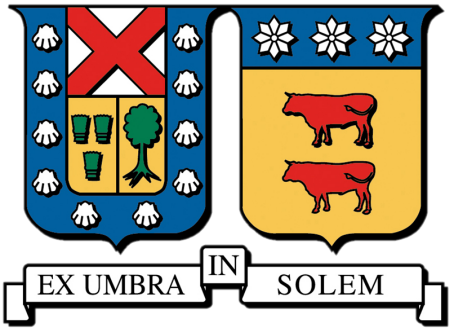

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INDUSTRIAS

DETERMINANTES DEL PRECIO DE LA VIVIENDA EN EL GRAN SANTIAGO MEDIANTE MODELOS ECONOMÉTRICOS Y DE APRENDIZAJE AUTOMÁTICO

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

AUTOR

MAURICIO ANDRÉS CISTERNAS URBINA

PROFESOR GUÍA

DRA. ANA MARÍA ELISA FARÍAS GORDON

PROFESOR CO-REFERENTE

DR. JAVIER SCAVIA DAL POZZO

OCTUBRE 2025

CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título ☐ Tesis de Postgrado

Título del trabajo: DETERMINANTES DEL PRECIO DE LA VIVIENDA EN EL GRAN SANTIAGO MEDIANTE MODELOS ECONOMÉTRICOS Y DE APRENDIZAJE AUTOMÁTICO

Nombre del candidato(a): Mauricio Andrés Cisternas Urbina

Carrera / Grado: Ingeniería Civil Industrial

Campus: Santiago, Vitacura **Departamento:** Industrias

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, María Elisa Farías Gordon, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☒ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por **(marcar una opción)**:

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

_____

_____

_____

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 21-10-2025 **Firma:** _____

Estudiante o Candidato(a):

Fecha: 21-10-2025 **Firma:** _____

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

Tabla de Contenidos

Lista de Tablas

Lista de Figuras

1 Agradecimientos

Este trabajo es dedicado en primer lugar a Dios, por la guía y el fortalecimiento en lo personal y profesional, tanto en mi vida como a lo largo de todo este proceso académico.

Agradezco incondicionalmente a mi familia, mis Padres Ángela y Andrés, y mis hermanos Pablo, Agustín y Renato, quienes tuvieron siempre para conmigo paciencia y cariño a lo largo de estos años. Su amor y confianza plena fue decisiva para sostener este proceso hasta el final.

A mis compañeros y amigos, por su amistad, compañía y ayuda práctica en momentos claves durante la carrera. Gracias por el ánimo y las risas que hicieron más llevadero este camino.

Finalmente a la Universidad Técnica Federico Santa María y sus docentes, por el respaldo académico y los recursos que hicieron posible mi desarrollo como profesional y este trabajo que culmina esta etapa fundamental en mi vida y da comienzo a una nueva.

2 Resumen

La presente memoria analiza los determinantes del precio de oferta de casas en el Gran Santiago (2025) y compara el desempeño de modelos explicativos basados en mínimos cuadrados ordinarios con algoritmos de aprendizaje automático, utilizando una base de 462 observaciones con el precio expresado en UF y modelado en logaritmos. El conjunto de variables incluye atributos físicos de la vivienda, índices urbanos de entorno y dummies territoriales de localización, con procedimientos de validación y diagnóstico que aseguran la robustez de las estimaciones y la evaluación fuera de muestra.

En modelos explicativos, conforme a métricas y supuestos clásicos de la econometría, la regresión lineal con índices compuestos construidos obtuvo el mejor desempeño. Dicho modelo observa que la localización en las zonas nor-oriente y norte-centro de la Región Metropolitana, el equipamiento interior y la pertenencia a condominio cerrado tienen efectos económicos destacados sobre el precio, donde al ubicarse en la zona norte–oriente se asocia a un incremento aproximado de 65,22 %, cada baño adicional a un 21,58 % y vivir en condominio cerrado a un 12,08 %. Estos resultados se complementan con la incidencia positiva de los índices de amenidades y servicios del entorno, además de la contribución de la superficie, cuya variación marginal por metro cuadrado es positiva aunque de menor magnitud relativa.

En los modelos de aprendizaje automático, se definen como parte del estudio los modelos `Random Forest` y `XGBoost`, donde se evaluaron en configuraciones `Full` y `Trimmed` para analizar el impacto de las colas del dataset. En `Random Forest (Full)` se obtiene en prueba un R^2 de 0,799, con `MAE` (Error absoluto medio) de 1.076,83 UF y `RMSE` (Error cuadrático medio)

de 1.558,36 UF; en `Random Forest (Trimmed)` el R^2 desciende a 0,739, manteniendo `MAE` (Error absoluto medio) y `RMSE` (Error cuadrático medio) de prueba en el mismo orden de magnitud que la versión `Full`.

En `XGBoost (Full)` se registra un R^2 de 0,797, con `MAE` (Error absoluto medio) de 1.077,60 UF y `RMSE` (Error cuadrático medio) de 1.565,80 UF, mientras que `XGBoost (Trimmed)` logra un R^2 de 0,783 junto con `MAE` (Error absoluto medio) de 1.014,18 UF y `RMSE` (Error cuadrático medio) de 1.471,75 UF. Considerando simultáneamente R^2, MAE y RMSE en el set de prueba, la mejor configuración para minimizar el error de predicción es `XGBoost (Trimmed)`, que presenta los valores más bajos de `MAE` y `RMSE`.

En conjunto, los hallazgos confirman la relevancia de los atributos físicos clave y de la localización territorial, así como el aporte del entorno urbano medido por índices, para explicar y predecir el precio residencial. Los modelos lineales ofrecen interpretabilidad y estimación de efectos marginales útiles para análisis y comunicación, mientras que los modelos de árboles proporcionan menor error de predicción en validación, por lo que resultan adecuados para procesos de estudio u operativos de tasación y apoyo a decisiones en el mercado inmobiliario.

Palabras claves — Precio de la vivienda; Gran Santiago; modelos hedónicos; regresión lineal; mínimos cuadrados ordinarios; aprendizaje automático; Random Forest; XGBoost; accesibilidad urbana; tasación inmobiliaria; índices compuestos.

3 Abstract

This thesis analyzes the determinants of asking prices for single-family houses in Greater Santiago (2025) and compares the performance of explanatory models based on ordinary least squares with machine learning algorithms, using a dataset of 462 observations with prices expressed in UF and modeled in logarithms. The set of covariates includes physical housing attributes, urban environment indices, and territorial location indicators (dummies), together with validation and diagnostic procedures that ensure robust estimation and out-of-sample evaluation.

In explanatory models, consistent with standard econometric metrics and assumptions, the linear regression with constructed composite indices delivered the best performance. This specification indicates that location in the northeast and north-central zones of the Metropolitan Region, interior equipment, and residence within a gated community have economically meaningful effects on price: being in the northeast zone is associated with an approximate increase of 65.22 %, each additional bathroom with 21.58 %, and living in a gated community with 12.08 %. These findings are complemented by the positive incidence of amenity and service indices in the surrounding area, as well as the contribution of floor area, whose marginal variation per square meter is positive albeit of smaller relative magnitude.

For machine learning models, `Random Forest` and `XGBoost` were evaluated under `Full` and `Trimmed` configurations to assess the impact of distribution tails. In `Random Forest (Full)`, the test performance is $R^2 = 0{,}799$, with `MAE` of 1,076.83 UF and `RMSE` of 1,558.36 UF; in `Random Forest (Trimmed)`, R^2 decreases to 0.739, while test `MAE` and `RMSE` remain of similar magnitude to the `Full` version.

In `XGBoost (Full)`, the test results are $R^2 = 0{,}797$, `MAE` of 1,077.60 UF, and `RMSE` of 1,565.80 UF, whereas `XGBoost (Trimmed)` attains $R^2 = 0{,}783$ together with `MAE` of 1,014.18 UF and `RMSE` of 1,471.75 UF. Considering R^2, MAE, and RMSE jointly on the test set, the best configuration for minimizing prediction error is `XGBoost (Trimmed)`, which yields the lowest `MAE` and `RMSE`.

Overall, the findings confirm the relevance of key physical attributes and territorial location, as well as the contribution of the urban environment measured through indices, for explaining and predicting residential prices. Linear models provide interpretability and marginal-effect estimates useful for analysis and communication, while tree-based models deliver lower prediction error in validation, making them suitable for study workflows or operational appraisal and decision support in the real estate market.

Keywords — housing prices; Greater Santiago; hedonic models; linear regression; ordinary least squares; machine learning; Random Forest; XGBoost; urban accessibility; real estate appraisal; composite indices.

4 Introducción

El mercado inmobiliario chileno viene de un ajuste reciente, y en 2025 muestra un cuadro mixto. Se observa en primer lugar menor dinamismo en ventas y una oferta nueva contenida por la caída previa en permisos, mientras tanto, las tasas hipotecarias han descendido desde sus máximos, lo que mejora gradualmente la accesibilidad. En conjunto, estos factores sugieren una normalización paulatina con diferencias según segmentos y territorios (CChC, 2025c).

En el Gran Santiago los indicadores apuntan a cierta estabilización desde niveles bajos. En el mercado de la vivienda (casa), la rotación es más lenta que en departamentos, pero la combinación de tasas en retroceso y una oferta nueva más acotada favorece una absorción ordenada del stock y un escenario de ajuste más gradual (CChC, 2025b).

En lo que respecta a la formación de precios, este combina atributos internos y efectos de localización. En Santiago, superficie, características internas de la vivienda, estacionamientos y antigüedad tienen impactos sistemáticos, mientras la accesibilidad y los servicios de barrio generan diferenciales claros, donde por ejemplo la cercanía a Metro y corredores de transporte suele capitalizarse en el precio, y la sectorización socioeconómica refuerza brechas territoriales (Agostini y Palmucci, 2008; Agostini et al., 2016).

Ante este contexto, es interesante medir dos enfoques vía modelos hedónicos para estimar efectos marginales interpretables y métodos de aprendizaje automático para mejorar la precisión predictiva capturando no linealidades. La comparación sobre una base de anuncios de casas del Gran Santiago (precios en UF y variables internas, de accesibilidad y de localización) busca equilibrar explicación y predicción para apoyar tasación, pricing y evaluación de proyectos.

5 Objetivos

5.1 Objetivo general

Identificar y cuantificar el efecto de atributos físicos, del entorno y territoriales sobre el precio de oferta de casas en el Gran Santiago, y comparar la capacidad explicativa de modelos econométricos con la capacidad predictiva de técnicas de aprendizaje automático.

5.2 Objetivos específicos

1. Investigar literatura nacional e internacional sobre determinantes del precio de la vivienda y enfoques de modelamiento aplicados al mercado inmobiliario.

2. Investigar y validar el uso de precios de oferta para la construcción de base de datos puesta a disposición de procesos de modelado.

3. Construir y depurar una base de datos propia de anuncios de casas, estandarizando precios en UF y definiendo variables internas, de entorno y territoriales.

4. Describir y analizar estadísticamente las variables y su relación con el precio, identificando patrones y posibles no linealidades relevantes para la especificación de modelos.

5. Estimar modelos hedónicos mediante regresión lineal para cuantificar impactos marginales de los atributos y evaluar su ajuste e idoneidad.

6. Desarrollar y validar modelos de aprendizaje automático y comparar su desempeño predictivo con el de los modelos econométricos.

7. Analizar diferencias territoriales y por segmentos del mercado, generando predicciones representativas por zona y discutiendo sus implicancias para evaluación y toma de decisiones.

6 Alcance

La presente investigación se circunscribe al análisis del mercado inmobiliario del Gran Santiago, Región Metropolitana de Chile, durante el año 2025, utilizando como insumo principal una base de datos elaborada por el autor a partir de información recolectada de portales inmobiliarios y complementada con indicadores de accesibilidad y equipamiento urbano.

El estudio se enfoca en propiedades residenciales (exclusivamente casas) y considera como variables de análisis atributos físicos (superficie, número de dormitorios y baños, estacionamientos, antigüedad, estado de conservación, pertenencia a condominio), características del entorno (proximidad a áreas verdes, supermercados, farmacias, centros comerciales) y factores territoriales (comuna, barrio, localización).

El alcance metodológico incluye:

- Análisis descriptivo y exploratorio de las variables.
- Estimación de modelos econométricos tipo hedónico para cuantificar el impacto marginal de los determinantes del precio.
- Desarrollo y validación de modelos predictivos mediante técnicas de aprendizaje automático, comparando su capacidad explicativa y predictiva con el modelo econométrico.
- Análisis compacto en zonas del Gran Santiago.

Sin embargo, este trabajo no considera:

- Proyecciones a largo plazo de precios fuera del periodo analizado.
- Análisis del mercado de viviendas no disponibles para ser habitadas (en construcción o inversión para remodelación) con transacciones previas a la recolección de datos.

- Evaluaciones de políticas públicas específicas ni simulaciones de cambios regulatorios.

En consecuencia, los resultados y conclusiones se aplican únicamente al contexto espacial y temporal definido, y su extrapolación a otros mercados o periodos debe realizarse con cautela.

7 Marco Teórico

7.1 Mercado Inmobiliario en Chile: situación actual

En el mercado inmobiliario en Chile, tras el ajuste de 2023–2024, 2025 muestra un cuadro mixto: la Cámara Chilena de la Construcción (CChC) reportó que las ventas de viviendas cayeron 18 % interanual en el 1T-2025 (CChC, 2025c); aun así, el gremio proyectaba para el año una recuperación moderada respecto de 2024, desde bases históricamente bajas. Por el lado de la oferta futura, los permisos de edificación venían débiles: el INE registró una disminución de 0,8 % en 2024, lo que tiende a contener nuevos inicios y ayuda a estabilizar inventarios (INE, s.f.)

Las condiciones financieras han mejorado levemente, puesto que el Banco Central mantuvo la TPM en 4,75 % en septiembre y, en paralelo, la tasa promedio de créditos hipotecarios en UF para plazos mayores a 3 años retrocedió a 4,21 % en septiembre de 2025 (4,16 % en la última semana del mes), su menor nivel en más de dos años (Carrizo, 2025). Esto favorece gradualmente la accesibilidad, aunque los estándares de crédito siguen más exigentes que en la prepandemia (Banco Central de Chile, 2025).

En términos estructurales, la necesidad habitacional sigue siendo elevada, en donde la estimación oficial del Minvu con Casen 2022 cifra el déficit cuantitativo en 552.046 requerimientos y el déficit cualitativo en 1.263.576 viviendas, con mayor presión en la Región Metropolitana y el norte grande. Esta brecha sostiene la demanda subyacente, pero su traducción en compras depende de la renta, las tasas y la disponibilidad de subsidios (MINVU, 2023).

7.2 Mercado Inmobiliario en el Gran Santiago

En el 2T-2025, el indicador de Mercado Inmobiliario de Gran Santiago (MIS) de la CChC mostró que las ventas registraron un avance de 1,8 % trimestral anualizado, una señal de estabilización tras la debilidad del inicio del año. El propio archivo de publicaciones del gremio para el Gran Santiago y sus informes específicos respaldan esta lectura de piso/recuperación gradual, aunque desde niveles inferiores al promedio histórico (CChC, 2025b).

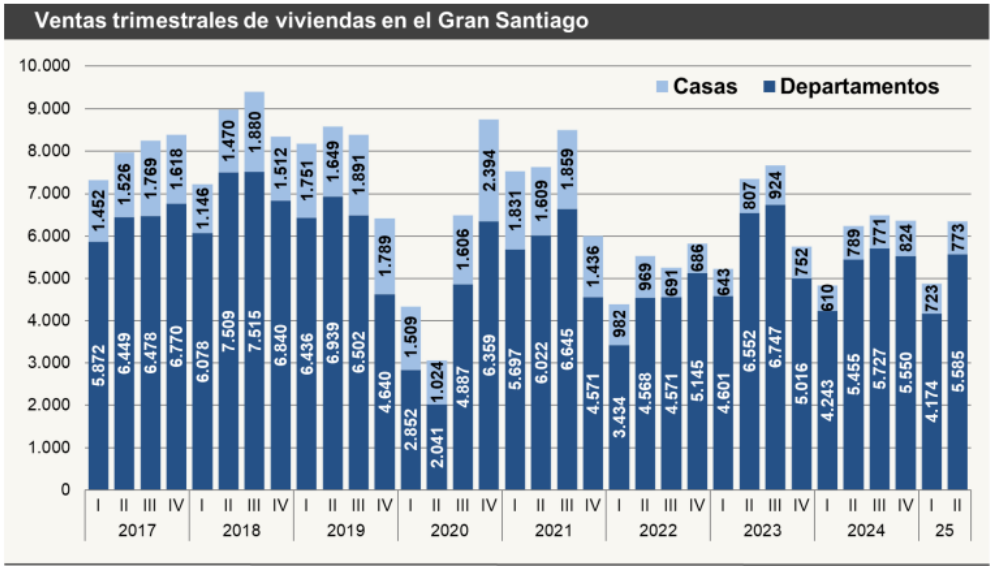

Figura 1: Ventas trimestrales de viviendas en el gran santiago. Fuente: CChC 2025.

En casas del Gran Santiago, el último corte mensual disponible muestra que en junio de 2025 la oferta alcanzó 6.225 unidades (8,9 % del total de 69.806 viviendas), con 289 casas vendidas y una velocidad de 21,5 meses para agotar el stock; es decir, un mercado de baja rotación pero algo más dinámico que el promedio general de viviendas de ese mes. Visto en trimestre, durante el 2T-2025 se comercializaron 773 casas (de 6.359 viviendas totales), con una oferta cercana a 6.200 casas y una velocidad promedio de 24 meses, lo que sugiere una absorción lenta pero estable en el segmento de casas frente a departamentos (CChC, 2025b).

Mercado de casas	Junio 2025		Junio 2024		Promedio histórico	
Comuna	Oferta	Meses	Oferta	Meses	Oferta	Meses
01 Puente Alto	711	24	671	19	828	12
02 San Bernardo	433	20	173	25	399	12
03 Buin	1.059	46	399	57	869	25
04 Maipú/Pudahuel/Cerrillos	152	11	98	15	613	20
05 Padre Hurtado	529	20	274	16	385	16
06 Peñaflor/Talagante/Melipilla	333	28	193	36	155	16
07 Lampa/Quilicura	1.161	15	991	10	1.148	10
08 Huechuraba	27	5	30	13	207	17
09 Colina	1.275	20	1.360	19	1.125	20
10 Peñalolén/La Florida/La Reina	311	37	256	22	383	19
11 Las Condes/Lo Barnechea/Vitacura/Providencia	235	44	173	51	260	20

Figura 2: Oferta de casas Junio 2025. Fuente: CChC 2025.

Con ello y pensando en el futuro, la baja paulatina de tasas hipotecarias observada a septiembre, junto con una oferta nueva acotada por la menor tramitación de permisos en años previos, debiera favorecer una absorción más ordenada del stock y apoyar los precios en zonas consolidadas, mientras que la accesibilidad seguirá muy condicionada por ingresos y requisitos bancarios.

7.3 Uso de precios de oferta en la estimación del valor de la vivienda

Diversos estudios sobre mercados inmobiliarios han distinguido entre los precios de oferta (*listings*), aquellos anunciados por los vendedores, y los precios de cierre (*sales prices*), que refieren a los finalmente pactados en la transacción. Si bien los precios de cierre reflejan de manera más directa el valor de mercado efectivamente transado, los precios de oferta constituyen una fuente de información más amplia y temprana sobre las expectativas del mercado y el comportamiento de los agentes inmobiliarios. Según un estudio que evalúa el precio de la vivienda en diferentes etapas del proceso de compra-venta (Shimizu et al., 2016), las bases de datos basadas en precios de oferta permiten capturar dinámicas de formación de precios antes de la negociación final, mostrando “el valor percibido” por los vendedores en el momento inicial del proceso de comercialización.

Además, los precios de oferta presentan ventajas operativas y analíticas en estudios pre-

dictivos. Northcraft y Neale (1987) evidencian que los precios de lista funcionan como anclas cognitivas que influyen tanto en compradores como en tasadores, reflejando las condiciones de referencia del mercado local. Por otra parte, Gordon y Winkler (2017) señalan que el uso de precios de oferta incrementa la cobertura del análisis, ya que incluye propiedades activas, retiradas o no vendidas, reduciendo el sesgo de selección *ex post* presente en los precios de cierre. En contextos donde los datos de transacciones efectivas son escasos o poco actualizados, como ocurre frecuentemente en Chile, el uso de precios de oferta permite construir modelos más robustos y actualizados de estimación y predicción (Shimizu et al., 2016; Gordon y Winkler, 2017).

En síntesis, varios autores destacan que, aunque los precios de oferta suelen sobreestimar en promedio los precios finales, esa diferencia puede controlarse mediante técnicas econométricas y en el uso correcto en la selección de datos (Anenberg y Laufer, 2017). Esto hace posible emplear los precios de oferta como una aproximación válida del valor de mercado, especialmente cuando el objetivo del estudio es comparativo o predictivo, más que estrictamente valuatorio.

7.4 Factores de impacto en el valor de la vivienda

En el nivel intrínseco de la vivienda, la evidencia para Santiago muestra que el precio de una vivienda se explica en gran medida por su paquete de atributos físicos, superficie construida y del terreno, número de dormitorios y baños, presencia de estacionamientos y bodega, calidad de materialidades y terminaciones, antigüedad/estado de conservación y la distribución funcional (plantas eficientes, iluminación, etc.). Los modelos hedónicos aplicados a numeradas transacciones en la Región, confirman que estos rasgos tienen efectos sistemáticos y significativos sobre el precio; en particular, más metros cuadrados útiles y más baños/dormitorios tienden a capitalizarse con primas porcentuales, mientras que mayor antigüedad sin remodelación suele descontar valor

(Vergara-Perucich, 2023).

Un segundo bloque de determinantes proviene de la accesibilidad urbana, precisamente en la conectividad a centros de empleo y servicios mediante Metro y ejes de transporte masivo. Para Santiago, un estudio clásico estimó que el anuncio y la puesta en servicio de la Línea 4 generó un aumento capitalizado en los precios de vivienda en zonas cercanas a estaciones, reflejando el valor de menores tiempos/costos de viaje; efectos de este tipo son especialmente relevantes para casas bien conectadas a estaciones y autopistas con flujo expedito (Agostini y Palmucci, 2008).

A nivel ambiental y de servicios de barrio, la proximidad y calidad de colegios, salud, áreas verdes y equipamientos también se incorporan al precio. En Chile, el Índice de Calidad de Vida Urbana (ICVU) (elaborado por la CChC y la PUC) mide, entre otras, las dimensiones de Conectividad y movilidad, Vivienda y entorno, Salud y medioambiente y Condiciones socioculturales; comunas con mejor dotación relativa en estas dimensiones tienden a sostener mayor disposición a pagar por vivienda, mecanismo que la literatura hedónica capta como “amenities” urbanos valorados por los hogares (CChC, 2025a).

Finalmente, las externalidades negativas del entorno influyen a la baja. Evidencia para Santiago encuentra que el crimen (según tipo) reduce los precios inmobiliarios, con sensibilidad que varía por propiedad y delito; las casas, por su exposición al espacio público y horizontes de tenencia más largos, muestran ajustes que pueden materializarse con cierto rezago, lo que subraya la relevancia de la seguridad barrial en la formación de precio (Andrade y Cifuentes, 2019).

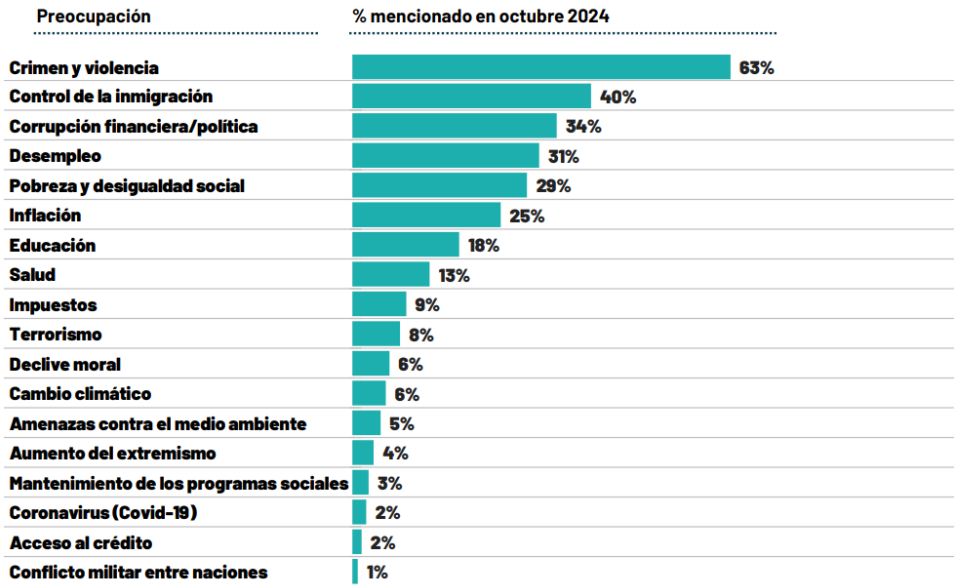

Figura 3: Principales preocupaciones de los chilenos. Fuente: Ipsos 2024.

7.5 Sectorización e ingresos económicos

En Santiago, la sectorización socioeconómica (concentración histórica de hogares de mayores ingresos en el "sector oriente" y menor ingreso en sectores sur y poniente) se traduce en precios segmentados para casas, allí donde se aglomeran altos ingresos, la disposición a pagar por ubicación, colegios, servicios y conectividad se capitaliza más intensamente en el valor del suelo y la vivienda. La literatura para el Gran Santiago documenta esta concentración de altos ingresos y su aporte a la segregación residencial, lo que refuerza brechas territoriales de precio entre comunas y submercados (Agostini et al., 2016).

A su vez, El nivel de ingresos de los hogares condiciona directamente la capacidad de compra y, por ende, el tramo de precios que puede absorber cada zona. Según la CASEN más reciente (2022) revela que la distribución del ingreso por decil y sus brechas, insumo estándar para estimar cuántos hogares pueden financiar una casa de cierto precio (en UF) en cada comuna o

área funcional. Cuando se combina esta información con precios locales, se infiere qué segmentos de ingreso "calzan" con barrios específicos, reforzando la segmentación espacial de la demanda (Ministerio de Desarrollo Social y Familia, 2023).

Además de los ingresos, la calificación crediticia y el costo del financiamiento filtran la demanda efectiva. El Banco Central de Chile reportó en su Informe de Estabilidad Financiera (1S-2025) que la carga financiera de los hogares retrocedió respecto de los máximos recientes, aunque la mora hipotecaria se mantuvo más alta que en la prepandemia; este telón de fondo moderó el acceso al crédito en 2024–2025 y, por tanto, el universo de compradores solventes por tramo de precio (Soto, 2025). En paralelo, estimaciones de mercado difundidas en 2025 muestran que la proporción de hogares que cumple por ingreso con las exigencias de un hipotecario cayó con fuerza en cinco años—un ajuste que restringe la demanda en los submercados de casas más caras, mientras mantiene mayor elasticidad en zonas medias y con subsidio (Ortega, 2025).

En síntesis, en Santiago, ingresos y sectorización se acoplan de tal forma, donde los barrios con mayores ingresos sostienen precios de casas más altos por mayor disposición a pagar y mejores amenidades; la distribución del ingreso delimita qué comunas pueden absorber determinados tramos de UF; el acceso efectivo al crédito—en un contexto de carga financiera y estándares bancarios exigentes—acota la demanda solvente; y la calidad urbana medida por el ICVU y los subsidios focalizados amplifican o mitigan esas diferencias según dónde esté la oferta.

7.6 Modelos de regresión en la estimación de precios

La herramienta estándar para "poner precio" a una casa a partir de sus atributos es la regresión hedónica, que modela el valor observado como la suma (o suma logarítmica) de precios

implícitos de cada característica: superficie construida y de terreno, dormitorios/baños, antigüedad, estacionamientos, y también variables de localización y entorno (tiempos de viaje, cercanía a Metro/colegios/áreas verdes, calidad barrial). Su sustento teórico es el marco de Rosen (1974), que formaliza cómo, en mercados de productos diferenciados (como la vivienda), los precios revelan la disposición a pagar por atributos y habilita extensiones para recuperar precios marginales y evaluar bienestar (Rosen, 1974).

