Problem set 3

Big Data y Machine Learning para economía aplicada

Lina María Alvarez Ardila, Sara Rocio Rojas Gomez

27/05/2025

Profesores: Ignacio Sarmiento Barbieri , Gustavo Adolfo Castillo Alvarez y Julián David Rojas Aguilar

Repositorio:
https://github.com/lialvareza1/Problem-Set-3---Lina-Alvarez-Ardila-y-Sara-Roc-o-Rojas.git

Introducción

Predecir el valor de una vivienda en una ciudad como Bogotá no es una tarea sencilla. En zonas como Chapinero, donde conviven distintos estratos, tipos de vivienda y niveles de acceso a servicios, el precio de una propiedad depende de muchos factores que no siempre son evidentes a simple vista. Este trabajo busca construir un modelo que permita predecir los precios de publicación de viviendas en Chapinero usando técnicas de machine learning, a partir de información disponible en anuncios, variables espaciales y otras características relevantes.

Contar con este tipo de modelos puede ser útil para muchas personas. Desde compradores e inversionistas que quieren tomar decisiones mejor informadas, hasta desarrolladores, entidades financieras y plataformas inmobiliarias interesadas en ofrecer herramientas de estimación automática de precios más precisas. Además, entender cómo se forman los precios en una localidad como Chapinero también puede ayudar a identificar desigualdades o tendencias de valorización en el territorio.

Aunque los modelos de precios no son nuevos, los enfoques hedónicos han sido ampliamente utilizados para descomponer el precio en función de atributos como área, número de habitaciones o ubicación, estos enfoques clásicos suelen tener limitaciones importantes. Varios estudios han demostrado que los precios de la vivienda no sólo dependen de variables internas, sino también del contexto espacial y temporal en el que están ubicadas. Por ejemplo, Chang y Wang (2025) proponen un modelo espacio-temporal que permite capturar la evolución de los precios en distintos distritos a lo largo del tiempo, mientras que Zali et al. (2025) desarrollan una red neuronal ponderada geográfica y temporalmente que incorpora tanto las características de la vivienda como su entorno y el momento de la transacción, mostrando mejoras sustanciales en la precisión. Otros trabajos destacan el valor de incorporar técnicas no paramétricas como los modelos de eficiencia metafrontier (Lozano et al., 2024), que permiten capturar diferencias de contexto entre zonas.

Un caso ampliamente conocido que ilustra los riesgos de usar modelos predictivos sin un buen control es el de la empresa Zillow en Estados Unidos. Su sistema de estimación automatizada de precios, conocido como Zestimate, logró una enorme visibilidad, pero también fue criticado por generar errores de predicción importantes, especialmente en contextos de alta variabilidad local. Esto llevó a que la empresa incurriera en pérdidas millonarias cuando intentó usar esos precios para comprar viviendas directamente del mercado (ver Chang & Wang, 2025; Han & Chun, 2025). Este ejemplo pone en evidencia que incluso con grandes volúmenes de datos y modelos complejos, es fundamental tener en cuenta las dinámicas espaciales y validar adecuadamente los resultados. Es así como el aprendizaje automático ha mostrado una metodología para el desarrollo de modelos más precisos, flexibles y capaces de integrar distintas fuentes de información, superando en muchos casos las limitaciones de los modelos hedónicos tradicionales.

Una línea de investigación reciente explora la combinación de técnicas de deep learning y algoritmos de ensamble para la predicción de precios inmobiliarios. Un ejemplo representativo de esta tendencia es el trabajo de BoostingCNNSOS (2025), que introduce un modelo híbrido que integra redes neuronales convolucionales (CNN) con métodos de boosting optimizados mediante metaheurísticas. Aplicado a más de 998.000 registros de transacciones inmobiliarias, este modelo superó en desempeño a métodos populares como XGBoost, LightGBM y LSTM, destacándose por su estabilidad y capacidad predictiva (Farahzadi et.al., 2025). Este tipo de metodologías muestran que la combinación inteligente de múltiples modelos puede capturar de forma más efectiva las complejas relaciones no lineales que caracterizan al mercado de vivienda, especialmente en zonas urbanas heterogéneas y puede ser útil para el desarrollo de los modelos a plantear en el presente texto.

Otro aporte importante proviene del uso de fuentes no convencionales de datos para enriquecer los modelos de predicción. Collante y Ríos (2023) desarrollaron un modelo de nowcasting del Índice de Precios de Vivienda Nueva (IPVN) en Colombia, utilizando información de Google Trends junto con variables macroeconómicas tradicionales como tasas de interés y desempleo. Los autores demostraron que la regresión Lasso logró mayor precisión que modelos clásicos como ARIMA. Este resultado es particularmente valioso para el caso que nos ocupa, pues en nuestro proyecto se contempla el uso de variables textuales extraídas de anuncios inmobiliarios, cuyo valor explicativo puede ser capturado con mayor eficacia por modelos penalizados o no paramétricos y los datos obtenidos de Google Trends pueden ser de gran ayuda.

