Predicción de avalúos catastrales en el municipio de Dagua-Valle del Cauca utilizando modelos de aprendizaje automático.

Stefania Hurtado González

Juan José Marín

Pontificia Universidad Javeriana Cali
Facultad de Ingeniería
Ingeniería de Sistemas y Computación
Proyecto de Grado
Cali, 2025

Pontificia Universidad Javeriana Cali
Facultad de Ingeniería
Ingeniería de Sistemas y Computación
Proyecto de Grado

Predicción de avalúos catastrales en el municipio de Dagua-Valle del Cauca utilizando modelos de aprendizaje automático.

Stefania Hurtado González
Juan José Marín

*d*irectores:

Dra. Gloria Inés Álvarez
Dr. Diego Linares

Cali, 2025

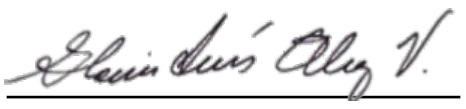
Santiago de Cali, Enero 19 de 2025

Señores
Pontificia Universidad Javeriana Cali.
Dr. Gerardo M. Sarria
Director Carrera de Ingeniería de Sistemas y Computación.
Cali.

Cordial Saludo

Por medio de la presente me permito informarle que los estudiantes de Ingeniería de Sistemas y Computación Stefania Hurtado González (cod: 8961789) y Juan José Marin (cod: 8947785) trabajan bajo mi dirección en el proyecto de grado titulado "Predicción de avalúos catastrales en el municipio de Dagua-Valle del Cauca utilizando modelos de aprendizaje automático.".

Atentamente,



Dra. Gloria Inés Álvarez

Dr. Diego Linares

Santiago de Cali, Enero 19 de 2025

Señores
Pontificia Universidad Javeriana Cali.
Dr. Gerardo M. Sarria
Director Carrera de Ingeniería de Sistemas y Computación.
Cali.

Cordial Saludo

Nos permitimos presentar a su consideración el proyecto de grado titulado "Predicción de avalúos catastrales en el municipio de Dagua-Valle del Cauca utilizando modelos de aprendizaje automático." con el fin de cumplir con los requisitos exigidos por la Universidad para llevar a cabo el proyecto de grado y posteriormente optar al título de Ingeniero de Sistemas y Computación.

Al firmar aquí, damos fe que entendemos y conocemos las directrices para la presentación de trabajos de grado de la Facultad de Ingeniería probadas el 26 de Noviembre de 2009, donde se establecen los plazos y normas para el desarrollo del anteproyecto y del trabajo de grado

Atentamente,

Stefania Hurtado González
Código: 8961789

Juan José Marin
Código: 8947785

Summary

The cadastral appraisal is a key element in territorial management, used to calculate property taxes and provide information on property values. In the municipality of Dagua, Valle del Cauca, the cadastral valuation process faces challenges related to the lack of data systematization and reliance on manual methods, which limits its accuracy and updates.

This study develops a predictive model based on machine learning techniques to improve the estimation of cadastral appraisals. During the research, eight supervised models were evaluated, including Random Forest, K-Nearest Neighbors, Gradient Boosting Machines, and Neural Networks. Random Forest was selected for its outstanding performance, achieving a determination coefficient (R^2) of 87.15 % and lower error metrics compared to the others.

The methodological development included data preparation, selection of relevant features such as land area, geoeconomic zone, and block/neighborhood, and hyperparameter optimization. Additionally, a graphical user interface was implemented for specialized users, such as cadastral management officials or technicians, providing access to the necessary information. This interface requests specific data required to make accurate predictions, such as the physical characteristics of the property and its location, ensuring that the estimates are useful and relevant in a professional context.

The results show that integrating machine learning into cadastral valuation can improve the accuracy and efficiency of the process. Specifically, the Random Forest model stood out for its performance compared to other evaluated models. With a mean absolute error (MAE) of 14.5M, the model achieved a determination coefficient (R^2) of 87.15 %, indicating a high capacity to explain the variability of cadastral values in the test set, providing a practical and replicable approach in other territorial contexts.

Keywords: Avalúo catastral, aprendizaje automático, Random Forest,Regresión Lineal, Redes neuronales, K-Nearest Neighbors, Gradient Boosting Machines, predicción, gestión territorial.

Índice

1. Descripción del Problema

1.1. Planteamiento del Problema

Dagua, un municipio ubicado en la región del Valle del Cauca en Colombia, cuenta con una población de 39.665 habitantes y enfrenta un desafío crítico relacionado con la precisión y eficiencia en la valoración catastral. Esta problemática impacta tanto a los propietarios de bienes inmuebles en la localidad como a las entidades responsables de gestionar la tributación predial y planificar el desarrollo urbano. La falta de exactitud en la valoración catastral puede ocasionar discrepancias en los impuestos prediales y una distribución desigual de la carga fiscal.

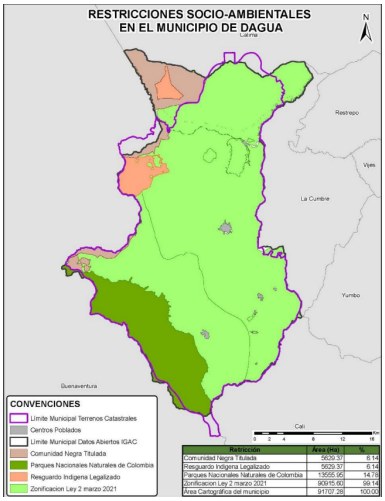


Figura 1: Cartografía de Dagua

En Colombia, el Instituto Geográfico Agustín Codazzi (IGAC) es la entidad encargada de la recolección de información cartográfica, geodésica y geográfica, conforme a las disposiciones establecidas por esta institución [1]. Además, la Unidad administrativa especial de catastro (UAEC) es responsable de la valoración catastral, actualización y la conservación del catastro multipropósito. En el municipio de Dagua, perteneciente al departamento del Valle del Cauca, el proceso de evaluación del avalúo catastral se lleva a cabo a través de una serie de etapas estructuradas, orientadas a precisar el valor de los predios en cuestión.

En Dagua, donde el catastro se implementó en el año 2006 y cuenta con 24,774 predios

registrados en el 2023 [2], se enfrentan limitaciones significativas en las metodologías actuales para resolver el desafío de la valoración catastral. Estas limitaciones están influenciadas por diversos factores, como la falta de sistematización de datos, la carencia de un enfoque basado en tecnología vanguardista y la ineficacia en la identificación de zonas homogéneas geoeconómicas. Además, la escasez de herramientas inteligentes y la insuficiente implementación de políticas públicas agravan la imprecisión en las valoraciones catastrales. La falta de instrumentos de planificación urbana adecuados añade una capa adicional de complejidad debido a que son esenciales para ordenar el uso del suelo, regular las condiciones de su transformación y conservación e incluso logran promover un desarrollo sostenible, denotando así una dificultad aún mayor de mantener actualizada y precisa la valoración catastral.

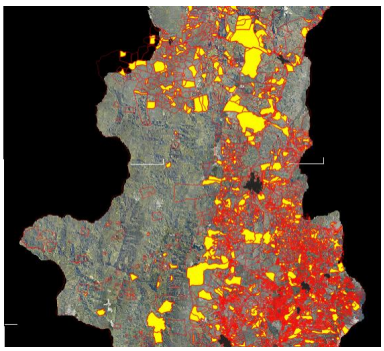


Figura 2: Predios urbanos y rurales, Municipio de Dagua

Como primera instancia, se emprende la fase de identificación predial, donde se recolecta información física y legal de los terrenos, abarcando elementos como dimensiones y construcciones presentes en dicha localidad. No obstante, la ausencia de una sistematización adecuada de estos datos abarca un desafío significativo que marca la inexactitud y mala gestión del proceso.

Posteriormente, se procede con la delimitación de las zonas homogéneas geoeconómicas, agrupando predios con similitudes particulares. Subsecuentemente, se establecen los valores unitarios correspondientes a las diversas categorías de construcciones, teniendo en cuenta una serie de variables tales como materiales utilizados, estándares de calidad, contexto urbanístico y características arquitectónicas, entre otros.

El aprendizaje automático se destaca como una herramienta valiosa para abordar estos desafíos. Su principal beneficio es su habilidad para realizar una clasificación más precisa y eficaz de las zonas geoeconómicas homogéneas, lo que conduce a una valoración

catastral más acertada de los bienes inmuebles. Al integrar tecnologías avanzadas como los Sistemas de Información Geográfica (SIG) y la Inteligencia Artificial, se mejora la recopilación, organización y análisis de datos. Esto perfecciona la determinación de los valores unitarios en áreas específicas, resultando en una evaluación catastral más precisa y transparente.

Como última instancia del proceso, este involucra la liquidación de los avalúos catastrales, no obstante, este paso se encuentra sujeto a un margen de error considerable debido a los factores mencionados anteriormente. Esta problemática se manifiesta con particular relevancia en el municipio de Dagua, específicamente en relación con la última actualización reflejada catastralmente realizada en el año 2023.

La carencia de un enfoque tecnológicamente avanzado obstaculiza una clasificación eficiente y precisa de las zonas mencionadas, comprometiendo la fiabilidad de los valores catastrales. La deficiente implementación de la política pública y la falta de herramientas inteligentes pueden ocasionar errores significativos en los avalúos.

Sin embargo, la integración de tecnologías innovadoras como los sistemas de información geográfica (SIG), la inteligencia artificial y el aprendizaje automático brinda perspectivas alentadoras. Estas herramientas tienen el potencial de agilizar el proceso de recolección y organización de datos, así como de mejorar la determinación de valores unitarios en áreas específicas [3], contribuyendo así a una evaluación más precisa de los bienes inmuebles en el municipio de Dagua.

En definitiva, estas soluciones tecnológicas ofrecen la oportunidad de optimizar el proceso de valoración y mitigar errores, impulsando una gestión transparente y eficiente de la propiedad en la localidad. Asimismo, la introducción de herramientas técnicas destinadas a monitorear las fluctuaciones del mercado inmobiliario y ajustar los valores unitarios, promueve una distribución más equitativa de la carga fiscal en el ámbito nacional.

1.1.1. Formulación

¿Cómo predecir el valor del avalúo catastral para un predio urbano y rural en el municipio de Dagua usando técnicas de aprendizaje automático?

1.1.2. Sistematización

- ¿Cómo preparar un dataset para aplicar un modelo de aprendizaje automático?

- Qué modelos de aprendizaje automático se utilizan, junto con sus métodos de entrenamiento correspondientes, para predecir el avalúo catastral en el municipio de Dagua
- ¿Cómo optimizar el desempeño de los modelos?
- ¿Cómo evaluar el desempeño de los modelos construidos?

1.2. Objetivos

1.2.1. Objetivo General

Desarrollar un modelo que permita predecir el avalúo catastral del municipio de Dagua Valle del Cauca, utilizando técnicas de aprendizaje automatico.

1.2.2. Objetivos Específicos

- Preparar la base de datos catastral del municipio de Dagua utilizando técnicas de exploración de datos.
- Entrenar varios modelos de aprendizaje automatico apropiados para predecir el avalúo catastral en el municipio de Dagua.
- Optimizar los modelos de predicción entrenados.
- Evaluar la eficacia de los modelos construidos.

1.3. Justificación

La implementación de un modelo basado en técnicas de aprendizaje automático para predecir el avalúo catastral en el municipio de Dagua, Valle del Cauca, se presenta como una solución innovadora. El uso del aprendizaje automático es fundamental debido a su capacidad para manejar grandes volúmenes de datos y descubrir patrones complejos que pueden no ser tan evidentes. Este modelo de predicción permitirá organizar y analizar la información catastral disponible, mejorando la precisión del proceso. Además, permitirá identificar y analizar patrones y relaciones significativas entre las variables catastrales

específicas de esta localidad, lo que conducirá a una valuación más precisa y justa de los predios junto con su actualización.

No obstante, es importante reconocer el papel fundamental de los reconocedores catastrales en este proceso. [4]Estos profesionales son esenciales para proporcionar datos precisos y actualizados sobre las propiedades, que son la base para cualquier modelo predictivo eficaz. La labor de los reconocedores catastrales, al inspeccionar y documentar las características físicas y legales de los predios, garantiza que el modelo de aprendizaje automático cuente con información verificada y confiable. Su trabajo en terreno complementa el análisis automatizado, asegurando que las predicciones del avalúo catastral reflejen fielmente la realidad de las propiedades en Dagua.

Asimismo, el modelo de predicción mejora la actualización del proceso. Es importante contar con la información actualizada y precisa para la toma de decisiones en la planificación territorial y el desarrollo urbano y rural de Dagua, ya que se debe contar con una base de datos sólida para la correcta gestión transparente de la propiedad en el municipio. Esta solución beneficiará tanto a los propietarios de los predios, ya que podrán obtener una valoración más justa y precisa, como a los actores gubernamentales y privados involucrados en el mercado inmobiliario local, los cuales tomarán decisiones más informadas basadas en datos precisos y actualizados, esto es fundamental para la correcta implementación y los resultados efectivos del modelo predictivo utilizado en este caso, ya que permite que el modelo de predicción aprenda de tendencias hasta cambios recientes en el territorio, específicamente desde el año 2020, mejorando así su capacidad predictiva y su relevancia para la planificación y gestión y desarrollo de Dagua, Valle del Cauca. La adopción del catastro multipropósito y la base normativa [5] permitirá una representación más precisa de la realidad territorial, incluyendo la informalidad detectada [6] y la discrepancia de datos entre el DANE y la base del IGAC. Este enfoque integral y actualizado es esencial para una actualización catastral efectiva, como lo establece la resolución 70 de 2011 [7], y para la consecución de una gestión catastral que respalde el desarrollo sostenible de Dagua.[8]

La viabilidad de esta solución se basa en la disponibilidad de datos catastrales y la capacidad del aprendizaje automático para manejar y analizar estos datos. Además, nuestra contribución personal a esta solución será la aplicación de nuestros conocimientos y habilidades en aprendizaje supervisado, para así, desarrollar, optimizar modelos predictivos y la interpretación de los resultados para lograr informar las decisiones de valoración catastral.

Por lo tanto, la adopción de un sistema de predicción del avalúo catastral utilizando tecnicas de aprendizaje automático en el municipio de Dagua representaría una oportunidad estratégica para mejorar la precisión, eficiencia y escalabilidad del proceso de valoración catastral, beneficiando tanto a los propietarios, como a los actores gubernamentales y privados involucrados en el mercado inmobiliario local.

1.4. Delimitaciones y Alcances

Para el enfoque del proyecto:

- Se proponen técnicas de preparación y exploración de datos adecuadas para distinguir las correlaciones entre las variables que se encuentran en la base de datos catastral.
- Se toma la decisión de utilizar un modelo de aprendizaje automatico en específico con el objetivo de predecir el avalúo catastral.
- Mediante la implementación de estrategias fundamentales de optimización, se busca mejorar el modelo de predicción de avalúo catastral en el municipio de Dagua.
- El modelo se entrena seleccionando las variables en la base de datos del IGAC (Instituto Geográfico Agustín Codazzi) enfocado solo en los datos referentes al municipio de Dagua.
- La evaluación del desempeño de los modelos seleccionados de aprendizaje automatico desarrollados, se realizará aplicando un conjunto de métricas y técnicas (error cuadrático medio, error absoluto medio, coeficiente de determinación, proporción precision, recall, F1-score, validación cruzada, grid search o perceptrón). Este proceso es esencial para determinar la precisión, capacidad y confiabilidad de los modelos en el contexto del proyecto.
- El modelo tendrá un alcance limitado ya que tomará en cuenta únicamente los datos disponibles del año 2020 hasta Enero del 2024, correspondientes a la base de datos del IGAC específicamente los datos del municipio de Dagua, lo que puede limitar el alcance del análisis a la información existente.

2. Marco de referencia

2.1. Áreas Temáticas

De acuerdo al Sistema de Clasificación de Computación ACM (CCS), las áreas temáticas que abarca el proyecto son:

- Computing methodologies - Machine learning - Neural networks
- Computing methodologies - Artificial intelligence - Natural language processing - Machine translation
- Information systems - Data management systems
- Social and professional topics - User characteristics - People with disabilities
- Information systems - Information systems applications
- Human-centered computing

2.2. Marco Teórico

El marco teórico establece los fundamentos conceptuales y técnicos necesarios para la predicción de avalúos catastrales en el municipio de Dagua, Valle del Cauca. Este capítulo se centra en definir y analizar los conceptos esenciales que sustentan el estudio, proporcionando claridad sobre los procesos y herramientas utilizadas. Entre los temas clave a desarrollar se encuentran el avalúo catastral, la gestión de datos territoriales y las metodologías de aprendizaje automático que permiten mejorar la precisión y eficiencia en la valoración de predios.

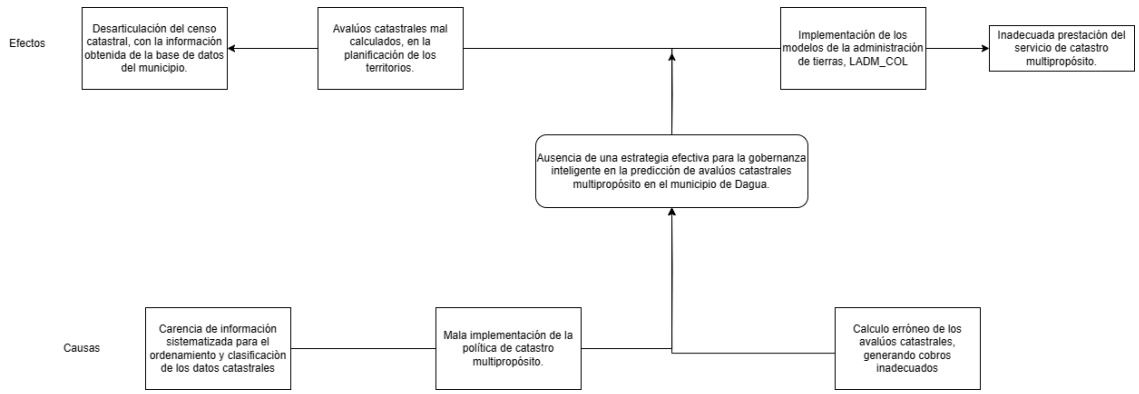


Figura 3: Arbol de problemas

Para abordar el problema planteado, es crucial comprender las bases del avalúo catastral como herramienta de gestión territorial y fiscal. Este proceso, crucial en la sociedad contemporánea, evidencia la importancia del control fiscal al reflejar el aumento del capital de ciertos individuos en el territorio [9]. Además, resulta esencial explorar el uso de modelos avanzados de aprendizaje automático. Estos incluyen técnicas predictivas que procesan grandes volúmenes de datos para generar resultados precisos y confiables. También se enfatizará en el rol de los datos territoriales y su adecuada gestión. Estos conceptos serán desarrollados en detalle a lo largo del capítulo, sentando las bases para la implementación de las estrategias propuestas.

