Impacto de la actualización catastral en el avalúo catastral: Evaluación mediante la implementación de Machine Learning

Andrés Felipe Pinilla Jiménez

Asesora

Dayana Alejandra Barrera Buitrago

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2024

Nota de Aceptación

_____

Nombre Director de Trabajo de Grado

_____ _____

Jurado Jurado

2024

Agradecimientos

A mis padres, Alberto y Nelly, quienes me enseñaron a terminar lo que he empezado.

Resumen

Colombia cuenta con un catastro a nivel municipal que, en términos generales, se encuentra desactualizado. En el marco de los acuerdos de paz firmados en 2016 con las FARC, en cumplimiento del punto 1 – Reforma Rural Integral - el gobierno actual (2022-2026) ha impulsado la actualización de este catastro como parte del enfoque multipropósito para la formalización de la propiedad rural. Este estudio analiza los efectos de la actualización catastral sobre el avalúo en 89 municipios actualizados entre 2016 y 2024. Para ello, se compara la información catastral de estos municipios antes y después del proceso de actualización.

El análisis utiliza una metodología basada en machine learning que combina técnicas de segmentación mediante HDBSCAN y modelos de regresión con XGBoost. La segmentación permitió identificar grupos homogéneos de predios según características compartidas, mientras que los modelos predictivos se aplicaron a cada grupo para estimar los incrementos en el avalúo catastral derivados de la actualización. Las variables consideradas incluyen los valores catastrales y las características físicas – extensión, área construida y zona - de los predios antes y posterior a la actualización. Este enfoque metodológico demostró ser suficiente en la identificación de patrones de actualización y en la predicción de los incrementos en avalúo en los predios ubicados en la zonas urbanas.

***Palabras claves*:** Actualización Catastral, Impuesto Predial, Política Tributaria, Modelos de Precios Hedónicos

Abstract

Colombia has a municipal-level cadastre that is outdated. As part of the peace agreements signed in 2016 with the FARC, specifically under Point 1 – Comprehensive Rural Reform – the current government (2022-2026) has promoted the update of this cadastre as part of the multipurpose cadastre approach to formalize rural property. This study analyzes the effects of cadastral updates on property valuation in 89 municipalities updated between 2016 and 2024. To achieve this, cadastral information from these municipalities before and after the update process is compared.

The analysis employs a machine learning-based methodology that combines segmentation techniques using HDBSCAN and regression models with XGBoost. Segmentation identified homogeneous groups of properties based on shared characteristics, while predictive models were applied to each group to estimate increases in cadastral valuation resulting from the update. Variables considered include cadastral values and physical characteristics—such as land size, constructed area, and location—before and after the update. This methodological approach proved effective in identifying update patterns and predicting valuation increases for properties located in urban areas.

***Keywords*:** Cadastral Update, Property Tax, Tax Policy, Hedonic Pricing Models.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Introducción

El uso del suelo constituye una forma clave de datos catastrales. El término "uso del suelo" varía entre países, y existen diversas formas de registrar esta información. Sin embargo, la importancia de disponer de datos precisos sobre el tipo de uso del suelo ha sido resaltada repetidamente en la literatura sobre el tema. Este tipo de información es fundamental para el cálculo de impuestos, respalda la planificación urbana, influye en el valor de los bienes inmuebles y afecta los procedimientos de gestión territorial. La razón principal para una gestión sostenible de las ciudades radica en la verificación sistemática y exhaustiva de la credibilidad de los datos utilizados para la base impositiva, especialmente la información sobre el uso del suelo contenida en el catastro. Para el caso de Polonia, y para todo territorio que carezca de información catastral actualizada, las investigaciones realizadas revelaron que, en muchos casos, estos datos son poco confiables y están desactualizados, lo que ha generado importantes pérdidas fiscales en los presupuestos de las unidades de gobierno local a lo largo de los años (Cienciała, Sobolewska-Mikulska, & Sobura, 2021).

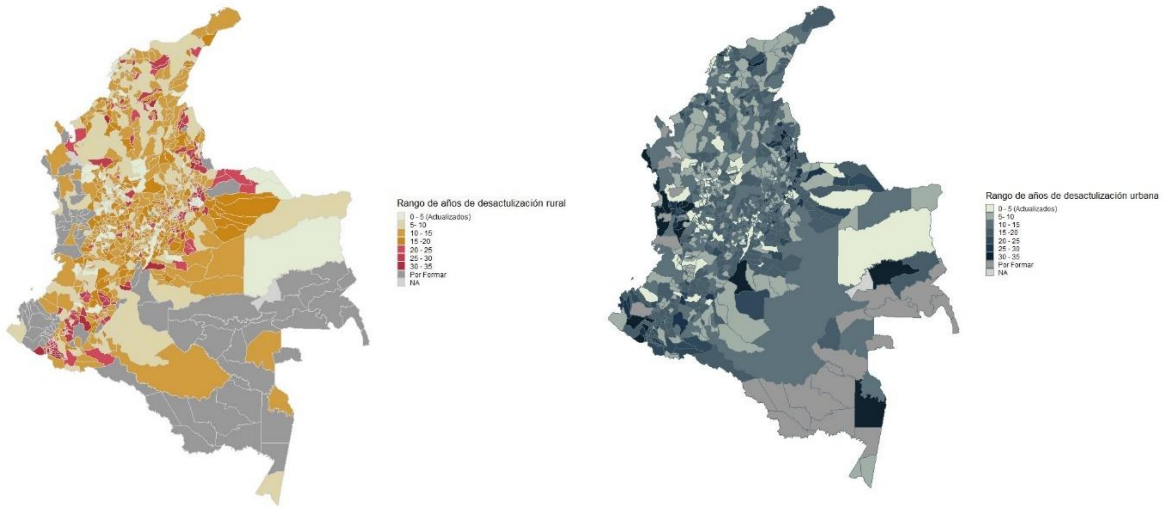
El *Acuerdo final para la terminación del conflicto y la construcción de una paz estable y duradera* (2016) plantea la necesidad de abordar el catastro como un elemento clave para resolver los problemas de tierras derivados del conflicto armado en Colombia. En este sentido, el acuerdo estipula en su primer punto la implementación de un Sistema General de Información Catastral, que sea integral y multipropósito, con el objetivo de que en un plazo máximo de siete años se complete la formación y actualización del catastro rural, integrando el registro de los inmuebles rurales y ejecutándose dentro del marco de autonomía de los municipios (Gobierno Nacional y FARC-EP, 2016, p. 19).

Un municipio que no cuenta con información catastral actualizada carece de un recurso fundamental para la adecuada planificación territorial y la definición de políticas públicas que impulsen el desarrollo social y económico de sus comunidades. Una base de datos integrada con el registro público de la propiedad inmueble, digitalizada e interoperable con otros sistemas de información territorial aporta mecanismos para una asignación más eficiente de los recursos públicos y el fortalecimiento fiscal de los territorios. Entre los desafíos en la implementación de la política catastral del país, se encuentra la descoordinación en los arreglos institucionales actuales respecto a la actualización catastral. En la actualidad, el IGAC es responsable de la actualización catastral de 863 municipios del país bajo un enfoque multipropósito. Sin embargo, la responsabilidad de gestionar información sectorial recae en diversas entidades públicas, y los gobiernos subnacionales son los encargados de planificar y ejecutar políticas de desarrollo territorial. Además, existe un bajo aprovechamiento de la información catastral como insumo en el diseño e implementación de políticas públicas sectoriales y territoriales, así como en su potencial para generar valor económico (DNP, 2020).

Según datos del Instituto Geográfico Agustín Codazzi (IGAC), en 905 municipios del país el avalúo catastral representa apenas el 19% de su valor comercial real. En promedio, el rezago catastral en las zonas urbanas de estos municipios es de 15 años, mientras que en las áreas rurales asciende a 16 años.

Figura 1

Rezago Catastral a Nivel Municipal por Zona



Nota. Tomado de la *Exposición de motivos del Proyecto de Ley 292C de 2023* (p. 18), Congreso de Colombia, 2023.

Al cierre de 2023, solo el 9% de los municipios del país cuentan con un catastro actualizado. Un 24% de ellos presentan una desactualización de entre 5 y 10 años, mientras que el 30% acumula un rezago de entre 10 y 15 años. Por otro lado, el 19% de los municipios se encuentra desactualizado entre 15 y 20 años, y un 12% tiene un atraso superior a 20 años. Además, existe un 6% de municipios que aún no cuentan con un catastro formado.

Tabla 1

Rezago Catastral

Estado del catastro	Avalúo 2023*	Recaudo Predial 2023*	Cantidad de municipios	Porcentaje de municipios
Actualizado	1.438.011.542	7.608.141	96	9%
Desactualizado 5-10 años	341.743.614	2.153.574	261	24%
Desactualizado 10-15 años	155.270.161	952.167	336	30%
Desactualizado 15-20 años	33.117.902	188.633	211	19%
Desactualizado 20-25 años	3.602.153	19.086	48	4%
Desactualizado 25-30 años	3.231.668	16.888	80	7%
Desactualizado 30-35 años	227.160	1.229	9	1%
Por formar	14.198.061	65.739	61	6%
Total	**1.989.402.261**	**11.005.457**	**1.102**	**100%**

Nota. Tomado de la *Exposición de motivos del Proyecto de Ley 292C de 2023* (p. 17), Congreso de Colombia, 2023.

Justificación

En Colombia, el marco normativo en materia catastral, definido por leyes como la Ley 44 de 1990 y la Ley 1450 de 2011, establece la obligación de actualizar los catastros periódicamente —cada cinco años— con el objetivo de reflejar las condiciones reales del mercado inmobiliario. Este mandato responde a la necesidad de garantizar una mayor equidad fiscal y una mejor distribución de los ingresos municipales, basándose en valores prediales más precisos y actualizados.

En este contexto, los modelos hedónicos de precios de bienes inmuebles se utilizan cada vez más en los procesos de valoración masiva de propiedades, aplicando especificaciones econométricas para obtener valoraciones automáticas con fines impositivos. La precisión predictiva de estos modelos es crucial, ya que influye directamente en los ingresos fiscales de las autoridades locales (Lozano-Gracia & Anselin, 2012).

La creciente disponibilidad de datos catastrales y transaccionales, junto con los avances en tecnologías de procesamiento de datos, presenta una oportunidad única para la incorporación de técnicas de machine learning (ML) en este ámbito. Los algoritmos de ML permiten manejar información con alta dimensionalidad, capturar relaciones no lineales, predicciones más precisas y patrones complejos en los datos, superando en muchos casos las limitaciones de los enfoques econométricos tradicionales.

Por tanto, este proyecto se justifica al proponer el desarrollo de un modelo basado en ML para predecir los incrementos en los avalúos catastrales derivados de las actualizaciones periódicas. Dicho modelo no solo facilitará a las autoridades municipales y nacionales la planificación de futuras actualizaciones catastrales, sino que también permitirá una mejor anticipación y gestión de los efectos fiscales y sociales asociados. Este aporte tiene el potencial de ser un pequeño

insumo para la eficiencia del sistema catastral colombiano, mejorar la equidad tributaria y contribuir al fortalecimiento de las finanzas públicas locales.

Objetivos

Objetivo General

Desarrollar un modelo de machine learning basado en la segmentación por clusterización para predecir el impacto de la actualización catastral en los valores catastrales de los predios, utilizando datos de 87 municipios actualizados entre 2016 y 2024.

Objetivos Específicos

Analizar los datos pre y post actualización catastral para identificar patrones de incremento en los valores catastrales, organizando la información relevante para el entrenamiento del modelo de machine learning.

Implementar un proceso de clusterización que permita segmentar los predios en grupos homogéneos con base en variables clave como destino económico, área construida y área de terreno, para mejorar la precisión de las predicciones del modelo.

Diseñar y evaluar un modelo predictivo utilizando técnicas de machine learning, específicamente XGBoost, para estimar los incrementos en el avalúo catastral, identificando las variables más influyentes en estos cambios.

Validar el modelo predictivo desarrollado mediante métricas de desempeño apropiadas, asegurando su capacidad de generalización a municipios no actualizados.

Marco de Referencia

En concordancia con Duarte (2022), el catastro recopila información sobre los predios, como su ubicación, área, valor catastral y situación legal, lo cual resulta esencial para la implementación de políticas públicas y proyectos de infraestructura. No obstante, el catastro rural está desactualizado en un 95%, lo que impacta negativamente en la eficiencia de los proyectos locales y regionales. El reto del catastro multipropósito radica en actualizar la regulación, descentralizar la ejecución y contratar gestores catastrales, transformándolo en un servicio público que involucre de manera directa a los propietarios con el fin de mejorar los resultados a mediano plazo.

La actualización catastral se convierte en un insumo crucial para el crecimiento económico del país, ya que variables como el impuesto predial, la generación de recursos propios de los municipios y la reducción de la dependencia de transferencias del gobierno central, así como la incorporación de gestores catastrales como nuevos actores en la gestión, son fundamentales para aprovechar las oportunidades de crecimiento (Andrade, 2023). Además, la información generada por la actualización catastral es clave para la aplicación de políticas públicas. Gallego et al. (2014) encuentran, por ejemplo, que se lograría mayor progresividad en el subsidio a los servicios públicos domiciliarios si este se basara en el avalúo catastral, en lugar de la actual estratificación.

Colombia presenta un alto grado de descentralización de los ingresos fiscales, según Enache (2021). Las autoridades municipales tienen autonomía para cobrar impuestos territoriales, entre ellos el Impuesto Predial Unificado, regulado por la Ley 44 de 1990. El fortalecimiento de las finanzas de las entidades territoriales se enmarca en el proceso de descentralización que atraviesa el país (OECD, 2019).

Ahmad, Brosio y Jiménez (2019) señalan que los países de América Latina recaudan en promedio solo el 0,5% del PIB a través del impuesto a la propiedad, cifra considerablemente inferior a la de otras regiones. En comparación, Colombia muestra un desempeño ligeramente mejor, con una recaudación del 0,8% del PIB, aunque se estima que el país podría incrementar esta cifra significativamente hasta alcanzar el 2,7% del PIB, si se implementan las reformas necesarias para mejorar el sistema tributario y superar los obstáculos actuales.

El bajo recaudo del IPU en Colombia se debe principalmente a la falta de actualización catastral de las propiedades y sus valores. Las tasas efectivas del impuesto están muy por debajo del máximo legal permitido (5x1000 frente a 16x1000), y tanto las propiedades urbanas como rurales están subregistradas y subvaloradas.

La actualización catastral guarda una relación positiva con el PIB per cápita y la competencia política, mientras que presenta una relación negativa con las tasas de pobreza y las transferencias del gobierno central. Además, se ha estimado que el recaudo potencial del IPU podría alcanzar el 1,5% del PIB, en comparación con el 0,6% actual (Iregui, Melo B., & Ramos F., 2005; Sánchez & España-Eljaiek, 2013). Para lograr una mejora en el recaudo, es necesario contar con un catastro actualizado, en el que las propiedades estén debidamente registradas y valoradas, y mejorar la comunicación entre la Alcaldía y el Concejo Municipal.

Un incremento desmesurado del Impuesto Predial Unificado (IPU) puede generar obstáculos en la implementación de la actualización catastral. Un ejemplo de ello se dio durante la actualización catastral de Cartagena en 2009, cuando muchos contribuyentes optaron por no pagar el IPU correspondiente a ese año y presentaron reclamaciones masivas ante el IGAC, solicitando la revisión de los avalúos catastrales. La mayoría exigía la aplicación del beneficio del límite del impuesto establecido en el artículo 6 de la Ley 44 de 1990. Esta situación generó

gran descontento entre la ciudadanía, lo que culminó en un paro en febrero de 2010 como protesta contra la Actualización Catastral de la ciudad (Martínez & Marín, 2015).

Estado del Arte

Sánchez y España (2013) señalan que el PIB per cápita y la tasa de pobreza influyen significativamente en la cantidad y el valor de los bienes inmuebles. Un mayor PIB per cápita está relacionado con un aumento en el número de propiedades por habitante, tanto urbanas como rurales, y con precios más altos de estos activos, mientras que una mayor tasa de pobreza se asocia negativamente con el valor y el número de inmuebles, sobre todo en zonas urbanas. La concentración de la propiedad de la tierra, medida mediante el índice de Gini, muestra efectos mixtos: se correlaciona positivamente con el valor de las propiedades en áreas rurales, pero negativamente en el número de propiedades en zonas urbanas, sugiriendo una concentración de inversión en las áreas rurales. Asimismo, la falta de actualización catastral afecta negativamente el número de propiedades registradas y su valor; cada año sin actualización reduce los precios de las propiedades urbanas y rurales en un 5.5% y 4%, respectivamente. Estos resultados subrayan la importancia de las actualizaciones catastrales para reflejar de manera precisa el valor y cantidad de propiedades en los municipios.

Buitrago & García (2023) examinan la influencia de las regulaciones en el mercado inmobiliario en Bogotá, analizando el impacto de los precios comerciales de los predios frente a la directriz del decreto 562 de 2014, el cual buscaba reglamentar la altura de los edificios construidos en la capital. El objeto del análisis es evaluar un ex ante – ex post del impacto de la regulación en el mercado inmobiliario, para ello se plantea un modelo de diferencia en diferencias (DiD, por sus siglas en ingles), donde la variable a explicar es el logaritmo natural del precio de los predios, mientras que las variables explicativas son la dummy de afectación – si el

predio se encuentra bajo el decreto 562 de 2014 o no, una dummy temporal del decreto donde es 1 si corresponde al periodo de implementación o 0 si es el caso contrario, un vector de variables temporales donde se encuentran el promedio de la área construida en la zona en m2, características socioeconómicas y densidad poblacional. Finalmente se presenta un vector de variables invariantes dentro de las cuales se evalúan la provisión de bienes públicos como la cercanía de los parques públicos, acceso a vías, transporte y facilidades tales como centros de seguridad, salud y educación. Los resultados muestran que la regulación incremento los precios de los predios bajo el efector del decreto 562 de 2014 entre el 16.4% y el 33%.

