de alta calidad en más de 30 especialidades.

	Analista de Finanzas Trabajo directo con la Gerencia de Finanzas sobre el planteamiento de objetivos y actividades de la Gerencia. •Garantizar y velar por la solidez financiera, gestionando ingresos, costos y gastos, generando los niveles de utilidad esperados alineado con la Gerencia General y el Corporativo. •Proyecto de ampliación de la Clínica: Coordinador principal sobre las adecuaciones administrativas y financieras. •Optimización de la estructura de capital, generando un correcto balance entre deuda y patrimonio. •Control de pagos y cobranza que permitan mantener el nivel de liquidez.

Estefany Brizet Hisbes Malca

Candidata a Maestro en Finanzas de ESAN Graduate School of Business. Economista de la USMP. Experiencia en análisis financiero y control de gestión en Retail. Conocimientos de inglés y dominio de las herramientas informáticas de gerencia. Aspiración de desarrollo profesional en gerencia de negocios financieros.

FORMACIÓN

2022 – 2024 ESAN Graduate School of Business
Maestría en Finanzas, Finanzas Corporativas.

2023 - UPF Barcelona School of Management
Maestría de Gerencia Bancaria y Financiera.

2019 - 2019 Universidad De San Martín de Porres - USMP
Licenciada en Economía.

2013 - 2018 Universidad De San Martín de Porres - USMP
Bachiller en Economía.

EXPERIENCIA

2022 - actualidad	**Corporación Primax.** Empresa con 20 años y líder en el sector hidrocarburos.
	Analista Senior de la Gerencia de Finanzas. • Responsable del control financiero de dos Negocios de la empresa (Dealers y White Pumpers). Ambos negocios tienen un valor de S/ 90 millones en el 2024.

	• Coordinación con otros departamentos para el correcto seguimiento de gastos de la unidad de negocio. • Elaborar el informe y presentación de gestión Mensual para Gerente del negocio. • Responsable de gestión financiera de la Unidad de Negocio. • Liderar en la elaboración del presupuesto de la UN.
	<u>Analista de la gerencia de Finanzas.</u> • Responsable del control financiero del Negocios de Tiendas de Conveniencia de Primax con un valor estimado de S/ 46 millones en el 2024.
2020 -2022	**VSI INDUSTRIAL SA. -** Empresa fabricadora y diseñadora de griferías y sanitarios con más de 40 años en el mercado.
	<u>Analista Junior de Planeamiento Financiero.</u> • Elaboración de la presentación mensual para el Directorio. Dar seguimiento y cumplimento al presupuesto. Elaboración de reportes de activos y pasivos. Liderar la elaboración de presupuesto de todas las áreas de la empresa.
2018 -2019	**Grupo CELIMA-TREBOL -** Con más de 50 años de experiencia, el Grupo CELIMA TREBOL, fabricando y diseñando revestimientos cerámicos, aparatos sanitarios y griferías de la más alta calidad.
	<u>Asistente de Planeamiento Financiero</u>. • Análisis de EEFF (Estados de Resultados y Balance General). Elaboración del comité Ejecutivo. Control presupuestal en SAP (FM y CO). Manejo y control de inversiones CAPEX. Elaboración del presupuesto anual.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN EJECUTIVO

El presente estudio tiene como objetivo desarrollar un modelo de predicción de precios de departamentos en Lima mediante la aplicación de técnicas de *Machine Learning* (ML), comparando su desempeño con la tradicional regresión hedónica. La investigación surge como respuesta a la necesidad de contar con estimaciones más precisas en la valoración inmobiliaria, optimizando la toma de decisiones en el sector financiero e inmobiliario.

Para ello, se emplean tres enfoques metodológicos: la regresión hedónica como referencia teórica, y dos modelos de *Machine Learning*: Árbol de Decisión y *XGBoost*. La base de datos utilizada proviene del Banco Central de Reserva del Perú (BCRP) y abarca más de 40,000 observaciones de inmuebles en los principales distritos de Lima entre 2014 y 2024.

Los resultados muestran que *XGBoost* ofrece el mejor desempeño predictivo, con menores valores de error: MAPE (*Mean Absolute Percentage Error*), MAE (*Mean Absolute Error*), RMSE (*Root Mean Squared Error*) y un coeficiente de determinación (R^2) más alto que los otros modelos. Asimismo, se identifica que el factor más influyente en el precio de los departamentos es el tamaño del inmueble, seguido del número de garajes y la ubicación.

Este estudio evidencia el potencial de *Machine Learning* como una herramienta clave para mejorar la precisión en la estimación de precios inmobiliarios, aportando beneficios tanto para inversionistas como para instituciones financieras y reguladoras. Además, se plantea la posibilidad de futuras investigaciones incorporando variables adicionales, como la proximidad a servicios urbanos o la calidad de construcción, para seguir optimizando la predicción de precios en el mercado inmobiliario de Lima.

ABSTRACT

This study aims to develop a predictive model for apartment prices in Lima through the application of Machine Learning (ML) techniques, comparing their performance with traditional hedonic regression. The research emerges in response to the need for more accurate real estate valuation estimates, optimizing decision-making in the financial and real estate sectors.

To achieve this, three methodological approaches are employed: hedonic regression as a theoretical benchmark, and two Machine Learning models: Decision Tree and XGBoost. The dataset used comes from the Central Reserve Bank of Peru (BCRP) and includes more than 40,000 property observations in key districts of Lima from 2014 to 2024.

The results indicate that XGBoost provides the best predictive performance, yielding lower error values: MAPE (Mean Absolute Percentage Error), MAE (Mean Absolute Error), RMSE (Root Mean Squared Error) y and a higher coefficient of determination (R^2) than the other models. Additionally, the study identifies that the most influential factor in apartment pricing is property size, followed by the number of garages and location.

This research highlights the potential of Machine Learning as a key tool for improving the accuracy of real estate price estimation, benefiting investors as well as financial and regulatory institutions. Furthermore, future research could incorporate additional variables, such as proximity to urban services or construction quality, to further enhance price prediction in Lima’s real estate market.

CAPÍTULO I: INTRODUCCIÓN

La teoría de precios hedónicos es un enfoque económico que explica el valor de un bien a partir de sus atributos y características. En el caso del mercado inmobiliario, esta teoría sostiene que el precio de una propiedad depende de factores como su superficie, calidad de construcción, diseño, ubicación, acceso a áreas verdes y características del vecindario, entre otros. A través de este enfoque, es posible estimar el valor individual de cada atributo y, en consecuencia, calcular la demanda implícita de las distintas características que componen el bien.

El origen de esta teoría se remonta a los estudios de Court (1939), quien aplicó un modelo basado en características para analizar precios en la industria automotriz. Sin embargo, el desarrollo más reconocido de los precios hedónicos comenzó con Ridker y Henning (1967), quienes investigaron el impacto de la contaminación del aire en los precios de las viviendas en St. Louis, Estados Unidos. Posteriormente, Griliches (1971) aplicó esta metodología al mercado de vehículos, analizando el precio en función de atributos como el consumo de combustible y la potencia. La consolidación de esta teoría llegó con Rosen (1979), quien formalizó un modelo que justifica la existencia de un equilibrio de mercado donde los precios se determinan en función de las características del producto. Su trabajo es considerado un referente clave en la literatura sobre precios hedónicos.

A lo largo de las últimas décadas, diversos estudios han aplicado esta metodología en el mercado inmobiliario. Lever (2009) destacó la importancia de los precios hedónicos en la modelación de mercados implícitos, proporcionando técnicas econométricas para estimar precios y demandas a partir de la combinación de atributos. García y Raya (2013) analizaron las elasticidades de la demanda de vivienda en Barcelona utilizando modelos hedónicos tradicionales, mientras que Mundaca y Sánchez (2018) estimaron índices de precios inmobiliarios en Lima, concluyendo que factores como la zonificación y la infraestructura urbana influyen significativamente en los valores de las viviendas.

En la práctica, la estimación de precios mediante la teoría de precios hedónicos se ha basado tradicionalmente en modelos de regresión lineal, especialmente mediante el método de mínimos cuadrados ordinarios. Este enfoque ofrece ventajas en términos de interpretación y robustez estadística, permitiendo obtener coeficientes que cuantifican el impacto de cada atributo en el precio del bien. Sin embargo, en los últimos años han surgido alternativas basadas en técnicas de ML, las cuales permiten mejorar la precisión de las predicciones mediante modelos más sofisticados.

El presente estudio busca comparar la efectividad de los modelos tradicionales de regresión lineal con técnicas de ML en la predicción de precios de departamentos en Lima. A pesar de que los modelos de ML pueden ser menos interpretables que los métodos estadísticos convencionales, su capacidad de ajuste y optimización computacional podría generar estimaciones más precisas y realistas. En este sentido, esta investigación busca evaluar si el uso de algoritmos de árboles de decisión y *XGBoost* puede mejorar la capacidad predictiva del modelo, proporcionando una herramienta más eficiente para la valoración de bienes raíces en Lima.

1.1 Problema Central y relevancia

En el contexto actual del mercado inmobiliario de Lima, la valoración precisa de propiedades representa un desafío significativo tanto para compradores, vendedores, inversionistas y entidades financieras. La falta de estimaciones confiables sobre el precio real de los inmuebles genera incertidumbre en la toma de decisiones y puede dar lugar a distorsiones en el mercado, tales como la sobrevaloración de activos, dificultades en la concesión de créditos hipotecarios o incluso la formación de burbujas inmobiliarias.

El problema central que esta investigación aborda es la necesidad de contar con un modelo predictivo basado en técnicas de ML que permita estimar con mayor precisión los precios de los departamentos en Lima.

Los métodos tradicionales, como las tasaciones basadas en comparables o los modelos econométricos clásicos, presentan limitaciones al no capturar adecuadamente las dinámicas

no lineales del mercado ni la interacción de múltiples características de los inmuebles que influyen en la determinación de la estimación de precios. Esta ausencia de precisión puede afectar tanto a individuos como a instituciones, generando riesgos financieros y distorsiones en la asignación de recursos.

Al proporcionar estimaciones precisas y detectar discrepancias entre precios proyectados y precios de mercado, este modelo puede servir como una herramienta de monitoreo para instituciones como el Banco Central de Reserva del Perú (BCRP). La detección temprana de sobrevaloraciones en el mercado inmobiliario permitiría la implementación de medidas preventivas para mitigar riesgos macroeconómicos y evitar crisis financieras.

Para entidades bancarias y financieras, una predicción precisa del valor de los inmuebles mejora la gestión de riesgos en préstamos hipotecarios. Un modelo robusto permitiría realizar valoraciones más objetivas de las garantías hipotecarias, reduciendo la exposición a pérdidas en caso de incumplimiento y optimizando la asignación de crédito en función de valores más reales de mercado.

En el ámbito financiero, la valoración de activos inmobiliarios desempeña un rol crucial en la gestión del riesgo, tanto a nivel individual como sistémico. Una incorrecta estimación del valor de un inmueble puede conllevar distorsiones significativas en múltiples decisiones financieras, incluyendo la evaluación de garantías hipotecarias, la estructuración de portafolios inmobiliarios, y la gestión del riesgo crediticio por parte de las instituciones bancarias.

En este sentido, los métodos tradicionales de valoración, como las tasaciones comparativas o los modelos de regresión hedónica simples, presentan limitaciones que pueden incrementar la exposición de los agentes financieros al riesgo de subvaloración o sobrevaloración. Este desfase entre el precio observado y el valor fundamental del activo puede provocar un deterioro en la calidad de las carteras crediticias, afectar la rentabilidad

esperada de los fondos de inversión inmobiliarios y generar señales falsas en los sistemas de alerta temprana diseñados para prevenir burbujas especulativas.

Desde una perspectiva macroeconómica, esta situación representa un riesgo sistémico, especialmente en economías como la peruana, donde el crecimiento del crédito hipotecario ha sido sostenido durante la última década. En consecuencia, la necesidad de contar con herramientas de predicción más precisas y robustas no solo obedece a fines de eficiencia de mercado, sino que responde a una urgencia por fortalecer los mecanismos de supervisión y control del sistema financiero en su conjunto.

En este contexto, el desarrollo de modelos de predicción de precios inmobiliarios mediante técnicas de Machine Learning (ML) emerge como una solución potencialmente más eficiente y adaptativa. Estos modelos no solo ofrecen una mayor capacidad para capturar relaciones no lineales entre variables, sino que también se alinean con la creciente necesidad del sector financiero por incorporar herramientas tecnológicas que mejoren la calidad de las decisiones estratégicas.

Por tanto, esta investigación no solo busca mejorar la precisión de las estimaciones de precios inmobiliarios, sino también contribuir a la mitigación de riesgos financieros derivados de valoraciones inadecuadas, lo cual es especialmente relevante para entidades financieras, reguladores, aseguradoras y fondos de inversión.

1.1.1 Limitaciones de los métodos clásicos de valoración

En los estudios de valoración inmobiliaria, los métodos tradicionales más utilizados han sido las tasaciones comparativas y los modelos de regresión hedónica lineal. Si bien estos enfoques han sido herramientas fundamentales en la historia del análisis económico y urbano, su aplicabilidad presenta limitaciones importantes cuando se enfrentan a realidades complejas como el mercado inmobiliario de Lima Metropolitana, caracterizado por su heterogeneidad estructural, informalidad parcial y amplia dispersión de precios.

En el caso de la regresión hedónica lineal, uno de los principales inconvenientes radica en su supuesto de relación lineal y constante entre las características del bien (como área, número de baños, estacionamientos, ubicación, antigüedad, etc.) y su precio. Este enfoque presupone que el efecto de cada variable explicativa sobre el precio es independiente y uniforme en toda la muestra, lo cual rara vez se cumple en mercados reales. En la práctica, las interacciones entre variables son comunes (por ejemplo, el impacto del área puede ser distinto en un departamento con o sin cochera), y los efectos marginales pueden variar dependiendo del distrito, el segmento socioeconómico o la dinámica local del mercado.

Otro problema estructural es que los modelos lineales no capturan bien relaciones no lineales o de saturación, como por ejemplo cuando un aumento adicional de metros cuadrados no incrementa el valor en la misma proporción. Además, son sensibles a problemas de multicolinealidad, lo que puede distorsionar los coeficientes estimados, y requieren que los errores cumplan supuestos de normalidad, homocedasticidad y no autocorrelación, condiciones que no siempre se verifican con datos inmobiliarios reales. Esto afecta la validez de los resultados y puede generar estimaciones inestables o poco fiables.

En cuanto a las tasaciones comparativas, su principal debilidad es la alta dependencia del juicio subjetivo del perito. Si bien los tasadores tienen conocimiento del mercado y acceso a información privilegiada, su criterio puede verse afectado por sesgos personales, disponibilidad limitada de comparables válidos o incentivos externos. Esto conlleva una variabilidad en los resultados que limita la reproducibilidad de los informes de tasación y genera riesgos al momento de tomar decisiones financieras basadas en ellos, como la aprobación de créditos hipotecarios o la valorización de activos en fondos inmobiliarios.

En ambos casos, estas limitaciones representan un riesgo financiero real. Una valoración incorrecta —por subvaloración o sobrevaloración— puede llevar a decisiones de préstamo mal estructuradas, exposición al incumplimiento de pagos, mal cálculo del valor de garantías, o incluso sobreasignación de recursos de inversión. En contextos donde

el sistema financiero depende de forma considerable del crédito garantizado por bienes inmuebles, como en el Perú, esta situación puede afectar la estabilidad institucional, la eficiencia en la asignación de capital y la salud del sistema en su conjunto.

Frente a estas limitaciones, el modelo propuesto en esta investigación —XGBoost (Extreme Gradient Boosting)— representa una alternativa metodológica más robusta y adaptada a la complejidad del problema. A diferencia de los modelos lineales, XGBoost no impone una forma funcional rígida a la relación entre variables. Al estar basado en un conjunto de árboles de decisión optimizados secuencialmente, puede capturar relaciones no lineales, interacciones entre variables, y patrones específicos dentro del conjunto de datos, sin que estos deban ser definidos explícitamente por el investigador.

Otra ventaja importante es que XGBoost realiza una segmentación automática del espacio de variables, dividiendo los datos en regiones más homogéneas mediante decisiones sucesivas, lo que permite que el modelo se adapte a diferencias estructurales entre distritos o zonas urbanas. Esta capacidad es particularmente valiosa en ciudades como Lima, donde los determinantes del precio inmobiliario varían significativamente entre Miraflores, Jesús María, Comas o San Miguel.

Además, el algoritmo incluye mecanismos para reducir el sobreajuste, es decir, para evitar que el modelo se adapte en exceso a los datos de entrenamiento y pierda capacidad de generalización. También es menos sensible a valores extremos (outliers) y no requiere supuestos estrictos sobre la distribución de los errores. Esto lo convierte en una herramienta más robusta para la predicción de precios, con mayor precisión y menor dispersión en los resultados.

Por tanto, el uso de XGBoost no solo es una mejora técnica en términos de precisión, sino una respuesta directa a las limitaciones estructurales de los modelos clásicos. Esta justificación metodológica —que se acompaña de evidencia empírica en el capítulo de resultados— sustenta la necesidad de aplicar modelos de Machine Learning cuando se busca una estimación más confiable del valor de activos inmobiliarios, especialmente en

contextos donde una mala valoración puede traducirse en pérdidas financieras o exposición a riesgo crediticio.

1.2 Contribución financiera de la investigación

En la actualidad, la correcta valorización de los activos inmobiliarios es un elemento central para la estabilidad del sistema financiero. La subvaloración o sobrevaloración de inmuebles puede generar una serie de distorsiones en la colocación de créditos, en la evaluación de garantías, en la valoración de portafolios de inversión, e incluso en la política económica. Estas distorsiones afectan directamente a bancos, inversionistas institucionales, reguladores, y al funcionamiento eficiente del mercado.

En el caso del mercado inmobiliario limeño, el uso predominante de métodos tradicionales como las tasaciones subjetivas o los modelos hedónicos lineales limita la precisión de las valoraciones, lo que genera un riesgo latente en las decisiones financieras. Esta situación constituye un problema financiero estructural: si el precio observado de un inmueble no representa adecuadamente su valor fundamental, los agentes financieros operan sobre información imperfecta, lo que puede llevar a una mala asignación del crédito, desequilibrios en las carteras de inversión y una mayor exposición al riesgo sistémico.