2.2.1. Avalúo catastral

Es un procedimiento riguroso y sistemático realizado por la autoridad catastral para estimar el valor oficial de inmuebles, incluyendo terrenos y edificaciones. Este valor se calcula con base en un análisis detallado y estadístico del mercado inmobiliario local, tomando como referencia los valores comerciales, pero sin superarlos en ningún caso [10].

Para determinar el avalúo catastral de cada predio, se agregan los valores asignados por separado al terreno y a las edificaciones existentes. Su cálculo se expresa en la fórmula:

$$\text{Avalúo Catastral} = \text{Valor m}^2C \times \text{m}^2C + \text{Valor m}^2T \times \text{m}^2T$$

- m^2_C: Área construida en metros cuadrados.

- m_T^2: Área del terreno en metros cuadrados.
- Valor m_C^2: Valor por metro cuadrado de la construcción.
- Valor m_T^2: Valor por metro cuadrado del terreno.

Adicionalmente, con esta informacion, se puede calcular el valor integral de la construcción utilizando la fórmula:

$$\text{Valor m}^2 \text{ Integral de la Construcción} = \frac{\text{Avalúo Catastral}}{\text{m}^2 \text{ de la Construcción}}$$

Etapas claves del avalúo catastral

- Identificación de los predios: Incluye la recopilación de datos físicos, jurídicos y económicos mediante herramientas tecnológicas y georreferenciación.

Clasificación en zonas homogéneas

- Zonas Homogéneas Físicas (ZHF): Áreas con características similares en topografía, infraestructura vial, servicios públicos, uso del suelo y tipo de edificaciones.

- Zonas Homogéneas Geoeconómicas (ZHG): Representan áreas con valores unitarios consistentes según el mercado inmobiliario especificamente del municipio de Dagua, Valle del Cauca.

- Asignación de valores unitarios: Basada en análisis de mercado y características específicas del predio.

- Liquidación del avalúo: Cálculo final utilizando las fórmulas mencionadas anteriormente en la sección de avalúo catastral.

Metodología de avalúos catastrales

Aqui se mencionan las variables clave para la determinacion del avalúo catastral.

- Clase del suelo: Son las actividades permitidas según la normatividad de la la región.

- Uso actual del suelo: Clasificación según destinaciones residenciales, comerciales, industriales, entre otras.

- Tipología de construcciones: Calificación basada en materiales, diseño y estado de conservación.

- Levantamiento de información: Recolección en la zona mediante métodos directos e indirectos.

2.2.2. Número predial nacional

El Número Predial Nacional (NPN) se encuentra diseñado bajo una estructura compuesta por 12 categorías que exactamente cuenta con treinta dígitos, cuya finalidad es sistematizar y detallar la información asociada a cada predio de manera lógica y jerárquica. Estas categorías integran aspectos fundamentales del predio, incluyendo su ubicación geográfica, características físicas, atributos constructivos y condiciones jurídicas. Esta configuración garantiza una identificación única y precisa del inmueble, optimizando su gestión, análisis y administración a nivel catastral y territorial [11].

Por ejemplo, las primeras categorías, como el departamento, municipio, zona y sector, establecen la ubicación exacta del predio dentro del territorio nacional. A continuación, elementos como el barrio, la manzana o vereda, y el terreno, definen su delimitación específica dentro de una comunidad o zona, ya sea rural o urbana. Finalmente, las categorías asociadas a la condición de propiedad y a la construcción detallan aspectos concretos relacionados con el uso, la estructura y la naturaleza jurídica del predio.

Figura 4: Número predial nacional

Esta estructura trasciende su función administrativa, convirtiéndose en una herramienta clave para optimizar procesos como la planificación territorial, la valoración catastral, la recaudación de impuestos y la toma de decisiones estratégicas. La segmentación de la información en categorías permite un acceso preciso y detallado a los datos de cada predio, lo cual es esencial para entidades públicas, propietarios, empresas y cualquier organización dedicada a la gestión del territorio o al análisis de datos catastrales.

2.2.3. Aprendizaje automático

El aprendizaje automático es una disciplina de la inteligencia artificial que se dedica a la creación de sistemas capaces de aprender y mejorar su desempeño a partir de la experiencia, sin ser programados de manera explícita para cada tarea. Esta área investiga y desarrolla algoritmos que pueden procesar datos y, a través de ellos, perfeccionar sus habilidades. De esta forma, los sistemas pueden realizar predicciones y tomar decisiones con mayor precisión de una forma independiente. Aunque el enfoque principal de este trabajo recae en el aprendizaje supervisado, que utiliza datos etiquetados para entrenar modelos predictivos, es importante mencionar los otros dos enfoques principales: el aprendizaje no supervisado, que identifica estructuras o patrones en datos no clasificados, y el aprendizaje por refuerzo, que optimiza el desempeño de los modelos mediante un sistema de recompensas y penalizaciones basado en la interacción con el entorno. Lo anterior nos da lugar a un concepto que se manejará a lo largo de la investigación:

- **Aprendizaje supervisado:** En este enfoque, los modelos se entrenan con datos previamente etiquetados y clasificados, lo que les permite aprender a predecir resultados basándose en esos ejemplos.

  Los métodos de aprendizaje supervisado se dividen en dos categorías principales, dependiendo del propósito de la tarea:

  1. **Clasificación:** Este enfoque se utiliza para asignar cada dato a una clase o categoría específica. Por ejemplo, en el ámbito del avalúo catastral, podría implicar categorizar predios según su uso, como residencial", çomercial.º ïndustrial".
  2. **Regresión:** En este método, el objetivo es prever valores continuos a partir de datos de entrada. Un caso específico sería calcular el valor de un predio tomando en cuenta factores como su ubicación geográfica, extensión del terreno o características de construcción.

2.2.4. Métricas Utilizadas

- **R^2 (Coeficiente de Determinación):** Es el indicador de que nos muestra que tan bien el modelo se ajusta los datos.

  Formula

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

- **RMSE (Root Mean Squared Error):** Es una métrica que penaliza los errores grandes, lo que lo hace bastante sensible a valores o datos atípicos.

  Formula

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

- **MAE(Mean Absolute Error):** Es el promedio de los errores absolutos entre los valores predichos y los valores reales, esta metrica no es sensible a datos atípicos.

  Formula

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

2.2.5. Métodos de Regresión

La regresión es una técnica utilizada en el aprendizaje supervisado para modelar relaciones entre variables independientes y una variable dependiente continua. En el ámbito de la predicción de avalúos catastrales, esta metodología permite generar estimaciones basadas en datos históricos y características específicas de cada propiedad, ofreciendo resultados consistentes y útiles para el análisis.

Para este proyecto, se seleccionaron y emplearon los siguientes algoritmos de regresión, cada uno con características particulares que los hacen adecuados para esta investigación:

1. Regresión lineal
2. Regresión multiple de 3 grados
3. Gradient boosting machines
4. Random Forest
5. Ensamble (Random Forest - Gradient Boosting)
6. K-Nearest Neighbors (KNN)
7. Maquina de soporte vectorial

2.2.6. Regresión lineal

La regresión lineal es una técnica utilizada para representar y analizar cómo una variable dependiente se relaciona con una o más variables independientes, aplicándose tanto en el estudio de datos experimentales como en la predicción de valores futuros [12].En su esencia, la regresión lineal supone que los cambios en una variable afectan de manera lineal a la otras, permitiendo cuantificar cómo los cambios en una variable afectan a otra de manera proporcional. Bajo este contexto, la línea de regresión, también conocida como línea de mejor ajuste, desempeña un papel fundamental [13]. Esta línea representa la relación matemática entre las variables y se ajusta para minimizar la discrepancia entre los valores proyectados por el modelo y los valores observados en

los datos reales. La precisión de esta línea de ajuste es esencial para comprender la tendencia general de los datos y evaluar la calidad del modelo de regresión.

En este trabajo, además de la regresión lineal, se ha utilizado la regresión lineal múltiple, que permite modelar relaciones más complejas al incluir múltiples variables independientes en el análisis. Cada tipo de regresión será descrito y analizado en detalle en las siguientes secciones, explorando su relevancia y aplicación en el contexto de la predicción de avalúos catastrales.

Por otro lado, pero no menos importante, la regresión lineal simple puede expresarse como:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Donde:

- y: Variable dependiente.
- x: Variable predictora.
- β_0: Término constante.
- β_1: Coeficiente de regresión que indica el cambio en y por cada unidad de cambio en x.
- ϵ: Término de error que representa la variación no explicada por el modelo.

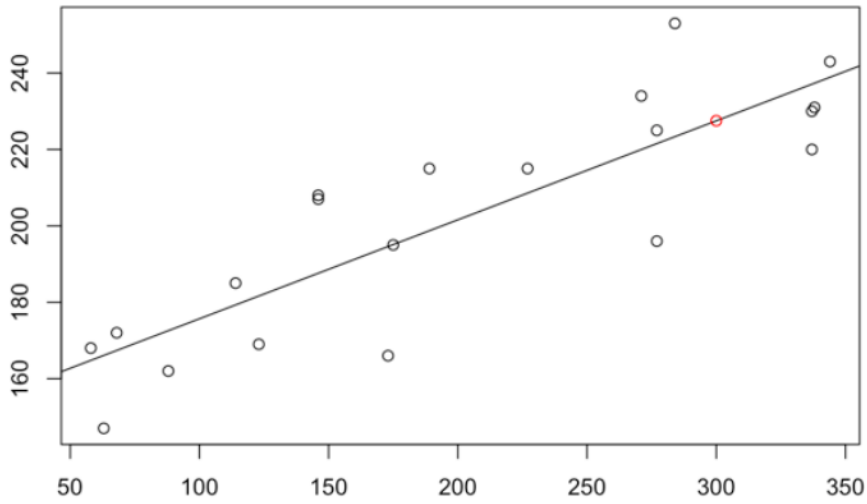


Figura 5: Regresión lineal [14]

Como se evidencia en la gráfica de predicciones 5, se observa una relación lineal entre las variables predichas y los valores reales. Los puntos representan las observaciones individuales, mientras que la línea recta corresponde al ajuste del modelo de regresión lineal, estas expresiones permiten modelar y predecir con precisión la relación entre las variables y son fundamentales en el análisis de datos para la predicción de avalúos catastrales.

2.2.7. Regresión lineal múltiple tres grados

La regresión lineal múltiple de tercer grado es una herramienta estadística diseñada para analizar relaciones complejas entre una variable dependiente y varias independientes, al incluir términos lineales, cuadrados y cúbicos en su modelo. En el ámbito del avalúo catastral, esta técnica resulta clave para identificar cómo factores como el tamaño del terreno, la ubicación geoeconómica o las características físicas del predio influyen en su valor. Al integrar términos polinomiales, el modelo puede capturar patrones no lineales que reflejan de manera más precisa la realidad [15].

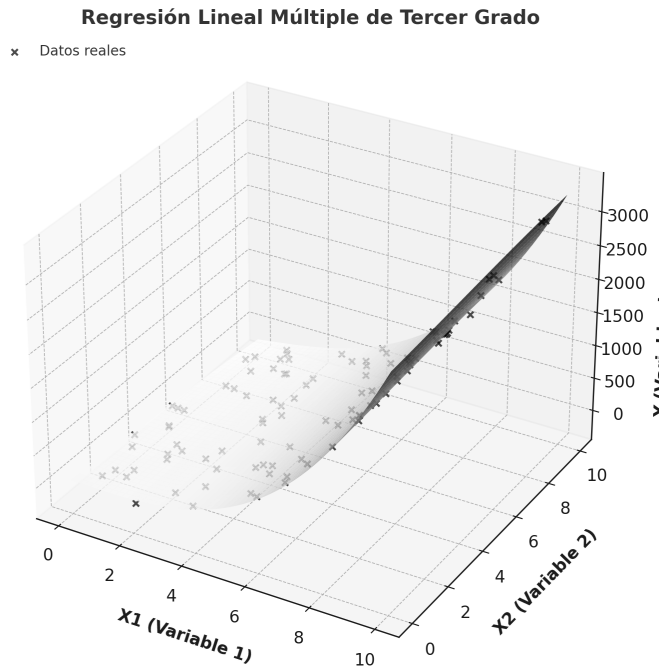


Figura 6: Regresión lineal multiple tres grados [14]

En la anterior gráfica 6 se presenta una regresión lineal múltiple de tercer grado, donde las variables independientes X_1 y X_2 se utilizan para predecir los valores de la variable dependiente Y. Los puntos marcados representan los datos reales, mientras que la superficie curva muestra el modelo ajustado. Como se observa, la relación entre las variables sigue un patrón no lineal, capturado adecuadamente por la regresión polinómica de tercer grado representada en la gráfica. Esta representación permite visualizar cómo el modelo ajustó los datos en un espacio tridimensional.

Durante el desarrollo de este proyecto, también se probó la regresión lineal múltiple de segundo grado, pero finalmente se optó por la de tercer grado debido a su capacidad para representar de manera más precisa las relaciones complejas presentes en los datos y a los mejores resultados obtenidos en las métricas de evaluación. El éxito de este modelo radica en un correcto ajuste y validación, utilizando métricas como el error cuadrático medio (MSE) para evitar el sobreajuste y garantizar predicciones confiables.

En el caso de la regresión lineal múltiple, que considera múltiples variables independientes, la ecuación se utiliza de esta manera:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

Donde:

- $x_1, x_2, \ldots, x_n$: Variables independientes.
- $\beta_1, \beta_2, \ldots, \beta_n$: Coeficientes de regresión asociados a cada variable independiente.

2.2.8. Gradient boosting machine

Es un modelo supervisado basado en técnicas de ensamble que combina múltiples predictores débiles, generalmente árboles de decisión poco profundos, para construir un modelo robusto y preciso. Su enfoque se desarrolla de manera iterativa, donde cada árbol adicional se entrena específicamente para reducir los errores residuales generados por los árboles anteriores. Este proceso se guía mediante el cálculo del gradiente negativo de métricas relevantes, como el error cuadrático medio en problemas de regresión, permitiendo al modelo concentrarse en corregir las predicciones más erróneas. Cada árbol realiza contribuciones incrementales, reguladas por un parámetro de aprendizaje, lo que garantiza un aumento gradual en la complejidad del modelo y reduce el riesgo de sobreajuste.

A continuación, se describen los pasos clave de su formulación matemática y entendimiento base del algoritmo [16].

- **Función objetivo y predicción inicial**

El objetivo principal es minimizar una función de pérdida $L(y, F(x))$, donde y representa los valores reales, $F(x)$ es el modelo predictivo, y x denota las características de entrada.

El modelo comienza con una predicción inicial, usualmente el valor medio de la variable dependiente en problemas de regresión:

$$F_0(x) = \arg\min_c \sum_{i=1}^{N} L(y_i, c), \tag{1}$$

donde N es el número de muestras en los datos.

- **Iteración del modelo**

En cada iteración m, el modelo entrena un nuevo predictor débil $h_m(x)$ para corregir los errores residuales generados por el modelo acumulado hasta ese punto $F_{m-1}(x)$. Esto se logra mediante la aproximación del gradiente negativo de la función de pérdida:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}, \tag{2}$$

donde r_{im} son los residuales calculados para cada muestra i en la iteración m.

- **Actualización del modelo**

El nuevo predictor $h_m(x)$ se ajusta para aproximar los residuales r_{im}. Una vez entrenado, el modelo se actualiza como:

$$F_m(x) = F_{m-1}(x) + \nu \cdot h_m(x), \tag{3}$$

donde $\nu \in (0, 1]$ es el parámetro de aprendizaje que regula la contribución incremental de cada predictor.

- **Predicción final**

El modelo final después de M iteraciones se expresa como:

$$F(x) = F_0(x) + \sum_{m=1}^{M} \nu \cdot h_m(x). \tag{4}$$

- **Función de pérdida típica**

Se utiliza el error cuadrático medio (MSE):

$$L(y, F(x)) = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - F(x_i)\right)^2. \tag{5}$$

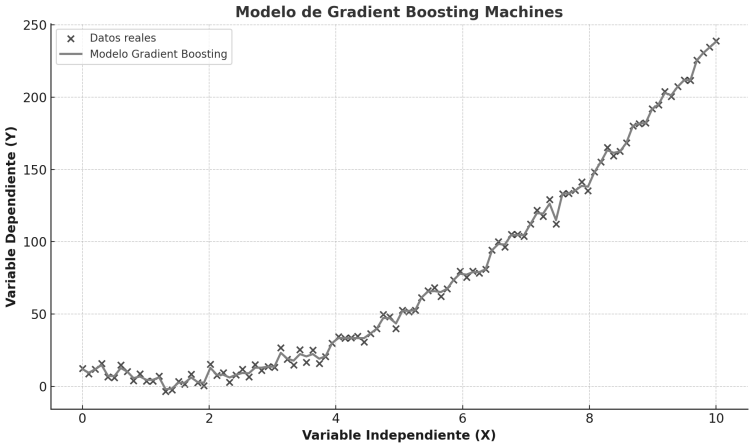


Figura 7: Gradient boosting machines [14]

Como se evidencia en el gráfico 7 se muestra como el modelo de Gradient Boosting Machines se ajusta a los datos. Los puntos ($\times$) representan los datos reales, mientras que la línea sólida ilustra las predicciones generadas por el modelo. Se observa cómo el modelo logra capturar la tendencia no lineal de la relación entre la variable independiente (X) y la variable dependiente (Y), proporcionando un ajuste preciso a lo largo de todo el rango de datos. Este comportamiento evidencia la capacidad del modelo para manejar relaciones complejas entre las variables.

2.2.9. Random Forest

El modelo de Random Forest se presenta como una herramienta bastante útil en el análisis estadístico y de predicción, especialmente para el estudio de variables complejas en datos tanto espaciales como no espaciales. Este enfoque metodológico permite identificar patrones no lineales y relaciones intrincadas dentro de grandes conjuntos de datos heterogéneos. Su capacidad para manejar un volumen significativo de datos y capturar interacciones multifactoriales lo convierte en una solución poderosa para áreas de aplicación como la econometría, la geografía y la planificación urbana.

En el caso específico del municipio de Dagua, Valle del Cauca, el modelo de Random Forest adquiere un valor particular en la predicción de avalúos catastrales. A diferencia de técnicas más simples, como la regresión lineal, este modelo se basa en la generación y combinación de múltiples árboles de decisión para representar y analizar relaciones complejas entre las características de los terrenos y sus valores catastrales. Esto le

permite superar limitaciones inherentes a otros métodos, proporcionando una visión más completa y detallada de los factores que influyen en el avalúo.

Esta metodología no solo considera atributos individuales de los predios, tales como el área, el uso económico y el estado de construcción, sino que también incorpora variables contextuales relevantes. Entre estas se destacan las infraestructuras disponibles y las características de las propiedades circundantes. Cada árbol generado en el modelo contribuye de manera incremental a la predicción final, mitigando el impacto de posibles sesgos y garantizando una estimación más precisa y confiable.

