PREDICCIÓN DEL VALOR DEL SUELO EN BOGOTÁ EN ZONAS DE ADOPCIÓN DE PLANES PARCIALES USANDO TÉCNICAS DE APRENDIZAJE AUTOMÁTICO.

Angela Nieto Gómez CC: 52.810.695
Angela Rodríguez Narváez CC: 1.085.301.721
Angela Segura Herrera CC: 53.080.071

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director
David Arango Londoño
CC: 1.130.586.950

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, JUNIO 2026

FICHA RESUMEN
TRABAJO DE GRADO DE MAESTRÍA

TÍTULO: Predicción del valor del suelo en Bogotá en zonas de adopción de planes parciales usando técnicas de aprendizaje automático

1. ÉNFASIS: N/A

2. TIPO DE PROYECTO: Aplicado

3. ÁREA DE TRABAJO: Ordenamiento territorial

4. ESTUDIANTES: Angela Nieto Gómez, Angela Rodríguez Narváez y Angela Segura Herrera

5. CORREO ELECTRÓNICO: angienietog@javerianacali.edu.co , angelar92@javerianacali.edu.co y asegurah@javerianacali.edu.co

6. DIRECCIÓN Y TELÉFONO:

Av Cll 68 No 60 – 97 To.10 Ap.301 Bogotá D.C., cel. 3017493053.

Calle 9 # 40-04 Edificio Atures Pasto (Nariño), cel.3173795095.

Kr 64 A 22-41 To.4 Ap.1101 Bogotá D.C., cel.3177719708.

7. DIRECTOR: David Arango Londoño

8. VINCULACIÓN DEL DIRECTOR: Docente Catedrático

9. CORREO ELECTRÓNICO DEL DIRECTOR: david.arango@javerianacali.edu.co

10. CODIRECTOR (Si aplica): No aplica

11. GRUPO O EMPRESA QUE LO AVALA (Si aplica): No aplica

12. OTROS GRUPOS O EMPRESAS: No aplica

13. PALABRAS CLAVE (al menos 5): valor del suelo, planes parciales, aprendizaje automático, Machine Learning, ciencia de datos, predicción.

14. ODS QUE APLICA EL PROYECTO (Agenda 2030): Industria, innovación e infraestructura, Ciudades y comunidades sostenibles.

15. FECHA DE INICIO: 07/07/2025

RESUMEN

La falta de análisis sobre el cambio del valor del suelo tras la adopción de planes parciales ha favorecido procesos de especulación y ha generado desacuerdos entre los actores involucrados. Estas situaciones pueden derivar en modificaciones de los planes, así como retrasos en su formulación, adopción y en el reparto equitativo de cargas y beneficios.

El presente proyecto tuvo como objetivo analizar y predecir el comportamiento del valor del suelo en Bogotá en zonas de adopción de planes parciales, utilizando técnicas de aprendizaje automático, a partir de variables de accesibilidad, físicas, socioeconómicas, de uso del suelo, temporales y asociadas con la intervención de Planes Parciales. Para ello, se realizó un proceso de integración, depuración y transformación de información proveniente de diferentes fuentes geográficas y catastrales, disponibles en los portales de datos abiertos del Distrito Capital de Bogotá, permitiendo construir modelos predictivos orientados a identificar patrones asociados a la variación del valor del suelo.

Durante el estudio se realizó un análisis de impacto utilizando el método de Diferencias en Diferencias (DiD), con el fin de evaluar el comportamiento del valor del suelo antes y después de la adopción de planes parciales en Bogotá. Posteriormente, se evaluaron diferentes metodologías de aprendizaje automático, incluyendo Regresión Lineal Múltiple, Máquinas de Soporte Vectorial (SVM), XGBoost y Random Forest. Los resultados obtenidos evidenciaron que el modelo Random Forest presentó el mejor desempeño predictivo, alcanzando el mayor coeficiente de determinación y los menores errores de predicción entre los modelos evaluados, permitiendo representar de manera más adecuada la complejidad espacial y económica asociada al comportamiento del valor del suelo (metro cuadrado por manzana) en Bogotá.

Posteriormente, mediante un proceso de inferencia espacial, el modelo seleccionado fue aplicado sobre la totalidad de las manzanas de la ciudad, con el propósito de identificar patrones territoriales asociados al comportamiento del valor del suelo y ampliar la interpretación visual de los resultados más allá de las zonas inicialmente utilizadas para el entrenamiento del modelo, correspondientes a las manzanas donde se localizaron los planes parciales adoptados y sus áreas de influencia. Lo anterior facilitó el análisis exploratorio de tendencias espaciales relacionadas con la dinámica urbana y la adopción de estos instrumentos de planificación territorial, así como el desarrollo de una herramienta de visualización para comunicar de manera clara e interactiva los resultados obtenidos, constituyendo un insumo potencial para futuros procesos de planificación, análisis territorial y gestión del suelo en la ciudad de Bogotá.

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE ILUSTRACIONES

1. INTRODUCCIÓN

Los planes de ordenamiento territorial constituyen uno de los principales instrumentos de planificación urbana en Colombia, a través de ellos se orienta la organización de las ciudades, que se encuentran en constante cambio. Dentro de estos instrumentos, los planes parciales cumplen un papel fundamental al permitir el desarrollo y la gestión del suelo bajo principios definidos por la Ley 388 de 1997, entre ellos la distribución equitativa de cargas y beneficios entre los diferentes actores involucrados en los procesos de urbanización y renovación urbana.

En Bogotá, los planes parciales han adquirido una relevancia creciente dentro de los procesos de desarrollo urbano y consolidación territorial, especialmente en áreas sujetas a renovación, expansión o transformación urbana. No obstante, el comportamiento del valor del suelo asociado a la adopción de estos instrumentos aún presenta importantes dificultades de análisis e interpretación, debido a la complejidad de las dinámicas urbanas, económicas, normativas y espaciales que intervienen en su formación.

La falta de análisis desde la ciencia de datos relacionados con el comportamiento del valor del suelo tras la adopción de instrumentos de gestión como los planes parciales ha dificultado la comprensión de sus efectos reales sobre el mercado inmobiliario urbano. Esta situación favorece procesos de especulación y genera desacuerdos entre actores públicos y privados, derivando en modificaciones de los planes parciales y, en consecuencia, en retrasos en su formulación, adopción y ejecución. Adicionalmente, estas problemáticas afectan los procesos asociados al reparto equitativo de cargas y beneficios, uno de los principios fundamentales del ordenamiento territorial establecidos en la Ley 388 de 1997.

A pesar de que actualmente existe una creciente disponibilidad de información geográfica, catastral y normativa en los portales de datos abiertos del Distrito Capital, son limitados los estudios que integran enfoques de ciencia de datos, análisis espacial y aprendizaje automático para analizar el comportamiento del valor del suelo y los posibles efectos asociados a la adopción de planes parciales en Bogotá. Lo anterior refleja debilidades en el uso de herramientas analíticas y predictivas que permitan aprovechar el potencial de estas fuentes de información para apoyar procesos de planificación territorial y toma de decisiones.

Por lo anterior, el presente trabajo de grado tuvo como objetivo analizar y predecir el comportamiento del valor del suelo en Bogotá en las manzanas de adopción de planes parciales y sus áreas de influencia, utilizando técnicas de aprendizaje automático a partir de variables de accesibilidad, físicas, socioeconómicas, de uso del suelo, temporales y asociadas con la intervención de planes parciales. Para ello, se desarrolló un proceso de integración, depuración y transformación de información proveniente de diferentes fuentes geográficas y catastrales disponibles en los portales de datos abiertos del Distrito Capital de Bogotá, siguiendo como referencia metodológica la metodología CRISP-DM para proyectos de ciencia de datos.

Durante el desarrollo del estudio se realizó inicialmente un análisis de impacto utilizando el método de Diferencias en Diferencias (DiD), con el propósito de evaluar el comportamiento del valor del suelo antes y

después de la adopción de planes parciales en las zonas de estudio. Posteriormente, se implementaron diferentes metodologías de aprendizaje automático, incluyendo Regresión Lineal Múltiple, Máquinas de Soporte Vectorial (SVM), XGBoost y Random Forest, con el fin de comparar su desempeño predictivo sobre el valor del suelo (metro cuadrado por manzana) en la ciudad de Bogotá.

Finalmente, se desarrolló un proceso de inferencia espacial y se construyó una herramienta de visualización orientada a representar territorialmente los resultados obtenidos y facilitar el análisis exploratorio de patrones espaciales asociados al comportamiento del valor del suelo en Bogotá.

2. DEFINICIÓN DEL PROBLEMA

2.1. PLANTEAMIENTO DEL PROBLEMA

El valor catastral del suelo en Bogotá está influenciado por múltiples factores físicos, económicos, socioespaciales y normativos. Entre estos últimos, instrumentos de gestión del suelo como los planes parciales generan un cambio estructural en el territorio al renovar o promover nuevos desarrollos urbanos y reconfigurar el uso del suelo. Se considera que la normatividad alrededor del aprovechamiento del suelo genera valor cuando se materializa este aprovechamiento debido a que este instrumento direcciona el comportamiento del mercado[1]; no obstante, el análisis del cambio en el valor del suelo en el contexto de adopción de estos planes no ha sido un área ampliamente investigada desde una perspectiva de ciencia de datos que permita validar estas afirmaciones.

La Ley 388 de 1997[2] establece la distribución equitativa de las cargas y los beneficios como uno de los principios del ordenamiento del territorio, por lo que los planes parciales deben incluir los mecanismos para garantizar el reparto equitativo de las cargas y los beneficios vinculados al mejor aprovechamiento de los inmuebles, esto es, la forma en que se equilibran los costos de la infraestructura (cargas) con el valor generado por el plan (beneficios) para asegurar su viabilidad financiera. Una de las principales problemáticas alrededor de estos planes y del reparto equitativo de las cargas y los beneficios se encuentra en las variaciones del valor del suelo generadas por la especulación, debido a la falta de análisis sobre el cambio en su valor real.

En Bogotá, en el marco de los Decretos Distritales 190 de 2004[3] y 555 de 2021[4] de los Planes de Ordenamiento Territorial (POT) de Bogotá D.C, se instituyeron los Planes Parciales como instrumento para desarrollar y complementar las disposiciones del POT. Para su aprobación, los planes parciales deben surtir una fase de información pública, convocando a los propietarios y vecinos, para que éstos expresen sus recomendaciones y observaciones. Muchas situaciones relacionadas con esta fase han desencadenado modificaciones en diferentes planes parciales y por ende dilaciones en su formulación o adopción, lo cual ha estado directamente relacionado con la valoración del suelo, ya que esta impacta directamente la negociación de cargas y beneficios, la gestión de la plusvalía y la adquisición de tierras generando desacuerdo entre los actores involucrados.

En este contexto, a pesar de que la información sobre el valor catastral y del precio del suelo en Bogotá se ha recopilado históricamente por diversas fuentes, contando actualmente con una creciente disponibilidad de información catastral, normativa y geoespacial del suelo, no se encontraron análisis globales alrededor de los cambios en su valor derivados de todos los planes parciales adoptados, ni de la predicción del valor futuro en estas zonas en particular.

En este sentido, dada la información disponible, se reconoció la debilidad en el uso de la ciencia de datos para el desarrollo de modelos predictivos sobre el valor del suelo a partir de técnicas de aprendizaje automático, que permitieran identificar las variables explicativas de estas variaciones en las zonas donde se han adoptado planes parciales, situación que ha limitado la toma de decisiones institucionales y particulares de manera ágil y concertada. De continuar esta problemática no habrá mejoras en la concertación de los actores que intervienen en estas decisiones, dificultando y retrasando la adopción de estos instrumentos de gestión territorial y, a su vez, limitando el desarrollo de la ciudad conforme con lo planificado en su POT.

2.2. FORMULACIÓN DEL PROBLEMA

¿Cómo predecir el comportamiento del valor del suelo en Bogotá con técnicas de aprendizaje automático, en zonas donde se han adoptado planes parciales?

Preguntas de sistematización:

1. ¿Cómo pueden integrarse las fuentes de datos disponibles para construir una base de datos útil para el modelo predictivo?

2. ¿Cuáles técnicas de aprendizaje automático son más adecuadas para predecir el valor del suelo, teniendo en cuenta el efecto de adopción los planes parciales?

3. ¿Cuál modelo presenta el mejor desempeño o ajuste respecto al valor del suelo?

4. ¿Cómo visualizar un patrón espacial de variación sobre el valor del suelo tras la adopción de planes parciales?

3. OBJETIVOS DEL PROYECTO

3.1. OBJETIVO GENERAL

Construir un modelo que permita predecir el valor de suelo en Bogotá en zonas de adopción de planes parciales, utilizando técnicas de aprendizaje automático.

3.2. OBJETIVOS ESPECÍFICOS

1. Preparar los datos relevantes para la construcción del modelo.
2. Desarrollar modelos de aprendizaje automático para predecir el valor del suelo, teniendo en cuenta el

efecto de la adopción de los planes parciales.

3. Evaluar los modelos desarrollados y seleccionar el que mejor desempeño o ajuste tenga respecto al valor del suelo.

4. Elaborar una herramienta que permita la visualización de patrones espaciales de variación sobre el valor del suelo tras la adopción de planes parciales a partir del modelo desarrollado.

3.3. RESULTADOS ESPERADOS

- Capítulo de análisis de las variables estadísticamente significativas en la predicción del valor del suelo en las zonas de influencia de los planes parciales.

- Modelos predictivos del valor del suelo teniendo en cuenta el efecto de la adopción de los planes parciales utilizando técnicas de aprendizaje automático.

- Capítulo de análisis de los modelos desarrollados y selección del mejor modelo predictivo del valor del suelo.

- Herramienta de visualización resaltando los patrones espaciales identificados tras la predicción del valor del suelo en Bogotá a partir de la adopción de planes parciales para orientar la toma de decisiones.

4. MARCO DE REFERENCIA

4.1. MARCO TEÓRICO

4.1.1. Valor del Suelo en Bogotá

El valor del suelo urbano depende de múltiples factores, entre los que destacan sus características físicas, económicas y jurídicas. Comprender estos elementos es fundamental en el análisis del valor del suelo y en procesos como el catastro y la planificación urbana. Por lo anterior, es necesario abordar algunas definiciones para comprender cómo se establece dicho valor en contexto de los planes parciales, así como identificar la entidad responsable de su determinación en la ciudad de Bogotá.

La Ley 1753 de 2015 [5] ordena la implementación del catastro nacional con enfoque multipropósito y del Sistema Nacional de Gestión de Tierras; la cual es formalizada en la Ley 1955 de 2019[6] y reglamentada parcialmente con el Decreto 148 de 2020[7], que tiene como objetivo garantizar la calidad de la información catastral de los bienes inmuebles del país y presenta algunos conceptos fundamentales sobre el valor de suelo, avalúo catastral y comercial, catastro, entre otros (Título 2, Capítulo 1).

En Bogotá, la Unidad Administrativa Especial de Catastro Distrital (UAECD), es la entidad encargada de la actualización del censo catastral y su cartografía oficial (Acuerdo 004 de 2021 - Art. 4)[8]. Además, se encuentra facultada para prestar el servicio público de gestión y operación catastral multipropósito, contando con un observatorio inmobiliario catastral. En este contexto, la información sobre el valor del suelo en Bogotá, generada

y publicada por la UAECD, es un insumo fundamental para el desarrollo de este proyecto de grado.

4.1.2 Instrumentos de planificación urbana: Planes Parciales

La Ley 388 de 1997 (Art. 9)[2], establece como obligación de los municipios adoptar los Planes de Ordenamiento Territorial (POT). Uno de sus principios fundamentales es la distribución equitativa de las cargas y los beneficios. En su componente general se encuentran las Normas Urbanísticas (Art. 15) y dentro de ellas las **Normas urbanísticas estructurales** que son "*Las que establecen directrices para la formulación y adopción de planes parciales*."[2]. En este sentido, el artículo 19 de esta Ley, define los Planes Parciales de la siguiente manera:

> "***Planes parciales.*** *Los planes parciales son los instrumentos mediante los cuales se desarrollan y complementan las disposiciones de los planes de ordenamiento, para áreas determinadas del suelo urbano y para las áreas incluidas en el suelo de expansión urbana, además de las que deban desarrollarse mediante unidades de actuación urbanística, macroproyectos u otras operaciones urbanas especiales, de acuerdo con las autorizaciones emanadas de las normas urbanísticas generales, en los términos previstos en la presente Ley.* (…)"[2]

Dentro de los aspectos que se incluyen en el plan parcial se encuentran los estímulos a propietarios e inversionistas para facilitar los procesos de concertación y garantizar el reparto equitativo de las cargas y los beneficios vinculados al aprovechamiento de los inmuebles, por lo que se hace indispensable analizar el valor del suelo, facilitando la concertación y evitando demoras en la adopción de los planes.

Además, es importante mencionar que la Ley 388 de 1997, Capítulo IV, establece las clases de suelo que se pueden encontrar en un POT, donde se identifica el suelo urbano, rural y de expansión urbana; al interior de estas clases podrán establecerse las categorías de suburbano y de protección. El suelo urbano está constituido por las áreas del territorio destinadas por el POT a usos urbanos, que cuenten con infraestructura vial y redes primarias de energía, acueducto y alcantarillado, donde es viable la urbanización y edificación, y el suelo de expansión urbana corresponde a la porción del territorio destinada a la expansión urbana, que se habilitará para el uso urbano durante la vigencia del POT [2].

Desde la expedición de los Decretos 190 de 2004[3] y 555 de 2021[4] de la Alcaldía Mayor de Bogotá D.C., mediante el cual se adopta la revisión general del POT de Bogotá D.C, se han adoptado más de 50 planes parciales y, aunque han pasado casi 18 años desde la adopción del primer plan parcial [9], sigue siendo un reto analizar el valor del suelo en estos desarrollos.

La Secretaría Distrital de Planeación (SDP), encargada de la planeación y gestión del suelo en Bogotá, lidera el ordenamiento territorial mediante instrumentos y estrategias[10], entre los que se encuentran los planes parciales de desarrollo, siendo la entidad que brinda los insumos relacionados con ordenamiento territorial y planes parciales, necesarios en el desarrollo de este proyecto.

4.1.3 Datos abiertos y su rol en la ciencia de datos

Los datos abiertos son de gran importancia en la Ciencia de Datos, ya que la democratización de los datos facilita el acceso a la información y permite que se obtengan resultados confiables y verificables.

En Colombia, la Ley 1712 de 2014[11] sobre Transparencia y Acceso a la Información Pública Nacional, define los datos abiertos como información primaria disponible en formatos estándar e interoperables que permiten su acceso y reutilización. Esta normativa dispone que las entidades, tanto públicas como privadas con funciones públicas, deben poner dicha información a disposición de la ciudadanía de manera libre, para su reutilización bajo licencia abierta y sin restricciones legales para su aprovechamiento.

Ahora bien, dentro de un gran número de datos abiertos existentes, se encuentran los datos espaciales, que permiten observar gráficamente las dinámicas del territorio y asociar a ellas información alfanumérica de gran importancia para análisis urbanísticos, sociales, ambientales, entre otros.

En Bogotá, se cuenta con la Infraestructura de Datos Espaciales para el Distrito Capital (IDECA)[12], que dispone información geoespacial actualizada y de alta calidad, entre la cual se encuentran datos dispuestos por la Secretaría Distrital de Planeación (SDP) y la Unidad Administrativa Especial de Catastro Distrital (UAECD), orientados al análisis del valor del suelo y a la regulación del ordenamiento territorial.

4.1.4 Ciencia de Datos y Modelado Predictivo

La ciencia de datos es una disciplina que integra matemáticas, computación y conocimiento especializado en un área determinada. A través de la estadística aplicada, el procesamiento de datos y el Machine Learning, se construye conocimiento valioso a partir de los datos, con el objetivo de orientar la toma de decisiones estratégicas.

Machine Learning

El Machine Learning o aprendizaje automático, es una de las disciplinas de la Ciencia de Datos definida como un conjunto de métodos que pueden detectar automáticamente patrones en los datos y con el uso de estos, descubrir patrones de datos futuros, que apoyen la toma de decisiones en las organizaciones[13]. Los tipos de implementación de Machine Learning se clasifican en aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo.

Aprendizaje supervisado: en este tipo de problema, "*se enseña o entrena al algoritmo a partir de datos que ya vienen etiquetados con la respuesta correcta. Cuanto mayor es el conjunto de datos más el algoritmo puede aprender sobre el tema* (...)"[14]. Una vez termina el entrenamiento, se ingresan nuevos datos de entrada para predecir el resultado con la experiencia adquirida en el entrenamiento.

El aprendizaje supervisado, puede dividirse en problemas de clasificación o de regresión, el primero orientado principalmente a variables de categóricas y el segundo a variables numéricas. Los algoritmos de clasificación más comunes son: clasificadores lineales, máquinas soporte vectorial (SVM), árboles de decisión, redes neuronales**,** k vecinos más cercanos (K-Nearest Neighbors), Naive Bayes, random forest (bosque aleatorio) y los algoritmos de regresión son principalmente la lineal y la logística.

Aprendizaje no supervisado: en este tipo de problema, "*el algoritmo es entrenado usando un conjunto de datos que no tiene ninguna etiqueta; en este caso, nunca se le dice al algoritmo lo que representan los datos. La idea*

es que el algoritmo pueda encontrar por sí solo patrones que ayuden a entender el conjunto de datos. (...)"[14], se establecen patrones y a medida que se va repitiendo el proceso ya se esperan ciertos resultados de acuerdo con el comportamiento que ha tenido anteriormente. Los algoritmos de agrupamiento más comunes son los de segmentación o clustering, las reglas de asociación o los análisis de componentes principales (ACP).

Análisis predictivo

Según algunos autores, el análisis predictivo se puede definir como *"la tecnología que aprende de la experiencia para predecir el futuro comportamiento de individuos para tomar mejores decisiones"*[15].

El análisis predictivo puede basarse en métodos estadísticos tradicionales o en técnicas de Machine Learning. Mientras los primeros se enfocan en probar hipótesis, el Machine Learning construye conocimiento desde los datos mediante algoritmos o modelos matemáticos que identifican patrones y permiten hacer predicciones o integrarse en modelos más complejos.[16]. Al respecto,

> "*Para la creación del modelo predictivo se utilizan unidades de muestra disponibles con atributos conocidos y un comportamiento conocido, a este conjunto de datos se le denomina conjunto de entrenamiento. Por otro lado, se utilizará una serie de unidades de otra muestra con atributos similares, pero de las cuales no se conoce su comportamiento, a este conjunto de datos se le denomina conjunto de prueba*."[17].

En este tipo de análisis es necesario contar con datos de calidad y en una cantidad suficiente para establecer patrones de comportamiento e inducir al conocimiento con técnicas computacionales, esenciales al habilitar el análisis de datos con el que se pueden descubrir correlaciones entre variables y posteriormente identificar los patrones de comportamiento para crear un modelo predictivo, susceptible de aplicarse a individuos concretos.

Técnicas aplicables al análisis predictivo

Entre las diferentes técnicas de Machine Learning, hay algunas que son específicas de clasificación y otras de regresión, sin embargo, la mayoría de las técnicas funcionan con ambos. Entre las técnicas más utilizadas en el análisis predictivo se seleccionaron la regresión lineal, máquinas de soporte vectorial, random forest y XGBoost para el desarrollo de los objetivos 2 y 3, que se explican a continuación.

- Regresión Lineal

  Su propósito es estimar valores futuros o clasificar datos basados en patrones aprendidos. Puede utilizarse como descriptiva, pero su enfoque es predictivo. Estas se clasifican en: regresión lineal, logística y otras regresiones como Polinómica, Ridge y Lasso[16].

- Máquinas de soporte vectorial (SVM)

  Algoritmo de aprendizaje supervisado utilizado principalmente en tareas de clasificación. Su objetivo es encontrar el hiperplano óptimo que separe las clases de datos, maximizando la distancia entre ellas dentro de un espacio de N dimensiones[18]. Se entrenan con datos históricos para aprender una frontera óptima que separe distintas clases o prediga valores numéricos. Las SVM son principalmente modelos predictivos,

pero pueden ayudar a entender qué datos influyen más en la toma de decisiones.

Las SVM con enfoque de regresión (SVR), corresponden a una adaptación donde la variable a predecir es continua. En este caso se busca ajustar una función que deje la mayoría de los puntos dentro de una banda de tolerancia correspondiente a los vectores de soporte. Esta técnica tiene la ventaja de funcionar bien con datos no lineales, observaciones lejanas y alta dimensionalidad, no obstante, es costoso computacionalmente y su interpretación no es sencilla.

- Random Forest (Bosque aleatorio)

Están formados por un conjunto de árboles de decisión individuales (definidos más adelante), cada uno entrenado con una muestra ligeramente distinta de los datos de entrenamiento generada mediante bootstrapping. La predicción de una nueva observación se obtiene agregando y promediando los resultados de las predicciones de todos los árboles individuales que forman el modelo.

Este modelo tiene diferentes ventajas entre las que se encuentran su adaptabilidad para manejar predictores numéricos y categóricos sin necesidad de crear variables dummy o one-hot encoding, además no necesitan el cumplimiento de distribuciones dado que es un método no paramétrico, no necesitan estandarización, no se influencian fácilmente por outliers, permiten obtener una predicción aún en ausencia del valor de un predictor (pero con menor precisión), seleccionan automáticamente los predictores e identifican de manera eficiente los más importantes. No obstante, entre sus limitaciones se encuentran su sensibilidad frente a datasets desbalanceados, su tendencia a favorecer a los predictores continuos y tienen dificultad para extrapolar por fuera del rango de los predictores observados en el entrenamiento [19].

- XGBoost (Extreme Gradient Boosting)

Es un algoritmo de aprendizaje automático basado en árboles de decisión que mejora progresivamente el modelo al corregir los errores en cada iteración. Para ello, construye modelos de forma secuencial, donde cada nuevo árbol se enfoca en las observaciones con mayor error. Además, incorpora técnicas de regularización, paralelización y muestreo aleatorio, lo que permite mejorar la precisión, reducir el sobreajuste y aumentar la eficiencia computacional.

Al utilizar árboles de decisión como base, el modelo divide recursivamente los datos con el fin de minimizar una función de pérdida y capturar relaciones complejas entre variables. Gracias a su capacidad para modelar relaciones no lineales y manejar grandes volúmenes de datos, es ampliamente utilizado en problemas de predicción, como la estimación del valor del suelo[20].

Adicionalmente, existen otras técnicas de Machine Learning para el análisis predictivo, como las que se presentan en la Tabla I:

Tabla I.
Técnicas aplicables al análisis predictivo

Técnica de ML	Descripción
Árbol de Decisión	Es un modelo predictivo que se conforma por reglas binarias (si/no), a partir de las cuales se consigue repartir las observaciones en función de sus atributos y predecir así el valor de la variable respuesta. Toma la estructura de un árbol para tomar decisiones y se utiliza tanto para clasificación como regresión [19].
Redes Neuronales Artificiales	Se inspiran en el cerebro humano y son entrenadas para aprender patrones complejos y realizar predicciones. Se utilizan principalmente cuando no se conoce la naturaleza exacta de la relación entre los valores de entrada y de salida[21]. Algunos usos son procesamiento de imágenes, lenguaje natural y juegos.
K- Vecinos más cercanos (K - Nearest Neighbors) (KNN):	Es una de las técnicas de clasificación y regresión más sencillas utilizadas en Machine Learning, basándose en la similitud entre datos, sin embargo, es más utilizada en clasificación, partiendo del supuesto que se pueden encontrar puntos similares el uno cerca del otro. [22]

4.1.5 Metodología de proyecto de ciencia de datos. (CRISP-DM)

La metodología CRISP-DM se desarrolla en 6 etapas iterativas que todo el tiempo se soportan en los objetivos del negocio y se aborda en las fases que se describen en la Tabla II y se presentan gráficamente en la Fig. 1:

Tabla II.
Aporte de trabajos previos que abordan el mismo problema o uno similar

Técnica de ML	Descripción
Entendimiento del Negocio	Según IBM[23] en esta etapa se busca traducir las necesidades del negocio en un problema concreto de minería de datos y desarrollar un plan preliminar para alcanzar dichos objetivos. Esta etapa aborda las siguientes actividades: determinar los objetivos de negocio, evaluar el contexto, determinar los objetivos de minería de datos y generar el plan de proyecto.
Comprensión de los Datos	Consiste en recolectar, caracterizar, describir y explorar la información necesaria, asegurando su relevancia y calidad para el análisis.
Preparación de los Datos	Esta etapa es clave para garantizar la calidad y coherencia de la información antes de aplicar modelos predictivos. Según IBM[23], hasta el 70% del tiempo total de un proyecto de ciencia de datos se dedica a esta fase, lo que resalta su importancia para el éxito del proyecto. Las actividades principales son: selección, limpieza, construcción, integración y formateo de datos.
Modelado	En esta etapa se busca determinar el modelo más adecuado para resolver la hipótesis u objetivo del proyecto. Para esto, se analizan aspectos como la calidad de los datos, las técnicas de modelado, la relación entre variables, el impacto costo-beneficio y el flujo de trabajo. Debido a que pueden existir múltiples modelos viables, es necesario seleccionar y evaluar los más apropiados, eligiendo el que ofrezca mejores resultados en términos de precisión y otros aspectos influyentes.
Evaluación	Se orienta a comprobar que el modelo cumpla con las expectativas del negocio mediante métricas de calidad y desempeño. Sus resultados pueden implicar ajustes o un retorno a fases anteriores si estos no son satisfactorios.[23]

Despliegue	Etapa en la que el modelo se pasa a ambiente productivo y se realiza su implementación para que los usuarios finales puedan utilizarlo. En general, la etapa de despliegue incluye dos tipos de actividades[23]: planificación y control del despliegue de los resultados y finalización de tareas de presentación como la producción de un informe final y la revisión de un proyecto.

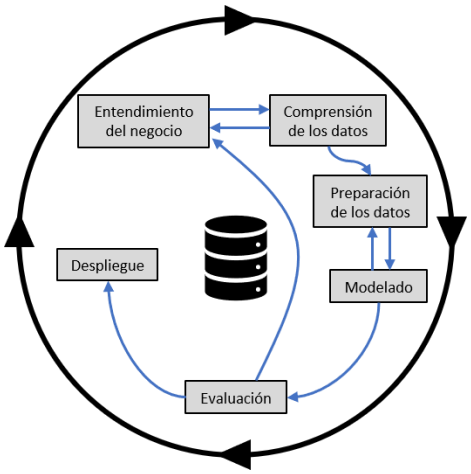


Fig. 1. Etapas de la metodología CRISP-DM.
Adaptado de [24]

4.1.6 Evaluación de impacto

La evaluación de impacto proporciona información sobre los efectos producidos por una intervención, que puede ser un proyecto pequeño, un programa amplio, un conjunto de actividades o una política, su objetivo es determinar y cuantificar cómo una intervención afecta los resultados de interés. Los efectos esperados pueden ser positivos o negativos, intencionados o no intencionados, directos o indirectos.

Dentro de las evaluaciones es clave identificar el grupo de tratamiento o de participantes en el programa o afectados por la intervención y el grupo de control o comparación que hace referencia a los que no participaron o no se afectaron por la intervención pero que tienen características estadísticamente idénticas al grupo de tratamiento. Lo anterior porque si ambos grupos son idénticos, la diferencia en resultados se explica por el efecto de la intervención [25].

Para determinar los efectos causales (relaciones causa-efecto) de una intervención existen diferentes métodos, entre ellos, algunos diseñados para datos de panel (como es el caso de este proyecto), que analizan los efectos causales teniendo en cuenta la endogeneidad de los datos observacionales, como el método de Diferencias en Diferencias, el enfoque clásico de valor añadido [26] y el Diseño de Regresión Discontinua [27].

