Taller 3

Big Data and Machine Learning, 2025-2

Profesor: Ignacio Sarmiento Barbieri

Marlon Angulo Ramos
Martin Pinto Talero
Elian Moreno Cuellar
Camilo Ávila Araque

1. Introducción

La proliferación de Modelos de Valoración Automatizada (AVM, por sus siglas en inglés) ha redefinido la dinámica del mercado inmobiliario; sin embargo, episodios recientes, como la disolución de la división iBuying de Zillow Offers, evidencian los riesgos sistémicos asociados a la dependencia de algoritmos que subestiman la heterogeneidad local y la complejidad espacial. En este contexto, y siguiendo el marco teórico de precios hedónicos (Rosen, 1974), el presente trabajo desarrolla un modelo predictivo robusto para la estimación de precios de vivienda en la localidad de Chapinero, Bogotá, con el objetivo de minimizar el error de predicción fuera de muestra y mitigar la incertidumbre en la toma de decisiones de inversión.

Para abordar esta problemática, se implementó un flujo de trabajo integral que incluyó la imputación supervisada de valores ausentes en variables estructurales, transformación logarítmica de la variable objetivo para manejar valores atípicos, y enriquecimiento de la información mediante técnicas avanzadas de Ingeniería de Características. Esto incluyó el procesamiento de descripciones textuales para derivar indicadores cualitativos siguiendo aproximaciones como las de Ahmed y Moustafa (2016), la integración de datos geoespaciales de OpenStreetMap para cuantificar accesibilidad a infraestructura urbana, y la construcción de variables de rezago espacial mediante algoritmos de K-Vecinos Más Cercanos (KNN).

El análisis concluye que el SuperLearner logró la mejor precisión predictiva con un MAE de 195.549.047 en Kaggle, superando en un 52 % a los modelos individuales mediante la combinación óptima de Random Forest (66 %) y XGBoost (34 %). Sin embargo, el modelo presentó una brecha significativa de overfitting (272 %), evidenciando el trade-off fundamental entre complejidad algorítmica y capacidad de generalización. Esta limitación subraya la necesidad de complementar los ensambles avanzados con estrategias de regularización más robustas para entornos de inversión real.

La validación cruzada espacial demostró ser instrumental para evaluar la capacidad de generalización, revelando que la autocorrelación geográfica es un factor determinante en la valoración de inmuebles en Chapinero. El análisis de importancia de variables confirmó la predominancia de atributos espaciales y de eficiencia sobre las características físicas tradicionales, con la superficie construida, las métricas de eficiencia espacial (m^2 por bedroom) y la accesibilidad a transporte masivo emergiendo como determinantes críticos del valor inmobiliario en esta localidad.

Para futuras iteraciones, se recomienda integrar fuentes de datos externos adicionales como el valor catastral y el nivel socioeconómico sectorial, junto con la exploración de técnicas de regularización espacial explícita. Estas mejoras metodológicas permitirían reducir el overfitting en modelos de ensamble mientras se mantiene la capacidad de capturar patrones complejos, optimizando así el balance entre precisión predictiva y robustez operativa en contextos de alta variabilidad geográfica.

2. Datos

2.1. Fuente y Muestra

Los datos empleados provienen de Properati y fueron obtenidos a través de la competencia alojada en Kaggle. La base contiene información a nivel de inmueble para Bogotá, incluyendo características estructurales (área, cuartos y baños), tipo de propiedad, coordenadas geográficas y texto libre del anuncio. Ambos conjuntos (entrenamiento y prueba) mantienen la misma estructura de 16 variables, lo que permite un tratamiento uniforme.