Además, la implementación de Random Forest en este contexto permite evaluar y jerarquizar la importancia de cada variable en la predicción de los valores catastrales. Esto resulta particularmente útil para identificar los factores determinantes que influyen en el mercado inmobiliario de Dagua, tales como la ubicación geográfica, las características físicas del terreno y el acceso a infraestructuras.

A continuación, se presentan los pasos relevantes y a tener en cuenta para la formulación matemática y la comprensión fundamental del algoritmo [17]

- **Construcción del bosque de árboles**

El objetivo principal de Random Forest es minimizar una función de pérdida $L(y, \hat{y})$, donde y representa los valores reales, $\hat{y}$ son las predicciones generadas por el modelo, y las características de entrada están representadas por x. Para lograr esto, el modelo entrena múltiples árboles de decisión T_b utilizando diferentes subconjuntos de datos seleccionados mediante muestreo.

- **Predicción del modelo**

Para un modelo con B árboles de decisión, la predicción final se obtiene mediante:

- **Regresión:** Promedio de las predicciones individuales:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x). \tag{6}$$

- **Clasificación:** Votación mayoritaria entre los árboles:

$$\hat{y} = \arg\max_{c} \sum_{b=1}^{B} I(T_b(x) = c), \tag{7}$$

donde $I(\cdot)$ es la función indicadora que toma el valor 1 si el árbol predice la clase c, y 0 en caso contrario.

- **Importancia de las variables**

La importancia de una variable j se evalúa mediante la disminución promedio de la impureza (criterio como el índice Gini o la entropía) cuando se utiliza la variable para dividir los nodos en los árboles:

$$I_j = \frac{1}{B} \sum_{b=1}^{B} \sum_{t \in T_b} \Delta i_t \cdot I(v_t = j), \tag{8}$$

donde Δi_t es la reducción de impureza en el nodo t, v_t es la variable utilizada en la división, y $I(\cdot)$ es la función indicadora.

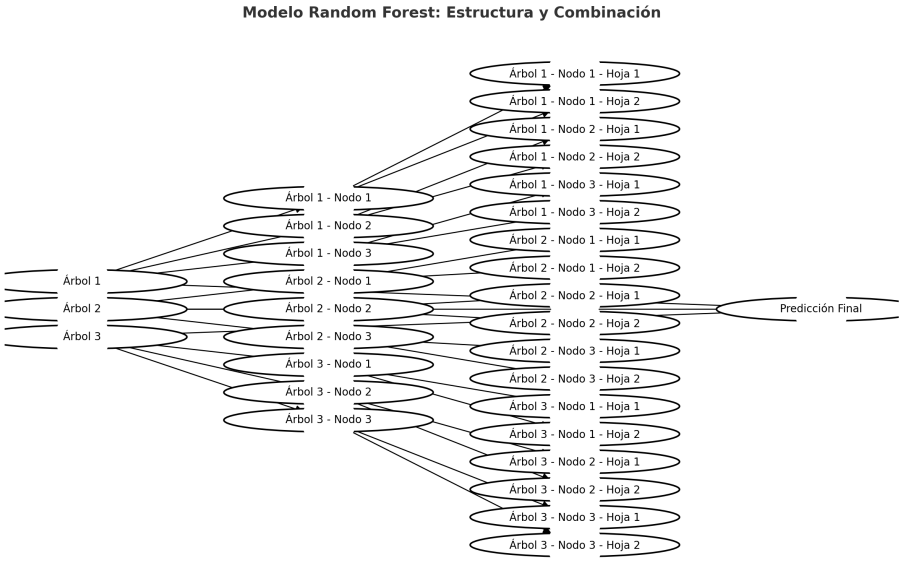


Figura 8: Random Forest [14]

Se aprecia en la previa ilustración 8 la estructura y combinación de un modelo Random Forest. El modelo está compuesto por múltiples árboles de decisión independientes

(Árbol 1, Árbol 2, Árbol 3, etc.), cada uno dividido en nodos y hojas. Cada árbol genera una predicción parcial basada en los datos que le corresponden. Estas predicciones parciales se combinan posteriormente para producir una predicción final, que resulta de la regresión. Esta representación destaca cómo el modelo Random Forest utiliza múltiples árboles para reducir el riesgo de sobreajuste y mejorar la precisión, al aprovechar la diversidad de las predicciones individuales.

2.2.10. Ensamble (Random Forest - Gradient Boosting)

Es un enfoque supervisado que combina las fortalezas de dos métodos de ensamble para crear un modelo robusto y eficiente, ideal para la predicción catastral y contextos de investigación. Por un lado, utiliza las predicciones de Random Forest, que entrena múltiples árboles de decisión de forma independiente, lo que le permite manejar datos diversos y ruidosos, como características geográficas, áreas de terreno y variables socioeconómicas. Por otro lado, incorpora Gradient Boosting, que ajusta los árboles de manera secuencial, enfocándose en corregir los errores más significativos y capturando relaciones no lineales y complejas entre las variables del predio y su valor. Esta combinación es especialmente útil en investigaciones catastrales, ya que permite mejorar la precisión en la predicción del valor de los predios, ajustándose a las particularidades de los datos, mientras controla la complejidad del modelo y reduce el riesgo de sobreajuste, garantizando resultados confiables y generalizables.

Por otro lado,en esta seccion, se mencionan pasos claves de elementos matemáticos que se deben tener en cuenta para este tipo de modelo ensamble [18]:

- **Predicción de Random Forest**

  El modelo utiliza matemáticamente las mismas fórmulas descritas en la Sección 2.2.9, combinando las fortalezas de Random Forest para crear un modelo robusto y eficiente.

- **Predicción de Gradient Boosting**

  Tal como se detalla en la Sección 2.2.8, el modelo emplea las mismas bases matemáticas descritas para Gradient Boosting, integrando sus capacidades de ajuste secuencial para corregir errores significativos y mejorar la precisión del modelo.

- **Combinar Random Forest y Gradient Boosting**

Las predicciones de ambos métodos se combinan ponderadamente para aprovechar sus ventajas:

$$\hat{y} = \alpha \cdot \hat{y}_{RF} + (1 - \alpha) \cdot \hat{y}_{GB}, \tag{9}$$

donde $\hat{y}_{RF}$ es la predicción de Random Forest, $\hat{y}_{GB}$ es la predicción de Gradient Boosting, y $\alpha \in [0, 1]$ es un coeficiente que controla la contribución relativa de cada modelo.

- Función de pérdida típica

Nos estamos enfrentando a un problemaa de regresión, por ende, en la función de pérdida se utliza el error cuadrático medio (MSE):

$$L(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2, \tag{10}$$

donde y_i son los valores reales y $\hat{y}_i$ son las predicciones combinadas.

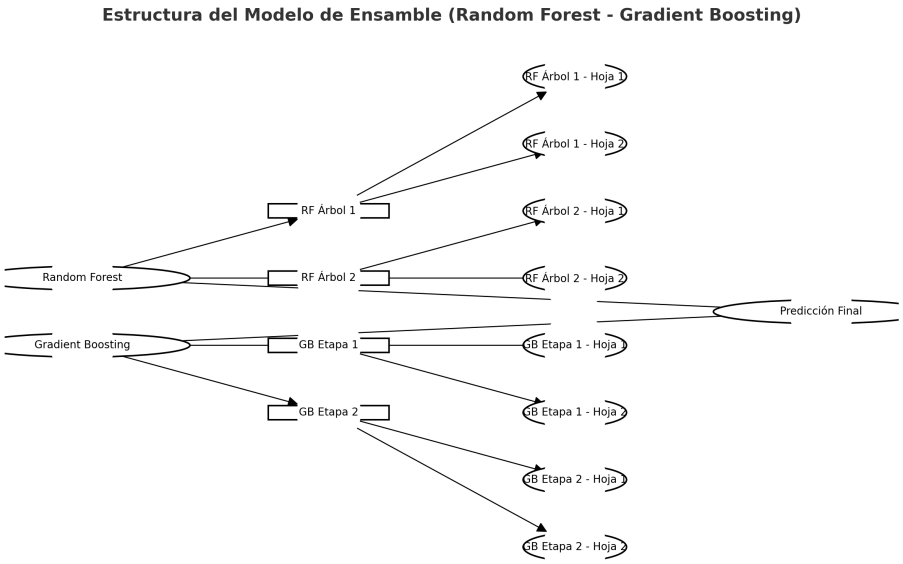


Figura 9: Emsamble Random Forest - Gradient Boosting [14]

El anterior gráfico 9 muestra el proceso que conlleva la predicción en conjunto para los modelos, se aprecia la similaridad que tiene con el modelo de Random Forest.

2.2.11. K-Nearest Neighbors (KNN)

El algoritmo K-Nearest Neighbors (KNN) se posiciona como una herramienta versátil y eficaz dentro del aprendizaje supervisado, capaz de abordar problemas de clasificación y regresión. Su funcionamiento se basa en la proximidad entre observaciones en un espacio de características, lo que lo convierte en una opción ideal para analizar datos en contextos donde las relaciones locales tienen un papel crítico en la precisión de las predicciones. En el caso específico de los avalúos catastrales, el KNN permite identificar patrones entre propiedades que comparten similitudes en atributos físicos y ubicación geográfica, bajo la premisa de que dichas similitudes se reflejan directamente en su valor catastral [19].

Este modelo es particularmente relevante en situaciones donde intervienen variables determinantes como el destino económico del predio, el uso del suelo, el área de terreno y construcción, y características espaciales como la manzana o vereda. Estas variables, al ser integradas en el modelo, no solo contribuyen a capturar relaciones evidentes, sino que también permiten explorar interacciones complejas y sutiles en los datos. Esto refuerza su capacidad para representar de manera precisa las dinámicas que influyen en los avalúos, consolidándolo como una herramienta valiosa en proyectos de análisis predictivo que requieren manejar diversidad y complejidad en las características.

Desde un enfoque técnico, el modelo KNN basa su operación en la identificación de los K vecinos más cercanos a un punto de interés dentro de un espacio de características, empleando métricas de distancia como la euclidiana o Manhattan. Cada vecino aporta información clave a la predicción, ya sea a través de un promedio ponderado inversamente proporcional a la distancia o mediante un promedio simple. Esta estructura algorítmica le permite capturar relaciones no lineales, lo cual es especialmente ventajoso en el análisis de avalúos catastrales, donde los datos suelen ser heterogéneos y altamente contextuales.

La elección del modelo KNN en este proyecto se justificó no solo por su simplicidad y adaptabilidad, sino también por su capacidad para manejar un conjunto amplio y diverso de características, incluidas variables categóricas transformadas y normalizadas para garantizar su compatibilidad con las métricas de distancia. La optimización del parámetro k, mediante técnicas como la validación cruzada, resultó crucial para equilibrar el sesgo y la varianza, evitando tanto el sobreajuste como la omisión de patrones relevantes. Este enfoque, combinado con un preprocesamiento adecuado de los

datos, permitió que el modelo se destacara entre las opciones evaluadas, demostrando un rendimiento robusto y consistente en la predicción de avalúos catastrales, así como una capacidad notable para adaptarse a las complejidades inherentes de este tipo de análisis.

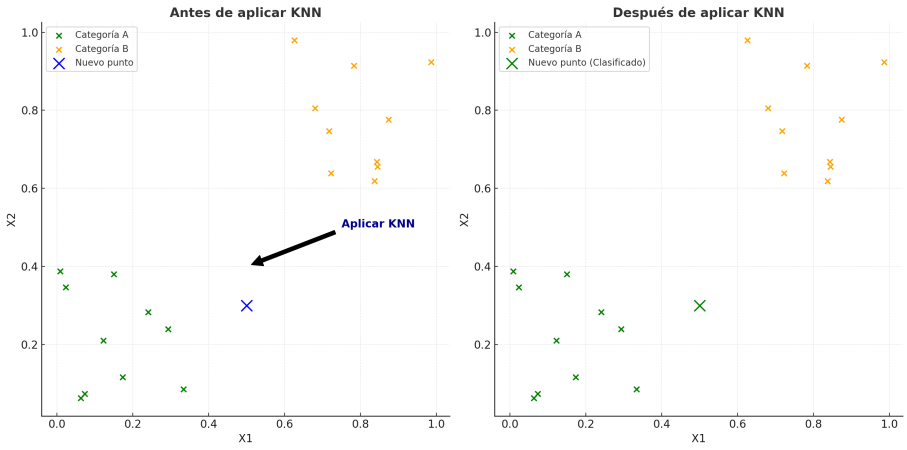


Figura 10: KNN [14]

La representación visual 10 muestra cómo el modelo KNN clasifica un nuevo punto según la proximidad a sus vecinos más cercanos, medida por una métrica de distancia. Es importante destacar que este modelo se fundamenta en la definición de proximidad entre puntos a través de una métrica de distancia. A continuación, se presentan definiciones específicas relacionadas con este modelo en el contexto matemático [20].

- **Distancia Euclidiana:**

$$d(x, x_i) = \sqrt{\sum_{j=1}^{d} (x_j - x_{ij})^2}. \tag{11}$$

- **Distancia Manhattan:**

$$d(x, x_i) = \sum_{j=1}^{d} |x_j - x_{ij}|. \tag{12}$$

- **Predicción para problemas de regresión**

La predicción $\hat{y}$ para un punto de interés x se calcula como el promedio de los valores asociados a los k vecinos más cercanos:

$$\hat{y} = \frac{1}{k} \sum_{i \in \mathcal{N}_k(x)} y_i, \tag{13}$$

donde $\mathcal{N}_k(x)$ denota el conjunto de índices de los k vecinos más cercanos a x.

- Predicción para problemas de clasificación

La predicción corresponde a la clase mayoritaria entre los k vecinos más cercanos:

$$\hat{y} = \arg\max_{c \in \mathcal{C}} \sum_{i \in \mathcal{N}_k(x)} I(y_i = c), \tag{14}$$

donde $\mathcal{C}$ es el conjunto de clases posibles e $I(\cdot)$ es una función indicadora que toma el valor 1 si $y_i = c$, y 0 en caso contrario.

2.2.12. Maquina de soporte vectorial

Las Máquinas de soporte vectorial, son algoritmos supervisados diseñados para encontrar un hiperplano óptimo en un espacio multidimensional, con el fin de separar o aproximar datos de manera eficiente. En el ámbito del avalúo catastral, este modelo se utiliza para predecir valores continuos, como el valor aproximado de un predio, mediante el ajuste de una función que minimiza el error entre los valores observados y estimados. Una de sus ventajas es su capacidad para manejar relaciones bastante complejas y no lineales entre variables como el área del terreno, la ubicación geográfica o la zona económica. Esto se logra utilizando funciones de transformación conocidas como kernels, que proyectan los datos a un espacio de mayor dimensión, facilitando la identificación de relaciones más lineales. Este enfoque ha demostrado ser altamente efectivo en aplicaciones como la predicción del área construida en entornos urbanos, resaltando su importancia en estimaciones vinculadas al avalúo catastral [21].

Matemáticamente, el objetivo principal de una Máquina de Soporte Vectorial es encontrar un hiperplano $w \cdot x + b = 0$ que separe los datos de dos clases o ajuste una función a los datos continuos en regresión, maximizando el margen entre los puntos más cercanos a dicho hiperplano. Para ello, se debe comprender y dar solución a el siguiente problema de optimización [22]:

$$\min_{w,b} \frac{1}{2}|w|^2 \text{ sujeto a } \begin{cases} y_i(w \cdot x_i + b) \geq 1 & \forall i, \text{en clasificación, } |y_i - (w \cdot x_i + b)| \leq \epsilon \\ \forall i, \text{en regresión.} \end{cases} \tag{15}$$

- w representa el vector de pesos que define la orientación del hiperplano.
- b es el término independiente que ajusta la posición del hiperplano.
- y_i son los valores reales (en regresión) o las etiquetas de clase (en clasificación).
- ϵ controla la tolerancia al error en regresión.

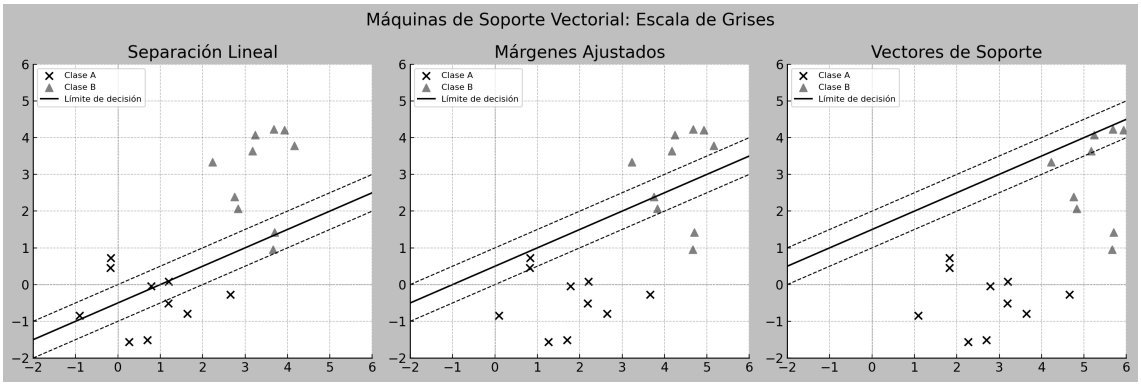


Figura 11: Maquina de soporte vectorial [14]

La figura 11 de las máquinas de soporte vectorial ilustra diferentes etapas en el proceso de clasificación: la separación lineal, el ajuste de márgenes y la identificación de los vectores de soporte. Estas etapas son fundamentales para comprender cómo este modelo maneja datos no linealmente separables, cuando los datos no lo son, se utiliza una función de transformación $\phi(x)$ que proyecta los datos a un espacio de mayor dimensión. Este modelo utiliza una función kernel $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ para calcular el producto interno entre los datos proyectados sin necesidad de realizar la transformación de manera explícita.

2.2.13. Redes neuronales

Las redes neuronales son sistemas computacionales diseñados para resolver problemas complejos mediante el procesamiento de datos a través de múltiples capas interconectadas. Estas capas, formadas por nodos o unidades computacionales, trabajan en conjunto para modelar patrones y relaciones presentes en los datos. Cada nodo aplica una función de activación que introduce no linealidades en el modelo, lo que permite a las redes neuronales aproximar funciones complejas y adaptarse a problemas no lineales. En el ámbito del aprendizaje automático, estas redes se destacan por su capacidad para aprender representaciones jerárquicas de los datos, haciéndolas altamente efectivas en tareas como clasificación, regresión y detección de patrones anómalos.

Cuando las redes neuronales incluyen múltiples capas ocultas, se enmarcan dentro del aprendizaje profundo, una subdisciplina que se enfoca en la resolución de problemas complejos mediante el análisis de grandes volúmenes de datos. Según lo expuesto en [23], estas redes pueden clasificarse en cuatro enfoques principales, cada uno diseñado para abordar diferentes tipos de problemas. En este trabajo, se implementa uno de estos enfoques, optimizando la capacidad de generalización y el rendimiento en escenarios con alta dimensionalidad y relaciones complejas entre las variables.

