Imágenes de Canva educativo

Estimación de valores de referencia de financiamiento hipotecario en México mediante aprendizaje automático: modelo basado en datos administrativos del SNIIV (2016 a 2025)

Estimation of mortgage financing reference values in Mexico using machine learning: model based on SNIIV administrative data (2016 to 2025)

Fabián Espinoza-Garza[1]*, Yobani Martínez-Ramírez[2], Alan Ramírez-Noriega[2]

RESUMEN

La estimación de valores de referencia del financiamiento hipotecario puede aportar información valiosa para corroborar los resultados de la valuación de propiedades. El objetivo de este estudio fue estimar el monto promedio del crédito hipotecario en México por combinación única de variables socioeconómicas, geográficas y temporales, es decir, valores de referencia agregados y no montos de créditos individuales. Se emplearon tres modelos de aprendizaje automático (AA): bosques aleatorios (Random Forest, RF), árboles de decisión (Decision Tree, DT) y regresión lineal (Linear Regression, LR), a partir de datos administrativos del Sistema Nacional de Información e Indicadores de Vivienda (SNIIV) del período 2013 a 2025, que comprende 4 515 891 registros. Se analizó el período 2016 a 2025, dado que la UMA (Unidad de Medida y Actualización), empleada para clasificar el valor de la vivienda, opera desde 2016. La depuración comprendió la supresión de valores faltantes, el filtrado por modalidad del crédito, el truncamiento por percentiles, la detección de anomalías mediante Isolation Forest y el filtrado por rangos normativos de UMA, con lo que se obtuvieron 829 542 registros que, agrupados por combinaciones únicas de atributos, produjeron 139 069 observaciones. La estabilidad de los modelos se comprobó mediante validación cruzada de cinco pliegues y su desempeño se evaluó en un conjunto de prueba con la raíz del error cuadrático medio (RMSE), el error porcentual absoluto medio (MAPE) y el coeficiente de determinación (R^2). El modelo RF mostró el mejor rendimiento, con R^2 de 0.970, RMSE de 187 250 pesos y MAPE de 11.48 %, con variabilidad mínima entre pliegues. Este estudio establece que el AA, en especial el RF, puede complementar los métodos tradicionales de valuación, si bien, su uso pleno exige incluir variables físicas del inmueble e incorporarse al marco regulatorio vigente.

PALABRAS CLAVE: aprendizaje automático, bosques aleatorios, SNIIV, estimación, financiamiento hipotecario.

ABSTRACT

Estimating mortgage financing reference values can provide valuable information to corroborate the results obtained in property valuation. The objective of this study was to estimate the average mortgage loan amount in Mexico for each unique combination of socioeconomic, geographic, and temporal variables, that is, aggregated reference values rather than individual loan amounts. Three machine learning (ML) models were used: random forests (RF), decision trees (DT), and linear regression (LR), based on administrative data from the National Housing Information and Indicators System (SNIIV) for the 2013 to 2025 period, comprising 4 515 891 records. The 2016–2025 period was analyzed, since the UMA (unit of measure and update), used to classify housing value, has been in operation since 2016. Data refinement included removing missing values, filtering by loan modality, percentile truncation, anomaly detection using Isolation Forest and filtering by UMA normative ranges, yielding 829 542 records which, grouped into unique attribute combinations, produced 139 069 observations. Model stability was verified through five-fold cross-validation, and performance was evaluated on a test set using the root mean square error (RMSE), the mean absolute percentage error (MAPE), and the coefficient of determination (R^2). The RF model showed the best performance, with an R^2 of 0.970, an RMSE of 187 250 pesos, and a MAPE of 11.48 %, with minimal variability across folds. This study establishes that ML, especially RF, can complement traditional valuation methods, although its full implementation requires including physical property variables and its integration into the current regulatory framework.

KEYWORDS: machine learning, random forest, SNIIV, estimation, mortgage financing.

***Correspondencia:** fabianespinozagarza@uaim.edu.mx**/Fecha de recepción:** 17 de septiembre de 2025**/Fecha de aceptación:** 3 de septiembre de 2026**/Fecha de publicación:** 21 de septiembre de 2026.

[1]Universidad Autónoma Indígena de México, Unidad Mochis, Fuente de Cristal 2334, Fuentes del Bosque, Los Mochis, Sinaloa, México, C. P. 81229. [2]Universidad Autónoma de Sinaloa, Facultad de Ingeniería Mochis, Los Mochis, Sinaloa, México, C. P. 81210.

https://doi.org/10.29059/cienciauat.v20i2.2058

INTRODUCCIÓN

El Aprendizaje Automático (AA) (ML, por sus siglas en inglés: Machine Learning), es una rama de la inteligencia artificial que desarrolla sistemas que aprenden y mejoran a partir de la experiencia, sin una programación específica (Jordan y Mitchell, 2015), lo que permite emplear algoritmos para extraer información de los datos y realizar predicciones basadas en ellos.

Aplicar estas nuevas herramientas tecnológicas a la valuación inmobiliaria permite aprovechar la riqueza de datos disponibles (por ejemplo, en el segmento de valor de vivienda se consideran rangos que incluyen datos como superficie construida, número de recámaras y calidad constructiva; por otra parte, se considera si es vivienda usada o nueva y la localización del inmueble), empleando diversas fuentes de información relevantes, que sería complicado lograr con los métodos tradicionales (Jafary y col., 2024; Stang y col., 2023). Sin embargo, es necesario considerar que la eficiencia de los algoritmos de AA depende precisamente de que la información destacada esté disponible. Al respecto, el Sistema Nacional de Información e Indicadores de Vivienda (SNIIV) proporciona una base de datos con registros de financiamientos para la vivienda en México; información muy útil para su análisis, contando con registros desde 2013 (SNIIV, 2026). Incluye diversos atributos relevantes, así como, el valor final y clasificaciones de valor expresados en UMAs (Unidad de Medida y Actualización), referencia económica que se actualiza cada año y se emplea para sustituir el uso del salario mínimo como base de cálculo desde el año 2016, de acuerdo con el Instituto Nacional de Estadística y Geografía (INEGI, 2025a). Por lo que su uso, teóricamente, es viable para entrenar modelos de AA y desarrollar un algoritmo capaz de estimar el importe de financiamientos que permitan inferir los precios de referencia del valor de viviendas en México.

El valor de un bien inmueble se ve influido por diversos factores, los cuales son considerados en los métodos tradicionales de cálculo, como la ubicación, metros cuadrados (m^2) de construcción, atributos particulares de la vivienda, oferta y demanda, entre otros (Jafary y col., 2024; Stang y col., 2023). Actualmente, tecnologías como los modelos de valuación automatizada (AVM, por sus siglas en inglés: Automated Valuation Models), los modelos hedónicos de precios y los algoritmos de AA, junto con los datos históricos disponibles, han mejorado el análisis de datos del mercadoinmobiliario, permitiendo estimaciones más precisas (Verma y col., 2023). Estas tecnologías ya operan de manera complementaria, y en ciertos procesos, alterna a las instancias tradicionales de avalúo. En Estados Unidos, los AVM se emplean en el origen y renovación de créditos hipotecarios (Glumac y Des-Rosiers, 2021; Kok y col., 2017); en los Países Bajos, la valuación catastral masiva anual es apoyada con modelos automatizados; también se usan en el Reino Unido y Corea del Sur (Glumac y Des-Rosiers, 2021; Hong y Kim, 2022). Sin embargo, la existencia de modelos de AA más avanzados y la mayor disponibilidad de datos, permite desarrollar algoritmos precisos y eficientes (Kok y col., 2017), pero su adopción requiere un análisis profundo para lograr una aceptación generalizada, representando así un desafío en la implementación de nuevas metodologías aplicables a la valuación inmobiliaria (Glumac y Des-Rosiers, 2021; Valier, 2020).

Un estudio que ha aprovechado parcialmente los datos administrativos del SNIIV es el realizado por Barrientos-Matute y Ruvalcaba-Gómez (2023), sin embargo, no emplearon modelos de AA y se utilizó solo en una zona específica de México.

La valuación se entiende como un procedimiento de carácter técnico y metodológico que, mediante un análisis integral de los aspectos físicos, económicos, sociales, jurídicos y de mercado, permite estimar el monto, expresado en términos monetarios, de las variables cuantitativas y cualitativas que inciden en el valor de un bien, de acuerdo con el Instituto de

Administración y Avalúos de Bienes Nacionales (INDAABIN, 2025). En este sentido, las Normas Internacionales de Valuación emitidas por el Consejo de Normas Internacionales de Valuación (IVSC, por sus siglas en inglés: International Valuation Standards Council) establecen un marco normativo común que favorece la transparencia, coherencia y credibilidad en las prácticas de valuación a nivel global (IVSC, 2022). Sin embargo, la valuación inmobiliaria, basada en métodos normados y ampliamente aceptados, es un tema de debate tanto en la práctica profesional como en el ámbito académico, debido a la discrepancia entre los valores estimados en las valuaciones y los precios reales de las transacciones (Kok y col., 2017).

El modelo de valuación inmobiliaria en México, aunque reglado y estructurado, muestra limitaciones en su aplicación y alcance, por lo que existe una necesidad de una mayor integración con las normas internacionales y un enfoque multidisciplinario (Salas-Tafoya, 2018). En el contexto mexicano, la regulación de los servicios de valuación recae principalmente en la Sociedad Hipotecaria Federal (SHF) y el Instituto de Administración y Avalúos de Bienes Nacionales (INDAABIN) (Salas-Tafoya, 2018). Pero, la implementación de otros enfoques, como los que ofrece el AA a través de modelos como el de bosques aleatorios (RF, por sus siglas en inglés: Random Forest), árboles de decisión (DT, por sus siglas en inglés: Decision Tree) y regresión lineal (LR, por sus siglas en inglés: Linear Regression), podría mejorar la eficiencia, la precisión y la consistencia de los procesos de valuación.

