Instituto Tecnológico Autónomo de México

Valuación de propiedades residenciales en México usando *Machine Learning*

Caso

que para obtener el título de

Maestro en Ciencia de Datos

Presenta

Román Alberto Vélez Jiménez

Ciudad de México 2024

«Con fundamento en los artículos 21 y 27 de la Ley Federal del Derecho de Autor y como titular de los derechos moral y patrimonial de la obra titulada "**Confianza institucional, inclusión social, intensidad religiosa y percepción tecnológica como factores de innovación para el crecimiento económico**", otorgo de manera gratuita y permanente al Instituto Tecnológico Autónomo de México y a la Biblioteca Raúl Bailléres Jr., la autorización para que fijen la obra en cualquier medio, incluido el electrónico, y la divulguen entre sus usuarios, profesores, estudiantes o terceras personas, sin que pueda percibir por tal divulgación una contraprestación.»

Fecha

Román Alberto Vélez Jiménez

Índice

1. Introducción

El mercado inmobiliario, conocido como *real estate*, es un sector económico que abarca la compra, venta y alquiler de bienes inmuebles. Este mercado presenta particularidades que lo distinguen de otros sectores, como la heterogeneidad de los activos, la baja liquidez y la asimetría de la información. Por ejemplo, no existen dos propiedades totalmente idénticas; aunque dos casas compartan ubicación, tamaño y antigüedad, pueden diferir en su orientación, distribución interna y calidad de los acabados, lo que complica la comparación entre propiedades [McDonald and McMillen, 2010]. En cuanto a la liquidez, los bienes inmuebles son activos que no pueden venderse o comprarse con la misma rapidez que otros activos financieros, como acciones o bonos, lo que genera una oferta relativamente inelástica y, por ende, un aumento más pronunciado en los precios [Ahlfeldt and Liao, 2024]. Asimismo, la asimetría de información es un factor relevante en la compra-venta de inmuebles, ya que los vendedores poseen mayor conocimiento sobre la propiedad que los compradores, lo que provoca valoraciones subjetivas [Akerlof, 1978].

En este contexto, la información adquiere un rol central, ya que los precios de los inmuebles varían en función de factores como la ubicación, el tamaño y la antigüedad, entre otros. Por tanto, una valoración precisa de una propiedad requiere un análisis exhaustivo de la información disponible [Real Estate Market, 2024a].

En México, el sector inmobiliario es crucial para la economía, representando el 12.48 % del PIB nacional en 2022, de acuerdo con datos de la Asociación Mexicana de Profesionales Inmobiliarios (AMPI) [Real Estate Market, 2024b]. Por ejemplo, durante el segundo trimestre de 2024, este sector aportó 2.8 billones de pesos, lo que representa un incremento del 1 % en comparación con el trimestre anterior según la Secrertaría de Economía de México [de Economía, 2024]. La importancia de este sector en la economía mexicana se refleja en la creación de empleo, la atracción de inversión extranjera y la generación de riqueza [de Economía, 2024]. En particular, el submercado inmobiliario residencial ha adquirido mayor relevancia, con una apreciación bruta del 87.9 % y nominal del 26.7 % entre 2015 y 2023 [Guide, 2024]. Parte de este crecimiento se debe al aumento en los costos de materiales y mano de obra, lo que incrementó el precio de la vivienda nueva en un 10.9 % en 2023 con respecto al año anterior. Además, la apreciación de zonas turísticas, como Quintana Roo y Baja California Sur, ha contribuido al crecimiento del mercado, con incrementos del 15.5 % y 16.7 %, respectivamente, en los últimos dos años [Guide, 2024], [Research, 2024]. En consecuencia, es esencial contar con un índice de precios de la vivienda que sea preciso, transparente y su metodología sea vanguardista para asegurar la calidad de dicho índice.

En 2002, la Sociedad Hipotecaria Federal (SHF) implementó un mecanismo oficial para monitorear los precios del mercado inmobiliario residencial [Real Estate Market, 2024a]. Este índice utiliza un modelo hedónico [Rosen, 1974], que considera características estructurales como la superficie construida, número de recámaras y baños, ubicación y antigüedad para estimar el valor de una propiedad [Guerrero Espinosa, 2005]. Aunque esta metodología fue innovadora en su momento, no ha sido actualizada en más de dos décadas. Dado que las relaciones entre las variables y el precio de la vivienda han cambiado y han surgido nuevas variables que afectan los precios, como por ejemplo el valor de las propiedades comparables, la metodología actual del índice puede no reflejar con precisión la realidad del mercado inmobiliario. Esto podría ocasionar que las políticas públicas derivadas del índice no sean adecuadas, lo que llevaría a distorsiones en el mercado.

1.1. Motivación

El índice de la SHF es una herramienta crucial utilizada por el Banco de México para la toma de decisiones en política monetaria, por el Instituto Nacional de Estadística y Geografía (INEGI) para el cálculo de la inflación, y por la Comisión Nacional Bancaria y de Valores (CNBV) en la regulación de las instituciones financieras. La relevancia de este índice en la formulación de políticas públicas y privadas es incuestionable.

Recientemente, la empresa Metrics Analytics puso a disposición del Instituto Tecnológico Autónomo de México (ITAM) una base de datos que contiene los avalúos certificados en México desde 2019 hasta 2023, realizados por peritos valuadores acreditados. Esta base de datos es la más completa y actualizada en el país, ya que incluye todos los avalúos realizados en este periodo. Los precios de venta reflejados en estos avalúos representan estimaciones objetivas del valor de mercado de las propiedades, a diferencia de los precios subjetivos que pueden encontrarse en los listados de portales inmobiliarios. La base cuenta con más de un millón de registros y más de 30 variables relevantes para el análisis.

La importancia de este proyecto radica en la relevancia del sector inmobiliario para la economía mexicana y en la necesidad urgente de actualizar la metodología del índice de precios de la vivienda. El objetivo principal es desarrollar un modelo predictivo que mejore la precisión de las estimaciones de precios en comparación con el tradicional modelo lineal hedónico. Este modelo servirá como base para una futura investigación orientada a la creación de un nuevo índice de precios de la vivienda, aprovechando las interacciones entre las variables y los precios identificadas mediante valores de Shapley. La aplicación de Ciencia de Datos es esencial para este proyecto, dado que los datos de avalúos presentan una alta dimensionalidad y relaciones no lineales entre las variables y el precio de la vivienda. Además, la heterogeneidad de las propiedades y la asimetría de la información en el mercado hacen de la predicción de precios un reto complejo.

Históricamente, el modelo hedónico, desarrollado por S. Rosen en 1974, ha sido el enfoque predominante para estimar los precios de las viviendas. Este modelo se basa en las características físicas y de entorno de las propiedades para predecir su valor. Aunque su simplicidad y capacidad de interpretación lo hacen atractivo, presenta desventajas significativas, como su alta sensibilidad a la especificación funcional de las variables y a la colinealidad entre ellas [Rosen, 1974]. En Estados Unidos, el modelo Case-Shiller ha ganado popularidad debido a su robustez al utilizar el precio histórico de las propiedades como principal insumo para predecir su valor futuro [Case and Shiller, 1987]. No obstante, su implementación en México es inviable debido a la falta de un registro continuo de precios de las propiedades a lo largo del tiempo. En la década de los 90, los modelos de ecuaciones estructurales también ganaron terreno al permitir modelar relaciones complejas entre variables observadas y latentes, aunque su complejidad dificulta su interpretación [Bollen, 1989]. En la actualidad, los modelos de aprendizaje automático han demostrado ser los más eficientes para la predicción de precios de viviendas, al poder capturar relaciones no lineales entre variables y manejar grandes volúmenes de datos [Rey-Blanco et al., 2024]. Este último enfoque será adoptado en el presente proyecto para predecir el precio de la vivienda en México.

1.2. Problemática

El desafío central de este proyecto es predecir el precio de una vivienda típica en México, denotado como ϖ, dado un conjunto de variables X que describen sus características, incluyendo su ubicación y el momento en que se realizó el avalúo. Por "vivienda típica"se

entienden las clases de propiedades como departamentos y casas habitación, excluyendo propiedades comerciales, oficinas y terrenos. Además, se descartan aquellas propiedades cuya superficie vendible sea menor a 30 m^2, cuyo valor de construcción, terreno o mercado sea igual a cero, y aquellas cuyo valor total concluido no esté alineado con el valor de mercado (se detallará el criterio de similaridad en la sección de Metodología). Se excluyen también las propiedades clasificadas como "únicas"según la tipología de la SHF.

Según la literatura [Ho et al., 2021], es más eficiente modelar el precio por metro cuadrado vendible, $\pi := \varpi/m^2$, en lugar del precio total, ya que los errores en la estimación de π tienden a ser multiplicativos. Por tanto, el problema a resolver es estimar la función f que describe la relación:

$$\pi = f(X) \times \varepsilon, \tag{1}$$

donde ε representa el error, que podría estar influido por sesgos cognitivos del valuador. Los cuales aunque presentan un campo de interés para futuras investigaciones, no serán abordados en este proyecto.

La mejor aproximación $\hat{f}$ será aquella que minimice el error porcentual medio absoluto (MAPE, por sus siglas en inglés) en la predicción del precio por metro cuadrado. El MAPE se define como:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\pi_i - \hat{\pi}_i}{\pi_i}\right|, \tag{2}$$

donde n es el número de observaciones, y π_i y $\hat{\pi}_i$ representan el precio por metro cuadrado real y el estimado, respectivamente. Cabe destacar que el MAPE es aplicable tanto para π como para ϖ, ya que ambos son proporcionales.

