Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física
Maestría en Ciencias de Datos



Estimador de precios de productos inmobiliarios (Escenario Departamentos)

TESIS para obtener el **GRADO** de
MAESTRO EN CIENCIAS DE DATOS

Tesis presentada por:
Daniel Martínez Escobosa

Director de tesis:
Dr. Alberto De Obeso Orendain

Tlaquepaque, Jalisco, Septiembre, 2026

Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física

Hoja de aprobación de la Maestría en Ciencias de Datos

Título de la tesis: **Estimador de precios de productos inmobiliarios (Escenario Departamentos)**

Autor: **Daniel Martínez Escobosa**

Tesis aprobada para completar todos los requisitos del grado de Maestría en Ciencias de Datos.

Director de tesis, **Dr. Alberto De Obeso Orendain**

Lector de tesis, **Mc. Juan Carlos Martínez Alvarado**

Lector de tesis, **Dr. Iván Esteban Villalón Turrubiates**

Asesora académica, **Dra. Rocío Carrasco Navarro**

Tlaquepaque, Jalisco, Septiembre, 2026

Real Estate Product Price Estimator (Apartment Scenario)

Daniel Martínez Escobosa

Abstract

Historically, the real estate sector has played an important role as a driver of the Mexican economy. Despite technological advances in the industry, defining sales prices remains a complex task that can influence the financial and commercial performance of a real estate project. This work aims to develop a Machine Learning solution to estimate market prices for vertical real estate products (apartments), using information corresponding to the Guadalajara market from DIME and the main characteristics of the properties.

To achieve this objective, a process of data exploration, preparation, and variable selection was carried out, followed by the training and comparison of different regression algorithms. As a result, a Ridge Regression model was selected, achieving a coefficient of determination (R^2) of approximately 0.859 on the test set, an RMSE of $723,726.15 MXN, and an MAE of $550,099.67 MXN. Subsequently, the analytical prototype was transformed into a reproducible project by separating the data preparation, training, and evaluation stages, as well as by incorporating version control tools and mechanisms for tracking artifacts and metrics.

Finally, a visual application was developed to capture the characteristics of an apartment, review the information provided, and obtain a price estimate. As a complementary feature, an optional generative artificial intelligence layer was incorporated to interpret natural language descriptions and transform them into the structured variables required by the application. This layer does not calculate or modify the estimated price, since that function is performed exclusively by the Ridge Regression model. The resulting solution provides a support tool for real estate analysis and price-related decision-making, without replacing a detailed market study or a formal real estate appraisal.

Keywords: *Data Science, Machine Learning, Real Estate Market, Ridge Regression, Generative Artificial Intelligence.*

Estimador de precios de productos inmobiliarios (Escenario Departamentos)

Daniel Martínez Escobosa

Resumen

Históricamente, el sector inmobiliario ha demostrado ser un detonador activo de la economía mexicana. Pese a los avances tecnológicos en la industria, la definición de precios de venta continúa siendo una tarea compleja que puede influir en el desempeño financiero y comercial de un proyecto inmobiliario. Este trabajo tiene como objetivo desarrollar una solución de Machine Learning para estimar precios de mercado de productos inmobiliarios verticales (departamentos), utilizando información correspondiente a la plaza Guadalajara de DIME y las principales características de los productos.

Para ello se realizó un proceso de exploración, preparación y selección de variables, seguido del entrenamiento y comparación de distintos algoritmos de regresión. Como resultado se seleccionó un modelo Ridge Regression, que obtuvo un coeficiente de determinación R^2 de aproximadamente 0.859 en el conjunto de prueba, un RMSE de $723,726.15 MXN y un MAE de $550,099.67 MXN. Posteriormente, el prototipo analítico se transformó en un proyecto reproducible mediante la separación de las etapas de preparación, entrenamiento y evaluación, así como el uso de herramientas de control de versiones y seguimiento de artefactos y métricas.

Finalmente, se desarrolló una aplicación visual para capturar las características de un departamento, revisar la información y obtener una estimación de precio. Como complemento se incorporó una capa opcional de inteligencia artificial generativa que interpreta descripciones en lenguaje natural y las transforma en las variables estructuradas requeridas por la aplicación. Esta capa no calcula ni modifica el precio estimado, ya que dicha función corresponde exclusivamente al modelo Ridge Regression. El resultado constituye una herramienta de apoyo para el análisis inmobiliario y la toma de decisiones relacionadas con precios, sin sustituir un estudio de mercado detallado ni una valuación inmobiliaria formal.

Palabras clave: *Ciencia de Datos, Machine Learning, Mercado Inmobiliario, Regresión Ridge, Inteligencia Artificial Generativa.*

Tabla de Contenidos

Página

1 Introducción 14

1.1 Antecedentes 14

1.1.1 Sector inmobiliario de vivienda en México 14

1.1.2 Cambios del mercado inmobiliario: tendencia a la verticalización y sus ventajas 15

1.1.3 Dinámica de la vivienda vertical en México durante el periodo analizado 17

1.1.4 Desafíos y modernización del sector inmobiliario 19

1.2 Justificación 20

1.3 Problema 21

1.4 Hipótesis 22

1.5 Objetivos 22

1.5.1 Objetivo general 22

1.5.2 Objetivos específicos 23

1.6 Aportación tecnológica a la empresa 23

1.7 Estructura del documento 24

2 Estado del arte 26

2.1 PropTech 26

2.2 Soluciones PropTech relacionadas con la estimación de valores inmobiliarios 27

2.2.1 Caso PropTech: Avaclick 27

2.2.2 Caso PropTech: Zillow 28

3 Marco teórico/conceptual 30

3.1 Tipos de bienes inmuebles 30

3.2 Valoración de bienes inmuebles 30

3.3 Técnicas estadísticas y modelos predictivos 31

3.3.1 Análisis de correlación 31

3.3.2 Análisis de componentes principales (PCA) 32

3.3.3 Regresión lineal múltiple 33
3.3.4 Regresión Ridge 34
3.3.5 Regresión Lasso 34
3.3.6 Árboles de decisión 35
3.3.7 Bolsa de árboles de decisión 36
3.3.8 Red Neuronal 37
3.3.9 Métricas de evaluación 39
3.4 Inteligencia artificial generativa y extracción estructurada de información 40
4 Metodología 43
4.1 Procedencia y características de los datos 43
4.2 Exploración y preparación de los datos 45
4.3 Estructura final de los datos y variables del modelo 51
4.4 Industrialización y reproducibilidad del modelo 53
4.5 Inferencia y validación de datos de entrada 55
4.6 Aplicación visual para usuarios de negocio 56
4.7 Integración de inteligencia artificial generativa 58
4.8 Arquitectura de la solución 59
5 Resultados y conclusiones 61
5.1 Resultados y discusión 61
5.1.1 Comparación y selección de modelos 61
5.1.2 Verificación funcional de la solución 63
5.1.3 Discusión de resultados 64
5.2 Conclusiones 67
5.3 Trabajo futuro 69

Lista de Figuras

Página

Figura 1.1: Participación anual PIB de vivienda..... 15
Figura 1.2: Mercado potencial de compra de vivienda por generación..... 15
Figura 1.3: Cantidad de hijos por mujer..... 16
Figura 1.4: Evolución trimestral del volumen de ventas totales al mes en México..... 17
Figura 1.5: Evolución trimestral del precio por metro construido en México..... 18
Figura 1.6: Evolución trimestral del número de proyectos inmobiliarios verticales vigentes en México..... 19
Figura 2.1: Ejemplo de reporte Avaclick..... 28
Figura 2.2: Zestimate de una propiedad..... 29
Figura 3.1: Componente principal Z1 dirección mayor varianza..... 32
Figura 3.2: Esquema del árbol de decisión..... 35
Figura 3.3: Esquema del modelo de bolsa de árboles de decisión..... 37
Figura 3.4: Representación neurona artificial..... 37
Figura 3.5: Capas de una red neuronal..... 38
Figura 4.1: Síntesis del proceso de obtención e integración de información DIME. Elaboración propia con base en [30]..... 44
Figura 4.2: Matriz de correlación de Spearman variables del conjunto de datos preparado..... 50
Figura 4.3: Interfaz de captura del TOG Price Predictor..... 57
Figura 4.4: Arquitectura general del TOG Price Predictor..... 60
Figura 5.1: Precio real frente al precio estimado por Ridge en el conjunto de prueba..... 62
Figura 5.2: Extracción estructurada de las características mediante IA generativa..... 63
Figura 5.3: Formulario revisado y resultado de la estimación del TOG Price Predictor..... 64
Figura 5.4: Distribución de los residuos del modelo Ridge en el conjunto de prueba..... 65

Lista de Tablas

Página

Tabla 4.1: Variables utilizadas para la estimación de precios. 52
Tabla 4.2: Etapas del pipeline reproducible. 54
Tabla 5.1: Comparación de desempeño de los modelos evaluados. 61

Dedicado a..

A mi familia, quienes me impulsaron a dar lo mejor de mí para concluir satisfactoriamente mis estudios y titulación de la maestría. Agradezco a mis profesores y directores del trabajo de obtención de grado por sus enseñanzas dentro y fuera del aula que me permitieron robustecer mi formación académica y profesional. También agradezco al ITESO por el otorgamiento de la beca fundadora.

1 Introducción

En este capítulo se presenta el contexto que da origen al proyecto, partiendo de la importancia del sector inmobiliario y de la evolución de la vivienda vertical en México. Posteriormente se describe la problemática identificada dentro de la empresa, la justificación para abordar el ejercicio mediante ciencia de datos, las hipótesis y los objetivos planteados, y la aportación tecnológica esperada. El detalle y descripción de las principales variables inmobiliarias consideradas en el trabajo se presentan en el Apéndice A.

1.1 Antecedentes

1.1.1 Sector inmobiliario de vivienda en México

El sector inmobiliario es aquel dedicado de manera general al desarrollo de vivienda, centros comerciales, oficinas y edificaciones de usos mixtos. Abarca toda la actividad sobre bienes inmuebles: tierra, edificios, derechos aéreos sobre la tierra y los derechos subterráneos debajo de la tierra. Al ser posesiones ancladas al suelo, también se les suele denominar bienes raíces en Latinoamérica [1].

Como industria, el sector inmobiliario ha demostrado ser un detonador activo de la economía mexicana, siendo un componente esencial de esta. Su importancia económica se puede observar con tres indicadores muy importantes: su participación en el Producto Interno Bruto (PIB) del país, la generación de empleo y la inversión [2].

Según cifras del Instituto Nacional de Estadística y Geografía (INEGI), el sector inmobiliario de vivienda tiene una elevada participación en el PIB total del país aportando como se indica en la Figura 1.1, entre 5.8% y 6.0% año con año entre 2011 y 2020 [3]. Asimismo, el INEGI señala que durante el 2020 las actividades económicas del sector vivienda generaron 2,278,417 puestos de trabajo, equivalentes al 5.5% del total nacional [4]. Mientras que, en inversión, la Asociación Mexicana de Capital Privado estima que en 2020 se invirtieron 483 mil millones de pesos en los principales giros de la industria inmobiliaria, donde el 53.2% correspondió únicamente al sector vivienda, y el resto fue ocupado por la suma de los segmentos turístico, comercial, industrial y oficinas [5].

Concepto	2011	2012	2013	2014	2015	2016	2017	2018	2019 p	2020 p
PIB de la vivienda	890,492	917,073	892,025	937,738	976,348	1,017,953	1,033,925	1,046,148	1,046,816	951,523
Proporción respecto al total nacional (Porcentaje)	6	5.9	5.7	5.8	5.9	6	6	5.9	5.9	5.8

Figura 1.1: Participación anual PIB de vivienda.

Asimismo, a futuro se prevé que esta industria siga siendo un motor de la economía, ya que de acuerdo con información de la Sociedad Hipotecaria Federal (SHF), se estima que la población nacional pasará de 112.3 millones en 2010 a 128.4 millones en 2040, representando una tasa de crecimiento medio anual de 0.45% y que traerá consigo un incremento en la demanda de vivienda con un factor de 1.5 (486 mil viviendas al año) entre 2010 y 2040, pasando de 28.6 a 43.2 millones [6].

1.1.2 Cambios del mercado inmobiliario: tendencia a la verticalización y sus ventajas

Según estudios de mercado de 4S Real Estate, de 2015 a 2020 se identifica que las preferencias de los compradores de vivienda han evolucionado, inclinándose claramente hacia la vivienda vertical como principal opción de compra, pasando de un 19 a 57% [7]. Dicho crecimiento exponencial se atribuye a tres principales factores:

Uno, el incremento poblacional que existe en México ha venido acompañado inherentemente de una mayor proporción de compradores generación Millennial respecto a Generación X como se muestra en la Figura 1.2, siendo una clase que a más temprana edad tiene independencia del hogar familiar y un estilo de vida más dinámico con enfoque global en el que además se posterga la decisión de contraer matrimonio (iniciar una familia). En adición, de 2020 a 2030, se calcula se invertiría la proporción entre generaciones Millennial y Alphas como mercados potenciales de la compra de vivienda (primera compra), pasando de 56% y 25%, a 43% y 57% [7] respectivamente como lo muestra la siguiente figura.

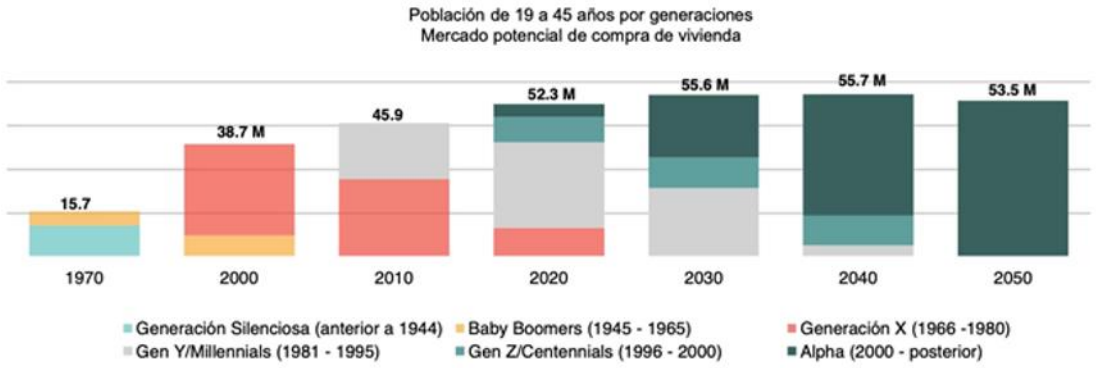


Figura 1.2: Mercado potencial de compra de vivienda por generación.

Dos, según señalan las Naciones Unidas, el número promedio de hijos que tiene cada mujer ha ido a la baja. En particular para México, en los últimos 30 años, ha disminuido de 3.5 a 2.1 hijos (ver Figura 1.3) [7]. Este cambio de configuración familiar permite a las nuevas generaciones necesitar menos espacio y que incluso estén dispuestas a sacrificarlo a cambio de otros factores que consideran más prioritarios.

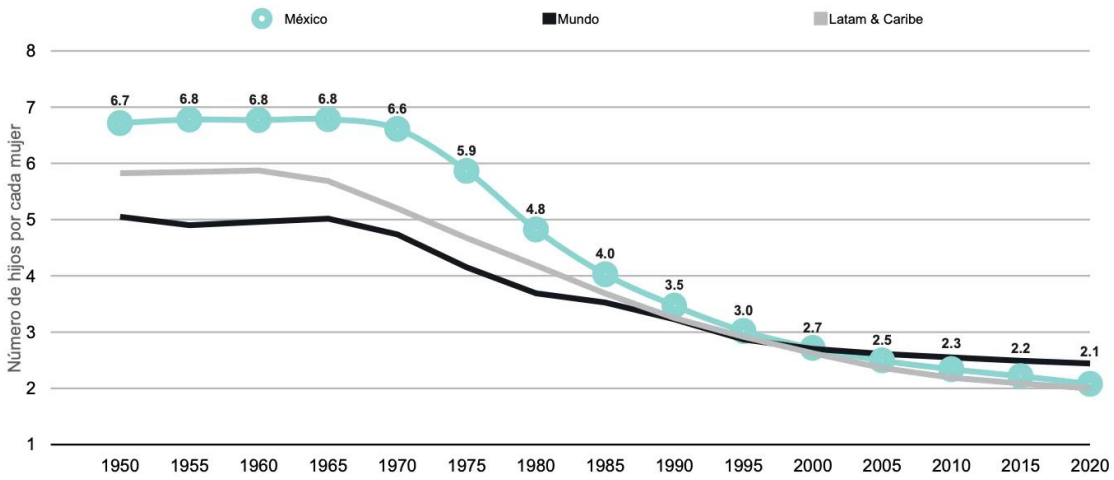


Figura 1.3: Cantidad de hijos por mujer.

Tres, existe una valorización creciente por atributos de movilidad y cercanía a servicios, centros de trabajo y lugares de entretenimiento, priorizando el tiempo de traslados. Esto según estudios realizados por 4S Real Estate, que arrojaron que más del 60% de los prospectos que dan prioridad a comprar un departamento en lugar de una casa, lo hacen por estos motivos [7]. En adición, existe una percepción de mayor seguridad viviendo más cerca de las zonas centrales de la mancha urbana y de preferencia por evitar utilizar el transporte público al considerarle deficiente.

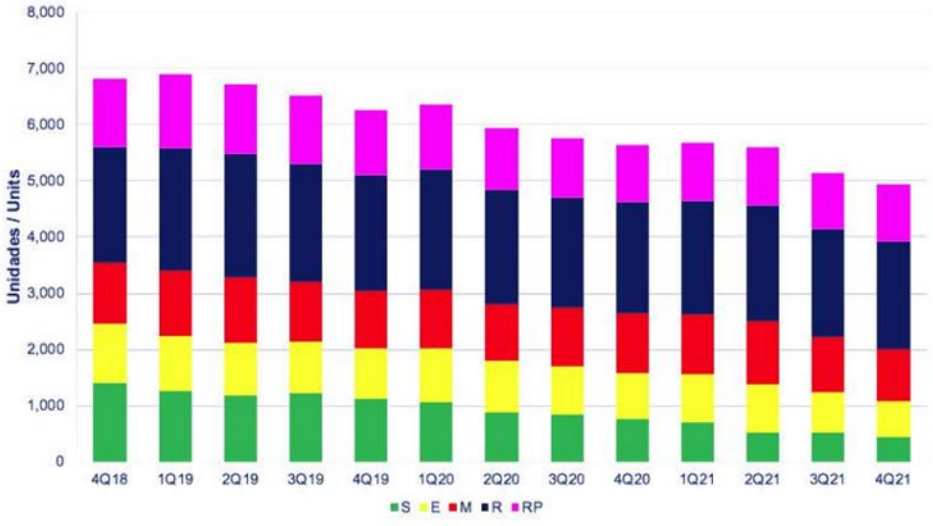
Asimismo, hay otra serie de ventajas implícitas en el desarrollo y adquisición de bienes inmuebles verticales tanto para desarrolladores inmobiliarios como para compradores. Al promoverse la densificación en la mancha urbana, al desarrollador le es más factible hacer uso de la infraestructura disponible y mejorarla, haciendo que los costos sean más eficientes y pudiendo impactar positivamente al consumidor a través de menores precios. Por otro lado, bajo una perspectiva económica y estratégica, aquellas propiedades en desarrollos verticales dentro de la mancha urbana, se sabe que adquieren mayor plusvalía a que si estuvieran situadas en la periferia donde en ocasiones no la logran. Si bien este último punto depende de la dinámica de cada ciudad, adquirir una vivienda vertical genera mayor confianza en los consumidores para recuperar su dinero en el caso que quieran establecer un modelo de renta, y a la vez favorece al

desarrollador inmobiliario en acelerar la venta a sabiendas de este beneficio y con ello recuperar más rápido su inversión.

1.1.3 Dinámica de la vivienda vertical en México durante el periodo analizado

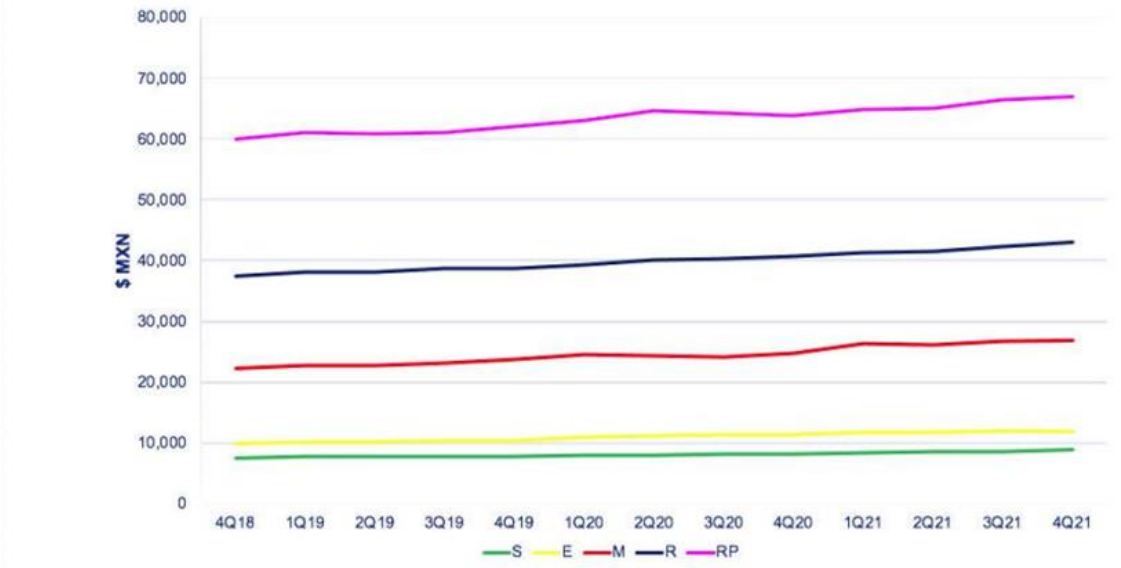
De acuerdo con Softec, empresa dedicada a la investigación del sector inmobiliario, tres indicadores relevantes para describir el comportamiento del mercado inmobiliario vertical en México durante el periodo analizado son el volumen de unidades vendidas, el precio promedio por metro cuadrado construido y la cantidad de proyectos vigentes.