El objetivo de este trabajo fue estimar el monto de financiamiento hipotecario promedio aprobado en México mediante los modelos de AA, RF, DT y LR, entrenados con datos del SNIIV de 2016 a 2025, que cubren todo el territorio nacional, además de comparar su capacidad predictiva, con el propósito de evaluar su posible uso como instrumento de referencia complementario a los métodos de valuación tradicionales.

MATERIALES Y MÉTODOS

La metodología utilizada en el presente estudio fue cuantitativa (Hernández-Sampieri, 2014), enfocada en el desarrollo y evaluación experimental de un modelo de estimación del monto de financiamiento para vivienda. El enfoque de la investigación incluye limpieza, procesamiento de datos, ingeniería de características y evaluación de los modelos de AA (Mody y col., 2023).

El procedimiento metodológico consistió en las siguientes fases: (1) revisión de la literatura enfocada a modelos de AA aplicados a la valuación inmobiliaria; (2) adquisición y preprocesamiento de datos: limpieza, tratamiento de valores atípicos y selección de variables pertinentes; (3) implementación y entrenamiento de modelos de AA; (4) evaluación comparativa del rendimiento a través de las métricas Coeficiente de Determinación (R^2), Raíz del Error Cuadrático Medio (RMSE, por sus siglas en inglés: Root Mean Square Error) y Error Porcentual Absoluto Medio (MAPE, por sus siglas en inglés: Mean Absolute Percentage Error), empleando una partición de 80/20 para entrenamiento y prueba, validación cruzada para confirmar estabilidad del modelo y análisis de importancia de variables del modelo seleccionado. Las cuatro fases estructuran los apartados siguientes de esta sección, como se indica en los subtítulos correspondientes.

Fase 1. Revisión de la literatura

Se partió de la revisión sistemática de la literatura (RSL) sobre modelos de AA aplicados a la valuación inmobiliaria publicada por Espinoza-Garza y col. (2024); dicha revisión se emplea aquí únicamente como antecedente y su protocolo metodológico completo (bases de datos, cadenas de búsqueda y criterios de inclusión) puede consultarse en la fuente citada, por lo que no se replica en el presente artículo. La selección de los modelos de RF y LR se basó en dicha revisión sistemática, donde reportaron que estos modelos destacaron en un análisis comparativo entre 15 modelos incluidos en el estudio, que emplearon información con características similares a la

base de datos del SNIIV. Los modelos de AA que analizan y comparan, se centran en el procesamiento de datos estructurados, considerando atributos cuantitativos y categóricos de bienes inmuebles, y no en imágenes u otras fuentes de información, lo que la hace pertinente para los objetivos planteados.

Fase 2. Adquisición y preprocesamiento de datos

Fuente de datos

Se consideró la información pública procedente del SNIIV que cuenta con información del periodo 2013 a 2025, y consiste en un total de 4 515 891 registros de financiamiento para vivienda en México y 16 variables. El SNIIV es la plataforma oficial de la Secretaría de Desarrollo Agrario, Territorial y Urbano (SEDATU) que concentra los registros administrativos de los financiamientos para vivienda otorgados por los organismos que integran el sistema (SNIIV, 2026). Cada registro corresponde a un crédito autorizado y describe sus atributos temporales (año y mes de autorización), geográficos (entidad federativa y municipio), socioeconómicos del acreditado (sexo, rango de edad y rango de ingresos) y financieros (organismo, modalidad, destino, tipo, monto y clasificación del valor de la vivienda expresada en UMA), lo que la convierte en la fuente pública más completa sobre el financiamiento formal de vivienda en el país.

Variables del modelo

Se inició con 16 variables, considerando variables numéricas: monto de financiamiento aprobado (monto), número de créditos autorizados (acciones) y año (año); y categóricas, entre las que destacan la entidad federativa (clave_entidad), el municipio (cve_mun, clave_municipio), nombre de entidad o municipio (entidad/municipio), el organismo y modalidad del crédito (organismo, modalidad, destino, tipo), características del beneficiario (sexo, edad_rango, ingresos_rango) y de la vivienda (vivienda_valor), así como el mes de autorización (mes). Sin embargo, la base de datos final utilizada para realizar los entrenamientos de los modelos se constituyó de 7 variables (Tabla 1), resultado de eliminar 9 de las 16 variables originales: (1) el mes, (2) el nombre de la entidad federativa, (3) el nombre del municipio, (4) el sexo, (5) el rango de edad, (6) el organismo financiador, (7) el destino del crédito, (8) el tipo de financiamiento y (9) el número de acciones. Dichas variables se descartaron en función de tres criterios: (1) redundancia, cuando la información ya estaba contenida en otra variable; (2) ausencia de valor predictivo, que no inciden directamente en el monto del crédito; e (3) irrelevancia para la valuación del inmueble, excluyendo variables demográficas del acreditado cuya relación con el valor de la vivienda no es directa.

Adquisición y depuración de datos

Los datos para el análisis y modelado se obtuvieron y prepararon, considerando como fuente de datos los proporcionados por el SNIIV, que incluye instituciones que otorgan créditos para vivienda, como Comisión Nacional Bancaria y de Valores (CNBV), Banco Nacional del Ejército, Fuerza Aérea y Armada (BANJERCITO), Comisión Federal de Electricidad (CFE), Comisión Nacional de Vivienda (CONAVI), Fondo Nacional de Habitaciones Populares (FONHAPO), Fondo de la Vivienda del Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado (FOVISSSTE), HABITAT MEXICO, Instituto del Fondo Nacional de la Vivienda para los Trabajadores (INFONAVIT), Instituto de la Vivienda (INVI), Instituto de Seguridad Social para las Fuerzas Armadas (ISSFAM), Petróleos Mexicanos (PEMEX) y SHF.

Se incluyeron créditos autorizados en México de 2013 a 2025, con un promedio de 347 376 registros por año y un total de 4 515 891 registros disponibles (SNIIV, 2026). Los archivos anuales de datos abiertos del SNIIV empleados en esta versión se descargaron el 24 de febrero de 2026 y contienen los créditos autorizados hasta el 31 de diciembre de 2025, último corte anual disponible; dado que la SEDATU actualiza periódicamente los archivos publicados, todas las cifras reportadas corresponden a ese corte y son reproducibles con los scripts del repositorio del proyecto.

Tabla 1. Datos depurados de la base de datos del SNIIV y sus características.
Table 1. Refined data from the SNIIV database and its characteristics.

Nombre del campo	Descripción	Tipo	Valores	Rol	Criterio
Año	Año del crédito	Numérica discreta	2013 a 2025	Explicativa	Capta tendencias temporales y ajuste UMA
Clave_entidad	Clave INEGI de entidad federativa	Categórica ordinal	1 a 32	Explicativa	Ubicación regional
cve_mun Clave_municipio	Clave INEGI de municipio	Categórica ordinal	Variable	Explicativa	Ubicación local
Modalidad	Modalidad del crédito hipotecario	Categórica nominal	1 = Vivienda nueva *2 = Mejoramiento 3 = Vivienda usada *4 = Otros programas	Explicativa	Determina tipo de inversión
Ingresos_rango	Rango de ingresos del acreditado en UMAs	Categórica ordinal	1 = 2.6 o menos 2 = 2.61 a 4.00 3 = 4.01 a 6.00 4 = 6.01 a 9.00 5 = 9.01 a 12.00 6 = más de 12 (UMAs mensuales)	Explicativa	Correlacionado con segmento de vivienda
Vivienda_valor	Segmento de valor de vivienda según UMAs	Categórica ordinal	1 = Económica 2 = Popular 3 = Tradicional 4 = Media; 5 = Residencial 6 = Residencial plus	Explicativa	Variable de mayor importancia predictiva
Monto_promedio	Monto promedio del crédito hipotecario (MXN)	Numérica continua	Variable	Variable objetivo	Variable a predecir

*Modalidades excluidas en la depuración.

El procesamiento de la base de datos se llevó a cabo utilizando el lenguaje de programación Python (versión 3.10), junto con la biblioteca especializada Pandas (McKinney, 2010).

Las variables seleccionadas cubren tres dimensiones clave para la estimación del monto de financiamiento. La temporal, representada por año, captura la evolución del mercado inmobiliario en términos de inflación y política habitacional. La geográfica, mediante las claves de entidad federativa y municipio, refleja las diferencias territoriales en el valor de la vivienda. Cabe precisar que estas claves no se emplean como magnitudes numéricas: en los modelos de árboles (DT y RF) actúan como categorías de partición territorial, ya que cada división sucesiva del árbol agrupa conjuntos de entidades o municipios sin presuponer un orden ni una distancia entre las claves. Su función es permitir que el modelo capture diferencias territoriales en los montos, asociadas, por ejemplo, a los costos del suelo

y de construcción de cada territorio, no establecer una relación numérica directa con la variable objetivo. Por esta razón, las claves se incluyen como predictores de partición, pero se excluyen de la interpretación sustantiva del análisis de correlación, cuyo coeficiente carece de significado para códigos de identificación sin orden territorial. Por último, la dimensión socioeconómica, integrada por el rango de ingresos del beneficiario y el valor de la vivienda, caracteriza el segmento de mercado del inmueble, mientras que modalidad complementa esta dimensión al determinar las condiciones del financiamiento otorgado. Las variables de carácter administrativo o asociadas al perfil demográfico del acreditado se excluyeron por no contribuir directamente a la estimación del monto del acreditado.