Diferencias en Diferencias (DiD)

DiD es uno de los métodos más utilizados en estudios de evaluación de impacto para estimar el efecto de una política o proyecto (tratamiento) implementado de forma no aleatoria, comparando los resultados en el grupo de tratamiento con un grupo de control, con datos después de la implementación del proyecto. Busca evaluar

el impacto después de controlar los factores que difieren entre los dos grupos. "*Si la diferencia después-antes en el grupo de control se resta de la misma diferencia en el grupo de tratamiento, se logran dos cosas. Primero, si otros cambios que ocurren con el tiempo también están presentes en el grupo de control, entonces estos factores se controlan cuando la diferencia después-antes del grupo de control se resta de la estimación del impacto. Segundo, si hay características importantes que son determinantes de los resultados y que difieren entre los grupos de tratamiento y control, entonces, siempre que estas diferencias entre tratamiento y control sean constantes en el tiempo, su influencia se elimina al estudiar los cambios a lo largo del tiempo. De manera importante, este último punto también se aplica a las diferencias entre tratamiento y control en características inescrutables e invariantes en el tiempo (ya que se restan)*."[28]

La estimación de modelos basados en DiD depende del supuesto de tendencias paralelas, que establece que el grupo de tratamiento, en ausencia del proyecto, habría seguido la misma tendencia temporal que el grupo de control (para la variable de resultado de interés). Factores observables y no observables pueden causar que el nivel de la variable de resultado difiera entre tratamiento y control, pero esta diferencia (en ausencia del proyecto en el grupo de tratamiento) debe ser constante a lo largo del tiempo. Dado que el supuesto es fundamentalmente inverificable, se puede respaldar utilizando varios períodos de datos previos al proyecto, mostrando que los grupos de tratamiento y control exhiben un patrón similar en los períodos previos al proyecto. Si este es el caso, la conclusión de que el impacto estimado proviene del tratamiento en sí, y no de una combinación de otras fuentes (incluyendo aquellas que causan las diferentes tendencias previas), se vuelve más creíble. [28]

Enfoque clásico de valor añadido

Este enfoque busca manejar la heterogeneidad no observada de los individuos o las diferencias entre las unidades de estudio que no han sido medidas o registradas en los datos, pero que influyen en los resultados, centrándose en los resultados de las unidades de estudio a lo largo del tiempo, mediante el control de los resultados previos de la misma unidad al estimar el efecto de algún insumo o intervención en los resultados actuales. Si sucede que el resultado observado es el mismo para una unidad en dos momentos diferentes, este enfoque estima el efecto de la intervención específica condicionado al resultado anterior. Este tipo de modelos utilizan información transversal de las mismas unidades en diferentes momentos de tiempo para eliminar los efectos fijos o efectos de cualquier factor invariante en el tiempo [26]

Diseño de Regresión Discontinua (RDD)

El RDD se basa en encontrar umbrales adecuados, de modo que los grupos que se encuentren por encima (o por debajo) dada su condición de tratamiento sean equivalentes en todos los demás aspectos a los grupos que se encuentren por debajo (o por encima).[29]

En el RDD puede explorarse el uso de límites geográficos como discontinuidades de regresión, dónde la variable de asignación es la distancia a un límite, comparando individuos a ambos lados de este. En este diseño de discontinuidad geográfica se comparan dos áreas adyacentes correspondientes a una de control y una de tratamiento, dónde la probabilidad de recibir el tratamiento salta de forma discontinua a lo largo de la frontera

que separa ambas zonas. En este caso a diferencia del RDD estándar, el efecto del tratamiento se define como una curva en un plano en lugar de un parámetro único[27].

Algunas limitaciones de esta aplicación son que el uso de una medida de distancia como única variable en la definición de frontera puede ser insuficiente ya que se ignora la naturaleza espacial de los datos, en la que dos individuos se pueden encontrar a la misma distancia del límite y a su vez encontrarse en puntos geográficos distantes dadas las condiciones de accesibilidad. Otra limitante corresponde a la necesidad de conocer detalladamente la zona geográfica de aplicación para evaluar la consistencia del límite dados los factores socioeconómicos preexistentes [27].

4.2. ANTECEDENTES

4.2.1 Trabajos previos que abordan el mismo problema o uno similar

Tabla III.
Aporte de trabajos previos que abordan el mismo problema o uno similar

Estudio	Aporte
Método automático para la predicción del avalúo comercial de un inmueble en la ciudad de Bogotá **Autores:** Viviane Téllez, Duyber Martínez[30]	Realizaron la predicción de avalúos comerciales de viviendas en Bogotá a través de técnicas de Machine Learning para optimizar la precisión, el tiempo y el costo de los procesos de peritaje inmobiliario, para lo cual se siguieron pasos similares a los planteados en este proyecto, tales como la construcción del dataset, análisis, selección del modelo y de medidas de desempeño. Realizaron modelos de regresión lineal, árboles de decisión, random forest y redes neuronales profundas para cada estrato socioeconómico de las viviendas y encontraron que en los estratos 2, 3 y 4 los modelos de árboles de decisión y random forest tenían mejor resultado en las medidas de desempeño R^2, RMSE y CV, mientras que en los estratos 5 y 6 los modelos que mejor ajustaban los precios de vivienda eran la regresión lineal y random forest.
Modelo de predicción de precios de viviendas en el municipio de Rionegro para apoyar la toma de decisiones de compra y venta de propiedad raíz **Autora:** Yuri Grajales[31]	Desarrolló un modelo para predecir los precios de vivienda en Rionegro (Antioquia) utilizando técnicas de Machine Learning, siguiendo la metodología CRISP-DM. Utilizó modelos de regresión lineal, árboles de decisión de regresión, máquinas de soporte vectorial, random forest y Gradient Boosting Machine (GBM) y los evaluó con las métricas de Error Porcentual Absoluto Medio y Coeficiente de determinación R2. Se encontró que los modelos Random Forest y GBM tuvieron el mejor desempeño en las métricas de evaluación y estabilidad en el proceso de validación cruzada con 10 conjuntos de prueba. Concluyó que las variables más relevantes para determinar el precio de la vivienda son: área total, área privada, área construida y estrato.
Norma y precio del suelo urbano en la ciudad de Bogotá	Analizaron si el comportamiento del precio del suelo en Bogotá respondía a las normas urbanísticas existentes o a la dinámica de oferta

Autores: Alex Araque, Lina Cantor, Paloma Moreno[1]	y demanda. Estimaron un modelo de Mínimos Cuadrados Ordinarios del logaritmo del metro cuadrado del suelo rural en función del área del terreno, la distancia del perímetro urbano, área construida y usos del suelo. Encontraron que existe una relación inversa entre la distancia al perímetro urbano y el tamaño de los predios con relación al precio del metro cuadrado del suelo rural, así mismo, identificaron que el uso residencial tiene un mayor precio en comparación con los demás usos. Los autores concluyeron que la norma urbanística no impide que se generen expectativas de precio del suelo rural por su posible incorporación al suelo urbano, porque esto responde a las dinámicas de crecimiento poblacional que generaron una expansión desordenada en Usaquén, Suba, Engativá, Fontibón y Usme, así como a la presión en el suelo rural de la frontera de la ciudad por la demanda de suelos de uso residencial.
Análisis del valor del suelo de predios inmersos en plan parcial – caso de estudio Plan Parcial Tres Quebradas – Operación Nuevo Usme en Bogotá **Autores:** Adriana Granados, Gonzalo Higuera. [32]	Realizaron un análisis del cambio en el valor del suelo derivado de la delimitación del Plan Parcial Tres Quebradas en la localidad de Usme de la ciudad de Bogotá, particularmente en los predios del barrio Puerta al Llano Rural y un análisis de las condiciones normativas, catastrales y socioeconómicas del Plan de Ordenamiento Zonal de Usme utilizando información del catastro distrital y de los valores históricos de los predios involucrados en este Plan Parcial, del barrio analizado, en diferentes años entre 2003 y 2020. Pese a que no lograron el objetivo de comparar el valor histórico de los predios con las ofertas comerciales y aunque este estudio no utiliza herramientas de ciencia de datos para su análisis, es importante porque reconoce la importancia de la identificación de los valores de mercado del suelo antes del anuncio de un plan parcial porque permite a las instituciones involucradas reconocer el valor real del suelo evitando la planificación territorial basada en la especulación; además es útil para que los profesionales avaluadores usen esta información en sus estudios de valor del suelo y establezcan límites razonables e identifiquen las variables que determinan el valor de las propiedades.
Análisis del valor del suelo del Plan Parcial Algarra en el municipio de Zipaquirá **Autor:** Nelson León[33]	Realizó un análisis del valor del suelo en el Plan Parcial Algarra del municipio de Zipaquirá para 5 Unidades de Actuación Urbanística (UAU) que comprendían áreas de actividad residencial, institucional y de comercio. A partir de la comparación del valor del área bruta del terreno antes y después de la modificación del POT, encontró que los valores de las 5 UAU incrementaron su valor entre un 10,6% y 29,6%. Si bien este estudio, no utiliza métodos en ciencia de datos para analizar el cambio en el valor del suelo en contexto de la implementación del Plan Parcial Algarra, encuentra evidencias de que existe un cambio en el valor del suelo explicado por este instrumento

	de gestión.
Evaluar el reajuste de tierras en Colombia: la distribución equitativa entre actores. **Autor:** Ricardo Daza[34]	Realizó un análisis de la diferencia en el valor de la tierra con la técnica de valor residual, que analiza el valor inicial y final después de contabilizar las ganancias y costos relacionados (cargas y beneficios), para analizar la capacidad de los actores de compartir el incremento en el valor del suelo dentro de las áreas donde se han adoptado planes parciales. Su análisis indica que el valor del suelo se incrementó hasta entre 61% y 1479%, comparando con su valor inicial. En este estudio también se evidencia la valorización resultante de la adopción de este instrumento de gestión del suelo.
Estudio de valor para un área del suelo de expansión contiguo a la comuna N. 7 en el municipio de Villavicencio **Autor:** Fabián Gómez[35]	Propuso determinar el valor actual del suelo en un área de expansión a partir de información primaria y un estudio de mercado. Se identificó que entre las variables que influyen en el precio de los predios, la más importante es la cercanía a un eje vial principal, porque concentra actividades comerciales y de uso residencial. Otra variable es la extensión de los terrenos; también se identificó que los terrenos ubicados en condominios tienen un mayor valor. Este estudio es relevante en el sentido de identificar variables que influyen en el precio del suelo y a su vez orientar este proyecto teniendo en cuenta estas variables en las fuentes de datos a utilizar.

Algunas de las investigaciones presentadas en la Tabla III no utilizan modelos de Machine Learning, sin embargo, aportan al propósito de este proyecto porque permiten identificar que existe un cambio en el valor del suelo derivado de la adopción de planes parciales, el cual es generado, entre otras variables, por las expectativas de la demanda y el uso potencial del suelo (especulación). Lo anterior es coherente con el planteamiento del problema presentado, donde se identificó que la raíz de los problemas en la adopción de planes parciales se deriva del desconocimiento del valor real del suelo.

Las investigaciones que si utilizan técnicas de Machine Learning, aportan a esta investigación para identificar cuáles modelos pueden ser útiles para la predicción con información de variables similares a las disponibles en las bases de datos consideradas en este proyecto.

4.2.2 Trabajos previos que abordan problemas diferentes al de interés con técnicas que pueden aplicar al problema actual

Tabla IV.
Aporte de trabajos previos que abordan técnicas que pueden aplicar al problema actual

Estudio	Aporte
Modelo econométrico para la estimación del valor del suelo rural en los municipios de la	Se propuso un modelo econométrico para la estimación del valor por hectárea del suelo rural en determinados municipios de Cundinamarca,

jurisdicción de la agencia catastral de Cundinamarca. **Autora:** Isabella González Méndez [36]	donde utilizó variables como el valor del suelo (variable respuesta), uso, disponibilidad de agua, número de hectáreas, coordenadas geográficas, grado de pendientes, valor potencial (índice numérico), acidez, erosión hídrica y potencial. Se seleccionó el modelo de KNN (K-Nearest Neighbors) por tener la mejor matriz de pesos relativos para el modelo espacial y realizó una transformación de Box-Cox del valor del suelo porque la variable respuesta no seguía una distribución normal. Aplicó un modelo clásico de regresión lineal donde el logaritmo natural del valor del suelo está en función de las variables más relevantes seleccionadas: número de hectáreas, pendiente y valor potencial.
Análisis del mercado de tierras con técnicas complejas **Autor:** Luis Vilches Blázquez[37]	Realizó un análisis exploratorio de las características del mercado de tierras de los municipios de Ibagué y Ortega en el Tolima, siguiendo los pasos de pre-procesamiento de datos, modelación y visualización de resultados, similar a la metodología CRISP-DM, aunque no se menciona de manera explícita. Utilizó modelos de clasificación (árbol de decisión y Naïve Bayes), regresión (lineal y no lineal) y clustering (K-means) y realizó una validación cruzada y hold-out para evitar elegir un modelo sobreajustado. Se analizaron las variables de categoría predominante del tamaño de la propiedad, piso térmico, clase agrológica, categoría predominante tradición, pendiente y destino económico. Se encontró que los modelos seleccionados eran adecuados para tener una aproximación y entendimiento de las relaciones y patrones implícitos del mercado de tierras rurales en Colombia.
Determinación de valores de terreno de suelo urbano en Bogotá de acuerdo con las ofertas inmobiliarias de predios NPH **Autor:** William Alexander Melo Cerón[38]	Analizó las variables que mejor describen el comportamiento del valor del suelo urbano en Bogotá con un modelo de regresión lineal múltiple y presentó una distribución espacial de los valores predichos por el modelo y de sus variaciones en el marco geográfico de la ciudad. Las variables explicativas del modelo, seleccionadas después del análisis de autocorrelación, fueron: valor de referencia, área de terreno, área construida, estrato, actividad económica, tipo de vía de acceso, ubicación del predio (norte-sur) y topografía. Probó diferentes modelos de regresión lineal y de árbol de decisión y seleccionó el primero porque explica la variabilidad de los datos en un 65% con un menor valor en los errores. No obstante, el estudio concluyó que el modelo no tiene la eficiencia suficiente para determinar el valor del suelo urbano, pero se puede usar como marco de referencia de un rango de precios para procesos de actualización catastral. Con ArGis 10.6.1 generó salidas gráficas de la distribución espacial de los valores unitarios reales y predichos del suelo concluyendo que el modelo atenúa la variabilidad real del suelo y que se presentan variaciones altas positivas en estratos altos y variaciones altas negativas en estratos 1 y 2.
Evaluación del impacto de la Ley 388 de 1997	Realizó una evaluación de impacto con un modelo de precios hedónicos

y sus instrumentos sobre el mercado del suelo en las principales ciudades del país **Autor:** Departamento Nacional de Planeación (DNP) [39]	para estimar el efecto marginal de los instrumentos de la Ley 388 de 1997 sobre el mercado del suelo para 3 ciudades de Colombia. En Bogotá el mejor modelo obtenido utiliza variables dummy para el estrato, localidad e interés cultural del inmueble con distancias a centros comerciales y otros equipamientos; en Medellín, las variables con mejor poder explicativo fueron la densidad máxima establecida para el polígono normativo, altura máxima permitida y dotación de metros cuadrados por habitante para uso residencial. En Cartagena, el modelo se especificó de manera distinta porque su fuente de datos fue catastral y se encontró que ninguna de las normas de uso establecidas tiene peso significativo sobre el precio del metro cuadrado.
Propuesta para la valuación masiva del suelo urbano. aplicación espacial del algoritmo Quantile Regression Forest **Autores:** Rocío Cerino, Juan Pablo Carranza, Mario Piumetto, María Bullano, Vania Caffarati, Federico Monzani [40]	Realizaron una valuación masiva del suelo en Córdoba (Argentina), con la técnica de aprendizaje automático Quantile Regression Forest para predecir el valor de la tierra. Este algoritmo permite predecir la mediana, siendo más consistente frente a datos atípicos. Incluyeron variables como las muestras del valor de la tierra en diferentes localidades, así como otras variables agrupadas en las categorías de distancia (a redes viales, cursos de agua, zonas de mayor o menor categoría) y entorno (porcentaje de metros cuadrados edificados frente al total de metros cuadrados del terreno, porcentaje de tierras baldías, entre otras), utilizando métodos cartográficos, bases de datos catastrales y satelitales. Se evaluó la capacidad predictiva del algoritmo con el método de validación cruzada. Con esta metodología realizaron la predicción de la valuación masiva de tierra urbana y su estimación de la distribución espacial de la consistencia de esta estimación.

Los estudios anteriores aportan a esta investigación porque indican los posibles modelos, variables y test estadísticos a utilizar en análisis relacionados con el valor del suelo o de la vivienda. El estudio de Melo[38] aporta ideas de visualización en un contexto geográfico, mientras que el informe de DNP [39] permite identificar que las variables que explican el precio del suelo en el contexto normativo de los instrumentos de gestión de la Ley 388 de 1997 son diferentes para cada ciudad, lo cual es importante si se requiere replicar el modelo de este estudio en otros contextos dentro de Colombia.

El estudio de Cerino et al[40], pese a que no es realizado en Colombia, es relevante porque obtiene predicciones con buena precisión a partir de un conjunto de datos limitado (1.031 observaciones de mercado) para predecir el valor sobre 92.242 parcelas urbanas, mostrando gráficamente la ubicación geográfica de las estimaciones con mayor variabilidad de manera transparente para orientar la toma de decisiones sobre los sectores donde se puede fortalecer la estimación. La anterior técnica servirá como referente de posibles modelos a utilizar en contexto de los objetivos propuestos.

5. METODOLOGÍA

Para el desarrollo de este proyecto se implementó la metodología CRISP-DM, que se compone de las siguientes fases: entendimiento del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y presentación de resultados. Si bien las fases de entendimiento del negocio y comprensión de los datos se abordaron durante el anteproyecto, con el fin de garantizar su viabilidad, en los siguientes apartados se sintetizarán todas las actividades desarrolladas en cada fase del proyecto.

5.1. ENTENDIMIENTO DEL NEGOCIO

En Bogotá, en el marco de los Decretos Distritales 190 de 2004[3] y 555 de 2021[4] se instituyeron los Planes Parciales como instrumento para desarrollar y complementar las disposiciones del Plan de Ordenamiento Territorial (POT). Sin embargo, existen dilaciones en su formulación o adopción, particularmente en la fase de consulta pública, lo cual ha estado directamente relacionado con la valoración del suelo que impacta directamente la negociación de cargas y beneficios, la gestión de la plusvalía y la adquisición de tierras generando desacuerdo entre los actores involucrados.

Aun cuando Bogotá D.C. cuenta con la IDECA [12] que pone a disposición del público un gran volumen de información espacial actualizada y de muy buena calidad relativa a la valoración del suelo, se encontró que esta fuente no ha sido utilizada para analizar el valor del suelo en contexto de los planes parciales y con estos resultados evitar dilaciones en la fase de información pública para agilizar su aprobación.

Teniendo en cuenta la problemática y la disponibilidad de información, se consideró relevante explorar la plataforma de información Geográfica https://www.ideca.gov.co/ y La Plataforma de Datos abiertos https://datosabiertos.bogota.gov.co/ [41] para identificar la suficiencia de información para cubrir la necesidad de explicar las variaciones en el valor del suelo en las zonas donde se han adoptado planes parciales, y de esta manera desarrollar el modelo predictivo para facilitar la toma de decisiones institucionales y particulares de manera ágil y concertada.

En este proceso de exploración se encontraron alrededor de 2.057 conjuntos de datos actualizados, estandarizados e interoperables de la ciudad de Bogotá pertenecientes a 70 entidades de la administración distrital. Comprendiendo la necesidad de analizar la valoración del suelo en contexto de la influencia de los planes parciales se seleccionaron principalmente, los provenientes de la SDP y la UAECD que eran los más adecuados dado que estas entidades se enfocan en aspectos valuatorios del suelo y normativos referentes a la ordenación del territorio.

Además de la pertinencia de las bases se aseguró contar con datos del valor del suelo antes y después de la adopción de los diferentes planes parciales dentro del período mencionado, con el fin de tener los insumos necesarios para el modelo predictivo propuesto. Considerando la disponibilidad de información y el objetivo del modelo, se precisó que no será parte del alcance del proyecto identificar las zonas óptimas para la adopción de futuros planes parciales, ni hacer recomendaciones específicas sobre su selección teniendo en cuenta que estos análisis responden a múltiples requisitos de la administración distrital.

5.2. COMPRENSIÓN DE LOS DATOS

5.2.1. Gestión y recolección de datos:

La recolección de datos para el presente proyecto se realizó a partir de información geográfica y estadística proveniente de entidades distritales, disponible en plataformas oficiales de datos abiertos del Distrito Capital. Como resultado de este proceso, se recopilaron 41 bases de datos correspondientes al periodo 2012–2025, las cuales fueron organizadas en nueve grupos de información según su relación con la variable dependiente del estudio (valor del suelo por metro cuadrado a nivel de manzana) y las variables explicativas consideradas para el modelamiento.

Las fuentes de información fueron obtenidas principalmente a través de la IDECA y del portal oficial de Datos Abiertos de Bogotá. Posteriormente, la información fue organizada y almacenada de manera estructurada en un repositorio digital, con el fin de garantizar su trazabilidad, consulta y adecuado procesamiento durante las diferentes etapas de preparación, integración, análisis y modelamiento de datos.

Los grupos de información definidos para el estudio corresponden a: i) avalúo catastral por manzana, ii) base de datos catastral para Bogotá D.C., iii) estratificación socioeconómica por manzana, iv) mapa de referencia de Bogotá D.C., v) mediana del valor de referencia de terreno por manzana, vi) mediana del valor integral por grupo económico, vii) información asociada al POT Decreto 555 de 2021, viii) predios liquidados con efecto plusvalía y ix) seguimiento de planes parciales de desarrollo.

La organización general de las fuentes de información utilizadas se presenta en la Fig. 2.

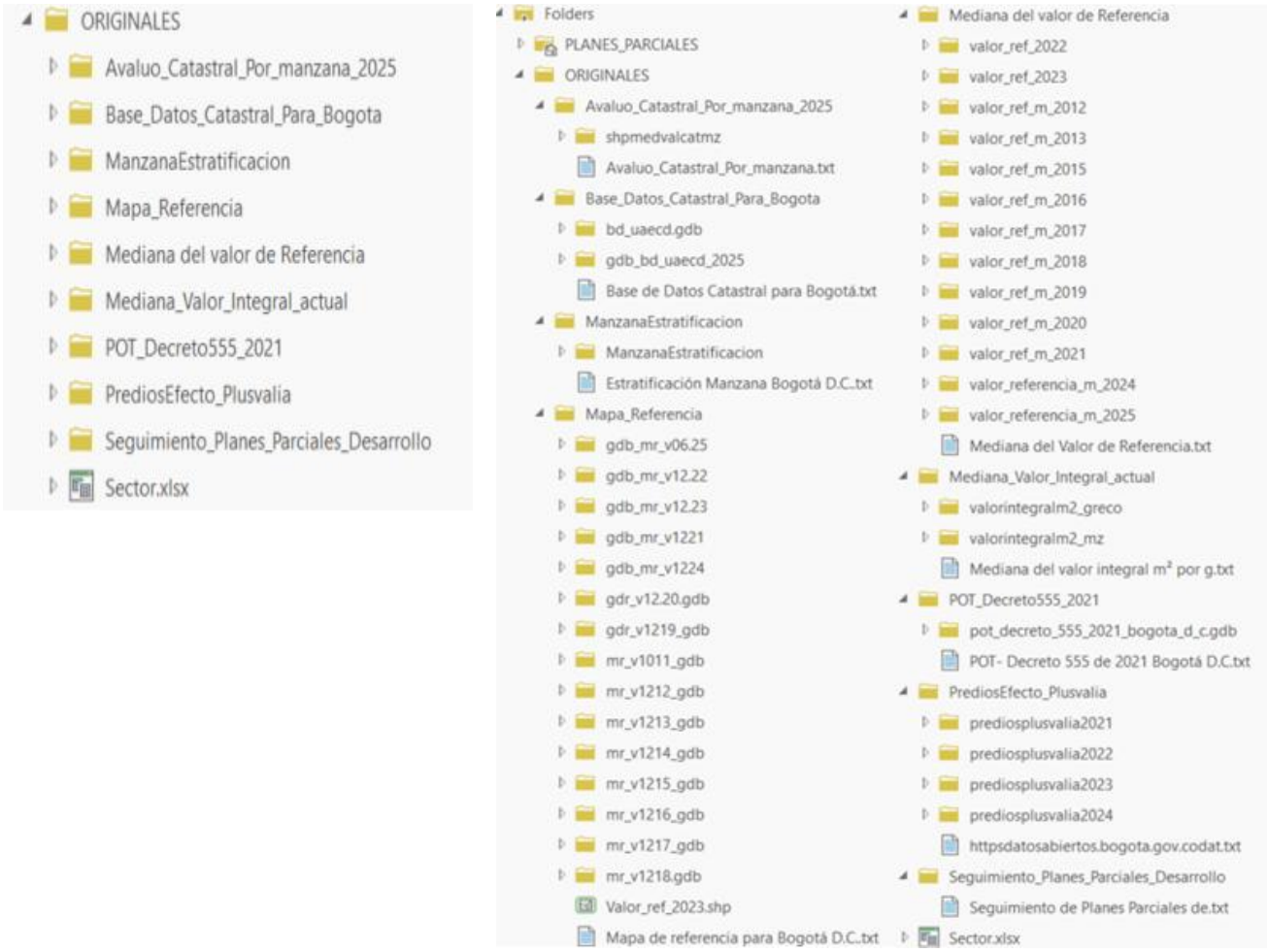


Fig. 2. Estructura de organización de la información.
Fuente: elaboración propia a partir de datos de IDECA.

5.2.2. Descripción de los datos

Las fuentes de información utilizadas en el presente estudio integran información catastral, territorial, normativa y socioeconómica, permitiendo caracterizar espacial y temporalmente el comportamiento del valor del suelo a nivel de manzana en Bogotá D.C. Estas fuentes constituyen el insumo principal para las etapas de preparación, integración y análisis de datos desarrolladas en la investigación.

La Tabla V. presenta un resumen de las principales fuentes de información utilizadas, indicando su entidad de origen, descripción general y propósito dentro del estudio.

El detalle técnico de las fuentes de información, incluyendo estructura de geodatabases, capas geográficas, atributos y organización de los conjuntos de datos empleados en el análisis, se presenta en el siguiente anexo:

Anexo A. Descripcion_de_los_datos.

Tabla V. Descripción de los datos

Fuente de información	Entidad	Descripción	Cobertura temporal	Uso principal
Avalúo catastral por manzana	IDECA – UAECD	Shapefile con información agregada del avalúo catastral a nivel de manzana urbana y de expansión urbana	2025	Validación y referencia catastral
Base de Datos Catastral Bogotá D.C.	UAECD	Geodatabase con información espacial y alafanumérica de las unidades catastrales (predios, construcciones, manzanas, sectores catastrales)	2025	Validación y referencia catastral
Estratificación Manzana Bogotá D.C.	IDECA	Shapefile con clasificación socioeconómica del territorio a nivel de manzana	2025	Variable socioeconómica
Mapa de referencia Bogotá D.C.	IDECA	14 bases de datos espaciales que contienen cartografía base y capas asociadas a la gestión catastral	2012–2025	Construcción de variables territoriales y soporte de integración espacial
Mediana del valor de referencia de terreno por manzana	UAECD	Shapefiles con Información histórica del valor referencial del suelo por manzana	2012–2025	Variable objetivo del modelo
Mediana del valor integral m² por grupo económico	UAECD	Valor integral del suelo según actividad económica predominante	2024	Análisis complementario
POT – Decreto 555 de 2021	Secretaría Distrital de Planeación	Información normativa y territorial del ordenamiento urbano	2021	Variables de accesibilidad y planificación
Predios liquidados con efecto plusvalía	Bogotá D.C.	Registro de predios sujetos a contribución por plusvalía	2021–2024	Contextualización urbanística
Seguimiento de Planes Parciales	Secretaría Distrital de Planeación	Información descriptiva y normativa de planes parciales adoptados	2001–2025	Caracterización de planes parciales

Fuente: elaboración propia a partir de datos de IDECA. El detalle técnico y las referencias completas de las fuentes utilizadas se presentan en el Anexo A.

En el presente estudio, las fuentes de información recopiladas constituyeron el insumo base para la construcción de variables derivadas relacionadas con accesibilidad, intervención urbana y características territoriales del entorno, utilizadas posteriormente durante la etapa de preparación de datos.

5.2.3. Selección de datos

Una vez recopiladas las fuentes de información, se realizó un proceso de exploración, revisión y validación de los atributos disponibles en cada conjunto de datos, con el propósito de identificar las variables pertinentes para el desarrollo del modelo predictivo.

La selección de datos se efectuó considerando criterios de pertinencia temática, disponibilidad, consistencia de la información, cobertura temporal y nivel de desagregación espacial. A partir de este proceso se identificaron variables asociadas al valor del suelo, características físicas del territorio, condiciones socioeconómicas, accesibilidad e información relacionada con instrumentos de planificación urbana.

La Tabla VI presenta el resumen de los principales datos seleccionados, indicando el conjunto de datos de origen, nivel espacial y propósito dentro del estudio.

Tabla VI. Variables y conjuntos de datos seleccionados para el análisis.

Conjunto de datos	Capa / Tabla	Variable	Justificación
Mediana del Valor de Referencia de Terreno por Manzana - Bogotá D.C.	Valor_Ref_M_año	MANCODIGO	Permite integrar la información entre diferentes bases de datos a nivel de manzana.
Mediana del Valor de Referencia de Terreno por Manzana - Bogotá D.C.	Valor_Ref_M_año	VREF (valor del suelo m^2)	Representa el valor comercial del suelo a nivel manzana, siendo la variable objetivo del estudio.
Mediana del Valor de Referencia de Terreno por Manzana - Bogotá D.C.	Valor_Ref_M_año	Anno	Permite analizar la evolución temporal del valor del suelo.
Mediana del Valor de Referencia de Terreno por Manzana - Bogotá D.C.	Valor_Ref_M_año	Shape_Area	Incorpora el tamaño de la unidad espacial como factor asociado al valor del suelo, en metros cuadrados.
Estratificación Manzana Bogotá D.C.	ManzanaEstratificacion	Estrato	Captura las condiciones socioeconómicas del entorno.
Mapa de referencia para Bogotá D.C.	SCat	Código De Sector	Permite estructurar el análisis territorial a una escala intermedia.
Mapa de referencia para Bogotá D.C.	Loca	Código De Localidad	Representa un nivel de agregación territorial superior.
Mapa de referencia para Bogotá D.C.	Lote	Lotcodigo	Se utiliza para el conteo de predios por manzana, al considerarse una variable relevante en la caracterización del valor del suelo.
Mapa de referencia para Bogotá D.C.	Uso	Tipo de uso	Permite caracterizar funcionalmente los inmuebles y su incorporación como variable explicativa en el modelo. Usos a nivel predial.
POT- Decreto 555 de 2021 Bogotá D.C.	Clasificación del suelo	Clase_Suelo	Incorpora el componente normativo del uso del suelo, indicando si es urbano, rural o de expansión urbana.

POT- Decreto 555 de 2021 Bogotá D.C.	Plan parcial	Codigo_Plan_Parcial	Permite identificar de manera única cada plan parcial.
POT- Decreto 555 de 2021 Bogotá D.C.	Plan parcial	Tipo	Permite clasificar los planes según su tipo de desarrollo.
POT- Decreto 555 de 2021 Bogotá D.C.	Plan parcial	Nombre	Facilita la identificación descriptiva de cada plan parcial.
POT- Decreto 555 de 2021 Bogotá D.C.	Plan parcial	Estado	Garantiza que el instrumento cuenta con validez normativa. Estado = “Adoptado”
POT- Decreto 555 de 2021 Bogotá D.C.	Plan parcial	Fecha_Acto_Administrativo	Permite ubicar temporalmente la adopción del plan parcial: Fecha del acto administrativo posteriores al 1 de enero de 2012,
POT- Decreto 555 de 2021 Bogotá D.C.	Parque	Identificador Único Parque	Permite asociar cada unidad espacial con los parques a gran escala próximos.
POT- Decreto 555 de 2021 Bogotá D.C.	RedMetro	Identificador tramo Red Metro	Permite identificar obras de infraestructura de gran impacto que se han o se están construyendo. Metro
POT- Decreto 555 de 2021 Bogotá D.C.	Red_infraestructura_Corredor_transporte	Identificador tramo Transmilenio	Permite identificar obras de infraestructura de gran impacto que se han o se están construyendo. Transmilenio Av 68.
POT- Decreto 555 de 2021 Bogotá D.C.	Red_ferrea	Identificador red férrea	Permite identificar obras de infraestructura de gran impacto que se han o se están construyendo. Regiotram.
POT- Decreto 555 de 2021 Bogotá D.C.	Amenazas_Riesgos_Mitigacion	Suelo_proteccion_riesgo	Permite identificar áreas del territorio que han sido clasificadas como suelo de protección por condiciones de riesgo. Puede explicar valores bajos del suelo o valores atípicos en el límite inferior.

Fuente: elaboración propia a partir de datos de IDECA.

5.3. PREPARACIÓN DE LOS DATOS

Dentro de la preparación de los datos relevantes para la construcción del Modelo, que corresponde al primer objetivo propuesto, se desarrollaron las siguientes actividades:

5.3.1 Integración de los datos

Una vez identificados los datos de interés, se llevaron a cabo los procesos necesarios para la integración de la información, dado que los datos se encontraban distribuidos en distintos conjuntos y estructuras.