De acuerdo al análisis elaborado por Gene Towle, socio director de Softec, por medio de la Figura 1.4, la evolución trimestral de ventas totales de productos verticales al mes para los sectores social, económico, medio, residencial y residencial plus ha venido en decremento paulatino durante los últimos 13 trimestres debido principalmente a una contracción económica de toda la industria derivada de los efectos de la pandemia mundial por covid-19, el cual tuvo su origen en el primer trimestre del 2020, y que ha generado menor inventario en producción por parte de los desarrolladores e incertidumbre en los consumidores. Asimismo este dato es posible aislarlo y no atribuirlo a un incremento desproporcionado del precio por metro unitario construido observando la Figura 1.5 [8].



Clasif	4Q18	1Q19	2Q19	3Q19	4Q19	1Q20	2Q20	3Q20	4Q20	1Q21	2Q21	3Q21	4Q21
S	1,394	1,261	1,190	1,230	1,132	1,060	878	838	756	706	523	515	445
E	1,060	986	927	912	893	953	923	862	818	849	866	734	646
M	1,081	1,155	1,160	1,055	1,025	1,045	1,010	1,036	1,061	1,067	1,114	978	903
R	2,065	2,174	2,192	2,094	2,044	2,133	2,027	1,966	1,988	2,015	2,054	1,911	1,927
RP	1,217	1,329	1,250	1,227	1,154	1,158	1,090	1,051	1,011	1,037	1,050	999	1,011
Grand Total	6,816	6,905	6,718	6,519	6,247	6,348	5,928	5,753	5,634	5,675	5,607	5,136	4,932

Figura 1.4: Evolución trimestral del volumen de ventas totales al mes en México.

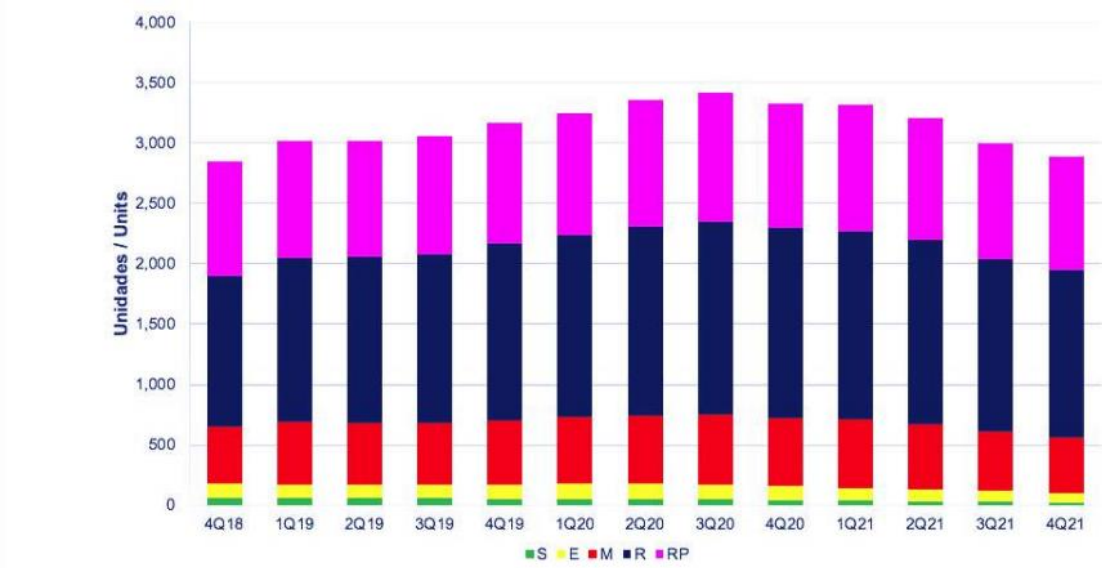


Clasif	4Q18	1Q19	2Q19	3Q19	4Q19	1Q20	2Q20	3Q20	4Q20	1Q21	2Q21	3Q21	4Q21
S	$7,446	$7,653	$7,742	$7,675	$7,735	$7,937	$8,010	$8,201	$8,219	$8,366	$8,497	$8,564	$8,776
E	$9,832	$10,047	$10,187	$10,404	$10,383	$10,831	$11,055	$11,340	$11,288	$11,718	$11,808	$11,847	$11,865
M	$22,327	$22,791	$22,736	$23,165	$23,638	$24,488	$24,312	$24,128	$24,659	$26,395	$26,200	$26,748	$26,768
R	$37,369	$37,998	$38,113	$38,789	$38,644	$39,335	$40,189	$40,362	$40,750	$41,378	$41,588	$42,260	$42,970
RP	$59,983	$61,016	$60,898	$61,164	$62,136	$63,063	$64,707	$64,353	$63,935	$64,801	$65,016	$66,420	$66,895
Grand Total	$35,590	$37,246	$37,109	$37,241	$37,446	$38,177	$39,138	$39,137	$39,407	$40,448	$41,081	$42,237	$43,740

Figura 1.5: Evolución trimestral del precio por metro construido en México.

Respecto a la cantidad de proyectos inmobiliarios con productos verticales en México, se observa que en los últimos 5 trimestres del periodo de análisis ha decrecido la cantidad de proyectos en todos los segmentos, pese a alcanzar un máximo local en el 3Q20 con 3,417 desarrollos vigentes (ver Figura 1.6). No obstante, se cierra en el 4Q21 con 50 proyectos más que los registrados en el 4Q18 (1.4% más). Adicionalmente es posible observar que no se mantiene la proporción de proyectos residenciales y residenciales plus de inicio a fin del periodo, toda vez que crece paulatinamente de 77.3% a 80.5%, indicando que no se crean más proyectos de estos segmentos, pero sí se agotan en mayor proporción los inventarios y proyectos de los segmentos social, económico y medio por los efectos económicos de la pandemia mundial.

De esta manera es posible deducir parcialmente que el mercado de desarrollo vertical durante el periodo analizado está muy competido (principalmente en los sectores residencial y residencial plus), y por tanto es importante diferenciarse y estudiar detenidamente las decisiones que se toman al interior como desarrollador para lograr las metas comerciales y subsistir en la industria.



Clasif	4Q18	1Q19	2Q19	3Q19	4Q19	1Q20	2Q20	3Q20	4Q20	1Q21	2Q21	3Q21	4Q21
S	65	62	62	57	56	55	52	51	46	39	30	30	26
E	114	112	111	112	118	125	125	124	113	104	100	88	80
M	467	514	510	507	528	552	560	571	563	566	538	495	455
R	1,257	1,361	1,378	1,406	1,472	1,506	1,569	1,607	1,573	1,563	1,527	1,429	1,393
RP	943	968	961	974	999	1,012	1,053	1,064	1,034	1,047	1,017	958	932
Grand Total	2,846	3,017	3,022	3,056	3,173	3,250	3,359	3,417	3,329	3,319	3,212	3,000	2,886

Figura 1.6: Evolución trimestral del número de proyectos inmobiliarios verticales vigentes en México.

1.1.4 Desafíos y modernización del sector inmobiliario

Una de las mayores preocupaciones del desarrollador y comercializador inmobiliario es que no se haya estudiado y analizado correctamente al mercado y que la política de precios definida para determinado producto no esté diseñada acorde a ello, acarreando dificultades para desplazar inventario e impactando negativamente en la rentabilidad del negocio. Por tanto, el establecimiento de precios de mercado es una actividad clave dentro del mercado inmobiliario para evitar principalmente las sobrevaluaciones, mismas que Min Zhu, subdirector del Fondo Monetario Internacional en 2014 [9], catalogó su detección como un arte más que como una ciencia, al necesitar dictámenes y análisis detallados.

El uso de nuevas tecnologías y la ciencia de datos es cada vez más necesario en una industria que busca romper con prácticas tradicionales y transformarse aceleradamente en un segmento PropTech, adaptado a las necesidades de consumo. En este sentido, el reto principal que se afronta en el sector inmobiliario es que no se suelen generar y trabajar datos a enormes niveles por el alto costo que representa su adquisición. No obstante, eso no ha impedido la incorporación de la inteligencia artificial en el sector de manera importante en los últimos años con aplicaciones que abarcan el estudio de propiedades en renta para sugerir valores de mercado; mercadotecnia automatizada para buscar tener una mejor visión de los consumidores y evaluar su grado de

interés de compra; o el estudio de comportamiento de consumidores en el tiempo para ofrecerle nuevas soluciones o productos.

Con este trabajo se busca proveer a una empresa desarrolladora inmobiliaria de una herramienta de apoyo que estime el precio de mercado de propiedades verticales nuevas con base en sus características, con la finalidad de complementar los procesos de definición de precios y toma de decisiones mediante información de mercado. No se busca sustituir los estudios de mercado, que pueden incluir análisis más detallados, ni obviar la opinión de los especialistas comerciales de la empresa, sino incorporar una referencia cuantitativa adicional y contribuir al proceso de transformación digital de la organización.

1.2 Justificación

La definición y actualización de precios de productos inmobiliarios es una actividad relevante para un desarrollador, ya que una desviación importante respecto a las condiciones del mercado puede afectar el ritmo de venta, el desplazamiento del inventario y los resultados financieros de un proyecto [10]-[11]. Por esta razón, contar con información objetiva que complemente el conocimiento y experiencia de los equipos comerciales puede ayudar a reducir la incertidumbre durante este proceso.

Para el caso de la empresa en la cual se desarrolló este trabajo, ya se había realizado la contratación de información especializada del mercado inmobiliario mediante una base DIME de Softec correspondiente a Guadalajara. Debido a que este tipo de información tiene un costo, se comercializa por plaza y puede actualizarse de manera trimestral, resultaba conveniente explorar si los datos disponibles podían ser aprovechados más allá de su consulta directa y utilizados como insumo para desarrollar modelos de ciencia de datos que generaran información adicional para la toma de decisiones.

El desarrollo del proyecto se planteó entonces para evaluar si, a partir de las características físicas, comerciales y de ubicación contenidas en esta información, era posible construir una referencia estimada del precio de mercado de departamentos. El objetivo no consiste en sustituir los estudios de mercado, las valuaciones inmobiliarias formales ni el conocimiento de los especialistas comerciales, sino en incorporar una referencia cuantitativa adicional que pueda complementar estos análisis y disminuir el trabajo manual requerido para estudiar diferentes productos y escenarios.

Adicionalmente, el ejercicio se encuentra alineado con el interés de la organización por avanzar hacia una toma de decisiones más orientada a datos. En caso de comprobar que la información disponible permite generar resultados útiles, el ejercicio también aporta elementos para valorar el aprovechamiento de futuras actualizaciones de la base DIME de Guadalajara y estudiar posteriormente la aplicación de una metodología similar con información de otras plazas donde la empresa tenga presencia.

De esta manera, el valor del proyecto no se limita únicamente a obtener un modelo predictivo, sino también a mostrar una forma de transformar una fuente de información ya disponible en la organización en una solución reproducible y de fácil utilización para usuarios de negocio. Esto permite establecer un punto de partida sobre el cual puedan desarrollarse posteriormente nuevas capacidades analíticas y herramientas que apoyen la evolución de la empresa hacia una operación cada vez más basada en datos.

1.3 Problema

El problema práctico general consiste en que la definición y actualización de precios de venta de productos inmobiliarios no es una tarea intuitiva ni sencilla para un desarrollador inmobiliario. Para el caso de la empresa en la cual se desarrolló este trabajo, estas decisiones se apoyaban principalmente en el conocimiento y experiencia de los equipos comerciales, así como en análisis de mercado realizados de forma manual. Este proceso podía requerir una cantidad importante de horas-hombre, principalmente cuando era necesario analizar varios productos, realizar incrementos de precio o evaluar la mezcla de productos de un nuevo proyecto.

El autor desarrolló este trabajo desde el área de Transformación Digital de la organización, lo que permitió abordar una problemática existente dentro de sus procesos comerciales y explorar el aprovechamiento de información inmobiliaria que la empresa ya tenía disponible.

Como parte de la búsqueda de alternativas para mejorar este proceso, la empresa contrató una base DIME de Softec correspondiente a Guadalajara. Esta información representa un levantamiento especializado del mercado inmobiliario que se comercializa por plaza o ciudad y se actualiza de forma trimestral. Además de utilizar la información para el análisis puntual del mercado, existía interés en evaluar si era posible obtener mayor valor de ella mediante técnicas de ciencia de datos y, con base en los resultados, determinar la conveniencia de continuar adquiriendo actualizaciones posteriores.

De esta manera, una de las preguntas que dio origen al proyecto fue determinar qué información y capacidad predictiva podía obtenerse a partir de una fuente de datos que la empresa ya había adquirido. En caso de resultar útil, la misma metodología podría servir como punto de partida para aprovechar futuras actualizaciones del DIME de Guadalajara, por ejemplo, mediante la actualización o reentrenamiento del modelo, y posteriormente estudiar su aplicación con información de otras ciudades en las que la empresa tuviera presencia.

El estudio se concentró particularmente en departamentos debido a la creciente verticalización del mercado inmobiliario de Guadalajara y a que, dentro de la empresa, esta división presentaba una mayor necesidad de contar con referencias adicionales para el análisis y actualización de precios. Esto permitió delimitar el problema a un tipo de producto específico antes de considerar posibles extensiones hacia otros mercados o productos inmobiliarios, como la vivienda horizontal.

Adicionalmente, la empresa buscaba avanzar hacia una forma de trabajo más orientada a datos. Por esta razón, el reto no consistía únicamente en desarrollar un modelo predictivo, sino también en encontrar una forma de transformar la información disponible en una referencia cuantitativa, reproducible y sencilla de utilizar por usuarios de negocio sin conocimientos técnicos especializados.

Por tanto, el problema abordado en este trabajo consiste en determinar si la información inmobiliaria disponible en el DIME de Guadalajara puede ser aprovechada mediante técnicas de ciencia de datos para construir una solución que estime precios de departamentos y que sirva como una referencia adicional en los procesos de análisis y toma de decisiones comerciales de la empresa.

1.4 Hipótesis

Las hipótesis de trabajo planteadas para el desarrollo del proyecto son las siguientes:

1. Las características físicas, comerciales y de ubicación contenidas en la información DIME de Guadalajara permiten construir un modelo de regresión capaz de estimar el precio de departamentos con un coeficiente de determinación R^2 de al menos 0.85 sobre datos no utilizados durante el entrenamiento.
2. La comparación de distintos algoritmos de regresión permite identificar un modelo con capacidad de generalización adecuada para el conjunto de datos estudiado, considerando de manera conjunta su desempeño sobre datos de prueba y la magnitud de los errores de estimación.
3. El modelo seleccionado puede integrarse en un proceso reproducible y en una aplicación visual que permita a usuarios no técnicos capturar las características de un departamento y obtener una estimación de precio, manteniendo separadas la captura de información y la lógica responsable de realizar la predicción.

1.5 Objetivos

1.5.1 Objetivo general

Desarrollar una solución basada en ciencia de datos que permita aprovechar información inmobiliaria disponible en la empresa para estimar precios de mercado de productos inmobiliarios verticales (departamentos) utilizando información correspondiente a la plaza Guadalajara de DIME, buscando alcanzar un coeficiente de determinación R^2 de al menos 0.85 en datos de prueba y generar una referencia objetiva que complemente los procesos actuales de análisis y facilite la toma de decisiones comerciales mediante una herramienta que pueda ser utilizada por usuarios no técnicos.

1.5.2 Objetivos específicos

- Explorar y preparar la información inmobiliaria disponible en la base DIME de Guadalajara, identificando variables con aportación para la estimación del precio de mercado de departamentos e incluyendo características físicas, comerciales y de ubicación, con la finalidad de evaluar el valor que puede obtenerse de esta fuente mediante técnicas de ciencia de datos.
- Construir y comparar diferentes modelos de regresión, evaluando su desempeño sobre datos no utilizados durante el entrenamiento mediante las métricas R^2, RMSE y MAE, y seleccionar el modelo que presente el mejor desempeño para el caso de estudio.
- Convertir el prototipo analítico en un proceso reproducible, separando las etapas de preparación de datos, entrenamiento, evaluación e inferencia, de tal forma que los resultados puedan repetirse y actualizarse de manera controlada.
- Desarrollar una herramienta visual que permita a usuarios no técnicos ingresar las características de un departamento, revisar la información proporcionada y obtener de forma sencilla una estimación de precio generada por el modelo seleccionado.
- Incorporar de manera complementaria una capa de inteligencia artificial generativa que facilite la captura de información mediante lenguaje natural, manteniendo la estimación del precio exclusivamente a cargo del modelo de regresión y requiriendo la revisión del usuario sobre la información extraída.
- Identificar las principales limitaciones del ejercicio y proponer líneas de evolución que permitan robustecer la solución y evaluar posteriormente su aplicación con nuevas actualizaciones de información, otras plazas o diferentes tipos de productos inmobiliarios.

1.6 Aportación tecnológica a la empresa

El desarrollo de este trabajo busca aportar a la empresa una primera aplicación práctica de ciencia de datos sobre información inmobiliaria que ya se encontraba disponible dentro de la organización. Más allá de la construcción de un modelo predictivo, se busca establecer una forma de transformar estos datos en información que pueda ser utilizada de manera sencilla para apoyar procesos comerciales y de análisis.

Las principales aportaciones identificadas son:

- Aprovechar con mayor profundidad la información contenida en la base DIME de Softec contratada por la empresa, utilizando sus variables para generar una referencia cuantitativa del precio de mercado de departamentos y contar con mayores elementos para valorar el aprovechamiento de futuras actualizaciones de esta fuente.
- Apoyar a los equipos comerciales y de análisis en la revisión de precios de productos existentes o en la evaluación inicial de nuevos productos, complementando el conocimiento de los especialistas con una estimación generada a partir de información real de mercado. La

herramienta no busca sustituir el criterio comercial ni los estudios de mercado, sino funcionar como una referencia adicional para la toma de decisiones.

- Disminuir parte del trabajo manual requerido para evaluar diferentes configuraciones de producto, permitiendo modificar características físicas, comerciales y de ubicación y obtener de manera ágil una nueva estimación. Esto puede ser especialmente útil durante las primeras etapas de análisis y diseño de nuevos proyectos inmobiliarios [12].
- Transformar el prototipo analítico en un proceso reproducible, con etapas independientes para la preparación de información, entrenamiento, evaluación e inferencia. Esto permite que el ejercicio no dependa únicamente del código utilizado durante su desarrollo inicial y establece una base para futuras actualizaciones del modelo.
- Facilitar el uso de la solución por usuarios de negocio sin conocimientos técnicos especializados mediante una aplicación visual para la captura, revisión y estimación de precios. De manera complementaria, se incorporó una capa opcional de inteligencia artificial generativa para facilitar la captura de características de un inmueble mediante lenguaje natural, sin intervenir en el cálculo del precio realizado por el modelo de regresión.
- Establecer un punto de partida para continuar explorando el uso de ciencia de datos dentro de la organización. Si futuras actualizaciones de información mantienen resultados adecuados, la metodología puede servir como referencia para estudiar nuevas versiones del DIME de Guadalajara y posteriormente evaluar su aplicación en otras plazas o tipos de productos inmobiliarios.

De esta manera, la aportación tecnológica del proyecto no se limita al desempeño de un algoritmo, sino que incluye la construcción de un proceso que permite pasar de una fuente de datos especializada a una herramienta de apoyo para usuarios de negocio. Esto se encuentra alineado con el interés de la empresa por avanzar hacia una operación con mayor uso de datos en sus procesos de toma de decisiones y permite mostrar un posible camino de evolución para posteriores iniciativas de transformación digital.

En conjunto, el problema identificado y los objetivos planteados delimitan el proyecto como un ejercicio de aprovechamiento de información inmobiliaria ya disponible en la organización. El propósito central es pasar de un proceso apoyado principalmente en análisis manual y experiencia comercial a una referencia adicional basada en datos, sin sustituir el conocimiento de los especialistas. A partir de este planteamiento, los siguientes capítulos presentan los antecedentes tecnológicos y las técnicas utilizadas para desarrollar y evaluar la solución.