La variable modalidad, si bien de naturaleza administrativa, se conservó porque contiene el tipo de operación inmobiliaria (adquisición de vivienda nueva, usada o suelo), condición que incide directamente en el monto del financiamiento y no está contenida en otras variables. No obstante, por tratarse de una variable categórica nominal, reducida a dos categorías tras la depuración (vivienda nueva y usada), no se incluye en la matriz de correlación, ya que el coeficiente de rangos carece de sentido para categorías sin un orden intrínseco; su aporte se evalúa únicamente a través de los modelos, donde actúa como predictor de partición. Se buscó asegurar la calidad y consistencia de los datos mediante la identificación y el tratamiento de valores faltantes, así como el filtrado de valores atípicos. Además, se eliminaron los datos que tenían atributos con valores faltantes, reduciendo los registros a 3 237 394 observaciones. Posteriormente, se conservaron únicamente los créditos de vivienda nueva o usada, considerando nuevas las viviendas recién construidas que nunca habían sido habitadas, y se descartaron los correspondientes a mejoramientos y otros programas; dicho filtro de modalidad constituyó la reducción principal de esta etapa, con 622 748 registros eliminados (de 3 237 394 a 2 614 646 registros).

Enseguida, se eliminaron los valores extremos mediante truncamiento por percentiles, eliminando el 0.01 % superior e inferior de monto del conjunto ya filtrado, lo que suprimió 11 180 registros adicionales (de 2 614 646 a 2 603 466 registros). Esta cifra, incluida en los 11 180 registros del truncamiento que se muestran en la Figura 1, supera el 0.02 % nominal del truncamiento porque el percentil inferior del monto coincide con 0.00 pesos: al aplicarse la desigualdad estricta, el filtro descarta también los 10 906 registros con monto nulo y 12 con monto negativo, carentes de sentido económico.

Además, se aplicó el algoritmo Isolation Forest (contamination = 0.000 01, n_estimators = 100, random_state = 42), que identificó y eliminó 27 registros atípicos multivariados (< 0.001 % del conjunto). En conjunto, el filtro de modalidad, el truncamiento por percentiles y el Isolation Forest eliminaron 633 955 registros (622 748 + 11 180 + 27), resultando un conjunto final depurado de 2 603 439 registros.

Posteriormente, se excluyeron los registros anteriores a 2016 (años 2013 a 2015), dado que la UMA opera como unidad de referencia para la clasificación de vivienda únicamente desde ese año, quedando 2 108 109 registros. A continuación, se aplicó un doble filtro por rangos normativos de UMA, con un límite inferior y un límite superior definidos por año y por clase de valor de la vivienda, conforme a los rangos oficiales de la clasificación por UMA del SNIIV (SNIIV, 2026). El propósito de este filtro fue conservar únicamente los créditos cuyo monto es consistente con la clase de valor de vivienda declarada en el propio registro: los montos inferiores al límite de su clase corresponden, por ejemplo, a cofinanciamientos parciales o errores de captura, mientras que los superiores resultan incompatibles con la clase declarada. De los 2 108 109 registros del periodo 2016 a 2025, se excluyeron 1 278 567 registros en total (60.6 %), 696 270 registros (33.03 %) por situarse debajo del límite inferior y 582 297 registros (27.62 %)

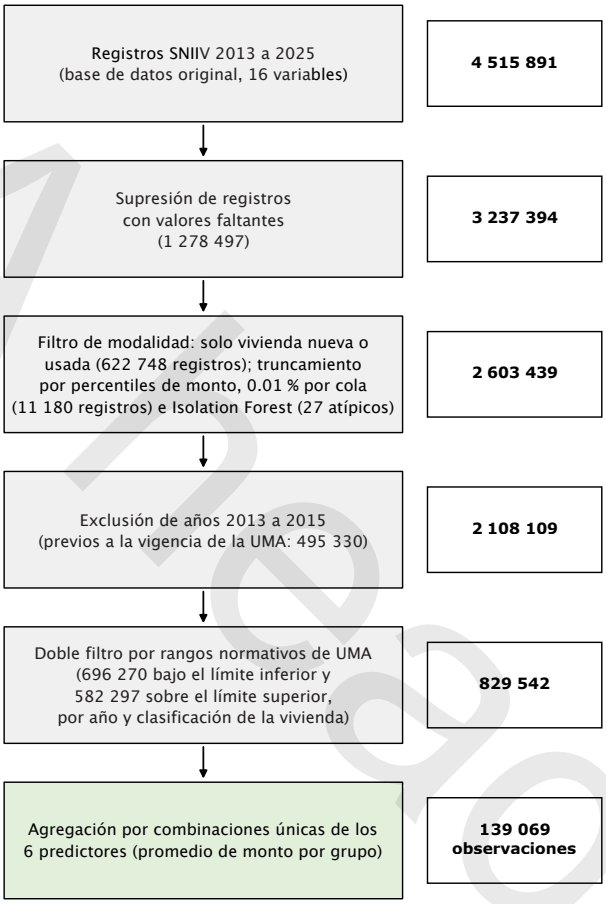


Figura 1. Cascada de depuración de la base de datos del SNIIV: registros conservados tras cada etapa, desde los 4 515 891 registros descargados hasta las 139 069 observaciones agregadas empleadas en el modelado.
Figure 1. Data-cleaning cascade of the SNIIV database: records retained after each stage, from the 4 515 891 downloaded records to the 139 069 aggregated observations used for modeling.

por superar el límite superior de su clase y año; conformando el conjunto de análisis con 829 542 registros de alta calidad.

Más adelante, se agruparon los datos considerando el promedio de las combinaciones posibles de sus atributos. Se calculó un valor promedio para cada año, usando solo las observaciones que tenían los mismos valores en los atributos de modalidad, entidad federativa, municipio, rango de ingresos y valor de la vivienda. En otras palabras, se calculó el promedio de los cambios para cada grupo de datos que tenían características similares en esas variables, consolidando en una sola observación todos los créditos que comparten idéntica combinación de dichos atributos.

Sobre cada grupo resultante se aplicó una función de agregación que calcula el monto crediticio promedio, reduciendo así el conjunto de datos a valores representativos por segmento, conformándose finalmente una base de datos con 139 069 observaciones agregadas (Figura 1). Cada combinación agrupó en promedio 5.96 créditos (mediana = 3; rango: 1 a 81); el 34.3 % de las combinaciones quedó integrado por un solo crédito, el 33.0 % de 2 a 5, el 27.0 % de 6 a 20 y el 5.6 % por más de 20 créditos.

Así mismo, se calculó la matriz de correlación y se visualizó mediante un mapa de calor para identificar la correlación entre variables y determinar la relación entre la etiqueta y sus atributos. Además, se generó un diagrama de dispersión geográfico de los registros, utilizando las coordenadas de cada municipio y diferenciando por segmento de valor de la vivienda, con el propósito de verificar la cobertura territorial de la base de datos. Ambas visualizaciones se elaboraron empleando las funciones que ofrece la biblioteca Matplotlib (Hunter, 2007). Es importante resaltar que, con el propósito de verificar la distribución geográfica de los registros, se incorporaron temporalmente las coordenadas de cada municipio (latitud y longitud) mediante el catálogo geoestadístico (INEGI, 2025b). Una vez concluida la revisión visual, estas columnas fueron eliminadas de la base de datos definitiva.

Análisis exploratorio y correlación de variables

Se realizó un análisis del coeficiente de correlación de Spearman utilizando la fórmula descrita por Restrepo y González (2007). Para la interpretación de la fuerza de asociación se siguió a Carranza-Rodríguez y col. (2024), considerando el valor absoluto del coeficiente: nula (0), pequeña (> 0 y ≤ 0.3), mediana

(> 0.3 y ≤ 0.5) y grande (> 0.5 y < 1) y completa (1). El análisis evaluó las relaciones bilaterales entre el monto de financiamiento aprobado, el año, las claves de entidad federativa y municipio, la modalidad del crédito, el rango de ingresos del beneficiario y el valor de la vivienda.

Construcción de la variable objetivo

Para cada combinación única de los 6 predictores se calculó el monto promedio aprobado (monto_promedio = media de monto por grupo), que constituye la variable objetivo del modelo. Este proceso de agregación transforma el problema de predecir créditos individuales en predecir un valor de referencia de mercado para cada perfil combinado de condiciones. Para garantizar la reproducibilidad de todos los procesos: partición de datos, validación cruzada, entrenamiento de modelos y detección de anomalías, se fijó la semilla aleatoria en random_state = 42 en todos los pasos.

Fase 3. Implementación y entrenamiento de los modelos de aprendizaje automático

En la fase de experimentación, se generaron tres modelos de regresión utilizando los algoritmos de aprendizaje automático detallados anteriormente (RF, DT y LR).

Modelos de AA

El modelo DT es un método supervisado no paramétrico utilizado para clasificación y regresión (Pedregosa y col., 2011). Este método construye el modelo realizando una serie de particiones binarias sucesivas, denominadas nodos. En su variante de regresión, que es la empleada en este estudio, la impureza de cada nodo se midió con la varianza de la variable objetivo dentro del nodo, es decir, con el error cuadrático medio (MSE, por sus siglas en inglés: Mean Squared Error) de las observaciones respecto a la media del nodo; en cada nodo se busca la división que produzca los nodos hijos más homogéneos posibles, esto es, la que minimice el MSE combinado de los nodos resultantes. La impureza del nodo t se calculó mediante la fórmula (1) (Breiman y col., 1984; Hastie y col., 2009):

$$MSE(t) = \frac{1}{N_t} \sum_{i \in D_t} (y_i - \overline{y_t})^2 \quad (1)$$

Donde:

MSE (t) = impureza del nodo t: promedio de los errores cuadráticos respecto a la media del nodo.