2.2.14. Perceptrón multicapa

El modelo de perceptrón multicapa (MLP) es una herramienta valiosa para mejorar la precisión de las predicciones en diversos contextos y aplicaciones. El MLP se basa en una arquitectura de red neuronal artificial compuesta por varias capas, incluyendo una capa inicial que recibe los datos de entrada, capas intermedias que procesan la información mediante cálculos y transformaciones (comúnmente llamadas capas ocultas), y una capa final que genera los resultados. Las neuronas dentro de cada capa están interconectadas, formando una red densa que permite al modelo aprender relaciones no lineales complejas a partir de los datos.

[24].

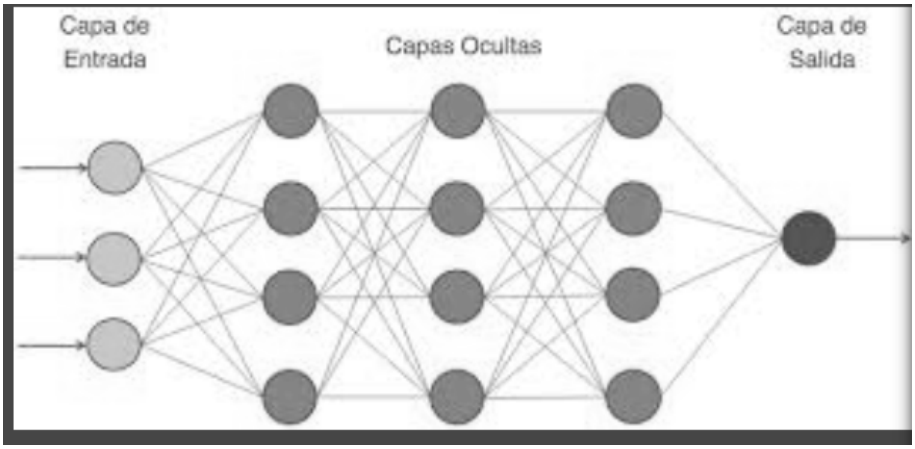


Figura 12: Perceptrón multicapa [14]

Matemáticamente, el Perceptrón Multicapa se basa en funciones de activación que transforman los datos a medida que pasan por las capas de la red como se muestra en la figura 12. Dados los datos de entrada $X = x_1, x_2, \ldots, x_n$, el modelo aplica pesos y sesgos en cada capa para calcular los valores intermedios de las neuronas [25].

- En la capa l, el valor de activación para la neurona j está dado por:

$$a_j^{(l)} = \sigma\left(\sum_i w_{ij}^{(l)} a_i^{(l-1)} + b_j^{(l)}\right), \tag{16}$$

  donde:

  - $a_j^{(l)}$ es la salida de la neurona j en la capa l.
  - $w_{ij}^{(l)}$ son los pesos que conectan la neurona i de la capa $(l-1)$ con la neurona j de la capa l.
  - $b_j^{(l)}$ es el sesgo asociado a la neurona j en la capa l.
  - $\sigma(\cdot)$ es la función de activación, como la sigmoide $\sigma(x) = \frac{1}{1+e^{-x}}$, ReLU máx$(0, x)$ o tanh $\tanh(x)$.

- El aprendizaje de este modelo se basa en la minimización de una función de pérdida L, la cual mide el error entre las salidas predichas $\hat{y}$ y los valores reales y. Esto se logra mediante el método de retropropagación, que ajusta los pesos y sesgos utilizando el gradiente descendente:

$$\Delta w_{ij}^{(l)} = -\eta \frac{\partial L}{\partial w_{ij}^{(l)}}, \tag{17}$$

donde η es la tasa de aprendizaje.

- La salida final de la red, en la capa de salida L, se calcula como:

$$\hat{y}_k = \sigma\left(\sum_j w_{jk}^{(L)} a_j^{(L-1)} + b_k^{(L)}\right), \tag{18}$$

donde $\hat{y}_k$ es la predicción para la clase o el valor objetivo k.

Al lograr aplicar este tipo de modelo bajo el contexto mencionado anteriormente, se puede aprovechar sus capacidades para capturar relaciones no lineales entre las diversas características de las propiedades y su valor catastral. Por ejemplo, al identificar interacciones complejas entre características como tamaño de la propiedad, ubicación geográfica, características estructurales y servicios disponibles[24], lo que podría no ser capturado de manera efectiva por modelos lineales más simples mencionados anteriormente, no obstante, el MLP puede adaptarse para manejar datos geoespaciales, como la ubicación de las inmuebles en un mapa, lo que permite incorporar información espacial en el proceso de predicción. denotando gran utilidad en el contexto de la valoración catastral en el municipio de Dagua.

2.3. Antecedentes

En los últimos años, la estimación del avalúo catastral ha sido un tema de creciente interés en la investigación y en la implementación de metodologías más eficientes. Diversos estudios han explorado soluciones automatizadas para la clasificación y valoración de terrenos, en especial en áreas rurales y urbanas. A continuación, se presentan algunos de los trabajos más relevantes en el ámbito de la tasación catastral, los cuales abordan desde métodos tradicionales hasta la implementación de tecnologías emergentes, como el aprendizaje automático y la teledetección.

Uno de los trabajos relevantes es la propuesta de una metodología automatizada para la clasificación física de terrenos rurales, enfocada en la valoración catastral dentro de un catastro multipropósito. En este estudio se utilizó el análisis espacial y la teledetección para definir variables dinámicas, como la accesibilidad, conectividad y aprovechamiento de fuentes hídricas. Esta metodología permitió reemplazar los métodos tradicionales, reduciendo la subjetividad en la valoración. La implementación de esta metodología en el municipio de San Francisco, Cundinamarca, mostró una mejora considerable en la eficiencia y precisión, con resultados validados en campo y comparados con estudios previos del IGAC.[26].

Por otro lado, en un enfoque para la valoración de tierras urbanas, se identificó que, en Latinoamérica, los valores catastrales generalmente no están actualizados debido a la complejidad del proceso de tasación masiva. En este estudio se propone un método basado en el uso de datos abiertos y gratuitos, como los obtenidos de OpenStreetMap y la Comisión Europea, para crear variables relacionadas con las características del terreno y el grado de urbanización. Estos datos se utilizan para entrenar modelos de aprendizaje automático que estiman el valor catastral, demostrando que este enfoque puede reducir la complejidad, los costos y el tiempo necesarios en el proceso de tasación masiva [27].

Asimismo, otro trabajo relevante en la mejora de los modelos de valoración inmobiliaria se centra en el uso de funciones de tendencia spline para reflejar los cambios en los precios de bienes raíces a lo largo del tiempo. Se argumenta que, para obtener modelos de valoración actualizados, es crucial que estos reflejen las tendencias actuales del mercado. En este estudio se demostró que las funciones spline son efectivas para modelar dichas tendencias, comparando su precisión con la de otros métodos tradicionales y algoritmos de aprendizaje automático [28].

Por último, un estudio sobre la comparación de algoritmos expertos con modelos de aprendizaje automático para la valoración inmobiliaria resalta la efectividad de los algoritmos basados en comparaciones de ventas cuando se dispone de pocos datos. Este análisis comparativo se realizó utilizando datos reales de un sistema catastral y registros de ventas de inmuebles, y se evaluó el rendimiento de varios modelos de regresión orientados a la valoración inmobiliaria [29]

3. Desarrollo de la solución

Para la implementación de un modelo adecuado para la predicción del avalúo catastral en el municipio de Dagua, el desarrollo se realizó en Python, aprovechando su amplia gama de bibliotecas para el análisis de datos y el modelado predictivo. Se comenzó con la selección y evaluación de diferentes enfoques algorítmicos que pudieran abordar las características del conjunto de datos, que incluye variables tanto continuas como categóricas, así como relaciones espaciales que deben ser consideradas en la predicción de los valores catastrales. Con el objetivo de encontrar la mejor opción, se propuso la creación y evaluación de varios modelos, los cuales fueron entrenados utilizando los datos disponibles y luego ajustados mediante la búsqueda de hiperparámetros.

Además, se emplearon herramientas de visualización de datos para realizar un análisis exploratorio previo, lo cual permitió comprender mejor las relaciones entre las variables y visualizar patrones relevantes. Esto incluyó el uso de gráficos y representaciones visuales que facilitaron la comprensión de la distribución y correlación de las variables dentro del conjunto de datos.

A través de la optimización de parámetros clave, se generaron modelos adicionales que permitieron realizar un análisis comparativo entre ellos. Estos modelos fueron evaluados mediante diversas métricas de rendimiento, como el error cuadrático medio (RMSE) y la precisión de las predicciones, utilizando una carpeta de datos de prueba ("Test"). Con base en los resultados obtenidos, se seleccionó el modelo con el mejor desempeño, que luego fue implementado en el prototipo funcional para la predicción del avalúo catastral.

Para llevar a cabo este proyecto, se utilizaron bibliotecas como Pandas, Numpy, Matplotlib y Scikit-learn, entre otras, las cuales fueron fundamentales para la manipulación, visualización y modelado de datos, así como para la optimización

de hiperparámetros.

4. Preparación de datos

El procesamiento inicial de datos constituye una etapa fundamental en proyectos de análisis predictivo, ya que define la calidad y precisión de los resultados [30]. En el caso del modelo de avalúo catastral para el municipio de Dagua, esta fase se enfocó en preparar, limpiar y estructurar la información recopilada para garantizar su consistencia y relevancia. Entre las tareas realizadas se incluyen la imputación de valores faltantes, la cual se llevó a cabo mediante metodologías estadísticas, tales como la media y la mediana, según el contexto de cada variable. Este enfoque permite asegurar que los datos imputados no alteren la integridad del conjunto de datos original. Además, se realizó el manejo de registros duplicados, la corrección de inconsistencias y la normalización de variables. También se ajustaron valores atípicos que podrían afectar la estabilidad y rendimiento del modelo predictivo. La división y extracción de información a partir del número de predio representó una etapa fundamental en el procesamiento inicial de los datos. Este identificador, proporcionado por el IGAC, contenía múltiples dimensiones de información que requerían ser analizadas y estructuradas para su uso en el modelo predictivo. El proceso implicó un análisis riguroso, apoyado en investigación, charlas técnicas con expertos del IGAC y pruebas iterativas, para identificar y abstraer las características más relevantes contenidas en este número único. Estas acciones transformaron el conjunto de datos en una base confiable y lista para las siguientes etapas del proyecto.

El procesamiento inicial también abarcó la segmentación del conjunto de datos en categorías específicas, como terrenos, construcciones y predios, lo que permitió un análisis más detallado de las relaciones entre las variables.Como resultado de este exhaustivo proceso de preparación de datos, se obtuvo un conjunto de datos depurado, estructurado y enriquecido, que integró tanto información estadística como visual crucial para las etapas posteriores de modelado predictivo. En esta fase, se aplicaron técnicas avanzadas de análisis exploratorio de datos (EDA), como análisis de correlación, visualización de distribuciones y detección de patrones multivariantes, que permitieron identificar relaciones significativas entre las distintas variables. Este enfoque facilitó una mejor comprensión de la estructura

subyacente de los datos y de cómo las variables interaccionan entre sí, lo cual es fundamental para un modelo predictivo robusto.

El proceso de enriquecimiento de los datos fue importante, ya que no solo se extrajeron insights sobre las características que influyen en la valoración catastral, sino que también se procedió a la corrección y ajuste de anomalías, incluyendo la imputación de valores faltantes y la eliminación de valores atípicos mediante técnicas adecuadas para cada tipo de variable. Esto permitió optimizar la calidad y la representatividad del conjunto de datos, lo cual redujo significativamente los posibles sesgos derivados de errores en los datos originales.

La depuración de datos realizada en esta etapa fue crucial para garantizar la integridad y fiabilidad del modelo predictivo. La limpieza de los datos, mediante la eliminación de inconsistencias y la normalización de variables, fortaleció la capacidad del modelo para generar predicciones precisas y coherentes. De esta manera, la fase de preparación de datos no solo estableció una base sólida para la construcción del modelo, sino que también alineó los datos con los requisitos técnicos de los algoritmos de Machine Learning. Como resultado, los modelos predictivos fueron entrenados con un conjunto de datos que representa de manera fiel las complejidades y variabilidad inherentes a las características y dinámicas del sistema estudiado, mejorando significativamente su capacidad para generalizar y realizar predicciones precisas en contextos y escenarios reales.

4.1. Procesamiento inicial de datos

El procesamiento inicial de los datos representó una de las etapas más críticas del proyecto, requiriendo un esfuerzo significativo para garantizar la correcta ejecución de cada parte del proceso [30]. Este paso no solo definió el punto de partida para el desarrollo del modelo de avalúo catastral, sino que también estableció los cimientos para la calidad y precisión de los resultados obtenidos. Durante esta fase, se realizaron múltiples pruebas y ajustes en cada etapa del procesamiento con el fin de asegurar que los datos fueran consistentes, relevantes y perfectamente adecuados para las siguientes fases del proyecto.

La etapa inicial del procesamiento de datos se centró en la carga y el análisis de las cuatro bases de datos proporcionadas por la unidad administrativa especial de catastro, las bases de datos proporcionadas del municipio de Dagua incluyen información detallada sobre terrenos, construcciones, destinos y zonas

geoeconómicas. La base de datos de terrenos **r1t** [31] contiene información sobre las características físicas de los predios, como su área y ubicación. La base de datos de construcciones **r1c** [32] almacena detalles sobre edificaciones dentro de los terrenos, incluyendo su uso y distribución. La base de datos de destino **r1d** [33] asocia cada predio con su uso específico y su valor catastral, permitiendo analizar la relación entre estos factores. Finalmente, la base de datos de zonas **r2b** [34] incorpora información geoeconómica para la valoración de los predios, estas contienen en total 116.227 registros. En el siguiente gráfico 13 se puede apreciar la proporción de cada uno.

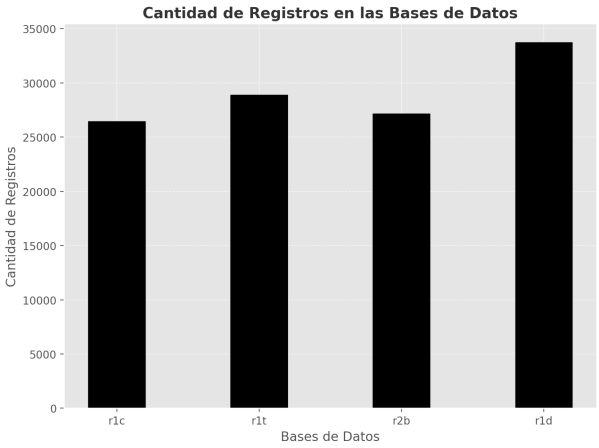


Figura 13: Proporción de bases de datos [14]

Este paso fue especialmente crítico, ya que implicó abordar cuatro conjuntos de datos independientes, cada uno con estructuras, formatos y características específicas. Fue necesario realizar un análisis exhaustivo y aplicar pruebas iterativas para identificar las correspondencias y relaciones entre los datos de cada conjunto. Tras un análisis riguroso, se concluyó que las cuatro bases de datos representaban diferentes dimensiones de una misma entidad global. Este hallazgo fue el resultado de un proceso minucioso de investigación, que permitió establecer una base sólida para los pasos subsecuentes.

4.1.1. Análisis y eliminación de características

Continuando con el procesamiento inicial de datos se realizó la eliminación de características en varias bases de datos relacionadas con avalúo catastral, siguiendo

criterios de relevancia, concordancia y redundancia. El objetivo principal fue optimizar la calidad y consistencia de los datos para su uso en análisis posteriores.

En este proyecto se procesaron cuatro bases de datos principales: `destino` (`r1d`), `terrenos` (`r1t`), `construcciones` (`r1c`) y `r2` (`r2b`). Se realizó una limpieza exhaustiva eliminando columnas que contenían datos redundantes, nulos o irrelevantes para el propósito del análisis. En la base de datos `r1t`, por ejemplo, se eliminaron características como metadatos (`created_at`, `updated_at`, `id`) y columnas no utilizadas relacionadas con áreas y geometrías (`geom`, `area_privada`, entre otras). De manera similar, en `r1c` se descartaron columnas irrelevantes, incluyendo identificadores y detalles específicos que no contribuían al análisis (`altura`, `observacion`, `nombre_destino`, entre otras). En el caso de `r1d`, solo se conservaron columnas esenciales como `nro_predial_completo`, `destino`, `area_terreno`, `area_construida` y `avaluo_catastral`, las cuales fueron procesadas y renombradas para asegurar consistencia. Se consolidaron datos redundantes, asegurando que cada tipo de información estuviera presente únicamente en la base de datos correspondiente. Por ejemplo, los datos relacionados con terrenos se mantuvieron en `r1t` y no en otras bases. Además, se descartaron todas las características nulas o con valores no válidos, lo que contribuyó a la consistencia y precisión de los datos finales.

El resultado de este trabajo fueron bases de datos depuradas con altas consistencias, en la que todos los registros coincidían perfectamente entre las diferentes tablas. Esto garantizó que, por ejemplo, los terrenos se correspondieran con predios existentes y que no hubiera construcciones asociadas a terrenos no válidos. La preparación de los datos en esta etapa fue crucial para garantizar la calidad y precisión en los análisis posteriores relacionados con los avalúos catastrales.

4.1.2. Filtrado y reestructuración de las bases de datos

El filtrado y reestructuración de las bases de datos representó una etapa esencial dentro del procesamiento inicial, enfocada en depurar y seleccionar únicamente los registros más relevantes para el desarrollo del modelo. Tras completar el proceso de carga, análisis y eliminación de características redundantes o irrelevantes, se establecieron criterios específicos para garantizar el manejo y relevancia de los datos utilizados en etapas posteriores. Este paso no solo permitió optimizar el conjunto de datos, sino que también sentó las bases para el correcto entrenamiento

de los modelos predictivos.