1.7 Estructura del documento

El resto del documento se organiza de la siguiente manera. El Capítulo 2 presenta el estado del arte y revisa el concepto de PropTech y algunas soluciones relacionadas con la estimación de

valores inmobiliarios. El Capítulo 3 desarrolla el marco teórico y conceptual necesario para comprender las técnicas estadísticas, los modelos predictivos y los conceptos de inteligencia artificial generativa utilizados en el trabajo. El Capítulo 4 describe la metodología seguida para obtener y preparar los datos, construir y evaluar los modelos, transformar el prototipo en un proceso reproducible y desarrollar la aplicación visual. Finalmente, el Capítulo 5 presenta los resultados obtenidos y su discusión, incluyendo la comparación de los modelos evaluados, la selección de Ridge Regression y la verificación funcional de la solución; posteriormente se presentan las principales conclusiones y las líneas de trabajo futuro derivadas de las limitaciones y oportunidades identificadas. Como complemento, el Apéndice A incluye un glosario de las principales variables inmobiliarias utilizadas en el proyecto, mientras que el Apéndice B presenta las abreviaciones empleadas a lo largo del documento.

2 Estado del arte

En este capítulo se presenta el contexto tecnológico en el cual se ubica el trabajo, partiendo del concepto de PropTech y de la incorporación de nuevas tecnologías al sector inmobiliario. Posteriormente se revisan dos soluciones relacionadas con la estimación de precios de propiedades, Avaclick y Zillow, con la finalidad de identificar algunos antecedentes y diferencias respecto al enfoque desarrollado en este proyecto.

2.1 PropTech

El mundo ha ido avanzando hacia una era donde la adopción de nuevas tecnologías y plataformas es necesaria para la supervivencia de las industrias. En lo que respecta al sector inmobiliario, se vive un proceso acelerado de transición hacia un segmento PropTech (Property Technology o Real Estate Technology), que es el proceso de transformación digital que vive el mercado inmobiliario en todos sus campos y cuyo propósito es el de la implementación de la tecnología e innovación para digitalizar, reinventar y optimizar los servicios dentro del sector. Aspectos como la inversión, la gestión, la comercialización, la financiación o el análisis de datos, han variado radicalmente desde que la digitalización alcanzó al sector inmobiliario [13]-[14].

Si bien no es un término nuevo, pues fue acuñado a inicios del siglo, su evolución dentro del sector puede describirse mediante cuatro oleadas de innovación tecnológica, cada una asociada con tendencias propias del ecosistema tecnológico de su época. A la fecha, las distintas oleadas continúan teniendo presencia dentro del sector [15]-[17].

- PropTech 1.0: fue la primera oleada y se originó en 2007. El mundo online era utilizado simplemente como una herramienta de búsqueda, y por ello la primera etapa se enfocó en generar visibilidad digital para el agente o propietario. Tuvo poca audiencia debido a la falta de conocimiento técnico y rigidez para aceptar la tecnología en el sector inmobiliario.
- PropTech 2.0: surge en 2008 pero floreció en 2014. Trata sobre el universo de ideas y posibilidades para transformar procesos y segmentos específicos del mercado que podrían beneficiarse simplemente transponiendo conceptos tecnológicos existentes como economía compartida, fintech, analítica, construcción inteligente, eficiencia energética y otros segmentos de tecnología. Esta etapa es clave y tiene una alta relación con el método de estudio de este trabajo pues se relaciona con las facilidades de procesamiento y almacenamiento de datos.
- PropTech 3.0: surge en 2019. Trata sobre profundizar y mejorar las experiencias de usuario y administración. Marca el gran salto hacia tecnologías emergentes como IoT, smart lifts, drones y realidad aumentada y blockchain.

- PropTech 4.0: se plantea como una etapa enfocada en aprovechar y combinar tecnologías existentes para optimizar procesos, estructuras y costos dentro del sector inmobiliario [15]-[17].

Internacionalmente, la adopción de soluciones PropTech ha ganado relevancia. De acuerdo con el reporte consultado [15], entre 2018 y 2020 se alcanzó un valor acumulado de 18 mil millones de dólares y un crecimiento de 63%, a pesar de la crisis sanitaria. Esto refleja una tendencia creciente hacia la incorporación de tecnología en distintos procesos de la industria inmobiliaria, particularmente en organizaciones que buscan mejorar el aprovechamiento de información y la toma de decisiones.

2.2 Soluciones PropTech relacionadas con la estimación de valores inmobiliarios

Para contextualizar soluciones relacionadas con el objeto de estudio de este proyecto, se presentan a continuación dos casos de empresas PropTech que generan estimaciones de valores inmobiliarios mediante el uso de información de mercado y características de las propiedades.

2.2.1 Caso PropTech: Avaclick

Avaclick es una PropTech mexicana enfocada en generar opiniones de valor para propiedades inmobiliarias. A partir de información básica del inmueble, la plataforma permite obtener una estimación de precio de venta y renta, precio por metro cuadrado y referencias de propiedades comparables disponibles en el mercado [18]. Su servicio está dirigido tanto a particulares como a profesionales inmobiliarios que buscan contar con una referencia para procesos de compra, venta, renta o inversión.

La propia plataforma distingue este tipo de estimación de un avalúo inmobiliario formal, por lo que su resultado funciona como una referencia de apoyo para el usuario y no como un sustituto de una valuación realizada por un profesional [18]. Para este trabajo, Avaclick representa un antecedente relevante al mostrar cómo una herramienta digital puede transformar información de una propiedad y de su mercado en una referencia de precio de fácil consulta para usuarios no técnicos.

A diferencia del enfoque de este proyecto, la solución desarrollada en este trabajo parte de una base DIME contratada por la empresa y busca construir y evaluar un modelo predictivo propio para departamentos utilizando información correspondiente a la plaza Guadalajara de DIME. Esto permite conocer las métricas obtenidas por el modelo y analizar posteriormente su posible actualización con nuevas versiones de la información disponible. A continuación, se muestra en la Figura 2.1 un ejemplo de un reporte de Avaclick:

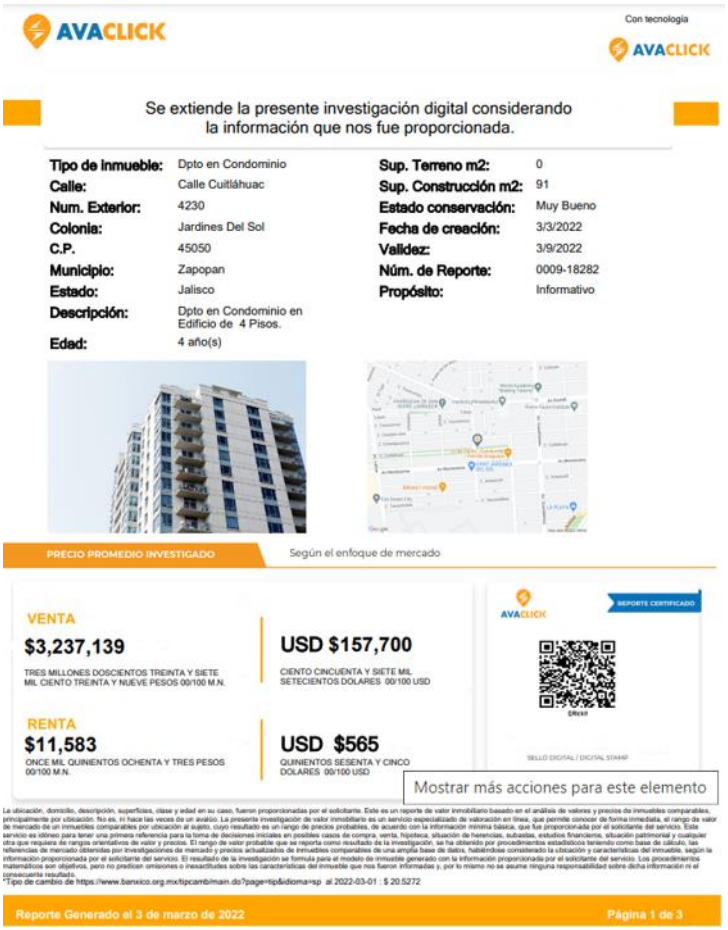


Figura 2.1: Ejemplo de reporte Avaclick.

2.2.2 *Caso PropTech: Zillow*

Zillow es una empresa tecnológica estadounidense enfocada en el mercado inmobiliario en línea. Una de sus herramientas más conocidas es Zestimate, un modelo que genera una estimación del valor de mercado de una vivienda a partir de diferentes fuentes de información, entre las que se incluyen registros públicos, información de listados inmobiliarios, datos proporcionados por usuarios, características físicas de la propiedad, ubicación y tendencias del mercado [19].

Actualmente, Zillow utiliza modelos estadísticos y de Machine Learning, incluyendo una arquitectura basada en redes neuronales, para generar estas estimaciones [19]. La empresa señala expresamente que el Zestimate no constituye un avalúo y que su precisión depende, entre otros factores, de la cantidad y calidad de información disponible para la propiedad y su zona.

Como referencia de su desempeño, Zillow reporta para Estados Unidos una tasa mediana de error cercana al 1.83% para viviendas que se encuentran en el mercado y de 7.01% para viviendas fuera del mercado [19]. Estas diferencias muestran la importancia que tiene la disponibilidad de información reciente y representativa para la estimación del valor de una propiedad.

Para este trabajo, Zillow constituye un ejemplo de cómo los modelos de Machine Learning pueden ser utilizados como apoyo para generar estimaciones inmobiliarias a partir de diferentes características y fuentes de datos. No obstante, su alcance, cantidad de información y mercado de operación son diferentes a los del presente proyecto, por lo que sus métricas no se utilizan como un criterio de comparación directa con el modelo desarrollado.

A continuación, se muestra en la Figura 2.2 un ejemplo de Zestimate:

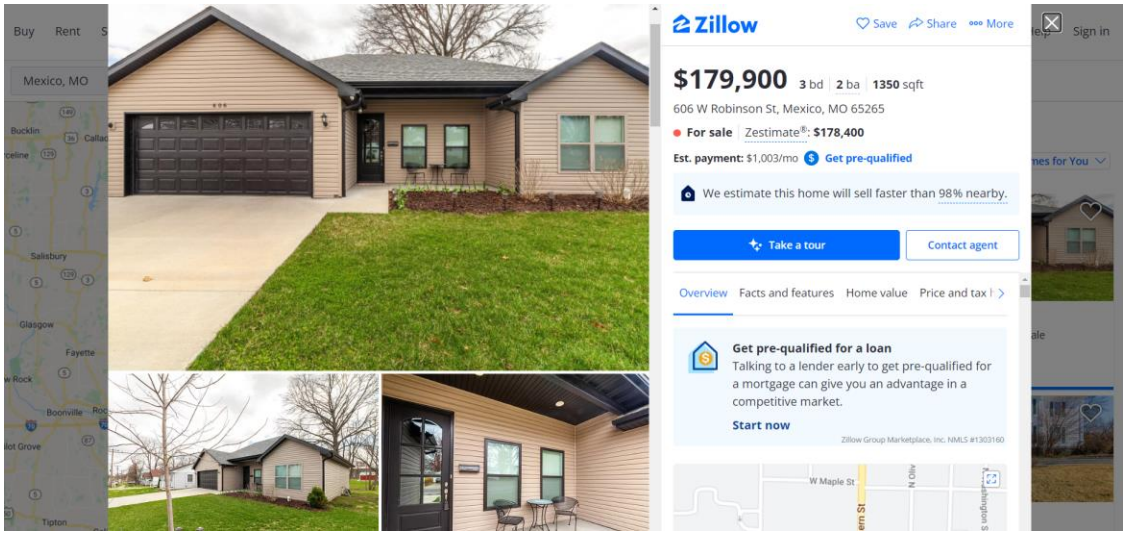


Figura 2.2: Zestimate de una propiedad.

Los casos revisados muestran que el uso de tecnología y datos para generar referencias de precios inmobiliarios ya tiene aplicaciones dentro del sector PropTech. Sin embargo, las soluciones existentes pueden responder a diferentes mercados, fuentes de información y objetivos de uso. En este trabajo se plantea un ejercicio enfocado específicamente en departamentos y en el aprovechamiento de información correspondiente a la plaza Guadalajara de DIME, buscando construir una solución propia que pueda ser utilizada como apoyo en los procesos de análisis de la organización.

3 Marco teórico/conceptual

En este capítulo se presentan los principales conceptos que sirven como fundamento para el desarrollo del proyecto. Primero se describen algunos conceptos relacionados con los bienes inmuebles y su valoración. Posteriormente se revisan las técnicas estadísticas y los modelos de regresión considerados durante el desarrollo, así como las métricas utilizadas para evaluar su desempeño. Finalmente, se introducen los conceptos necesarios para comprender el uso complementario de inteligencia artificial generativa dentro de la aplicación desarrollada.

3.1 Tipos de bienes inmuebles

Los 4 tipos de bienes inmuebles más importantes que existen son [1]:

- Bienes inmuebles residenciales: son el tipo de bien raíz más popular. Su característica principal es que son construidos para ser habitados y comúnmente son vistos como inversión. Conglomeran varios tipos de productos como: viviendas unifamiliares, condominios, casas adosadas, departamentos, dúplex, áticos, lofts.
- Bienes inmuebles comerciales: hacen referencia a todas aquellas propiedades en las cuales se realiza una actividad comercial, tales como: edificios de oficinas, locales, centros comerciales, hoteles y hospedaje, almacenamiento, instalaciones de atención médica, iglesias, museos.
- Bienes inmuebles industriales: incluyen edificios industriales y propiedades, así como almacenes. Los edificios se pueden utilizar para investigación, producción, almacenamiento y distribución de productos. Algunos edificios que distribuyen bienes se consideran bienes inmuebles comerciales.
- Bienes inmuebles tierra: incluyen terrenos baldíos de uso agrícola o rancho, así como terrenos que no tienen ninguna propiedad comercial o residencial desarrollada. A menudo son comprados como inversión.

En este trabajo se aborda el tipo de bien inmueble residencial, ya que está contextualizado para departamentos (desarrollos verticales).

3.2 Valoración de bienes inmuebles

Los bienes raíces pueden tener diferentes tasaciones según el objetivo que tengan. A continuación, se explican los diferentes tipos de valoración y su enfoque [20]:

- Valor de mercado: es el que un tercero está dispuesto a pagar. Regularmente se obtiene mediante un estudio de mercado. También, es el precio al que podría venderse el inmueble, mediante contrato privado entre un vendedor voluntario y un comprador independiente.

- Valor real comercial: Lo calcula un perito valuador (profesional) que analiza todos los factores de la propiedad y su contexto, creando un documento oficial que muestra y justifica dicho valor.
- Valor hipotecario: es la tasación que realizan los bancos teniendo en cuenta la posibilidad futura de comerciar con el inmueble, condiciones del mercado normales y locales, uso en el momento de la tasación y usos alternativos correspondientes.
- Precio de venta o de lista: es el precio al que decide vender el vendedor. Por lo regular se busca que sea cercano al valor de mercado.
- Valor fiscal: método que emplea Hacienda para calcular el valor de un bien inmueble y que sirve de base de cálculo para el pago de impuestos en caso de compraventa o herencia.
- Valor catastral: lo decide cada municipio en función de diferentes factores como la ubicación de la vivienda y se utiliza para calcular el Impuesto Sobre Bienes Inmuebles (IBI). Suele ser menor que el precio de mercado.
- Valor de suelo: precio del suelo de la finca o de la materia prima utilizada para la construcción (aunque no tenga espacio libre por construir se puede obtener este valor).

3.3 *Técnicas estadísticas y modelos predictivos*

3.3.1 *Análisis de correlación*

La correlación es una medida de asociación que permite cuantificar el grado y sentido de la relación lineal entre dos variables. El coeficiente de correlación de Pearson es una de las medidas más utilizadas para analizar este tipo de relación [21].

Para estudiar la relación lineal existente entre dos variables continuas es necesario disponer de parámetros que permitan cuantificar dicha relación. Uno de estos parámetros es la covarianza, que indica el grado de variación conjunta de dos variables aleatorias:

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{N-1} \tag{3.1}$$

siendo $\bar{x}$ e $\bar{y}$ las medias de cada variable y x_i e y_i el valor de las variables para la observación i.

La covarianza depende de las escalas en que se miden las variables estudiadas, por lo tanto, no es comparable entre distintos pares de variables. Para poder hacer comparaciones se estandariza la covarianza, generando lo que se conoce como coeficientes de correlación. Existen diferentes tipos, de entre los que destacan el coeficiente de Pearson, Rho de Spearman y Tau de Kendall. Todos ellos varían entre +1 y -1. Siendo +1 una correlación positiva perfecta y -1 una correlación negativa perfecta.

La correlación de Pearson, es la más común. Funciona bien con variables cuantitativas que tienen una distribución normal (e incluso se le considera robusto a pesar de la falta de normalidad), aunque es más sensible a los valores extremos que las otras dos alternativas.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{3.2}$$

3.3.2 Análisis de componentes principales (PCA)

El Análisis de Componentes Principales (PCA por sus siglas en inglés) es un método estadístico de aprendizaje no supervisado que permite reducir la dimensionalidad de un conjunto de datos mediante nuevas variables denominadas componentes principales, buscando conservar la mayor proporción posible de la variabilidad contenida en los datos [22].

Las componentes principales se obtienen por combinación lineal de las variables originales. Se pueden entender como nuevas variables obtenidas al combinar de una determinada forma las variables originales. La primera componente principal de un grupo de variables (X_1,X_2,…,X_p) es la combinación lineal normalizada de dichas variables que tiene mayor varianza:

$$Z_1 = \varphi_{11}X_1 + \varphi_{21}X_2 + \cdots + \varphi_{p1}X_p \tag{3.3}$$

Esta relación puede observarse gráficamente en la Figura 3.1.

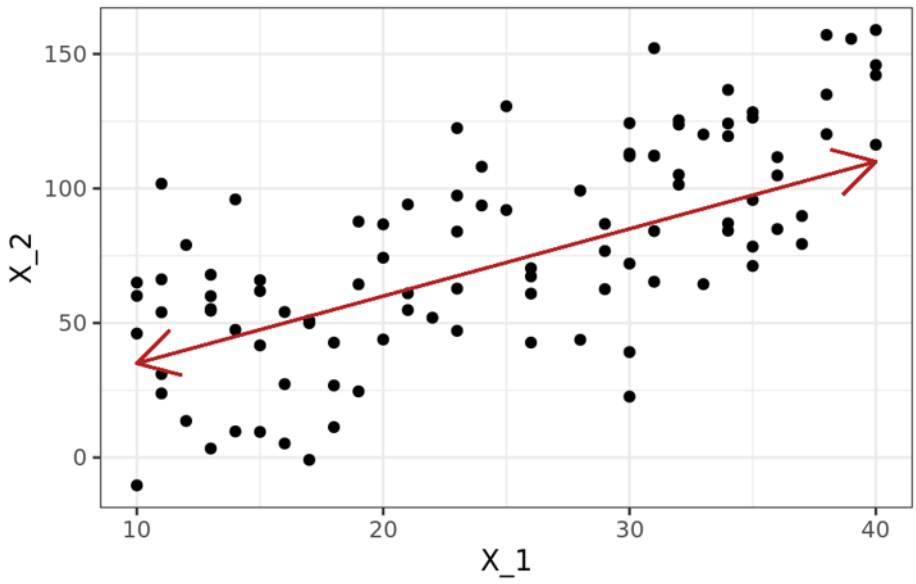


Figura 3.1: Componente principal Z1 dirección mayor varianza.

Que la combinación lineal sea normalizada implica que:

$$\sum_{j=1}^{p} \varphi_{j1}^2 = 1 \tag{3.4}$$

Los términos φ_11, …, φ_1p reciben el nombre de loadings y son los que definen a la componente. Por ejemplo, φ_11 es el loading de la variable X1 de la primera componente principal. Los loadings pueden interpretarse como el peso o importancia que tiene cada variable en cada componente y, por lo tanto, ayudan a conocer la información que recoge cada una de las componentes.

Para conocer cuánta información es capaz de capturar cada una de las componentes principales obtenidas, se recurre a la proporción de varianza explicada por cada una de ellas. Asumiendo que las variables se han normalizado para tener media cero, la varianza total presente en el set de datos se define como:

$$\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2 \tag{3.5}$$

y la varianza explicada por la componente m es:

$$\frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \varphi_{jm} x_{ij} \right)^2 \tag{3.6}$$

Tanto la proporción de varianza explicada, como la proporción de varianza explicada acumulada son dos valores de gran utilidad a la hora de decidir el número de componentes principales a utilizar para reducción de dimensionalidad.

3.3.3 Regresión lineal múltiple

La regresión lineal múltiple permite generar un modelo lineal en el que el valor de la variable dependiente o respuesta (Y) se determina a partir de un conjunto de variables independientes llamadas predictores (X1, X2, X3…). Los modelos de regresión múltiple pueden emplearse para predecir el valor de la variable dependiente o para evaluar la influencia que tienen los predictores sobre ella [23].

Los modelos lineales múltiples siguen la siguiente ecuación:

$$Y_i = (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni}) + e_i \tag{3.7}$$

En donde:

β_0: es la ordenada en el origen, el valor de la variable dependiente Y cuando todos los predictores son cero.

β_1: se conocen como coeficientes parciales de regresión. Representan el efecto promedio que tiene el incremento en una unidad de la variable predictora X_i sobre la variable dependiente Y, manteniéndose constantes el resto de las variables.

e_i: es el residuo o error, la diferencia entre el valor observado y el estimado por el modelo.

3.3.4 Regresión Ridge

La regresión Ridge es una extensión de la regresión lineal múltiple que incorpora un término de regularización durante el ajuste del modelo. Este método fue propuesto por Hoerl y Kennard como una alternativa para problemas en los que existe una relación importante entre las variables predictoras, situación que puede provocar inestabilidad en los coeficientes estimados mediante mínimos cuadrados [24].