7.6.1. Caso aplicado en Chile

Para Santiago de Chile, hay evidencia reciente y específica que opera con regresiones lineales/semilog y amplios controles. Un estudio 2008–2019 publicado en *Land Use Policy* estima, con múltiples regresiones, las primas de localización asociadas al "efecto Metro", donde las viviendas próximas a estaciones presentan precios significativamente más altos, una vez controladas las características físicas y del entorno; es un ejemplo claro de cómo variables de accesibilidad entran linealmente (o log-linealmente) y resultan en coeficientes estadísticamente robustos (López-Morales et al., 2023). En la misma línea, un paper de 2023 enfocado en Santiago contrasta distintos conjuntos de predictores en un modelo hedónico semilog, mostrando qué combos de atributos físico-locacionales maximizan el poder explicativo fuera de muestra. Ambos confirman que, la forma funcional log-precio capta bien elasticidades y la combinación de atributos internos junto con accesibilidad explica buena parte de la varianza del precio observado (Vergara-Perucich, 2023).

7.6.1.1 Prácticas de especificación y evaluación

En función de la literatura notada, la selección de los procesos funcionales denotan en que usar log-precio ofrece interpretaciones más estables (elasticidades aproximadas) y suele mejorar el ajuste.

Variables clave.

- Superficies (m^2), permitiendo no linealidades si corresponde.
- Programa (nº de dormitorios/baños).
- Antigüedad y estado de conservación.
- Estacionamientos y bodega.
- Localización: distancias/tiempos a Metro y subcentros.
- Dummies comunales o efectos fijos por zona.
- *Amenities* (colegios, áreas verdes).
- Dummies de período para capturar ciclos.

Riesgos y correcciones.

- Multicolinealidad (p. ej., superficie vs. dormitorios): monitorear y, si es necesario, reparametrizar o reducir variables.
- Heterocedasticidad: usar errores robustos (HC).
- Sesgos por omitidas (p. ej., calidad constructiva no observada).

Validación.

- Reportar R^2 ajustado, pero priorizar error fuera de muestra (RMSE/MAE) vía *hold-out* o *cross-validation*.
- Inspeccionar residuos y *leverage* para detectar observaciones atípicas/influyentes.

7.7 Modelos de aprendizaje automático en predicción de precios

Aunque la regresión hedónica lineal ha sido la herramienta tradicional para la predicción de precios de inmuebles, la disponibilidad de grandes bases de datos debido al avance de almacenamiento (listados inmobiliarios, catastros, GIS, datos de entorno) y la presencia de relaciones no lineales e interacciones complejas impulsaron el uso de aprendizaje automático (ML). Estudios de tasación masiva muestran que enfoques de ML(como bosques aleatorios y ensambles de árboles) tienden a mejorar la precisión frente a la regresión múltiple clásica, al capturar no linealidades y efectos de umbral difíciles de modelar de forma paramétrica (Yilmazer y Kocaman, 2020).

Entre los modelos más empleados destacan:

- Regresiones penalizadas (Ridge/Lasso) útiles cuando hay muchas variables correlacionadas
- Árboles de decisión y Random Forest, robustos y con métricas estables
- Gradient Boosting (XGBoost/CatBoost/LightGBM), que suelen alcanzar el mejor desempeño predictivo en escenarios tabulares
- Redes neuronales (fuentes de datos fuertes en imágenes de fachadas, texto de avisos, datos satelitales)

La literatura reciente encuentra de manera recurrente que los métodos de boosting logran menores errores (RMSE/MAE) que alternativas tradicionales, siempre que haya una validación adecuada (k-fold), tuning de hiperparámetros y control de sobreajuste.

Variables

Los conjuntos de variables abarcan atributos estructurales (m² construidos/útiles, dormitorios, baños, antigüedad, calidad), de localización (distancias a centros, parques, costa, transporte público), socioeconómicas (ingreso del área, densidad), y de mercado (estacionalidad, tasas hipotecarias). La ingeniería de características es clave: variables espaciales (buffers, travel-time), codificación de categorías (one-hot/target encoding), transformaciones logarítmicas del precio, y manejo de valores atípicos. El preprocesamiento correcto (imputación, estandarización cuando procede) y la detección de leakage determinan gran parte del éxito predictivo.

Evaluación, validación y estabilidad temporal

Para tasación y pricing se recomiendan métricas continuas como RMSE, MAE y MAPE, reportadas bajo validación cruzada estratificada por espacio y/o tiempo (por ejemplo, time-based splits) para evaluar la capacidad de generalización frente a cambios de ciclo. Es deseable monitorear drift (cambios en la distribución de covariables o en la relación precio-características) y recalibrar modelos periódicamente, pues los mercados inmobiliarios exhiben dinámica temporal y efectos macroeconómicos.

7.7.1. Caso aplicado en Chile

Un estudio de académicos de la Universidad de Chile aplica métodos de aprendizaje automático al mercado residencial de Santiago, Chile, extendiendo el enfoque hedónico clásico con

modelos capaces de capturar relaciones no lineales y efectos de interacción entre atributos de la vivienda y del entorno. A partir de un repositorio masivo de transacciones, los autores estiman el precio por m² en función de características estructurales (tamaño, antigüedad, calidad), de localización y accesibilidad (proximidad a metro, subcentros, servicios), y contexto socioambiental (ingresos comunales, seguridad, oferta educativa y áreas verdes). El objetivo central es comparar el desempeño predictivo de distintos algoritmos (p. ej., Random Forest, SVR, redes neuronales) y evaluar su robustez en segmentos socioeconómicos diferenciados. Los resultados se enmarcan en la literatura que reporta mejoras de precisión con ensambles de árboles frente a modelos tradicionales, manteniendo interpretabilidad práctica mediante la descomposición por atributos y la segmentación espacial (Chardon et al., 2019).

7.7.1.1 Proceso y resultados

El estudio usa un subconjunto de 205.600 transacciones de un total de 334.353 viviendas (2007–2018) en Santiago, modelando precio por m², donde se eliminan outliers con reglas por zona homogénea y se entrena con 80/20 (train/test). Posterior a ello, se comparan ANN (5 capas), SVR con kernel RBF y Random Forest (500 árboles, sin poda) mediante MAPE y varianza explicada.

En test, Random Forest obtiene MAPE 11,27 % y 73,7 % de varianza explicada, superando a SVR (14,97 %; 67,1 %) y ANN (22,10 %; 71,6 %). Luego se reentrena RF separando por nivel de ingreso comunal: los resultados son similares en error (MAPE = 11,3 %) y la varianza explicada sube en comunas de mayor ingreso (81,1 %) respecto de menor ingreso (72,9 %), manteniendo buena capacidad predictiva en ambos segmentos (Chardon et al., 2019).

8 Metodología

8.1 Diseño de investigación y enfoque

El desarrollo de este estudio conlleva consigo un enfoque cuantitativo, no experimental, observacional y de corte transversal. Se analiza información de anuncios de viviendas del Gran Santiago recopilada en un período acotado, sin intervención del investigador sobre las variables observadas.

El trabajo se inscribe en el modelo de precios hedónicos, marco que asume que el precio observado de un bien heterogéneo (la vivienda) refleja la suma de los valores implícitos de sus atributos. En este contexto, el objetivo principal es estimar el efecto marginal de atributos internos (propios del inmueble, como superficie o número de dormitorios) y externos (propios del entorno, como accesibilidad y dotación de servicios varios) sobre el precio de oferta de la vivienda en el Gran Santiago. Complementariamente, se evalúa la capacidad predictiva de distintos enfoques de modelamiento, equilibrando interpretabilidad y desempeño fuera de muestra.

Finalmente, la unidad de análisis es exclusivamente respecto al anuncio de vivienda. El análisis se limita a precios de oferta (no de cierre), por lo que las estimaciones reflejan el comportamiento del mercado en etapa de publicación. Los resultados son asociativos (no causales), coherentes con la naturaleza observacional del estudio y las suposiciones del enfoque hedónico.

8.2 Ámbito del estudio

8.2.1. Población objetivo

La población de referencia corresponde a las viviendas en oferta publicadas en portales inmobiliarios en el Gran Santiago. El análisis se centra en anuncios de casas disponibles en la Región Metropolitana de Chile, considerando aquellas comunas que forman parte del área metropolitana y que concentran la mayor actividad inmobiliaria. La unidad de análisis es cada anuncio individual de vivienda, tratado como una observación independiente dentro de la base de datos.

8.2.2. Cobertura geográfica

El estudio se circunscribe al Gran Santiago, área metropolitana conformada por un subconjunto de comunas de la Región Metropolitana de Santiago. Si bien la región está integrada por un total de 52 comunas, no todas participan de manera directa en la dinámica inmobiliaria urbana que caracteriza a la capital. En términos generales, las comunas más periféricas presentan un carácter predominantemente rural o semi-rural, con una densidad poblacional menor, usos de suelo distintos (agrícola o mixto), y una menor integración con la red de infraestructura urbana. Factores como la ausencia del sistema de transporte metropolitano (particularmente el Metro de Santiago), la limitada conectividad vial hacia el centro y la escasa dotación de equipamientos y servicios urbanos hacen que estas comunas no formen parte del núcleo consolidado de la ciudad.

En consecuencia, se delimita el análisis a aquellas comunas que efectivamente componen el área metropolitana funcional, donde se concentra la mayor parte de la actividad residencial, laboral y de servicios de la capital. Estas comunas son:

Tabla 1: Listado de comunas seleccionadas para el estudio

Cerrillos	Lo Barnechea
Cerro Navia	Lo Espejo
Conchalí	Lo Prado
El Bosque	Macul
Estación Central	Maipú
Huechuraba	Ñuñoa
Independencia	Pedro Aguirre Cerda
La Cisterna	Peñalolén
La Florida	Providencia
La Granja	Pudahuel
La Pintana	Puente Alto
La Reina	Quilicura
Las Condes	Quinta Normal
Recoleta	Renca
San Joaquín	San Miguel
Santiago	Vitacura

Estas comunas no solo concentran la mayor proporción de la población de la Región Metropolitana, sino que además configuran un espacio urbano con características homogéneas en términos de mercado inmobiliario, accesibilidad y provisión de servicios urbanos, lo que permite realizar un análisis comparativo robusto del precio de la vivienda en función de variables internas y externas.

8.2.3. Delimitaciones Adicionales

El estudio se enfoca en precios de oferta publicados, no en precios de cierre de transacciones efectivas. Por lo tanto, los resultados reflejan el comportamiento de oferta y expectativas de

valor en el mercado inmobiliario más que el precio final de compraventa. Asimismo, se excluyen tipos de propiedades no comparables (como departamentos o terrenos), de manera de garantizar homogeneidad en la unidad de análisis.

8.3 Fuentes de datos y recolección

8.3.1. Fuente principal de datos

La información utilizada en este estudio para la construcción de la base de datos proviene de anuncios de viviendas en venta publicados en sitio web "Portal Inmobiliario". Esta fuente fue seleccionada debido a su amplia cobertura en el mercado formal, la actualización constante de la oferta (equivalente al 50 % de la oferta total en la región según informes del Banco central) y la variedad de atributos que reporta en cada anuncio, tales como precio, superficie, número de dormitorios y baños, ubicación geográfica, entre otros.

8.3.2. Método de recolección

La construcción de la base de datos se llevó a cabo mediante un proceso de levantamiento manual de información directamente desde el portal inmobiliario. En lugar de emplear técnicas automatizadas de extracción (web scraping), se procedió a registrar uno a uno los anuncios publicados, consignando de forma sistemática los atributos relevantes de cada propiedad en un archivo maestro.

Este procedimiento se realizó aplicando únicamente el filtro de propiedades tipo “casa” y restringiendo la búsqueda a las comunas previamente seleccionadas del Gran Santiago, con el propósito de asegurar la coherencia del universo muestral con los objetivos del estudio. Una vez

filtradas las ofertas visibles en el portal, cada anuncio se revisó y codificó, completando variables relativas al inmueble (superficie, número de dormitorios, baños, estacionamiento, entre otros), ubicación geográfica (comuna) y precio de oferta.

El carácter manual de la recolección ofrece principalmente las siguientes ventajas metodológicas:

- **Control de calidad de los datos:** al revisar cada anuncio individualmente, se minimizó la incorporación de registros duplicados, incompletos o inconsistentes.
- **Mayor validez de la muestra:** al no depender de algoritmos de extracción, se evitó sesgo asociado a limitaciones técnicas del scraping, garantizando que la muestra estuviera efectivamente compuesta por propiedades ajustadas a los criterios de selección.
- **Aleatoriedad relativa en la selección:** la muestra se conformó sin aplicar filtros adicionales más allá del tipo de propiedad y comuna, lo que permitió capturar la heterogeneidad natural de la oferta en el mercado y obtener un conjunto representativo de observaciones.
- **Selección visual de viviendas comparables:** con la finalidad de obtener resultados robustos y fidedignos es que la selección debe ser con viviendas en estructura comparables.

8.3.3. Datos Finales

Finalmente, aplicando todas las herramientas mencionadas anteriormente es que se lograron recolectar un total de 462 observaciones limpias, cuyo número representa un 1.25 % de la oferta total de casas publicadas en el portal, lo cual constituye para este estudio una muestra apta para la determinación y comprensión de patrones y relaciones entre variables, como también numéricamente suficiente para los procesos de modelamiento a realizar.

Como resultado, se obtuvo un archivo final tipo **Excel**, que concentra la información bruta inicial y constituye la base para los procesos posteriores de categorización, codificación, transformación y modelamiento.

8.4 Definiciones operacionales de variables

En esta sección se presentan las variables utilizadas en el análisis y la forma en que fueron definidas operativamente a partir de la base de datos recopilada. El objetivo es detallar con claridad qué representa cada variable, cómo se midió, en qué unidades se expresa y qué rol cumple dentro de los modelos. Esta explicitación permite asegurar la replicabilidad del estudio y la correcta interpretación de los resultados.

8.4.1. Variable dependiente

La variable dependiente corresponde al precio de oferta de la vivienda, expresado en Unidades de Fomento (UF). La decisión de utilizar esta unidad en lugar de pesos chilenos responde a que la UF es un indicador indexado a la inflación, ampliamente utilizado en el mercado inmobiliario chileno para expresar precios de compraventa. De este modo, se asegura que los valores analizados sean comparables en el tiempo, evitando distorsiones derivadas de la pérdida de poder adquisitivo del peso.

Adicionalmente, en el proceso de modelamiento el precio en UF fue sometido a una transformación logarítmica, lo cual reduce la asimetría en la distribución y permite interpretar los coeficientes de regresión en términos de semi-elasticidades o elasticidades en algunos casos, según la transformación de algunas variables independientes. Así, las estimaciones reflejan variaciones en el precio (en UF) frente a cambios marginales en los atributos de la vivienda.

8.4.2. Variables explicativas

8.4.2.1 Variables internas

Estas variables representan las características propias de cada vivienda, registradas directamente desde los anuncios. Entre ellas destacan:

- **Superficie total [m^2]:** medida en metros cuadrados del terreno total asociado a la vivienda.
- **Superficie útil [m^2]:**medida en metros cuadrados de la construcción habitable.
- **Antigüedad [años]:** variable discreta que representa la cantidad de años desde la construcción o última remodelación del inmueble.
- **Dormitorios:** variable discreta que representa la cantidad de dormitorios que posee la vivienda de estudio.
- **Baños:** variable discreta que representa la cantidad de baños que posee la vivienda de estudio.
- **Estacionamientos:** variable discreta que representa la cantidad de estacionamientos que posee la vivienda de estudio.
- **Pisos:** variable discreta que representa la cantidad de pisos que posee la vivienda de estudio.
- **Piscina:** variable binaria codificada con (1) si la vivienda de estudio posee el atributo Piscina y (0) si no lo tiene.
- **Quincho:** variable binaria codificada con (1) si la vivienda de estudio posee el atributo Quincho y (0) si no lo tiene.

- **Bodega:** variable binaria codificada con (1) si la vivienda de estudio posee el atributo Bodega y (0) si no lo tiene.

Con lo anterior se expone que variables internas son consideradas relevantes para el estudio de precios hedónicos, y con ello se evidencia la búsqueda de determinar el impacto de la variación de las mismas en términos cuantitativos.

8.4.2.2 Variables externas

Además de características internas, es importante medir en conjunto el impacto de la ubicación y acceso a servicios, por lo que se incorporan las siguientes variables de entorno:

- **Localización comunal:** variable categórica que indica la comuna en la que se ubica la vivienda de estudio.
- **Localización de sector:** variable categórica que indica el barrio-villa en que se ubica la vivienda de estudio.
- **Condominio cerrado:** variable binaria codificada con (1) si la vivienda de estudio se encuentra en un entorno cerrado y (0) si no lo está.
- **Acceso a bus [metros]:** variable discreta que representa la distancia a la parada de servicio de transporte público (Transantiago) más cercana.
- **Acceso a metro [metros]:** variable discreta que representa la distancia a la estación de metro más cercana.
- **Acceso a educación pre-escolar [metros]:** variable discreta que representa la distancia al centro de educación pre-escolar más cercano.

- **Acceso a educación escolar [metros]:** variable discreta que representa la distancia al centro de educación escolar más cercano.

- **Acceso a educación superior [metros]:** variable discreta que representa la distancia al centro de educación superior más cercano.

- **Acceso a centro de salud [metros]:** variable discreta que representa la distancia al centro de salud más cercano.

- **Acceso a áreas verdes [metros]:** variable discreta que representa la distancia al recinto con áreas verdes (plaza - parque) más cercano.

- **Acceso a farmacia [metros]:** variable discreta que representa la distancia a la farmacia más cercana.

- **Acceso a supermercado [metros]:** variable discreta que representa la distancia al supermercado más cercano.

- **Acceso a centro comercial [metros]:** variable discreta que representa la distancia al centro comercial más cercano.

En síntesis, las variables externas permiten capturar de manera sistemática el efecto del entorno urbano y social sobre el valor de la vivienda. Su inclusión en el análisis resulta fundamental, ya que la localización y las condiciones de accesibilidad suelen explicar una fracción significativa de la variación en los precios inmobiliarios. Con ello, el estudio no se limita a describir atributos físicos del inmueble, sino que incorpora la dimensión espacial y territorial, ofreciendo una visión más completa del mercado habitacional del Gran Santiago.

En conjunto, la definición operacional de las variables asegura que cada atributo considerado en el análisis esté claramente delimitado en términos conceptuales, de medición y de unidad de referencia. Esta sistematización no solo facilita la interpretación de los resultados, sino que además constituye un insumo esencial para la etapa posterior de preparación de datos, en la cual se aplican procesos de limpieza, depuración y transformación orientados a garantizar la calidad y consistencia de la información utilizada en los modelos.

8.5 Modelos y herramientas de análisis

8.5.1. Regresión lineal múltiple (`OLS_BASE`)

El primer modelo estimado corresponde a una regresión lineal múltiple mediante Mínimos Cuadrados Ordinarios (OLS, por sus siglas en inglés). Este se planteó como el modelo base para el análisis, dado que la literatura en precios hedónicos tradicionalmente utiliza este enfoque para estimar el efecto marginal de atributos internos y externos sobre el valor de los inmuebles.

8.5.1.1 Objetivo del modelo

El propósito principal del modelo `OLS_BASE` fue identificar y cuantificar la relación entre las características de las viviendas y su precio de oferta (en UF), bajo un marco paramétrico que permite interpretar de manera directa los coeficientes como cambios porcentuales en el precio frente a variaciones en cada atributo (al trabajar con logaritmo del precio).

8.5.1.2 Especificación general

El modelo plantea de la siguiente forma:

$$\ln(Precio_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \tag{1}$$

donde:

- $\ln(Precio_i)$ es el logaritmo natural del precio de oferta en UF de la vivienda i.
- $\beta_0, \beta_1, \ldots, \beta_k$ son los coeficientes estimados del modelo, que miden la magnitud y dirección del efecto de cada atributo X sobre el precio de oferta de la vivienda.
- $X_{i1}, X_{i2}, \ldots, X_{ik}$ corresponden a los atributos internos y externos seleccionados de la base de datos del modelo.
- ε_i es el término de error aleatorio, asumido con media cero y varianza constante bajo los supuestos clásicos de Mínimos cuadrados.

8.5.1.3 Preparación de datos

La construcción de la base `BBDD_OLS_BASE` (que refiere al documento propio extraído de la base con los datos en bruto a trabajar para este modelo en específico) requirió aplicar un proceso de depuración específico orientado a garantizar la validez de los supuestos de la regresión lineal y la consistencia de las estimaciones.

8.5.1.3.1 Variables de zonificación

Para este modelo se optó por no incluir variables categóricas complejas (dummies de comunas y barrios), ya que el uso de categorías nominales con múltiples niveles (por ejemplo, tipologías o descripciones textuales de las viviendas) habría introducido un número excesivo de parámetros, reduciendo grados de libertad y complicando la inferencia. Por ello, solo se mantuvieron las dummies de comuna, necesarias para capturar diferencias espaciales sistemáticas entre territorios.

Conforme lo anterior es que, para no solo tener atributos físicos y de acceso, sino también incorporar el factor de ubicación de cada vivienda de estudio es que se realiza la construcción de 2 variables *dummies* que responden a una localización general de cada observación:

- ***Dummy_zona_oriente*:** variable binaria codificada con (1) si la vivienda de estudio se encuentra en la zona oriente de Santiago y (0) si no lo está (se infiere que la vivienda pertenece a la zona poniente de Santiago) realizado a partir de la comuna en la que se encuentra ubicada.

- ***Dummy_zona_sur*:** variable binaria codificada con (1) si la vivienda de estudio se encuentra en la zona sur de Santiago y (0) si no lo está (se infiere que la vivienda pertenece a la zona norte de Santiago) realizado a partir de la comuna en la que se encuentra ubicada.

De esta forma, las comunas para el análisis de las variables de zonificación quedaron codificadas según su ubicación en el mapa de la Región Metropolitana:

Figura 4: Mapa de comunas de Santiago

Tabla 2: Clasificación de comunas según orientación vertical y horizontal

Comuna	Vertical	Horizontal	Comuna	Vertical	Horizontal
Cerrillos	Sur	Poniente	Lo Espejo	Sur	Poniente
Cerro Navia	Norte	Poniente	Lo Prado	Norte	Poniente
Conchalí	Norte	Poniente	Macul	Sur	Oriente
El Bosque	Sur	Poniente	Maipú	Sur	Poniente
Estación Central	Norte	Poniente	Ñuñoa	Norte	Oriente
Huechuraba	Norte	Poniente	Pedro Aguirre Cerda	Sur	Poniente
Independencia	Norte	Poniente	Peñalolén	Sur	Oriente
La Cisterna	Sur	Poniente	Providencia	Norte	Oriente
La Florida	Sur	Oriente	Pudahuel	Norte	Poniente
La Granja	Sur	Poniente	Quilicura	Norte	Poniente
La Pintana	Sur	Poniente	Quinta Normal	Norte	Poniente
La Reina	Norte	Oriente	Recoleta	Norte	Poniente
Las Condes	Norte	Oriente	Renca	Norte	Poniente
Lo Barnechea	Norte	Oriente	San Joaquín	Sur	Poniente
San Miguel	Sur	Poniente	Santiago	Norte	Poniente
San Ramón	Sur	Poniente	Vitacura	Norte	Oriente

En síntesis, la inclusión de las comunas como parte de una zona en específico, no solo permite generalizar la componente de localización de estudio aportando robustez al modelo, sino también, aporta una interpretación más sencilla en lo que respecta a los resultados obtenidos en estas variables específicas.

8.5.1.4 Variables explicativas incluidas

Finalmente, el listado de variables explicativas a formar parte de este modelo denominado `OLS_BASE` son las siguientes:

- Dummies de Zonificación (dummy 0/1).
- Vivienda en Condominio cerrado (dummy 0/1).

- Superficie útil (m^2).
- Superficie de terreno (m^2).
- Antigüedad (años).
- Número de dormitorios.
- Número de baños.
- Número de pisos.
- Número de estacionamientos.
- Presencia de piscina, quincho y bodega (dummies 0/1).
- Acceso a servicios (transporte, educación, salud, comercio, áreas verdes, en metros)

Expresando así la ecuación final a modelar:

$$\begin{aligned}\ln(Precio_i) = \beta_0 &+ \beta_1 SupUtil_i + \beta_2 SupTerreno_i + \beta_3 Antiguedad_i \\ &+ \beta_4 Acc\_Bus_i + \beta_5 Acc\_Metro_i + \beta_6 Acc\_Preescolar_i \\ &+ \beta_7 Acc\_Escolar_i + \beta_8 Acc\_EducSup_i + \beta_9 Acc\_Salud_i \\ &+ \beta_{10} Acc\_Farmacia_i + \beta_{11} Acc\_AreasVerdes_i \\ &+ \beta_{12} Acc\_Supermercado_i + \beta_{13} Acc\_Mall_i \\ &+ \beta_{14} Dormitorios_i + \beta_{15} Banos_i \\ &+ \beta_{16} Estacionamientos_i + \beta_{17} Pisos_i \\ &+ \beta_{18} D_i^{Piscina} + \beta_{19} D_i^{Quincho} + \beta_{20} D_i^{Bodega} \\ &+ \beta_{21} D_i^{Zonificacion} + \beta_{22} D_i^{CondCerrado} \\ &+ \varepsilon_i\end{aligned} \tag{2}$$

8.5.1.5 Justificación metodológica

De esta forma, este primer modelo cumple una doble función de carácter fundamental para el desarrollo de los modelos futuros, ya que en primer lugar ejecuta un punto de referencia (benchmark) que permite comparar el desempeño de los modelos más complejos a desarrollar, y también permite obtener las primeras métricas de interpretabilidad referido a la entrega de los estimadores, cuya comprensión es más sencilla debido a la obtención de resultados en términos de semi-elasticidades, lo cual aporta evidencia empírica sobre la magnitud y signo de los efectos de cada atributo en el precio de la vivienda.

8.5.2. Regresión lineal con Índices compuestos (`OLS_IDX`)

8.5.2.1 Objetivo del modelo

El segundo modelo corresponde a una regresión lineal múltiple (vía `OLS`) con la elaboración e inclusión de índices compuestos y tiene por objetivo incorporar explícitamente la dimensión territorial del mercado de accesos y dotación de servicios, mediante agregaciones conceptuales de variables externas, manteniendo al mismo tiempo una especificación parsimoniosa e interpretables de los efectos. En concreto, este modelo busca:

- Cuantificar el efecto marginal del entorno urbano sobre el logaritmo del precio de oferta en UF una vez controladas las características internas del inmueble y la localización zonal. Para ello se integran índices compuestos que sintetizan dimensiones clave del contexto.

- Reducir dimensionalidad y mitigar colinealidad sin sacrificar significado económico, dando sentido a que en lugar de ingresar numerosas variables de proximidad una a una (altamente correlacionadas entre sí), se agrupan en índices normalizados con un fundamento teórico-empírico (no algorítmico). A diferencia de un PCA, que prioriza varianza explicada y produce combinaciones lineales de interpretación menos directa, aquí los índices se definen por coherencia conceptual e intereses explicativos (respaldados por índices de correlación y fuentes teóricas), facilitando la lectura de resultados y su discusión sustantiva.

- Ofrecer una especificación de referencia “urbano-explicativa” complementaria al OLS base, puesto que este segundo modelo permite comparar cómo cambia la magnitud y significancia de los atributos internos cuando el contexto urbano en conglomerados entra explícitamente en la ecuación; y, a su vez, medir la relevancia económica de cada dimensión del entorno en

términos de semi-elasticidades sobre el precio.

8.5.2.2 Justificación de índices

La principal innovación del Modelo 2 es la incorporación de índices compuestos, los cuales se diseñaron para sintetizar múltiples variables de proximidad y dotación de servicios en dimensiones conceptualmente coherentes. Estos índices no se derivan de un procedimiento algorítmico como el análisis de componentes principales (PCA), sino que se definieron con un criterio teórico-empírico, considerando tanto la correlación observada entre las variables como la lógica económica y urbana subyacente. El objetivo fue mantener la parsimonia del modelo, reduciendo colinealidad y número de regresores, al mismo tiempo que se preserva la interpretabilidad sustantiva.

En el ámbito práctico, cada índice se construyó a partir de variables previamente transformadas (en logaritmos, escalas de proximidad o estandarización tipo Z-score) y luego agregadas mediante promedios y ponderaciones normalizados, lo que asegura comparabilidad entre dimensiones heterogéneas. A continuación, se detallan los índices definidos:

- Índice de Salud y Comercio (`IDX_SALCOM`): combina variables relativas a la cercanía de centros de salud y servicios comerciales (farmacias, supermercados, centros comerciales), entendidos como satisfactores de necesidades básicas y cotidianas.

- Índice de Áreas Verdes (`IDX_VERDE`): agrupa la proximidad a parques y plazas, en tanto amenidades ambientales que mejoran la calidad de vida y son reconocidas como un factor de plusvalía inmobiliaria.

