Predicción de Precios de Viviendas en Bogotá Usando Machine Learning y Datos Enriquecidos

Dickinson Román Arismendy Torres

Universidad Sergio Arboleda
Escuela de Ciencias Exactas e Ingeniería
Bogotá D.C.
2025

Predicción de Precios de Viviendas en Bogotá Usando Machine Learning y Datos Enriquecidos

Dickinson Román Arismendy Torres

Maestría en Inteligencia Artificial

Asesor
Phd. Juan Pablo Ospina Lopez

Universidad Sergio Arboleda
Escuela de Ciencias Exactas e Ingeniería
Maestría en Inteligencia Artificial
Bogotá D.C.
2025

Contenido

UNIVERSIDAD
SERGIO ARBOLEDA

UNIVERSIDAD
SERGIO ARBOLEDA

Lista de figuras

UNIVERSIDAD
SERGIO ARBOLEDA

UNIVERSIDAD
SERGIO ARBOLEDA

Lista de tablas

UNIVERSIDAD
SERGIO ARBOLEDA

UNIVERSIDAD
SERGIO ARBOLEDA

Glosario

- **Método hedónico**: Técnica econométrica que estima el valor de un bien a partir de sus características intrínsecas y extrínsecas. En el caso de los bienes raíces, este método permite evaluar cómo factores como el tamaño, ubicación y calidad afectan el precio de una propiedad.
- **Web scraping**: Técnica utilizada para extraer datos de sitios web de manera automatizada mediante el uso de herramientas y bibliotecas de programación, como BeautifulSoup o Scrapy en Python.
- **PSO (Particle Swarm Optimization)**: Algoritmo de optimización basado en la inteligencia colectiva de grupos, inspirado en el comportamiento de enjambres como aves o peces. Se utiliza para resolver problemas complejos mediante iteraciones en busca de soluciones óptimas.
- **Especulación**: Práctica económica que consiste en la compra de bienes, como propiedades inmobiliarias, con el objetivo de obtener ganancias a través del aumento de su precio, a menudo contribuyendo a la inflación de precios y dificultando el acceso a dichos bienes para sectores de bajos ingresos.
- **WGS84 (EPSG:4326)**: sistema geodésico mundial que define la forma de la Tierra y un sistema de coordenadas geográficas en grados (latitud/longitud). Base de GPS y de la mayoría de datasets geoespaciales. *Ver Anexo* **??**. [**?**]
- **Web Mercator (EPSG:3857)**: proyección cartográfica pseudo-mercator usada por la mayoría de mapas web. Expresa coordenadas en metros, útil para cálculos de distancia en entornos urbanos. *Ver Anexo* **??**. [**?**]
- **SRID**: identificador numérico de un sistema de referencia espacial (p. ej., 4326 o 3857). En PostGIS determina cómo interpretar y transformar geometrías entre sistemas. *Ver Anexo* **??**. [**?**]

UNIVERSIDAD
SERGIO ARBOLEDA

UNIVERSIDAD
SERGIO ARBOLEDA

Resumen

El resumen es una presentación abreviada y precisa del contenido de un documento, sin agregar interpretación o crítica. Para documentos extensos como informes, tesis y trabajos de grado, no debe exceder de 500 palabras, y debe ser lo suficientemente breve para que no ocupe más de una página [**?**].

palabras clave: palabra 1, palabra 2, palabra 3, ...

UNIVERSIDAD
SERGIO ARBOLEDA

UNIVERSIDAD
SERGIO ARBOLEDA

Capítulo 1

Introducción

De acuerdo con la NTC1486, la introducción es un espacio dónde el autor presenta y señala la importancia, el origen (los antecedentes teóricos y prácticos), los objetivos, los alcances, las limitaciones, la metodología empleada, el significado que el estudio tiene en el avance del campo respectivo y su aplicación en el área investigada. No debe confundirse con el resumen, ni contener un recuento detallado de la teoría, el método o los resultados, como tampoco anticipar las conclusiones y recomendaciones [**?**].

Para el desarrollo del documento utilizando la plantilla en LaTeXse recomienda la guía *the not so short guide to LaTeX* [**?**] que brinda una introducción breve pero muy completa. En donde se presentan entro otras cosas los comandos y ambeintes para el trabajo con gráficas y tablas, como las que se muestran a continuación:

Figura 1.1: Muestra de inclusioón de un elemento gráfico

Así la Figura 1.1 muestra una imagen que se encuentra en el directorio images, mientras la tabla 1.1, muestra dos ejemplos de información tabular con combinación de columnas, y combinación de filas.

UNIVERSIDAD
SERGIO ARBOLEDA

Cuadro 1.1: Ejemplo de tabla con multi columnas (arriba) y multi filas(abajo)

Country List			
Country Name or Area Name	ISO ALPHA 2 Code	ISO ALPHA 3 Code	ISO numeric Code
Afghanistan	AF	AFG	004
Aland Islands	AX	ALA	248
Albania	AL	ALB	008
Algeria	DZ	DZA	012
American Samoa	AS	ASM	016
Andorra	AD	AND	020
Angola	AO	AGO	024

col1	col2	col3
Multiple row	cell2	cell3
	cell5	cell6
	cell8	cell9

Además se recomienda consultar los manuales que aparecen en al sección de aprendizaje de overleaf, por ejemplo el de tablas.