La transición del filtrado a la reestructuración de datos marcó un punto crucial en el procesamiento de la información. **Se definió como criterio principal la inclusión de predios con un máximo de cinco terrenos y aquellos que no poseían construcciones o que tenían hasta cinco construcciones.** Este criterio, además, destaca por su flexibilidad, permitiendo su ajuste a cualquier límite deseado lo que abre la posibilidad de realizar pruebas adicionales y adaptar el enfoque según los requerimientos del análisis futuro. En el contexto técnico, este proceso también se evaluó con un límite reducido de tres terrenos y tres construcciones, proporcionando una perspectiva adicional al lector para explorar diferentes configuraciones y su impacto en los resultados. Los registros que excedían estas condiciones fueron eliminados, ya que representaban un impacto significativo en la amplitud de la base de datos consolidada. Este impacto se manifiesta en términos de complejidad computacional y dimensionamiento de los datos, lo cual podría perjudicar el rendimiento y la precisión de los modelos de aprendizaje automático. En este contexto, se destaca que cada construcción asociada a un predio añadía siete características adicionales a la base de datos. Por ejemplo, un predio con dos construcciones incrementaba en catorce las columnas del conjunto de datos, mientras que cada terreno adicional añadía dos características. Un predio con dos terrenos y dos construcciones, por tanto, generaba un aumento total de dieciocho columnas. Estas expansiones, aunque necesarias para describir adecuadamente las propiedades, incrementaban exponencialmente la complejidad del conjunto de datos. La inclusión de un número excesivo de características afecta negativamente el rendimiento de los modelos por diversos factores. En primer lugar, un aumento en la cantidad de datos puede llevar al fenómeno conocido como "maldición de la dimensionalidad", donde los modelos requieren un mayor número de observaciones para generalizar correctamente, lo que a menudo resulta en un sobre ajuste y una menor capacidad de predicción en datos nuevos [35]. Además, un mayor número de características implica un incremento significativo en los costos computacionales, tanto en términos de tiempo de entrenamiento como de capacidad de almacenamiento [36]. Asimismo, no todas las características adicionales aportan información relevante para el modelo, y su inclusión puede introducir ruido en los datos, dificultando la identificación de patrones significativos [37].

La decisión de establecer un límite en el número de terrenos y construcciones res-

pondió a la necesidad de equilibrar la diversidad y representatividad de los datos con la eficiencia en el procesamiento. Este enfoque permitió reducir el tamaño y la redundancia del conjunto de datos, preservando únicamente los registros que aportaban valor al análisis. Al mismo tiempo, garantizó que la base de datos resultante fuera operativamente viable y adecuada para las exigencias de los modelos de aprendizaje automático empleados en el proyecto. Este filtrado no solo optimizó la calidad de los registros seleccionados, sino que también aseguró que los datos utilizados en las etapas posteriores fueran consistentes y representativos de las características relevantes para el modelo de avalúo catastral.

4.1.3. Unión de bases de datos

Como parte integral de el procesamiento inicial de datos la unión de las bases de datos tuvo como objetivo integrar la información proveniente de cuatro bases de datos, cada una con estructuras y características específicas, en un único conjunto de datos consolidado y coherente. La integración fue clave para capturar las múltiples características asociadas a los predios, garantizando que la información fuera consistente y adecuada para los modelos predictivos. Se realizó un análisis exhaustivo de cada una de las bases de datos ademas de este se identificaron las claves primarias, relaciones entre las bases y los registros en común, considerando el criterio principal aplicado durante el filtrado previo. Se utilizó como eje el identificador único de predio, que permitió vincular de manera precisa los datos relevantes. También se corrigieron inconsistencias, se eliminaron registros duplicados y se descartaron variables irrelevantes para asegurar la compatibilidad y calidad del conjunto de datos final. Este análisis inicial permitió establecer una estructura clara que facilitó la integración de los datos. La metodología incluyó la normalización de formatos, la eliminación de inconsistencias y la consolidación de los datos en una estructura uniforme que contiene 22553 registros, representando un 90 %

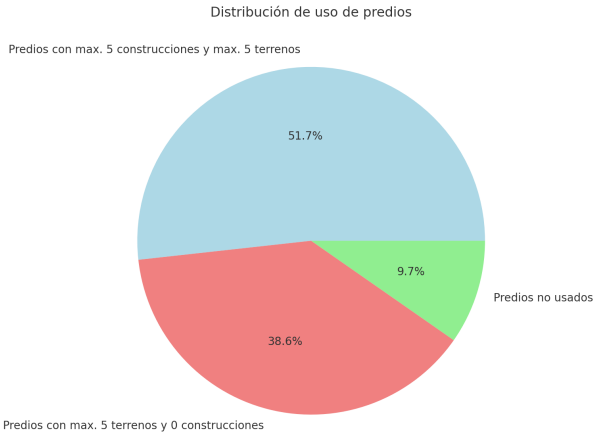


Figura 14: Gráfico de distribución de predios (datos capturados) [14]

de el total de predios que existían en el municipio de Dagua en el año 2024 como se aprecia en el gráfico circular 14. El identificador único del predio aseguró que los registros combinados fueran válidos y representativos de todas las dimensiones consideradas. El impacto de la unión fue significativo tanto en términos de calidad como de manejabilidad del conjunto de datos. En el contexto del proyecto, la unión de las bases de datos marcó un hito importante al permitir una visión integrada y consistente de la información disponible. Este proceso sentó las bases para un análisis eficiente y un entrenamiento preciso de los algoritmos de aprendizaje automático. La calidad del conjunto de datos consolidado fue determinante para garantizar la robustez y confiabilidad del modelo de avalúo catastral, demostrando la importancia de una integración rigurosa y bien planificada.

4.1.4. División del número de predio

Una parte crucial de esta fase fue la división del número de predio, un identificador proporcionado por el IGAC, del cual se derivaron múltiples características esenciales para el análisis. Mediante un proceso riguroso que incluyó investigación, charlas con expertos del IGAC y pruebas iterativas, se lograron abstraer inicialmente 16 características relevantes. Estas características incluían atributos como el departamento, municipio, zona (rural o urbana), sector, manzana/vereda, terreno, y otros detalles específicos sobre la propiedad, como el número de edificio, piso o unidad de propiedad horizontal. Sin embargo, durante el análisis de datos, se determinó que solo seis de estas características eran críticas para

el modelo predictivo y para la funcionalidad de la aplicación. En consecuencia, desde la etapa de división se tomó la decisión de abstraer únicamente estas seis características más relevantes: zona, sector, manzana/vereda, terreno, condición de propiedad, y tipo de registro.

La importancia de esta decisión se refleja directamente en el modelo final, utilizado por la aplicación. Este modelo requiere que el usuario introduzca únicamente la manzana/vereda para predecir el avalúo de un predio en el municipio de Dagua. Esta simplificación no solo optimiza la eficiencia del sistema, sino que también facilita la experiencia del usuario al reducir la cantidad de información necesaria para realizar predicciones precisas.

4.2. Análisis de datos

El análisis de datos para la predicción del avaluó catastral en el municipio de Dagua se fundamentó en la selección de una base de datos que contiene 22,477 registros, representando exactamente el 90.3 % de los predios en la región. Este porcentaje refleja el alto grado de representatividad de los datos utilizados, lo que permite construir un modelo robusto y confiable. La decisión de emplear esta base responde a la necesidad de contar con un volumen significativo de datos como se aprecia en el grafico 15, una condición esencial para garantizar la precisión en los modelos de Machine Learning.

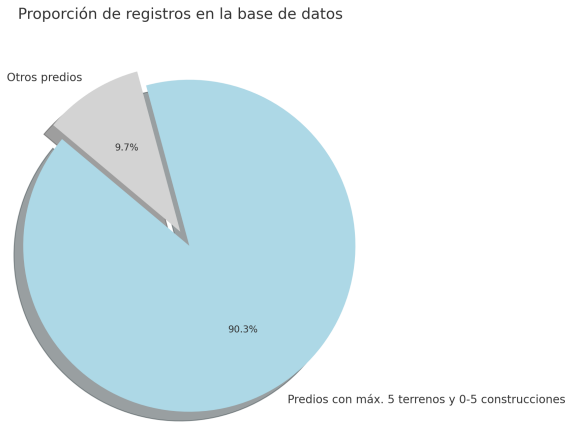


Figura 15: Cantidad de datos obtenidos [14]

La base de datos incluye predios con características más complejas, tales como

propiedades que abarcan hasta cinco terrenos y cinco construcciones por predio, lo que proporciona una diversidad considerable en las características evaluadas. Esta variabilidad en los datos es crucial para capturar un espectro amplio de escenarios y comportamientos en la región de Dagua, enriqueciendo el análisis y la capacidad predictiva del modelo.

4.2.1. Características clave de los datos utilizados

Número de habitaciones, baños, locales y pisos: Estas variables permiten describir la capacidad funcional del predio y su potencial de uso. Por ejemplo, un mayor número de baños y locales suele asociarse a propiedades con un propósito más comercial o residencial de alto valor. El número de pisos, por su parte, indica el nivel de verticalidad de la construcción, una característica relevante en zonas urbanas.

Áreas totales, incluyendo área del terreno y área construida: Estas son variables continuas que definen las dimensiones físicas del predio. El área del terreno mide el espacio total de la propiedad, mientras que el área construida delimita la superficie efectiva utilizada para edificación. Ambas son fundamentales para determinar el valor catastral como se puede apreciar en el gráfico de importancia de variables para el modelo de Random Forest 17 el cual fue realizado posteriormente durante el desarrollo de la investigación,ademas propiedades con mayor superficie construida y terrenos extensos suelen presentar un incremento en su valorización debido a su mayor utilidad o potencial.

Clasificación del uso del predio: Este atributo describe la función principal del predio, categorizándolo en residencial, comercial, industrial, o mixto; los predios comerciales e industriales generalmente poseen un valor superior debido a su mayor rentabilidad y demanda en el mercado. Por ejemplo, propiedades comerciales ubicadas en zonas estratégicas como áreas urbanas o comerciales suelen tener precios más elevados. En el caso de predios mixtos, su versatilidad para combinar actividades residenciales y comerciales les otorga un valor agregado. Los usos residenciales, aunque menos valorados en términos generales, también aportan insights relevantes al modelo, especialmente en zonas de desarrollo urbano, esta información es crucial para establecer diferencias entre propiedades que, pese a compartir características estructurales similares, tienen fines y valores completamente distintos.

Estado físico del predio: La condición física de un predio es una variable cualitativa que influye significativamente en su valoración.

Propiedades en buen estado: Estas suelen tener mayor demanda y, por ende, un valor catastral más alto. Esto incluye propiedades recién construidas, renovadas, o bien mantenidas.

Propiedades deterioradas o en estado de abandono: Estas tienden a presentar valores catastrales reducidos debido a los costos adicionales necesarios para reparaciones o reconstrucciones. Este atributo permite al modelo incorporar la depreciación o apreciación derivada de factores como el tiempo, el mantenimiento y las condiciones climáticas que afectan el estado de la propiedad.

Contexto geográfico: Las zonas físicas y geoeconómicas son variables contextuales que proporcionan información sobre la ubicación y las características económicas del área donde se encuentra cada predio.

Zonas físicas: Estas describen características del entorno, como si el predio está ubicado en áreas urbanas, rurales, planas, o montañosas. Por ejemplo, los terrenos ubicados en zonas rurales tienden a tener menores costos por metro cuadrado que aquellos en áreas urbanizadas.

Zonas geoeconómicas: Estas clasifican los predios según el nivel socioeconómico de la región en la que están situados. Predios en zonas de alto desarrollo económico o comercial tienden a tener valores más altos debido a factores como acceso a infraestructura, proximidad a servicios esenciales, y desarrollo urbano.

4.2.2. Conclusión del análisis de datos

El análisis de los datos para la predicción del avalúo catastral en el municipio de Dagua evidencia la relevancia de integrar variables estructurales, contextuales y cualitativas previamente descritas, y cómo estas influyen de manera directa en la valoración catastral de los predios. Estas características, explicadas detalladamente en las secciones anteriores, no solo mejoran la precisión de los modelos predictivos basados en Machine Learning, sino que también profundizan en la comprensión técnica y contextual de los factores que determinan el valor de las propiedades en la región.

La combinación de atributos físicos del predio, como el número de habitaciones, baños, locales y pisos, junto con el área del terreno y el área construida, per-

mite capturar con precisión las dimensiones y capacidades estructurales de cada propiedad. Estos elementos son fundamentales para modelar las variaciones en el mercado catastral, ya que reflejan las diferencias en funcionalidad, diseño y potencial de uso entre los predios.

Asimismo, la clasificación del uso del predio proporciona una categorización detallada entre usos residenciales, comerciales, industriales y mixtos, destacando cómo estas tipologías impactan de manera diferenciada en la valoración económica. Por ejemplo, los predios con fines comerciales o industriales, ubicados en zonas estratégicas, presentan un mayor valor debido a su potencial de rentabilidad y desarrollo económico.

El contexto geográfico, que incluye las zonas físicas y geoeconómicas, aporta un marco crítico para entender la influencia de la ubicación en el valor de las propiedades. Las propiedades situadas en zonas urbanas o regiones de alto desarrollo económico reflejan un aumento en su valor, debido al acceso a infraestructura, servicios y mercados dinámicos. Por el contrario, los predios rurales o en zonas con menor desarrollo presentan valores relativamente bajos, lo que confirma la importancia de esta dimensión en el análisis.

Este enfoque integrador destaca la capacidad del modelo para capturar las complejidades del mercado catastral, convirtiéndose no solo en una herramienta para predicción, sino también en un recurso estratégico para la toma de decisiones. La posibilidad de emplear este modelo en la planificación de políticas urbanísticas, optimización de recursos y diseño de estrategias económicas reafirma su aplicabilidad en escenarios prácticos.

Finalmente, la inclusión de una **base de datos** [38] diversa y representativa, que abarca el 90.3 % de los predios en Dagua y contiene las características mas esenciales para el contexto de avalúo catastral, subraya el rigor técnico de esta investigación. Este enfoque garantiza que los resultados obtenidos sean escalables y aplicables a otros municipios y regiones con características similares, posicionando este modelo como una solución innovadora y adaptable para abordar retos de valoración catastral en diferentes contextos.

4.3. Selección de características para los modelos

Este enfoque se centró en elegir las más relevantes, permitiendo a los modelos procesar la información con mayor precisión y eliminar posibles sesgos generados por diferencias en las escalas de las variables. Además, este proceso aseguró que los resultados reflejaran fielmente las relaciones entre las características de entrada y el valor catastral, evitando distorsiones en el análisis.

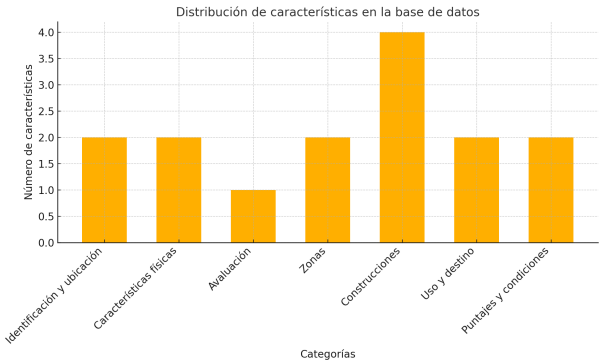


Figura 16: Distribución de las 46 características [14]

Se trabajó con una base de datos que contiene el 90 % de los registros totales (**22,477 predios**) del municipio de Dagua. El número total de columnas y caracteristicas iniciales fue de 46.

4.3.1. Métodos de selección de características

Se probaron diferentes métodos con cada uno de los modelos para determinar cuál ofrecía un mejor desempeño. Entre estos, los métodos **SelectKBest**, R**ecursive Feature Elimination (RFE)** y **Regularización Lasso (L1)** se destacaron por su consistencia en los resultados. Lasso, en particular, mostró ser efectivo para manejar características correlacionadas, eliminando aquellas irrelevantes las cuales pudimos determinar por sus correlaciones y el gráfico de variables de importancia de variables para el modelo de Random Forest 17, esto nos aseguró la retención de las más influyentes en la predicción del valor catastral. Esto resultó en un modelo de Random Forest más simplificado.

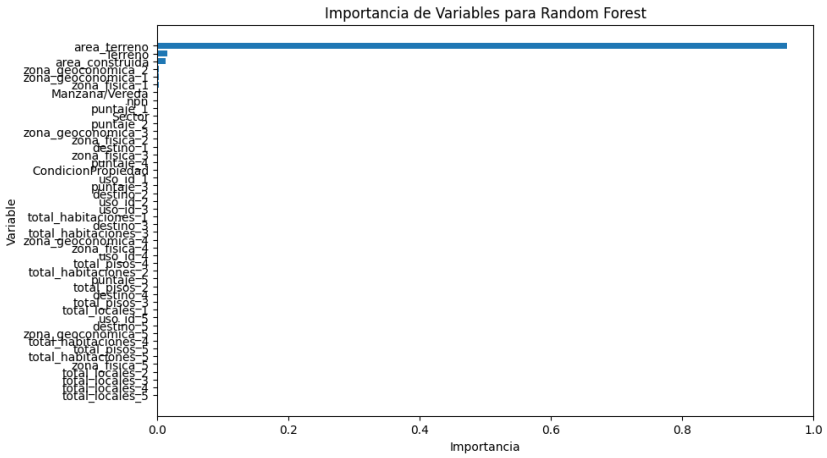


Figura 17: Importancia de variables para el modelo de Random Forest [14]

Resultados de SelectKBest

- **Random Forest**
  - **Características seleccionadas:** puntaje_1, zona_fisica_1, zona_geoconomica_1, area_terreno, area_construida, Manzana/Vereda
    Número de características seleccionadas: 6

- **KNN**
  - **Características seleccionadas:** zona_fisica_1, zona_geoconomica_1, area_terreno, area_construida
    Número de características seleccionadas: 4

- **Ensemble (Random Forest + Gradient Boosting)**
  - **Características seleccionadas:** puntaje_1, zona_fisica_1, zona_geoconomica_1, area_terreno, area_construida, Manzana/Vereda
    Número de características seleccionadas: 6

- **Gradient Boosting**
  - **Características seleccionadas:** puntaje_1, zona_fisica_1, zona_geoconomica_1, area_terreno, area_construida
    Número de características seleccionadas: 5

Resultados de RFE

- RFE fue aplicado a varios modelos para determinar su eficacia, pero se observó que, debido a las limitaciones del método en datasets grandes, su desempeño fue menor comparado con SelectKBest y Lasso

Resultados de Lasso

- Lasso permitió construir modelos más simplificados y eficientes. Su capacidad para manejar características altamente correlacionadas y eliminar redundancias es lo que lo destaca referete a otros métodos.

La selección de características se llevó a cabo mediante métodos iterativos aplicados a los datos de entrenamiento, validación y prueba. Entre los enfoques utilizados, destaca la correlación directa con el objetivo, que permitió identificar variables altamente relacionadas con el valor catastral. Esto ayudó a reducir el conjunto inicial de características, manteniendo únicamente aquellas con una relación significativa. Adicionalmente, se emplearon métodos como SelectKBest con pruebas estadísticas, como f_regression y mutual_inforegression, que identificaron características basadas en la fuerza de su relación con la variable objetivo.