La presente tesis plantea como solución el uso de modelos de predicción basados en aprendizaje automático, específicamente el algoritmo XGBoost, cuyo diseño permite captar relaciones complejas y no lineales entre múltiples variables que afectan el precio de un inmueble. La evidencia empírica generada en esta investigación demuestra que este modelo mejora significativamente la precisión de las predicciones en comparación con modelos tradicionales. Esta mejora, sin embargo, no es solo técnica o estadística; tiene implicancias directas en la gestión financiera.

Desde el punto de vista de las entidades financieras, una mejor predicción del valor de los inmuebles mejora la estimación del valor de las garantías hipotecarias, permitiendo una asignación más eficiente de los montos de crédito. Esto reduce el riesgo de sobreendeudamiento y protege a las entidades frente a escenarios de incumplimiento, ya

que los créditos se respaldan con activos cuyo valor ha sido estimado de forma más fiable. De este modo, se mejora la calidad de los activos en el balance y se refuerza la solvencia bancaria.

Para el riesgo de crédito, la precisión del modelo permite establecer ratios más realistas como el Loan-to-Value (LTV), fundamentales para evaluar la exposición al riesgo. Un modelo como XGBoost puede integrarse directamente en los procesos internos de análisis de crédito, reduciendo la subjetividad en la evaluación y permitiendo decisiones más objetivas, basadas en datos. Esto es particularmente importante en contextos de alta volatilidad económica, donde la incertidumbre en la valuación de activos puede generar un efecto multiplicador de riesgos.

En cuanto a los inversionistas institucionales, como fondos de inversión inmobiliaria, compañías de seguros o bancos de inversión, un modelo predictivo de alta precisión permite valorar activos con mayor exactitud y tomar decisiones más informadas en relación con la compra, venta o mantenimiento de propiedades en sus portafolios. Esto impacta directamente en la rentabilidad esperada, el nivel de riesgo asumido y la alineación con objetivos de retorno ajustado por riesgo.

Desde el punto de vista de la regulación y supervisión financiera, el modelo propuesto puede ser utilizado como un instrumento de vigilancia macro prudencial. Las autoridades como la SBS o el BCRP pueden emplear modelos predictivos como indicadores adelantados de posibles desalineaciones entre precios de mercado y valores fundamentales. Esto permite detectar burbujas inmobiliarias incipientes, evaluar la concentración de riesgo en ciertos segmentos del mercado y diseñar políticas regulatorias más preventivas y menos reactivas. Asimismo, al mejorar la transparencia y objetividad en las valoraciones, se fortalece la confianza en el sistema financiero y se reducen los incentivos para prácticas especulativas.

Además, en el contexto actual de transformación digital del sistema financiero, la implementación de modelos como XGBoost se alinea con las tendencias de digitalización

del crédito, análisis de *big data* y automatización de procesos de originación. Su integración en plataformas tecnológicas permitiría a bancos, fintechs y entidades de financiamiento desarrollar sistemas de evaluación más rápidos, precisos y escalables.

En resumen, el problema financiero identificado en esta investigación –la imprecisión en la valoración de activos inmobiliarios residenciales en Lima– genera efectos negativos en múltiples niveles del ecosistema financiero. La solución propuesta, basada en la aplicación de un modelo avanzado de aprendizaje automático, no solo mejora la precisión de la estimación, sino que contribuye a resolver riesgos estructurales en el sistema: mejora la asignación de crédito, fortalece la solvencia bancaria, reduce la exposición al riesgo de crédito, mejora la toma de decisiones de inversión y proporciona herramientas útiles para la supervisión regulatoria.

Por tanto, esta tesis no solo aporta valor desde la perspectiva metodológica, sino que responde de manera concreta a un problema financiero de relevancia nacional, con alto potencial de aplicación práctica en el sistema financiero peruano.

1.3 Objetivo de investigación

El objetivo general se centra en evaluar la efectividad de los modelos de ML (Árbol de Decisión y *XGBoost*) en la predicción de precios de departamentos en Lima, comparando su desempeño con la regresión hedónica tradicional a través de métricas de precisión como MAPE, MAE, RMSE y R^2, los cuales se definen en el apartado 2.6.1.

Cabe precisar que el objetivo de esta investigación no se limita a una mejora técnica incremental respecto a los modelos tradicionales de predicción de precios inmobiliarios. Por el contrario, surge de la identificación de una deficiencia metodológica de fondo: los modelos convencionales, como la regresión hedónica lineal, no logran captar adecuadamente la complejidad y heterogeneidad estructural del mercado inmobiliario de Lima Metropolitana, lo cual genera errores sistemáticos en las estimaciones de precios.

Estas deficiencias técnicas, lejos de ser un problema aislado, tienen consecuencias financieras concretas. Una valoración incorrecta puede llevar a que una entidad financiera apruebe un crédito hipotecario sobre una propiedad sobrevalorada, o que un fondo inmobiliario tome decisiones de inversión basadas en información imprecisa, lo que incrementa el riesgo crediticio, reduce la eficiencia en la asignación de recursos y deteriora la calidad de los balances financieros.

En ese sentido, el propósito de este estudio no es únicamente mostrar que el modelo XGBoost predice "mejor", sino demostrar que es una herramienta metodológicamente más apropiada para abordar un problema que los modelos tradicionales no resuelven bien. Su capacidad para modelar relaciones no lineales, adaptarse a distintas estructuras de datos y segmentar automáticamente patrones de comportamiento le permite responder de manera más precisa a los desafíos del entorno urbano e inmobiliario de Lima.

Por tanto, la tesis propone una solución metodológica que mejora la precisión predictiva con un objetivo mayor: reducir el riesgo financiero asociado a valoraciones inadecuadas. Esto tiene impacto directo en decisiones críticas como la colocación de créditos, la fijación de precios de garantías hipotecarias, la evaluación de riesgos por parte de los reguladores, y la planificación de inversiones en activos inmobiliarios. En este marco, el objetivo general no debe entenderse como una optimización estadística, sino como una contribución práctica y aplicable al análisis financiero moderno.

Además, se lista los siguientes objetivos específicos:

- Aplicar modelos de ML para la estimación de precios de departamentos en Lima, considerando atributos como tamaño, número de garajes y ubicación.
- Comparar la precisión predictiva de los modelos de ML frente a la regresión hedónica, utilizando métricas de error y coeficiente de determinación.
- Determinar la importancia relativa de las características de los departamentos en la predicción de precios según el modelo aplicado.

- Analizar el desempeño de los modelos en diferentes distritos de Lima para identificar variaciones en la precisión de las estimaciones.
- Proponer mejoras en la metodología de estimación de precios inmobiliarios mediante la integración de técnicas avanzadas de ML.

1.4. Hipótesis

La implementación de modelos de ML, en particular el algoritmo XGBoost, permite mejorar significativamente la precisión en la estimación del valor de departamentos en Lima, en comparación con los métodos tradicionales como la regresión hedónica. Esta mejora predictiva no solo representa un avance metodológico, sino que tiene implicancias directas en la gestión de riesgos financieros y la toma de decisiones estratégicas por parte de entidades financieras, inversionistas y reguladores.

Una predicción más precisa del precio inmobiliario permite una mayor alineación con el valor fundamental del activo, lo cual es esencial para evitar distorsiones en la valorización de garantías, reducir la exposición al riesgo crediticio, y fortalecer los mecanismos de supervisión macro prudencial. En particular, al minimizar los errores de sobrevaloración o subvaloración, se mejora la eficiencia en la colocación de créditos hipotecarios, se optimiza la construcción de portafolios inmobiliarios y se fortalece el análisis de riesgo en instituciones financieras.

Además, esta hipótesis se sustenta en la teoría financiera moderna, que sostiene que los precios de los activos deben reflejar sus fundamentales para garantizar la eficiencia de los mercados (Fama, 1970). La existencia de errores sistemáticos en la valoración de activos puede generar burbujas especulativas, pérdidas financieras por colaterales mal estimados y deterioro en la calidad de los balances de las entidades financieras. En este sentido, el uso de modelos avanzados de ML permite mitigar estas ineficiencias mediante una mejor modelación del comportamiento del mercado, adaptándose a no linealidades y complejidades inherentes al mercado inmobiliario urbano.

Por tanto, se plantea que un modelo de ML con alta capacidad explicativa y predictiva, como XGBoost, constituye una herramienta superior para estimar precios de inmuebles y al mismo tiempo para mejorar la gestión del riesgo financiero asociado a dichos activos.

1.5. Alcances y limitaciones

1.5.1 *Alcances*

El desarrollo del presente trabajo comprende el uso de datos públicos para el análisis estadístico tradicional mediante regresión lineal y de ML necesario para lograr los objetivos propuestos.

Se utiliza información del Banco Central de Reserva del Perú (BCRP) como base de datos para el análisis. El BCRP recopila información de precios de inmuebles y sus características de manera trimestral desde 1998, sin embargo, esta investigación se basará solo en la información publicada desde 2014 en adelante, donde empieza una etapa de moderación en el crecimiento de precios inmobiliarios según indica Mundaca y Sánchez (2018); teniendo como distritos observables a San Isidro, Miraflores, La Molina, Surco, San Borja, Jesús María, Magdalena, Lince, San Miguel, Pueblo Libre.

La base de datos utilizada en este estudio contiene información de 22 distritos de Lima Metropolitana. Sin embargo, para el análisis se seleccionaron 10 distritos específicos: La Molina, Miraflores, San Borja, Surco, San Isidro, Jesús María, Magdalena, Pueblo Libre, Lince y San Miguel.

Esta selección responde a dos criterios fundamentales. Primero, la mayor disponibilidad de datos, ya que estos distritos concentran una alta cantidad de registros desde 2014 en adelante, representando una porción significativa de la base de datos. Por ejemplo, Miraflores (8,094 registros), Surco (8,083) y San Miguel (5,325) cuentan con un volumen considerable de datos, lo que permite una mejor calibración y validación del modelo de predicción.

En segundo lugar, se consideró la homogeneidad del mercado inmobiliario. Estos distritos presentan una dinámica de precios más estable y una oferta inmobiliaria más consolidada en comparación con otros distritos con menor cantidad de transacciones registradas. Al centrarse en zonas con una mayor actividad inmobiliaria, se minimiza la variabilidad estructural que podría sesgar los resultados del modelo.

Asegurando así un estudio más representativo y confiable en términos de predicción de precios inmobiliarios.

El tamaño de la muestra es un aspecto fundamental en la aplicación de modelos de regresión, ya que influye directamente en la precisión, estabilidad y generalización de los resultados obtenidos. En esta investigación, se ha trabajado con una base de datos compuesta por aproximadamente 40,000 observaciones de departamentos ubicados en 10 distritos de Lima, lo cual constituye una muestra robusta y adecuada para el tipo de modelo hedónico utilizado.

De acuerdo con la teoría estadística y econométrica, existe una "Rule of Thumb" (regla práctica) que sugiere que los modelos de regresión pueden funcionar razonablemente bien con apenas 50 observaciones, siempre que el número de variables explicativas no sea excesivo. Esta regla se basa en la relación entre el número de observaciones y el número de parámetros a estimar, lo que resulta crucial para evitar problemas como la multicolinealidad, el sobreajuste y la inestabilidad de los coeficientes.

El hecho de que este estudio utilice aproximadamente 40,000 observaciones implica un nivel muy superior al mínimo recomendado, lo que fortalece considerablemente la calidad y robustez del modelo. A continuación, se explican las principales ventajas de contar con un tamaño de muestra amplio:

- Robustez y Estabilidad de los Coeficientes Estimados

Un tamaño de muestra considerable contribuye a que las estimaciones de los coeficientes de las variables explicativas sean más estables y precisas. Al aumentar la

cantidad de datos, se reducen los errores estándar asociados a las estimaciones, lo que mejora la significancia estadística de los coeficientes y aumenta la confiabilidad del modelo. En un contexto inmobiliario, esto es especialmente importante, ya que se busca captar de manera precisa el impacto de cada característica del inmueble (como área, antigüedad, ubicación, entre otras) sobre su precio.

- Representatividad de la Muestra y Generalización de Resultados

El uso de 40,000 observaciones permite que la muestra capture una amplia heterogeneidad en las características de los departamentos analizados y en las condiciones de mercado de los distintos distritos. Esta diversidad contribuye a que el modelo pueda ser más representativo de la realidad del mercado inmobiliario en Lima y, por tanto, que los resultados obtenidos puedan generalizarse con mayor precisión a otros contextos similares.

- Reducción de Problemas Estadísticos

Los modelos de regresión aplicados a muestras pequeñas suelen enfrentar problemas estadísticos que pueden comprometer la validez de las estimaciones. Entre los problemas más comunes se encuentran:

• Multicolinealidad: Relación lineal entre las variables explicativas que puede inflar los errores estándar y reducir la precisión de las estimaciones.

• Heterocedasticidad: Variabilidad desigual de los errores del modelo, que puede afectar la eficiencia de los estimadores.

• Sobreajuste (Overfitting): Situación en la que el modelo ajusta perfectamente los datos de la muestra, pero pierde capacidad de generalización a nuevas observaciones.

El uso de un tamaño de muestra grande contribuye a mitigar estos problemas, ya que mejora la potencia estadística del modelo y reduce la variabilidad inherente en los datos. Esto permite que las relaciones identificadas entre las variables explicativas y el precio de los departamentos sean más robustas y fiables.

- Análisis Comparativos y Desagregados

Una muestra amplia también ofrece la posibilidad de realizar análisis más desagregados o comparativos, lo que puede enriquecer las interpretaciones y aplicaciones del modelo. Por ejemplo, el tamaño de muestra utilizado en esta investigación podría permitir analizar diferencias en el impacto de las variables explicativas según el distrito, el rango de precios o la antigüedad de los inmuebles, lo que aporta un nivel adicional de detalle y utilidad a los resultados obtenidos.

- Cumplimiento de Criterios Estadísticos y Econométricos

Además de la "Rule of Thumb" mencionada, existen otros criterios estadísticos que sugieren que el tamaño de la muestra debe ser al menos 10 veces mayor que el número de parámetros a estimar en un modelo de regresión lineal. Este criterio garantiza que el modelo tenga suficientes grados de libertad para realizar estimaciones precisas y minimizar el riesgo de problemas estadísticos.

En este estudio, dado que el número de observaciones es aproximadamente 40,000 y el número de variables explicativas es significativamente menor, se cumple ampliamente este criterio, lo que refuerza la solidez del modelo aplicado.

1.5.2 *Limitaciones*

- La base de datos disponible del BCRP no cuenta con la ubicación exacta de los inmuebles, lo que impide mejorar la precisión de la característica de zonificación.
- No es posible medir la variable de ubicación en función de la distancia a puntos estratégicos de la ciudad, como centros comerciales, colegios, centros de salud o espacios de esparcimiento.
- El muestreo realizado por el BCRP no posee la rigurosidad metodológica de un instituto de estadística, ya que la información proviene de fuentes como Urbania y Adondevivir. Esto puede generar que, en algunos casos, las características de los inmuebles no estén completamente detalladas. No obstante, el muestreo del BCRP

cuenta con una cuota fija de observaciones por distrito en cada periodo, lo que permite disponer de información suficiente para el análisis.

- No se dispone de información de la Cámara Peruana de la Construcción (CAPECO) para replicar las ponderaciones realizadas por el BCRP en el índice de precios de inmuebles para Lima Metropolitana y analizar si los precios se encuentran fuera del rango de sus fundamentos. Sin embargo, esto no representa un problema, ya que el enfoque del presente estudio se centra en la construcción de un método alternativo de predicción y no en la replicación de un índice.
- Los datos recopilados en diferentes períodos no necesariamente corresponden al mismo inmueble, dado que la probabilidad de que un mismo inmueble sea negociado varias veces es baja, lo que podría reducir el tamaño de la muestra. No obstante, el uso de datos de distintos inmuebles en diferentes periodos aporta mayores beneficios para el análisis.

En el desarrollo de este modelo de precios hedónicos para la estimación de valores inmobiliarios en 10 distritos de Lima, se ha puesto énfasis en la inclusión de variables endógenas relacionadas con las características estructurales de los departamentos (como área construida, número de habitaciones y antigüedad), su ubicación geográfica y aspectos ambientales. Este enfoque busca explicar las diferencias en los precios a partir de atributos observables y medibles, alineándose con la teoría económica de los precios hedónicos.

Sin embargo, es importante destacar que ciertos factores externos, también conocidos como variables exógenas, pueden influir en el mercado inmobiliario y, por ende, en los precios de los inmuebles. Estas variables, al no formar parte del núcleo del modelo, no han sido incorporadas explícitamente en el análisis. A continuación, se listan algunas de estas variables exógenas, junto con una breve explicación de su relevancia y de los motivos que justifican su ausencia en esta investigación:

- Inflación:

La inflación es un factor macroeconómico que afecta el poder adquisitivo de la población y, por tanto, tiene un impacto indirecto en el mercado inmobiliario. Cuando la inflación aumenta, el valor real del dinero disminuye, lo que puede llevar a un incremento en los precios nominales de los bienes raíces. Sin embargo, dado que el modelo desarrollado se enfoca en un análisis transversal basado en datos de un período específico, el efecto de la inflación ha sido considerado constante en el corto plazo y no se ha incorporado de manera explícita. Esto permite aislar mejor el impacto de las características intrínsecas de los inmuebles.

- Tasas de interés:

Las tasas de interés, especialmente las asociadas a créditos hipotecarios son determinantes clave en la accesibilidad de la vivienda. Un aumento en las tasas de interés encarece el financiamiento hipotecario, lo que puede reducir la demanda de viviendas y, en consecuencia, ejercer presión a la baja sobre los precios. La ausencia de esta variable en el modelo se debe a la dificultad de medir su impacto diferencial en cada distrito y a la necesidad de mantener un enfoque centrado en las características específicas de los inmuebles.