Finalmente, estudios recientes han enfatizado la importancia de incorporar explícitamente las dimensiones espaciales y temporales en los modelos predictivos de los precios de las viviendas. Mattera y Franses (2025) proponen un modelo de factores latentes estructurado mediante clustering espaciotemporal, aplicado a series de precios de vivienda en Estados Unidos. A través de un procedimiento iterativo de estimación de factores globales y específicos por clúster, el modelo logra capturar patrones locales diferenciados que se pierden en enfoques agregados. Los resultados muestran que los factores específicos por zona geográfica mejoran notablemente la capacidad predictiva, especialmente cuando se trata de contextos urbanos con fuerte heterogeneidad interna. Esta evidencia resulta especialmente pertinente para un entorno como Chapinero, donde conviven distintos estratos, densidades y patrones de desarrollo urbano en áreas geográficamente cercanas.

Estos antecedentes respaldan la decisión metodológica de este trabajo: construir un modelo que integre variables estructurales, contextuales, textuales y espaciales mediante técnicas avanzadas de aprendizaje automático. Al evaluar distintas metodologías y atender explícitamente la dimensión espacial, se busca no solo mejorar la precisión predictiva, sino también aportar una mirada más rica sobre los determinantes territoriales de los precios de vivienda en Bogotá.
Con base en estas experiencias y aportes de la literatura, este proyecto propone construir un modelo de predicción de precios de vivienda que combine variables estructurales, espaciales y contextuales, evaluando distintas metodologías de machine learning y prestando especial atención a la validación geográfica de los resultados.

Descripción de los datos y adecuación al problema

La base de datos utilizada para este ejercicio corresponde a una muestra de inmuebles en Bogotá, dividida en dos subconjuntos: uno de entrenamiento y otro de prueba. Cada observación contiene información a nivel de propiedad, incluyendo variables numéricas y categóricas como precio, área total y construida, número de habitaciones y baños, tipo de propiedad, entre otras. Además, se cuenta con las coordenadas geográficas (latitud y longitud), lo cual permitió enriquecer la información con variables espaciales externas.

Para preparar la base de datos de cara al objetivo de predicción, se realizó un proceso exhaustivo de limpieza y validación. En primer lugar, se eliminaron registros con inconsistencias lógicas, como propiedades con más baños que habitaciones. Luego se filtraron valores extremos en variables como precio o superficie, y finalmente se imputaron datos faltantes utilizando la mediana agrupada por tipo de propiedad o número de habitaciones. Este enfoque permitió mantener la coherencia entre observaciones sin introducir sesgos evidentes.

[1]Además, se incorporaron nuevas variables que capturan aspectos del entorno geográfico inmediato de cada propiedad:

dist_colegio: a partir de datos de la Secretaría de Educación de Bogotá, se calculó la distancia geodésica entre cada inmueble y el colegio más cercano.

dist_parque : utilizando información oficial de parques y escenarios deportivos, se identificó el parque más próximo a cada propiedad, calculando distancias desde los centroides de los polígonos, previamente corregidos.

estrato_manzana: se cruzaron las coordenadas de los inmuebles con una capa de manzanas catastrales para extraer el estrato. En casos con valores faltantes, se imputó el valor modal de las manzanas cercanas dentro de un radio de 50 metros.

num_gastrobares_500m: con base en datos de la Infraestructura de Datos Espaciales de Bogotá (IDECA), se contó el número de establecimientos de gastronomía y bares en un radio de 500 metros alrededor de cada inmueble.

num_airbnb_1500m: Esta información se obtuvo de la página AIRDNA, en la que se lograron extraer datos espaciales de la ubicación de Airbnbs en Bogotá. Con las coordenadas exactas logramos unir las bases y conseguir el número de airbnbs en un perímetro de 1.5 km, esto se logró con el uso de paquetes como “jsonlite” y “sf”.

[2]También se aprovechó el contenido textual de los anuncios, específicamente la variable description, que incluye una descripción libre del inmueble. A partir de este campo se construyeron variables cuantitativas como:

---

[1] Todos los datos espaciales fueron transformados al sistema de referencia WGS84 (CRS 4326) para asegurar la homogeneidad en las mediciones.

[2] Para ello, se estandarizó el texto (conversión a minúsculas) y se aplicaron expresiones regulares para detectar la presencia de estas palabras.

longitud_texto: número total de caracteres.

num_palabras: número total de palabras.

