Comparación de técnicas para la estimación del valor comercial de predios en la ciudad de Bogotá.

MAGISTER EN MATEMÁTICAS APLICADAS Y CIENCIAS DE LA COMPUTACIÓN

Raul Andres Rodriguez Trujillo

Dirección:
Nelson Alirio Cruz Gutierrez
Jurgen Daniel Toloza Delgado

Universidad del Rosario
Escuela de Ingeniería, Ciencia y Tecnología
Maestría en Matemáticas Aplicadas y Ciencias de la Computación.

DEDICATORIA

A las dos mujeres más importantes de mi vida, porque sin uds no seria lo mismo. La victoria no tendría el mismo sabor sin ustedes.

AGRADECIMIENTOS

Quiero expresar mi más sincero agradecimiento al profesor Nelson Alirio Cruz, director de este trabajo de grado, por su invaluable guía, su paciencia y por compartir generosamente su conocimiento a lo largo de este proceso. Su acompañamiento constante y su enseñanza han sido fundamentales para la culminación de este proyecto.

Extiendo también un especial reconocimiento al profesor Jurgen Daniel Toloza, co-director del trabajo, por sus aportes, observaciones y orientación que enriquecieron significativamente el desarrollo del estudio.

Agradezco a la Unidad Administrativa Especial de Catastro Distrital, por brindar el acceso a la información necesaria y apoyar el desarrollo de este trabajo, así como a todas las personas que, de una u otra forma, compartieron su experiencia, conocimientos y disposición para aportar al avance de esta investigación.

Finalmente, agradezco a la Universidad del Rosario por ser el espacio académico que me permitió crecer como profesional, formarme críticamente y llevar a cabo este trabajo con los más altos estándares de calidad.

Abstract

Español

Este trabajo de grado tiene como objetivo comparar distintas técnicas de estimación del valor comercial de predios residenciales en Bogotá, a partir de una base de datos construida y depurada con más de 21.000 registros de ofertas inmobiliarias en Bogotá. La metodología empleada contempla una evaluación comparativa entre enfoques tradicionales y modernos de modelación, incluyendo modelos aditivos generalizados (GAM), Random Forest (RF) y LightGBM (LGBM) mediante métricas como el MAPE y R^2. El desarrollo se realizó usando R por su caracteristica de código abierto permitiendo la exploración de otras alternativas y el amplio desarrollo que se encuentra al rededor de los modelos GAM. Tras aplicar los modelos a una muestra representativa de predios y evaluar su desempeño según métricas de error y capacidad predictiva, se encontró que LGBM obtuvo los mejores resultados para predios en propiedad horizontal, mientras que RF fue más preciso en predios en propiedad no horizontal.

Inglés

This undergraduate thesis aims to compare different techniques for estimating the commercial value of residential properties in Bogotá, based on a database constructed and refined with more than 21,000 records of real estate listings in Bogotá. The methodology involves a comparative evaluation of both traditional and modern modeling approaches, including Generalized Additive Models (GAM), Random Forest (RF), and LightGBM (LGBM) through metrics such as MAPE and R^2. The analysis was conducted using R, taking advantage of its open-source nature, which enables the exploration of alternative methods and benefits from the extensive development around GAM models. After applying the models to a representative sample of properties and assessing their performance using error metrics and predictive accuracy, the results showed that LGBM performed best for horizontally owned properties, while RF was more accurate for non-horizontal properties.

Índice

Lista de tablas

Lista de figuras

1. INTRODUCCIÓN

El Decreto 148 de 2020 reglamentó aspectos fundamentales relacionados con la prestación del servicio público de gestión catastral, incorporando la posibilidad de utilizar métodos directos o indirectos para la captura de información, en el marco de los procesos de barrido predial masivo. En particular, dentro de los métodos indirectos se contempla el uso de modelos estadísticos, geoestadísticos y econométricos, así como el análisis de *big data* mediante técnicas de *machine learning*.

Tradicionalmente, las actividades del componente económico, a partir de las cuales se obtienen los avalúos, han requerido una considerable inversión de recursos, tanto en términos de personal como de logística. Estas actividades incluyen la captura de ofertas inmobiliarias, la realización de avalúos puntuales seleccionados a partir de una muestra, y la liquidación de los valores predio a predio conforme a reglas y condiciones previamente definidas. En el sistema SECOP II se encuentran contratos asociados a estas labores por valores hasta los $1 800 000 000.

En este contexto, la adopción de métodos indirectos representa una oportunidad significativa para optimizar la operación catastral, al permitir una reducción sustancial en el uso de recursos humanos y logísticos en los procesos de actualización o conservación catastral. En este sentido, el uso de métodos indirectos basados en modelos estadísticos permite reducir tanto los tiempos de entrega de los productos asociados al componente económico como los costos operativos, promoviendo ejercicios de valoración más eficientes, reproducibles y respaldados en análisis de datos objetivos. En este esquema, el papel de los avaluadores puede reorientarse hacia funciones de control de calidad, encargándose de validar los resultados obtenidos a través de los modelos.

Actualmente, la Unidad Administrativa Especial de Catastro Distrital estima el avalúo de los inmuebles mediante técnicas de valoración masiva que aprovechan, como insumo principal, la información transaccional de ofertas inmobiliarias. Para los predios en propiedad horizontal, se emplea un modelo econométrico hedónico específicamente un modelo lineal generalizado (GLM), en el cual el valor se explica a partir de las características físicas y del entorno. En el caso de la propiedad no horizontal, se usa la suma del componente de la construcción obtenido por medio del método de reposición y el componente del terreno por medio de la zonas homogeneas físicas.

Además de estas metodologías tradicionales, existen múltiples técnicas estadísticas y de machine learning que pueden complementar y, en ciertos casos, superar el desempeño de los enfoques basados en GLM y método valuatorio residual. Por lo que se exploraron tres de ellas: los Modelos Aditivos Generalizados (GAM), que permiten capturar relaciones no lineales y efectos suaves de las variables explicativas; el *Random Forest*, capaz

de capturar relaciones complejas y mitigar el sobreajuste mediante la agregación de múltiples árboles; y *LightGBM* , un algoritmo de boosting que optimiza la velocidad y eficiencia de entrenamiento al mismo tiempo que conserva alta precisión predictiva. La teoria alrededor de estas técnicas es bastante robusta y cuentan con implementaciones en lenguajes de programación en R y Python.

En este sentido, el presente estudio propone comparar de forma sistemática los Modelos Aditivos Generalizados (GAM), el *Random Forest* y *Light Gradient Boosting Machine* (LightGBM), en virtud de las fortalezas y limitaciones de cada enfoque. Para dar cumplimiento a los objetivos planteados, este trabajo se organiza en seis capítulos principales: en los capítulos 2 y 3 se presentan los objetivos general y específicos y se expone la justificación del estudio; el capítulo 4 desarrolla el marco teórico y el estado del arte de las técnicas evaluadas; en el capítulo 5 se detalla la metodología empleada, abarcando la preparación de la base de datos y los criterios de evaluación; el capítulo 6 expone los resultados y hallazgos obtenidos con cada modelo; y, por último, el capítulo 7 reúne las conclusiones derivadas del análisis y plantea recomendaciones para investigaciones futuras.

2. OBJETIVOS

2.1. Objetivo general

Comparar diferentes técnicas de estimación del valor comercial de los predios residenciales en Bogotá.

2.2. Objetivos específicos

1. Construir, depurar y validar una base de datos de ofertas inmobiliarias en Bogotá, que sirva como insumo para el proceso de estimación.

2. Desarrollar modelos estadísticos y de *machine learning* orientados a la estimación del valor comercial de los predios.

3. Evaluar el desempeño de las técnicas empleadas mediante métricas adecuadas de validación y precisión, usando el promedio del error absoluto MAPE y el coeficiente de determinación R^2.

4. Obtener estimaciones del valor comercial de los predios residenciales en Bogotá utilizando los modelos con mejor desempeño.

3. PROBLEMA Y JUSTIFICACIÓN

Los métodos para la aplicación y cálculo de los avalúos de los inmuebles en Colombia cuentan con sustento normativo, principalmente a partir de la Resolución 620 de 2008 del IGAC, el Decreto 148 de 2020 y la Resolución 1040 de 2023 del IGAC, las cuales establecen que:

"El método de comparación o de mercado, es una técnica valuatoria que busca establecer el valor comercial del bien, a partir del estudio de las ofertas o transacciones recientes, de bienes semejantes y comparables al del objeto de avalúo. Tales ofertas o transacciones deberán ser clasificadas, analizadas e interpretadas para llegar a la estimación del valor comercial." [1][2].

Esto quiere decir que para acciones de valoración de los inmuebles se permite el uso de fuentes que contengan información transaccional reciente, como lo indica la técnica de comparación de mercado para la obtención del avalúo.

En línea con lo establecido y como lo dispone [3], los siguientes artículos mencionan y permiten el uso de una gran variedad de técnicas y herramientas para realización el barrido predial masivo:

"ARTÍCULO 2.2.2.2.5. Barrido predial masivo. Es el conjunto de estrategias, actividades y acciones orientadas a conseguir la identificación de las características físicas, jurídicas y económicas de los predios sobre un territorio determinado. El barrido predial masivo comprende diferentes maneras de intervención en el territorio, incluyendo, entre otros, métodos directos e indirectos de captura de información, esquemas colaborativos, uso de registros administrativos, modelos geoestadísticos y econométricos y demás procedimientos técnicos, herramientas tecnológicas e instrumentos de participación comunitaria con enfoque territorial, así como el uso de otras fuentes de información del territorio que permitan obtener los datos necesarios para establecer la línea base de información catastral multipropósito en un municipio, igual que para su mantenimiento y actualización permanente. Los productos derivados de las actividades de barrido predial masivo deberán cumplir con las especificaciones técnicas definidas por la autoridad reguladora."

"ARTÍCULO 2.2.2.2.6. Métodos de recolección de información. Los procesos catastrales podrán adelantarse mediante la combinación de los siguientes métodos:

a) Métodos directos: Aquellos que requieren una visita de campo con el fin de recolectar la realidad de los bienes inmuebles.

b) Métodos indirectos: Son aquellos métodos de identificación física, jurídica y económica de los bienes inmuebles a través del uso de imágenes de sensores remotos, integración

de registros administrativos, modelos estadísticos y econométricos, análisis de Big Data y demás fuentes secundarias como los observatorios inmobiliarios, para su posterior incorporación en la base catastral.

c). Métodos declarativos y colaborativos: Son los derivados de la participación de la comunidad en el suministro de información que sirva como insumo para el desarrollo de los procesos catastrales. Los gestores catastrales propenderán por la adopción de nuevas tecnologías y procesos comunitarios que faciliten la participación de los ciudadanos."

Por lo anterior, se propone en el presente trabajo de grado, a partir de un análisis comparativo, exponer algunas técnicas de estimación del avalúo comercial, de tal manera que sirvan como referente para la valoración de los inmuebles utilizando la información transaccional de ofertas inmobiliarias. Este estudio surge principalmente con la motivación de presentar una gama más amplia de técnicas de estimación, diferentes al método tradicional de modelos lineales generalizados actualmente empleados por la Unidad Administrativa Especial de Catastro Distrital. Este modelo se presenta como una alternativa para la valoración de inmuebles en propiedad horizontal, considerando que, en la actualidad, el proceso se realiza mediante métodos basados en tablas de valores definidos a partir del criterio de personal experto, lo que implica un elevado costo en términos de capital humano.

4. MARCO TEÓRICO Y ESTADO DEL ARTE

Con el paso de los años, la valoración de viviendas ha experimentado una constante evolución. Impulsada por el crecimiento económico y el otorgamiento de subsidios, la compra de vivienda nueva y usada en la región se ha visto estimulada. Este dinamismo ha generado que la estimación puntual del valor comercial sea, en muchas ocasiones, costosa y demorada.

Aunque la aplicación de métodos de avalúo masivos está permitida, se han explorado diversas técnicas de estimación, tales como métodos de regresión lineales y no lineales. Estos métodos, entre otros, han demostrado eficacia en la estimación del valor de las propiedades y en su adaptación a distintos tipos de viviendas y mercados regionales; sin embargo, su implementación en campo suele ser costosa y, además, no incorporan información geográfica como latitud y longitud.

La elección del método de estimación más adecuado depende de las necesidades específicas de la valoración y del contexto del mercado inmobiliario. En este documento se abordarán métodos de regresión como los modelos aditivos generalizados (GAM), *Random Forest* y métodos de *boosting* como *LightGBM* . Además, se considerará la inclusión de algún otro método semiparamétrico o computacional dentro de la comparación, apoyándose en trabajos previos como los de [4], [5], [6], [7] y [8], que servirán de referencia para la elaboración del ejercicio.

Para el desarrollo de los objetivos, se empleará el software R y las librerías relacionadas con las técnicas mencionadas. A continuación, se presenta una breve descripción de cada una de estas técnicas.

1. Modelos Aditivos Generalizados (GAM)

Los modelos aditivos generalizados (GAM), propuestos originalmente por [9], pueden definirse de la siguiente forma, según se describe en [10]:

$$g(\mu_i) = A_i\theta \ + \ f_1(x_{1i}) \ + \ f_2(x_{2i}) \ + \ f_3(x_{3i}, x_{4i}) \ + \ \cdots + f_j(x_{ij})$$

donde $\mu_i \equiv \mathbb{E}[Y_i]$ y $Y_i \sim \text{EF}(\mu_i, \phi)$. Aquí, Y_i es la variable de respuesta y $\text{EF}(\mu_i, \phi)$ denota una distribución de la familia exponencial con media μ_i y parámetro de escala ϕ. La fila A_i corresponde a la matriz de diseño para los componentes estrictamente paramétricos, θ es el vector de parámetros asociado, y las funciones f_j son funciones suaves de las covariables x_k.

La idea de los GAM surge como una generalización de los modelos lineales generalizados, GLM, al incluir funciones suaves como alternativa para modelar características no lineales, [9] propuso el algoritmo de *backfitting* para la estimación. Este algoritmo tiene la ventaja de permitir representar las funciones f_i de un modelo aditivo,

$Y = f_1(X) + f_2(X) + \cdots + f_i(X) + \epsilon$, mediante prácticamente cualquier técnica de suavizado o modelado. La desventaja radica en que la estimación del grado de suavizado es difícil de integrar directamente en el modelo. De acuerdo con [10], la función suave en el modelo GAM se puede representar mediante una base de expansión, cada una con una penalización asociada que controla la suavidad de la función. La estimación puede realizarse mediante métodos de regresión penalizada y, a partir de los datos, es posible estimar el grado de suavizado adecuado para las funciones f_i utilizando validación cruzada generalizada o la maximización de la verosimilitud marginal. El parámetro de suavizado se denota como λ, y este controla el equilibrio entre un buen ajuste del modelo y la suavidad de las funciones.