Para ello, se definió una base de datos de trabajo, a partir de la cual se realizó la consolidación de la información, generando una única capa geográfica tipo polígono, con la geometría de las manzanas catastrales de Bogotá, que integra los atributos relevantes del análisis. Esta capa reúne tanto la variable a predecir como los datos que posteriormente darán lugar a las variables explicativas del modelo, a partir de las cuarenta y un (41) bases de datos recolectadas.

La consolidación de los datos se realizó mediante procesos de análisis, depuración, estandarización, unión y homologación de la información tanto espacial como alfanumérica, garantizando la correspondencia entre las diferentes fuentes de información e identificando claramente el año de dichas fuentes.

La Tabla VII. resume los principales procesos de integración y transformación desarrollados para la construcción de la base consolidada de análisis espacial. La descripción metodológica detallada de cada procedimiento se presenta en el siguiente anexo.

Anexo B. Integracion_de_Datos

Tabla VII. Integración de los datos.

Num.	Proceso	Descripción metodológica resumida	Herramientas / Procesos	Resultado generado	Figura
I	Integración base Mediana del Valor de Referencia de Terreno 2012–2025	Consolidación histórica espacial de las capas de valores de referencia de terreno utilizando como geometría base las manzanas del año 2012.	Project, Join Field, Alter Field	Base espacial integrada “DatosManz2012_2025.gdb”	Ver Anexo A– Fig. 1
II	Cantidad de predios por manzana	Generación del conteo de predios por manzana para cada vigencia anual.	Field Calculation, Join Field	Variables de conteo de predios por manzana (2012–2025)	Ver Anexo A – Fig. 2
III	Estrato por manzana	Integración del estrato y código de zona provenientes de la estratificación por manzana.	Join Field	Variables socioeconómicas por manzana	—
IV	Clasificación del suelo	Incorporación de la clasificación del suelo definida en el POT mediante selección espacial.	Select by Location, Join Field	Clasificación del suelo por manzana	—
V	Clasificación valor manzana	Identificación de continuidad histórica de la información de precios por manzana.	Field Calculation, Reglas categóricas	Variable ClasificacionValorMz	Ver Anexo A– Fig. 3
VI	Uso del suelo	Determinación del uso predominante por área para cada manzana y vigencia anual a partir de información predial.	Field Calculation, Calculate Geometry, Summarize, Statistics, Join, Recodificación de atributos, Estandarización	Variables de uso predominante del suelo por vigencia	Ver Anexo A– Fig. 4-5
VII	Intervención	Identificación de manzanas intervenidas por planes parciales adoptados.	Select by Location, Field Calculation	Variable binaria Intervenido	—

VIII	Sector	Construcción del identificador de sector a partir de los seis primeros dígitos del MANCODIGO.	Field Calculation	Variable Sector	—
IX	Suelo de protección por riesgo	Identificación de manzanas ubicadas en zonas clasificadas como suelo de protección por riesgo	Select by Location, Field Calculation	Variable binaria de riesgo	—
X	Distancia a infraestructura y parques	Cálculo de distancias mínimas a obras de infraestructura y parques de gran escala.	Near	Variables de distancia mínima a infraestructura y parques	—
XI	Distancia a planes parciales	Obtención de distancia a los dos planes parciales más cercanos y ranking de proximidad.	Near Table, Join	Variables de proximidad a planes parciales	Ver Anexo A – Fig. 6
XII	Atributos de plan parcial	Integración de atributos normativos y administrativos de los planes parciales.	Join	Base de datos consolidada final para análisis estadístico y modelamiento	

Fuente: elaboración propia a partir de datos de IDECA

Posteriormente, la información fue depurada y organizada, y la tabla resultante fue exportada a formato Excel, obteniendo una base de datos compuesta por 87.841 registros en formato ancho, la cual fue utilizada para el desarrollo del análisis descriptivo y estadístico en el entorno de programación Python (Google Colab), la cual se encuentra bajo el siguiente anexo:

Anexo C. Base_datos_integrada_2012_2025

5.3.2 Análisis exploratorio de los datos

Una vez consolidada la base de datos espacial, y previo a los procesos de limpieza y procesamiento de la información, se realizó un análisis exploratorio inicial mediante la visualización de las manzanas de Bogotá en función de su distancia a los planes parciales. Para ello, las distancias fueron clasificadas en rangos, permitiendo representar espacialmente los niveles de proximidad de cada manzana frente a estos instrumentos de planificación urbana.

Esta representación cartográfica permite identificar la distribución espacial de las manzanas según su grado de cercanía o lejanía a los planes parciales, evidenciando la existencia de zonas con alta proximidad y otras significativamente alejadas, así como áreas de la ciudad donde se observa una mayor concentración de planes parciales. De esta manera, se obtiene una visión general del alcance espacial de estos instrumentos y de la variabilidad en las distancias dentro del territorio.

Este análisis constituye un primer acercamiento para dimensionar la relación espacial entre las manzanas y los planes parciales, sirviendo como insumo para la formulación de hipótesis y la posterior construcción de variables relacionadas con accesibilidad y proximidad en el modelo. La visualización obtenida se presenta en la Fig. 3.

Como complemento a este análisis espacial, se desarrolló un ejercicio exploratorio con el fin de evaluar de

manera preliminar si la adopción de planes parciales influye en el comportamiento del valor promedio del metro cuadrado a nivel de manzana.

Para ello, se seleccionaron de manera aleatoria dos planes parciales adoptados con posterioridad al año 2012, considerando que la base de datos inicia en este año y que el objetivo es analizar el comportamiento del precio antes y después de la adopción del plan.

Se tomó un umbral preliminar de 500 metros, el cual se adoptó como criterio de proximidad en este análisis, con el fin de delimitar un área de influencia y facilitar la identificación de posibles efectos espaciales en el entorno de los planes parciales. En este sentido, se seleccionaron aleatoriamente algunas manzanas ubicadas dentro del área de influencia, definidas como aquellas situadas a una distancia menor o igual a 500 metros y de manera complementaria, se seleccionaron manzanas ubicadas fuera del área de influencia, es decir, localizadas a distancias mayores de 500 metros, para conformar un grupo de comparación sin intervención directa.

Este ejercicio cualitativo permitió realizar una aproximación preliminar al posible efecto de la adopción de los planes parciales sobre el valor del metro cuadrado por manzana en Bogotá. El análisis se desarrolló en Microsoft Excel, mediante la construcción de gráficos de series temporales del valor del suelo para las manzanas seleccionadas. Los resultados obtenidos se presentan en la Fig. 4.

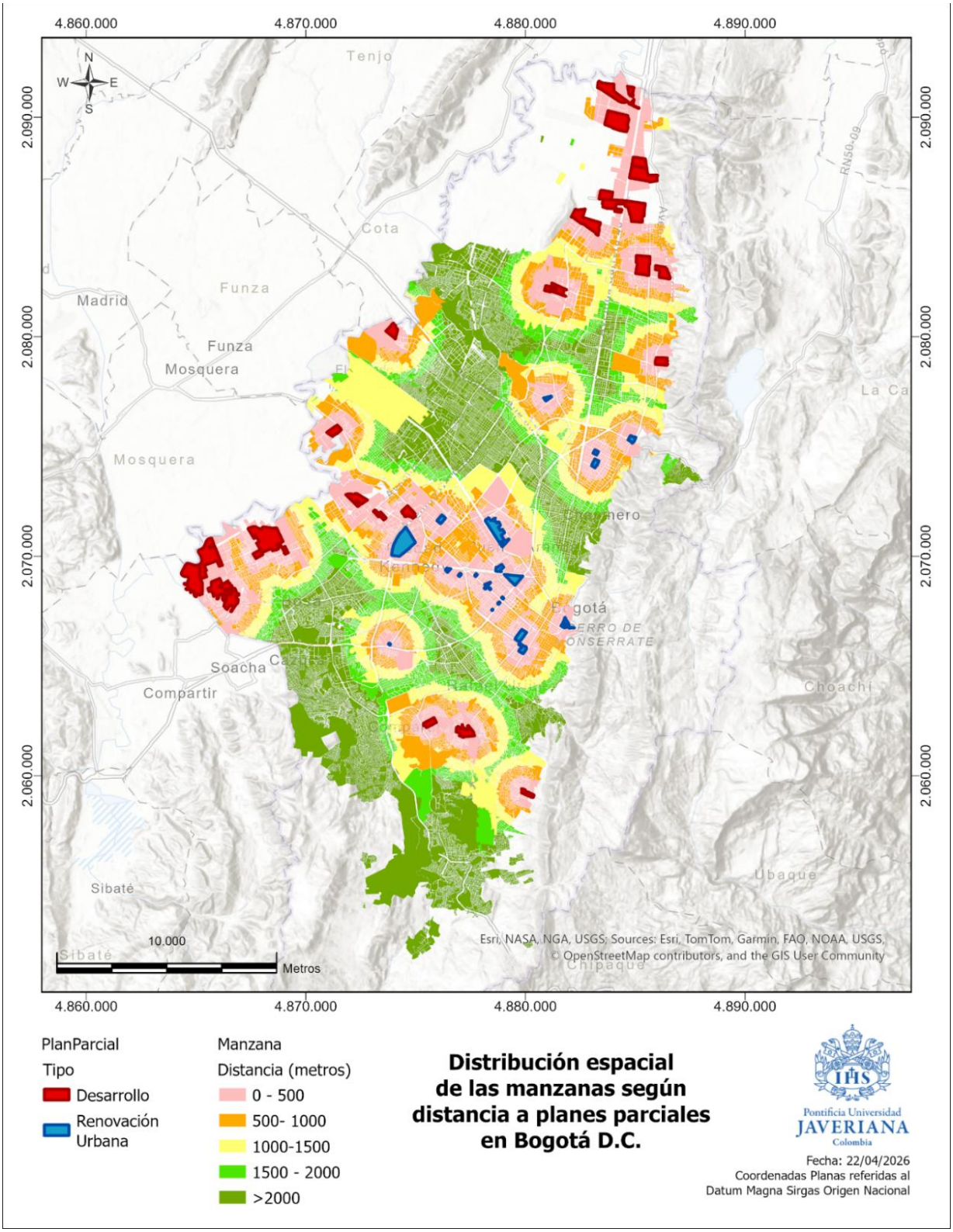


Fig. 3. Distribución espacial de las manzanas según distancia a planes parciales en Bogotá D.C.

Fuente: elaboración propia a partir de datos de IDECA.

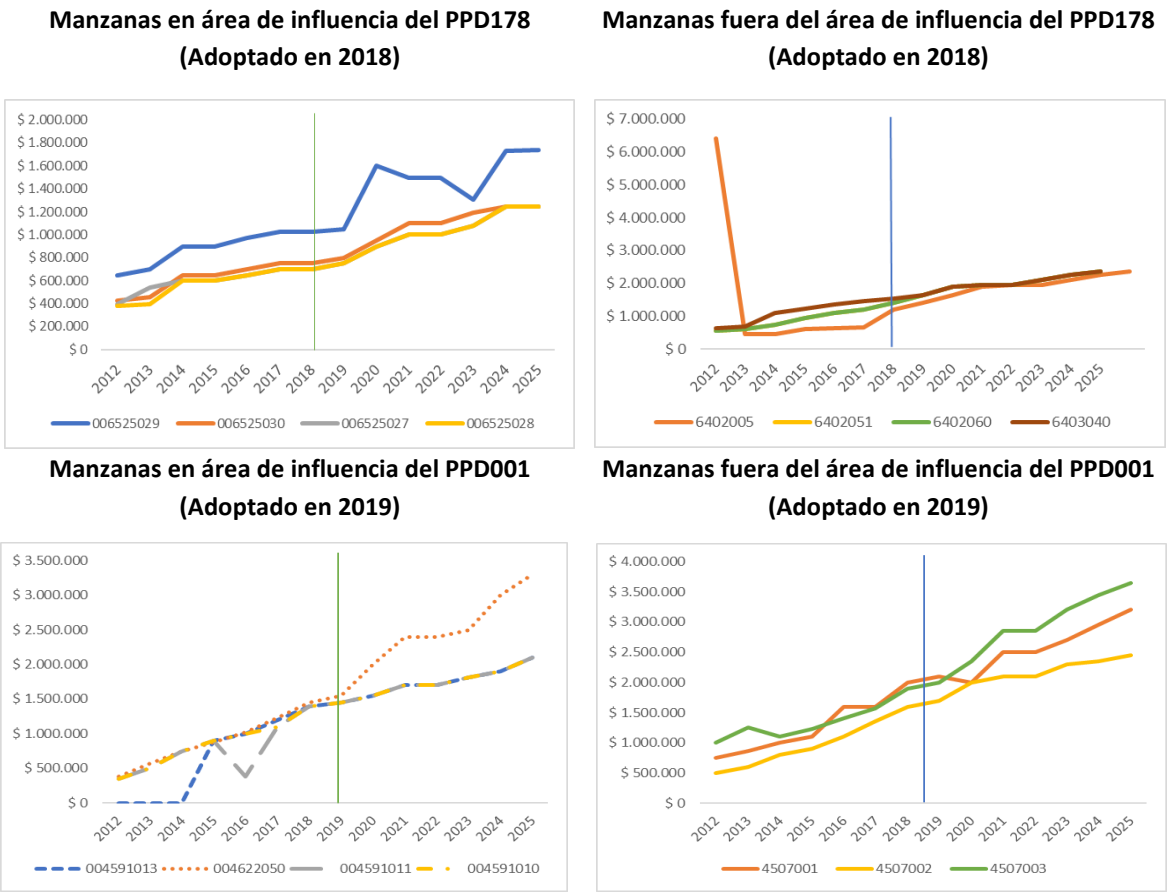


Fig. 4. Comparación del efecto de la adopción de los planes parciales sobre el valor del suelo por manzana: (a) Manzanas en área de influencia del PPD178; (b) Manzanas fuera del área de influencia del PPD178; (c) Manzanas en área de influencia del PPD001; (d) Manzanas fuera del área de influencia del PPD001.

Fuente: elaboración propia a partir de datos de IDECA.

De manera preliminar, los resultados anteriores permitieron validar que en los casos donde las manzanas se encuentran dentro del área de influencia de los planes parciales, se presenta un cambio incremental del valor del suelo (precio del metro cuadrado por manzana), lo cual permite validar parcialmente la hipótesis de que el desarrollo y adopción de planes parciales en Bogotá genera un efecto de valorización sobre el valor del metro cuadrado. No obstante, este ejercicio fue ampliado y replicado para todos los planes parciales, con métodos estadísticos y de evaluación de impacto que se describirán en secciones posteriores.

5.3.3 Informe estadístico

Descripción de los datos

Para el desarrollo del análisis descriptivo, se trabajó inicialmente con la base de datos exportada desde *ArcGIS Pro*, denominada Anexo C. Base_datos_integrada_2012_2025, la cual integra la información espacial y alfanumérica previamente procesada. Esta base fue cargada y procesada en el entorno de programación Python (Google Colab).

La base presenta una estructura en formato ancho (wide format), en la que cada registro corresponde a una unidad de análisis a nivel de manzana, identificada mediante el código MANCODIGO, y las variables asociadas se organizan en columnas. En particular, las variables relacionadas con el valor del suelo (V_REF), así como otras características temporales y físicas, se encuentran desagregadas por año en diferentes columnas, lo que permite observar la evolución temporal de cada manzana dentro de una misma fila. La base de datos cuenta con un total de 87.840 registros y 83 atributos, incluyendo variables de tipo numérico, categórico y temporal, así como campos derivados del procesamiento espacial, como distancias a planes parciales y a obras de infraestructura.

Si bien esta estructura facilita la visualización inicial de la información, más adelante para el desarrollo de análisis temporales y de modelación es necesario transformar los datos a un formato largo (long format), en el cual cada fila representa una observación por manzana y año.

Procesamiento y limpieza de la base de datos

A partir de la base de datos en formato ancho, se llevó a cabo el procesamiento y limpieza de la información, con el fin de garantizar la consistencia, calidad y adecuación de los datos para su uso en el desarrollo de los modelos predictivos. Para ello, se realizaron una serie de acciones orientadas a la depuración, transformación y validación de la base de datos, las cuales se describen a continuación.

1. Se transformaron las observaciones en las que el valor "<Null>" se encontraba almacenado como texto, reemplazándolas por valores nulos (NaN), con el fin de permitir su correcta lectura y procesamiento posterior.

2. Se transformó el tipo de variable de las columnas con prefijos de “Area” (a partir de 2019), "V_REF" y “Distancia_” como "Object" a su formato correcto "float" considerando que corresponden a variables numéricas que contienen valores decimales. De manera similar, se transformaron las columnas con el prefijo "CntPrd" de "object" a "int" dado que corresponde al conteo de predios por manzana. Esta última variable presenta valores nulos en diferentes manzanas, lo cual se explica porque en alguno de los años específicos la manzana no existía y por tanto no se reporta número de predios vinculados a su código de manzana (MANCODIGO).

3. Se transformó la variable MANCODIGO a "object" dado que es un identificador de las manzanas y no es una variable con la cual se realizan operaciones.

4. La variable “Intervenido” fue convertida a tipo booleano, dado que representa un clasificador binario que indica si una manzana fue o no fue intervenida por un plan parcial.

5. Se eliminaron las variables "Shape_Length.1" y "Shape_Area.1", toda vez que fueron generadas automáticamente por el software ArcGis Pro, aparecían en 4 columnas distintas y no tenían relevancia para el análisis del proyecto.

6. Se eliminaron las variables "Nombre" y "Número Acto Administrativo", dado que no aportan información relevante para el modelo.

7. Se generó la variable “Anno_Adopcion_Plan” a partir de la variable "Fecha Acto Administrativo" y se eliminó esta última.

Por otra parte, se identificó que las variables como ESTRATO, RANKING_CERCANIA_PP tenían tipo "int" o "float" pero que conceptualmente son categóricas ordinales y debían transformarse como factores. Sin embargo, por facilidad del procesamiento se decidió realizar este paso posterior a la transformación de la base de datos a formato largo para que los modelos planteados posteriormente las interpreten adecuadamente. Este procesamiento se encuentra en el script de Google Colab:

Anexo D. Analisis_exploratorio_parte1

Depuración de Información

Dado que se identificaron valores nulos en diferentes columnas asociadas al prefijo “V_REF”, se realizó un proceso de depuración de la información sobre la base en formato ancho, el cual facilita este tipo de tratamiento. Para ello, se llevaron a cabo una serie de acciones orientadas a garantizar la completitud y consistencia de los datos, con el fin de obtener una base robusta y adecuada para la fase de modelación.

1. Identificación de manzanas con datos incompletos

   Se identificaron y cuantificaron las manzanas (MANCODIGO) que presentaban más de seis valores nulos en la serie temporal, para lo cual, se utilizó la variable ClasificacionValorMz, construida a partir del conteo de valores nulos en la serie de tiempo de cada manzana.

   Esta variable permitió clasificar las manzanas según la continuidad de la información, identificando aquellas con datos completos y aquellas con comportamientos discontinuos.

2. Evaluación de manzanas con pocos valores nulos

   Las manzanas que presentaban menos de seis valores nulos en la variable V_REF fueron reservadas para análisis posterior, una vez transformada la base a formato largo, para evaluar la viabilidad de imputar los valores faltantes o eliminar únicamente las observaciones incompletas.

3. Inconsistencias en la integridad espacial

   Se identificaron manzanas con inconsistencias en la integridad espacial, correspondientes a casos en los cuales un mismo identificador (MANCODIGO) se encontraba asociado a varios polígonos o presentaba geometrías duplicadas. Estos casos fueron analizados mediante análisis espacial en *ArcGIS Pro*, identificando tanto errores topológicos (sobreposiciones y duplicidad de geometrías) como errores temáticos (asignación incorrecta del identificador).

   Se evidenció además que la base predial del catastro distrital presenta una depuración a partir del año 2018, en la cual se corrigen estas inconsistencias. Con base en ello, se ajustó la información de años

anteriores, lo cual garantiza la correspondencia entre geometría e identificador. Un ejemplo de este tipo de inconsistencias se presenta en la Fig. 5.

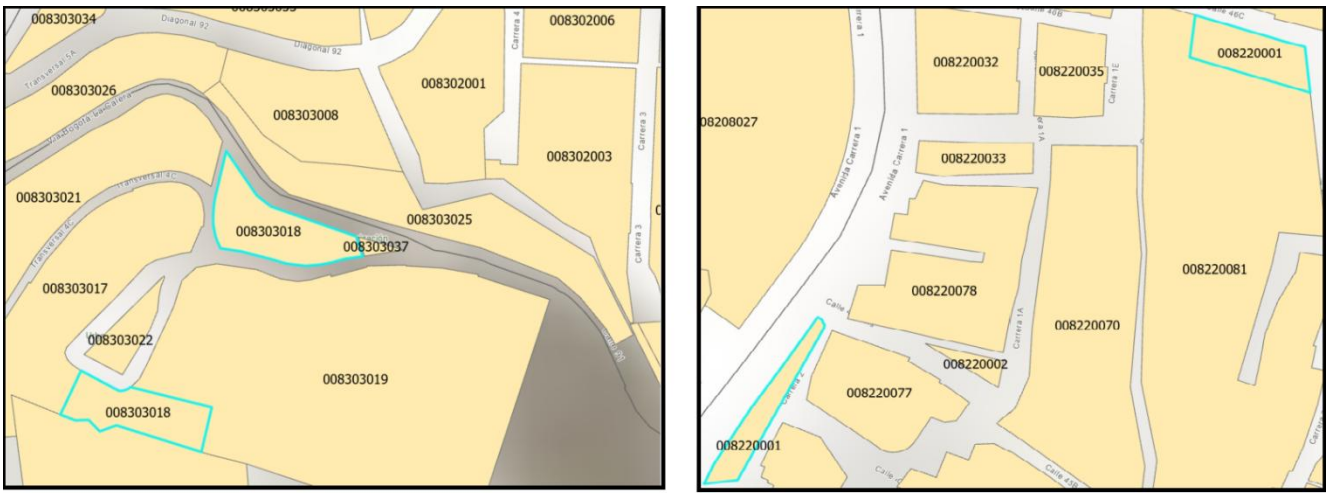


Fig. 5. Ejemplo donde múltiples polígonos comparten el mismo identificador (MANCODIGO), previo a la depuración de la información.
Fuente: elaboración propia a partir de datos de IDECA.

4. Filtrado de planes parciales por coherencia temporal

Se identificó que algunas manzanas podían estar asociadas a diferentes planes parciales a distancias similares. No obstante, varios de estos planes fueron adoptados antes de 2012. Dado que el análisis inicia en este año, se excluyeron dichos planes, para garantizar la coherencia temporal y la correcta evaluación de su efecto sobre el valor del suelo. (Ver Fig. 6).

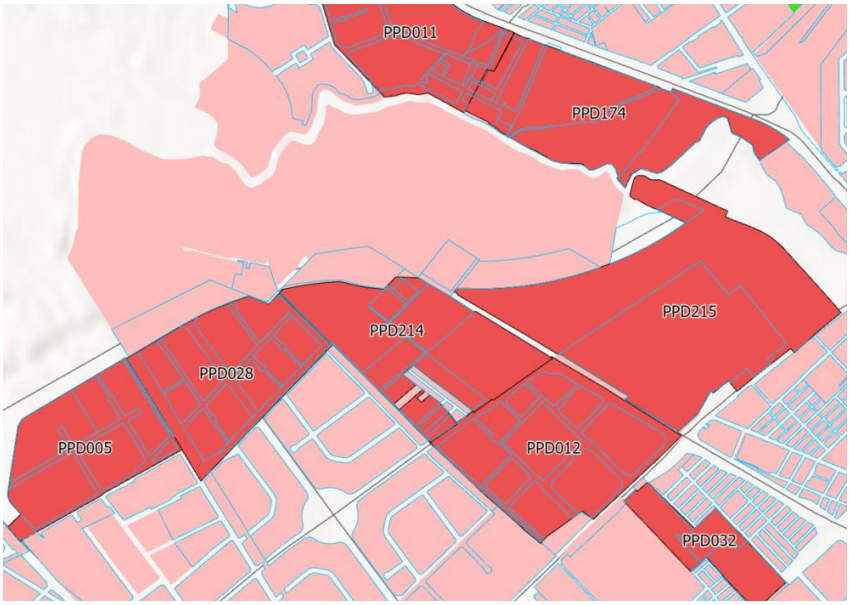


Fig. 6. Ejemplo Planes Parciales Contiguos.
Fuente: elaboración propia a partir de datos de IDECA.

5. Clasificación de planes parciales según tipo de suelo

   Se identificaron tres categorías de suelo: urbano, rural y de expansión urbana y se evidenció que algunos planes parciales se localizan parcialmente en más de una categoría. Para resolver esta situación, se aplicó un criterio de predominancia espacial, para lo cual se asignó a cada plan una única categoría que representara más del 50% de su área. En consecuencia, cada plan parcial fue asociado a una única clase de suelo, garantizando consistencia en el análisis.

6. Reintegración y procesamiento en Python

   Una vez realizados los procesos de depuración y ajuste en *ArcGIS Pro*, el Feature dataset fue actualizado e integrado nuevamente. Posteriormente, la base de datos fue exportada y cargada en el entorno de programación Python (Google Colab), con el fin de continuar con las etapas de procesamiento, transformación y análisis.

7. Eliminación de manzanas con alta proporción de datos faltantes

   Se eliminaron las manzanas que presentaban más de la mitad de la serie temporal con valores nulos, debido a la inviabilidad de imputación. Como resultado, se excluyeron 1.237 manzanas.

8. Transformación a formato largo (long format)

   La base fue transformada a formato largo para las variables temporales, permitiendo que cada fila representara una observación por manzana y año. Durante este proceso, se creó la variable “Anno”, extrayendo el año de variables como “V_REF”, “Area”, “Uso” y “CntPrd”.

   Como resultado, el dataset pasó de 87.804 filas y 76 variables a 1.193.694 observaciones y 26 variables.

9. Ajuste de tipos de datos

   Este cambio transformó las columnas con identificador “V_REF”, “Area” y “CntPrd” nuevamente a tipo “object”, las cuales fueron transformadas a “float” (V_REF, Area) e “int”, respectivamente. Respecto a la variable “CntPrd” se reemplazaron los valores NaN con “0” para poder transformar a tipo “int”.

10. Eliminación de valores iniciales inconsistentes

    A partir de los clasificadores definidos en la variable ClasificacionValorMz, se identificaron las observaciones cuya serie temporal no iniciaba en el año 2012, sino en periodos posteriores. Con base en este análisis, se detectaron registros en los cuales el valor de referencia del metro cuadrado (V_REF) tomaba valores iguales a cero en los primeros años de la serie. No se consideró metodológicamente adecuado asumir que los valores iniciales son equivalentes a los observados en periodos posteriores, por lo cual se optó por eliminar dichas observaciones.

Ante la dificultad de imputar estos valores de manera consistente, se decidió eliminar únicamente las observaciones afectadas, sin excluir la totalidad de la manzana. Esta decisión se tomó fundamentada en que los modelos de Machine Learning no requieren necesariamente series completas y continuas para todas las variables del dataset.

Como resultado de este proceso, se eliminaron 2.852 filas de la base de datos.

11. Agrupación de la variable Uso

    Con relación a la variable "Uso", que representa los usos económicos predominantes por área a nivel de manzana, se identificó la existencia de aproximadamente 30 categorías diferentes. Con el fin de facilitar el análisis y mejorar la interpretabilidad del modelo, se realizó un proceso de agrupación de categorías, definiendo la variable "grupo_economico" a partir de características funcionales similares entre los distintos usos.

    Como resultado, se establecieron las siguientes categorías: *dotacional, industrial, bodegas, comercio y oficinas, residencial, depósitos y parqueaderos, y otros*.

Análisis e imputación de datos faltantes

1. Se identificaron valores nulos en grupo_economico y para su imputación se llevaron a cabo los siguientes pasos:

    a. Se calculó el Sector de cada manzana con los primeros 6 dígitos de la variable MANCODIGO.

    b. Se calcularon tres medidas de tendencia central (moda) para la variable grupo_economico:

        - grupo_moda: para cada manzana (MANCODIGO)
        - grupo_moda_sector: Por cada grupo económico y sector para cada año de análisis
        - grupo_moda_sector_sin_anno**:** Para cada sector y grupo económico sin tener en cuenta el año para luego realizar una unión por atributos (left join) de estas 3 variables sobre la base general.

    c. A partir de estas 3 modas se realizó la imputación teniendo en cuenta estos criterios:

        - Inicialmente imputar por la moda de la manzana.
        - Si no existía, se imputó por la moda del sector en el año correspondiente.
        - Como última opción si no se conoce ese valor, se imputó por la moda del sector independiente del año.

    d. En los casos donde no se encontró uso por ninguno de los criterios anteriores, se decidió eliminar estas manzanas porque no hay criterio para imputar los valores faltantes. En total se eliminaron 4 manzanas (88 filas) que corresponden a tres sectores, y en cada sector son las únicas manzanas

existentes.

2. Una vez revisada la base de datos, se identificaron valores faltantes en la variable de Conteo de Predios (“CntPrd”). Para esta variable, se definió reemplazar los valores por la moda de esta variable para cada manzana.

3. En cuanto a los valores faltantes de la variable “Area”, inicialmente se realizó un proceso de redondeo de decimales, con el fin de evitar diferencias mínimas entre años que no resultan significativas para el análisis, pero que podrían afectar la aplicación de condicionales y comandos en el procesamiento de la información. Para la imputación se siguieron estos criterios:

    a. Si toda la serie tenía el mismo valor se imputa ese valor en los campos vacíos. Con este criterio se imputaron 1.213 manzanas.

    b. Cuando las manzanas presentan cambios de área, se tuvieron en cuenta las siguientes condiciones:

        - Si el año anterior y el año posterior tenían la misma área se asignó el mismo valor;
        - Para el año 2022, las áreas presentaban diferencias, que podrían deberse al cambio de sistema de proyección cartográfica oficial para Colombia es MAGNASIRGAS / Origen-Nacional. En este caso se asignó el área menor, debido al comportamiento del nuevo sistema de proyección cartográfico.
        - Para el resto de los valores nulos en la variable se asignó el promedio de la serie.

4. En cuanto a los valores faltantes en la variable objetivo V_REF (valor del precio del metro cuadrado por manzana), se realizó el siguiente proceso:

    - Se identificaron los códigos de manzana (MANCODIGO) y se extrajeron en un dataset auxiliar

    - Se agrupó la base de datos por código de manzana (MANCODIGO) y ranking de cercanía a los planes parciales (NEAR_RANK). Al dividir el dataframe en subgrupos independientes la interpolación no se cruza en grupos distintos.

    - Se realizó una interpolación lineal de valores faltantes (NaN) en la columna V_REF, de forma independiente dentro de cada grupo definido por MANCODIGO y NEAR_RANK, y se guardó el resultado en una nueva columna llamada “V_REF_interp” que contiene los valores originales cuando estos existen y los interpolados donde antes había nulos.

      La interpolación lineal [x.interpolate(method="linear"], consiste en calcular una recta entre el valor anterior y el siguiente y asignar valores intermedios proporcionales. Con el parámetro “limit_direction = both”, se imputan los valores nulos que se presentan al principio o al final de la serie. Con este método se imputaron 10.794 observaciones.

- Se validó cuántos valores de la columna “V_REF_interp” quedaron nulos aún después de la imputación y se eliminó la manzana "001319088" que presentaba datos que no pudieron ser interpolados

Análisis y tratamiento de valores atípicos

1. *Ajuste de Variaciones atípicas en Valores de Referencia de Manzanas*

   Durante el análisis exploratorio de los datos, se evidenció que las manzanas localizadas en áreas clasificadas como “Suelo de Protección por Riesgo” presentan valores de metro cuadrado significativamente inferiores en comparación con el resto de las manzanas de la base de datos. En algunos casos, estos valores atípicos se concentran en agrupaciones espaciales que corresponden a barrios completos.

   A partir de esta identificación, se procedió a incorporar esta condición dentro del modelo mediante una variable espacial. Para ello se realizó una selección por localización en el software ArcGis Pro y se creó una columna nueva tipo booleano para indicar aquellas manzanas que cumplían dicha condición. Para hacerlo se tuvo en cuenta el feature Class "Suelo_proteccion_riesgo", ubicado en el Fetaure Dataset "Amenazas_Riesgos_Mitigacion", de la base de datos del POT de Bogotá.