El MAPE es una métrica adecuada para este problema debido a que el precio por metro cuadrado tiende a seguir una distribución sesgada a la derecha, lo que hace más útil estimar medidas cercanas a la moda de la distribución que a la media. Así, el MAPE es una métrica más robusta y fácil de interpretar, en comparación con el error cuadrático medio (MSE) o el error absoluto medio (MAE), especialmente para su explicación ante un público no especializado.

1.3. Objetivo

El objetivo de este proyecto es superar un *benchmark* basado en el promedio armónico de π por entidad federativa y año, estimado a partir del conjunto de entrenamiento. Dado que la ubicación es una de las variables más determinantes del precio de la vivienda, se pretende evaluar si los modelos hedónicos y los modelos predictivos propuestos pueden capturar más información que este promedio armónico [Heyman and Sommervoll, 2019]. Para validar la efectividad del modelo propuesto, se comparará el MAPE en el conjunto de prueba entre ambos enfoques.

La metodología empleada para alcanzar este objetivo sigue el proceso CRISP-DM (Cross-Industry Standard Process for Data Mining), que consta de seis etapas: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. En este proyecto, la comprensión del negocio se abordará mediante una revisión exhaustiva de la literatura y entrevistas con expertos del mercado inmobiliario. La comprensión de los datos se realizará mediante un análisis exploratorio, con el fin de identificar valores atípicos, redundancias, valores faltantes y áreas de oportunidad. La preparación de los datos implicará la limpieza de estos, tratando valores atípicos y datos faltantes, así como la ingeniería de variables para enriquecer las características en X. El modelado consistirá

en seleccionar el algoritmo de aprendizaje automático más adecuado para predecir π, empleando técnicas que puedan manejar grandes volúmenes de datos y ser entrenadas en una computadora personal. Finalmente, se interpretarán los resultados mediante la identificación de las variables más relevantes utilizando valores de Shapley.

La implementación de este proyecto y el código correspondiente se encuentran disponibles en el repositorio de GitHub: https://github.com/romanAVJ/mds-research-stay.

2. Metodología

La construcción del modelo predictivo para la estimación del precio de las viviendas se estructuró en varias fases clave: reconocimiento de la información disponible, extracción y limpieza de los datos, ingeniería de variables y desarrollo de modelos de aprendizaje automático. A continuación, se detallan estas etapas.

2.1. Extracción y Limpieza

La limpieza de datos es fundamental cuando se trabaja con millones de registros, ya que garantiza la calidad y precisión del análisis. Datos sin depurar, que contienen valores erróneos, duplicados o incompletos, pueden distorsionar los resultados, afectando la validez de los modelos predictivos y las conclusiones. En grandes volúmenes, estos problemas son difíciles de detectar manualmente, lo que hace imprescindible automatizar el proceso para evitar sesgos y errores sistemáticos.

Además, la limpieza de datos optimiza el rendimiento computacional al reducir el tamaño del conjunto de datos y mejorar la eficiencia en el procesamiento. Esto es crucial en análisis a gran escala, donde el tiempo de ejecución y la precisión del modelo dependen directamente de la calidad de los datos utilizados.

2.1.1. Eliminación de valores atípicos y erróneos

En esta subfase, se enfocó en identificar y eliminar registros con valores que no seguían patrones lógicos o esperados en las propiedades. Primero, se eliminaron aquellas propiedades cuyas características físicas presentaban inconsistencias, como casos en los que la suma de la superficie construida y accesoria era menor que la superficie vendible. Esto resultaba especialmente problemático en el análisis de bienes raíces, dado que las superficies accesorias, aunque no destinadas exclusivamente a la vivienda, contribuyen al valor de una propiedad (ej. tejados, bodegas, cuartos de servicio). Eliminar este tipo de registros mejoró la precisión en la estimación de precios.

Posteriormente, se eliminaron propiedades cuyas características financieras también presentaban desviaciones importantes. Específicamente, se enfocó en los logaritmos del valor concluido y del valor de mercado, eliminando aquellos fuera del intervalo de confianza del 99 %. Este método se basó en una regresión log-log que estratificaba los valores por tipo de inmueble, asegurando que solo se consideraran propiedades cuyo valor concluido no se desviara más de un 3.6 % respecto al valor de mercado. Esta depuración eliminó una cantidad considerable de registros que, de otra manera, podrían haber sesgado los resultados del modelo predictivo.

2.1.2. Validación geoespacial

La validación geoespacial fue una etapa crítica para asegurar la coherencia de las ubicaciones de las propiedades dentro del análisis ya que la ubicación es un factor relevante en la determinación del precio de una propiedad [Heyman and Sommervoll, 2019]. Se eliminaron registros cuyas coordenadas geográficas eran inconsistentes con la clave geográfica reportada, es decir, aquellos que se encontraban fuera de los límites del país o no coincidían con la zona geográfica correspondiente. Esto permitió reducir la posibilidad de errores en el análisis que pudieran surgir de propiedades mal ubicadas geográficamente.

Este análisis permitió mejorar la integridad de la base de datos en un 1.2 %, lo que puede parecer marginal, pero es crucial cuando se trabaja con millones de registros. Esta limpieza no solo asegura que el análisis geoespacial sea confiable, sino que también previene posibles anomalías en el modelo predictivo que puedan surgir por propiedades ubicadas en áreas irrelevantes o incorrectas.

2.1.3. Detección de propiedades duplicadas

Se consideró una propiedad P_i duplicada si compartía el mismo identificador con otra propiedad P_j y la distancia euclidiana promedio entre sus características era menor a un umbral τ. Las características incluidas en el cálculo fueron: coordenadas geográficas $\underline{x}$, fecha de avalúo t, superficie vendible m y precio de venta ϖ. Formalmente,

$$d(P_i) = \frac{1}{\#(P_i)} \sum_{j,k,j\neq k} ||\underline{x}_j - \underline{x}_k||_2 + ||t_j - t_k||_2 + ||m_j - m_k||_2 + ||\varpi_j - \varpi_k||_2. \quad (3)$$

Si la distancia promedio $d(P_i) < \tau$, se eliminan $\#(P_i) - 1$ propiedades duplicadas. En los casos en que las distancias fueran cero pero existía una diferencia en el tiempo de avalúo, se conservó la última propiedad en el tiempo. De esta manera, se identificaron 56,269 propiedades duplicadas (19.9 % del total), aunque solo el 1 % de los registros fue eliminado, ya que la mayoría presentaba diferencias significativas en sus características.

2.1.4. Imputación de valores faltantes

Con el fin de centrar el análisis en áreas de interés, se excluyeron propiedades ubicadas en zonas con baja densidad habitacional. Para ello, se empleó el algoritmo DBSCAN [Schubert et al., 2017], que agrupa puntos geográficos en clústeres basados en la densidad, identificando como ruido los puntos que no cumplen con los requisitos mínimos de vecindad. El espacio D estuvo constituido por las coordenadas geográficas de las propiedades, y los parámetros de radio ϵ y el número mínimo de puntos ν se determinaron mediante un análisis de sensibilidad, buscando minimizar el porcentaje de propiedades eliminadas sin perder la coherencia demográfica de las zonas estudiadas. Los datos sugirieron que $\epsilon = 1000$ m^2 y $\nu = 100$ propiedades eran los valores más adecuados. Esta etapa resultó en la eliminación del 8.7 % de los registros. Posteriormente, se incorporó la geometría espacial basada en el sistema H3 de Uber para realizar un análisis más granular en la geografía. El sistema H3 es una estructura de teselado geoespacial jerárquico que subdivide la superficie de la Tierra en hexágonos de área aproximadamente uniforme. En este caso, se utilizó una resolución 7 de H3, donde cada hexágono cubre un área promedio de 5.16 km^2. Dentro de cada hexágono, se eliminaron las propiedades cuyos log-precios estandarizados se encontraban a más de 5 desviaciones estándar de la media de esa zona, lo que permitió eliminar un 0.1 % adicional de los registros. Esta metodología ayudó a asegurar que las propiedades analizadas estuvieran concentradas en áreas geográficamente coherentes y cuyas propiedades tuvieran precios comparables localmente.

2.2. Ingeniería de Variables

Además de las variables proporcionadas por la base de datos de avalúos, fue necesario incorporar información macroeconómica del país para tipificar las distintas zonas del país año tras año, así como generar variables que describan la relación entre las propiedades y su entorno, como los comparables de una propiedad y el valor del suelo en la zona.

2.2.1. Tipificación de Zonas

Para tipificar las zonas del país, se recopiló la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) del INEGI, la cual contiene información sobre los ingresos y gastos de los hogares mexicanos. Esta encuesta se realiza cada dos años y tiene una granularidad a nivel municipal, sin embargo, hay municipios que no reportan información en ciertos periodos, por lo cual se interpoló la información con datos previos si estaban disponibles y, de lo contrario, con los datos de municipios vecinos. Asimismo, se extrajo del Sistema de Información Económica del Banco de México (SIE) la inflación y el salario mínimo real de cada año. También se obtuvo información de la Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO) sobre las regiones marinas, para identificar propiedades cercanas a la costa, ya que la conservación y el precio de una propiedad cerca de la costa difieren de las propiedades ubicadas en el interior del país.