2. Random Forest (RF)

Random Forest es una técnica de aprendizaje automático basada en el uso de árboles de decisión. Se construye mediante un método de ensamble, entrenando múltiples árboles de decisión sobre diferentes muestras del conjunto de entrenamiento. Este método de ensamble se denomina *bagging* [11]. Según [12], el método consiste en usar muchos conjuntos de entrenamiento, construir un modelo de predicción independiente para cada conjunto y luego promediar las predicciones resultantes, es decir, las funciones $f^{(1)}(x), f^{(2)}(x), \ldots, f^{(B)}(x)$ obtenidas de B muestras bootstrap[13] diferentes, para obtener un único modelo de baja varianza:

$$\hat{f}_{\text{avg}}(x) \;=\; \frac{1}{B}\sum_{b=1}^{B} f^{(b)}(x).$$

El algoritmo considera que en cada división del árbol se selecciona aleatoriamente un subconjunto de m predictores, usualmente $m \approx \sqrt{p}$, donde p es el número total de predictores. Esta selección aleatoria evita que todos los árboles se partan sobre el mismo predictor, reduciendo la correlación entre ellos y logrando que el promedio de sus predicciones resulte en un modelo más estable y con menor varianza.

Además, el *Random Forest* tiene la particularidad de que, al generar las muestras bootstrap para entrenar cada árbol, las observaciones no seleccionadas (conocidas como *out-of-bag*) se utilizan para estimar el error de predicción, denominado error fuera de la bolsa (OOB). Dado que estos datos no participan en el entrenamiento de cada árbol, esta técnica ofrece una estimación confiable del rendimiento del modelo sin requerir un conjunto de validación adicional.

Los parámetros para el modelo *Random Forest* son los siguientes:

- **n_tree:** Determina el número de árboles a entrenar. Un número mayor de árboles suele reducir la varianza del modelo, aunque incrementa el costo computacional.

- **mtry:** Controla la cantidad de predictores considerados en cada nodo. Generalmente, $m \approx \sqrt{p}$.
- **min_node_size:** Controla el tamaño mínimo de los nodos terminales (hojas). Limitar este tamaño restringe la profundidad y complejidad del árbol, ayudando a prevenir el sobreajuste.

3. Light Gradient Boosting Machine (LightGBM)

LightGBM es una técnica de aprendizaje automático que, al igual que *Random Forest* (RF), se basa en árboles de decisión. Al igual que RF, es un método de ensamble que emplea un esquema de *boosting* específico. Este método recibe el nombre de *Gradient Boosting Decision Trees*, pues funciona bajo el principio del *boosting*, que consiste en entrenar múltiples modelos de manera secuencial, donde cada nuevo modelo intenta corregir o reducir los errores de los modelos anteriores; en este caso, cada modelo es un árbol de decisión. Durante cada etapa, el nuevo árbol se ajusta para corregir los errores cometidos por el conjunto previo, utilizando como guía el gradiente de la función de pérdida elegida [14].

En particular, esta técnica se caracteriza por su enfoque *leaf-wise* para el crecimiento de los árboles, a diferencia del enfoque *level-wise* empleado por otras implementaciones como el *XGBoost*[15], la figura 1, ilustra este principio. Este enfoque permite optimizar el uso de memoria y alcanzar una mayor eficiencia en comparación con métodos tradicionales. De acuerdo con [16], el algoritmo implementa dos técnicas para aprovechar esta estructura: *Gradient-based One Side Sampling* (GOSS) y *Exclusive Feature Bundling* (EFB). GOSS prioriza las instancias con gradientes más altos para reducir la cantidad de datos sin sacrificar la precisión, mientras que EFB agrupa características mutuamente excluyentes para disminuir la dimensionalidad.

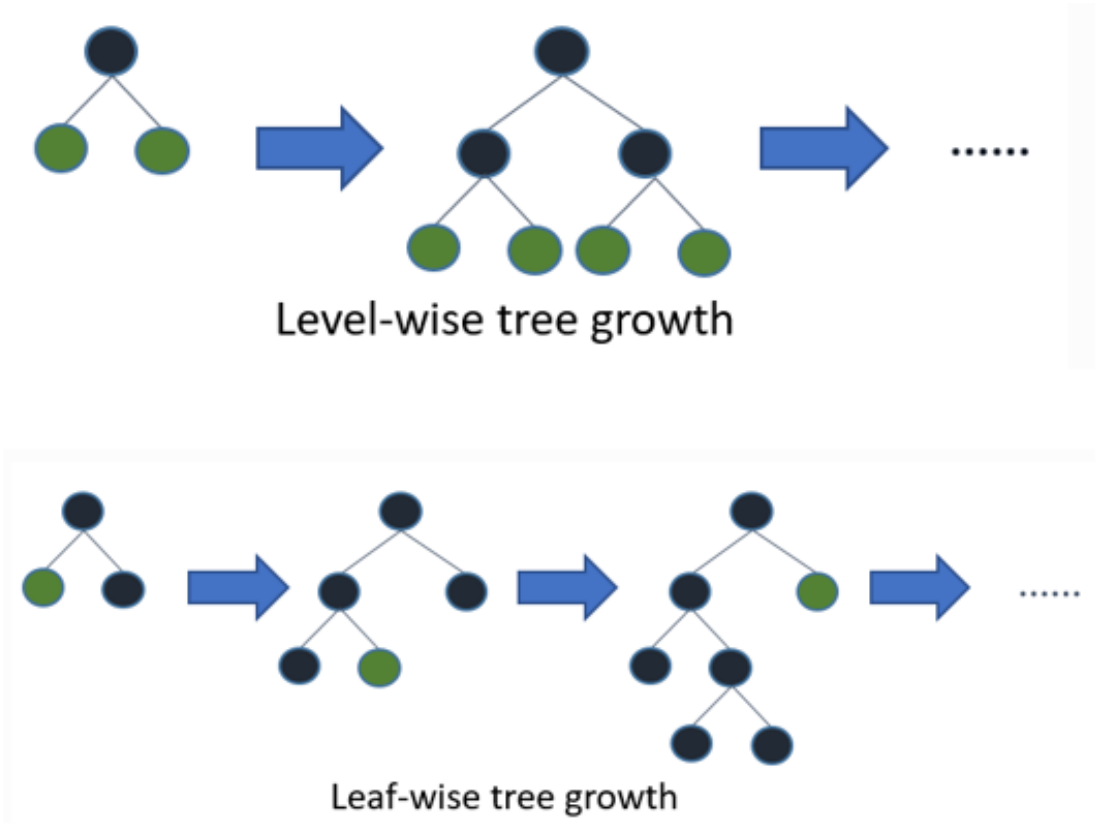

Figura 1: Diagrama del enfoque de arboles de decisión Level Wise y Leaf Wise.
Fuente: Lightgbm features.

Así, *LightGBM* surge como una alternativa para manejar grandes volúmenes de datos de forma rápida y eficiente en recursos computacionales, sin sacrificar precisión. Los parámetros del modelo *LightGBM* utilizados en este estudio son los siguientes[17]:

- **mtry:** Porcentaje de predictores sorteados en cada división. Controla la *submuestra de variables* y actúa como regularización de alto nivel.
- **min_n:** Tamaño mínimo de observaciones que debe contener una hoja. Evita hojas demasiado pequeñas y, por ende, el sobreajuste.
- **tree_depth:** Profundidad máxima permitida para cada árbol que crece de forma *leaf-wise*. Reduce la complejidad del modelo.
- **trees:** Número total de árboles (iteraciones de *boosting*).
- **learn_rate:** Tasa de aprendizaje; valores más bajos suavizan cada actualización, requiriendo generalmente más iteraciones.

- **loss_reduction:** Ganancia mínima necesaria para permitir una nueva división; previene particiones irrelevantes.
- **sample_size:** Fracción de filas usadas en cada iteración; activa el *submuestreo*, que reduce la varianza y acelera el ajuste.
- **stop_iter:** Número de iteraciones sin mejora en la métrica de validación tras el cual se detiene el entrenamiento anticipadamente.

Ademas de las técnicas anteriormente descritas, se exploraron otras técnicas para modelar la dependencia espacial en datos geográficos. Basados en modelos de tipo *Random Forest*, se han hecho estudios alrededor de este principio para ofrecer otras alternativas a la hora de modelar datos con información espacial. Los paquetes mencionados a continuación son propuestas desarrolladas en el software estadístico R [18]. Una aproximación destacada es el método implementado en el paquete `spatialML`, fundamentado en el enfoque de *Geographical Random Forest (GRF)*, que busca capturar la heterogeneidad espacial a través de múltiples modelos locales basados en bosques aleatorios [19]. Por su parte, el paquete `meteo` desarrolla el método *Random Forest Spatial Interpolation (RFSI)*, incorporando observaciones cercanas y distancias espaciales en la interpolación, destacando el algoritmo en su precisión y eficiencia computacional [20]. Asimismo, el paquete `RandomForestsGLS` introduce una aproximación que utiliza mínimos cuadrados generalizados (GLS) para manejar la dependencia espacial mediante procesos gaussianos, ofreciendo una alternativa para estimar efectos no lineales en presencia de correlación espacial [21].

Tabla 1: Comparación de métodos revisados y justificación de su uso en el estudio. Fuente: Elaboración propia.

Método	**Ventajas**	**Limitaciones**	**Justificación en este estudio**
Modelos Aditivos Generalizados (GAM)	Capturan relaciones no lineales; interpretables; permiten suavizadores para covariables continuas	Requieren selección cuidadosa del parámetro de suavizado; sensibles a extrapolaciones	Incluido: amplia aplicación previa en valoración inmobiliaria y buena interpretabilidad
Random Forest (RF)	Robusto a ruido y variables irrelevantes; bajo riesgo de sobreajuste; no requiere supuestos paramétricos fuertes	Menor interpretabilidad; posible pérdida de precisión en extrapolaciones	Incluido: probado desempeño en problemas de predicción inmobiliaria
LightGBM	Alta velocidad de entrenamiento; eficiente con grandes volúmenes de datos; buen manejo de datos desbalanceados	Requiere ajuste fino de hiperparámetros; menor interpretabilidad que GAM	Incluido: alta eficiencia y precisión, adecuado para datos con múltiples variables explicativas
Geographical Random Forest (GRF)	Captura heterogeneidad espacial; genera modelos locales	Documentación limitada; alta demanda computacional	No priorizado: falta de estabilidad y soporte en bibliotecas actuales
Random Forest Spatial Interpolation (RFSI)	Incorpora distancias espaciales; buena precisión en interpolación	Aplicación restringida a problemas de interpolación; requiere coordenadas de alta precisión	No priorizado: objetivo del estudio no centrado en interpolación espacial pura
RandomForestsGLS	Maneja correlación espacial mediante procesos gaussianos	Implementación compleja; tiempo de cómputo elevado	No priorizado: alta complejidad y poca evidencia en el dominio inmobiliario

No obstante, aunque estas técnicas son robustas y novedosas, su reciente desarrollo implica desafíos relacionados con la estabilidad y reproducibilidad al aplicarse en diferentes contextos y conjuntos de datos. Adicionalmente, la documentación disponible es aún limitada, basándose principalmente en los artículos de referencia, con pocos ejemplos prácticos, lo cual dificulta la interiorización del aprendizaje para nuevos usuarios. Además, no todas las funciones implementadas en los paquetes mencionados se encuentran completamente operativas. Por esta razón, en la Tabla 1 se presenta un resumen comparativo de los principales métodos revisados, destacando sus ventajas, limitaciones y la justificación de su inclusión o exclusión en el presente estudio.

Para la comparación del desempeño de los modelos se utilizaron el Error Porcentual Absoluto Medio (MAPE) y el coeficiente de determinación (R^2).

El **Mean Absolute Percentage Error** (MAPE)[22] es una métrica estadística que cuantifica el error relativo medio de un modelo de predicción en términos porcentuales, midiendo la magnitud promedio de los errores absolutos respecto a los valores observados. Formalmente, para un conjunto de N observaciones, con valores reales y_i y valores estimados $\hat{y}_i$, el MAPE se define como:

$$\text{MAPE} = \frac{100}{N} \sum_{i=1}^{N} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

donde:

- y_i es el valor real de la observación i,
- $\hat{y}_i$ es el valor estimado por el modelo,
- N es el número total de observaciones.

El MAPE es adimensional, presenta una interpretación directa como porcentaje de error medio, y es invariante a cambios de escala en la variable dependiente. No obstante, puede ser sensible cuando y_i se aproxima a cero, lo que debe ser considerado en su interpretación. Ademas permite expresar el error de predicción en términos porcentuales, lo que facilita su interpretación y comparación directa con el valor comercial de los inmuebles; esta métrica es ampliamente utilizada en el ámbito de valoración inmobiliaria por su claridad y aplicabilidad práctica [4, 5, 7].

El R^2, por su parte, mide la proporción de variabilidad explicada por el modelo y el ajuste del modelo.

El **coeficiente de determinación** (R^2) es una métrica que evalúa la proporción de la variabilidad de la variable respuesta que es explicada por el modelo de predicción. Para un conjunto de N observaciones, se define como:

$$R^2 = 1 - \frac{\sum_{t=1}^{N} (y_t - \hat{y}_t)^2}{\sum_{t=1}^{N} (y_t - \bar{y})^2}$$

donde y_t es el valor observado, $\hat{y}_t$ el valor predicho y $\bar{y}$ el promedio de los valores observados. En modelos como *Random Forest* y *LightGBM*, un R^2 elevado indica que el ensamble de árboles captura de forma efectiva la relación entre las variables predictoras y la respuesta. Sin embargo, un valor alto no garantiza ausencia de sobreajuste, por lo que debe interpretarse junto a otras métricas de validación.

En *Modelos Aditivos Generalizados*[10], además de un coeficiente de determinación ajustado ($r.sq$) calculado de forma análoga al R^2, se emplea la **devianza explicada**, como alternativa especialmente útil en el caso de datos que no siguen una distribución

normal. La **devianza** (D) mide la discrepancia entre el modelo ajustado y el modelo saturado y se define como:

$$D = 2\left[l(\hat{\beta}_{\max}) - l(\hat{\beta})\right]\phi$$

donde $l(\hat{\beta}_{\max})$ es la log-verosimilitud del modelo saturado, $l(\hat{\beta})$ es la log-verosimilitud del modelo ajustado y ϕ es el parámetro de escala. La **devianza explicada** se obtiene como la proporción de la devianza nula que es explicada por el modelo, y se interpreta como una medida análoga al R^2 pero adaptada a distribuciones no gaussianas.

En este contexto, el uso conjunto de MAPE y R^2 proporciona una evaluación complementaria que, en términos relativos, resulta más sencilla de interpretar para comparar el desempeño entre diferentes modelos o configuraciones. El MAPE ofrece una medida del error en porcentaje, lo que facilita la comunicación de los resultados a públicos no técnicos, mientras que el R^2 resume la proporción de variabilidad explicada por el modelo, brindando una perspectiva más global de su capacidad predictiva.

Si bien el $RMSE$ es una medida adecuada y ampliamente utilizada, especialmente en contextos donde los errores grandes deben penalizarse de forma más severa, su interpretación directa puede ser menos intuitiva cuando los valores de la variable respuesta presentan una alta dispersión o diferentes escalas. En tales casos, aunque el $RMSE$ sigue siendo útil para la comparación técnica entre modelos, puede no ser tan claro para la toma de decisiones prácticas, razón por la cual se privilegia la inclusión de métricas relativas como el MAPE para complementar el análisis.