Dummies por palabras clave, que capturan la presencia de términos como: "remodelado", "lujoso", "vista", y "amoblado".

Estas variables ayudan a capturar atributos cualitativos del inmueble que suelen estar asociados con la percepción de valor, pero que no siempre son visibles en las variables estructuradas tradicionales.

Estadística descriptiva

Justificación de las variables creadas a partir de la literatura

La inclusión de variables que capturan el entorno inmediato de cada propiedad se sustenta en abundante literatura que ha demostrado la relevancia del contexto urbano en la formación del precio de la vivienda. En particular, variables como la distancia al colegio más cercano y distancia a parques reflejan la accesibilidad a servicios educativos y recreativos, elementos valorados por las familias y asociados a calidad de vida urbana. Estudios como los de Zali et al. (2025) y Chang y Wang (2025) destacan que la cercanía a este tipo de infraestructuras genera una prima en el valor de los inmuebles, al representar atributos no incorporados en las características físicas de la propiedad pero sí en su valor percibido. En esta línea, la literatura sobre modelos hedónicos y modelos geográficamente ponderados (GWR, GTWR) también ha mostrado que la presencia de colegios, zonas verdes o centros de recreación, tiene efectos heterogéneos pero significativos en el precio del suelo y de la vivienda (Lozano et al., 2024).

Por otro lado, la variable de estrato socioeconómico, obtenida mediante la intersección espacial con manzanas catastrales, actúa como proxy del nivel de ingreso y las condiciones del entorno barrial, las cuales influyen tanto en la disposición a pagar como en la valorización futura. Esta aproximación ha sido recomendada en estudios de valuación que reconocen la heterogeneidad espacial del mercado inmobiliario, donde variables como el estrato permiten capturar diferencias estructurales no observables directamente en los atributos físicos del inmueble (Han & Chun, 2025).

Asimismo, la incorporación del número de gastrobares en un radio de 500 metros permite capturar dinámicas más recientes de transformación urbana y valorización por usos mixtos. Este tipo de establecimientos no sólo reflejan concentración de vida nocturna y comercio especializado, sino también procesos de gentrificación o modernización de ciertos sectores urbanos, como ha sido documentado en estudios de predicción inmobiliaria en contextos urbanos densos y dinámicos (Zali et al., 2025).

Finalmente, el aprovechamiento del campo de texto libre description responde a la creciente evidencia sobre el valor predictivo de la información no estructurada en modelos de precios. Variables como la longitud del texto o el número de palabras han sido utilizadas como indicadores de esfuerzo de marketing o detalle en la descripción, lo cual puede reflejar

propiedades con características más elaboradas o de mayor valor. Además, la detección de palabras clave como “remodelado”, “lujoso”, “vista” o “amoblado” permite identificar atributos cualitativos que, aunque no estén codificados en variables estructuradas, inciden directamente en la percepción del valor por parte de los potenciales compradores. En ese sentido, estas dummies actúan como indicadores del atractivo del inmueble y su potencial de valorización, lo que se alinea con hallazgos recientes sobre el uso de modelos text-as-data para análisis inmobiliarios (Han & Chun, 2025).

Justificación empírica

La cantidad de datos es suficiente para lograr caracterizar la población investigada y la variable de interés. Se puede observar en la Tabla #, como una primera mirada a los datos, un resumen de los valores promedio de variables clave en las bases de entrenamiento y prueba utilizadas para el modelado. Y se entiende que las propiedades en la base de entrenamiento tienen, en promedio, mayor superficie y número de habitantes que las de prueba. Esto sugiere diferencias estructurales entre ambos conjuntos que podrían afectar la generalización del modelo, por lo cual es fundamental una validación cuidadosa para evitar sesgos en las predicciones.

Table 1: Promedios de variables comparables en las bases de entrenamiento y prueba

[HTML]EFEFEF **Variable**	**Train**	**Test**
Precio (COP)	628149492.08	-
Superficie total (m^2)	128.79	115.36
Superficie cubierta (m^2)	123.57	106.70
Número de cuartos	3.28	2.39
Número de habitaciones	3.27	2.38
Número de baños	2.70	2.61

Ahora bien, en la caracterización de las variables de la base de entrenamiento se encuentra que: En los datos relacionados al precio, por ejemplo, se encuentra que la media general aproximada es de 628 millones y tiende a ser más alta entre más sube el estrato. Además, tal como se puede observar en la gráfica en el anexo 1, existe una cantidad alta de datos atípicos en la distribución de los precios en todos los estratos menos en el estrato 1, entre tanto, el estrato promedio por manzana es el 4.