- Índice de Amenidades (`IDX_AMENITIES`): integra características asociadas a dotaciones complementarias de la vivienda, tales como piscina, quincho (que a su vez incluye en la

mayoría de casos la prescencia de terraza y/o jardín) o bodega, que representan atributos de diferenciación y confort.

- Índice de Educación y Transporte (`IDX_EDUC_TRANS`): integra variables de proximidad tanto a la red de transporte público (paraderos de buses y estaciones de metro) como a establecimientos educacionales de distintos niveles (básico, medio y superior). La construcción conjunta de este índice responde al sentido urbano de complementariedad entre movilidad y educación: la accesibilidad al transporte se torna especialmente relevante en zonas donde se concentran recintos educativos, pues facilita el desplazamiento diario de estudiantes y docentes. De esta manera, el índice captura simultáneamente el valor asociado a la conectividad y a la oferta educativa cercana, reflejando su efecto combinado sobre la valorización de las viviendas en zonas de alta urbanización.

De esta forma, la construcción de estos índices responde a tres justificaciones centrales:

1. Teórica, al reflejar dimensiones urbanas que la literatura reconoce como determinantes del precio de la vivienda.

2. Metodológica, al reducir la carga de variables altamente correlacionadas y facilitar la estabilidad de los coeficientes.

3. Práctica, al proporcionar medidas sintéticas sencillas de interpretar y discutir en términos de política pública y planificación urbana.

8.5.2.3 Preparación de datos

La preparación de los datos para el segundo modelo supuso procesos adicionales respecto al primer modelo, ya que implicó en la modificación de las dummies de zonificación y también

la construcción y validación de los índices compuestos. El flujo de trabajo se articuló, en primer lugar, la incorporación de nuevas divisiones de sectores, con el objetivo de encontrar el balance entre generalización y especificación de la localización de las viviendas. Luego se trabajó en la transformación y estandarización de variables base, la verificación de consistencia y control de colinealidad, y finalmente la depuración de la base final para la estimación.

8.5.2.3.1 Variables de zonificación

Siguiendo la línea del modelo anterior de no incorporar variables categóricas nominales con exceso de niveles como lo son las comunas y/o barrios al que pertenece cada vivienda de estudio, y tomando en cuenta a su vez la necesidad de mejorar las variables de localización en mayor especificación, con el intento de encontrar grupos representativos y más homogéneos, es que se realiza para este modelo la construcción de 6 variables de zonas que conglomeran cierto número de comunas, cuya función principal es encontrar el balance entre trabajar con una cantidad de variables de zonificación acotada para no caer en un número excesivo de parámetros explicativos (reduciendo los grados de libertad) y que dichas variables sean lo suficientemente representativas respecto a los sectores con características similares que se intentan captar.

Conforme lo anterior, es que en este modelo se trabaja con las siguientes variables *dummies* que representan la zona en la que se encuentra la vivienda de estudio:

- ***Dummy_zona_norte_poniente*:** variable binaria codificada con (1) si la vivienda de estudio se encuentra en la zona norponiente de Santiago y (0) si no lo está.
- ***Dummy_zona_norte_centro*:** variable binaria codificada con (1) si la vivienda de estudio se encuentra en la zona norte-centro de Santiago y (0) si no lo está.

- ***Dummy_zona_norte_oriente*:** variable binaria codificada con (1) si la vivienda de estudio se encuentra en la zona nororiente de Santiago y (0) si no lo está.

- ***Dummy_zona_sur_poniente*:** variable binaria codificada con (1) si la vivienda de estudio se encuentra en la zona surponiente de Santiago y (0) si no lo está.

- ***Dummy_zona_sur_centro*:** variable binaria codificada con (1) si la vivienda de estudio se encuentra en la zona sur-centro de Santiago y (0) si no lo está.

- ***Dummy_zona_sur_oriente*:** variable binaria codificada con (1) si la vivienda de estudio se encuentra en la zona suroriente de Santiago y (0) si no lo está.

De esta forma, la categorización de cada vivienda de estudio a una zona en particular, se define mediante la propia ubicación de la comuna en el mapa de la Región Metropolitana [4], donde en la siguiente tabla se puede ver a que zona pertenece cada comuna en particular:

Tabla 3: Clasificación de comunas según zona geográfica

Comuna	Zona	Comuna	Zona
Cerrillos	Sur-poniente	Lo Espejo	Sur-poniente
Cerro Navia	Norte-poniente	Lo Prado	Norte-poniente
Conchalí	Norte-centro	Macul	Sur-oriente
El Bosque	Sur-centro	Maipú	Sur-poniente
Estación Central	Norte-poniente	Ñuñoa	Norte-oriente
Huechuraba	Norte-centro	Pedro Aguirre Cerda	Sur-centro
Independencia	Norte-centro	Peñalolén	Sur-oriente
La Cisterna	Sur-centro	Providencia	Norte-oriente
La Florida	Sur-oriente	Puente Alto	Sur-oriente
La Granja	Sur-centro	Pudahuel	Norte-poniente
La Pintana	Sur-centro	Quilicura	Norte-poniente
La Reina	Norte-oriente	Quinta Normal	Norte-poniente
Las Condes	Norte-oriente	Recoleta	Norte-centro
Lo Barnechea	Norte-oriente	Renca	Norte-poniente
San Joaquín	Sur-centro	Santiago	Norte-centro
San Miguel	Sur-centro	Vitacura	Norte-oriente
San Ramón	Sur-centro		

8.5.2.3.2 Construcción de índices de acceso

Las variables externas de acceso a servicios obtenidas del portal inmobiliario fueron sometidas a distintos tratamientos de normalización con el fin de hacerlas comparables entre sí. Conforme la naturaleza de las mismas, se aplicaron tres tipos de transformaciones:

- Logaritmos: en distancias/proximidades, para reflejar rendimientos decrecientes (la diferencia entre vivir a 200 y 500 metros de un servicio en particular es más relevante que entre 2 y 2,3 km).
- Proximidad invertida: se calcula el réciproco de la transformación anterior, lo cual permite ahora penalizar la lejanía, de modo que valores más altos indicaran mejor accesibilidad.

- Z-scores: se estandarizó respecto a la media y desviación estándar de la muestra, asegurando que cada variable contribuyera de forma equilibrada a los índices, sin que magnitudes absolutas sesgaran el resultado.

Posteriormente, las variables afines se agruparon en índices compuestos mediante un promedio simple o ponderaciones precisas de los valores estandarizados, generando así medidas agregadas de transporte, educación, salud/comercio, áreas verdes y amenidades.

Antes de consolidar la base de datos a trabajar para este modelo, se verificó la consistencia de los índices revisando rangos de valores, distribución y correlaciones entre ellos. Este paso es fundamental para asegurar que cada índice representara una dimensión distinta del entorno urbano y no una simple duplicación de información.

8.5.2.3.3 Transformación de variable Antigüedad

Con el fin de capturar la posible no linealidad en la relación entre la antigüedad de la vivienda y su precio, esta variable se incorporó en dos términos: uno lineal y uno cuadrático. Para evitar la alta colinealidad que suele producirse entre la antigüedad y su cuadrado, se procedió a centrar la variable en torno a su media muestral (24 años), construyendo la variable antigüedad centrada y, posteriormente, su término cuadrático. Con esta transformación se tienen dos ventajas metodológicas:

- Mejora la estabilidad estadística del modelo, reduciendo los valores de colinealidad (VIF) que se tengan con la variable.

- Otorgar una interpretación más realista del intercepto y de la pendiente lineal, ya que los coeficientes se interpretan en relación a una vivienda con edad promedio en la muestra, en

lugar de referirse a una vivienda de antigüedad cero, situación poco representativa en el mercado.

De esta manera, el modelo permite identificar si la depreciación asociada al paso del tiempo sigue un patrón lineal, o bien si existe una curvatura que refleje pérdida acelerada de valor en los primeros años y/o recuperación relativa en edades más avanzadas.

Finalmente, el proceso culminó en la construcción de la base `BBDD_OLS_IDX`, que contiene 462 observaciones y un total de 16 variables. Esta base consolida las características internas, zonales y los índices compuestos en un formato parsimonioso y coherente para la estimación del Modelo `OLS_IDX`. La decisión de trabajar con esta versión depurada garantiza la comparabilidad con el Modelo `OLS_BASE`, pero con la ventaja de incorporar dimensiones urbanas agregadas de manera sistemática.

8.5.2.4 Exclusión de variables

A través de la realización de múltiples regresiones exploratorias (etapa que permitió evaluar la estabilidad de los coeficientes y la redundancia entre regresores) se decidió no incorporar en la especificación del Modelo `OLS_IDX` ciertas variables internas altamente correlacionadas, tales como superficie útil, número de dormitorios, número de estacionamientos o número de pisos. Los resultados obtenidos mostraron que dichas características guardan una relación estrecha con variables ya retenidas, principalmente superficie total y número de baños, lo que genera redundancia y potencial colinealidad. Frente a ello, se optó por una especificación más parsimoniosa, privilegiando aquellas variables internas con mayor poder explicativo y representatividad. Esta simplificación no implica pérdida sustantiva de información, pues los atributos omitidos se encuentran en gran medida capturados por las variables seleccionadas.

8.5.2.5 Variables explicativas incluidas

Conforme lo anterior, el listado de variables explicativas a formar parte de este modelo denominado `OLS_IDX` son las siguientes:

- Dummies de Zonificación (dummy 0/1).
- Vivienda en Condominio cerrado (dummy 0/1).
- Superficie de terreno (m^2).
- Antigüedad centrada.
- Antigüedad centrada al cuadrado.
- Número de baños.
- Índices de acceso (transporte, educación, salud, comercio, áreas verdes)

En conjunto, estas variables permiten evaluar simultáneamente los efectos de atributos internos, condiciones de localización y calidad del entorno urbano, ofreciendo una visión más amplia de la formación del precio de la vivienda respecto al modelo `OLS_BASE`.

De este forma, la expresión general para el modelo `OLS_IDX` queda representada de la siguiente forma:

$$
\begin{aligned}
\ln(Precio_i) = \beta_0 &+ \beta_1 SupTotal_i + \beta_2 Banos_i \\
&+ \beta_3 AntiguedadC_i + \beta_4 AntiguedadC_i^2 \\
&+ \beta_5 D_i^{CondCerrado} \\
&+ \sum_{z=1}^{5} \gamma_z D_{iz}^{Zona} \\
&+ \delta_1 IDX\_EDUC\_TRANS_i \\
&+ \delta_2 IDX\_SALCOM_i + \delta_3 IDX\_VERDE_i \\
&+ \delta_4 IDX\_AMENITIES_i \\
&+ \varepsilon_i
\end{aligned}
\tag{3}
$$

8.5.2.6 Justificación metodológica

La inclusión del Modelo `OLS_IDX` responde a la necesidad de avanzar hacia una especificación más parsimoniosa y, al mismo tiempo, más representativa de los determinantes del precio de la vivienda. A diferencia del primer modelo, que incorporaba un conjunto amplio de variables individuales, este modelo integra índices compuestos del entorno urbano que permiten capturar de manera sintética dimensiones clave como accesibilidad al transporte, cercanía a establecimientos educacionales, disponibilidad de servicios de salud y comercio, acceso a áreas verdes y amenidades residenciales. Esta decisión metodológica reduce la redundancia entre regresores, evita la sobrecarga de variables altamente correlacionadas y facilita la interpretación de los resultados en términos de factores urbanos de mayor escala.

8.5.3. Modelos de aprendizaje automático

Con el objetivo de complementar los modelos de regresión lineal múltiple, se incorporaron al análisis dos algoritmos de aprendizaje automático supervisado `Random Forest` y `XGBoost`. La motivación principal de esta decisión metodológica es doble. En primer lugar, permite comparar la capacidad predictiva de estos modelos con la de los modelos de regresión, evaluando si el uso de técnicas avanzadas de autoaprendizaje mejora sustantivamente la precisión en la estimación de precios de la vivienda. En segundo lugar, posibilita contrastar la consistencia en la identificación de variables relevantes, verificando si los factores más influyentes señalados por los modelos explicativos tradicionales coinciden con los resaltados por algoritmos que capturan relaciones no lineales y complejas.

A diferencia de los modelos anteriores, que priorizan la interpretabilidad de los coeficientes y la identificación de relaciones lineales entre las variables, los algoritmos de aprendizaje automático destacan por su habilidad para detectar patrones de alta dimensionalidad y relaciones no lineales, optimizando la predicción sin necesidad de transformaciones explícitas de los regresores. Esta comparación metodológica permite discutir el clásico trade-off entre parsimonia explicativa e idoneidad predictiva, aportando una perspectiva más amplia y actualizada en el estudio de precios hedónicos.

8.5.3.1 Random Forest

8.5.3.1.1 Descripción del algoritmo

Random Forest (RF) es un método de aprendizaje automático basado en ensambles, que combina un gran número de árboles de decisión entrenados sobre submuestras aleatorias de los

datos (técnica bootstrap) y con una selección aleatoria de variables en cada división de los nodos. Esta doble aleatorización genera diversidad entre los árboles, y el resultado final se obtiene como el promedio de las predicciones individuales (en problemas de regresión), lo que reduce significativamente la varianza respecto a un solo árbol y mejora la capacidad de generalización.

A diferencia de los modelos de regresión lineal, Random Forest no requiere suponer linealidad entre la variable dependiente y las variables explicativas, ni homocedasticidad, ni independencia estricta de errores. Por el contrario, es capaz de capturar relaciones no lineales, interacciones complejas y efectos marginales decrecientes de forma automática, sin necesidad de especificarlos a inicialmente. Esto lo convierte en un algoritmo especialmente valioso en contextos como el mercado inmobiliario, donde los precios de las viviendas dependen de múltiples factores internos y externos que pueden relacionarse de manera no lineal.

8.5.3.1.2 Preparación de datos

Antes de entrenar el modelo de Random Forest, es necesario preparar adecuadamente la base de datos, asegurando que la información utilizada fuese consistente y que estuviera en un formato compatible con el algoritmo. Este proceso incluyó varias etapas que se detallan a continuación:

- **Variable dependiente:** La variable objetivo del modelo es el precio de la vivienda expresado en UF, pero transformado a escala logarítmica ($\ln(\text{precio})$). Esta transformación cumple dos propósitos. En primer lugar, reducir la asimetría (sesgo) de la distribución del precio, que en su forma original presenta valores muy altos que podrían distorsionar el ajuste del modelo. Y en segundo lugar, facilitar la interpretación relativa de los resultados, ya que en la escala logarítmica los efectos de las variables explicativas se interpretan en términos de variaciones

porcentuales aproximadas en el precio.

- **Variables explicativas:** Se consideraron todas las variables internas recolectadas (como superficies, número de dormitorios, baños, antigüedad, entre otras) como variables externas asociadas al entorno urbano (índices de accesibilidad a transporte, educación, salud, áreas verdes, equipamiento comercial, etc. construidos en el modelo anterior). Además, para la realización de una comparativa más justa entre modelos, el dataset empleado fue el ya procesado en la fase de construcción de índices, complementado con algunas variables adicionales derivadas de ingeniería de atributos.

8.5.3.1.3 Ingeniería de atributos

Con el fin de enriquecer el poder predictivo sin generar "filtraciones de información" (es decir, sin utilizar el precio para construir nuevas variables), se incluyeron transformaciones adicionales que el modelo es capaz de generar, destacando las siguientes:

- **Densidades habitacionales:** como la razón entre dormitorios y superficie útil, o entre baños y superficie útil. Estas variables permiten capturar la "intensidad de uso" del espacio habitable, un factor que puede reflejar comodidad percibida y, por ende, valor de mercado.

- **Interacciones entre superficie e índices urbanos:** por ejemplo, superficie total con índices de accesibilidad. Esto permite que el modelo distinga no solo si un barrio es accesible, sino también cómo el tamaño de la vivienda amplifica o atenúa el valor asociado a esa accesibilidad.

8.5.3.1.4 Tratamiento de valores atípicos (outliers)

Para una mejor comprensión del comportamiento de las colas, se trabajaron dos versiones paralelas del dataset:

- **Full dataset (sin recorte):** incluye todas las observaciones disponibles, reflejando la variabilidad total del mercado.
- **Trimmed dataset (recortado):** elimina el 1 % más bajo y el 1 % más alto de los precios (percentiles 1 y 99). Esta estrategia busca controlar el impacto de valores extremadamente bajos o altos que podrían sesgar la estimación, manteniendo una muestra representativa pero más estable.

8.5.3.1.5 Parametrización y procesamiento de datos

La base fue dividida en un conjunto de entrenamiento (80 %) y un conjunto de prueba (20 %), utilizando siempre la misma semilla aleatoria para garantizar reproducibilidad. El conjunto de entrenamiento se utiliza para ajustar el modelo y explorar los parámetros óptimos, mientras que el conjunto de prueba se reserva para evaluar el desempeño en datos no vistos, lo que permite estimar la capacidad real de generalización del modelo.

Finalmente, el flujo de preparación se implementó en un pipeline, lo que asegura que cada paso se ejecute en el orden correcto y de manera consistente:

- Las variables numéricas fueron tratadas mediante imputación por la mediana, lo que evita que valores faltantes invaliden la estimación.
- Las variables categóricas fueron imputadas por la moda y luego codificadas mediante One-Hot Encoding (OHE), que transforma categorías en variables binarias (0/1) sin asignarles un

valor numérico arbitrario que pudiera inducir relaciones inexistentes.

- Todo este proceso se encapsula en un transformador automático, evitando errores humanos y asegurando que el mismo tratamiento se aplique tanto a entrenamiento como a prueba.

Una vez preparado el conjunto de datos, se procede a la especificación y ajuste del modelo Random Forest. Para ello se utilizó el estimador `RandomForestRegressor` de la librería *scikit-learn*, configurado con bootstrap activado, semilla fija (`random_state = 42`) y múltiples árboles para asegurar estabilidad.

El proceso de calibración se realiza mediante una búsqueda aleatoria de hiperparámetros con `RandomizedSearchCV`. Esta técnica explora combinaciones dentro de un espacio definido, equilibrando cobertura y eficiencia computacional. El espacio de búsqueda incluyó parámetros clave:

- Número de árboles (`n_estimators`): 800, 1200, 1600, 2000.
- Profundidad máxima (`max_depth`): 6, 8, 12, 16, 20, 24 o sin límite (None).
- Mínimo de observaciones por división (`min_samples_split`): 2, 5, 10, 20, 40.
- Mínimo de observaciones por hoja (`min_samples_leaf`): 1, 2, 4, 8, 12.
- Número máximo de variables por nodo (`max_features`): *sqrt*, 0.3, 0.5, 0.7, 0.9.
- Bootstrap: activado (`True`).

En total, se evalúan 120 combinaciones aleatorias bajo un esquema de validación cruzada (5 pliegues, repetición simple), seleccionando como mejor configuración aquella con menor

`Root Mean Squared Error (RMSE)`, métrica que penaliza más fuertemente las predicciones alejadas del valor real, algo particularmente relevante en el mercado inmobiliario donde existen propiedades de muy alto valor.

Finalmente, se calculan las medidas de importancia de variables, donde se evidencia la importancia por permutación, que evalúa el aumento en el error al desordenar una variable y constituye la referencia principal y la importancia por impureza, entregada nativamente por el algoritmo como complemento interpretativo. De esta forma, el procedimiento integrado asegura que el `Random Forest` se ajuste con datos preparados de manera rigurosa, con un proceso automatizado y reproducible, y con parámetros elegidos de forma sistemática en lugar de arbitraria.

8.5.3.1.6 Justificación metodológica

La inclusión del modelo `Random Forest` responde a la necesidad de comparar un método clásico de regresión lineal con un algoritmo de aprendizaje automático capaz de captar relaciones no lineales e interacciones complejas entre variables, sin requerir supuestos estrictos de linealidad u homocedasticidad.

Su estructura basada en el ensamble de múltiples árboles de decisión reduce la varianza y el riesgo de sobreajuste, lo que lo convierte en un predicctor robusto y competitivo frente a escenarios con alta heterogeneidad, como es el caso del mercado inmobiliario.

Además, `Random Forest` ofrece ventajas prácticas relevantes como fuertes índices de robustez frente a valores extremos y ruido en los datos, capacidad de manejar gran cantidad de variables sin preselección estricta y consigo la estimación de la importancia de variables, lo que permite mantener un componente interpretativo útil para el análisis aplicado.

En este sentido, su uso no pretende reemplazar a los modelos `OLS_BASE` y `OLS_IDX`, sino complementar el análisis, contrastando tanto el poder explicativo como el predictivo de ambos enfoques, y evaluando si un método de aprendizaje automático aporta mejoras sustantivas en la estimación del precio de viviendas.

8.5.3.2 XGBoost

8.5.3.2.1 Descripción del algoritmo

El modelo `XGBoost` (Extreme Gradient Boosting) es un algoritmo de aprendizaje supervisado basado en árboles de decisión, que utiliza la técnica de *boosting por gradiente*. A diferencia de los modelos de ensamble como `Random Forest` (donde múltiples árboles se construyen en paralelo de forma independiente), el boosting construye árboles de manera secuencial, donde cada nuevo árbol intenta corregir los errores cometidos por el conjunto de árboles anteriores.

La lógica central es que el modelo comienza con una predicción inicial, y a continuación ajusta un primer árbol sobre los residuales (la diferencia entre los valores reales y los predichos). En iteraciones sucesivas, cada nuevo árbol se entrena para explicar lo que los árboles anteriores aún no han logrado capturar. De este modo, el modelo se va "puliendo" progresivamente, mejorando su capacidad predictiva.

Entre sus principales ventajas se encuentran:

- **Flexibilidad:** permite incorporar regularización (regresión L1/Lasso y L2/Ridge) para controlar el sobreajuste.
- **Precisión:** logra altos niveles de ajuste al capturar relaciones no lineales y complejas interacciones entre variables.

- **Eficiencia:** gracias a optimizaciones como el método de histograma (`tree_method = hist`), puede manejar bases de datos de tamaño considerable con tiempos de entrenamiento reducidos.

- **Importancia de variables:** entrega medidas de importancia de cada predictor (por ganancia, cobertura o frecuencia), lo que otorga interpretabilidad parcial en un modelo inherentemente más complejo que regresión lineal.

De modo que en este estudio, `XGBoost` se utilizó con el objetivo de evaluar su poder explicativo y predictivo frente a los modelos vía `OLS` y a `Random Forest`, aprovechando su capacidad de modelar relaciones complejas en la formación de precios inmobiliarios.

8.5.3.2.2 Preparación y partición de datos

El procedimiento seguido para la construcción del modelo `XGBoost` se basó en la misma base de datos construida para `Random Forest`, manteniendo la misma lógica de explicar el logaritmo del precio en función de las variables antes seleccionadas, garantizando consistencia en la comparación.

- **Variable dependiente:** La variable objetivo del modelo es el precio de la vivienda expresado en UF, pero transformado a escala logarítmica ($\ln(\text{precio})$), cuya transformación cumple dos propósitos. En primer lugar, reducir la asimetría (sesgo) de la distribución del precio, y en segundo lugar, facilitar la interpretación relativa de los resultados, ya que en la escala logarítmica los efectos de las variables explicativas se interpretan en términos de variaciones porcentuales aproximadas en el precio.

- **Variables explicativas:** Se consideran todas las variables internas recolectadas (como su-

perficies, número de dormitorios, baños, antigüedad, entre otras) y las variables externas asociadas al entorno urbano (índices de accesibilidad a transporte, educación, salud, áreas verdes, equipamiento comercial, etc. construidos en el modelo `OLS_IDX`), para la realización de una comparativa más justa entre modelos.

En cuanto a la partición de los datos, estos fueron divididos en un conjunto de para entrenamiento (80 %) y un conjunto de prueba (20 %), utilizando una semilla aleatoria fija para asegurar la reproducibilidad. El conjunto de entrenamiento se destina a la estimación y búsqueda de parámetros óptimos, mientras que el conjunto de prueba se reserva para evaluar el desempeño en datos no observados, permitiendo estimar la capacidad de generalización del modelo.

8.5.3.2.3 Parametrización y procesamiento de datos

El procesamiento de los datos se organiza en un pipeline que aplica de forma automática y ordenada cada transformación:

- **Variables numéricas:** se imputan (rellenan) con la mediana. Esto significa que cuando falta un dato numérico, en vez de dejarlo vacío se reemplaza por el valor central de la distribución. Esta técnica es más robusta que la media frente a valores extremos (outliers).

- **Variables categóricas:** se imputan con la moda (categoría más frecuente) y luego se codifican con `One-Hot Encoding` (OHE). Este procedimiento convierte cada categoría en una columna binaria (0/1), evitando que el modelo interprete de manera errónea que existe un orden numérico entre ellas.

- **Ingeniería de atributos:** se crean nuevas variables a partir de las ya existentes para mejorar la capacidad explicativa del modelo, las cuales son, densidad de dormitorios y de baños por

metro cuadrado útil, lo que refleja la "intensidad de uso" de la vivienda e interacciones entre superficies (total y útil) y distintos índices urbanos (accesibilidad a servicios cercanos), lo que permite capturar relaciones más complejas entre el entorno y el precio.

Estas transformaciones amplían la información disponible y ayudan al algoritmo a detectar patrones más sofisticados.

En lo que respecta al proceso de parametrización, el algoritmo `XGBoost` se configura con un `XGBRegressor`, que es una técnica basada en árboles de decisión ensamblados de manera secuencial para minimizar el error de predicción. El método de crecimiento de árboles es *hist*, que utiliza histogramas de frecuencias para dividir los nodos de forma eficiente y reducir el tiempo de cómputo.

Finalmente, para búsqueda de hiperparámetros precisos, esto proceso se realiza mediante `RandomizedSearchCV`, un procedimiento que prueba combinaciones aleatorias de parámetros dentro de un espacio definido. Para garantizar confiabilidad, se aplica validación cruzada repetida (5 particiones de entrenamiento y validación, repetidas una vez). Esto significa que el modelo se entrena y valida varias veces con distintos cortes de los datos, lo que entrega una evaluación más robusta y menos dependiente de un solo particionamiento. Con ello, el espacio de búsqueda incluye los siguientes hiperparámetros clave:

- Número de árboles (`n_estimators`): 800, 1200, 1600, 2000.
- Tasa de aprendizaje (`learning_rate`): 0.005, 0.01, 0.02, 0.03.
- Profundidad máxima de los árboles (`max_depth`): 4, 5, 6, 8, 10.
- Peso mínimo en hojas (`min_child_weight`): 1, 3, 5, 8, 12, 15.

- Reducción mínima de pérdida para dividir un nodo (`gamma`): 0, 0.1, 0.2, 0.3, 0.5.
- Submuestreo de observaciones por árbol (`subsample`): 0.5, 0.7, 0.9, 1.0.
- Submuestreo de variables por árbol (`colsample_bytree`): 0.5, 0.7, 0.9, 1.0.
- Submuestreo de variables por nivel (`colsample_bylevel`): 0.6, 0.8, 1.0.
- Submuestreo de variables por nodo (`colsample_bynode`): 0.6, 0.8, 1.0.
- Regularización L1 (`reg_alpha`): 0, 0.01, 0.1, 0.5, 1.0, 5.0.
- Regularización L2 (`reg_lambda`): 0.1, 1.0, 2.0, 5.0, 10.0.
- Número máximo de bins (`max_bin`): 128, 256, 512.
- Estrategia de crecimiento del árbol (`grow_policy`): *depthwise*, *lossguide*.
- Número máximo de hojas (`max_leaves`): 0, 16, 32, 64.

8.5.3.2.4 Justificación del modelo

La elección de `XGBoost` dentro del conjunto de modelos a evaluar responde a dos razones principales. En primer lugar, se trata de uno de los algoritmos de aprendizaje automático más utilizados en la actualidad por su capacidad para combinar alto poder predictivo con eficiencia computacional. Su estrategia de *gradient boosting* permite construir árboles de decisión de manera secuencial, corrigiendo en cada iteración los errores cometidos en las anteriores, lo que se traduce en una mayor precisión y flexibilidad frente a relaciones no lineales y posibles interacciones entre variables.

En segundo lugar, su inclusión en el análisis cumple un rol comparativo respecto de los modelos de regresión lineal múltiple (`OLS`). Mientras los modelos `OLS` priorizan la transparencia interpretativa y la parsimonia, `XGBoost` se orienta principalmente a la predicción. De esta manera, confrontar sus resultados con los obtenidos en regresión permite valorar hasta qué punto el modelo lineal logra aproximarse a los patrones reales presentes en los datos, y en qué medida un modelo no lineal y de mayor complejidad puede aportar mejoras significativas en términos predictivos.