2.2.2. Comparables

Para identificar los comparables de una propiedad, se generó un radio de búsqueda de 2.5 km alrededor de cada propiedad utilizando el algoritmo KDTree [Maneewongvatana and Mount, 1999]. La selección del radio se basó en la distancia promedio de caminata de 30 minutos, que es un indicador comúnmente utilizado en bienes raíces para definir la proximidad de una propiedad a servicios y áreas de interés [Murtagh et al., 2020]. Este algoritmo tiene una complejidad promedio de $O(\log n)$, lo que permite realizar el análisis para más de un millón de registros. Todas las propiedades que se encontraban dentro del radio de búsqueda se consideraron vecinos, sin embargo el nivel de comparabilidad se determinó mediante un análisis de similitud entre las propiedades. Por lo que un comparable es un vecino pero no todos los vecinos son comparables.

- **Similitud geográfica:** La similitud geográfica entre dos propiedades se refiere a la proximidad espacial entre ambas, medida en términos de sus coordenadas geográficas. Esto se puede cuantificar utilizando la distancia euclidiana entre las coordenadas normalizadas de longitud y latitud, lo cual proporciona una medida directa de cuán cercanas están físicamente dos propiedades. Esta similitud es útil para predecir precios porque, en bienes raíces, la ubicación es uno de los factores más determinantes en la valoración de una propiedad. Zonas cercanas tienden a compartir características sociodemográficas, infraestructura, y acceso a servicios, lo que impacta directamente en el valor. En este caso, la similitud geográfica cumple la definición matemática de distancia al utilizar una métrica basada en la norma L2.

  $$s_1(P_i, P_j) := ||\underline{x}_i - \underline{x}_j||_2,$$

  donde $\underline{x}_i$ y $\underline{x}_j$ son los vectores de coordenadas de las propiedades P_i y P_j, respectivamente.

- **Similitud topológica:** La topología de una propiedad refiere a la configuración del terreno sobre el cual está construida. La similitud topológica mide las diferencias en el área total de terreno entre dos propiedades como un porcentaje relativo, siendo útil para comparar propiedades en términos de tamaño de lote. Esta similitud es valiosa en el contexto de bienes raíces porque propiedades con terrenos más grandes suelen tener un mayor valor, debido a la posibilidad de expansión o uso adicional del suelo. Cuando

la diferencia porcentual de la superficie es mínima, las propiedades son topológicamente similares.

$$s_2(P_i, P_j) := \left| 1 - \frac{\zeta_j}{\zeta_i} \right|,$$

donde ζ_i y ζ_j representan las superficies de terreno de las propiedades P_i y P_j.

- **Similitud de superficie construida:** Esta similitud se refiere a la comparación del área construida entre dos propiedades. Se calcula como la diferencia porcentual relativa entre las superficies construidas de ambas propiedades, esta comparación es vital ya que el tamaño del área habitable influye directamente en el precio. Propiedades con superficies construidas similares suelen ofrecer el mismo nivel de utilidad y funcionalidad para el propietario, lo que se refleja en el valor de mercado.

$$s_3(P_i, P_j) := \left| 1 - \frac{c_j}{c_i} \right|,$$

donde c_i y c_j representan la superficie construida de las propiedades P_i y P_j.

- **Similitud de características:** Esta métrica toma en cuenta varias características categóricas que definen la funcionalidad de una propiedad, tales como la presencia de elevador, el número de recámaras, baños, y estacionamientos, entre otros. Al tratarse de variables que afectan directamente la comodidad y utilidad del inmueble, estas características son determinantes en la evaluación comparativa entre dos propiedades. Se utiliza la norma L2 de las diferencias. En este caso, esta medida cumple la definición de distancia.

$$s_4(P_i, P_j) := \sqrt{(e_i - e_j)^2 + (l_i - l_j)^2 + (\ell_i - \ell_j)^2 + (r_i - r_j)^2 + (b_i - b_j)^2 + (p_i - p_j)^2},$$

donde e indica la presencia de elevador, l es el nivel de la propiedad, ℓ es la vida útil de la propiedad, r es el número de recámaras, b es el número de baños, y p es el número de cajones de estacionamiento.

- **Similitud temporal:** Esta similitud mide la diferencia en el tiempo entre las fechas de avalúo o construcción de dos propiedades a lo largo de 730 días, lo que equivalente a dos años. Se calcula como el cuadrado de la diferencia en días normalizado en un rango de 0 a 1. En el contexto inmobiliario, las propiedades construidas o valuadas en fechas cercanas suelen haber estado sujetas a condiciones de mercado similares, lo cual es un factor relevante para la comparación de precios. Esta medida de similitud puede ser tratada como una distancia en el tiempo.

$$s_5(P_i, P_j) := \left(\frac{t_i - t_j}{730} \right)^2,$$

donde t_i y t_j son las fechas de avalúo de las propiedades P_i y P_j, respectivamente.

Es importante señalar que s_k se define como una función de similitud y no como una distancia, ya que no cumple con la definición matemática de distancia. Por ejemplo, la similitud de la superficie construida se define como la diferencia porcentual relativa. Observamos que:

$$s(c_i, c_j) = \frac{c_i - c_j}{c_i} = \frac{c_i - c_j}{c_j} = s(c_j, c_i) \quad \iff c_i = c_j,$$

lo que no siempre se cumple.

Para medir que tan comparable es una propiedad con otra, se define una similitud general entre dos propiedades como el promedio ponderado de las similitudes anteriores. Las ponderaciones asociadas a cada tipo de similitud, denotadas como λ_k, están diseñadas para cumplir con la condición $\sum_{k=1}^{5} \lambda_k = 1$, lo que garantiza que el total de las ponderaciones sume uno. Dichas ponderaciones fueron determinadas mediante el juicio de expertos en *real estate*, quienes, a partir de su conocimiento y experiencia, asignaron la relevancia relativa de cada tipo de similitud en la predicción de precios de propiedades. Entonces la medida de similitud entre las propiedades P_i y P_j esta dada por

$$s_{ij} = \sum_{k=1}^{5} \lambda_k s_k(P_i, P_j). \tag{4}$$

Una vez calculada la similitud entre propiedades, se estima la media y la varianza del logaritmo del precio por metro cuadrado de los comparables de una propiedad utilizando el ponderador w_j del comparable P_j de la propiedad P_i. Este ponderador se define como:

$$w_j := \frac{e^{-\xi s_{ij}}}{\sum_{k=1}^{n_i} e^{-\xi s_{ik}}}, \tag{5}$$

donde ξ es un hiperparámetro de decaimiento. Suponiendo que $\log \pi \sim \mathcal{N}(\mu, \sigma^2)$ y que P_i tiene n_c comparables, el estimador de máxima verosimilitud para μ y σ^2 está dado por [Casella and Berger, 2024]:

$$\hat{\mu} = \sum_{j=1}^{n_c} w_j \log \pi_j \quad \text{y} \quad \hat{\sigma}^2 = \frac{1}{1 - \sum_k w_k^2} \sum_{j=1}^{n_c} w_j (\log \pi_j - \hat{\mu})^2, \tag{6}$$

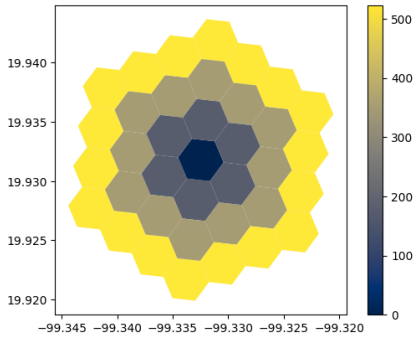
donde π_j es el precio por metro cuadrado de la propiedad P_j. De esta manera, se obtiene la media y la varianza de los precios por metro cuadrado de los comparables.

2.2.3. Valor del terreno

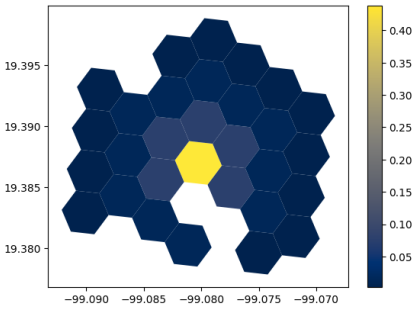
El valor del terreno se estimó utilizando la topología H3 de Uber, junto con métodos de suavizamiento y expansión geoespacial. La topología H3 emplea una rejilla hexagonal que cubre la superficie terrestre y que puede subdividirse en hexágonos de tamaños más pequeños. La metodología consistió en tres etapas: obtener estadísticas del logaritmo del precio por metro cuadrado del terreno ($\log \pi_\zeta$) para las propiedades dentro del hexágono h_i, suavizar estas estadísticas usando anillos de hexágonos y, finalmente, expandir las estadísticas a los hexágonos vecinos.

En la primera etapa, se estima la media y la varianza del logaritmo del precio por metro cuadrado del terreno para las propiedades dentro del hexágono h_i, utilizando los estimadores clásicos de máxima verosimilitud $\tilde{\mu}_i$ y $\tilde{\sigma}_i^2$.

En la segunda etapa, se suavizan las estadísticas del logaritmo del precio por metro cuadrado del terreno para los hexágonos vecinos. Primero, se define la distancia entre los hexágonos h_i y h_j como la distancia euclidiana entre sus centros. Dado que el hexágono es una figura regular, la distancia entre dos hexágonos se basa en la distancia entre sus centros. Todos los hexágonos h_k en el mismo anillo alrededor de h_i tienen la misma distancia, es decir, $d_{ij} = d_{ik}$ para todo $j, k \in \text{Ring}_r(h_i)$. Geométricamente, el hexágono genera una teselación perfecta [Puu and Weidlich, 2007]. Esto se ilustra en la Figura 1a.