N_t = número total de observaciones en el nodo t.

D_t = conjunto de observaciones que integran el nodo t.

y_i = valor observado de la variable objetivo (monto promedio) para la observación i del nodo t.

$\overline{y_t}$ = media de la variable objetivo en el nodo t.

En cada nodo, el árbol elige la variable y el punto de corte que producen la mayor reducción del MSE ponderado de los nodos hijos respecto al nodo padre; la predicción de cada nodo terminal es la media de las observaciones que contiene.

A su vez, RF está conformado por un conjunto de árboles de decisión (DT); su predicción se obtuvo promediando los resultados de las predicciones de cada árbol individual del cual está compuesto. Por otra parte, el modelo LR se deriva de una técnica estadística utilizada para predecir el resultado de una variable de respuesta, empleando un número de variables que la explican conforme a la fórmula (2) (Maulud y Abdulazeez, 2020):

$$y = \beta_0 + \beta_1 x + \varepsilon \quad (2)$$

Donde:

y = la variable dependiente.

β_0 = el intercepto del modelo.

β_1 = el coeficiente de regresión o pendiente que mide el efecto de x sobre y.

ε = el término de error aleatorio que recoge la variabilidad no explicada por el modelo.

La regresión lineal múltiple se calculó con la fórmula (3) (Maulud y Abdulazeez, 2020):

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (3)$$

Donde:
$\widehat{\beta}$ = el vector de coeficientes estimados del modelo de regresión múltiple.
X = la matriz de variables explicativa.
X^T = la transpuesta de la matriz X.
y = el vector de los valores observados de la variable dependiente.
$(X^T X)^{-1}$ = la matriz inversa del producto $X^T X$.

Configuración de hiperparámetros

El modelo RF fue configurado con los siguientes hiperparámetros: n_estimators = 200, max_depth = 15, min_samples_split = 5, min_samples_leaf = 2, max_features = 'sqrt'. El modelo DT se entrenó sin restricción de profundidad (criterion = 'squared_error'). La LR se ajustó por mínimos cuadrados ordinarios sin regularización. Todos los modelos se inicializaron con random_state = 42 para reproducibilidad. Los valores de los hiperparámetros del RF se establecieron de manera heurística, partiendo de los valores por defecto de la biblioteca y de configuraciones recomendadas en la literatura, verificando su efecto mediante ensayos preliminares sobre el conjunto de entrenamiento; no se realizó una búsqueda sistemática de hiperparámetros (grid search o random search con validación cruzada; Bergstra y Bengio, 2012), lo que se reconoce como una oportunidad de mejora.

Implementación computacional

Tanto el entrenamiento como la evaluación del modelo se llevó a cabo utilizando las métricas disponibles en la biblioteca Scikit-Learn, versión 1.4 (Pedregosa y col., 2011). Posteriormente, se realizó la implementación, creando un entorno con código ejecutable en el lenguaje Python, versión 3.10, mediante la plataforma Colab (Colaboratory), la cual permitió escribir y ejecutar código en Python directamente desde el navegador, sin necesidad de configuraciones adicionales (Google, 2025). De esta manera, se crearon resultados reproducibles y se generó la documentación de los procesos realizados. Por último, se guardó el modelo previamente entrenado y validado mediante las métricas mencionadas, y fue así como se llevaron a cabo pruebas que permitieron verificar empíricamente los resultados obtenidos posteriormente.

La experimentación se realizó en Google Colaboratory (Google, 2025), empleando un entorno CPU estándar con Python 3.10 y Scikit-Learn 1.4. El tiempo total de entrenamiento de los tres modelos (RF, DT y LR) sobre el conjunto de entrenamiento fue de aproximadamente 15 min. El entorno de Colab fue seleccionado por su accesibilidad, gratuidad y compatibilidad con los paquetes de cómputo científicos utilizados.

Fase 4. Evaluación comparativa del rendimiento

Diseño experimental: partición de datos y validación cruzada

Para el diseño experimental se efectuó la división de los datos. Los registros del conjunto de datos se dividieron en dos subconjuntos mutuamente excluyentes: 80 % para entrenamiento y 20 % para prueba, con semilla aleatoria fija para garantizar la reproducibilidad. La validación cruzada de cinco pliegues (k-fold cross-validation, k = 5) se incluyó en el diseño experimental con un objetivo concreto: probar la estabilidad del modelo, al comprobar que el rendimiento observado no es consecuencia de una partición concreta de los datos, sino una propiedad genérica del algoritmo frente a distintas divisiones del conjunto de entrenamiento. Con este fin, el conjunto de entrenamiento (80 % de los datos) se dividió en cinco subconjuntos mutuamente excluyentes y exhaustivos; en cada iteración, se utilizaron cuatro subconjuntos para entrenar el modelo y el restante para evaluar su desempeño, rotando sistemáticamente hasta agotar las cinco combinaciones posibles. Así, el R^2 medio entre pliegues se toma como estimador de la consistencia interna del modelo. Dicho procedimiento es importante diferenciarlo de la evaluación final de desempeño predictivo, la cual se hizo únicamente sobre el conjunto de prueba independiente (20 % reservado desde el principio), que no participó en ningún proceso de entrenamiento o ajuste. De esta manera, las métricas constituyen una eva-

luación fuera de la muestra, libre de sesgo por sobreajuste.

Métricas de evaluación del rendimiento

El rendimiento de los modelos se midió mediante el coeficiente de determinación (R^2) que mide la proporción de la varianza de la variable dependiente explicada por el modelo, de modo que valores más altos reflejan un mejor ajuste al promedio de la variable objetivo. Es importante resaltar que el R^2 mide la proporción de varianza explicada y no constituye por sí mismo una medida de precisión predictiva en valores absolutos; para ese propósito se emplean el RMSE expresado en las mismas unidades que la variable objetivo y MAPE expresado en porcentaje. En particular, el RMSE cuantifica la magnitud promedio de los errores de predicción, penalizando en mayor medida los errores grandes, por lo tanto, valores más bajos indican mayor precisión del modelo y el MAPE cuantifica el error relativo promedio entre los valores reales y los estimados, expresado en términos porcentuales, de manera que, menores valores denotan una mayor exactitud en las predicciones. Se calculan con las fórmulas (4), (5) y (6), respectivamente (Gunes, 2023):

R^2 se calcula como se muestra en la fórmula (4):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

Donde:
n: número de observaciones.
y_i: valor real de la observación i.
$\hat{y}_i$: valor predicho por el modelo para la observación i.
$\bar{y}$: media aritmética de todos los valores reales.
$\sum (y_i - \hat{y}_i)^2$: suma de los errores cuadrados.
$\sum (y_i - \bar{y})^2$: suma total de cuadrados.

El RMSE resultante siempre es ≥ 0 y el valor más cercano a 0 es el ideal, se calcula mediante la fórmula (5):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2} \quad (5)$$

Donde:
n: número de observaciones o tamaño de la muestra.
y_i: valor real observado.
$\hat{y}_i$: valor predicho por el modelo.
$(y_i - \hat{y}_i)$: error de predicción.

MAPE mide el error promedio de predicción, el resultado se expresa en porcentaje mediante la multiplicación por 100, y su valor ideal es el más cercano a 0, mediante la fórmula (6):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (6)$$

Donde:
n: número de observaciones.
y_i: valor real observado.
$\hat{y}_i$: valor predicho por el modelo.
$\left|\frac{y_i - \widehat{y_i}}{y_i}\right|$: valor absoluto del error relativo.

Empleando las métricas antes mencionadas se midió el rendimiento de los modelos y estos se ajustaron para mejorar los resultados obtenidos, de acuerdo con varios autores como Deppner y col. (2023), Jung y col. (2022), Matey y col. (2022), Nazarov y Yarmatov (2023).

Consideraciones éticas

Los datos utilizados en este estudio son de acceso público y anónimos, publicados por el SNIIV. No contienen identificadores personales que permitan la individualización de los acreditados. Así mismo, el código fuente y el flujo de procesamiento están disponibles públicamente en el repositorio de GitHub: https://github.com/buhodigital/Estimaci-n-vivienda-modelos-AA

RESULTADOS

Análisis de correlación

Los resultados mostraron una correlación positiva grande entre el valor de la vivienda y el monto del crédito (ρ = 0.94) (Figura 2), magnitud favorecida por el filtro de rangos normativos de UMA, que acota el monto según la clase de la vivienda, lo cual indica que, a mayor valor de la vivienda, se otorgan montos de financiamiento más altos (Carranza-Rodríguez