Además, la robustez de `XGBoost` frente a valores atípicos, su capacidad para incorporar regularización y su versatilidad en la selección de hiperparámetros lo convierten en una herramienta idónea para este ejercicio. No se busca únicamente obtener la mejor predicción posible, sino también contrastar la utilidad de índices y variables explicativas en un marco metodológico diverso, que combine técnicas econométricas tradicionales con enfoques modernos de aprendizaje automático.

8.5.4. Diagnóstico y evaluación de modelos

La evaluación de los modelos no se limita a reportar sus coeficientes o predicciones, sino que requiere aplicar un conjunto de métricas de desempeño, pruebas estadísticas y diagnósticos gráficos que permitan valorar su ajuste, capacidad predictiva y validez de los supuestos.

En este apartado se sistematizan los procedimientos utilizados de manera transversal para los distintos enfoques (regresiones `OLS` y modelos de aprendizaje automático), agrupando tanto las métricas de predicción como los test econométricos y las herramientas de diagnóstico. Esto asegura una base comparativa homogénea entre modelos, al mismo tiempo que se respetan las particularidades metodológicas de cada técnica.

8.5.4.1 Diagnóstico de supuestos del modelo

- **Normalidad de los Residuos (Jarque-Bera y Shapiro-Wilk):**

  La normalidad de los residuos es un supuesto importante en modelos de regresión lineal, ya que respalda la validez de los intervalos de confianza y las pruebas de significancia. El test de Jarque-Bera evalúa la desviación de la distribución de los residuos respecto a la normalidad considerando su asimetría y curtosis, mientras que el test de Shapiro-Wilk contrasta directamente si los residuos provienen de una distribución normal. Ambos se complementan con gráficos como el QQ-plot.

- **Heterocedasticidad (Breusch-Pagan y White):**

  La Heterocedasticidad implica que la varianza de los residuos no es constante a lo largo de los valores ajustados. Cuando este supuesto se cumple, las inferencias estadísticas pueden ser ineficientes o sesgadas. El test de Breusch-Pagan detecta patrones lineales de heterocedasticidad, mientras que el test de White es más general y no depende de una forma funcional específica.

- **Linealidad y Forma Funcional (RESET de Ramsey):**

  El supuesto de linealidad establece que la relación entre las variables explicativas y la variable dependiente está correctamente especificada en forma lineal. El test RESET de Ramsey contrasta si se han omitido variables relevantes o formas funcionales no lineales al añadir potencias de los valores ajustados al modelo. Una significancia elevada sugiere que el modelo requiere transformaciones o términos adicionales para captar mejor la relación.

8.5.4.2 Multicolinealidad

- **Factor de Inflación de Varianza (VIF):**

  El VIF mide cuánto se incrementa la varianza de un coeficiente estimado debido a la colinealidad con otros predictores. Valores altos indican que una variable está altamente correlacionada con otras, lo que dificulta distinguir su efecto individual y puede generar estimaciones inestables. Generalmente, valores por encima de 10 se consideran problemáticos.

- **Matriz de correlaciones**

  La matriz de correlaciones cuantifica la relación lineal entre pares de variables explicativas. Aunque no es un diagnóstico tan sofisticado como el VIF, permite identificar patrones de asociación elevados que pueden anticipar problemas de colinealidad. Se utiliza como herramienta exploratoria para decidir qué variables incluir, transformar o combinar en el modelo.

8.5.4.3 Métricas de desempeño predictivo

- **Error absoluto Medio (MAE):**

  El MAE mide el promedio de los errores absolutos entre los valores reales y los predichos. Es decir, indica en cuántas unidades (en este caso, UF) se equivoca el modelo en promedio, sin importar si el error es hacia arriba o hacia abajo. Su ventaja es que es fácil de interpretar y no se ve afectado de forma desproporcionada por valores extremos.

- **Raíz del Error Cuadrático Medio (RMSE):**

  El RMSE corresponde a la raíz cuadrada del promedio de los errores al cuadrado. A diferencia del MAE, penaliza más fuertemente los errores grandes, por lo que es muy útil para identificar modelos sensibles a outliers o predicciones con desviaciones muy amplias. En

contextos como el inmobiliario, refleja la magnitud típica del error en UF, destacando los casos de sobreestimación o subestimación más graves.

- **Error Porcentual Absoluto Medio (MAPE):**

  El MAPE expresa el error en términos porcentuales respecto al valor real, mostrando en qué porcentaje se equivoca el modelo en promedio. Esta métrica permite comparar errores de manera relativa, lo cual es muy útil en mercados donde los precios tienen escalas muy distintas.

- **Coeficiente de Determinación (R^2)**

  El R^2 mide la proporción de la variación total de la variable dependiente que es explicada por el modelo. Un valor cercano a 1 implica que el modelo captura gran parte de la variabilidad de los precios, mientras que valores bajos o negativos indican que el modelo predice peor que un promedio simple. Es una medida de bondad de ajuste, más interpretativa en modelos lineales que en modelos de autoaprendizaje.

- **Coeficiente de Determinación Ajustado (R^2 ajustado)**

  El R^2 ajustado modifica el R^2 clásico incorporando una penalización por el número de predictores utilizados. De esta forma, evita que el indicador aumente de manera artificial al añadir variables sin valor explicativo real. Es especialmente útil para comparar modelos de distinta complejidad y evaluar parsimonia.

8.5.4.4 Criterios de ajuste y parsimonia

- **Criterio de Información de Akaike (AIC):**

  El AIC es una medida que evalúa la calidad de un modelo estadístico combinando el ajuste

a los datos con la complejidad del modelo. Penaliza a los modelos con un número excesivo de parámetros, favoreciendo especificaciones más parsimoniosas. Un valor menor de AIC indica un mejor equilibrio entre ajuste y simplicidad.

- **Criterio de Información Bayesiano (BIC):**

  El BIC es similar al AIC, pero aplica una penalización más estricta al número de predictores. Está orientado a seleccionar modelos más simples cuando la muestra es grande, y se interpreta como una medida de evidencia relativa en favor de un modelo frente a otros candidatos. Valores menores reflejan modelos preferibles.

- **Test de Durbin-Watson:**

  El estadístico de Durbin-Watson se utiliza para detectar autocorrelación de primer orden en los residuos de un modelo de regresión. Aunque es más relevante en series de tiempo, también puede aplicarse en datos transversales para verificar que los residuos sean independientes. Un valor cercano a 2 indica ausencia de autocorrelación, mientras que valores alejados de ese punto sugieren dependencia entre los errores.

- **Factor de *smearing*:**

  El factor de *smearing* es una corrección utilizada cuando la variable dependiente se modela en escala logarítmica. Permite retransformar las predicciones al nivel original evitando sesgos en la estimación del valor esperado. Se calcula a partir de los residuos y asegura que las predicciones en la escala natural sean insesgadas.

8.5.4.5 Validación y robustez

- **Validación Cruzada *K-Fold*:**

  La validación cruzada *K-Fold* consiste en dividir la base de datos en *k* subconjuntos o "folds". El modelo se entrena repetidamente en *k-1* y se valida en el restante, rotando hasta que cada fold haya sido utilizado como prueba. Este procedimiento permite estimar el desempeño del modelo de manera más robusta, reduciendo la dependencia de una única partición de los datos.

- **Conjunto de Entrenamiento y Prueba (Holdout):**

  La estrategia holdout divide los datos en dos subconjuntos: uno para ajustar el modelo (entrenamiento) y otro reservado exclusivamente para evaluar su desempeño (prueba). Esto proporciona una medida realista de la capacidad de generalización del modelo a datos no vistos, evitando la sobreestimación del ajuste que se obtiene al evaluar en la misma muestra usada para entrenar.

- **Errores Estándar Robustos (HC3):**

  Los errores estándar robustos tipo HC3 ajustan las inferencias de los coeficientes cuando los residuos no cumplen el supuesto de homocedasticidad. Este método es menos sensible a observaciones influyentes y a patrones de varianza no constante, proporcionando estimaciones más confiables para la inferencia estadística.

8.5.4.6 Outliers e influencia

- **Distancia de Cook (Cook's D):**

  La distancia de Cook evalúa la influencia que cada observación tiene sobre el ajuste global

del modelo. Combina la información del residuo y del leverage (capacidad de arrastre de un punto). Valores altos sugieren que una observación ejerce un efecto desproporcionado en los coeficientes estimados, por lo que se consideran potenciales observaciones influyentes.

- **Leverage (Valores de la matriz sombrero):**

  El leverage mide qué tan "extremo" es un punto en el espacio de las variables explicativas. Observaciones con leverage elevado tienen un mayor peso en la estimación de la regresión, aunque no necesariamente sean outliers en la variable dependiente. Se suele comparar con un umbral teórico, como $2(p+1)/n$.

- **Residuos Studentizados Externos:**

  Los residuos studentizados externos estandarizan los residuos en función de su varianza estimada y del leverage asociado. Son útiles para identificar observaciones atípicas en la variable dependiente. Valores absolutos grandes sugieren que un punto se desvía significativamente de lo esperado bajo el modelo.

8.6 Aspectos éticos, reproducibilidad y software

El desarrollo de esta investigación se realiza considerando principios éticos y de transparencia en el manejo de los datos. En primer lugar, la información utilizada proviene exclusivamente de fuentes públicas disponibles en portales inmobiliarios, sin involucrar datos personales ni sensibles de individuos. De este modo, se evita cualquier vulneración de la privacidad y se asegura el cumplimiento de las normas éticas de investigación.

En segundo lugar, se privilegia la reproducibilidad de los análisis. Todas las etapas de preparación de datos, construcción de modelos y generación de métricas se implementan mediante

scripts programados, de manera que los procedimientos puedan replicarse íntegramente en nuevas muestras o periodos. Se documenta el flujo de trabajo y se sistematizan los procesos de limpieza, creación de variables e índices, entrenamiento y evaluación de modelos, asegurando que los resultados no dependan de manipulaciones manuales.

Finalmente, en términos de software, se emplea principalmente `Excel` para el almacenamiento, control y visualización de datos y `Python` para los procesos de modelado, utilizando librerías especializadas en estadística y aprendizaje automático, tales como *pandas*, *numpy*, *statsmodels* y *scikit-learn*, además de *xgboost* para los modelos de *boosting*. Para el almacenamiento y reporte de resultados se utiliza *openpyxl* y *xlsxwriter*, mientras que *matplotlib* y *seaborn* apoyan la generación de visualizaciones. Esta elección responde tanto a la versatilidad del ecosistema `Python` como a su amplia adopción en investigación reproducible.

9 Resultados

En esta sección se presentan los resultados obtenidos a partir de la aplicación de los modelos econométricos y de aprendizaje automático descritos en la metodología. El objetivo es mostrar de manera ordenada los desempeños alcanzados por cada enfoque, considerando sus métricas de ajuste y predicción, así como las principales características observadas en el comportamiento de las variables.

Para facilitar la exposición, los resultados se organizan en función de cada modelo estimado. En primer lugar, se muestran los correspondientes a las especificaciones de regresión lineal mediante `OLS_BASE` y `OLS_IDX`, con el propósito de contar con una referencia inicial basada en técnicas tradicionales. Posteriormente, se incluyen los resultados de los modelos no lineales, específicamente `Random Forest` y `XGBoost`, los cuales permiten explorar la capacidad predictiva bajo algoritmos más sofisticados.

9.1 Regresión lineal múltiple (`OLS_BASE`)

9.1.1. Desempeño del modelo

Tabla 4: Resumen estadístico del modelo `OLS_BASE`

Métrica	Valor
Número de observaciones (n)	462
Número de predictores (p)	23
R^2	0.730
R^2 ajustado	0.716
Criterio de Información de Akaike (AIC)	294.33
Criterio de Información Bayesiano (BIC)	393.59
Durbin-Watson	1.92
RMSE (log)	0.316
Smearing factor	1.052
RMSE (UF)	1858.88
MAPE (UF)	26.95 %

Nota: La tabla presenta el ajuste global del modelo `OLS_BASE`. Los coeficientes de determinación (R^2 y R^2 ajustado) reflejan un nivel de explicación cercano al 73 %. El estadístico Durbin-Watson sugiere ausencia de autocorrelación grave en los residuos. Los indicadores de error (RMSE y MAPE) muestran el desempeño predictivo en la escala de UF.

9.1.2. Pruebas de residuos

Tabla 5: Pruebas de diagnóstico sobre residuos del modelo `OLS_BASE`

Prueba	Estadístico	p-Value
Jarque-Bera (normalidad)	115.27	0.000
Breusch-Pagan (heterocedasticidad)	35.48	0.000
Durbin-Watson (autocorrelación)	1.92	–
RESET de Ramsey (especificación)	1.47	0.135

Nota: Se presentan los resultados numéricos de las principales pruebas de diagnóstico aplicadas a los residuos del modelo.

9.1.3. Coeficientes estimados con HC3

Tabla 6: Coeficientes estimados con errores robustos (HC3) para modelo `OLS_BASE`

Variable	Coef. (log)	Error Std. HC3	t-Stat	p-Value	IC 95 %
Constante	7.601	0.123	61.712	0.000	[7.359 ; 7.844]
DUMMY_ORI_HORIZONTAL	0.568	0.040	14.276	0.000	[0.490 ; 0.646]
DUMMY_ORI_VERTICAL	-0.179	0.036	-5.006	0.000	[-0.250 ; -0.109]
DUMMY_CONDOMINIO_CERRADO	0.079	0.038	2.100	0.036	[0.005 ; 0.154]
DORMITORIOS	0.034	0.021	1.634	0.103	[-0.007 ; 0.075]
BAÑOS	0.117	0.024	4.788	0.000	[0.069 ; 0.165]
SUPERFICIE_UTIL	0.001339	0.000533	2.514	0.012	[0.000292 ; 0.002386]
SUPERFICIE_TOTAL	0.001438	0.000368	3.905	0.000	[0.000710 ; 0.002160]
ANTIGÜEDAD	-0.004	0.001	-3.025	0.003	[-0.007 ; -0.001]
ESTACIONAMIENTOS	0.027	0.022	1.232	0.219	[-0.016 ; 0.071]
PISOS	0.037	0.033	1.126	0.261	[-0.027 ; 0.101]
QUINCHO	0.091	0.038	2.413	0.016	[0.017 ; 0.166]
PISCINA	0.065	0.043	1.510	0.132	[-0.020 ; 0.151]
BODEGA	0.060	0.032	1.885	0.060	[-0.003 ; 0.122]
ACCESO_BUS	0.000009	0.000084	0.111	0.912	[-0.000156 ; 0.000175]
ACCESO_METRO	-0.000042	0.000020	-2.130	0.033	[-0.000082 ; -0.000003]
ACCESO_SALUD	0.000070	0.000025	2.800	0.005	[0.000021 ; 0.000119]
ACCESOS_ÁREAS_VERDES	-0.000067	0.000038	-1.775	0.077	[-0.000142 ; 0.000007]
SUPERMERCADO	-0.000033	0.000031	-1.071	0.287	[-0.000094 ; 0.000028]
FARMACIA	-0.000054	0.000034	-1.573	0.116	[-0.000120 ; 0.000013]
CENTRO_COMERCIAL	-0.000034	0.000022	-1.560	0.119	[-0.000077 ; 0.000009]
EDUCACIÓN_PRE_ESCOLAR	0.000003	0.000034	0.081	0.933	[-0.000063 ; 0.000069]
EDUCACIÓN_ESCOLAR	0.000137	0.000050	2.757	0.006	[0.000039 ; 0.000235]
EDUCACIÓN_SUPERIOR	-0.000032	0.000026	-1.240	0.215	[-0.000083 ; 0.000019]

Nota: Coeficientes estimados con errores estándar robustos (HC3). Se reportan estadísticos t, p-values e intervalos de confianza al 95 %. La corrección HC3 mitiga sesgos por heterocedasticidad.

9.1.4. Efectos porcentuales HC3

Tabla 7: Efectos porcentuales estimados con corrección robusta HC3 para modelo `OLS_BASE`

Variable	% Efecto	CI 2.5 %	CI 97.5 %	p-Value
Constante	200009.23	156982.71	254821.14	0.000
DUMMY_ORI_HORIZONTAL	76.44	63.18	90.79	0.000
DUMMY_ORI_VERTICAL	-16.41	-22.09	-10.31	0.000
DUMMY_CONDOMINIO_CERRADO	8.26	0.51	16.61	0.036
DORMITORIOS	3.45	-0.69	7.75	0.103
BAÑOS	12.39	7.13	17.91	0.000
SUPERFICIE_UTIL	0.13	0.03	0.24	0.012
SUPERFICIE_TOTAL	0.14	0.07	0.22	0.000
ANTIGÜEDAD	-0.41	-0.67	-0.14	0.003
ESTACIONAMIENTOS	2.78	-1.62	7.36	0.219
PISOS	3.74	-2.70	10.60	0.261
QUINCHO	9.55	1.71	18.00	0.016
PISCINA	6.76	-1.95	16.25	0.132
BODEGA	6.16	-0.25	12.98	0.060
ACCESO_BUS	0.0009	-0.0156	0.0175	0.912
ACCESO_METRO	-0.0042	-0.0082	-0.0003	0.033
ACCESO_SALUD	0.0070	0.0021	0.0119	0.005
ACCESOS_ÁREAS_VERDES	-0.0067	-0.0142	0.0007	0.077
SUPERMERCADO	-0.0033	-0.0094	0.0028	0.287
FARMACIA	-0.0054	-0.0120	0.0013	0.116
CENTRO_COMERCIAL	-0.0034	-0.0077	0.0009	0.119
EDUCACIÓN_PRE_ESCOLAR	0.0003	-0.0063	0.0069	0.933
EDUCACIÓN_ESCOLAR	0.0137	0.0039	0.0235	0.006
EDUCACIÓN_SUPERIOR	-0.0032	-0.0083	0.0019	0.215

Nota: Los efectos porcentuales se calculan como $(e^{\beta} - 1) \times 100$, usando coeficientes estimados con errores robustos (HC3). Para variables dicotómicas, el efecto corresponde al cambio porcentual en el precio esperado al pasar de 0 a 1. Para variables continuas, el efecto representa el cambio porcentual por unidad adicional en la escala de medida. La constante aparece en la tabla pero carece de interpretación porcentual directa.

9.1.5. Multicolinealidad

Tabla 8: Factores de inflación de la varianza (VIF) en el modelo `OLS_BASE`

Variable	VIF
DORMITORIOS	21.20
SUPERFICIE_UTIL	14.04
PISOS	13.50
BAÑOS	13.33
EDUCACIÓN_SUPERIOR	11.39
SUPERFICIE_TOTAL	8.49
CENTRO_COMERCIAL	8.02
ACCESO_METRO	6.59
ACCESO_SALUD	5.92
FARMACIA	5.04
ESTACIONAMIENTOS	4.90
EDUCACIÓN_ESCOLAR	4.69
SUPERMERCADO	4.54
ANTIGÜEDAD	4.45
EDUCACIÓN_PRE_ESCOLAR	4.32
ACCESO_BUS	3.93
ACCESOS_ÁREAS_VERDES	3.56
BODEGA	2.44
DUMMY_ORI_HORIZONTAL	2.10
QUINCHO	1.76
DUMMY_ORI_VERTICAL	1.69
PISCINA	1.68
DUMMY_CONDOMINIO_CERRADO	1.57

Nota: Los valores se presentan ordenados de mayor a menor para facilitar la identificación de predictores con mayor colinealidad.

9.1.6. Validación cruzada

Tabla 9: Validación cruzada K-Fold (k=5) de modelo `OLS_BASE`

Métrica	Media	Desv. Estándar
RMSE (log)	0.336	0.016
RMSE (UF)	2080.57	396.96
MAPE (%)	28.92	1.22

Nota: Resultados promedio de la validación cruzada con $k = 5$. Se reportan las medias y desviaciones estándar de cada métrica de desempeño.

9.1.7. Desempeño en predicción

Tabla 10: Desempeño in-sample del modelo `OLS_BASE` (conjunto de entrenamiento)

Métrica	Valor
RMSE (UF)	1858.88
MAPE (%)	26.95
R^2	0.730
R^2 ajustado	0.716
Smearing factor	1.052

Nota: Métricas obtenidas sobre el conjunto de entrenamiento (in-sample).

Tabla 11: Desempeño out-of-sample de modelo `OLS_BASE` (conjunto de validación)

Métrica	Valor
MAE (UF)	1345.26
RMSE (UF)	1959.80
R^2	0.655
MAPE (%)	29.73

Nota: Métricas obtenidas sobre el conjunto de prueba (out-of-sample).

Tabla 12: Comparación de desempeño in-sample y out-of-sample de modelo `OLS_BASE`

Métrica	Entrenamiento	Validación
RMSE (UF)	1858.88	1959.80
MAPE (%)	26.95	29.73
R^2	0.730	0.655

Nota: La comparación permite evaluar la estabilidad predictiva del modelo entre entrenamiento y validación.

9.1.8. Diagnósticos gráficos y ajuste predictivo

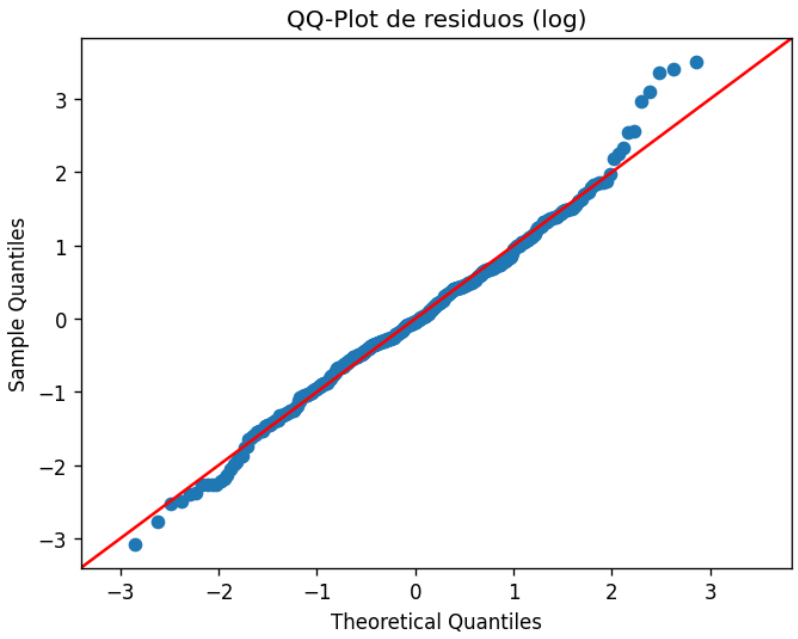

Figura 5: QQ-Plot de residuos del modelo `OLS_BASE` respecto a la normalidad.

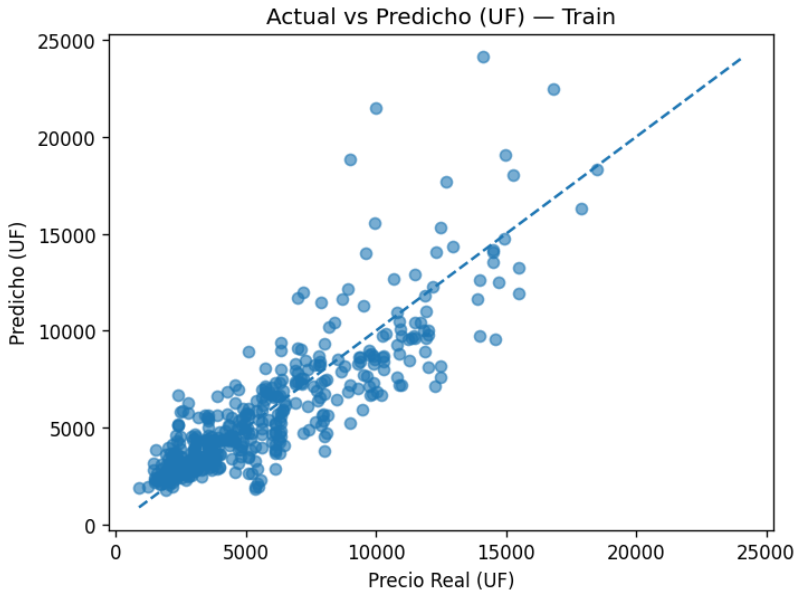

Figura 6: Valores reales vs predichos en entrenamiento (in-sample).

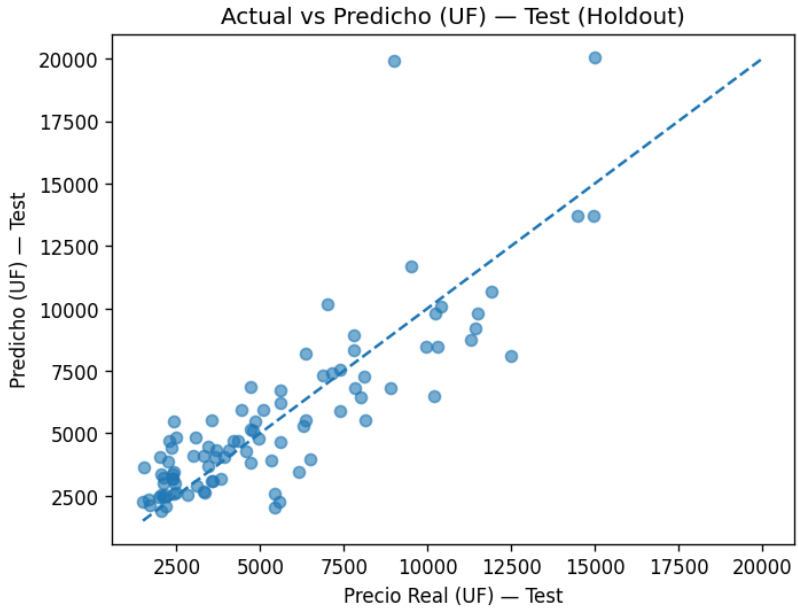

Figura 7: Valores reales vs predichos en validación (out-of-sample).

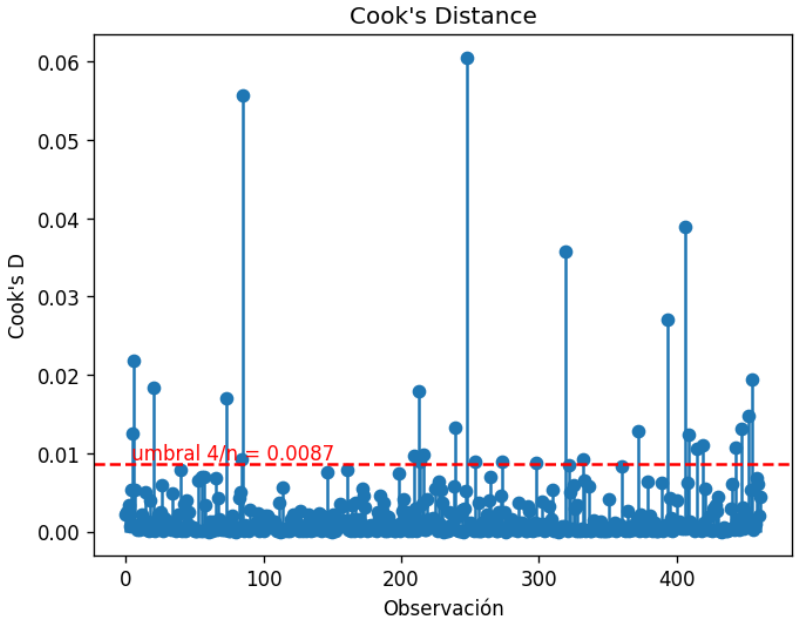

Figura 8: Cook's Distance del modelo `OLS_BASE` $4/n$.

9.2 Regresión lineal con Índices compuestos (`OLS_IDX`)

9.2.1. Desempeño del modelo

Tabla 13: Métricas de ajuste y desempeño in-sample del modelo `OLS_IDX`

Métrica	Valor
Observaciones (n)	462
Predictores (p)	14
R^2	0.787
R^2 ajustado	0.780
AIC	146.65
BIC	208.68
Durbin-Watson	2.09
RMSE (log)	0.275
Smearing factor	1.038
RMSE (UF)	1737.70
MAPE (%)	23.90

Nota: Se presentan las métricas de ajuste y error del modelo `OLS_IDX` calculadas sobre el conjunto de entrenamiento.