   Posteriormente se exportó el análisis espacial a una hoja de Excel, para cargarla como insumo en el script de limpieza e integrarla con la base de trabajo. Utilizando este insumo se unió la información desde la tabla insumo “riesgo” hacia la base de trabajo (df_resultado) usando MANCODIGO como llave agregando la columna “Proteccion_Riesgo”.

   Posteriormente se convirtió la columna de clasificación de booleano para facilitar el filtro de información. Se eliminaron de la base las manzanas que tenían la clasificación de riesgo y que además tuvieran un valor por metro cuadrado inferior a $10.000. De esta manera se eliminaron las manzanas que tienen protección por riesgo, pero además tienen un valor de referencia atípicamente bajo.

   Con la ayuda del análisis espacial se detectó que las manzanas con códigos "004602002", "005622008", "201105019", "009131013" presentaban valores atípicos muy bajos en determinados años que podían afectar el análisis, por lo cual se realizó un filtro manual de la información eliminando los atípicos específicos de la serie de datos de estas manzanas.

2. *Valores atípicos de Número de predios por manzana para la serie de tiempo*

   Se analizaron las diferencias relativas en el conteo de predios con diferentes niveles de variación para capturar posibles datos atípicos. Sin embargo, en las manzanas identificadas con estas variaciones se analizaron otras variables como el área y el uso del suelo y estos cambios se explicaban por cambios en el uso del suelo y/o área del predio, por lo que se decidió conservar la información de esta variable sin realizar ninguna imputación o eliminación de manzanas.

3. *Identificar Valores atípicos de Área de manzana para la serie de tiempo*

   Se analizaron los cambios en esta variable y se encontró que la información era coherente, ya que cuando no hay variaciones del conteo de predios, las áreas permanecen igual, lo que indica que no hay valores atípicos en esta variable.

Pasos finales de limpieza y preparación del dataset

1. Asignar Localidad a partir del sector catastral, tomando como insumo el archivo Excel generado con los códigos de las localidades para cada sector. Para los valores nulos que quedaron en esta variable, se asignó el valor 19, correspondiente a Ciudad Bolívar. Finalmente se transformó esta variable a número (int). Adicionalmente, se validó espacialmente con el feature class Loca (del feature dataset Entidad Territorial), mediante la herramienta Select by Location de ArcGIS Pro.

2. Se eliminaron las siguientes columnas de variables creadas a lo largo de los procesos anteriores que se habían conservado para realizar las verificaciones de cálculo e imputación correctos:: "Area", "sector", "Localidad", "MomentoI", "grupo_moda", "grupo_moda_sector", "grupo_moda_sector_sin_anno", "grupo_economico_anterior", "Area_antes", "Area_imputada_puente", "V_REF", "Fue_imputado", "CntPrd_moda", "Uso_", "ClasificacionValorMz"

Este procesamiento se encuentra en el script de Google Colab bajo el nombre:

Anexo E. Analisis_exploratorio_parte2

Como resultado de estos procesos, se obtuvo una base de datos final, compuesta por 1.152.634 registros y 24 variables, lista para la etapa de modelación:

Anexo F. Base_datos_integrada_modelacion

5.3.4 Análisis de Impacto de los Planes Parciales sobre el Valor del Suelo

De manera complementaria a los objetivos propuestos, se realizó un análisis de impacto causal mediante el método de Diferencias en Diferencias (DiD) para evaluar el efecto de los Planes Parciales sobre el valor del suelo en Bogotá. Utilizar este enfoque tenía el propósito de responder a la pregunta: ¿qué habría ocurrido sin la intervención (plan parcial)? Esta metodología tiene supuestos distintos a los modelos de predicción, cuyo objetivo es estimar el valor del suelo a partir de variables observables. En este sentido, ambos análisis son complementarios; por una parte, el método de DiD permite validar la hipótesis de impacto diferencial del instrumento, mientras que los modelos predictivos, presentados en los objetivos siguientes, buscan generalizar la estimación del precio a nuevas observaciones.

En contexto de lo anterior, el análisis de impacto buscó evaluar si la implementación de planes parciales en la ciudad de Bogotá generaba un efecto causal sobre el valor del suelo (V_REF_Mz) en su área de influencia. El enfoque metodológico utilizado fue un modelo de Diferencias en Diferencias (DiD) con efectos fijos, en el que

se estima el efecto promedio de la implementación de planes parciales sobre las manzanas de influencia.

Adicionalmente, considerando que los planes parciales pueden responder a dinámicas urbanas diferenciadas, se incorpora en el análisis la distinción según el tipo de plan, clasificándolos en planes de desarrollo y planes de renovación urbana. Esta diferenciación permite evaluar posibles efectos heterogéneos sobre el valor del suelo, reconociendo que cada tipo de intervención puede generar impactos distintos en el territorio. Tras un análisis exploratorio se observó en los planes de Renovación, la tendencia de un mayor aumento en el valor del suelo transcurrido un año de la adopción del plan parcial, mientras que en los planes de desarrollo este aumento es menor y se dificulta su valoración al observarse un estancamiento sostenido en el valor del suelo como consecuencia del efecto de la pandemia. Los resultados muestran que los planes de renovación presentan valores promedio del suelo más altos antes del plan, asociados a zonas con mayor consolidación urbana, mientras que los planes de desarrollo muestran mayores incrementos porcentuales, sugiriendo un mayor potencial de aumento de valor en áreas en proceso de consolidación.

Partiendo de los 500 metros definidos como influencia del plan parcial, a nivel exploratorio se definió el grupo de control o comparación como de las manzanas en el área de influencia, como las manzanas de iguales características en estrato y grupo económico a una distancia mayor a 2000 metros. Las variables de estrato y grupo económico se seleccionan a partir de la moda de las manzanas de influencia. Posteriormente, se incorporan al análisis, variables asociadas con la proximidad a infraestructura de gran impacto, debido a que estas variables también influyen en el valor del suelo.

Se observa que en general las manzanas tienen una tendencia creciente constante, por lo cual se puede afirmar que su comportamiento se explica por la influencia de proyectos de infraestructura que se han desarrollado a lo largo del periodo de estudio que, de la misma forma que los planes parciales, generan incremento en el valor del metro cuadrado por manzana.

Este procesamiento se encuentra en el script de Google Colab:

Anexo G. Analisis_estadistico_exploratorio_impacto

Análisis de impacto

Con base en el análisis exploratorio realizado previamente, se implementa un modelo de Diferencias en Diferencias (DiD) con efectos fijos para evaluar el impacto causal de la adopción de planes parciales sobre el valor de referencia por metro cuadrado a nivel de manzana con los siguientes parámetros:

Manzanas tratadas

- Manzanas intervenidas
- Manzanas a una distancia menor de 500 metros de un plan parcial.

Manzanas de control

- Misma localidad del plan
- Mismo estrato (moda por plan)
- Mismo grupo económico (moda por plan)

- Lejos de:
    - Plan parcial (>500m)
    - Infraestructura estratégica (>500m de metro, parques metropolitanos, Transmilenio Av 68, red férrea)

Descripción de los datos:

- Observaciones: 1.152.634
- Unidades espaciales (manzanas): 41.370
- Manzanas intervenidas: 222
- Planes parciales: 47
- Periodo: 2012 – 2025
- Años de adopción: múltiples cohortes (2013–2024)
- Variable dependiente: V_REF_Mz - valor de referencia del suelo por manzana
- Distribución de manzanas intervenidas por tipo de plan y grupo

Tabla VIII.

Distribución de manzanas intervenidas por tipo de plan y grupo.

Tipo_Plan	Control	Intervenido
Desarrollo	25.596	124
Renovación	15.552	98

Fuente: elaboración propia a partir de datos de IDECA

En la Fig. *7*Fig. *7*. Brecha anual en Log-Precio: Manzanas de influencia - Manzanas de control., se realiza una aproximación visual en términos del logaritmo del valor del suelo, con el fin de establecer la brecha entre manzanas tratadas y de control. Se realiza la transformación logarítmica con el fin de reducir el sesgo en el valor del suelo.

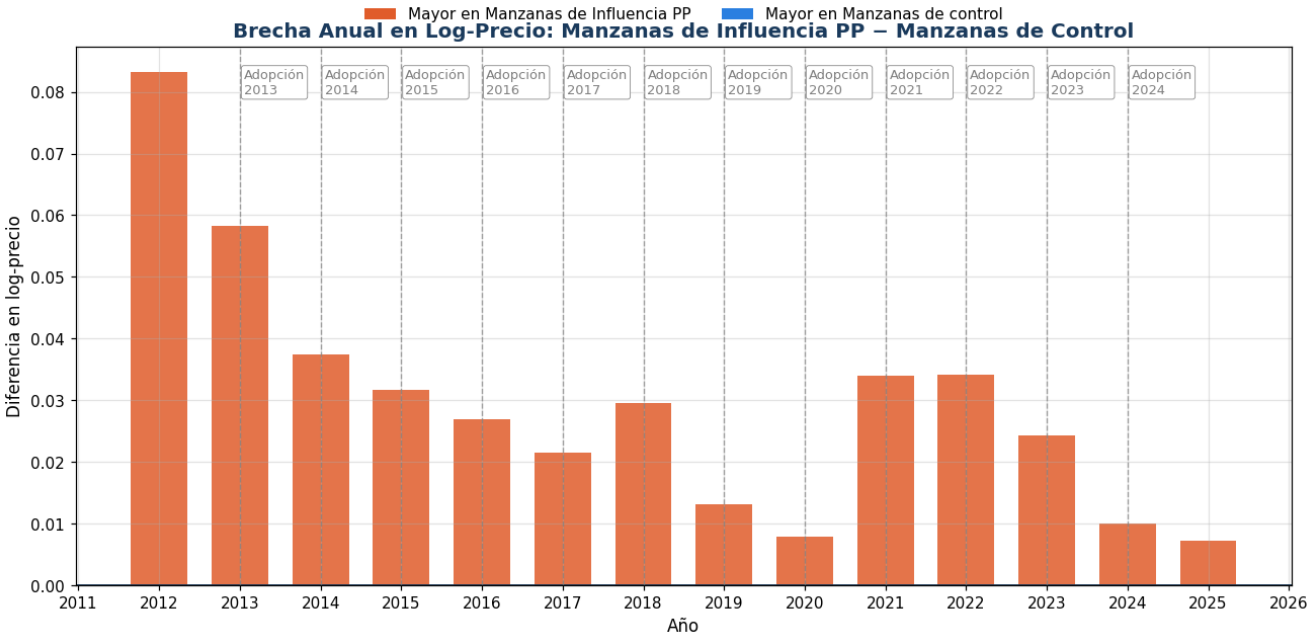


Fig. 7. *Brecha anual en Log-Precio: Manzanas de influencia - Manzanas de control.*
Fuente: elaboración propia a partir de datos de IDECA.

Se observa que el valor del suelo en las manzanas tratadas o de influencia es consistentemente mayor que en las de control. En términos porcentuales se interpreta que la diferencia de 0,01 corresponde a un 1% y así sucesivamente. En el año 2012, se presenta la mayor diferencia con un valor de 8,2% superior en las manzanas de influencia, mientras que en el año 2020 (año de pandemia) la diferencia es mínima. El carácter escalonado puede hacer referencia a la heterogeneidad de las manzanas tratadas, a la maduración de intervenciones realizadas en años anteriores o a la diferencia frente a la cantidad de intervenciones realizadas de un año a otro.

Validación del supuesto de Tendencias Paralelas

El supuesto principal del modelo de DiD indica que en ausencia de la intervención, manzanas tratadas y de control habrían seguido tendencias paralelas en los precios. Para validarlo empíricamente se estima un análisis de evento (event study), que permite verificar si los coeficientes pre-tratamiento son estadísticamente iguales a cero.

Tendencias Antes y Después de la Adopción

Al alinear todas las manzanas respecto a su año de adopción (t=0) se promedian las trayectorias para las manzanas de tratamiento y control. En la *Fig. 8*, se observa que en el análisis de evento el grupo de tratamiento esta siempre por encima del grupo de control. Previo a la adopción de los planes se observan diferencias en las pendientes y en la evolución de la brecha entre ambos grupos, lo que sugiere un cumplimiento imperfecto del supuesto de tendencias paralelas. Posterior a la adopción, ambas series continúan creciendo, con una ligera ampliación de la brecha a favor de las manzanas de influencia. No obstante, dado el comportamiento en el periodo pretratamiento, estos resultados no se pueden atribuir al efecto de la intervención.

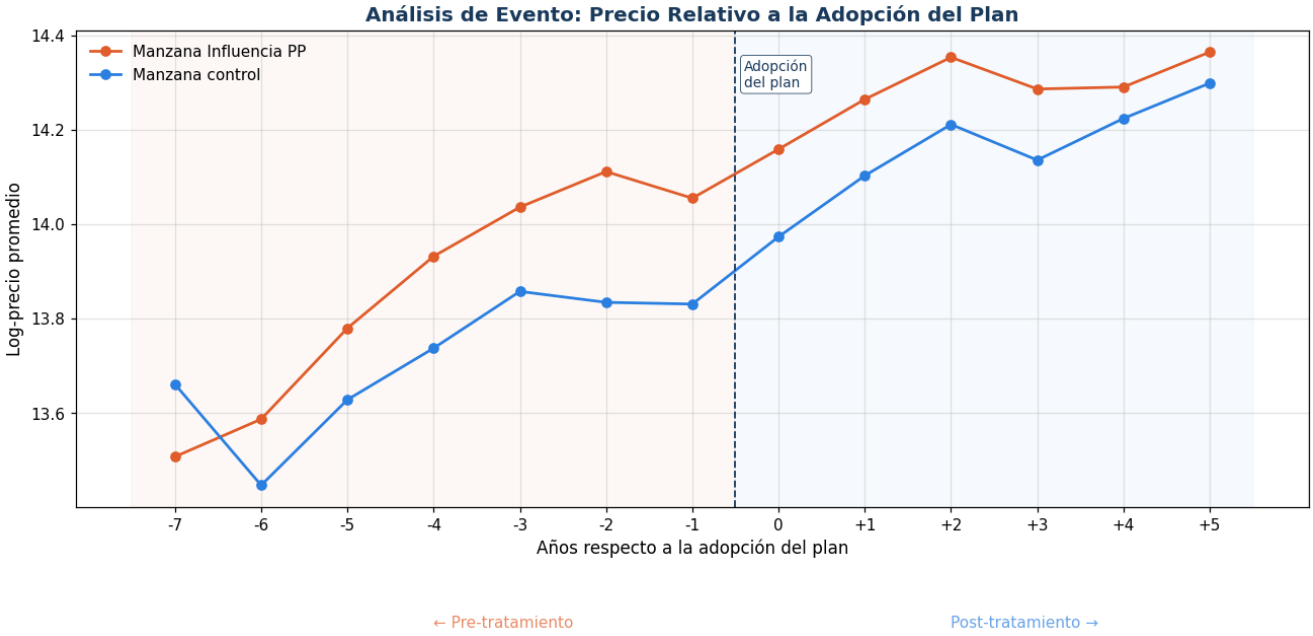


Fig. 8. Evolución del precio relativo del suelo antes y después de la adopción de planes parciales.
Fuente: elaboración propia a partir de datos de IDECA.

Estimación del análisis de evento (evento study)

$$log(P_{it}) = \alpha_i + \gamma_t + \sum_{k \neq -1} \beta_k \cdot 1(t - T_i = k) \cdot Tratado_i + \varepsilon_{it}$$

Dónde:

α_i → efectos fijos por manzana

γ_t → efectos fijos por año

k → años relativos a la adopción

Referencia: $k = -1$ → todo se interpreta respecto a ese año

β_k mide la diferencia en log-valor entre tratamiento y control en el periodo k, relativa al año inmediatamente anterior a la adopción.

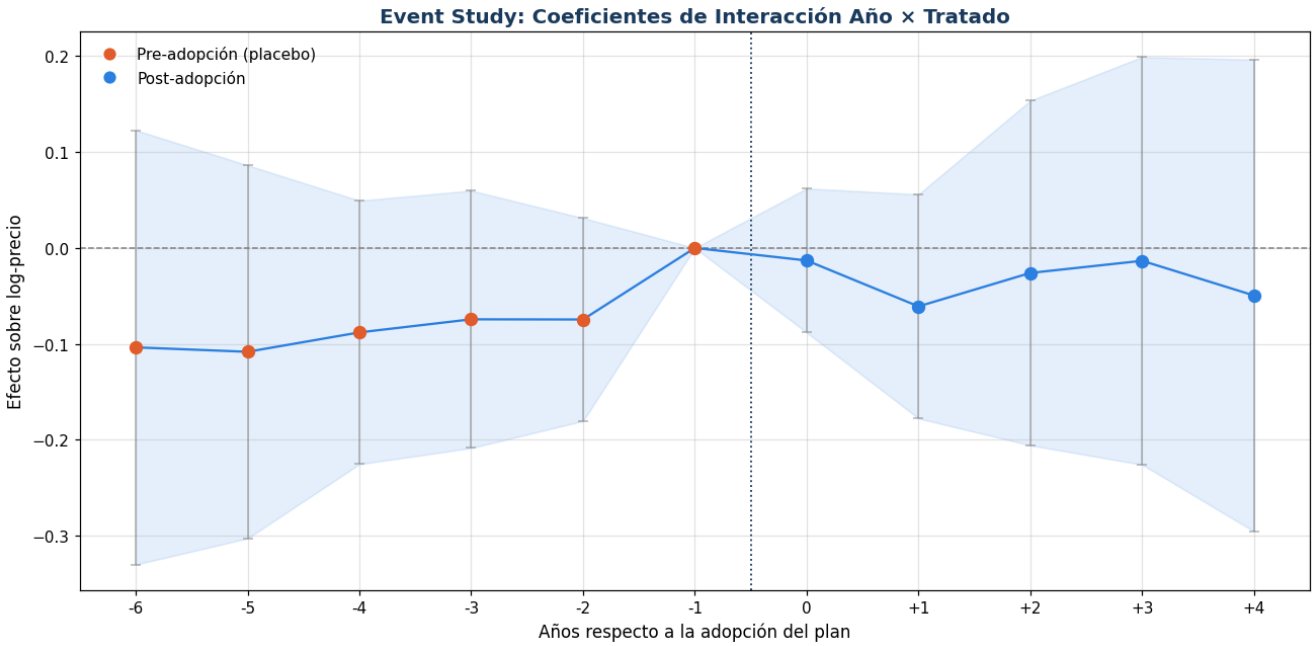


Fig. 9. Estimación del event study: coeficientes de interacción año × tratado
Fuente: elaboración propia a partir de datos de IDECA.

La Fig. 9 muestra que los coeficientes fluctúan alrededor del cero, es decir, el cero está dentro de los intervalos de confianza. Por tanto, el supuesto de tendencias paralelas se valida indicando que antes de cada plan, los precios de las manzanas tratadas y de control evolucionaban de forma similar. Por otra parte, los coeficientes de los planes en el periodo post-adopción, al ser cercanos a cero y con mayor variabilidad pueden indicar que el efecto de los planes parciales no se percibe posiblemente por sus extensos periodos de implementación.

Teniendo en cuenta que algunas manzanas pueden encontrarse en el área de influencia de múltiples planes que a su vez pudieron ser adoptados en diferentes periodos de tiempo, y que el efecto puede ser variable entre planes, las comparaciones pueden introducir pesos negativos en el promedio afectando su interpretación, obteniendo un efecto cercano a cero o incluso negativo, aunque el efecto real de cada plan sea positivo.

Estimación de modelos causales: Diferencias en Diferencias (DiD)

Coeficiente de interés:

$$\beta_{DiD} = \left(\bar{Y}_{tratado,post} - \bar{Y}_{tratado,pre}\right) - \left(\bar{Y}_{control,post} - \bar{Y}_{control,pre}\right)$$

Término de interacción que captura el cambio adicional en el precio de las manzanas intervenidas después de la adopción del plan, descontando tendencias generales del mercado inmobiliario. El modelo especificado es:

$$\log(V\_REF\_Mz_{it}) = \alpha + \beta_{DiD}(Tratado_i \times Post_{it}) + \gamma X_{it} + \mu_i + \lambda_t + \varepsilon_{it}$$

Dónde μ_i son efectos fijos por manzana y λ_t efectos fijos por año.

Modelo 1: OLS básico (modelo DiD sin efectos fijos)

$$\log(V\_REF\_Mz_{it}) = \alpha + \beta_{DiD}(Tratado_i \times Post_{it}) + \varepsilon_{it}$$

Estima el efecto promedio del tratamiento usando una especificación lineal básica.

Modelo 2: OLS con controles estructurales

$$\log(V\_REF\_Mz_{it}) = \alpha + \beta_{DiD}(Tratado_i \times Post_{it}) + \gamma X_{it} + \varepsilon_{it}$$

Dónde:

$$\boldsymbol{X_{it}} \begin{bmatrix} Tipo\_plan_i \\ \mathbf{log}(\text{Area\_Plan}_i + 1) \\ \text{grupo\_economico}_i \\ \text{Estrato}_i \\ \mathbf{log}(Distancia\_Plan_{it} + 1) \\ \mathbf{log}(\text{Area\_Mz}_i + 1) \end{bmatrix}$$

Incorpora variables que capturan características socioeconómicas, geométricas y de localización respecto al plan parcial.

Modelo 3: Efectos fijos de manzana (FE)

$$\log(V\_REF\_Mz_{it}) = \beta_{DiD}(Tratado_i \times Post_{it}) + \mu_i + \varepsilon_{it}$$

Introduce efectos fijos por manzana.

Modelo 4: Efectos fijos de doble vía (TWFE)

$$\log(V\_REF\_Mz_{it}) = \alpha + \beta_{DiD}(Tratado_i) + \mu_i + \lambda_t + \varepsilon_{it}$$

Corresponde al DiD estándar con panel que introduce efectos fijos por manzana y efectos fijos por año.

Modelo 5: TWFE con controles

$$\log(V\_REF\_Mz_{it}) = \alpha + \beta_{DiD}(Tratado_i \times Post_{it}) + \gamma X_{it} + \mu_i + \lambda_t + \varepsilon_{it}$$

Dónde:

$$\boldsymbol{X_{it}} \begin{bmatrix} \mathbf{log}(Distancia\_Plan_{it} + 1) \\ \mathbf{log}(\text{Area\_Mz}_i + 1) \end{bmatrix}$$

Mantiene la estructura TWEF agregando controles relevantes. Dado que el modelo incorpora efectos fijos por

manzana, todas las variables invariantes en el tiempo quedan absorbidas por dichos efectos, únicamente se incluyen aquellas que pueden presentar variación temporal.

Comparación de modelos

Tabla IX.
Comparación del estimador DiD en los modelos

Modelo	Coef. DiD	Err. Estándar	IC 95% Inf	IC 95% Sup	p-valor	Sig.	Impacto %	R^2
M1: OLS básico	0,0574	0,0322	-0,0057	0,1204	0,075	·	5,9	0,16
M2: OLS + controles	0,0526	0,0268	0,0000	0,1052	0,05	·	5,4	0,417
M3: FE manzana	0,0787	0,0595	-0,0378	0,1953	0,186		8,19	0,3848
M4: TWFE	0,0005	0,0496	-0,0968	0,0978	0,992		0,05	0,0
M5: TWFE + controles	0,0003	0,0496	-0,0969	0,0975	0,995		0,03	0,0

Notas:

· p < 0,1

Impacto % = (exp(β) – 1) x 100

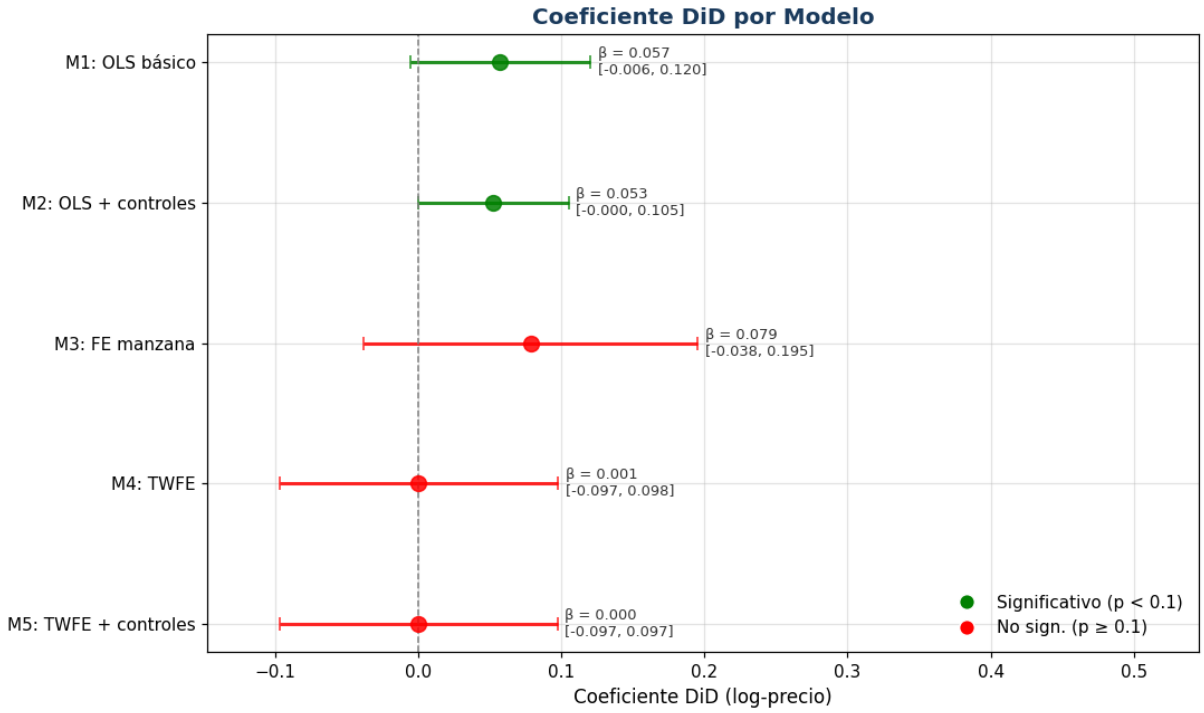


Fig. 10. Comparación visual del coeficiente DiD en los modelos con IC al 90%.
Fuente: elaboración propia a partir de datos de IDECA.

Hallazgos

Los resultados iniciales sugieren un efecto positivo de los planes parciales sobre el valor del suelo. Sin embargo, al controlar por heterogeneidad no observable mediante efectos fijos de manzana y efectos temporales, el

efecto desaparece completamente. Esto indica que las diferencias observadas en modelos iniciales responden a características estructurales preexistentes y no a un impacto causal atribuible a la implementación de los planes parciales.

El coeficiente β positivo, sugiere que en promedio las manzanas tratadas presentan un mayor aumento del valor del suelo posterior a la adopción del plan, respecto a las manzanas de control. No obstante, no corresponde a un efecto que genera exclusivamente el plan parcial, sino que se pueden estar capturando dinámicas de valorización preexistentes en las zonas intervenidas. (Ver Tabla IX y Fig. 10).

El modelo 4 al incorporar efectos fijos por manzana y año, presenta un β muy cercano a cero. Lo cual indica que no hay un cambio diferencial atribuible a la intervención y la valorización observada ya estaba ocurriendo. Esto se confirma con el modelo 5 dónde controles adicionales no cambian nada.

Los resultados evidencian que los planes parciales no se implementan aleatoriamente, se ubican en zonas con dinámicas específicas. Adicionalmente, tras la implementación de un plan no se presenta un salto estructural en el valor del precio, por el contrario, al parecer el mercado se anticipa al proceso de valorización manteniendo una continuidad en la tendencia de aumento del valor del suelo en el tiempo.

Igualmente, el impacto depende del estado del territorio al momento de la adopción del plan parcial. Los planes parciales de Renovación Urbana se suelen implementar sobre zonas consolidadas, mientras que los planes parciales de Desarrollo se implementan en zonas con poco o nulo desarrollo urbanístico y por ende menos valorizadas.

El detalle sobre el desarrollo del análisis de DiD se encuentra en el siguiente anexo

Anexo H. Analisis_de_Impacto

Adicionalmente, el procesamiento se encuentra en el siguiente script de Google Colab:

Anexo I. Analisis_impacto_planes_parciales

6. MODELADO

6.1. Selección de modelos de Aprendizaje Automático

Para abordar el Objetivo 2 que consiste en el desarrollo de modelos de aprendizaje automático (Machine Learning) para predecir el valor del suelo teniendo en cuenta el efecto de la adopción de los planes parciales, se consideró que este análisis tiene diferentes puntos de vista entre ellos el estadístico, espacial, económico y de ciencia de datos aplicada a la variable objetivo.

El propósito de los modelos es predecir el valor del suelo del metro cuadrado por manzana en Bogotá, incorporando variables físicas, espaciales, socioeconómicas y temporales, asociadas a la adopción de planes

parciales. La estrategia metodológica busca captar dos tipos de efecto: efecto directo, sobre las manzanas donde se localizan los planes parciales y el efecto indirecto (spillover), sobre las manzanas colindantes o cercanas, que pueden verse afectadas por cambios en accesibilidad, expectativas del mercado, transformación urbana y dinámica inmobiliaria

En este sentido la elección de los modelos debe permitir el contraste entre la interpretabilidad y la capacidad predictiva, porque no se trata únicamente de predecir el valor del suelo, sino de establecer relaciones e interacciones con las variables que pueden influir en su valor, incluyendo la adopción de planes parciales.

Por tal razón, se identificó que los modelos a desarrollar debían cumplir con los criterios de capacidad predictiva, explicativa y de robustez frente a problemas de colinealidad, relaciones complejas y ruido en los datos. Por ello se seleccionaron los siguientes modelos de aprendizaje supervisado:

- **Regresión Lineal Múltiple:** es el modelo básico que permite cuantificar la influencia de las variables independientes respecto al valor del metro cuadrado por manzana, bajo el supuesto de que las relaciones entre estas variables son lineales. Se seleccionó porque es un modelo que tiene alta interpretabilidad, sin embargo, se debe considerar que podrían existir relaciones no lineales que este modelo no logrará capturar.

- **Random Forest:** es un modelo versátil que captura las relaciones no lineales entre variables sin especificarlas, además evalúa múltiples relaciones y combinaciones entre variables eligiendo las que más explican la variación del precio y su ventaja radica en ser robusto frente a los problemas de colinealidad que normalmente pueden tener estas variables y también en su capacidad de trabajar con variables categóricas codificadas y numéricas. Si bien es un modelo de mayor complejidad, no pierde capacidad de interpretación.

- **XGBoost:** al igual que el modelo random forest, permite capturar relaciones complejas y no lineales, pero también discontinuas, además es sensible a cambios marginales y tiene un buen desempeño con variables categóricas codificadas y numéricas.

- **Máquinas de Soporte Vectorial:** son útiles en este contexto porque pueden modelar relaciones no lineales sin sobreajustar los modelos y se adaptan a datasets con alta dimensionalidad. No obstante, su capacidad interpretativa es menor que en los otros modelos.

6.2. Entendimiento de los datos

Previo a la implementación de los modelos predictivos, se realizó un proceso de entendimiento de los datos, con el fin de analizar la estructura del dataset previamente explorado, analizado y depurado, e identificar la naturaleza de las variables y sus posibles relaciones.

Inicialmente, se llevó a cabo una revisión general de la base de datos, identificando el número de registros, atributos y tipos de datos asociados a cada variable. Este proceso permitió verificar que el trabajo realizado en

el análisis exploratorio haya sido exitoso, viendo la clasificación de las variables en numéricas, categóricas y temporales, consistencia y completitud.

Posteriormente, se realizó un análisis descriptivo de las variables numéricas mediante medidas de tendencia central y dispersión, y de distribución y frecuencia para las variables categóricas, lo cual permitió identificar la distribución de los datos y la presencia de valores atípicos. De manera complementaria, se emplearon visualizaciones como diagramas de caja (boxplots) para evaluar la variabilidad y detectar posibles outliers en variables relevantes como el valor del suelo (V_REF_Mz) y las variables explicativas del modelo.

Adicionalmente, se llevó a cabo un análisis de correlación entre variables numéricas, con el objetivo de identificar relaciones lineales y posibles problemas de multicolinealidad.

Finalmente, este proceso permitió comprender la estructura y características del dataset, constituyéndose en la base para la definición de transformaciones, selección de variables y construcción de los modelos predictivos, garantizando la coherencia entre los datos y los objetivos del análisis.

1. Revisión General

La base de datos final está compuesta por 1.152.634 observaciones y 25 atributos, lo cual permitió contar con un volumen suficiente de información para el desarrollo y validación de los modelos de aprendizaje supervisado.