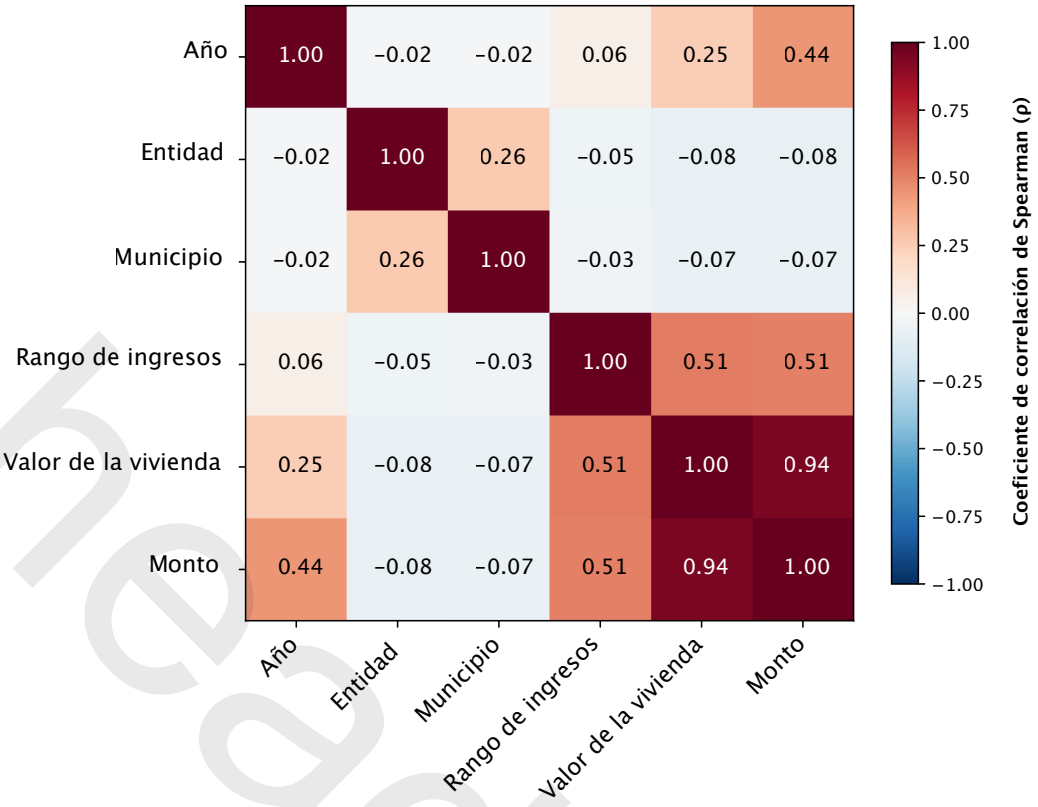


Figura 2. Matriz de correlación de Spearman entre las cinco variables predictoras ordinales y el monto de financiamiento aprobado, calculada sobre la base depurada corregida (n = 829 542 registros).

Figure 2. Spearman correlation matrix between the five ordinal predictor variables and the approved financing amount, computed on the corrected refined dataset (n = 829 542 records).

y col., 2024). También se encontraron correlaciones positivas grandes (Figura 2) entre el rango de ingresos y el valor de la vivienda y rango de ingresos y el monto (ρ = 0.51 en ambos casos), lo que sugiere que mayores ingresos del acreditado se asocian con un monto de crédito más alto, para obtener un valor más elevado del índice del valor de la vivienda. El año y el monto mostraron una correlación mediana (ρ = 0. 44), lo que podría asociarse a la tasa de inflación anual. Además, se reportaron correlaciones pequeñas entre el año y el valor de la vivienda (ρ = 0.25), lo que es congruente con el desplazamiento gradual de la composición de los segmentos de vivienda a lo largo del periodo analizado; y entre las claves de entidad y municipio (ρ = 0.26), asociación esperable por la estructura jerárquica de las claves geográficas del INEGI, que no se interpreta sustantivamente. En contraste, la relación del monto con las claves de entidad y municipio presentaron coeficientes cercanos a cero (ρ ≈ −0.08 y −0.07), que tampoco se interpretan sustantivamente: al ser códigos de identificación sin un orden territorial, estas variables participan en los modelos únicamente como categorías de partición territorial. La variable modalidad se excluyó de la matriz de correlación por tratarse de una variable categórica nominal, y se conservó solo como predictor del modelo. Todas las correlaciones de Spearman reportadas en esta sección, fueron estadísticamente significativas (P ≤ 0.001), lo cual es esperable dado el tamaño de la muestra (n = 829 542 registros).

En la Tabla 2 se observan las 10 entidades con mayor número de registros de créditos de vivienda del conjunto depurado. El Estado de México, Nuevo León, Veracruz, Jalisco y Gua-

Tabla 2. Entidades federativas con mayor número de registros en el conjunto depurado (n = 829 542).
Table 2. Federal entities with the largest number of records in the refined dataset (n = 829 542).

Entidad federativa	Registros (n)	Porcentaje (%)
México	73 575	8.87
Nuevo León	57 128	6.89
Veracruz	56 921	6.86
Jalisco	52 861	6.37
Guanajuato	39 104	4.71
Coahuila	38 317	4.62
Ciudad de México	36 392	4.39
Puebla	34 312	4.14
Tamaulipas	30 498	3.68
Sonora	29 923	3.61
Resto de las entidades (22)	380 511	45.86
Total	**829 542**	**100.00**

najuato concentraron la mayor cantidad de registros acumulando en conjunto el 33.7 % del total. La distribución geográfica se muestra en la Figura 3, sobre los contornos del territorio nacional y de las entidades federativas. A mayor valor del índice de valor de la vivienda, mayor tamaño, precio y equipamiento de la vivienda; estas se clasifican del 1 al 6 desde "económica" a "residencial plus" (SNIIV, 2026).

Evaluación y comparación de modelos

El modelo RF obtuvo el mejor rendimiento, con un coeficiente de determinación R^2 de 0.982 sobre el conjunto de entrenamiento (80 %) y R^2 de 0.970 sobre el conjunto de prueba independiente (20 %), lo que indica que, a partir exclusivamente de variables socioeconómicas y geográficas, es posible explicar el 97.0 % de la variabilidad en los montos promedio de financiamiento hipotecario. Asimismo, RF alcanzó el menor MAPE (11.48 %) y el menor RMSE (187 250 pesos), superando en ambas métricas a DT y LR (Tabla 3).

La Figura 4a ilustra la distribución de registros antes de aplicar los rangos normativos, mostrando la dispersión de montos fuera de los límites UMA, al contrastarla con la Figura 4b que presenta la distribución resultante tras la delimitación, donde la concentración de datos dentro de los rangos válidos es notoriamente mayor. Esta comparación visual complementa los resultados de dispersión del modelo RF presentados en la Figura 5: la depuración por rangos normativos reduce la dispersión de los montos, lo que se refleja en una mejor predicción.

En la Figura 5a se muestra la dispersión entre los valores reales y los predichos por el modelo RF en el conjunto de entrenamiento, con R^2 = 0.982. Se observa una mayor concentración de observaciones en el rango de valores bajos, segmento donde el modelo tiene mayor precisión, lo cual es coherente con la distribución asimétrica de los datos. El comportamiento se mantiene en el conjunto de prueba mostrado en la Figura 5b, con R^2 = 0.970, lo que confirma la capacidad de generalización del modelo sobre la base depurada mediante los rangos normativos de UMA.

Finalmente, la validación cruzada k-fold confirmó la robustez y estabilidad de los tres modelos desarrollados (Tabla 4). Las desviaciones estándar del R^2 entre pliegues fueron bajas en todos los casos: RF (σ = 0.000 3), LR (σ = 0.001 1) y DT (σ = 0.001 4). Para el caso particular del RF, los valores de R^2 por pliegue oscilaron en un rango de 0.970 a 0.971, indicando una gran consistencia frente a distintas particiones del conjunto de datos. De