5. METODOLOGÍA

Para el desarrollo de los objetivos, se propone seguir el siguiente esquema:

1. **Preparación, depuración y análisis descriptivo de la información:**

   A partir de la información de ofertas inmobiliarias suministrada por la Unidad Administrativa Especial de Catastro Distrital, se construyó una base de datos a nivel de predio que integra variables físicas como el área construida, el área del terreno, la antigüedad y la calificación del inmueble con variables geográficas, tales como las coordenadas del lote correspondiente. Estos datos se encuentran disponibles en el portal de Datos Abiertos de Bogotá [23]. Una vez consolidada la base de datos, se realizará un control de consistencia en términos del valor comercial registrado. Esta etapa incluirá además la elaboración de gráficos descriptivos para las variables que serán utilizadas en el proceso de modelado. Adicionalmente, se realizará la imputación de datos faltantes mediante el método de los **K-nearest neighbors (KNN)**, basado en el principio de que las observaciones con características similares tienden a presentar valores próximos entre sí. Este enfoque permite aprovechar la información de vecinos cercanos para estimar los valores faltantes de manera precisa y coherente con la estructura de los datos. La implementación se realizará utilizando la función correspondiente del paquete `VIM` en R [24], que ofrece una forma flexible y eficiente de aplicar KNN en conjuntos de datos con múltiples variables. Además permite estimar los datos faltantes de manera precisa sin asumir distribuciones paramétricas, respetando la estructura multivariante del conjunto de datos.

2. **Evaluación y selección de técnicas:**

   Partiendo de la idea de que el avalúo puede expresarse en función de sus características, se propone el siguiente modelo:

   $$Y_i = f(\mathbf{X}_i; \boldsymbol{\theta}) + \varepsilon_i,$$

   donde Y_i representa el avalúo del predio i, $\mathbf{X}_i$ es el vector de características, $\boldsymbol{\theta}$ el vector de parámetros y ε_i el término de error. Esta expresión servirá como guía para la estimación del avalúo, independientemente de la técnica empleada.

   Se evaluaron técnicas estadísticas, como los modelos aditivos generalizados (*GAM*), y métodos de *machine learning*, como *Random Forest* y *LightGBM*, utilizando métricas de desempeño tales como el error porcentual absoluto medio (MAPE) y el coeficiente de determinación (R^2). Para los modelos *Random Forest* y *LightGBM* se exploraron diferentes combinaciones mediante una búsqueda en grilla. En el

caso de *Random Forest*, se ajustó el número de árboles y la cantidad de predictores, mientras que para *LightGBM* se optimizó la profundidad máxima del modelo. Estos parámetros se definieron a través de un ajuste de hiperparámetros con validación cruzada sobre la muestra de entrenamiento, seleccionando la combinación que maximizó el desempeño y evitó el sobreajuste. La configuración final se determinó priorizando la capacidad de generalización en la muestra de prueba. Adicionalmente, se aplicaron transformaciones a las variables con el objetivo de obtener la mejor estimación posible del valor comercial.

3. **Análisis comparativo:**

   En esta etapa se realizará un análisis comparativo que permitirá identificar las fortalezas y debilidades de cada técnica, basándose en la evaluación previa. Esta comparación no solo cuantificará el rendimiento de los métodos, sino que también destacará sus ventajas y limitaciones en diferentes contextos y tipos de propiedades. Por ejemplo, algunas técnicas podrían ser más adecuadas para la predicción de precios en propiedades residenciales en propiedad horizontal, mientras que otras podrían funcionar mejor para inmuebles en propiedad no horizontal. Así, se obtendrá un panorama claro del comportamiento de cada método bajo diversas condiciones y tipos de datos inmobiliarios.

4. **Modelado final:**

   Finalmente, se procederá a la estimación definitiva del valor comercial utilizando la base completa de predios residenciales disponible en el portal de datos abiertos Bogotá [23]. Las técnicas que hayan demostrado un mejor desempeño para predios en propiedad horizontal y no horizontal serán aplicadas sobre el conjunto total de datos residenciales de Bogotá. Las estimaciones obtenidas serán comparadas con los valores calculados por la Unidad Administrativa Especial de Catastro Distrital (UAECD), que servirán como referencia para evaluar la precisión y validez de los modelos seleccionados.

6. RESULTADOS Y DISCUSIÓN

En esta sección se expone el proceso de análisis y modelado estadístico llevado a cabo con el fin de explorar y comprender las relaciones existentes entre las variables del estudio. El análisis comienza con una descripción detallada de los datos utilizados, incluyendo las principales variables consideradas, su significado y relevancia dentro del contexto de la investigación. Posteriormente, se detalla el procedimiento implementado para la imputación de datos faltantes y la identificación de valores atípicos. En lo sucesivo, se empleará la denominación *Casas* para referirse a los predios en propiedad no horizontal (NPH) y la de *Apartamentos* para los predios en propiedad horizontal (PH).

6.1. Descripción de la base datos

La base de datos consta de un total de 21549 ofertas inmobiliarias capturadas durante octubre del año 2022 y agosto del 2023 por parte de la Unidad Administrativa Especial de Catastro Distrital, compuesta por 20 variables. Estas ofertas cubren gran parte del territorio distrital, representando diversos sectores del distrito. Incluyen información de todos los estratos socioeconómicos, desde los más bajos hasta los más altos. Las ofertas corresponden a propiedades de tipo residencial, distribuidas principalmente entre apartamentos y casas.

Se presentan a continuación las definiciones de las variables utilizadas.

- **código barrio**: código único compuesto por seis dígitos, utilizado para la identificación del sector.
- **código manzana**: código de dos dígitos que identifica la manzana catastral a la que pertenece el predio.
- **código predio**: número que identifica el predio dentro de la manzana, compuesto por tres dígitos.
- **código construcción**: número de edificación dentro de una propiedad horizontal (PH) o el número de mejora de un predio no propiedad horizontal (NPH). Este código está compuesto por 13 dígitos: los seis primeros corresponden al sector catastral, los dos siguientes al número de manzana, los dos siguientes al número del predio, y los tres últimos a la edificación.
- **código resto**: código de 18 dígitos que identifica el piso donde se localiza la unidad y la ubicación dentro de dicho piso. Los seis primeros dígitos corresponden al sector catastral, seguidos por dos del número de manzana, dos del número del predio, tres de la edificación y cinco del código resto.

- **área construida**: área construida del predio, expresada en metros cuadrados.
- **área terreno**: superficie del terreno del predio, expresada en metros cuadrados.
- **puntaje**: puntaje total que corresponde a la calificación del predio. El puntaje se resume en una medida otorgada por un experto avaluador basada en los acabados del predio.
- **edad**: estado de la estructura del predio, basado en la antigüedad de la edificación.
- **clase predio**: clasificación del predio según su régimen, ya sea propiedad horizontal o no propiedad horizontal.
- **zhf**: código de la zona homogénea física (zhf), que define un espacio geográfico con características similares en términos de norma de uso del suelo, actividad económica, topografía, vías y servicios públicos. Se extrae únicamente el componente topográfico, tomando la posición 6, donde 1 es plano, 2 es empinado y 3 inclinado.
- **coordenada x**: coordenada geográfica plana que representa la longitud del predio, referenciada en el sistema EPSG:6247.
- **coordenada y**: coordenada geográfica plana que representa la latitud del predio, referenciada en el sistema EPSG:6247.
- **cantidad habitaciones**: número de habitaciones del predio.
- **cantidad baños**: número de baños del predio.
- **garajes**: número de garajes del predio.
- **código localidad**: código que identifica la localidad administrativa a la que pertenece el predio.
- **estrato**: código del estrato socioeconómico del predio, con valores entre 0 y 6. La estratificación clasifica la población en grupos con características socioeconómicas similares, basándose en aspectos físicos de las viviendas, entorno y el contexto urbanístico o rural.
- **valor**: valor del inmueble en pesos, basado en la oferta inmobiliaria capturada.
- **valor terreno**: valor del terreno en pesos, extraído de la oferta inmobiliaria.

Ademas de las variables anteriomente mencionadas, se construyen variables tales como, el índice de construcción, definido como la proporción entre el área construida y el área de terreno. Por último, la variable objetivo para el estudio es el valor del inmueble, denotada como "valor".

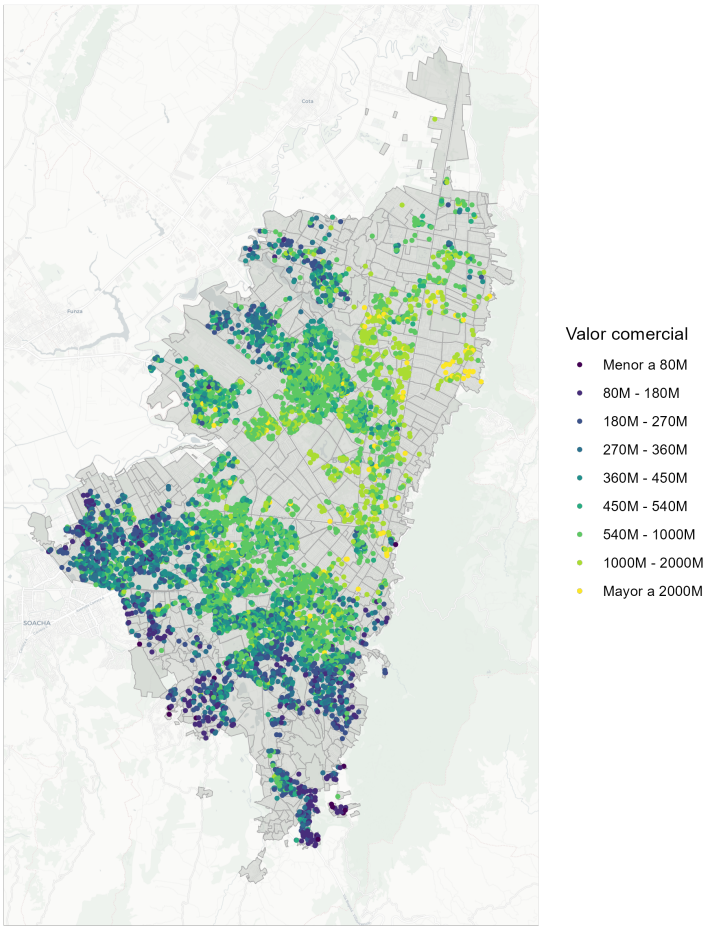

Figura 2: Distribución de ofertas para las casas por rangos de Valor del inmueble.
Fuente: Elaboración propia.

La figura 2 presenta la distribución espacial del valor comercial de las casas o propiedades no horizontales en Bogotá, clasificado en nueve rangos de precios que van desde valores inferiores a 80 millones de pesos hasta superiores a 2000 millones. Cada punto en el mapa representa un predio de este tipo, coloreado de acuerdo con el rango de valor correspondiente. Se observa un patrón geográfico claro: los valores más bajos (tonos oscuros) predominan en el suroccidente y parte del occidente de la ciudad, mientras que los valores intermedios a altos (tonos verdes y amarillos) se concentran en sectores del nororiente y zonas centrales. La presencia de valores elevados en el nororiente y áreas centrales puede asociarse con factores como la alta demanda inmobiliaria, el acceso a servicios, la infraestructura urbana y las condiciones socioeconómicas predominantes. Por el contrario, los valores más bajos en el sur y suroccidente reflejan áreas con menor valorización relativa, posiblemente relacionadas con menor densidad de desarrollos de alto valor, menores niveles socioeconómicos y menor presión de demanda en el mercado inmobiliario.

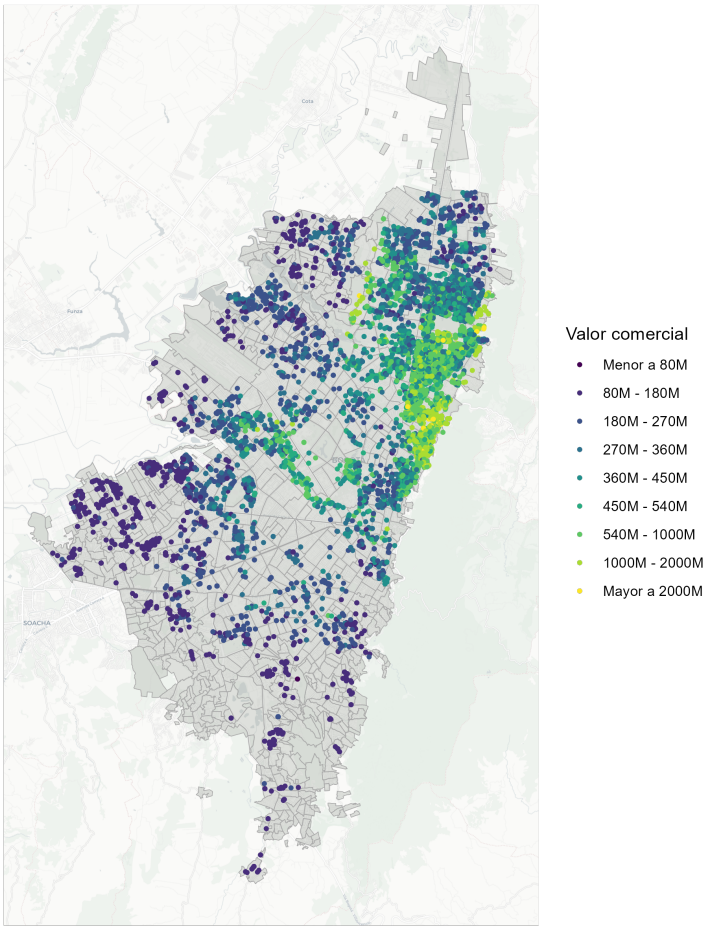

Figura 3: Distribución de ofertas para los apartamentos por rangos de Valor del inmueble.
Fuente: Elaboración propia.

La figura 3 muestra la distribución espacial del valor comercial de los apartamentos o propiedades en régimen de propiedad horizontal en Bogotá, clasificado en nueve rangos de precios que van desde valores inferiores a 80 millones de pesos hasta superiores a 2000 millones. A diferencia de las casas, cuya presencia es más dispersa en el territorio, los apartamentos se concentran principalmente en el eje oriente–nororiente y en áreas puntuales del occidente. Los valores más bajos (tonos oscuros) predominan en sectores periféricos del occidente y sur, mientras que los valores intermedios y altos (tonos verdes y amarillos) se agrupan en corredores residenciales de mayor densidad, especialmente en el nororiente y zonas centrales. Esta distribución responde, en gran medida, a la localización de desarrollos inmobiliarios verticales, la cercanía a centros de empleo y servicios, así como a la concentración de estratos socioeconómicos medios y altos en estas áreas.

6.2. Depuración e imputación de datos faltantes

Para el ejercicio se han definido criterios específicos para la selección de inmuebles con el objetivo de garantizar que los datos representen propiedades típicas del mercado y evitar sesgos causados por valores atípicos. En primer lugar, se incluyeron inmuebles con valores inferiores a 3 000 millones de pesos, dado que inmuebles por encima de este umbral son casos extremos que no representan el comportamiento general del mercado. Adicionalmente, se estableció un límite de área de terreno de 400 m^2 para excluir propiedades atípicas, que pueden distorsionar el análisis, así como también observaciones con áreas construidas superiores a 650 m^2. Finalmente, se limitaron los datos a propiedades con antigüedad inferior a 100 años o con estrato 0, con el fin de descartar inmuebles que podrían ser considerados patrimonio histórico o cultural, los cuales tienen características particulares que no reflejan el comportamiento típico del mercado inmobiliario. En total, luego de aplicar los criterios de exclusión, se eliminaron aproximadamente el 6 % de los datos por considerarse atípicos.