Se verificó que no existen valores nulos en ninguna de las variables, dado que cada una presenta el mismo número de registros (1.152.634 non-null), lo cual confirma la efectividad de los procesos previos de limpieza del dataset. Adicionalmente, se observa que, aunque los tipos de datos se encuentran correctamente definidos a nivel técnico, es necesario realizar una adecuación conceptual de las variables previo al modelado, ejemplo variable Estrato.

Esta estructura permite una correcta manipulación de la información y facilita su posterior uso en los modelos predictivos. En general, la base de datos presenta condiciones adecuadas de completitud, consistencia y estructura, lo que garantiza su idoneidad para el desarrollo de modelos de aprendizaje automático.

2. Análisis Descriptivo de Variables

A partir del análisis de medidas de tendencia central y dispersión, se evidencian varios aspectos relevantes del comportamiento de los datos.

La variable dependiente V_REF_Mz presenta alta dispersión, con valores que oscilan entre aproximadamente $5.000 y $29.900.000 por m², con una media de $1.501.500 pesos y una mediana de $1.200.000 pesos, lo que evidencia una distribución asimétrica hacia la derecha, asociada a la presencia de valores altos en ciertas zonas del territorio. Este comportamiento es consistente con la heterogeneidad del mercado del suelo en Bogotá, donde coexisten zonas con valores muy bajos y otras con valores significativamente elevados.

La variable Conteo_Predios también presenta alta variabilidad, donde se observan manzanas con un solo predio y otras con más de 700 predios, lo que indica diferencias importantes en la densidad predial entre sectores.

En general, la presencia de valores extremos y la amplitud en los rangos de varias variables sugieren la necesidad de aplicar transformaciones (como las logarítmicas), para mejorar la estabilidad y el desempeño de los modelos predictivos.

3. Análisis de correlación entre variables numéricas

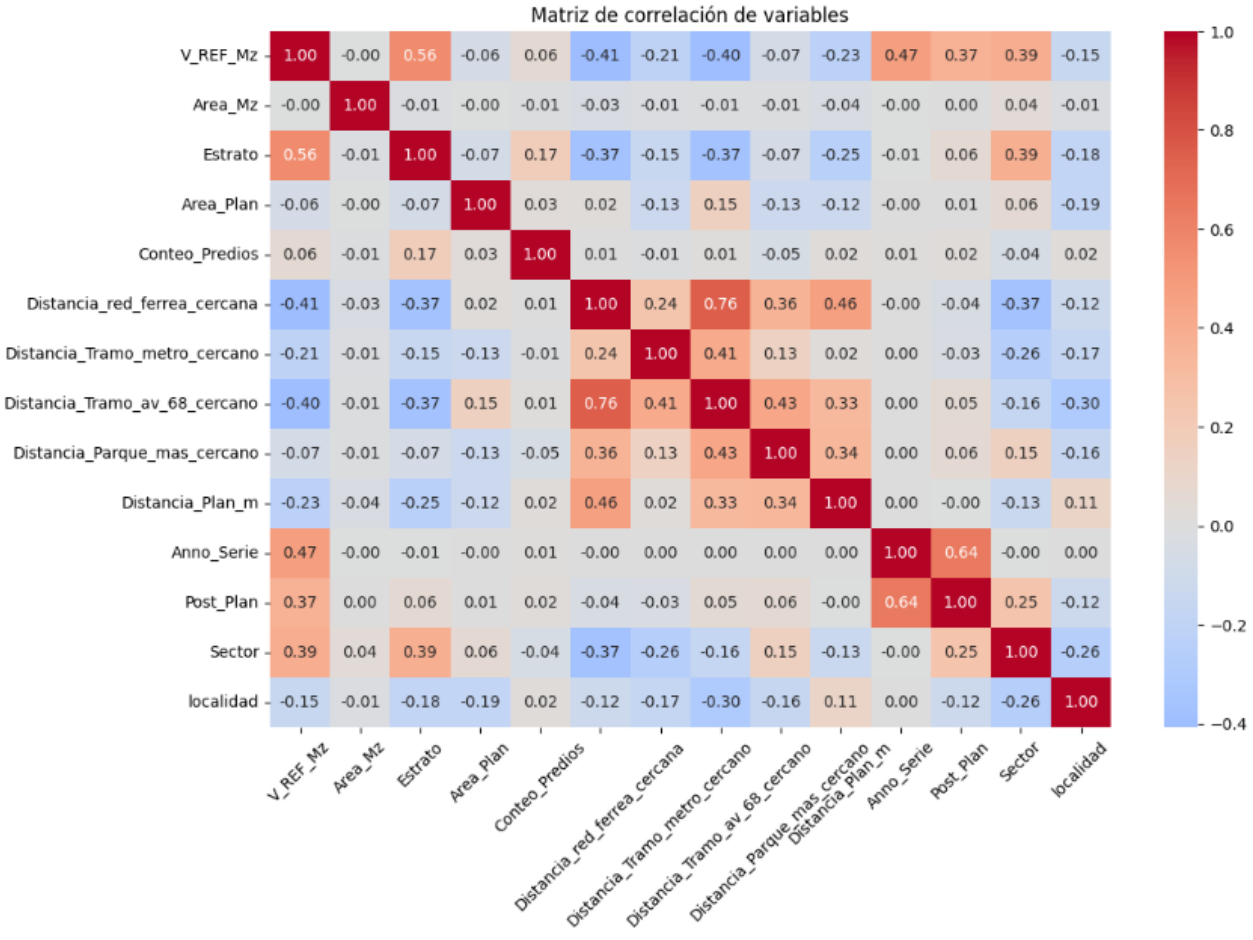


Fig. 11. *Matriz de correlación.*
Fuente: elaboración propia a partir de datos de IDECA.

En la matriz de correlación (Ver Fig. 11) se puede observar que la variable estrato socioeconómico presenta la mayor asociación positiva con el valor del suelo (0.56), lo cual confirma su relevancia como determinante en el valor del suelo en Bogotá. Asimismo, se observa una correlación positiva con la variable temporal Anno_serie (0.47), lo cual es coherente, ya que el valor del suelo tiende a crecer en el tiempo.

En contraste, las variables de distancia a infraestructuras urbanas presentan correlaciones negativas moderadas,

destacándose la distancia a la red férrea (-0.41) y a la Avenida 68 (-0.40), lo cual muestra que a mayor distancia a obras de infraestructura importantes que permiten accesibilidad a las manzanas su valor disminuye.

Adicionalmente, se identifican correlaciones relativamente altas entre algunas variables explicativas, particularmente entre diferentes medidas de distancia, lo que sugiere posibles problemas de multicolinealidad. Para lo cual en algunos modelos se debe realizar una cuidadosa selección de variables, para evitar redundancias.

Finalmente, variables como el área de la manzana y el número de predios presentan correlaciones cercanas a cero, lo que sugiere una baja relación lineal directa con el valor del suelo, aunque podrían aportar valor en modelos no lineales.

La presencia de multicolinealidad entre variables explicativas es especialmente relevante en modelos como la regresión lineal, dado que estos son sensibles a la redundancia y a la escala de las variables. En contraste, modelos basados en árboles, como Random Forest y XGBoost, son más robustos frente a variables altamente correlacionadas, debido a su estructura de partición jerárquica, lo que reduce la necesidad de procesos estrictos de selección de variables [42].

6.3. Diseños de prueba

En este numeral se describe el diseño general de prueba utilizado para el desarrollo de los modelos predictivos implementados en el presente estudio.

Para ello, se definió una estrategia común de partición de los datos en conjuntos de entrenamiento y validación, con el fin de evaluar la capacidad predictiva de los modelos sobre información no utilizada durante su ajuste. Considerando la naturaleza temporal de la información, se optó por un esquema de partición basado en el tiempo (split temporal), con el propósito de preservar la estructura de la serie, evitar problemas de fuga de información (data leakage) y garantizar una evaluación más realista del desempeño predictivo. En este sentido, se ordenó el conjunto de datos cronológicamente y se estableció un conjunto de entrenamiento compuesto por las observaciones correspondientes al periodo 2012–2022, y un conjunto de validación con los datos del periodo 2023–2025, correspondiente a una proporción aproximada del 80% y 20%, respectivamente.

Teniendo en cuenta lo anterior, se implementaron y evaluaron las cuatro técnicas de modelaje predictivo seleccionadas: Regresión Lineal Múltiple, Random Forest, Máquinas de Soporte Vectorial con enfoque de Regresión (SVR) y XGBoost, que comprenden diferentes enfoques, desde modelos lineales tradicionales hasta métodos de aprendizaje automático más avanzados.

Para la evaluación del desempeño del modelo, se utilizaron métricas como el coeficiente de determinación (R^2), el error cuadrático medio (RMSE) y el error medio absoluto (MAE). Estas métricas de evaluación se mantuvieron consistentes entre modelos, con el fin de garantizar la comparabilidad de los resultados.

En cuanto a la Preparación de los datos, a continuación, se indica la secuencia de actividades realizadas en general, con el fin de optimizar el desempeño de los modelos:

1. Se realizó la transformación de variables categóricas a formato binario, incluyendo las variables *ClaseSuelo*, *Estrato*, *Tipo_Plan* y *grupo_economico*. Para ello se empleó la técnica de One-hot encoding, la cual consiste en transformar cada categoría en una variable binaria independiente. Esto permite representar información cualitativa en forma numérica sin imponer un orden entre categorías, evitando sesgos en la estimación del modelo, para facilitar su correcta interpretación. [43]

   Dependiendo del modelo implementado, se utilizó la opción drop_first = True o False, con el fin de evitar problemas de multicolinealidad en modelos lineales o conservar la totalidad de la información en modelos basados en aprendizaje automático.

2. Posteriormente, se aplicó una transformación logarítmica a las variables con distribuciones asimétricas y presencia de valores extremos, tales como valor de referencia, distancias, áreas y conteo de predios. Esto permitió reducir la asimetría de las distribuciones y mitigar la influencia de valores extremos, obteniendo variables más estables y adecuadas para el modelado estadístico. En la Fig. 12. Como ejemplo se presenta transformación logarítmica aplicada a la variable a predecir V_REF_Mz.

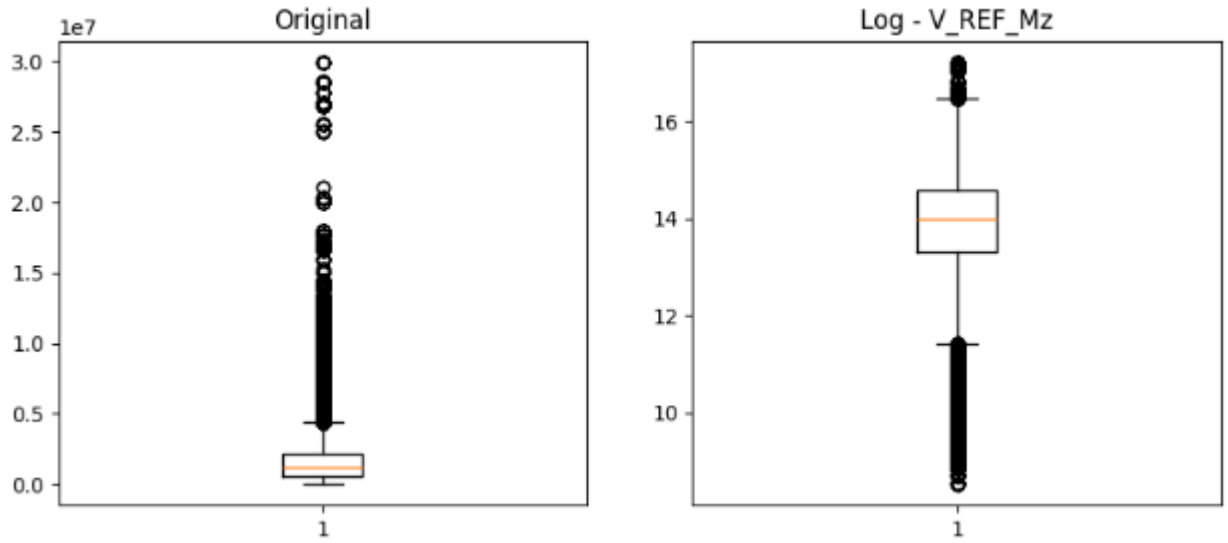


Fig. 12. *Ejemplo de varible V_REF_Mz con transformación logarítmica.*
Fuente: elaboración propia a partir de datos de IDECA.

3. Posteriormente, en aquellos modelos que lo requerían por ser técnicas sensibles a la escala de los datos, se realizó la normalización de las variables numéricas independientes, con el fin de garantizar que todas las variables se encuentren en una escala comparable y mejorar la estabilidad numérica de los modelos. En este proceso se excluyeron las variables dummy, la variable objetivo y los identificadores. La implementación específica de este paso se detalla en las subsecciones correspondientes [44].

4. Finalmente, se realizó la selección de las manzanas que tienen influencia de planes parciales, clasificándolas de la siguiente manera:

- **Manzana Influencia PP:** Manzanas que tienen influencia del plan parcial, es decir que se localizan a menos de 500 metros del Plan parcial.
- **Manzana PP:** Manzanas en la cual se localiza el plan parcial.

5. Creación de variables auxiliares

La dinámica temporal del mercado del suelo es capturada con la variable "Anno_Serie", la cual permite incorporar la tendencia general de los precios a lo largo del tiempo.

Adicionalmente, se creó la variable dummy de intervención "Post_Plan", que permite diferenciar los periodos anteriores y posteriores a la implementación del plan parcial.

Finalmente, se definió la variable "Tratamiento" como el producto entre la variable "Intervenido" y "Post_Plan", de tal forma que toma el valor de 1 para aquellas manzanas con influencia del plan parcial en el periodo posterior a su adopción, y 0 en caso contrario. Esta variable permite capturar el efecto diferencial asociado a la implementación del plan parcial sobre el valor del suelo.

Para algunos modelos fue necesaria la creación de variables adicionales, las cuales se describen en las secciones correspondientes.

6. Selección de variables

A continuación, en la Tabla X se describen de manera general las variables explicativas consideradas en los modelos. Estas variables podrán ser depuradas, transformadas o complementadas según los requerimientos específicos de cada modelo. Para facilitar su interpretación, se agruparon en las siguientes categorías de variables: accesibilidad, físicas, temporales e intervención, socioeconómicas y de uso del suelo.

Tabla X.

Selección general de variables para entrenamiento de los modelos.

Categoría	Variable	Descripción
Variables de accesibilidad (log)	Distancia_Plan_m_log	Distancia al plan parcial más cercano (transformada en log)
	Distancia_Parque_mas_cercano_log	Distancia al parque más cercano (log)
	Distancia_Tramo_metro_cercano_log	Distancia al tramo de metro más cercano (log)
	Distancia_Tramo_av_68_cercano_log	Distancia al tramo de metro más cercano (log)
	Distancia_red_ferrea_cercana_log	Distancia al tramo de metro más cercano (log)
	Ranking_cercania_Plan	Ranking_cercania_Plan
Variables físicas (log)	Area_Mz_log	Área de la manzana (log)
	Area_Plan_log	Área del plan parcial (log)
	Conteo_Predios_log	Número de predios por manzana (log)
Variables temporales e intervención	Anno_Serie	Año de la serie (tendencia del mercado)
	Anno_adopcion_plan	Año en que se adoptó el Plan Parcial
	Post_Plan	Variable dummy que indica periodo posterior al plan
	Intervenido	Variable dummy que indica si la manzana está intervenida

Categoría	Variable	Descripción
	Tratamiento	Interacción entre Intervenido y Post_Plan (efecto del plan)
Variables Socioeconómicas y de uso del Suelo	Estrato_1 a Estrato_6	Variables dummy del estrato socioeconómico (una categoría base excluida)
	ClaseSuelo_R	Suelo rural
	ClaseSuelo_U	Suelo urbano
	grupo_economico_comercio_y_oficinas	Uso comercial y de oficinas
	grupo_economico_depositos_y_parqueaderos	Uso de depósitos y parqueaderos
	grupo_economico_dotacional	Uso dotacional
	grupo_economico_industrial	Uso industrial
	grupo_economico_otros	Otros usos
	grupo_economico_residencial	Uso residencial
	Grupo_economico_bodegas	Bodegas
	Tipo_Plan_Renovacion	Plan Parcial de Renovación
	Tipo_Plan_Desarrollo	Plan Parcial de Desarrollo
	Sector	División territorial escala Intermedia
	Localidad	División territorial escala Mayor

Las particularidades en la configuración y evaluación de cada modelo se presentan en las subsecciones siguientes.

6.4. Construir los modelos seleccionados.

6.4.1. Regresión Lineal Múltiple

Para la estimación del modelo, se aplicó el proceso de normalización de las variables numéricas independientes descrito en la sección de Diseños de prueba, mediante el uso de la función *StandardScaler* de la librería *scikit-learn*, la cual transforma los datos para que presenten media cero y desviación estándar unitaria.

En este modelo se incorporó el criterio de influencia espacial definido previamente, considerando tanto las manzanas directamente intervenidas como aquellas ubicadas dentro del área de influencia del plan parcial (500 metros). En una etapa inicial, este criterio se aplicó mediante el filtrado del conjunto de datos, con el fin de delimitar el área de análisis.

Posteriormente, esta lógica fue incorporada dentro del modelo a través de variables específicas, tales como Intervenido y Tratamiento, las cuales permiten capturar el efecto directo de la intervención, así como su dimensión temporal. De esta manera, el modelo no solo consideró la delimitación espacial inicial, sino que también integró explícitamente estas dinámicas dentro de su estructura explicativa.

En este sentido, para capturar de manera integral los principales determinantes del valor del suelo, el conjunto de variables explicativas del modelo se conformó por variables de dimensiones espaciales, físicas, socioeconómicas y temporales como: Anno_Serie, Post_Plan, Intervenido, Tratamiento estrato, grupo económico, sector y localidad.

A partir del análisis de la matriz de correlación, se evidenció la presencia de alta correlación entre algunas variables, particularmente en las variables de distancia. En este contexto, se decidió implementar un proceso de selección automática de variables con el fin de evitar problemas de multicolinealidad, reducir el ruido y prevenir el sobreajuste de algunos modelos.

En particular, para el modelo de regresión lineal, se empleó el modelo de regularización Lasso (Least Absolute Shrinkage and Selection Operator), el cual incorpora un término de penalización que minimiza algunos coeficientes a cero, lo que permite identificar las variables que mejor explican el modelo. Este enfoque resulta especialmente útil cuando existen gran cantidad de variables y presencia de multicolinealidad, como en el análisis del valor del suelo [45].

En este sentido, aunque tanto la regresión lineal múltiple como el modelo Lasso pertenecen a la familia de modelos lineales, este último se emplea como una herramienta complementaria para la selección de variables.

Tabla XI.
Coeficientes estimados del modelo Lasso.

Variable	Coeficiente
Anno_Serie	0,4591
Estrato_3	0,2822
Estrato_4	0,1715
Estrato_2	0,1541
Estrato_6	0,1513
Sector	0,1312
Estrato_5	0,1206
Conteo_Predios_log	0,0769
ClaseSuelo_U	0,0249
grupo_economico_comercio_y_oficinas	0,0075
grupo_economico_depositos_y_parqueaderos	-0,0028
Anno_adopcion_plan	-0,0092
Area_Mz_log	-0,0157
ClaseSuelo_R	-0,0223
Distancia_Plan_m_log	-0,0265
localidad	-0,0481
Distancia_Tramo_metro_cercano	-0,057
Distancia_red_ferrea_cercana	-0,0784
Estrato_1	-0,1058
grupo_economico_otros	-0,1418
Distancia_Tramo_av_68_cercano	-0,1429

Fuente: elaboración propia a partir de datos de IDECA.

Los resultados del modelo Lasso reflejan que las variables que mejor explican valor del suelo corresponden principalmente a la dimensión temporal, la estratificación socioeconómica y la localización espacial. En particular, la variable “Anno_Serie” presenta el mayor coeficiente positivo, lo que refleja que el mercado del suelo tiende a crecer a lo largo del tiempo. Asimismo, las variables asociadas al estrato muestran una relación positiva con el valor del suelo, indicando que estratos más altos se asocian con mayores valores. (Ver Tabla XI).

En este sentido, aunque tanto la regresión lineal múltiple como el modelo Lasso pertenecen a la familia de modelos lineales, este último en este estudio se emplea como una herramienta complementaria para la selección de variables.

Como resultado de lo anterior, las variables de distancia presentaron coeficientes negativos, lo que sugiere que la accesibilidad a infraestructuras urbanas es un determinante importante del valor del suelo. En relación con las variables “Post_Plan” y “Tratamiento”, estas no fueron seleccionadas por el modelo Lasso, lo cual puede deberse a su alta correlación con la variable temporal. No obstante, dichas variables se mantienen conceptualmente relevantes, ya que permiten capturar el efecto diferencial asociado a la implementación de los planes parciales bajo un enfoque de diferencias en diferencias desarrollado en la sección correspondiente al análisis de impacto. En cuanto a la selección de variables, no se incluyeron las siguientes: “Distancia_Tramo_av_68_cercano_log” y “Distancia_red_ferrea_cercana_log”, debido a la alta correlación que presentaban con otras variables de accesibilidad, así como a el bajo poder explicativo evidenciado en el modelo Lasso.

De acuerdo con lo establecido en la sección de Diseños de prueba, se realizó una partición temporal de los datos, utilizando las observaciones correspondientes al periodo 2012–2022 para el conjunto de entrenamiento y las del periodo 2023–2025 para el conjunto de prueba. Como resultado, se obtuvo un total de 82,537 registros, de los cuales 64,853 (78.57%) fueron destinados al entrenamiento del modelo y 17,684 (21.43%) al conjunto de prueba.

Una vez seleccionadas las variables y realizado la partición temporal de los datos se implementó un modelo de regresión lineal múltiple utilizando la función LinearRegression de la librería scikit-learn. Posteriormente, el modelo fue ajustado con el conjunto de entrenamiento, con el fin de estimar la relación entre las variables explicativas y la variable objetivo.

Posteriormente, se extrajeron los coeficientes estimados del modelo, incluido el intercepto, con el fin de analizar la influencia y dirección del efecto de cada variable sobre el valor del suelo. La Fig. 13 presenta los coeficientes estandarizados obtenidos en el modelo de regresión lineal múltiple, permitiendo identificar la importancia relativa de las variables dentro de la explicación del comportamiento del valor del suelo. En términos generales, los resultados evidencian que el valor del suelo se encuentra principalmente asociado a factores estructurales, como la estratificación socioeconómica, la dinámica temporal y las condiciones de localización, por encima de efectos específicos relacionados con la implementación de planes parciales.

Los resultados evidencian que las variables asociadas a la estratificación socioeconómica presentan los mayores coeficientes positivos, especialmente en los estratos altos (Estrato_4, Estrato_5 y Estrato_6), indicando una relación directa entre mayores niveles socioeconómicos y mayores valores del suelo. Asimismo, la variable Anno_Serie refleja una tendencia creciente del mercado del suelo a lo largo del tiempo.

En cuanto a las variables de intervención, Intervenido presenta un efecto positivo, mientras que Tratamiento registra un coeficiente negativo, sugiriendo que el efecto posterior a la adopción del plan parcial no necesariamente genera incrementos adicionales en el valor del suelo al controlar las demás variables del

modelo.

Las variables de accesibilidad presentan efectos mixtos. Distancia_Tramo_metro_cercano_log evidencia un efecto negativo, indicando que la proximidad a infraestructura de transporte se asocia con mayores valores del suelo, mientras que la cercanía al plan parcial presenta una influencia limitada.

Respecto a las características físicas, Conteo_Predios_log presenta un efecto positivo, mientras que Area_Mz_log y Area_Plan_log muestran efectos negativos, sugiriendo que mayores densidades prediales se reaccionan con mayores valores unitarios del suelo.

Finalmente, ClaseSuelo_U presenta un efecto positivo y ClaseSuelo_R un efecto negativo significativo, evidenciando diferencias entre suelo urbano y rural. Por su parte, las variables del grupo económico presentan en su mayoría coeficientes negativos frente a la categoría base utilizada en el modelo.

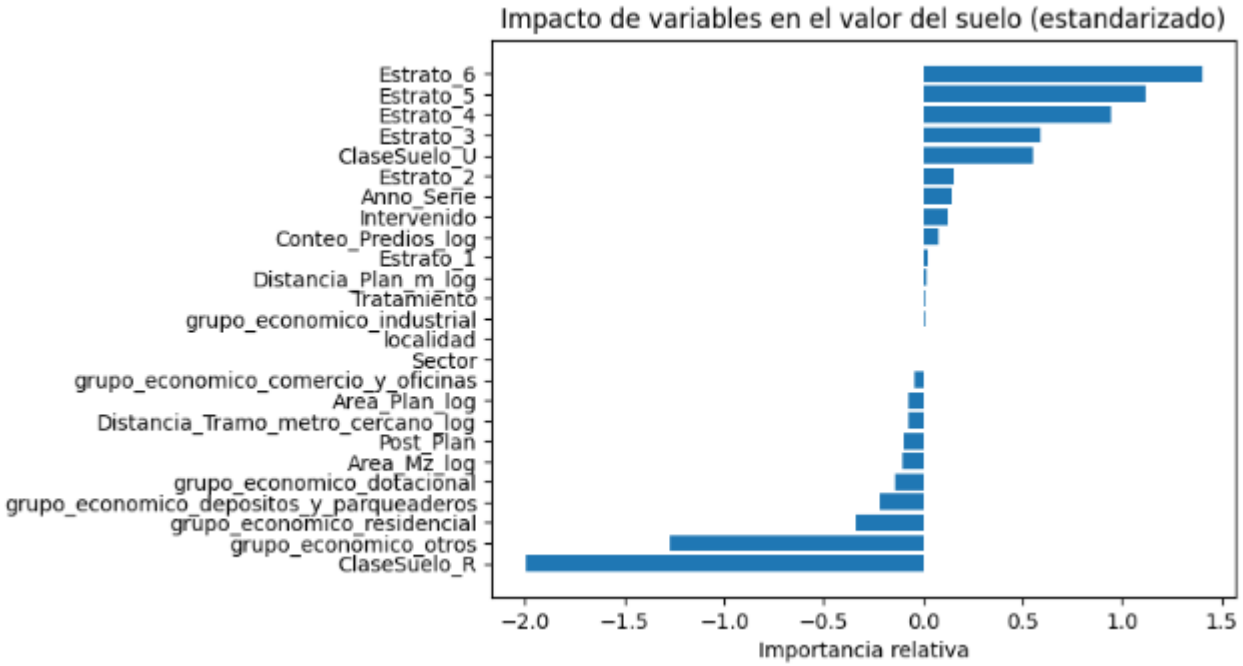


Fig. 13. *Impacto de las variables en el Valor de suelo. Regresión lineal múltiple*

Fuente: elaboración propia a partir de datos de IDECA.

Evaluación de Supuestos del Modelo

Una vez estimado el modelo, se realizaron predicciones sobre el conjunto de prueba con el fin de evaluar su capacidad predictiva. Para ello, se utilizaron las métricas estándar definidas: MSE, RMSE y R^2. (Ver Tabla XII)

Tabla XII.

Métricas de Evaluación del Modelo de Regresión Lineal Múltiple.

Métrica	Entrenamiento	Prueba	Interpretación
MSE	0,2986	0,3582	Error cuadrático medio del modelo en escala logarítmica.
RMSE	0,5464	0,5985	Magnitud promedio del error del modelo, penalizando en mayor medida los errores grandes.
MAE	0,3603	0,385	Error absoluto promedio entre los valores observados y predichos en escala logarítmica.
R^2	0,6199	0,2904	Proporción de la variabilidad del valor del suelo explicada por el modelo.
MAE aproximado (%)	43,38%	46,97%	Diferencia porcentual aproximada promedio entre los valores observados y predichos.
RMSE aproximado (%)	72,70%	81,94%	Evidencia la presencia de algunas predicciones con errores relativamente altos.

Fuente: elaboración propia a partir de datos de IDECA.

Los resultados de la regresión lineal múltiple muestran que el modelo tiene una capacidad explicativa moderada para predecir el valor del suelo en las manzanas de Bogotá. En prueba, el coeficiente de determinación obtenido (R^2 = 0,2904) indica que el modelo explica aproximadamente el 29,04 % de la variabilidad del valor del suelo. Asimismo, el MAE aproximado porcentual (46,97 %) evidencia que, en promedio, las predicciones presentan una diferencia cercana al 47 % respecto a los valores reales, mientras que el RMSE aproximado porcentual (81,94 %) sugiere la presencia de algunas predicciones con errores considerablemente altos. En conjunto, estos resultados indican que el modelo logra captar tendencias generales del comportamiento del mercado del suelo, aunque presenta limitaciones para representar completamente la complejidad espacial y económica del fenómeno analizado.

Este procesamiento se encuentra en el script de Google Colab:

Anexo J. Modelo_Regresion_Lineal_Multiple

6.4.2. XGBoost (Extreme Gradient Boosting)

Dado que las relaciones que explican el valor del suelo suelen ser no lineales y presentan heterogeneidad espacial, se empleó el algoritmo XGBoost (Extreme Gradient Boosting), que está basado en técnicas de boosting sobre árboles de decisión. Este enfoque permite capturar de manera eficiente interacciones complejas entre variables espaciales, temporales y urbanísticas, superando las limitaciones de los modelos lineales tradicionales.

XGBoost construye de manera iterativa un conjunto de árboles de decisión, donde cada nuevo árbol busca corregir los errores del modelo anterior mediante la optimización de una función de pérdida regularizada. Esta característica lo convierte en un modelo robusto frente al sobreajuste y altamente eficaz para modelar relaciones no lineales y patrones complejos en los datos [46]. Para el desarrollo de este modelo no se aplicaron procesos de normalización o estandarización a las variables numéricas, dado que es un algoritmo basado en

árboles de decisión que no es sensible a la escala de los datos, a diferencia de la regresión lineal o las SVM.

Preparación de variables espaciales y temporales

Como parte de la preparación específica para el modelo XGBoost, se incorporaron variables espaciales y temporales orientadas a capturar tanto el efecto directo de los planes parciales como posibles efectos indirectos o de influencia espacial sobre las áreas cercanas.

Además de las variables previamente definidas (Intervenido, Post_Plan y Tratamiento), se construyeron las variables Vecina y Vecina_Post, con el fin de representar las manzanas no intervenidas localizadas dentro del área de influencia de 500 metros y los posibles efectos spillover posteriores a la adopción del plan parcial.

Posteriormente, se realizó una validación de consistencia sobre las variables derivadas del tratamiento y de influencia espacial, verificando que Post_Plan, Tratamiento, Vecina y Vecina_Post cumplieran correctamente sus definiciones lógicas. No se identificaron inconsistencias ni traslapes entre las categorías analizadas.

Transformación y selección de variables

La variable objetivo V_REF_Mz fue transformada mediante la función log(1+x) con el fin de reducir la asimetría de la distribución y mejorar la estabilidad del modelo. De igual manera, se aplicaron transformaciones logarítmicas a variables continuas relacionadas con accesibilidad, tamaño y conteo predial.

La selección final de variables se fundamentó en criterios teóricos, espaciales y urbanos, agrupando variables de accesibilidad, intervención, influencia espacial, características físicas y variables socioeconómicas y de uso del suelo. En este contexto, se consideró que modelos basados en árboles, como XGBoost, pueden capturar relaciones no lineales e interacciones complejas entre variables, por lo que no se eliminaron variables únicamente con base en correlaciones bivariadas.

Asimismo, las variables categóricas fueron transformadas mediante One-Hot Encoding utilizando la opción drop_first=False, debido a que este tipo de modelos no presenta sensibilidad a problemas de multicolinealidad.

Construcción de matrices de modelado

Para el entrenamiento del modelo, se construyeron las matrices de variables explicativas (X) y variable objetivo (y), definiendo como variable objetivo V_REF_Mz_log. Posteriormente, se verificó que todas las variables fueran numéricas y aptas para el modelado. Finalmente, el conjunto de datos fue dividido en entrenamiento y prueba siguiendo el criterio temporal descrito anteriormente, preservando así la secuencia temporal de los datos y evitando fuga de información, como se explicó en la sección de Diseños de Prueba.

Definición del modelo base XGBoost

Se definió un modelo base XGBoost como configuración inicial para la estimación, buscando equilibrar capacidad predictiva, control del sobreajuste y eficiencia computacional. Posteriormente, se realizó un proceso de optimización de hiperparámetros con el fin de evaluar mejoras en el desempeño del modelo y fortalecer su

capacidad de generalización.

Entrenamiento del modelo

Una vez definido el modelo base, se procedió a su entrenamiento utilizando el conjunto de datos de entrenamiento previamente construido. Durante este proceso, el modelo aprende las relaciones entre las variables explicativas y la variable objetivo a partir de la minimización iterativa del error.