A diferencia de la regresión lineal convencional, Ridge penaliza el tamaño de los coeficientes mediante regularización L2. De esta manera, el modelo busca minimizar simultáneamente el error de predicción y el tamaño de sus coeficientes. Su función objetivo puede representarse como:

$$\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2\right\} \tag{3.8}$$

donde λ es un parámetro de regularización mayor o igual a cero. Cuando λ es igual a cero, el modelo corresponde a una regresión lineal sin regularización. Conforme aumenta su valor, se incrementa la penalización aplicada a los coeficientes y estos tienden a reducir su magnitud.

Esta reducción permite disminuir la sensibilidad del modelo ante variables correlacionadas y controlar su varianza, aunque introduce cierto sesgo en la estimación [24]. A diferencia de otros métodos de regularización, Ridge generalmente no elimina variables al llevar sus coeficientes exactamente a cero, sino que limita su magnitud y conserva su participación dentro del modelo.

Por estas características, Ridge resulta útil cuando se dispone de varias variables que pueden contener información relacionada entre sí y se busca obtener un modelo lineal con mayor estabilidad y capacidad de generalización.

3.3.5 Regresión Lasso

La regresión Lasso (Least Absolute Shrinkage and Selection Operator) es otro método de regresión lineal regularizada. Fue propuesto por Tibshirani y, al igual que Ridge, busca controlar la complejidad del modelo mediante la incorporación de una penalización sobre los coeficientes de regresión [25].

La principal diferencia respecto a Ridge se encuentra en el tipo de regularización utilizada. Mientras Ridge aplica una penalización L2 basada en el cuadrado de los coeficientes, Lasso utiliza una penalización L1 basada en sus valores absolutos. Su función objetivo puede representarse como:

$$\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right\} \tag{3.9}$$

donde λ controla el grado de regularización aplicado al modelo. Conforme aumenta su valor, se incrementa la penalización sobre los coeficientes.

Una característica particular de Lasso es que algunos coeficientes pueden reducirse exactamente a cero, por lo que además de regularizar el modelo puede realizar de manera implícita una selección de variables [25]. Esta propiedad lo diferencia de Ridge, que generalmente reduce la magnitud de los coeficientes sin eliminarlos por completo.

Por estas características, Lasso resulta útil cuando se busca disminuir la complejidad del modelo e identificar un subconjunto de variables con mayor participación en la predicción. En este trabajo se consideró como una de las alternativas de regresión regularizada para comparar su desempeño con otros modelos.

3.3.6 Árboles de decisión

Un árbol de decisión es un conjunto de oraciones incrustadas de la forma "si, entonces" que dividen el conjunto de datos de los predictores. Las particiones en el conjunto de predictores se utilizan para estimar la salida del modelo como se ejemplifica en la Figura 3.2 [26].

La estructura de un árbol se basa principalmente en tres puntos principales:

- El predictor o predictores que se utilizarán y el punto de partición del conjunto de datos.
- La profundidad o complejidad del árbol.
- La ecuación de predicción en los nodos terminales u hojas del árbol.

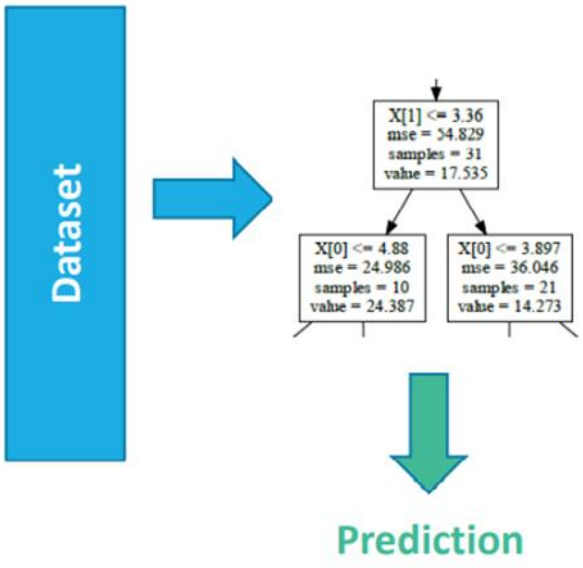


Figura 3.2: Esquema del árbol de decisión.

Para un problema de regresión, se requiere encontrar o proponer un modelo para lograr un mapeo entre dos conjuntos de datos.

$$Y = f(X) \tag{3.10}$$

donde Y y X son el conjunto de variables de salida o a estimar, y el conjunto de predictores o variables de entrada, respectivamente. La función f(·) será aproximada por el árbol de decisión.

Para la construcción de un árbol de decisión, se considera el conjunto de datos completo S, donde $X, Y \in S$. El conjunto de datos S se divide mediante una búsqueda exhaustiva de todas las particiones posibles para crear dos subconjuntos de datos $S1$ y $S2$, de tal forma que se minimiza un criterio de varianza conjunta, $L(X_{split})$.

$$L(x_{pj,split}) = \frac{1}{n_1}\sum_{i\in S_1}(y_i - \bar{y}_1)^2 + \frac{1}{n_2}\sum_{i\in S_2}(y_i - \bar{y}_2)^2 \tag{3.11}$$

Donde, $x_{pj,split}$ es un valor de partición en el predictor pj y los parámetros $\bar{y}_1$ y $\bar{y}_2$ son los promedios de las muestras de la variable de salida pertenecientes a los respectivos subconjuntos $S1$ y $S2$.

El valor de la función $L(x_{pj,split})$, depende de la mejor $x_{pj,split}$, que se selecciona para minimizar la varianza en la variable de salida Y. La búsqueda de la mejor partición debe realizarse de forma exhaustiva en todos los predictores x_{pj}, y solo se elige una partición de entre todos los predictores.

Este procedimiento se realiza recursivamente en los subconjuntos $S1$ y $S2$.

La función de predicción en cada hoja del árbol se define como el valor promedio de las muestras de salida del subconjunto de datos Sj correspondientes a la j-ésima rama o nodo final. La predicción $\hat{y}_i$ correspondiente a una muestra x_i se define como:

$$\hat{y}_{i.j} = \bar{y}_j \tag{3.12}$$

3.3.7 Bolsa de árboles de decisión

El ensamblaje de modelos es una estrategia que combina múltiples estimadores con la finalidad de obtener una predicción más estable. El método Bagging utiliza muestras generadas mediante bootstrapping para entrenar varios modelos del mismo tipo. En el caso de una bolsa de árboles de decisión, cada árbol genera una predicción para la misma observación y las predicciones se promedian para obtener el resultado final (ver Figura 3.3). De esta forma, la combinación de distintos árboles permite reducir la varianza de la predicción y mejorar su estabilidad [26].

La construcción de una bolsa de árboles se hace de forma sencilla:

1. Generar un subconjunto de datos a partir de los datos originales.
2. Entrenar un árbol de decisiones no podado basado en el subconjunto de datos.
3. Repetir los pasos 1 y 2 hasta que se formen m árboles de decisión.

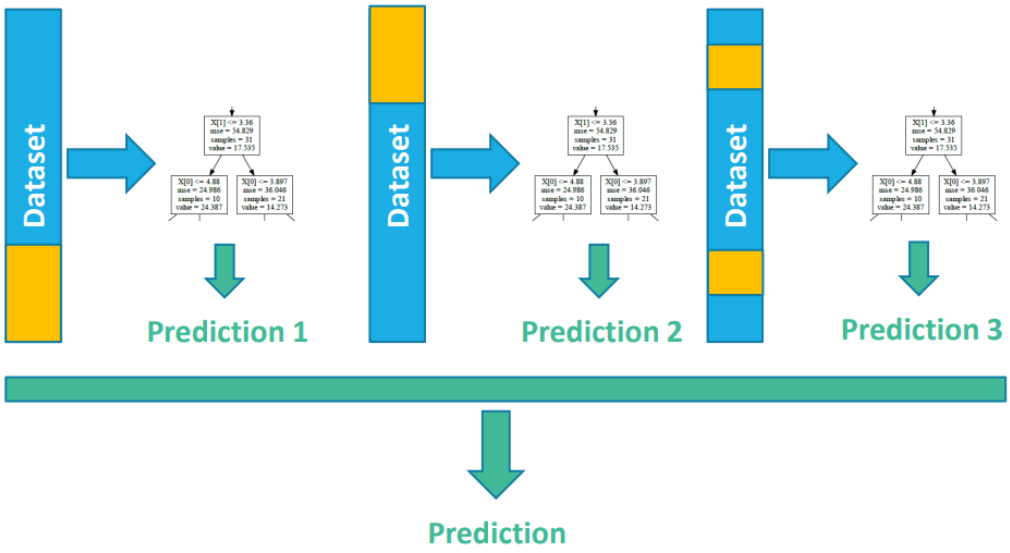


Figura 3.3: Esquema del modelo de bolsa de árboles de decisión.

3.3.8 *Red Neuronal*

Una red neuronal artificial es un modelo altamente no lineal y muy versátil que se usa comúnmente para resolver problemas de regresión o clasificación [26].

La forma en la que una neurona artificial procesa una señal se representa en dos etapas, tal como se muestra en la Figura 3.4 y se explica a continuación:

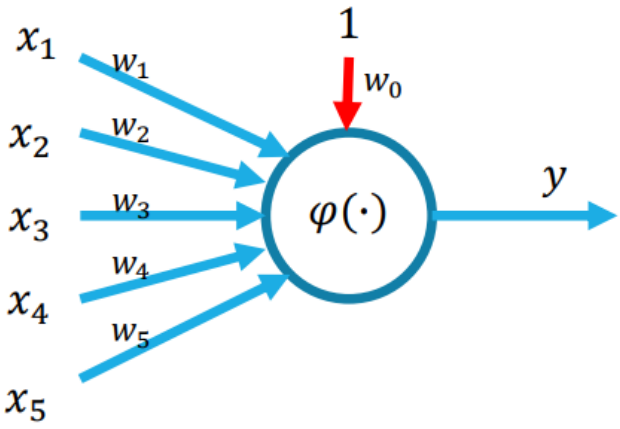


Figura 3.4: Representación neurona artificial.

Combinación de las entradas. Todas las señales que recibe la neurona deben ser consideradas para generar una respuesta. Cada señal recibida tiene asociada una constante w, que permite manipular la importancia de cada señal recibida o el aporte de información de cada entrada. Su representación matemática se da por la siguiente ecuación:

$$v = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5 \tag{3.13}$$

Respuesta neuronal. Considerando todas las señales de entrada, la neurona genera una señal de respuesta y=φ(v) por medio de una función de activación, la cual es diferenciable y ayuda a resolver el problema de optimización en el entrenamiento. Las funciones de activación más comunes son:

$$\text{Lineal: } \varphi(v) = v \tag{3.14}$$

$$\text{Sigmoidal: } \varphi(v) = \frac{1}{1+e^{-v}} \tag{3.15}$$

$$\text{Tangente hiperbólica: } \varphi(v) = \tanh(v) \tag{3.16}$$

Ahora bien, partiendo de la teoría de una neurona artificial, una red neuronal es la interconexión de múltiples neuronas para resolver un problema complejo que una sola neurona no podría. Su organización, como se visualiza en la Figura 3.5, se realiza en capas para simplificar el análisis: de entrada, ocultas y de salida.

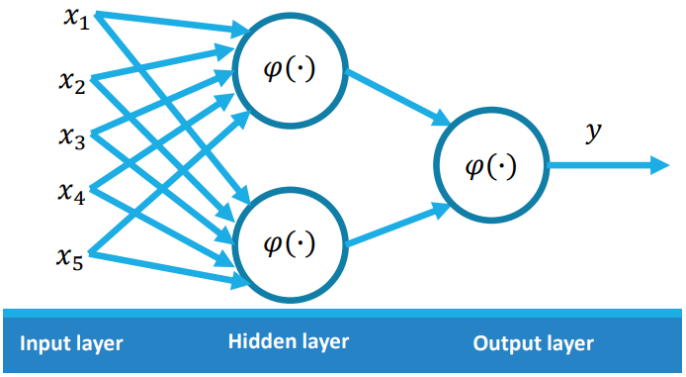


Figura 3.5: Capas de una red neuronal.

Intuitivamente, se puede ver que una red neuronal proporciona un modelo no lineal y esto puede ajustarse por medio de los pesos de cada capa.

Para un problema de regresión, se usa que la función de activación en la capa de salida sea la lineal, asimismo la evaluación de una red neuronal se realiza mediante una función de error o función de costo J.

$$J = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.17}$$

Donde, y_i es la salida deseada, $\hat{y}_i$ es la salida de la red neuronal para una entrada x_i y n es el número de salidas o neuronas en la capa de salida.

Por tanto, la evaluación de la red neuronal para un conjunto de muestras m se puede escribir como:

$$J = \frac{1}{2nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(y_{i,j} - \hat{y}_{i,j}\right)^2 \tag{3.18}$$

Se considera que un modelo neuronal ya tiene aprendido el comportamiento de los datos a ajustar cuando se tiene un valor mínimo de la función J. Para lograr esto es necesario encontrar los pesos W que minimicen la evaluación de J para todas las muestras de un conjunto de datos.

Ahora bien, cuando se aplica la función de costo J con un modelo neuronal, los pesos W tienen una relación no lineal a la salida y esto hace que J sea una función no convexa. Es decir, no hay una única solución W que minimiza la función de costo J. Como ayuda ante ello se utiliza un algoritmo de aprendizaje cuya función es la minimización de J modificando los parámetros W de la red neuronal.

Durante el proceso de aprendizaje o entrenamiento de una red neuronal, la actualización de los pesos se realiza iterativamente siguiendo la siguiente regla:

$$W_k = W_{k-1} + \Delta W_{k-1} \tag{3.19}$$

Los algoritmos de entrenamiento de redes neuronales más conocidos son:

a) Gradiente descendente
b) Método de Newton
c) Gauss-Newton
d) Levenberg-Marquardt

3.3.9 Métricas de evaluación

Para evaluar y comparar el desempeño de los modelos de regresión se utilizaron tres métricas: el coeficiente de determinación R^2, el Error Absoluto Medio (MAE) y la Raíz del Error Cuadrático Medio (RMSE). El uso conjunto de estas métricas permite observar tanto la capacidad del modelo para explicar la variabilidad presente en los datos como la magnitud de los errores obtenidos durante la predicción [27].

El **coeficiente de determinación R^2** mide la proporción de la variabilidad de la variable objetivo que puede ser explicada por el modelo. Su valor máximo es 1 y valores cercanos a este límite indican un mejor ajuste entre los valores observados y los estimados. Un valor de cero representa un desempeño equivalente a utilizar como predicción constante el valor promedio de la variable objetivo, mientras que también pueden obtenerse valores negativos cuando el desempeño del modelo es inferior a esta referencia [27].

El coeficiente de determinación puede expresarse como:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.20}$$

donde y_i representa el valor observado, $\hat{y}_i$ el valor estimado por el modelo y $\bar{y}$ el promedio de los valores observados.

En este trabajo, R^2 se utiliza como la principal referencia para conocer la capacidad predictiva general de los modelos y comparar su desempeño sobre el conjunto de datos de prueba. Sin embargo, esta métrica se complementa con medidas de error expresadas directamente en pesos mexicanos para facilitar la interpretación de los resultados.

El **Error Absoluto Medio o MAE (Mean Absolute Error)** calcula el promedio de la diferencia absoluta entre los valores observados y los valores estimados por el modelo [27]. Se define mediante la siguiente expresión:

$$MAE = (1/n)\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.21}$$

donde n representa la cantidad de observaciones evaluadas, y_i el valor observado y $\hat{y}_i$ el valor estimado.

Una ventaja de esta métrica es que conserva las mismas unidades de la variable que se busca predecir. En el caso de este trabajo, el MAE se expresa en pesos mexicanos y permite interpretar de manera directa la magnitud promedio del error cometido por el modelo. Un valor menor de MAE representa un mejor desempeño.

La **Raíz del Error Cuadrático Medio o RMSE (Root Mean Squared Error)** se obtiene a partir de la raíz cuadrada del promedio de los errores elevados al cuadrado [27]. Su expresión es:

$$RMSE = \sqrt{(1/n)\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.22}$$

Al elevar los errores al cuadrado, esta métrica otorga mayor peso a las desviaciones de mayor magnitud. Al aplicar posteriormente la raíz cuadrada, el resultado vuelve a expresarse en las mismas unidades de la variable objetivo [27]. Para este proyecto, el RMSE se expresa también en pesos mexicanos.

El análisis conjunto de R^2, MAE y RMSE permite comparar los modelos desde diferentes perspectivas. R^2 permite observar la proporción de variabilidad explicada, mientras que MAE y RMSE permiten conocer la magnitud de los errores de estimación. Por esta razón, la selección del modelo final no se realiza únicamente a partir de una sola métrica, sino considerando de manera conjunta su comportamiento sobre los datos de entrenamiento y prueba.

3.4 Inteligencia artificial generativa y extracción estructurada de información

La inteligencia artificial generativa comprende modelos capaces de generar o transformar contenido a partir de la información proporcionada como entrada. Dentro de este campo, los

modelos de lenguaje de gran escala o LLM (Large Language Models) pueden interpretar instrucciones y producir texto a partir del contexto recibido. Estas capacidades permiten utilizarlos no solamente para generar contenido, sino también como una interfaz para interpretar información expresada mediante lenguaje natural [28].

En aplicaciones de ciencia de datos, esta capacidad puede ser utilizada como complemento de modelos predictivos tradicionales. En lugar de sustituir al modelo encargado de realizar una predicción numérica, un modelo generativo puede utilizarse como una capa previa para interpretar información proporcionada por el usuario y transformarla a una estructura que posteriormente pueda ser consumida por otros componentes de una aplicación.

Esta separación es relevante para el presente trabajo, ya que la inteligencia artificial generativa se utiliza únicamente para facilitar la captura de las características de un inmueble. La estimación del precio continúa siendo responsabilidad del modelo de regresión seleccionado, manteniendo separadas la interpretación del lenguaje natural y la predicción numérica.

Los **modelos de lenguaje de gran escala** son sistemas entrenados sobre grandes cantidades de información y diseñados para procesar y generar secuencias de texto. A partir del contexto proporcionado en una instrucción, estos modelos pueden realizar diferentes tareas relacionadas con lenguaje, entre ellas clasificación, extracción, resumen, generación de contenido y transformación de información entre distintos formatos [28].

Una característica relevante de estos modelos es la posibilidad de interactuar con ellos mediante lenguaje natural, disminuyendo la necesidad de que un usuario conozca previamente la estructura técnica de la información requerida por una aplicación. De esta forma, una descripción escrita por una persona puede ser interpretada para identificar dentro de ella determinados conceptos o valores.

Sin embargo, la generación de una respuesta por parte de un modelo de lenguaje no implica que su contenido sea necesariamente correcto. Los modelos generativos pueden producir información incorrecta o interpretar de manera inadecuada una instrucción, por lo que sus resultados deben ser considerados dentro de un proceso que incluya mecanismos de validación y, cuando sea necesario, revisión por parte del usuario [28].

La **extracción estructurada de información** consiste en transformar información que originalmente se encuentra en un formato no estructurado, como una descripción escrita en lenguaje natural, en un conjunto de campos definidos que puedan ser procesados posteriormente por un sistema.

En el caso de los modelos de lenguaje, es posible definir previamente la estructura que se espera obtener como respuesta. Una alternativa para ello consiste en utilizar esquemas que establezcan los campos requeridos y el tipo de información asociado a cada uno. OpenAI denomina Structured Outputs a un mecanismo mediante el cual la salida de un modelo puede restringirse

para ajustarse a un esquema JSON definido por el desarrollador [29]. Entre sus casos de uso se encuentra precisamente la extracción de datos estructurados a partir de información no estructurada.

El cumplimiento de una estructura definida no garantiza por sí mismo que los valores contenidos en ella sean correctos. Por esta razón, la información extraída debe ser sometida posteriormente a reglas de validación y, cuando el contexto de la aplicación lo requiera, a la revisión del usuario antes de ser utilizada por un modelo predictivo.

Para el presente proyecto, este enfoque permite recibir una descripción de un departamento en lenguaje natural y transformarla en las variables que requiere la aplicación. Cuando alguna característica no se encuentra expresada en la descripción, ésta debe ser identificada como faltante y posteriormente completada o confirmada por el usuario. Una vez revisada la información, las variables pueden continuar hacia el proceso de validación y posteriormente al modelo encargado de realizar la estimación del precio.

Los conceptos revisados en este capítulo establecen la base teórica para el desarrollo de la solución propuesta. Las técnicas de análisis de datos permiten estudiar y preparar la información inmobiliaria, mientras que los distintos modelos de regresión proporcionan alternativas para estimar el precio a partir de las características disponibles. A su vez, las métricas R^2, MAE y RMSE permiten evaluar y comparar cuantitativamente su desempeño. Finalmente, la inteligencia artificial generativa se incorpora como un componente complementario para facilitar la interacción del usuario con la herramienta, manteniendo separada la interpretación del lenguaje natural del modelo responsable de realizar la estimación del precio. A partir de estos fundamentos, en el siguiente capítulo se describe la metodología utilizada para preparar los datos, construir y evaluar los modelos, así como para transformar el ejercicio analítico en una solución reproducible y utilizable por usuarios de negocio.