(a) Distancia en metros entre hexágonos por niveles de anillos.



(b) Pesos w_j de hexágonos por niveles de anillos. Notar que $\sum_j w_j = 1$ independientemente de que existan hexágonos h_j con $n_j = 0$.

Figura 1: Comparación de distancias y pesos entre hexágonos.

El peso del hexágono h_j en el suavizamiento de las estadísticas del hexágono h_i se define como:

$$\omega_j := \frac{e^{-\nu d_{ij}}}{\sum_{k \in \text{Rings}_r(h_i)} e^{-\nu d_{ik}}}, \tag{7}$$

donde ν es un hiperparámetro de decaimiento, n_j es el número de propiedades dentro del hexágono h_j, y $\text{Rings}_r(h_i)$ es el conjunto de hexágonos vecinos a h_i en r anillos, tales que $n_j > 0$. Un ejemplo de esto se muestra en la Figura 1b. De esta manera, el estimador de máxima verosimilitud para la media y la varianza del logaritmo del precio por metro cuadrado del terreno para los hexágonos vecinos se da por [Snedecor and Cochran, 1989]:

$$\begin{aligned}
\hat{\mu}_i &= \frac{\sum_{j \in \{j:\sigma_j^2 > 0\}} w_j \frac{n_j}{\sigma_j^2} \tilde{\mu}_j + \sum_{j \in \{j:\sigma_j^2 = 0\}} w_j \tilde{\mu}_j}{\sum_{j \in \{j:\sigma_j^2 > 0\}} w_j \frac{n_j}{\sigma_j^2} + \sum_{j \in \{j:\sigma_j^2 = 0\}} w_j}, \\
\hat{\sigma}_i^2 &= \frac{\sum_{j \in \{j:\sigma_j^2 > 0\}} w_j \left[(n_j - 1)\sigma_j^2 + n_j(\tilde{\mu}_j - \hat{\mu}_i)^2\right] + \sum_{j \in \{j:\sigma_j^2 = 0\}} w_j n_j (\tilde{\mu}_j - \hat{\mu}_i)^2}{\sum_j w_j n_j}.
\end{aligned} \tag{8}$$

Finalmente, para los hexágonos h_i en los que no se pudo estimar las estadísticas debido a la falta de información durante el suavizamiento, se expande la estimación usando $\rho > r$ anillos de hexágonos y ponderando las estadísticas obtenidas con un factor de penalización γ definido como:

$$\gamma_j := \frac{e^{-\nu d_{ij}}}{\sum_{k \in h_i} e^{-\nu d_{ik}}}. \tag{9}$$

La diferencia entre ω_j y γ_j radica en que ω_j considera solo los hexágonos vecinos a h_i que contienen propiedades, mientras que γ_j considera todos los hexágonos vecinos, independientemente de si contienen propiedades o no, por lo que $\sum_j \gamma_j < 1$.

Así, el estimador heurístico para la media y la varianza $(\check{\mu}_i, \check{\sigma}_i^2)$ del logaritmo del precio por metro cuadrado del terreno en los hexágonos sin estadísticas estimadas se define como:

$$\begin{aligned}
\check{\mu}_i &= \gamma_i \hat{\mu}_i + \phi(1 - \gamma_i)\text{mín}(\hat{\mu}), \\
\check{\sigma}_i^2 &= \gamma_i \hat{\sigma}_i^2 + \phi(1 - \gamma_i)\text{máx}(\hat{\sigma}^2),
\end{aligned} \tag{10}$$

donde mín($\hat{\mu}$) y máx($\hat{\sigma}^2$) son el mínimo de las medias y el máximo de las varianzas de los hexágonos en una ciudad, respectivamente, y $\phi \in (0, 1]$ es un factor de penalización.

2.3. Modelos de Aprendizaje de Máquina

Para la calibración de los modelos de aprendizaje de máquina se emplearon las librerías de Python `scikit-learn` y `catboost` [Buitinck et al., 2013] [Dorogush et al., 2018]. La elección de estas librerías se debe a que ambas son ampliamente reconocidas por su capacidad para manejar conjuntos de datos de gran tamaño y por ofrecer modelos de estado del arte en diversas tareas de clasificación y regresión. `scikit-learn` es una biblioteca versátil que proporciona una amplia gama de algoritmos de aprendizaje de máquina y herramientas para preprocesamiento, selección de características y validación cruzada, lo que la convierte en una opción ideal para la construcción de pipelines de modelado.

Por otro lado, `catboost` es un modelo de boosting de gradiente que ha demostrado un rendimiento excepcional en términos de precisión y tiempo de estimación, sobre todo en problemas con datos categóricos y de gran volumen. Los modelos de boosting, como `catboost`, son especialmente útiles en este tipo de tareas, ya que construyen múltiples árboles de decisión de manera secuencial, ajustándose a los errores de los árboles anteriores, lo que permite mejorar el rendimiento predictivo. Además, `catboost` ofrece una serie de optimizaciones, como el manejo eficiente de variables categóricas sin necesidad de un preprocesamiento intensivo y la reducción de overfitting gracias a su implementación específica de boosting.

Si bien no son los únicos métodos disponibles para predecir precios en el sector *real estate*, `scikit-learn` y `catboost` destacan por su bajo costo computacional en comparación con otros enfoques más complejos y demandantes de recursos. Además, ofrecen una gran robustez y flexibilidad en diversos tipos de datos, lo que los hace ideales para problemas donde la eficiencia y la precisión son claves. Lo cual, permite obtener modelos de alta calidad sin incurrir en tiempos de cómputo excesivos, algo especialmente relevante al trabajar con bases de datos que contienen millones de observaciones, como el utilizado en este estudio.

2.3.1. Selección de Variables

Para la selección de las variables más importantes se empleó un modelo de regresión lineal con regularización LASSO [Tibshirani, 1996]. Matemáticamente, se define como:

$$\underline{\hat{\beta}} = \text{argmin}_{\underline{\beta}} \left\{ \frac{1}{2n} \|\underline{y} - X\underline{\beta}\|_2^2 + \lambda \|\underline{\beta}\|_1 \right\}, \tag{11}$$

donde $\underline{y}$ es el vector de precios de las propiedades, X es la matriz de características de las propiedades, y λ es el hiperparámetro de regularización. La regularización LASSO penaliza las variables que no aportan información al modelo, resultando en la eliminación de variables con coeficientes $\beta_i = 0$.

Además, se utilizó un modelo de `catboost` para identificar las variables más importantes mediante los valores de Shapley [Lundberg and Lee, 2017]. Los valores de Shapley en el contexto de modelos de aprendizaje de máquina se definen como:

$$\phi_i = \frac{1}{n_X!} \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n_X - |S| - 1)!}{n_X!} \left[\underline{f}(S \cup \{i\}) - \underline{f}(S) \right], \tag{12}$$

donde n_X es el total de características en el modelo, i.e. $n_X = \#(N)$, N es el conjunto de todas las características, y S representa un subconjunto de N que no incluye la característica

i. La función $\underline{f}(S)$ denota la predicción del modelo usando únicamente las características en el subconjunto S, mientras que $\underline{f}(S \cup \{i\})$ es la predicción del modelo usando las características en el subconjunto S más la característica i. El valor ϕ_i representa el valor de Shapley de la característica i.

La interpretación de los valores de Shapley es que, si $\phi_i > 0$, la característica i tiene un impacto positivo en la predicción del modelo. Por el contrario, si $\phi_i < 0$, la característica i tiene un impacto negativo. Este valor refleja la contribución promedio de una característica al valor final de la predicción, considerando todas las posibles combinaciones en las que esa característica podría aparecer junto con las otras.

2.3.2. Modelos de Regresión

La tarea de regresión consiste en predecir un valor continuo, como el precio de una propiedad, a partir de un conjunto de características o variables independientes. En términos matemáticos, se busca aproximar una función desconocida f, que relaciona las variables independientes X con la variable dependiente y. Esta aproximación se denota como $\hat{f}$, donde el objetivo es que $\hat{f}(X) \approx y$, minimizando el error entre los valores predichos $\hat{y}$ y los valores observados y. En el caso específico de este estudio, y representa el precio por metro cuadrado de una propiedad, y las variables X incluyen características como la ubicación, superficie construida, número de habitaciones, entre otros atributos relevantes del inmueble. A lo largo del proceso, se probaron varios modelos de regresión, cada uno con el fin de ajustar lo mejor posible la relación entre X y y, evaluando su rendimiento en términos de precisión predictiva y otros criterios de desempeño.

1. **Modelo de Promedios**

   Este modelo simple se define como:

   $$\hat{f}(\varphi, t) := \frac{n}{\sum_{i \in (\varphi, t)} 1/y_i}, \tag{13}$$

   donde n es el número de propiedades en la entidad federativa φ y en el año t. Este modelo sirve como referencia (benchmark) para comparar el rendimiento de otros modelos.

2. **Modelo de Regresión Lineal**

   Este modelo se define como:

   $$\hat{f}(\underline{x}) = \underline{\beta}^T \underline{x}, \tag{14}$$

   donde $\underline{\beta}$ es el vector de coeficientes de la regresión y $\underline{x}$ es el vector de características de la propiedad [Montgomery and Runger, 2010]. Este modelo supone que la relación entre las características y el precio de la propiedad es lineal. Debido a que π presenta efectos multiplicativos, se utilizó el logaritmo del precio por metro cuadrado de la propiedad, es decir, $\log \pi$.