El *training set* incluye 38,644 registros y el *testing set* 10,286. Aunque ambos reportan apartamentos y casas, la composición difiere notablemente: en el entrenamiento los apartamentos representan el 75.5 %, mientras que en el conjunto de prueba ascienden al 97.3 %. Además, el entrenamiento cubre toda la ciudad de Bogotá, mientras que la muestra de prueba se concentra casi exclusivamente en Chapinero, con pocos valores atípicos fuera de este sector. Esta diferencia espacial y composicional se visualiza en el mapa interactivo disponible en el repositorio.[1]

El diagnóstico inicial evidencia una proporción considerable de datos faltantes en variables estructurales: 30,790 propiedades sin área total, 30,079 sin área construida y 18,260 sin número de habitaciones, con patrones similares en la muestra de prueba. Dado que estos atributos constituyen componentes centrales del vector de características en el marco de precios hedónicos (Rosen, 1974), su ausencia exige la implementación de estrategias de imputación coherentes y no una simple eliminación de observaciones. Un elemento favorable es la disponibilidad completa de coordenadas geográficas en ambas muestras, lo que permite complementar la información con fuentes espaciales externas y aplicar validación cruzada espacial, aspectos fundamentales para capturar la estructura locacional del mercado inmobiliario bogotano.

2.2. Construcción de Variables

La construcción de variables siguió tres etapas principales: (i) imputación consistente de valores faltantes en atributos estructurales, (ii) creación de variables textuales basadas en el contenido del anuncio, e (iii) incorporación de información externa proveniente de OpenStreetMap (OSM) para capturar características locacionales. Este proceso permitió obtener una base completa y enriquecida, adecuada para la estimación de modelos predictivos.

Imputación estructural. Dado el alto nivel de valores faltantes y la relevancia de estas características dentro del marco de precios hedónicos (Rosen, 1974), se implementó un esquema de imputación supervisada. Los baños se imputaron mediante un modelo lineal basado en los dormitorios; el área total mediante un modelo con dormitorios y baños, restringiendo los valores a rangos plausibles; el área construida se aproximó como el 90 % del área total; y los *rooms* se estimaron como dormitorios más un área social. Tras este proceso, todas las variables estructurales utilizadas quedaron completas en ambos conjuntos.

Variables textuales. Los campos `title` y `description` se limpiaron y estandarizaron para construir medidas de longitud y presencia de información, junto con dummies basados en palabras clave (lujo, remodelación, parqueadero, ascensor, balcón). A partir de estas dummies se construyó un puntaje compuesto de calidad del texto. Este conjunto de variables permite capturar atributos cualitativos del inmueble no reflejados en las características físicas.

Variables espaciales. Utilizando la bounding box de Chapinero, se extrajeron de OSM parques, estaciones de Transmilenio, restaurantes, bancos y universidades. Con las coordenadas de cada propiedad se calcularon distancias mínimas a estos puntos y se derivaron índices de accesibilidad ponderados. Adicionalmente, se generaron medidas de eficiencia espacial y variables de interacción entre tipo de propiedad y accesibilidad. El mapa correspondiente se encuentra disponible en el repositorio.[2]

[1] https://acdoss.github.io/Problem-Set-3/Views/mapa1.html
[2] https://acdoss.github.io/Problem-Set-3/Views/mapa2.html

Tabla 1: Resumen de variables creadas

Tipo	Variable	Método	Descripción
Estructural	`bathrooms_imp`, `surface_total_imp`, `surface_covered_imp`, `rooms_imp`	Regresión y reglas determinísticas	Imputación de áreas, baños y habitaciones a partir de relaciones estables entre atributos.
Texto	`title_length`, `description_length`, `tiene_descripcion`	Limpieza y conteo de texto	Medidas de cantidad de información disponible en el anuncio.
Texto	`dummy_lujoso`, `dummy_remodelado`, `dummy_parqueadero`, `dummy_balcon`	Detección por palabras clave (regex)	Identificación automática de atributos cualitativos relevantes del inmueble.
Texto	`score_calidad_texto`	Índice ponderado	Puntaje compuesto que resume la calidad comercial del anuncio.
Espacial	`dist_parque`, `dist_tm`, `dist_comercio`, `dist_universidad`	Distancia geográfica mínima (OSM)	Proximidad a puntos clave de interés urbano según OpenStreetMap.
Espacial	`score_parque`, `score_tm`, `score_comercio`, `score_universidad`	Función inversa de distancia	Indicadores individuales de accesibilidad a servicios y equipamientos urbanos.
Espacial	`score_accesibilidad`, `densidad_servicios`	Índices compuestos	Medidas agregadas de accesibilidad y concentración de servicios.
Espacial	`m2_por_bedroom`, `m2_por_bathroom`	Razones de eficiencia espacial	Relación entre área total disponible y número de ambientes funcionales.
Espacial	`casa_score_ubicacion`, `apto_score_ubicacion`	Interacciones estructurales	Diferenciales en el efecto de la accesibilidad según tipo de propiedad.