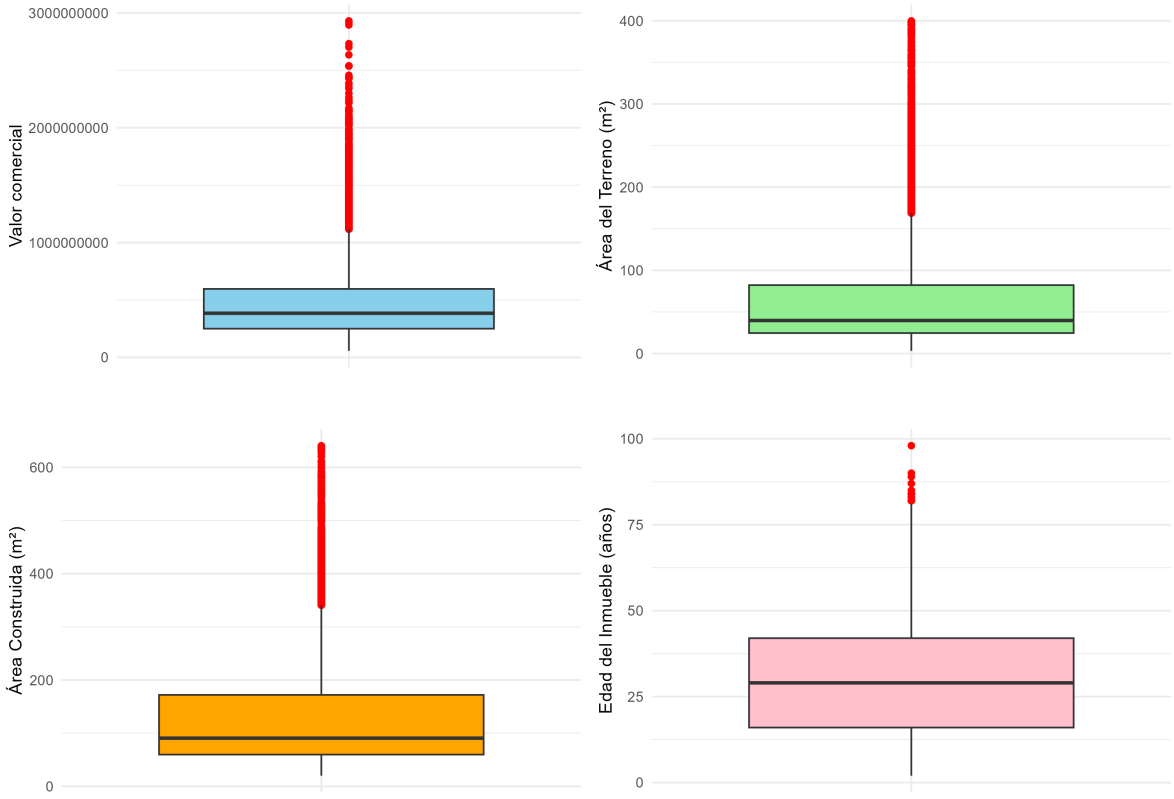

Figura 4: Observaciones sin valores extremos para el valor comercial, Área de terreno ,Área construida y Edad. Fuente: Elaboración propia.

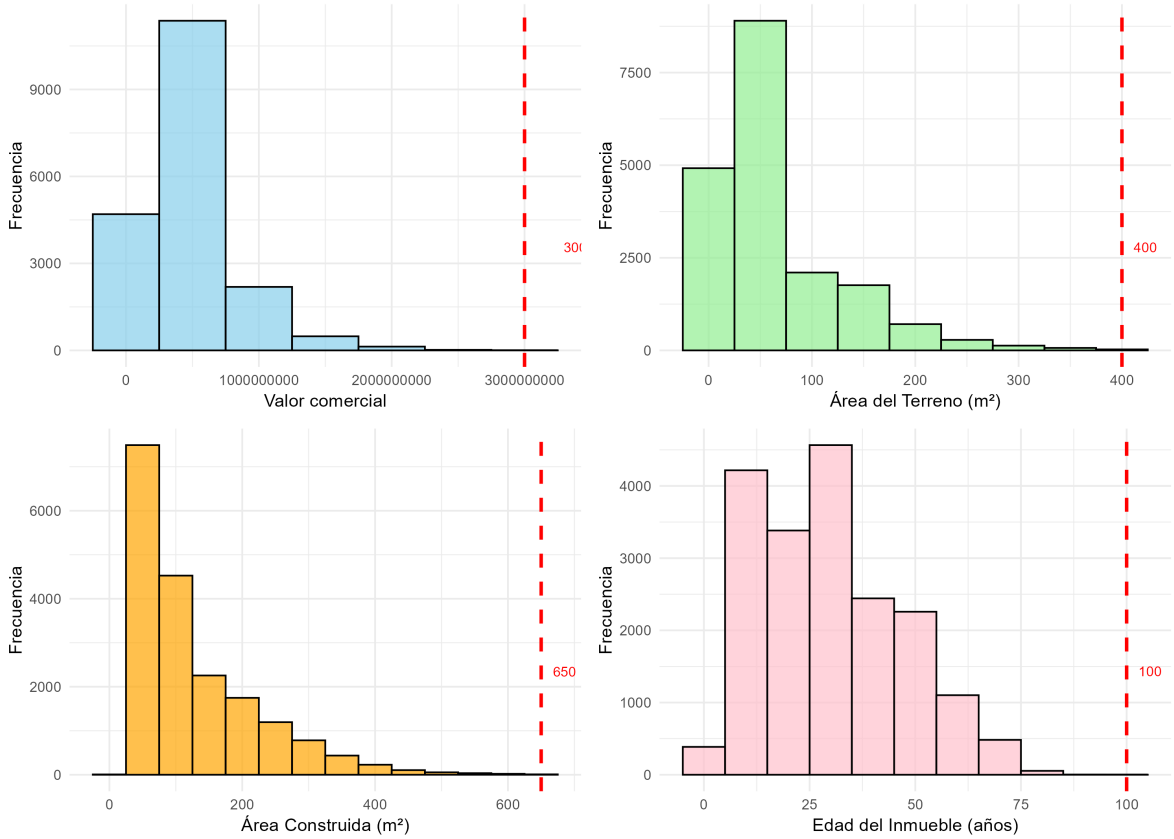

Figura 5: Distribución del valor comercial, Área de terreno, Área construida y Edad. Fuente: Elaboración propia.

Los gráficos realizados muestran las distribuciones de las variables. En la figura 5 se revela cómo los datos se concentran dentro de los límites establecidos y cómo los valores atípicos se encuentran situados a la derecha de los cortes, representados por las líneas verticales. Esto es corroborado desde los diagramas de cajas de la figura 4, donde los puntos individuales son referenciados fuera de los límites de las cajas.

En relación con el manejo de los datos faltantes, se identificó que las variables *cantidad baños*, *cantidad habitaciones* y *garajes* presentan valores ausentes. Dado que estas variables se consideran relevantes para el proceso de modelación, se implementó el método de imputación KNN, propuesto en la librería VIM [24]. Esta técnica aprovecha la información disponible en variables relacionadas, como *área construida* y *valor comercial*, para estimar los valores faltantes de manera precisa. Para este propósito, se utilizó un valor predeterminado de $k = 5$, es decir, los 5 registros más parecidos, y la mediana como función de agregación, lo que permite reducir sesgos y garantizar estimaciones más centradas y representativas. La elección del KNN se debe por su capacidad para capturar relaciones locales entre observaciones similares y flexibilidad al no reque-

rir suposiciones paramétricas destacandose en su efectividad en datos multivariantes heterogéneos, como los atributos de predios residenciales.

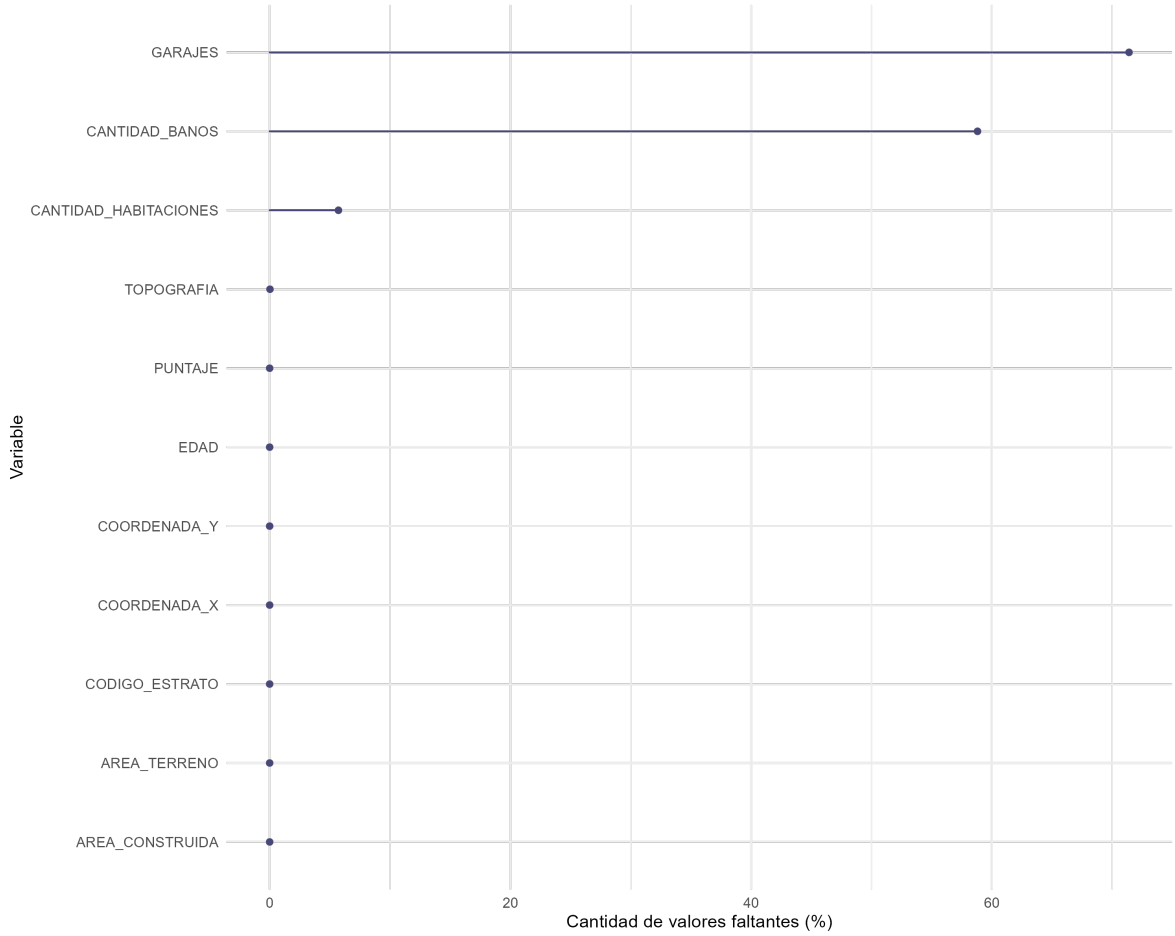

Figura 6: Distribución de valores faltantes por variable. Fuente: Elaboración propia.

Aunque la proporción de valores a imputar en las variables *cantidad baños* y *garajes* es considerablemente alta (superior al 80 % en ambos casos), solo se considerará una de estas variables en el proceso de modelación, priorizando aquella que muestre un mejor ajuste y una mayor contribución al desempeño del modelo. La Figura 6 muestra, en términos porcentuales, la magnitud de los datos faltantes para cada variable incluida en el análisis. Se observa que, además de las dos variables mencionadas, *cantidad habitaciones* presenta un porcentaje reducido de valores ausentes, mientras que el resto de variables, tales como *topografía*, *puntaje*, *edad*, *coordenadas*, *código de estrato*, *área de terreno* y *área construida*, prácticamente no presentan valores faltantes.

6.3. Construcción del Modelo

Finalizado el proceso de depuración, se obtuvo una base para modelación de 18896 observaciones, distribuidas en 12437 apartamentos y 6459 casas. Durante la exploración

se detectó que no existen ofertas inmobiliarias para el estrato 1 en apartamentos, lo cual tiene sentido dado que son inmuebles que no se ofertan con frecuencia. Para iniciar el proceso de modelación, se tuvieron en cuenta algunas consideraciones relacionadas con el comportamiento de los tipos de inmuebles que predominan en el mercado inmobiliario. La separación de los datos en apartamentos y casas responde a las particularidades que estos tipos de inmuebles presentan en el mercado. En el caso de los apartamentos, al tratarse de propiedades horizontales, el área de terreno no es un factor determinante para su valoración, dado que dicha área es compartida entre las unidades y su impacto en el valor comercial es generalmente homogéneo. Por otro lado, en las casas, que suelen ser propiedades no horizontales, el área de terreno es una variable a tener en cuenta en la determinación del valor del inmueble, debido al uso exclusivo que el inmueble tiene sobre el área, así como a la versatilidad de uso y la posibilidad de expansión en futuras edificaciones. Por lo tanto, en la Tabla 2 se muestran las variables seleccionadas para los modelos de apartamentos y casas.

Variable	**Casas (NPH)**	**Apartamentos (PH)**
AREA_CONSTRUIDA	X	X
AREA_TERRENO	X	
COORDENADA_X	X	X
COORDENADA_Y	X	X
PUNTAJE	X	X
EDAD	X	X
TOPOGRAFIA	X	X
CODIGO_ESTRATO	X	X

Tabla 2: Variables empleadas en los modelos de apartamentos y casas.
Fuente: Elaboración propia.

Para la comparación de las tres técnicas de valoración a partir de la estimación del avalúo de apartamentos y casas, se realiza una partición inicial de los datos en una muestra de entrenamiento y otra de prueba, asignando el 90 % de los datos a la muestra de entrenamiento y el 10 % restante a la de prueba. Con el fin de garantizar que ambas muestras sean representativas de la población original, se empleó como variable de control la combinación *Estrato–Localidad*, preservando así la distribución espacial en ambas particiones.

La estructura que se utilizará para el modelo GAM, incluirá las coordenadas geográficas, el área construida y el área de terreno como variables que entran al modelo como variables suaves, para capturar relaciones no lineales respecto a la variable respuesta, mientras que las demás variables se incluirán como efectos lineales. El modelo GAM se complementará utilizando la distribución Gamma con enlace logarítmico, lo que garantiza valores definidos y positivos para la variable respuesta.

En cuanto a los modelos *Random Forest* y *LightGBM*, con el fin de capturar adecua-

damente el efecto espacial, se incorporaron las coordenadas geográficas tanto de forma independiente como mediante su interacción. Las demás variables no tendrán ninguna transformación. Por último, se aplicará la transformación del logaritmo natural a la variable respuesta. Además, para evitar sobreajuste en la selección del modelo, se realizará validación cruzada con 10 folds.