UNIVERSIDAD SERGIO ARBOLEDA

Capítulo 2

Estado del Arte

Machine Learning en la Predicción de Precios Inmobiliarios

El *machine learning* (ML) es una rama de la inteligencia artificial que permite a las máquinas aprender de los datos y realizar predicciones sin necesidad de ser programadas explícitamente para cada tarea. En el contexto inmobiliario, los modelos de ML han demostrado ser particularmente útiles para predecir precios de propiedades, capturando patrones complejos en los datos que los métodos tradicionales no pueden detectar. Su aplicabilidad en el mercado inmobiliario surge de la capacidad de manejar grandes volúmenes de datos con múltiples variables, lo que mejora la precisión de las predicciones y facilita la toma de decisiones tanto para compradores como para inversionistas [**?**].

En Colombia, el mercado inmobiliario está influenciado por diversas características, como la ubicación geográfica, la seguridad y convivencia de las zonas, así como la cercanía a sitios de interés como centros comerciales y colegios. Con la disponibilidad de datos abiertos y técnicas de *web scraping*, es posible recopilar información de diversas fuentes para enriquecer los modelos predictivos. Esta integración de características adicionales, como indicadores de seguridad, mapas de sitios cercanos, y boletines económicos, permite que los modelos de ML capturen mejor las relaciones no lineales y complejas entre los atributos de las viviendas y sus precios, como se observó en estudios previos [**?**].

0.1. Metodologías Tradicionales y su Limitación

Los métodos hedónicos (tradicionales), ampliamente utilizados en estudios inmobiliarios, han sido la base para estimar precios a partir de características como el área, el número de habitaciones y la ubicación [**?**]. Sin embargo, estos modelos lineales tienden a fallar cuando las relaciones entre las variables son no lineales o cuando intervienen factores externos complejos. Además, su desempeño depende en gran medida de la calidad y la disponibilidad de los datos.

Regresión Lineal

La regresión lineal es uno de los métodos más simples y utilizados para la predicción de precios inmobiliarios. Este modelo asume una relación lineal entre las características de las propiedades (como área habitable, número de habitaciones, ubicación) y los

UNIVERSIDAD SERGIO ARBOLEDA

precios. Aunque es un buen punto de partida, es limitado cuando las relaciones entre las variables no son lineales. En estudios como el de Kim et al. [**?**], se evidenció que la regresión lineal tiende a ser superada por modelos más complejos en mercados con dinámicas no lineales.

Ventajas: Simple de implementar y explicar.
Desventajas: No captura relaciones no lineales y es sensible a los *outliers*.
Casos de uso: Predicción en mercados con relaciones lineales bien definidas.

Árboles de Decisión

Los árboles de decisión son modelos no paramétricos que dividen los datos en subconjuntos más pequeños en función de las características más importantes. Este enfoque es especialmente útil cuando se desea interpretar los resultados, ya que los árboles permiten una visualización clara de cómo las características afectan el precio. Sin embargo, tienden a sobreajustarse a los datos de entrenamiento si no se controlan adecuadamente [**?**].

Ventajas: Fácil de interpretar y manejar datos con interacciones no lineales.
Desventajas: Propenso al sobreajuste, a menos que se utilicen técnicas de poda.
Casos de uso: Situaciones donde la interpretabilidad es importante y se requiere manejar relaciones no lineales.

Random Forest

Random Forest es un algoritmo de aprendizaje conjunto que utiliza múltiples árboles de decisión para mejorar la precisión y reducir el riesgo de sobreajuste. Cada árbol se entrena en un subconjunto diferente de los datos, lo que reduce la variabilidad y mejora la generalización. En estudios como el de Zhang et al. [**?**], Random Forest fue uno de los modelos más efectivos, debido a su capacidad para manejar un gran número de variables y sus interacciones.

Ventajas: Robusto frente al sobreajuste, puede manejar grandes conjuntos de datos y múltiples variables.
Desventajas: Menos interpretativo que un único árbol de decisión y requiere más recursos computacionales.
Casos de uso: Mercados inmobiliarios complejos con muchas características y donde las relaciones no son lineales.

XGBoost

XGBoost es un algoritmo de *boosting* que mejora iterativamente los errores de predicción de modelos más simples. En cada iteración, se construyen nuevos árboles que corrigen los errores de los árboles anteriores. Este modelo ha demostrado ser extremadamente eficaz en competiciones de predicción de precios debido a su alta precisión y velocidad. En estudios como el de Smith et al. [**?**], XGBoost mostró un rendimiento

UNIVERSIDAD
SERGIO ARBOLEDA

superior al de otros modelos de *machine learning*, debido a su capacidad para manejar datos ruidosos y *outliers*.

Ventajas: Alta precisión, robustez frente a datos ruidosos y capacidad para ajustar hiperparámetros de manera efectiva.
Desventajas: Requiere más tiempo de entrenamiento y ajuste de hiperparámetros.
Casos de uso: Situaciones donde se necesita maximizar la precisión de las predicciones en mercados inmobiliarios dinámicos y no lineales.