De igual manera, en la gráfica 2 de los anexos, se presenta la distribución espacial del precio promedio por manzana en Bogotá, con un enfoque particular en la localidad de Chapinero, pues es nuestra localidad de interés. En el panel izquierdo, se observa una marcada heterogeneidad en los precios, con concentraciones más altas en el centro-norte y zonas nororientales de la ciudad, especialmente en la localidad de Suba y Usaquén

En contraste, las áreas del sur y occidente presentan precios promedio significativamente menores, lo que refleja patrones de segregación socioespacial en el precio de los inmuebles. El panel derecho, centrado en la localidad de Chapinero según su croquis oficial, evidencia una estructura interna similar: las manzanas del norte y nororiente de la localidad ,asociadas históricamente a barrios de estrato alto, exhiben mayores niveles de valorización, mientras

que las zonas limítrofes al sur presentan valores más bajos. Sin embargo, la cantidad de los datos presentados en Chapinero están limitados y se observa una heterogeneidad importante.

Y por último, es posible justificar la inclusión de las variables dummy en el modelo pues capturan atributos cualitativos que inciden en la percepción de valor de un inmueble. Como se observa en la gráfica 3 de los anexos, las propiedades que presentan estas características tienden a tener precios promedio más altos y una mayor dispersión hacia valores elevados o un cambio en el comportamiento del precio. Aunque las diferencias medianas no son extremas, sí existen patrones sistemáticos que pueden ayudar a fortalecer la predicción del modelo.

Resultados:

El modelo que presentó el mejor desempeño predictivo fue un Random Forest ajustado sobre la base de entrenamiento utilizando validación cruzada tradicional. Este modelo pertenece a la familia de métodos de ensamble, y se caracteriza por construir múltiples árboles de decisión a partir de diferentes subconjuntos de datos y predictores, promediando posteriormente sus resultados. La forma funcional estimada puede expresarse como:

$$\widehat{price}_i = f(x_i\,;\,\theta)$$

donde xi representa el vector de características observadas para la propiedad i, y θ denota el conjunto de parámetros internos optimizados del modelo, incluyendo el número de predictores seleccionados en cada división del árbol, el tamaño mínimo de los nodos terminales y la regla de partición utilizada.

Desde un enfoque teórico, Random Forest constituye una estrategia eficaz para predecir precios en mercados inmobiliarios dada su capacidad para capturar relaciones no lineales complejas y modelar interacciones entre variables sin requerir una especificación funcional previa. A diferencia de los modelos lineales tradicionales, que exigen supuestos fuertes sobre la forma de los datos, Random Forest es un método no paramétrico que aprende directamente de la estructura empírica del conjunto de datos (Breiman, 2001). Su robustez frente a valores atípicos, su tolerancia a la multicolinealidad y su buen desempeño en contextos con muchos predictores lo convierten en una opción adecuada cuando se trabaja con información heterogénea, como es común en bases de datos inmobiliarias (Kuhn & Johnson, 2013). Además, puede manejar sin dificultad tanto variables numéricas como categóricas, lo que permite aprovechar al máximo la diversidad informativa disponible.

En términos metodológicos, se evaluaron dos estrategias complementarias de validación cruzada. La primera consistió en una validación cruzada tradicional de cinco particiones aleatorias estratificadas, que permite estimar la capacidad predictiva del modelo sobre datos no observados sin considerar la dimensión geográfica. La segunda fue una validación cruzada espacial, implementada mediante la división del área geográfica en una grilla de 5x5 celdas definidas por las coordenadas de latitud y longitud. Cada celda se utilizó como conjunto de prueba en una iteración, lo que garantiza independencia espacial entre entrenamiento y

validación y permite detectar posibles sobreajustes por cercanía espacial (Roberts et al., 2017).

La selección de hiperparámetros se realizó mediante una búsqueda en malla sobre un conjunto de combinaciones de mtry, min.node.size y splitrule, utilizando el paquete ranger en R para una implementación computacionalmente eficiente del algoritmo. El modelo se entrenó para predecir el valor de la variable price, utilizando como predictores todas las variables disponibles, excluyendo identificadores y las coordenadas geográficas utilizadas en la partición espacial.

Los resultados muestran que el modelo entrenado bajo validación cruzada tradicional presenta un rendimiento claramente superior, con un RMSE de aproximadamente 165 millones de pesos colombianos y un coeficiente de determinación R^2de 0.6811. Esto sugiere que el modelo logra explicar cerca del 68% de la variación en los precios observados. En contraste, al aplicar validación cruzada espacial, el desempeño se reduce considerablemente: el RMSE asciende a 3.03 millones y el R^2 cae a 0.1669. Esta diferencia evidencia una importante autocorrelación espacial que no es completamente absorbida por las variables explicativas, y que tiende a inflar artificialmente la precisión del modelo cuando se evalúa sobre observaciones cercanas al conjunto de entrenamiento.