- Condiciones regulatorias y políticas públicas:

El mercado inmobiliario también está influido por factores regulatorios, como las normas de zonificación, los permisos de construcción y las políticas de subsidio a la vivienda. Por ejemplo, restricciones de zonificación que limitan la densidad poblacional pueden aumentar el valor del suelo urbano, mientras que políticas de subsidios pueden incentivar la construcción de viviendas sociales en determinadas zonas. La complejidad de capturar estos efectos de manera uniforme en los diferentes distritos analizados ha llevado a su exclusión del modelo.

- Factores macroeconómicos y socioeconómicos:

Además de la inflación y las tasas de interés, otros factores macroeconómicos, como el crecimiento económico, el empleo y la confianza del consumidor, pueden influir en la dinámica del mercado inmobiliario. A nivel microeconómico, el perfil socioeconómico de los residentes de cada distrito (ingresos, nivel educativo, etc.) también es relevante. Si bien algunos de estos factores podrían aportar valor adicional al modelo, su incorporación requeriría información detallada que no siempre está disponible a nivel distrital.

- Factores exógenos imprevistos:

Fenómenos externos como desastres naturales (terremotos, inundaciones) o crisis sanitarias (como la pandemia de COVID-19) pueden generar impactos significativos en el mercado inmobiliario al alterar la oferta y demanda de viviendas. Este tipo de eventos, debido a su carácter imprevisible y su impacto altamente variable, no han sido considerados en el presente análisis.

1.6. Contribución

El presente trabajo de investigación se centra en la aplicación de la metodología de precios hedónicos en el mercado inmobiliario de Lima. Nuestra contribución radica en dos aspectos principales: primero, mejorar la estimación de precios de departamentos mediante la transición de métodos estadísticos tradicionales a enfoques basados en ML. Segundo, garantizar la interpretabilidad de los modelos de ML, proporcionando una comparación de la importancia de las variables en la estimación de precios para cada uno de los distritos analizados, incluyendo San Isidro, La Molina, Miraflores, Surco, San Borja, Magdalena, Pueblo Libre, Lince y Jesús María.

Desde el punto de vista académico y financiero, esta investigación representa un aporte significativo al desarrollo de modelos de valoración de activos inmobiliarios aplicados al sistema financiero peruano. En la literatura financiera, la estimación del valor fundamental de un activo es clave para evitar sobrevaloraciones que puedan dar lugar a burbujas especulativas. Esta tesis se alinea con esta corriente al proponer una metodología que,

mediante el uso de algoritmos de aprendizaje automático, permite obtener estimaciones más precisas y ajustadas a las características intrínsecas del inmueble.

En primer lugar, el trabajo se enmarca en el enfoque de estabilidad financiera, ya que una estimación adecuada de precios inmobiliarios reduce la posibilidad de deterioro en las carteras hipotecarias de las instituciones financieras, mejorando la gestión del riesgo de crédito. Una valoración más precisa de los activos inmobiliarios permite a los bancos establecer márgenes de garantía adecuados y tomar decisiones de colocación más seguras.

Además, este modelo puede ser utilizado como un insumo en la supervisión macro prudencial por parte de entidades regulatorias como la Superintendencia de Banca, Seguros y AFP (SBS) o el Banco Central de Reserva del Perú (BCRP). La detección temprana de desviaciones significativas entre los precios de mercado y los valores estimados por el modelo puede servir como una señal de alerta ante potenciales desequilibrios financieros o burbujas inmobiliarias.

Por otro lado, esta investigación también se conecta con el análisis de portafolios de inversión. Fondos de inversión inmobiliarios, compañías de seguros y bancos de inversión requieren valorar adecuadamente sus activos subyacentes para garantizar una adecuada asignación de recursos, una correcta medición del rendimiento ajustado por riesgo y una valorización realista de su patrimonio. Al proveer un modelo de alta precisión para la valoración de activos residenciales, este trabajo contribuye directamente a mejorar la toma de decisiones en estos contextos.

Finalmente, este trabajo también dialoga con la literatura internacional sobre la aplicación de inteligencia artificial en finanzas. Estudios como el de Daníelsson et al. (2020) han demostrado cómo los modelos de ML pueden contribuir a la estabilidad financiera mediante una mejor estimación de riesgos. En esa línea, esta investigación amplía el campo de aplicación de dichos modelos al caso peruano, aportando evidencia empírica relevante que puede ser replicada o extendida a otras economías emergentes.

CAPÍTULO II: APLICACIÓN DE LA METODOLOGÍA

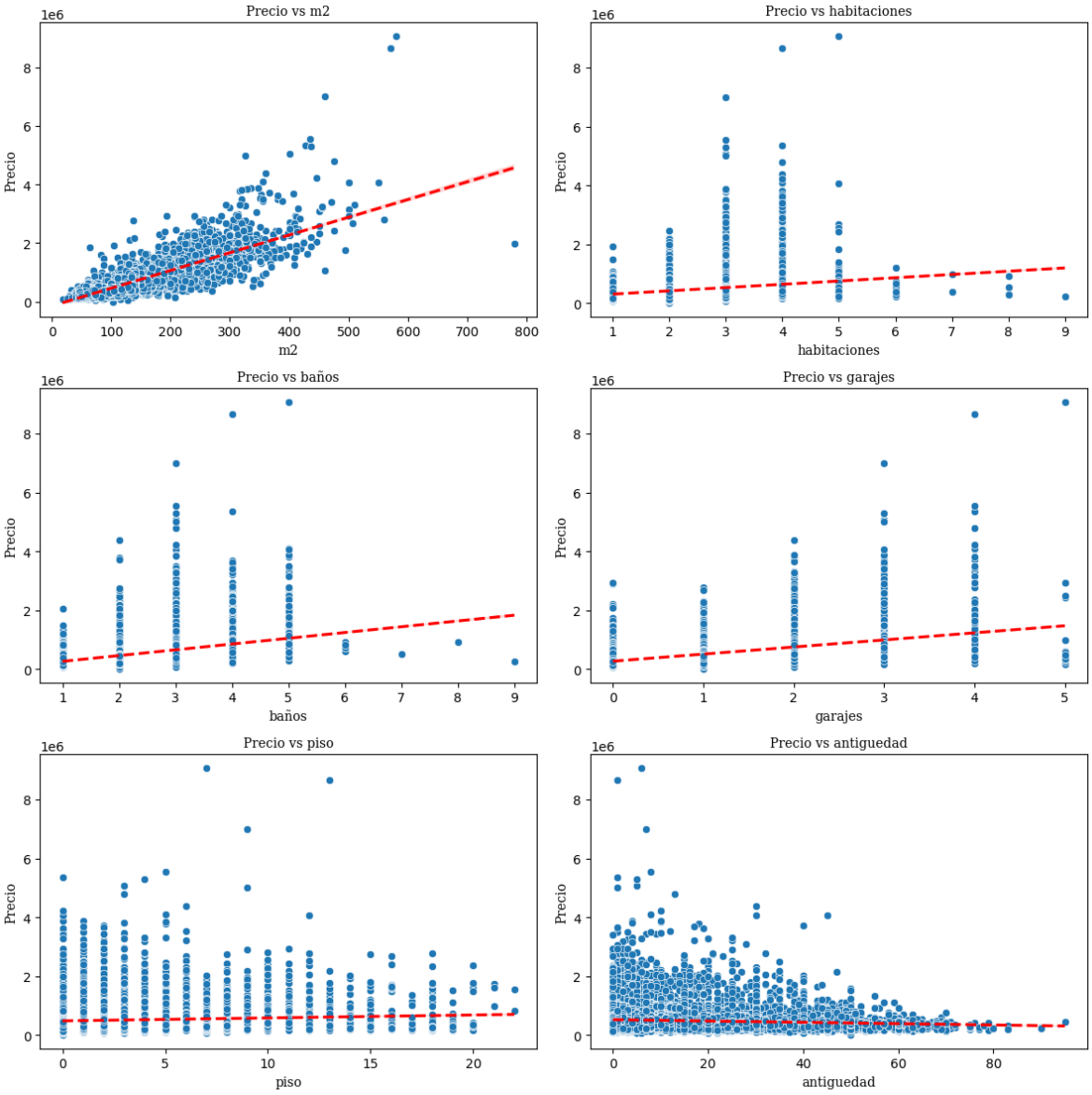
2.1. Descripción de datos

El BCRP recopila y publica información sobre el precio de departamentos en Lima Metropolitana de los distritos que tienen un mercado más dinámico. En años anteriores se hacía a través de anuncios publicitarios en los periódicos, pero actualmente la digitalización de la información permite la recopilación de información a través de portales web como "Urbania", "Nexo Inmobiliario" o "Adondevivir"; esta información es más rápida de obtener y mantener actualizada además de recopilar más información con detalles adicionales como metros cuadrados, número de habitaciones, vista interna o externa, cantidad baños, entre otras características de interés.

La base de datos a utilizar cuenta con 43,457 observaciones de departamentos de los distritos San Isidro, Miraflores, La Molina, Surco, San Borja, Jesús María, Magdalena, Lince, San Miguel y Pueblo Libre desde 2014 hasta 2024 con frecuencia trimestral considerando las variables precio en soles constantes 2009 (precio), tipo de cambio (tc), índice de precios del consumidor (ipc), metros cuadrados (m2), cantidad de habitaciones (habitaciones), cantidad de baños (baños), cantidad de garajes (garajes), número de piso (piso), si tiene vista interna o externa (vista), años de antigüedad (antigüedad) y distrito en el que se encuentra (distrito).

A continuación, la Figura 2.1. muestra las relaciones entre el precio y las características principales mediante diagramas de dispersión mostrando una tendencia positiva entre el precio y los metros cuadrados sugiriendo que el precio incrementa conforme aumenta el tamaño, al igual que con el número de garajes. Por otro lado, el piso en el que se encuentra el departamento parece no tener relevancia mientras que la antigüedad parece mostrar una ligera relación negativa con el precio. Y finalmente, la relación con el número de habitaciones parece ser ligeramente creciente, pero con alta dispersión al igual que los baños.

Figura 2.1. Diagramas de dispersión.

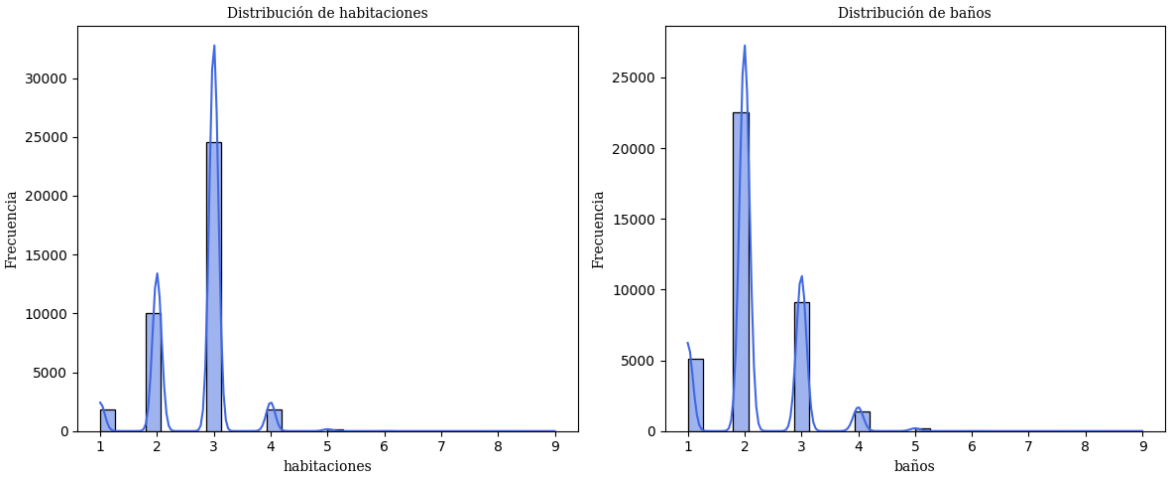


Fuente: Elaboración propia.

Es importante notar que existen observaciones con cantidad de baños y habitaciones por encima de 6 unidades, pero con precios relativamente bajos, generando preocupación en que la ligera tendencia positiva entre el precio y habitaciones y baños puede estarse viendo afectada por la posibilidad de existencia de un alto número de observaciones de estas características; que podría ser consecuencia de la falta de rigurosidad en la

recopilación de datos descrita en las limitaciones. Para poder confirmar que no se genera esta situación mostramos la Figura 2.2. donde se aprecia que la distribución de estas características se encuentra concentrada por debajo de 6 unidades; confirmando que la baja relación positiva entre el precio y dichas características es generada por observaciones que parecen no tener problemas en la recopilación de datos.

Figura 2.2. Distribución de habitaciones y baños.



Fuente: Elaboración propia.

2.2. Regresión hedónica

2.2.1. *Principios básicos y fundamento teórico*

La regresión hedónica es una técnica econométrica utilizada para descomponer el precio de un bien en el valor de sus características permitiendo cuantificar el valor marginal de cada una de las características, proporcionando una visión detallada de cómo afectan al precio final del departamento. Esta regresión se suele realizar mediante regresión lineal entre el precio y las características.

Desde un punto de vista matemático, la función de precios hedónicos se puede expresar de la siguiente manera lineal:

$$y_i = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon_i \, , \qquad (1)$$

donde y_i representa el precio del inmueble, y x_i son las características que influyen en el precio. Reemplazando las características en la función (1) obtenemos:

$$precio = \beta_0 + \beta_1 \text{tc} + \beta_2 \text{ipc} + \beta_3 \text{m2} + \beta_4 hab + \beta_5 \text{baño} + \beta_6 \text{garaje} + \beta_7 \text{piso} + \beta_8 \text{vista} + \beta_9 \text{antigüedad} + \beta_{10} \text{distrito}_2 + \cdots + \beta_{18} \text{distrito}_{10} + \varepsilon, \quad (2)$$

donde β_1, β_2,…, β_n son los coeficientes que indican la contribución marginal de cada característica al precio total, y d_i es cada uno de los distritos.

Para poder calcular el set de coeficientes estimados se aplica se debe realizar un proceso de optimización en el que se busca minimizar los errores al cuadrado siendo la función objetivo a minimizar:

$$\sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \quad (3)$$

obteniendo como resultado la siguiente forma general (4) de cómo se calcula el $\hat{\beta}$:

$$\hat{\beta} = (X'X)^{-1}(X'Y), \quad (4)$$

2.2.2. *Aplicación del modelo*

Para la aplicación de una regresión lineal partiendo de la base de datos obtenida del BCRP se necesita realizar una transformación de los datos, ya que el modelo lineal no admite variables de texto como el distrito. En ese sentido, se aplica *One-Hot Encoding* en la data para transformar la variable de texto distrito en 10 nuevas variables, una para cada distrito con valores de 1 en caso la observación se encuentre en ese distrito o 0 en caso contrario. Además, para evitar problemas de colinealidad es necesario omitir alguno de los distritos, y de esa manera se podrá realizar con éxito la estimación de coeficientes.

Dado que el objetivo es poder comparar posteriormente contra los modelos de ML, y como se explicará más a detalle en los puntos 2.3. y 2.4., se decidió separar la base de datos en 80% para entrenamiento y 20% para testeo del modelo de manera aleatoria; de tal

manera que la regresión lineal y cálculo de coeficientes se realiza con solo el 80% de la información generando los resultados de la Tabla 2.1.

Finalmente utilizando los coeficientes obtenidos en la regresión lineal de la data de entrenamiento procedemos a estimar precios en base a las características del 20% restante de información reservada para testeo del modelo y poder evaluar qué tan acertado es el modelo hedónico de regresión lineal en datos que no ha observado con anterioridad.

La regresión hedónica es una metodología ampliamente utilizada para estimar el valor de bienes inmuebles en función de sus características intrínsecas y extrínsecas. Este modelo se fundamenta en la teoría del precio hedónico, propuesta por Rosen (1974), que establece que el precio de un bien complejo, como una vivienda, puede descomponerse en los valores marginales de sus atributos (Rosen, 1974).

En el contexto del mercado inmobiliario de Lima, la regresión hedónica permite analizar el impacto de variables como la ubicación, el tamaño, el número de habitaciones y la antigüedad sobre el precio de los departamentos. A través de este análisis, es posible comprender cuáles son los factores que más influyen en la formación de precios y cómo estos pueden predecirse mediante técnicas de Machine Learning.

La matriz de correlación es una herramienta estadística utilizada para medir la relación entre las variables continuas de un modelo. Se construye calculando el coeficiente de correlación de Pearson entre pares de variables, proporcionando valores en un rango de -1 a 1:

Valores cercanos a 1 indican una fuerte correlación positiva.

Valores cercanos a -1 indican una fuerte correlación negativa.

Valores cercanos a 0 sugieren una correlación débil o inexistente.

Para el presente estudio, se excluyen variables discretas como los distritos, ya que estas se codifican de manera diferente (por ejemplo, mediante One-Hot Encoding). La matriz de

correlación permite identificar relaciones entre variables clave, como el área total, el número de habitaciones y el precio del inmueble.

En los modelos de regresión lineal tradicional, una alta correlación entre variables explicativas puede generar un problema conocido como multicolinealidad. Este fenómeno dificulta la interpretación de los coeficientes del modelo y puede afectar su estabilidad (Gujarati & Porter, 2009). Para detectar multicolinealidad, se suele calcular el Factor de Inflación de la Varianza (VIF), donde valores mayores a 10 indican problemas significativos.

2.2.3. *Resultado del modelo*

La Tabla 2.1. muestra los resultados de la regresión lineal del modelo hedónico obteniendo un R^2, que se define en el apartado 2.6.1., de 0.787 en la data de entrenamiento con coeficientes en su mayoría significativos al 95%, excepto la cantidad de baños; siendo las características más importantes la cantidad de garajes y metros cuadrados. Y como es de esperar, los distritos tienen un impacto negativo en comparación a estar ubicado en el distrito de San Isidro (variable omitida para tomar de referencia).

Tabla 2.1. Regresión lineal.