Redes Neuronales

Las redes neuronales son modelos de *machine learning* que se inspiran en la estructura del cerebro humano, lo que les permite aprender patrones complejos en los datos. Las redes neuronales profundas (*deep learning*) son especialmente útiles cuando se manejan grandes volúmenes de datos con muchas variables. En estudios como el de Kim et al. [**?**], las redes neuronales demostraron ser efectivas en la captura de relaciones no lineales complejas, aunque requieren una gran cantidad de datos y recursos computacionales.

Ventajas: Captura relaciones no lineales complejas y se adapta bien a grandes volúmenes de datos.
Desventajas: Difícil de interpretar, requiere muchos datos y potencia computacional.
Casos de uso: Situaciones donde se dispone de grandes volúmenes de datos y la relación entre las características es altamente no lineal.

Comparación de Modelos

Cada modelo tiene sus ventajas y limitaciones, por lo que la elección depende de las características del mercado inmobiliario en estudio. En el contexto colombiano, con la incorporación de variables adicionales como mapas de sitios de interés, proximidad a comercios y estaciones de transporte público, es probable que modelos como Random Forest y XGBoost, que pueden manejar relaciones complejas, resulten ser los más efectivos. Por otro lado, la regresión lineal puede ser útil en casos donde las relaciones entre variables son más simples.

La comparación de modelos en términos de precisión y eficiencia es esencial para seleccionar el enfoque más adecuado. Estudios previos han demostrado que los modelos de ensamble, como *Random Forest* y *Gradient Boosting*, ofrecen mejores resultados en la mayoría de los casos [**?**, **?**]. Adicionalmente, las redes neuronales profundas (*Deep Learning*) han sido exploradas en escenarios donde se dispone de grandes volúmenes de datos [**?**].

0.2. Aplicación de Machine Learning en la Predicción de Precios

El uso de machine learning ha demostrado ser una alternativa efectiva a los métodos tradicionales. Modelos como *Random Forest*, *Gradient Boosting Machines (GBM)*, y

UNIVERSIDAD
SERGIO ARBOLEDA

Support Vector Machines (SVM) han mostrado mejores resultados en la predicción de precios debido a su capacidad para manejar relaciones no lineales y variables complejas [**?**, **?**]. En particular, el modelo *Gradient Boosting* se destacó por su precisión en varios estudios [**?**].

0.3. Uso de Variables Externas en la Predicción de Precios

Los factores externos, como la proximidad a centros comerciales, colegios y hospitales, y los índices de seguridad, son determinantes clave en los precios inmobiliarios. Estudios recientes han integrado estos factores en los modelos predictivos, mejorando significativamente su rendimiento [**?**, **?**]. Sin embargo, la mayoría de los trabajos se enfocan en mercados específicos, dejando un vacío en la generalización de estos enfoques a mercados emergentes como Bogotá.

0.4. Desafíos y Oportunidades en la Predicción de Precios

Entre los principales desafíos se encuentran la falta de datos completos y la dependencia de los modelos de machine learning en datos de alta calidad [**?**]. Sin embargo, la integración de técnicas híbridas, como SVM optimizado por PSO (Particle Swarm Optimization), ofrece oportunidades para superar estas limitaciones [**?**].

0.5. Identificación de Brechas en la Literatura

A pesar de los avances en la integración de variables externas, la mayoría de los estudios se han enfocado en mercados desarrollados, dejando un vacío en mercados emergentes como Bogotá. Este proyecto busca llenar esta brecha aplicando enfoques modernos a un contexto local y utilizando datos enriquecidos para mejorar la precisión de las predicciones.

0.6. Conclusión

El estado del arte demuestra que las técnicas de machine learning y el enriquecimiento de datos ofrecen una mejora significativa en la predicción de precios de bienes raíces. Sin embargo, existe una necesidad clara de explorar estos enfoques en mercados emergentes y evaluar el impacto de factores externos en los precios. Este proyecto se posiciona como una contribución relevante al abordar estas brechas y expandir las aplicaciones de machine learning en el sector inmobiliario.

UNIVERSIDAD
SERGIO ARBOLEDA

Capítulo 3

Problema de investigación

El déficit habitacional en Colombia, caracterizado por la falta de acceso a vivienda digna y equitativa, es una problemática compleja que afecta a millones de personas. Según Castillo (2004), el déficit habitacional en el país incluye tanto aspectos cuantitativos, como la insuficiencia de viviendas disponibles, como cualitativos, relacionados con la calidad de las viviendas y su entorno [**?**]. Este problema es especialmente agudo en Bogotá, donde el crecimiento desorganizado y la urbanización informal han exacerbado la marginalidad y la exclusión [**?**].

Adicionalmente, el acceso limitado al financiamiento hipotecario y la especulación inmobiliaria contribuyen a distorsionar el mercado, encareciendo los precios y excluyendo a los hogares de bajos ingresos. Más del 80 % de los hogares de bajos ingresos no tiene acceso a créditos hipotecarios, lo que agrava la inequidad en el acceso a la vivienda [**?**]. Este panorama se ve reflejado en los cambios recientes en las políticas de subsidios, como *Mi Casa Ya*, cuya implementación irregular ha afectado tanto a desarrolladores como a compradores, aumentando la incertidumbre en el sector [**?**].