Laskin, Gadassina y Zaitseva (2021) analizan el valor catastral como herramienta para monitorear el valor de mercado de bienes inmuebles en diferentes zonas de Rusia, con especial énfasis en la área metropolitana de San Petersburgo. Para su evaluación tratan de explicar el comportamiento de los precios del mercado inmobiliario expresado en rublos por m2, las variables explicativas son la valoración catastral, las características físicas del predio – tamaño en metros cuadrados, habitaciones, baños etc -, antigüedad, destinación del predio (residencial, industrial, comercial etc), cercanía a escuelas, parques y centros de salud, finalmente se tiene en cuenta las condiciones del mercado local. Para su estimación usan un modelo de distribución log -normal de dos dimensiones. El principal resultado es la propuesta metodológica del modelo predictivo de los valores comerciales del mercado inmobiliario.

Guadalajara et al. (2021) exploran cómo el valor catastral del suelo urbano y las características del vecindario afectan la valoración hipotecaria promedio de las viviendas en Valencia, España. Para ello toman diferentes variables agregadas de barrios urbanos como son el valor promedio de la propiedad, el valor promedio del metro cuadrado, el promedio de edad de los propietarios, proporción de predios con aire acondicionado, proporción de predios con aire

acondicionado, proporción de predios con ascensor; además se toman los rangos de edad de los habitantes del barrio, medios de transporte, servicios financieros, densidad de carros por 100 habitantes. Las metodologías de estimación usadas son de análisis espacial – modelo autorregresivo espacial - y regresión OLS. Ambas metodologías revelan ser adecuadas para la predicción del valor de la vivienda, donde el modelo autorregresivo es preferido dado que corrige los problemas de autocorrelación en las variables espaciales.

Lozano-Gracia y Anselin (2012) evaluaron las estimaciones de valores catastrales en Bogotá para determinar su precisión, utilizando datos de la Unidad de Análisis Económico del Catastro Distrital (UAECD). El modelo considera diversas características físicas del predio, como tipo de techo, materiales de construcción, estructura, pisos, baños, y proximidad a parques, áreas protegidas, centros comerciales y servicios de seguridad.

Para analizar el rendimiento predictivo de diferentes especificaciones de modelos, cada uno se estima 100 veces con distintas submuestras, generando variabilidad en las predicciones. La precisión se evalúa mediante la mediana del error porcentual absoluto promedio y el porcentaje de predicciones que caen dentro de un 10% y 20% del valor real. Se comparan dos enfoques: (1) especificaciones de un modelo base que incluyen efectos fijos de submercados, características tradicionales de vecindario y proximidad a instalaciones calculadas mediante GIS; y (2) regímenes espaciales, donde se genera un conjunto de estimaciones específico para cada submercado. Con una base de más de 14,000 propiedades residenciales en Bogotá, el análisis explora el impacto de considerar la heterogeneidad de submercados y el uso de zonas homogéneas (HZ) frente a variables de distancia calculadas mediante GIS. Los resultados indican que los modelos que incorporan variables de distancia son tan efectivos, o incluso levemente superiores, a los modelos con HZ, lo cual podría reducir costos en su implementación

automatizada. Además, los modelos que incorporan heterogeneidad mediante submercados espaciales (como el modelo 6-SMK basado en estratos socioeconómicos) superan a aquellos que aplican efectos fijos. Sin embargo, la variabilidad en la sobreestimación entre estratos plantea inquietudes de equidad, ya que el modelo 6-SMK mostró mejor desempeño en los estratos bajos y una ligera desventaja en el más alto.

Zhang et al. (2015) proponen un modelo mejorado de error espacial para la valoración masiva de bienes inmuebles comerciales, utilizando Shenzhen como estudio de caso donde se evalúa la opción de un impuesto sobre los bienes inmuebles. Para ello, la evaluación masiva es una técnica eficaz para establecer la base imponible. Para alcanzar una alta precisión y reducir los costos de valoración, se propone un marco innovador para la evaluación de bienes raíces comerciales, que integra un modelo de error espacial (SEM), matemáticas difusas y econometría. El modelo SEM convencional se modifica para adaptarse a la evaluación masiva de propiedades comerciales, incorporando un enfoque de matemáticas difusas para especificar la matriz de pesos espaciales (SWM). Además, mediante herramientas econométricas, se analizan los impactos de factores inherentes y de ubicación en los precios de bienes raíces comerciales. La variable dependiente es el precio unitario de los inmuebles comerciales, mientras que las variables independientes incluyen el área, que muestra una contribución positiva pero no significativa; el frente de calle, que presenta una relación positiva significativa con el precio unitario; el ancho, cuyo efecto no es significativo; y la altura, que influye positivamente en el precio. Además, la tasa de vacancia tiene una relación negativa con el precio unitario, mientras que la profundidad muestra un impacto negativo debido a la limitación en la disposición del espacio. Por último, el nivel comercial, que refleja las ventajas de localización, está positivamente correlacionado de manera significativa con el precio unitario, destacando como un factor clave en la valoración de

los inmuebles comerciales. Los experimentos demuestran que los resultados de valoración son precisos y que mejora la consistencia entre diferentes objetos de evaluación. Asimismo, la introducción de matemáticas difusas permite ampliar la construcción de la SWM de una variable única a múltiples variables.

Wang, Li, y Yu (2020) muestran que el modelo tradicional de regresión lineal para la evaluación masiva de propiedades resulta insuficiente frente a grandes volúmenes de datos, características complejas de las viviendas y altos requerimientos de precisión. Por ello, es necesario incorporar características espaciotemporales para desarrollar un modelo más efectivo. Este estudio analiza el núcleo urbano de Beijing, utilizando datos de transacciones reales a nivel comunitario de 2014, 2016 y 2018. Se comparan tres modelos: regresión múltiple (MRA) con mínimos cuadrados ordinarios (OLS), regresión ponderada geográficamente (GWR) y regresión ponderada geográfica y temporal (GTWR). La variable a explicar es el precio promedio de una vivienda en determinada comunidad, mientras que las variables independientes son las propiedades estructurales de la comunidad – ubicación, antigüedad, características físicas del predio como los baños, el tamaño del predio, habitaciones y materiales, y distancia – con un radio de 2km – frente a transporte y servicios públicos como educación, comercio y seguridad. El modelo GTWR, con un R^2 ajustado de 0.8192, supera a los demás, evidenciando que los precios de vivienda son sensibles a factores espaciales y temporales. Este enfoque ofrece una herramienta eficiente para la planificación, gestión del suelo, tributación, seguros y finanzas, destacando las características espaciales de los parámetros relacionados con precios en áreas densamente pobladas. Aunque el modelo GTWR es eficaz para la evaluación masiva con datos comunitarios multianuales, presenta limitaciones. El uso de datos a nivel comunitario puede perder información relevante de transacciones individuales, y aplicar modelos locales con

grandes volúmenes de datos puede afectar su estabilidad y eficiencia. Además, los datos de alquiler, más frecuentes y estables en submercados de vivienda, podrían ofrecer una perspectiva valiosa sobre el valor habitacional.

Tabla 2

Revisión de Literatura con Modelos Lineales y Espaciales

Trabajo	Variable a Explicar	Variables Explicativas	Modelo	Universo
Sánchez y España (2013)	Valor catastral del predio	PIB per cápita, tasa de pobreza, índice de Gini de concentración de la tierra, densidad poblacional, informalidad de los predios, actualización catastral	Efectos fijos	937 municipios
Buitrago & García (2023)	Logaritmo natural del valor comercial del predio	Dummy de afectación del decreto 562 de 2014, dummy temporal, variables temporales: promedio área construida en m2, características socioeconómicas y densidad poblacional; variables atemporales: acceso a parques públicos, transporte, vías, seguridad, educación y salud.	Diferencias en diferencias	837,505 predios en Bogotá en 2017
Laskin, Gadassina y Zaitseva (2021)	Logaritmo del precio comercial del predio	Valor catastral en el periodo anterior, características físicas del predio, destinación y uso del predio, y proximidad a bienes públicos.	Distribución log – normal de dos dimensiones	
Guadalajara et al. (2021)	Logaritmo natural del precio comercial de la vivienda	Valor promedio de la propiedad, el valor promedio del metro cuadrado, el promedio de edad de los propietarios, proporción de predios con aire acondicionado, proporción de predios con aire acondicionado, proporción de predios con ascensor; además se toman los rangos de edad de los habitantes del barrio, medios de transporte, servicios financieros,	Modelo autorregresivo espacial, OLS	Valencia, España. 2017

Trabajo	Variable a Explicar	Variables Explicativas	Modelo	Universo
		densidad de carros por 100 habitantes		
Lozano-Gracia y Anselin (2012)	Valor comercial a nivel zona	Características físicas de construcción del predio, como el tipo de techo, el material de construcción, estructura, pisos, baños, cercanía a los parques, cercanía a áreas protegidas, cercanía a centros comerciales, centros seguridad, entre otros servicios públicas.	OLS	14,079 predios en Bogotá D.C.
Zhang et al. (2015)	Precio unitario	Área, Frente de calle, Ancho, Altura, Tasa de vacancia, Profundidad, Nivel comercial, Localización	SEM-MA	236 ofertas en el Área comercial de Huaqiang - Shenzen, China. 2012
Wang, Li, y Yu (2020)	Precio promedio de una vivienda en determinada comunidad (Log)	las propiedades estructurales de la comunidad – ubicación, antigüedad, características físicas del predio como los baños, el tamaño del predio, habitaciones y materiales, distancia – con un radio de 2km – frente a transporte, educación, comercio y seguridad	GTWR	3064 ofertas en la Zona antigua de Beijing, China. 2014, 2016, 2018.

En el uso de técnicas de Machine Learning para valoraciones masivas se destaca, Ali et al. (2020) emplean técnicas de aprendizaje automático para delimitar vecindarios a partir de datos de tasación geocodificados. A diferencia de los códigos postales y zonas censales, que son estáticos y no capturan las dinámicas del mercado inmobiliario, este enfoque identifica vecindarios como agrupaciones de propiedades con características y precios similares. Se aplicó

un algoritmo de agrupamiento espacial basado en densidad jerárquica (HDBSCAN) - este enfoque se basa únicamente en la distancia entre propiedades y sus comparables, sin necesidad de incluir todas las características físicas o demográficas de las propiedades - para generar estas delimitaciones a partir de datos de tres condados con alta actividad de tasaciones: en Los Ángeles, San Diego y Orange County, en el sur de California, utilizando datos de CoreLogic entre 2014 y 2018. El modelo utiliza filtros espaciales para mapear conexiones entre propiedades y agrupar aquellas con mayor densidad de vínculos. Como variable dependiente se usa el valor de tasación por metro cuadrado, mientras que las independientes son la distancia geográfica entre propiedades tasadas y comparables.

Los vecindarios generados explican mejor las variaciones en características inmobiliarias, como el precio por metro cuadrado, en comparación con códigos postales y zonas censales. Además, estos vecindarios son dinámicos, con capacidad de expandirse o contraerse según los cambios en los submercados de vivienda. Aunque el modelo no abarca todas las propiedades debido a limitaciones en los datos y propiedades clasificadas como "ruido" por el algoritmo.

Antipov y Pokryshevskaya (2012) emplearon el modelo Random Forest con enfoque segmentado para la valoración masiva de apartamentos residenciales, complementado con un enfoque basado en CART para diagnósticos del modelo. El objeto de estudio fue de 2848 apartamentos de dos habitaciones, con área de hasta 160 m² y precios de hasta 30 millones de rublos en San Petersburgo, Rusia durante la primavera de 2010. La muestra final se dividió en 2695 observaciones para entrenamiento y 150 para pruebas. Utiliza como variable dependiente el precio del apartamento, representado tanto como el precio total en miles de rublos como el precio por metro cuadrado. Las variables independientes incluyen características físicas y funcionales de los apartamentos, como el área total y habitable, el tamaño de las habitaciones principales, el

área de la cocina y el tipo de unidad de baño. También se consideran factores estructurales, como el número de pisos del edificio, el piso donde se encuentra el apartamento y el tipo de construcción. El modelo Random Forest presentó un error medio absoluto porcentual (MAPE) menor al 9.8% para apartamentos con área menor a 61.5 m², mientras que para los de mayor tamaño el MAPE fue del 19.4%. En distritos específicos, el MAPE fue del 12.9%, mientras que en otros distritos alcanzó el 23.6%.

Park y Bae (2015) desarrollaron un modelo predictivo de precios de vivienda utilizando algoritmos de machine learning, como C4.5, RIPPER, Naïve Bayesian y AdaBoost, basado en un conjunto de datos de 5359 casas adosadas en el condado de Fairfax, Virginia ENTRE 2004 - 2008. La variable dependiente fue la categorización del precio de cierre en relación con el precio de lista (mayor o menor), mientras que las variables independientes incluyeron características físicas (número de baños, dormitorios, tipo de calefacción, tamaño del lote, entre otras), calificaciones de escuelas públicas, tasas hipotecarias y factores temporales, como el mes de listado. Los resultados mostraron que el algoritmo RIPPER superó consistentemente a los demás modelos en términos de precisión, demostrando su potencial para asistir a vendedores y agentes inmobiliarios en la toma de decisiones basadas en la valoración de precios de propiedades.

Peterson and Flanagan (2009) compararon los modelos hedónicos lineales tradicionales con redes neuronales artificiales (ANN) utilizando una muestra de 46,467 propiedades residenciales transaccionadas entre 1999 y 2005 en el condado de Wake, Carolina del Norte. La variable dependiente fue el precio de venta de las propiedades, mientras que las variables independientes incluyeron más de 18 características de las propiedades, tales como el número de baños, pisos, y variables categóricas como el código de ubicación. Los resultados demostraron que las ANN superaron significativamente a los modelos lineales en términos de precisión,

generando menores errores de valoración y extrapolando mejor en entornos de precios volátiles. Este estudio destaca que las ANN, al manejar no linealidades complejas y resolver problemas asociados con variables categóricas y colinealidad, representan una alternativa eficiente frente a los métodos de valoración tradicionales.

Sharma et al. (2024) abordaron el problema de predicción de precios de viviendas como una tarea de regresión, utilizando el conjunto de datos de viviendas de Ames City, Iowa, compuesto por 2930 registros y 82 variables. La variable dependiente fue el precio de las viviendas, mientras que las variables independientes incluyeron características como la calidad general de la casa (*Overall Qual*), el área habitable del piso principal (*Gr Liv Area*), el tamaño del garaje en términos de autos (*Garage Cars*), y el área total del sótano (*Total Bsmt SF*). El estudio comparó múltiples modelos de machine learning, como regresión lineal múltiple, perceptrón multicapa, bosque aleatorio, soporte vectorial y XGBoost, aplicando además técnicas de ajuste de hiperparámetros mediante *GridSearchCV*. Los resultados identificaron a XGBoost como el modelo más preciso para la predicción de precios, con un error cuadrático medio (MSE) de 0.001, destacando la importancia de las características mencionadas en el valor predictivo del modelo.

Tabla 3

Revisión de Literatura con Aplicación de Técnicas de Aprendizaje Automático

Trabajo	Variable a Explicar	Variables Explicativas	Modelo	Universo
Ali et al. (2020)	Precio por metro cuadrado	Distancia geográfica entre propiedades tasadas y comparables	HDBSCAN	Los Ángeles, San Diego y Orange County, en el sur de California, EE. UU. 2014 - 2018

Trabajo	Variable a Explicar	Variables Explicativas	Modelo	Universo
Park y Bae (2015)	Categorización del precio de cierre	Número de baños, dormitorios, tipo de calefacción, tamaño del lote, entre otras), calificaciones de escuelas públicas, tasas hipotecarias y factores temporales	C4.5, RIPPER, Naïve Bayesian y AdaBoost	5359 casas adosadas en el Condado de Fairfax, Virginia, EE. UU.. 2004 - 2008
Antipov y Pokryshevskaya (2012)	Precio por metro cuadrado/Precio total del apartamento	Área total y habitable, tamaño de las habitaciones principales, área de la cocina, tipo de unidad de baño, número de pisos del edificio, piso donde se encuentra el apartamento y el tipo de construcción	Random Forest segmentado	2848 apartamentos en San Petersburgo, Rusia. 2010.
Peterson and Flanagan (2009)	Precio de venta de vivienda	Características de las propiedades, tales como el número de baños, pisos, y variables categóricas como el código de ubicación	ANN	46,467 propiedades residenciales transaccionadas en el condado de Wake, Carolina del Norte, EE. UU.. 1999 - 2005
Sharma et al. (2024)	Precio comercial unitario de vivienda	Calidad general de la casa, área habitable del piso principal, tamaño del garaje en términos de autos y el área total del sótano, entre otros.	Regresión Lineal, Multi-Layer Perceptron, Random Forest, SVR, XGBoost	2930 registros de Ames City, Iowa, EE. UU..

Marco Teórico

En el marco de este estudio, se busca identificar patrones de incremento del avalúo catastral utilizando la base de datos de los 87 municipios ya actualizados con enfoque

multipropósito entre 2016 y 2024. Estos patrones permitirán pronosticar el potencial incremento del avalúo catastral en municipios que aún no han sido actualizados, aprovechando técnicas avanzadas de análisis de datos fundamentadas en la literatura existente.

El trabajo de Ali et al. (2020) destaca la utilidad de las técnicas de clusterización para segmentar propiedades en grupos homogéneos con base en características similares. Si bien dicho estudio utiliza datos espaciales para identificar agrupaciones dinámicas de propiedades, en el presente trabajo, la clusterización se fundamenta en variables específicas a nivel predio como el destino económico, el área construida y el área del terreno. Estas variables son relevantes para identificar patrones comunes en el comportamiento del avalúo catastral, permitiendo generar grupos comparables que reflejen de manera adecuada las dinámicas de mercado dentro de los municipios analizados. Este primer momento de clusterización es clave para estructurar los datos y mejorar la precisión de las predicciones en etapas posteriores, adaptándose a la realidad del mercado inmobiliario colombiano.