4 Metodología

En este capítulo se describe la metodología seguida para desarrollar la solución propuesta, desde la obtención y preparación de la información inmobiliaria hasta la construcción, evaluación e implementación del modelo predictivo. Posteriormente se presenta la transformación del prototipo analítico en un proceso reproducible, así como el desarrollo de los componentes de inferencia, validación y aplicación visual. Finalmente, se describe la incorporación de una capa opcional de inteligencia artificial generativa para facilitar la captura de información mediante lenguaje natural.

4.1 Procedencia y características de los datos

Los datos utilizados en este trabajo provienen de DIME (Dinámica del Mercado Inmobiliario), una fuente de información elaborada y comercializada por Softec para el análisis del mercado inmobiliario en México. El acceso a esta información se obtuvo en el marco de la relación comercial de la empresa con Softec. Para este caso se contó con información correspondiente a la plaza de Guadalajara.

Los reportes DIME se elaboran de manera periódica y se comercializan por plaza o ciudad. Para la versión final del proyecto se utilizó el archivo correspondiente a Guadalajara 4Q22, es decir, información de la plaza de Guadalajara correspondiente al cuarto trimestre de 2022. El archivo de datos original no es una fuente pública y no forma parte del repositorio público del proyecto; únicamente el código, la configuración del proceso y los artefactos que no contienen la información original se mantienen bajo control de versiones. Para reproducir completamente el proceso desde los datos originales es necesario contar previamente con acceso autorizado a dicha fuente.

Para efectos de este trabajo, la referencia al mercado de Guadalajara corresponde a la cobertura incluida en la plaza Guadalajara del DIME utilizado. Por ello, el alcance geográfico del conjunto de datos está determinado por la información disponible en dicha fuente y por los criterios de preparación aplicados posteriormente, y no por una delimitación administrativa específica de la Zona Metropolitana de Guadalajara.

La base DIME presenta una estructura tabular que combina campos comunes del desarrollo con bloques de información correspondientes a distintos prototipos inmobiliarios. En el archivo Guadalajara 4Q22, una misma fila puede contener información de hasta tres prototipos, identificados mediante los bloques A, B y C, por lo que la estructura original no corresponde directamente a una observación por producto. Entre los campos disponibles se encuentran variables cualitativas, como la clasificación, el tipo de producto, la colonia y el municipio, así como variables cuantitativas y temporales relacionadas con superficie, terraza, recámaras,

estacionamientos, niveles, precios, tiempo en comercialización, unidades, inventario y fechas de actualización y entrega. Esta estructura fue reconfigurada posteriormente para obtener una observación independiente por producto, como se describe en la sección 4.2. La descripción de las principales variables disponibles se presenta con mayor detalle en el Apéndice A.

De acuerdo con la metodología de levantamiento descrita por Softec en el material utilizado para este trabajo, la información se obtiene mediante distintas actividades de investigación de mercado, entre ellas visitas a proyectos inmobiliarios, encuestas y levantamientos a desarrolladores, revisión de medios publicitarios y consulta de información pública. Una de las técnicas utilizadas consiste en realizar recorridos físicos de las zonas de interés e identificar la oferta existente mediante un esquema de "comprador misterioso" [30].

La información obtenida mediante los recorridos se complementa con otras fuentes disponibles, como sitios de desarrolladores, medios de difusión y fuentes gubernamentales. A partir de la metodología descrita por Softec, en la Figura 4.1 se presenta una síntesis de elaboración propia del proceso general de obtención e integración de información [30].

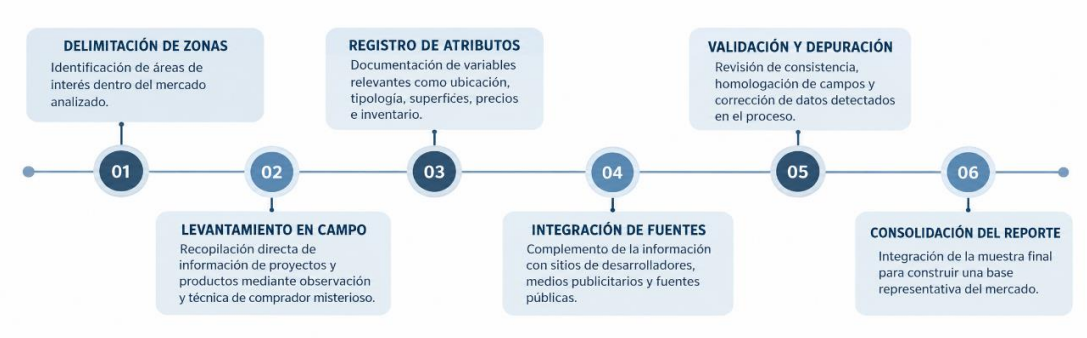


Figura 4.1: Síntesis del proceso de obtención e integración de información DIME. Elaboración propia con base en [30].

Entre las fuentes de información consideradas por Softec para complementar sus reportes se encuentran:

- Base de datos habitacional de Softec
- Censo Nacional de Población y Vivienda
- Directorio Estadístico de Unidades Económicas
- Consejo Nacional de Población
- Secretaría de Turismo (Datatur)

Para el desarrollo del modelo no se utilizó directamente la totalidad de campos disponibles en el archivo original. La información fue sometida posteriormente a un proceso de exploración,

filtrado, transformación y selección de variables con la finalidad de obtener el conjunto de datos utilizado para el entrenamiento de los modelos. Este proceso se describe en la siguiente sección.

4.2 Exploración y preparación de los datos

El proceso de exploración y preparación de los datos tuvo como finalidad transformar la información contenida en el reporte DIME Guadalajara 4Q22 en un conjunto de datos adecuado para el entrenamiento y evaluación de los modelos de regresión. Durante esta etapa se revisó la estructura original de la información, se homologaron variables y categorías, se aplicaron criterios de limpieza y filtrado y se generaron características necesarias para representar los productos inmobiliarios considerados en el estudio.

La preparación se desarrolló inicialmente como parte del prototipo analítico y posteriormente fue incorporada como una etapa independiente y reproducible dentro del pipeline del proyecto. En la versión final, esta lógica se concentra en la etapa prepare_data, cuyo resultado constituye el conjunto de datos preparado que posteriormente se divide en muestras de entrenamiento y prueba.

El archivo original Guadalajara 4Q22 contiene 288 registros y 57 columnas. Sin embargo, esta estructura no corresponde directamente a una observación por producto inmobiliario, debido a que una misma fila puede contener información de distintos prototipos de vivienda. Por esta razón, además de las actividades de limpieza y selección de variables, fue necesario reestructurar la información antes de utilizarla para el modelado.

a) Análisis exploratorio. Como primera etapa se revisó la estructura general de la información mediante estadísticas descriptivas y la inspección de las variables disponibles. El archivo Guadalajara 4Q22 contiene originalmente 288 registros y 57 columnas; sin embargo, cada renglón puede incluir información de hasta tres prototipos de vivienda identificados mediante los bloques A, B y C.

Para llevar la información a una estructura adecuada para el análisis, los datos comunes de cada desarrollo fueron combinados de forma independiente con cada uno de los bloques de producto A, B y C. Los registros resultantes se integraron posteriormente en una sola tabla, buscando que cada renglón representara un producto inmobiliario y que sus características se encontraran organizadas por columnas. Durante esta reestructuración se descartaron aquellos bloques que no contaban con la información necesaria para formar una observación completa.

Una vez reestructurada la información, se homologó la variable correspondiente al tipo de producto con la finalidad de delimitar el conjunto de datos al segmento de vivienda vertical considerado en este trabajo. Para efectos del ejercicio, las observaciones identificadas originalmente como departamentos o lofts se agruparon bajo una misma categoría de modelado, mientras que las tipologías correspondientes a vivienda horizontal fueron excluidas. De esta

manera, el conjunto utilizado en las etapas posteriores quedó concentrado en los productos verticales que constituyen el objeto de estudio.

También se revisó la distribución de los productos verticales por municipio. Después de la reestructuración y homologación del tipo de producto, y antes de aplicar los filtros posteriores, la mayor concentración de observaciones correspondía a Guadalajara y Zapopan, con 296 y 204 registros, respectivamente. Tlaquepaque, Tlajomulco de Zúñiga y Tonalá presentaban 14, 13 y 10 observaciones; Jocotepec 6, Chapala 3 y El Salto 1. Chapala y El Salto eran los dos municipios con menor representación y fueron excluidos de la versión final del conjunto de datos. Esta selección quedó definida de manera explícita dentro de los parámetros de preparación del pipeline y no mediante una regla automática basada en un número mínimo de observaciones por municipio.

Adicionalmente, durante la exploración inicial se revisaron variables relacionadas con las características físicas del producto, su comportamiento comercial y su ubicación. Esta revisión permitió identificar la necesidad de homologar algunas categorías, generar variables adicionales y establecer criterios de negocio para filtrar observaciones antes del entrenamiento de los modelos. Estas transformaciones se describen en los siguientes incisos.

b) Ingeniería y transformación de características. Como parte de la preparación de los datos se realizaron transformaciones sobre algunas variables con la finalidad de obtener una representación adecuada para los modelos de regresión.

Una de estas transformaciones correspondió a la clasificación asignada por Softec a cada producto. Debido a que esta variable presenta un orden entre categorías, se convirtió a una escala numérica ordinal de la siguiente forma: S = -2, E = -1, M = 0, R = 1 y RP = 2. Esta codificación permite conservar el orden relativo de las categorías dentro del conjunto de datos. Para efectos de la solución, la clasificación se considera una característica conocida del segmento del producto al momento de plantear el escenario de estimación, por lo que se incorpora como una variable de entrada junto con las demás características del departamento y del desarrollo.

Adicionalmente, se generó la variable months_to_delivery, que representa la cantidad de meses que faltaban para la entrega estimada del inmueble al momento de la actualización de cada registro. Para su cálculo se obtuvo la diferencia entre la fecha de entrega y la fecha de actualización registrada por Softec. Cuando la fecha de entrega era anterior a la fecha de actualización, el valor se estableció en cero, interpretando que el producto ya se encontraba entregado al momento del levantamiento.

Esta variable permite incorporar al modelo una característica relacionada con la etapa en la que se encontraba el desarrollo al momento de observar su precio. Esto resulta relevante debido a que un producto puede encontrarse en preventa, en proceso de construcción o ya terminado, condiciones que pueden estar asociadas con diferentes momentos de su estrategia comercial.

Finalmente, las variables de fecha utilizadas para construir esta característica fueron eliminadas del conjunto de datos destinado al entrenamiento, conservándose únicamente la medida derivada en meses.

c) Valores nulos. Durante la reestructuración de los bloques de producto A, B y C se identificaron observaciones que no contenían la información necesaria para conformar un registro completo. En estos casos se optó por descartar las observaciones incompletas en lugar de aplicar técnicas de imputación.

Posteriormente, una vez homologados los nombres y tipos de las variables, se realizó una nueva revisión de completitud antes de continuar con las etapas de filtrado e ingeniería de características. Como resultado de este proceso, el conjunto de datos preparado utilizado para el modelado no presenta valores nulos.

Esta decisión permitió trabajar únicamente con observaciones que contaran con la información requerida por el modelo, evitando introducir valores estimados artificialmente en las variables de entrada.

d) Tratamiento de valores extremos. Durante la exploración de la información se identificaron productos con características alejadas del tipo de desarrollo que la empresa consideraba dentro de su mercado objetivo. Para delimitar el ejercicio, se incorporaron criterios definidos con apoyo de los equipos de negocio de la organización, considerando que productos con superficies o precios por metro cuadrado considerablemente superiores a los rangos habituales tendrían un mercado potencial más reducido y no correspondían al tipo de proyectos que la empresa contemplaba desarrollar.

Para la versión final del proceso se conservaron únicamente observaciones con una superficie interior menor a 250 m^2 y un precio por metro cuadrado menor a $90,000 MXN. Estos valores se definieron como parámetros del proceso de preparación, de tal forma que los criterios utilizados para filtrar la información quedaran explícitos y pudieran ser reproducidos posteriormente.

El precio por metro cuadrado utilizado en este filtrado corresponde a una variable disponible originalmente en el reporte DIME y se empleó únicamente como apoyo para identificar observaciones fuera del rango considerado para el estudio. Una vez aplicado este criterio, dicha variable fue eliminada del conjunto de características utilizado para entrenar el modelo, evitando utilizar como predictor una medida directamente relacionada con la variable objetivo precio.

Estos filtros corresponden al contexto del mercado y a los criterios de negocio considerados durante el desarrollo del ejercicio en 2022, por lo que no deben interpretarse como límites generales o permanentes del mercado inmobiliario. En particular, el umbral de $90,000 MXN por metro cuadrado responde a las condiciones observadas en ese momento y podría requerir actualización al utilizar información de periodos posteriores. De manera similar, una propiedad

con características fuera de los rangos considerados requeriría una revisión adicional antes de utilizar la solución para estimar su precio.

e) Asimetría. Como parte del análisis exploratorio se evaluó la forma de las distribuciones de las variables numéricas mediante el cálculo de asimetría (skewness) y curtosis. Esta revisión se realizó con la finalidad de identificar variables cuya distribución pudiera beneficiarse de alguna transformación antes de ser utilizadas para el modelado.

De manera experimental se evaluó la transformación Box-Cox sobre las variables superficie interior (sqm), meses en venta (months_in_sale) y unidades proyectadas del plan maestro (master_plan_units). Debido a que esta transformación requiere valores estrictamente positivos, cuando alguna de las variables contenía valores iguales o menores a cero se aplicó previamente un desplazamiento para llevar todos sus valores al dominio requerido. Para cada variable, la implementación de SciPy estimó el parámetro λ correspondiente y posteriormente se calcularon nuevamente la asimetría y curtosis con el propósito de observar el efecto de la transformación.

Este ejercicio se utilizó como parte del análisis exploratorio y no como una transformación definitiva del conjunto de datos. Para evitar modificar la información original durante las pruebas se trabajó sobre una copia de los datos y, una vez concluido el análisis, el proceso de preparación para el modelado continuó utilizando las variables sin la transformación Box-Cox. Por esta razón, dicha transformación no forma parte del pipeline reproducible final.

f) División y escalamiento de los datos. Una vez concluida la preparación del conjunto de datos, se separó la variable objetivo correspondiente al precio actual del producto, identificada originalmente en los bloques A, B y C del reporte DIME como Precio Actual A, Precio Actual B y Precio Actual C, y homologada durante el procesamiento bajo el nombre interno price. Posteriormente, los registros se dividieron en dos conjuntos independientes: 80% para entrenamiento y 20% para prueba, utilizando una semilla aleatoria fija (random_state = 42) con la finalidad de mantener reproducible la partición. A partir de los 536 registros preparados se obtuvieron 428 observaciones para entrenamiento y 108 para prueba.

Después de realizar esta separación se aplicó un escalamiento mediante StandardScaler de la librería scikit-learn. Se utilizaron dos escaladores independientes: uno para las variables predictoras y otro para la variable objetivo. En ambos casos, los parámetros de escalamiento se estimaron únicamente con la información del conjunto de entrenamiento; posteriormente, esos mismos parámetros se utilizaron para transformar el conjunto de prueba.

Realizar el escalamiento después de la partición evita que información estadística del conjunto de prueba participe durante la preparación de los datos de entrenamiento, reduciendo así el riesgo de fuga de información (data leakage). La transformación estandariza las variables con respecto a la media y desviación estándar obtenidas en el conjunto de entrenamiento, facilitando además la aplicación posterior de modelos sensibles a diferencias de escala entre las variables.

Finalmente, los escaladores de las variables predictoras y de la variable objetivo se conservaron como artefactos del proyecto junto con los nombres y el orden de las características, permitiendo reutilizar exactamente las mismas transformaciones en las etapas posteriores de evaluación e inferencia.

g) Selección de características. Como parte del desarrollo experimental se utilizaron diferentes herramientas para analizar la aportación y posible redundancia de las variables disponibles antes de construir los modelos. Para ello se revisó la relación de las variables numéricas con el precio actual del producto y se complementó el análisis mediante componentes principales (PCA) y correlaciones entre las características.

En primer lugar, se analizó la relación lineal entre las variables numéricas y el precio actual del producto, variable posteriormente homologada bajo el nombre interno price. Adicionalmente, se realizó un PCA sobre las variables independientes previamente escaladas. Para este ejercicio se configuró el análisis para conservar el 95% de la varianza explicada acumulada, revisando posteriormente los pesos de las variables dentro de los componentes principales. El propósito fue utilizar esta información como apoyo para comprender la estructura del conjunto de datos e identificar posibles variables candidatas a ser descartadas. Los componentes principales obtenidos se utilizaron únicamente con fines exploratorios y no sustituyeron a las variables originales utilizadas posteriormente para entrenar los modelos.

También se estudió la relación entre las variables numéricas mediante una matriz de correlación de Spearman y un agrupamiento jerárquico basado en correlación. Este análisis permitió ordenar variables con comportamientos similares y facilitar la identificación de posibles redundancias. A diferencia del planteamiento realizado inicialmente, en la versión final del ejercicio no se utilizó un procedimiento automático que eliminara variables hasta alcanzar un umbral fijo de correlación; la selección se apoyó conjuntamente en los resultados exploratorios y en criterios relacionados con el problema de negocio.

Algunas variables fueron descartadas debido a la información que representaban o a su relación con otras características. Por ejemplo, las unidades vendidas (sold_units) presentaban una correlación aproximada de 0.96 con el total de unidades del desarrollo (total_units), por lo que se conservó esta última junto con otras variables que permiten representar el avance comercial. El precio por metro cuadrado (price_per_sqm) fue útil para el análisis y tratamiento de valores extremos, pero no se utilizó como predictor debido a que depende directamente del precio que se busca estimar y, por tanto, no constituye una característica independiente conocida para realizar una nueva estimación. Asimismo, el número de baños (baths) fue descartado al presentar información estrechamente relacionada con la superficie interior del producto (sqm).

A partir de este proceso se conservó un conjunto de características físicas, comerciales y de ubicación para la construcción de los modelos. La composición definitiva de las variables de entrada se presenta posteriormente al describir el conjunto de datos utilizado por la solución final.

Como complemento visual, la Figura 4.2 presenta la matriz de correlación de Spearman de las principales variables numéricas y ordinales conservadas en el conjunto de datos preparado final.

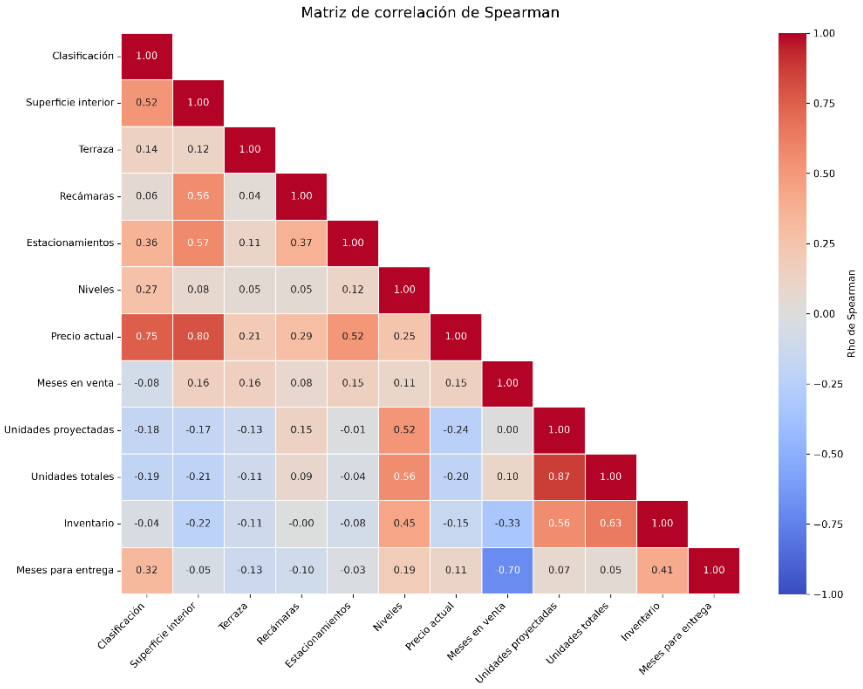


Figura 4.2: Matriz de correlación de Spearman variables del conjunto de datos preparado.

En la Figura 4.2 se observa que las asociaciones monotónicas más altas con el precio actual corresponden a la superficie interior (ϱ = 0.805), la clasificación (ϱ = 0.751) y el número de estacionamientos (ϱ = 0.516). Entre las variables explicativas también destacan la relación entre las unidades proyectadas y las unidades totales del desarrollo (ϱ = 0.87), así como la relación inversa entre los meses en venta y los meses para entrega (ϱ = -0.70). Estos resultados muestran que algunas características contienen información relacionada entre sí; sin embargo, en la versión final del ejercicio no se aplicó una eliminación automática basada únicamente en un umbral de correlación, sino que la selección consideró conjuntamente los resultados exploratorios y la utilidad de las variables para el problema de estimación.

h) Construcción de modelos. A partir del conjunto de entrenamiento preparado en las etapas anteriores se construyeron y evaluaron distintas alternativas de regresión con la finalidad de comparar enfoques lineales, modelos basados en árboles y un modelo no lineal. Durante esta etapa experimental se consideraron seis algoritmos: regresión lineal, árbol de decisión, bolsa de árboles (Bagging), red neuronal mediante MLPRegressor, regresión Ridge y regresión Lasso.