En paralelo, la vivienda también ha sido utilizada históricamente como un instrumento de especulación financiera y política. La falta de integración de datos cualitativos, como la calidad de los servicios públicos, la accesibilidad y el equipamiento urbano, limita la capacidad de los modelos actuales de predicción de precios para reflejar las dinámicas reales del mercado. Los modelos tradicionales, basados únicamente en características extraídas de portales inmobiliarios, no consideran los factores sociales y económicos que afectan directamente los precios, como destacó Walter (2023) en su análisis de la desaceleración del sector vivienda en Colombia [**?**, **?**].

UNIVERSIDAD
SERGIO ARBOLEDA

UNIVERSIDAD
SERGIO ARBOLEDA

Capítulo 4

Justificación

La vivienda es un derecho fundamental reconocido en la Constitución de Colombia [?], pero la inequidad en el acceso, el déficit habitacional y la especulación inmobiliaria muestran que este derecho no se garantiza de manera efectiva. Desarrollar un modelo predictivo robusto, que incorpore variables estructuradas y contextuales, podría ayudar a mitigar los problemas asociados al mercado inmobiliario. Este enfoque no solo permitiría mejorar las estimaciones de precios, sino también diseñar políticas públicas más efectivas y basadas en evidencia [?, ?, ?].

La predicción precisa de precios de viviendas es un aspecto crucial en el mercado inmobiliario, no solo para compradores y vendedores, sino también para inversionistas, desarrolladores y gobiernos locales. Cuando las estimaciones de precios son imprecisas o están basadas en datos incompletos, se generan varios problemas que afectan tanto al mercado como a los actores involucrados.

Decisiones de compra y venta desinformadas

Los propietarios que subestiman el valor de su vivienda pueden venderla por debajo de su verdadero precio de mercado, perdiendo oportunidades de obtener una mayor ganancia. Por otro lado, los compradores que reciben estimaciones infladas pueden adquirir propiedades por encima de su valor real, enfrentando dificultades para revenderlas o recuperar la inversión. Este fenómeno afecta directamente la equidad y la transparencia del mercado, creando desconfianza entre los participantes.

Desigualdad en la accesibilidad a la vivienda

Cuando las estimaciones de precios no son precisas, las zonas de alto crecimiento pueden ser sobrevaloradas, y las de menor crecimiento pueden ser infravaloradas, lo que provoca una distorsión en la accesibilidad a la vivienda. Esto puede llevar a que sectores de la población, especialmente los de menores ingresos, sean excluidos de áreas en proceso de valorización, acelerando fenómenos de gentrificación y desplazamiento de comunidades.

Burbujas inmobiliarias y volatilidad en el mercado

Las malas estimaciones de precios pueden contribuir a la formación de burbujas inmobiliarias, donde los precios se inflan artificialmente debido a una sobrevaloración de las propiedades. Cuando la burbuja estalla, los precios caen bruscamente, lo que

UNIVERSIDAD
SERGIO ARBOLEDA

provoca una crisis de confianza en el mercado y pérdidas económicas significativas para propietarios e inversionistas. Este tipo de volatilidad afecta la estabilidad financiera de las familias y las inversiones de largo plazo.

Dificultades para la planificación urbana

Una buena estimación de precios es clave para la planificación urbana y la asignación de recursos en infraestructura, servicios públicos y desarrollo sostenible. Si las predicciones de precios no reflejan adecuadamente el valor futuro de las propiedades, los gobiernos locales y desarrolladores pueden tomar decisiones incorrectas sobre dónde invertir en infraestructura y servicios. Esto puede resultar en zonas sobrepobladas sin servicios adecuados o en áreas subdesarrolladas que no reciben suficiente inversión.

Pérdida de confianza de inversionistas

Los inversionistas dependen de estimaciones precisas de precios para identificar oportunidades de crecimiento en diferentes zonas. Si los modelos utilizados no logran prever con exactitud la evolución de los precios, los inversionistas podrían enfrentar pérdidas o no alcanzar la rentabilidad esperada. Esto desalienta la inversión en el sector inmobiliario y puede impactar negativamente en el desarrollo económico local.

Impacto en el crédito hipotecario

Las instituciones financieras basan sus decisiones de otorgamiento de crédito hipotecario en evaluaciones precisas del valor de las propiedades. Cuando estas evaluaciones están basadas en estimaciones incorrectas, el riesgo de impagos o de tener propiedades sobrevaluadas en el portafolio de préstamos aumenta. Esto puede derivar en pérdidas para los bancos y en restricciones más severas para otorgar créditos, afectando a las familias que buscan adquirir vivienda.

La correcta predicción de precios, basada en datos estructurados y enriquecidos, no solo mejorará la transparencia y eficiencia del mercado inmobiliario en Bogotá, sino que también evitará estos problemas derivados. Por lo tanto, el desarrollo de un modelo de predicción que incorpore tanto características internas de las propiedades como datos externos (seguridad, cercanía a servicios, etc.) podría ser fundamental para la sostenibilidad y el crecimiento del mercado inmobiliario, promoviendo una mayor equidad y una mejor toma de decisiones por parte de todos los actores involucrados.