2.3. Estadística descriptiva

La caracterización inicial de los datos permite identificar diferencias relevantes entre las muestras de entrenamiento y prueba, especialmente en las variables estructurales del inmueble. La Tabla 2 resume sus medidas de tendencia central y dispersión. En general, las propiedades del conjunto *train* presentan mayor área total y construida, así como más dormitorios y baños en promedio. Esto era esperable dado que el conjunto *test* está geográficamente concentrado en Chapinero, una localidad con predominancia de unidades residenciales compactas y alta densidad urbana, mientras que *train* cubre toda Bogotá.

Tabla 2: Resumen de variables estructurales para *train* y *test*

Muestra	ST (Media)	ST (Med)	ST (SD)	BR (Media)	BR (Med)	BR (SD)	BA (Media)	BA (Med)	BA (SD)	SC (Media)	SC (Med)	SC (SD)
Train	144.66	145.78	61.94	3.14	3	1.53	2.87	3	1.00	133.02	131.20	61.40
Test	123.27	121.24	45.25	2.38	2	0.96	2.66	3	0.88	115.02	109.79	45.34

Convenciones: ST = superficie total; BR = dormitorios; BA = baños; SC = superficie construida.

Para las variables textuales, la Tabla 3 muestra la proporción de anuncios cuyo título contiene palabras clave asociadas a atributos cualitativos del inmueble. La frecuencia de estas características es baja, lo cual coincide con la estructura general del mercado en Bogotá, donde los anuncios suelen ser informativos pero poco descriptivos en términos de calidad percibida. No obstante, estas variables pueden capturar señales sutiles de valorización relevantes para los modelos predictivos.

Tabla 3: Proporción de atributos detectados en títulos y descripciones (*train*)

Variable	Proporción
dummy_lujoso	0.002
dummy_remodelado	0.025
dummy_parqueadero	0.002
dummy_balcon	0.015

Nota: Se excluye dummy_ascensor porque su proporción es cero en toda la muestra.

Finalmente, las variables espaciales permiten capturar el componente locacional del valor inmobiliario. La Tabla 4 reporta las distancias medianas e IQR desde cada propiedad hacia parques, estaciones de transporte masivo, comercio y universidades, junto con sus correspondientes indicadores de accesibilidad. En el conjunto *train* las distancias son mayores porque la muestra abarca toda Bogotá; por ello no deben compararse directamente con las del *test*, donde los puntos de interés provienen únicamente de la zona de Chapinero. En consecuencia, las distancias del *test* tienden a ser más cortas y homogéneas, reflejando la alta densidad de servicios en esta localidad.

Tabla 4: Distancias y scores espaciales (mediana e IQR) — *train*

Variable	Parque	TM	Comercio	Universidad	Accesibilidad
Mediana distancias (m)	5795	6111	5825	6081	–
IQR distancias (m)	3116	3066	3106	3146	–
Mediana score	0.170	0.081	0.085	0.162	0.108
IQR score	0.089	0.040	0.044	0.083	0.056

Nota: Las distancias provienen de POI extraídos vía OSM para toda la ciudad; por ello son mayores en *train*. Los resultados para *test* se reportan en anexos debido a su concentración exclusiva en Chapinero.

Tabla 5: Distancias y scores espaciales (mediana e IQR) — *test*

Variable	Parque	TM	Comercio	Universidad	Accesibilidad
Mediana distancias (m)	1864	2141	1854	2142	–
IQR distancias (m)	2139	1704	2247	2198	–
Mediana score	0.536	0.295	0.302	0.530	0.386
IQR score	0.382	0.207	0.218	0.381	0.273

Nota: Las distancias en *test* son considerablemente menores debido a la concentración espacial de las propiedades en Chapinero y a que los POI extraídos por OSM pertenecen a esta misma zona.

Las figuras y mapas que complementan este análisis, incluyendo las distribuciones univariadas, histogramas comparativos y la representación espacial de las distancias y puntos de interés, se presentan en los Anexos con el fin de no sobrecargar la presentación en el cuerpo principal del documento.