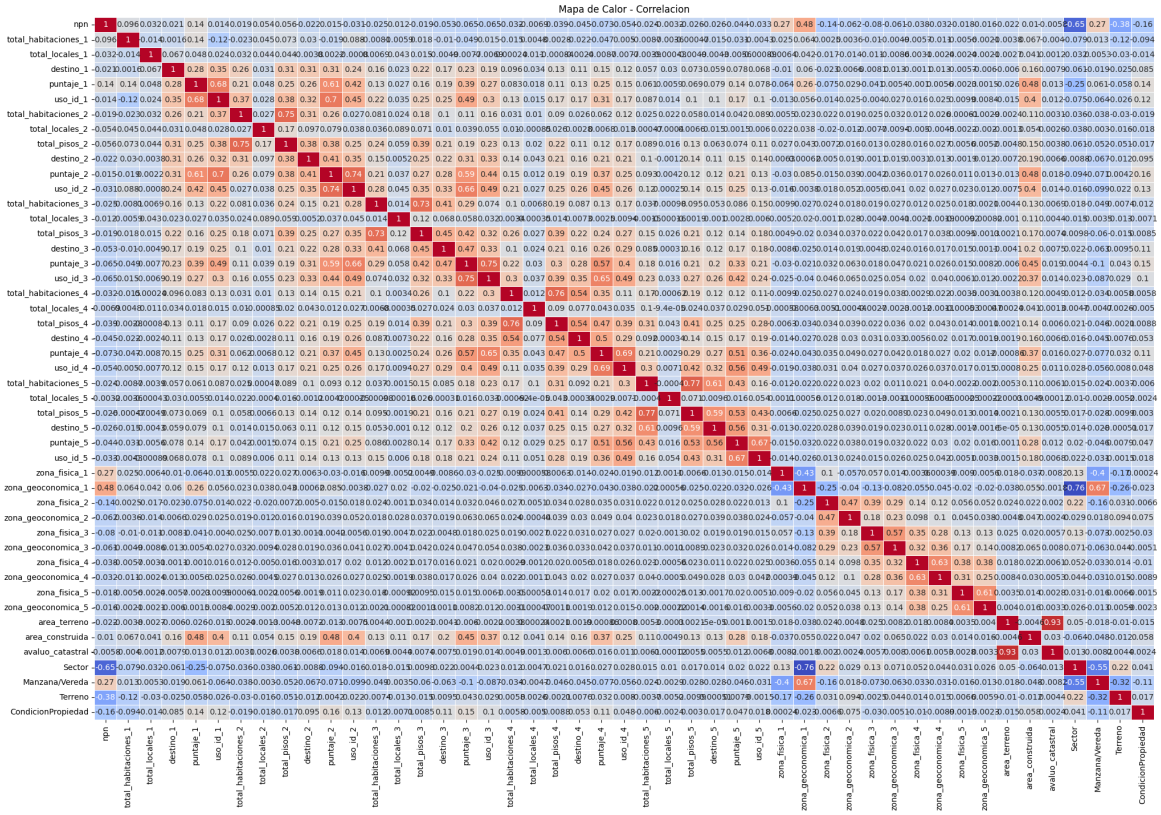


Figura 18: Correlación [14]

Entre los métodos más tradicionales, como el análisis de correlación entre variables como se aprecia en el gráfico de correlación 22, destacó la Regularización Lasso (L1) como el enfoque más efectivo para la selección de características en el modelo de Random Forest debido a que con 6 características igualó la precisión de escoger las 46 características iniciales, se pudo notar que al escoger un numero diferente a 6 el modelo tendió a cometer errores en sus predicciones, la precisión comentada anteriormente se pudo evidenciar en el entrenamiento y resultados previos en la investigación 24.

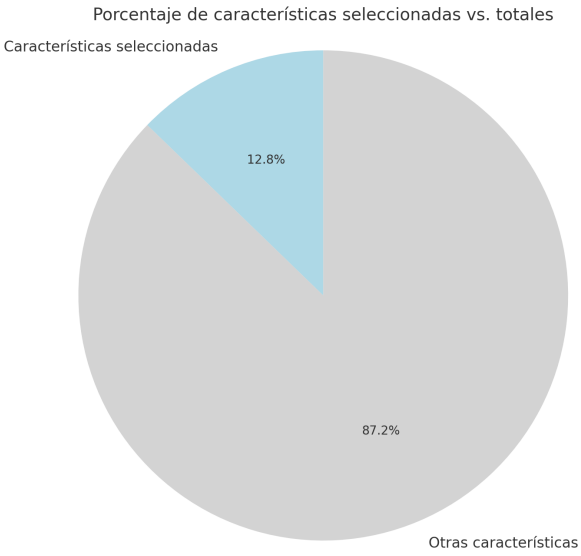


Figura 19: Proporción de características seleccionadas [14]

El usuario final se ve beneficiado de este metodo debido a que en la interfaz, debe digitar una cantidad muy baja con respecto a las características totales, esto se aprecia en el gráfico 19. Ademas este método permitió identificar las variables con mayor impacto en la predicción, eliminando aquellas con coeficientes cercanos a cero y reteniendo únicamente las más relevantes. En este proyecto, las características seleccionadas mediante Lasso incluyeron:

- **Área del terreno**: Una variable crucial para calcular el valor catastral.
- **Área construida**: Indicador directo de la infraestructura del predio.
- **Zona física y geoeconómica**: Reflejan el contexto del predio en términos de ubicación y desarrollo económico.
- **Manzana/Vereda**: Factor relevante que agrupa predios por su localización específica.
- **Puntaje 1**: Una métrica específica que mostró una correlación significativa con el valor catastral.

La capacidad de Lasso para manejar características correlacionadas fue especialmente valiosa, ya que permitió seleccionar una representación óptima en casos de alta multicolinealidad, descartando variables redundantes. Además, la flexibilidad para ajustar el nivel de regularización garantizó un equilibrio adecuado entre la complejidad del modelo y su desempeño predictivo.

Aunque otros métodos, como **SelectKBest** y **RFE**, también ofrecieron valor en etapas iniciales, Lasso sobresalió por su capacidad para integrar la selección de características en el proceso de modelado. Esto simplificó el análisis al generar un modelo más interpretable y eficiente.

4.4. Escalado de características

El escalado de las características seleccionadas complementó este proceso, asegurando que variables como el área del terreno y el área construida contribuyeran equitativamente al aprendizaje de los modelos. Esto fue esencial en modelos como **Random Forest** y **MLP**, donde diferencias en la escala pueden influir en la asignación de importancia y en la convergencia. Este enfoque integral, liderado por Lasso y respaldado por el escalado adecuado, sentó las bases para construir un sistema predictivo sólido y confiable para el avalúo catastral de los predios en Dagua.

5. Entrenamiento de modelos

Para la predicción de avalúos catastrales, se optó por implementar una serie de modelos supervisados de aprendizaje automático debido a su capacidad para adaptarse a datos tabulares y capturar patrones complejos en las características de las propiedades.

index	npn	total_habitaciones_1	total_locales_1	destino_1	puntaje_1	uso_id_1	total_habitaciones_2
0	76233000100000001000100000000	0	0	3	60	83	0
1	76233000100000001000200000000	3	0	1	31	12	0
2	76233000100000001000300000000	0	0	21	40	85	0
3	76233000100000001000400000000	2	0	1	30	12	3
4	76233000100000001000600000000	0	0	9	50	91	0

Figura 20: Previsualización base de datos [14]

uso_id_2	total_habitaciones_3	total_locales_3	total_pisos_3	destino_3	puntaje_3	uso_id_3	total_habitaciones_4	total_locales_4
83	0	0	0	2	40	94	0	0
0	0	0	0	0	0	0	0	0
94	4	0	1	63	46	14	0	0
14	0	0	0	21	40	85	0	0
85	0	0	0	63	62	14	0	0

Figura 21: Continuación de previsualización base de datos [14]

En el transcurso y evolución de la construcción de cada uno de los modelos, se presentaron grandes desafíos que demandaron atención cuidadosa en distintas etapas del proceso. Uno de los aspectos más relevantes fue el tratamiento de los datos, un paso esencial para garantizar su calidad y consistencia. Asimismo, se realizó una evaluación de los modelos para identificar cuál era el más adecuado en términos de precisión, capacidad de generalización y desempeño.

Por otro lado, entre los procesos destacados se encuentra el escalado de datos, que jugó un papel crucial en algunos contextos y modelos específicos. Esta técnica se utilizó como una herramienta clave para transformar las características del conjunto de datos, ajustándolas a un rango uniforme. Este ajuste permitió asegurar que todas las variables contribuyeran de manera equilibrada al proceso de aprendizaje, eliminando posibles sesgos generados por diferencias en magnitudes y optimizando el rendimiento de los modelos en el manejo de datos complejos. La transformación de los datos mediante el escalado no solo mejoró la calidad del aprendizaje, sino que también permitió realizar evaluaciones más homogéneas y representativas.

En este contexto, la partición de datos desempeñó un papel fundamental. Se utilizó una estrategia de división precisa, como se visualiza en la Figura 18, donde el 81 % de los datos se destinó al entrenamiento (X-Train), el 9 % se utilizó para la validación (X-Val) y el 10 % restante se reservó para la prueba final (X-Test).

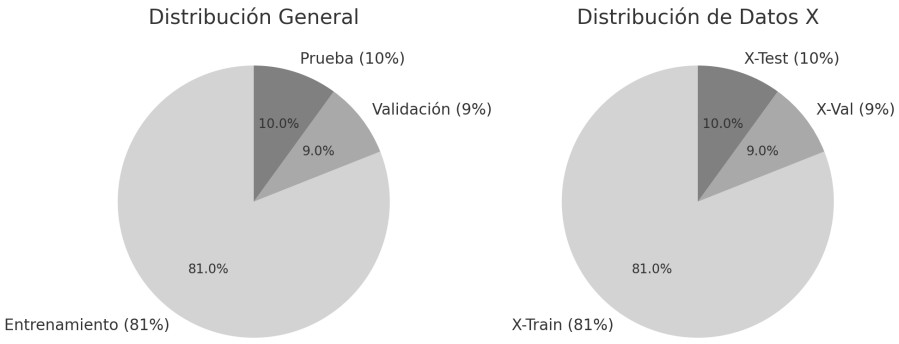


Figura 22: Partición de datos [14]

Esta estrategia asegura que la mayoría de los datos estén disponibles para entrenar el modelo, mientras que una proporción menor se dedica a validar su desempeño y evaluar su capacidad de generalización con datos completamente nuevos.

Esta separación permite medir con precisión y efectividad las predicciones tanto en los datos accesibles para el modelo como en aquellos que nunca antes había visto. La correcta partición de los datos garantiza que los modelos se entrenen con información de alta calidad, proporcionando un marco robusto para evaluar y comparar resultados.

Para optimizar el desempeño de los modelos y mejorar su capacidad para identificar patrones significativos, se realizó un proceso iterativo de selección y escalado de características ampliado en la secciones **4.3** y **4.4**.

Además, el entrenamiento de los modelos incluyó un análisis detallado de los datos para garantizar que las características seleccionadas fueran las más relevantes y aportaran el mayor valor al proceso predictivo, esta selección, permitió reducir la dimensionalidad del conjunto de datos, facilitando la identificación de las variables más importantes para la predicción del valor catastral. Esto mejoró la eficiencia de los modelos y optimizó los tiempos de entrenamiento, al tiempo que se preservaba la precisión en las predicciones.

Por último, es importante resaltar que el enfoque integral adoptado durante el entrenamiento aseguró que los modelos pudieran generalizar de manera efectiva y proporcionar resultados confiables. Este proceso no solo optimizó el rendimiento de los algoritmos en términos de precisión y generalización, sino que también permitió establecer un marco de trabajo replicable para futuras investigaciones relacionadas con la predicción de avalúos catastrales. El uso del escalado, en particular, demostró ser una herramienta esencial para mejorar la capacidad de los algoritmos de procesar datos complejos, garantizando que las evaluaciones fueran justas y representativas.

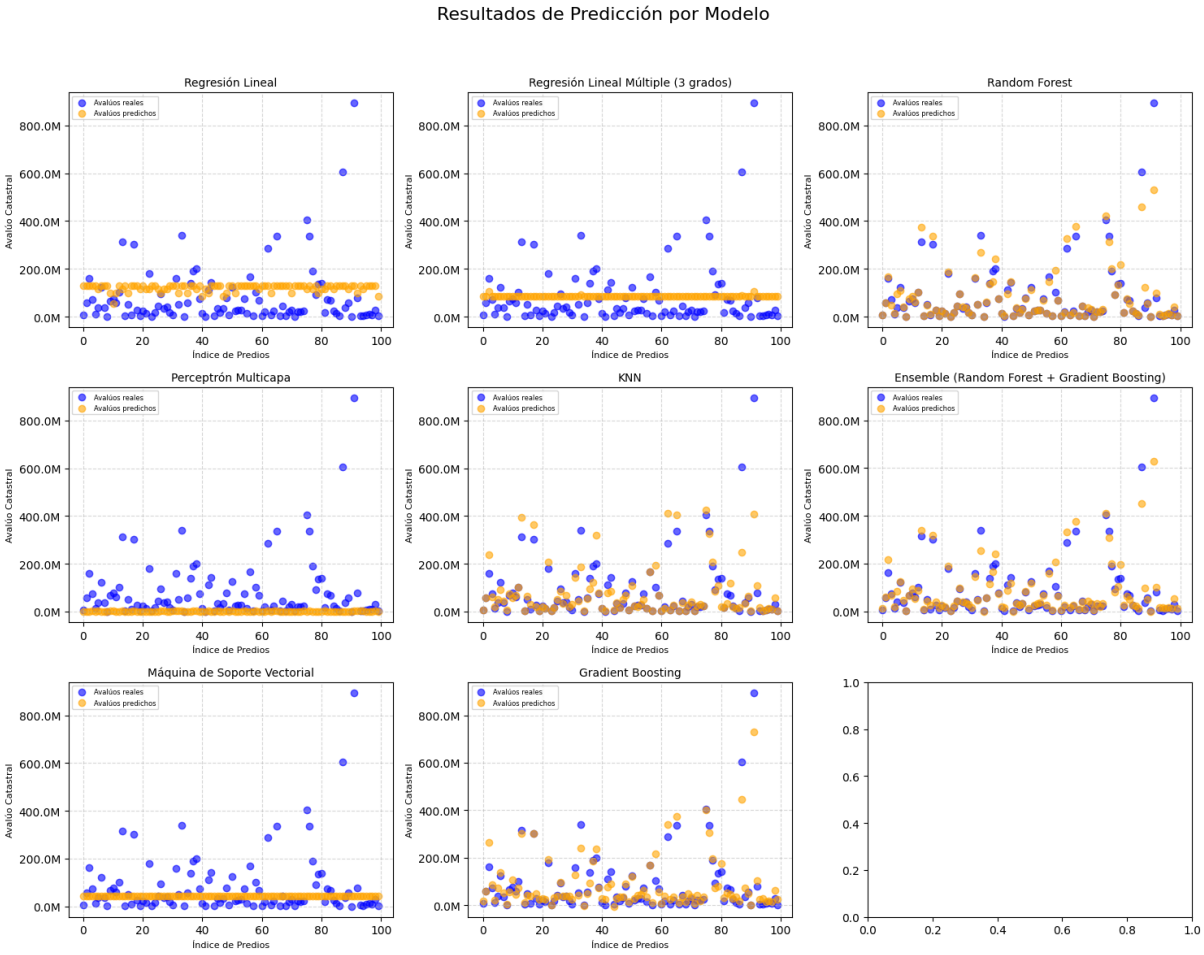


Figura 23: Gráficos comparación prueba base [14]

La gráfica comparativa presentada en la figura 23 respalda estos hallazgos, mostrando cómo random forest, perceptrón multicapa y K-nearest neighbors (KNN) se destacan entre los ocho modelos evaluados. Los puntos naranjas representan los valores predichos por los modelos, mientras que los puntos azules corresponden a los valores reales. En el caso de random forest, los puntos predichos están significativamente más alineados con los valores reales, indicando su alta precisión. Perceptrón multicapa también mostró un ajuste consistente, particularmente en los predios con valores medianos, mientras que KNN destacó por capturar tendencias generales con una dispersión controlada.

La figura 24, que muestra las métricas de resultados de las pruebas base, confirma que tanto como el modelo de Random forest y el modelo de KNN sobresalieron en términos de precisión, capacidad de generalización y desempeño global.

	Modelo	MSE	RMSE	MAE	R2
0	Regresión Lineal	20502695511.5M	143.2M	100.8M	-6.63%
1	Regresión Lineal Múltiple (3 grados)	16978488409.6M	130.3M	80.5M	11.70%
2	Random Forest	2471292297.6M	49.7M	14.5M	87.15%
3	Perceptrón Multicapa	27060706567.9M	164.5M	89.0M	-40.73%
4	KNN	6381879612.2M	79.9M	29.2M	66.81%
5	Ensemble (Random Forest + Gradient Boosting)	2538316692.5M	50.4M	18.8M	86.80%
6	Máquina de Soporte Vectorial	21590927124.3M	146.9M	73.7M	-12.29%
7	Gradient Boosting	3780927634.7M	61.5M	25.6M	80.34%

Figura 24: Métricas de resultados de pruebas base [14]

Como resultado del proceso de entrenamiento, los modelos que son únicos en su técnica que obtuvieron los mejores resultados fueron Random Forest, y K-Nearest Neighbors (KNN), ya que la comparacion de los ocho modelos se logran destacar por su superioridad en términos de precisión, capacidad de generalización y desempeño global.

El modelo Random Forest logró un MSE de 24,712,922.7, con un RMSE de 49.7M y un MAE de 14.5M, alcanzando un R^2 de 87.15 %, mientras que KNN obtuvo un MSE de 6,381,879,612.2, un RMSE de 79.9M, un MAE de 22.9M y un R^2 de 66.81 %.

Sin embargo, al analizar los tiempos de entrenamiento que se encuentran en la siguiente figura 25, KNN fue el más eficiente, con 0.01 segundos, siendo el modelo más rápido en comparación con los demás. Random Forest, en cambio, requirió 49.39 segundos, un tiempo considerablemente mayor pero aún dentro de un rango aceptable para su alto desempeño.

Modelo	Tiempo de Entrenamiento (s)
KNN	0.01
Regresión Lineal	0.34
Regresión Lineal Múltiple (3 grados)	0.67
Gradient Boosting	6.65
Ensemble (Random Forest + Gradient Boosting)	30.87
Máquina de Soporte Vectorial	31.6
Random Forest	49.39
Perceptrón Multicapa	258.84

Figura 25: Tiempos de entrenamiento [14]

En contraste, modelos como Perceptrón Multicapa presentaron tiempos de entrenamiento significativamente más altos (258.84 segundos), lo que puede impactar la escalabilidad del modelo en escenarios con grandes volúmenes de datos.

Estos resultados sugieren que Random Forest es una excelente opción cuando se prioriza precisión y generalización, siendo una alternativa eficiente en contextos donde el tiempo de procesamiento es un factor crítico.

6. Optimización de modelos

6.1. Selección de mejores modelos

La selección de estos modelos fue el resultado de un análisis técnico, basado en métricas cuantitativas obtenidas de los conjuntos de entrenamiento, validación y test, donde se incluyó:

1. Selección de características específicas adaptadas a las particularidades de cada modelo, optimizando su capacidad de aprendizaje.
2. Técnicas de escalado de variables, aplicadas individualmente a cada modelo para garantizar un equilibrio en la contribución de las características.
3. Ajustes de hiperparámetros, realizados mediante una metodología sistemática de mejora continua para cada modelo.