Para entender el comportamiento de la variable respuesta, se ilustra en la Figura 7 la distribución según el tipo de inmueble.

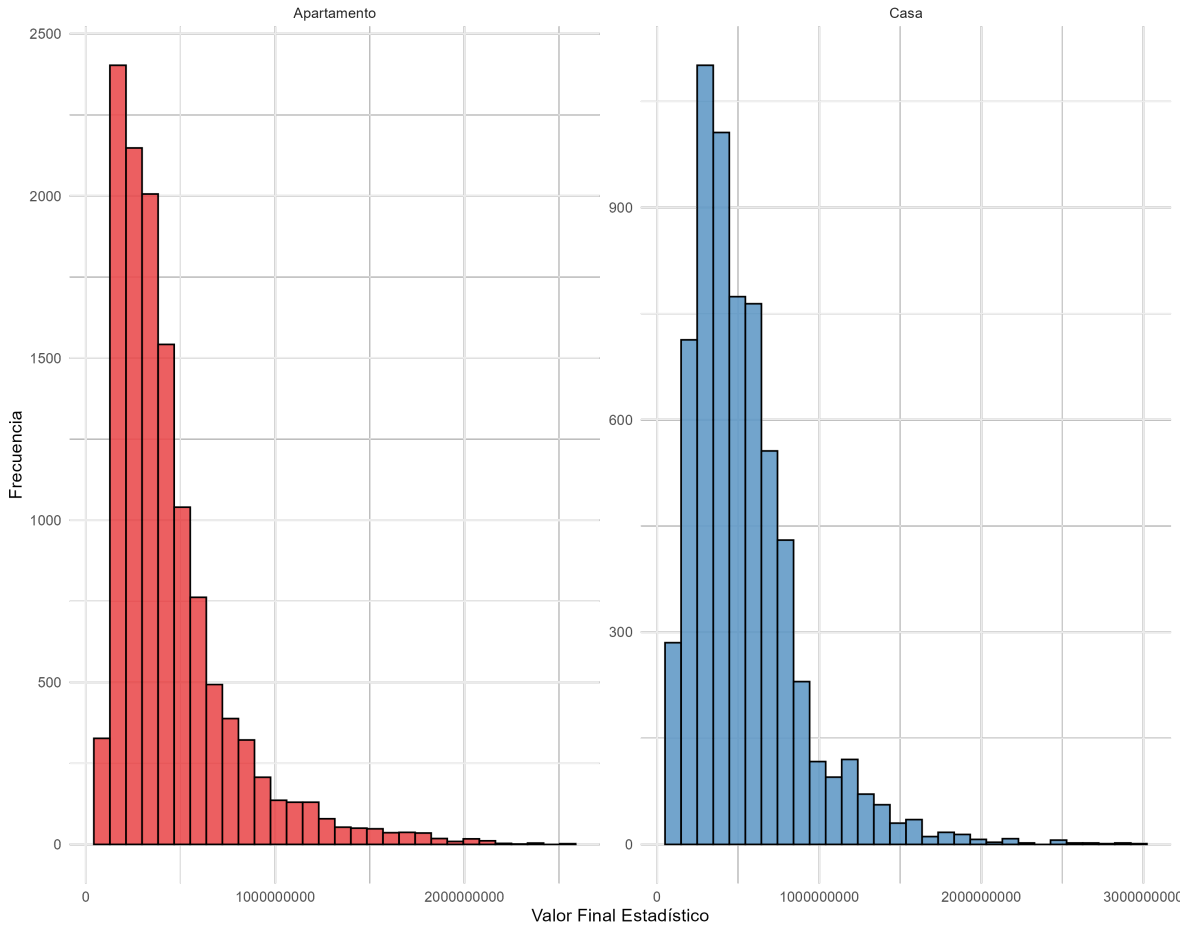

Figura 7: Distribución del valor comercial por tipo de inmueble. Fuente: Elaboración propia.

6.3.1. Apartamentos

Modelo GAM

Para los apartamentos, siguiendo la estructura del modelo GAM, la configuración del modelo es la siguiente.

- **Variable dependiente:** *VALOR*, que representa el valor comercial del inmueble.

- **Términos no paramétricos:**
  - `s(COORDENADA_X, COORDENADA_Y, bs = "gp", k = 50)`: Capturan las relaciones espaciales mediante un suavizamiento proveniente de un proceso gaussiano (`gp`) de dimensión 50.
  - `s(AREA_CONSTRUIDA, k = 10)`: Modela la relación no lineal entre el área construida y el valor comercial, usando una base de dimensión 10.
- **Términos paramétricos:**
  - `PUNTAJE, EDAD, TOPOGRAFIA, CODIGO_ESTRATO`.
- **Familia y enlace:**
  - El modelo utiliza una distribución Gamma, acorde a la naturaleza de la variable respuesta. Para garantizar que los valores sean siempre positivos, se utiliza el enlace logarítmico (`log`).
- **Método de ajuste:**
  - Se emplea el método de máxima verosimilitud restringida (REML), para prevenir el sobreajuste en los términos suavizados.

El modelo ajustado para apartamentos, permite interpretar los coeficientes en términos multiplicativos tras aplicar la transformación exponencial (e^{β}), en ese modo se puede saber en cuantas unidades al aumentar el valor de una unidad por ejemplo en la variable puntaje incrementa en el valor del inmueble. Lo cual destaca al momento de la interpretación y explicación del valor de los inmuebles teniendo en cuenta las variables que inciden en su cálculo.

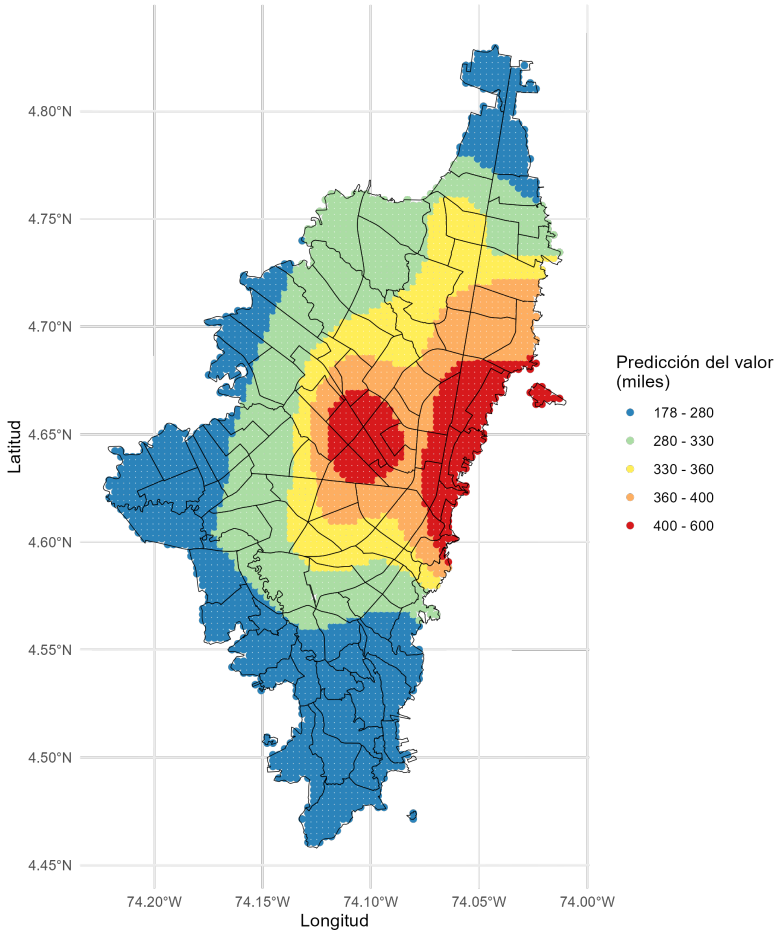

Figura 8: Efecto marginal de las coordenadas sobre la predicción del valor de los apartamentos - Modelo GAM. Fuente: Elaboración propia.

En la Figura 8 se presenta el efecto marginal de la ubicación geográfica sobre la estimación del avalúo de los apartamentos, cuando las coordenadas se incorporan al modelo *GAM* mediante un término suavizado bidimensional s(COORDENADA_X, COORDENADA_Y), capaz de capturar variaciones espaciales no lineales. En esta representación, las demás variables del modelo se mantienen constantes. Se observa que la zona centro-nororiental registra las estimaciones más elevadas, con valores superiores a 400 y hasta 600 millones de pesos. Valores intermedios, entre 360 y 400 millones, se proyectan hacia el nororiente y centro-oriente. A partir de este eje, las predicciones disminuyen gradualmente: primero a un rango de 330 a 360 millones en una franja circundante, luego a 280 a 330 millones en un cinturón más periférico, y finalmente a 178 a 280 millones en las zonas sur y suroccidental. Este patrón revela un comportamiento espacial continuo, lo cual sugiere una influencia clara de la localización geográfica en la determinación del avalúo. Esta conclusión se refuerza al contrastar los resultados con la distribución geográfica de las ofertas inmobiliarias mostrada en la Figura 3.

Modelo Random Forest

Mediante la estructura del modelo *Random Forest*, se busca identificar la combinación óptima de hiperparámetros que minimice el error de predicción. Para ello, se implementa una búsqueda en rejilla (*grid search*) sobre dos hiperparámetros principales: **mtry**, que representa el número de variables seleccionadas aleatoriamente en cada división del árbol, y **min_n**, que indica el número mínimo de observaciones requerido en un nodo para que se realice una nueva partición. Durante el proceso de ajuste, se fijaron algunos parámetros del modelo: el número total de árboles se estableció en 500, y el tamaño de la muestra utilizada en cada árbol se mantuvo constante. Como resultado de la búsqueda, y utilizando una semilla de aleatorización de 54321, se determinó que la combinación de hiperparámetros que ofreció el mejor desempeño fue $mtry = 4$ y $min\_n = 5$.

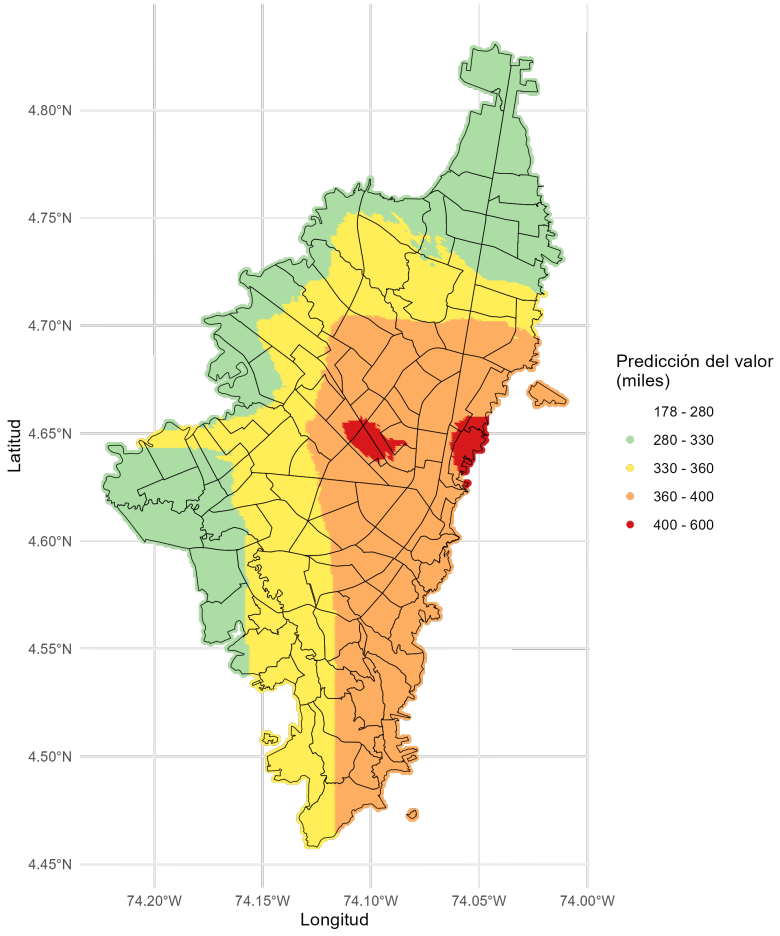

Figura 9: Efecto marginal de las coordenadas sobre la predicción del valor de los apartamentos - Modelo Random Forest. Fuente: Elaboración propia.

La Figura 9 presenta el efecto marginal de la ubicación geográfica sobre el valor predicho de los apartamentos, cuando las coordenadas de latitud y longitud se incorpo-

ran al modelo *Random Forest* tanto como predictores independientes como mediante su interacción. Dado que cada árbol en el bosque aleatorio particiona el espacio de predictores mediante divisiones binarias en los nodos, la agregación de múltiples árboles genera una segmentación espacial compuesta por regiones rectangulares con valores promedio diferenciados, lo que otorga al mapa una apariencia de bloques adyacentes.

El modelo predice los valores más elevados, superiores a los 400 millones de pesos, en las zonas central y oriental de la ciudad, rodeadas por una banda continua con valores entre 360 y 400 millones que se extiende de norte a sur. En el noroccidente y parte del centro-occidente, las predicciones descienden al intervalo de 330 a 360 millones, abarcando también el sur urbano. Finalmente, el extremo norte y varios sectores del occidente presentan valores entre 280 y 330 millones. Cabe destacar que el rango inferior, de 178 a 280 millones, no se evidencia en ninguna de las regiones definidas por el modelo, lo que sugiere que, si bien las coordenadas geográficas permiten capturar adecuadamente zonas de precios medios y altos, el modelo presenta limitaciones para identificar áreas con valores bajos, posiblemente debido a una menor representación de estas en el conjunto de entrenamiento o a una mayor heterogeneidad no explicada por las variables disponibles.