HDBSCAN (Agrupamiento Espacial Basado en Densidad Jerárquica) es un algoritmo de aprendizaje automático no supervisado que utiliza una estructura jerárquica para generar agrupamientos planos, basados en la estabilidad de los clústeres (McInnes et al., 2017). HDBSCAN también considera la existencia de predios con diferentes densidades en los datos y puede identificar ciertos puntos como ruido, es decir, puntos que no pertenecen a ningún grupo y deben ser excluidos. El algoritmo funciona con un único parámetro: el número mínimo de observaciones necesarias para definir un clúster (Cesario et al., 2007). Otro aspecto clave es su capacidad para capturar jerarquías y segmentaciones naturales en los datos. Por ejemplo, los municipios pueden generar patrones específicos debido a diferencias en densidad urbana, rural, y tamaño, lo que requiere segmentación local. HDBSCAN identifica clústeres independientes

dentro de cada municipio respetando estas estructuras particulares. Asimismo, la clasificación por tipo de predio (urbano o rural), por el tipo de actualización (rural y urbana, solo urbana o rural) y por destino económico (comercio, habitacional, industria etc.) se beneficia de la flexibilidad del algoritmo para agrupar según patrones específicos de densidad.

HDBSCAN también destaca en contextos donde las distribuciones de los datos no son normales y existen outliers, a diferencia de los algoritmos de clústeres más comunes como el *K-Means* (Hou et al., 2016). Por ejemplo, la valoración de predios presenta colas largas similares a las del análisis de ingresos, donde unos pocos predios concentran valores extremadamente altos. El algoritmo, al no asumir normalidad en las variables, es robusto frente a estas distribuciones asimétricas y permite tratar los valores atípicos como ruido, protegiendo la integridad de los clústeres estimados.

Por otro lado, Sharma et al. (2024) demuestran que los modelos de machine learning basados en XGBoost tienen un mayor potencial predictivo que otras técnicas, como la regresión lineal, el bosque aleatorio o las redes neuronales. Este modelo destaca por su capacidad para manejar relaciones complejas y no lineales entre variables, optimizar el ajuste de hiperparámetros y reducir el error en las predicciones. En el presente estudio, XGBoost se adopta para modelar los patrones de cambio en el avalúo catastral y generar pronósticos precisos sobre los incrementos esperados en municipios no actualizados, considerando variables clave como la localización, las características del predio y las dinámicas de mercado.

El método de Extreme Gradient Boosting (XGBoost) ha adquirido una gran relevancia en el ámbito del aprendizaje automático gracias a su destacado rendimiento y versatilidad. Según Chen and Guestrin (2016), XGBoost se presenta como una implementación escalable y eficiente del método de gradient tree boosting, diseñado para abordar problemas a gran escala del mundo

real utilizando un mínimo de recursos. Su popularidad se debe a su capacidad para resistir el sobreajuste, una característica fundamental para garantizar que el modelo pueda generalizar correctamente al trabajar con nuevos datos (Demir & Sahin, 2023).

El XGBoost puede ser presentado como un problema de optimización, donde la función objetivo es:

$$f_{obj} = \sum_{i=1}^{n} L(y_i, y_{'i}) + \sum_{k=1}^{K} \theta(f_k)$$

Donde K es el número de árboles, $L(y_i, y_{'i})$ es la función de perdida y f es el termino de regularización que es usado para controlar la complejidad y evitar el sobreajuste.

Para expresar la valor de la predicción en el momento i^{th} de la muestra después de la iteración t^{th}:

$$y_{'i}{}^{t} = y_{'i}{}^{t-1} + f_t(X_i)$$

Entre las métricas discutidas para analizar la calidad de las estimaciones de las regresiones XGBoost, se encuentran el R-cuadrado, el error absoluto medio (MAE), y el error cuadrático medio (MSE)

El valor de R-cuadrado mide qué tan bien se ajusta el modelo a los datos y su precisión al predecir muestras no vistas. Se calcula utilizando la fórmula:

$$R^2 = 1 - \frac{SSR}{SSM}$$

donde SSR representa la suma de los errores al cuadrado de la línea de regresión, y SSM es la suma de los errores al cuadrado de la línea media.

Por otro lado, el error cuadrático medio (MSE) se calcula evaluando la diferencia al cuadrado entre los valores reales y los predichos. Un MSE más bajo indica un mejor desempeño del modelo. La fórmula para calcular el MSE es:

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

Finalmente, el error absoluto medio (MAE) cuantifica la diferencia promedio entre los valores predichos y observados. Se calcula sumando los errores absolutos y dividiendo por el tamaño de la muestra:

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

Un MAE más bajo implica que el modelo realiza predicciones más cercanas a los valores reales.

Dada la naturaleza de los datos trabajados, caracterizados por la no normalidad en su distribución y la presencia de observaciones extremas (A1), se opta por utilizar el rango intercuartílico (IQR) para efectos de inferencia. El IQR, como medida robusta de dispersión, es particularmente útil en contextos donde los datos presentan atipicidades, ya que utiliza la mediana como medida central de tendencia y elimina problemas asociados con valores atípicos (Whaley, 2005). Su cálculo se define como la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1):

$$IQR = Q1 - Q3$$

Donde Q1 y Q3 representan los percentiles 25 y 75 respectivamente. El gráfico 2.2.1 ilustra visualmente la composición de un gráfico de cajas, destacando sus componentes principales, como el rango intercuartílico, la mediana – percentil 50 –, los bigotes, que se expresan de la siguiente manera:

$$L_{inferior} = Q1 - 1.5\,IQR$$

$$L_{superior} = Q3 + 1.5\,IQR$$

Donde $L_{inferior}$ y $L_{superior}$ hacen referencia al bigote por debajo del percentil 25 y del percentil 75, respectivamente. Por su parte los datos atípicos se presentan como:

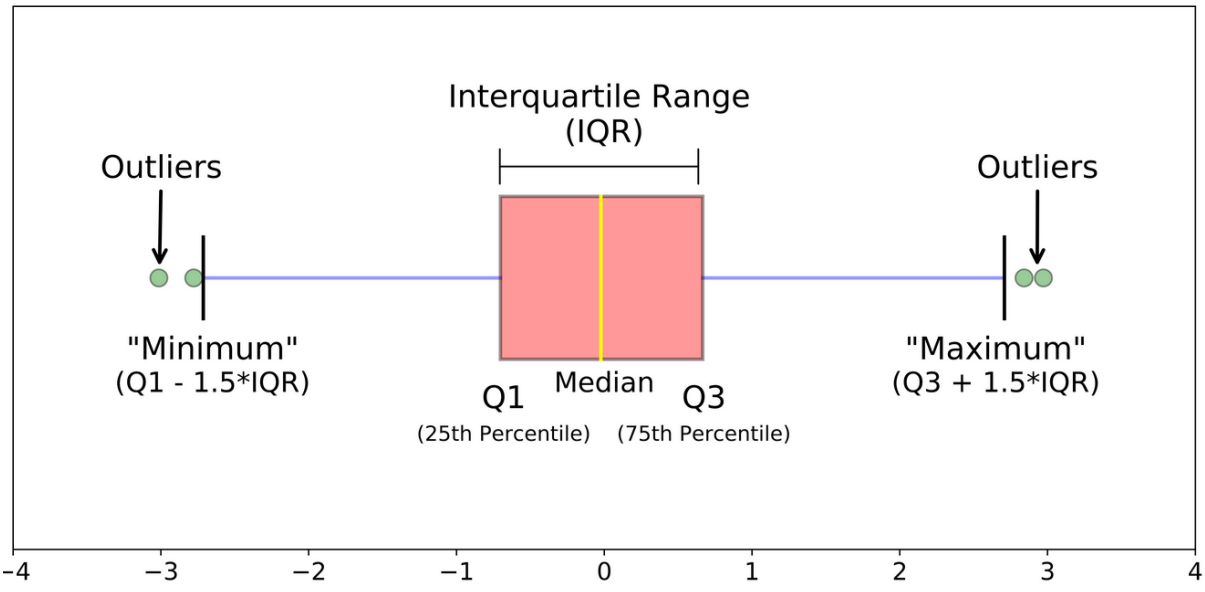
$$x < L_{inferior}$$

$$x > L_{superior}$$

En donde $x < L_{inferior}$ hacen referencia a los datos atípicos inferiores y $x > L_{superior}$ representan los datos atípicos superiores. Estos elementos tienen como propósito reflejar la variabilidad de los datos dentro de un rango esperado.

Figura 2

Rango Intercuartílico



Nota. Tomado de Box Plot, por P. Sharma, 2019, *Data Science Unwind.* https://datascienceunwind.wordpress.com/2019/10/03/box-plot/. Copyright 2019 por P. Sharma.

Al integrar estas metodologías, este trabajo propone un enfoque que combina la segmentación mediante clusterización con la predicción basada en machine learning. Este diseño no solo permite capturar las particularidades de los patrones de incremento en los municipios ya

actualizados, sino que también proporciona una herramienta para estimar los incrementos en el avalúo catastral en contextos aún pendientes de actualización.

Marco conceptual

En concordancia por lo dispuesto en el Anexo 1 de la resolución 1040 de 2023 emitida por el IGAC, máxima autoridad catastral en Colombia, a continuación, se relacionan los conceptos sobre los cuales se trabajará en este escrito:

- Catastro: Inventario o censo de los bienes inmuebles localizados en el territorio nacional, de dominio público o privado, independiente de su tipo de tenencia, el cual debe estar actualizado y clasificado con el fin de lograr su identificación jurídica, física y económica con base en criterios técnicos y objetivos. (Instituto Geográfico Agustín Codazzi [IGAC], 2023, Anexo 1, p. 2).
- Catastro multipropósito: Aquel en el que la información que se genere a partir de su implementación, debe servir como un insumo fundamental en la formulación e implementación de diversas políticas públicas, contribuyendo a brindar una mayor seguridad jurídica, la eficiencia del mercado inmobiliario, el desarrollo y el ordenamiento territorial, integrada con el registro público de la propiedad inmueble, digital e interoperable con otros sistemas de información del territorio, y que provea instrumentos para una mejor asignación de los recursos públicos y el fortalecimiento fiscal de los territorios (Instituto Geográfico Agustín Codazzi [IGAC], 2023, Anexo 1, p. 2).
- Base de datos catastral: Es el compendio de la información geográfica y alfanumérica estructurada, que se almacena y se gestiona en un sistema informático, referente a los aspectos físicos, jurídicos y económicos de los predios inscritos en el catastro. Debe ser

interoperable con el registro de la propiedad inmueble y con otros sistemas de administración del territorio (Instituto Geográfico Agustín Codazzi [IGAC], 2023, Anexo 1, p. 1).

- Información catastral: Características físicas, jurídicas y económicas de los predios. Dicha información constituirá la base catastral y deberá ser incorporada por los gestores catastrales en el Sistema Nacional de Información Catastral - SINIC o en la herramienta tecnológica que haga sus veces (Instituto Geográfico Agustín Codazzi [IGAC], 2023, Anexo 1, p. 3).

- Valor comercial de un inmueble: Precio más probable por el cual un predio se transaría en un mercado en donde el comprador y el vendedor actuarían libremente con el conocimiento de las condiciones físicas y jurídicas que afectan el bien; el cual se sustenta a través de métodos de valoración como: comparación o mercado, capitalización de rentas o ingresos, costo de reposición y/o técnica residual. Independientemente del método valuatorio, debe contar con soporte económico acorde con la teoría de valor.

- Avalúo catastral: Valor de un predio resultante de un ejercicio técnico en desarrollo de los procesos catastrales, que, en ningún caso, podrá ser inferior al 60% del valor comercial o superar el valor de este último (Instituto Geográfico Agustín Codazzi [IGAC], 2023, Anexo 1, p. 1).

- Propietario: Titular del derecho real de dominio o propiedad en virtud de un acto o negocio jurídico válido inscrito en el Registro de Instrumentos Públicos (Instituto Geográfico Agustín Codazzi [IGAC], 2023, Anexo 1, p. 7).

- Proceso de formación catastral: Es el conjunto de actividades destinadas a identificar, recoger e incorporar en la base de datos catastral, por primera vez, la información

física, jurídica y económica de la totalidad de los predios que conforman un territorio objetivo para la gestión catastral (Instituto Geográfico Agustín Codazzi [IGAC], 2023, Anexo 1, p. 6).

- Proceso de actualización catastral: Es el conjunto de actividades destinadas a identificar, recoger, incorporar o rectificar en la base de datos catastral los cambios o inconsistencias en la información catastral en sus componentes físicos, jurídicos y económicos, en un territorio objetivo, durante un período determinado (Instituto Geográfico Agustín Codazzi [IGAC], 2023, Anexo 1, p. 8).
- Proceso de conservación catastral: Se entiende por el conjunto de acciones tendientes a mantener actualizada la base catastral de forma permanente, mediante la identificación, recolección e incorporación de los cambios en la información de un bien inmueble. La conservación catastral podrá realizarse a solicitud de parte o de oficio (Instituto Geográfico Agustín Codazzi [IGAC], 2023, Anexo 1, p. 8).
- Predio: Inmueble con o sin título registrado, no separado por otro predio, con o sin unidades de construcción y vinculado con personas naturales o jurídicas, según su relación de tenencia: propietario, poseedor u ocupante (Instituto Geográfico Agustín Codazzi [IGAC], 2023, Anexo 1, p. 6).
- Mutación catastral: Cambios que se presentan en los componentes físico, jurídico o económico de un predio (Instituto Geográfico Agustín Codazzi [IGAC], 2023, Anexo 1, p. 5).
- Información económica: Corresponde al avalúo catastral del inmueble, el cual deberá guardar relación con los valores de mercado (Instituto Geográfico Agustín Codazzi [IGAC], 2023, Anexo 1, p. 4).
- Información física: Corresponde a la representación geométrica, la identificación de la cabida, los linderos y las construcciones de un inmueble. La identificación física no implica

necesariamente el reconocimiento de los linderos del predio in situ (Instituto Geográfico Agustín Codazzi [IGAC], 2023, Anexo 1, p. 4).

- Información jurídica: Identificación de la relación jurídica de tenencia entre el sujeto activo del derecho, sea el propietario, poseedor u ocupante, con el inmueble. Esta calificación jurídica no constituye prueba ni sanea los vicios de la propiedad (Instituto Geográfico Agustín Codazzi [IGAC], 2023, Anexo 1, p. 4).
- Mercado inmobiliario: Es la interacción de agentes cuyas decisiones mantienen o modifican la oferta, demanda y precio de bienes inmuebles en un ámbito geográfico y tiempo determinado (Instituto Geográfico Agustín Codazzi [IGAC], 2023, Anexo 1, p. 4).
- Número predial nacional: Código numérico de treinta (30) dígitos, que se le asigna a cada predio para su identificación en la base de datos catastral de acuerdo con la estructura definida por el IGAC (Instituto Geográfico Agustín Codazzi [IGAC], 2023, Anexo 1, p. 5).
- Destinación económica: Clasificación para fines estadísticos y económicos que se da a cada inmueble en su conjunto -terreno, unidades de construcción-, en el momento de la identificación predial de conformidad con la actividad predominante que en él se desarrolle (Instituto Geográfico Agustín Codazzi [IGAC], 2023, Anexo 1, p. 2).
- Valoración masiva: Es el proceso de determinar el valor de un grupo de propiedades en una fecha específica, utilizando datos comunes, métodos estandarizados y pruebas estadísticas (Eckert, 1990).
- Bienes raíces o inmuebles: Las cosas que no pueden transportarse de un lugar a otro; como las tierras y minas, y las que adhieren permanentemente a ellas, como los edificios, los árboles (Artículo 656 del Código Civil).

Marco Legal

El marco normativo que regula el catastro y los impuestos a la propiedad raíz en Colombia está fundamentado en diversas leyes que articulan su alcance, aplicación y administración. La Ley 44 de 1990, en su artículo 1, establece los principios fundamentales del Impuesto Predial Unificado, definiendo su naturaleza como un gravamen de orden municipal, cuya base gravable se determina por el avalúo catastral o autoavalúo del inmueble. Paralelamente, otras disposiciones, como la Ley 1450 de 2011, introducen lineamientos para la formación y actualización catastral, mientras que la Ley 14 de 1983 fija el rol del Instituto Geográfico Agustín Codazzi (IGAC) como autoridad nacional en materia catastral. Este marco legal se complementa con disposiciones constitucionales y normas específicas, como la Ley 1995 de 2019, que introduce el modelo de catastro multipropósito,

La Ley 44 de 1990 regula la normatividad relacionada con el catastro y los impuestos a la propiedad raíz. En su artículo 1 define el Impuesto Predial Unificado, estableciendo su alcance y naturaleza.

Los procesos de formación y actualización catastral están normados por el artículo 24 de la Ley 1450 de 2011, el cual estipula:

"Las autoridades catastrales tienen la obligación de formar los catastros o actualizarlos en todos los municipios del país dentro de períodos máximos de cinco (5) años, con el fin de revisar los elementos físicos o jurídicos del catastro originados en mutaciones físicas, variaciones de uso o de productividad, obras públicas o condiciones locales del mercado inmobiliario."

Asimismo, el parágrafo del mismo artículo determina que:

"El avalúo catastral de los bienes inmuebles fijado para los procesos de formación y actualización catastral a que se refiere este artículo no podrá ser inferior al sesenta por ciento (60%) de su valor comercial."

La entrada en vigencia de los avalúos catastrales, fijada para el 1 de enero de cada año, está respaldada por el artículo 8 de la Ley 14 de 1983. Esta misma ley, en su artículo 12, establece al Instituto Geográfico Agustín Codazzi (IGAC) como la máxima autoridad en materia catastral a nivel nacional.

Por otro lado, los ajustes anuales del avalúo catastral en procesos de conservación están definidos en el artículo 6 de la Ley 242 de 1995.

En el ámbito constitucional, el artículo 317 de la Constitución Política de Colombia autoriza exclusivamente a los municipios a gravar el Impuesto Predial Unificado, cuya base gravable corresponde al avalúo catastral o autoavalúo del inmueble, en concordancia con el artículo 3 de la Ley 44 de 1990. Por su parte, el artículo 2 de esta misma ley señala que el Impuesto Predial Unificado es un impuesto del orden municipal, cuya administración, recaudo y control corresponden a los municipios respectivos.

Finalmente, el artículo 1 de la Ley 1995 de 2019 establece que los avalúos catastrales deben regirse por lo dispuesto en el modelo de catastro multipropósito.

Metodología

El objetivo principal del proyecto es proporcionar una herramienta basada en machine learning que permita a las administraciones municipales tener un indicativo para anticipar el impacto de la actualización catastral en los valores prediales.