3. **Modelo de Regresión con Árbol de Decisión**

   Este modelo se define como:

   $$\hat{f}(\underline{x}) = \sum_{i=1}^{n_{\text{leaf}}} \beta_i \mathbb{I}(\underline{x} \in \text{leaf}_i), \tag{15}$$

donde β_i es el coeficiente de la hoja i y $\mathbb{I}$ es la función indicadora. Este modelo permite una relación no lineal entre las características y el precio de la propiedad, ajustándose a los datos de entrenamiento de manera más flexible que el modelo de regresión lineal. Es especialmente útil para incorporar características geoespaciales como la longitud y latitud de la propiedad, lo cual no es posible con un modelo de regresión lineal [Breiman, 2017].

4. **Modelo** `CatBoost`

   La definición formal del modelo `catboost` se encuentra en [Dorogush et al., 2018]. La principal característica de `catboost` es ajustar un árbol de decisión a los residuos del árbol anterior y continuar con este proceso iterativamente. Este enfoque permite una adaptación más flexible a los datos de entrenamiento en comparación con los modelos de regresión lineal y de árbol de decisión. Además, `catboost` maneja variables categóricas sin necesidad de codificación adicional, lo cual es ventajoso dado el gran número de variables categóricas como el tipo de propiedad, la clase de la propiedad y el uso de la propiedad.

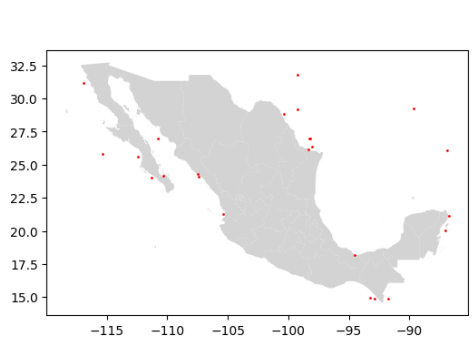
   Se decidió no utilizar otros modelos de árboles como `XGBoost` o `LightGBM` debido a que `catboost` gestiona variables categóricas de manera más eficiente y pertenecen a la misma familia de modelos. Además, los modelos de boosting han demostrado ser superiores en la mayoría de los casos [Chen and Guestrin, 2016] y presentan limitaciones al escalar a millones de registros, lo que limita la aplicabilidad de modelos como bosques aleatorios, máquinas de soporte vectorial o bayesianos.

3. Resultados

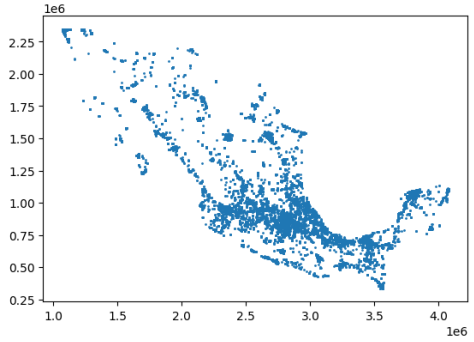
En esta sección se presentan los resultados obtenidos tras la exploración de datos, la ingeniería de variables, la selección de características y la evaluación de modelos de aprendizaje automático.

3.1. Exploración de Datos

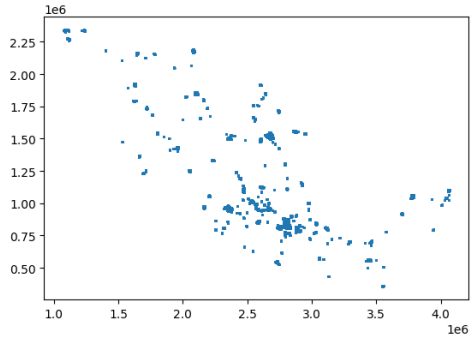
La base de datos proporcionada por la Dra. María de las Mercedes Adamuz Peña inicialmente contenía 1,048,575 propiedades y 38 variables. Tras la limpieza de datos e ingeniería de variables, se conservó el 81.34 % de los datos, resultando en un total de 852,913 propiedades y 90 variables. La mayor parte de la eliminación de datos se debió a la depuración de datos geoespaciales, que implicó quitar propiedades con coordenadas geográficas incorrectas y propiedades ubicadas en áreas con baja densidad de población, como se describe en la sección de limpieza de datos (ver figura 2).



(a) Propiedades fuera de México.



(b) Propiedades en zonas de baja densidad.

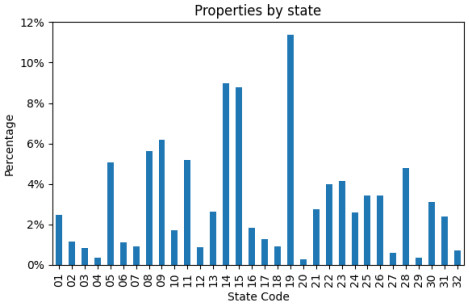


(c) Conjuntos convexos de propiedades en zonas de alta densidad.

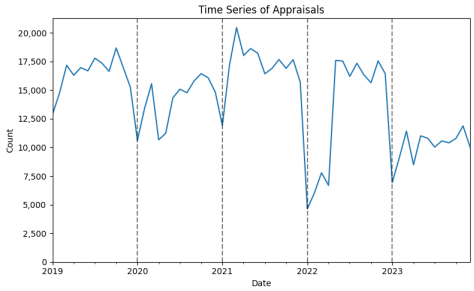
Figura 2: Limpieza de datos geoespaciales.

El 76 % de las propiedades son casas y el restante 24 % son departamentos. La categoría de inmuebles más frecuente es la de *interés social*, que representa el 56 % del total, mientras que las categorías *Residencial* y *Residencial Plus* son las menos comunes, representando solo el 1.1 %. El estado con el mayor número de propiedades es Nuevo León (estado 19), con un

11.5 % del total, mientras que los estados con menor número de propiedades son Campeche (estado 4), Oaxaca (estado 20) y Tlaxcala (estado 29), con menos del 1 % del total (ver figura 3a). Además, la serie temporal del número de avalúos muestra un quiebre en 2023, lo cual podría indicar problemas con la calidad de los datos. Se observa que al inicio de cada año hay una disminución sustancial en el número de avalúos (ver figura 3b).



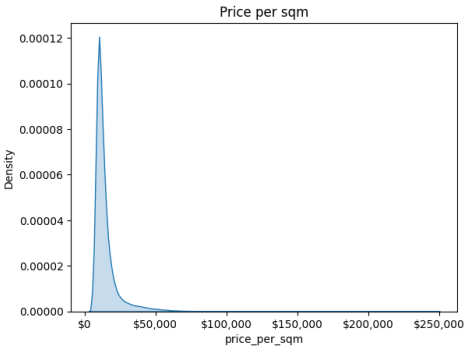
(a) Número de avalúos por entidad federativa.



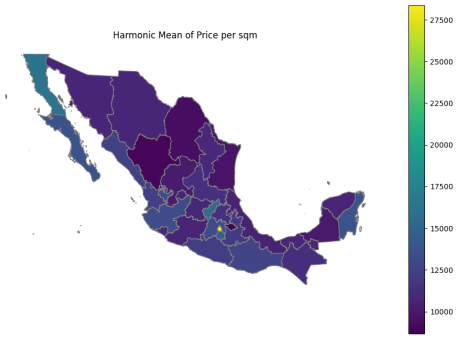
(b) Serie temporal de avalúos.

Figura 3: Gráficas de avalúos por entidad federativa y serie temporal.

El precio por metro cuadrado (π) varía de \$4,700 a \$248,687 pesos, con una mediana de \$11,895, una media de \$14,505 y una desviación estándar de \$9,100 (ver figura 4a). Por otro lado, el promedio armónico de π por entidad federativa varía de \$9,000 a \$27,400, siendo la Ciudad de México la entidad federativa con el precio más alto por metro cuadrado (ver figura 4b).



(a) Distribución de precios por metro cuadrado.



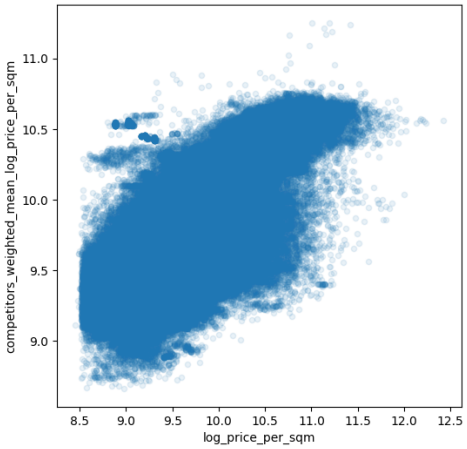
(b) Distribución del promedio armónico de precios por metro cuadrado por entidad federativa.

Figura 4: Gráficas de precios por metro cuadrado.