La regresión lineal se utilizó como modelo base para contar con un punto de referencia frente a alternativas de mayor complejidad. Posteriormente se evaluó un árbol de decisión y un modelo

de bolsa de árboles con el propósito de explorar relaciones no lineales entre las características y el precio. También se incorporó una red neuronal multicapa como una alternativa adicional de modelado no lineal. Finalmente, se evaluaron Ridge y Lasso para analizar modelos lineales regularizados, particularmente relevantes ante la presencia de información relacionada entre algunas de las variables predictoras.

Durante la experimentación se utilizaron técnicas de validación cruzada y búsqueda de hiperparámetros de acuerdo con las características de cada modelo. Por ejemplo, la regresión lineal fue evaluada mediante validación cruzada, mientras que para Bagging, Ridge y Lasso se implementaron búsquedas de hiperparámetros mediante GridSearchCV. Por esta razón, el procedimiento de ajuste no fue idéntico para todos los algoritmos, sino que se adaptó a las características y parámetros de cada alternativa.

El desempeño de los modelos se comparó posteriormente sobre el conjunto de prueba que permaneció separado del entrenamiento. Como resultado de esta evaluación se seleccionó Ridge Regression como el modelo definitivo de la solución. A partir de esta selección, la versión reproducible del proyecto conserva únicamente el entrenamiento de Ridge y realiza la búsqueda de su parámetro de regularización mediante GridSearchCV y validación cruzada con particiones KFold. El modelo con el mejor resultado de validación cruzada se guarda posteriormente como artefacto para las etapas de evaluación e inferencia.

4.3 Estructura final de los datos y variables del modelo

Una vez concluidas las etapas de reestructuración, limpieza, filtrado e ingeniería de características, se obtuvo un conjunto de datos preparado con 536 observaciones y 18 columnas, sin valores nulos. De estas columnas, una corresponde a la variable objetivo, es decir, el precio actual del producto, mientras que las 17 restantes constituyen las características utilizadas por el modelo.

Aunque el modelo recibe 17 columnas predictoras, estas representan conceptualmente 12 características del producto y del desarrollo. Once se incorporan directamente como variables numéricas u ordinales, mientras que el municipio constituye una variable categórica que se transforma internamente en seis indicadores booleanos mediante codificación dummy. De esta forma, una sola característica de ubicación genera las columnas correspondientes a Guadalajara, Jocotepec, Tlajomulco de Zúñiga, Tlaquepaque, Tonalá y Zapopan.

La variable objetivo corresponde al campo Precio Actual disponible originalmente para los prototipos A, B y C del reporte DIME. Durante la reestructuración estos campos se homologaron bajo el nombre interno price. De manera similar, otras variables provenientes de los diferentes bloques de producto fueron homologadas bajo un mismo identificador para facilitar su procesamiento. El código de preparación realiza, por ejemplo, las equivalencias entre Área y sqm, Terraza y terrace, Recámara y bhk, Cajón y park_u, y Niveles y levels.

La Tabla 4.1 presenta las características conceptuales utilizadas para generar una estimación, su identificador dentro del proyecto y su función general. Las definiciones completas de las variables provenientes del reporte DIME se encuentran también en el Apéndice A. La información original describe, entre otros campos, el área, terraza, recámaras, cajones de estacionamiento, niveles, meses en venta, unidades totales e inventario.

Tabla 4.1: Variables utilizadas para la estimación de precios.

Variable para el usuario	Nombre interno	Tipo en el modelo	Descripción
Clasificación Softec	classification	Ordinal numérica	Categoría de segmentación del producto, codificada de S = -2 a RP = 2
Superficie interior	sqm	Numérica	Metros cuadrados de construcción del prototipo
Terraza	terrace	Numérica	Metros cuadrados de terraza o patio
Recámaras	bhk	Numérica	Número de recámaras del producto
Cajones de estacionamiento	park_u	Numérica	Número de cajones incluidos
Niveles	levels	Numérica	Número de niveles del desarrollo o torre
Meses en venta	months_in_sale	Numérica	Tiempo transcurrido desde el inicio de comercialización
Unidades proyectadas	master_plan_units	Numérica	Unidades consideradas en el plan del desarrollo
Unidades totales	total_units	Numérica	Unidades correspondientes a la etapa disponible
Inventario	inventory	Numérica	Unidades disponibles para venta
Meses para entrega	months_to_delivery	Numérica derivada	Tiempo entre la fecha de actualización y la fecha prevista de entrega
Municipio	town → 6 dummies	Categórica transformada	Ubicación municipal del desarrollo

La variable municipio no permanece como una columna denominada town en el conjunto utilizado por Ridge. Durante la preparación se transforma en seis variables indicadoras: Guadalajara, Jocotepec, Tlajomulco de Zúñiga, Tlaquepaque, Tonalá y Zapopan. Por esta razón, las 12 características conceptuales capturadas para realizar una estimación se convierten en 17 características de entrada para el modelo. Al agregar la variable objetivo price, el conjunto preparado queda compuesto por las 18 columnas reportadas anteriormente.

La variable objetivo es el precio actual del producto (price), expresado en pesos mexicanos, correspondiente al precio reportado en DIME para el producto en comercialización al momento del levantamiento. Este valor se utiliza en el trabajo como referencia para la estimación de precios de mercado, sin que ello implique que corresponda al precio final de cierre de una operación. La variable objetivo no forma parte de las características proporcionadas al modelo durante la inferencia.

4.4 Industrialización y reproducibilidad del modelo

Una vez concluida la etapa experimental y seleccionado el enfoque de modelado, el desarrollo se reorganizó con la finalidad de separar las principales actividades del proceso y facilitar su reproducción. El prototipo inicial concentraba en un mismo archivo las actividades de exploración, preparación, entrenamiento y evaluación. Para la versión final, estas tareas se dividieron en etapas independientes con su propia responsabilidad.

Para administrar esta secuencia se utilizó Data Version Control (DVC), mediante el archivo dvc.yaml. El pipeline quedó compuesto por cuatro etapas principales: prepare_data, split_data, train y evaluate. DVC registra para cada una de ellas sus dependencias, parámetros y productos de salida, permitiendo identificar cuándo un cambio en los datos, el código o la configuración requiere volver a ejecutar una o más etapas.

La primera etapa, prepare_data, parte del archivo original Guadalajara 4Q22 y ejecuta las actividades de reestructuración, limpieza, filtrado e ingeniería de características descritas anteriormente. Como resultado genera el archivo prepared_data.csv, que constituye el conjunto de datos preparado para el modelado. Entre los parámetros utilizados en esta etapa se encuentran los municipios excluidos, los límites de superficie y precio por metro cuadrado, la codificación de la clasificación Softec y las columnas que deben descartarse.

La etapa split_data recibe el conjunto preparado y realiza la separación entre entrenamiento y prueba utilizando una proporción de 80% y 20%, respectivamente, con una semilla fija (random_state = 42). En esta misma etapa se ajustan los escaladores utilizando exclusivamente el conjunto de entrenamiento y se generan como artefactos scaler_X.pkl, scaler_Y.pkl y feature_names.json. Estos últimos permiten conservar tanto las transformaciones necesarias como el nombre y orden de las características que posteriormente espera el modelo.

La tercera etapa, train, utiliza exclusivamente los datos de entrenamiento para ajustar el modelo Ridge. La búsqueda del parámetro de regularización se realiza mediante GridSearchCV, empleando validación cruzada KFold de 10 particiones, una semilla fija de 42 y el coeficiente de determinación R^2 como criterio de evaluación. Los valores de alpha evaluados se encuentran definidos en params.yaml, evitando mantener esta configuración dispersa dentro del código. El mejor estimador obtenido se almacena como modelo_final.pkl.

Finalmente, la etapa evaluate utiliza el modelo entrenado y los conjuntos de entrenamiento y prueba para calcular las métricas finales de desempeño. Los resultados se almacenan en reports/metrics.json, lo que permite consultar de manera consistente los valores obtenidos después de cada ejecución del pipeline.

Además de la definición de las etapas en dvc.yaml, los principales parámetros del proceso se centralizaron en params.yaml. Esto permite modificar criterios como la partición de los datos, los filtros empleados durante la preparación o los valores de regularización evaluados sin alterar directamente la lógica de los scripts. En conjunto, esta estructura facilita la trazabilidad de los cambios y reduce el riesgo de que una nueva ejecución dependa de modificaciones manuales no documentadas.

El código, los archivos de configuración y los elementos no confidenciales del proyecto se mantienen bajo control de versiones con Git y se encuentran disponibles en el repositorio público del TOG Price Predictor [31]. Sin embargo, el archivo original Guadalajara 4Q22 contiene información comercial de acceso privado y, por esta razón, no forma parte del repositorio. Para reproducir completamente la preparación desde los datos originales es necesario contar previamente con dicho archivo en la ubicación esperada por el pipeline. Las etapas que integran este proceso, junto con sus principales entradas, actividades y salidas, se resumen en la Tabla 4.2.

Tabla 4.2: Etapas del pipeline reproducible.

Etapa	Entrada principal	Proceso	Salida principal
prepare_data	Guadalajara 4Q22.xlsx	Reestructuración, limpieza, filtros e ingeniería de características	prepared_data.csv
split_data	prepared_data.csv	Partición 80/20 y escalamiento	Train, test, escaladores y nombres de variables
train	Datos de entrenamiento	Ridge + GridSearchCV + KFold	modelo_final.pkl
evaluate	Modelo, train y test	Cálculo de métricas	metrics.json

Esta separación permite volver a ejecutar únicamente las etapas afectadas cuando cambian sus dependencias o parámetros, manteniendo una relación explícita entre los datos utilizados, el código ejecutado y los artefactos generados.

4.5 *Inferencia y validación de datos de entrada*

Una vez entrenado y almacenado el modelo, se desarrolló un componente independiente para realizar estimaciones sobre nuevos productos inmobiliarios. El proceso de inferencia utiliza el modelo Ridge previamente entrenado, los escaladores correspondientes a las variables predictoras y a la variable objetivo, así como el archivo que conserva los nombres y el orden de las características utilizadas durante el entrenamiento.

Para realizar una nueva estimación se capturan las 12 características conceptuales descritas en la sección anterior. A partir de ellas se construye una observación con exactamente las 17 columnas esperadas por el modelo. La clasificación Softec se transforma mediante la misma codificación ordinal empleada durante la preparación de los datos y el municipio se representa activando la variable indicadora correspondiente. El orden final de las columnas se obtiene directamente del artefacto feature_names.json, reduciendo el riesgo de que una nueva observación sea enviada al modelo con una estructura diferente a la utilizada durante su entrenamiento.

Antes de ejecutar la estimación se aplican validaciones orientadas a identificar inconsistencias lógicas y casos que se encuentren fuera del dominio considerado para el modelo. Por ejemplo, el inventario disponible no puede ser mayor que el número total de unidades del desarrollo. Asimismo, la superficie interior debe ser menor a 250 m^2, de acuerdo con el criterio de delimitación aplicado durante la preparación del conjunto de datos. Cuando una entrada incumple alguno de estos criterios, la predicción se bloquea hasta que el usuario corrija la información.

Además de estas validaciones, se comparan los valores numéricos capturados con los rangos observados en el conjunto de entrenamiento. Debido a que los archivos utilizados por el modelo contienen variables estandarizadas, estos rangos se recuperan previamente a su escala original utilizando el escalador ajustado durante el entrenamiento. Si una característica se encuentra por debajo del mínimo o por encima del máximo observado, el sistema genera una advertencia indicando que la estimación implica una extrapolación.

A diferencia de los errores de validación, una advertencia de extrapolación no bloquea automáticamente la estimación. Su propósito es informar al usuario que el modelo está siendo utilizado sobre una combinación de características que se encuentra fuera de los valores observados durante el entrenamiento y que, por tanto, el resultado debe interpretarse con mayor precaución.

Una vez superadas las validaciones bloqueantes, las 17 características se transforman utilizando el mismo escalador de variables predictoras empleado durante el entrenamiento. El modelo Ridge genera posteriormente una predicción en la escala normalizada de la variable objetivo y, finalmente, se aplica la transformación inversa mediante el escalador correspondiente al precio para expresar el resultado nuevamente en pesos mexicanos.

Como medida adicional de consistencia, los municipios utilizados durante una estimación deben pertenecer a las categorías disponibles para el modelo. En la aplicación visual esta condición se controla mediante una lista de opciones válidas y, adicionalmente, el componente de inferencia rechaza un municipio desconocido en caso de ser utilizado de manera independiente.

Este procedimiento permite conservar durante la inferencia las mismas transformaciones, nombres de variables y reglas generales utilizadas durante el entrenamiento. De esta forma, la generación de una nueva estimación no depende de reconstruir manualmente las características ni de aplicar nuevamente transformaciones fuera del proceso definido por la solución.

4.6 Aplicación visual para usuarios de negocio

Con la finalidad de facilitar el uso del modelo por personas sin conocimientos técnicos de programación, se desarrolló una aplicación visual mediante Streamlit. La interfaz permite capturar las características del departamento y del desarrollo, ejecutar las validaciones descritas anteriormente y obtener una estimación sin necesidad de interactuar directamente con los scripts de Python o con los artefactos del modelo.

El formulario organiza la información en dos grupos principales. El primero corresponde a las características del departamento, incluyendo municipio, clasificación Softec, superficie interior, superficie de terraza, número de recámaras y cajones de estacionamiento. El segundo concentra información del desarrollo, como número de niveles, meses en venta, unidades totales, unidades proyectadas, inventario disponible y meses para entrega. Para variables categóricas como municipio y clasificación se utilizan listas con las opciones reconocidas por el modelo, mientras que las variables numéricas se capturan mediante campos con restricciones básicas de formato y valores mínimos.

Antes de generar una estimación, la aplicación envía los valores capturados al componente de validación. Cuando se detecta una inconsistencia considerada bloqueante, se informa al usuario y no se realiza la predicción hasta que los datos sean corregidos. Por otra parte, cuando los valores se encuentran fuera de los rangos observados durante el entrenamiento, la aplicación permite continuar, pero muestra una advertencia indicando que el modelo está realizando una extrapolación y que el resultado debe interpretarse con mayor precaución. La Figura 4.3 muestra la interfaz utilizada para capturar y revisar estas características antes de generar la estimación.



Figura 4.3: Interfaz de captura del TOG Price Predictor.

Una vez procesada la información, la interfaz presenta como resultado principal el precio estimado en pesos mexicanos. Adicionalmente se calcula un precio estimado por metro cuadrado interior dividiendo la predicción obtenida entre la superficie capturada. Este segundo valor se presenta únicamente como una referencia derivada del resultado del modelo y no constituye una variable utilizada para generar la predicción. La aplicación también permite consultar un resumen de las características que fueron utilizadas para realizar la estimación.

Para proporcionar contexto sobre el desempeño del modelo, la aplicación consulta directamente el archivo reports/metrics.json y muestra las métricas R^2, MAE y RMSE obtenidas durante la evaluación. En particular, el MAE se presenta como una referencia del error absoluto promedio observado en el conjunto de prueba y se aclara expresamente que no representa un intervalo de confianza para una predicción individual. Esta distinción busca evitar que el usuario interprete la métrica de evaluación como un rango garantizado alrededor del precio estimado.

La interfaz también permite consultar los rangos mínimos y máximos observados en el conjunto de entrenamiento para las variables numéricas. Esta información complementa las advertencias de extrapolación y permite al usuario conocer el dominio de datos sobre el cual fue entrenado el modelo. Finalmente, la aplicación señala que el resultado corresponde a una estimación estadística y que no sustituye una valuación inmobiliaria formal.

La aplicación puede utilizarse mediante captura completamente manual. De manera adicional se incorporó una opción para interpretar descripciones en lenguaje natural mediante inteligencia artificial generativa; sin embargo, esta funcionalidad actúa únicamente como apoyo para completar el formulario y no modifica el modelo encargado de estimar el precio. Su funcionamiento se describe en la siguiente sección.

4.7 *Integración de inteligencia artificial generativa*

Como complemento a la captura manual de información se incorporó una capa opcional de inteligencia artificial generativa con la finalidad de facilitar la interacción con la herramienta. Esta funcionalidad permite que el usuario describa un departamento y su desarrollo mediante lenguaje natural y que la información proporcionada sea transformada en las variables estructuradas requeridas por la aplicación. La inteligencia artificial generativa no participa en el cálculo del precio, ya que la estimación continúa siendo realizada exclusivamente por el modelo Ridge Regression.

La implementación se desarrolló utilizando la API de OpenAI y una salida estructurada definida mediante un esquema de datos. En el código, dicho esquema se representa mediante la clase PropertyExtraction de Pydantic e incluye las 12 características conceptuales requeridas por la herramienta: municipio, clasificación Softec, superficie interior, terraza, recámaras, cajones de estacionamiento, niveles, meses en venta, unidades totales, unidades proyectadas, inventario y meses para entrega. Este enfoque corresponde al uso de salidas estructuradas descrito previamente en la sección 3.4 [29].

Todos los campos forman parte del esquema, pero pueden regresar con valor nulo cuando la información no está presente en la descripción del usuario. Las instrucciones enviadas al modelo establecen explícitamente que únicamente debe extraer información proporcionada en el texto, que no debe inventar valores faltantes, realizar supuestos sobre las características del inmueble ni calcular o sugerir un precio. Para la clasificación Softec se estableció adicionalmente que sólo puede extraerse cuando el usuario proporciona de manera explícita el código correspondiente, evitando inferirla a partir de la ubicación, superficie, descripción o precio del producto.

Los municipios y clasificaciones permitidos no se mantienen como listas independientes dentro de la capa generativa, sino que se obtienen de las opciones disponibles para el modelo predictivo. Después de recibir la respuesta estructurada se realiza una validación determinística adicional: el municipio se normaliza únicamente cuando corresponde con alguna de las categorías reconocidas y la clasificación se convierte al formato esperado y se conserva sólo cuando pertenece a las opciones válidas. En caso contrario, estos valores se consideran faltantes.

Una vez obtenida la extracción, la aplicación revisa cada campo antes de incorporarlo al formulario. Los valores ausentes se identifican como información faltante y los valores que no cumplen las condiciones básicas de formato se descartan. Los campos que no fueron identificados por la inteligencia artificial no son completados automáticamente; conservan el valor que ya se encontraba en el formulario para que el usuario pueda revisarlo o modificarlo. La interfaz informa qué variables fueron obtenidas de la descripción y cuáles requieren atención manual.

Cuando la descripción proporcionada se encuentra incompleta, la aplicación exige que el usuario confirme haber revisado y completado manualmente los campos faltantes antes de permitir la estimación. Después de esta revisión, la información continúa por las mismas

validaciones determinísticas descritas en la sección 4.5 y posteriormente se construyen las características de entrada para Ridge. De esta manera, la descripción proporcionada por el usuario se utiliza únicamente para facilitar la captura inicial de información. Una vez extraídos los datos, el usuario puede revisarlos y completar los campos faltantes antes de continuar con las validaciones de entrada. Sólo después de este proceso la información es enviada al modelo Ridge, que es el encargado de generar la estimación del precio.

La integración se diseñó además como un componente opcional. Si el servicio de inteligencia artificial no se encuentra disponible, ocurre algún error durante la interpretación o el usuario simplemente decide no utilizarlo, la aplicación permite continuar mediante la captura manual de las características. Por lo tanto, la disponibilidad de la capa generativa no constituye una dependencia para ejecutar el modelo de estimación de precios.

4.8 Arquitectura de la solución

La solución final se estructuró separando el proceso utilizado para construir y actualizar el modelo del proceso empleado para generar nuevas estimaciones. Esta separación permite distinguir las actividades que requieren acceso a la base histórica y entrenamiento de aquellas que se ejecutan cada vez que un usuario consulta un nuevo producto inmobiliario.

El primer flujo corresponde a la preparación y entrenamiento. Parte del archivo DIME Guadalajara 4Q22 y utiliza el pipeline administrado mediante DVC para ejecutar de manera secuencial la preparación de los datos, la separación de los conjuntos de entrenamiento y prueba, el entrenamiento del modelo Ridge y su evaluación. Como resultado se generan los artefactos necesarios para realizar estimaciones posteriores, entre ellos el modelo entrenado, los escaladores de las variables predictoras y del precio, el archivo con los nombres de las características y las métricas de evaluación.

El segundo flujo corresponde a la inferencia. En este caso, el usuario proporciona las características de un departamento y de su desarrollo mediante el formulario de la aplicación. La información pasa primero por las reglas de validación y posteriormente se transforma a la misma estructura utilizada durante el entrenamiento. A continuación, se aplica el escalador de las variables de entrada, se ejecuta el modelo Ridge y la predicción resultante se transforma nuevamente a pesos mexicanos para presentarla al usuario. Este procedimiento utiliza los artefactos generados durante el entrenamiento, por lo que no es necesario volver a entrenar el modelo cada vez que se realiza una consulta.

La capa de inteligencia artificial generativa se integra únicamente como una alternativa para facilitar la captura de información. Cuando el usuario decide utilizarla, la descripción en lenguaje natural se convierte en datos estructurados que posteriormente pueden ser revisados y modificados en el mismo formulario. A partir de ese punto, tanto la captura manual como la asistida por inteligencia artificial siguen el mismo proceso de validación e inferencia. Por lo tanto,

el componente generativo permanece separado del cálculo del precio y su disponibilidad no condiciona el funcionamiento del modelo predictivo.

La arquitectura general de la solución y la relación entre estos componentes se resumen en la Figura 4.4.