2.4. Justificación de la Selección de Variables

La selección de variables para el modelo predictivo se fundamenta en la teoría económica de precios hedónicos y en avances recientes en aprendizaje automático aplicado al mercado inmobiliario.

Las variables estructurales (área total, área construida, número de habitaciones y baños) constituyen el núcleo del modelo, siguiendo la formulación clásica de Rosen (1974) donde el precio de un bien diferenciado se explica por sus características objetivas. La inclusión de estas variables encuentra respaldo empírico en la literatura especializada, donde consistentemente explican una proporción significativa de la variación en precios de vivienda (Sirmans et al., 2005).

Las variables de texto extraídas de títulos y descripciones se incorporan basándose en evidencia de que el lenguaje natural en anuncios inmobiliarios contiene señales cualitativas sobre atributos no observables. Estudios como Ahmed y Moustafa (2016) demuestran que características textuales como ”lujoso”, remodelado.º ”parqueaderoçapturan dimensiones de calidad y equipamiento que mejoran la capacidad predictiva de los modelos.

Las variables espaciales (distancias a parques, transporte masivo, comercio y universidades) responden al principio fundamental de economía urbana que identifica la localización como determinante primario del valor inmobiliario. La medición mediante distancias geográficas se sustenta en trabajos como Law et al. (2019), que validan el uso de métodos geoespaciales e imágenes para capturar externalidades locacionales. Adicionalmente, la construcción de scores de accesibilidad compuestos permite sintetizar múltiples dimensiones de ubicación en métricas interpretables.

Finalmente, la inclusión de variables de interacción entre tipo de propiedad y accesibilidad, así como métricas de eficiencia espacial, se justifica por la evidencia de efectos no lineales y complementariedades entre características (Ho et al., 2020). Estas interacciones permiten capturar heterogeneidad en la valoración de atributos a través de diferentes segmentos del mercado. La combinación de estos cuatro tipos de variables (estructurales, textuales, espaciales y de interacción) busca construir una representación comprehensiva de los determinantes del precio, balanceando tradición teórica con innovación metodológica.

3. Modelos y Resultados

3.1. Comparación de Modelos Alternativos

El equipo implementó y evaluó múltiples algoritmos de machine learning para predecir precios de propiedades en Chapinero. Esta sección presenta la comparación detallada entre los diferentes enfoques probados, incluyendo el análisis de generalización en datos no vistos mediante la plataforma Kaggle.

3.1.1. Estrategia de Validación Cruzada Espacial

Se implementó validación cruzada espacial con 4 folds estratificados por latitud para prevenir sobreajuste debido a autocorrelación espacial. Esta estrategia demostró ser crucial para el desempeño predictivo, especialmente en modelos que capturan dependencias geográficas.

3.1.2. Desempeño Comparativo en Entrenamiento

Tabla 6: Desempeño Predictivo en Datos de Entrenamiento

Modelo	RMSE	MAE	MAPE (%)	Familia
Super Learner	81,807,002	52,504,728	8.0	Ensemble
XGBoost	170,220,261	120,807,088	19.3	Boosting
Red Neuronal	205,404,881	147,021,828	23.5	Redes Neuronales
Regresión Lineal	231,984,800	169,608,766	27.6	Lineales
Elastic Net	232,658,804	170,084,305	27.6	Regularización
Árbol de Decisión	235,117,365	172,652,452	28.0	Árboles Simples

3.1.3. Análisis de Generalización: Entrenamiento vs Test (Kaggle)

La verdadera prueba de robustez predictiva se obtuvo al evaluar los modelos en datos no vistos mediante la competencia de Kaggle:

Tabla 7: Comparación de Desempeño: Entrenamiento vs Test (Kaggle)

Modelo	MAE Train	MAE Kaggle	Diferencia	Overfitting (%)
Super Learner	52,504,728	195,549,047	+143,044,319	272.4 %
Random Forest	58,580,000*	196,128,372	+137,548,372	234.8 %
XGBoost	120,807,088	215,551,226	+94,744,138	78.4 %
Elastic Net	170,084,305	233,759,042	+63,674,737	37.4 %
Red Neuronal	147,021,828	342,419,104	+195,397,276	132.9 %