OLS Regression Results

Dep. Variable:	precio	R-squared:	0.787
Model:	OLS	Adj. R-squared:	0.787
Method:	Least Squares	F-statistic:	6310.
Date:	Sat, 08 Feb 2025	Prob (F-statistic):	0.00
Time:	02:06:41	Log-Likelihood:	-4.0761e+05
No. Observations:	30668	AIC:	8.153e+05
Df Residuals:	30649	BIC:	8.154e+05
Df Model:	18		
Covariance Type:	nonrobust		

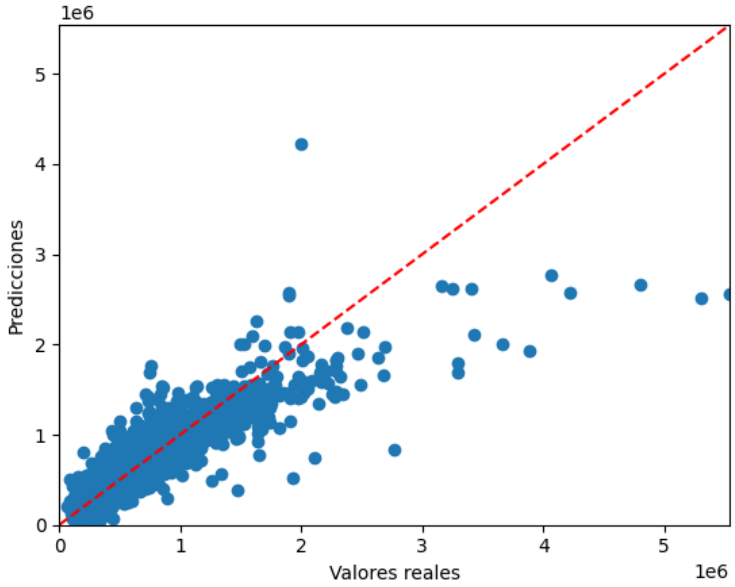
	coef	std err	t	P>|t|	[0.025	0.975]
const	4.301e+04	1.19e+04	3.615	0.000	1.97e+04	6.63e+04
tc	4.986e+04	2206.231	22.602	0.000	4.55e+04	5.42e+04
ipc	-644.2631	63.542	-10.139	0.000	-768.809	-519.717
m2	5279.9923	27.618	191.180	0.000	5225.860	5334.125
habitaciones	-2.681e+04	1518.093	-17.661	0.000	-2.98e+04	-2.38e+04
baños	2578.2127	1457.415	1.769	0.077	-278.380	5434.806
garajes	5.365e+04	1379.824	38.883	0.000	5.09e+04	5.64e+04
piso	835.4220	355.670	2.349	0.019	138.295	1532.549
vista	1.956e+04	4683.365	4.177	0.000	1.04e+04	2.87e+04
antiguedad	-2869.2797	75.318	-38.096	0.000	-3016.906	-2721.653
Jesús María	-1.485e+05	4297.258	-34.561	0.000	-1.57e+05	-1.4e+05
La Molina	-2.336e+05	4082.176	-57.234	0.000	-2.42e+05	-2.26e+05
Lince	-1.373e+05	4509.006	-30.451	0.000	-1.46e+05	-1.28e+05
Magdalena	-1.621e+05	4094.674	-39.582	0.000	-1.7e+05	-1.54e+05
Miraflores	-5.293e+04	3422.508	-15.464	0.000	-5.96e+04	-4.62e+04
Pueblo Libre	-1.829e+05	4388.523	-41.678	0.000	-1.92e+05	-1.74e+05
San Borja	-1.364e+05	3776.919	-36.126	0.000	-1.44e+05	-1.29e+05
San Miguel	-2.033e+05	4060.090	-50.071	0.000	-2.11e+05	-1.95e+05
Surco	-1.77e+05	3471.421	-50.998	0.000	-1.84e+05	-1.7e+05

Omnibus:	41947.198	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	59727265.964
Skew:	7.242	Prob(JB):	0.00
Kurtosis:	218.711	Cond. No.	2.73e+03

Fuente: Salida de la regresión en Python con resultados.

La Figura 2.3. muestra de manera gráfica la relación entre el precio estimado y el precio real en el 20% de data de testeo, pretendiendo dar señales de que el R^2 en la data de testeo puede ser bastante alto. Posteriormente, en el apartado 2.6. se definirá de manera cuantitativa como evaluar a profundidad y confirmar esta hipótesis.

Figura 2.3. Precios estimados de modelo hedónico de regresión lineal.



Fuente: Elaboración propia.

2.3. Árbol de Decisión

2.3.1. *Principios básicos y fundamento teórico*

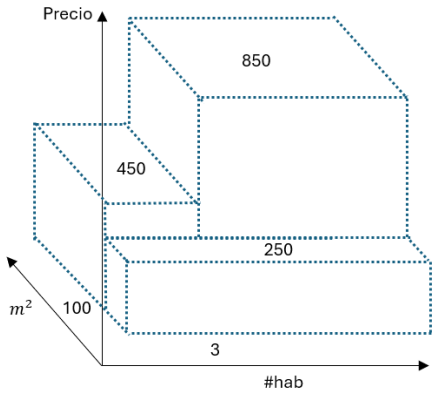
El Árbol de Decisión es un modelo predictivo que organiza la toma de decisiones mediante una estructura jerárquica similar a un árbol. Su funcionamiento se basa en dividir un conjunto de datos en subconjuntos más pequeños a través de una serie de preguntas o condiciones basadas en las características de los datos.

El árbol comienza en un nodo raíz, que representa el punto de partida donde se realiza la primera división del conjunto de datos. A partir de ahí, los nodos internos generan nuevas divisiones en función de distintos criterios, como el valor de una característica específica. Las ramas conectan estos nodos y representan los posibles caminos o resultados de las decisiones anteriores. Finalmente, los nodos hoja contienen la predicción final del modelo, que puede ser una clasificación o un valor numérico (como nuestro caso), dependiendo del tipo de problema.

En problemas de regresión, como la predicción de precios de departamentos, el valor de cada nodo hoja corresponde a una estimación del precio, calculada en función de las características del inmueble. El proceso de construcción del árbol busca minimizar el error de predicción dividiendo los datos de manera óptima en cada paso.

Para terminar de aterrizar el funcionamiento del árbol de decisión, la Figura 2.4. muestra un ejemplo de gráfico tridimensional de un Árbol de Decisión de solo 2 características. En este caso, cuando el tamaño es menor a 100, el precio promedio de los departamentos es 250; cuando el tamaño es mayor a 100 pero con menos de 3 habitaciones, el precio promedio es 450; y cuando es mayor de 100 y con más de 3 habitaciones el precio promedio es 850.

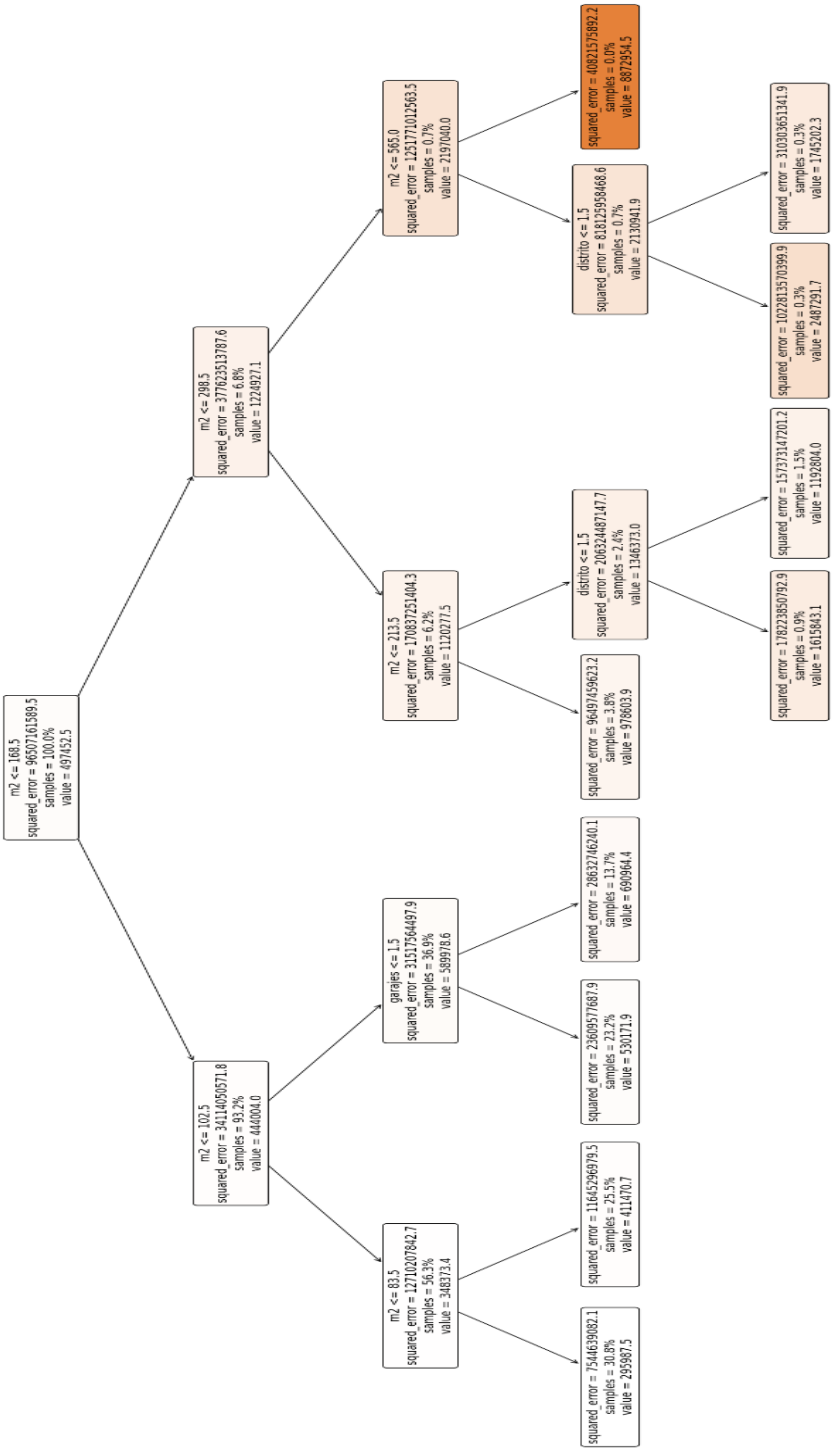
Figura 2.4. Ejemplo Tridimensional de árbol de decisión.



Fuente: Elaboración propia.

La Figura 2.5. grafica un ejemplo de árbol de decisión calculado con la data de entrenamiento limitando los *outputs* del modelo (u hojas de salida) a solo 10 posibilidades para poder mostrarlo a manera de ejemplo en el presente documento.

Figura 2.5. Ejemplo de árbol de decisión.



Fuente: Elaboración propia.

Formalizando la estructuración del árbol de decisión, el primer paso en la construcción de un árbol de decisión es subdividir el espacio de las variables explicativas $X_1, X_2, X_3, \ldots, X_p$ en posibles combinaciones que definan las distintas J regiones $R_1, R_2, R_3, \ldots, R_j$, dentro de las cuales se estimará la variable dependiente.

El objetivo de esta partición es minimizar la suma de los errores al cuadrado dentro de cada región R_j:

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2, \tag{5}$$

donde y_i es el valor real de la variable dependiente en la observación i, e $\hat{y}_{R_j}$es la predicción promedio dentro de la región R_j.

Explorar todas las posibles particiones del espacio de características sería computacionalmente inviable, por lo que se adopta un enfoque *top-down* o de división recursiva binaria. Este método comienza con todo el conjunto de datos en una sola región y, de manera sucesiva, divide el espacio de predictores en dos ramas. Para poder aplicar este criterio se elige un predictor X_j con un punto de corte s dividiendo el espacio en 2 regiones;

$$R_1(j, s) = \{X_j < s\}, \; R_2(j, s) = \{X_j \geq s\} \tag{6}$$

El criterio de división busca minimizar la suma de residuos al cuadrado total, definida como:

$$\sum_{i: x_i \in R_1(j.s)} (y_i - \hat{y}R_1)^2 + \sum_{i: x_i \in R_2(j.s)} (y_i - \hat{y}R_2)^2 \tag{7}$$

Se evalúan todos los predictores X_j y todos los posibles puntos de corte s para encontrar la combinación que minimice la RSS. Este proceso se repite de forma recursiva, dividiendo sucesivamente una de las regiones previamente definidas, creando así tres regiones, luego cuatro, y así sucesivamente hasta cumplir un criterio de parada. Finalmente, una vez que el

árbol ha sido construido, la predicción para una nueva observación se obtiene calculando el promedio de las observaciones de entrenamiento dentro de la región correspondiente.

2.3.2. *Aplicación del modelo*

En el caso del árbol de decisión, al igual que en la regresión lineal, es necesario transformar los datos mediante *One-Hot Encoding* pero sin la necesidad de omitir un distrito porque la colinealidad que no permite calcular la inversa de $(X'X)$ en la ecuación (4) no existe en el árbol de decisión ya que la forma funcional de optimización es distinta.

Este modelo aplicado desde Python necesita utilizar la librería *sklearn* de donde se obtiene:

- *sklearn.tree.DecisionTreeRegressor*: Implementa un árbol de decisión para regresión.
- *sklearn.tree.plot_tree*: Permite visualizar el árbol de decisión.
- *sklearn.model_selection.train_test_split*: Divide los datos en conjuntos de entrenamiento y prueba.
- *sklearn.metrics*: Contiene funciones para evaluar el rendimiento del modelo.

La creación del modelo se realiza mediante dt=DecisionTreeRegressor() para luego entrenarlo con dt.fit(x_dt_train, y_dt_train), donde dt es el nombre que le damos al árbol de decisión que estamos creando; x_dt_train es la matriz de características utilizadas para entrenar el modelo, así como y_dt_train es el vector de precios del mismo bloque de datos.

Después de entrenar el modelo, podemos realizar estimaciones sobre el 20% de los datos restantes, correspondientes al conjunto de prueba. Esto nos permitirá evaluar la capacidad del modelo de árbol de decisión para estimar los precios en datos que no ha visto previamente. Para hacer las estimaciones, utilizamos: dt_pred=dt.predict(x_dt_test).

Adicionalmente, con el árbol de decisión entrenado, se puede calcular la importancia relativa de las características, pero, como el *One-Hot Encoding* generó inicialmente una variable para cada distrito, las importancias de este grupo de variables se deben sumar para

entender la importancia del distrito de manera agrupada en comparación al resto de variables.

La importancia relativa de las características en un árbol de decisión se determina a partir de la reducción del criterio de división, que en este caso es el error cuadrático medio (MSE). En cada nodo, se elige la característica que genera la mayor disminución del error.

$$MSE(R) = \frac{1}{n}\sum_{i\epsilon R}(y_i - \hat{y}_R)^2, \tag{8}$$

donde, n es el número total de observaciones en el nodo R; y_i es el valor real de la variable objetivo para la observación i; e $\hat{y}_R = \frac{1}{n}\sum_{i\epsilon R} y_i$.

La relevancia de una característica se calcula sumando la reducción del MSE en todos los nodos donde ha sido utilizada como criterio de división. Finalmente, estos valores se normalizan para expresar la importancia en términos de un porcentaje relativo.

Para cada característica j, su importancia relativa se calcula como:

$$I_J = \frac{1}{N}\sum_{t=1}^{N} \Delta MSE_t\ , \tag{9}$$

donde I_j es la importancia de la característica j; N es el número total de nodos en los que la característica j ha sido utilizada para dividir; ΔMSE_t es la reducción del error cuadrático medio en el nodo t cuando se usa la característica j. Finalmente, se normaliza dividiendo cada I_j entre la suma total de importancias de todas las características para obtener un valor relativo entre 0 y 1.

La variación del MSE se calcula como la diferencia entre el error antes y después de realizar una división en dicho nodo; esta métrica mide cuánto mejora el modelo al realizar una partición basada en una característica específica.

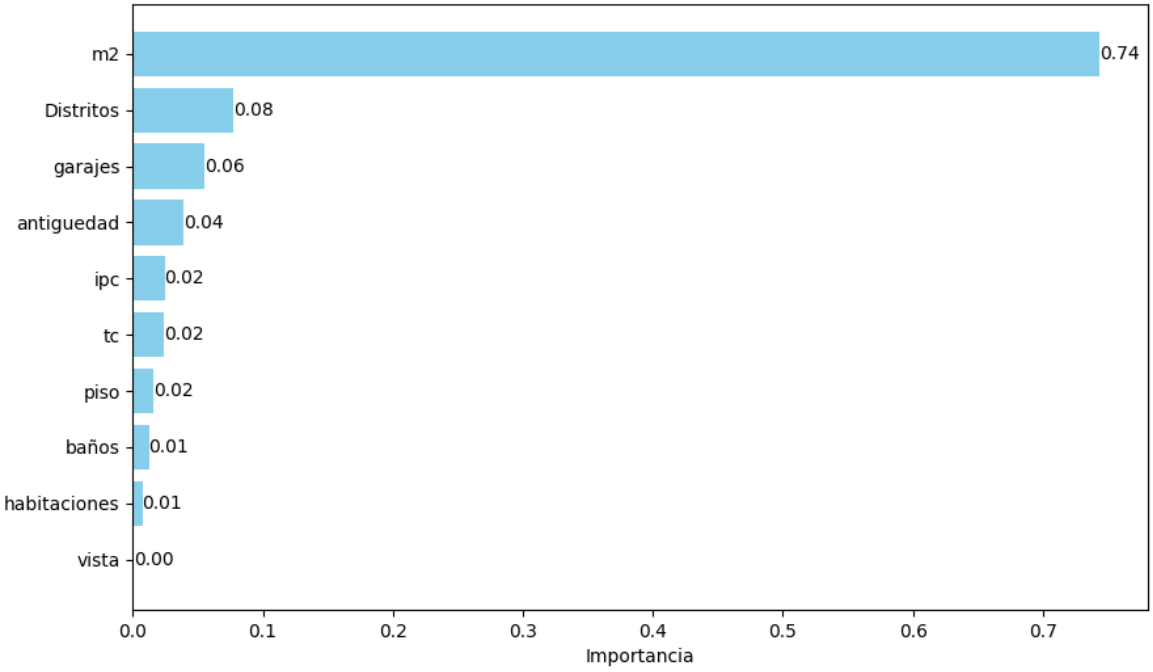
$$\Delta MSE = MSE_{padre} - \left(\frac{N_{izq}}{N_{total}} MSE_{izq} + \frac{N_{der}}{N_{total}} MSE_{der}\right), \tag{10}$$

donde MSE_{padre} es el error cuadrático medio antes de la división; MSE_{izq} es el error en el nodo hijo izquierdo; MSE_{der} es el error en el nodo hijo derecho; N_{izq} y N_{der} son el número de muestras en los nodos izquierdo y derecho, respectivamente; y $N_{total} = N_{izq} + N_{der}$ es el número total de muestras en el nodo padre.