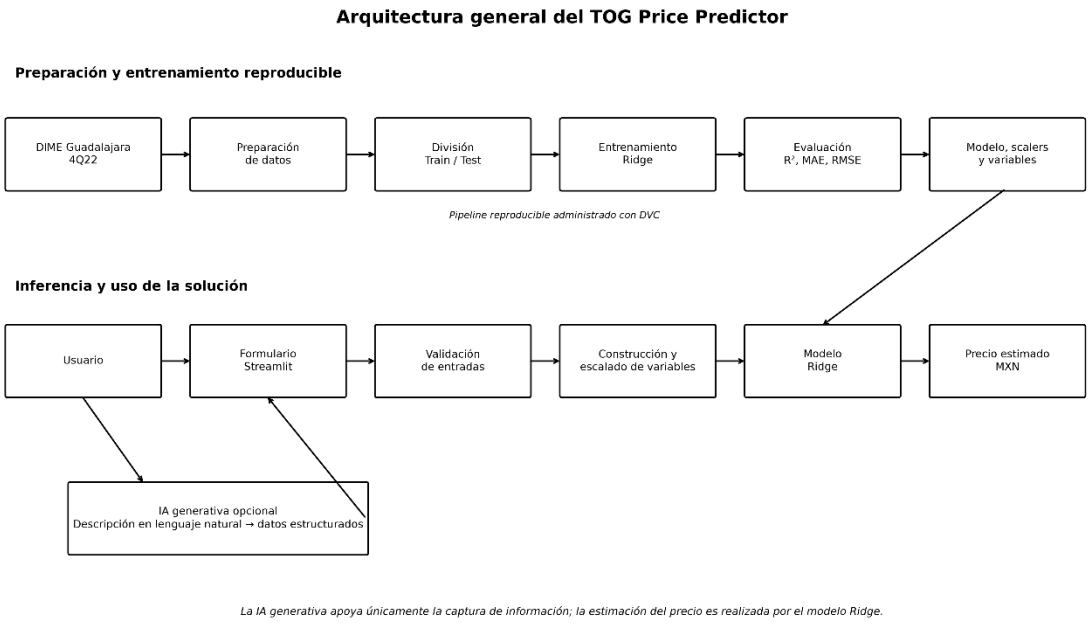


Figura 4.4: Arquitectura general del TOG Price Predictor.

Esta organización permite mantener separados los componentes de preparación y entrenamiento, la lógica de inferencia, las validaciones y la interfaz de usuario. También facilita que alguno de estos elementos pueda modificarse posteriormente sin concentrar nuevamente todo el funcionamiento de la solución en un solo archivo.

En conjunto, la metodología descrita en este capítulo permitió transformar la información inmobiliaria disponible en una solución reproducible para la estimación de precios. El proceso comprendió desde la reestructuración y preparación de los datos hasta la selección del modelo, su industrialización y la construcción de una interfaz orientada a usuarios de negocio. La incorporación de validaciones y de una capa opcional de inteligencia artificial generativa buscó facilitar el uso de la herramienta sin modificar la responsabilidad del modelo predictivo en el cálculo del precio. A partir de esta metodología, en el siguiente capítulo se presentan los resultados obtenidos y se analiza el desempeño de la solución desarrollada.

5 Resultados y conclusiones

5.1 Resultados y discusión

En este capítulo se presentan y discuten los principales resultados obtenidos durante el desarrollo y evaluación de la solución propuesta. Primero se compara el desempeño de los modelos de regresión considerados durante la etapa experimental y se analiza la selección de Ridge Regression como modelo final. Posteriormente se presentan las métricas obtenidas sobre los datos de prueba y los resultados de la integración del modelo dentro de la aplicación desarrollada. A partir de estos resultados se analizan los principales hallazgos, alcances y limitaciones del ejercicio, y se revisa el cumplimiento del objetivo general y de los objetivos específicos, considerando tanto el desempeño predictivo como la transformación del prototipo en una solución reproducible y utilizable por usuarios de negocio. Finalmente, se presentan las principales conclusiones del trabajo y se plantean posibles líneas de evolución derivadas de las limitaciones y oportunidades identificadas durante el proyecto.

5.1.1 Comparación y selección de modelos

A partir del conjunto de datos preparado se entrenaron y compararon seis alternativas de regresión utilizando la misma partición de entrenamiento y prueba. La Tabla 5.1 presenta el desempeño obtenido por cada modelo mediante R^2, RMSE y MAE. También se incluye la diferencia entre el R^2 de entrenamiento y prueba como una referencia para observar posibles diferencias en la capacidad de generalización de los modelos.

Tabla 5.1: Comparación de desempeño de los modelos evaluados.

Modelo	R^2 Train	R^2 Test	RMSE Test (MXN)	MAE Test (MXN)	Brecha R^2
Ridge Regression	0.87234	0.85886	$723,726.15	$550,099.67	0.01347
Lasso Regression	0.87215	0.85818	$725,465.49	$550,056.49	0.01396
Regresión lineal	0.87482	0.84971	$746,816.56	$569,668.62	0.02510
Bagging Trees	0.96962	0.84009	$770,361.94	$522,503.73	0.12953
Red neuronal (MLP)	0.90716	0.80705	$846,202.79	$599,456.30	0.10011
Árbol de decisión	0.99993	0.73385	$993,838.35	$665,438.17	0.26608

Los resultados muestran que Ridge Regression obtuvo el mayor coeficiente de determinación sobre el conjunto de prueba, con un R^2 de 0.8589, seguido por Lasso Regression con 0.8582 y la regresión lineal con 0.8497. Ridge obtuvo además el menor RMSE en prueba, con $723,726.15

MXN. En el caso del MAE, Bagging presentó el valor más bajo con $522,503.73 MXN, mientras que Lasso y Ridge obtuvieron resultados muy cercanos entre sí, con $550,056.49 MXN y $550,099.67 MXN, respectivamente.

La relación entre los precios reales y los estimados por Ridge en el conjunto de prueba se muestra en la Figura 5.1.

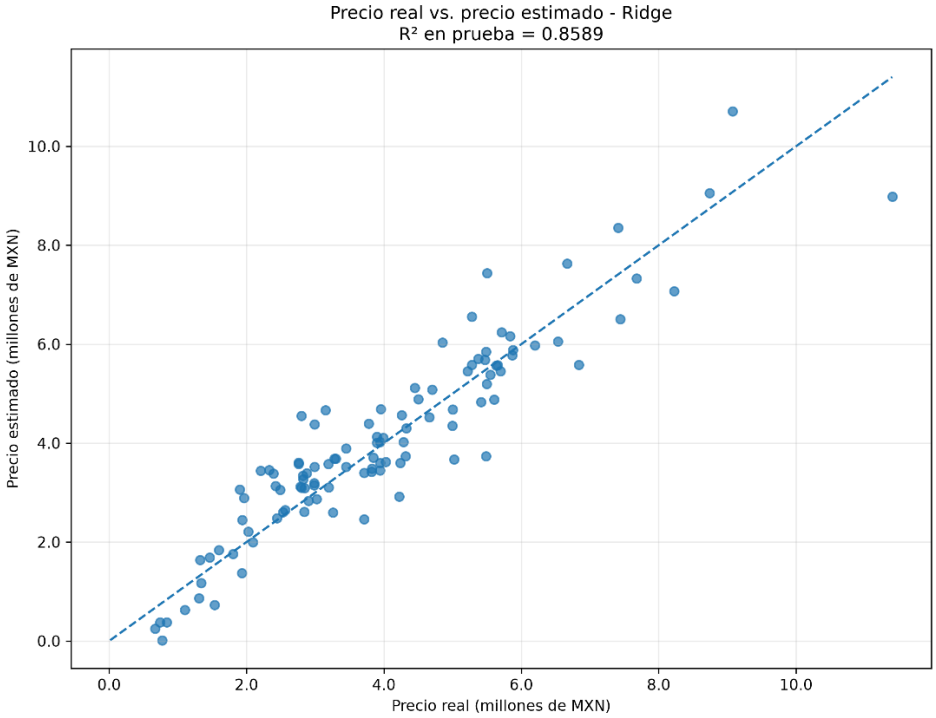


Figura 5.1: Precio real frente al precio estimado por Ridge en el conjunto de prueba.

En contraste, los modelos basados en árboles presentaron diferencias mayores entre entrenamiento y prueba. El árbol de decisión alcanzó un R^2 de 0.9999 en entrenamiento, pero disminuyó a 0.7339 en prueba. Bagging redujo esta diferencia y alcanzó un R^2 de 0.8401 en prueba, aunque mantuvo una brecha superior a la observada en los modelos lineales regularizados. La red neuronal obtuvo un R^2 de 0.8071 en prueba.

Considerando conjuntamente el desempeño sobre datos no utilizados durante el entrenamiento, el error obtenido y la diferencia entre entrenamiento y prueba, se seleccionó Ridge Regression como modelo final. Durante la búsqueda de hiperparámetros, el mejor valor de regularización fue $\alpha = 20$, obteniéndose un R^2 promedio de 0.8537 mediante validación cruzada sobre los datos de entrenamiento.

El resultado seleccionado supera el criterio establecido en el objetivo general de alcanzar un R^2 de al menos 0.85 sobre datos de prueba. Sin embargo, las métricas de error muestran que todavía existe una diferencia relevante entre algunos precios reales y sus estimaciones, por lo que el

resultado debe entenderse como una referencia cuantitativa para apoyar el análisis comercial y no como una valuación exacta de cada inmueble.

5.1.2 Verificación funcional de la solución

Con la finalidad de verificar el funcionamiento integrado de la solución, se realizó una prueba utilizando un mismo caso de control mediante la captura asistida por inteligencia artificial generativa. Se proporcionó una descripción correspondiente a un departamento ubicado en Zapopan, con clasificación Softec R, 85 m² de superficie interior, sin terraza, dos recámaras, un cajón de estacionamiento, ocho niveles en el desarrollo, 12 meses en venta, 100 unidades totales, 100 unidades proyectadas, un inventario de 30 unidades y 12 meses para la entrega.

A partir de esta descripción, la capa generativa identificó correctamente las 12 características requeridas por la aplicación, como se muestra en la Figura 5.2. Los valores extraídos fueron incorporados al formulario para su revisión y, una vez ejecutadas las validaciones de entrada, el modelo Ridge generó una estimación de $3,941,128 MXN, equivalente a $46,366 MXN por m² interior, como se observa en la Figura 5.3.

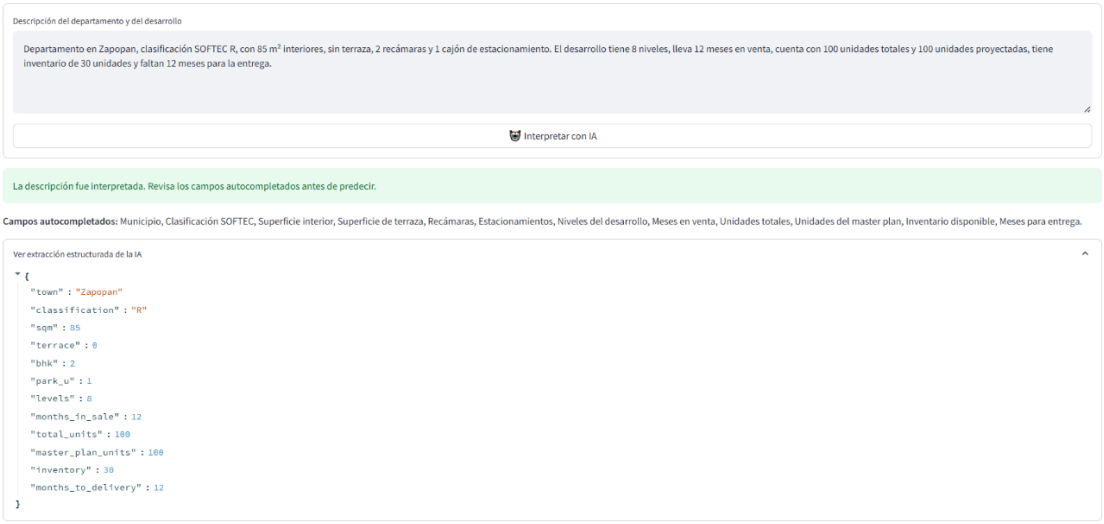


Figura 5.2: Extracción estructurada de las características mediante IA generativa.

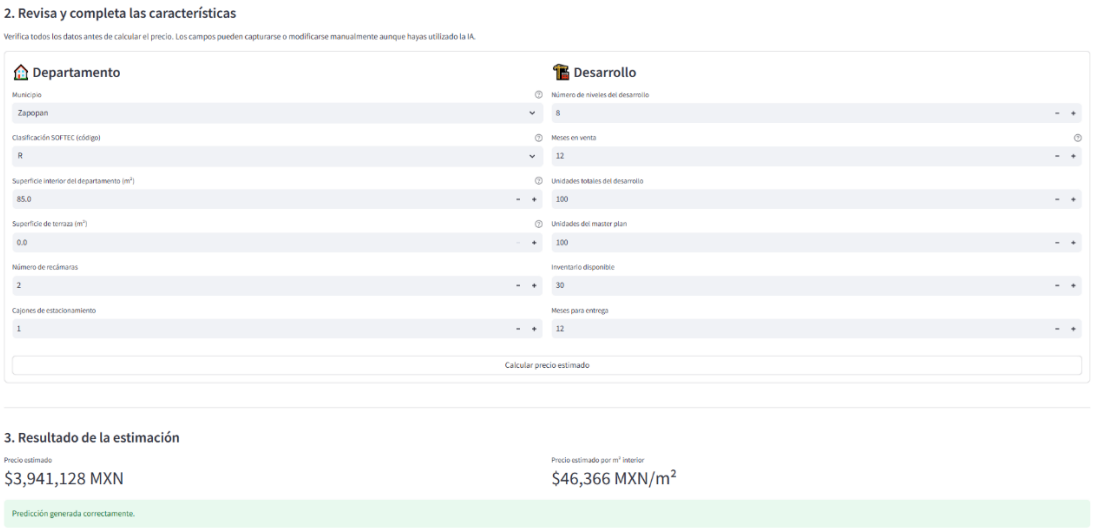


Figura 5.3: Formulario revisado y resultado de la estimación del TOG Price Predictor.

En este caso de control, el resultado obtenido después de utilizar la captura asistida por inteligencia artificial coincidió con el generado previamente mediante la captura manual de las mismas características. Esto permite comprobar que la capa generativa modifica únicamente la forma en que se incorporan los datos al formulario y no el procedimiento utilizado para estimar el precio. La prueba corresponde a una verificación funcional de la integración y no a una evaluación independiente de la precisión de la inteligencia artificial generativa.

5.1.3 Discusión de resultados

Los resultados obtenidos muestran que los modelos lineales presentaron un comportamiento competitivo para el conjunto de datos estudiado. Ridge, Lasso y la regresión lineal alcanzaron valores de R^2 en prueba cercanos entre sí, aunque los dos modelos regularizados presentaron una menor diferencia entre entrenamiento y prueba. En particular, Ridge obtuvo un R^2 de 0.8589 en prueba y una diferencia de 0.0135 respecto al conjunto de entrenamiento, lo que indica un comportamiento estable entre ambas particiones para este caso de estudio.

La comparación también muestra que la selección del modelo no debe realizarse utilizando una sola métrica. Bagging obtuvo el menor MAE en prueba, con $522,503.73 MXN, pero su R^2 disminuyó de 0.9696 en entrenamiento a 0.8401 en prueba y su RMSE fue mayor al obtenido por Ridge. Esta combinación de un MAE relativamente bajo y un RMSE más alto es compatible con la presencia de algunas predicciones con errores de mayor magnitud, ya que el RMSE penaliza con más fuerza este tipo de diferencias. De manera similar, el árbol de decisión presentó un R^2 prácticamente igual a uno sobre entrenamiento y de 0.7339 sobre prueba, evidenciando una diferencia considerable entre ambos conjuntos.

Ridge y Lasso obtuvieron resultados particularmente cercanos. Lasso alcanzó un MAE de $550,056.49 MXN frente a $550,099.67 MXN de Ridge, una diferencia de aproximadamente $43

MXN, mientras que Ridge presentó un R^2 ligeramente superior, un menor RMSE y una menor diferencia entre el R^2 de entrenamiento y prueba. Considerando conjuntamente estas métricas y el resultado de la validación cruzada, se seleccionó Ridge como modelo definitivo de la solución.

El R^2 de 0.8589 obtenido por Ridge supera el valor de 0.85 planteado como referencia en el objetivo general. Este resultado indica que, para los datos utilizados en el ejercicio, una parte importante de la variabilidad observada en los precios puede ser representada mediante las características disponibles. Sin embargo, el MAE de aproximadamente $550 mil MXN y el RMSE de aproximadamente $724 mil MXN muestran que todavía pueden existir diferencias relevantes entre el precio observado y el estimado para productos individuales. Por esta razón, la predicción debe utilizarse como una referencia adicional para el análisis y no como un valor exacto o una valuación inmobiliaria formal.

Como complemento a las métricas globales de desempeño, se analizaron los residuos obtenidos por el modelo Ridge sobre el conjunto de prueba. Para este análisis, el residual se calculó como la diferencia entre el precio real y el precio estimado, por lo que los valores positivos corresponden a casos en los que el modelo subestimó el precio y los negativos a casos en los que lo sobreestimó. La distribución obtenida se presenta en la Figura 5.4.

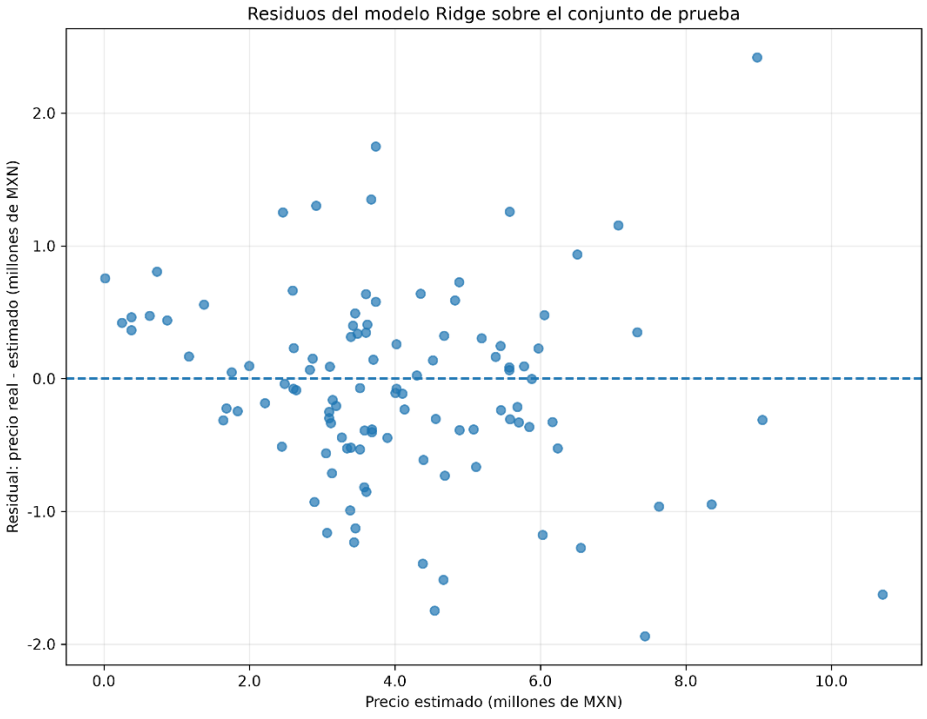


Figura 5.4: Distribución de los residuos del modelo Ridge en el conjunto de prueba.

Los residuos se distribuyen tanto por encima como por debajo de cero en la región donde se concentra la mayor parte de las observaciones, aunque también se presentan algunos errores de mayor magnitud. En los productos con precios estimados más altos se observa una mayor

dispersión en algunos casos, si bien el número de observaciones en ese rango también es menor. Este comportamiento muestra que, aun cuando el desempeño global del modelo resulta adecuado para el objetivo planteado, pueden existir productos individuales cuya estimación presente diferencias relevantes respecto al precio observado.

Otro aspecto relevante corresponde a las características disponibles para representar la ubicación. En la solución final esta información se incorpora a nivel de municipio, por lo que el modelo puede distinguir entre las zonas municipales presentes en el conjunto de entrenamiento, pero no captura directamente diferencias más específicas entre colonias, vialidades o entornos particulares dentro de un mismo municipio. Esto representa una limitación, ya que dos productos con características similares pueden presentar diferencias de precio asociadas con factores de ubicación que no se encuentran representados en las variables actuales.

Los criterios aplicados durante la preparación también deben interpretarse dentro del alcance del ejercicio. Los límites de superficie y precio por metro cuadrado utilizados para delimitar la muestra fueron definidos considerando el tipo de producto que la empresa consideraba relevante para sus posibles desarrollos y las condiciones del mercado observadas en 2022. Por tanto, estos valores no deben asumirse como límites permanentes del mercado. Una actualización de la información requeriría revisar nuevamente estos criterios antes de reentrenar el modelo.

La temporalidad de la información constituye otra limitación importante. El modelo fue construido a partir del reporte Guadalajara 4Q22, por lo que representa las relaciones existentes en la información disponible para ese periodo. Los cambios posteriores en precios, costos, oferta, demanda o composición de los productos pueden modificar dichas relaciones. En consecuencia, el desempeño obtenido no permite asumir que el modelo mantendrá las mismas métricas al utilizarse con información correspondiente a años posteriores sin realizar previamente una actualización y una nueva evaluación.