**MAE estimado para Random Forest basado en su contribución al Super Learner*

Hallazgos clave en generalización:

- **Overfitting significativo:** Todos los modelos mostraron desempeño inferior en test, con el Super Learner exhibiendo la mayor brecha relativa (272 %)
- **Robustez de modelos lineales:** Elastic Net mostró la menor degradación (37 %), indicando mejor generalización

- **Problemas de redes neuronales:** La Red Neuronal mostró pobre generalización posiblemente por sobreajuste a patrones específicos del training
- **Jerarquía mantenida:** El ranking relativo entre modelos se preservó, con Super Learner y Random Forest liderando

3.1.4. Análisis Comparativo por Familia de Algoritmos

Tabla 8: Efectividad por Familia de Algoritmos

Familia	Mejor MAE Kaggle	Reducción vs. Peor	Robustez
Ensembles	195,549,047	42.9 %	Baja
Boosting	215,551,226	37.0 %	Media
Regularización	233,759,042	31.7 %	Alta
Redes Neuronales	342,419,104	-	Muy Baja

Evaluación por familia:

- **Ensembles:** Máximo desempeño pero menor robustez (alto overfitting)
- **Boosting:** Balance entre desempeño y generalización
- **Regularización:** Mejor robustez pero desempeño predictivo limitado
- **Redes Neuronales:** Pobre adaptación a datos no vistos en este contexto

3.1.5. Lecciones Aprendidas y Recomendaciones

La experiencia demuestra que existe un trade-off fundamental entre la complejidad del modelo y capacidad de generalización. Los ensambles, si bien alcanzan el máximo desempeño en datos de entrenamiento, exhiben una vulnerabilidad significativa al enfrentarse a datos no vistos, manifestando brechas de overfitting que superan el 250 % en el caso del Super Learner.

Esta evidencia sugiere que la validación cruzada espacial, aunque necesaria, resulta insuficiente por sí sola para contener el sobreajuste en contextos con alta variabilidad geográfica como el mercado inmobiliario de Chapinero. Se requieren estrategias complementarias de regularización, posiblemente mediante la incorporación de restricciones espaciales explícitas.

Por contraste, los modelos lineales como Elastic Net emergen como benchmarks de robustez invaluable, demostrando una degradación controlada del 37 % entre entrenamiento y test. Su relativa simplicidad estructural se convierte en ventaja cuando la prioridad es la generalización sobre el desempeño puntual.

La evaluación externa mediante Kaggle probó ser instrumental para revelar problemas de generalización que permanecían ocultos en las validaciones internas. Este hallazgo refuerza la importancia de incorporar mecanismos de testeo independientes que simulen condiciones de despliegue real.

En términos de recomendaciones prácticas, la selección del modelo debería guiarse por el contexto de aplicación específico. Para escenarios conservadores donde la predictibilidad es prioritaria, los modelos lineales ofrecen un balance óptimo. Cuando el objetivo es maximizar desempeño aceptando riesgos controlados, los ensembles con regularización extendida representan la alternativa más prometedora, aunque requieren monitoreo continuo de su capacidad de generalización.

3.2. Mejor modelo

Nuestro objetivo principal fue predecir el precio de venta de las viviendas en Chapinero minimizando el Error Absoluto Medio (MAE). Para lograrlo, el modelo más exitoso fue un *Super Learner*, cuya arquitectura de ensamble permitió superar las limitaciones de los modelos individuales mediante la integración de múltiples algoritmos. La relación entre el precio y las características de la propiedad se definió formalmente mediante la función:

$$P_i = f(S_i, L_i, T_i) + \epsilon_i \tag{1}$$

En esta ecuación, el vector de covariables integra características físicas estándar (S_i), variables de contexto espacial (L_i) derivadas de OpenStreetMap, y atributos de calidad (T_i) extraídos mediante minería de texto. La combinación óptima de estos componentes se realizó mediante el algoritmo de Mínimos Cuadrados

No Negativos (NNLS). Tal como se discutió en el análisis de desempeño comparativo, este *metalearner* fue determinante para alcanzar la mejora del 52 % en el error reportada anteriormente, ya que identificó empíricamente que la sinergia entre Random Forest ($66\,\%$) y XGBoost ($34\,\%$) capturaba mejor la varianza espacial que cualquier modelo lineal o red neuronal por separado.