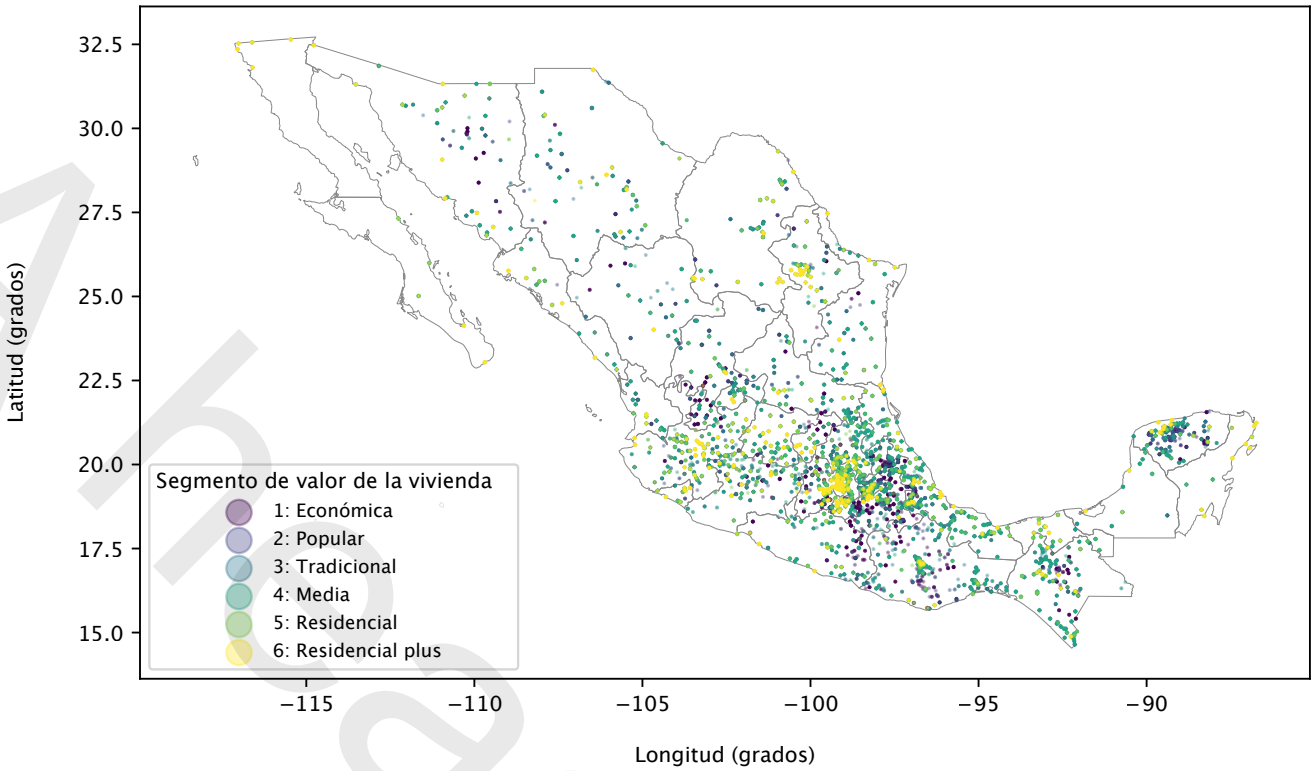


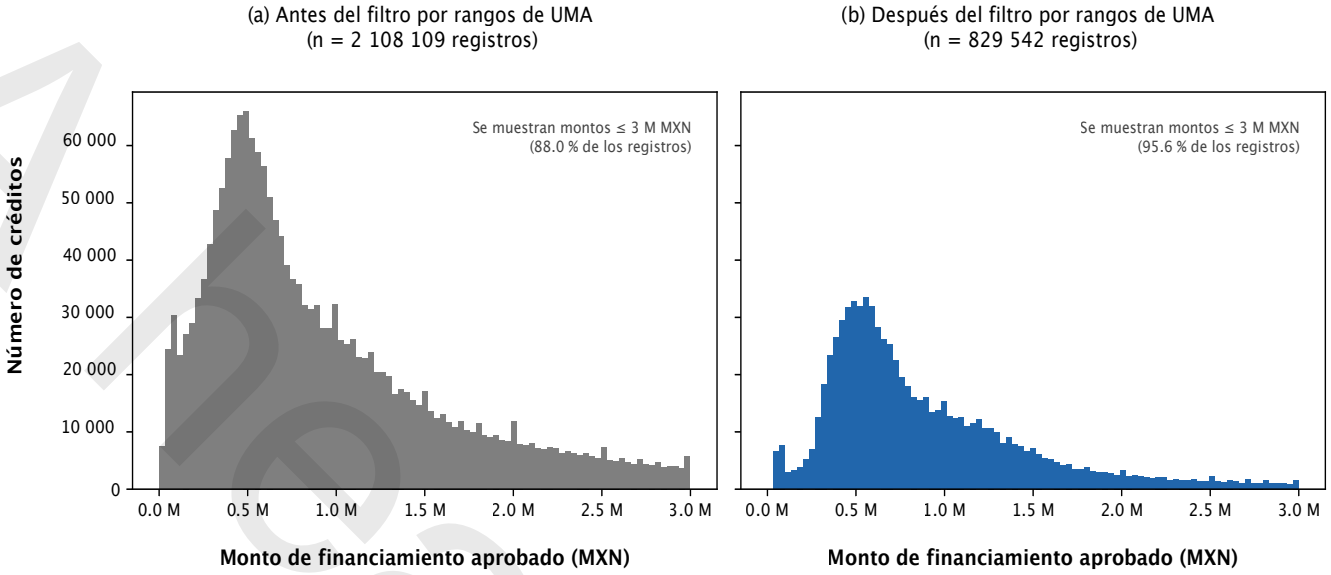
Figura 3. Distribución geográfica de los créditos hipotecarios por municipio (n = 829 542 registros), coloreados por segmento de valor de la vivienda, sobre el contorno del territorio nacional y de las entidades federativas.
Figure 3. Geographic distribution of mortgage loans by municipality (n = 829 542 records), colored by housing value segment, over the outline of the national territory and state boundaries.

Tabla 3. Comparativa de métricas R^2 de entrenamiento y prueba, MAPE y RMSE.
Table 3. Comparison of R^2 (training and test), MAPE and RMSE metrics.

Modelo	R^2 entrenamiento	R^2 prueba	MAPE	RMSE
RF	0.982	0.970	11.48 %	187 250
DT	1.000	0.941	14.15 %	262 104
LR	0.703	0.704	74.95 %	587 520

este modo, RF se considera el modelo más equilibrado de los tres evaluados, al combinar el mayor poder explicativo con la menor variabilidad entre iteraciones, lo cual refuerza la fiabilidad de sus estimaciones para su aplicación a la valoración de créditos hipotecarios.

El análisis de importancia de variables del modelo RF, calculado como la reducción media de impureza del nodo (t) (MSE) atribuible a cada predictor en el conjunto de los 200 árboles (Pedregosa y col., 2011), indicó que el valor de la vivienda concentra el 86.6 % de la importancia total, seguido del año (6.5 %) y del rango de ingresos del acreditado (4.2 %); la clave de municipio, la modalidad y la clave de entidad federativa aportaron menos del 1 % cada una (0.9 %, 0.9 % y 0.8 %, respectivamente). Este resultado es congruente con la matriz de correlación (Figura 2), en la que el



Ambos paneles comparten la misma escala en el eje Y

Figura 4. Distribución del monto de financiamiento aprobado (a) antes y (b) después del filtro por rangos normativos de UMA.
Figure 4. Distribution of the approved financing amount (a) before and (b) after the UMA normative-range filter.

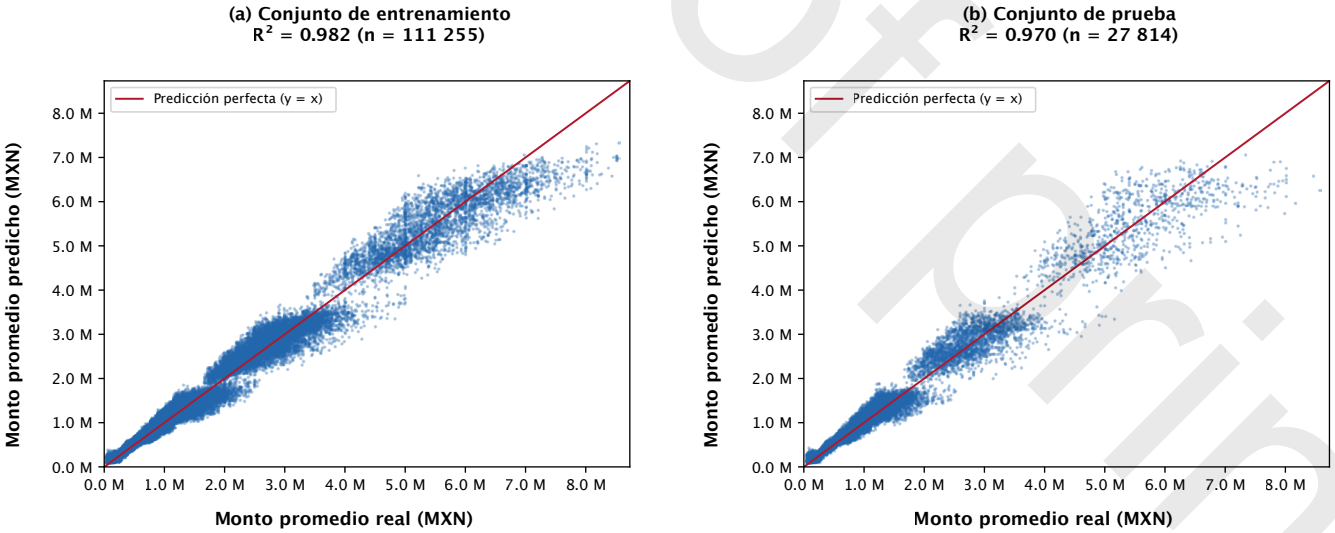


Figura 5. Valores reales frente a predichos del modelo Bosques Aleatorios (RF) en los conjuntos de (a) entrenamiento ($R^2 = 0.982$) y (b) prueba ($R^2 = 0.970$); la línea roja indica la predicción perfecta ($y = x$).
Figure 5. Actual versus predicted values of the Random Forest (RF) model for the (a) training set ($R^2 = 0.982$) and (b) test set ($R^2 = 0.970$); the red line indicates perfect prediction ($y = x$).

Tabla 4. Resultados de la validación cruzada k-fold (k = 5) del R^2 por modelo.
Table 4. K-fold (k = 5) cross-validation results of R^2 per model.

Modelo	R^2 promedio (CV k = 5)	σ	Rango de R^2 por pliegue
RF	0.970	0.000 3	0.970 a 0.971
DT	0.943	0.001 4	0.940 a 0.941
LR	0.703	0.001 1	0.702 a 0.705

σ : Desviación estándar

valor de la vivienda, el año y el rango de ingresos presentaron las asociaciones de mayor magnitud con el monto.

DISCUSIÓN

Los resultados obtenidos demostraron que el algoritmo RF fue consistentemente superior a los modelos de LR y DT para la estimación del monto de financiamiento de vivienda a partir de datos del SNIIV seleccionados.

Su nivel de capacidad explicativa fue alto, con un R^2 de 0.970 sobre el conjunto de prueba independiente, que indica que el modelo explica el 97.0 % de la variabilidad en los montos de financiamiento promedio; y considerando que la base de datos utilizada proviene de registros con codificación categórica ordinal, sin incluir variables físicas continuas propias del bien inmueble como superficie de construcción, número de habitaciones, antigüedad, entre otras, que representan los predictores más frecuentes en otros métodos de valuación como el enfoque comparativo de mercado, el enfoque de costos y el enfoque de ingresos o capitalización de rentas (INDAABIN, 2025; IVSC, 2022), así como los modelos hedónicos de precios (Kok y col., 2017; Valier, 2020).