De manera similar, los resultados obtenidos no demuestran que el modelo entrenado para Guadalajara pueda utilizarse directamente en otras ciudades. Cada plaza puede presentar estructuras de precio, segmentos y condiciones de mercado diferentes. La metodología y el pipeline desarrollados sí pueden servir como base para realizar ejercicios posteriores con información de otras plazas, pero sería necesario preparar los datos correspondientes, volver a entrenar los modelos y evaluar su desempeño antes de considerar su utilización.

Además del desempeño predictivo, el proyecto permitió convertir el ejercicio experimental en una solución que puede ser utilizada sin interactuar directamente con el código. La aplicación visual facilita la captura y revisión de las variables, muestra advertencias cuando una observación se encuentra fuera de los rangos utilizados durante el entrenamiento y presenta las métricas del modelo como contexto para interpretar la estimación. La capa de inteligencia artificial generativa se incorporó como una alternativa adicional para capturar información mediante lenguaje natural, aunque su función se limita a estructurar los datos proporcionados por el usuario y no interviene en el cálculo del precio.

Finalmente, la separación del proyecto en etapas reproducibles permite que futuras actualizaciones puedan evaluarse de manera más controlada. Si la empresa dispone posteriormente de nuevas versiones del DIME de Guadalajara, será posible volver a ejecutar el proceso de preparación, entrenamiento y evaluación y comparar los nuevos resultados con los obtenidos en este trabajo. Esto resulta especialmente importante considerando que tanto las condiciones del mercado como el dominio de los productos de interés pueden modificarse con el tiempo.

En conjunto, los resultados muestran que Ridge cumple el criterio de desempeño planteado y que la solución puede integrarse en una herramienta reproducible y utilizable por usuarios de negocio. Al mismo tiempo, los errores observados, la temporalidad de la información y el nivel de detalle disponible para representar la ubicación delimitan el alcance de las estimaciones, por lo que su uso debe mantenerse como una referencia de apoyo al análisis comercial.

5.2 Conclusiones

El objetivo general del trabajo se considera cumplido, ya que fue posible desarrollar una solución basada en ciencia de datos que aprovecha la información inmobiliaria disponible en la base DIME de Guadalajara para estimar precios de departamentos. El modelo Ridge seleccionado obtuvo un R^2 de 0.8589 sobre el conjunto de prueba, superando el valor de 0.85 establecido como referencia en el objetivo general. Este resultado estuvo acompañado de un RMSE de $723,726.15 MXN y un MAE de $550,099.67 MXN, por lo que la solución debe entenderse como una referencia cuantitativa adicional para apoyar el análisis y no como una valuación exacta de cada inmueble.

Respecto al primer objetivo específico, fue posible explorar y preparar la información de DIME Guadalajara, transformando su estructura original en un conjunto de datos utilizable para modelado. A partir de este proceso se conservaron características físicas, comerciales y de ubicación. La ubicación pudo incorporarse a nivel de municipio; sin embargo, no se utilizó información más específica como colonia, por lo que esta dimensión permanece como una de las principales oportunidades para robustecer el modelo.

Para el segundo objetivo se construyeron y compararon seis alternativas de regresión: regresión lineal, árbol de decisión, Bagging, red neuronal, Ridge y Lasso. Los resultados mostraron que Ridge presentó el mayor R^2 en prueba y el menor RMSE, además de una diferencia reducida entre su desempeño en entrenamiento y prueba. Aunque Bagging obtuvo un MAE menor, presentó una brecha considerablemente mayor entre ambas particiones. Por esta razón, la selección se realizó considerando de manera conjunta R^2, RMSE, MAE y el comportamiento de generalización, y no únicamente una métrica aislada.

El tercer objetivo también se cumplió al transformar el prototipo inicial en un proceso reproducible. Las actividades de preparación de datos, separación de entrenamiento y prueba, entrenamiento y evaluación se dividieron en etapas independientes y se incorporaron

mecanismos para controlar parámetros, artefactos y métricas. Esto permite repetir el proceso de manera más controlada y establece una base para evaluar posteriormente nuevas versiones de la información.

El cuarto objetivo también se cumplió mediante el desarrollo de una aplicación visual que permite capturar las características de un departamento y obtener la estimación generada por el modelo sin necesidad de interactuar directamente con el código. La herramienta incluye validaciones sobre las entradas y advertencias cuando una observación se encuentra fuera de los rangos utilizados durante el entrenamiento, facilitando que el usuario tenga mayor contexto al interpretar el resultado.

Respecto al quinto objetivo, se incorporó de manera complementaria una capa de inteligencia artificial generativa para interpretar descripciones escritas en lenguaje natural y convertirlas en las variables requeridas por el formulario. Esta funcionalidad se mantuvo separada del cálculo del precio: la inteligencia artificial únicamente apoya la captura de información, mientras que la estimación continúa siendo realizada por Ridge. Además, cuando la descripción se encuentra incompleta, la aplicación requiere que el usuario revise y complete la información antes de continuar.

El sexto objetivo se atendió mediante la identificación de las principales limitaciones de la solución y de las líneas posibles para su evolución. El modelo fue construido con información correspondiente a Guadalajara 4Q22, utiliza ubicación a nivel de municipio y se encuentra condicionado por las variables y el tipo de producto presentes en esa fuente. Por tanto, sus resultados no permiten afirmar que el mismo modelo pueda utilizarse directamente con información actualizada, en otras ciudades o para otros tipos de vivienda. Estas extensiones requerirían nuevas bases, reentrenamiento y una evaluación independiente.

Más allá del resultado de una métrica específica, la principal aportación del trabajo consiste en demostrar que información especializada que ya se encontraba disponible en la empresa puede transformarse en una herramienta práctica para apoyar decisiones comerciales. El proyecto permitió pasar de un ejercicio exploratorio a una solución que integra preparación de datos, modelado, reproducibilidad, validaciones e interfaz de usuario, manteniendo al mismo tiempo las limitaciones del modelo visibles para quien utiliza la herramienta.

En este sentido, la solución no pretende sustituir el criterio de los especialistas comerciales ni los estudios de mercado utilizados por la organización. Su función es proporcionar una referencia adicional basada en datos que pueda complementar dichos procesos y servir como punto de partida para continuar evaluando el aprovechamiento de información inmobiliaria dentro de la empresa.

5.3 *Trabajo futuro*

A partir de los resultados y limitaciones identificados, una primera línea de trabajo consiste en actualizar el ejercicio con nuevas versiones de la información DIME de Guadalajara. El modelo desarrollado utiliza información correspondiente al cuarto trimestre de 2022, por lo que una actualización permitiría evaluar si las relaciones encontradas entre las variables se mantienen con el paso del tiempo y si el desempeño del modelo continúa siendo adecuado. Este proceso también requeriría revisar nuevamente algunos criterios utilizados durante la preparación de los datos, como los límites de superficie y precio por metro cuadrado, ya que fueron establecidos para el contexto del mercado analizado en ese momento.

Una segunda oportunidad consiste en incorporar variables que permitan representar con mayor detalle la ubicación y el entorno de los desarrollos. En este trabajo fue posible utilizar el municipio, pero no se contó con suficiente información para modelar de manera adecuada diferencias a nivel colonia. Futuras versiones podrían complementarse con información proveniente de otras fuentes, incluyendo características como cercanía a parques, centros educativos, servicios de salud, zonas comerciales, vialidades o indicadores relacionados con seguridad y accesibilidad. La utilidad de estas variables tendría que evaluarse mediante nuevas pruebas antes de incorporarlas al modelo.

También resulta de interés evaluar la metodología en otras plazas donde exista información suficiente. Esto no implica utilizar directamente el modelo entrenado para Guadalajara, ya que las condiciones y estructuras de precio pueden ser diferentes entre ciudades. La alternativa sería utilizar el proceso desarrollado como base para preparar la información de cada plaza, volver a entrenar los modelos y comparar sus resultados. En una etapa posterior también podría estudiarse la conveniencia de construir un modelo que incorpore información de varias ciudades, siempre que se cuente con una cantidad y calidad de datos suficientes para representar adecuadamente las diferencias entre mercados.

Otra extensión corresponde al desarrollo de un ejercicio similar para vivienda horizontal. El presente trabajo se concentró en departamentos para delimitar el problema inicial, pero la empresa también desarrolla productos de vivienda horizontal. Por esta razón, sería posible estudiar una solución equivalente para casas, considerando que las variables relevantes y la estructura de los datos podrían ser diferentes y requerirían un nuevo proceso de selección y evaluación.

La disponibilidad de información histórica de varios periodos permitiría además explorar problemas distintos a la estimación del precio actual. Por ejemplo, podría estudiarse la evolución de precios por proyecto, producto o competidor y evaluar técnicas orientadas a estimar precios futuros. Este tipo de análisis requeriría contar con múltiples cortes temporales comparables, condición que no se encuentra cubierta por la base utilizada en el presente trabajo.

De igual forma, una posible mejora sería incorporar información comercial que no se encuentra disponible en la fuente utilizada, particularmente descuentos, promociones o esquemas de pago. El precio analizado en este trabajo corresponde al precio reportado para los productos en comercialización, pero el valor final de una operación puede verse afectado por este tipo de condiciones. Contar con dicha información permitiría explorar modelos orientados a estimar no sólo el precio publicado, sino también referencias más cercanas a las condiciones comerciales efectivas.

Finalmente, aunque la solución ya cuenta con un pipeline reproducible y una aplicación funcional, una siguiente etapa consistiría en evaluar su incorporación dentro de un entorno corporativo de uso continuo. Esto implicaría definir mecanismos de actualización de datos, reentrenamiento, monitoreo de métricas, control de accesos y responsables del mantenimiento del modelo. La solución desarrollada proporciona una base para esta evolución, pero estas actividades forman parte de una posible implementación posterior y no del alcance del presente trabajo.

Referencias

[1] Remax, "El sector inmobiliario: qué es y cómo funciona" [Online document] Mar. 2021, Available at HTTP: https://www.franquiciaremax.es/el-sector-inmobiliario-que-es-y-como-funciona/

[2] R.E. Market, "Importancia del real estate en la economía" [Online document] Ago. 2020, Available at HTTP: https://realestatemarket.com.mx/mercado-inmobiliario/29655-importancia-del-real-estate-en-la-economia

[3] INEGI, "Vivienda" [Online document] Dic. 2021, Available at HTTP: https://www.inegi.org.mx/temas/viviendasat/

[4] Redacción Wiggot, "2.2 millones de empleos genera el desarrollo de vivienda en México" [Online document] Mar. 2022, Available at HTTP: https://wiggot.com/archivos/2-2-millones-de-empleos-genera-el-desarrollo-de-vivienda-en-mexico/#:~:text=Sector%20inmobiliario%20y%20su%20importancia,Construcci%C3%B3n

[5] Redacción Wiggot, "El sector inmobiliario y su importancia en la economía nacional" [Online document] Dic. 2020, Available at HTTP: https://wiggot.com/archivos/el-sector-inmobiliario-y-su-importancia-en-la-economia-nacional/

[6] Procuraduría Federal del Consumidor, "El sector inmobiliario en México" [Online document] Sep. 2012, Available at HTTP: https://www.gob.mx/profeco/documentos/el-sector-inmobiliario-en-mexico?state=published

[7] 4S Real Estate, "The inssights: gran reportaje de veticalización versión México 2021" [Online document] Jun. 2021, Available at HTTP: https://inmobiliare.com/por-que-se-esta-verticalizando-mexico/

[8] G. Towle, "Nueva normalidad inmobiliaria 2022" Softec, [Online document] (3), (2022 Sum), Available at HTTP: https://softec.com.mx/home/images/descargas/2022/220315SoftecBanamex.pdf

[9] M. Zhu, "Los mercados inmobiliarios, la estabilidad financiera y la economía " Fondo Monetario Internacional, [Online conference] (6), (2014 Sum), Available at HTTP: https://www.imf.org/es/News/Articles/2015/09/28/04/53/sp060514

[10] M2M, "¿Cuánto vale mi casa? La importancia de poner el precio de venta correcto para vender con éxito" [Online document] May. 2020, Available at HTTP: https://www.m2mteam.es/post/cu%C3%A1nto-vale-mi-casa-la-importancia-de-poner-el-precio-de-venta-correcto-para-vender-con-%C3%A9xito

[11] G. Madrid, “La importancia de poner el precio correcto a un inmueble agiliza la venta” [Online document] Nov. 2020, Available at HTTP: https://gozamadrid.com/2020/11/01/la-importancia-de-poner-el-precio-correcto-a-un-inmueble-agiliza-la-venta/

[12] I. Inmobiliaria, “Las claves de la expansión para los desarrolladores inmobiliarios” [Online document] May. 2021, Available at HTTP: https://www.inversioninmobiliariacr.com/es/marketing/item/2403-las-claves-de-la-expansion-para-los-desarrolladores-inmobiliarios

[13] I. Cruz, “Proptech: la digitalización agita el sector inmobiliario” [Online document] Mar. 2018, Available at HTTP: https://www.bbva.com/es/proptech-digitalizacion-pone-patas-sector-inmobiliario/

[14] D. González, “¿Cómo ha avanzado la industria PropTech en México?” [Online document] Dic. 2020, Available at HTTP: https://inmobiliare.com/como-ha-avanzado-la-industria-proptech-en-mexico/

[15] Apto, “Proptech: la evolución de la industria inmobiliaria hacia una era digital" [Online document] Nov. 2018, Available at HTTP: https://apto.mx/wp-content/uploads/2021/11/Apto_Reporte_Industria-Inmobiliaria.pdf

[16] Apac, “PropTech: the next big thing in real estate" [Online document] Feb. 2022, Available at HTTP: https://apacentrepreneur.com/proptech-the-next-big-thing-in-real-estate/

[17] PropTech, “A las puertas de la tercera generación PropTech: de 1.0 al 3.0" [Online document] Feb. 2018, Available at HTTP: https://proptech.es/las-puertas-la-tercera-generacion-proptech-1-0-3-0/

[18] Avaclick, “Conoce el precio de cualquier propiedad en México,”[Online document], Available at HTTP: https://avaclick.com/Home/InicioN

[19] Zillow, “What is a Zestimate?,” [Online document], Available at HTTP: https://www.zillow.com/zestimate/

[20] Housell, “¿Cuál es el valor real de una vivienda? Diferentes métodos para calcularlo" [Online document] Ene. 2019, Available at HTTP: https://www.housell.com/blog/casa-valoracion-tasacion/

[21] J. Amat, “Correlación lineal y regresión lineal" [Online document] Jun. 2016, Available at HTTP: https://www.cienciadedatos.net/documentos/24_correlacion_y_regresion_lineal

[22] J. Amat, “Análisis de componentes principales (Principal Component Analysis, PCA) y t-SNE" [Online document] Jun. 2017, Available at HTTP: https://www.cienciadedatos.net/documentos/35_principal_component_analysis

[23] J. Amat, “Introducción a la regresión lineal múltiple" [Online document] Jul. 2016, Available at HTTP: https://www.cienciadedatos.net/documentos/25_regresion_lineal_multiple.html

[24] A. E. Hoerl y R. W. Kennard, “Ridge Regression: Biased Estimation for Nonorthogonal Problems,” Technometrics, vol. 12, no. 1, pp. 55-67, 1970.

[25] R. Tibshirani, “Regression Shrinkage and Selection Via the Lasso,” Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, pp. 267-288, 1996.

[26] M. Kuhn y K. Johnson, Applied Predictive Modeling, New York, NY: Springer, 2013. DOI: 10.1007/978-1-4614-6849-3.

[27] Scikit-learn developers, “Metrics and scoring: quantifying the quality of predictions,” Scikit-learn 1.4.2 documentation, 2024, Available at HTTP: https://scikit-learn.org/1.4/modules/model_evaluation.html

[28] OpenAI, “GPT-4 Technical Report,” arXiv:2303.08774, 2023.

[29] OpenAI, “Introducing Structured Outputs in the API,” [Online document], Aug. 2024, Available at HTTP: https://openai.com/index/introducing-structured-outputs-in-the-api/

[30] SOFTEC, S.C., DIME Habitacional Guadalajara, Año 33, No. 3862, noviembre de 2022. Publicación comercial.

[31] D. Martínez Escobosa, “TOG Price Predictor,” GitHub repository, 2026, Available at HTTP: https://github.com/DanielMartinezEs/tog_price_predictor.

Apéndice A. Glosario de variables

A continuación, se presentan algunas de las principales variables contenidas en los reportes DIME utilizados en este trabajo. Estas variables describen características asociadas tanto a los productos inmobiliarios como a los desarrollos a los que pertenecen. Para cada una se indica el tipo de información que contiene y una breve descripción de su significado. Las definiciones se apoyan en el glosario incluido en el DIME Habitacional Guadalajara utilizado como fuente de información [30].

- Clasificación (cualitativa): categoría de segmentación asignada por Softec al producto inmobiliario. Los códigos utilizados en este trabajo corresponden a S = Social, E = Económica, M = Media, R = Residencial y RP = Residencial Plus.
- Tipo (cualitativa): tipología de la vivienda. Por ejemplo, departamento, loft, town house, etcétera.
- Nombre (cualitativa): nombre del proyecto o desarrollo inmobiliario.
- Promotor (cualitativa): nombre de la empresa promotora o desarrolladora asociada al proyecto inmobiliario.
- Dirección (cualitativa): domicilio en donde está ubicado el desarrollo inmobiliario.
- Colonia (cualitativa): colonia dentro de la plaza en la que está ubicado el desarrollo inmobiliario.
- Municipio (cualitativa): municipio de la entidad en la que está situado el desarrollo inmobiliario.
- Área (cuantitativa): superficie construida habitable asociada al prototipo, expresada en metros cuadrados.
- Terraza (cuantitativa): cantidad de metros cuadrados de terraza o patio del prototipo.
- Terreno (cuantitativa): cantidad total de terreno sobre el cual está desplantada la construcción.
- Recámaras (cuantitativa): número de habitaciones que tiene la propiedad.
- Baños (cuantitativa): número de baños que tiene la propiedad.
- Alcoba (cualitativa): detalla si la unidad cuenta o no con alcoba.
- Cuarto de servicio (cuantitativa): cantidad de cuartos de servicio en la unidad.
- Cajón (cuantitativa): número de espacios de estacionamiento considerados para el prototipo en el precio de oferta.
- Niveles (cuantitativa): cantidad de pisos o plantas reportados para el inmueble o torre. Nota: para los casos de departamentos o lofts se refiere a la cantidad de niveles en la torre.

- Precio actual (cuantitativa): precio de oferta observado en el corte más reciente de información.
- Precio inicial (cuantitativa): precio de oferta registrado al comienzo de la comercialización del producto.
- Valor M2 (cuantitativa): relación entre el precio reportado y la superficie construida del producto, expresada en pesos por metro cuadrado.
- Absorción (cuantitativa): velocidad mensual de desplazamiento de unidades registrada por el desarrollo.
- Meses en venta (cuantitativa): tiempo transcurrido en meses desde el inicio de la comercialización hasta la fecha correspondiente al corte de información.
- Meses de inventario (cuantitativa): estimación del tiempo necesario para desplazar las unidades disponibles considerando el ritmo de venta observado.
- Éxito comercial (cuantitativa): indicador porcentual que relaciona el desplazamiento mensual de unidades con el tamaño del desarrollo.
- Unidades proyectadas (cuantitativa): total de unidades contempladas en el plan general del desarrollo (master plan), incluyendo etapas presentes o futuras.
- Unidades totales (cuantitativa): cantidad de unidades correspondientes a la etapa disponible al momento del levantamiento.
- Inventario (cuantitativa): cantidad de unidades que permanecen disponibles para comercialización en el desarrollo.
- Vendidas (cuantitativa): número de unidades vendidas al corte de la fecha de actualización.
- Fecha de alta (temporal): fecha en la que el desarrollo fue incorporado a los registros utilizados por Softec.
- Fecha de inicio (temporal): fecha reportada para el comienzo de la venta o preventa del desarrollo.
- Fecha de actualización (temporal): fecha última de medición en la cual se capturó información del producto o desarrollo habitacional.
- Fecha de entrega (temporal): fecha en la cual se terminó o terminaría la obra del inmueble para hacer la entrega física al cliente (escriturarle).
- Financiamiento primera opción (cualitativa): institución financiera reportada como primera alternativa de financiamiento ofrecida al comprador (en el caso de que aplique el método de pago).

Apéndice B. Abreviaciones

- API: Application Programming Interface (Interfaz de Programación de Aplicaciones)
- DIME: Dinámica del Mercado Inmobiliario (DIME Habitacional de Softec)
- DVC: Data Version Control
- IA: Inteligencia Artificial
- INEGI: Instituto Nacional de Estadística y Geografía
- LLM: Large Language Model (Modelo de Lenguaje de Gran Escala)
- MAE: Mean Absolute Error (Error Absoluto Medio)
- MLP: Multi-Layer Perceptron (Perceptrón Multicapa)
- MXN: Peso mexicano
- PCA: Principal Component Analysis (Análisis de Componentes Principales)
- PIB: Producto Interno Bruto
- R^2: Coeficiente de determinación
- RMSE: Root Mean Squared Error (Raíz del Error Cuadrático Medio)
- SHF: Sociedad Hipotecaria Federal
- TOG: Trabajo de Obtención de Grado
- IoT: Internet of Things (Internet de las Cosas)
- JSON: JavaScript Object Notation