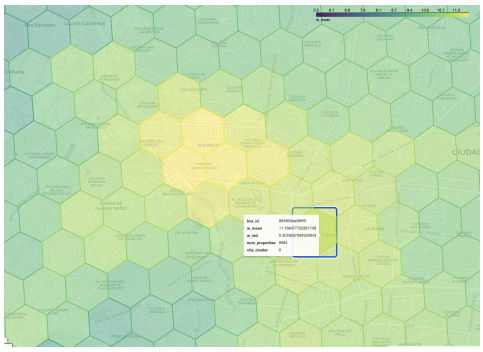
La correlación entre el promedio del logaritmo del precio por metro cuadrado de los comparables ($\log \pi_c$), definidos en la sección anterior como las propiedades que son similares en características a la propiedad objetivo, y el logaritmo del precio de dicha propiedad ($\log \pi$) es del 63 %. En cambio, la correlación entre el promedio del logaritmo del precio por metro cuadrado de los vecinos geoespaciales, es decir, las propiedades ubicadas a una distancia de

hasta 2.5 km, es solo del 47 %. Esto sugiere que ponderar la influencia de los comparables en función de su similitud con la propiedad objetivo proporciona más información valiosa para predecir su precio por metro cuadrado que simplemente considerar las propiedades cercanas geográficamente. En consecuencia, utilizar comparables basados en características relevantes, más allá de la proximidad geográfica, resulta ser un enfoque más robusto para la estimación del valor de una propiedad. El 14 % de los inmuebles no tenían competencia, por lo que se imputó su valor utilizando el log precio del terreno ($\log \pi_\zeta$). El proceso de búsqueda de comparables y cálculo de vecinos tomó 12 minutos, realizando más de 40,000,000 de comparaciones (ver figura 5a).

La distribución geoespacial de $\log \pi_\zeta$ muestra que los terrenos con mayor valor se encuentran en Polanco, Ciudad de México, y Puerto Cancún, Quintana Roo (ver figura 5b). Se utilizaron hexágonos de resolución 7, con un área de 0.73 km^2, para aproximar las manzanas. En total, se emplearon 31,392 hexágonos para cubrir las zonas densas de México, y el cálculo de precios en la zona se completó en menos de 10 minutos.



(a) Relación entre $\log \pi$ y $\log \pi_c$.



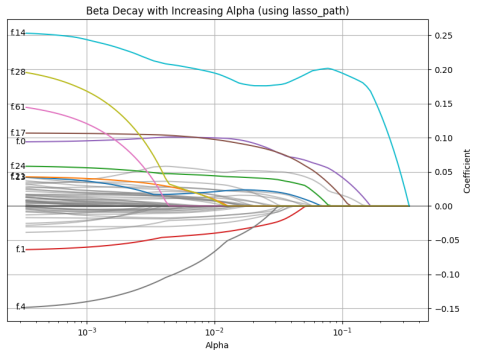
(b) Distribución geoespacial de $\log \pi_\zeta$ en la Ciudad de México.

Figura 5: Gráficas de precios por metro cuadrado de comparables y zonas.

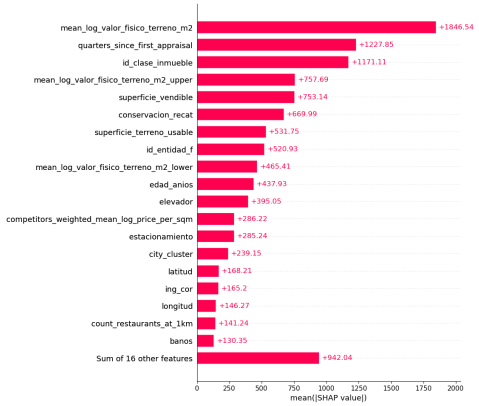
3.2. Selección de Variables

Tras el análisis exploratorio de los datos, se seleccionaron las variables más relevantes para la predicción del precio por metro cuadrado, considerando su relevancia económica y causal. Inicialmente se consideraron 35 variables de las 90 disponibles. Mediante un análisis de los caminos de las betas en un modelo de regresión lineal con distintas penalizaciones LASSO, se identificaron las variables más relevantes: el log precio del metro cuadrado del terreno ($\log \pi_\zeta$), la clase de inmueble, el número de trimestres transcurridos desde el avalúo y la edad de la propiedad (ver figura 6a).

El modelo `CatBoost` no altera significativamente el orden de importancia de las variables, pero añade los intervalos de confianza de $\log \pi_\zeta$ y la entidad federativa (ver figura 6b) como predictores importantes.



(a) Caminos de las betas en un modelo de regresión lineal con penalización LASSO.



(b) Importancia de las variables en el modelo `CatBoost` usando valores de Shapley.

Figura 6: Selección de variables. Las variables coloreadas en 6a son las 10 variables con mayor valor absoluto de β. Estas variables incluyen el log precio del metro cuadrado del terreno, la superficie usable del terreno, si la propiedad es un departamento, trimestres transcurridos, la clase de inmueble, el estado de conservación, el log precio por metro cuadrado de los comparables, la edad del inmueble en años y la superficie vendible, respectivamente.

3.3. Modelos

Se entrenaron cuatro modelos de aprendizaje automático para la predicción del precio por metro cuadrado: regresión lineal, regresión lineal con las variables que usa SHF para el modelo hedónico, árbol de decisión y `CatBoost`. Los modelos se evaluaron en un conjunto de prueba que representa el 10 % de los datos, estratificado por el tipo de inmueble (casa o departamento). El modelo `CatBoost` se entrenó con el 81 % de los datos, debido a la selección de un subconjunto de datos para la validación y evitar el sobreajuste. Los resultados de los modelos se presentan en la Tabla 1. No se observó sobreajuste en el mejor 90 % de las predicciones para ninguno de los modelos, incluyendo el `CatBoost`.

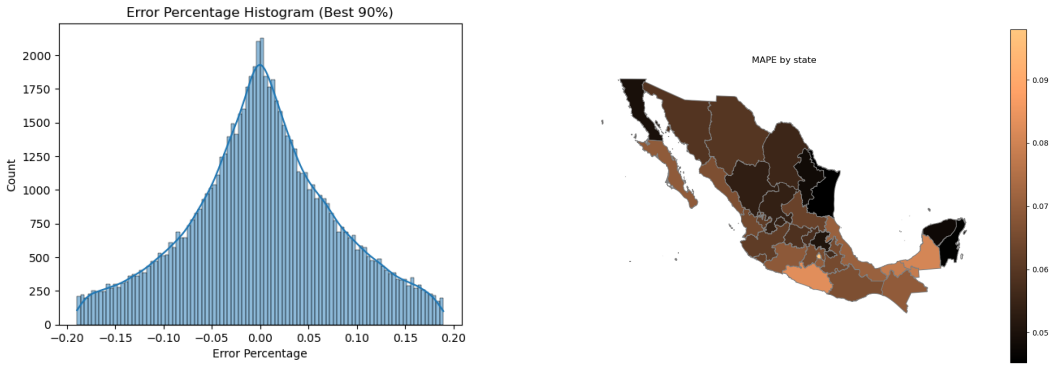
Modelo	n	n_x	MAPE (%)	MAE	RMSE	R^2
Baseline	767,621	160	17.45	2,570.38	3,870.86	0.6432
Regresión Lineal SHF	767,621	8	20.25	3,186.84	4,976.50	0.5469
Regresión Lineal	767,621	43	11.47	1,680.41	2,458.57	0.9057
Árbol de Decisión	767,621	43	8.59	1,268.39	1,934.88	0.9506
CatBoost	690,859	26	6.00	902.22	1,424.68	0.9718

Cuadro 1: Comparación de modelos en el conjunto de prueba para el mejor 90 % de las predicciones.

El modelo `CatBoost` demostró el mejor rendimiento en el conjunto de prueba, con un MAPE de 6.00 %, lo que representa un error significativamente menor en comparación con el *Baseline*, que obtuvo un MAPE de 17.45 % en el mejor 90 % de las predicciones, evidenciando un error casi 3 veces mayor.

Cabe destacar las métricas del modelo *Regresión Lineal SHF*, ya que es el modelo actualmente utilizado para la construcción del índice de precios de la vivienda en México. Este modelo se basa en ocho variables: el logaritmo de la superficie vendible, el logaritmo de la superficie construida, el logaritmo del ingreso promedio corriente per cápita del municipio donde se encuentra la propiedad, el número de baños, el número de medios baños, el total de pisos de la vivienda, el número de recámaras, la disponibilidad de estacionamiento y el tipo de inmueble (Casa, Casa en Condominio o Departamento).

3.4. Bondad de Ajuste



(a) Histograma de errores porcentuales.

(b) MAPE por entidad federativa.

Figura 7: Bondad de ajuste del modelo en el mejor 90 % de las predicciones del conjunto de prueba.

El mejor modelo obtuvo un MAPE del 6.00 % en el mejor 90 % de las predicciones y un MAPE del 8.30 % en el total del conjunto de entrenamiento. Se observaron algunos errores extremos superiores ± 50 % del valor real. Los errores para departamentos y casas son estadísticamente similares. En cuanto a las entidades federativas, los errores son mayores en la Ciudad de México, Guerrero y Campeche, mientras que Tamaulipas presenta el menor error con un 4.53 %. No obstante, los errores en las zonas costeras son los más altos, con un promedio del 11.35 % para el mejor 90 % de las predicciones (ver figura 7).

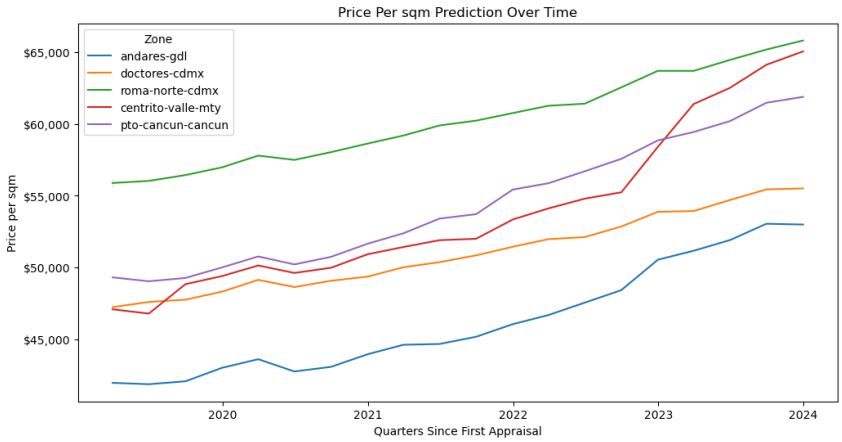
No se observó un sesgo sistemático en los errores por covariable, excepto en las propiedades ubicadas a 500 metros de la costa. Se analizó detalladamente si existe un sesgo relacionado con los trimestres transcurridos desde el avalúo, ya que esta variable refleja la plusvalía de la propiedad. Sin embargo, al ajustar un modelo de regresión lineal con respecto al error porcentual, no se rechazó la hipótesis nula de que el error es independiente de los trimestres transcurridos.