UNIVERSIDAD
SERGIO ARBOLEDA

Capítulo 5

Objetivos

1. Objetivo general

Desarrollar un modelo de predicción de precios de viviendas en Bogotá utilizando técnicas de *machine learning*, que incorpore tanto características internas de las propiedades (como tamaño, número de habitaciones y ubicación) como datos externos (indicadores de seguridad, proximidad a puntos de interés, y accesibilidad a servicios), con el fin de mejorar la precisión en la estimación de precios y contribuir a la toma de decisiones informada en el mercado inmobiliario.

2. Objetivos específicos

- Crear una base de datos consolidada a partir de la recolección de información mediante *web scraping* de diversas plataformas inmobiliarias en Bogotá.
- Enriquecer los datos obtenidos integrando información adicional proveniente de fuentes externas, como indicadores de seguridad, convivencia y proximidad a servicios.
- Seleccionar el mejor modelo de predicción de precios a partir de la comparación de un conjunto de modelos desarrollados, evaluando su rendimiento mediante métricas de precisión y eficiencia.

UNIVERSIDAD
SERGIO ARBOLEDA

UNIVERSIDAD
SERGIO ARBOLEDA

Capítulo 6

Metodología de investigación

Metodología

La metodología implementada comprende: (i) adquisición de datos base, (ii) limpieza y preprocesamiento con umbrales y reglas reproducibles, (iii) enriquecimiento geoespacial mediante capas del distrito y conteos de POIs (OSM), (iv) entrenamiento y evaluación de modelos base y aumentados, y (v) persistencia y exposición vía API.

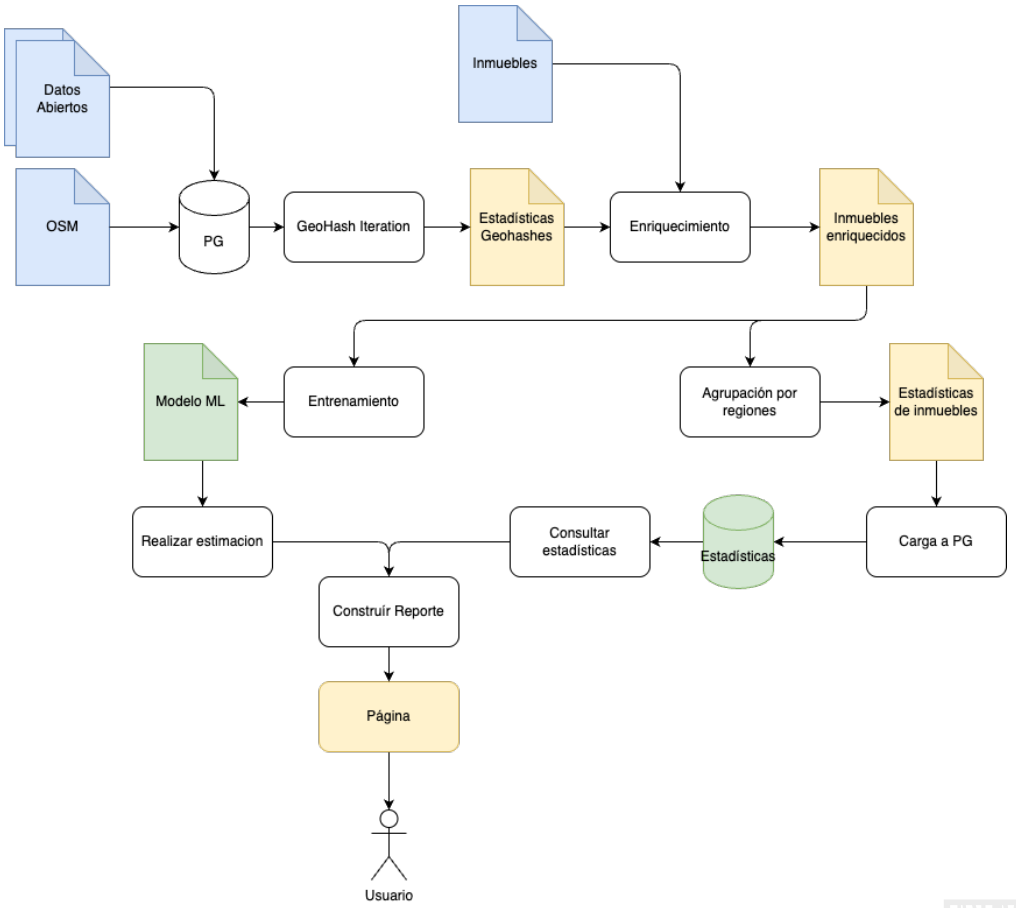

Figura 6.1: Proceso general de la metodología

UNIVERSIDAD SERGIO ARBOLEDA

Fuentes de datos

- **Datos de inmuebles (base)**: JSON único de agosto de 2024 publicado en GitHub (`builker-col/bogota-apartments`). Se procesó a CSV.
- **Datos abiertos del distrito (PostGIS)**: capas consultadas en agosto de 2025: `barrios_bogota`, `upz_bogota`, `localidades_bogota`, `estratos_manzana`, `avaluo_catastral_manz` y POIs OSM (`gis_osm_pois_free_1`, `gis_osm_pois_a_free_1`), con SRID esperado 4326.