El desempeño del modelo RF es coherente con lo reportado por Stang y col. (2023), quienes indicaron que los modelos de AA, especialmente los de ensamble como RF, gradient boosting y XGBoost, superan sistemáticamente a la regresión lineal empleada en AVM, sobre todo en situaciones en las que los predictores tienen relaciones no lineales, como las que se presentan entre la localización, la superficie construida, la antigüedad y la calidad constructiva del inmueble; y una gran cantidad de variables por cada registro. Sin embargo, cabe resaltar una importante diferencia metodológica en los estudios revisados por los autores mencionados, como los de Ho y col. (2021), Kumkar y col. (2018) y Singh y col. (2020), en los cuales disponen de variables físicas detalladas (superficie, número de habitaciones, calidad constructiva) que les permiten alcanzar MAPE menores al 15 %. Con el filtrado por rangos normativos de UMA, el MAPE obtenido (11.48 %) resulta comparable al de dichos estudios, aun cuando no se dispone de variables físicas del inmueble. Esto resalta la relevancia de enriquecer las bases de datos públicas disponibles en México con atributos físicos del inmueble.

El MAPE del 11.48 % necesita una interpretación contextualizada. Este nivel de error porcentual absoluto sería moderadamente elevado si se aplicara una valuación directa sobre registros individuales, pero es consistente con la naturaleza del conjunto de datos, ya que los montos del SNIIV tienen una alta variabilidad intrínseca entre los municipios y los segmentos de valor, y el modelo trabaja con promedios agregados por combinación de atributos. Investigaciones comparables realizadas en mercados con un nivel de detalle de datos similar reportaron valores de MAPE dentro de rangos entre el 20 % y 40 % (Matey y col., 2022), por lo tanto, el resultado del presente estudio se encuentra incluso por debajo de esos parámetros.

Asimismo, Hong y Kim (2022) demostraron que la combinación de múltiples AVM mejora la precisión sobre el uso de un modelo individual. Aunque en el presente estudio no se combinaron modelos distintos, sino que RF,

DT y LR se entrenaron y evaluaron de forma independiente, dicho hallazgo es coherente con la ventaja observada del RF, que constituye en sí mismo un ensamble de árboles de decisión, sobre los modelos individuales DT y LR. En el contexto de México implica considerar el marco regulatorio vigente. Como señala Salas-Tafoya (2018), la valuación de bienes inmuebles de propiedad nacional está regulada por el INDAABIN y la de bienes de propiedad particular por la SHF, a través de las "Reglas de carácter general que establecen la metodología para la valuación". En este sentido, es necesario analizar la viabilidad de incluir metodologías de AA dentro de los marcos normativos, no solo como apoyo, sino como herramientas complementarias válidas para el proceso de valuación. De este modo, se permitiría avanzar hacia un modelo híbrido, en el que los métodos normados tradicionales convivan con algoritmos de nueva generación.

Un aspecto adicional es la estabilidad del modelo, ya que la validación cruzada arrojó un R^2 medio de 0.970 +/- 0.000 3 para el modelo RF (Tabla 4), con una diferencia mínima respecto al R^2 del conjunto de prueba independiente (0.970: Tabla 3). Esta concordancia entre esquemas de evaluación, junto con la baja dispersión entre pliegues (0.970 a 0.971: Tabla 4), sugiere que el modelo generaliza adecuadamente a datos no vistos de la misma distribución y que su desempeño es estable ante distintas particiones de los datos.

En el modelo RF, el valor de la vivienda fue el predictor de mayor contribución explicativa (86.6 %), seguido del año (6.5 %) y del rango de ingresos del acreditado (4.2 %). Dicho comportamiento fue congruente con los valores obtenidos en las correlaciones de Spearman de estas variables con el monto (ρ = 0.94, 0.44 y 0.51, respectivamente). Este resultado concuerda con la teoría económica del mercado hipotecario: los montos de préstamo reflejan tanto la capacidad de pago del solicitante como el segmento de mercado al que pertenece la vivienda (Linneman y Wachter, 1989). La importancia del año manifiesta la actualización de los montos a lo largo del período, asociada a la inflación y a los cambios en la política crediticia de los organismos (Ling y col., 1998).

El presente estudio presenta algunas limitaciones que deben considerarse al interpretar los resultados. La variable objetivo (monto) es el monto de financiamiento aprobado por las instituciones del sistema SNIIV, no el valor catastral ni el precio de mercado del inmueble; aunque ambas medidas están correlacionadas, el monto de financiamiento incorpora efectos de la política crediticia de cada organismo y los techos de crédito institucionales, lo que introduce ruido en la estimación del valor puro del inmueble. El modelo no incorpora variables macroeconómicas (tasas de interés hipotecarias, índices de inflación, crecimiento del Producto Interno Bruto estatal) que pueden tener un impacto en el valor de la vivienda. La cobertura del SNIIV está limitada al financiamiento formal (INFONAVIT, FOVISSSTE, banca privada y otros organismos registrados), excluyendo el mercado informal de vivienda que representa una proporción significativa de las transacciones inmobiliarias en México (Tellman y col., 2021). El modelo estima el monto promedio por combinación de predictores y no el monto individual de un crédito; por tanto, no debe aplicarse directamente a registros individuales. Los numerosos registros de cada municipio y la posible relación entre el rango de ingresos del beneficiario y el valor de la vivienda no fueron analizados de manera formal en este estudio. La dimensión geográfica representada por las claves de entidad y municipio aporta información complementaria, identificando las diferencias territoriales en el valor del suelo y los costos de construcción entre las regiones de México sin embargo, en el modelo LR las claves de entidad y municipio se incorporan como valores numéricos, lo que carece de sentido territorial y contribuye a su menor desempeño; en los modelos de árboles este problema se mitiga porque las divisiones sucesivas tratan las claves como agrupaciones de territorios.

También debe considerarse que el doble filtro por rangos normativos de UMA excluyó el 60.6 % de los registros del periodo 2016 a 2025 (696 270 por el límite inferior y 582 297 por el superior) para asegurar la consistencia entre el monto y la clase de vivienda declarada; esta depuración acota la varianza de la variable objetivo y favorece la magnitud de la correlación entre el valor de la vivienda y el monto, por lo que las métricas reportadas corresponden al universo de créditos consistentes con la norma y podrían no generalizarse a los registros excluidos. Los hiperparámetros del modelo RF se fijaron de manera heurística y no mediante una búsqueda sistemática con validación cruzada, por lo que no se garantiza que la configuración empleada sea la óptima. La comparación de modelos se realizó mediante métricas de desempeño sobre un conjunto de prueba único y validación cruzada de cinco pliegues. No se realizaron pruebas de significancia estadística formal entre modelos (p. ej. prueba Friedman o intervalos de confianza bootstrap del RMSE), lo que constituye una limitación metodológica que trabajos futuros deberían abordar.

Por otra parte, los tres factores complementarios que hacen que el RF sea superior al DT y al LR en este conjunto de datos son los siguientes. En primer lugar, los predictores combinan variables genuinamente ordinales (rango de ingresos, segmento de valor de la vivienda) con claves nominales de alta cardinalidad codificadas numéricamente (entidad y municipio); en conjunto generan interacciones no lineales y agrupamientos territoriales que LR no puede capturar mediante una función lineal global, mientras que los modelos de árboles tratan estas claves como categorías de partición. Esta explicación es consistente con una de las limitaciones del estudio: el problema central no es la ordinalidad de las variables en sí, sino la codificación numérica de variables nominales de alta cardinalidad, cuyo efecto se mitiga en los modelos de árboles. En segundo lugar, la estrategia de ensamble de RF, que promedia las predicciones de múltiples árboles entrenados sobre diferentes submuestras, reduce la varianza de las estimaciones con respecto a un único DT, que tiende más al sobreajuste. En tercer lugar, la alta dimensionalidad implícita de las variables geográficas (32 entidades × más de 2 400 municipios) hace que los algoritmos que aprenden mediante divisiones progresivas de los datos, como RF, sean más adecuados que métodos globales como LR. Estos factores explican la brecha de R^2 entre RF y LR, así como el mejor desempeño en RMSE (Tabla 3).

CONCLUSIONES

En el conjunto de prueba independiente (20 %, fuera de muestra), el modelo RF obtuvo el mejor desempeño de los tres algoritmos evaluados. También demostró ser estable y robusto según la validación cruzada de cinco pliegues, con un margen de diferencia mínimo entre su desempeño en el entrenamiento (80 %) y el obtenido en el conjunto de prueba independiente, lo que descarta un sobreajuste significativo. Los resultados mostraron un ajuste y una precisión alta del modelo RF al estimar el monto de financiamiento promedio de vivienda, en términos de varianza explicada según R^2 y de error absoluto según RMSE y MAPE, a partir de las variables seleccionadas de la base de datos del SNIIV. Sin embargo, al considerar las variables que pueden ser incluidas en un avalúo, se hace evidente la importancia de incorporar más factores para que los modelos de AA capten mejor las características que influyen en el valor de una vivienda, generando estimaciones más precisas y, por lo tanto, más confiables. Los modelos de AA y, en particular, el modelo RF, pueden servir como herramientas de validación rápida para detectar montos atípicos y actuar como referencia de consulta durante el proceso de un avalúo formal. Por otro lado, al integrar la dimensión temporal (2016 a 2025) con la clasificación por UMA, el modelo captura la dinámica de precios ajustada por inflación, lo que constituye una alternativa de bajo costo computacional para el monitoreo continuo de las tendencias del mercado hipotecario formal. Cabe aclarar que, en ninguno de estos casos, el modelo sustituye la valua-

ción presencial ni el dictamen de un valuador certificado. Los modelos construidos son herramientas con potencial para complementar los métodos convencionales aplicados a la valuación inmobiliaria, al ser de acceso abierto, reproducibles y actualizables a partir de los datos anuales del SNIIV. En la medida en que las relaciones entre las variables del modelo se mantengan, condición que debería verificarse mediante un reentrenamiento periódico, el sistema tiene potencial para: a) proporcionar estimaciones de referencia del financiamiento hipotecario formal a nivel municipal; (b) apoyar el monitoreo de tendencias del mercado hipotecario formal; y (c) servir como insumo exploratorio en el análisis de políticas habitacionales. No obstante, su uso como instrumento definitivo de valuación requiere una validación adicional fuera de la muestra temporal.

