Making Money with ML? *“It’s all about location location location!!!”*

Daniel Felipe Cortes Cendales

Jorge Esteban Gómez Carmona

Nicolas Alberto Gonzales Gort

Sebastian Marín Lombo

30 de Octubre del 2023

**Palabras Clave:**

**Clasificación JEL:**

**Link del repositorio:** <https://github.com/BigData-MachineLearning/P-Set2.git>

# Introducción

Debido al alto grado de fluctuación en el mercado inmobiliario, los inversores, propietarios de viviendas, tasadores, asesores fiscales y otros participantes del mercado inmobiliario, como prestamistas hipotecarios y aseguradoras, especulan constantemente con los precios de la vivienda. No es de extrañar que exista una gran cantidad de investigaciones sobre el tema. Por ejemplo, los investigadores en el campo de la construcción combinaron métodos de aprendizaje automático con el conocimiento profesional de la industria de la construcción. En la industria de la construcción se utilizan muchos sistemas inteligentes y muchos de ellos han logrado buenos beneficios económicos y sociales.

La predicción tradicional del precio de la vivienda se basa en comparaciones de costos y precios de venta. Es difícil hacer pronósticos ya que los factores que afectan el mercado inmobiliario varían desde los factores socioeconómicos (por ejemplo, tasa de criminalidad per cápita, acceso al transporte, ingreso promedio y nivel educativo) hasta características específicas de la casa (por ejemplo, pies cuadrados, número de dormitorios, estilo de construcción, fecha de construcción). última remodelación). Por lo tanto, el desarrollo y la disponibilidad de varios modelos de predicción del precio de la vivienda pueden desempeñar un papel útil para llenar un vacío de información que puede mejorar la eficiencia del mercado inmobiliario.

Un informe reciente del Zillow Group, un popular sitio web de bases de datos sobre viviendas, indica que los vendedores y compradores de casas recurren cada vez más a la investigación en línea para estimar el precio de la vivienda antes de contactar a los agentes inmobiliarios. Investigar por tu cuenta cuánto vale la casa que te interesa puede resultar complicado por múltiples motivos. Como se mencionó, hay muchos factores que influyen en el precio potencial de una casa, lo que hace que sea más complicado para un individuo decidir cuánto vale una casa por sí solo sin ayuda externa, lo que puede llevar a que las personas tomen decisiones mal informadas sobre si comprar o vender sus casas y sobre qué precios son razonables.

[Resumen de nuestros resultados]

# Revisión de Literatura

## Uso de patrones de datos para predecir

La búsqueda de patrones en los datos se ha estudiado durante mucho tiempo en muchos campos, incluida la estadística, el reconocimiento de patrones y el análisis exploratorio de datos. El análisis de datos puede proporcionar más conocimiento sobre una empresa al ir más allá de los datos almacenados explícitamente para obtener conocimiento. sobre el negocio. Aquí es donde el *Machine Learning* tiene beneficios obvios para cualquier empresa.

Grudinitski, Shilling & Quang Do (1995), aplicando un análisis de redes neuronales , proporcionan evidencia que resalta la importancia de emplear datos predictivos para entrar a los mercados hipotecarios, allí encuentran, por ejemplo, que las características del patrimonio neto, el estado civil y el nivel educativo de un prestatario o si un co-prestatario está involucrado contribuyen de manera significativa a la capacidad de la red neuronal para determinar la elección de una hipoteca. Por su parte, Zhang & Zhou (2004), emplean estrategias de predicción en el contexto de a aplicación financiera desde la perspectiva tanto técnica como de aplicación. Además, comparan diferentes técnicas de clasificación y estimación para sus propósitos. De igual manera, discuten cuestiones importantes de minería de datos involucradas en aplicaciones financieras específicas. Finalmente, destacan una serie de desafíos y tendencias para futuras investigaciones en esta área.

Kaiha, Copeland et al. (2006) Presentan información relacionada con el estado actual de los bienes raíces en los EE. UU. Muestran que, la caída de la vivienda a nivel nacional fue más evidente la caída del precio de venta medio de la viviendas existentes, estimando a través de algoritmos de clasificación caídas en los precioso de alrededor de $225.000 US. Adicional a ello, muestran evidencia a través de predicciones que habrá una crisis inmobiliaria en la cual los activos residenciales se reducirán ampliamente. Finalmente, Kntrimas & Verikas (2011) empleando regresiones lineales de Mínimos Cuadrados Ordinarios, construyen modelos predictivos para evaluar el desempeño que tendrán los impuestos sobre el valor de los inmuebles en EE. UU. Su método se compara con enfoques de inteligencia computacional: regresión con máquina de vectores de soporte (SVM), perceptrón multicapa (MLP) y un comité de predictores. Obteniendo resultados útiles encaminados a una mayor recaudación tributaria sobre el valor de los bienes.

## Predicción de precios de las viviendas.

Parte importante de este artículo es predecir el precio de las viviendas en Bogotá, para lo cual se sabe que empleando métodos asociados al uso de bases de datos con información de las viviendas. Por ejemplo, Bahía (2013), proporciona pronósticos independientes del mercado inmobiliario sobre los precios de las viviendas mediante el uso de técnicas de minería de datos. La idea principal era construir el modelo de red neuronal utilizando dos tipos de redes neuronales. Primero, la red neuronal de avance (FFBP) y, en segundo lugar, la red neuronal de avance en cascada (CFBP). Luego compararon los dos modelos para encontrar la predicción de mejor rendimiento de los precios de la vivienda. Los autores estimaron el valor medio de las viviendas ocupadas por sus propietarios en los suburbios de Boston, teniendo en cuenta 13 atributos del vecindario en una muestra de 506 puntos de datos.