Fase 1 Comprensión del Negocio

Este tipo de análisis puede ser de importancia para la planificación fiscal a nivel local, dado que el avalúo catastral sirve de base gravable al IPU, la magnitud del incremento de esta impactará los ingresos provenientes del IPU, lo cual representan una parte fundamental de las finanzas públicas territoriales. El éxito del proyecto se medirá por la capacidad del modelo para generar proyecciones precisas que informen la toma de decisiones estratégicas en la actualización catastral. La comprensión de los recursos disponibles, como la información presente en las bases catastrales es clave, así como la identificación de supuestos y limitaciones, en términos de la calidad de los datos y las posibles dinámicas distintivas entre municipios.

Fase 2 Comprensión de los Datos

Para la consecución de un modelo de machine learning que tenga como resultado un pronóstico de un incremento del avalúo catastral producto de una actualización catastral es necesario tener como insumo la información de las bases catastrales de los 87 municipios actualizados con enfoque multipropósito desde 2016 hasta 2024, esto implica que para hacer la distribución training set – test set es de suma importancia tener los valores catastrales antes de actualización y después de actualización. El momento de la actualización juega un papel clave, además de conocer el alcance de cada actualización, dado que no todas las actualizaciones catastrales necesariamente abarcan la totalidad de predios en un municipio, puede que la actualización se realice solo a alguna de las zonas: urbano o rural, además también puede darse

el caso de que la actualización sea una parcialidad de predios, a modo de ejemplo se puede dar el que el gestor catastral solo logre actualizar el 50% de predios rurales o urbanos de un municipio.

La combinación de estas dos bases de datos es fundamental para el análisis, ya que nos permite identificar la zona de actualización y segmentar los datos pre y post actualización, lo que posteriormente permitirá el análisis comparativo y exploratorio de los cambios en los avalúos prediales. Además, se lleva a cabo un análisis exploratorio de los datos para evaluar su calidad, detectando posibles errores, datos faltantes o inconsistencias entre los distintos años y municipios (Ver apéndice A).

A parte de lo anterior es importante volver a precisar que el universo de 87 municipios actualizados entre 2016 – 2024 corresponden a aquellos municipios bajo gestión catastral del Instituto Geográfico Agustín Codazzi – IGAC, dado que en Colombia a través de la ley 1995 de 2019 *Plan Nacional de Desarrollo 2018-2022* "Pacto por Colombia, pacto por la equidad" se descentraliza las funciones de la gestión catastral a diferentes actores, incluyendo al sector privado, cada uno manejando sus sistemas de información de manera independiente, lo cual conlleva a que no exista al momento de este ejercicio una base de datos catastral consolidada de todos los predios a nivel nacional, si no que esta información se encuentra fragmentada de acuerdo a cada gestor catastral. Con base a lo anterior se usa la información del IGAC, entidad rectora en materia catastral del país y quien posee la mayor muestra de municipios actualizados.

Las bases catastrales del IGAC se conocen como Registro 1 y Registro 2, la primera contiene información física, jurídica y económica de cada predio, mientras la segunda brinda información de la tipología de construcción. Dado que el objetivo del presente trabajo es una predicción del ajuste de los valores catastrales, es decir de una parte de la información económica, la base de interés a manejar el Registro 1 – R1.

Fase 3 Preparar los Datos para el Modelado

Para preparar los datos para el modelado, primero es necesario conocer la información que se va a manejar, la siguiente tabla contiene un ejemplo de un predio que aparece en un R1:

Tabla 4

Ejemplo de Una Observación – Predio – en la Base Catastral R1

NUMERO_PREDIAL	TIPO_REGISTRO	NUMERO_ORDEN	TOTAL_REGISTROS	NOMBRE	TIPO_DOCUMENTO	NUMERO_DOCUMENTO
00010000000X000X000000000	1	001	001	Pepito Perez	N	100000XXX000000
DIRECCION	DESTINO_ECONOMICO	AREA_TERRENO_M2	AREA_CONSTRUIDA_M2	VALOR_AVALUO	VIGENCIA	NUMERO_PREDIAL_ANTERIOR
Marte	D	1156250	1648	1270065000	01012024	000X0001000X000

Nota. Adaptado de *Registro tipo 1 – R1,* IGAC, 2024

Como se puede observar en la tabla anterior, las bases catastrales contienen información confidencial – jurídica - como lo son el nombre del propietario del predio (NOMBRE), el tipo de documento (TIPO_DOCUMENTO), el número de este (NUMERO_DOCUMENTO), y la ubicación del predio (DIRECCION). Con lo anterior, un primer paso es eliminar estas variables para no manejar información confidencial, además de que no añaden valor agregado al presente trabajo. A parte de las variables anteriormente mencionadas también se procede a eliminar las columnas TIPO_REGISTRO, NUMERO_ORDEN, y TOTAL REGISTROS dado que no contienen datos que sean de utilidad, y como parte del procedimiento de ETL se recomienda manejar la dimensionalidad de la información de la manera más optima posible.

Las bases catastrales R1 además de contener a manera de observación las características de un predio también tienen en cuenta los propietarios de un predio, es decir que en cada fila un

predio se puede repetir varias veces según el número de propietarios, la siguiente tabla ilustra un ejemplo:

Tabla 5

Ejemplo de Un Predio con Dos Propietarios

NUMERO_PREDIAL	TIPO_REGISTRO	NUMERO_ORDEN	TOTAL_REGISTROS	NOMBRE	TIPO_DOCUMENTO	NUMERO_DOCUMENTO
00010000000X000X000000000	1	001	001	Pepito Perez	N	100000XXX000000
DIRECCION	DESTINO_ECONOMICO	AREA_TERRENO_M2	AREA_CONSTRUIDA_M2	VALOR_AVALUO	VIGENCIA	NUMERO_PREDIAL_ANTERIOR
Marte	D	1156250	1648	1270065000	01012024	000X0001000X000
NUMERO_PREDIAL	TIPO_REGISTRO	NUMERO_ORDEN	TOTAL_REGISTROS	NOMBRE	TIPO_DOCUMENTO	NUMERO_DOCUMENTO
00010000000X000X000000000	1	001	001	Juanito Perez	N	100X00XXX000000
DIRECCION	DESTINO_ECONOMICO	AREA_TERRENO_M2	AREA_CONSTRUIDA_M2	VALOR_AVALUO	VIGENCIA	NUMERO_PREDIAL_ANTERIOR
Marte	D	1156250	1648	1270065000	01012024	000X0001000X000

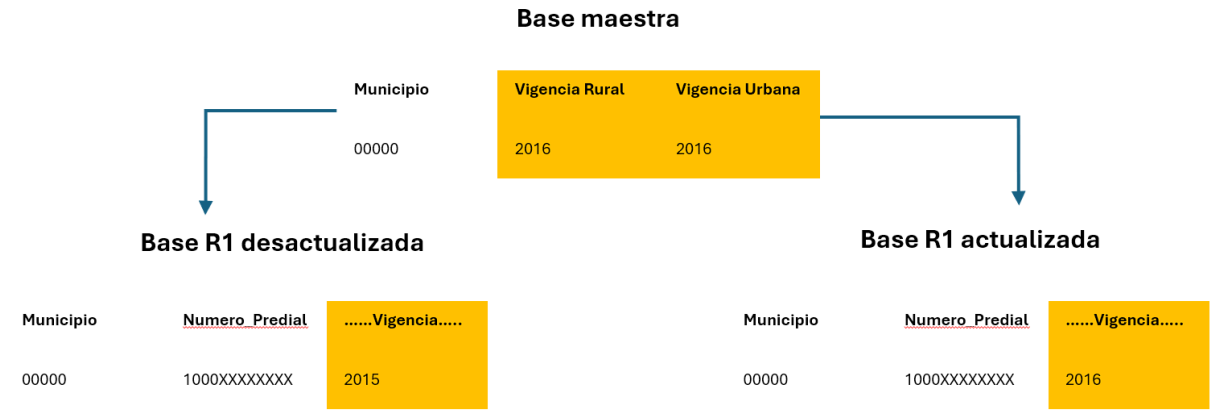
Nota. Adaptado de *Registro tipo 1 – R1*, IGAC, 2024

Dado que las bases catastrales contienen más de una observación por predio, y que nuestro ejercicio busca proyectar el incremento del avalúo por predio, es primordial eliminar estos duplicados por predio según el propietario. Para lograr esto se usa la variable NUMERO_PREDIAL que es el identificador de cada predio, eliminando todos los posibles duplicados de cada predio, si es el caso.

Después de eliminar la información confidencial y redundante de las bases, y hacer el tratamiento de los duplicados se procede a identificar el momento y alcance de una actualización catastral y como se representaría en la base. Para el primer caso se debe tener en cuenta la información disponible en la *Base Maestra* del IGAC, la cual contiene a nivel municipal una caracterización catastral, como son cantidad de predios, área construida, avalúo catastral, vigencia de la zona rural y vigencia de la zona urbana. La variable de interés son la vigencia rural y la vigencia urbana, que muestran el año de la última y vigente actualización de la base catastral del municipio segregado por zona. Por ejemplo, para un municipio A si la vigencia de la zona rural es 2016 y de la zona urbana es 2016 quiere decir que la actualización se realizó en 2016 abarcando el total del municipio, por lo tanto, para los efectos de nuestro ejercicio debe tomarse la base catastral vigencia 2015 – antes de actualización – frente a la vigencia 2016 – después de actualización. La siguiente ilustración muestra un ejemplo de un predio perteneciente al municipio hipotético con código DIVPOLA "00000" para identificar la base desactualizada y actualizada a través de la Base Maestra teniendo como variable clave la Vigencia.

Figura 3

Vinculación de las Bases para Determinar el Momento y Alcance de la Actualización



Habiendo identificado el momento de la actualización según la información disponible en la base maestra del IGAC, se procede a agrupar las bases catastrales desde 2015 hasta 2024, abarcando nueve vigencias catastrales. Con estas bases, se filtran los municipios según el año de actualización. Para este ejercicio, se identificaron 89 momentos distintos de actualización catastral.

Las actualizaciones catastrales pueden abarcar diferentes zonas de un municipio, lo que significa que un mismo municipio puede tener dos momentos de actualización durante el periodo mencionado. Por ejemplo, si la zona rural de un municipio se actualizó en 2016 y la zona urbana en 2020, se generan dos registros: uno para la actualización de la zona rural en 2016 y otro para la actualización de la zona urbana en 2020. A continuación, se presentan los municipios, el año de actualización y el alcance de esta:

Tabla 6

Municipios Actualizados Bajo Gestión Catastral IGAC: 2016 – 2024

Municipio (Código DIVIPOLA)	Año	Alcance de Actualización (Zona)
8137	2016	Urbano y Rural
8675	2016	Urbano y Rural
8770	2016	Urbano y Rural
13430	2016	Urbano y Rural
13440	2016	Urbano y Rural
13620	2016	Urbano y Rural
15047	2016	Urbano y Rural
15226	2016	Urbano y Rural
15822	2016	Urbano y Rural
23079	2016	Urbano y Rural
23350	2016	Urbano y Rural
68705	2016	Urbano y Rural

Municipio (Código DIVIPOLA)	Año	Alcance de Actualización (Zona)
23500	2017	Urbano y Rural
41615	2017	Urbano y Rural
15001	2018	Urbano y Rural
15367	2018	Urbano y Rural
15806	2018	Urbano y Rural
15837	2018	Urbano y Rural
25181	2018	Urbano y Rural
47460	2018	Urbano y Rural
73268	2018	Urbano y Rural
54673	2019	Urbano y Rural
68755	2019	Urbano y Rural
13836	2020	Urbano y Rural
20400	2020	Urbano y Rural
41396	2020	Urbano y Rural
70508	2020	Urbano y Rural
73449	2020	Urbano y Rural
15092	2023	Urbano y Rural
15114	2023	Urbano y Rural
15215	2023	Urbano y Rural
15537	2023	Urbano y Rural
15723	2023	Urbano y Rural
15755	2023	Urbano y Rural
15757	2023	Urbano y Rural
15790	2023	Urbano y Rural
19001	2023	Urbano y Rural
25817	2023	Urbano y Rural
41016	2024	Urbano y Rural
41306	2024	Urbano y Rural
50683	2024	Urbano y Rural

Municipio (Código DIVIPOLA)	Año	Alcance de Actualización (Zona)
81065	2024	Urbano y Rural
99624	2024	Urbano y Rural
99524	2024	Urbano y Rural
52240	2018	Rural
54405	2019	Rural
54874	2019	Rural
63401	2019	Rural
81001	2021	Rural
81220	2021	Rural
99773	2021	Rural
91407	2022	Rural
91536	2022	Rural
91798	2022	Rural
25295	2023	Rural
13212	2024	Rural
13248	2024	Rural
13654	2024	Rural
23678	2024	Rural
44279	2024	Rural
50287	2024	Rural
50577	2024	Rural
73067	2024	Rural
73616	2024	Rural
68745	2018	Urbano
19318	2019	Urbano
27001	2019	Urbano
15759	2020	Urbano
50313	2020	Urbano
99773	2020	Urbano

Municipio (Código DIVIPOLA)	Año	Alcance de Actualización (Zona)
25295	2021	Urbano
66045	2021	Urbano
66075	2021	Urbano
66088	2021	Urbano
66318	2021	Urbano
66383	2021	Urbano
66440	2021	Urbano
66572	2021	Urbano
66687	2021	Urbano
13212	2022	Urbano
13248	2022	Urbano
25612	2022	Urbano
23678	2023	Urbano
63401	2023	Urbano
73616	2023	Urbano
73873	2023	Urbano
81065	2023	Urbano
85250	2023	Urbano
19075	2024	Urbano

Durante la filtración se crean dos conjuntos de datos llamados “R1_actualizado” y “R1_desactualizado”, el primerio contiene la información de cada municipio en el momento anterior a la actualización catastral, el segundo por su parte contiene la información posterior a la actualización catastral. Se crean además la variable TIPO_ACTUALIZACION, que identifica el alcance de la actualización.

Tabla 7

Cantidad de Registros en la Base Actualizada y Desactualizada

Base	Observaciones
R1 actualizado	995.664
R1 desactualizado	825.545

Como se observa en la tabla anterior, la actualización catastral revelo la existencia de 170.119 predios adicionales. Esto se debe a la ponderación de los nuevos predios identificados y la cancelación de aquellos que ya no existen.

Sin embargo, los análisis sobre el impacto de la actualización catastral en el incremento del avalúo solo pueden realizarse considerando los predios que están presentes tanto en la base actualizada como en la desactualizada, es decir, aquellos que se pueden rastrear en ambas bases. Esto implica que la dinámica asociada a los predios nuevos o eliminados tras la actualización catastral queda fuera del alcance de este estudio.

El siguiente paso consistió en unificar las bases R1 actualizada y R1 desactualizada en una única base que contuviera exclusivamente los predios rastreables mediante su número predial. Como resultado, se obtuvo un universo de 568.833 observaciones. La disminución en la cantidad de observaciones se debe, en gran medida, a que una proporción significativa de predios experimentó cambios en su nomenclatura predial tras la actualización, lo cual puede estar relacionado con modificaciones en la delimitación del perímetro urbano, así como con procesos de fragmentación o fusión de predios.

Una vez que los datos han sido depurados y transformados, se integraron en un único conjunto de datos que incluye tanto los avalúos previos como los posteriores a la actualización, además incorpora información sobre las características físicas y económicas de los predios. Para facilitar el entendimiento y el análisis del dataset unificado, se presenta a continuación un

diccionario de datos que describe las principales variables incluidas, especificando su nombre, tipo, descripción y las unidades de medida correspondientes:

Tabla 8

Diccionario de Datos de la Base Unificada

Variable	Tipo	Descripción	Unidades
DESTINACION_ECONOMICA_DES	Categórica	Destino económico del predio antes de la actualización catastral.	No aplica
DESTINACION_ECONOMICA_ACT	Categórica	Destino económico del predio después de la actualización catastral.	No aplica
TIPO.y	Categórica	Alcance de la actualización catastral: Rural, Rural y urbana, o Urbana.	No aplica
tipo_predio.x	Categórica	Tipo de predio: rural o urbano.	No aplica
AREA_CONSTRUIDA_DES	Numérica	Área construida del predio antes de la actualización catastral.	Metros cuadrados
AREA_CONSTRUIDA_ACT	Numérica	Área construida del predio después de la actualización catastral.	Metros cuadrados
AREA_TERRENO_ACT	Numérica	Área de terreno del predio después de la actualización catastral.	Metros cuadrados
AVALUO_DES	Numérica	Valoración catastral del predio antes de la actualización catastral.	Pesos colombianos
AVALUO_ACT	Numérica	Valoración catastral del predio después de la actualización catastral.	Pesos colombianos

Finalizando, se aplicó una transformación logarítmica a los valores del avalúo catastral, dado el comportamiento de colas largas observado en la distribución de estos valores, donde pocos predios concentran una proporción significativa del avalúo total (Ver Apendice A). Este enfoque permitió estabilizar la varianza, mejorar la normalidad de los datos y optimizar su comportamiento en el modelamiento.

Con la construcción y depuración del dataset unificado, que incluye las variables clave para el análisis de los avalúos y las características asociadas a los predios antes y después de la actualización catastral, se dispone de una base sólida y estructurada para abordar el siguiente paso del estudio. Este conjunto de datos permite explorar las relaciones entre las variables y evaluar los efectos de la actualización catastral. A partir de este punto, se procederá al modelamiento, donde se implementarán técnicas estadísticas y econométricas para analizar y explicar las dinámicas observadas en los avalúos y su relación con las características de los predios y su contexto.

Fase 4 Modelamiento

Para abordar la predicción de los efectos de la actualización catastral en los valores prediales, se implementó un enfoque basado en la segmentación y modelado con aprendizaje automático. Inicialmente, se aplicó la metodología de clustering HDBSCAN, que permitió identificar grupos de predios con características homogéneas. La evaluación de la calidad del clustering se llevó a cabo mediante el índice de Davies-Bouldin (que mide la compacidad y separación de los clusters, donde valores más bajos indican mejor calidad), el índice de Calinski-Harabasz (que evalúa la varianza intra e inter-cluster, favoreciendo valores más altos para clusters bien definidos), y métricas adicionales como el índice ARI (Adjusted Rand Index, utilizado para medir la similitud entre clústeres generados y etiquetas de referencia, considerando

coincidencias ajustadas al azar) para comparar los clústeres originales con los generados mediante bootstrap y etiquetas aleatorias.