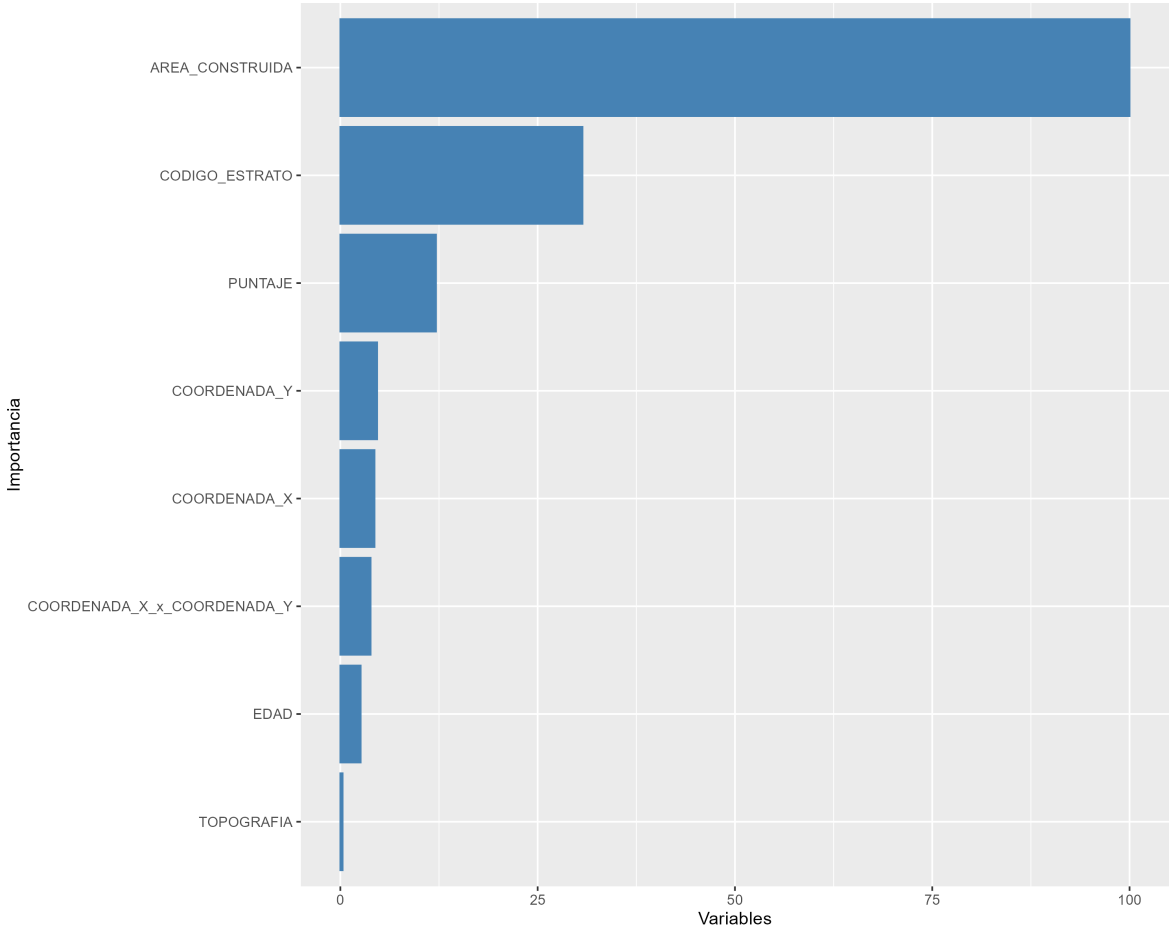

Figura 10: Variables importantes en el modelo RF de apartamentos
Fuente: Elaboración propia.

La Figura 10 muestra que las coordenadas geográficas tienen un rol secundario en la reducción del error de predicción en el modelo *Random Forest*: cada una de ellas (latitud, longitud e interacción) contribuye con menos del 5 % a la ganancia total en precisión, en contraste con variables estructurales como el *Área construida* o el *Estrato*, que presentan una importancia sustancialmente mayor. Este patrón concuerda con lo observado en el mapa espacial generado por el modelo, en el que se identifican con claridad las zonas de mayor valor, pero se observa una menor capacidad para capturar adecuadamente las áreas de menor precio.

Estos resultados sugieren que, si bien el componente espacial aporta información útil, su influencia en el modelo es limitada en comparación con los factores estructurales. En consecuencia, la estimación del valor comercial de los inmuebles está predominantemente determinada por características físicas y socioeconómicas, mientras que la localización geográfica actúa como un modificador de segundo orden dentro del proceso predictivo.

Modelo LightGBM

En el caso del modelo *LightGBM*, se llevó a cabo un ajuste de hiperparámetros mediante búsqueda en malla (*grid search*), considerando los parámetros **mtry**, **min_n** y **tree_depth**. El número total de árboles se fijó en 1.000, y se estableció una tasa de aprendizaje (*learn_rate*) de 0.0038. Para evitar sobreajuste y acelerar el proceso de entrenamiento, se implementó un criterio de parada anticipada (*early stopping*) que detiene el proceso si no se observa mejora en el rendimiento durante 20 iteraciones consecutivas. La combinación de hiperparámetros que mostró el mejor desempeño fue: **mtry** = 6, **min_n** = 4 y **tree_depth** = 8.

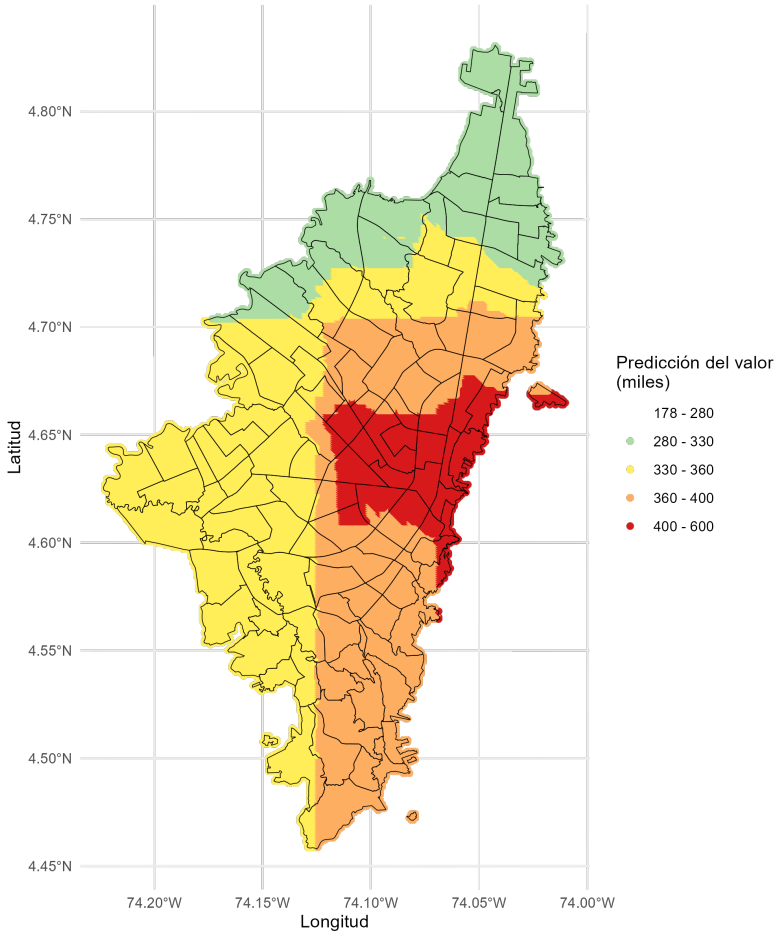

Figura 11: Efecto marginal de las coordenadas sobre la predicción del valor de los apartamentos - Modelo LGBM. Fuente: Elaboración propia.

La Figura 11 presenta el efecto marginal de las coordenadas geográficas sobre el valor predicho de los apartamentos cuando las demás covariables del modelo *LightGBM* se mantienen constantes. Al igual que en el modelo *Random Forest*, las coordenadas geográficas se incorporan tanto de forma independiente como a través de su interacción. El mapa revela una zona bien delimitada de altos avalúos en el centro-oriente de la

ciudad, con estimaciones superiores a los 400 millones de pesos. Esta área se encuentra rodeada por una franja intermedia, entre 360 y 400 millones, que recorre el eje central de la ciudad hacia el sur. Hacia el nororiente y en el extremo norte, las predicciones disminuyen a un rango de 330 a 360 millones, mientras que en la porción septentrional más alejada las estimaciones se sitúan entre 280 y 330 millones. Al igual que en el modelo *Random Forest*, el intervalo inferior de 178 a 280 millones no se manifiesta en ninguna zona del mapa, lo que sugiere que el modelo no logra identificar adecuadamente las áreas con avalúos más bajos. Dado que *LightGBM* se basa en árboles de decisión, la representación espacial adopta una estructura escalonada compuesta por pequeños rectángulos, reflejo de las divisiones jerárquicas en el espacio de predictores.

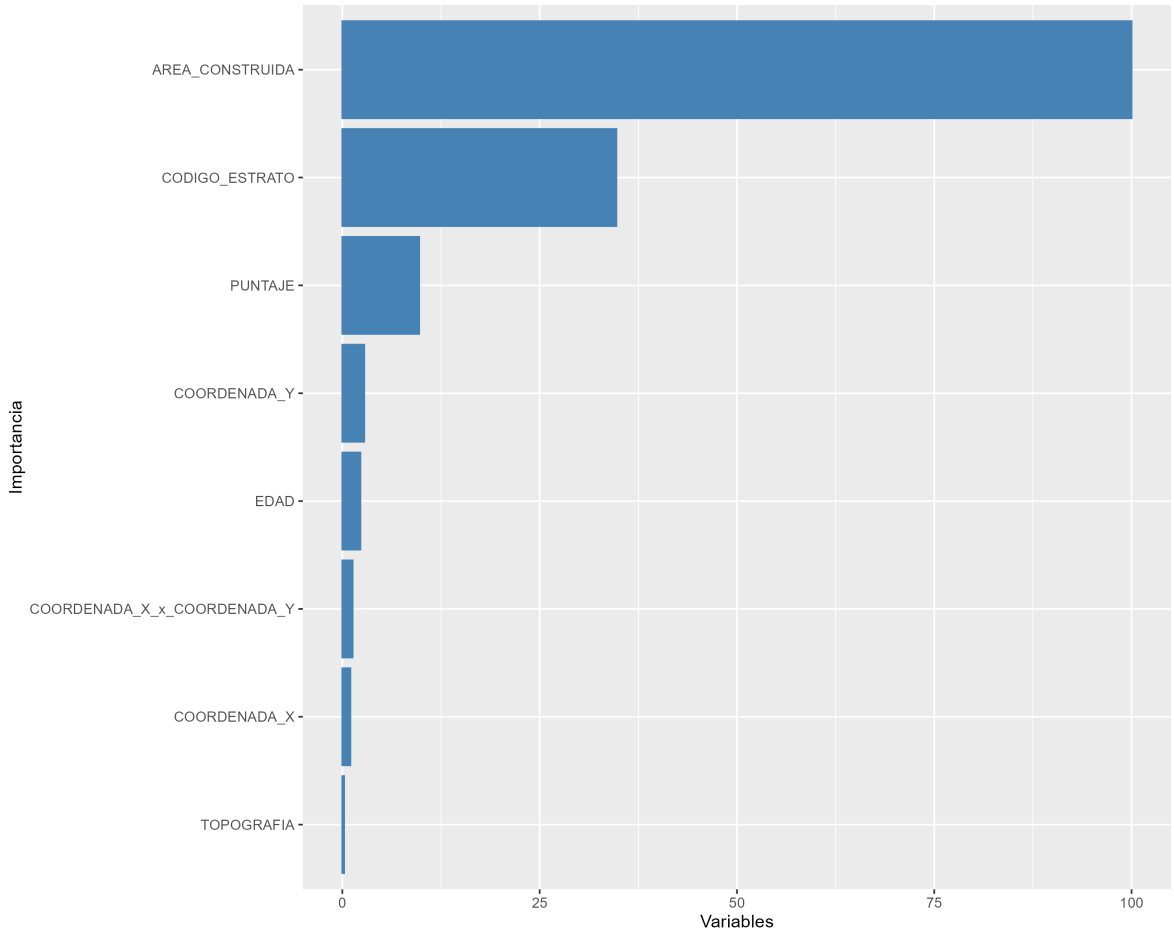

Figura 12: Variables importantes en el modelo LGBM de apartamentos.
Fuente: Elaboración propia.

La Figura 12 evidencia que las variables *Área construida* y *Estrato* concentran la mayor parte de la reducción del error en el modelo *LightGBM*, posicionándose como los factores estructurales más relevantes para la predicción del valor de los apartamentos. En contraste, las coordenadas geográficas Latitud ($COORDENADA\_Y$), Longitud ($COORDENADA\_X$) y su interacción aportan individualmente menos del cinco por

ciento a la ganancia total en precisión. Si bien su contribución es significativa desde el punto de vista espacial, permanece subordinada al efecto de las variables físicas del inmueble, que resultan determinantes en el rendimiento del modelo.

6.3.2. Modelo Casas

Modelo GAM

Siguiendo una estructura análoga a la utilizada para los apartamentos, el modelo *GAM* ajustado para las casas se conFigura de la siguiente manera:

- **Variable dependiente:** *VALOR*, que representa el valor comercial del inmueble.
- **Términos no paramétricos:**
  - `s(COORDENADA_X, COORDENADA_Y, bs = "gp", k = 50)`: Captura las variaciones espaciales mediante un suavizamiento bidimensional basado en un proceso gaussiano (`gp`) con base de dimensión 50.
  - `s(AREA_CONSTRUIDA)`: Modela la relación no lineal entre el área construida y el valor comercial del inmueble, utilizando una base de dimensión cinco (por defecto).
  - `s(AREA_TERRENO)`: Captura la relación no lineal entre el área del terreno y el valor del inmueble, también con una base de dimensión cinco.
- **Términos paramétricos:**
  - `PUNTAJE, EDAD, TOPOGRAFIA, CODIGO_ESTRATO`. Estas variables se incluyen como efectos lineales, asumiendo una relación aproximadamente lineal con la variable dependiente.
- **Familia y enlace:**
  - Se adopta una distribución Gamma, apropiada para variables continuas positivas, junto con un enlace logarítmico (`log`) para garantizar predicciones positivas.
- **Método de ajuste:**
  - Se utiliza el método de máxima verosimilitud restringida REML para ajustar el modelo, lo que contribuye a controlar el sobreajuste en los términos suavizados.

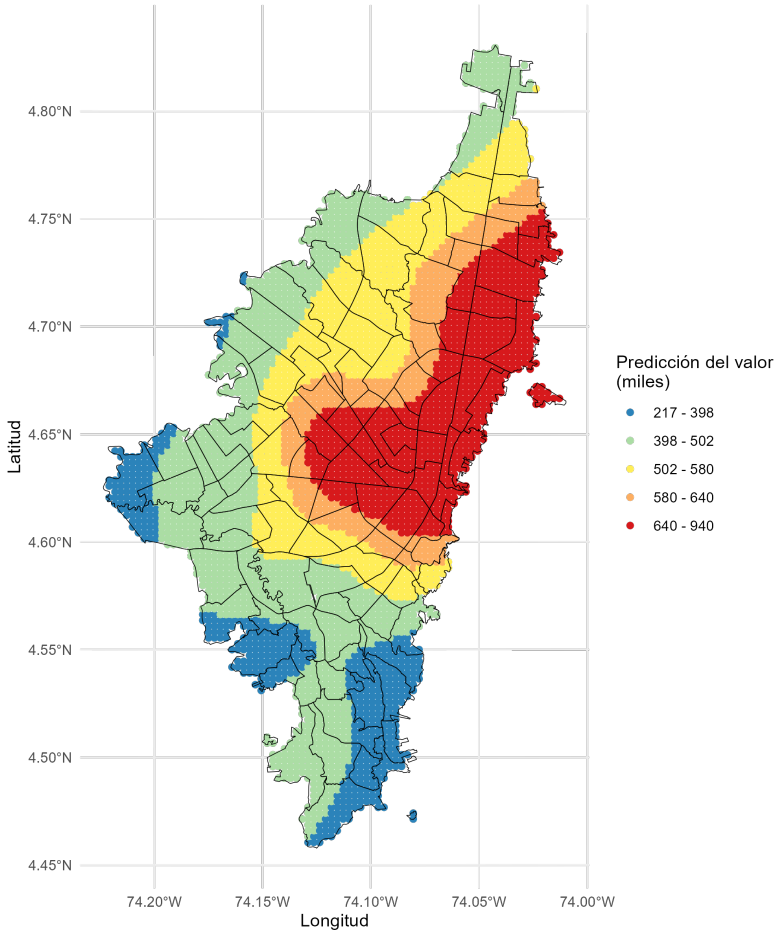

Figura 13: Efecto marginal de las coordenadas sobre la predicción del valor de los casas - Modelo GAM. Fuente: Elaboración propia.