La selección de los tres modelos más destacados se fundamentó en una estrategia de diversificación diseñada para maximizar la robustez y el desempeño global del sistema predictivo. Este enfoque combinó algoritmos con características complementarias, garantizando un análisis integral de los datos desde distintas perspectivas.

El Random Forest fue elegido por su capacidad de manejar datos con alta dimensionalidad y su resistencia a la sobreajuste. Este modelo de ensamble utiliza múltiples árboles de decisión entrenados con subconjuntos aleatorios de datos y de variables, lo que permite capturar relaciones complejas y reducir la varianza en las predicciones. En el contexto de los datos catastrales, esta capacidad resulta particularmente útil, ya que permite modelar relaciones no lineales y gestionar la posible redundancia en las variables. Además, Random Forest proporciona una interpretabilidad considerable al calcular la importancia de cada característica en las predicciones, lo que facilita su aplicabilidad en escenarios donde la comprensión del modelo es clave. No obstante, su costo computacional puede ser elevado cuando se emplean grandes cantidades de árboles, lo que requiere un equilibrio adecuado en la configuración de hiperparámetros.

Por otro lado, el algoritmo K-Nearest Neighbors (KNN) fue seleccionado por su simplicidad y efectividad en la clasificación y regresión de datos catastrales. Su principio de funcionamiento, basado en la similitud entre observaciones mediante distancias euclidianas u otras métricas, permite realizar predicciones basadas en los vecinos más cercanos dentro del espacio de características. Esta metodología es especialmente útil cuando se trabaja con datos que presentan estructuras espaciales o geográficas, como es el caso de la información catastral. Sin embargo, KNN es altamente sensible a la escala de los datos, por lo que se requirió una normalización adecuada para evitar sesgos en las predicciones. Además, su costo computacional puede volverse prohibitivo en conjuntos de datos muy grandes, ya que la búsqueda de los vecinos más cercanos requiere un almacenamiento y procesamiento significativo.

El Perceptrón Multicapa fue elegido por su capacidad de representar relaciones no lineales complejas dentro del conjunto de datos. De los ocho algoritmos probados, siete emplean metodologías de regresión, mientras que el Perceptrón Multicapa es el único basado en redes neuronales. Esta diferencia es clave, ya que los datos catastrales con los que se trabaja presentan patrones altamente no lineales, los cuales pueden ser difíciles de modelar con técnicas tradicionales de regresión.

La arquitectura de una red neuronal permite identificar y modelar interacciones complejas entre variables, lo que mejora la capacidad predictiva cuando se trata de datos con relaciones intrincadas.

Además, este modelo es particularmente sensible a valores atípicos y a datos de gran magnitud. Debido a la naturaleza de los datos catastrales, que pueden incluir valores extremos en variables como extensión de terrenos o avalúo, se requirió una estrategia de escalado adecuada para evitar que estos valores dominen el aprendizaje del modelo. Este fue uno de los factores más determinantes para su correcta implementación. Sin una normalización o estandarización de los datos, el Perceptrón Multicapa podría verse afectado negativamente, dificultando la convergencia del entrenamiento y afectando su desempeño.

En resumen, la incorporación de Random Forest, KNN y Perceptrón Multicapa dentro del conjunto de modelos seleccionados se justificó por su capacidad de capturar diferentes aspectos de los datos, su complementariedad y su adaptabilidad a los patrones intrínsecos de la información catastral. La combinación de estos modelos garantizó una solución robusta y eficiente para el problema de análisis predictivo.

6.2. Métodos de búsqueda de hiperparámetros

Para optimizar los modelos utilizados en la predicción del avalúo catastral de los predios en Dagua, se emplearon las técnicas de búsqueda de hiperparámetros como Grid Search, tras aplicar GridSearch para explorar combinaciones de hiperparámetros, se utilizaron métodos adicionales como RandomizedSearchCV y técnicas avanzadas para ampliar la búsqueda a partir de los mejores resultados obtenidos inicialmente. Este enfoque permitió explorar rangos más amplios y valores alternativos que GridSearch no contempló, optimizando los resultados en casos específicos. El criterio seguido fue iterativo: si los nuevos métodos mejoraban el desempeño, se adoptaban los nuevos hiperparámetros; de lo contrario, se mantenían los mejores resultados previos. En lo que respecta de Grid Search, este es un método que evalúa todas las combinaciones posibles dentro de un espacio de búsqueda predefinido, lo que garantiza precisión, reproducibilidad y transparencia al identificar configuraciones óptimas [39]. Su enfoque sistemático es particularmente útil en modelos como Random Forest y Perceptrón Multicapa (MLP), donde los hiperparámetros afectan significativamente el desempeño.

Por otro lado, tenemos el otro método empleado llamado Randomized Search, este proporciona una alternativa más eficiente al explorar combinaciones de hiperparámetros de manera aleatoria, permitiendo cubrir rápidamente un espacio amplio de posibilidades.

6.2.1. Métodos de búsqueda avanzados

Se implementaron técnicas avanzadas para la optimización de hiperparámetros, las cuales son optimización bayesiana, optuna y automatización mediante hyperOpt. Estas metodologías permitieron explorar y ajustar de manera eficiente los hiperparámetros, logrando mejoras significativas en los resultados de algunos modelos [40].El uso combinado de grid search, randomized Search, optimizacion bayesiana,optuna y automatización mediante hyperOpt asegura una optimización eficiente y una evaluación robusta de los modelos de predicción. Su aplicación incrementó la precisión de las predicciones y optimizó el proceso de ajuste, contribuyendo al desarrollo de modelos más robustos y confiables.

Ahora bien, la optimización bayesiana utiliza un modelo probabilístico para explorar de manera eficiente el espacio de hiperparámetros, guiando la búsqueda hacia configuraciones prometedoras basadas en información obtenida en iteraciones previas [41]. Dicho método tuvo un impacto positivo en modelos como el perceptrón multicapa, donde permitió ajustar parámetros clave como el número de capas ocultas y la tasa de aprendizaje de manera precisa, optimizando el desempeño sin necesidad de exploraciones exhaustivas. Sin embargo, en el caso de KNN, la optimización no mostró mejoras significativas debido a la simplicidad del modelo y su menor dependencia de múltiples hiperparámetros, no obstante, a pesar de estas limitaciones, este método destacó por su capacidad de reducir tiempos de cómputo y mejorar la comprensión del efecto de los hiperparámetros en los resultados.

Por otro lado, optuna se destacó como un marco avanzado para la optimización de hiperparámetros, integrando estrategias adaptativas que priorizan configuraciones prometedoras, bajo este contexto, este método resultó especialmente efectiva en el modelo perceptron multicapa, donde optimizó parámetros como el número de capas ocultas, funciones de activación y tasas de regularización, lo que se tradujo en mejoras notables en métricas como R^2 y RMSE en validación y test; se debe tener tambien en c cuenta , que aunque en Random Forest las mejoras

fueron más limitadas, Optuna demostró su flexibilidad al adaptarse a los espacios de búsqueda y ofrecer configuraciones bien fundamentadas.

Para finalizar, tambien se tuvo en cuenta la automatización con hyperOpt, ya que emplea algoritmos probabilísticos para ajustar hiperparámetros de manera eficiente, adaptándose a espacios de búsqueda más complejos, es importante destacar que este método tuvo un impacto notable en KNN, donde exploró parámetros como el número de vecinos, métricas de distancia y ponderación, logrando configuraciones que incrementaron la consistencia y precisión en validación y test. En Random Forest, HyperOpt contribuyó a identificar pequeños ajustes que mejoraron la generalización del modelo. Sin embargo, en MLP, no superó los resultados obtenidos con Optuna, debido a la mayor sensibilidad de este modelo a la configuración de hiperparámetros complejos que requieren exploraciones más detalladas.

En conjunto, estas técnicas avanzadas de optimización proporcionaron herramientas clave para identificar las configuraciones más sólidas y precisas en los modelos seleccionados (Random Forest, Perceptron multicapa, y KNN).

6.2.2. Análisis optimización de modelos

Los resultados de esta etapa, denota el comportamiento distinto en cuanto a su rendimiento y capacidad de generalización, después de un detallado proceso de optimización de hiperparámetros.

En primer lugar, el modelo de K-Nearest Neighbors (KNN) mostró un rendimiento pre-optimización de 68.42 %. Sin embargo, después de las optimizaciones, que incluyeron técnicas como Grid Search, randomizer, optimus, hyperOpt y optimización bayesiana, se obtuvo una ligera disminución en el desempeño, alcanzando 67.75 %. Este descenso se atribuye a la naturaleza probabilística del algoritmo, que puede verse afectada por las condiciones del conjunto de datos y los parámetros seleccionados, lo que ocasiona variaciones en el rendimiento en los datos de prueba [42]. Por otro lado, el modelo de random forest presentó un rendimiento pre-optimización sobresaliente, con una puntuación de 84.33 %. Sin embargo, las optimizaciones aplicadas, incluyendo todas las técnicas mencionadas, no lograron mejorar su desempeño. A pesar de este estancamiento en la optimización, se mantuvo como el mejor modelo en términos de rendimiento. Esto indica que, sin necesidad de ajustes, random forest ya presentaba una excelente capacidad

de generalización y predicción, superando los otros modelos incluso sin modificaciones, a continuación se encuentran los mejores resultados para cada modelo encontrados durante cada metodo.

Modelo	Hiperparámetros Buscados	Resultados Finales (Hiperparámetros)	Método de Optimización
Random Forest (GridSearchCV)	{'n_estimators': [40, 50], 'max_depth': [None, 30], 'min_samples_split': [2, 4]}	{'max_depth': 30, 'min_samples_split': 2, 'n_estimators': 40}	GridSearchCV
Random Forest (RandomizedSearchCV)	{'n_estimators': [40, 50, 100, 200], 'max_depth': [None, 10, 20, 30, 40], 'min_samples_split': [2, 4, 6, 8]}	{'n_estimators': 200, 'min_samples_split': 2, 'max_depth': 40}	RandomizedSearchCV
KNN (GridSearchCV)	{'n_neighbors': [1, 3, 5, 10, 15, 20, 30, 50], 'weights': ['uniform', 'distance'], ...}	{'algorithm': 'ball_tree', 'leaf_size': 30, 'metric': 'manhattan', ...}	GridSearchCV
KNN (RandomizedSearchCV)	{'n_neighbors': [1, 3, 5, 10, 15, 20, 30, 50], 'weights': ['uniform', 'distance'], ...}	{'metric': 'manhattan', 'n_neighbors': 2, 'weights': 'distance'}	RandomizedSearchCV
KNN (Bayesiana)	{'n_neighbors': (1, 50), 'weights': [...], 'metric': [...]}	OrderedDict([('metric', 'manhattan'), ('n_neighbors', 1), ('weights', 'uniform')])	Bayesiana
Perceptrón Multicapa (GridSearchCV)	{'hidden_layer_sizes': [(100, 50), (100, 50, 25), (150, 75)], 'activation': [...], ...}	{'activation': 'tanh', 'alpha': 0.01, 'hidden_layer_sizes': (150, 75), ...}	GridSearchCV
Perceptrón Multicapa (RandomizedSearchCV)	{'hidden_layer_sizes': [(100, 50), (100, 50, 25), (150, 75), (100, 100, 50)], ...}	Varios hiperparámetros con errores, mejor score R^2: 0.561916	RandomizedSearchCV
Perceptrón Multicapa (Bayesiana)	{'hidden_layer1': (50, 200), 'hidden_layer2': (20, 100), 'hidden_layer3': (50, 200), ...}	Mejores hiperparámetros bayesianos para MLP	Bayesiana
Random Forest (Optuna)	{'n_estimators': Int(50, 300), 'max_depth': Categorical([None, 10, 20, 30, 40, 50]), ...}	{'n_estimators': 215, 'max_depth': 40, 'min_samples_split': 15, ...}	Optuna
KNN (Optuna)	{'n_neighbors': Int(1, 50), 'weights': Categorical(['uniform', 'distance']), ...}	{'n_neighbors': 2, 'weights': 'uniform', 'metric': 'euclidean'}	Optuna
Perceptrón Multicapa (Optuna)	{'hidden_layer_sizes': Categorical([(50,), (100,), (100,50), (150,)]), ...}	{'hidden_layer_sizes': (50,), 'activation': 'logistic', 'solver': 'lbfgs', ...}	Optuna
KNN (Hyperopt)	{'n_neighbors': quniform(2, 5, 1), 'weights': choice(['uniform', 'distance']), ...}	Mejores combinaciones: n_neighbors=2, weights='distance', metric='manhattan', ...	Hyperopt
Random Forest (Hyperopt)	{'n_estimators': quniform(150, 250, 10), 'max_depth': quniform(30, 50, 5), ...}	Mejores combinaciones de RF: n_estimators ≈ 230, max_depth ≈ 40, ...	Hyperopt
Perceptrón Multicapa (Hyperopt)	{'hidden_layer1': uniform(50, 200), 'hidden_layer2': uniform(20, 100), ...}	Mejores hiperparámetros MLP bayesianos: {'alpha': ~0.023, 'hidden_layer1': 76, ...}	Hyperopt

Figura 26: Resultados de métodos de búsqueda de hiperparametros [14]

En el anterior gráfico 26 se encuentran el rango de búsqueda y los resultados de cada búsqueda de hiperparametros realizada para cada modelo.

Finalmente, el modelo de perceptrón multicapa, aunque inicialmente presentó un bajo rendimiento de 21.7 %, fue el único modelo en el que las optimizaciones y ajustes de hiperparámetros mostraron una mejora significativa. Con los ajustes realizados, el modelo logró una puntuación de 42.33 %, lo que, aunque aún inferior a los otros modelos, representó una mejora considerable. A pesar de no ser el modelo más preciso, el perceptrón multicapa demostró ser el único que aprovechó eficazmente las optimizaciones de hiperparámetros, lo que resalta su capacidad para adaptarse a los datos de manera más efectiva que KNN y random forest bajo las condiciones experimentales.

Modelos	Hiperparámetros	Método seleccionado
Random forest	n_estimators=100, random_state=42	Ninguno
KNN	n_neighbors=5, weights='distance'	Ninguno
Perceptrón multicapa	hidden_layer_sizes: (68, 88, 82), alpha: 0.06938, learning_rate_init: 0.09123, max_iter: 1000, solver: adam, random_state: 42	Hyperopt

Cuadro 1: Selección de métodos de optimización para diferentes modelos

En conclusión, aunque random forest se destacó por su rendimiento superior, el perceptrón multicapa mostró el mayor progreso después de la optimización como se puede ver en el cuadro 1 lo que subraya la importancia de un análisis detallado de cada modelo y la relevancia de los ajustes de hiperparámetros. Todos los resultados presentados corresponden al promedio de los resultados de los modelos obtenidos en los conjuntos de entrenamiento, validación y prueba, brindando una evaluación más equilibrada y robusta del rendimiento de los modelos.

7. Evaluación de resultados

La evaluación de los resultados se llevó a cabo mediante la prueba de los tres modelos seleccionados, utilizando conjuntos de entrenamiento idénticos para garantizar la comparabilidad. Cada modelo fue refinado con los métodos específicos aplicados durante su optimización, maximizando su desempeño y permitiendo un análisis detallado de sus capacidades. Posteriormente, se analizaron las métricas de rendimiento y se examinaron las gráficas generadas para conseguir una visión general de los resultados. Finalmente, tras este proceso de evaluación, se seleccionó Random Forest como el modelo más adecuado, destacándose por su desempeño superior en términos de precisión, capacidad de generalización y robustez en la predicción del avalúo catastral, a continuación se muestra el cuadro 2 con los resultados de los 3 modelos.

Modelo	MSE	RMSE	MAE	R2
Random forest	2471292297.6M	49.7M	14.5M	87.15 %
KNN	6265851401.8M	79.2M	28.4M	67.41 %
Perceptrón multicapa	9302803570.8M	96.5M	54.9M	51.62 %

Cuadro 2: Resultados de los mejores 3 modelos

Random Forest: Este modelo presentó el mejor desempeño entre los evaluados, destacándose en todas las métricas clave. En el conjunto de validación, obtuvo un coeficiente de determinación (R^2) de 81.51 %, lo que evidencia su capacidad para explicar gran parte de la variabilidad en los datos de entrenamiento. En el conjunto de prueba, este valor aumentó a 87.15 %, demostrando una notable capacidad de generalización a datos no vistos. Además, los valores de MSE

(2,471,292,297.6M), RMSE (49.7M) y MAE (14.5M) fueron los más bajos en comparación con los otros modelos, lo que indica que Random Forest produce predicciones significativamente más precisas y con menor error.

K-Nearest Neighbors (KNN): Aunque este modelo mostró un desempeño razonable, no logró alcanzar los niveles de precisión y generalización observados en Random Forest. En el conjunto de validación, obtuvo un R^2 de 69.43 %, y en el conjunto de prueba, este valor descendió a 67.41 %, reflejando dificultades para capturar relaciones complejas en los datos. Sus métricas de error, incluyendo MSE (6,265,851,401.8M), RMSE (79.2M) y MAE (28.4M), fueron considerablemente más altas, lo que indica menor precisión y una mayor desviación en sus predicciones.

Perceptrón Multicapa (MLP): A pesar de los esfuerzos realizados para optimizar este modelo, sus métricas estuvieron muy por debajo de las obtenidas por los otros modelos. En el conjunto de validación, el R^2 fue de 33.04 %, y aunque aumentó a 51.62 % en el conjunto de prueba, el modelo no demostró una generalización adecuada a nuevos datos. Además, los valores de MSE (9,302,803,570.8M), RMSE (96.5M) y MAE (54.9M) fueron los más altos, lo que evidenció un rendimiento insuficiente en términos de precisión y confiabilidad.

Random Forest fue seleccionado como el modelo final debido a su sobresaliente desempeño en términos de precisión, robustez y capacidad de generalización. Su R^2 alto en los conjuntos de validación (81.51 %) y prueba (87.15 %) demuestra que es capaz de explicar con precisión las relaciones complejas entre las características de entrada y el valor catastral, incluso en datos nuevos. Además, sus métricas de error significativamente más bajas (MSE, RMSE y MAE) consolidan su posición como la opción más confiable y precisa para las tareas de predicción. Aunque KNN mostró un desempeño aceptable, no alcanzó la precisión ni la capacidad de generalización necesarias para este proyecto. Por otro lado, el Perceptrón Multicapa (MLP) fue uno de los modelos que más evolucionó a lo largo del proyecto, mejorando significativamente tras aplicar escalado a los datos y realizar una búsqueda profunda de hiperparámetros. Estos ajustes incluyeron la optimización del número de neuronas, tasas de aprendizaje y funciones de activación. Sin embargo, a pesar de estas mejoras, el MLP no logró superar a Random Forest ni a KNN en métricas clave como R^2, MSE y RMSE, evidenciando limitaciones en su capacidad de generalización. En consecuencia, aunque mostró progreso notable, no alcanzó el nivel de desempeño requerido para ser

seleccionado como modelo final.