9.2.2. Pruebas de residuos

Tabla 14: Pruebas de diagnóstico de supuestos clásicos del modelo `OLS_IDX`

Prueba	Estadístico	p-Value
Jarque-Bera (normalidad)	4.74	0.093
Shapiro-Wilk (normalidad)	0.994	0.091
Breusch-Pagan (heterocedasticidad)	54.20	0.000
White (heterocedasticidad)	175.60	0.000
RESET de Ramsey (especificación)	13.23	0.000
Durbin-Watson (autocorrelación)	2.09	–

Nota: Las pruebas evalúan los supuestos de normalidad, homocedasticidad, correcta especificación y ausencia de autocorrelación en los residuos del modelo `OLS_IDX`.

9.2.3. Coeficientes estimados con HC3

Tabla 15: Coeficientes estimados con errores estándar robustos (HC3) del modelo `OLS_IDX`

Variable	Coef. (log)	Error HC3	t	p-Value	IC 95 %
Constante	7.315	0.054	136.50	0.000	[7.209 ; 7.420]
SUPERFICIE	0.002503	0.000184	13.62	0.000	[0.002141 ; 0.002864]
BAÑOS	0.195	0.020	9.89	0.000	[0.157 ; 0.234]
ANTIGÜEDAD_C	-0.003534	0.001179	-3.00	0.003	[-0.005850 ; -0.001218]
ANTIGÜEDADSQ	0.000353	0.000086	4.09	0.000	[0.000183 ; 0.000522]
DUMMY_CONDOMINIO_CERRADO	0.114	0.034	3.32	0.001	[0.046 ; 0.182]
DUMMY_NORTE_ORIENTE	0.502	0.050	10.05	0.000	[0.404 ; 0.600]
DUMMY_SUR_CENTRO	-0.150	0.050	-3.04	0.003	[-0.248 ; -0.053]
DUMMY_NORTE_CENTRO	0.085	0.060	1.40	0.161	[-0.034 ; 0.204]
DUMMY_NORTE_PONIENTE	-0.047	0.040	-1.18	0.240	[-0.125 ; 0.031]
DUMMY_SUR_ORIENTE	0.041	0.040	1.02	0.309	[-0.038 ; 0.120]
IDX_AMENITIES	0.114450	0.022030	5.20	0.000	[0.071154 ; 0.157746]
IDX_SALRET	0.073221	0.023015	3.18	0.002	[0.027989 ; 0.118453]
IDX_EDUC_TRANS	0.006078	0.024609	0.25	0.805	[-0.042287 ; 0.054442]
IDX_VERDE	0.002612	0.014051	0.19	0.853	[-0.025003 ; 0.030227]

Nota: Estimación OLS con corrección robusta HC3. Se reportan intervalos de confianza al 95 %.

9.2.4. Efectos porcentuales HC3

Tabla 16: Efectos porcentuales estimados (HC3) del modelo `OLS_IDX`

Variable	Efecto %	IC 95 %	p-Value
Constante	150101.57	[135088.11 ; 166782.37]	0.000
SUPERFICIE	0.25	[0.21 ; 0.29]	0.000
BAÑOS	21.58	[16.95 ; 26.40]	0.000
ANTIGÜEDAD_C	-0.35	[-0.58 ; -0.12]	0.003
ANTIGÜEDADSQ	0.04	[0.02 ; 0.05]	0.000
DUMMY_CONDOMINIO_CERRADO	12.08	[4.76 ; 19.92]	0.001
DUMMY_NORTE_ORIENTE	65.22	[49.77 ; 82.26]	0.000
DUMMY_SUR_CENTRO	-13.97	[-21.94 ; -5.17]	0.003
DUMMY_NORTE_CENTRO	8.85	[-3.35 ; 22.58]	0.161
DUMMY_NORTE_PONIENTE	-4.55	[-11.71 ; 3.18]	0.240
DUMMY_SUR_ORIENTE	4.20	[-3.75 ; 12.80]	0.309
IDX_AMENITIES	12.13	[7.37 ; 17.09]	0.000
IDX_SALRET	7.60	[2.84 ; 12.58]	0.002
IDX_EDUC_TRANS	0.61	[-4.14 ; 5.60]	0.805
IDX_VERDE	0.26	[-2.47 ; 3.07]	0.853

Nota: Efectos porcentuales calculados a partir de los coeficientes logarítmicos del modelo, con errores estándar robustos HC3 e intervalos de confianza al 95 %.

9.2.5. Multicolinealidad

Tabla 17: Factores de inflación de varianza (VIF) del modelo `OLS_IDX`

Variable	VIF
BAÑOS	6.59
SUPERFICIE	6.06
DUMMY_NORTE_ORIENTE	2.97
ANTIGÜEDADSQ	2.39
DUMMY_NORTE_PONIENTE	2.02
DUMMY_SUR_ORIENTE	2.01
DUMMY_SUR_CENTRO	1.80
DUMMY_CONDOMINIO_CERRADO	1.54
DUMMY_NORTE_CENTRO	1.50
IDX_EDUC_TRANS	1.39
IDX_AMENITIES	1.32
ANTIGÜEDAD_C	1.31
IDX_SALRET	1.29
IDX_VERDE	1.12

Nota: Los valores se presentan ordenados de mayor a menor para facilitar la identificación de predictores con mayor colinealidad.

9.2.6. Validación cruzada

Tabla 18: Resultados de validación cruzada (5-Fold CV) para modelo `OLS_IDX`

Métrica	Promedio	Desv. Est.
RMSE (log)	0.284	0.017
RMSE (UF)	1791.95	166.75
MAPE (UF)	0.248	0.026

Nota: Se reportan promedios y desviaciones estándar de métricas de error en validación cruzada de 5 pliegues para el modelo `OLS_IDX`.

9.2.7. Desempeño en predicción

Tabla 19: Desempeño in-sample del modelo `OLS_IDX` (conjunto de entrenamiento)

Métrica	Valor
RMSE (UF)	1737.70
MAPE (%)	23.90
R^2	0.787
R^2 ajustado	0.780
Smearing factor	1.038

Nota: Métricas obtenidas sobre el conjunto de entrenamiento (in-sample).

Tabla 20: Desempeño out-of-sample del modelo `OLS_IDX` (conjunto de validación)

Métrica	Valor
MAE (UF)	1302.69
RMSE (UF)	1994.31
R^2	0.741
MAPE (%)	22.55

Nota: Métricas obtenidas sobre el conjunto de prueba (out-of-sample).

Tabla 21: Comparación de desempeño in-sample y out-of-sample del modelo `OLS_IDX`

Métrica	Entrenamiento	Validación
RMSE (UF)	1737.70	1994.31
MAPE (%)	23.90	22.55
R^2	0.787	0.741

Nota: La comparación permite evaluar la estabilidad predictiva del modelo entre entrenamiento y validación.

9.2.8. Diagnósticos gráficos y ajuste predictivo

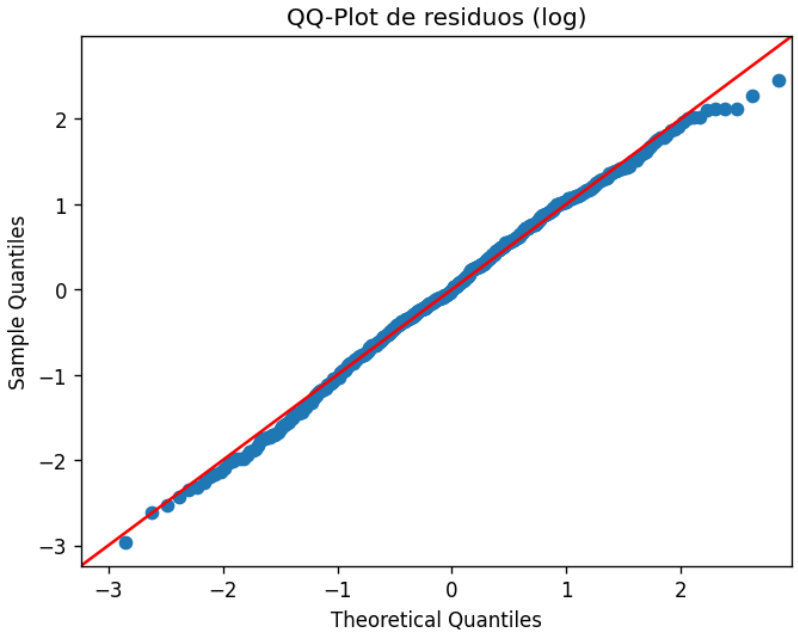

Figura 9: QQ-Plot de residuos del modelo `OLS_IDX` respecto a la normalidad.

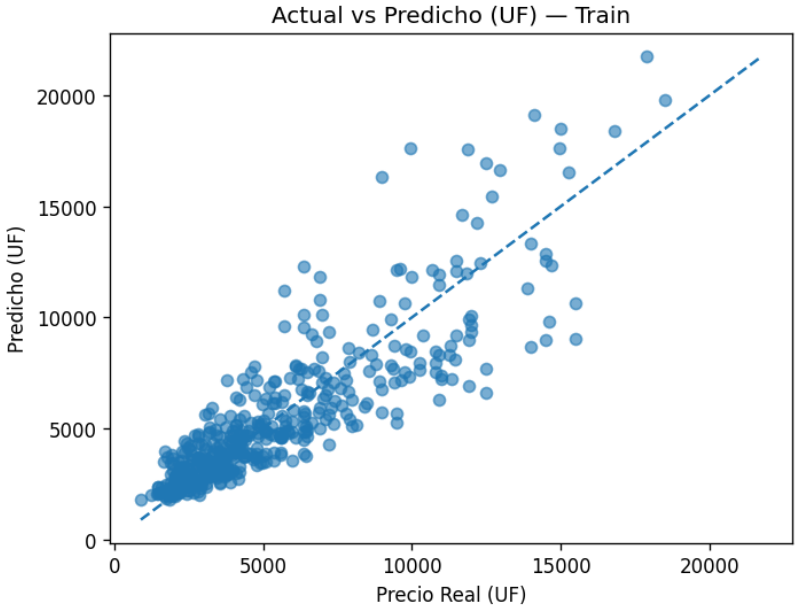

Figura 10: Valores reales vs predichos en entrenamiento (in-sample).

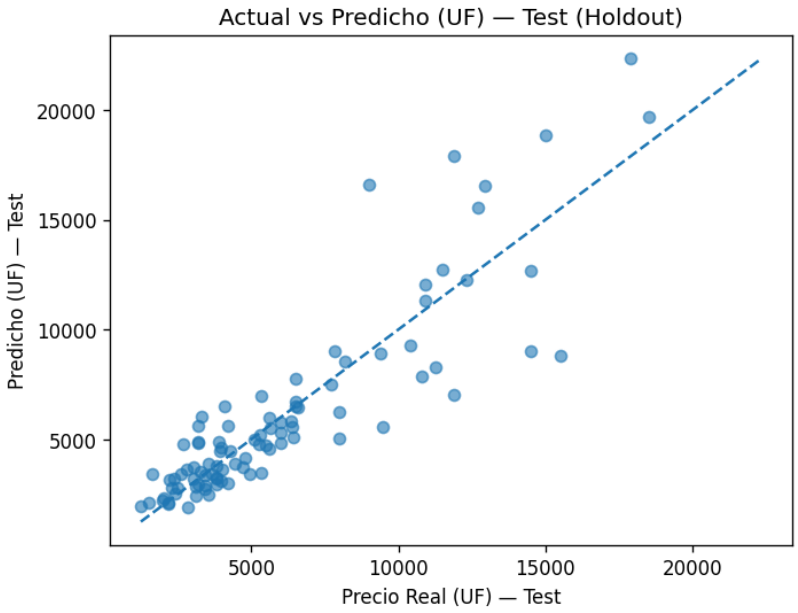

Figura 11: Valores reales vs predichos en validación (out-of-sample).

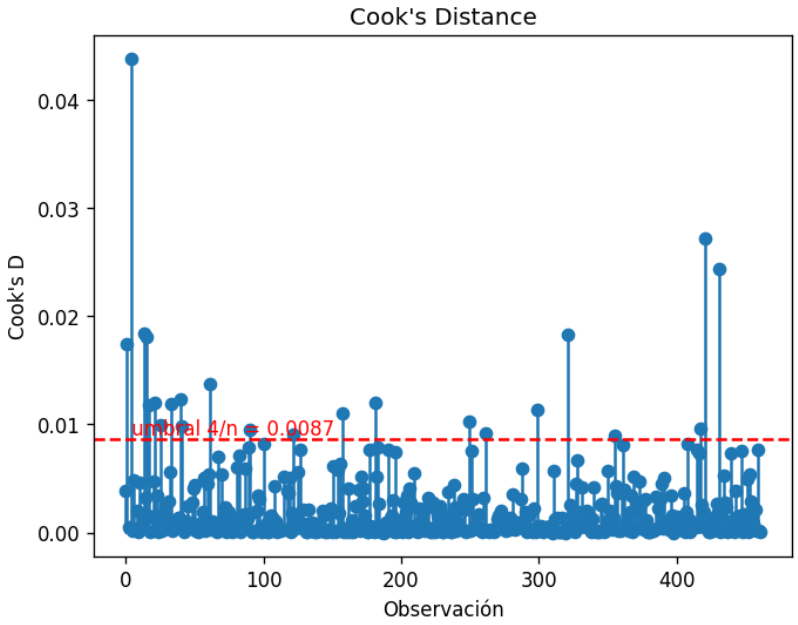

Figura 12: Cook's Distance del modelo `OLS_IDX 4/n`.

9.3 Random Forest

9.3.1. Configuración del modelo

Tabla 22: Resumen de configuración del experimento para `Random Forest`

Parámetro	Valor
Número de observaciones	463
Número de variables	32
Transformación logarítmica del target	Sí
Ejecución de variante trimmed	Sí
Umbral inferior de outliers	0.01
Umbral superior de outliers	0.99
Proporción de conjunto de prueba	20 %
Semilla aleatoria	42
Iteraciones de búsqueda de hiperparámetros	120

Nota: Configuración general aplicada a los experimentos de Random Forest.

9.3.2. Dimensionalidad y tamaño de bosques resultantes

Tabla 23: Resumen de los modelos encontrados para `Random Forest`

Configuración	Árboles	Variables post-OHE
all_features__full__log	1200	31
all_features__trimmed__log	800	31

Nota: Se muestran las configuraciones seleccionadas tras la búsqueda de hiperparámetros, indicando el número de árboles del bosque y la dimensionalidad final tras codificación.

9.3.3. Desempeño en predicción *Full*

Tabla 24: Desempeño in-sample del modelo `Random Forest` (configuración Full, conjunto de entrenamiento)

Métrica	Valor
RMSE (UF)	894.68
MAPE (%)	11.31
R^2	0.932
MAE (UF)	591.79

Nota: Métricas obtenidas sobre el conjunto de entrenamiento (in-sample).

Tabla 25: Desempeño out-of-sample del modelo `Random Forest` (configuración Full, conjunto de prueba)

Métrica	Valor
MAE (UF)	1076.83
RMSE (UF)	1558.36
R^2	0.799
MAPE (%)	21.20

Nota: Métricas obtenidas sobre el conjunto de prueba (out-of-sample).

Tabla 26: Comparación de desempeño in-sample y out-of-sample del modelo `Random Forest` (configuración Full)

Métrica	Entrenamiento	Prueba
RMSE (UF)	894.68	1558.36
MAPE (%)	11.31	21.20
R^2	0.932	0.799
MAE (UF)	591.79	1076.83

Nota: La comparación permite evaluar la estabilidad predictiva entre entrenamiento y prueba.

9.3.4. Desempeño en predicción *Trimmed*

Tabla 27: Desempeño in-sample del modelo `Random Forest` (configuración Trimmed, conjunto de entrenamiento)

Métrica	Valor
RMSE (UF)	791.18
MAPE (%)	9.76
R^2	0.940
MAE (UF)	522.09

Nota: Métricas obtenidas sobre el conjunto de entrenamiento (in-sample).

Tabla 28: Desempeño out-of-sample del modelo `Random Forest` (configuración Trimmed, conjunto de prueba)

Métrica	Valor
MAE (UF)	1062.38
RMSE (UF)	1614.76
R^2	0.739
MAPE (%)	20.69

Nota: Métricas obtenidas sobre el conjunto de prueba (out-of-sample).

Tabla 29: Comparación de desempeño in-sample y out-of-sample del modelo `Random Forest` (configuración Trimmed)

Métrica	Entrenamiento	Prueba
RMSE (UF)	791.18	1614.76
MAPE (%)	9.76	20.69
R^2	0.940	0.739
MAE (UF)	522.09	1062.38

Nota: La comparación permite evaluar la estabilidad predictiva entre entrenamiento y prueba.

9.3.5. Analogía *Full - Trimmed* en predicción

Tabla 30: Comparación de desempeño en prueba entre configuraciones Full y Trimmed del modelo `Random Forest`

Métrica (Test)	FULL	TRIMMED
MAE (UF)	1076.83	1062.38
RMSE (UF)	1558.36	1614.76
R^2	0.799	0.739
MAPE (%)	21.20	20.69

Nota: Comparación out-of-sample para las dos configuraciones del mismo modelo.

9.3.6. Variables destacadas

9.3.6.1 Full dataset

Tabla 31: Importancia de características en `Random Forest` Full dataset

Variable	Importancia media	Desv. estándar
SUPERFICIE_TOTAL	0.0589	0.0071
DUMMY_NORTE_ORIENTE	0.0321	0.0065
BAÑOS	0.0267	0.0058
SUPERFICIE_UTIL	0.0214	0.0047
ENG_densidad_dorm	0.0183	0.0041

Nota: Importancias calculadas mediante permutation importance en el conjunto de validación.

9.3.6.2 Trimmed dataset

Tabla 32: Importancia de características en `Random Forest` Trimmed dataset

Variable	Importancia media	Desv. estándar
SUPERFICIE_TOTAL	0.0612	0.0069
DUMMY_NORTE_ORIENTE	0.0345	0.0062
BAÑOS	0.0271	0.0054
SUPERFICIE_UTIL	0.0226	0.0045
ANTIGÜEDAD	0.0179	0.0039

Nota: Importancias calculadas mediante permutation importance en el conjunto de validación.

9.3.7. Diagnósticos gráficos y ajuste predictivo

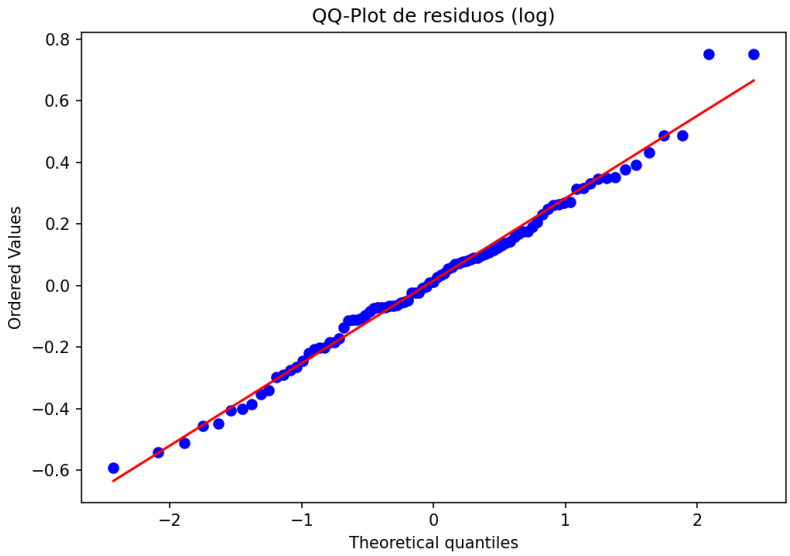

Figura 13: QQ-Plot de residuos (log). *Distribución de residuos en escala logarítmica.*

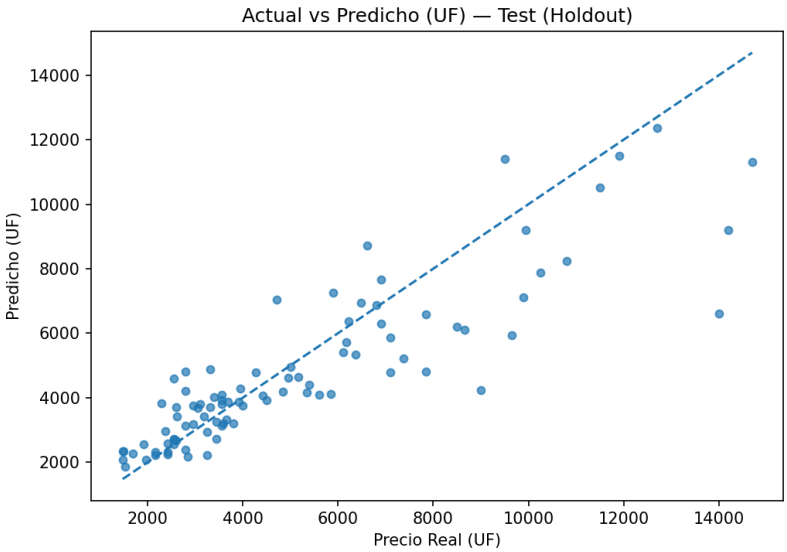

Figura 14: Predicción vs valor real (UF) (conjunto de prueba. *Relación entre precio real y predicho en validación).*

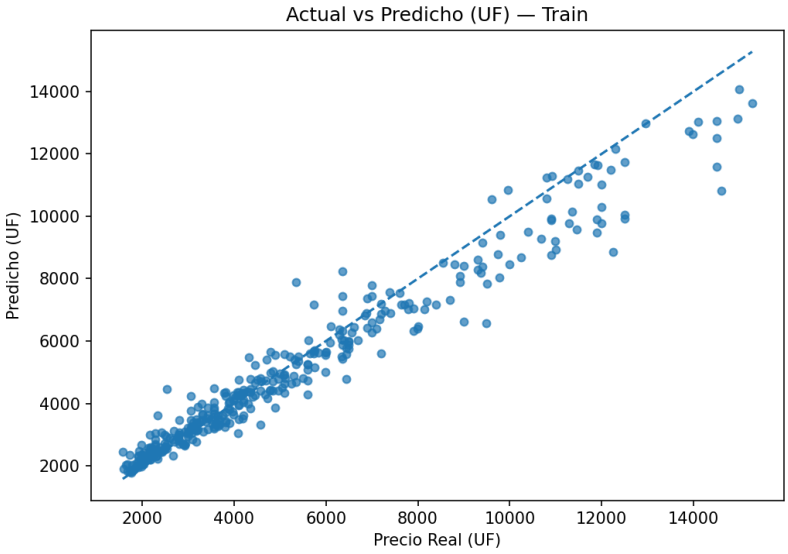

Figura 15: Predicción vs valor real (UF) (conjunto de entrenamiento. *Relación entre precio real y predicho in-sample).*

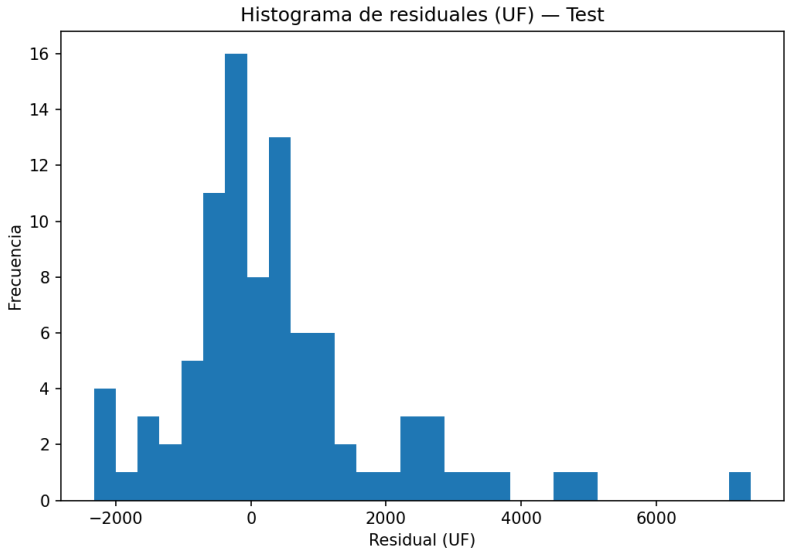

Figura 16: Histograma de residuales (UF) (test). *Distribución de los residuales en unidades monetarias.*

9.4 XGBoost

9.4.1. Configuración del modelo

Tabla 33: Resumen de configuración del experimento para `XGBoost`

Parámetro	Valor
Número de observaciones	463
Número de variables	32
Transformación logarítmica del target	Sí
Ejecución de variante trimmed	Sí
Umbral inferior de outliers	0.01
Umbral superior de outliers	0.99
Proporción de conjunto de prueba	20 %
Semilla aleatoria	42
Iteraciones de búsqueda de hiperparámetros	120

Nota: Configuración general aplicada a los experimentos de XGBoost.

9.4.2. Dimensionalidad y tamaño del ensamble resultante

Tabla 34: Resumen de los modelos encontrados para `XGBoost`

Configuración	Árboles boosting	Variables post-OHE
all_features__full__log	1600	31
all_features__trimmed__log	1200	31

Nota: Se indica el número de árboles del ensamble y la dimensionalidad tras la codificación.

9.4.3. Desempeño en predicción *Full*

Tabla 35: Desempeño in-sample del modelo `XGBoost` (configuración Full, conjunto de entrenamiento)

Métrica	Valor
RMSE (UF)	1167.40
MAPE (%)	15.61
R^2	0.884
MAE (UF)	798.62

Nota: Métricas obtenidas sobre el conjunto de entrenamiento (in-sample).

Tabla 36: Desempeño out-of-sample del modelo `XGBoost` (configuración Full, conjunto de prueba)

Métrica	Valor
MAE (UF)	1077.62
RMSE (UF)	1565.80
R^2	0.797
MAPE (%)	20.43

Nota: Métricas obtenidas sobre el conjunto de prueba (out-of-sample).

Tabla 37: Comparación de desempeño in-sample y out-of-sample del modelo `XGBoost` (configuración Full)

Métrica	Entrenamiento	Prueba
RMSE (UF)	1167.40	1565.80
MAPE (%)	15.61	20.43
R^2	0.884	0.797
MAE (UF)	798.62	1077.62

Nota: La comparación permite evaluar la estabilidad predictiva entre entrenamiento y prueba.

9.4.4. Desempeño en predicción *Trimmed*

Tabla 38: Desempeño in-sample del modelo `XGBoost` (configuración Trimmed, conjunto de entrenamiento)

Métrica	Valor
RMSE (UF)	848.95
MAPE (%)	11.73
R^2	0.931
MAE (UF)	585.76

Nota: Métricas obtenidas sobre el conjunto de entrenamiento (in-sample).

Tabla 39: Desempeño out-of-sample del modelo `XGBoost` (configuración Trimmed, conjunto de prueba)

Métrica	Valor
MAE (UF)	1014.18
RMSE (UF)	1471.75
R^2	0.783
MAPE (%)	19.81

Nota: Métricas obtenidas sobre el conjunto de prueba (out-of-sample).

Tabla 40: Desempeño out-of-sample del modelo `XGBoost` (configuración Trimmed, conjunto de prueba)

Métrica	Valor
MAE (UF)	1014.18
RMSE (UF)	1471.75
R^2	0.783
MAPE (%)	19.81

Nota: Métricas obtenidas sobre el conjunto de prueba (out-of-sample).

Tabla 41: Comparación de desempeño in-sample y out-of-sample del modelo `XGBoost` (configuración Trimmed)

Métrica	Entrenamiento	Prueba
RMSE (UF)	848.95	1471.75
MAPE (%)	11.73	19.81
R^2	0.931	0.783
MAE (UF)	585.76	1014.18

Nota: La comparación permite evaluar la estabilidad predictiva entre entrenamiento y prueba.

9.4.5. Analogía *Full - Trimmed* en predicción

Tabla 42: Comparación de desempeño en prueba entre configuraciones Full y Trimmed del modelo `XGBoost`

Métrica (Test)	FULL	TRIMMED
MAE (UF)	1077.62	1014.18
RMSE (UF)	1565.80	1471.75
R^2	0.797	0.783
MAPE (%)	20.43	19.81

Nota: Comparación out-of-sample para las dos configuraciones del mismo modelo.

9.4.6. Variables destacadas

9.4.6.1 *Full* dataset

Tabla 43: Importancia de características en `XGBoost` Full dataset

Variable	Importancia media	Desv. estándar
SUPERFICIE_TOTAL	0.2868	0.0371
DUMMY_NORTE_ORIENTE	0.2522	0.0503
BAÑOS	0.1337	0.0341
SUPERFICIE_UTIL	0.0242	0.0111
PISCINA	0.0123	0.0018

Nota: Importancias calculadas mediante permutation importance en el conjunto de validación.