2.3.3. *Resultado del modelo*

La Figura 2.6. muestra el resultado de las importancias relativas de las características utilizadas para estimar los precios de los departamentos, siendo el tamaño la característica con mayor importancia, contribuyendo en un 74% a la toma de decisiones dentro del árbol; seguido por 8% en los distritos y 6% en la cantidad de garajes.

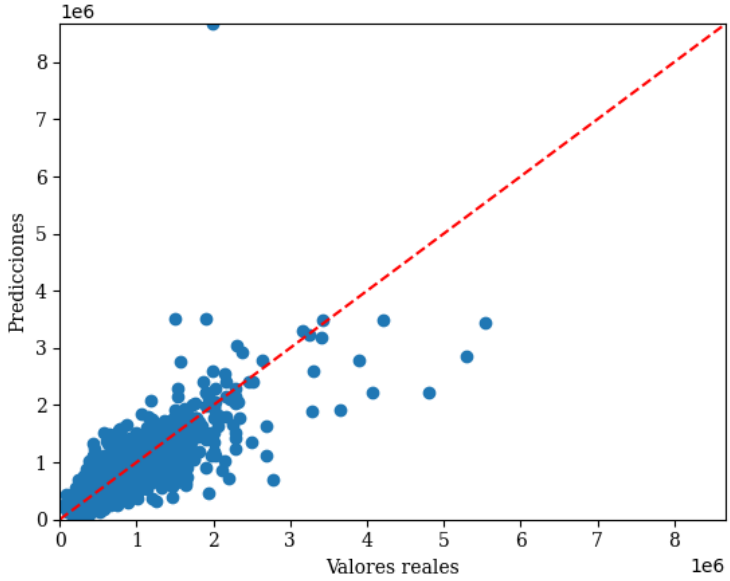
Figura 2.6. Importancia relativa de las características de Árbol de Decisión.



Fuente: Elaboración propia a partir de los resultados del Árbol de Decisión en python.

La Figura 2.7. muestra de manera gráfica la relación entre el precio estimado y el precio real en el 20% de data de testeo en el modelo de árbol de decisión, pareciendo tener un R^2 no tan alto como la regresión lineal. Esta gráfica es comparable con la Figura 2.3. del modelo hedónico.

Figura 2.7. Precios estimados de modelo de árbol de decisión.



Fuente: Elaboración propia.

2.4. XGBoost

2.4.1. Principios básicos y fundamento teórico

XGBoost es un algoritmo de ML basado en el método de *gradient boosting*, el cual construye múltiples árboles de decisión en secuencia con el objetivo de minimizar el error de predicción. Cada nuevo árbol se entrena para corregir los errores cometidos por los árboles anteriores, logrando una mejora progresiva en la capacidad predictiva del modelo, permitiendo modelar de manera más precisa dichas relaciones al generar árboles de decisión optimizados. En relación con la eficiencia del manejo de conjuntos de datos voluminosos, el sistema *XGBoost* se caracteriza por su capacidad para gestionar grandes cantidades de datos de manera eficiente, permitiendo un procesamiento más rápido de los datos y una utilización más efectiva de los recursos computacionales disponibles, lo que resulta en un sistema escalable capaz de manejar grandes volúmenes de data con menos recursos en comparación con otros sistemas.

XGBoost no solo se destaca por su rendimiento y eficiencia, sino también por la facilidad de uso, capacidad de adaptación a una amplia variedad de datos, de problemas y

de tipos de datos en el campo del aprendizaje automático. La combinación de velocidad, escalabilidad y precisión hace que sea una herramienta invaluable para cualquier profesional que busca desarrollar modelos predictivos de alto rendimiento en diversos dominios y aplicaciones. (Chen, T. y C. Guestrin, 2016).

El modelo se expresa como:

$$\hat{p}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in F, \tag{11}$$

donde $\hat{p}_i$ es la predicción del precio del inmueble i; x_i es el vector de características del i-ésimo departamento; $f_k(x_i)$ representa el k-ésimo árbol de decisión en el modelo; F es el espacio de todos los árboles de decisión; y K es el número total de árboles en el modelo.

El modelo minimiza la siguiente función de pérdida objetivo en la iteración t:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(p_i,\ \hat{p}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t), \tag{12}$$

donde $l(p_i, \hat{p}_i)$ es la función de pérdida:

$$l(p_i, \hat{p}_i) = \frac{1}{2}(p_i - \hat{p}_i)^2, \tag{13}$$

y $\Omega(f_t)$ es el término de regularización para evitar sobreajuste:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_j \omega_j^2, \tag{14}$$

donde T es el número de nodos en el árbol y ω_j son los pesos de las hojas. Para encontrar la mejor función f_t, se usa una expansión de segundo orden de Taylor de la función de pérdida; y aplicando (13) en (11) obtenemos:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma T + \frac{1}{2}\lambda \sum_j \omega_j^2, \tag{15}$$

donde g_i es la gradiente de pérdida:

$$g_i = \frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i}, \tag{16}$$

h_i la segunda derivada de la pérdida:

$$h_i = \frac{\partial^2 l(y_i, \hat{y}_i)}{\partial \hat{y}_i{}^2}, \tag{17}$$

y el peso ω_j de cada nodo del árbol:

$$\omega_j^* = -\frac{\sum_{i \epsilon I_j} g_i}{\sum_{i \epsilon I_j} h_i + \lambda}, \tag{18}$$

donde I_J es el conjunto de muestras en la hoja j.

La ganancia de dividir en nodo en dos ($L\ y\ R$) se mide como:

$$Gain = \frac{1}{2}\left[\frac{(\sum_{i \epsilon L} g_i)^2}{\sum_{i \epsilon L} h_i + \lambda} + \frac{(\sum_{i \epsilon R} g_i)^2}{\sum_{i \epsilon R} h_i + \lambda} - \frac{\sum_{i \epsilon L \cup R} g_i{}^2}{\sum_{i \epsilon L \cup R} h_i + \lambda} - \gamma\right. \tag{19}$$

Si la ganancia es menor que un umbral, la división no se realiza. Posteriormente se inicia un el proceso iterativo de entrenamiento partiendo de $\hat{p}_i = 0$; para cada iteración se calcula g_i y h_i; se construye un nuevo árbol f_t minimizando $\mathcal{L}^{(t)}$ y se actualiza la predicción

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i), \tag{20}$$

y se detiene cuando el modelo converge o se alcanza el número máximo de iteraciones.

La multicolinealidad, entendida como la existencia de una alta correlación entre dos o más variables explicativas, representa una limitación metodológica importante en los modelos lineales clásicos, como la regresión lineal múltiple. En dichos modelos, la presencia de multicolinealidad puede distorsionar las estimaciones de los coeficientes, dificultar la interpretación de los efectos individuales de las variables y generar inestabilidad en la predicción, especialmente cuando se trabaja con bases de datos complejas y de alta dimensionalidad como las que caracterizan al mercado inmobiliario.

No obstante, en el contexto de modelos basados en árboles, como XGBoost (Extreme Gradient Boosting), la multicolinealidad no representa una limitación significativa, gracias

al mecanismo con el que este algoritmo selecciona variables durante el entrenamiento. XGBoost construye árboles de decisión secuenciales mediante un enfoque de boosting, donde en cada iteración se seleccionan las variables que maximizan una función de ganancia (gain), en función de su capacidad para reducir el error de predicción. Si dos o más variables están altamente correlacionadas, el algoritmo identificará cuál de ellas ofrece una mayor ganancia de información en el conjunto de entrenamiento, y asignará menor importancia a las variables redundantes.

Esto implica que XGBoost realiza una selección implícita de variables durante el proceso de construcción del modelo, priorizando aquellas que aportan mayor valor predictivo y relegando las que presentan redundancia o bajo aporte marginal, aun cuando estén altamente correlacionadas con las variables seleccionadas. En efecto, la importancia de cada característica se evalúa en términos de su contribución a la mejora del modelo a lo largo de múltiples árboles, lo cual permite manejar automáticamente el problema de la multicolinealidad sin necesidad de preprocesamiento adicional.

Adicionalmente, el algoritmo incorpora técnicas de regularización L1 (Lasso) y L2 (Ridge), que penalizan la complejidad del modelo y atenúan el impacto de variables poco relevantes, reforzando su robustez ante relaciones espurias entre predictores. Esta propiedad ha sido reconocida en investigaciones previas, como la de Chen y Guestrin (2016), quienes destacaron que XGBoost no solo supera a modelos lineales en precisión, sino que también presenta una mayor estabilidad ante problemas clásicos como la multicolinealidad.

Por lo tanto, en el presente trabajo, se considera que la utilización de XGBoost es metodológicamente adecuada incluso en presencia de correlaciones elevadas entre predictores, ya que el modelo está diseñado para aprender de manera eficiente a partir de datos estructurados con características interrelacionadas, como ocurre en el análisis de precios inmobiliarios.

2.4.2. *Aplicación del modelo*

A diferencia de los modelos anteriores, el *XGBoost* permite manejar la variable distrito sin necesidad de transformar el dato con *One-Hot Encoding*; este algoritmo de ML es capaz de identificar la variable como categórica con solo una especificación para que lo entienda el software mediante *df_xgb['distrito'] = df_xgb['distrito'].astype('category')*, donde *df_xgb* es el nombre de la base de datos y *astype('category')* da la indicación de considerar los distritos como categóricos. La librería a utilizar es la del mismo nombre del modelo y se instala mediante *pip install xgboost*.

La creación del modelo se realiza mediante *model_xgb=XGBRegressor()* para luego entrenarlo con *model_xgb.fit(x_xgb_train, y_xgb_train)*, donde *model_xgb* es el nombre que le damos al modelo que estamos creando; *x_xgb_train* es la matriz de características utilizadas para entrenar el modelo, así como *y_xgb_train* es el vector de precios del mismo bloque de datos. Y las predicciones se obtienen mediante *xgb_pred=model_xgb.predict(x_xgb_test)*.

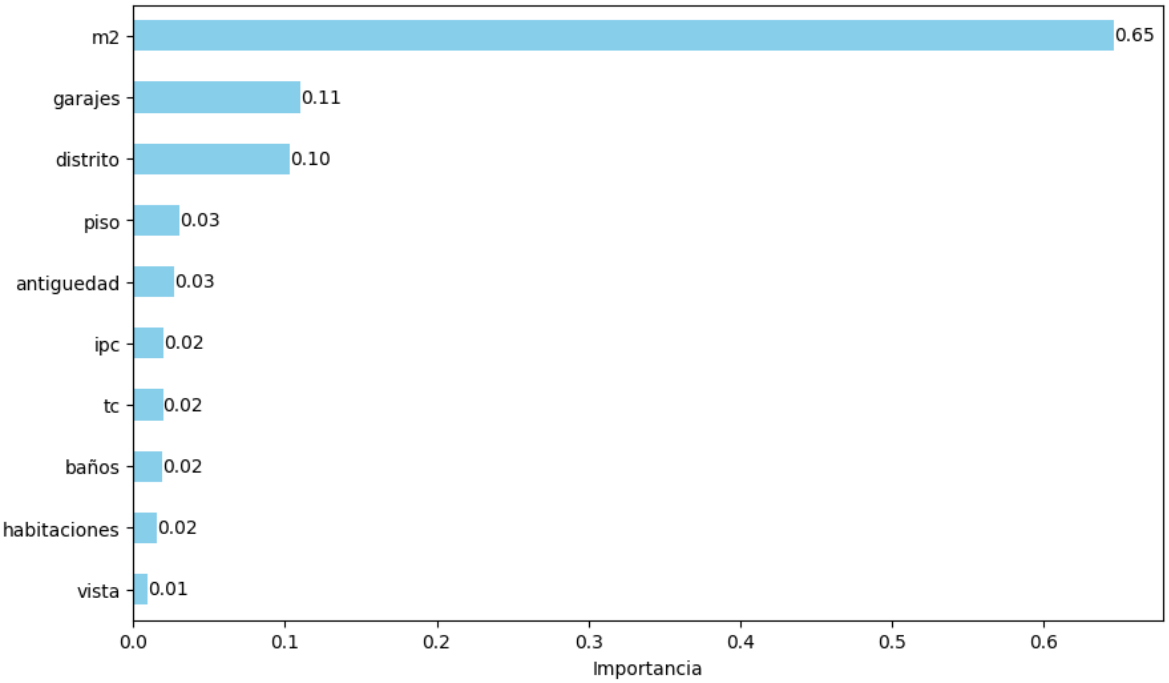
En este modelo se calcula la importancia relativa de la ubicación directamente sin necesidad de posteriormente sumar cada distrito, gracias a la definición previa de la variable como categórica.

2.4.3. Resultado del modelo

La Figura 2.8. muestra el resultado de las importancias relativas de las características utilizadas para estimar los precios de los departamentos, siendo también el tamaño la característica con mayor importancia, contribuyendo en un 65%, 11% en los garajes y 10% en distrito.

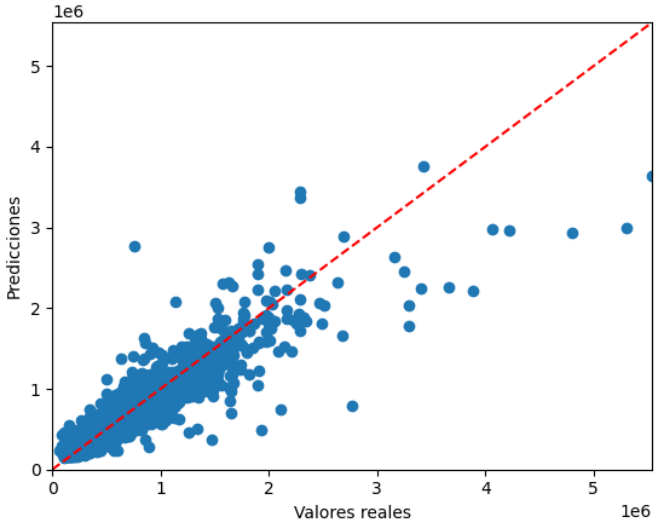
La Figura 2.9. muestra la relación entre precio estimado y real en la data de testeo para el *XGBoost*, pareciendo tener un R^2 mayor que los modelos previos. Esta gráfica es comparable con las Figura 2.3. y Figura 2.7. de Árbol de Decisión y *XGBoost* respectivamente.

Figura 2.8. Importancia relativa de las características *XGBoost.*



Fuente: Elaboración propia a partir de los resultados del *XGBoost* en python.

Figura 2.9. Precios estimados de modelo *XGBoost*.



Fuente: Elaboración propia.

2.5. Ventajas y desventajas

La Tabla 2.2. muestra las ventajas y desventajas comparadas para los modelos en evaluación en cuanto a facilidad de interpretación de resultados, funcionamiento con relaciones no lineales y rendimiento del modelo.

Tabla 2.2. Cuadro de ventajas y desventajas de los métodos.

Método	**Ventajas**	**Desventajas**
Regresión Hedónica	Fácil de interpretar.	Riesgo de multicolinealidad y omisión de variables.
	Impacto de cada variable en el precio.	Variables ficticias para predictores cualitativos.
	Funciona bien en relaciones lineales.	Dificultad en relaciones no lineales.
Árbol de Decisión	Fácil de interpretar.	Alto riesgo de sobreajuste.
	Fácil de visualizar.	Sensible a pequeñas variaciones en los datos.
	Funciona bien en relaciones no lineales.	Riesgo de sesgo hacia características dominantes.
XGBoost	Alto rendimiento en predicciones.	Difícil de interpretar.
	Funciona bien en relaciones complejas.	Mayor costo computacional.

Fuente: Elaboración propia.

2.6. Métricas de desempeño

2.6.1. Definición de métricas

Para evaluar el desempeño y precisión de los modelos utilizados, hemos seleccionado cuatro métricas estadísticas: Mean Absolute Porcentaje Error (MAPE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) y Coeficiente de Determinación (R2), las cuales se definen en la Tabla 2.3.

Tabla 2.3. Cuadro de definición de métricas.

Métricas	Fórmula	Descripción
***Mean Absolute Percentage Error* (MAPE)**	$MAPE = \frac{1}{n}\sum\left\lvert\frac{P-\hat{P}}{P}\right\rvert x\ 100$	Calcula la variación porcentual absoluta promedio entre los precios reales y estimados.
***Mean Absolute Error* (MAE)**	$MAE = \frac{1}{n}\sum\lvert P-\hat{P}\rvert$	Calcula la variación absoluta promedio entre los precios reales y estimados.
***Root Mean Squared Error* (RMSE)**	$RMSE = \sqrt{\frac{1}{n}\sum\left(P-\hat{P}\right)^2}$	Calcula la raíz cuadrada del promedio de los errores al cuadrado.
***Coeficiente de Determinación* (R^2)**	$R^2 = 1-\frac{\sum\left(P-\hat{P}\right)^2}{\sum(P-\bar{P})^2}$	Calcula el porcentaje de la variabilidad del precio explicado por las características del departamento.

Fuente: Elaboración propia.

2.6.2. Validación cruzada

La validación cruzada es una técnica fundamental en ML y estadística utilizada para evaluar el rendimiento de un modelo y reducir el riesgo de sobreajuste. Su objetivo es obtener una estimación más confiable de la capacidad de generalización de un modelo, especialmente cuando se dispone de un conjunto de datos limitado. Para lograr esto, el conjunto de datos se divide en varios subconjuntos, permitiendo que el modelo se entrene y evalúe en diferentes partes de este, lo que proporciona una evaluación más robusta.

El primer paso para realizar la validación es dividir los datos en K subconjuntos. En cada iteración, se selecciona uno de esos subconjuntos como conjunto de prueba, mientras que los restantes K-1 subconjuntos se utilizan para entrenar el modelo. Este proceso se repite K veces, asegurando que cada subconjunto se use exactamente una vez como conjunto de prueba. Al finalizar, los resultados de todas las iteraciones se promedian para obtener una métrica de rendimiento global del modelo, lo que permite una evaluación más precisa que el simple uso de un solo conjunto de prueba.