Posteriormente, para cada clúster identificado, se construyeron modelos de regresión utilizando XGBoost, un algoritmo de alto rendimiento especialmente adecuado para tareas predictivas con múltiples variables, como es el caso nuestro de variables categóricas y numéricas. La validación de los modelos incluyó una estrategia de validación cruzada k-fold (para dividir los datos en múltiples subconjuntos y evaluar el modelo de forma robusta en diferentes particiones, asegurando una buena generalización).

Las métricas empleadas para evaluar el desempeño de los modelos fueron el Mean Absolute Error (MAE) (que mide el error promedio absoluto entre los valores predichos y los reales, fácil de interpretar en la misma escala de la variable objetivo), el Mean Squared Error (MSE) (que penaliza errores grandes al elevarlos al cuadrado, destacando desviaciones significativas), y el coeficiente de determinación (R^2) (que indica qué proporción de la variabilidad en los datos es explicada por el modelo, donde valores más cercanos a 1 son deseables).

Adicionalmente, se realizaron pruebas de normalidad de los residuos mediante el test de Shapiro-Wilk (para evaluar si los errores siguen una distribución normal, una suposición clave en muchos modelos), junto con la visualización de histogramas y gráficos de densidad (para inspeccionar visualmente la distribución de los errores) y comparar los valores predichos con los observados (Ver Apéndice B). Estos análisis aseguraron que las predicciones de cada modelo fueran consistentes y adecuadas para cada clúster.

Este enfoque combinado de clustering y modelado permitió captar las heterogeneidades presentes en los datos y desarrollar predicciones más precisas y contextualizadas sobre los efectos de la actualización catastral.

Fase 5 Evaluación

En la fase de evaluación, se validaron los modelos desarrollados tanto a nivel de clústeres como globalmente, para asegurar su precisión y utilidad en el contexto del proyecto. Para ello, se utilizaron métricas de desempeño como el MAE, el MSE y el coeficiente de determinación (R^2), que permitieron medir la capacidad predictiva de los modelo.

Además, se llevó a cabo un análisis de sensibilidad para identificar las variables más relevantes en los cambios del avalúo catastral. Esto incluyó el análisis de la importancia de las variables generadas por los modelos XGBoost, a través de la función *feature,* lo cual proporcionó información sobre los factores que influyen en los resultados (Ver Apéndice B).

Finalmente, se analizaron los resultados obtenidos a través de gráficos de comparación entre valores observados y predichos, diferenciados por clúster, así como histogramas y densidades de los residuos para evaluar el ajuste de los modelos. En el caso del clustering, se revisaron las métricas de validación interna y externa, como el índice de Davies-Bouldin y el ARI, para asegurar que los agrupamientos reflejaran las estructuras reales de los datos.

Como resultado de esta evaluación, se verificó que los modelos cumplieran con los objetivos del proyecto, y se identificaron posibles ajustes o mejoras necesarias antes de avanzar hacia la implementación final de los modelos predictivos. Este proceso garantiza que las predicciones realizadas sean confiables y útiles para el análisis de los efectos de la actualización catastral.

Fase 6 Despliegue

El modelo predictivo será implementado para generar proyecciones de los municipios que no han sido actualizados catastralmente. Estas proyecciones se utilizarán para estimar los incrementos futuros en los avalúos catastrales según grupos de predios con características

símiles, permitiendo a las autoridades municipales tener un insumo que sirva de indicador para anticipar los efectos fiscales de una potencial actualización catastral. Finalmente se generará un informe final, que incluirá un análisis a detalle de los resultados, potenciales mejoras y líneas de investigación, para la toma de decisiones a nivel municipal.

Resultados

Clústeres por metodología HDSCAN

La tabla 9 y la figura 4 muestran la caracterización de cada clúster. En total surgieron 7 clústeres que abarcan el 94.39% de predios, solo el 5.61% termino como ruido. El Clúster 4 cuenta con la mayor cantidad de predios (318,991) y un uso predominantemente habitacional antes y después de la actualización (99.85% y 99.83%, respectivamente) y urbano (92.24%). El Clúster 5 presenta el mayor avalúo promedio antes de la actualización ($96,124,100), con un uso exclusivamente comercial (100%) y una predominancia urbana del 96.60%. Por su parte, el Clúster 3 sobresale tanto por el mayor incremento absoluto en avalúo promedio (de $17,525,000 a $142,476,800) como por su transformación hacia actividades agrícolas realizadas en la principalmente zona urbana (85%), con el 98.78% de los predios dedicados a este fin después de la actualización. El Clúster 6 se caracteriza por su enfoque exclusivamente agropecuario (100%) y su naturaleza predominantemente rural (82.28%). Finalmente, el clúster 2 representa de manera preminente los lotes de engorde

Tabla 9

Caracterización de Clústeres HDBSCAN

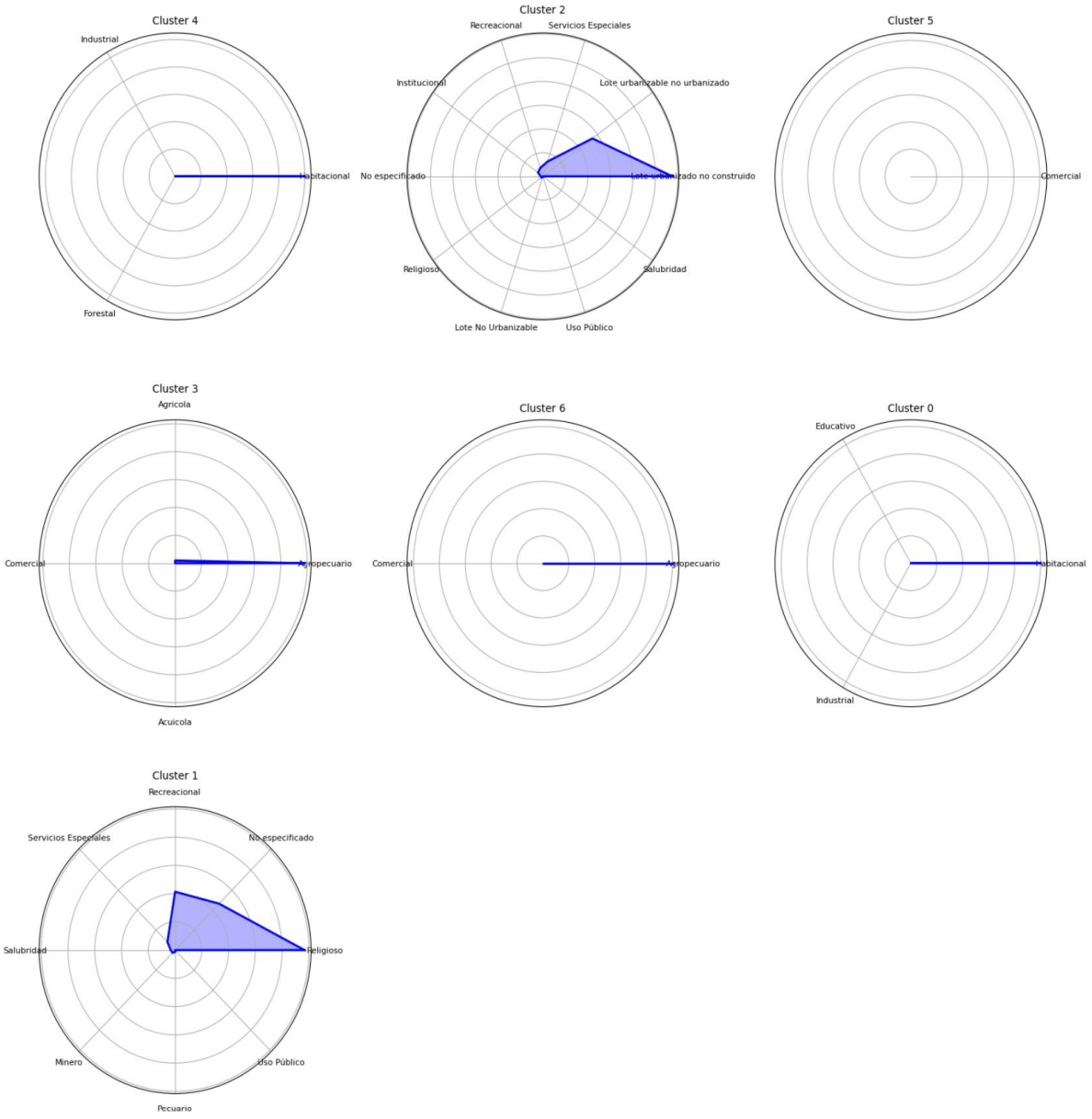
Clúster	Cantidad de Predios	Promedio de Área Construida (Antes)	Promedio de Área Construida (Después)	Promedio de Área de Terreno	Promedio de Avalúo (Antes)	Promedio de Avalúo (Después)	Destino Económico (Antes)	Destino Económico (Después)	Tipo de Actualización	Zona
4	318,991	84.2	97.9	592.7	$28,655,880	$54,373,330	Habitacional (99.85%), Industrial (0.14%),	Habitacional (99.83%), Industrial (0.14%),	RURAL Y URBANA (65.18%), URBANA (30.39%),	Urbano (92.24%), Rura

Clúster	Cantidad de Predios	Promedio de Área Construida (Antes)	Promedio de Área Construida (Después)	Promedio de Área de Terreno	Promedio de Avalúo (Antes)	Promedio de Avalúo (Después)	Destino Económico (Antes)	Destino Económico (Después)	Tipo de Actualización	Zona
							Forestal (0.005%)	Educativo (0.02%)	RURAL (4.43%)	l (7.76%)
2	42,457	12.4	20.3	1767.1	$24,318,980	$47,480,550	Lote urbanizado (57.58%), Lote urbanizable (27.18%), Servicios Especiales (6.57%)	Lote urbanizado (71.54%), Lote urbanizable (23.64%), Institucional (2.69%)	RURAL Y URBANA (70.01%), URBANA (26.28%), RURAL (3.70%)	Urbano (83.25%), Rural (16.75%)
5	12,770	139.6	147.4	534.2	$96,124,100	$143,132,700	Comercial (100%)	Comercial (100%)	RURAL Y URBANA (59.87%), URBANA (38.83%), RURAL (1.30%)	Urbano (96.60%), Rural (3.40%)
3	12,602	38.8	69.7	314078.6	$17,525,000	$142,476,800	Agropecuario (97.98%), Agrícola (1.96%), Comercial (0.05%)	Agrícola (98.78%), Agroforestal (0.73%), Acuícola (0.49%)	RURAL Y URBANA (97.88%), RURAL (1.79%), URBANA (0.32%)	Urbano (84.96%), Rural (15.04%)
6	127,089	30.6	47.0	239522.9	$15,006,290	$46,468,400	Agropecuario (99.99%), Comercial (0.0016%)	Agropecuario (100%)	RURAL Y URBANA (87.44%), RURAL (9.26%), URBANA (3.30%)	Rural (82.28%), Urbano (17.72%)

Clúster	Cantidad de Predios	Promedio de Área Construida (Antes)	Promedio de Área Construida (Después)	Promedio de Área de Terreno	Promedio de Avalúo (Antes)	Promedio de Avalúo (Después)	Destino Económico (Antes)	Destino Económico (Después)	Tipo de Actualización	Zona
0	10,951	10.5	6.4	598.7	$5,927,313	$27,818,720	Habitacional (99.83%), Educativo (0.10%), Industrial (0.06%)	Lote urbanizado (87.59%), Lote urbanizable (10.09%), Lote No Urbanizable (1.42%)	RURAL Y URBANA (92.90%), URBANA (6.82%), RURAL (0.27%)	Urbano (83.95%), Rural (16.05%)
1	12,041	42.4	54.8	13663.4	$27,032,590	$73,568,600	Religioso (48.37%), No especificado (23.25%), Recreacional (20.67%)	Religioso (48.36%), No especificado (23.25%), Recreacional (20.68%)	RURAL Y URBANA (54.77%), URBANA (35.35%), RURAL (9.87%)	Urbano (71.57%), Rural (28.43%)
Ruido	31,932 (5.61%)									

Figura 4

Ilustración Radial con la Proporción de Predios en los Clústeres Según Destino Económico

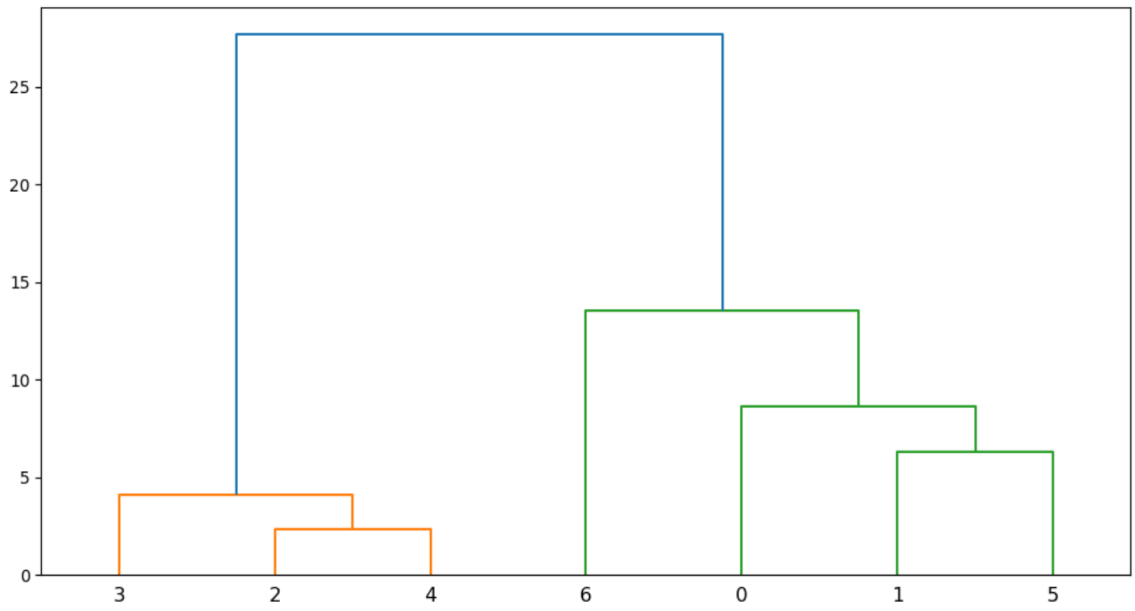


El dendrograma proporciona una representación jerárquica de las relaciones entre los clústeres obtenidos en el análisis. Los clústeres que se unen en niveles bajos, cerca del eje X, comparten mayores similitudes, mientras que aquellos que se agrupan a alturas mayores son más

distintos entre sí. Asimismo, los clústeres prominentes, que aparecen unidos a alturas significativamente mayores, destacan por su carácter distintivo. Es importante señalar que el ruido y los valores atípicos, excluidos del cálculo, no se reflejan en los centroides ni en el análisis. El grafico X muestra que existen dos agrupaciones predominantes, los clústeres, 5, 1 , 0 y 6; mientras que por otro lado se encuentran los clústeres 4, 2 y 3. Las parejas de clústeres 5 - 1 y 2 – 4 presentan mayor similitud. En el caso de la pareja 5 – 1 se debe principalmente a la proporcionalidad del tipo de actualización que para ambos casos muestra una proporción del 55 – 59 % de actualización completa mientras que alrededor del 36 – 38 % de los predios de estos clústeres fueron actualizados en la zona urbana. Por su parte la pareja 2 – 4 que representan los predios habitacionales e industriales de mayoría urbana y los lotes de engorde también de mayoría urbana muestran similitud en la diversidad de destinos económicos que se encuentran en ellos, un avalúo promedio similar y también una proporción de predios bajo un tipo de actualización similar: 65 – 70% completa y 26 – 30% urbana.

Figura 5

Dendrograma Basado en Centroides del Clúster HDBSCAN



En el análisis previo se presentaron los resultados de los clústeres generados, destacando sus características principales y patrones diferenciadores. A continuación, en la tabla 10, se expone la validación de dichos clústeres, que permite evaluar la calidad y consistencia de la estimación. Los resultados de esta validación reflejan una agrupación satisfactoria y características claramente diferenciadas. El índice Davies-Bouldin (DB) para los clústeres reales es de 1.7, un valor bajo que indica clústeres compactos y bien definidos. En contraste, el índice DB de los clústeres aleatorios es significativamente más alto (528.07), lo que evidencia su baja calidad. La diferencia de -526.37 en el índice DB (Real - Aleatorio) refuerza la superioridad de los clústeres reales sobre los generados aleatoriamente. Además, el índice de Calinski-Harabasz (CH), con un valor notablemente alto de 270,568.51, confirma una fuerte cohesión interna y una separación clara entre clústeres. Por último, el índice de Rand Ajustado (ARI) entre los clústeres originales y el bootstrap es de 0.26, indicando una estabilidad moderada en la evaluación de consistencia. En conjunto, estos indicadores validan de manera robusta la efectividad del modelo de clustering HDBSCAN empleado en el presente trabajo.

Tabla 10

Validación de Clústeres HDBSCAN

Métrica	Resultado	Inferencia
Índice de Davies-Bouldin (DB)	1.7	1.7 es un valor bajo, lo que indica que los clústeres son compactos y bien separados.
Índice de Davies-Bouldin (Aleatorio)	528.07	528.07 es un valor alto, lo que sugiere que los clústeres aleatorios son mucho peores que los reales.
Diferencia DB (Real - Aleatorio)	-526.37	-526.37 es una diferencia significativa, lo que confirma que los clústeres reales son significativamente mejores que los aleatorios.
Índice de Calinski-Harabasz (CH)	270568.51	270,568.51 es un valor muy alto, lo que indica que los clústeres son compactos internamente y bien separados entre sí.
ARI entre clústeres originales y bootstrap	0.26	0.26 indica una estabilidad moderada en los clústeres al comparar los originales con el bootstrap.