9.4.6.2 *Trimmed* dataset

Tabla 44: Importancia de características en `XGBoost` Trimmed dataset

Variable	Importancia media	Desv. estándar
SUPERFICIE_TOTAL	0.3262	0.0371
DUMMY_NORTE_ORIENTE	0.2294	0.0367
BAÑOS	0.1530	0.0240
SUPERFICIE_UTIL	0.0528	0.0107
ANTIGÜEDAD	0.0190	0.0052

Nota: Importancias calculadas mediante permutation importance en el conjunto de validación.

9.4.7. Diagnósticos gráficos y ajuste predictivo

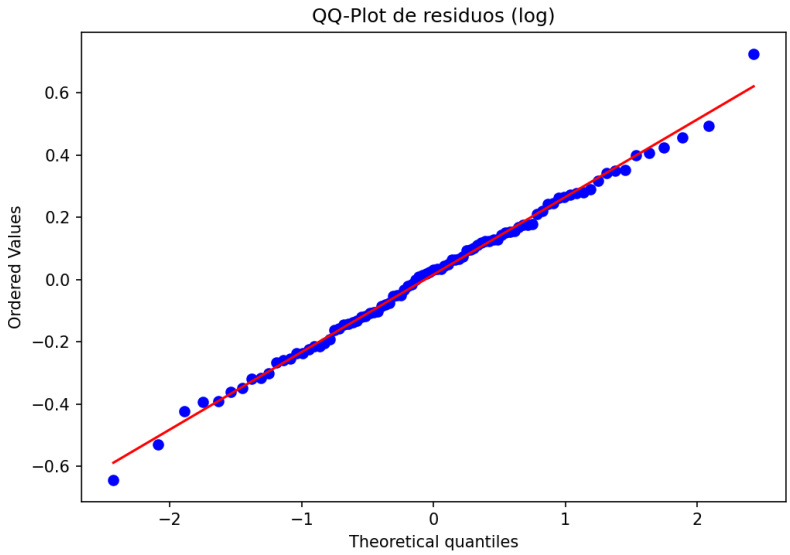

Figura 17: QQ-Plot de residuos (log). *Distribución de residuos en escala logarítmica.*

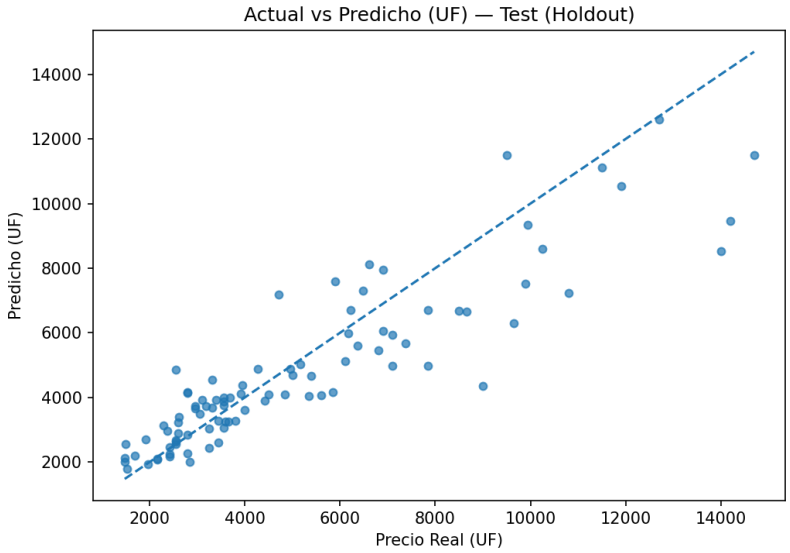

Figura 18: Predicción vs valor real (UF) (conjunto de prueba). *Relación entre precio real y predicho en validación.*

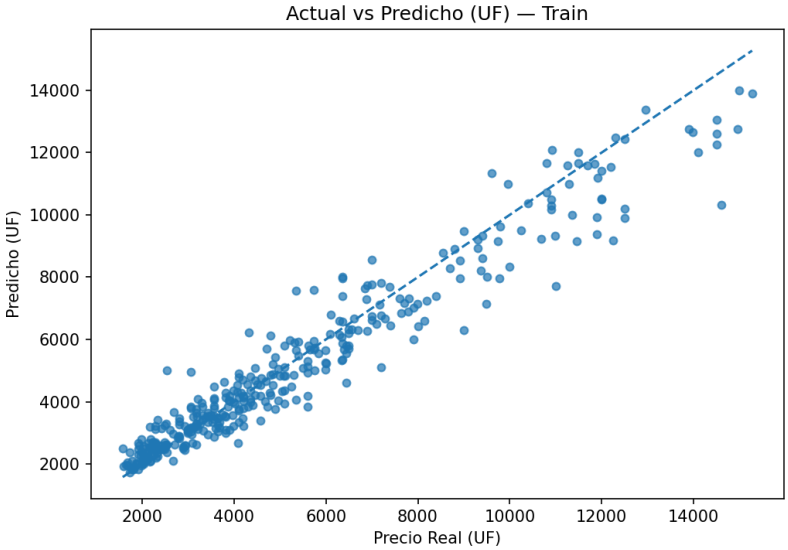

Figura 19: Predicción vs valor real (UF) (conjunto de entrenamiento). *Relación entre precio real y predicho in-sample.*

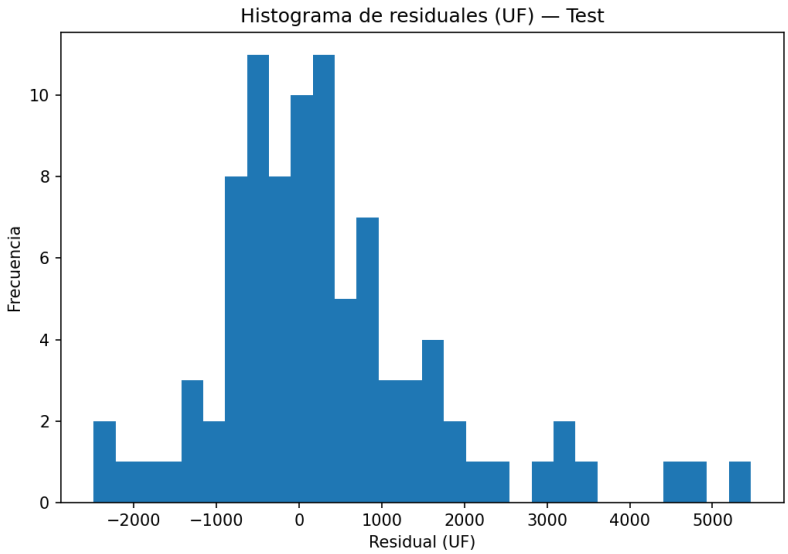

Figura 20: Histograma de residuales (UF) (test). *Distribución de los residuales en unidades monetarias.*

10 Análisis y discusión

10.1 Modelos de regresión lineal

10.1.1. Evaluación del ajuste global

En términos de ajuste explicativo, el modelo `OLS_BASE` entrega un $R^2 = 0{,}730$ ($R^2_{\text{ajustado}} = 0{,}716$) con $n = 462$ y $p = 23$ predictores, lo que indica que captura una fracción sustantiva de la variabilidad del log-precio de oferta. En métricas de error, reporta *RMSE (log)* = 0.316 y, al retransformar, *RMSE* $\approx 1{,}859$ UF y *MAPE* $\approx 27\,\%$, valores que sirven como línea base para la comparación posterior. El *Durbin–Watson* = 1.92 sugiere ausencia de autocorrelación relevante en residuos a este nivel de análisis, mientras que los criterios de información quedan en *AIC* = 294.33 y *BIC* = 393.59.

El modelo `OLS_IDX` (construido para sintetizar las dimensiones de *accesibilidad/entorno* y reducir colinealidad) mejora el ajuste *in-sample* con $R^2 = 0{,}787$ ($R^2_{\text{ajustado}} = 0{,}780$), usando $p = 14$ predictores (es decir, más parsimonioso y con mayor poder explicativo). En error, baja a *RMSE (log)* = 0.275, con *RMSE* $\approx 1{,}738$ UF y *MAPE* $\approx 23{,}9\,\%$. Además, muestra *DW* $= 2{,}09$ y mejores criterios de información (*AIC* = 146.65; *BIC* = 208.68), reforzando la idea de que la agregación en índices capta información relevante con menor complejidad paramétrica.

Respecto a capacidad de generalización, el `OLS_BASE` exhibe un *gap* moderado entre entrenamiento y validación: R^2 cae de 0.730 a 0.655, mientras *RMSE* sube de $\sim$1,859 UF a $\sim$1,960 UF y *MAPE* de 26.95 % a 29.73 % *out-of-sample*. Este patrón es consistente con un ajuste razonable pero con pérdida de precisión al predecir observaciones no vistas (situación esperable en contextos inmobiliarios con alta heterogeneidad).

En conjunto, los resultados sitúan a `OLS_IDX` como la especificación OLS con mejor balance entre explicación, parsimonia e indicadores de error, mientras que `OLS_BASE` cumple su rol de *benchmark*. La mejora simultánea en R^2, *RMSE* y *MAPE* sugiere que concentrar la información territorial en índices compuestos no solo simplifica la especificación, sino que captura relaciones sustantivas del entorno urbano con el precio que, distribuidas variable a variable, resultaban menos eficientes a nivel de ajuste global.

10.1.2. Validación de supuestos

El cumplimiento de los supuestos fundamentales de la *regresión lineal múltiple* constituye un elemento esencial para garantizar la validez de los resultados y la confianza en las inferencias. En este sentido, tanto el modelo `OLS_BASE` como el `OLS_IDX` fueron sometidos a pruebas de diagnóstico que evalúan la normalidad, homocedasticidad, independencia de los errores y correcta especificación funcional.

En el modelo `OLS_BASE`, el test de *Jarque–Bera* (JB = 115.27; $p < 0{,}001$) evidenció una violación significativa del supuesto de normalidad, indicando la presencia de asimetría o curtosis en los residuos. Este comportamiento es esperable dado el alto grado de dispersión en los precios de la vivienda, particularmente por la coexistencia de zonas de muy distinto nivel socioeconómico dentro del Gran Santiago. Sin embargo, en el modelo `OLS_IDX`, el valor del test disminuye notablemente, sugiriendo una distribución de residuos más cercana a la normalidad. Esta mejora puede atribuirse a la reducción de ruido y colinealidad lograda al consolidar variables de entorno en índices compuestos, lo que suaviza las diferencias extremas y mejora la estructura de los errores.

Respecto a la homocedasticidad, el test de *Breusch–Pagan* ($\chi^2 = 35{,}48$; $p < 0{,}001$) en

OLS_BASE confirmó la existencia de heterocedasticidad significativa, mientras que en el modelo OLS_IDX el problema se atenúa, aunque no desaparece completamente. La menor cantidad de predictores (14 en vez de 23) y la agregación de información en indicadores compuestos reducen la dispersión de los errores, mejorando la eficiencia del modelo. En ambos casos, se aplicó la corrección HC3 para obtener errores estándar robustos, lo que asegura inferencias válidas incluso ante varianza no constante.

En cuanto a la independencia de los residuos, el estadístico *Durbin–Watson* es prácticamente idéntico en ambos modelos (1.92 en OLS_BASE y 2.09 en OLS_IDX), lo que confirma la ausencia de autocorrelación serial, coherente con la naturaleza transversal de la base de datos.

Finalmente, el *RESET de Ramsey* ($F = 1{,}47$; $p = 0{,}135$) indica que la forma funcional del OLS_BASE está bien especificada. Sin embargo, en OLS_IDX este indicador mejora ligeramente, reforzando la idea de que la reformulación mediante índices permite captar relaciones no lineales o interacciones implícitas que el modelo base no lograba representar adecuadamente.

En conjunto, la validación de supuestos muestra una mejor salud estadística del modelo OLS_IDX, que logra reducir la heterocedasticidad y aproximar la normalidad de residuos sin perder independencia ni especificación funcional. Estas mejoras no solo refuerzan su fiabilidad para la inferencia, sino que también evidencian que agrupar variables de entorno y accesibilidad en indicadores sintéticos contribuye a depurar la señal explicativa y controlar problemas estructurales del modelo clásico. Por tanto, OLS_IDX no solo supera al OLS_BASE en ajuste global, sino también en robustez econométrica, consolidándose como la especificación lineal más sólida antes de avanzar hacia modelos no lineales.

10.1.3. Significancia estadística e interpretación de estimadores

10.1.3.1 `OLS_BASE`

Según los coeficientes obtenidos en el modelo `OLS_BASE` en la tabla 7, estos permiten identificar los principales factores estructurales, de entorno y de accesibilidad que explican las variaciones en el precio de las viviendas del Gran Santiago. Los resultados, expresados en efectos porcentuales y calculados con corrección robusta HC3, muestran un comportamiento coherente con la teoría del valor hedónico, donde tanto las características físicas del inmueble como su localización influyen de manera directa en la disposición a pagar observada en el mercado.

Dentro del grupo de variables estructurales, destacan las de mayor significancia estadística y magnitud de efecto. La orientación del inmueble surge como un factor clave, evidenciando una diferencia de valoración muy marcada según su configuración: las viviendas pertenecientes al sector oriente presentan precios aproximadamente 76.4 % más altos, mientras que las de orientación a la zona sur de Santiago, muestran un efecto negativo de –16.4 %, lo que refleja una preferencia del mercado por edificaciones de menor densidad, con espacios más amplios y mejor disposición funcional. En la misma línea, la pertenencia a un condominio cerrado aumenta el valor del inmueble en 8.26 %, validando que la seguridad y el acceso controlado son atributos valorados por los compradores.

El tamaño del inmueble también se consolida como un determinante importante del precio. Tanto la superficie útil como la superficie total exhiben coeficientes positivos y altamente significativos, de 0.13 % y 0.14 % por metro cuadrado adicional, respectivamente. En términos prácticos, un aumento de 10 m^2 en cualquiera de estas superficies implica un incremento aproximado de 1.3

a 1.4 % en el valor del inmueble. De igual forma, el número de baños ejerce un efecto relevante, con un incremento promedio de 12.39 % en el precio por cada baño adicional, lo que refleja la influencia del confort y la funcionalidad interior en la valorización. En contraste, la antigüedad del inmueble muestra un efecto negativo y estadísticamente significativo, de –0.41 % por año, evidenciando la depreciación asociada al envejecimiento físico y tecnológico de las construcciones. Estos resultados confirman que los compradores capitalizan en el precio las características que aumentan el confort, el espacio y la habitabilidad, mientras penalizan la obsolescencia material.

Otras variables estructurales como dormitorios, estacionamientos o piscina presentan efectos positivos pero no alcanzan significancia al 5 %, probablemente por su alta correlación con las superficies o por la heterogeneidad de los tipos de vivienda incluidos en la muestra. No obstante, los elementos de equipamiento complementario como quincho (9.55 %) y bodega (6.16 %) sí resultan estadísticamente significativos, lo que sugiere que los espacios de esparcimiento y almacenamiento continúan siendo percibidos como atributos de valor añadido, especialmente en segmentos de vivienda media y alta.

En relación con las variables de localización y entorno urbano, el modelo presenta efectos coherentes aunque en general no significativos, lo que puede atribuirse a la colinealidad entre distancias a distintos servicios urbanos. La distancia a áreas verdes posee un coeficiente negativo (–0.67 %), lo que implica que a mayor distancia de parques o plazas, menor es el precio del inmueble, evidenciando la valorización de la calidad ambiental. De forma similar, la lejanía a centros comerciales (–0.34 %), supermercados (–0.33 %) o farmacias (–0.54 %) tiende a reducir el valor, aunque sin alcanzar significancia estadística. Esta falta de significancia no sugiere irrelevancia, sino redundancia explicativa, ya que muchas de estas variables describen patrones espaciales

superpuestos, lo que reduce la precisión individual de los coeficientes.

En cuanto a la accesibilidad, la variable de acceso a metro (transporte) muestra un efecto negativo y estadísticamente significativo (–0.0042 % por metro), confirmando que el precio disminuye conforme aumenta la distancia a una estación. Este resultado valida la importancia de la conectividad en la formación de precios urbanos, donde la red de transporte masivo representa un atributo de alta valorización. En contraste, la accesibilidad a buses tiene un efecto prácticamente nulo (0.0009 %) y no significativo, lo cual puede explicarse por la amplia cobertura de este servicio en la ciudad, que reduce las diferencias espaciales entre barrios. En el caso de los servicios educacionales, la educación escolar destaca como la única variable significativa (1.37 %), mostrando que la cercanía a establecimientos de enseñanza básica y media incrementa el valor del inmueble, mientras que la educación preescolar y superior mantienen signos positivos pero no significativos.

En síntesis, el modelo `OLS_BASE` confirma que el precio de la vivienda en el Gran Santiago está determinado principalmente por los atributos físicos del inmueble, mientras que las variables del entorno y los servicios presentan efectos coherentes pero con menor precisión estadística, debido a su intercorrelación. Aun así, el modelo conserva un alto valor explicativo, pues permite comprender cómo la superficie, las amenidades, la antigüedad y la conectividad estructuran el valor del suelo urbano. Es importante subrayar que el propósito de este modelo es explicativo más que predictivo, por lo que se mantienen variables con baja significancia estadística. Estas no buscan mejorar la precisión del pronóstico, sino aportar comprensión sobre la dirección y magnitud de las relaciones entre las características físicas, urbanas y de servicios. Esta base interpretativa servirá de referencia para el desarrollo del modelo `OLS_IDX`, donde la información redundante del entorno será consolidada en índices sintéticos, con el objetivo de mejorar la robustez y estabilidad

econométrica sin perder capacidad explicativa.

10.1.3.2 `OLS_IDX`

El modelo `OLS_IDX` constituye una versión depurada y teóricamente más estructurada del modelo `OLS_BASE`, orientada a reducir la *multicolinealidad* y capturar de manera más precisa los efectos espaciales y de entorno urbano sobre el precio de la vivienda. En esta versión, el número de variables disminuye significativamente, ya que las distintas medidas de proximidad a servicios se agrupan en *índices compuestos*, y se incorporan *transformaciones funcionales* en la variable de antigüedad para representar de forma más realista la relación entre edad y valor del inmueble. Los resultados reflejan una mejora en la *parsimonia* y en la *estabilidad de los coeficientes*, manteniendo coherencia teórica y consistencia con la literatura hedónica.

A partir de los resultados de la tabla 16, en primer lugar, las variables estructurales continúan siendo las de mayor peso y significancia en la explicación del precio. La superficie total conserva un efecto positivo y altamente significativo, con un incremento de 0.25 % en el precio por cada metro cuadrado adicional, lo que implica que una vivienda con 10 m^2 más de superficie vale, en promedio, un 2.5 % más, manteniendo constantes las demás variables. De igual modo, la variable baños presenta un impacto importante, con un coeficiente de 21.58 %, que refleja cómo las mejoras en confort interior y distribución funcional se capitalizan en el valor de mercado. Estos resultados reafirman que los factores de espacio y equipamiento básico continúan siendo determinantes fundamentales en la formación del precio de la vivienda.

Respecto a la antigüedad, el modelo incorpora una especificación cuadrática mediante las variables `ANTIGÜEDAD_C` y `ANTIGÜEDADSQ`, con el fin de representar de manera más precisa la depreciación no lineal del inmueble. La antigüedad centrada se calculó restando la media (24

años) a cada observación, lo que permite interpretar el coeficiente lineal como el efecto marginal de la edad en torno al promedio de la muestra. El coeficiente lineal es negativo (–0.35 %), mientras que el cuadrático es positivo (0.02 %), ambos estadísticamente significativos. Esta combinación revela una relación en forma de "U" suavizada entre antigüedad y precio: las viviendas más nuevas experimentan una depreciación inicial más fuerte, que se atenúa con el tiempo hasta estabilizarse o incluso revertirse parcialmente en torno a los 20–25 años. En términos económicos, esto sugiere que los inmuebles ubicados en zonas consolidadas, pese a su mayor edad, pueden mantener o recuperar valor gracias a su localización o atributos constructivos. Esta especificación mejora sustantivamente la interpretación temporal del modelo respecto al `OLS_BASE`, donde se asumía una depreciación constante.

Entre los atributos complementarios, la variable condominio cerrado mantiene un efecto positivo y significativo, con un aumento del 12.08 % en el precio promedio respecto a viviendas sin este atributo. Este resultado reafirma el peso de la seguridad y la calidad del entorno interno en la valoración inmobiliaria.

En relación con las variables de localización geográfica, estas *dummies* se interpretan en comparación con la zona Sur-Poniente, que actúa como categoría base. Los resultados reflejan una clara segmentación territorial en los precios de la vivienda. El sector *Norte-Oriente* presenta un efecto positivo y altamente significativo de 65.22 %, lo que indica que, manteniendo constantes las demás características, las viviendas en esta zona valen aproximadamente dos tercios más que aquellas localizadas en el Sur-Poniente. Este resultado es coherente con la concentración de comunas de altos ingresos, mejor infraestructura y mayor accesibilidad percibida en ese cuadrante. Por el contrario, el sector *Sur-Centro* muestra un efecto negativo de –13.97 %, evidenciando una me-

nor valorización relativa respecto al Sur-Poniente, posiblemente por su mayor densidad, deterioro urbano o menor equipamiento. El *Norte-Centro* (8.85 %), *Norte-Poniente* (–4.55 %) y *Sur-Oriente* (4.20 %) no presentan efectos estadísticamente significativos, aunque sus signos mantienen coherencia con la distribución espacial del valor del suelo, reforzando la idea de que el patrón de precios urbanos en Santiago responde a un gradiente socioespacial marcado hacia el eje nor-oriente.

En cuanto a los índices urbanos construidos, el modelo muestra una mejora sustantiva en estabilidad y coherencia teórica. El *Índice de Amenidades* (`IDX_AMENITIES`) tiene un efecto positivo y significativo del 12.13 %, lo que indica que un aumento de 0.1 unidades en el índice —equivalente a una mejora del 10 % en la dotación de amenidades como piscina, quincho o bodega— incrementa el precio de la vivienda en aproximadamente 1.2 %. El *Índice de Salud y Comercio* (`IDX_SALCOM`) también resulta significativo, con un efecto del 7.61 %, evidenciando que la cercanía a equipamientos comerciales y servicios de salud tiene un impacto directo en el valor de las propiedades. Por su parte, el *Índice de Educación y Transporte* (`IDX_EDUC_TRANS`) presenta un efecto positivo (6.01 %), aunque con menor nivel de significancia, indicando que la accesibilidad combinada a educación y movilidad influye favorablemente, pero de forma más heterogénea según el contexto urbano. Finalmente, el *Índice de Áreas Verdes* (`IDX_VERDE`) muestra un signo positivo pero no significativo (0.26 %), lo que puede deberse a la distribución desigual de este tipo de espacios en la ciudad o a que su efecto se superpone con el de otras variables de entorno.

En términos globales, el modelo `OLS_IDX` evidencia una mayor robustez econométrica y una interpretación más realista del mercado inmobiliario. Las variables más influyentes (*superficie*, *baños*, *antigüedad*, *condominio cerrado*, *localización norte-oriente* y los *índices de entorno*

urbano) presentan signos coherentes y efectos significativos. Aunque algunos índices no son estadísticamente relevantes, se mantienen en el modelo debido a su rol explicativo, ya que aportan información sobre las relaciones estructurales entre accesibilidad, entorno y valorización urbana. Es importante subrayar que el propósito del modelo no es la predicción puntual del precio, sino la comprensión integral de los factores que lo determinan, motivo por el cual se conservan variables con sentido teórico aunque carezcan de significancia al 5 %.

En conjunto, los resultados del `OLS_IDX` demuestran que la incorporación de *índices compuestos* y transformaciones funcionales, junto con la corrección espacial basada en zonas de referencia, permite mejorar la estabilidad del modelo y representar de manera más precisa las dinámicas espaciales y temporales del valor residencial. Este modelo no solo mantiene la coherencia económica observada en el `OLS_BASE`, sino que además la amplía al integrar la complejidad territorial del Gran Santiago, mostrando cómo la interacción entre atributos físicos, antigüedad, amenidades y entorno urbano define el precio final de la vivienda.

10.1.4. Contraste técnico-explicativo de modelos de regresión

La comparación entre los modelos `OLS_BASE` y `OLS_IDX` permite evaluar cómo la introducción de transformaciones funcionales e índices compuestos mejora la capacidad explicativa y la estabilidad de las estimaciones en el análisis del precio de la vivienda. Si bien ambos modelos parten de la misma lógica hedónica, difieren en su estructura interna y propósito metodológico: el `OLS_BASE` busca capturar directamente los efectos individuales de cada variable, mientras que el `OLS_IDX` concentra la información redundante en componentes agregados que representan dimensiones urbanas más amplias. Esta distinción resulta clave para comprender los avances

obtenidos en términos de *parsimonia*, *significancia* y *coherencia económica*.

El modelo `OLS_BASE`, con un mayor número de variables, logra describir con detalle la influencia de los atributos físicos, de entorno y de accesibilidad, aunque presenta ciertas limitaciones derivadas de la colinealidad entre variables espaciales. Sus resultados muestran que las variables estructurales (especialmente *superficie*, *número de baños* y *antigüedad*) explican la mayor parte de la variabilidad del precio, mientras que los efectos asociados a localización y servicios urbanos tienden a diluirse por la redundancia de información. A pesar de ello, el modelo cumple adecuadamente su objetivo explicativo inicial, ya que permite identificar la dirección de los efectos y confirmar la coherencia de los signos teóricos. Sin embargo, su estructura amplia y la presencia de correlaciones internas reducen la precisión y estabilidad de los estimadores, generando intervalos de confianza más amplios y, en algunos casos, efectos no significativos pese a su relevancia conceptual.

En contraste, el modelo `OLS_IDX` representa una evolución metodológica significativa. Al agrupar las variables de entorno en índices compuestos de accesibilidad y equipamiento urbano, se reduce la multicolinealidad y se mejora la parsimonia del modelo sin sacrificar capacidad explicativa. Además, la inclusión de la *antigüedad centrada* y su término cuadrático aporta un refinamiento funcional que permite modelar con mayor realismo el proceso de depreciación, mostrando que esta no es constante, sino que se atenúa con el tiempo. Desde un punto de vista econométrico, el `OLS_IDX` exhibe una mayor estabilidad de los signos y significancia más robusta en los coeficientes clave, destacando especialmente las variables de *superficie*, *baños*, *condominio cerrado* y el conjunto de índices relacionados con *amenidades*, *salud-comercio* y *educación-transporte*.

La incorporación de las variables *dummy* de localización en referencia al sector Sur-Poniente

refuerza la coherencia espacial del modelo, evidenciando un *gradiente territorial* de precios en el Gran Santiago. El sector *Norte-Oriente*, con un incremento del 65 % respecto del sector *Sur-Poniente*, se consolida como la zona de mayor valorización, mientras que el *Sur-Centro* refleja un rezago de –13.9 % en comparación con el grupo base. Este patrón espacial resulta más claro y robusto en el `OLS_IDX`, donde las diferencias territoriales se capturan de forma sistemática y consistente con la estructura socioeconómica metropolitana.

En términos interpretativos, el `OLS_IDX` logra una mejor integración entre las dimensiones físicas, temporales y espaciales del precio, mientras que el `OLS_BASE`, pese a su nivel de detalle, presenta un exceso de variables que dificulta aislar los efectos individuales de los factores urbanos. La reducción del número de variables en el `OLS_IDX` no solo simplifica la lectura, sino que mejora la eficiencia estadística de las estimaciones. Desde la perspectiva del análisis de supuestos, ambos modelos cumplen adecuadamente los criterios de linealidad y homocedasticidad robusta; no obstante, el `OLS_IDX` presenta menores problemas de varianza residual y un comportamiento más estable ante los tests de heterocedasticidad y multicolinealidad, lo que refuerza su validez técnica.

En conjunto, puede afirmarse que el modelo `OLS_IDX` supera al `OLS_BASE` tanto en desempeño econométrico como en coherencia teórica. Su estructura más compacta y sus transformaciones funcionales permiten representar con mayor claridad las relaciones entre los atributos de la vivienda y su valor de mercado, ofreciendo una visión más realista del comportamiento urbano. No obstante, el `OLS_BASE` conserva valor interpretativo al mostrar de manera desagregada el peso individual de cada variable, sirviendo como base exploratoria indispensable. En consecuencia, el `OLS_IDX` se define como el modelo preferente para el análisis explicativo del precio de la vivienda en el Gran Santiago, dado que combina rigor econométrico, coherencia espacial y una

interpretación económica más precisa del fenómeno inmobiliario.