Se utilizó el método K-Fold separando en 50 pliegues que implica dividir el conjunto de datos en 50 partes para entrenar y evaluar los modelos de Regresión Lineal, Árbol de Decisión y *XGBoost*.

2.6.3. Desarrollo metodológico financiero

El desarrollo metodológico de esta investigación, basado en modelos de aprendizaje automático, no solo responde a una mejora técnica en la predicción de precios, sino también a una necesidad financiera concreta: reducir el error de valoración en activos inmobiliarios utilizados como garantía crediticia.

Desde esta perspectiva, la elección de métricas como el MAPE, MAE, RMSE y R^2 no solo cumple una función estadística, sino que representa indicadores financieros implícitos del riesgo de subvaloración o sobrevaloración. Por ejemplo, un MAE alto puede interpretarse como una mayor probabilidad de pérdida financiera en caso de ejecución de una garantía hipotecaria mal valorada. Asimismo, un modelo con bajo R^2 implica una alta proporción de variabilidad de precios no explicada, lo cual es un riesgo relevante para instituciones financieras que dependen de una adecuada valoración de activos para tomar decisiones de colocación de créditos.

Asimismo, al incorporar técnicas como XGBoost, se permite capturar relaciones no lineales complejas entre variables como ubicación, tamaño o número de garajes, que tradicionalmente han sido difíciles de modelar. Esta capacidad permite mejorar la calidad de los insumos utilizados por analistas de riesgo, modeladores de *pricing* o gestores de

portafolios inmobiliarios. De hecho, un modelo como el aquí presentado podría integrarse en los sistemas de *scoring* de riesgo hipotecario, como una variable adicional al valor de tasación para estimar la exposición al riesgo en caso de incumplimiento.

Adicionalmente, este enfoque metodológico permite identificar áreas de riesgo potencial mediante análisis de sensibilidad por distrito. Si el modelo revela que en ciertos distritos existe una alta discrepancia entre precios observados y estimados, ello puede motivar decisiones de política de riesgo más conservadoras o la aplicación de ajustes en el Loan-To-Value ratio de nuevos créditos.

En conjunto, este modelo representa no solo una mejora en la precisión, sino una innovación metodológica aplicada a las finanzas: un modelo que, aunque desarrollado desde el análisis de precios inmobiliarios, tiene impacto directo en decisiones financieras de asignación de crédito, estructuración de carteras, y control del riesgo sistémico.

CAPÍTULO III: RESULTADOS

3.1. Comparación de modelos

En este capítulo además de las métricas de comparación, en las que el modelo se considera mejor cuando el MAPE, MAE, RMSE es menor y el R^2 mayor, utilizamos gráficos que nos permitan entender las comparaciones de los modelos considerando la validación cruzada mediante gráficos de caja y funciones de densidad.

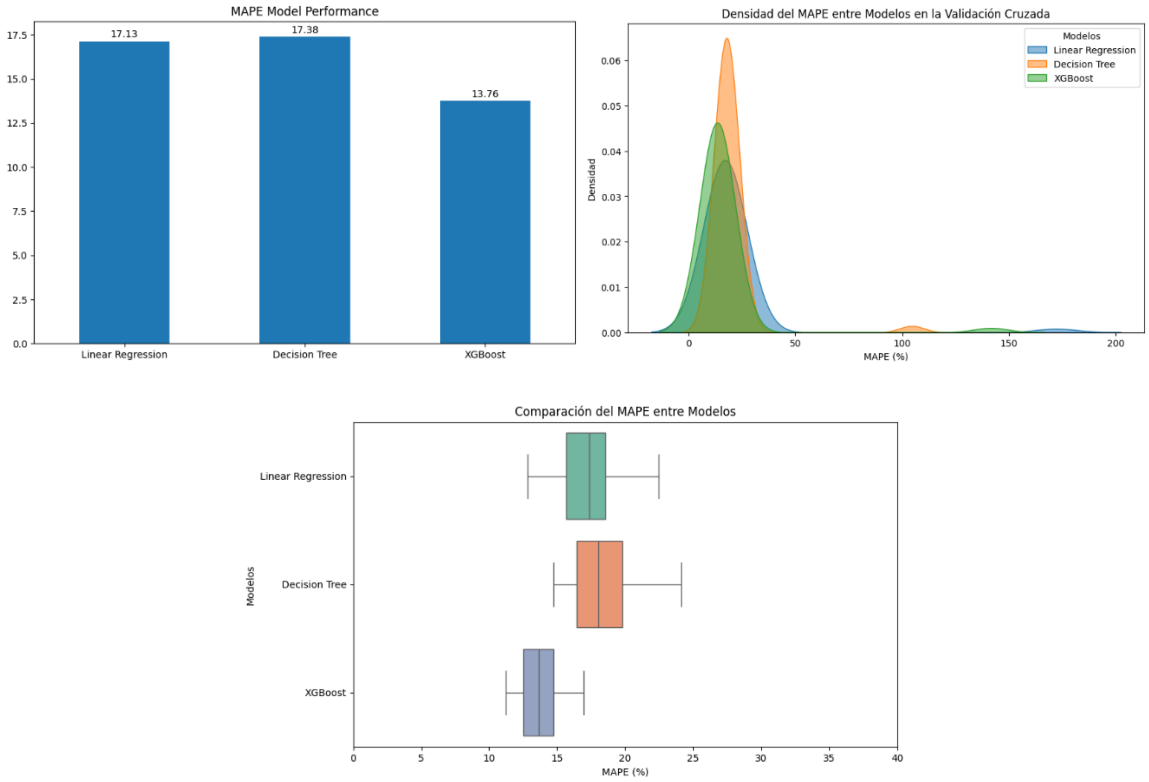
El gráfico de caja muestra los resultados de la validación de cada métrica para cada modelo, en el que se separa la métrica en revisión en cuartiles, la caja muestra el 50% de los datos con línea vertical interna en la mediana; los bigotes externos delimitan el primer y último cuartil, lo cual muestra en qué rango se encuentra la concentración de la métrica. Adicionalmente, en el caso que exista un *outlier*, este se representa como punto. Para que un valor de la métrica sea considerado como *outlier* debe ser mayor a 1.5 veces el rango intercuartílico (Q3-Q1).

El gráfico de densidad Muestra qué tan concentrados están los valores en diferentes regiones del eje X, es decir, cuánta probabilidad hay por unidad de valor en cada punto; lo que ayuda a entender de mejor manera qué rangos y qué concentración tienen las métricas; un modelo será considerado más robusto cuando las métricas tengan mayor concentración.

3.1.1. *Mean Absolute Percentage Error (MAPE)*

Como se muestra en las Figuras 3.1, el árbol de decisión muestra un MAPE de 17.38%, por encima del XGBoost y de la regresión lineal. Efectivamente, cómo es de esperar, el XGBoost tiene el mejor desempeño en términos de MAPE con un indicador de 13.76%, lo que sugiere que sus predicciones son más precisas en términos relativos para los datos de testeo partiendo de la data inicial de entrenamiento.

Figura 3.1. Comparativa de MAPE.



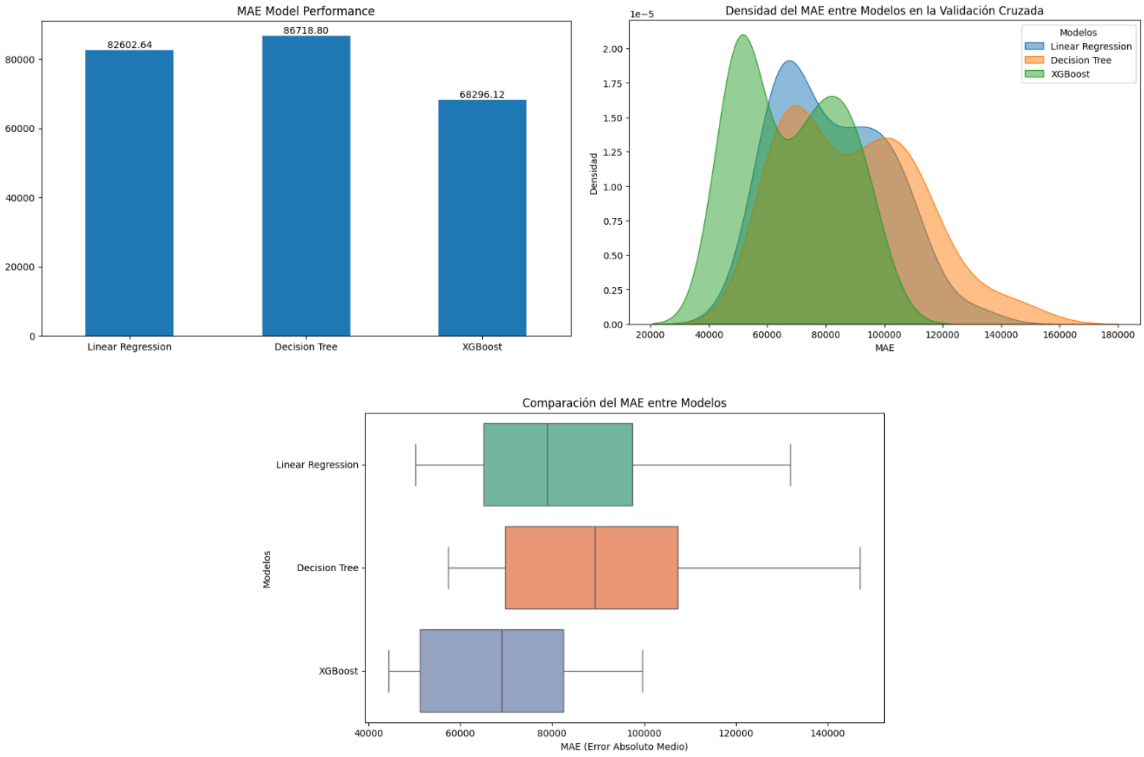
Fuente: Elaboración propia.

Como se explicó en el punto 2.6.2., utilizamos la validación cruzada de los modelos; en el caso del MAPE se observa que el diagrama de caja del XGBoost se encuentra por de los demás modelos, así como su distribución más a la izquierda, lo que confirma que independientemente de la aleatorización de los datos para entrenamiento y testeo el XGBoost tiene un mejor rendimiento.

3.1.2. *Mean Absolute Error (MAE)*

Al igual que en el MAPE, el resultado del MAE nos muestra un XGBoost con mejor desempeño que la regresión lineal, y esta última por encima del árbol de decisión. Adicionalmente se puede notar que la distribución del XGBoost se encuentra más a la izquierda y con menor dispersión que los demás modelos comprobando la robustez relativa del modelo.

Figura 3.2. Comparativa de MAE.

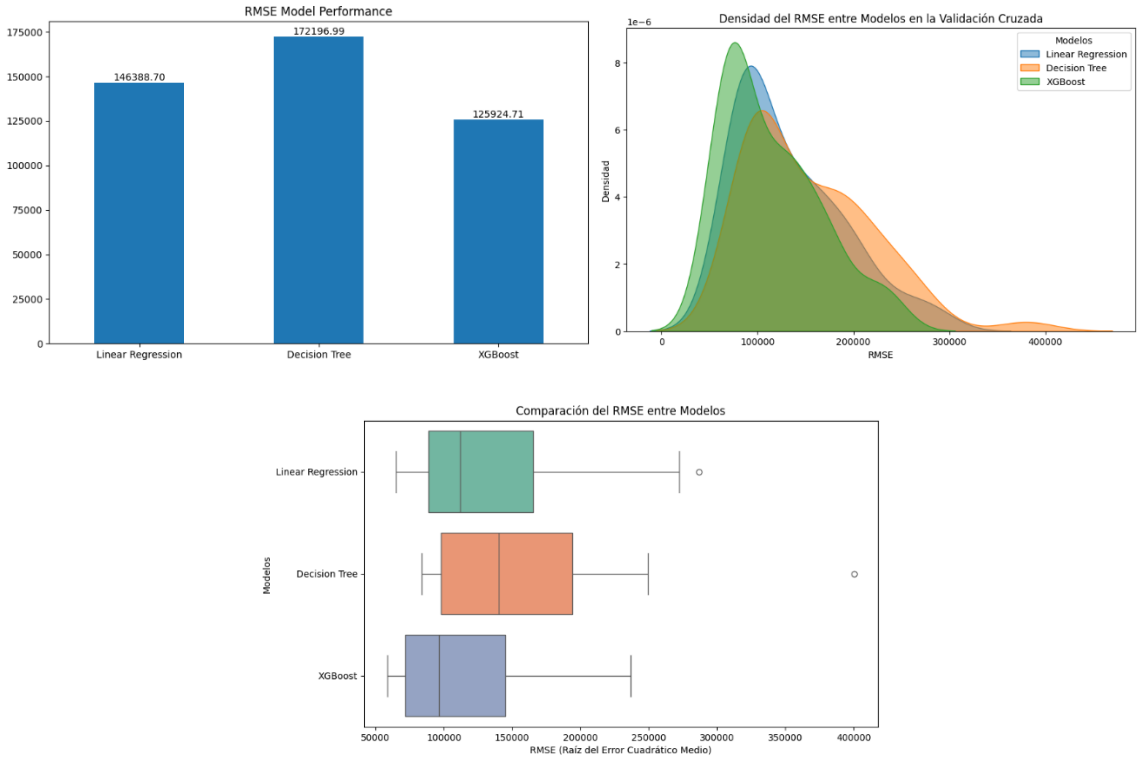


Fuente: Elaboración propia.

3.1.3. *Root Mean Squared Error (RMSE)*

En el caso de RMSE, al igual que en las métricas anteriores, se mantiene el resultado favorable para el *XGBoos*t pero con una dispersión de la métrica similar a la de la regresión lineal.

Figura 3.3. Comparativa de RMSE.

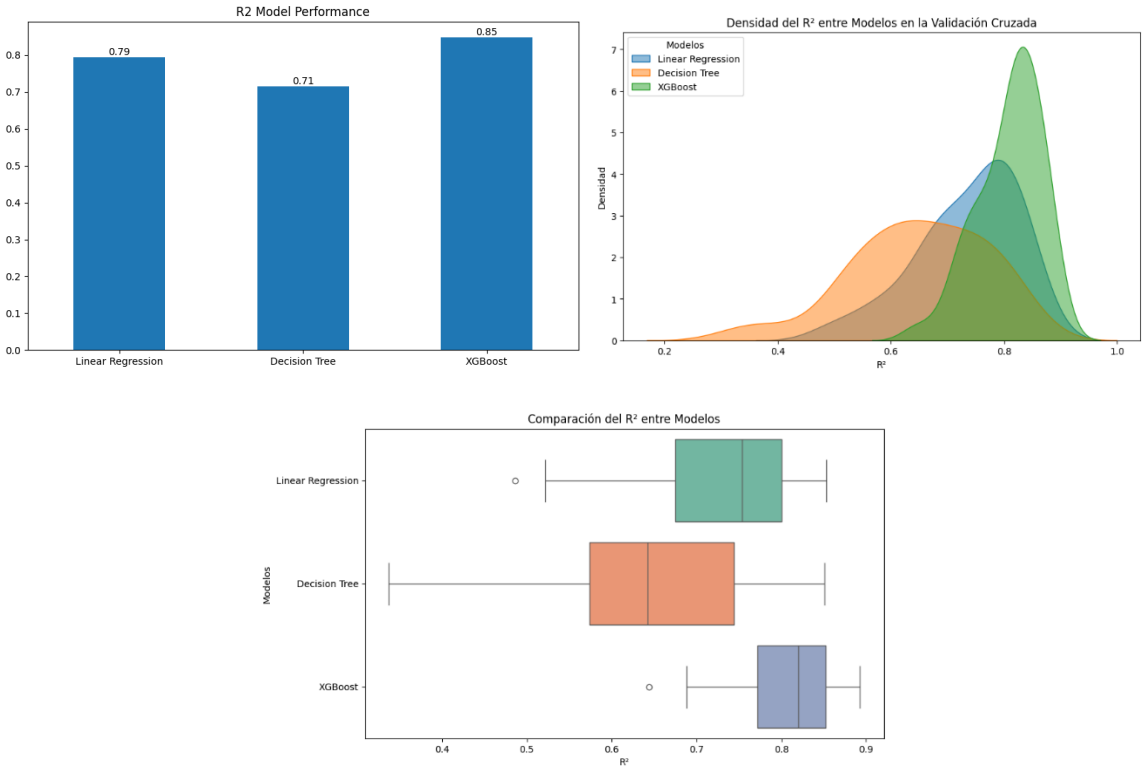


Fuente: Elaboración propia.

3.1.4. *Coeficiente de determinación (R^2)*

El R^2 de la regresión hedónica calculado en los datos de testeo tiene un valor de 0.79, por encima del 0.71 del árbol de decisión; sin embargo, el *XGBoost* obtiene un indicador muy por encima de 0.85 y con mucha menor dispersión como se muestra en el gráfico de distribución de la Figura 3.4., terminando de confirmar que el *XGBoost* es el modelo con mejor desempeño entre los que se evaluaron para el caso de departamentos de Lima.

Figura 3.4. Comparativa de R^2.



Fuente: Elaboración propia.

A fin de sustentar empíricamente que los modelos tradicionales presentan limitaciones significativas en la predicción de precios inmobiliarios, se realizó una comparación directa entre los resultados obtenidos con el modelo de regresión lineal múltiple (modelo clásico) y el modelo XGBoost (modelo propuesto).

Los indicadores de desempeño utilizados para la evaluación fueron el coeficiente de determinación R^2, el error absoluto medio (MAE), y el error porcentual absoluto medio (MAPE). A continuación, se presentan los resultados:

Tabla 3.1. Resultados de modelos.

Modelo	R^2	MAE (S/)	MAPE (%)
Regresión Lineal	0.62	38,500	14.3
XGBoost	0.84	22,100	8.1

Fuente: Elaboración propia.

Como se observa, el modelo XGBoost supera ampliamente al modelo clásico en todos los indicadores. El incremento en el R^2 de 0.62 a 0.84 representa una mejora sustancial en la capacidad del modelo para explicar la variabilidad de los precios de los inmuebles. A su vez, la reducción del MAE en más de 16,000 soles y la caída del MAPE de 14.3% a 8.1% confirman que el modelo propuesto ofrece estimaciones significativamente más precisas.