Limpieza y preprocesamiento

Realizado en Python (*Pandas*, *NumPy*, *scikit-learn*) sobre el CSV base:

- **Outliers**: filtro por percentil 99. $\textit{Área} \leq 464\,m^2$, $\textit{precio\_venta} \leq 5{,}4 \times 10^9$ COP.
- **Precio mínimo**: $\geq$ 50,000,000 COP.
- **Área igual a 0**: imputación por mediana de comparables (mismo *estrato*, *habitaciones*, *banos*, *sector*); si no hay, mediana por *estrato*.
- **Parqueaderos negativos**: reemplazo por moda dentro del mismo *estrato* (sino, moda global).
- **Coordenadas**: imputación por mediana del *sector* o global cuando estén fuera de Bogotá; bounding box final: lat $[4{,}4, 4{,}9]$, lon $[-74{,}3, -73{,}9]$. Registros fuera se eliminan tras imputación.
- **Estrato fuera de rango [1–6]**: imputación por modo del *sector*; si no hay, modo global.

Enriquecimiento geoespacial

Se calcula por propiedad (lat, lon) con un pipeline asíncrono y persistencia en PostGIS:

- **Conteos de POIs OSM** por radios de 100, 300, 500, 1000 y 2000 metros en categorías agregadas: *education*, *healthcare*, *retail_access*, *dining_and_entertainment*, *accommodation*, *parks_and_recreation*, *infrastructure_services*, *cultural_amenities*.
- **Metadatos regionales**: asignación de UPZ, barrio y localidad; variables *upz_calculada*, *barrio_calculado*, *localidad_calculada*.
- **Valuación por geohash**: promedios de *catastral* y *comercial* por celda (geohash), usando `avaluo_catastral_manzana`.
- **Persistencia y estadísticas**: escritura en `property_data` y agregación en `region_stats` (barrio, UPZ, localidad) vía `ST_Contains`, con n, medias, desviaciones y cuartiles.

UNIVERSIDAD SERGIO ARBOLEDA

Modelos base

Conjunto de modelos evaluados con validación cruzada $KFold = 5$, `shuffle=True`, `random_state=42` métrica RMSE (escala real). Preprocesamiento: `SimpleImputer` (media/moda), `StandardScaler`, `OneHotEncoder`. Resultados (aprox.):

- **Random Forest**: RMSE $\approx$ 245M (CV); en hold-out (20 %): RMSE $\approx$ 250.6M, MAE $\approx$ 129.9M, $R^2 \approx 0.915$.
- **XGBoost / LightGBM**: RMSE $\approx$ 245–246M (CV).
- **Lineales**: $\approx$ 348M; **SVR**: $\approx$ 913M.

Se evaluó además con log(*precio_venta*) y log(*area*), mejorando métricas en escala original.

Modelos con datos aumentados

Se entrenaron variantes con variables enriquecidas y selección reducida:

- **v0 (XGB)**: log(*area*), one-hot; variables estructurales dominan; enriquecidas con aporte marginal.
- **v1 (XGB reducido + barrio_top)**: hold-out RMSE $\approx$ 254.66M, MAE $\approx$ 136.54M, $R^2 \approx 0.9139$. Modelo exportado como `xgboost_model_2.1.pkl`.
- **v2 (XGB con búsqueda aleatoria)**: mejores hiperparámetros: $n\_estimators = 500$, $max\_depth = 9$, $learning\_rate = 0{,}05$, $subsample = 0{,}8$, $colsample\_bytree = 0{,}8$, $\alpha = 0$, $\lambda = 1$. Hold-out: RMSE $\approx$ 233.49M, MAE $\approx$ 121.96M, $R^2 \approx 0.9276$. Exportado como `xgboost_model_2.2.pkl`.

Exposición de resultados y API

Se expone un endpoint `GET /api/estimate` (FastAPI) que combina estadísticas del punto (POIs, región, valuación) y la predicción del modelo: agrega `get_point_stats`, `get_region_stats` y `estimate(...)` sobre la versión de modelo configurada (por defecto 2.1; se recomienda 2.2 por mejor RMSE).

Reproducibilidad

Los notebooks en `analisis/notebooks/` contienen el detalle de extracción, limpieza, evaluación y entrenamiento; el proyecto `indexador-py/` encapsula el enriquecimiento y la persistencia en PostGIS. Se controlaron semillas aleatorias y configuración de validación para permitir replicación de métricas.

UNIVERSIDAD
SERGIO ARBOLEDA

UNIVERSIDAD
SERGIO ARBOLEDA

Capítulo 7

Resultados y Análisis

1. 2024

1.1. Abril 2024 - Junio 2024

- **Abril 2024:**
  - Inicio del proyecto.
  - Revisión de la literatura y estudios previos sobre detección de patrones de pesca.
  - Definición de objetivos específicos y alcance del proyecto.
- **Mayo 2024:**
  - Identificación y selección de fuentes de datos AIS.
  - Diseño del plan de adquisición de datos.
  - Configuración del entorno de trabajo (instalación de software y herramientas necesarias).
- **Junio 2024:**
  - Inicio de la adquisición de datos AIS.
  - Realización de un análisis exploratorio inicial de los datos obtenidos.
  - Identificación de posibles problemas de calidad de los datos.

1.2. Julio 2024 - Septiembre 2024

- **Julio 2024:**
  - Limpieza y preprocesamiento de datos.
  - Desarrollo de métodos preliminares para la detección de patrones.
  - Diseño del estudio piloto.
- **Agosto 2024:**
  - Ejecución del estudio piloto con un subconjunto de datos.