Pese a esta caída en desempeño bajo validación espacial, el modelo elegido para generar las predicciones finales fue el ajustado con validación cruzada tradicional, dado que proporciona un mejor ajuste global y una menor tasa de error promedio

Cuadro 2: Comparación de desempeño del modelo Random Forest

Métrica	Validación cruzada normal	Validación cruzada espacial
RMSE	165,051,250	3,030,914,07
R^2	0.6811	0.1669

Comparación entre el modelo final y otros modelos evaluados

Durante el proceso de modelado se estimaron múltiples especificaciones algorítmicas que incluían modelos lineales, árboles de decisión, ensambles, redes neuronales y meta-modelos combinados. Cada modelo fue evaluado bajo dos esquemas de validación: una validación cruzada aleatoria tradicional y una validación cruzada espacial estructurada en una grilla geográfica. El objetivo fue identificar no solo cuál era el modelo con menor error absoluto, sino también cuáles eran sus fortalezas relativas en términos de robustez predictiva y generalización espacial.

El modelo con mejor desempeño fue un Random Forest ajustado con validación cruzada tradicional, obteniendo un RMSE de 165,5 millones COP. Este resultado se explica por varias razones. En primer lugar, Random Forest es un algoritmo de ensamble basado en la construcción de múltiples árboles de decisión sobre subconjuntos aleatorios de datos y predictores. Esta técnica permite reducir el riesgo de sobreajuste, manejar eficazmente

interacciones no lineales y estabilizar la predicción frente a ruido en los datos (Breiman, 2001). Adicionalmente, este modelo tolera bien la inclusión de variables irrelevantes o colineales gracias a su proceso aleatorio de selección de predictores en cada split.

El modelo Stacking, que combinó XGBoost y Random Forest, alcanzó un RMSE de 180,5 millones, lo que lo posicionó como el segundo mejor. Su desempeño competitivo se explica por la diversidad algorítmica de sus componentes: mientras XGBoost optimiza márgenes de error de manera aditiva con un objetivo basado en gradientes (Chen & Guestrin, 2016), Random Forest introduce diversidad a través de la aleatorización estructural. Sin embargo, este stacking no logró superar al Random Forest por sí solo, posiblemente porque ambos modelos comparten mecanismos de ensamblaje similares y la combinación no aportó una ganancia neta adicional. Además, la capa meta-modelo utilizada para ensamblar sus predicciones pudo no haber sido lo suficientemente sofisticada como para capitalizar las diferencias entre ambos modelos.

El tercer modelo en desempeño fue el Gradient Boosting (XGBoost), con un RMSE de 186,1 millones. A pesar de su capacidad para generar predicciones altamente precisas, especialmente en contextos con variables no lineales y jerarquías complejas, XGBoost es particularmente sensible al ajuste de hiperparámetros. Si no se realiza una búsqueda exhaustiva (e.g., en profundidad del árbol, tasa de aprendizaje, número de iteraciones), su rendimiento puede decrecer notablemente. En este caso, aunque se utilizaron parámetros razonables, es posible que la combinación no haya sido óptima, lo que resultó en un desempeño ligeramente inferior al de Random Forest.

Por su parte, el Super Learner que combinó modelos logit y CART obtuvo un RMSE de 267,1 millones. Esta técnica, desarrollada por van der Laan, Polley y Hubbard (2007), busca encontrar una combinación convexa óptima de modelos base minimizando el error de predicción mediante validación cruzada. En efecto, logró mejorar sobre los modelos que lo componen por separado, lo cual valida su fundamento teórico. No obstante, su performance limitado frente a modelos de ensamble más complejos se debe a la baja capacidad predictiva de sus componentes individuales, que no lograron modelar adecuadamente las relaciones no lineales y espaciales entre los atributos de las propiedades.

Los modelos CART, regresión logística (Logit) y la red neuronal MLP presentaron desempeños significativamente más débiles, con RMSEs entre 289 y 307 millones. CART, aunque interpretable, es conocido por su inestabilidad: pequeñas variaciones en los datos pueden producir árboles completamente distintos. A pesar de su simplicidad y velocidad de entrenamiento, CART tiende a sobreajustar en ausencia de poda efectiva (Hastie, Tibshirani & Friedman, 2009). La regresión logística, por su parte, es lineal en los parámetros y no captura bien las interacciones ni las no linealidades implícitas en el precio de los inmuebles, especialmente cuando se incluyen variables derivadas de texto y geolocalización. Finalmente, la red neuronal utilizada (MLP con una capa oculta de cinco nodos) era estructuralmente limitada: el número de neuronas y capas no permitía modelar patrones complejos ni profundizar en la no linealidad de los datos, lo cual es coherente con su bajo desempeño.