DECLARACIÓN DE CONFLICTO DE INTERESES

Los autores declararon no tener conflictos de intereses de ningún tipo.

REFERENCIAS

Barrientos-Matute, J. P. y Ruvalcaba-Gómez, E. A. (2023). Metodología para explorar el valor público de datos abiertos ofertados en portales web de vivienda. *Economía, Sociedad y Territorio*, 23(72), 405-431. https://doi.org/10.22136/est20232002

Bergstra, J. & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, 281-305.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Wadsworth.

Carranza-Rodríguez, A. M., Carranza-Monzón, D. L. y León-Luyo, S. L. (2024). Aplicación de las escalas de medición ordinal para interpretar coeficientes de la correlación en investigación científica. *Revista Científica Searching De Ciencias Humanas y Sociales*, 5(1), 45-56. https://doi.org/10.46363/searching.v5i1.4

Deppner, J., von-Ahlefeldt-Dehn, B., Beracha, E., & Schaefers, W. (2023). Boosting the Accuracy of Commercial Real Estate Appraisals: An Interpretable Machine Learning Aproach. *Journal of Real Estate Finance and Economics,* 71, 1-38. https://doi.org/10.1007/s11146-023-09944-1

Espinoza-Garza, F., Martínez-Ramírez, Y., Ramírez-Noriega, A. y Álvarez-Sánchez, I. N. (2024). Una revisión sistemática de la literatura sobre la precisión de modelos de aprendizaje automático aplicados a la tasación de bienes raíces. *Revista de Investigación en Tecnologías de la Información*, 12(28), 4-16. https://doi.org/10.36825/RITI.12.28.002

Glumac, B. & Des-Rosiers, F. (2021). Practice briefing – Automated valuation models (AVMs): their role, their advantages and their limitations. *Journal of Property Investment & Finance*, 39(5), 481-491.

Google (2025). *Google Colab*. [En línea]. Disponible en: https://colab.research.google.com/. Fecha de consulta: 8 de junio de 2025.

Gunes, T. (2023). Model agnostic interpretable machine learning for residential property valuation. *Survey Review*, 56(399), 525-540. https://doi.org/10.1080/00396265.2023.2293366

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (Second edition). Springer.

Hernández-Sampieri, R. (2014). Metodología de la investigación (Sexta edición). Mc Graw-Hill.

Ho, W. K. O., Tang, B. S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research*, 38(1), 48–70. https://doi.org/10.1080/09599916.2020.1832558

Hong, J. & Kim, W. S. (2022). Combination of machine learning-based automatic valuation models for residential properties in south korea. *International Journal of Strategic Property Management*, 26(5), 362-384. https://doiorg/10.3846/ijspm.2022.17909

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90-95. https://doi.org/10.1109/MCS

E.2007.55

INDAABIN, Instituto de Administración y Avalúos de Bienes Nacionales (2025). Metodologías de los servicios valuatorios regulados por INDAABIN. [En línea]. Disponible en: https://www.gob.mx/indaabin/documentos/metodologias-de-caracter-tecnico-24208. Fecha de consulta: 17 de mayo de 2025.

INEGI, Instituto Nacional de Estadística y Geografía (2025a). Unidad de Medida y Actualización (UMA). [En línea]. Disponible en: https://www.inegi.org.mx/temas/uma/. Fecha de consulta: 27 de febrero de 2025.

INEGI, Instituto Nacional de Estadística y Geografía (2025b). Catálogo Único de Claves de Áreas Geoestadísticas Estatales, Municipales y Localidades. [En línea]. Disponible en: https://www.inegi.org.mx/app/ageeml/. Fecha de consulta: 17 de enero de 2025.

IVSC, International Valuation Standards Council (2022). Normas Internacionales de Valuación. [En línea]. Disponible en: https://ivsc.org/. Fecha de consulta: 16 de septiembre de 2025.

Jafary, P., Shojaei, D., Rajabifard, A., & Ngo, T. (2024). Automated land valuation models: A comparative study of four machine learning and deep learning methods based on a comprehensive range of influential factors. *Cities*, 151, 105115. https://doi.org/10.1016/j.cities.2024.105115

Jordan, M. I. & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255-260. https://doi.org/10.1126/science.aaa8415

Jung, J., Kim, J., & Jin, C. (2022). Does machine learning prediction dampen the information asymmetry for non-local investors? *International Journal of Strategic Property Management*, 26(5), 345-361. https://doi.org/10.3846/ijspm.2022.17590

Kok, N., Koponen, E. L., & Martínez-Barbosa, C. A. (2017). Big data in real estate? From manual appraisal to automated valuation. *Journal of Portfolio Management*, 43(6), 202-211. https://doi.org/10.3905/jpm.2017.43.6.202

Kumkar, P., Madan, I., Kale, A., Khanvilkar, O., & Khan, A. (2018). Comparison of Ensemble Methods for Real Estate Appraisal. [En línea]. Disponible en: 10.1109/ICICT43934.2018.9034449. Fecha de consulta: 19 de enero de 2025.

Ling, D. C. & McGill, G. A. (1998). Evidence on the demand for mortgage debt by owner-occupants. *Journal of Urban Economics*, 44(3), 391-414.

Linneman, P. & Wachter, S. (1989). The impacts of borrowing constraints on homeownership. *Real Estate Economics*, 17(4), 389-402. https://doi.org/10.1111/1540-6229.00499

Matey, V., Chauhan, N., Mahale, A., Bhistannavar, V., & Shitole, A. (2022). Real Estate Price Prediction using Supervised Learning. *2022 IEEE Pune Section International Conference, PuneCon 2022*. https://doi.org/10.1109/PuneCon55413.2022.10014818

Maulud, D. & Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, 1(2), 140-147. https://doi.org/10.38094/jastt1457

McKinney, W. (2010). Data structures for statistical computing in Python. In S. van-der-Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 51-56). https://doi.org/10.25080/Majora-92bf1922-00a

Mody, P., Motiramani, M., & Singh, A. (2023). Enhancing Real Estate Market Insights through Machine Learning: Predicting Property Prices with Advanced Data Analytics. 2023 4th IEEE Global Conference for Advancement in Technology, *GCAT 2023*. https://doi.org/10.1109/GCAT59970.2023.10353243

Nazarov, F. M. & Yarmatov, S. (2023). Optimization of Prediction Results Based on Ensemble Methods of Machine Learning. *Proceedings - 2023 International Russian Smart Industry Conference, SmartIndustryCon* 2023, 181-185. https://doi.org/10.1109/SmartIndustryCon57312.2023.10110726

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). *Scikit-learn: Machine learning in Python. Journal of Machine Learning Research*, 12, 2825-2830.

Restrepo, L. F. & González, J. (2007). De Pearson a Spearman. *Revista Colombiana de Ciencias Pecuarias*, 20(2), 183-192.

Salas-Tafoya, J. M. (2018). Transdisciplinariedad valuatoria. Hacia una construcción sistémica para la valuación inmobiliaria/Valuable transdisciplinarity. Towards a systemic construction for real estate valuation. *RICEA Revista Iberoamericana de Contaduría, Economía y Administración*, 6(12), 348-369. https://doi.org/10.23913/ricea.v6i12.109

Singh, A., Sharma, A., & Dubey, G. (2020). System Assurance Engineering and Management. [En línea]. Disponible en: https://nam04.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdoi.org%2F10.1007%2Fs13198-020-00946-3&data=05%7C02%7Cienciauat%40uat.edu.mx%7C122c5e065f544d8b4a2108df14df5799%7C725ab307b77c4f669168376b1c7f9990%7C0%7C0%7C639252621672908005%7CUnknown%7CTWFpbGZsb3d8eyJFbXB0eU1hcGkiOnRydWUsIlYiOiIwLjAuMDAwMCIsIlAiOiJXaW4zMiIsIkFOIjoiTWFpbCIsIldUIjoyfQ%3D%3D%7C0%7C%7C%7C&sdata=YozkQviF84tkQrSz%2FMzOaeKfzIZGZRkG9w8B6%2FVIFho%3D&reserved=0. Fecha de consulta: 19 de enero de 2025.

SNIIV, Sistema Nacional de Información e Indicadores de vivienda (2025). Datos abiertos de financiamiento en México. [En línea]. Disponible en: https://sniiv.sedatu.gob.mx/Reporte/Datos_abiertos. Fecha de consulta: 31 de agosto de 2026.

Stang, M., Krämer, B., Nagl, C., & Schäfers, W. (2023). From human business to machine learning—methods for automating real estate appraisals and their practical implications. *Zeitschrift Für Immobilienökonomie*, 9(2), 81-108. https://doi.org/10.1365/s41056-022-00063-1

Tellman, B., Eakin, H., Janssen, M. A., de-Alba, F., & Turner, B. L. (2021). The role of institutional entrepreneurs and informal land transactions in Mexico City's urban expansion. *World Development*, 140, 105374. https://doi.org/ 10.1016/j.worlddev.2020.105374

Valier, A. (2020). Who performs better? AVMs versus hedonic models. *Journal of Property Investment & Finance*, 38(3), 213-225.

Verma, P. K., Arya, S., & Asbe, C. (2023). Predicting future housing prices: a machine learning approach. *Multidisciplinary Science Journal*, 5, 2023ss0206. https://doi.org/10.31893/multiscience.2023ss0206