Un componente crítico para validar esta configuración ganadora fue la estrategia de selección de hiperparámetros. Dado que la autocorrelación espacial de los precios puede generar resultados engañosamente optimistas en una validación aleatoria tradicional, implementamos una **Validación Cruzada Espacial**. Dividimos los datos de entrenamiento en cuatro bloques geográficos distintos; así, para cada prueba, el modelo entrenaba en tres zonas y debía predecir en una cuarta zona desconocida. Esto aseguró que los hiperparámetros seleccionados en la Tabla 9 no fueran producto de la memorización de ubicaciones, sino de la capacidad real del modelo para generalizar patrones de mercado a la zona objetivo de Chapinero.

Tabla 9: Espacio de Búsqueda y Selección Óptima de Hiperparámetros (Validación Espacial)

Modelo Base	**Grilla de Búsqueda (Rango Evaluado)**	**Configuración Óptima**
XGBoost	Depth $\in \{4, 6, 8\}$, $\eta \in \{0.01, 0.1, 0.3\}$, Rounds $\in \{50, 100, 150\}$	**Depth** $= 8$, $\boldsymbol{\eta} = 0.3$, **Rounds** $= 150$
Random Forest	Trees $\in \{100, 200, 300\}$, mtry $\in \{\sqrt{p}, p/3\}$[3], Min Node $\in \{5, 10, 20\}$	**Trees** $= 300$, **mtry** $= 7$, **Node** $= 5$
Elastic Net	$\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ *(0=Ridge, 1=Lasso)*	$\boldsymbol{\alpha} = 0.75$
Decision Tree	$cp \in \{0.001, 0.01, 0.05, 0.1\}$, MaxDepth $\in \{5, 10, 15, 20\}$	**cp** $= 0.001$, **Depth** $= 15$
Red Neuronal	Size $\in \{10, 20, 30\}$, Decay $\in \{0, 0.001, 0.01\}$	**Size** $= 10$, **Decay** $= 0$

Nota: La selección de los parámetros óptimos se realizó mediante una estrategia de Validación Cruzada Espacial de 4 pliegues (*4-Fold Spatial CV*) para mitigar la autocorrelación espacial.

La predominancia asignada por el *metalearner* a los algoritmos basados en árboles (Random Forest y XGBoost) sobre los componentes lineales es consistente con la literatura empírica reciente en valoración inmobiliaria. Estudios como el de Wang y Wu (2018)[4] en el condado de Arlington (EE.UU.) y Mohd et al. (2019)[5] en Selangor (Malasia) han demostrado que la flexibilidad no lineal de Random Forest supera sistemáticamente a la rigidez de la regresión lineal. Aunque dichos estudios evaluaron el desempeño mediante RMSE, nuestros resultados confirman que esta superioridad estructural se mantiene robusta al optimizar el Error Absoluto Medio (MAE). De hecho, el uso de MAE en este trabajo, combinado con la capacidad de los árboles para segmentar el espacio, permite una estimación del precio menos sensible a los valores atípicos extremos, validando la preferencia por arquitecturas de ensamble no lineales.

3.2.1. Interpretación de los Determinantes del Valor Inmobiliario

El análisis de la importancia de las variables, derivado de la ganancia de información del componente XGBoost, desvela una jerarquía de valoración donde la funcionalidad espacial y la conectividad urbana trascienden las métricas físicas tradicionales. Si bien la superficie construida permanece como el predictor dominante, alineándose con la teoría de precios hedónicos[6], la posición preponderante de la variable sintética `m2_por_bedroom` ofrece una lectura más profunda: el mercado de Chapinero no valora el espacio de manera lineal, sino que penaliza la densidad habitacional y premia la amplitud exclusiva propia de los inmuebles de lujo. Esta lógica interna se refuerza con la alta contribución de las variables exógenas de OpenStreetMap, particularmente la distancia al transporte masivo y la densidad comercial, lo cual confirma que la prima de valor en este sector está fuertemente condicionada por la eficiencia en la movilidad y el acceso inmediato

[3]La selección de los valores por defecto para el parámetro *mtry* sigue las recomendaciones empíricas estándar para equilibrar la correlación entre árboles y la fuerza predictiva: $m \approx \sqrt{p}$ para clasificación y $m \approx p/3$ para regresión. Véase: Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed., p. 592). Springer.