Cabe señalar que bajo validación espacial, todos los modelos experimentaron un deterioro en su capacidad predictiva. Random Forest pasó de 165 a 303 millones de RMSE, mientras que

Logit alcanzó un valor de 330 millones. Esta diferencia ilustra la existencia de autocorrelación espacial en los datos: los precios de propiedades cercanas tienden a parecerse, y la validación aleatoria ignora esta dependencia espacial, inflando artificialmente las métricas de desempeño (Roberts et al., 2017). El único modelo cuya diferencia entre validaciones fue marginal fue la red neuronal, lo cual podría explicarse por su bajo poder predictivo en ambos escenarios más que por una mayor robustez espacial.

La tabla siguiente resume los resultados obtenidos en cada caso:

Modelo	RMSE Tradicional	RMSE Espacial
Random Forest	165536972	303069792
Stacking (XGBoost + Random Forest)	180543469	295511207
Gradient Boosting (XGBoost)	186142103	313173166
Super Learner (Logit + CART)	267095259	297389367
Árboles de Decisión (CART)	289216774	323763344
Red Neuronal (MLP)	289882887	287997461
Regresión Logística (Logit)	307124740	330866936

El mejor modelo fue aquel que logró equilibrar poder predictivo y robustez frente a la complejidad y ruido de los datos. Los resultados también evidencian que, más allá del algoritmo elegido, la correcta selección de variables, el ajuste de hiperparámetros y la validación geográficamente informada son aspectos clave para construir modelos confiables en contextos espaciales urbanos.

Conclusiones y Recomendaciones

Este estudio permitió explorar, comparar y evaluar el desempeño de diferentes modelos de aprendizaje automático para la predicción de precios inmobiliarios en el barrio Chapinero de Bogotá, empleando una base de datos estructurada con variables estructurales, geográficas y textuales. Los resultados obtenidos evidencian que los métodos de ensamble, particularmente el Random Forest, ofrecen ventajas significativas frente a enfoques más simples o lineales, tanto en términos de precisión como de capacidad para modelar relaciones complejas entre predictores.

El modelo de Random Forest, ajustado con validación cruzada tradicional, fue el que presentó el mejor desempeño predictivo, logrando explicar una proporción sustancial de la varianza en los precios observados. Este resultado se explica por su estructura interna, basada en la agregación de múltiples árboles de decisión construidos sobre muestras y subconjuntos aleatorios de predictores, lo que le otorga una notable capacidad para manejar la heterogeneidad de los datos, modelar interacciones no lineales y ser robusto frente a variables irrelevantes o colineales. A pesar de este buen desempeño, el análisis también evidenció que su capacidad de generalización disminuye bajo esquemas de validación espacial, lo que pone en evidencia la presencia de autocorrelación geográfica en los datos y el riesgo de sobreajuste cuando no se controla por esta dimensión.

Modelos de ensamblaje más sofisticados, como el stacking entre XGBoost y Random Forest, así como el Super Learner que combinó logit y CART, mostraron desempeños competitivos, aunque sin superar al Random Forest puro. Esto sugiere que, en contextos donde los modelos base ya son altamente performantes, las ganancias marginales de combinar múltiples predictores pueden ser limitadas si no se incorpora suficiente diversidad algorítmica o si el meta-modelo no es lo suficientemente potente. Por el contrario, modelos más simples, como la regresión logística, los árboles de decisión únicos y las redes neuronales con arquitecturas reducidas, se vieron claramente superados. En estos casos, las limitaciones estructurales del modelo, la falta de capacidad para capturar no linealidades y la ausencia de procesos de regularización o poda efectiva contribuyeron a un mayor error de predicción.

Un aspecto central que emerge del análisis es la importancia de utilizar esquemas de validación cruzada espacial cuando se trabaja con datos georreferenciados. La comparación entre validaciones tradicionales y espaciales evidenció que la mayoría de los modelos sobreestiman su desempeño si no se controla la dependencia espacial. Esta observación refuerza la necesidad de incorporar la dimensión geográfica no solo como variable explicativa, sino también como criterio metodológico al momento de evaluar la generalización de los modelos.