Optimización Hiperparámetros

Con el fin de mejorar el desempeño del modelo, se llevó a cabo un proceso de optimización de hiperparámetros mediante una búsqueda en grilla (Grid Search) (Ver Tabla XIII), que muestra los valores más representativos del modelo base y del modelo optimizado. Este procedimiento permite evaluar distintas combinaciones de parámetros con el objetivo de identificar la configuración que minimiza el error de predicción.

La búsqueda se realizó sobre un conjunto de parámetros clave del modelo XGBoost, incluyendo la profundidad máxima de los árboles (max_depth), la tasa de aprendizaje (learning_rate), el número de estimadores (n_estimators), así como los parámetros de muestreo (subsample y colsample_bytree).

Tabla XIII.
Comparación de hiperparámetros del modelo XGBoost base y optimizado.

Parámetro	***Modelo Base***	***Modelo Optimizado***
colsample_bytree	*0,80*	*1.00*
learning_rate	*0,05*	*0,10*
max_depth	*6*	*8*
min_child_weight	*5*	*No definido*
n_estimators	*300*	*300*
random_state	*42*	*No definido*
reg_alpha	*0,50*	*No definido*
reg_lambda	*1,00*	*No definido*
subsample	*0,80*	*0,80*

Fuente: elaboración propia.

Predicción

Una vez definidos los hiperparámetros del modelo, se realizaron predicciones sobre los conjuntos de entrenamiento y prueba con el propósito de evaluar la capacidad del modelo para estimar el valor del suelo tanto en datos conocidos como en observaciones no utilizadas durante el proceso de entrenamiento.

Debido a la naturaleza temporal de la información analizada, se implementó un esquema de validación temporal, evitando mezclar registros de años futuros dentro del entrenamiento. Este enfoque permite simular

un escenario más cercano a una aplicación real de predicción, para garantizar que el modelo sea evaluado sobre periodos posteriores a aquellos utilizados para su ajuste.

Para la optimización de hiperparámetros, el conjunto de entrenamiento fue dividido internamente en un subconjunto de entrenamiento y otro de validación, respetando el orden cronológico de los datos. Posteriormente, el modelo final fue reentrenado con la totalidad del conjunto de entrenamiento y evaluado sobre el conjunto de prueba.

Evaluación del desempeño

La evaluación del modelo se realizó sobre la variable objetivo log (V_REF_Mz), correspondiente a la transformación logarítmica del valor de referencia del suelo por manzana. Esta transformación permitió reducir la asimetría presente en la distribución de los valores y estabilizar la varianza de la variable dependiente.

Para medir el desempeño predictivo se utilizaron las métricas estándar de regresión definidas: MAE, RMSE y el R^2. Estas métricas fueron calculadas tanto para el conjunto de entrenamiento como para el conjunto de prueba, permitiendo evaluar la capacidad de ajuste y generalización del modelo. La Tabla XIV presenta la comparación de desempeño entre el modelo XGBoost base y el modelo optimizado en escala logarítmica.

Tabla XIV.
Comparación de métricas de desempeño entre el modelo XGBoost base y optimizado. Escala logarítmica

Métrica	Modelo Base	Modelo Optimizado
MAE Train	0,12466690	0,05174413
RMSE Train	0,19801124	0,07770608
R^2 Train	0,95008671	0,99213739
MAE Test	0,17709471	0,14467029
RMSE Test	0,25441669	0,25701203
R^2 Test	0,87179004	0,87065774

Fuente: elaboración propia a partir de datos de IDECA.

Los resultados evidencian que ambos modelos presentan un desempeño muy similar tanto en entrenamiento como en prueba. En el conjunto de prueba, el modelo alcanzó un coeficiente de determinación (R^2) cercano a 0,87, indicando que aproximadamente el 87% de la variabilidad observada en el valor del suelo en escala logarítmica puede ser explicada por las variables incluidas en el modelo.

Asimismo, los valores de MAE y RMSE obtenidos en prueba reflejan errores de predicción relativamente bajos considerando la complejidad espacial y temporal del fenómeno analizado. Esto sugiere que el modelo posee una adecuada capacidad de generalización y logra capturar de manera consistente los patrones asociados al comportamiento del valor del suelo en las manzanas de Bogotá.

En el conjunto de entrenamiento, el modelo optimizado presentó métricas ligeramente inferiores respecto al

modelo base; sin embargo, las diferencias observadas son mínimas y no representan cambios significativos en el desempeño global del modelo. De igual forma, las métricas obtenidas en prueba permanecieron prácticamente iguales entre ambas configuraciones, indicando que el proceso de optimización de hiperparámetros no produjo mejoras sustanciales en la capacidad predictiva final.

Estos resultados indican que la configuración inicial del modelo XGBoost ya ofrecía un desempeño robusto para el problema analizado y que la estructura de los datos, así como la calidad y capacidad explicativa de las variables utilizadas, tuvieron una influencia más relevante en el rendimiento del modelo que los ajustes adicionales de hiperparámetros.

Evaluación del modelo mediante transformación exponencial de métricas logarítmicas

Con el fin de facilitar la interpretación de los resultados obtenidos en escala logarítmica, se realizó una transformación exponencial de las métricas de error. La Tabla XV. presenta el MAE y RMSE expresados como porcentajes aproximados respecto a los valores observados.

En el conjunto de prueba, el modelo alcanzó un MAE aproximado de 19,37 % y un RMSE de 28,97 %, indicando un desempeño adecuado en la estimación del valor del suelo. Aunque se presentan mayores errores en algunas observaciones extremas, este comportamiento es coherente con la heterogeneidad y complejidad del mercado inmobiliario urbano. En entrenamiento, las métricas fueron ligeramente menores (MAE: 13,28 % y RMSE: 21,90 %), sin evidenciar diferencias excesivas frente al conjunto de prueba, lo que sugiere una adecuada capacidad de generalización.

Tabla XV.

Evaluación de Desempeño Modelo XGBoost mediante transformación aproximada a porcentaje desde escala logarítmica.

Validación transformación aproximada a porcentaje desde escala logarítmica
MAE aproximado en prueba (%): 19,37%
RMSE aproximado en prueba (%): 28,97%
MAE aproximado en entrenamiento (%): 13,28%
RMSE aproximado en entrenamiento (%): 21,90%

Fuente: elaboración propia a partir de datos de IDECA.

Validación: valores reales vs valores predichos

La *Fig. 14* compara los valores reales y predichos por el modelo XGBoost en escala logarítmica y real. En ambas representaciones, la cercanía de los puntos a la diagonal refleja un buen ajuste predictivo.

La Fig. 14(a) muestra que, en escala logarítmica, el modelo captura adecuadamente la tendencia general del valor del suelo, reduciendo parcialmente la dispersión de los datos. Por su parte, la Fig. 14(b), en escala real y excluyendo valores atípicos extremos, evidencia un mejor ajuste en los rangos centrales, aunque persisten diferencias en valores elevados. En general, los resultados confirman una adecuada capacidad predictiva del modelo para representar los patrones espaciales del valor del suelo en Bogotá.

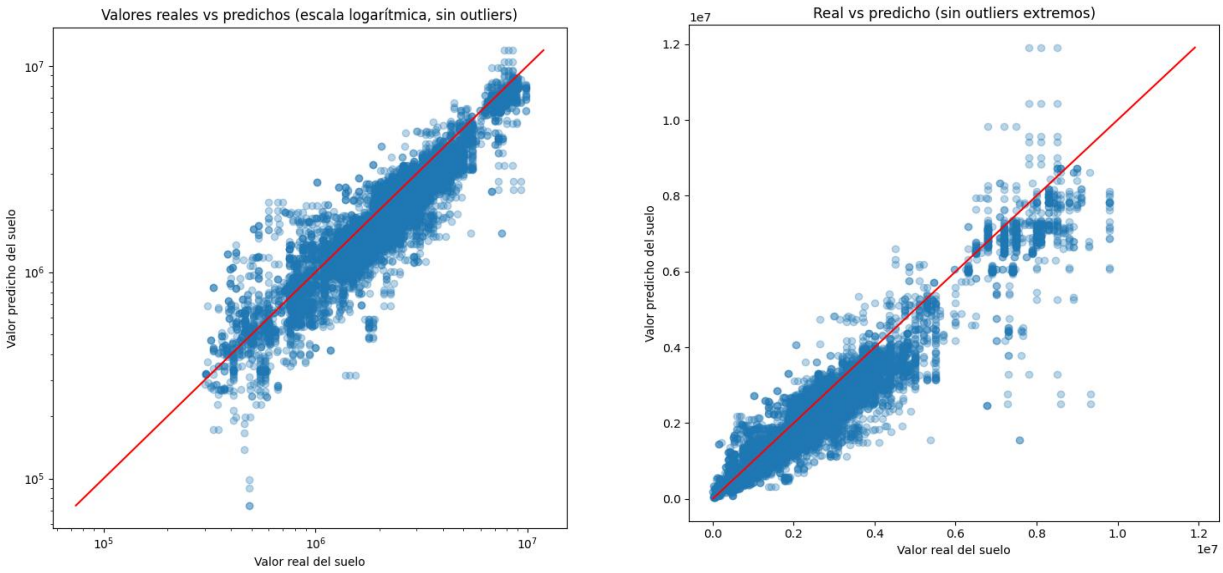


Fig. 14. *Comparación entre valores reales y predichos del valor del suelo mediante el modelo XGBoost: (a) escala logarítmica; (b) sin valores atípicos.*

Fuente: elaboración propia a partir de datos de IDECA.

Distribución del error

La *Fig. 15* presenta la distribución del error de predicción del modelo, calculado como la diferencia entre el valor real y el valor predicho, excluyendo valores extremos. La mayor concentración de errores se ubica alrededor de cero, lo que indica que el modelo no presenta un sesgo significativo y que la mayoría de las predicciones son cercanas a los valores observados.

No obstante, se observa una ligera asimetría hacia valores positivos, sugiriendo una moderada tendencia a la subestimación en algunas observaciones. Asimismo, la dispersión evidencia errores de mayor magnitud en ciertos casos, especialmente en valores altos del suelo, donde el mercado presenta mayor complejidad y variabilidad. En general, la distribución confirma un comportamiento estable y consistente del modelo.

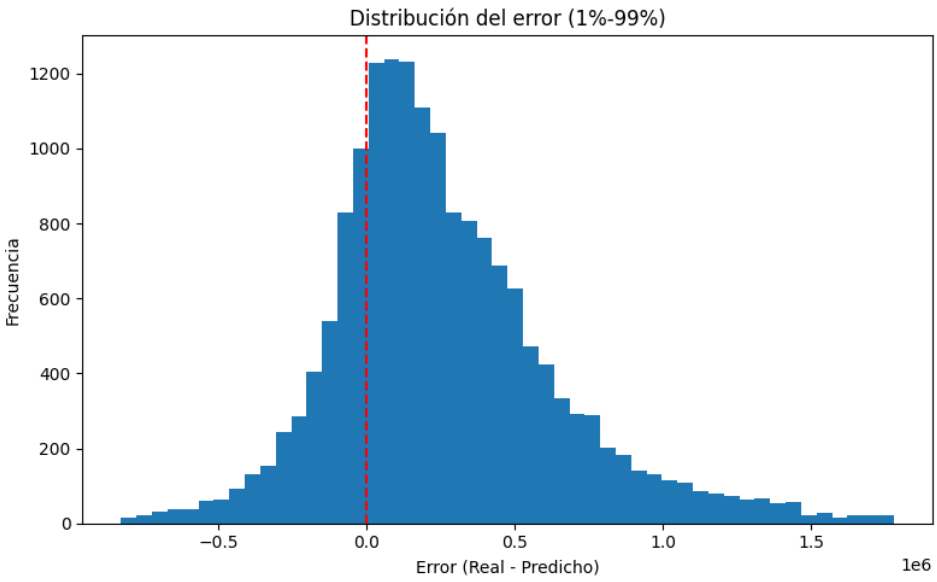


Fig. 15. *Distribución del error de predicción (valor real – valor predicho) del modelo XGBoost.*
Fuente: elaboración propia a partir de datos de IDECA.

Importancia de Variables

Con el fin de identificar las variables con mayor influencia en las predicciones del modelo, se calculó la importancia de variables generada por XGBoost. La *Fig. 16* presenta la relevancia relativa de las variables incluidas en el modelo.

Los resultados evidencian que las variables asociadas al tipo de plan parcial presentan una alta relevancia predictiva en el modelo, lo que sugiere que las características normativas y urbanísticas de los planes parciales influyen en el comportamiento del valor del suelo. Asimismo, variables socioeconómicas y temporales, como el estrato y el año de la serie (*Anno_Serie*), también muestran una contribución importante en la estimación del valor del suelo.

Por otra parte, variables de accesibilidad y localización, como las distancias a infraestructura vial y sistemas de transporte, presentan aportes moderados al modelo. En contraste, variables directamente relacionadas con la intervención, como *Intervenido*, *Vecina* y *Post_Plan*, presentan una importancia menor frente a variables estructurales y espaciales, lo que sugiere que los efectos de los planes parciales podrían estar siendo capturados indirectamente por otras variables urbanas, socioeconómicas y espaciales con mayor capacidad explicativa o que las características estructurales del plan importan más que su ejecución o intervención.

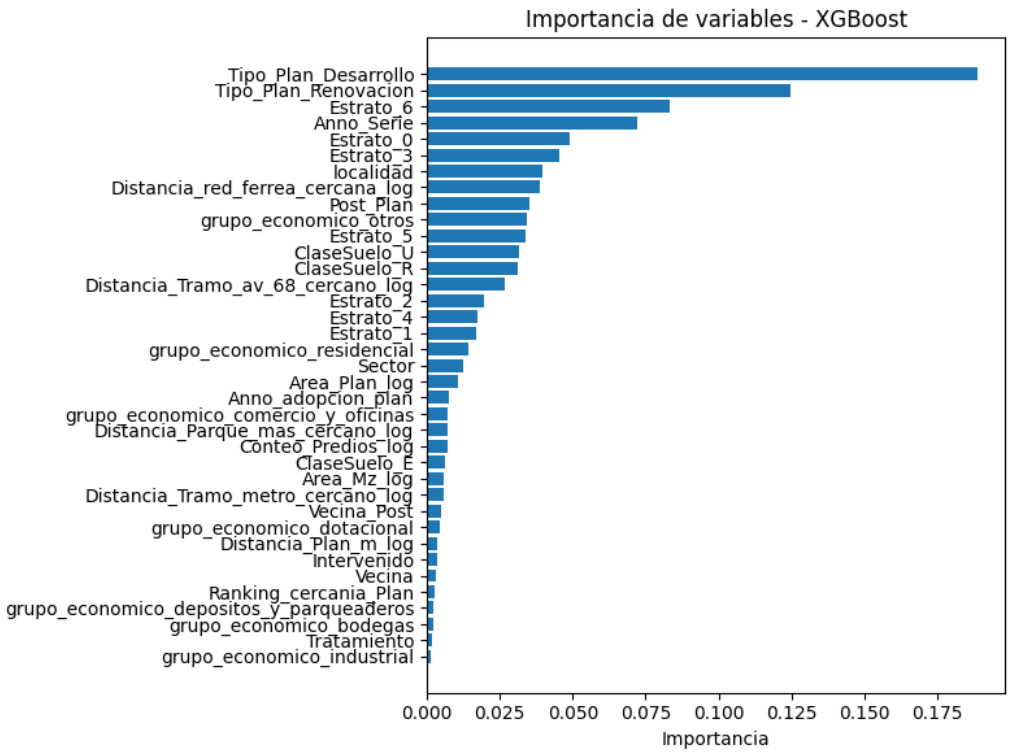


Fig. 16. *Importancia de Variables. Modelo XGBoost.*

Fuente: elaboración propia a partir de datos de IDECA.

Este procesamiento se encuentra en el script de Google Colab:

Anexo K. Modelo_XGBoost

6.4.3. Máquinas de Soporte Vectorial con enfoque de Regresión (SVR)

Teniendo en cuenta que las SVR son sensibles a la escala de las variables, se incorpora en el proceso de normalización descrito en el numeral 6.3 Diseños de prueba, la variable Anno_Serie, ya que al tener una escala diferente domina el cálculo y distorsiona el modelo.

Para el proceso de normalización se utiliza la función RobustScaler() que usa la mediana y el rango intercuartílico (IQR) en lugar de la media y la desviación estándar, generando una transformación más robusta frente a datos atípicos que persisten incluso después de la transformación logarítmica.

Con el fin de estimar el impacto del plan parcial sobre el valor del suelo, se incorpora una variable de interacción denominada Tratamiento, definida como el producto entre la variable de intervención espacial (Intervenido) y la variable temporal (Post_Plan). Esta variable permite identificar el efecto diferencial del plan parcial en las zonas intervenidas, controlando simultáneamente por tendencias generales del mercado (Anno_Serie) y por cambios temporales comunes a todas las observaciones (Post_Plan).

Al realizar el Split temporal descrito en el numeral “6.3 Diseños de prueba” se obtiene la siguiente distribución

de los datos:

- Entrenamiento: 64.853 registros
- Prueba: 17.684 registros

Optimización de hiperparámetros

Hiperparámetros del SVR: Núcleo (kernel), parámetro de penalización (C), ancho de la zona de tolerancia ε- (épsilon), coeficiente del núcleo (gamma).

Teniendo en cuenta el costo computacional requerido para la ejecución del modelo SVR y su correspondiente optimización de hiperparámetros, se implementó la siguiente estrategia:

1. Muestra estratificada: se seleccionó una muestra representativa del 25% que mantuvo la proporción en cada año sobre los datos de entrenamiento.

2. Búsqueda de hiperparámetros sobre muestra: Sobre la muestra estratificada se implementó la técnica RandomizedSearchCV que permite explorar el espacio de hiperparámetros mediante muestreo aleatorio. Se exploró en un rango amplio de C (de 1 hasta 1000) y de ε (de 0.001 hasta 0.5) en escala logarítmica. Se usó validación cruzada con 5 folds y 30 combinaciones aleatorias, en este caso se utilizó la función TimeSeriesSplit() que agrega datos progresivamente en cada fold, explora el espacio de hiperparámetros de forma aleatoria sobre distribuciones continuas en escala logarítmica para C y ε, respetando el orden temporal en cada fold.

3. Búsqueda fina: se refina el resultado obtenido anteriormente, con una segunda búsqueda de hiperparámetros en un rango estrecho sobre los mejores parámetros encontrados. Se usó validación cruzada con 5 folds y 20 combinaciones aleatorias.

Los mejores hiperparámetros encontrados fueron:

Tabla XVI.
Mejores hiperparámetros modelo SVR

C	62.03412970559334
Épsilon	0.003988699590700804
Gamma	scale
kernel	rbf

Fuente: elaboración propia a partir de datos de IDECA.

Dónde,

C – controla el balance entre ajuste y generalización.
Épsilon (ε) - Margen de tolerancia.
scale = 1/(n_features × var(X)). Controla el alcance de influencia de cada punto.
rbf (Radial Basis Function), adecuado para relaciones no lineales complejas.

Entrenamiento del modelo

Se realiza el entrenamiento con los mejores hiperparámetros obteniendo las siguientes métricas sobre el desempeño predictivo en el conjunto de prueba.

Tabla XVII.
Métricas de desempeño predictivo modelo SVR

	Escala log	Aproximado (%)
MSE	0,1147	
RMSE	0,3387	0,403
MAE	0,2168	0,242
R^2	0,7727	

Fuente: elaboración propia a partir de datos de IDECA.

El modelo explica el 77,27% de la variabilidad del logaritmo del valor del suelo. Dado que la variable dependiente fue transformada mediante logaritmo, las métricas de error se interpretan en términos relativos. El RMSE sugiere un error multiplicativo aproximado del 40%, mientras que el MAE indica un error promedio cercano al 24% en la predicción del valor del suelo.

La Fig. *17* presenta la relación entre los valores reales y los valores predichos por el modelo en escala logarítmica. La línea roja representa el ajuste perfecto, donde los valores predichos coinciden exactamente con los valores observados. Si bien la mayoría de los puntos se agrupan alrededor de la diagonal, se identifican dispersiones asociadas a los errores de predicción, los cuales presentan su mayor magnitud en el extremo inferior, indicando que el modelo presenta mayor dificultad para estimar con precisión valores atípicos bajos.

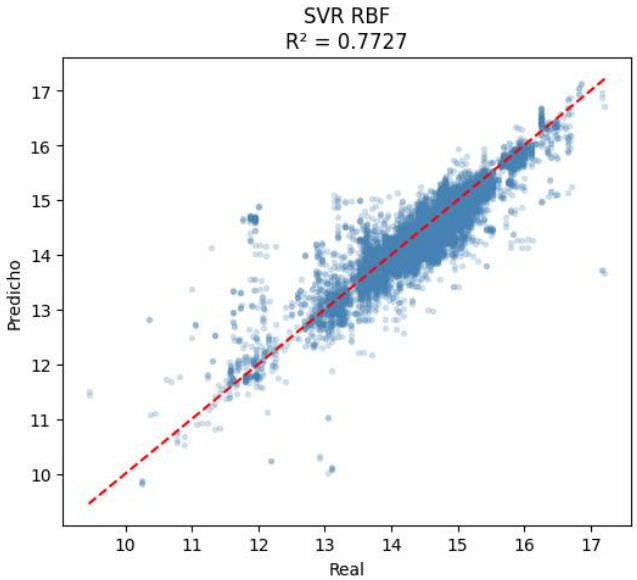


Fig. 17. *Comparación entre valores reales y predichos del valor del suelo mediante el modelo SVR.*
Fuente: elaboración propia a partir de datos de IDECA.

Importancia de Variables

La Fig. *18* presenta a importancia de las variables, medida en términos de cuánto cae el R^2 del modelo cuando se aleatoriza cada variable. Cuanto más larga la barra, más importante es la variable.

Las variables más importantes se representan en rojo por encima de 0,1 y corresponden a aquellas que sostienen el modelo. Dentro de estas variables se encuentran las distancias a Av 68, red férrea, parque metropolitano y línea del metro, confirmando que la accesibilidad a infraestructura urbana es un determinante importante en el valor del suelo. También resulta relevante el área del plan parcial, de modo que planes más grandes generan mayor expectativa de transformación urbana y por tanto mayor aumento en el valor del suelo. También, se incluyen la gran mayoría de estratos socioeconómicos, reflejando que este factor estructural tiene una alta importancia en la definición del valor del suelo.

Las variables de importancia media se representan en azul entre 0,02 y 0,1, estas contribuyen al modelo, pero en menor medida. Se destaca la presencia de la variable Intervenido, indicando que estar dentro del área de implementación de un plan parcial, tiene efecto sobre el valor, el cual es menor comparado con las características estructurales de localización y estratificación.

Finalmente, las variables de importancia baja o nula se representan en color gris, por debajo de 0,02. Que Tratamiento y Post_Plan estén aquí no significa que el plan parcial no tenga efecto causal, sino que para predecir el valor del suelo, estas variables aportan poco una vez que el modelo ya conoce la localización, el estrato y las distancias a infraestructura.

Cabe aclarar que este modelo no está en la capacidad de medir el efecto aislado de una variable independiente sobre el valor del suelo, corresponde a un modelo optimizado para predecir el valor del suelo. Considerando la alta heterogeneidad de los datos se considera que el modelo proporciona una predicción razonable sobre el valor del suelo por metro cuadrado.

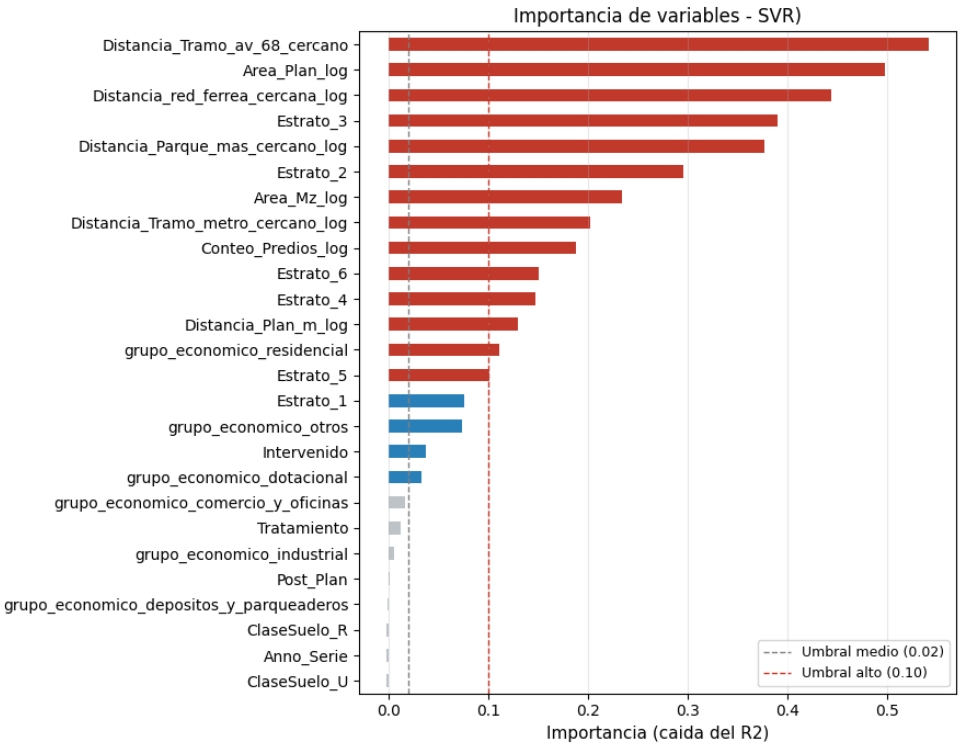


Fig. 18. *Importancia de Variables. Modelo SVR.*
Fuente: elaboración propia a partir de datos de IDECA.

Este procesamiento se encuentra en los siguientes scripts de *Google Colab:*

Anexo L. Modelo_SVM_parte1
Anexo M. Modelo_SVM_parte2

6.4.4. Random Forest

Considerando que el análisis del precio del metro cuadrado por manzana en Bogotá puede depender de relaciones entre variables que no tengan un comportamiento lineal, se eligió el método de aprendizaje automático supervisado random forest para evaluar y capturar estas relaciones y encontrar las variables que explicaran el precio de metro cuadrado por manzana, en contexto de la adopción de planes parciales. También se seleccionó este modelo por su capacidad de comprender la lógica de variables que son categóricas ordinales o que se encuentran codificadas.

Preparación de los datos:

Teniendo en cuenta que el modelo random forest identifica automáticamente la multicolinealidad y que no requiere la evaluación de supuestos como homocedasticidad y normalidad, se desarrollaron los pasos mencionados en la sección 6.3, excluyendo o realizando de manera adicional algunas actividades que ajustaron los datos de acuerdo con la lógica del modelo, para eliminar las situaciones que puedan ser causa de fallo o deficiencia en el proceso de aprendizaje, como se detalla a continuación:

1. Previo a la transformación de variables al formato binario con one-hot encoding (drop_first = True), se creó una copia de las variables para incluirlas en el modelo en su formato original. Este proceso se realizó para: *Estrato*, *Grupo económico, Clase Suelo, Tipo Plan* y las diferentes variables de distancia y área.

2. Codificación de variables categóricas anteriores con LabelEncoder, aplicado para las variables en formato original: *Estrato, Grupo económico, Clase Suelo y Tipo Plan.*

3. Creación de variables auxiliares según lo descrito en el diseño de pruebas general.

4. Agrupación en listas de las variables para utilizar en el modelo, según el diseño de pruebas general. Se adicionó la lista *vars_orig* donde se organizaron las variables originales de distancia, área, estrato, grupo económico, tipo plan, etc.

5. Filtrar la información de las manzanas del plan parcial y las que se encuentran en área de influencia del plan parcial de acuerdo con el diseño de pruebas general.

Con los pasos anteriores se desarrollaron 2 modelos random forest. El primero con los hiperparámetros que el modelo definió por defecto y el segundo utilizando la optimización de hiperparámetros con la librería *Optuna* que utiliza un algoritmo de optimización bayesiana para explorar la mejor combinación de hiperparámetros en el menor tiempo posible. Se eligió esta optimización, toda vez que se encontró una experiencia de su aplicación en un modelo de random forest de predicción de precios de vivienda [47].

Diseño Modelo 1: Random forest con hiperparámetros por default

1. Se siguieron los pasos para división del dataset, conforme se describió en el diseño de pruebas general.

2. Se definieron la variable dependiente y las independientes.

    - y = V_REF_Mz_log (valor m2 manzana en logaritmo)
    - X = Listas de variables generadas (vars_log, vars_tiempo, vars_suelo, vars_orig)

Entrenamiento del modelo: set de entrenamiento y validación cruzada con expanding window

Se entrenó el modelo de random forest realizando una validación cruzada de ventana expansiva (Expanding Window), que utiliza rangos de años del set de entrenamiento para predecir el año futuro. Se aplicó este método toda vez que la validación cruzada con k-folds generaría la predicción con el conocimiento de los datos a futuro toda vez que la división de los conjuntos se realiza de manera aleatoria bajo el supuesto de que cada observación es independiente e idénticamente distribuida, lo cual no aplica en una serie temporal. Esto implicaría tomar datos de cualquier año, incluyendo observaciones de los años a predecir [48].

El método de ventana expansiva funciona expandiendo gradualmente el set de entrenamiento, de manera que la división de los conjuntos de datos siempre realizará la prueba en datos futuros. Además, este método es adecuado entre los existentes para series de tiempo en los casos en los que se identifican patrones persistentes

y estables [48]. Como se ha observado a lo largo del análisis exploratorio y de DiD, los precios tienen una tendencia creciente continua a lo largo del tiempo. Los resultados del entrenamiento bajo este método se presentan en la siguiente tabla:

Tabla XVIII.
Resultado Random Forest con Expanding Window Cross-Validation.

Fold	Train_years	Test_year	R2	RMSE	RMSE (aproximado %)	MAE	MAE (aproximado %)
0	2012-2014	2015	90%	0,25	28%	0,20	22%
1	2012-2015	2016	92%	0,23	25%	0,17	19%
2	2012-2016	2017	94%	0,19	20%	0,13	14%
3	2012-2017	2018	92%	0,22	24%	0,16	17%
4	2012-2018	2019	93%	0,20	22%	0,13	14%
5	2012-2019	2020	92%	0,21	23%	0,15	16%
6	2012-2020	2021	88%	0,25	29%	0,14	15%
	Promedio		92%		23%		15%

Fuente: elaboración propia a partir de datos de IDECA.

De acuerdo con lo anterior, el R^2 de todas las divisiones realizadas en la validación cruzada para series de tiempo indican que el modelo explica en promedio el 92% de la variabilidad de los datos, con excepción de la división N° 6, que indica que el modelo explica un 88% de la variabilidad de los datos.

En cuanto al RMSE y MAE, para su correcta interpretación se deben transformar a una escala interpretable con la función exponencial, considerando que los valores de la variable se encuentran en logaritmo. Los resultados de la columna RMSE aproximado (%) indican que en promedio el error de las predicciones está en 23%, mientras que el MAE aproximado (%) indica que la desviación de las predicciones respecto al valor real es en promedio del 15%

Las diferencias entre los resultados de R^2, RMSE y MAE entre las distintas divisiones realizadas demuestran que existe una variabilidad temporal en los precios del metro cuadrado por manzana.

Posterior a este paso, se realizó el entrenamiento del modelo sobre todo el set definido para el entrenamiento, con los hiperparámetros definidos automáticamente por la función.

```
modelf.fit(X_train, y_train)
```

```
RandomForestRegressor
RandomForestRegressor(random_state=42)
```

```
RMSE Train: 0.060948796362102785
MAE Train: 0.02728582785969043
R2 Train: 0.9952710276419731
```

```
# Hiperparámetros que utiliza el modelo por defecto
model.get_params()
```

```
{'bootstrap': True,
 'ccp_alpha': 0.0,
 'criterion': 'squared_error',
 'max_depth': None,
 'max_features': 1.0,
 'max_leaf_nodes': None,
 'max_samples': None,
 'min_impurity_decrease': 0.0,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'monotonic_cst': None,
 'n_estimators': 100,
 'n_jobs': None,
 'oob_score': False,
 'random_state': 42,
 'verbose': 0,
 'warm_start': False}
```

Fig. 19. *Resultados Random Forest sobre todo el set de entrenamiento y descripción de hiperparámetros por defecto.*
Fuente: elaboración propia a partir de datos de IDECA.

Los resultados de la *Fig. 19* indican que el modelo explica el 99% de la variabilidad de los datos, con un error del 6% sobre todo el set de entrenamiento, lo cual representa un ajuste muy favorable respecto a los datos originales. Sin embargo, la capacidad predictiva del modelo se evalúa realmente sobre el set de prueba, que corresponde a los datos no conocidos por el algoritmo.