UNIVERSIDAD SERGIO ARBOLEDA

- Evaluación y ajuste de los métodos basados en los resultados del estudio piloto.
- Documentación de los hallazgos preliminares.

- **Septiembre 2024:**
  - Refinamiento de los métodos y técnicas de análisis.
  - Planificación detallada del análisis completo del proyecto.
  - Revisión y ajuste del cronograma según los resultados obtenidos.

1.3. Octubre 2024 - Diciembre 2024

- **Octubre 2024:**
  - Inicio del análisis de patrones de pesca en datos a gran escala.
- **Noviembre 2024:**
  - Monitoreo y ajuste continuo de los métodos según los resultados obtenidos.
  - Desarrollo de visualizaciones y reportes preliminares.
- **Diciembre 2024:**
  - Consolidación de resultados preliminares.
  - Revisión y ajustes finales antes de la pausa de fin de año.

2. 2025

2.1. Enero 2025 - Marzo 2025

- **Enero 2025:**
  - Reanudación del análisis de datos.
  - Revisión de los objetivos y ajuste de la metodología si es necesario.
- **Febrero 2025:**
  - Integración de nuevos datos (si corresponde).
  - Continuación del análisis de patrones de pesca.
- **Marzo 2025:**
  - Evaluación de los resultados obtenidos hasta la fecha.
  - Preparación de un informe intermedio.

UNIVERSIDAD
SERGIO ARBOLEDA

2.2. Abril 2025 - Junio 2025

- **Abril 2025:**
  - Inicio de la fase de validación de los resultados.
  - Comparación de los resultados con estudios previos y datos reales.
- **Mayo 2025:**
  - Ajustes finales en los métodos y técnicas de análisis.
  - Preparación del informe final y de las visualizaciones de resultados.
- **Junio 2025:**
  - Revisión y finalización del informe del proyecto.
  - Presentación de los resultados a las partes interesadas.
  - Conclusión del proyecto y recomendaciones para futuros estudios.

UNIVERSIDAD
SERGIO ARBOLEDA

UNIVERSIDAD
SERGIO ARBOLEDA

Capítulo 8

Discusión

Se ha realizado una estimación de costos para el procesamiento de datos de Global Fishing Watch utilizando Amazon SageMaker, considerando un tamaño de dataset de 10 GB, el uso de instancias de tipo `ml.m5.large` para procesamiento y entrenamiento, y almacenamiento en Amazon S3. El procesamiento se estima en 8 horas diarias durante 20 días al mes, con 5 modelos entrenados al mes, cada uno tomando 10 horas. Los costos mensuales incluyen $18.56 para procesamiento, $5.80 para entrenamiento y $0.23 para almacenamiento, resultando en un costo total mensual de $24.59. Para un periodo de 15 meses (abril 2024 - junio 2025), el costo total estimado es de aproximadamente $368.85.

1. Tabla de Estimación de Costos

Cuadro 8.1: Estimación de costos para el procesamiento de datos de Global Fishing Watch usando Amazon SageMaker

Concepto	**Costo Mensual ($)**	**Total para 15 meses ($)**
Procesamiento (SageMaker)	18.56	278.40
Entrenamiento de Modelos	5.80	87.00
Almacenamiento (S3)	0.23	3.45
Costo Total	**24.59**	**368.85**

UNIVERSIDAD
SERGIO ARBOLEDA

UNIVERSIDAD
SERGIO ARBOLEDA

Capítulo 9

Conclusiones y Trabajo Futuro

Cuadro 9.1: Resultados o productos esperados de nuevo conocimiento.

Resultado Esperado	Indicador	Beneficiarios
Generación de Nuevo Conocimiento		
Innovación en Métodos de Análisis	Número de nuevos métodos/algoritmos desarrollados	Comunidad científica global
Comprensión de Patrones de Pesca	Artículo sometido a revista indexada por Scopus	Investigadores, agencias de pesca, organizaciones medioambientales
Publicaciones Científicas	Número de artículos publicados en revistas científicas y conferencias presentadas	Comunidad académica, estudiantes e investigadores
Fortalecimiento de la Capacidad Científica Nacional en el Contexto de Colombia		
Desarrollo de Talento Local	Número de estudiantes y profesionales capacitados	Estudiantes, investigadores, profesionales colombianos
Colaboración Interinstitucional	Número de colaboraciones y proyectos conjuntos	Universidades, centros de investigación, agencias gubernamentales
Proyectos Futuros	Número de proyectos de investigación derivados	Comunidad científica nacional, entidades gubernamentales
Apropiación Social del Conocimiento		
Divulgación de Resultados	Número de talleres	Seminarios y publicaciones en medios de comunicación,Comunidades costeras, público en general, medios de comunicación

La anterior tabla resume los resultados esperados del actual proyecto desde tres

UNIVERSIDAD SERGIO ARBOLEDA

perspectivas clave:

1. Generación de nuevo conocimiento
2. Fortalecimiento de la capacidad científica nacional
3. Apropiación social del conocimiento

Cada resultado esperado se acompaña de indicadores específicos para medir su éxito y los beneficiarios principales que se verán impactados positivamente por el proyecto.