En conclusión, la elección de Random Forest como modelo final se fundamenta en su capacidad para generar predicciones precisas y generalizables, superando significativamente a los otros enfoques evaluados. Este modelo no solo cumple con los requisitos técnicos del proyecto, sino que también ofrece una solución confiable y robusta para la predicción del avalúo catastral en los predios analizados.

8. Interfaz gráfica

Para el diseño de la interfaz, el modelo se serializó utilizando el mismo conjunto de datos empleado durante su entrenamiento, asegurando que los resultados fueran consistentes con las métricas obtenidas en las pruebas iterativas realizadas a lo largo del proyecto. Una vez finalizado el proceso de serialización y validación, el modelo fue alojado en Google Drive y desplegado en una aplicación diseñada específicamente para pruebas por parte de usuarios con conocimientos especializados en las características de los predios, como los expertos del IGAC. Esta aplicación permite ingresar las características necesarias para el modelo, logrando un nivel de precisión del 85 % en las predicciones, según las evaluaciones realizadas durante las pruebas iterativas realizadas durante el proyecto.

La aplicación “Estimador de Avalúo Catastral” mostrada en la figura 27, fue diseñada para calcular de manera precisa y eficiente el avalúo catastral de predios en el municipio de Dagua, empleando

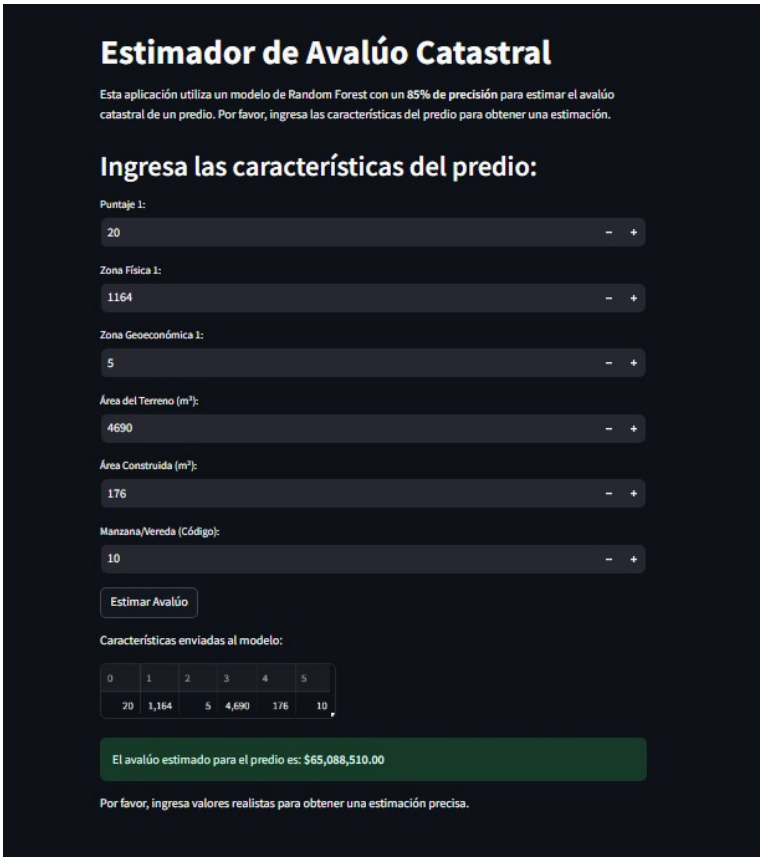


Figura 27: Interfaz de la aplicación multiplataforma [14]

un modelo de aprendizaje automático basado en Random Forest, previamente entrenado y optimizado. Este modelo, con una precisión del 85 %, se serializó utilizando Joblib y se integró en una interfaz interactiva desarrollada con la librería Streamlit, la cual permite una experiencia de usuario simple y efectiva. La aplicación no solo destaca por su simplicidad, sino también por ser multiplataforma, lo que garantiza su accesibilidad desde cualquier dispositivo con conexión a Internet. Además, la integración con `pyngrok` asegura la creación de túneles HTTP estables y seguros, facilitando el acceso remoto a la aplicación incluso cuando se ejecuta en entornos locales. El túnel genera un enlace público que permite a los usuarios interactuar con la herramienta en tiempo real. La aplicación requiere que el usuario, típicamente un funcionario del IGAC, ingrese características clave del predio, como el puntaje asociado, la zona física y geoeconómica, el área del terreno, el área construida y el código de manzana o vereda. Estos datos se procesan en tiempo real, generando una predicción monetaria precisa del avalúo catastral.

Streamlit fue elegido por su capacidad para facilitar la integración de modelos de aprendizaje automático, su simplicidad para construir interfaces dinámicas, su claridad en la presentación de resultados y su compatibilidad multiplataforma. Esta herramienta ofrece importantes beneficios para los funcionarios del IGAC, al automatizar los cálculos, garantizar resultados consistentes y proporcionar un recurso accesible, estable y portátil, útil tanto en campo como en oficina, mejorando significativamente los procesos de avalúo y gestión catastral.

9. Conclusiones

Los objetivos planteados al inicio del proyecto fueron cumplidos de manera efectiva, logrando resultados significativos en cada etapa del desarrollo:

Se preparó la base de datos catastral del municipio de Dagua utilizando técnicas de exploración de datos, garantizando la calidad y consistencia de los registros.

Se entrenaron varios modelos de aprendizaje automático, adecuados para la predicción del avalúo catastral en el municipio de Dagua, evaluados en función de su rendimiento y precisión.

Los modelos fueron optimizados mediante el ajuste de hiperparámetros, logrando mejoras significativas en su desempeño.

Finalmente, se evaluó la eficacia de los modelos construidos, utilizando métricas como la precisión y el error cuadrático medio (MSE).

Este análisis se enfocó en el problema de la predicción precisa de los avalúos catastrales, una tarea esencial para optimizar procesos de valoración predial y mejorar la toma de decisiones en sectores públicos e incluso privados. Nuestra propuesta de solución frente a este reto consistió en desarrollar un modelo predictivo utilizando técnicas avanzadas de aprendizaje automático, con énfasis en la selección de características relevantes y el preprocesamiento adecuado de los datos.

Para la construcción de los modelos, se utilizó un conjunto de datos compuesto por variables relacionadas con las características físicas y geoeconómicas de los predios ubicados en el municipio de Dagua. Los algoritmos empleados, como Random Forest, Support Vector Machines y Redes Neuronales, fueron evaluados mediante métricas clave, destacándose Random Forest con una precisión del 87.15 % y un

MSE reducido tras la aplicación de selección de características y escalado de datos.

El análisis de los resultados confirmó que la combinación de selección de características y escalado de datos tuvo un impacto significativo en el rendimiento de los modelos. Las variables como “área construida”, “zona física”, “zona geoeconómica”, “área de terreno”, “manzana-vereda” y “puntaje” demostraron una alta correlación con los valores catastrales, siendo determinantes en las predicciones.

La elección del modelo final consideró tanto la capacidad de interpretación como la aplicabilidad en ese, el modelo de perceptrón multicapa comenzó con un R extsuperscript2 muy bajo, incluso negativo. Sin embargo, a medida que pasó por cada uno de los procesos de preprocesamiento y optimización, su R extsuperscript2 y otras métricas mejoraron de manera constante. Esto le permitió posicionarse como uno de los tres mejores modelos seleccionados, aunque no alcanzó el desempeño del modelo Random Forest, que se mantuvo como el mejor a lo largo de todo el análisis.

En síntesis, este proyecto reafirma el potencial del aprendizaje automático para abordar problemas complejos como la predicción de avalúos catastrales. Los resultados obtenidos no solo cumplieron los objetivos planteados, sino que también establecieron un marco robusto y replicable para investigaciones futuras en el área inmobiliaria.

10. Trabajos futuros

Este proyecto está enfocado en la utilización de datos específicos en un entorno catastral, teniendo en cuenta que incluye información clave y detallada de cada uno de los predios. Se debe considerar que la información utilizada como referencia para el posible entrenamiento del modelo proviene de un análisis exhaustivo y de la combinación de múltiples bases de datos manejadas por la Unidad Administrativa Especial de Catastro, en conjunto con la información del IGAC. Esto permite obtener detalles fundamentales sobre los predios ubicados en el municipio de Dagua, Valle.

Es importante resaltar que este modelo no incluye información recolectada sobre las fluctuaciones económicas de los inmuebles en el sector, ya que desde el inicio

de la investigación se especificó que los datos fueron recolectados y obtenidos exclusivamente de fuentes oficiales gubernamentales especializadas para la toma de decisiones a nivel departamental.

En cuanto al enfoque y alcance de la investigación, este proyecto se centra en el desarrollo de un modelo predictivo para los avalújos catastrales mediante el uso de técnicas avanzadas de aprendizaje automático. Aunque el alcance está delimitado al municipio de Dagua, Valle, este estudio establece una base técnica que podría adaptarse a contextos más amplios. Se recomienda, como trabajo futuro, la extensión del modelo a un nivel nacional mediante la integración de datos adicionales y la evaluación de nuevas estrategias de selección de características que permitan ajustar el modelo a las particularidades de cada región.

Si bien este proyecto tiene un enfoque delimitado, las metodologías empleadas, como el preprocesamiento de datos y la selección de variables clave, ofrecen una estructura técnica robusta para incrementar la eficacia de las predicciones en diversos contextos. Además de abordar los desafíos actuales de predicción, este trabajo establece un cimiento técnico que puede facilitar el desarrollo de modelos predictivos más avanzados y escalables, con potencial de aplicación tanto a nivel departamental como nacional.

Referencias

[1] I. G. A. Codazzi, "Avalúos técnicos: Determinación del valor de bienes muebles e inmuebles," *Mandatario Avalúos*, 2024. [Online]. Available: https://www.igac.gov.co/es/contenido/mandatario-avaluos

[2] U. administrativa especial de catastro Gobernación del Valle del Cauca, "Apoyo en el desarrollo de las actividades necesarias para la actualización catastral con enfoques multipropósito en los diferentes municipios focalizados, para la implementación de la política pública de catastro multipropósito," *The Conversation*, 2022. [Online]. Available: https://repository.udistrital.edu.co/bitstream/handle/11349/30965/TRABAJO%20DE%20GRADO%20VANESSA%20TORRES%20RUIZ%20DOC%20FINAL.pdf?sequence=2&isAllowed=y

[3] A. E. C. FAJARDO, "Propuesta metodológica para calcular el avalúo de un predio empleando redes neuronales artificiales," 2014. [Online]. Available: https://repositorio.unal.edu.co/bitstream/handle/unal/54464/51964148.2015.pdf.pdf?sequence=1

[4] V. C. T. Ruiz, "Unidad administrativa especial de catastro," *The Conversation*, 2023. [Online]. Available: https://www.valleavanza.com/gestor-catastral/

[5] "ISO19152:2012 (LADM-COL)," Proyecto Catastro Multipropósito, 15 de Noviembre de 2017, Colombia. [Online]. Available: https://www.proadmintierra.info/wp-content/uploads/2018/04/ladmcol.pdf

[6] D. A. N. de Estadística (DANE), "Censo nacional de población y vivienda 2018," https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/censo-nacional-de-poblacion-y-vivenda-2018, 2018.

[7] "RESOLUCIÓN NÚMERO 0070 DE 2011," Instituto Geográfico "Agustín Codazzi", 4 de Febrero del 2011, Colombia.

[8] V. A. S.A.S. (2023) Informe final. [Online]. Available: https://drive.google.com/file/d/1zyI4APIjCDEO0jtqcFPh4U195XcoDDdO/view?usp=sharing

[9] E. Bermúdez-Martínez, "La reglamentación del avalúo catastral y las razones para el incremento del impuesto predial," *Revista de Investigación, Desarrollo e Innovación*, vol. 12, no. 24, 2015. [On-

line]. Available: https://repository.ucatolica.edu.co/entities/publication/fca5c928-2152-4807-9c94-bea3fa9da997

[10] D. Dorado, "El catastro multipropósito: Reflexiones alrededor de su potencialidad y aplicación," *Revista de Investigación, Desarrollo e Innovación*, vol. 12, no. 24, 2022. [Online]. Available: https://repositoriocdim.esap.edu.co/handle/123456789/26087

[11] Instituto Geográfico Agustín Codazzi (IGAC). (2025) ¿qué es el número predial nacional y ya está vigente en todo el país? [Online]. Available: https://antiguo.igac.gov.co/es/contenido/que-es-el-numero-predial-nacional-ya-esta-vigente-en-todo-el-pais

[12] A. Khuri, "Introducción al análisis de regresión lineal, quinta edición por douglas c. montgomery, elizabeth a. peck, g. geoffrey vining," *Revista Estadística Internacional*, vol. 81, pp. 318–319, 2013. [Online]. Available: https://doi.org/10.1111/insr.12020_10

[13] J. I. López-Naranjo, "Análisis de regresión y su aplicación en las ciencias agropecuarias," *Revista de Investigación, Desarrollo e Innovación*, vol. 12, no. 24, 2022. [Online]. Available: https://ri.ujat.mx/bitstream/200.500.12107/3983/1/Ana%CC%81lisis%2Bde%2Bregresio%CC%81n%2Baplicado.pdf

[14] J. J. M. Stefania Hurtado, "Modelo de avalúo catastral en el municipio de dagua," Disponible en: https://colab.research.google.com/drive/1YTk85ZGjE-dJPiwi0Jp2UYvIPBDmLiAh?usp=sharing, 2025, accedido: enero 7, 2025.

[15] U. de Granada, "Modelos de regresión lineal múltiple," https://www.ugr.es/~montero/matematicas/regresion_lineal.pdf, n.d., accedido el 9 de enero de 2025.

[16] M. L. C. Team. (2025) Gradient boosting machines - machine learning course. [Online]. Available: https://mlcourse.ai/book/topic10/topic10_gradient_boosting.html

[17] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

[18] A. U. or Not Listed, "A generalized decision tree ensemble based on the neuralnetworks architecture: Distributed gradient boosting forest

(dgbf)," *arXiv preprint arXiv:2402.03386*, 2024. [Online]. Available: https://arxiv.org/abs/2402.03386

[19] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine Learning*, vol. 6, no. 1, pp. 37–66, 1991. [Online]. Available: https://doi.org/10.1007/BF00153759

[20] P. Hitek. (2025) Modelado matemático del algoritmo knn (k-nearest neighbors). [Online]. Available: https://panamahitek.com/modelado-matematico-del-algoritmo-knn-k-nearest-neighbors/?utm_source=chatgpt.com#google_vignette

[21] J. F. C. Castañeda, "Utilización de las máquinas con vectores de soporte para regresión m2 de construcción en bogotá," https://www.academia.edu/57641925/Utilizaci%C3%B3n_de_las_m%C3%A1quinas_con_vectores_de_soporte_para_regresi%C3%B3n_m2_de_construcci%C3%B3n_en_Bogot%C3%A1, 2021.

[22] H. Adams, E. Farnell, and B. Story, "Support vector machines and radon's theorem," *arXiv preprint arXiv:2011.00617*, 2020.

[23] D. Calvo. (2025) Clasificación de redes neuronales artificiales. [Online]. Available: https://www.diegocalvo.es/clasificacion-de-redes-neuronales-artificiales/#google_vignette

[24] E. Sánchez and A. Alanís, "Redes neuronales: Conceptos fundamentales y aplicaciones a control automático," *Madrid: Pearson Educación*, 2006.

[25] F. Ceballos, L. E. Muñoz, and J. Moreno Cadavid, "Selección de perceptrones multicapa usando aprendizaje bayesiano," *Redalyc: Tecnologías de la información y las comunicaciones*, vol. 5, no. 2, pp. 19–32, 2012.

[26] Q. O. S. Patricia, "Metodología de clasificación física para el avalúo masivo de terrenos de predios rurales en un catastro multipropósito," 2017. [Online]. Available: https://repositorio.unal.edu.co/handle/unal/63062

[27] J. P. Carranza, M. A. Piumetto, C. M. Lucca, and E. Da Silva, "Mass appraisal as affordable public policy: Open data and machine learning for mapping urban land values," *Land Use Policy*, vol. 119, p. 106211, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0264837722002381

[28] B. Trawiński, Z. Telec, J. Krasnoborski, M. Piwowarczyk, M. Talaga, T. Lasota, and E. Sawiłow, "Comparison of expert algorithms with machine lear-

ning models for real estate appraisal," in *2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, 2017, pp. 51–54.

[29] M. Jarosz, M. Kutrzyński, T. Lasota, M. Piwowarczyk, Z. Telec, and B. Trawiński, "Machine learning models for real estate appraisal constructed using spline trend functions," in *Intelligent Information and Database Systems*, N. T. Nguyen, K. Jearanaitanakij, A. Selamat, B. Trawiński, and S. Chittayasothorn, Eds. Cham: Springer International Publishing, 2020, pp. 636–648.

[30] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Amsterdam, Netherlands: Morgan Kaufmann, 2011.

[31] J. J. Marín and S. Hurtado, "Base de datos r1t," February 2024. [Online]. Available: https://drive.google.com/file/d/1-B0xJGdR78q4n5QGtZmJ1iRYdQVHvzbl/view?usp=sharing

[32] ——, "Base de datos r1c," February 2024, google Drive, [Accessed: Feb. 11, 2025]. [Online]. Available: https://drive.google.com/file/d/1-9D-sx10YOMRlld6fPyfSs967-1AfLpc/view?usp=sharing

[33] ——, "Base de datos r1d," February 2024. [Online]. Available: https://docs.google.com/spreadsheets/d/1L_fbZ_139W0vN-h5AD3UvI81lz-QKdcs/edit?usp=sharing&ouid=109231314181582979565&rtpof=true&sd=true

[34] ——, "Base de datos r2b," February 2024. [Online]. Available: https://docs.google.com/spreadsheets/d/1-C1p6wqb51ituvpn1-r58T2zL826ZwR5/edit?usp=sharing&ouid=109231314181582979565&rtpof=true&sd=true

[35] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[36] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[37] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[38] J. J. Marín and S. Hurtado, "Base de datos resultante," Diciembre 2024. [Online]. Available: https://docs.google.com/spreadsheets/d/1NQUqvCY6vfk6YNp8QmYOkYQcXyBeGpVR/edit?usp=sharing&ouid=109231314181582979565&rtpof=true&sd=true

[39] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. Feb, pp. 281–305, 2012.

[40] Desconocido, "Hyperopt: ajuste de hiperparámetros basado en optimización bayesiana," https://ichi.pro/es/hyperopt-ajuste-de-hiperparametros-basado-en-optimizacion-bayesiana-140338828128041 2023, accedido: 10-ene-2025.

[41] ——, "Optimización bayesiana para hiperparámetros," https://cienciadedatos.net/documentos/62_optimizacion_bayesiana_hiperparametros, 2023.

[42] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.