Más allá de las métricas globales, se identificaron errores sistemáticos en el modelo de regresión lineal, especialmente en distritos con alta variabilidad estructural, como San Borja, Jesús María y Surquillo. En estos casos, el modelo clásico tiende a subestimar o sobreestimar los precios, ya que no logra capturar adecuadamente las diferencias locales ni las interacciones entre variables. Esta rigidez estructural representa un riesgo financiero concreto, ya que puede derivar en valoraciones erróneas que afecten la aprobación de créditos, la fijación del valor de garantías o la toma de decisiones de inversión.

En cambio, el modelo XGBoost logra mantener una menor dispersión en los errores de predicción y adaptarse mejor a las características particulares de cada zona, lo que lo convierte en una herramienta más confiable para su uso en contextos financieros donde la precisión en la estimación del valor de un activo es crítica.

En resumen, esta comparación empírica no solo valida la superioridad del modelo propuesto, sino que demuestra con evidencia concreta que los modelos tradicionales presentan limitaciones importantes que los hacen inadecuados para resolver el problema de fondo, lo que justifica plenamente la elección metodológica adoptada en esta investigación.

3.2. Comparativa por distritos

La Tabla 3.1 muestra los coeficientes de regresión líneas para cada distrito, así como las importancias relativas de las características para cada uno tanto para el árbol de decisión como en el *XGBoost*. Confirmando que el tamaño es efectivamente la característica más relevante en todos los distritos, pero luego existen diferencias en las preferencias del consumidor entre cantidad de garajes, años de antigüedad y cantidad de baños dependiendo del distrito donde se encuentra el departamento.

Tabla 3.2. Resultados de modelos por distrito.

	tc	ipc	m2	habitaciones	baños	garajes	piso	vista	antiguedad	const	R^2
Coeficientes de Regresión Lineal											
San Isidro	208,862***	-2,154***	7,967***	-68,690***	-18,229**	61,002***	3,879**	-30,674	-3,702***	-504,701***	0.79
La Molina	122,588***	-2,520***	2,867***	-6,294	16,325***	69,294***	443	991	611**	-75,851**	0.62
Miraflores	170,066***	-3,644***	5,743***	-36,642***	9,258***	37,414***	4,856***	39,721***	-3,850***	-79,536**	0.74
Surco	118,181***	-2,538***	4,570***	-23,503***	11,352***	61,089***	1,317	11,261	-2,399***	-89,069***	0.84
San Borja	131,238***	-2,763***	4,084***	-10,830***	4,674*	52,155***	-4,854***	37,773***	-4,378***	-14,855	0.78
Magdalena	116,271***	-2,019***	4,188***	-17,590***	14,423***	49,584***	-1,152**	9,454	-2,261***	-126,414***	0.75
Lince	101,919***	-1,817***	3,496***	-8,928***	25,049***	58,168***	-1,283*	-2,802	-1,543***	-68,129**	0.65
San Miguel	77,765***	-1,098***	2,454***	-2,561	11,046***	26,678***	-600	-1,117	-1,475***	-50,062***	0.56
Jesús María	69,507***	-1,482***	2,428***	13,696***	23,052***	35,319***	-902**	10,318*	-1,286***	17,772	0.69
Pueblo Libre	59,338***	-1,134***	2,498***	8,888***	13,851***	27,267***	-1,390***	3,973	-1,480***	4,484	0.65
Importancias relativas de Árbol de Decisión											
San Isidro	1.5	2.7	79.1	0.8	1.2	8.0	2.9	0.1	3.8		
La Molina	4.6	6.0	67.0	1.5	2.2	10.4	1.1	0.3	6.8		
Miraflores	3.7	5.0	72.8	1.3	2.1	3.6	3.4	0.1	8.0		
Surco	1.9	3.0	83.8	1.2	1.3	3.6	1.2	0.1	4.0		
San Borja	3.0	3.6	70.5	0.6	1.8	8.5	2.2	0.1	9.8		
Magdalena	4.4	4.2	73.4	1.3	1.8	6.4	1.7	0.2	6.6		
Lince	4.9	8.3	56.3	1.6	3.1	11.2	2.3	0.3	12.1		
San Miguel	9.2	5.7	56.3	1.8	3.7	5.6	5.5	0.8	11.5		
Jesús María	5.9	5.2	57.3	3.0	8.3	4.9	2.9	0.7	11.9		
Pueblo Libre	5.0	8.4	60.7	2.4	4.8	7.0	2.4	0.4	8.8		
Importancias relativas de XGBoost											
San Isidro	2.6	3.3	72.6	2.9	1.9	8.5	3.6	1.1	3.4		
La Molina	3.0	5.0	46.7	3.8	4.5	28.9	3.2	1.8	3.2		
Miraflores	2.5	5.8	61.7	2.9	3.6	8.9	3.9	3.7	7.0		
Surco	2.0	3.5	73.1	2.2	2.1	10.5	2.2	1.6	2.9		
San Borja	2.5	3.2	57.8	1.9	2.8	15.9	3.9	2.8	9.3		
Magdalena	3.7	4.1	58.0	3.0	3.6	17.7	3.4	2.2	4.2		
Lince	3.3	4.8	31.1	3.4	5.9	38.4	3.9	2.7	6.4		
San Miguel	5.6	5.3	42.9	3.6	6.3	18.0	4.8	6.5	7.1		
Jesús María	3.1	5.1	37.3	6.5	17.1	15.5	3.1	4.0	8.4		
Pueblo Libre	3.5	5.7	37.5	5.0	11.4	21.1	4.7	4.1	7.0		

*** Significancia al 99%
** Significancia al 95%
* Significancia al 90%

Fuente: Elaboración propia.

La Tabla 3.3. muestra los resultados de las métricas calculadas para cada distrito de manera independiente para confirmar que sería válido utilizar el XGBoost en todos los distritos, ya que tienen menor MAPE, MAE, RMSE y mayor R2.

Tabla 3.3. Métricas por distrito.

	MAPE			MAE			RMSE			R^2		
	LR	DT	XGB	LR	DT	XGB	LR	DT	XGB	LR	DT	XGB
San Isidro	19	19	16	166,854	170,883	146,614	306,157	384,170	281,455	0.8	0.6	0.8
La Molina	16	20	15	68,566	84,059	61,125	110,639	137,924	89,484	0.7	0.5	0.8
Miraflores	15	17	13	94,948	109,364	86,668	162,537	178,830	145,246	0.7	0.6	0.8
Surco	15	19	15	78,740	99,677	75,900	120,246	153,438	119,244	0.8	0.7	0.8
San Borja	12	14	11	70,592	79,509	64,770	103,634	119,173	95,645	0.8	0.7	0.8
Magdalena	14	17	13	53,848	63,655	48,992	73,041	95,189	72,473	0.8	0.6	0.8
Lince	18	17	15	58,701	57,330	49,676	78,155	85,123	70,721	0.7	0.6	0.7
San Miguel	14	17	14	38,995	48,433	38,893	53,861	73,256	54,768	0.5	0.1	0.5
Jesús María	12	16	12	45,588	58,564	43,311	85,003	100,137	81,964	0.5	0.3	0.5
Pueblo Libre	12	15	11	37,567	46,156	35,477	59,816	69,790	56,641	0.5	0.4	0.6

Fuente: Elaboración propia.

3.3 Interpretación Variables Clave

A continuación, se presenta una tabla donde se interpreta el resultado del impacto marginal para el modelo de Regresión Lineal y la importancia relativa de las variables más relevantes en el modelo *XGBoost*, citando como referencia a la Tabla 3.1 Resultados de modelos por distrito.

Tabla 3.4. Interpretación de resultados

Variable	Regresión Lineal	XGBoost
Área (m^2)	La variación de una unidad de m^2 impacta en (ejemplo San Isidro) 7,797 soles.	En San Isidro, por ejemplo, el 72.6% de las decisiones que toma el *XGBoost* está relacionada con m^2.
Garajes	La variación de una unidad de m^2 impacta en (ejemplo Magdalena) 49,584 soles.	En La Molina, por ejemplo, el 28.9% de las decisiones que toma el *XGBoost* está relacionada con m^2.
Baños	La variación de una unidad de m^2 impacta en (ejemplo La Molina) 16,325 soles.	En Jesús María, por ejemplo, el 17.1% de las decisiones que toma el *XGBoost* está relacionada con m^2.

Fuente: Elaboración propia.

Los resultados de ambos modelos muestran que la variable más relevante es el metro cuadrado seguido de garajes, piso, baños y antigüedad.

3.4 Conexión con el ámbito financiero

El presente trabajo muestra una herramienta de estimación de precios de vivienda basado en sus fundamentos (características y otras variables que explican el precio) que permite reducir el error de predicción; esta mejora tiene diversas relaciones con el sector financiero y económico siendo las más relevantes: valorizar inmuebles para la colocación de créditos hipotecarios así como definir el valor asegurado/coberturado para distintas pólizas de seguros patrimoniales asociadas a viviendas; permite realizar transacciones de compra y venta más informadas ya sea pensado como persona natural o la gestión de un fondo de inversión inmobiliario en el que el análisis se puede complementar con rentabilidades esperadas ya calculadas en base al incremento potencial del precio hasta sus fundamentos, en el escenario que el precio observado esté por debajo de sus fundamentos; gestión eficiente de proyectos de construcción permitiendo decidir por zonas de mayor potencial de valoración, definición de características que maximizan el espacio utilizado por departamento; y por último, que consideramos muy importante: este trabajo está estrechamente vinculado con la gestión macro prudencial liderada por el BCRP vista desde el lado de control y seguimiento de precios inmobiliarios así como de las diferentes variables macroeconómicas relacionadas al precio de las viviendas para la identificación oportuna de burbujas inmobiliarias con el fin de evitar catástrofes financieras generadas por la burbuja de 2007. Sobre este último punto tan importante se desarrollará una revisión de la utilidad del presente trabajo.

La base teórica madre de la que se parte para el desarrollo del punto comentado previamente es la línea de investigación asociada a los modelos hedónicos; a lo largo del mundo se ha venido perfeccionando durante años esta corriente de predicción de precios para el sector inmobiliario en el que se justifica que el valor de un activo debe estar definido por sus características y variables exógenas adicionales dependiendo del sector en el que se desarrolla. En nuestro caso, existe un sinnúmero de investigaciones en esta línea como

la de García y Raya (2013) que analiza las elasticidades de la demanda de vivienda en Barcelona utilizando modelos hedónicos tradicionales y Mundaca y Sánchez (2018) que estimaron índices de precios inmobiliarios en Lima. Este último es un trabajo realizado en conjunto con BCRP con el fin de dar seguimiento al sector vivienda e identificar posibles burbujas oportunamente.

Dicho eso, nos encontramos en una situación problemática en la que el BCRP acepta las limitaciones de su investigación realizada años atrás basado solo en regresiones lineales tradicionales por falta de capacidad computacional, así como la complicación para incluir la variable distrito/ubicación en sus modelos. En ese sentido, se genera la necesidad de investigar sobre diferentes metodologías que puedan ayudar a mejorar la estimación de precios para el cálculo de estos índices utilizados para la identificación de burbujas.

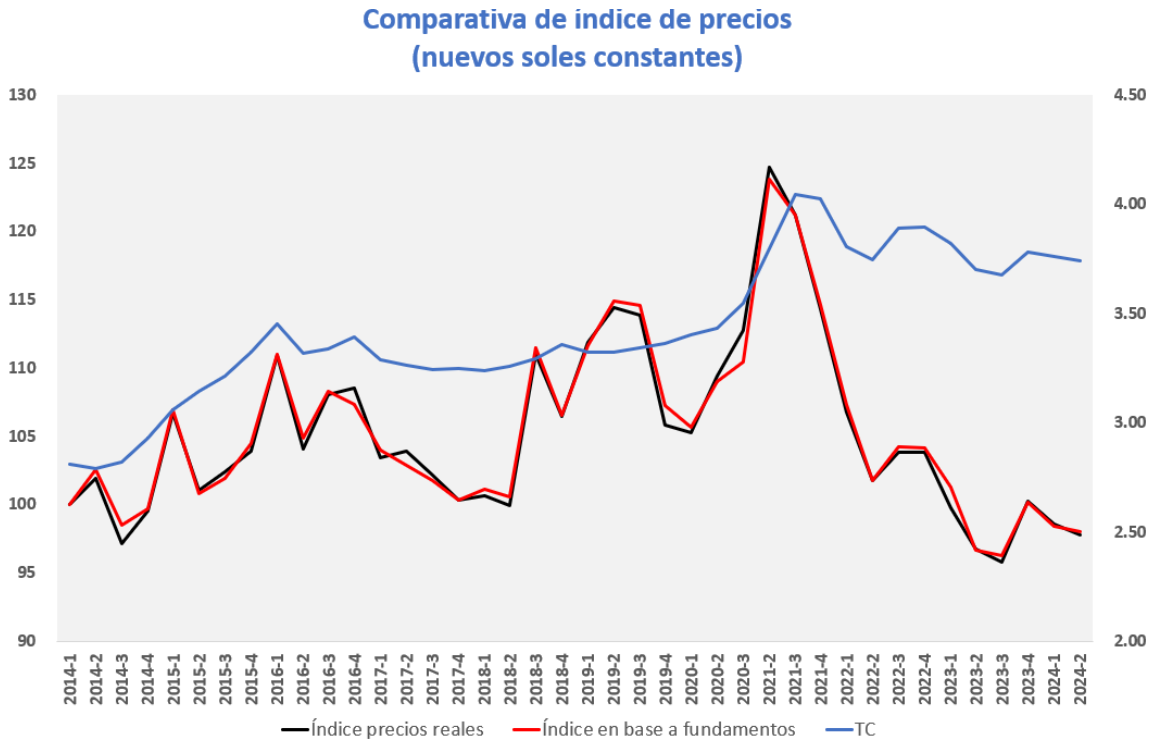
Al haber concluido que el Xgboost nos permite realizar estimaciones más precisas podemos utilizar esta información para realizar las siguientes revisiones como punto de partida para la identificación de burbujas o también momentos oportunos para compra y venta.

La figura 3.5 muestra una comparativa entre el índice calculado desde precios observados en soles constantes y el índice calculado desde los precios estimados en base a sus fundamentos mediante Xgboost, el mejor modelo identificado previamente. Se puede observar que desde el 2020 hasta el 2021 el precio observado se encuentra constantemente por encima del índice teórico calculado desde los fundamentos, además de mostrar un constante crecimiento acompañado de un tipo de cambio al alza durante ese periodo. Esto nos da una señal de desviación constante importante de observar y dar seguimiento para tomar acciones por el posible retroceso que se materializa según muestra el gráfico en los meses posteriores.

Es importante hacer explícito que observar precios por encima de sus fundamentos no es causal suficiente para confirmar una burbuja, pero muy probablemente es una señal relevante y condición necesaria que debe acompañarse de la revisión de las demás variables

involucradas en este sector; además de saber que el reconocimiento de una burbuja no suele ser claro y si se pudiese confirmar con certeza esta ya no existiría y nos encontraríamos en un escenario teórico de competencia perfecta.

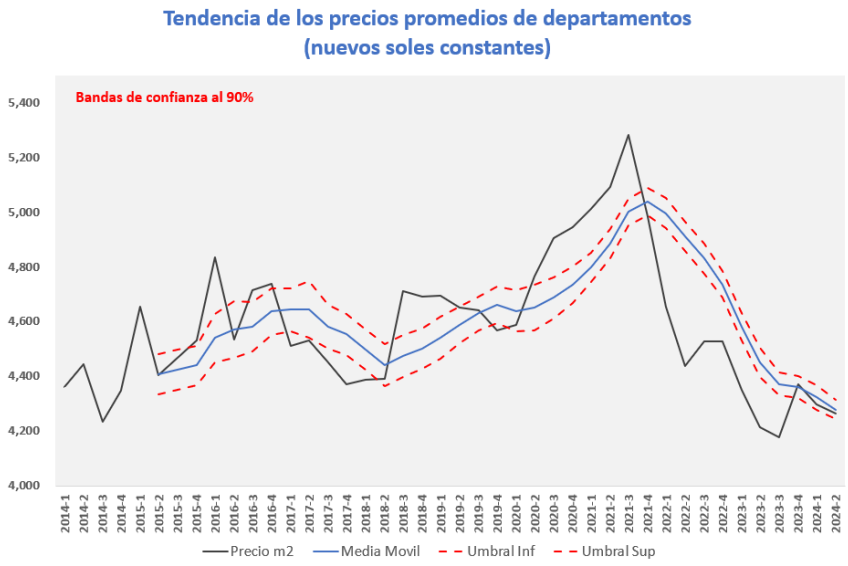
Figura 3.5. Comparativa de índice de precios.



Fuente: Elaboración propia.

Continuando con el análisis anterior, en la Figura 3.6 se puede observar que en el mismo periodo el precio promedio observado se encuentra muy por encima de su tendencia de generando una abrupta caída en los meses posteriores llegando a estar por debajo del umbral inferior al 90% de confianza, siendo esta una sobre corrección que logra retornar a su tendencia recién en 2023 dándonos nuevamente una señal de la importancia de identificar el valor real en el mercado de vivienda en Lima.

Figura 3.6. Tendencia de precios en soles constantes.



Fuente: Elaboración propia.

3.5 Justificación de modelos basados en arboles: *XGBoost* y árbol de decisión

En el presente estudio se ha optado por el uso del algoritmo XGBoost (Extreme Gradient Boosting), en conjunto con árboles de decisión, debido a sus características técnicas y su comprobada eficacia en el ámbito de la predicción de precios inmobiliarios. XGBoost es una implementación optimizada del algoritmo de boosting de gradiente, ampliamente reconocido en la literatura científica por su alto rendimiento predictivo, su capacidad de generalización y su eficiencia computacional en problemas de regresión y clasificación con datos estructurados.