3.5. Índice de Precios de la Vivienda

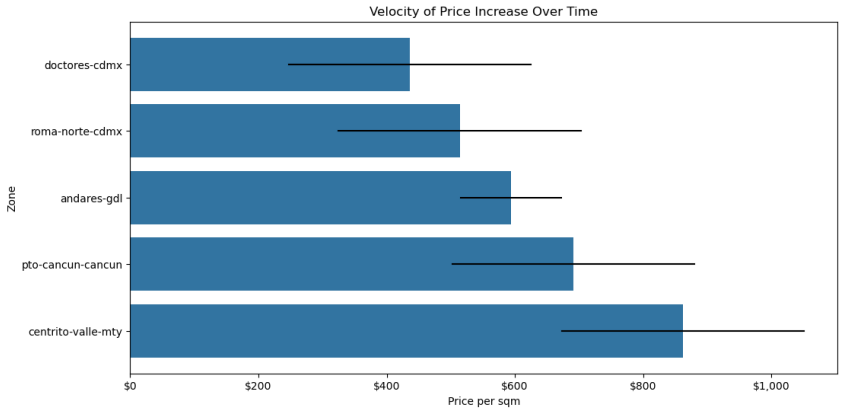
El objetivo principal de este estudio es lograr la predicción más precisa posible del precio por metro cuadrado de vivienda en México. Sin embargo, el modelo `CatBoost` también permite la construcción de un índice de precios de la vivienda mediante el uso de contrafactuales. Este índice se basa en la predicción del valor de una propiedad con características constantes, pero ubicada en diferentes zonas geográficas a lo largo del tiempo. La propiedad

seleccionada para el índice es un departamento nuevo de categoría residencial plus, con un área de 60 m^2, que incluye un baño completo y un medio baño, estacionamiento, elevador y vigilancia.

Para analizar la evolución temporal del precio por metro cuadrado, se utilizó el modelo `CatBoost` para predecir el precio de esta propiedad en diferentes ubicaciones geográficas y trimestres.



(a) Índice de Precios de la Vivienda en México.



(b) Velocidad de crecimiento de los precios por metro cuadrado.

Figura 8: Índice de Precios de la Vivienda en México y velocidad de crecimiento en zonas seleccionadas.

Las áreas seleccionadas para el análisis corresponden a los hexágonos *8848a20667fffff*, *884519b491fffff*, *884995ba3dfffff*, *884995ba27fffff*, y *8849ab4b45fffff*, que representan, respectivamente, las zonas de Centrito Valle en Monterrey, Puerto Cancún en Cancún, Roma Norte y Doctores en la Ciudad de México, y Andares en Guadalajara [1].

En la figura 8a se observa la evolución del precio por metro cuadrado en las diferentes ubicaciones geográficas. La figura 8b muestra la velocidad de crecimiento de los precios,

[1] Los hexágonos se pueden visualizar en https://h3geo.org/

es decir, la variación trimestral del precio por metro cuadrado en cada una de las zonas. Centrito Valle en Monterrey presenta la mayor velocidad de crecimiento, con un incremento de $862 pesos por metro cuadrado por trimestre. En contraste, la colonia Doctores en la Ciudad de México muestra el menor crecimiento, con $436 pesos por metro cuadrado por trimestre.

El modelo propuesto ofrece una herramienta flexible para visualizar la evolución del precio por metro cuadrado en los 31,392 hexágonos descritos en la sección de Metodología. Esto permite un análisis detallado de las dinámicas de precios no solo a nivel estatal o municipal, sino también en áreas específicas de interés.

4. Conclusiones

El trabajo presentado abarca los cinco puntos de la metodología CRISP-DM mencionada en la sección de objetivos: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado y evaluación [Wirth and Hipp, 2000]. En la comprensión del negocio, se identificó la necesidad de contar con un índice de precios de la vivienda preciso en México. En el análisis de los datos, se identificó la fuente de datos y otras fuentes que pudieran aportar información faltante a la original, se realizó una limpieza de los datos y se identificaron las variables más relevantes para la predicción del precio por metro cuadrado. En la preparación de los datos, se realizó una ingeniería de variables para enriquecer las variables X y se seleccionó el modelo de aprendizaje de máquina más adecuado para la predicción de π. En el modelado, se entrenó el modelo y se evaluó su desempeño en el conjunto de prueba. Finalmente, en la evaluación, se interpretaron los resultados del modelo para identificar las variables más relevantes en la predicción del precio por metro cuadrado.

4.1. Implicaciones

Los resultados obtenidos en este trabajo tienen implicaciones significativas para la construcción de un índice de precios de la vivienda en México. En primer lugar, se identificó que el valor del terreno es la variable más relevante en la predicción del precio por metro cuadrado. Esto sugiere que el índice de precios de la vivienda podría basarse en el valor del terreno en lugar del valor de la construcción, como lo hace el índice de SHF. Además, esto abre la posibilidad de explorar plusvalías en delimitaciones geográficas más específicas y no sólo en las divisiones político-administrativas del país. Por ejemplo, podríamos evaluar la plusvalía en áreas cercanas al Parque Fundidora en Monterrey, como se muestra en la figura 9.

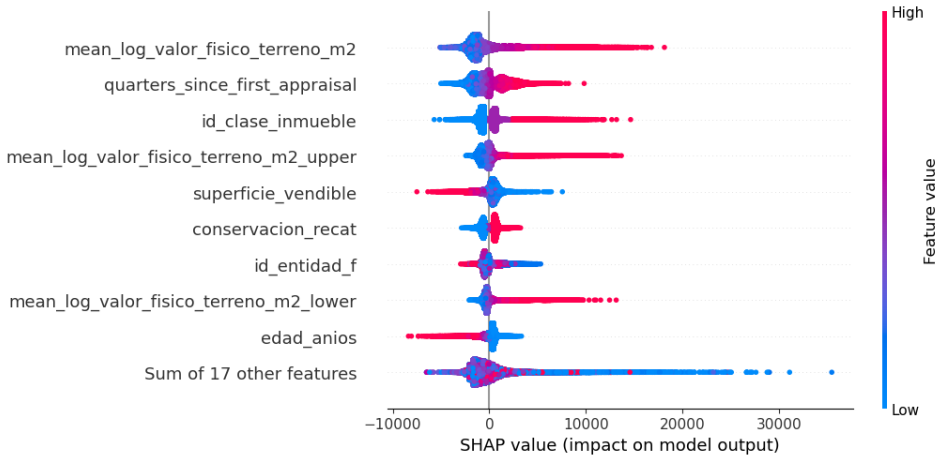
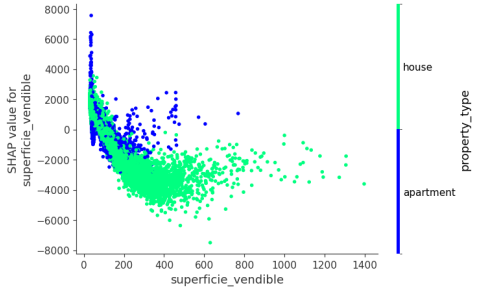


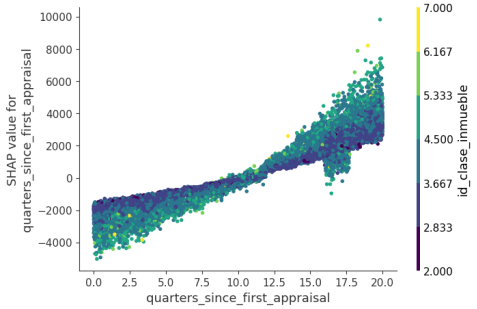
Figura 9: Importancia de las variables en el modelo CatBoost.

En segundo lugar, se identificaron formas funcionales de las variables independientes con respecto al precio por metro cuadrado, *ceteris paribus*, lo que permitirá construir un modelo econométrico con supuestos más realistas, como el crecimiento a tasas decrecientes del precio en relación con la superficie vendible, como se muestra en la figura 10a. En tercer lugar, se detectaron interacciones importantes entre variables, como la plusvalía en función de la clase

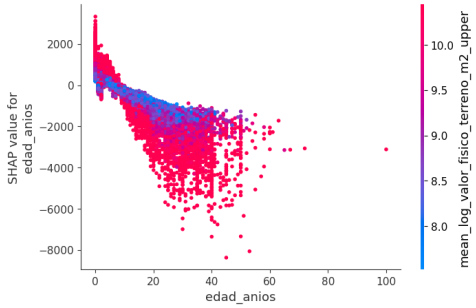
de inmueble según la SHF (ver figura 10b), o las diferentes pendientes de la depreciación de un inmueble dependiendo del valor del terreno (ver figura 10c). Estas interacciones proporcionan una base sólida para incorporar términos de interacción en el modelo econométrico y ofrecen una mayor comprensión de cómo se interrelacionan las variables.