De manera análoga a lo observado en el caso de los apartamentos, la Figura 13 presenta el efecto marginal de las coordenadas geográficas sobre la estimación del valor de las casas, manteniendo constantes las demás covariables del modelo *GAM*. Se observa una franja centro-oriental con los avalúos más elevados, donde se concentra una extensa zona con valores superiores a 640 millones de pesos. Esta área se conecta hacia el nororiente a través de una franja de valores intermedios-altos, 580 a 640 millones, seguida de un cinturón más amplio con predicciones entre 502 y 580 millones, que marca la transición hacia los sectores periféricos. En estas zonas externas, los valores estimados descienden gradualmente, primero al rango de 398 a 502 millones y, finalmente, al intervalo más bajo de 217 a 398 millones en la periferia sur y suroccidental de la ciudad. Al comparar este patrón con la distribución espacial de las ofertas, según la Figura 2, se constata que el suavizado aplicado sobre las coordenadas logra capturar adecuadamente las variaciones espaciales del mercado.

Modelo Random Forest

Siguiendo una configuración análoga a la empleada para los apartamentos, en el caso de las casas se ajustó un modelo *Random Forest* buscando la combinación óptima de hiperparámetros que minimizara el error de predicción. Para ello, se implementó una búsqueda en grilla sobre los parámetros **mtry** (número de variables consideradas en cada partición) y **min_n** (número mínimo de observaciones requeridas en un nodo para efectuar una división), manteniendo fijo el número de árboles en 500. La mejor combinación se obtuvo utilizando una semilla de replicación igual a 54321, y correspondió a $mtry = 5$ y $min\_n = 5$.

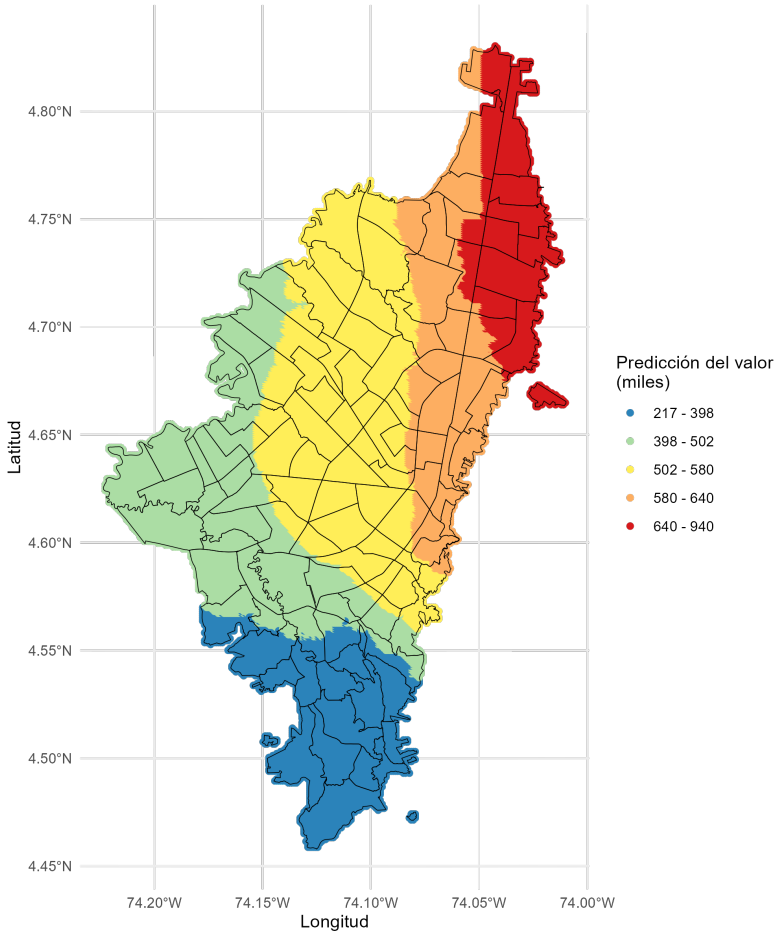

Figura 14: Efecto marginal de las coordenadas sobre la predicción del valor de las casas - Modelo Random Forest. Fuente: Elaboración propia.

La Figura 14 muestra el efecto marginal de la ubicación geográfica sobre la estimación del avalúo para las casas. El rango más bajo, entre 217 y 398 millones, predomina en las zonas sur y suroccidental, mientras que una franja de 398 a 502 millones cubre el noroccidente y parte del norte. En la zona céntrica de la ciudad, los valores se sitúan entre 502 y 580 millones, extendiéndose hacia el occidente, seguida por una banda que

se extiende de norte a sur a lo largo del eje oriente-centro. Finalmente, las predicciones más elevadas, entre 640 y 940 millones, se concentran en la parte noro-oriental. De manera análoga a lo observado en los apartamentos, el modelo *Random Forest* genera una superficie escalonada, producto del promedio de valores dentro de regiones rectangulares, característica propia de los modelos basados en árboles de decisión.

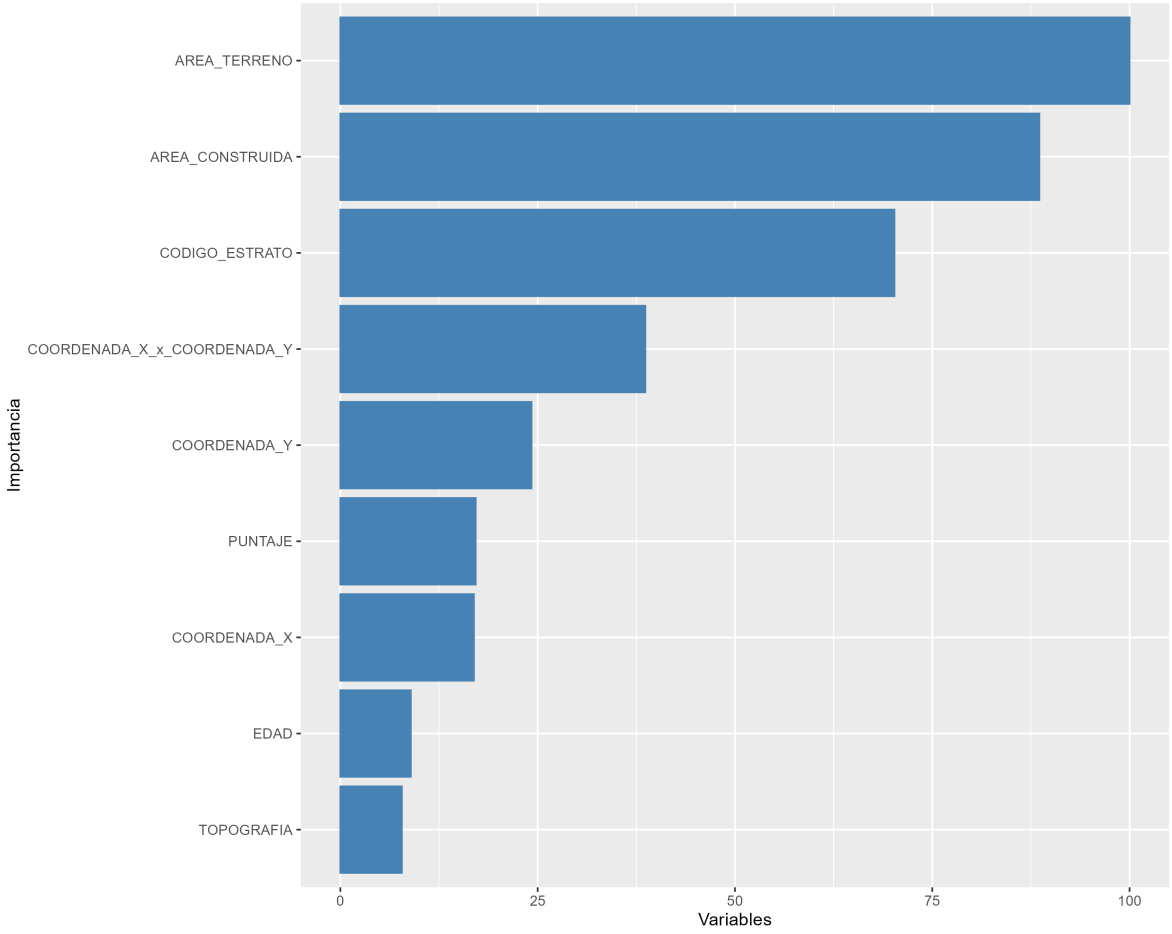

Figura 15: Variables importantes en el modelo RF de casas. Fuente: Elaboración propia.

La Figura 15 muestra que las variables *Área terreno*, *Área construida* y *Estrato* concentran la mayor parte de la reducción del error en el modelo. Por su parte, la latitud ($COORDENADA\_Y$), longitud ($COORDENADA\_X$) y la interacción latitud × longitud contribuyen en menor medida, siendo la interacción la que aporta la mayor reducción de error entre estos predictores espaciales. De manera análoga a lo observado en los apartamentos, la contribución de estas variables espaciales es significativa pero secundaria, destacando finalmente las variables físicas como las más relevantes en el modelo *Random Forest*.

Modelo LightGBM

Al igual que en el modelo para apartamentos, en el modelo *LightGBM* para casas se ajustaron los hiperparámetros mediante una búsqueda en grilla, incluyendo **mtry**, **min_n** y **tree_depth**. El número de árboles se fijó en 700 y se estableció una tasa de aprendizaje (*learn_rate*) de 0.0039. Además, se aplicó un criterio de parada anticipada (*early stopping*) tras 20 iteraciones consecutivas sin mejora en el rendimiento. La combinación óptima de hiperparámetros resultó ser: $mtry = 6$, $min\_n = 12$ y $tree\_depth = 7$.

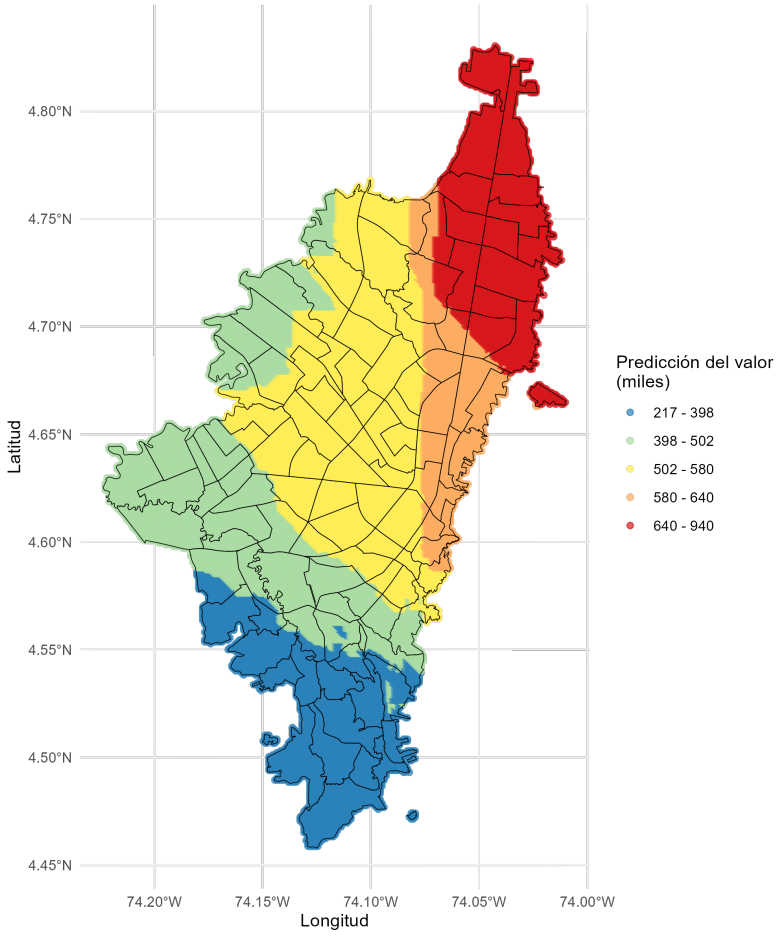

Figura 16: Efecto marginal de las coordenadas sobre la predicción del valor de las casas - Modelo LGBM. Fuente: Elaboración propia.

La Figura 16 presenta el efecto marginal de las coordenadas geográficas sobre el valor predicho de las casas. De manera similar a lo observado en los modelos de apartamentos con *RF* y *LightGBM*, se incorporan tanto las coordenadas independientes como su interacción. La zona sur y suroriente registra valores en el rango de 217 a 398 millones. A continuación, una franja que se extiende desde el suroriente hasta el occidente presenta valores entre 398 y 502 millones. En la zona central se observan valores entre 502 y 580 millones, extendiéndose hacia el noroccidente. Más al norte y partiendo desde

la zona centro-septentrional, predominan valores entre 580 y 640 millones. Finalmente, en la zona norte se concentran valores superiores a 640 millones. Visualmente, la representación de las coordenadas no muestra un comportamiento estrictamente escalonado, aunque la delimitación entre los rangos tiende a formar regiones de forma rectangular, característica típica de los modelos basados en árboles de decisión.

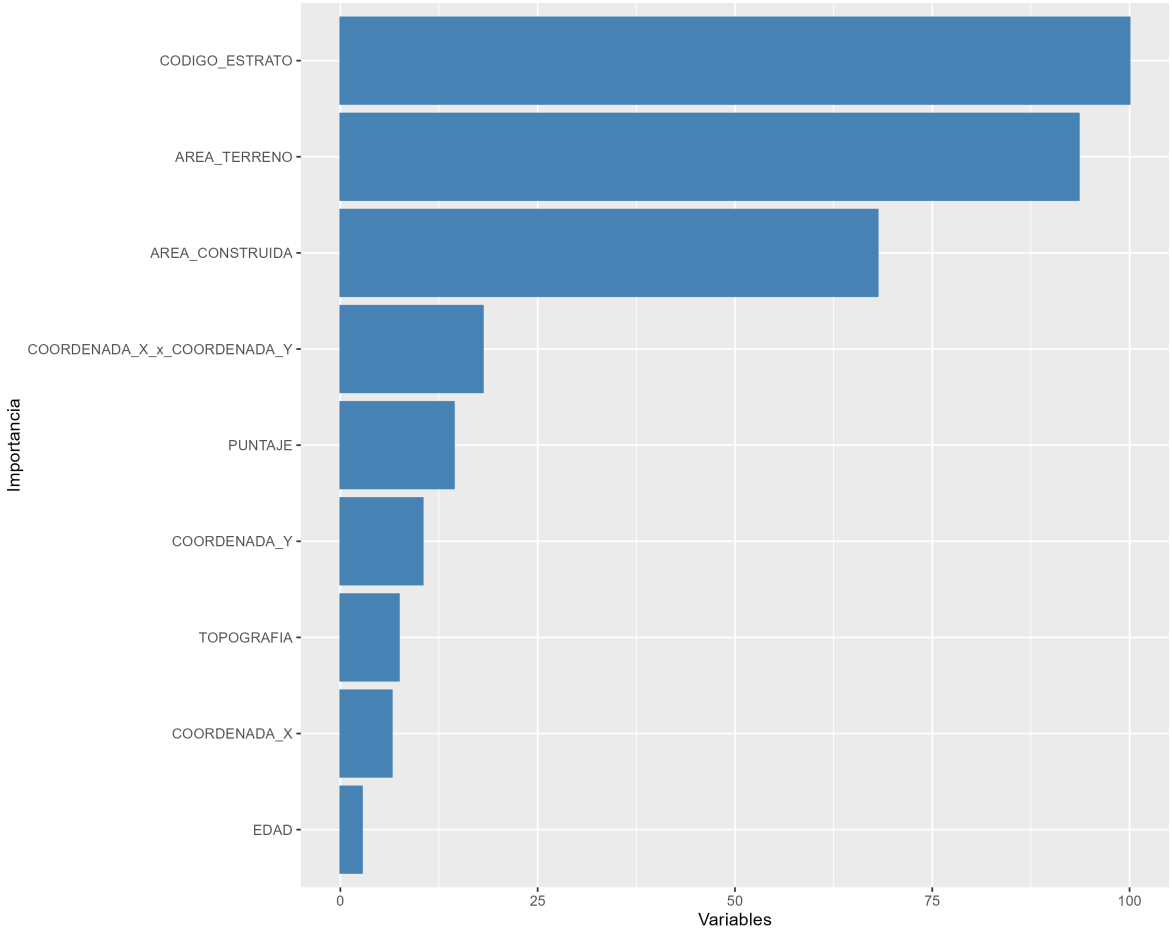

Figura 17: Variables importantes en el modelo LGBM de casas. Fuente: Elaboración propia.