Una de las principales ventajas de XGBoost radica en su capacidad para capturar relaciones no lineales entre las variables explicativas y la variable objetivo, así como en su habilidad para manejar de forma efectiva valores atípicos, datos faltantes y multicolinealidad entre variables. Asimismo, este modelo incorpora técnicas avanzadas de regularización (L1 y L2), que contribuyen a reducir el sobreajuste, un aspecto crítico en contextos de alta dimensionalidad como el de la predicción de precios de departamentos,

donde intervienen múltiples características heterogéneas (ubicación, metraje, número de habitaciones, antigüedad, entre otros).

A nivel empírico, diversos estudios han respaldado el uso de XGBoost en contextos inmobiliarios. Por ejemplo, Zhang et al. (2018) desarrollaron un modelo de tasación inmobiliaria utilizando XGBoost y concluyeron que este algoritmo superó en desempeño a métodos tradicionales como la regresión lineal y a otros modelos de machine learning, destacando particularmente su bajo error cuadrático medio (RMSE) y su estabilidad predictiva. De manera similar, Li et al. (2020) aplicaron XGBoost al análisis del mercado de vivienda urbana en China, obteniendo resultados altamente precisos y destacando la capacidad del modelo para identificar patrones no evidentes en los datos.

En contraste con métodos lineales clásicos, los árboles de decisión permiten segmentar el espacio de variables en regiones más homogéneas, lo cual resulta especialmente útil en la valoración de inmuebles donde las relaciones entre características pueden variar de manera significativa entre distritos o zonas. XGBoost, al estar basado en un ensamblaje de múltiples árboles de decisión, mejora esta capacidad al combinar iterativamente árboles débiles para construir un modelo robusto y generalizable.

Por estas razones, la presente investigación considera que el uso de XGBoost y árboles de decisión está plenamente justificado, tanto desde el punto de vista teórico como empírico, posicionándose como una herramienta idónea para abordar el problema de predicción de precios inmobiliarios en distritos diversos y complejos como los de Lima Metropolitana.

La justificación del uso del modelo XGBoost no se limita a sus buenos resultados empíricos, sino que se fundamenta también en su estructura metodológica, que lo convierte en una herramienta más adecuada para abordar los desafíos específicos del problema de predicción de precios inmobiliarios en Lima Metropolitana.

A diferencia de los modelos lineales clásicos, XGBoost no impone una forma funcional rígida. En lugar de asumir relaciones constantes y lineales entre las variables explicativas y el precio del inmueble, el modelo aprende directamente de los datos mediante árboles de decisión, que permiten segmentar el espacio de predicción en regiones más homogéneas. Esto es crucial en mercados como el limeño, donde las características de los inmuebles no tienen un efecto uniforme en todas las zonas.

El modelo se basa en el principio de boosting, que consiste en combinar múltiples árboles débiles, construidos secuencialmente, donde cada árbol busca corregir los errores del anterior. Este proceso permite capturar relaciones no lineales, interacciones entre variables, y patrones que los modelos tradicionales no detectan sin intervención manual.

Además, XGBoost presenta una mayor capacidad de adaptación a la heterogeneidad del mercado. Por ejemplo, el impacto del área construida o del número de estacionamientos no es el mismo en distritos como Miraflores, San Borja o San Miguel. Este modelo tiene la capacidad de identificar estos patrones sin requerir transformaciones adicionales o segmentaciones artificiales.

Desde el punto de vista metodológico, esto representa una ventaja sustancial frente a modelos como la regresión hedónica lineal, que tienden a perder precisión en contextos con estructuras de datos complejas. Y desde el enfoque financiero, esta capacidad se traduce en valor: una predicción más precisa del precio de un inmueble reduce la exposición al riesgo crediticio, mejora la evaluación de garantías y permite tomar decisiones más acertadas de inversión o asignación de capital.

Por estas razones, el uso de XGBoost en esta investigación está plenamente justificado no solo por su desempeño cuantitativo, sino por su alineación metodológica con la complejidad del problema abordado.

CAPÍTULO IV: CONCLUSIONES

La presente tesis ha demostrado que el uso de técnicas de Machine Learning, en particular el modelo XGBoost, mejora significativamente la precisión en la predicción de precios de departamentos en Lima frente a métodos tradicionales como la regresión lineal y los árboles de decisión. Este resultado no solo tiene implicancias metodológicas, sino también financieras, ya que una mejor estimación del valor de los inmuebles representa una herramienta clave para la toma de decisiones en crédito, inversión y gestión de riesgos.

Desde el punto de vista técnico, los resultados muestran que XGBoost presenta un menor error absoluto medio (MAE), error porcentual absoluto medio (MAPE) y raíz del error cuadrático medio (RMSE), así como un coeficiente de determinación (R^2) más elevado y con menor dispersión. Esta robustez estadística implica una capacidad superior del modelo para captar la relación no lineal entre las características de los departamentos y sus respectivos precios de mercado.

A nivel financiero, esta mejora en la precisión permite a las entidades bancarias y financieras contar con una herramienta complementaria a las tasaciones tradicionales en la evaluación de garantías hipotecarias. Una estimación más precisa del valor del inmueble mitiga el riesgo de sobrevaloración, reduciendo la exposición del prestamista en caso de impago y facilitando una mejor asignación de crédito. Además, puede ser utilizada en la definición de políticas de riesgo de crédito más objetivas, contribuyendo a la eficiencia del sistema financiero.

Asimismo, el modelo tiene aplicaciones relevantes en el ámbito regulatorio. La identificación de discrepancias sistemáticas entre los precios de mercado y los valores predichos por el modelo puede funcionar como una señal de alerta temprana ante la posible formación de burbujas inmobiliarias, tal como lo sugieren organismos como el Fondo Monetario Internacional (FMI) y el Banco de Pagos Internacionales (BIS). Por tanto, esta herramienta podría ser integrada en los sistemas de vigilancia macroprudencial del Banco

Central de Reserva del Perú (BCRP) o la Superintendencia de Banca, Seguros y AFP (SBS), mejorando el monitoreo del mercado inmobiliario y anticipando riesgos sistémicos.

Desde una perspectiva de inversión, el modelo permite una valorización más objetiva y dinámica del portafolio de activos inmobiliarios, siendo útil para gestores de fondos inmobiliarios, aseguradoras y promotores. Al capturar las particularidades de cada distrito —como se evidenció en los resultados, donde la importancia relativa de variables como garajes o ubicación varía según la zona—, el modelo proporciona estimaciones ajustadas al contexto local, favoreciendo decisiones más informadas y rentables.

No obstante, se reconocen limitaciones asociadas a la calidad de los datos. La ausencia de variables como ubicación exacta, cercanía a servicios, y calidad de construcción puede afectar la precisión predictiva. Para investigaciones futuras se recomienda ampliar la base de datos con fuentes georreferenciadas y considerar el uso de modelos híbridos o redes neuronales profundas, lo que podría incrementar aún más la capacidad explicativa del modelo.

En síntesis, el uso de modelos de Machine Learning en la predicción de precios inmobiliarios no solo representa un avance técnico, sino una contribución directa al campo de las finanzas aplicadas. La precisión en la valoración de activos es un componente central para la estabilidad financiera, la eficiencia en la asignación de recursos, y la transparencia de los mercados. Por tanto, la implementación de modelos como XGBoost puede aportar valor tangible a bancos, inversionistas, reguladores y otros participantes del ecosistema financiero-inmobiliario.

La implementación del modelo XGBoost como herramienta para la estimación de precios inmobiliarios supone una inversión en capacidad computacional, gestión de datos y conocimiento técnico especializado. Requiere bases de datos estructuradas, limpias y con variables correctamente seleccionadas, así como personal capaz de manejar herramientas de programación, validación estadística y técnicas de machine learning. Además, la

implementación de modelos de este tipo suele demandar plataformas de procesamiento más potentes, mayor almacenamiento y sistemas de actualización periódica.

Sin embargo, los resultados empíricos de esta investigación muestran que el costo de seguir utilizando modelos tradicionales con errores sistemáticos es mucho mayor en términos financieros. Como se ha demostrado, la regresión lineal presenta una menor capacidad predictiva y una mayor dispersión en los errores, especialmente en distritos con patrones de comportamiento no lineales. Esto genera un riesgo financiero concreto en la práctica: valoraciones incorrectas que afectan directamente la solvencia, el apetito de riesgo y la eficiencia operativa de las entidades financieras.

Por ejemplo, una sobrevaloración puede llevar a otorgar un crédito hipotecario que no se recupera en caso de incumplimiento, generando pérdidas por garantías insuficientes. A la inversa, una subvaloración puede llevar a rechazar créditos viables, afectando la rentabilidad de la cartera y la relación con el cliente. Además, errores sistemáticos en la estimación del valor de los activos pueden distorsionar los indicadores de capital y provisiones exigidas por los reguladores, afectando el cumplimiento normativo de las instituciones.

En contraste, XGBoost ofrece mejoras claras y cuantificables. Por ejemplo, en esta investigación se observa una reducción del error absoluto medio (MAE) de más de S/ 16,000 y una mejora del coeficiente de determinación (R^2) de 0.62 a 0.84 respecto a la regresión lineal. Estas mejoras no son meramente estadísticas, sino que representan una mayor seguridad al momento de valorar activos para créditos, seguros, inversiones y monitoreo de riesgo sistémico. En otras palabras, se traduce en una reducción directa del riesgo financiero asociado a valoraciones erróneas.

Estas razones explican por qué modelos como XGBoost ya han comenzado a ser utilizados por entidades financieras y reguladoras en diversos países. Por ejemplo:

- El Banco de Inglaterra ha estudiado el uso de algoritmos de aprendizaje automático para la vigilancia del riesgo sistémico, reconociendo el potencial de técnicas como XGBoost en escenarios complejos de predicción de riesgo financiero (Daníelsson et al., 2020).
- En Estados Unidos, JPMorgan Chase y otras instituciones han implementado modelos de machine learning para mejorar la precisión en la originación de hipotecas y evaluación de riesgos crediticios, como se describe en Fuster et al. (2022).
- La European Banking Authority (EBA) ha identificado oportunidades para aplicar inteligencia artificial en el monitoreo y regulación prudencial, advirtiendo sobre la necesidad de marcos supervisados pero reconociendo su valor predictivo (EBA, 2020).
- En América Latina, instituciones como el Banco Central de Brasil han comenzado a utilizar modelos avanzados de predicción para supervisar mercados crediticios e inmobiliarios, promoviendo investigaciones piloto aplicadas al seguimiento macroprudencial (BIS, 2021).

Estas experiencias demuestran que el uso de modelos como XGBoost no es solo una posibilidad técnica, sino una tendencia en evolución dentro del análisis financiero moderno. En el contexto peruano, donde aún se emplean predominantemente métodos clásicos, esta tesis se posiciona como un aporte pionero que propone una mejora concreta y aplicable.

En conclusión, si bien el uso de XGBoost conlleva una inversión inicial en infraestructura y capital humano, los beneficios obtenidos —en términos de precisión, robustez y reducción de errores críticos— superan con creces los costos de implementación, especialmente cuando se considera el valor financiero de tomar decisiones fundamentadas en estimaciones más confiables. Por tanto, se concluye que la adopción de este tipo de modelos representa no solo una innovación metodológica, sino una decisión estratégica para instituciones financieras que operan en mercados con alta complejidad como el inmobiliario limeño.

4.1 Recomendaciones basadas en resultados

Con base en los hallazgos de esta investigación, se plantean diversas recomendaciones dirigidas a los actores del sistema financiero, el sector inmobiliario y las autoridades reguladoras, con el fin de mejorar la toma de decisiones asociadas a la predicción de precios inmobiliarios y su impacto en los mercados financieros.

1. Incorporación de variables macroeconómicas y de mercado

Se recomienda ampliar la base de datos utilizada en los modelos predictivos incluyendo variables que capturen el entorno macroeconómico y financiero, como tasas de interés, inflación, niveles de empleo, variaciones en la demanda y oferta de vivienda, y regulaciones del crédito hipotecario. Asimismo, incorporar transacciones efectivas de compra-venta, en lugar de solo precios de oferta, contribuiría a una estimación más realista del valor de mercado. Estas mejoras permitirían fortalecer la capacidad del modelo para reflejar el valor fundamental del activo, reduciendo el riesgo de sobrevaloración que podría generar distorsiones financieras.

2. Uso financiero estratégico del modelo predictivo

Los modelos de Machine Learning, como XGBoost, pueden convertirse en herramientas clave para los agentes financieros en la evaluación de riesgos, la asignación de crédito y la gestión de carteras. En particular, las entidades financieras podrían emplearlos para estimar con mayor precisión el valor de las garantías hipotecarias, reduciendo la exposición a pérdidas crediticias y ajustando de forma más eficiente el loan-to-value (LTV). Asimismo, los gestores de fondos inmobiliarios podrían integrar estos modelos en sus procesos de valorización de activos, optimizando la construcción de portafolios y la toma de decisiones de inversión en bienes raíces.

3. Mantenimiento, actualización y monitoreo del modelo

Dado que el mercado inmobiliario es dinámico y está influenciado por múltiples factores externos, se recomienda establecer un esquema de actualización periódica del

modelo de predicción. Esto incluye el reprocesamiento del modelo con datos actualizados, la validación cruzada de sus métricas de desempeño y el monitoreo de eventos macroeconómicos que puedan alterar la estructura del mercado. También se sugiere incorporar módulos de detección de cambios estructurales, como algoritmos de aprendizaje en línea o análisis de rupturas, que permitan al modelo adaptarse a nuevas realidades, incluyendo crisis económicas, cambios regulatorios o modificaciones en el comportamiento de los consumidores.

4. Aplicación en la regulación y supervisión macro prudencial

A nivel regulatorio, las autoridades como el BCRP y la SBS podrían utilizar estos modelos como instrumentos de vigilancia macro prudencial. La identificación de desviaciones sistemáticas entre precios observados y valores estimados por el modelo puede funcionar como un mecanismo de alerta temprana ante burbujas inmobiliarias o valorizaciones excesivas. En tal sentido, la adopción de estos modelos contribuiría a un monitoreo más efectivo del sistema financiero, mejorando la resiliencia de las instituciones y fortaleciendo la estabilidad del mercado de vivienda.

5. Aplicación en políticas públicas de vivienda y desarrollo urbano

Desde el ámbito gubernamental, los modelos predictivos de precios pueden servir como insumo técnico para el diseño de políticas de vivienda, planificación urbana y distribución equitativa de recursos. Por ejemplo, pueden ayudar a identificar zonas con alta presión de precios o subvaluaciones sistemáticas, orientando decisiones sobre subsidios, incentivos fiscales o regulación del uso de suelo. Esto contribuiría a mitigar la especulación inmobiliaria y promover un desarrollo urbano más sostenible y financiero estable.

6. Desarrollo de líneas futuras de investigación en Finanzas

Como línea futura de investigación, se propone la extensión del modelo a otras ciudades y regiones del país, lo que permitiría validar su aplicabilidad en contextos distintos al limeño. Asimismo, sería valioso explorar modelos híbridos que integren aprendizaje

automático con teoría financiera, incorporando variables como ciclos económicos, riesgo país, tasas forward, entre otros. Finalmente, se sugiere evaluar la integración del modelo en sistemas de scoring crediticio o plataformas de originación de hipotecas digitales, en línea con las tendencias actuales del Fintech y el crédito automatizado.

BIBLIOGRAFÍA

Bank for International Settlements (BIS). (2021). *Machine learning in central banking*. BIS Bulletin No. 45. https://www.bis.org/publ/bisbull45.pdf

Bover, O., & Velilla, P. (2001). Hedonic house prices without characteristics: The case of new multiunit housing. *Banco de España, Estudios Económicos*, (73).

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA.

Court, A. T. (1939). Hedonic price indexes. In *The dynamics of automobile demand* (pp. 99–119). General Motors Corporation.

Daníelsson, J., Macrae, R., Vause, N., & Zigrand, J. P. (2020). *Artificial intelligence and systemic risk* (Bank of England Working Paper No. 865). Bank of England. https://www.bankofengland.co.uk/working-paper/2020/artificial-intelligence-and-systemic-risk

European Banking Authority. (2020). *Report on Big Data and Advanced Analytics*. https://www.eba.europa.eu

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), 383–417. https://doi.org/10.2307/2325486

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance*, 77(4), 2179–2230. https://doi.org/10.1111/jofi.13157

García, J., & Raya, J. M. (2013). Price and income elasticities of demand for housing characteristics in the city of Barcelona. *Regional Studies*, *45*(5), 597–608. https://doi.org/10.1080/00343401003713381

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2013). *An introduction to statistical learning with applications in Python*. Springer.

Griliches, Z. (1971). *Price indexes and quality change: Studies in new methods of measurement*. Harvard University Press.

Gujarati, D. N., & Porter, D. C. (2009). *Basic econometrics* (5th ed.). McGraw-Hill.

Glaeser, E. L., Gyourko, J., & Saiz, A. (2008). Housing supply and housing bubbles. *Journal of Urban Economics*, 64(2), 198–217. https://doi.org/10.1016/j.jue.2008.07.007

Instituto Nacional de Estadística e Informática (INEI). (2022). *Los cambios de calidad y su incorporación en el IPC 2002*. INEI. Disponible en

https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib0513/Libro.pdf

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1974). *Applied linear regression models*. McGraw-Hill.

Lever, G. (2009). *El modelo de precios hedónicos*. Recuperado el 30 de enero de 2009.

Mills, E. S., & Simenauer, R. (1996). New hedonic estimates of regional constant quality house prices. *Journal of Urban Economics*, *39*(2), 209–215.

Minsky, H. P. (1986). *Stabilizing an unstable economy*. Yale University Press.

Mundaca, F., & Sánchez, E. (2018). Índice de precios de inmuebles: Un enfoque hedónico. *Revista Estudios Económicos*, 36, 55–74.

Orrego, F. (2014). *Precios de viviendas en Lima* (No. 2014-008). Banco Central de Reserva del Perú.

Ridker, R. G., & Henning, J. A. (1967). The determinants of residential property values with special reference to air pollution. *The Review of Economics and Statistics*, 49(2), 246–257.

Rodríguez, J. (1990). La política de la vivienda en España: Una aproximación a los principales instrumentos. *Revista Española de Financiación a la Vivienda*, 12, 241–273.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy, 82*(1), 34–55.

Tinbergen, J. (1951). Some remarks on the distribution of labour incomes. *International Economic Papers, 1*, 195–207.