(a) Precio por metro cuadrado vs. superficie vendible.



(b) Plusvalía vs. clase de inmueble.



(c) Depreciación vs. valor del terreno.

Figura 10: Interacciones entre variables.

4.2. Comparativa con el Status Quo

En comparación con el índice de precios de la vivienda de la SHF, este modelo incluye el precio de la vivienda en función de la superficie vendible, la depreciación de la vivienda en función de la antigüedad y el valor del terreno, interacciones no lineales entre variables, tipificación de zonas costeras y urbanas, y la plusvalía en función de la clase de inmueble. Además, este modelo es más transparente y fácil de interpretar que el modelo de la SHF, ya que se basa en un enfoque de aprendizaje automático que permite identificar las variables más relevantes en la predicción del precio por metro cuadrado y las interacciones entre ellas. Por otro lado, el modelo de la SHF se basa en un enfoque econométrico que no permite identificar las interacciones entre variables y es menos transparente en términos de la importancia de las variables en la predicción del precio por metro cuadrado.

Un resultado secundario de este trabajo es la identificación de errores en la base de datos de la SHF, compartida por Metrics Analytics. En particular, el análisis preliminar de los datos reveló una falla en la extracción de información por parte del proveedor de datos, lo que al finalizar el presente trabajo conllevó al proveedor a realizar una nueva extracción de datos, con un aumento significativo en el número de observaciones, de 1,048,575 a 2,626,783.

Esto permitirá una mayor precisión en la predicción de los precios de la vivienda en México y una mayor representatividad de la muestra en futuras investigaciones.

4.3. Limitaciones y Futuras Investigaciones

Una limitación importante de este trabajo es que se basa en datos de avalúos de la SHF, que pueden no ser representativos de todo el mercado de la vivienda en México. Por ejemplo, aquellos inmuebles que se anuncian en portales inmobiliarios y no pasan por un proceso de avalúo, los cuales reflejan más las dinámicas de oferta y demanda y son de clases residenciales y residenciales plus, representan menos del 1.1 % de la muestra en esta base de datos.

El siguiente paso natural es construir un modelo econométrico que incorpore las formas funcionales y las interacciones identificadas en este trabajo. Este modelo permitirá estimar el precio por metro cuadrado de la vivienda en México con mayor precisión y transparencia que el modelo de la SHF. Además, se puede explorar la posibilidad de incorporar otras fuentes de datos, como portales inmobiliarios, para mejorar la representatividad del modelo. Para este modelo econométrico se sugiere la utilización de un modelo de regresión lineal con términos de interacción y transformaciones no lineales de las variables independientes. También puede ser de interés explorar modelos bayesianos o modelos dinámicos lineales para incorporar la información temporal de los datos.

Una vertiente de investigación interesante es la construcción de modelos que basen la predicción de los inmuebles únicamente en función de sus vecinos, los vecinos de los vecinos, y así sucesivamente, para identificar clusters de plusvalía en el país. Esto se puede lograr con redes neuronales basadas en grafos, como *Graph Convolutional Networks* (GCN) [Ferludin et al., 2022], que permiten modelar la estructura de los datos y capturar las relaciones entre los inmuebles, además de ser escalables a millones de observaciones.

El fenómeno de la plusvalía es crucial para la economía y la geografía, y su estudio puede aportar información valiosa para la toma de decisiones en el sector inmobiliario. Un índice de precios de la vivienda basado en técnicas avanzadas de aprendizaje automático y econometría permitirá a los tomadores de decisiones contar con información más precisa y transparente. La comprensión y análisis del aumento de precios de vivienda no solo ayudan a entender las dinámicas del mercado, sino que también ofrecen oportunidades para tomar decisiones estratégicas que pueden influir en la planificación urbana, la inversión inmobiliaria y el desarrollo económico regional.

Es importante tener en cuenta que el Índice de Vivienda elaborado por SHF es una herramienta crucial para el análisis de la estabilidad financiera en México, especialmente en los reportes del Banco de México (Banxico). Este índice proporciona datos sobre los precios de la vivienda, que son utilizados en los reportes de estabilidad financiera para evaluar las condiciones del mercado inmobiliario y su impacto en la economía. La información contenida en estos reportes permite a Banxico monitorear las dinámicas del sector inmobiliario, identificando riesgos potenciales que podrían afectar la estabilidad financiera general del país. Por ejemplo, al analizar las variaciones en el índice de precios de la vivienda, Banxico puede detectar tendencias de sobrecalentamiento o enfriamiento en el mercado, lo que a su vez influye en sus decisiones de política monetaria y regulación financiera. Además, el uso del índice de vivienda en los reportes de estabilidad financiera ayuda a proporcionar un marco más amplio para entender cómo las fluctuaciones en el sector inmobiliario pueden repercutir en el sistema financiero, afectando tanto a los consumidores como a las instituciones financieras. Es por ello que la precisión del índice de vivienda es de gran importancia para garantizar que las decisiones de política monetaria y regulación financiera se basen en

información confiable y actualizada.

Mejorar la precisión de estos índices puede transformar la forma en que se perciben y gestionan las inversiones en bienes raíces, impulsando un desarrollo más equitativo y sostenible.

Referencias

[Ahlfeldt and Liao, 2024] Ahlfeldt, G. and Liao, L. (2024). How does supply and demand affect the housing market? *LSE Online*. Accessed: 2024-09-16.

[Akerlof, 1978] Akerlof, G. A. (1978). The market for "lemons": Quality uncertainty and the market mechanism. In *Uncertainty in economics*, pages 235–251. Elsevier.

[Bollen, 1989] Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological methods & research*, 17(3):303–316.

[Breiman, 2017] Breiman, L. (2017). *Classification and regression trees*. Routledge.

[Buitinck et al., 2013] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., et al. (2013). Api design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*.

[Case and Shiller, 1987] Case, K. E. and Shiller, R. J. (1987). Prices of single family homes since 1970: New indexes for four cities.

[Casella and Berger, 2024] Casella, G. and Berger, R. (2024). *Statistical inference*. CRC Press.

[Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

[de Economía, 2024] de Economía, S. (2024). Real estate: Wages, production, investment, opportunities and complexity. *Data México*. Accessed: 2024-09-16.

[Dorogush et al., 2018] Dorogush, A. V., Ershov, V., and Gulin, A. (2018). Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.

[Ferludin et al., 2022] Ferludin, O., Eigenwillig, A., Blais, M., Zelle, D., Pfeifer, J., Sanchez-Gonzalez, A., Li, W. L. S., Abu-El-Haija, S., Battaglia, P., Bulut, N., et al. (2022). Tf-gnn: Graph neural networks in tensorflow. *arXiv preprint arXiv:2207.03522*.

[Guerrero Espinosa, 2005] Guerrero Espinosa, J. A. (2005). Elaboración de índice de precios de vivienda shf.

[Guide, 2024] Guide, G. P. (2024). Mexico's residential property market analysis 2024. *Global Property Guide*. Accessed: 2024-09-16.

[Heyman and Sommervoll, 2019] Heyman, A. V. and Sommervoll, D. E. (2019). House prices and relative location. *Cities*, 95:102373.

[Ho et al., 2021] Ho, W. K., Tang, B.-S., and Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research*, 38(1):48–70.

[Lundberg and Lee, 2017] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

[Maneewongvatana and Mount, 1999] Maneewongvatana, S. and Mount, D. M. (1999). Analysis of approximate nearest neighbor searching with clustered point sets. *arXiv preprint cs/9901013*.

[McDonald and McMillen, 2010] McDonald, J. F. and McMillen, D. P. (2010). *Urban economics and real estate: Theory and policy*. John Wiley & Sons.

[Montgomery and Runger, 2010] Montgomery, D. C. and Runger, G. C. (2010). *Applied statistics and probability for engineers*. John wiley & sons.

[Murtagh et al., 2020] Murtagh, E. M., Mair, J. L., Aguiar, E., Tudor-Locke, C., and Murphy, M. H. (2020). Outdoor walking speeds of apparently healthy adults: A systematic review and meta-analysis. *BMC Sports Science, Medicine and Rehabilitation*, 12:1–15.

[Puu and Weidlich, 2007] Puu, T. and Weidlich, W. (2007). *The stability of hexagonal tessellations*. Nomos Verlagsgesellschaft mbH & Co. KG.

[Real Estate Market, 2024a] Real Estate Market (2024a). Cómo se determinan los precios de las propiedades. Accessed: 2024-07-23.

[Real Estate Market, 2024b] Real Estate Market (2024b). Importancia del real estate en la economía. Accessed: 2024-07-23.

[Research, 2024] Research, B. (2024). Mexico real estate outlook. first semester 2024. *BBVA Research*. Accessed: 2024-09-16.

[Rey-Blanco et al., 2024] Rey-Blanco, D., Zofío, J. L., and González-Arias, J. (2024). Improving hedonic housing price models by integrating optimal accessibility indices into regression and random forest analyses. *Expert Systems with Applications*, 235:121059.

[Rosen, 1974] Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy*, 82(1):34–55.

[Schubert et al., 2017] Schubert, E., Sander, J., Ester, M., Kriegel, H. P., and Xu, X. (2017). Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21.

[Snedecor and Cochran, 1989] Snedecor, G. W. and Cochran, W. G. (1989). Statistical methods, 8thedn. *Ames: Iowa State Univ. Press Iowa*, 54:71–82.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

[Wirth and Hipp, 2000] Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1, pages 29–39. Manchester.