Regresión por XGBoost

La tabla presentada a continuación resume el comportamiento del incremento observado y predicho del avalúo catastral, desglosado por clúster, incluyendo el promedio, la mediana y el rango intercuartílico (IQR) de los incrementos. Dado que los avalúos presentan un sesgo a la derecha, con pocos predios concentrando la mayor parte del avalúo total y ciertos incrementos extremos que actúan como observaciones atípicas, el promedio puede no ser una medida representativa. Por esta razón, se incluye la mediana como indicador central más robusto y el IQR para reflejar la dispersión típica de los datos. Los rangos esperados de los incrementos, basados en el IQR (entre el primer y tercer cuartil), proporcionan una idea de dónde es más probable que se ubiquen los valores típicos de cada clúster, excluyendo la influencia de valores extremos.

El clúster 3 destaca como el grupo con el mayor incremento promedio observado (41.74%) y mediana (10.81%). Su rango intercuartílico (IQR) observado es de 21.14 puntos porcentuales, indicando que es factible que los incrementos para este clúster se encuentren entre 3.02% y 24.16%. Esto refleja una alta dispersión dentro del grupo, probablemente atribuida a valores atípicos o diferencias marcadas en los predios.

En contraste, el clúster 5 muestra el menor incremento promedio (0.68%) y mediana (0.16%), con un IQR observado de 0.59 puntos porcentuales. Esto sugiere que los incrementos típicos para este clúster se encuentran en un rango estrecho entre 0.00% y 0.59%, indicando mayor homogeneidad en este grupo.

Tabla 11

Resultados de la Estimación de Incremento del Avalúo Catastral Según Clústeres – Observado vs Pronosticado

Clúster	Promedio Incremento Observado	Promedio Incremento Predicho	Mediana Incremento Observado	Mediana Incremento Predicho	IQR Incremento Observado	IQR Incremento Predicho	Rango Incremento Observado (Q1, Q3)	Rango Incremento Predicho (Q1, Q3)
4	5.6	5.8	0.6	0.6	2.0	1.7	(0.045, 1.996)	(0.228, 1.960)
2	11.7	11.6	0.8	1.1	3.2	2.7	(0.000, 3.181)	(0.251, 2.956)
5	0.7	0.9	0.2	0.3	0.6	0.5	(0.000, 0.589)	(0.056, 0.535)
3	41.7	41.4	10.8	11.3	21.1	17.3	(3.022, 24.163)	(4.389, 21.680)
6	15.6	16.3	2.0	2.9	6.3	6.0	(0.030, 6.286)	(0.199, 6.214)
0	14.9	14.9	3.5	3.3	8.8	8.5	(1.281, 10.122)	(1.480, 9.958)
1	14.2	14.5	0.6	0.8	5.3	5.5	(0.000, 5.295)	(0.007, 5.470)

Para corroborar las estimaciones anteriores, a continuación, se desarrollará una validación de las estimaciones generadas por los modelos XGBoost aplicados a los clústeres definidos mediante el algoritmo HDBSCAN, con el objetivo de evaluar la precisión en la predicción del incremento del avalúo catastral tras la actualización. Para ello, se presenta un gráfico de dispersión que ilustra la relación entre los valores observados y los valores predichos por el modelo.

En la figura 6, el eje X representa los valores observados del incremento del avalúo, mientras que el eje Y muestra los valores pronosticados. Cada punto verde simboliza una observación específica, ubicada de acuerdo con sus valores observados y predichos. La línea

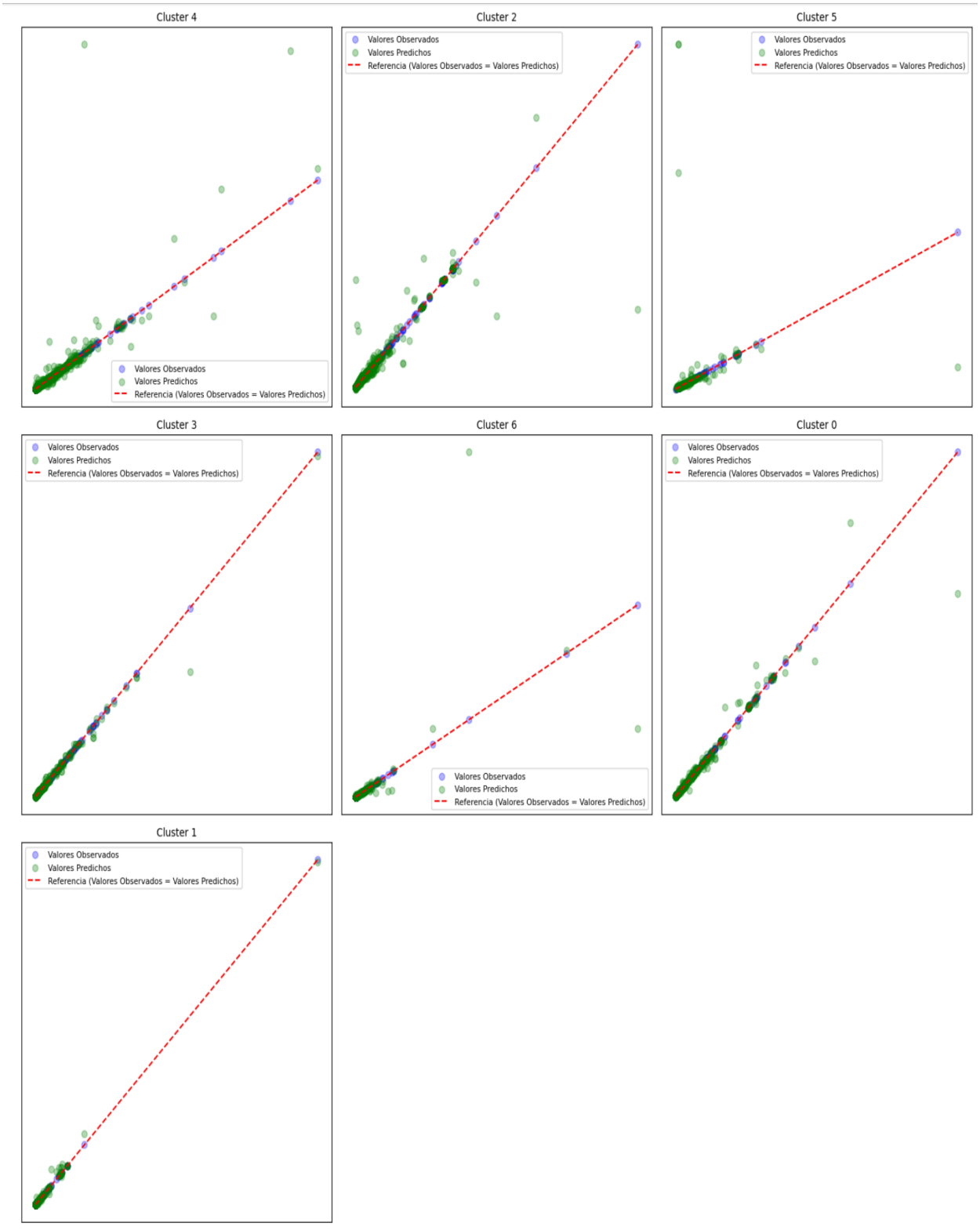
roja, con una pendiente de 45 grados, sirve como referencia ideal: si el modelo predijera de manera perfecta, todos los puntos se ubicarían exactamente sobre esta línea.

La cercanía de los puntos a la línea de referencia indica la precisión del modelo: puntos más próximos a la línea reflejan predicciones acertadas, mientras que puntos alejados evidencian discrepancias entre los valores observados y predichos. Además, la ubicación de los puntos respecto a la línea permite identificar posibles tendencias del modelo, como sobreestimaciones (puntos por encima de la línea) o subestimaciones (puntos por debajo).

Los clústeres con mejor desempeño son el 3, 1 y 0, ya que sus puntos están más próximos a la línea de referencia. En contraste, los clústeres 4 y 2 muestran puntos significativamente más alejados de dicha línea. Por otro lado, los clústeres 5 y 6 presentan pocas predicciones fuera de la línea de referencia, aunque estas tienden a estar considerablemente distantes. En términos generales, las estimaciones muestran un comportamiento satisfactorio para valores bajos; sin embargo, en el caso de incrementos más elevados, considerados como valores atípicos, las predicciones tienden a desviarse ligeramente.

Figura 6

Valores del Incremento del Avalúo Pronosticado Frente a Valores Observados



Ahora se presentan las métricas de validación del modelo de regresión XGBoost aplicadas a los diferentes clústeres con el propósito de evaluar el desempeño predictivo en cada grupo. Estas incluyen el Error Absoluto Medio (MAE), la prueba de validación cruzada, el Error Cuadrático Medio (MSE) y el coeficiente de determinación (R^2), proporcionando información sobre la precisión y capacidad explicativa del modelo. Los resultados muestran diferencias significativas entre los clústeres, con algunos presentando ajustes satisfactorios (R^2 cercanos a 1) y otros evidenciando limitaciones en la predicción (R^2 negativos).

Los clústeres 1, 3, y 0 destacan por su sólida capacidad predictiva, mientras que el clúster 5 y 6 muestra un comportamiento significativamente deficiente. Esto sugiere que el modelo se ajusta bien para ciertos grupos de datos, pero enfrenta dificultades en otros.

Tabla 12

Métricas de Evaluación de los Modelos de Regresión XGBoost por Clústeres

Clúster	MAE (Validación Cruzada)	MAE	MSE	R^2	Prueba de Normalidad (p-valor)	Interpretación
4	0.8477	0.83378	709.2562	0.6454	0	Desempeño moderado, con un R^2 aceptable, pero errores algo elevados y sin normalidad en los datos.
2	1.5104	1.63675	335.1517	0.8925	0	Buen ajuste del modelo, con errores relativamente bajos, R^2 alto, pero datos no normales.
5	0.3912	0.41	45.9254	-5.7427	0	Desempeño pobre, con un R^2 negativo y datos no normales, indicando que el modelo no explica bien.
3	43.397	4.08839	409.6618	0.9842	0	Excelente desempeño en R^2, pero con un MAE elevado y datos no normales.
6	11.5061	4.76706	694	-0.171	0	Predicciones poco confiables, con errores muy altos, R^2 negativo y datos no normales.
0	1.3929	1.06649	58.5502	0.9606	0	Buen desempeño, con errores bajos y un R^2 cercano a 1, aunque los datos no son normales.

Clúster	MAE (Validación Cruzada)	MAE	MSE	R^2	Prueba de Normalidad (p-valor)	Interpretación
1	2.2161	1.2939	59.5173	0.9945	0	Ajuste excepcional, con el R^2 más alto y errores bajos, aunque sin cumplir la normalidad en los datos.

En concordancia con las métricas de evaluación, las estimaciones del modelo son precisas y funcionales para predios agrícolas, agropecuarios y acuícolas ubicados en zonas urbanas con actualizaciones completas; predios destinados a usos religiosos y recreacionales en zonas urbanas; y predios habitacionales y educativos también en zonas urbanas con actualizaciones completas. Con mayor cautela, los resultados estimados pueden considerarse válidos para lotes de engorde bajo una actualización completa, así como para predios habitacionales e industriales urbanos, ya sea en el contexto de una actualización completa o exclusivamente urbana. La tabla 13 presenta la mediana y el rango intercuartílico de los incrementos en el avalúo catastral. Como se mencionó anteriormente, la mediana se considera un indicador más representativo que el promedio para este análisis, ya que es menos sensible a valores extremos. Por su parte, el rango intercuartílico describe los intervalos de incremento que podrían considerarse esperados o no sorprendentes para los grupos de predios analizados.

Tabla 13

Incremento del Avalúo Catastral Según Grupos de Predios

Clúster	Validez de Pronósticos	Mediana Incremento Pronosticado	Rango Esperado del Incremento Pronosticado (IQR)
1	Válidos para predios religiosos, recreacionales y no especificados en zonas urbanas.	0.8	Entre 0.07 y 5.470
3	Válidos para predios agrícolas, agropecuarios y acuícolas en zonas urbanas con actualizaciones completas.	11.3	Entre 4.389 y 21.680

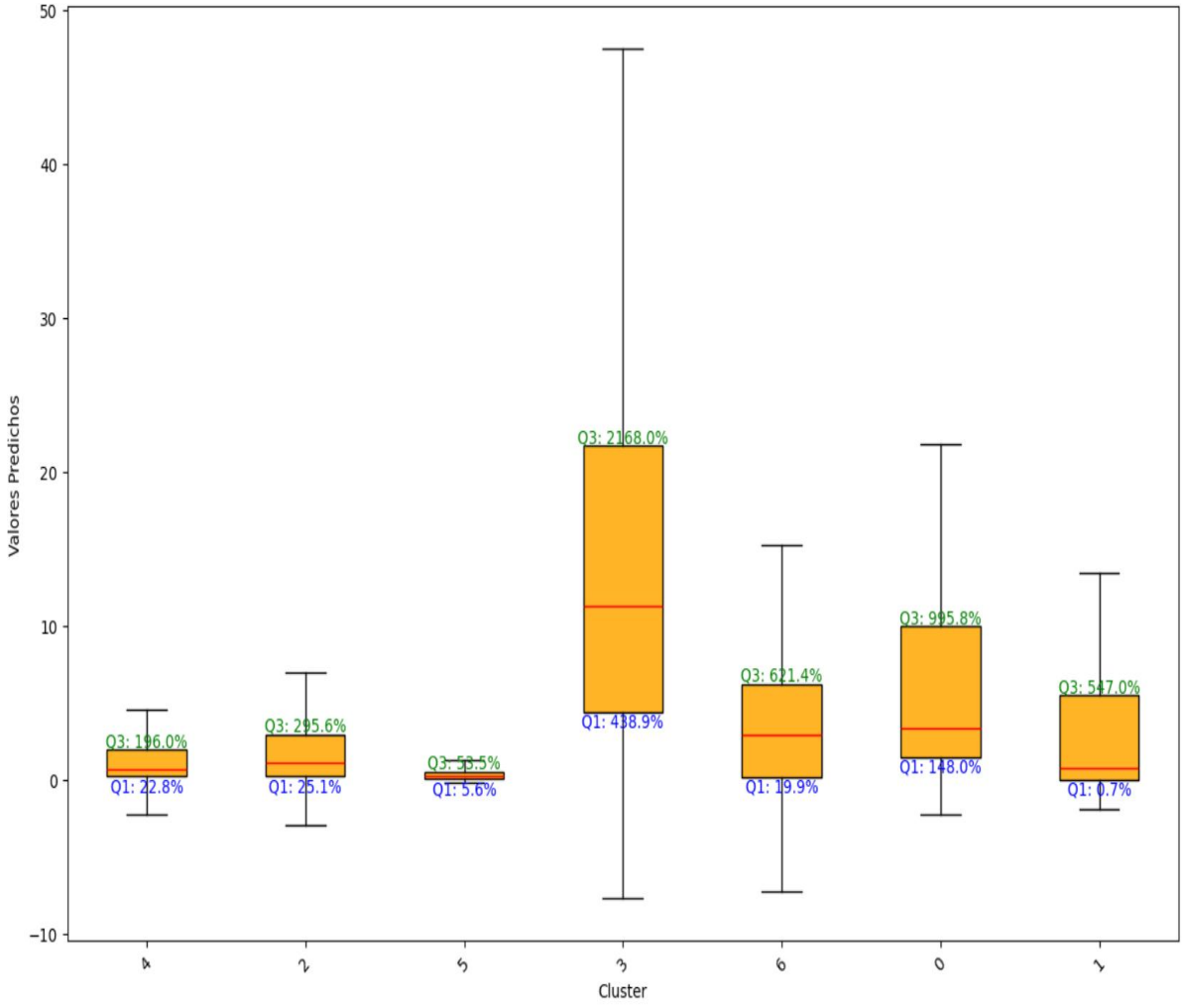
Clúster	Validez de Pronósticos	Mediana Incremento Pronosticado	Rango Esperado del Incremento Pronosticado (IQR)
0	Válidos para predios habitacionales y educativos en zonas urbanas con actualizaciones completas.	3.3	Entre 1.480 y 9.958
4	Válidos para predios habitacionales e industriales en zonas urbanas con actualizaciones completas o urbanas.	0.6	Entre 0.228 y 1.960
2	Razonablemente válidos para lotes urbanizados y urbanizables en zonas urbanas con actualizaciones completas.	1.1	Entre 0.251 y 2.956
6	No válidos para predios agropecuarios en zonas rurales.	-	-
5	No válidos para predios comerciales en zonas urbanas.	-	-

La figura 7 es un gráfico de cajas que muestra la distribución de los valores predichos para cada clúster generado en el análisis. Cada caja representa los valores comprendidos entre el primer cuartil (Q1) y el tercer cuartil (Q3), con una línea roja que indica la mediana. Los bigotes se extienden hasta el rango intercuartílico ajustado por 1.5 (IQR×1.5) para visualizar posibles valores extendidos. Además, las etiquetas junto a las cajas indican los valores de Q1 y Q3 expresados como porcentaje en cada clúster. Dado los resultados del grafico 3.2.1 los valores extremos o atípicos pronosticados son relegados en tanto estas estimaciones no son satisfactorias. El clúster 3 - predios agrícolas, agropecuarios y acuícolas en zonas urbanas con actualizaciones completas - presenta mayor varianza en su pronóstico y rangos de incrementos más altos, por el contrario, los clúster 2 - lotes urbanizados y urbanizables en zonas urbanas con actualizaciones completas – y 4 - predios habitacionales e industriales en zonas urbanas con actualizaciones completas o urbanas- presentan menor dispersión en el pronóstico. Finalmente, los clústeres 0 - predios habitacionales y educativos en zonas urbanas con actualizaciones completas - y 1-

predios religiosos, recreacionales y no especificados en zonas urbanas – muestran una dispersión media, las alta que los clústeres 2 y 4 pero menor al clúster 3.

Figura 7

Boxplot de Valores del Incremento del Avalúo Pronosticado Según Clúster



Conclusiones

Se desarrolló una base de datos que, tras un proceso de depuración, consolidó un universo de 568.833 predios, incluyendo información tanto previa como posterior a la actualización catastral. Este conjunto de datos fue empleado en el modelo de clusterización mediante la metodología HDBSCAN, permitiendo segmentar los predios en siete clústeres o agrupaciones con características similares. Dichas características incluyen el tamaño, el área construida, la zona, el tipo de actualización y el avalúo catastral antes y después de la actualización.

La posterior aplicación de modelos de regresión XGBoost para predecir incrementos en el avalúo catastral permitió evaluar la precisión y confiabilidad de las estimaciones en distintas categorías de predios.