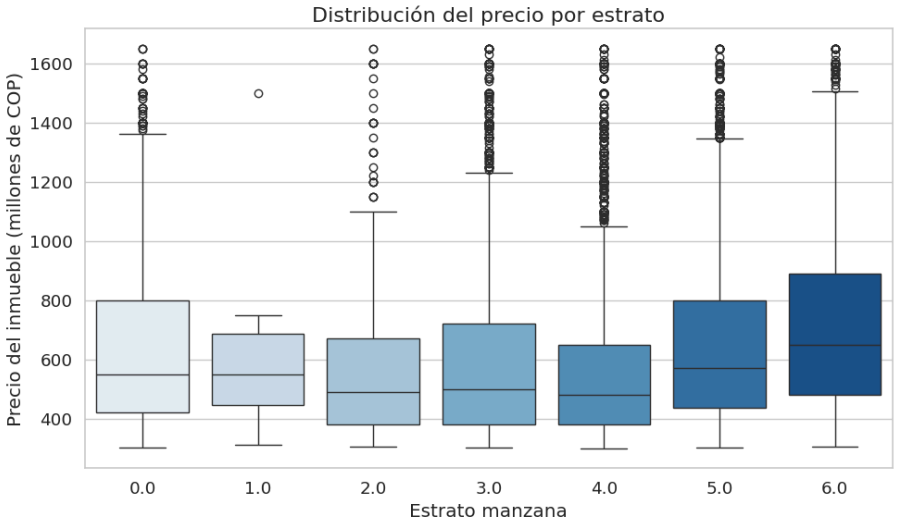
En términos metodológicos, se concluye que el éxito de un modelo no depende únicamente del algoritmo utilizado, sino de una combinación entre una adecuada selección de variables, una estrategia rigurosa de validación y un proceso de ajuste de hiperparámetros cuidadoso. Incluso modelos con alto potencial teórico, como las redes neuronales, pueden mostrar desempeños modestos si no se optimiza su estructura o si no se entrena con volúmenes de datos suficientes. Por esta razón, futuros desarrollos podrían considerar arquitecturas más profundas, el uso de técnicas de aprendizaje por transferencia o la incorporación de embeddings derivados de modelos de lenguaje natural para variables textuales.

Finalmente, los resultados obtenidos permiten extraer recomendaciones prácticas para contextos similares. En primer lugar, se recomienda priorizar modelos de ensamble como Random Forest o Boosting, que han demostrado ser eficaces y relativamente estables frente a problemas de sobreajuste. En segundo lugar, es imprescindible incluir validaciones espaciales al evaluar modelos con datos territoriales, ya que su omisión puede conducir a interpretaciones erróneas sobre la capacidad predictiva. En tercer lugar, se sugiere explorar de forma más sistemática combinaciones de modelos mediante ensamblajes o meta-modelos más complejos, en especial si se logra una mayor diversidad algorítmica y se cuenta con herramientas computacionales adecuadas para su implementación.

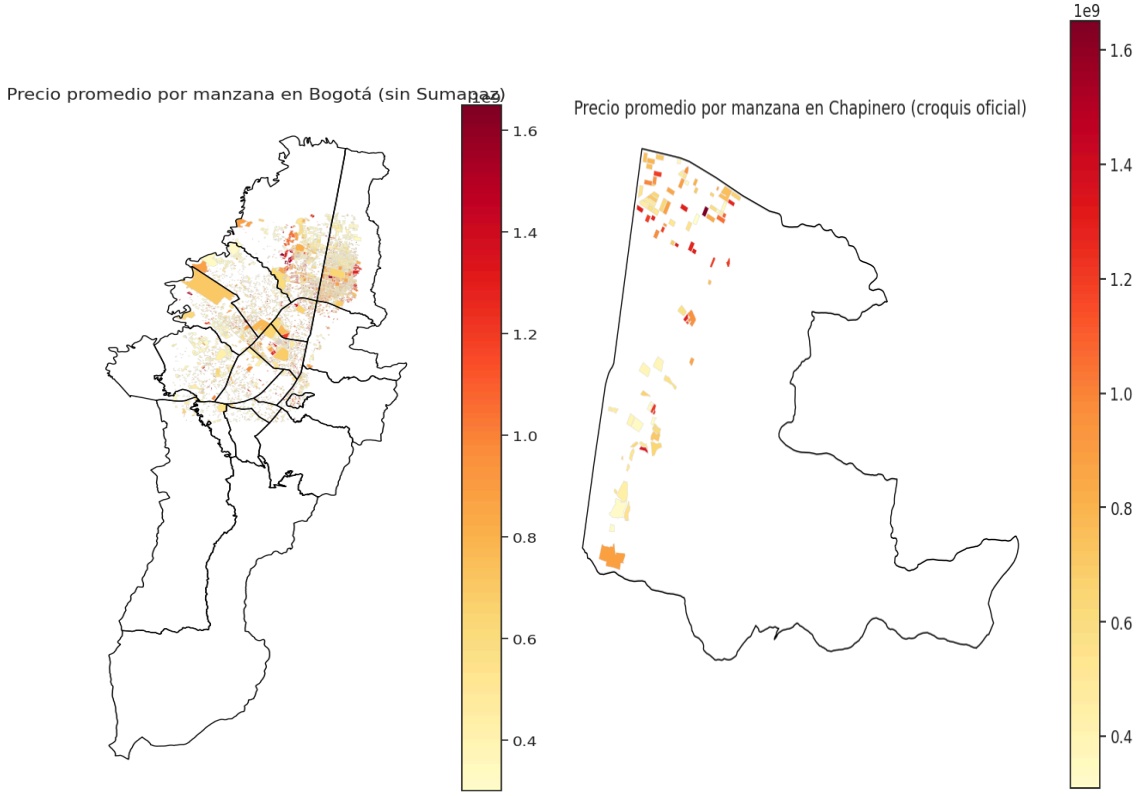
En conjunto, este ejercicio ilustra el valor de integrar herramientas modernas de aprendizaje automático con consideraciones teóricas del análisis espacial y de datos estructurados. El reto no consiste únicamente en alcanzar el mejor desempeño en una métrica determinada, sino en construir modelos confiables, interpretables y capaces de generalizar en escenarios reales de toma de decisiones, como es el caso del mercado inmobiliario urbano.