Al realizar las predicciones del modelo sobre el set de prueba se encuentra que se reduce el coeficiente de determinación R^2 a 91% y el RMSE incrementa el error a 24%. Al analizar el coeficiente de determinación para cada uno de los años de la predicción, se identifica que el modelo pierde capacidad predictiva a medida que aumenta el tiempo futuro de la predicción, como se presenta en la Tabla XIX.

Tabla XIX.
Resultados Random Forest sobre todo el set de prueba y descripción de coeficiente de determinación por año de predicción.

Métrica de Evaluación	**Resultado**
RMSE Test	0,21
RMSE Test aproximado (%)	24%
MAE Test	0,15
MAE Test aproximado (%)	17%
R2 Test:	91%

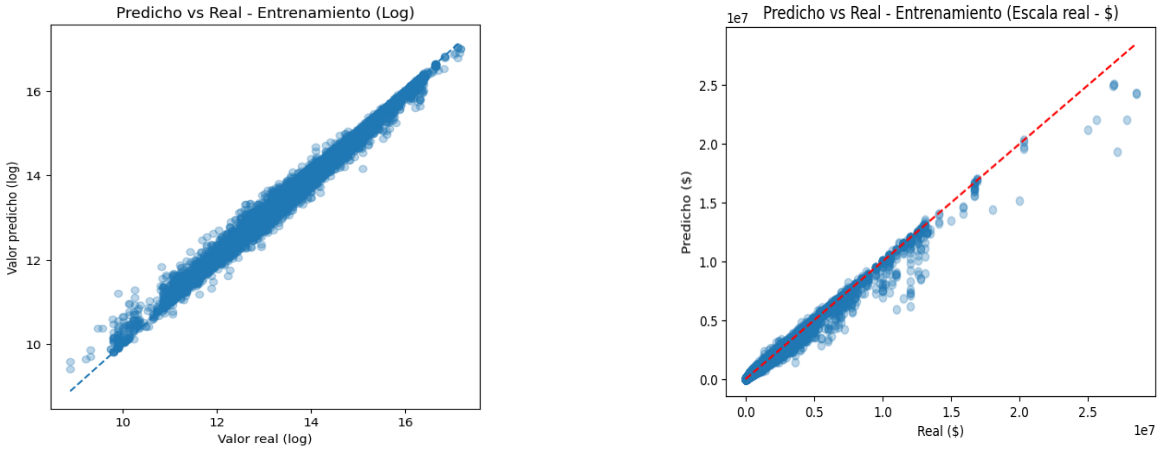
Anno_Serie	
2023	0.964576
2024	0.909366
2025	0.855489

Fuente: elaboración propia a partir de datos de IDECA

Gráficamente estos resultados se presentan en la Fig. 20, donde se presenta el ajuste de los valores predichos

frente a los reales tanto en escala logarítmica como en escala real, para el set de entrenamiento y de prueba, en los cuales se puede observar que el ajuste se reduce al predecir el valor del metro cuadrado por manzana en años que el algoritmo desconoce. No obstante, es un modelo con buen ajuste, cuya capacidad para explicar la variabilidad del modelo no es inferior al 90%.

Resultados valores predichos vs reales – Set de entrenamiento



Resultados valores predichos vs reales – Set de prueba

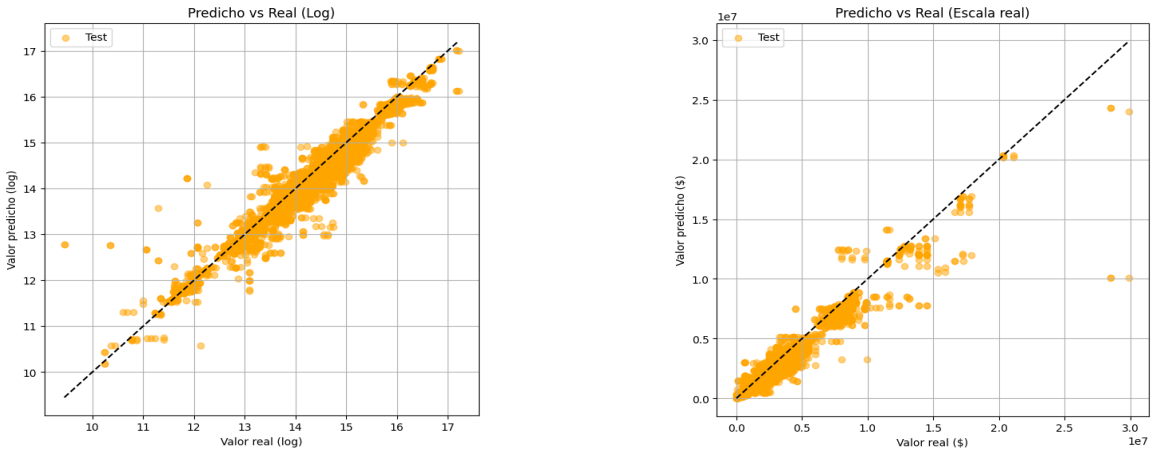


Fig. 20. Comparación entre valores observados y predichos del valor del suelo mediante el modelo Random Forest: (a) entrenamiento en escala logarítmica; (b) entrenamiento en escala real; (c) prueba en escala logarítmica; (d) prueba en escala real.
Fuente: elaboración propia a partir de datos de IDECA

Modelo Random Forest con ajuste de hiperparámetros

Al realizar la optimización del modelo con Optuna, se aplicaron 20 ensayos en los cuales se utilizaron diferentes combinaciones de los hiperparámetros y finalmente el modelo encontró que los que mejor optimizaban el modelo eran:

- n_estimators: 284
- max_depth: 20

- n_jobs = -1
- min_samples_leaf: 1

Al correr el modelo con esta optimización tanto para el set de entrenamiento como de prueba, se obtuvieron los resultados que se presentan en la Tabla XX

Tabla XX.
Resultado Random Forest con optimización de hiperparámetros set de entrenamiento y prueba.

Métrica de Evaluación	Resultado Set Entrenamiento	Resultado Set Prueba
RMSE	0,10	0,22
RMSE aproximado (%)	10%	25%
MAE	0,5	0,17
MAE aproximado (%)	5%	20%
R2 Test:	99%	90%

Fuente: elaboración propia a partir de datos de IDECA.

Como se observa, la optimización de hiperparámetros no mejora la capacidad explicativa del modelo del coeficiente R^2 sino que la reduce, aunque marginalmente. De igual manera el RMSE aumenta de 24% a 25% y el MAE de 17% a 20% respecto al modelo con los hiperparámetros automáticos.

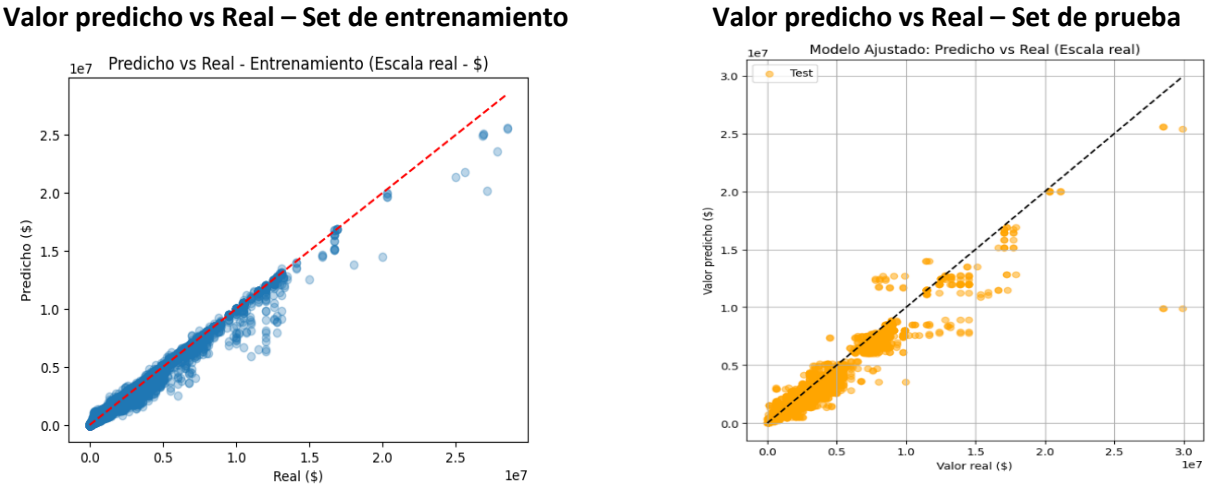


Fig. 21. Comparación entre valores observados y predichos del valor del suelo en escala real mediante el modelo Random Forest optimizado: (a) conjunto de entrenamiento; (b) conjunto de prueba.
Fuente: elaboración propia a partir de datos de IDECA

Al analizar gráficamente los resultados del ajuste de las predicciones en el set de entrenamiento y prueba en la escala real no se encuentran diferencias significativas respecto a los resultados encontrados en el modelo con los hiperparámetros automáticos.

Finalmente, utilizando el método de Importancia de variables basada en Random Forest (*Feature Importance*), se identificó la importancia de cada variable en la predicción del modelo. Cabe aclarar que este enfoque permite identificar los factores explican el precio del m² por manzana, pero es limitado porque no define si la relación es

positiva o negativa. El método construye la importancia de cada variable a partir de los siguientes pasos.

1. Suma cuánto reduce el error en todos los árboles
2. Se promedia
3. Se normaliza (suma total = 1)

En la Fig. 22 se presentan los resultados del método aplicado tanto para el modelo Random Forest con hiperparámetros automáticos y para el optimizado con Optuna, y se encontró que la variable *Estrato_orig* es la más relevante en la predicción del modelo, seguido del año de la serie y las distancias originales y logarítmicas al Tramo de la Av. 68 y la red férrea más cercana. No se observan diferencias entre ambos modelos, toda vez que el cambio de resultados fue marginal entre ellos.

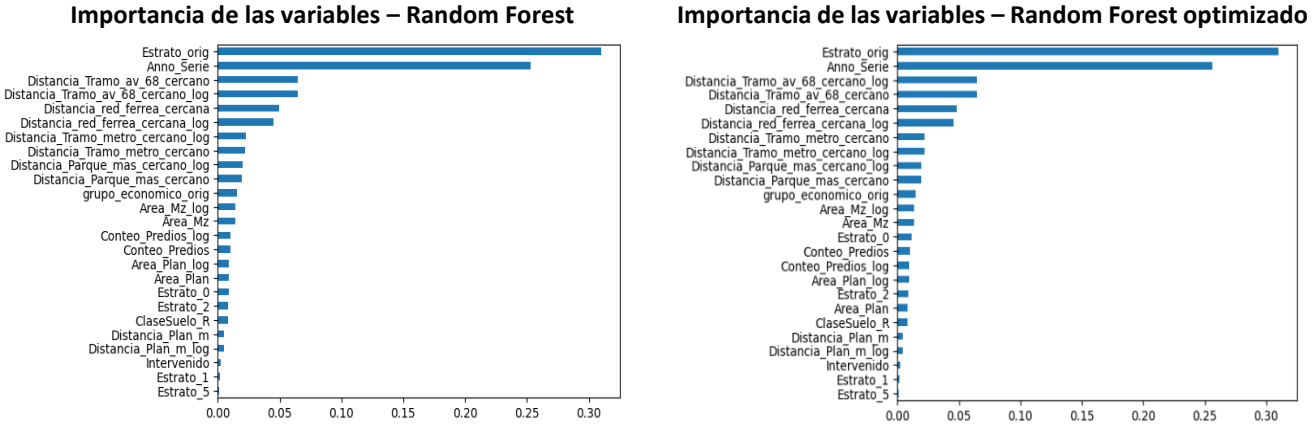


Fig. 22. Importancia relativa de las variables explicativas en los modelos Random Forest: (a) modelo original; (b) modelo optimizado.
Fuente: elaboración propia a partir de datos de IDECA.

Si bien las variables relacionadas con los planes parciales no se muestran como relevantes, esto es coherente con el comportamiento general del precio del suelo, dado que se identificó en el análisis de DiD que, aunque los Planes Parciales tienen un efecto positivo sobre el precio, este impacto es temporal y se manifiesta en máximo los 3 años siguientes a la adopción del plan; posterior a este periodo de tiempo, el comportamiento del precio del metro cuadrado por manzana se ve influenciado por las variables que determinan su precio, particularmente el estrato. Por lo tanto, en un análisis agregado de toda la serie temporal, esta influencia puede ser poco representativa respecto a otras variables que influyen de manera permanente en el precio.

De manera general, se puede concluir que el modelo presenta un muy buen desempeño predictivo, con un R^2 de 0,91 en el conjunto de prueba y un error relativo promedio cercano al 24%. Si bien el ajuste es muy alto en el conjunto de entrenamiento (R^2 = 0,99), la diferencia en el desempeño en el set de prueba no sugiere que el modelo tenga un d alto, dado que se mantiene por encima del 90%, por lo cual se considera que el modelo tiene alta capacidad de aprendizaje y permite una adecuada generalización de los resultados. Este procesamiento se encuentra en el script de Google Colab:

Anexo N. Modelo_Random_Forest

7. EVALUACIÓN

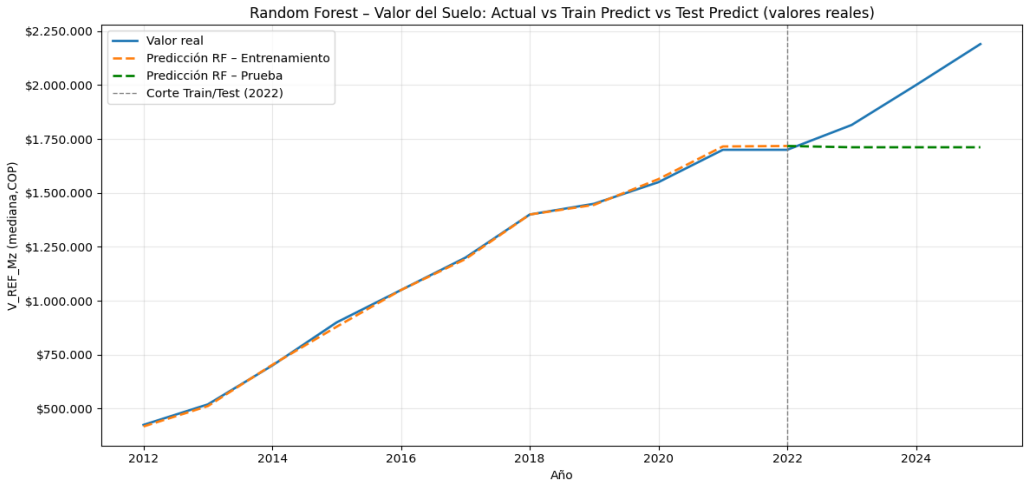
Teniendo en cuenta los resultados presentados previamente, en este capítulo se presentan de manera comparativa para los modelos realizados, los resultados gráficos de las predicciones en los conjuntos de entrenamiento y prueba, así como las métricas de evaluación definidas para validar la capacidad predictiva del modelo.

Cabe anotar que, si bien los modelos se desarrollaron con información conocida para todo el periodo de análisis, se predice el valor del suelo para observaciones que el modelo no vio durante el entrenamiento, aunque esas observaciones pertenezcan al mismo periodo histórico. Es decir que el alcance de los modelos es predictivo dentro del periodo analizado, pero no prospectivo.

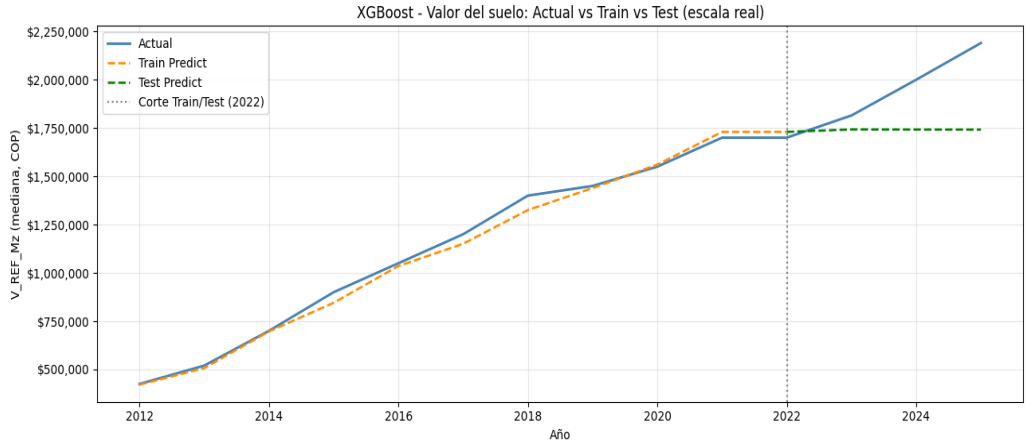
Adicionalmente, en los modelos aplicados se realizaron ajustes, variaciones u optimizaciones que permitieran mejorar los resultados de las métricas de evaluación, sin embargo, estos nuevos modelos usualmente consumían mayores recursos en el procesamiento de la información, sin lograr mejoras significativas en las métricas o, en algunos casos, reduciendo los resultados respecto a los modelos base aplicados.

La Fig. 23, muestra el comportamiento del valor del suelo a lo largo del periodo analizado, contrastando los valores reales de la base frente a las predicciones en el set de entrenamiento y prueba, para lo cual se utilizó la mediana del valor del metro cuadrado por manzana para cada año, con el fin de minimizar la influencia de valores atípicos.

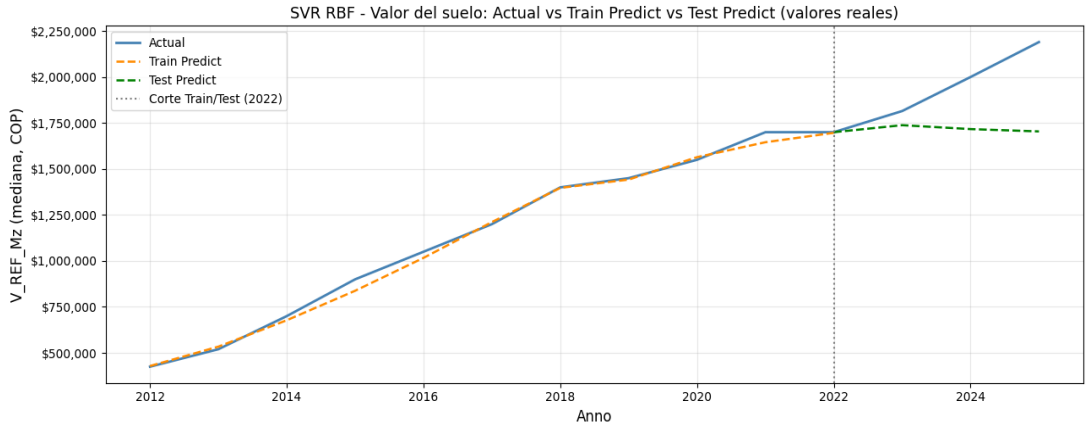
Random Forest



XGBoost



Máquinas de Soporte Vectorial



Regresión Lineal Múltiple

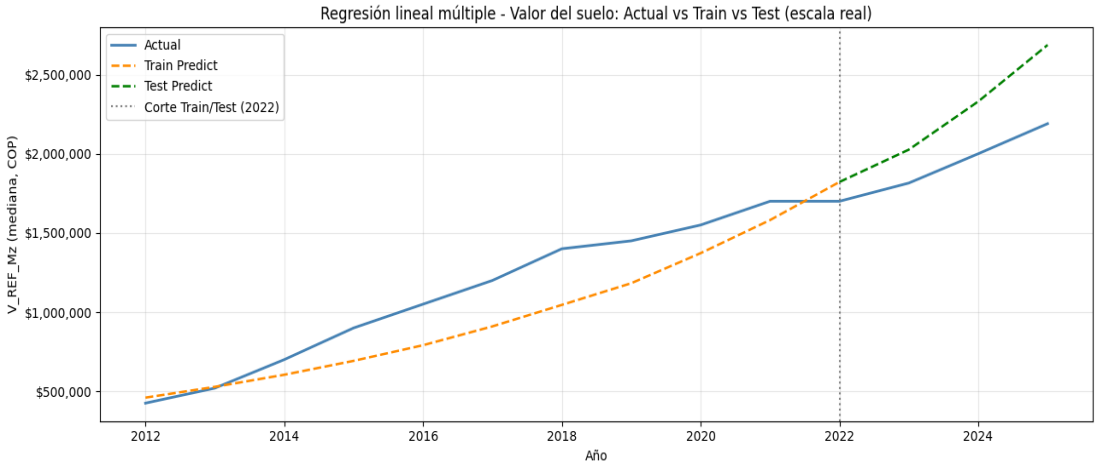


Fig. 23. Resultados de las predicciones en los conjuntos de entrenamiento y prueba: (a) Random Forest; (b) XGBoost (c) SVM (d) Regresión Lineal

Fuente: elaboración propia a partir de datos de IDECA.

De acuerdo con la Fig. 23 en el periodo inicial de entrenamiento (2012-2022), para todos los modelos, con excepción de la Regresión Lineal, la predicción (línea naranja) se ajusta en gran medida a los valores reales (línea azul) confirmando que los modelos capturaron en gran medida el comportamiento del valor del suelo. En el periodo de prueba (2022-2025), la predicción (línea verde) tiende a mantenerse plana, mientras que el valor del suelo aumenta, es decir que el modelo predice que el valor del suelo se estabiliza, cuando en realidad continúa creciendo.

Lo anterior es coherente, para los modelos de Random Forest, XGBoost y SVM considerando que la lógica para la construcción del gráfico implicaba el uso de medidas de tendencia central para agregar la información por año, convirtiendo un problema no lineal que fue evaluado de manera individual, es decir, observación por observación en una tendencia lineal.

De manera general, estos modelos no extrapolan tendencias, sino que repiten el último patrón aprendido o combinaciones de los datos de entrenamiento. Por esta razón, si bien en las predicciones sobre el set de entrenamiento los modelos ajustan de manera correcta a nivel micro (por cada observación), no extrapolan correctamente la tendencia temporal agregada dado que no reconocen estados temporales sino patrones.

En este sentido, los resultados evidencian que la capacidad de los modelos para capturar dinámicas de crecimiento temporal es limitada, lo cual se refleja en la estabilización de las predicciones en el periodo posterior al corte de entrenamiento. Lo anterior sugiere que esta brecha seguirá aumentando a lo largo del tiempo, por lo que se debe considerar el alcance temporal del modelo e interpretar con precaución los resultados del periodo de prueba.

Finalmente, la Fig. 23(d) evidencia que la regresión lineal múltiple presenta mayores limitaciones en la predicción del valor del metro cuadrado por manzana, pero si captura la tendencia creciente del valor del suelo en Bogotá. A lo largo del todo el periodo analizado se observan diferencias importantes entre los valores observados y predichos, especialmente en el conjunto de prueba, lo que refleja dificultades para representar la complejidad espacial, económica y temporal del mercado del suelo urbano.

En cuanto a las métricas de evaluación definidas para evaluar todos los modelos: coeficiente de determinación (R^2), el error cuadrático medio (RMSE) y el error absoluto medio (MAE), el MAE fue considerado relevante debido a que proporciona una medida directa y fácilmente interpretable del error promedio de predicción, además de presentar menor sensibilidad frente a observaciones atípicas. Esto resulta especialmente importante en el presente estudio, considerando la existencia de comportamientos heterogéneos y valores elevados del suelo en diferentes sectores de Bogotá.

Por su parte, el RMSE fue incluido debido a que penaliza en mayor medida los errores grandes de predicción, permitiendo identificar desviaciones significativas entre los valores reales y predichos. Esta métrica es ampliamente utilizada en modelos basados en árboles de decisión y técnicas de boosting, como XGBoost y Random Forest, debido a su capacidad para evaluar la precisión global del ajuste.

Finalmente, el coeficiente de determinación (R^2) fue utilizado por su facilidad de interpretación y por permitir

medir la proporción de variabilidad del fenómeno que logra ser explicada por el modelo. A diferencia de las métricas de error, el R^2 no depende de las unidades de medida de las variables, lo que facilita la comparación del desempeño entre diferentes modelos predictivos.

Con el fin de complementar el análisis visual presentado anteriormente, se realizó una comparación cuantitativa del desempeño predictivo de los modelos evaluados utilizando métricas de error y capacidad explicativa sobre el conjunto de prueba. Para ello, se emplearon el coeficiente de determinación (R^2), el Error Absoluto Medio (MAE) y la Raíz del Error Cuadrático Medio (RMSE), permitiendo comparar la capacidad de generalización y precisión de cada metodología de aprendizaje automático implementada. De esta manera, la Tabla XXI reúne los resultados de los modelos con el mejor desempeño de todas las variaciones de entrenamiento realizadas.

Tabla XXI.

Resultado comparado de las métricas de evaluación en set de prueba de los modelos de aprendizaje automático utilizados

Modelo Evaluado	**RMSE aproximado (%)**	**MAE aproximado (%)**	**R2**
Regresión Lineal Múltiple	81,9%	47,0%	29,0%
Máquina de Soporte Vectorial	40,3%	24,2%	77,3%
XGBoost	29,0%	19,4%	87,2%
Random Forest	24,6%	16,5%	91,1%

Fuente: elaboración propia a partir de datos de IDECA

Los resultados comparativos evidencian diferencias importantes en el desempeño predictivo de los modelos utilizados para estimar el valor del suelo por metro cuadrado por manzana para la ciudad de Bogotá. Los errores RMSE y MAE presentados corresponden a aproximaciones porcentuales derivadas de la transformación exponencial de las métricas obtenidas en escala logarítmica. A partir de las métricas presentadas en la Tabla XXI, se puede concluir lo siguiente:

- La Regresión Lineal Múltiple presentó el menor desempeño general, con un R^2 de 29,0 %, indicando que solo logra explicar una parte limitada de la variabilidad del valor del metro cuadrado por manzana, además de registrar los errores de predicción más altos (RMSE = 81,9 % y MAE = 47,0 %). Este modelo evidencia que las relaciones lineales simples no son suficientes para representar la complejidad espacial, temporal y económica del fenómeno analizado.

- La Máquina de Soporte Vectorial mejoró considerablemente los resultados, alcanzando un R^2 de 77,3 % y reduciendo de manera importante los errores de predicción (RMSE = 40,3 % y MAE = 24,2 %). Esto es coherente toda vez que este tipo de modelos capturan relaciones más complejas y no lineales entre las variables explicativas y el valor del metro cuadrado por manzana. No obstante, aún se observan errores de predicción relativamente significativos frente a modelos de mayor desempeño.

- Por su parte, XGBoost presentó mejor desempeño respecto a los dos modelos anteriores, con un R^2 de

87,2 % y errores menores en la predicción (RMSE = 29,0 % y MAE = 19,4%). Estos resultados evidencian que el modelo logra representar adecuadamente patrones complejos y relaciones no lineales presentes en los datos. Asimismo, la incorporación de estructuras basadas en árboles potenciados permitió mejorar significativamente la capacidad predictiva respecto a los modelos tradicionales y lineales.

- Finalmente, el modelo Random Forest fue el que obtuvo el mejor desempeño global para todas las métricas evaluadas, alcanzando el mayor poder explicativo (R^2 = 91,1 %) y los menores errores de predicción (RMSE = 24,6 % y MAE = 16,5 %). Esto indica que el modelo logra aproximarse de manera más precisa a los valores reales del suelo, consolidándose como el método más robusto y adecuado entre los evaluados para representar la dinámica espacial y económica del mercado del suelo en Bogotá.

Por lo expuesto, se elige el modelo Random Forest como el modelo con el mejor desempeño para la predicción del valor del suelo por metro cuadrado por manzana en la ciudad de Bogotá, al presentar el mayor coeficiente de determinación (R^2) y los menores errores de predicción entre los modelos evaluados.

Asimismo, el análisis realizado evidencia que los planes parciales no se distribuyen de manera aleatoria en el territorio, sino que tienden a localizarse en zonas con características urbanas, normativas, económicas y espaciales particulares. En este sentido, la predicción futura del comportamiento del valor del suelo en áreas donde aún no se han implementado planes parciales requiere identificar previamente dichas condiciones territoriales y los criterios asociados a la selección de estas áreas por parte de la administración distrital.

Las evaluaciones se encuentran en los respectivos scripts de *Google Colab* desarrollados para cada modelo.

8. VISUALIZACIÓN

Con el propósito de facilitar la interpretación territorial de los resultados obtenidos durante el proceso de modelamiento, se desarrolló una herramienta de visualización geoespacial interactiva orientada a resaltar los patrones espaciales identificados tras la predicción del valor del suelo en Bogotá a partir de la adopción de planes parciales. El visor fue concebido como un prototipo funcional de apoyo a la exploración espacial y a la comunicación de resultados derivados de técnicas de aprendizaje automático aplicadas al análisis urbano.

La herramienta integra las predicciones obtenidas mediante el modelo Random Forest, previamente seleccionado durante la etapa de modelamiento por presentar el mejor desempeño predictivo. Con fines de visualización e interpretación territorial, el modelo seleccionado fue aplicado posteriormente a la totalidad de las manzanas de Bogotá, permitiendo ampliar la lectura espacial de los patrones de variación del valor del suelo para toda la ciudad. Esta fase corresponde a un ejercicio de inferencia espacial orientado a facilitar la exploración visual de tendencias territoriales y no sustituye la evaluación realizada sobre el dominio principal del estudio.

La construcción de la herramienta se hizo inicialmente en QGIS, donde se integraron y prepararon las capas geográficas asociadas a las manzanas de Bogotá, localidades, zonas de influencia y planes parciales. Posteriormente, la base cartográfica fue exportada a entorno web mediante el complemento QGIS2Web, permitiendo generar una estructura inicial compatible con tecnologías web.

A partir de esta exportación, para la visualización espacial interactiva se utilizó la librería Leaflet, mientras que PapaParse fue empleado para la lectura y procesamiento de datos tabulares en formato CSV asociados a las predicciones del modelo. Adicionalmente, se utilizó Plotly para la generación de gráficos analíticos y métricas de desempeño integradas dentro del tablero. Sobre esta estructura el visor fue ajustado y personalizado en Visual Studio Code utilizando HTML, CSS y JavaScript, incorporando funcionalidades y componentes personalizados como filtros dinámicos, controles de capas geográficas, leyendas temáticas, ventanas emergentes (pop-ups) y herramientas de navegación espacial, orientadas a mejorar la experiencia de exploración espacial y la interacción con la información geográfica.

El visor permite explorar espacialmente variables asociadas a localidades, tipos de planes parciales, estratos socioeconómicos, grupos económicos y los valores reales y predichos del suelo por metro cuadrado de las manzanas de Bogotá. Asimismo, incorpora herramientas de filtrado y superposición de capas que facilitan la identificación de dinámicas de valorización y posibles patrones espaciales relacionados con la adopción de planes parciales en Bogotá.

El proceso de inferencia espacial desarrollado se encuentra documentado en el script de Google Colab presentado en el siguiente anexo:

Anexo O. Modelo_Random_Forest_visualizacion.