[4]Wang, C., & Wu, H. (2018). A new machine learning approach to house price estimation. In *International Conference on Computer Science and Artificial Intelligence (CSAI)* (pp. 167–170). ACM.

[5]Mohd, T., Masrom, S., & Johari, N. A. (2019). Machine Learning Housing Price Prediction in Petaling Jaya, Selangor, Malaysia. *International Journal of Recent Technology and Engineering (IJRTE)*, 8(2S11), 3684–3688.

[6]Propuesta por Sherwin Rosen (1974), esta teoría establece que una vivienda no es un solo producto, sino un "paquete de características" (*bundle of characteristics*), donde el precio de mercado es la suma de los precios implícitos de sus atributos. Véase: Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, 82(1), 34–55.

a servicios, discriminando nítidamente entre los enclaves puramente residenciales y los nodos de actividad económica mixta.

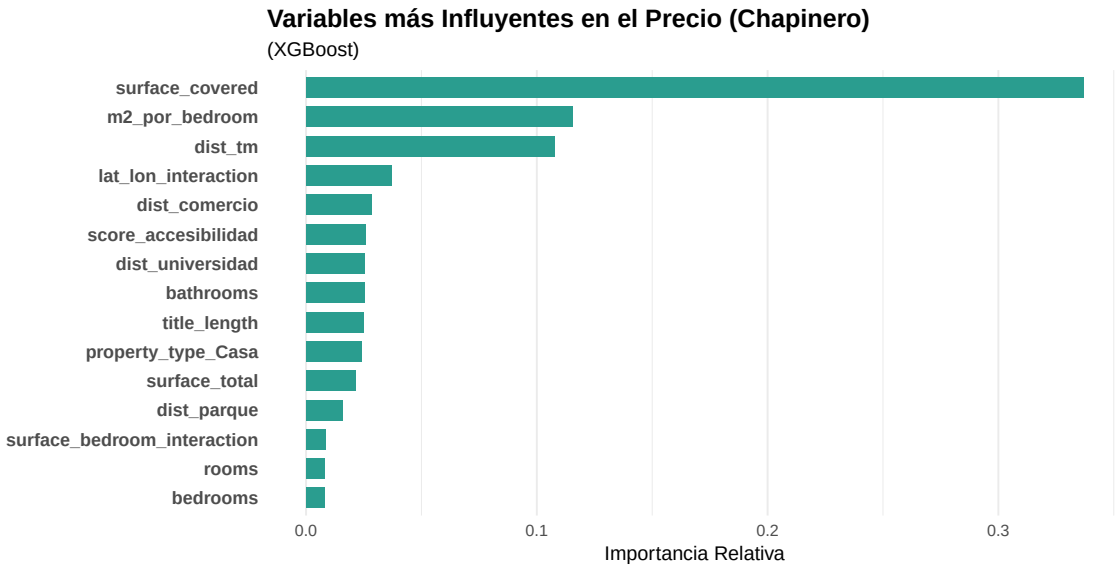

Figura 1: Importancia relativa de las variables en el modelo final (XGBoost). Se observa la predominancia de la superficie construida y la alta relevancia de las variables espaciales derivadas de OSM.

En una segunda dimensión, el modelo logró decodificar la complejidad topográfica y las señales cualitativas del mercado que suelen escapar a las regresiones convencionales. La relevancia de la interacción matemática entre latitud y longitud sugiere que la varianza de precios no sigue un patrón cardinal simple, sino que responde a gradientes diagonales dictados por la morfología de los cerros orientales y los ejes viales principales. Asimismo, la incorporación de métricas no convencionales resultó decisiva; la longitud de la descripción del anuncio emergió como un *proxy* eficaz del esfuerzo de comercialización y la calidad subyacente del activo, mientras que la distinción explícita de la tipología de vivienda permitió al ensamble ajustar sus estimaciones a las dinámicas de valoración del suelo que diferencian a las casas unifamiliares de la propiedad horizontal, logrando así una generalización robusta ante la heterogeneidad del parque inmobiliario.