Anexos:

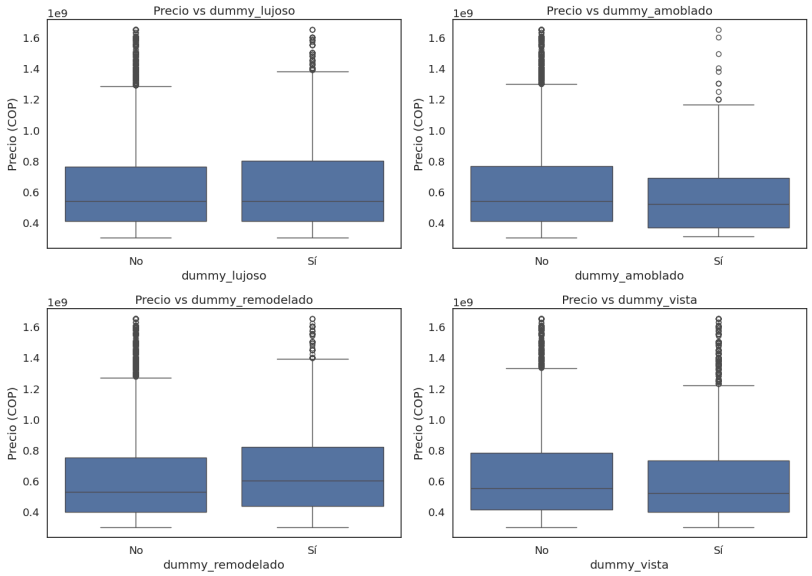
Gráfica 1 : distribución del precio por estrato



Gráfica 2: distribución espacial del precio promedio por manzana en Bogotá.



Gráfica 3: Boxplot de las variables dummies



Referencias

- Albouy, D., Christensen, P., & Sarmiento-Barbieri, I. (2020). The distributional effects of urban land use regulation. Journal of Urban Economics, 118, 103256.
- Bivand, R., Pebesma, E., & Gómez-Rubio, V. (2013). Applied Spatial Data Analysis with R (2nd ed.). Springer.
- Chang, Y.-M., & Wang, Y.-T. (2025). A spatial–temporal model of house prices in Northern Taiwan. Mathematics, 13(736), 1–24.
- Christensen, P., Sarmiento-Barbieri, I., & Timmins, C. (2022). Housing discrimination and the pollution exposure gap in the United States. Journal of the Association of Environmental and Resource Economists, 9(6), 1081–1117.
- Cliff, A. D., & Ord, J. K. (1973). Spatial Autocorrelation. Pion Limited.
- Han, E.-J., & Chun, S.-H. (2025). A long short-term memory model using kernel density estimation for forecasting apartment prices in Seoul City. Expert Systems with Applications, 283, 127748.
- Lozano, S., Gutiérrez, E., Klizentyte, K., & Susaeta, A. (2024). Efficient property value estimation for single‐family homes in central Florida. International Transactions in Operational Research, 32(5), 2952–2980.
- McMillen, D. P., Sarmiento-Barbieri, I., & Singh, R. (2019). Do light rail investments improve neighborhood accessibility? Evidence from the Charlotte Lynx system. Regional Science and Urban Economics, 75, 208–220.
- Polley, E., van der Laan, M., & Hubbard, A. (2013). Super learner in prediction. UC Berkeley Division of Biostatistics Working Paper Series.
- van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. Statistical Applications in Genetics and Molecular Biology, 6(1), Article 25.
- Zali, S., Pahlavani, P., Ghorbanzadeh, O., Khazravi, A., Ahmadlou, M., & Givekesh, S. (2025). Housing price modeling using a new geographically, temporally, and

characteristically weighted generalized regression neural network (GTCW-GRNN) algorithm. Buildings, 15(9), 1405.