Por su parte, el visor geoespacial interactivo puede consultarse en el siguiente anexo:

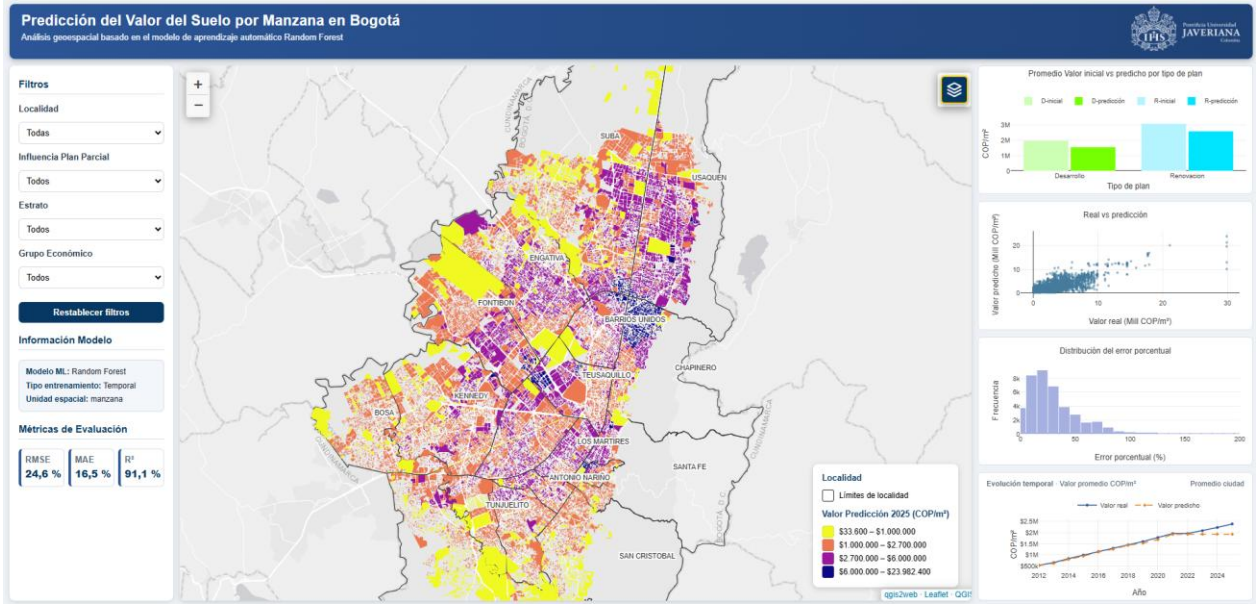
Anexo P. Tablero

El visor se ejecuta abriendo el archivo HTML “Index” a través del software libre Visual Studio Code, el cual no requiere licenciamiento, facilitando la consulta, divulgación e interpretación de los resultados por parte de diferentes tipos de usuarios, como se observa en la Fig. 24

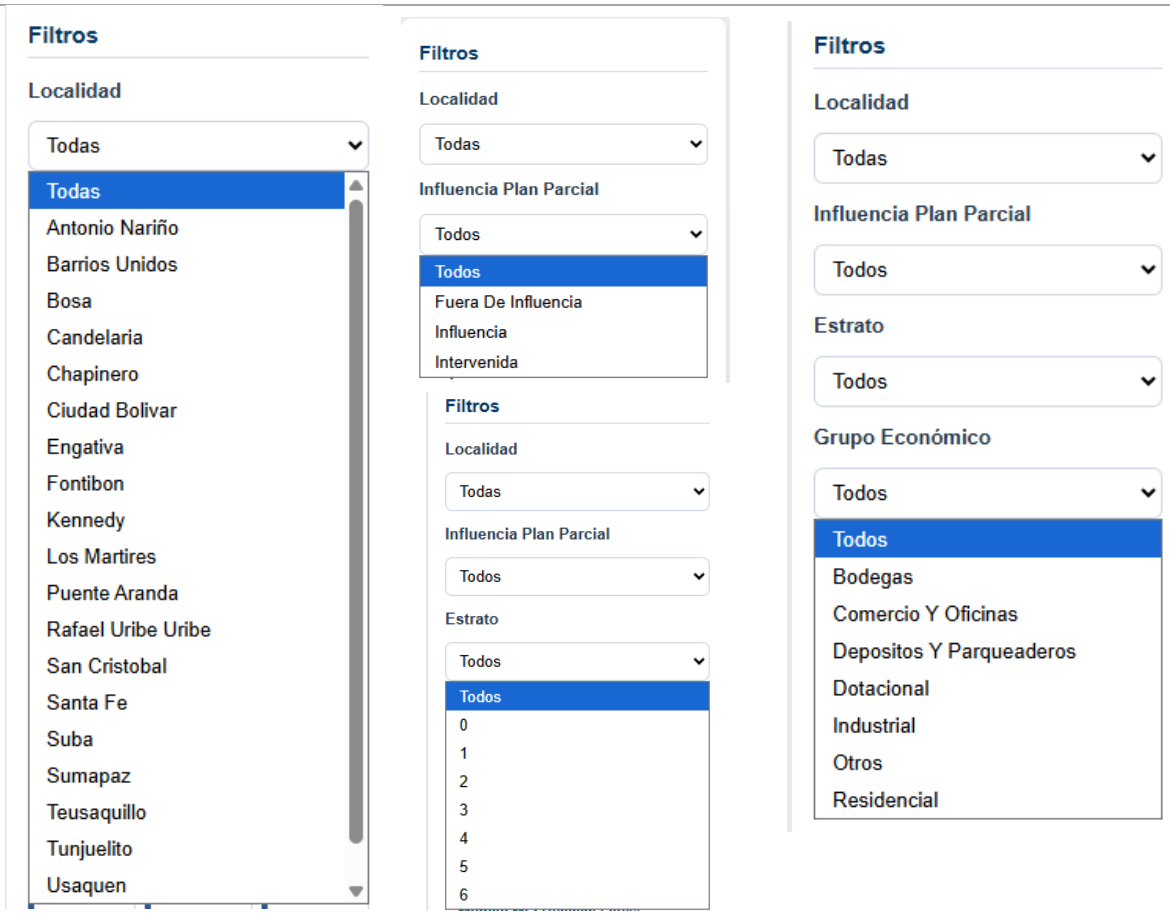
Igualmente, se puede acceder al video demostrativo del tablero a través del siguiente link [49]:

Tablero_Planes_Parciales.mp4

Estructura general Visor



Filtros del visor



Información del modelo y métricas de evaluación

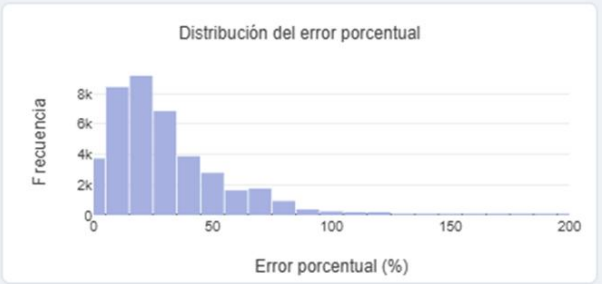
Información Modelo

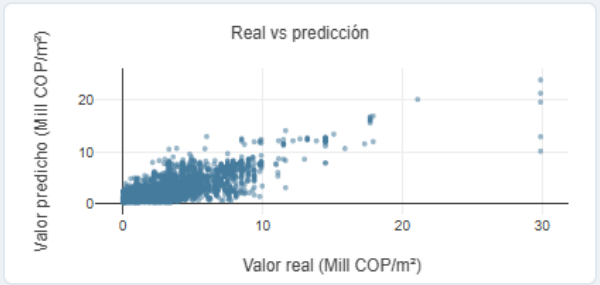
Modelo ML: Random Forest
Tipo entrenamiento: Temporal
Unidad espacial: manzana

Métricas de Evaluación

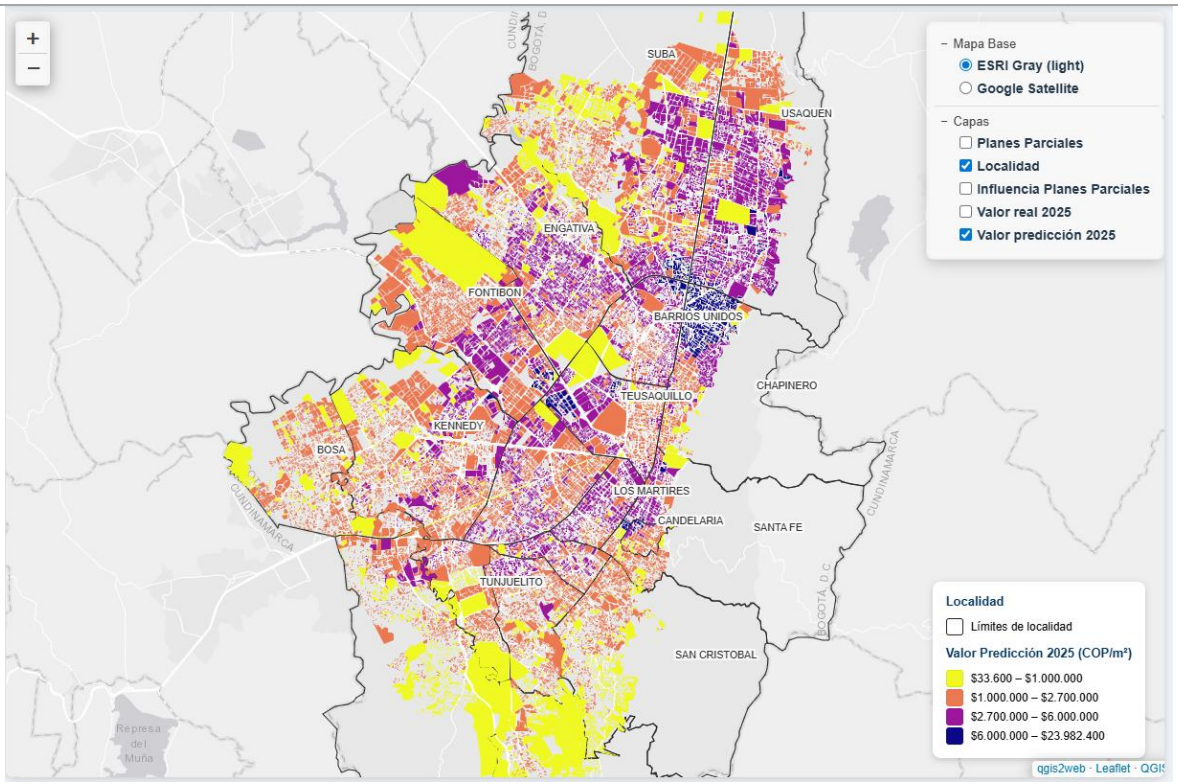
RMSE	MAE	R^2
24,6 %	16,5 %	91,1 %

Gráficos descriptivos e interactivos

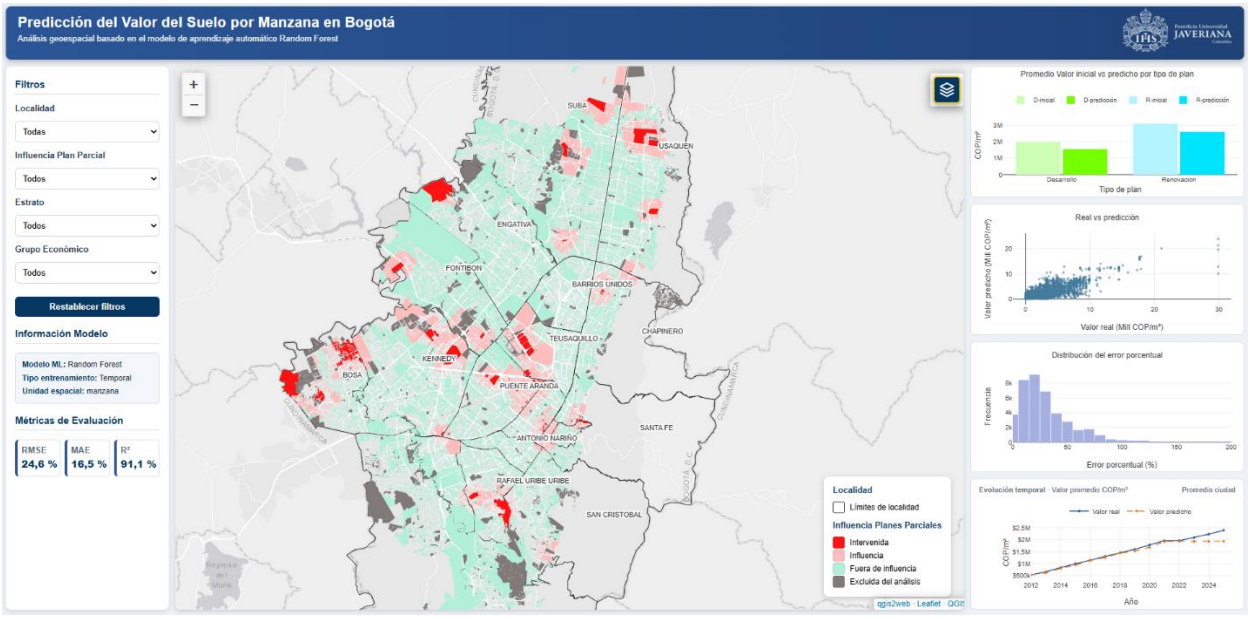




Mapa de Bogotá D.C., capas y leyendas



Mapa de Bogotá D.C., capa de Influencia Planes Parciales



Elementos interactivos: Pop ups

Elementos interactivos: filtros y gráficos dinámicos por manzana

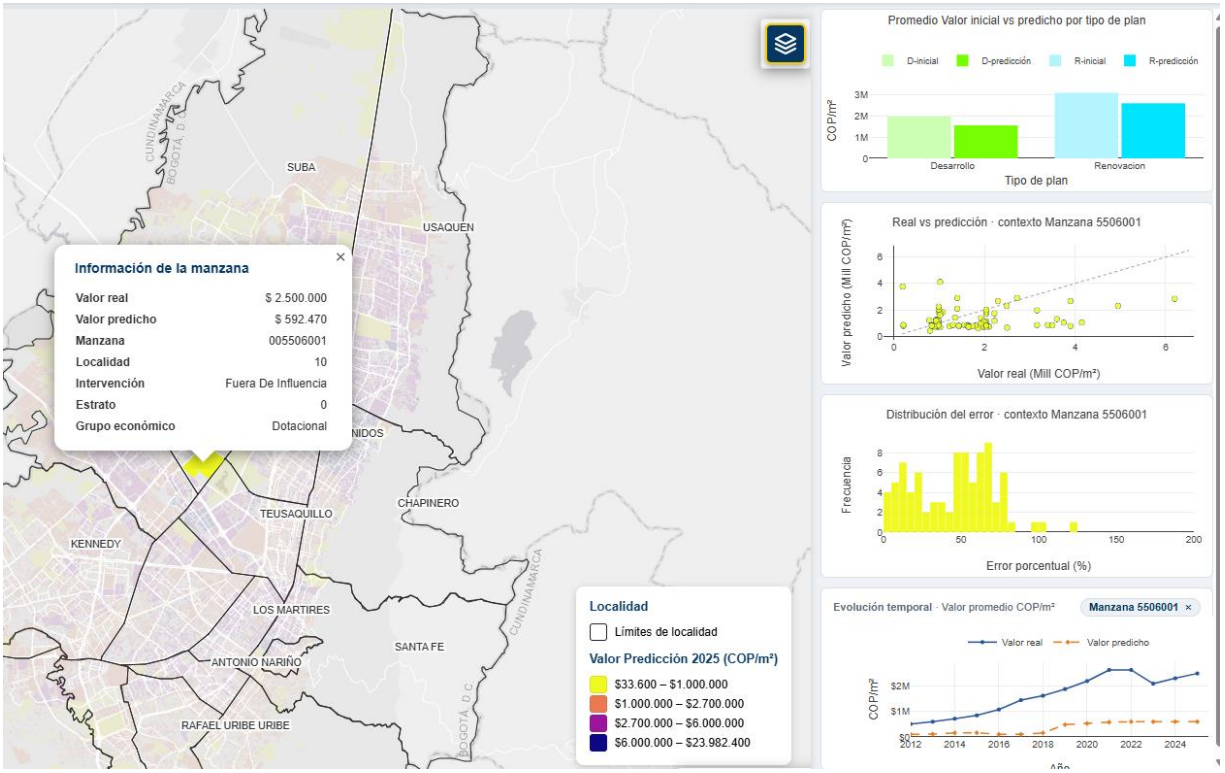


Fig. 24. Vista general del visor web: (a) Estructura general Visor; (b) Filtros del visor; (c) Información del modelo y métricas de evaluación; (d) Gráficos descriptivos e interactivos; (e) Mapa de Bogotá D.C., capas y leyendas; (f) Mapa de Bogotá D.C., capa de Influencia Planes Parciales (g) Elementos interactivos: Pop ups; (h) Elementos interactivos: filtros y gráficos dinámicos por manzana

Fuente: elaboración propia.

9. CONCLUSIONES Y TRABAJOS FUTUROS

9.1. Conclusiones

Las conclusiones generales del estudio evidencian que el precio del metro cuadrado por manzana para la ciudad de Bogotá está determinado principalmente por factores estructurales, socioeconómicos y de localización, más que por la implementación de planes parciales. En todos los modelos evaluados se identificó que variables asociadas al estrato socioeconómico, la localización urbana y la dinámica temporal del mercado presentaron una mayor capacidad explicativa sobre el comportamiento del valor del suelo, consolidándose como los principales determinantes del fenómeno analizado. Asimismo, los resultados mostraron que los planes parciales sí generan efectos sobre el valor del suelo; sin embargo, dichos impactos no son dominantes frente a la lógica estructural del mercado inmobiliario urbano.

En el análisis de Diferencias en Diferencias, la desaparición del efecto causal en los modelos con efectos fijos de doble vía (TWFE) indica que las diferencias observadas responden principalmente a características estructurales y dinámicas territoriales preexistentes. Más que interpretar los planes parciales como una variable causal aislada, este análisis permitió evidenciar que estos deben entenderse como parte de un conjunto de condiciones territoriales asociadas a procesos de transformación urbana.

El ejercicio comparativo permitió evidenciar que los modelos de aprendizaje automático, especialmente Random Forest y XGBoost, presentan un desempeño significativamente superior frente a la regresión lineal múltiple y las Máquinas de Soporte Vectorial, logrando capturar de mejor manera relaciones no lineales y patrones complejos presentes en la variable objetivo. Esto demuestra que la dinámica espacial del valor del suelo en Bogotá responde a interacciones complejas entre variables urbanas, temporales y socioeconómicas, que difícilmente pueden representarse mediante modelos lineales tradicionales.

De igual forma, el análisis permitió identificar que los efectos indirectos, de influencia o spillovers espaciales asociados a áreas vecinas de los planes parciales existen, aunque con una incidencia relativamente limitada frente a otros factores estructurales. También se evidenciaron diferencias entre los tipos de plan parcial, observándose que los planes de renovación urbana se ubican generalmente en zonas con precios por metro cuadrado más altos, mientras que los planes de desarrollo presentan mayores incrementos relativos de valorización después de la intervención.

Asimismo, los resultados indican que el modelo unificado mantiene un desempeño adecuado para representar las dinámicas generales del valor del suelo en Bogotá, pese a las diferencias identificadas entre tipos de plan. Esto respalda metodológicamente la utilización de un único modelo para capturar de manera integrada las dinámicas urbanas y espaciales de la ciudad.

Finalmente, la Infraestructura de Datos Espaciales del Distrito Capital (IDECA), dispone de un amplio volumen de información espacial actualizada, confiable y de alta calidad. Estos datos constituyeron un insumo fundamental para el desarrollo del presente proyecto y, representan una gran oportunidad para la implementación de nuevos enfoques y aplicaciones basadas en ciencia de datos, análisis espacial y técnicas de

aprendizaje automático orientadas a la toma de decisiones en el territorio.

9.2. Trabajos Futuros

Se sugiere que en trabajos futuros se definan las zonas potenciales u óptimas para la adopción de planes parciales, teniendo en cuenta los múltiples requisitos técnicos, normativos y territoriales exigidos por la administración distrital. En la medida en que dichas zonas puedan ser identificadas, se considera viable realizar ejercicios de predicción futura del valor del suelo en las posibles áreas de influencia asociadas, utilizando modelos con alta capacidad de generalización como el seleccionado en el presente estudio.

Adicionalmente, se propone incorporar nuevas variables relacionadas con equipamientos urbanos, infraestructura vial existente, variables ambientales, amenazas o condiciones de riesgo, así como dinámicas socioeconómicas más detalladas disponibles en los conjuntos de datos del Plan de Ordenamiento Territorial de Bogotá y otras fuentes de IDECA. Lo anterior, debido a que estos factores pueden influir significativamente en el comportamiento del valor del suelo en Bogotá.

Finalmente, se recomienda explorar modelos espacialmente explícitos que permitan capturar fenómenos de autocorrelación espacial y dependencias geográficas, así como desarrollar análisis temporales más detallados que permitan identificar posibles efectos anticipados asociados a dinámicas especulativas previas a la adopción de los planes parciales.

ANEXOS

Los anteriores anexos se pueden consultar en el siguiente sharepoint de OneDrive: Anexos

REFERENCIAS BIBLIOGRÁFICAS

[1] A. Araque, L. Cantor, y P. Moreno, “Norma y precio del suelo urbano en la ciudad de Bogotá”, *EURE*, vol. 48, nº 143, p. 22, 2021, doi: 10.7764/EURE.48.143.13.

[2] Congreso de Colombia, “Ley 388 de 1997. Por la cual se modifica la Ley 9a de 1989 y la Ley 3a de 1991 y se dictan otras disposiciones sobre desarrollo territorial”. 1997. Accedido: 18 de mayo de 2025. [En línea]. Disponible en: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=339

[3] Alcaldía Mayor de Bogotá D.C., “Decreto 190 de 2004. Por medio del cual se compilan las disposiciones contenidas en los Decretos Distritales 619 de 2000 y 469 de 2003”. 2004. Accedido: 18 de mayo de 2025. [En línea]. Disponible en: https://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=13935

[4] Alcaldía Mayor de Bogotá D.C., “Decreto 555 de 2021. Por el cual se adopta la revisión general del Plan de Ordenamiento Territorial de Bogotá D.C.” 2021. Accedido: 18 de mayo de 2025. [En línea]. Disponible en: https://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=119582

[5] Congreso de Colombia, “Ley 1753 de 2015. Por la cual se expide el Plan Nacional de Desarrollo 2014–2018 ‘Todos por un nuevo país’”. 2015. Accedido: 18 de mayo de 2025. [En línea]. Disponible en: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=61933

[6] Congreso de Colombia, “Ley 1955 de 2019. Por la cual se expide el Plan Nacional de Desarrollo 2018–2022 ‘Pacto por Colombia, pacto por la equidad’”. 2019. Accedido: 18 de mayo de 2025. [En línea]. Disponible en: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=93970

[7] Congreso de Colombia, “Decreto 148 de 2020. Por el cual se reglamentan parcialmente los artículos 27 y 28 de la Ley 1955 de 2019 en relación con el catastro multipropósito”. 2020. Accedido: 18 de mayo de 2025. [En línea]. Disponible en: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=105952

[8] Unidad Administrativa Especial de Catastro Distrital, “Acuerdo 004 de 2021. Por el cual se adopta el reglamento técnico para la definición y actualización del valor del suelo en el Distrito Capital”. 2021. Accedido: 18 de mayo de 2025. [En línea]. Disponible en: https://www.catastrobogota.gov.co/sites/default/files/archivos2020/Acuerdo_004_2021_Estatutos_Estructura.pdf

[9] Alcaldía Mayor de Bogotá D.C., “Decreto 452 de 2008. Por el cual se adopta el Plan Parcial ‘La Pampa’, ubicado en la localidad de Kennedy”. 2008. Accedido: 18 de mayo de 2025. [En línea]. Disponible en: https://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=34227

[10] Alcaldía Mayor de Bogotá D.C., “Decreto 432 de 2022. Por medio del cual se modifica la estructura organizacional de la Secretaría Distrital de Planeación y se dictan otras disposiciones”. 2022. Accedido: 18 de mayo de 2025. [En línea]. Disponible en: https://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=128740

[11] C. de Colombia, “Ley 1712 de 2014. Por medio de la cual se crea la Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional y se dictan otras disposiciones”. 2014. [En línea]. Disponible en: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=56882

[12] IDE de Bogotá, “IDECA”. 2025. Accedido: 18 de mayo de 2025. [En línea]. Disponible en: https://www.ideca.gov.co/sobre-ideca/la-ide-de-bogota

[13] K. P. Murphy, *Machine Learning. A Probabilistic Perspective*. Cambridge, Massachusetts: MIT Press, 2012. Accedido: 27 de mayo de 2025. [En línea]. Disponible en: https://raw.githubusercontent.com/kerasking/book-1/master/ML%20Machine%20Learning-A%20Probabilistic%20Perspective.pdf

[14] IAAR, “Libro online de IAAR. Introducción al Machine Learning”. 2025. Accedido: 18 de mayo de 2025. [En línea]. Disponible en: https://iaarbook.github.io/ML/

[15] M. B. Solutions, “Análisis y ciencia de datos: orígenes e historia”. 2025. Accedido: 18 de mayo de 2025.

[En línea]. Disponible en: https://www.mistralbs.com/blog/analisis-y-ciencia-de-datos-origenes-e-historia/
[16] J. Mora Pineda, “Modelos predictivos en salud basados en aprendizaje de máquina (machine learning)”. 2025. Accedido: 18 de mayo de 2025. [En línea]. Disponible en: https://www.elsevier.es/es-revista-revista-medica-clinica-las-condes-202-pdf-S0716864022001213
[17] C. Espino Timón, “Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso”. 2017. Accedido: 18 de mayo de 2025. [En línea]. Disponible en: https://openaccess.uoc.edu/bitstream/10609/59565/6/caresptimTFG0117mem%C3%B2ria.pdf
[18] IBM, “¿Qué son las máquinas de vectores de soporte (SVM)?” 2025. [En línea]. Disponible en: https://www.ibm.com/es-es/think/topics/support-vector-machine
[19] J. Amat, “Árboles de decisión, random forest, gradient boosting y C5.0”, Ciencia de Datos. Accedido: 19 de abril de 2026. [En línea]. Disponible en: https://cienciadedatos.net/documentos/33_arboles_decision_random_forest_gradient_boosting_c50
[20] “Chen and C. Guestrin, ‘XGBoost: A Scalable Tree Boosting System,’ en Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.”,
[21] D. J. Matich, “Redes Neuronales: Conceptos Básicos y Aplicaciones”. 2001. Accedido: 18 de mayo de 2025. [En línea]. Disponible en: https://www.frro.utn.edu.ar/repositorio/catedras/quimica/5_anio/orientadora1/monograias/matich-redesneuronales.pdf
[22] *G. James, D. Witten, T. Hastie and R. Tibshirani, An Introduction to Statistical Learning: With Applications in R, 2nd ed. New York, NY, USA: Springer, 2021.*
[23] IBM, “Guía de CRISP-DM de IBM SPSS Modeler”. 2025. Accedido: 18 de mayo de 2025. [En línea]. Disponible en: https://www.ibm.com/docs/es/SS3RA7_18.4.0/pdf/ModelerCRISPDM.pdf
[24] P. Chapman *et al.*, “CRISP-DM 1.0 Step-by-step data mining guide”.
[25] P. J. Gertler, S. Martinez, P. Premand, L. B. Rawlings, y C. M. J. Vermeersch, *Impact Evaluation in Practice*. 2011. doi: 10.1596/978-0-8213-8541-8.
[26] M. Schlotter, G. Schwerdt, y L. Woessmann, “Econometric methods for causal evaluation of education policies and practices: a non-technical guide”, *Educ. Econ.*, vol. 19, nº 2, pp. 109–137, 2011, doi: 10.1080/09645292.2010.511821.
[27] L. J. Keele y R. Titiunik, “Geographic Boundaries as Regression Discontinuities”, *Polit. Anal.*, vol. 23, nº 1, pp. 127–155, 2015, doi: 10.1093/pan/mpu014.
[28] A. Fredriksson y G. M. D. Oliveira, “Impact evaluation using Difference-in-Differences”, *RAUSP Manag. J.*, vol. 54, nº 4, pp. 519–532, oct. 2019, doi: 10.1108/RAUSP-05-2019-0112.
[29] “Methods using econometrics to evaluate the impact of trade promotion activities”, 2022.
[30] V. J. Téllez Buitrago y D. N. Martínez Sánchez, “Método automático para la predicción del avalúo comercial de un inmueble en la ciudad de Bogotá”. 2021. Accedido: 18 de mayo de 2025. [En línea]. Disponible en: https://repository.ucatolica.edu.co/server/api/core/bitstreams/a4db7bcd-edcb-4244-bf62-974db5e40294/content
[31] Y. V. Grajales Alzate, “Modelo de predicción de precios de viviendas en el municipio de Rionegro para apoyar la toma de decisiones de compra y venta de propiedad raíz”. 2019. Accedido: 18 de mayo de 2025. [En línea]. Disponible en: https://repository.upb.edu.co/bitstream/handle/20.500.11912/5285/Modelo%20predicci%C3%B3n%20precios%20viviendas.pdf?sequence=1&isAllowed=y
[32] A. P. Granados Jaimes y G. Higuera Alba, “Análisis del valor del suelo de predios inmersos en plan parcial – caso de estudio Plan Parcial Tres Quebradas – Operación Nuevo Usme en Bogotá”, Trabajo Final de Grado, Universidad Distrital Francisco José de Caldas, Bogotá D.C., 2020.
[33] N. E. León Gómez, “Análisis del valor del suelo del Plan Parcial Algarra en el municipio de Zipaquirá”,

Trabajo Final de Grado Especialización, Universidad Distrital Francisco José de Caldas, Bogotá D.C., 2022.
[34] R. A. Daza Hoyos, “Evaluar el reajuste de tierras en Colombia: la distribución equitativa entre actores”, *Cuad. Vivienda Urban.*, vol. 12, nº 23, 2019, doi: 10.11144/Javeriana.cvu12-23.ertc.
[35] F. H. G. Gómez Huertas, “Estudio de valor para un área del suelo de expansión contiguo a la comuna N. 7 en el municipio de Villavicencio”, Trabajo Final de Grado Especialización, Universidad Santo Tomás, Bogotá D.C., 2023.
[36] I. González Méndez, “Modelo econométrico para la estimación del valor del suelo rural en los municipios de la jurisdicción de la agencia catastral de Cundinamarca”. 2022. Accedido: 18 de mayo de 2025. [En línea]. Disponible en: https://repository.udistrital.edu.co/server/api/core/bitstreams/1ae8812b-d1bf-4f9f-89e0-222ad3a8c978/content
[37] L. M. Vilches-Blázquez, “Análisis del mercado de tierras con técnicas complejas”. 2025. Accedido: 18 de mayo de 2025. [En línea]. Disponible en: https://www.upra.gov.co/es-co/Documents/Mercado_Tierras/art%C3%ADculo%20analisis.pdf
[38] W. A. Melo Cerón, “Determinación de valores de terreno de suelo urbano en Bogotá de acuerdo con las ofertas inmobiliarias de predios NPH”. 2025. Accedido: 18 de mayo de 2025. [En línea]. Disponible en: https://repository.libertadores.edu.co/server/api/core/bitstreams/91e63b27-9f9d-45c2-a51b-e242967616c1/content
[39] Departamento Nacional de Planeación, “Evaluación del impacto de la Ley 388 de 1997 y sus instrumentos sobre el mercado del suelo en las principales ciudades del país”. 2013. Accedido: 18 de mayo de 2025. [En línea]. Disponible en: https://colaboracion.dnp.gov.co/CDT/Vivienda%20Agua%20y%20Desarrollo%20Urbano/POT%20y%20Mercado%20de%20Suelo%20-%20Econometr%C3%ADa.pdf
[40] R. M. Cerino, J. P. Carranza, M. A. Piumetto, M. E. Bullano, V. Caffarati Donalisio, y F. Monzani, “Propuesta para la valuación masiva del suelo urbano. aplicación espacial del algoritmo Quantile Regression Forest”, *Rev. Vivienda Ciudad*, vol. 8, 2021.
[41] Alcaldía Mayor de Bogotá, “Portal de Datos Abiertos Bogotá”. 2025. Accedido: 24 de agosto de 2025. [En línea]. Disponible en: https://datosabiertos.bogota.gov.co/
[42] *T. Hastie, R. Tibshirani y J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009.*
[43] *T. Hastie, R. Tibshirani y J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009.*
[44] *A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed. Sebastopol, CA, USA: O’Reilly Media, 2019, ch. 2.*
[45] “R. Tibshirani, ‘Regression shrinkage and selection via the Lasso,’ Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, pp. 267–288, 1996.”
[46] *T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2016, pp. 785–794.*
[47] “Optimization of random forest model using optuna”, Kaggle. Accedido: 20 de abril de 2026. [En línea]. Disponible en: https://www.kaggle.com/code/mustafagerme/optimization-of-random-forest-model-using-optuna
[48] P. Sun, “Cross-Validation in Time Series”, Medium. Accedido: 20 de abril de 2026. [En línea]. Disponible en: https://medium.com/@pacosun/respect-the-order-cross-validation-in-time-series-7d12beab79a1
[49] *A. Nieto, A. Rodríguez y A. Segura, “Visor geoespacial interactivo para Predicción del valor del suelo en Bogotá” OneDrive, [Video]. 2026.* [En línea Video]. Disponible en: https://javerianacaliedu-my.sharepoint.com/:v:/g/personal/asegurah_javerianacali_edu_co/IQBfEACCMlZVQKqXfy3o90jzAd9IgCN1Oq1

wj_A8MBCOQjI?nav=eyJyZWZlcnJhbEluZm8iOnsicmVmZXJyYWxBcHAiOiJPbmVEcml2ZUZvckJ1c2luZXNzIiwicmVmZXJyYWxBcHBQbGF0Zm9ybSI6IldlYiIsInJlZmVycmFsTW9kZSI6InZpZXciLCJyZWZlcnJhbFZpZXciOiJNeUZpbGVzTGlua0NvcHkifX0&e=NmFuWs