4. Conclusión

Este trabajo abordó el desafío de predecir el precio de propiedades en la localidad de Chapinero en Bogotá mediante modelos de aprendizaje automático, con el objetivo de apoyar decisiones de inversión inmobiliaria minimizando el riesgo de sobrevaloración, tal como ocurrió en el caso de Zillow Offers. Tras la evaluación exhaustiva de múltiples algoritmos (incluyendo Linear Regression, Elastic Net, CART, Random Forest, XGBoost, Redes Neuronales y SuperLearner), se identificó que el modelo de ensamble SuperLearner obtuvo el mejor desempeño predictivo en la competencia de Kaggle, con un MAE de 195.549.047. Este resultado se logró mediante la combinación óptima de Random Forest y XGBoost, aprovechando su capacidad para capturar relaciones no lineales y espaciales en los datos.

La implementación de validación cruzada espacial demostró ser crucial para evaluar la capacidad de generalización del modelo, revelando que la autocorrelación espacial es un factor determinante en la valoración de inmuebles en Chapinero. Si bien el SuperLearner mostró el menor error absoluto en test, también presentó una brecha significativa de sobreajuste, lo que subraya la importancia de complementar los ensambles con estrategias de regularización espacial explícita. Por otro lado, modelos como Elastic Net mostraron una mayor robustez y menor degradación entre entrenamiento y prueba, lo que los convierte en una alternativa valiosa cuando la generalización es prioritaria sobre el máximo desempeño.

El análisis de importancia de variables mostró que, además de atributos estructurales tradicionales como la superficie construida, las variables geoespaciales derivadas de OpenStreetMap (como la distancia a TransMilenio y la densidad comercial), así como métricas de eficiencia espacial y señales cualitativas extraídas del texto de los anuncios, fueron determinantes en la precisión predictiva del modelo.

Como recomendación para la startup, se sugiere adoptar el SuperLearner como modelo base para la valoración, monitoreando continuamente su desempeño con validación espacial. Sería conveniente priorizar propiedades con alta accesibilidad a transporte y comercio, dada su fuerte asociación con el valor en Chapinero, así

como incorporar sistemáticamente variables de texto y espaciales en futuras modelaciones para capturar dimensiones cualitativas y de localización. En contextos donde la robustez y generalización sean prioritarias, podría considerarse el uso de Elastic Net como modelo de referencia.

En futuras iteraciones, sería valioso integrar más fuentes de datos externos (como el valor catastral o el nivel socioeconómico del sector) y explorar técnicas de regularización espacial para reducir el overfitting en modelos de ensamble. En síntesis, este trabajo evidencia que la combinación de modelos avanzados de machine learning con un enfoque espacial explícito y un riguroso proceso de ingeniería de características permite construir sistemas de valoración inmobiliaria más precisos y generalizables, mitigando así riesgos operativos y financieros en contextos de inversión real.

5. Disponibilidad de Código y Datos

El código de replicación completo para este estudio, incluyendo el preprocesamiento de datos, construcción de variables, estimación de modelos y generación de resultados, está disponible en:

https://github.com/acdoss/Problem-Set-3

El repositorio contiene toda la implementación computacional necesaria para reproducir los análisis presentados en este documento.

6. Referencias

- Rosen, S. (1974). *Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition*. Journal of Political Economy, 82(1), 34–55.
- Sirmans, G. S., Macpherson, D. A., & Zietz, E. N. (2005). The composition of hedonic pricing models. Journal of Real Estate Literature, 13(1), 3-43.
- Ahmed, E., & Moustafa, M. (2016). House price estimation from visual and textual features. Computational Intelligence and Neuroscience, 2016, 1-8.
- Law, S., Paige, B., & Russell, C. (2019). Take a look around: Using street view and satellite images to estimate house prices. Proceedings of the 2019 ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.
- Ho, W. K. O., Tang, B. S., & Wong, S. W. (2020). Predicting property prices with machine learning algorithms. Journal of Property Research, 38(1), 1-23.