La Figura 17 muestra, al igual que en el modelo RF, que el *Área terreno*, *Área construida* y el *Estrato* son las variables que concentran la mayor parte de la reducción del error. La Latitud ($COORDENADA\_Y$), la Longitud ($COORDENADA\_X$) y la interacción Latitud × Longitud aportan cada una una proporción menor, siendo la interacción la que más reduce el error entre estas. Así, aunque las coordenadas contribuyen a la reducción del error, no constituyen el efecto principal en el modelo LGBM.

6.4. Métricas del Modelo

En esta sección se presentan las métricas de ajuste y rendimiento obtenidas para los tres algoritmos evaluados (*GAM*, *LightGBM* y *Random Forest*) sobre los predios de apartamentos y casas. Tras particionar los datos en una muestra de entrenamiento (90 %) y otra de prueba (10 %), se calcularon el error porcentual absoluto medio (MAPE) y el coeficiente de determinación R^2 para cada modelo, con el fin de comparar su precisión y capacidad de generalización.

Modelo	**MAPE (Train)**	**MAPE (Test)**	R^2
GAM	9.60	9.61	95 %
LIGHTGBM	8.57	9.18	97 %
RF	9.14	9.81	96 %

Tabla 3: Resultados del modelo para apartamentos. Fuente: Elaboración propia.

La Tabla 3 muestra, para los apartamentos, las métricas de rendimiento de los tres modelos. Se destaca que, como característica principal, los tres algoritmos generalizan de forma adecuada sobre nuevas observaciones: Las variaciones de MAPE son inferiores a un punto porcentual, lo que indica ausencia de sobreajuste.

Entre ellos, el modelo *LightGBM* sobresale por su desempeño: alcanza un MAPE de 8.6 % y 9.2 %, respectivamente, y explica el 97 % de la varianza.

Modelo	**MAPE (Train)**	**MAPE (Test)**	R^2
GAM	19.11	19.62	79 %
LIGHTGBM	16.91	18.90	84 %
RF	16.64	17.50	84 %

Tabla 4: Resultados del modelo para casas. Fuente: Elaboración propia.

Por otro lado, la Tabla 4 muestra, para las casas, las métricas de rendimiento de los tres modelos. Si bien los algoritmos generalizan de manera aceptable y las variaciones de MAPE no superan los 2.5 puntos porcentuales.

En este escenario, el modelo *Random Forest* sobresale por su equilibrio óptimo: alcanza un MAPE de 16.6 % y 18.9 %, respectivamente, y explica el 84 % de la varianza.

El mejor desempeño de *LightGBM* en apartamentos se atribuye a su capacidad de manejar grandes volúmenes de datos y capturar interacciones complejas entre variables, lo cual es posible debido a la mayor disponibilidad de registros para este tipo de inmueble. Por su parte, *Random Forest* mostró mejores resultados en casas probablemente porque su enfoque de agregación de múltiples árboles individuales proporciona mayor estabilidad y reduce el riesgo de sobreajuste, lo que es útil cuando se dispone de un número menor de observaciones, como ocurre con las casas.

6.5. Modelación final

Para la estimación del valor comercial de los predios de Bogotá, se aplicaron los modelos con mejor desempeño a todo el universo de apartamentos y casas, con el propósito de evaluar la precisión de las estimaciones respecto a las predicciones obtenidas por la Unidad Administrativa Especial de Catastro Distrital. Así, para el universo de apartamentos se aplicó el modelo *LightGBM* , y para las casas se utilizó el modelo *Random Forest*. Se empleó la variación porcentual como métrica de comparación, definida por la siguiente expresión:

$$\Delta_i = \frac{\widehat{V}_i - V_{i,\text{Catastro}}}{V_{i,\text{Catastro}}} \times 100\,\%,$$

Los resultados serán examinados de acuerdo con la distribución de la variación.

Percentil	Δ_i **Casas**	Δ_i **Apartamentos**
0	-76.88	-70.13
10	-7.22	-9.17
20	1.46	-5.41
30	7.97	-3.01
40	13.70	-1.01
50	19.58	0.87
60	26.21	2.89
70	34.51	5.27
80	47.21	8.34
90	75.79	13.31
100	15282.20	171.19

Tabla 5: Distribución de la variación porcentual por percentil. Fuente: Elaboración propia.

En la Tabla 5 se muestra la comparación de la variación por percentiles del precio observado. Para el modelo estimado de apartamentos *LightGBM* , entre los percentiles 10 y 90 se concentran variaciones entre -9.2 % y 13.3 %, lo que indica una baja dispersión respecto al valor reportado por la entidad. La mediana de las variaciones es inferior al 1 %, lo que representa un buen resultado considerando la heterogeneidad del mercado inmobiliario. En cuanto al modelo obtenido para casas mediante la estructura de *Random Forest*, se observa una mayor dispersión entre los percentiles 10 y 90, con variaciones que oscilan entre -7.2 % y 75.7 %. La mediana de las variaciones se sitúa alrededor del 19.6 %, reflejando una sobrestimación respecto al valor reportado por la entidad. Esto corrobora lo evidenciado en las métricas de rendimiento, por lo que las estimaciones para el universo de casas pueden presentar mayor imprecisión.

7. CONCLUSIONES

El presente estudio muestra que los algoritmos de aprendizaje automático, en particular *LightGBM* y *Random Forest*, demostraron ser herramientas fiables para estimar el valor comercial de predios residenciales en Bogotá. Estos modelos fueron construidos sobre una base de datos que fue unificada a lo largo del estudio usando fuentes de catastro y de datos abiertos sobre casas y apartamentos. En el caso de los apartamentos, *LightGBM* logró un MAPE de 8.6 % en entrenamiento y 9.2 % en prueba, explicando el 97 % de la variabilidad observada. Por su parte, *Random Forest* mostró mayor estabilidad y precisión al estimar el valor de las casas, obteniendo un MAPE de 16.6 % en entrenamiento y 17.5 % en prueba, así como un coeficiente de determinación del 84 %. Cabe destacar que las métricas de rendimiento para las casas no son las mejores, dado que la menor cantidad de ofertas disponibles para este tipo de predio limita el entrenamiento, en contraste con la mayor disponibilidad de información para apartamentos.

Asimismo, la comparación de los tres tipos de modelos GAM, *Random Forest* y *LightGBM*, permitió identificar fortalezas y limitaciones particulares en cada caso. Los modelos GAM destacan por su interpretabilidad y su capacidad para representar de forma suave las tendencias espaciales, aunque presentan menor poder predictivo en comparación con los enfoques basados en árboles. *Random Forest* ofrece predicciones robustas y estables, pero su naturaleza limita la captura fina del componente espacial. Por su parte, *LightGBM* combina eficiencia computacional y alta precisión en la predicción, aunque, al igual que *Random Forest*, no modela de manera óptima los efectos espaciales.

De este modo, y en función del contexto de uso y los objetivos profesionales, las tres técnicas pueden brindar estimaciones confiables del avalúo comercial y una representación sólida del mercado inmobiliario, incluso al extrapolarse a predios con otras características (por ejemplo, inmuebles comerciales o industriales). Para los inmuebles en propiedad no horizontal estos modelos brindan una alternativa para la valoración dado que disminuye visitar los predios y reducirá los costes de recurso humano en el levantamiento de información.

Por otro lado, en un contexto académico, este trabajo de grado ofrece un marco metodológico para comparar y evaluar técnicas de estimación de avalúos, resaltando la importancia de avanzar hacia metodologías más flexibles y adaptables a las particularidades de los datos geoespaciales. En este sentido, se invita a futuros investigadores a profundizar en enfoques híbridos que integren técnicas de *machine learning* con componentes espaciales, como los modelos de bosques aleatorios geográficos, interpolaciones espaciales basadas en aprendizaje automático y métodos que incorporan estructuras de

dependencia espacial.

En síntesis, se cumplió con los objetivos planteados: (i) se construyó y validó una base de datos confiable a partir de información catastral y de datos abiertos; (ii) se desarrollaron y ajustaron modelos GAM, *Random Forest* y *LightGBM* para la estimación del valor comercial; (iii) se evaluó su desempeño mediante MAPE y R^2, identificando a *LightGBM* como el de mayor precisión para apartamentos y a *Random Forest* para casas; y (iv) se generaron estimaciones del valor comercial de los predios residenciales en Bogotá usando los modelos con mejor desempeño.

8. REFERENCIAS

[1] Resolución 620 de 2008 - Instituto Geográfico Agustín Codazzi (IGAC), "Resolución 620 de 2008," Septiembre 2008, por la cual se establecen los procedimientos para los avalúos ordenados dentro del marco de la Ley 388 de 1997. Diario Oficial No. 47.124 del 26 de septiembre de 2008 - Consultado el 29 de Mayo de 2025. [Online]. Available: https://www.igac.gov.co/

[2] Resolución 1040 del 2023 - Instituto Geográfico Agustín Codazzi - IGAC, "Resolución 1040 del 2023," Agosto 2023, por medio de la cual se expide la Resolución Única de la Gestión Catastral Multipropósito EL DIRECTOR GENERAL DEL INSTITUTO GEOGRÁFICO AGUSTÍN CODAZZI (IGAC), en uso de sus facultades legales y reglamentarias, en especial, las otorgadas por los artículos 43 y 47 de la Ley 2294 de 2023; numerales 2, 3 y 20 del artículo 10 del Decreto número 846 de 2021 - Consultado el 29 de Mayo de 2025. [Online]. Available: https://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=149779

[3] Decreto 148 de 2020 - Presidencia de la República de Colombia, "Decreto 148 de 2020," Febrero 2020, por el cual se reglamentan parcialmente los artículos 79, 80, 81 y 82 de la Ley 1955 de 2019 y se modifica parcialmente el Título 2 de la Parte 2 del Libro 2 del Decreto 1170 de 2015. Diario Oficial No. 51.231 del 4 de febrero de 2020 - Consultado el 29 de Mayo de 2025. [Online]. Available: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=111230

[4] A. T. Medina-Giraldo, "Predicción de los precios de vivienda en la ciudad de medellín y el Área metropolitana," Master's thesis, Universidad Internacional de La Rioja, 2023. [Online]. Available: https://reunir.unir.net/handle/123456789/14630

[5] S. Barrios Caracas and K. S. Quinto Rodríguez, "Modelación del precio de la oferta de vivienda en venta de la ciudad de cali, considerando variables propias del activo y covariables de su entorno," 2021. [Online]. Available: https://bibliotecadigital.univalle.edu.co/server/api/core/bitstreams/4d33f08e-8958-44aa-a199-8aee5436750e/content

[6] J. Toloza-Delgado, O. Melo, and N. Cruz, "Joint spatial modeling of mean and non-homogeneous variance combining semiparametric sar and gamlss models for hedonic prices," *Spatial Statistics*, vol. 65, p. 100864, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2211675324000551

[7] N. Lozano-Gracia and L. Anselin, "Is the price right?: Assessing estimates of cadastral values for bogotá, colombia*," *Regional Science Policy and*

Practice, vol. 4, no. 4, pp. 495–509, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1757780223006066

[8] V. J. Téllez-Buitrago and D. N. Martínez-Sánchez, "Método automático para la predicción del avalúo comercial de un inmueble en la ciudad de bogotá," 2021. [Online]. Available: https://repository.ucatolica.edu.co/entities/publication/1c7a29cf-e871-413a-9317-d2835137c2b6

[9] T. Hastie and R. Tibshirani, "Generalized additive models," *Statistical science*, vol. 1, no. 3, pp. 297–310, 1986.

[10] S. Wood, *Generalized Additive Models: An Introduction with R*, 2nd ed. Chapman and Hall/CRC, 2017.

[11] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[12] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, 2nd ed., ser. Springer Texts in Statistics. New York, NY: Springer, 2021.

[13] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, 1st ed. Chapman and Hall/CRC, 1994.

[14] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001, 1999 Reitz Lecture.

[15] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," 2016. [Online]. Available: https://doi.org/10.1145/2939672.2939785

[16] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017, pp. 3149–3157. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[17] Microsoft and Contributors, "Lightgbm features," https://lightgbm.readthedocs.io/en/latest/Features.html, 2025, consultado el 29 de Mayo de 2025.

[18] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2025. [Online]. Available: https://www.R-project.org/

[19] S. Georganos, T. Grippa, A. N. Gadiaga, C. Linard, M. Lennert, S. Vanhuysse, N. Mboga, E. Wolff, and S. Kalogirou, "Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling," *Geocarto International*, vol. 36, no. 2, pp. 121–136, 2021.

[20] A. Sekulić, M. Kilibarda, G. B. M. Heuvelink, M. Nikolić, and B. Bajat, "Random forest spatial interpolation," *Remote Sensing*, vol. 12, no. 10, p. 1687, 2020.

[21] A. Saha, S. Basu, and A. Datta, "Random forests for spatially dependent data," *Journal of the American Statistical Association*, 2021.

[22] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International journal of forecasting*, vol. 22, no. 4, pp. 679–688, 2006.

[23] Unidad Administrativa Especial de Catastro Distrital, "Ordenamiento territorial - datos abiertos bogotá," https://datosabiertos.bogota.gov.co/group/ordenamiento-territorial?organization=uaecd&page=1, 2025, accedido: 21 de Marzo de 2024.

[24] A. Kowarik and M. Templ, "Imputation with the r package vim," *Journal of Statistical Software*, vol. 74, p. 1–16, 2016. [Online]. Available: https://www.jstatsoft.org/index.php/jss/article/view/v074i07