En los predios religiosos, recreacionales y no especificados, ubicados en zonas urbanas con actualización completa o urbana, el modelo mostró un desempeño sobresaliente, con incrementos pronosticados de una mediana del 80% y un rango esperado entre 7% y 547%, mostrando coherencia con los valores observados. Por otro lado, los predios agrícolas, agropecuarios y acuícolas en zonas urbanas con actualización completa presentaron los mayores incrementos estimados, con una mediana del 1,130% y un rango entre 439% y 2,168%, lo que evidencia la capacidad del modelo para capturar las características distintivas de estos predios.

En los predios habitacionales y educativos, ubicados también en zonas urbanas con actualización completa, el modelo predijo incrementos con una mediana del 330% y un rango esperado entre 148% y 996%. De manera similar, los predios habitacionales e industriales en zonas urbanas con actualización completa o urbana muestra potenciales incrementos con una mediana del 60% y un rango esperado entre 22.8% y 196%. Aunque las predicciones fueron aceptables, este grupo presentó un margen de error mayor en comparación con otros.

En el caso de los lotes urbanizados y urbanizables en zonas urbanas con actualización completa, las predicciones mostraron una mediana del 110% y un rango esperado entre 25.1% y 295.6%, con resultados adecuados, aunque con cierta variabilidad que podría mejorarse. Sin embargo, en los predios agropecuarios ubicados en zonas rurales con actualización completa, el modelo mostró un desempeño limitado, con incrementos pronosticados de una mediana del 290% y un rango entre 19.9% y 621.4%, reflejando dificultades del modelo para capturar las dinámicas de este grupo.

Finalmente, en los predios comerciales en zonas urbanas con actualización completa o urbana, el desempeño del modelo fue deficiente, indicando que el modelo no logró representar adecuadamente las características de estos predios.

Para la evaluación de la segmentación mediante HDBSCAN se usaron las métricas de Índice de Davies-Bouldin (DB), Índice de Davies-Bouldin (Aleatorio), Diferencia DB (Real - Aleatorio), Índice de Calinski-Harabasz (CH) y el ARI entre clústeres originales y Bootstrap. Los resultados de la evaluación de los clústeres mostraron que eran compactos y estables, uniformes y no aleatorios. Por su parte las métricas de evaluación de las regresiones XGBoost fueron el MAE, Validación Cruzada, R2 y la pruebas de normalidad de errores; se demostró un desempeño satisfactorio en grupos específicos, principalmente en zonas urbanas y predios con actualizaciones completas. No obstante, las limitaciones observadas en predios agropecuarios rurales y comerciales urbanos subrayan la necesidad de mayor información, enfoques complementarios o ajustes en el modelo para mejorar la confiabilidad en estos casos.

Aunque las técnicas empleadas en este trabajo mostraron resultados positivos en varios contextos, el análisis puede beneficiarse de la inclusión de información adicional en futuras investigaciones. Por ejemplo, contar con variables espaciales como coordenadas geográficas o

características del entorno de los predios podría mejorar la precisión de las predicciones, particularmente en zonas rurales donde el desempeño del modelo fue más limitado. Esto podría proporcionar una visión más completa de los factores que influyen en los avalúos catastrales en la zona rural.

Referencias

Ahmad, E., Brosio, G., & Jiménez, J. P. (2019). *Repensando el impuesto a la propiedad en América Latina: recaudación potencial, resistencia política y opciones de reforma*. *VIII* Jornadas Iberoamericanas de Financiación Local, Universidad Iberoamericana, México DF.

Ali, R. H., Graves, J., Wu, S., Lee, J., & Linstead, E. (2020). A machine learning approach to delineating neighborhoods from geocoded appraisal data. *ISPRS International Journal of Geo-Information, 9(7), 451*. https://doi.org/10.3390/ijgi9070451.

Andrade Pérez, F. W. (2023). Nuevo modelo de gestión catastral en Colombia: Herramienta para generar nuevas oportunidades de crecimiento económico. *Revista Estrategia Organizacional, 12*(1), 103–122.

Asamblea Nacional Constituyente. (1991). *Constitución Política de Colombia*. https://www.suin-juriscol.gov.co/viewDocument.asp?ruta=Constitucion/1687988.

Buitrago-Mora, D., & Garcia-López, M.-A. (2023). Real estate prices and land use regulations: Evidence from the Law of Heights in Bogotá. *Regional Science and Urban Economics, 101*, Article 103914. https://doi.org/10.1016/j.regsciurbeco.2023.103914.

Cesario, E., Manco, G., & Ortale, R. (2007). Top-down parameter-free clustering of high-dimensional categorical data. *IEEE Transactions on Knowledge and Data Engineering, 19*(12), 1607–1624.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785.

Cienciała, A., Sobolewska-Mikulska, K., & Sobura, S. (2021). Credibility of the cadastral data on land use and the methodology for their verification and update. *Land Use Policy, 102*, 105204. https://doi.org/10.1016/j.landusepol.2020.105204.

Congreso de la República de Colombia. (1983). *Ley 14 de 1983, por la cual se fortalecen los fiscos de las entidades territoriales y se dictan otras disposiciones* https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=267#14.

Congreso de la República de Colombia. (1990). *Ley 44 de 1990, por la cual se dictan normas sobre catastro e impuestos sobre la propiedad raíz, se dictan otras disposiciones de carácter tributario, y se conceden unas facultades extraordinarias* https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=283.

Congreso de la República de Colombia. (1995). *Ley 242 de 1995*, *por la cual se modifican algunas normas que consagran el crecimiento del índice de precios al consumidor del año anterior como factor de reajuste de valores, y se dictan otras disposiciones* https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=6078#6.

Congreso de la República de Colombia. (2011). *Ley 1450 de 2011*, *por la cual se expide el Plan Nacional de Desarrollo, 2010-2014* https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=43101.

Congreso de la República de Colombia. (2019). *Ley 1995 de 2019, por medio de la cual se dictan normas catastrales e impuestos sobre la propiedad raíz y se dictan otras disposiciones de carácter tributario territorial*. https://www.suin-juriscol.gov.co/viewDocument.asp?ruta=Leyes/30037810.

Congreso de la República de Colombia. (2023). *Proyecto de Ley 292C de 2023, por el cual se adoptan medidas en materia de Impuesto Predial Unificado, se modifica parcialmente la*

Ley 44 de 1990, se deroga la ley 1995 y se dictan otras disposiciones https://www.camara.gov.co/sites/default/files/2023-11/PL.292-2023C%20%28IMPUESTO%20PREDIAL%29.pdf.

Demir, S., & Sahin, E. K. (2023). An investigation of feature selection methods for soil liquefaction prediction based on tree-based ensemble algorithms using AdaBoost, gradient boosting, and XGBoost. *Neural Computing and Applications, 35*(4), 3173–3190. https://doi.org/10.1007/s00521-022-07356-9

Departamento Nacional de Planeación. (2020). *Actualización catastral con enfoque multipropósito*. https://proyectostipo.dnp.gov.co/images/pdf/CatastroMultiproposito/GUIA_METODOLOGICA_FINAL_3112020_VF_DANE_IGAC.pdf

Duarte, J. A. (2022). El catastro multipropósito como una construcción que parte de la comunidad: Propuesta para alcanzar una visión con propiedad. *Revista Equidad y Desarrollo, 36*(36), 1–180.

Eckert, J. K. (1990). *Property appraisal and assessment administration*. International Association of Assessing Officers.

Enache, C. (2021). *Sources of government revenue in the OECD* (Fiscal Fact No. 748). Tax Foundation. https://files.taxfoundation.org/20210210172143/Sources-of-Government-Revenue-in-the-OECD-2021.pdf.

Evgeny, A. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications, 39*(2), 1772–1778. https://doi.org/10.1016/j.eswa.2011.08.077.

Gallego, J., Gutiérrez, L., López, D., & Sepulveda, C. (2014). *Subsidios cruzados en servicios públicos domiciliarios basados en el avalúo catastral* (No. 12257). Universidad del Rosario.

Gobierno Nacional y FARC-EP. (2016). *Acuerdo final para la terminación del conflicto y la construcción de una paz estable y duradera*. https://bit.ly/2LUs0qp.

Guadalajara, N., López, M. Á., Iftimi, A., & Usai, A. (2021). Influence of the cadastral value of the urban land and neighborhood characteristics on the mean house mortgage appraisal. *Land, 10*(3), Article 250. https://doi.org/10.3390/land10030250.

Hou, J., Gao, H., & Li, X. (2016). DSets-DBSCAN: A parameter-free clustering algorithm. *IEEE Transactions on Image Processing, 25*(7), 3182–3193. https://doi.org/10.1109/TIP.2016.2558679.

Iregui, A., Melo, B. L., & Ramos, J. F. (2005). El Impuesto Predial en Colombia: Factores explicativos del recaudo. *Borradores de Economía, Banco de la República de Colombia*.

Laskin, M. B., Gadasina, L. V., & Zaytseva, E. A. (2021). The cadastral value as a tool for monitoring the real estate market value. *St. Petersburg University Journal of Economic Studies, 37*(1), 84–108. https://doi.org/10.21638/spbu05.2021.104.

Lozano-Gracia, N., & Anselin, L. (2012). Is the price right?: Assessing estimates of cadastral values for Bogotá, Colombia. *Annals of Regional Science*. https://doi.org/10.1111/j.1757-7802.2012.01062.x.

Martínez, I. I., & Marín, B. G. (2015). Impacto financiero en los contribuyentes del impuesto predial unificado ocasionado por la ausencia de regulación de gradualidad en los procesos de actualización catastral en Cartagena D.T. y C. *Revista Saber, Ciencia y Libertad, 10*(2), 147–158.

McInnes, L., Healy, J., & Astels, S. (2017). HDBSCAN: Hierarchical density-based clustering. *Journal of Open Source Software, 2*(11), 205. https://doi.org/10.21105/joss.00205.

OECD. (2019). *OECD multi-level governance studies: Asymmetric decentralisation: Policy implications in Colombia*. https://doi.org/20.500.12592/hg7mv0.

Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications, 42*(6), 2928–2934. https://doi.org/10.1016/j.eswa.2014.11.040.

Peterson, S., & Flanagan, A. (2009). Neural Network Hedonic Pricing Models in Mass Real Estate Appraisal. *Journal of Real Estate Research*, *31*(2), 147–164. https://doi.org/10.1080/10835547.2009.12091245.

Sánchez Torres, F., & España, I. (2013). *Estructura, potencial y desafíos del impuesto predial en Colombia*. Universidad de los Andes, Facultad de Economía, CEDE. http://hdl.handle.net/1992/8431.

Sharma, H., Harsora, H., & Ogunleye, B. (2024). An optimal house price prediction algorithm: XGBoost. *Analytics, 3*, 30–45. https://doi.org/10.3390/analytics3010003.

Sharma, P. (2019, octubre 3). Box plot. *Data Science Unwind*. https://datascienceunwind.wordpress.com/2019/10/03/box-plot.

Wang, D., Li, V. J., & Yu, H. (2020). Mass appraisal modeling of real estate in urban centers by geographically and temporally weighted regression: A case study of Beijing's core area. *ISPRS International Journal of Geo-Information, 9*(5), Article 304. https://doi.org/10.3390/ijgi9050304.

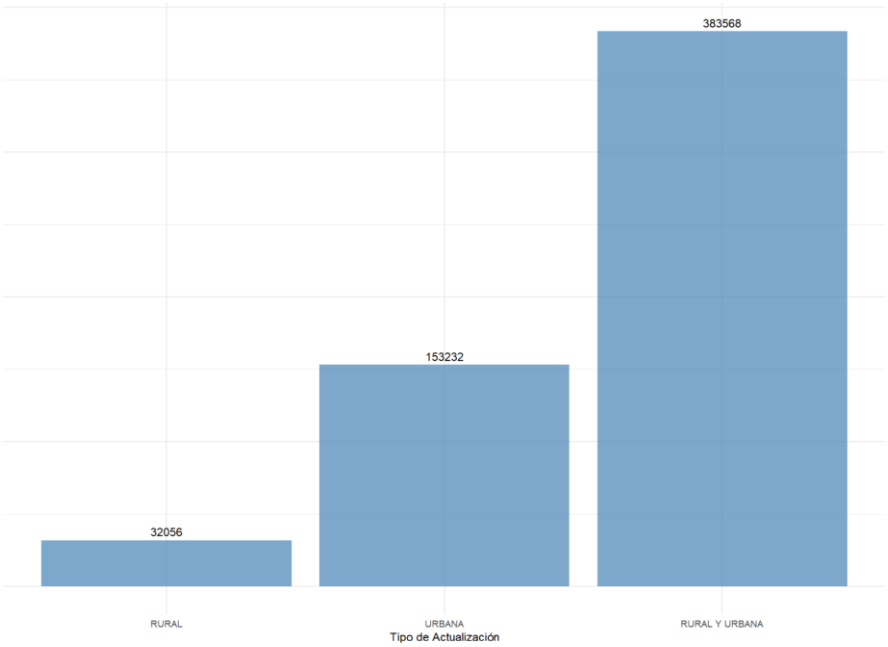
Whaley, D. L. (2005). *The interquartile range: Theory and estimation* [Tesis de maestría, East Tennessee State University]. Electronic Theses and Dissertations. https://dc.etsu.edu/etd/1030.

Zhang, R., Du, Q., Geng, J., Liu, B., & Huang, Y. (2015). An improved spatial error model for the mass appraisal of commercial real estate based on spatial analysis: Shenzhen as a case study. *Habitat International, 46*, 196–205. https://doi.org/10.1016/j.habitatint.2014.12.001.