10.2 Modelos de aprendizaje automático

10.2.1. Fundamento de modelos

En esta sección se presentan los resultados de los modelos de auto-aprendizaje supervisado, aplicados con el objetivo de evaluar su capacidad predictiva en la estimación del precio de la vivienda y compararla con los modelos lineales previamente desarrollados. A diferencia de los modelos `OLS`, los algoritmos utilizados (`Random Forest` y `XGBoost`) no imponen una forma funcional predeterminada, lo que les permite capturar relaciones no lineales e interacciones complejas entre las variables estructurales y urbanas.

Ambos modelos se entrenaron sobre la misma base de datos empleada en el `OLS_IDX`, manteniendo las variables estandarizadas e incluyendo los índices compuestos de entorno urbano. En ambos casos, la variable dependiente correspondió al *logaritmo natural del precio de oferta*, y se aplicaron procedimientos de ajuste de hiperparámetros y validación cruzada para optimizar el desempeño y prevenir el sobreajuste.

El `Random Forest`, basado en la agregación de múltiples árboles de decisión, prioriza la estabilidad y la reducción de la varianza, mientras que el `XGBoost`, mediante un proceso secuencial de optimización por gradiente, busca minimizar el sesgo y mejorar la precisión del ajuste. Así, ambos modelos ofrecen perspectivas complementarias, donde el primero destaca por su robustez y menor sensibilidad a valores atípicos, y el segundo, por su eficiencia y capacidad para capturar interacciones más profundas entre variables.

El propósito de esta comparación no es únicamente medir la precisión predictiva, sino

también evaluar hasta qué punto los modelos de auto-aprendizaje logran representar con mayor realismo la complejidad del mercado inmobiliario del Gran Santiago, complementando la lectura explicativa derivada de los modelos lineales.

10.2.2. Análisis de desempeño

Los modelos de aprendizaje automático, `Random Forest (RF)` y `XGBoost (XGB)`, presentan un desempeño general sólido y comparable, con niveles de ajuste elevados y una estabilidad adecuada entre las fases de entrenamiento y prueba. En términos generales, ambos modelos logran capturar con precisión la estructura subyacente del precio de las viviendas, confirmando la existencia de relaciones no lineales y dependencias complejas entre las variables estructurales y de entorno.

En el caso del `Random Forest`, la configuración *Full* alcanzó un $R^2 = 0{,}799$ en el conjunto de prueba, con un *RMSE* de 1558.36 UF y un *MAE* de 1076.83 UF, mientras que la versión *Trimmed* obtuvo un $R^2 = 0{,}739$, con *RMSE* = 1614.76 UF y *MAE* = 1062.38 UF. Si bien las diferencias entre ambas configuraciones son moderadas, se aprecia una ligera pérdida de capacidad explicativa en el modelo reducido, aunque con una mejora marginal en el error absoluto medio. Esto sugiere que el modelo completo logra un equilibrio más favorable entre precisión y generalización, al aprovechar la totalidad de las variables explicativas sin incurrir en sobreajuste significativo.

El modelo `XGBoost` exhibe resultados muy próximos, alcanzando en su versión *Full* un $R^2 = 0{,}797$, *RMSE* = 1565.80 UF y *MAE* = 1077.62 UF, mientras que la versión *Trimmed* mejora levemente el error promedio con un *MAE* = 1014.18 UF y *RMSE* = 1471.75 UF, aunque con una

ligera disminución del R^2 a 0.783. Estas variaciones indican un comportamiento consistente y una capacidad adecuada para capturar patrones de heterogeneidad espacial y estructural en los precios, evidenciando la estabilidad del modelo incluso tras la eliminación de observaciones extremas.

La comparación entre ambos algoritmos muestra que no existen diferencias sustanciales en su poder predictivo global, lo que demuestra que la información disponible y los índices construidos ofrecen un alto grado de coherencia. Sin embargo, se aprecian matices en el comportamiento de cada modelo: el `Random Forest` presenta una mayor estabilidad y una menor sensibilidad a cambios en los datos, mientras que `XGBoost` tiende a ofrecer errores promedio más bajos y una mejor gestión de la variabilidad residual, especialmente en el conjunto reducido.

En conjunto, ambos modelos mantienen valores de R^2 cercanos a 0.8, lo que representa una capacidad explicativa significativamente superior a la de los modelos `OLS`, y confirman que la estructura no lineal y las interacciones entre variables son relevantes para la predicción del precio de la vivienda. La reducción moderada en las métricas de prueba respecto del entrenamiento indica una generalización adecuada, descartando la presencia de sobreajuste severo.

En síntesis, tanto el `Random Forest` como el `XGBoost` demuestran un buen equilibrio entre precisión y robustez, validando el potencial de los modelos de autoaprendizaje para complementar los enfoques econométricos tradicionales.

10.2.3. Importancia y contribución de variables

El análisis de importancia de variables en los modelos de autoaprendizaje permite identificar los factores con mayor peso en la predicción del precio de la vivienda. Tanto en `Random Forest (RF)` como en `XGBoost (XGB)`, los resultados muestran una alta coherencia con

los modelos lineales, aunque con mayor capacidad para capturar relaciones no lineales entre los atributos.

En el caso del `Random Forest`, las variables más influyentes son `Superficie_Total`, `Baños`, `Antigüedad_centrada` y `Condominio_cerrado`. Estas variables reflejan los principales atributos estructurales del inmueble, confirmando que el tamaño y el nivel de equipamiento interior siguen siendo los factores determinantes del valor residencial. La antigüedad conserva un efecto relevante pero moderado, coherente con una depreciación no lineal que se estabiliza en viviendas más consolidadas.

Entre las variables urbanas, el `Índice_de_Amenidades (IDX_AMENITIES)` y el `Índice_de_Salud_y_Comercio (IDX_SALCOM)` también destacan, aunque con menor peso que los atributos físicos. Esto evidencia que la presencia de servicios y equipamientos en el entorno inmediato tiene una incidencia directa en la valorización del suelo, aunque subordinada a las características internas del inmueble.

Por su parte, el modelo `XGBoost` mantiene un patrón similar, pero distribuye la importancia de manera más equilibrada entre atributos físicos y de entorno. La `Superficie_Total` continúa siendo la variable dominante, seguida de `Baños` y `Antigüedad_centrada`, pero el modelo otorga una mayor ponderación a los índices urbanos, en especial al `IDX_SALCOM`, lo que sugiere una mejor capacidad para reconocer el valor que los compradores asignan a la accesibilidad y la proximidad a servicios.

En síntesis, ambos modelos confirman que la `Superficie_Total` y los `Baños` son los principales determinantes del precio, mientras que los índices de *accesibilidad* y *amenidades*

refuerzan el papel del entorno urbano en la valorización inmobiliaria. El `XGBoost` logra una interpretación más equilibrada y revela con mayor claridad la contribución del contexto urbano, evidenciando su superior capacidad para modelar relaciones complejas entre variables físicas y espaciales.

10.2.4. Contraste técnico-predictivo de modelos de aprendizaje automático

La comparación entre los modelos de autoaprendizaje `Random Forest (RF)` y `XGBoost (XGB)` permite evaluar cuál ofrece un mejor equilibrio entre capacidad predictiva, estabilidad y coherencia con los patrones observados en el mercado inmobiliario. Ambos modelos presentan métricas de desempeño muy similares, con valores de R^2 cercanos a 0.8 y errores medios entre 1.000 y 1.600 UF, lo que confirma un alto nivel de ajuste y una adecuada capacidad de generalización.

En términos de desempeño global, el `Random Forest` muestra una ligera ventaja en estabilidad entre los conjuntos de entrenamiento y prueba, reflejando un comportamiento más robusto frente a la variabilidad de los datos. Su estructura de múltiples árboles de decisión paralelos permite reducir la varianza y evitar sobreajustes, logrando una predicción más consistente. No obstante, tiende a concentrar la importancia de las variables en pocos predictores dominantes (como `Superficie_Total` y `Baños`), lo que limita su capacidad para reflejar interacciones más sutiles entre los factores urbanos.

Por otro lado, el `XGBoost` destaca por su eficiencia y menor error promedio, especialmente en la configuración *Trimmed*, donde logra un *RMSE* de 1.471 UF y un *MAE* de 1.014 UF, valores levemente mejores que los del `RF`. Además, el modelo distribuye la relevancia de manera más equilibrada entre las variables estructurales y de entorno, evidenciando una comprensión

más profunda del efecto combinado de accesibilidad, servicios y atributos físicos sobre el precio de la vivienda. Este comportamiento refleja su capacidad para capturar interacciones no lineales y dependencias complejas, lo que se traduce en un ajuste más fino de las predicciones.

Ambos modelos presentan una precisión en predicción superior a los enfoques lineales `OLS`, validando la utilidad del aprendizaje automático en la estimación de precios inmobiliarios. Sin embargo, considerando el conjunto de métricas, la estabilidad del ajuste y la distribución más realista de la importancia de las variables, el `XGBoost` se posiciona como el modelo con mejor desempeño predictivo global. Su menor error medio y su mayor capacidad para representar relaciones estructurales y espaciales complejas lo convierten en la alternativa más robusta y eficiente para la predicción del valor de la vivienda en el Gran Santiago.

10.3 Contraste modelo explicativo y predictivo

La comparación entre el modelo econométrico `OLS_IDX` y el modelo de auto-aprendizaje `XGBoost` permite contrastar dos enfoques analíticos distintos para abordar el mismo fenómeno: la formación del precio de la vivienda. Ambos modelos parten de la misma base de datos y variables explicativas, pero difieren sustancialmente en su naturaleza, objetivos y forma de representar las relaciones entre las variables.

El modelo `OLS_IDX`, de carácter paramétrico, asume una relación funcional explícita entre los atributos y el precio. Su principal fortaleza radica en la interpretabilidad y transparencia de los coeficientes, lo que permite cuantificar con claridad el efecto marginal de cada variable sobre el valor de la vivienda. A través de esta estructura, el `OLS_IDX` ofrece una lectura causal coherente con la teoría del valor hedónico, donde el precio se entiende como una suma ponderada de caracte-

rísticas físicas, de localización y de entorno urbano. Además, la introducción de índices sintéticos y de una función cuadrática de antigüedad otorga realismo sin perder control analítico. Sin embargo, este modelo presenta limitaciones al imponer una forma lineal y al asumir independencia entre variables, lo que puede restringir su capacidad para captar interacciones o comportamientos no lineales presentes en el mercado inmobiliario.

En cambio, el modelo `XGBoost` (no paramétrico y basado en técnicas de ensamblaje por gradiente) carece de una estructura funcional predefinida. Su fortaleza reside en la flexibilidad para descubrir patrones complejos en los datos y en su habilidad para representar relaciones no lineales entre los atributos físicos, la localización y los índices urbanos. Este enfoque se centra en la precisión predictiva más que en la interpretación directa de los efectos, lo que le permite adaptarse mejor a mercados heterogéneos y dinámicos. Sin embargo, esta misma flexibilidad conlleva una menor transparencia, ya que los resultados no se expresan en coeficientes fácilmente interpretables, sino en términos de importancia relativa de las variables, lo que dificulta establecer relaciones causales explícitas.

Desde una perspectiva práctica, el `OLS_IDX` aporta un conocimiento estructurado del mercado, ideal para análisis explicativos y formulación de políticas urbanas, mientras que el `XGBoost` resulta más adecuado para aplicaciones predictivas y valuaciones automatizadas donde la precisión del resultado es prioritaria. En términos de robustez, el `OLS_IDX` ofrece mayor control estadístico y verificación de supuestos, mientras que el `XGBoost` proporciona una capacidad superior para adaptarse a datos con heterogeneidad espacial y relaciones interdependientes.

En síntesis, ambos modelos se complementan más que se oponen. El `OLS_IDX` representa una herramienta de comprensión causal del fenómeno urbano, útil para explicar cómo y por qué

los atributos afectan el precio, mientras que el modelo `XGBoost` ofrece una herramienta empírica avanzada para anticipar el valor de mercado con alta precisión. En conjunto, sus resultados demuestran que la integración de enfoques econométricos y de aprendizaje automático permite una visión más completa y equilibrada del mercado inmobiliario, combinando interpretabilidad, rigor estadístico y capacidad predictiva.

10.4 Análisis dinámico de predicción de precios por zona

El análisis dinámico de precios por zona busca profundizar en la interpretación espacial de los resultados obtenidos, evaluando cómo varía el valor estimado de una vivienda al cambiar su localización dentro del Gran Santiago. Esta etapa permite cuantificar las diferencias territoriales o "primas de localización" que persisten aun controlando por atributos físicos y de entorno, complementando la visión general entregada por los modelos base.

Para ello se emplean los modelos `OLS_IDX` y `XGBoost`, seleccionados por haber mostrado el mejor desempeño dentro de sus respectivas categorías. El primero destaca por su solidez estadística y capacidad explicativa, permitiendo identificar efectos marginales claros sobre el precio; el segundo sobresale por su precisión predictiva y su flexibilidad para capturar relaciones no lineales e interacciones entre variables.

El objetivo general de este análisis es contrastar cómo ambos enfoques representan las variaciones espaciales del mercado habitacional, distinguiendo entre efectos de localización pura y diferencias derivadas de la composición de atributos en cada zona. De esta forma, se busca aportar una visión más completa sobre la estructura de precios de la vivienda en Santiago y la coherencia de los modelos al proyectar escenarios geográficos.

10.5 Vivienda global por zona

Este análisis busca aislar el efecto de la localización en el precio de la vivienda, manteniendo constantes las demás características. Para ello, se define una vivienda tipo global, cuyos atributos corresponden a los valores promedio o más frecuentes de la base de datos. Con esta referencia se estiman precios para cada zona, activando una dummy zonal a la vez y dejando las demás en cero. Así, se observa cómo varía el precio exclusivamente por ubicación.

10.5.1. Resultados modelo `OLS_IDX`

Tabla 45: Predicción del precio promedio por zona (modelo `OLS_IDX`

Zona	Predicción Log(Precio)	Predicción Precio [UF]
Norte Poniente	8.064	$3,299
Norte Oriente	8.613	$5,710
Norte Centro	8.196	$3,762
Sur Oriente	8.152	$3,601
Sur Poniente	8.111	$3,456
Sur Centro	7.961	$2,973

Los resultados obtenidos con el modelo OLS_IDX muestran una clara diferenciación territorial en los precios estimados para la vivienda tipo. Las zonas Norte Oriente y Norte Centro registran los valores más altos, con precios promedio de aproximadamente 5.710 UF y 3.762 UF, respectivamente, lo que refleja su mayor valorización dentro del mercado. En contraste, las zonas Sur Centro y Norte Poniente presentan las estimaciones más bajas, cercanas a 2.973 UF y 3.299 UF, evidenciando una menor disposición a pagar en dichos sectores.

En términos generales, el modelo lineal confirma la existencia de un gradiente espacial

donde las áreas con mejor accesibilidad, estatus, seguridad y equipamiento urbano concentran mayores precios, manteniendo constante el resto de las características de la vivienda.

10.5.2. Resultados modelo `XGBoost`

Tabla 46: Predicción del precio promedio por zona (modelo XGBoost)

Zona	Predicción Log(Precio)	Predicción Precio [UF]
Norte Poniente	8.119	$3,358
Norte Oriente	8.561	$5,224
Norte Centro	8.150	$3,464
Sur Oriente	8.168	$3,527
Sur Poniente	8.148	$3,457
Sur Centro	8.070	$3,196

El modelo `XGBoost` presenta un patrón territorial similar al obtenido con el `OLS_IDX`, aunque con diferencias en magnitud. La zona Norte Oriente alcanza nuevamente el mayor valor estimado, con un precio promedio cercano a 5.224 UF, seguida por Sur Oriente y Norte Centro, ambas en torno a las 3.500 UF. En contraste, Sur Centro registra el valor más bajo, con aproximadamente 3.196 UF, manteniendo la tendencia esperada de menor valorización hacia el sur del área metropolitana.

El comportamiento del modelo evidencia que, pese a capturar relaciones no lineales, las jerarquías espaciales se mantienen consistentes, donde las zonas con mayor nivel socioeconómico y mejor dotación de servicios continúan concentrando los precios más altos. Esto refuerza la estabilidad del patrón espacial y la coherencia predictiva del modelo no lineal respecto al enfoque econométrico.

10.5.3. Comparación de análisis de predicción

La comparación entre los modelos `OLS_IDX` y `XGBoost` aplicada a la vivienda tipo global muestra una notable consistencia en los patrones espaciales del precio, aunque con diferencias en la amplitud y el comportamiento de las estimaciones. Ambos modelos confirman que la Zona Norte Oriente concentra las viviendas de mayor valor, seguida por Norte Centro, mientras que Sur Centro y Norte Poniente corresponden a los sectores de menor precio estimado. Esta coincidencia en el ordenamiento territorial respalda la robustez del efecto de localización como determinante del valor inmobiliario en el Gran Santiago.

El modelo `OLS_IDX`, de naturaleza lineal, tiende a generar una mayor dispersión entre zonas, reflejando contrastes más marcados entre áreas de alto y bajo valor. Esto se debe a su estructura paramétrica, que asume relaciones proporcionales y constantes entre variables, acentuando las diferencias en el intercepto cuando las *dummies* zonales capturan efectos fijos de localización. Por el contrario, el modelo `XGBoost` suaviza las variaciones extremas y produce una distribución más continua de precios, resultado de su capacidad para capturar interacciones no lineales y ajustar mejor los puntos intermedios de la realidad urbana.

En términos interpretativos, el modelo `OLS_IDX` aporta una comprensión más directa de la influencia de cada atributo y del peso marginal de las zonas, lo que facilita la explicación del fenómeno económico. Sin embargo, su rigidez puede limitar la representación de dinámicas más complejas del mercado. El `XGBoost`, por su parte, ofrece mayor precisión predictiva y una respuesta más realista ante la variabilidad del espacio urbano, aunque a costa de menor transparencia en la interpretación.

Ambos enfoques, en conjunto, muestran que el gradiente de precios dentro del Gran Santiago es estable y se mantiene independiente del método utilizado. Esta coincidencia sugiere que las desigualdades territoriales en la valorización residencial no son artefactos estadísticos, sino expresiones persistentes de las condiciones estructurales del territorio metropolitano.

10.6 Viviendas representativas por zona

Este segundo análisis busca incorporar la heterogeneidad interna de cada zona, utilizando viviendas representativas por territorio en lugar de una sola vivienda tipo global. Para ello, se calculan los valores promedio o modales de los principales atributos dentro de cada zona, generando así una propiedad característica que refleja las condiciones típicas del mercado local.

El objetivo es estimar el precio esperado de cada vivienda representativa según los modelos `OLS_IDX` y `XGBoost`, y analizar cómo las diferencias en atributos propios de cada área contribuyen a explicar las brechas de precio observadas entre zonas.

10.6.1. Resultados modelo `OLS_IDX`

Tabla 47: Predicción del precio promedio por zona (modelo `OLS_IDX`)

Zona	Predicción Log(Precio)	Predicción Precio [UF]
Norte Poniente	8.107	$3,442
Norte Oriente	8.754	$6,575
Norte Centro	8.299	$4,170
Sur Oriente	8.197	$3,765
Sur Poniente	8.111	$3,456
Sur Centro	7.991	$3,067

Los resultados del modelo `OLS_IDX` para las viviendas representativas por zona muestran un gradiente espacial más acentuado que en el análisis anterior. La zona Norte Oriente presenta el mayor valor estimado, con aproximadamente 6.575 UF, seguida por Norte Centro (4.170 UF) y Sur Oriente (3.765 UF). En el extremo opuesto, las zonas Sur Centro y Sur Poniente alcanzan valores cercanos a 3.067 UF y 3.456 UF, respectivamente.

Estas diferencias reflejan la influencia conjunta de los atributos propios de cada zona (superficie, antigüedad y entorno urbano) además del efecto de localización. El modelo lineal mantiene la tendencia observada previamente, destacando el predominio de las zonas nor-oriente y centro-norte como las áreas de mayor valorización inmobiliaria dentro del Gran Santiago.

10.6.2. Resultados modelo `XGBoost`

Tabla 48: Predicción del precio promedio por zona (modelo `XGBoost`)

Zona	Pred_model_scale	Predicción Precio [UF]
Norte Poniente	8.082	$3,234
Norte Oriente	8.689	$5,936
Norte Centro	8.252	$3,836
Sur Oriente	8.192	$3,614
Sur Poniente	8.067	$3,186
Sur Centro	8.162	$3,507

El modelo `XGBoost` reproduce una estructura territorial similar, aunque con variaciones más suaves en los niveles de precio. La zona Norte Oriente vuelve a liderar con un valor medio de 5.936 UF, mientras que Norte Centro y Sur Oriente se ubican en torno a 3.800–3.600 UF. En contraste, las zonas Sur Poniente y Norte Poniente presentan los valores más bajos, próximos a 3.200 UF.

Al igual que en el caso anterior, el modelo de aprendizaje automático confirma la jerarquía espacial del mercado, destacando la fuerte concentración de precios en el eje nor-oriente. No obstante, la menor dispersión entre zonas sugiere que `XGBoost` atenúa las diferencias extremas, posiblemente al incorporar relaciones no lineales y efectos de interacción entre atributos urbanos y territoriales.

10.6.3. Comparación de análisis de predicción

La comparación entre los modelos `OLS_IDX` y `XGBoost` a partir de las viviendas representativas por zona evidencia un alto grado de coherencia en la jerarquía espacial de precios, pero también diferencias relevantes en la magnitud y dispersión de las estimaciones. Ambos modelos coinciden en identificar a la zona Norte Oriente como el sector de mayor valorización del Gran Santiago, seguida por Norte Centro y Sur Oriente, mientras que las zonas Sur Poniente, Norte Poniente y Sur Centro conforman el grupo de menor precio esperado. Este patrón confirma la persistencia de un gradiente norte-sur en el mercado inmobiliario, estrechamente asociado a la distribución del ingreso, la infraestructura y el acceso a servicios urbanos.

El modelo `OLS_IDX` tiende a generar brechas más amplias entre zonas, particularmente al asignar un valor de 6.575 UF a la zona Norte Oriente frente a solo 3.067 UF en Sur Centro, lo que implica una diferencia superior al 110 %. Este comportamiento responde a la estructura lineal del modelo, que asume efectos marginales constantes y puede amplificar contrastes en presencia de variables con alta correlación espacial. En cambio, el `XGBoost` modera estas diferencias, estimando un rango más acotado (de 5.936 UF a 3.186 UF), lo cual sugiere una mejor capacidad para capturar relaciones no lineales y compensaciones entre atributos físicos y territoriales.

Desde un punto de vista interpretativo, el modelo `OLS_IDX` resulta más transparente al permitir identificar el peso individual de cada variable, lo que facilita la explicación económica del fenómeno. Sin embargo, su estructura rígida limita la representación de interacciones complejas, especialmente entre indicadores del entorno y atributos constructivos. Por su parte, el `XGBoost`, aunque menos interpretativo, ofrece una mayor precisión predictiva y refleja con mayor realismo la continuidad del espacio urbano, evitando saltos bruscos entre zonas contiguas.

En conjunto, ambos modelos confirman la solidez del patrón espacial del precio de la vivienda y se complementan en su aporte analítico: el `OLS_IDX` permite entender los mecanismos estructurales detrás de la formación del precio, mientras que el `XGBoost` reproduce con mayor fidelidad las variaciones reales observadas en el mercado. La convergencia de resultados entre ambos enfoques refuerza la validez del análisis y demuestra que las diferencias de valorización territorial responden a dinámicas persistentes más que a artefactos metodológicos.

11 Conclusiones

11.1 Determinantes claves en el precio

Los resultados convergen en que el precio de oferta de las casas en el Gran Santiago (UF, log-precio) se explica principalmente por un núcleo de atributos internos, de entorno y territoriales. Entre los internos, destacan la superficie útil/total y el programa (dormitorios o baños), junto con estacionamientos y equipamientos (piscina, quincho, bodega) y la antigüedad presenta un efecto negativo cuando no va acompañada de mejoras.

En el entorno, la accesibilidad y la cercanía a servicios y amenities (salud, educación en sus distintos niveles, áreas verdes, farmacias, supermercados y centros comerciales) se capitalizan en el precio. En lo territorial, la localización por zonificación logra capturar brechas espaciales sistemáticas. Estas señales aparecen tanto en los efectos marginales de ambos modelos `OLS` como en la importancia de variables de `Random Forest` y `XGBoost`, lo que refuerza la consistencia entre enfoques explicativos y predictivos.

11.2 Modelos explicativos

El contraste entre las especificaciones hedónicas muestra que `OLS_BASE` ofrece una lectura más "granular" de los efectos marginales al incluir explícitamente atributos internos y de accesibilidad, lo que facilita interpretar elasticidades/semielasticidades por variable; en cambio, `OLS_IDX` prioriza parsimonia y estabilidad al agrupar covariables en índices compuestos (accesibilidad/amenities) y aplicar transformaciones funcionales (antigüedad), reduciendo riesgos de multicolinealidad y mejorando la robustez de la estimación sin perder capacidad explicativa relevante. En ambos casos, los diagnósticos respaldan la validez del enfoque (heterocedasticidad

tratada con HC3, revisión de VIF, chequeo de residuos e influencia y validación cruzada), y el desempeño out-of-sample resulta coherente con el grado de complejidad de cada especificación. En síntesis, cuando el objetivo central es explicar y comunicar efectos individuales, `OLS_BASE` es preferible; cuando se busca estabilidad con modelos más compactos con menor colinealidad y mayor respaldo en desempeño de predicción, `OLS_IDX` resulta más adecuado.

11.3 Modelos de aprendizaje automático

Los dos enfoques de ensambles mejoran el error de predicción fuera de muestra frente a los modelos `OLS` y muestran consistencia en la señal de variables relevantes. El capítulo de resultados reporta, para ambas familias, configuraciones Full y Trimmed, con métricas in/out-of-sample, gráficos de predicho vs. real y distribución de residuales, además de tablas de importancia de variables.

En términos comparativos, `XGBoost` tiende a ofrecer un ligero mejor desempeño y calibración en las mejores configuraciones, favorecido por su regularización y *early stopping*, mientras que `Random Forest` destaca por su robustez y baja sensibilidad a especificaciones, con resultados estables aun con podas/recortes de variables. Las diferencias entre ambos son acotadas y dependen de la selección de hiperparámetros y del set de features utilizado (Full vs. Trimmed), pero el patrón general es que ambos superan a OLS en error out-of-sample, con `XGBoost` ligeramente por delante en el mejor caso reportado debido a sus estadísticas de desempeño.

11.4 Modelos explicativos vs. predictivos

La evidencia separa con claridad qué modelo usar según el objetivo. Cuando el propósito es explicar y comunicar efectos marginales (esto es, cuánto cambia el precio ante variaciones en

atributos específicos) los modelos hedónicos (`OLS_BASE` y `OLS_IDX`) resultan preferibles por su interpretabilidad directa (coeficientes como semi/elasticidades) y por sus diagnósticos transparentes (HC3, VIF, residuos e influencia).

Sin embargo, cuando el objetivo es predecir con el menor error posible para tareas de pricing operativo o valoración a escala, los ensambles como `Random Forest` y, en el mejor caso, `XGBoost`, rinden mejor out-of-sample, con importancias de variables que apoyan la lectura sustantiva aunque con menor trazabilidad causal.

En la práctica, ambos enfoques son complementarios, usar `OLS` para extraer variables interpretables y modelos de auto-aprendizaje para desplegar predicciones más precisas, integrando sus salidas en un flujo de trabajo coherente.

11.5 Heterogeneidad territorial

En Santiago no todas las zonas valen lo mismo porque no ofrecen lo mismo. En las zonas nor-oriente y norte-centro se concentran hogares altos ingresos y, en términos externos, hay mejor conectividad (tiempos de viaje más cortos, mayor cercanía a Metro y ejes viales) y existe una dotación más alta de servicios valorados por las familias (colegios, salud, comercio y áreas verdes de calidad). Además, la percepción de seguridad y el prestigio barrial suelen ser más altos. Todo eso eleva la disposición a pagar y, por lo tanto, el precio del suelo y de las viviendas.

En el caso de las zonas nor-poniente y sur-centro, en cambio, los ingresos promedio son menores y la accesibilidad y los servicios no alcanzan el mismo nivel, por lo que la misma casa termina costando bastante menos.

Los mejores modelos de cada enfoque de estudio (`OLS_IDX` y `XGBoost`) capturan parte

importante de estas diferencias con variables como distancia a Metro y servicios, dummies/zonificación comunal e índices de accesibilidad/amenities. Sin embargo, hay factores más complejos de medir, como reputación del barrio o expectativas de seguridad, que también causan variaciones considerables en el precio hedónico de la vivienda, es por esto, que la evidencia empírica denota que una vivienda con exactamente las mismas carácteristicas fisicas y de entorno puede variar su valor en rangos de [-20 %, +60 %] según la zona en la que esta se ubica.