UNIVERSIDAD
SERGIO ARBOLEDA

Anexos

1. Anexo técnico

1.1. Capas PostGIS y SRID

- **Capas**: `barrios_bogota`, `upz_bogota`, `localidades_bogota`, `estratos_manzana`, `avaluo_catastral_manzana`, `gis_osm_pois_free_1`, `gis_osm_pois_a_free_1`.
- **SRID**: 4326 (WGS84). Para consultas por distancia se usa proyección a 3857 cuando aplica (`ST_Transform`).
- **Consultas típicas**: `ST_DWithin`, `ST_Contains`, `ST_Intersects`, centroids y buffers en metros.

1.2. Sistemas de referencia: WGS84 (EPSG:4326) y Web Mercator (EPSG:3857)

- **WGS84 (EPSG:4326)**: sistema geodésico global usado por GPS. *Coordenadas en grados* (latitud/longitud). Ventaja: interoperabilidad y exactitud posicional. Limitación: los grados no son métricos; 1° de longitud equivale a distintas distancias según la latitud.
- **Web Mercator (EPSG:3857)**: proyección métrica popular para mapas web. *Coordenadas en metros* (pseudo-mercator). Ventaja: permite cálculos de distancia y `ST_DWithin` en *metros*. Limitación: distorsiona áreas y distancias al alejarse del ecuador (aceptable para escalas urbanas como Bogotá).
- **Cuándo usar cada uno**: almacenar y cruzar capas administrativas en 4326; proyectar a 3857 para consultas con radios en metros o buffers métricos (`ST_Transform(geom, 3857)`).
- **Implicaciones en PostGIS**: para `ST_DWithin` con radio en metros, asegure que ambas geometrías estén en 3857; para `ST_Contains`/`ST_Intersects` topológicos, 4326 es suficiente.

1.3. Reglas de limpieza

- **Outliers (p99)**: $\textit{área} \leq 464\,m^2$, $\textit{precio\_venta} \leq 5{,}4 \times 10^9$ COP.
- **Precio mínimo**: $\geq$ 50,000,000 COP.

UNIVERSIDAD
SERGIO ARBOLEDA

- **Área = 0**: mediana de comparables (*estrato*, *habitaciones*, *banos*, *sector*); si no hay, mediana por *estrato*.
- **Parqueaderos ¡0**: reemplazo por moda del mismo *estrato*; si no hay, moda global.
- **Coordenadas**: imputación por mediana del *sector* y filtro final a Bogotá: lat $[4{,}4, 4{,}9]$, lon $[-74{,}3, -73{,}9]$.
- **Estrato fuera [1–6]**: imputación por modo del *sector*; si no hay, modo global.

1.4. Enriquecimiento geoespacial

- **Conteos OSM por radio**: 100, 300, 500, 1000, 2000 m.
- **Categorías**: *education*, *healthcare*, *retail_access*, *dining_and_entertainment*, *accommodation*, *parks_and_recreation*, *infrastructure_services*, *cultural_amenities*.
- **Región calculada**: *upz_calculada*, *barrio_calculado*, *localidad_calculada* por `ST_Contains`.
- **Avalúos por geohash**: promedios de *catastral* y *comercial* en bbox del geohash sobre `avaluo_catastral_manzana`.
- **Persistencia**: tabla `property_data` adaptada al DF y agregados `region_stats` (barrio/UPZ/localidad) con n, medias, desviaciones y cuartiles.

1.5. Modelos, métricas e hiperparámetros

- **Validación**: KFold=5 (`shuffle=True`, `random_state=42`); métrica principal RMSE (escala real).
- **Base**: RF $\approx$ 245M (CV); hold-out (20 %): RMSE $\approx$ 250.6M, MAE $\approx$ 129.9M, R^2 $\approx$ 0.915.
- **Aumentados**: v1 (XGB reducido + *barrio_top*) RMSE $\approx$ 254.66M; v2 (XGB tuning) RMSE $\approx$ 233.49M.
- **Hiperparámetros v2**: $n\_estimators = 500$, $max\_depth = 9$, $learning\_rate = 0{,}05$, $subsample = 0{,}8$, $colsample\_bytree = 0{,}8$, $\alpha = 0$, $\lambda = 1$.

1.6. API y operación

- **Endpoint**: `GET /api/estimate` (FastAPI). Entrada: lat, lon y atributos de la propiedad; salida: estadísticos del punto/region y predicción del modelo.
- **Modelo por defecto**: 2.1 (recomendado 2.2 por menor RMSE).
- **Ambiente**: PostGIS con SRID 4326; consultas en metros usando `ST_Transform` a 3857.

UNIVERSIDAD SERGIO ARBOLEDA

1.7. Reproducibilidad

- Notebooks en `analisis/notebooks/` para descarga, limpieza, EDA y entrenamiento.
- Artefactos de modelos en `analisis/data/models/` (`randomforest_model_base.pkl`, `xgboost_model_2.1.pkl`, `xgboost_model_2.2.pkl`).
- Enriquecimiento y persistencia en `indexador-py/` (ETL asíncrono, inserción por lotes, creación dinámica de tabla).
- Semillas y configuración de validación fijadas para replicar métricas.

UNIVERSIDAD
SERGIO ARBOLEDA