Apéndices

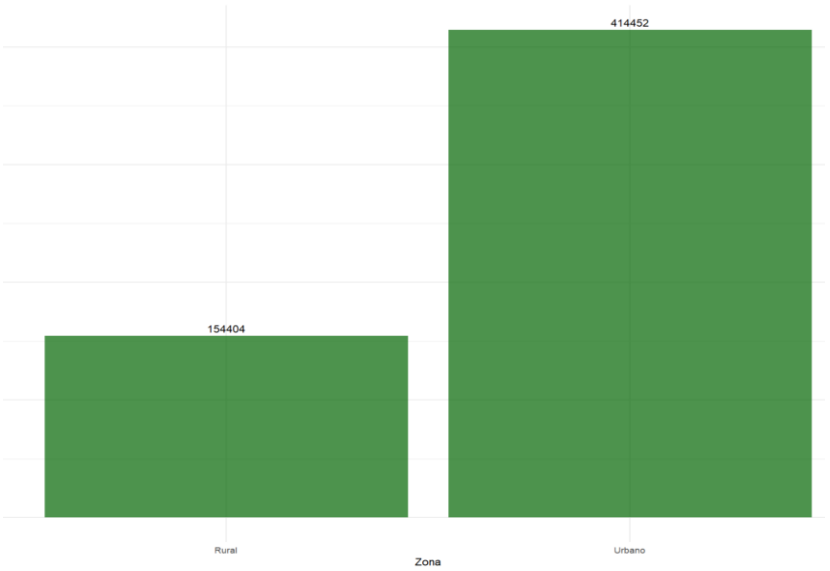
Apéndice A

Frecuencia de Predios Trazables por Tipo de Actualización



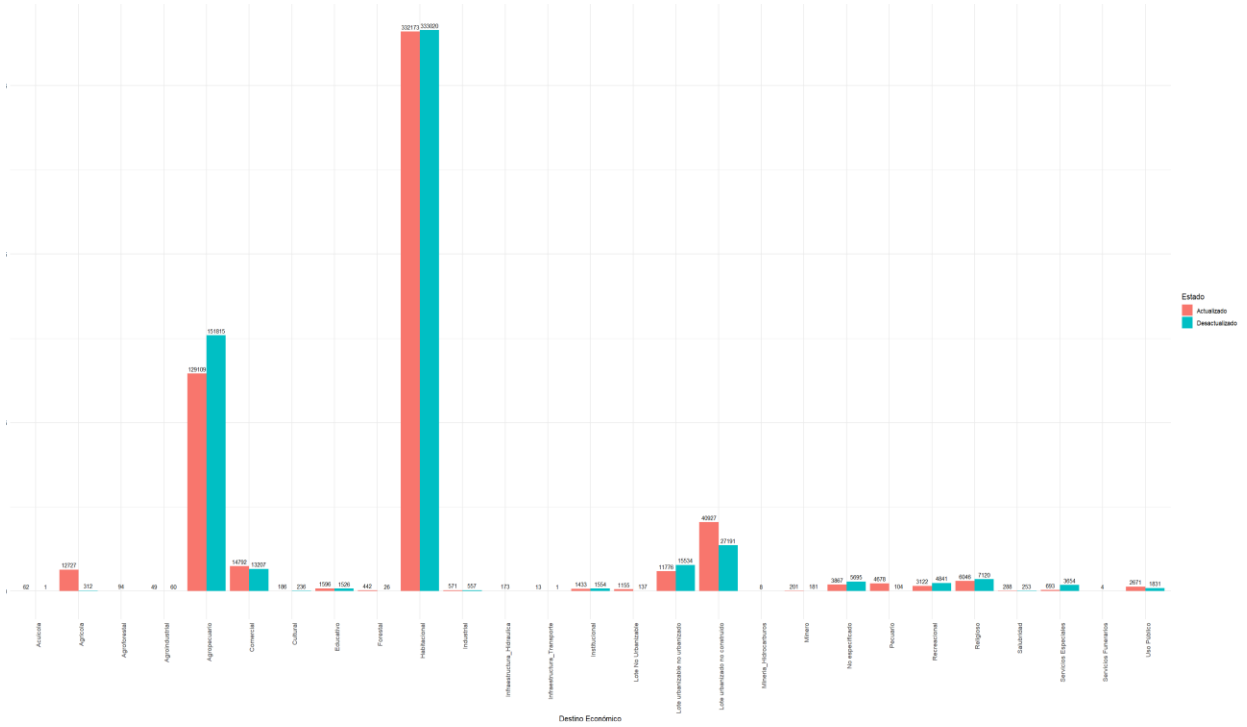
Apéndice B

Frecuencia de Predios Trazables por Zona



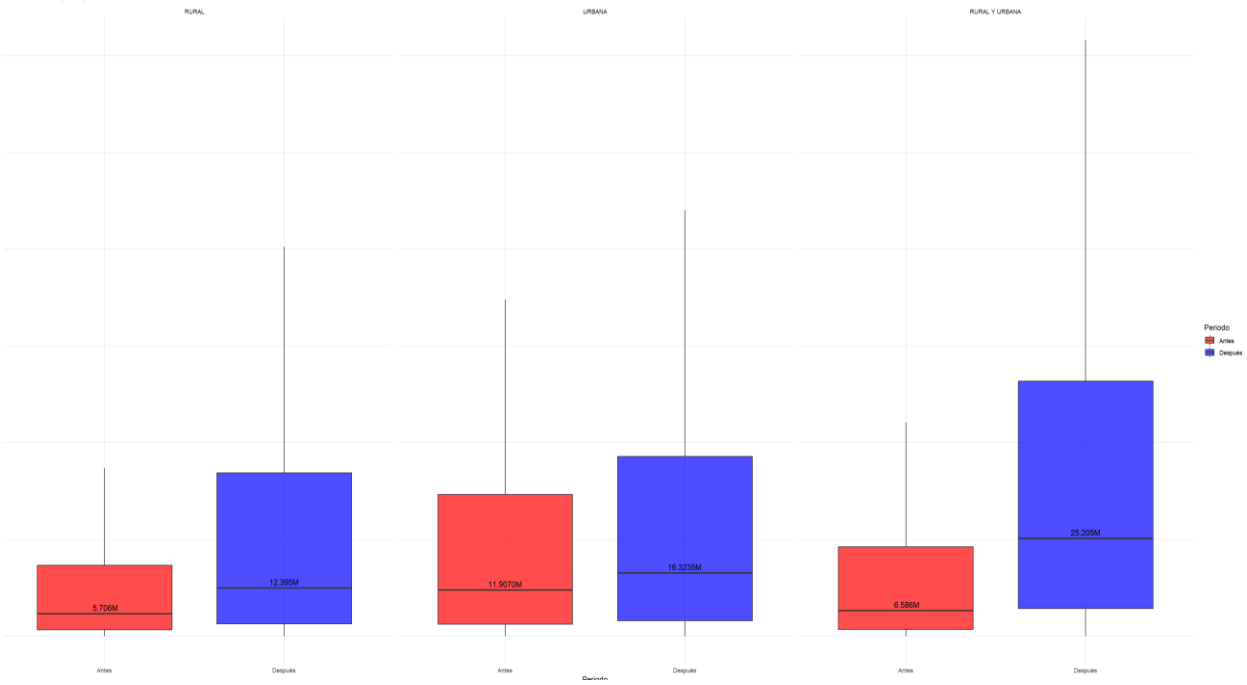
Apéndice C

Frecuencia de Predios Trazables por Destinación Económica



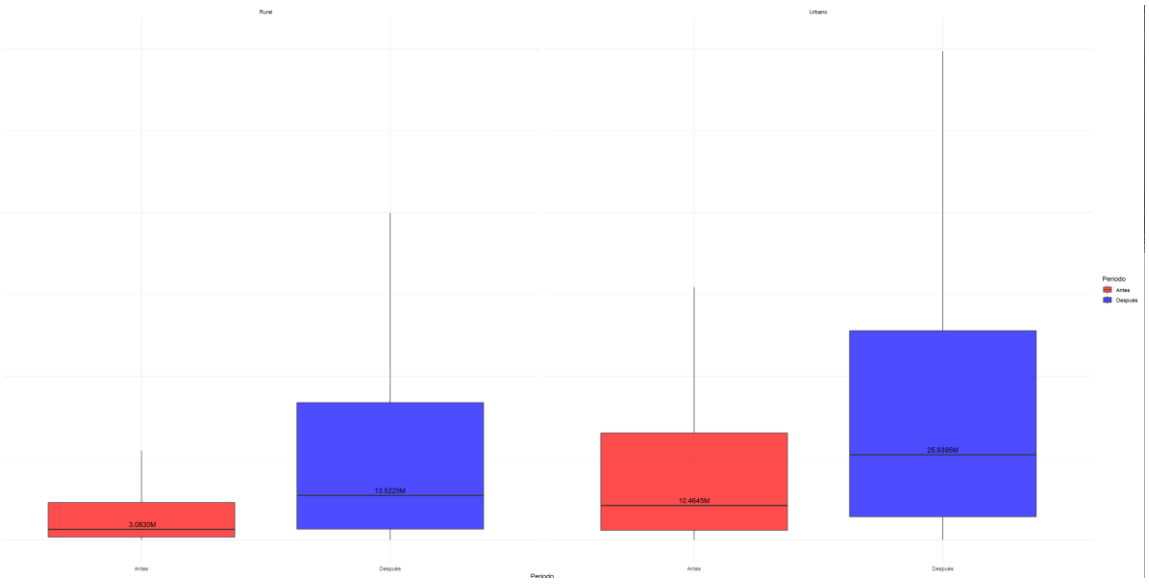
Apéndice D

Distribución del Avalúo Catastral Antes y Después de la Actualización Catastral Según el tipo de Actualización (Sin Datos Atípicos)



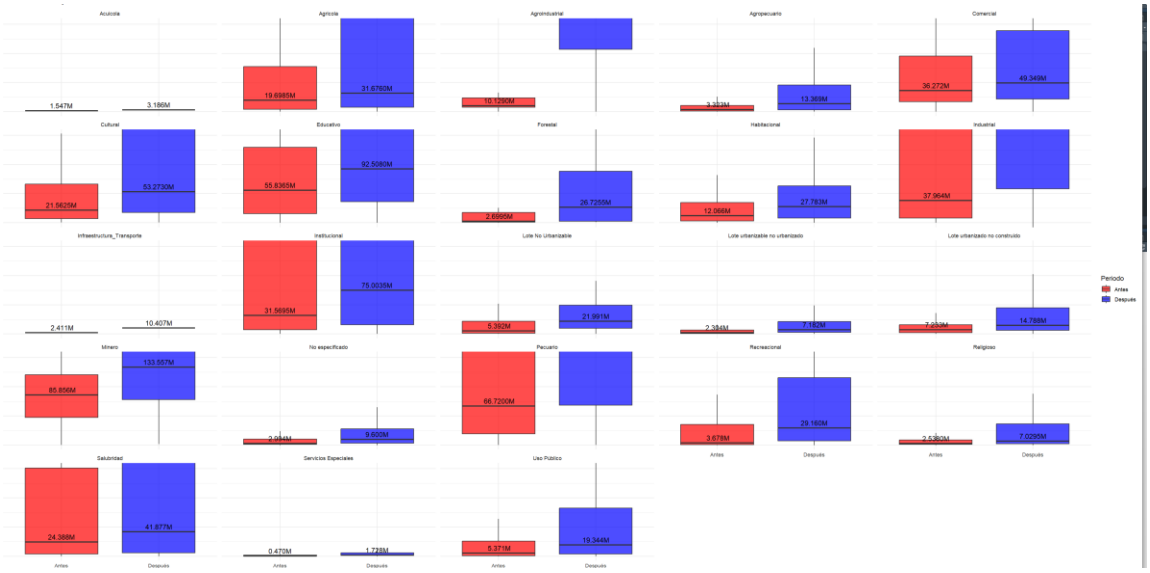
Apéndice E

Distribución del Avalúo Catastral Antes y Después de la Actualización Catastral Según la Zona (Sin Datos Atípicos)



Apéndice F

Distribución del Avalúo Catastral Antes y Después de la Actualización Catastral Según el Destino Económico (Sin Datos Atípicos)



Apéndice G

Estadísticas Descriptivas del Avalúo Catastral, el Área de Terreno y el Área Construida

Estadística	Área de Terreno (Antes)	Área de Terreno (Después)	Área Construida (Antes)	Área Construida (Después)	Avalúo (Antes)	Avalúo (Después)
Mínimo	-	-	-	-	-	-
1er Cuartil	78	78	-	-	1,916,000	5,700,000
Mediana	180	180	37	51	7,604,000	21,720,000
Media	143,700	105,100	69	87	29,690,000	66,450,000
3er Cuartil	1,450	1,432	90	108	26,560,000	58,700,000
Máximo	18,310,000,000	1,846,000,000	71,297	700,000	190,500,000,000	300,900,000,000

Apéndice H

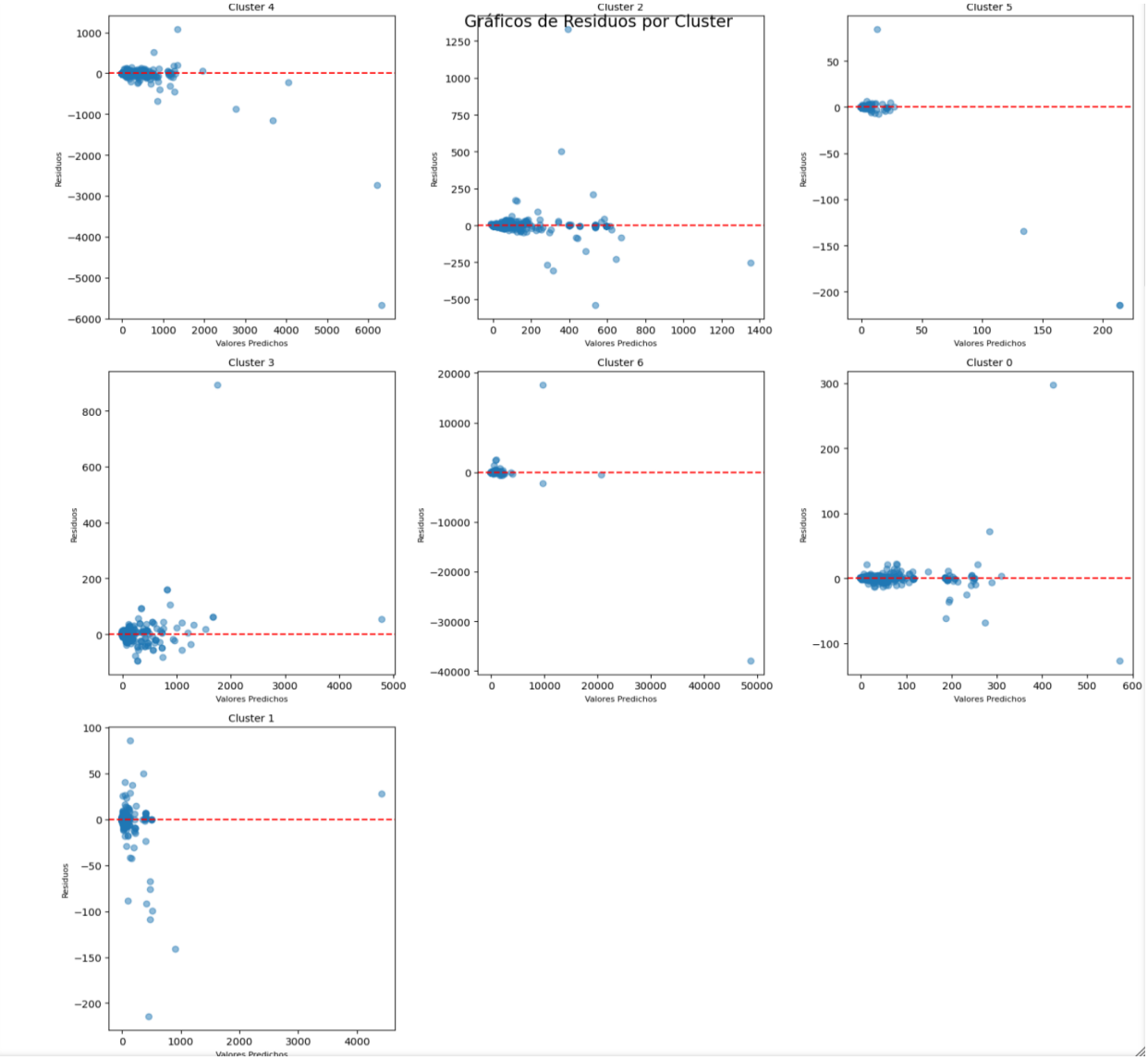
Correlación de Pearson del Área de Terreno, el Valor de m2 y el Área Construida Frente al Avalúo Catastral Según la Zona

Zona	Periodo	Variable	Correlación	Valor t	Grados de Libertad	p-Valor	Intervalo de Confianza (95%)
Rural	Desactualizado	Área Terreno vs Avalúo	0.576	276.75	154,402	< 2.2e-16	[0.572, 0.579]
Rural	Desactualizado	Área Construida vs Avalúo	0.397	170.17	154,402	< 2.2e-16	[0.393, 0.401]
Rural	Actualizado	Área Terreno vs Avalúo	0.014	5.48	154,402	4.30E-08	[0.009, 0.019]
Rural	Actualizado	Área Construida vs Avalúo	0.343	143.51	154,402	< 2.2e-16	[0.339, 0.347]
Urbano	Desactualizado	Área Terreno vs Avalúo	0.242	160.64	414,450	< 2.2e-16	[0.239, 0.245]

Zona	Periodo	Variable	Correlación	Valor t	Grados de Libertad	p-Valor	Intervalo de Confianza (95%)
Urbano	Desactualizado	Área Construida vs Avalúo	0.351	241.17	414,450	< 2.2e-16	[0.348, 0.353]
Urbano	Actualizado	Área Terreno vs Avalúo	0.051	32.94	414,450	< 2.2e-16	[0.048, 0.054]
Urbano	Actualizado	Área Construida vs Avalúo	0.745	717.95	414,450	< 2.2e-16	[0.743, 0.746]
Rural	Desactualizado	Valor por m² del Terreno vs Avalúo	0.073	27.982	146,112	< 2.2e-16	[0.0679, 0.0781]
Rural	Actualizado	Valor por m² del Terreno vs Avalúo	0.1	38.321	146,112	< 2.2e-16	[0.0947, 0.1048]
Urbano	Desactualizado	Valor por m² del Terreno vs Avalúo	0.035	21.392	373,806	< 2.2e-16	[0.0318, 0.0382]
Urbano	Actualizado	Valor por m² del Terreno vs Avalúo	0.625	489.93	373,806	< 2.2e-16	[0.6234, 0.6273]

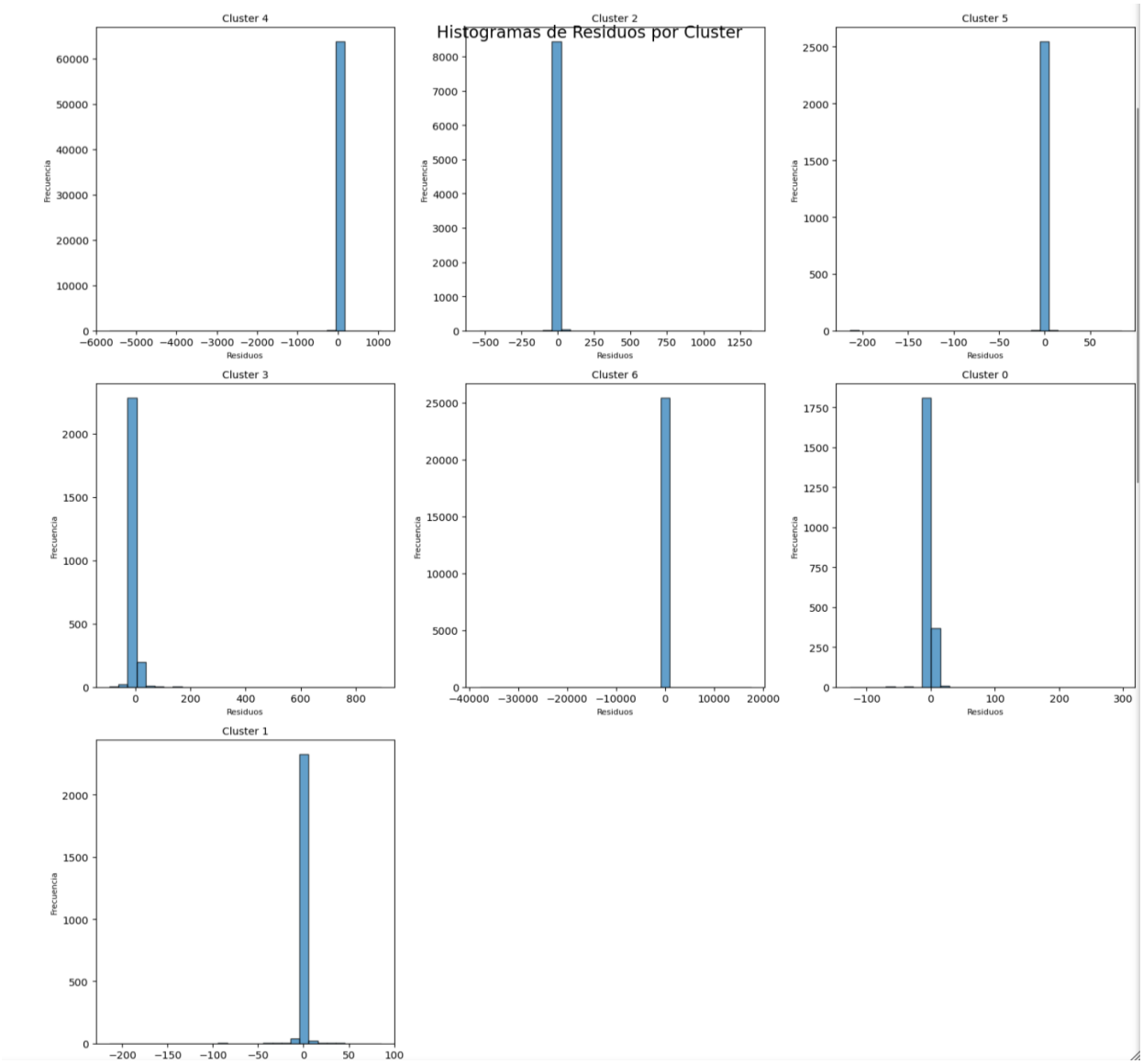
Apéndice I

Gráficos de Residuos del Modelo de Regresión XGBoost Para Cada Clúster



Apéndice J

Histograma de Residuos del Modelo de Regresión XGBoost para Cada Clúster



Apéndice K

Importancia de las variables para la estimación del modelo de regresión XGBoost para cada clúster