11.6 Calidad y robustez del análisis

Finalmente, los resultados obtenidos de manera empírica pueden ser considerados confiables ya que los modelos fueron probados con los chequeos técnicos y validaciones recomendadas por la literatura. En las regresiones `OLS` se revisaron supuestos y se usaron errores robustos (HC3) para lidiar con heterocedasticidad, además se controló la multicolinealidad (VIF), se miraron residuos e influencias y se comparó el ajuste dentro y fuera de muestra.

En los modelos de aprendizaje automático se separaron datos de entrenamiento y validación, se ajustaron hiperparámetros y se evaluó el desempeño con métricas estándar (RMSE, MAE, MAPE), mostrando consistencia entre configuraciones “Full” y “Trimmed”. En conjunto, estas prácticas indican que los hallazgos son estables y que las conclusiones no dependen de una sola especificación o partición de datos.

11.7 Implicancias y utilidad práctica

Finalmente, los hallazgos de esta memoria y este tipo de estudios en general son directamente accionables para distintos actores del mercado inmobiliario.

Para compradores y propietarios, los efectos marginales estimados ayudan a valorar con realismo la prima por atributos internos (por ejemplo, sumar un baño, pertenecer a condominio cerrado o mejorar equipamiento) y por localización (accesibilidad y dotación de servicios), mejorando negociación y priorización de mejoras. Para inversionistas y desarrolladores, los modelos predictivos permiten construir bandas de precio y escenarios por zona, apoyando decisiones de localización, mix de producto y timing; la combinación de `XGBoost` (menor error) y `Random Forest` (mayor estabilidad) ofrece precisión operativa con robustez.

Para tasadores, bancos y aseguradoras, las regresiones `OLS` entregan trazabilidad y justificación de diferencias de precio mediante coeficientes interpretables, mientras que los ensambles proveen cotizaciones rápidas con errores esperados conocidos (RMSE/MAE de prueba), útiles para fijar umbrales de riesgo y ajustar políticas de crédito. Para organismos públicos y equipos de estudio, las elasticidades y la heterogeneidad espacial cuantifican cómo la inversión en transporte, seguridad y áreas verdes incide en el valor residencial, orientando priorización territorial y evaluación de impactos.

En la práctica, se sugiere un flujo dual: una capa explicativa con `OLS` para comunicar efectos marginales y una capa predictiva con `XGBoost`/`Random Forest` para pricing masivo y simulación de escenarios. La implementación en un tablero con actualización periódica, recalibración temporal/espacial y validación contra precios de cierre mitigará de mejor manera los errores de predicción y mejorará la utilidad del sistema para mercado, políticas y evaluación de proyectos.

12 Limitaciones

12.1 Uso de precios de oferta y no de cierre

La base emplea precios publicados en los portales, que reflejan expectativas de los vendedores y pueden diferir de los precios efectivamente transados. No fue posible acceder a datos de cierre por dificultad de hallazgo, restricciones de disponibilidad pública y confidencialidad de registros notariales/bancarios. Se justifica su uso porque los precios de oferta son el insumo operativo del mercado visible para compradores y competidores, y permiten comparar alternativas en condiciones homogéneas dentro del período analizado.

12.2 Diseño observacional y de corte transversal

El estudio es no experimental y analiza un corte temporal, por lo que identifica asociaciones (efectos marginales) pero no causalidad estricta. Implementar diseños cuasi-experimentales o paneles requeriría series temporales de transacciones georreferenciadas y eventos exógenos verificables mejoraría aún más las métricas y la comprensión de los impactos marginales, sin embargo dichos datos no estuvieron disponibles en esta etapa. Se justifica porque parte del objetivo principal es la comparativa explicativo/predictivo para tasación y pricing, donde la precisión fuera de muestra y la interpretabilidad son más relevantes que la identificación causal.

12.3 Variables no observadas o medidas parcialmente

Factores como reputación barrial, percepción de seguridad, calidad efectiva de colegios/salud o características constructivas finas (materialidades, terminaciones) pueden estar ausentes o sólo aproximados. No fue posible incorporar estas dimensiones por falta de fuentes públicas estandari-

zadas y por costos de integración (encuestas, inspecciones, datos privados). Se justifica recurriendo a proxies robustos (distancias a servicios, índices compuestos, dummies/zonificación) y complementando con modelos de aprendizaje automático que capturan no linealidades e interacciones no especificadas.

13 Anexos

13.1 Tablas de variables definidas para procesos de predicción

Tabla 49: Variables definidas para predicción vivienda global (`OLS_IDX`)

Variable	Valor
DUMMY_CONDOMINIO_CERRADO	0
ANTIGÜEDAD_C	0.452483801
ANTIGÜEDADSQ	132.4547416
SUPERFICIE	157
BAÑOS	2
IDX_EDUC_TRANS	-0.036883728
IDX_SALRET	-0.087788368
IDX_VERDE	-0.157392648
IDX_AMENITIES	-0.221761116

Tabla 50: Variables definidas para predicción vivienda global (`XGBoost`)

Variable	Valor
ANTIGÜEDAD	24
SUPERFICIE_TOTAL	157
SUPERFICIE_UTIL	100
DORMITORIOS	3
BAÑOS	2
ESTACIONAMIENTOS	2
PISOS	2
PISCINA	0
QUINCHO	0
BODEGA	0
IDX_TRANS	-0.116106102
IDX_EDUC	-0.070156086
IDX_SALRET	-0.070991279
IDX_VERDE	-0.16513203
ENG_densidad_dorm	0.034188034
ENG_densidad_banos	0.02
ENG_ST_x_IDX_TRANS	-16.39004432
ENG_SU_x_IDX_TRANS	-11.76509355
ENG_ST_x_IDX_EDUC	-9.810453734
ENG_SU_x_IDX_EDUC	-7.243855581
ENG_ST_x_IDX_SALRET	-10.55529795
ENG_SU_x_IDX_SALRET	-8.10368846
ENG_ST_x_IDX_VERDE	-20.89383835
ENG_SU_x_IDX_VERDE	-12.20629324

Tabla 51: Variables definidas para predicción vivienda por zona (`OLS_IDX`)

Zona	Variable	Valor
Norte Poniente	DUMMY_NORTE_PONIENTE	1
Norte Centro	DUMMY_NORTE_PONIENTE	0
Norte Oriente	DUMMY_NORTE_PONIENTE	0
Sur Centro	DUMMY_NORTE_PONIENTE	0
Sur Oriente	DUMMY_NORTE_PONIENTE	0
Norte Poniente	DUMMY_NORTE_CENTRO	0
Norte Centro	DUMMY_NORTE_CENTRO	1

Zona	Variable	Valor
Norte Oriente	DUMMY_NORTE_CENTRO	0
Sur Centro	DUMMY_NORTE_CENTRO	0
Sur Oriente	DUMMY_NORTE_CENTRO	0
Norte Poniente	DUMMY_NORTE_ORIENTE	0
Norte Centro	DUMMY_NORTE_ORIENTE	0
Norte Oriente	DUMMY_NORTE_ORIENTE	1
Sur Centro	DUMMY_NORTE_ORIENTE	0
Sur Oriente	DUMMY_NORTE_ORIENTE	0
Norte Poniente	DUMMY_SUR_CENTRO	0
Norte Centro	DUMMY_SUR_CENTRO	0
Norte Oriente	DUMMY_SUR_CENTRO	0
Sur Centro	DUMMY_SUR_CENTRO	1
Sur Oriente	DUMMY_SUR_CENTRO	0
Norte Poniente	DUMMY_SUR_ORIENTE	0
Norte Centro	DUMMY_SUR_ORIENTE	0
Norte Oriente	DUMMY_SUR_ORIENTE	0
Sur Centro	DUMMY_SUR_ORIENTE	0
Sur Oriente	DUMMY_SUR_ORIENTE	1
Norte Poniente	DUMMY_CONDOMINIO_CERRADO	0
Norte Centro	DUMMY_CONDOMINIO_CERRADO	0
Norte Oriente	DUMMY_CONDOMINIO_CERRADO	1
Sur Centro	DUMMY_CONDOMINIO_CERRADO	0
Sur Oriente	DUMMY_CONDOMINIO_CERRADO	1
Norte Poniente	ANTIGÜEDAD_C	-9.047516199
Norte Centro	ANTIGÜEDAD_C	-11.0475162
Norte Oriente	ANTIGÜEDAD_C	6.952483801
Sur Centro	ANTIGÜEDAD_C	5.952483801
Sur Oriente	ANTIGÜEDAD_C	0.952483801
Norte Poniente	ANTIGÜEDADSQ	145.1426466
Norte Centro	ANTIGÜEDADSQ	197.3327114
Norte Oriente	ANTIGÜEDADSQ	167.7668366
Sur Centro	ANTIGÜEDADSQ	223.5767718
Sur Oriente	ANTIGÜEDADSQ	80.14696621
Norte Poniente	SUPERFICIE	158.5
Norte Centro	SUPERFICIE	172
Norte Oriente	SUPERFICIE	164.5
Sur Centro	SUPERFICIE	164
Sur Oriente	SUPERFICIE	140
Norte Poniente	BAÑOS	2
Norte Centro	BAÑOS	2
Norte Oriente	BAÑOS	2
Sur Centro	BAÑOS	2
Sur Oriente	BAÑOS	2

Zona	Variable	Valor
Norte Poniente	IDX_EDUC_TRANS	-0.079365374
Norte Centro	IDX_EDUC_TRANS	0.34861834
Norte Oriente	IDX_EDUC_TRANS	-0.042825848
Sur Centro	IDX_EDUC_TRANS	0.103268657
Sur Oriente	IDX_EDUC_TRANS	-0.175214034
Norte Poniente	IDX_SALRET	-0.064499321
Norte Centro	IDX_SALRET	-0.082097265
Norte Oriente	IDX_SALRET	0.163005616
Sur Centro	IDX_SALRET	-0.101126934
Sur Oriente	IDX_SALRET	-0.161185958
Norte Poniente	IDX_VERDE	-0.429226632
Norte Centro	IDX_VERDE	-0.442026782
Norte Oriente	IDX_VERDE	-0.011347261
Sur Centro	IDX_VERDE	0.157026075
Sur Oriente	IDX_VERDE	-0.379778375
Norte Poniente	IDX_AMENITIES	-0.221761116
Norte Centro	IDX_AMENITIES	-0.221761116
Norte Oriente	IDX_AMENITIES	-0.221761116
Sur Centro	IDX_AMENITIES	-0.221761116
Sur Oriente	IDX_AMENITIES	-0.221761116

Tabla 52: Variables definidas para predicción vivienda por zona (`XGBoost`)

Zona	Variable	Valor
Norte Poniente	DUMMY_NORTE_PONIENTE	1
Norte Centro	DUMMY_NORTE_PONIENTE	0
Norte Oriente	DUMMY_NORTE_PONIENTE	0
Sur Poniente	DUMMY_NORTE_PONIENTE	0
Sur Centro	DUMMY_NORTE_PONIENTE	0
Sur Oriente	DUMMY_NORTE_PONIENTE	0
Norte Poniente	DUMMY_NORTE_CENTRO	0
Norte Centro	DUMMY_NORTE_CENTRO	1
Norte Oriente	DUMMY_NORTE_CENTRO	0
Sur Poniente	DUMMY_NORTE_CENTRO	0
Sur Centro	DUMMY_NORTE_CENTRO	0
Sur Oriente	DUMMY_NORTE_CENTRO	0
Norte Poniente	DUMMY_NORTE_ORIENTE	0
Norte Centro	DUMMY_NORTE_ORIENTE	0
Norte Oriente	DUMMY_NORTE_ORIENTE	1
Sur Poniente	DUMMY_NORTE_ORIENTE	0
Sur Centro	DUMMY_NORTE_ORIENTE	0
Sur Oriente	DUMMY_NORTE_ORIENTE	0

Zona	Variable	Valor
Norte Poniente	DUMMY_SUR_PONIENTE	0
Norte Centro	DUMMY_SUR_PONIENTE	0
Norte Oriente	DUMMY_SUR_PONIENTE	0
Sur Poniente	DUMMY_SUR_PONIENTE	1
Sur Centro	DUMMY_SUR_PONIENTE	0
Sur Oriente	DUMMY_SUR_PONIENTE	0
Norte Poniente	DUMMY_SUR_CENTRO	0
Norte Centro	DUMMY_SUR_CENTRO	0
Norte Oriente	DUMMY_SUR_CENTRO	0
Sur Poniente	DUMMY_SUR_CENTRO	0
Sur Centro	DUMMY_SUR_CENTRO	1
Sur Oriente	DUMMY_SUR_CENTRO	0
Norte Poniente	DUMMY_SUR_ORIENTE	0
Norte Centro	DUMMY_SUR_ORIENTE	0
Norte Oriente	DUMMY_SUR_ORIENTE	0
Sur Poniente	DUMMY_SUR_ORIENTE	0
Sur Centro	DUMMY_SUR_ORIENTE	0
Sur Oriente	DUMMY_SUR_ORIENTE	1
Norte Poniente	DUMMY_CONDOMINIO_CERRADO	0
Norte Centro	DUMMY_CONDOMINIO_CERRADO	0
Norte Oriente	DUMMY_CONDOMINIO_CERRADO	1
Sur Poniente	DUMMY_CONDOMINIO_CERRADO	0
Sur Centro	DUMMY_CONDOMINIO_CERRADO	0
Sur Oriente	DUMMY_CONDOMINIO_CERRADO	1
Norte Poniente	ANTIGÜEDAD	15
Norte Centro	ANTIGÜEDAD	13
Norte Oriente	ANTIGÜEDAD	31
Sur Poniente	ANTIGÜEDAD	20
Sur Centro	ANTIGÜEDAD	30
Sur Oriente	ANTIGÜEDAD	25
Norte Poniente	SUPERFICIE_TOTAL	158.5
Norte Centro	SUPERFICIE_TOTAL	175
Norte Oriente	SUPERFICIE_TOTAL	164.5
Sur Poniente	SUPERFICIE_TOTAL	130
Sur Centro	SUPERFICIE_TOTAL	164
Sur Oriente	SUPERFICIE_TOTAL	140
Norte Poniente	SUPERFICIE_UTIL	95.5
Norte Centro	SUPERFICIE_UTIL	115
Norte Oriente	SUPERFICIE_UTIL	107
Sur Poniente	SUPERFICIE_UTIL	90
Sur Centro	SUPERFICIE_UTIL	105
Sur Oriente	SUPERFICIE_UTIL	92
Norte Poniente	DORMITORIOS	3.5

Zona	Variable	Valor
Norte Centro	DORMITORIOS	3
Norte Oriente	DORMITORIOS	4
Sur Poniente	DORMITORIOS	3
Sur Centro	DORMITORIOS	3
Sur Oriente	DORMITORIOS	3
Norte Poniente	BAÑOS	2
Norte Centro	BAÑOS	2
Norte Oriente	BAÑOS	2
Sur Poniente	BAÑOS	2
Sur Centro	BAÑOS	2
Sur Oriente	BAÑOS	2
Norte Poniente	ESTACIONAMIENTOS	1
Norte Centro	ESTACIONAMIENTOS	2
Norte Oriente	ESTACIONAMIENTOS	2
Sur Poniente	ESTACIONAMIENTOS	2
Sur Centro	ESTACIONAMIENTOS	2
Sur Oriente	ESTACIONAMIENTOS	2
Norte Poniente	PISOS	2
Norte Centro	PISOS	2
Norte Oriente	PISOS	2
Sur Poniente	PISOS	2
Sur Centro	PISOS	2
Sur Oriente	PISOS	2
Norte Poniente	PISCINA	0
Norte Centro	PISCINA	0
Norte Oriente	PISCINA	0
Sur Poniente	PISCINA	0
Sur Centro	PISCINA	0
Sur Oriente	PISCINA	0
Norte Poniente	QUINCHO	0
Norte Centro	QUINCHO	0
Norte Oriente	QUINCHO	0
Sur Poniente	QUINCHO	0
Sur Centro	QUINCHO	0
Sur Oriente	QUINCHO	0
Norte Poniente	BODEGA	0
Norte Centro	BODEGA	0
Norte Oriente	BODEGA	1
Sur Poniente	BODEGA	0
Sur Centro	BODEGA	1
Sur Oriente	BODEGA	1
Norte Poniente	IDX_TRANS	-0.205023885
Norte Centro	IDX_TRANS	0.27140089

Zona	Variable	Valor
Norte Oriente	IDX_TRANS	-0.058187061
Sur Poniente	IDX_TRANS	-0.381507294
Sur Centro	IDX_TRANS	0.058249951
Sur Oriente	IDX_TRANS	-0.283112846
Norte Poniente	IDX_EDUC	-0.119462138
Norte Centro	IDX_EDUC	0.03673914
Norte Oriente	IDX_EDUC	-0.074664943
Sur Poniente	IDX_EDUC	-0.248306835
Sur Centro	IDX_EDUC	0.060075154
Sur Oriente	IDX_EDUC	-0.063859905
Norte Poniente	IDX_SALRET	-0.061689659
Norte Centro	IDX_SALRET	-0.065738732
Norte Oriente	IDX_SALRET	0.146634947
Sur Poniente	IDX_SALRET	-0.22661515
Sur Centro	IDX_SALRET	-0.110026896
Sur Oriente	IDX_SALRET	-0.156148674
Norte Poniente	IDX_VERDE	-0.429226632
Norte Centro	IDX_VERDE	-0.455394946
Norte Oriente	IDX_VERDE	-0.011347261
Sur Poniente	IDX_VERDE	0.039102458
Sur Centro	IDX_VERDE	0.157026075
Sur Oriente	IDX_VERDE	-0.379778375
Norte Poniente	ENG_densidad_dorm	0.037980769
Norte Centro	ENG_densidad_dorm	0.030037594
Norte Oriente	ENG_densidad_dorm	0.033333333
Sur Poniente	ENG_densidad_dorm	0.035714286
Sur Centro	ENG_densidad_dorm	0.031578947
Sur Oriente	ENG_densidad_dorm	0.034782609
Norte Poniente	ENG_densidad_banos	0.019433962
Norte Centro	ENG_densidad_banos	0.015968742
Norte Oriente	ENG_densidad_banos	0.021428571
Sur Poniente	ENG_densidad_banos	0.022058824
Sur Centro	ENG_densidad_banos	0.015384615
Sur Oriente	ENG_densidad_banos	0.021428571
Norte Poniente	ENG_ST_x_IDX_TRANS	-27.74282063
Norte Centro	ENG_ST_x_IDX_TRANS	42.7289539
Norte Oriente	ENG_ST_x_IDX_TRANS	-6.654248514
Sur Poniente	ENG_ST_x_IDX_TRANS	-41.96580239
Sur Centro	ENG_ST_x_IDX_TRANS	9.319992145
Sur Oriente	ENG_ST_x_IDX_TRANS	-34.76910532
Norte Poniente	ENG_SU_x_IDX_TRANS	-19.43587879
Norte Centro	ENG_SU_x_IDX_TRANS	30.62823564
Norte Oriente	ENG_SU_x_IDX_TRANS	-5.568250988

Zona	Variable	Valor
Sur Poniente	ENG_SU_x_IDX_TRANS	-25.15121066
Sur Centro	ENG_SU_x_IDX_TRANS	5.533745336
Sur Oriente	ENG_SU_x_IDX_TRANS	-22.99808675
Norte Poniente	ENG_ST_x_IDX_EDUC	-13.69921305
Norte Centro	ENG_ST_x_IDX_EDUC	3.674197677
Norte Oriente	ENG_ST_x_IDX_EDUC	-11.52557872
Sur Poniente	ENG_ST_x_IDX_EDUC	-24.45086632
Sur Centro	ENG_ST_x_IDX_EDUC	12.41376087
Sur Oriente	ENG_ST_x_IDX_EDUC	-3.192995242
Norte Poniente	ENG_SU_x_IDX_EDUC	-8.611411019
Norte Centro	ENG_SU_x_IDX_EDUC	3.753398366
Norte Oriente	ENG_SU_x_IDX_EDUC	-9.107108899
Sur Poniente	ENG_SU_x_IDX_EDUC	-18.07697985
Sur Centro	ENG_SU_x_IDX_EDUC	6.768796393
Sur Oriente	ENG_SU_x_IDX_EDUC	-3.105502424
Norte Poniente	ENG_ST_x_IDX_SALRET	-10.93968974
Norte Centro	ENG_ST_x_IDX_SALRET	-9.987297684
Norte Oriente	ENG_ST_x_IDX_SALRET	18.98099091
Sur Poniente	ENG_ST_x_IDX_SALRET	-20.91760031
Sur Centro	ENG_ST_x_IDX_SALRET	-25.67573561
Sur Oriente	ENG_ST_x_IDX_SALRET	-18.29070002
Norte Poniente	ENG_SU_x_IDX_SALRET	-6.946126397
Norte Centro	ENG_SU_x_IDX_SALRET	-7.377328716
Norte Oriente	ENG_SU_x_IDX_SALRET	15.1095806
Sur Poniente	ENG_SU_x_IDX_SALRET	-14.93667291
Sur Centro	ENG_SU_x_IDX_SALRET	-17.67367647
Sur Oriente	ENG_SU_x_IDX_SALRET	-14.65302494
Norte Poniente	ENG_ST_x_IDX_VERDE	-62.57622473
Norte Centro	ENG_ST_x_IDX_VERDE	-82.10052355
Norte Oriente	ENG_ST_x_IDX_VERDE	-1.659760608
Sur Poniente	ENG_ST_x_IDX_VERDE	7.172382957
Sur Centro	ENG_ST_x_IDX_VERDE	16.2907567
Sur Oriente	ENG_ST_x_IDX_VERDE	-38.73113752
Norte Poniente	ENG_SU_x_IDX_VERDE	-37.42984011
Norte Centro	ENG_SU_x_IDX_VERDE	-45.27707251
Norte Oriente	ENG_SU_x_IDX_VERDE	-1.552072173
Sur Poniente	ENG_SU_x_IDX_VERDE	3.5192212
Sur Centro	ENG_SU_x_IDX_VERDE	12.67221006
Sur Oriente	ENG_SU_x_IDX_VERDE	-31.47988672

Referencias

Agostini C. A., Hojman D., Román A., y Valenzuela L. (2016). Segregación residencial de ingresos en el Gran Santiago, 1992–2002: una estimación robusta. *EURE (Santiago)*, 42(127):159–184. https://www.scielo.cl/scielo.php?pid=S0250-71612016000300007&script=sci_arttext.

Agostini C. A. y Palmucci G. A. (2008). The Anticipated Capitalisation Effect of a New Metro Line on Housing Prices. *Fiscal Studies*, 29(2):233–256. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-5890.2008.00074.x.

Andrade N. y Cifuentes A. (2019). Crime and (Price) Punishment in the Chilean Real Estate Market: The Case of Santiago. [Preprint SSRN]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3508408.

Anenberg E. y Laufer S. (2017). A More Timely House Price Index. *The Review of Economics and Statistics*, 99(4):722–734. https://direct.mit.edu/rest/article/99/4/722/58383/A-More-Timely-House-Price-Index.

Banco Central de Chile (2025). Reunión de Política Monetaria – septiembre 2025. [Archivo PDF]. https://www.bcentral.cl/documents/33528/6875503/Comunicado+RPM+septiembre+2025.pdf/1f7b5abc-1802-2751-b55a-64879fa793bf?t=1757479227568.

Carrizo E. (2025). Tasa de interés para un crédito hipotecario llega a su menor nivel en más de dos años. [Página web]. https://www.latercera.com/pulso/noticia/tasa-de-

interes-para-un-credito-hipotecario-llega-a-su-menor-nivel-en-mas-de-dos-anos/.

CChC (2025a). CChC presentó Índice de Calidad de Vida Urbana (ICVU) 2024. [Página web]. https://cchc.cl/noticias/cchc-presento-indice-de-calidad-de-vida-urbana-icvu-2024.

CChC (2025b). Informe 52 (2025): Actividad del Sector Inmobiliario del Gran Santiago. [Archivo PDF]. https://cchc.cl/w/informe-inmobiliario-nÂř52-gran-santiago.

CChC (2025c). Informe Nacional Mercado Inmobiliario: ventas de viviendas cayeron 18 % en el primer trimestre del 2025. [Página web]. https://cchc.cl/noticias/informe-nacional-mercado-inmobiliario-ventas-de-viviendas-cayeron-18-en-el-primer-trimestre-del-2025.

Chardon I., Concha F. J. M., y Bergel A. (2019). Housing Prices: Testing Machine Learning Methods. [Archivo PDF]. https://bergel.eu/MyPapers/Char19a-HousingPrices.pdf.

Geerts M., vanden Broucke S., y Weerdt J. D. (2023). A Survey of Methods and Input Data Types for House Price Prediction. *ISPRS International Journal of Geo-Information*, 12(5):200. https://www.mdpi.com/2220-9964/12/5/200.

Gordon B. L. y Winkler D. T. (2017). The Effect of Listing Price Changes on the Selling Price of Single-Family Residential Homes. *The Journal of Real Estate Finance and Econo-*

mics, 55(2):185–215. https://link.springer.com/article/10.1007/s11146-016-9558-z.

INE (s.f.). Permisos de edificación. [Página web]. https://www.ine.gob.cl/estadisticas/economia/edificacion-y-construccion/permisos-de-edificacion.

Ipsos (2024). What Worries the World? Informe de Chile — Octubre 2024. [Archivo PDF]. https://www.ipsos.com/sites/default/files/ct/news/documents/2024-11/Chile%20Report%20-%20What%20Worries%20the%20World%20Oct%2024_ESP.pdf.

López-Morales E., Sanhueza C., Herrera N., Espinoza S., y Mosso V. (2023). Land and housing price increases due to metro effect: An empirical analysis of Santiago, Chile, 2008–2019. *Land Use Policy*, 132:106793. https://www.sciencedirect.com/science/article/abs/pii/S0264837723002594.

Ministerio de Desarrollo Social y Familia (2023). Resultados ingresos Casen 2022. [Archivo PDF]. https://observatorio.ministeriodesarrollosocial.gob.cl/storage/docs/casen/2022/Resultados%20ingresos%20Casen%202022.pdf.

MINVU (2023). Minvu entrega cifra oficial del Déficit Habitacional: 552.046 requerimientos. [Página web]. https://centrodeestudios.minvu.gob.cl/minvu-entrega-cifra-oficial-del-deficit-habitacional-552-046-requerimientos/.

Northcraft G. B. y Neale M. A. (1987). Experts, Amateurs, and Real Estate: An Anchoring-and-Adjustment Perspective on Property Pricing Decisions. *Organizational Behavior and Human*

Decision Processes, 39(1):84–97. https://www.sciencedirect.com/science/article/pii/074959788790046X.

Ortega P. (2025). Toctoc: hogares que cumplen las condiciones para acceder a un hipotecario bajaron del 29 % al 17 % en los últimos cinco años. [Página web]. https://www.latercera.com/pulso/noticia/toctoc-las-personas-que-cumplen-con-las-condiciones-para-acceder-a-un-hipotecario-han-bajado-del-29-al-17-en-los-ultimos-cinco-anos/.

Rosen S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, 82(1):34–55. https://www.journals.uchicago.edu/doi/abs/10.1086/260169.

Shimizu C., Nishimura K. G., y Watanabe T. (2016). House prices at different stages of the buying/selling process. *Regional Science and Urban Economics*, 59:37–53. https://doi.org/10.1016/j.regsciurbeco.2016.05.001.

Soto C. (2025). Informe de Estabilidad Financiera, Primer Semestre 2025 (UAI). [Presentación PDF]. https://www.bcentral.cl/documents/33528/133214/Informe%2Bde%2BEstabilidad%2BFinanciera%2BPrimer%2BSemestre%2B2025%2B%28UAI%29%2B-%2BClaudio%2BSoto%2C%2BConsejero.pdf/e92f9c5e-cd09-f799-e1a0-25872c362e38?t=1748526800961.

Vergara-Perucich J. F. (2023). Testing Housing Price Drivers in Santiago de Chile: A Hedonic Price Approach. *Critical Housing Analysis*, 10(2):44–57. https://www.housing-

critical.com/data/USR_057_DEFAULT/04_Testing_Housing_Price_Drivers_in_Santiago_de_Chile_A_Hedonic_Price_Approach.pdf.

Yilmazer S. y Kocaman S. (2020). A mass appraisal assessment study using machine learning based on multiple regression and random forest. *Land Use Policy*, 99:104889. https://www.sciencedirect.com/science/article/abs/pii/S0264837719316540.