Otro caso es el de Mu et al (2014), quienes analizaron un conjunto de datos que contiene los valores de las casas en los suburbios de Boston y utilizando varios métodos de aprendizaje automático hicieron pronósticos asociados a sus precios. Con base en estas predicciones, los autores buscaban que las agencias gubernamentales y los promotores inmobiliarios pudieran tomar mejores decisiones sobre si iniciar o no desarrollos inmobiliarios en las regiones correspondientes. Los valores de las viviendas se pronosticaron utilizando métodos como: máquina de vectores de soporte (SVM), máquina de vectores de soporte de mínimos cuadrados (LSSVM) y métodos de mínimos cuadrados parciales (PLS). Estos algoritmos se comparan con los resultados previstos. Finalmente, utilizando múltiples características, pronosticaron los valores de las viviendas en Boston y se descubrió que los resultados de predicción de los diversos enfoques de aprendizaje automático variaban. Aunque existe una falta de linealidad grave en los datos, los resultados del experimento también mostraron que los métodos SVM y LSSVM fueron superiores al método PLS para abordar el problema de la no linealidad.

Y en última instancia, Hromada (2015) describe un software innovador que se puede utilizar para evaluaciones inmobiliarias y análisis de anuncios inmobiliarios publicados en Internet en la República Checa. El software recopila, analiza y evalúa sistemáticamente datos sobre los cambios en el mercado inmobiliario. Cada semestre, el software recopila más de 650.000 cotizaciones de precios sobre la venta o el alquiler de apartamentos, casas, propiedades comerciales y solares. Todos los anuncios inmobiliarios se almacenan continuamente en una base de datos de software y se analizan minuciosamente para determinar su credibilidad.

# Datos

## Datos brutos obtenidos

La información bruta se obtuvo de la página para finca raíz “*properati*”[[1]](#footnote-1) en posesión de la compañía, que contiene una muestra de datos individuales en Bogotá, que se muestra en la Tabla 1. *“Properati”* se presenta como una fuente valiosa para abordar el problema planteado, que consiste en predecir los precios de las viviendas en Bogotá.

La principal ventaja de esta base de datos radica en su amplitud, ya que contiene precios recientes de propiedades en toda la ciudad, lo que proporciona una visión completa del mercado inmobiliario. Además, esta base de datos incluye descripciones detalladas de las propiedades, lo que permite extraer información relevante para la solución del problema.

Sin embargo, es importante señalar que la base de datos presenta desafíos significativos. Pues se observan muchos valores faltantes, como se evidencia en la Tabla 2, lo que introduce incertidumbre en el análisis y modelado de datos; asimismo, se han identificado algunos errores de digitación o extracción durante el proceso de web scraping, lo que ha resultado en valores astronómicamente altos en algunas propiedades. Estos desafíos deben abordarse cuidadosamente en el proceso de limpieza y preparación de datos para garantizar que los modelos de predicción se desarrollen con información precisa y confiable.

Así, dentro la base de datos original se han identificado valores faltantes en varias variables que desempeñan un papel fundamental en la predicción de precios de viviendas. Específicamente, las variables "surface\_total" y "surface\_covered" muestran un alto número de valores faltantes, con 30,790 y 30,079 registros incompletos respectivamente. La superficie de una propiedad es un factor crucial en la determinación de su precio, por lo que la presencia de estos valores faltantes representa un desafío significativo. De igual manera, la variable "rooms" exhibe 18,260 valores faltantes, mientras que "bathrooms" tiene 10,071 valores faltantes, ambas variables, que describen las características de las viviendas, son relevantes para la predicción de precios. Adicionalmente, la variable "title" presenta 22 valores faltantes, y la variable "description" tiene 9 valores faltantes, ambas variables se utilizan para describir las propiedades y pueden contener información valiosa para el análisis.

Con esto en mente, la construcción de la base de datos implicó la incorporación de variables tanto a partir del texto descriptivo como de fuentes externas de datos espaciales.

En cuanto a las variables de texto, se incluyeron las siguientes:

1. **Dummy Parqueadero:** Esta variable se creó para medir si en la descripción de la propiedad se mencionaba la disponibilidad de un parqueadero. Tomaba el valor de 1 si se mencionaba un parqueadero y 0 en caso contrario.
2. **En qué piso es el apto:** A través del análisis del texto descriptivo, se extrajo el número de piso en el que se encuentra el apartamento, lo que puede ser un factor relevante en la valoración de la propiedad.
3. **Dummy Penthouse:** Esta variable se utilizó para identificar si la vivienda era un penthouse, lo que generalmente implica un estatus premium. Tomaba el valor de 1 si se mencionaba que la propiedad era un penthouse y 0 en caso contrario.

En cuanto a las variables externas, se incorporaron datos provenientes de fuentes abiertas:

1. **Ciclovías:** La distancia a la ciclovía más cercana se obtuvo a partir de los datos disponibles en el portal de datos abiertos de Bogotá, lo que puede ser relevante para aquellos interesados en la accesibilidad en bicicleta.
2. **Transmilenio:** La distancia a la estación de Transmilenio más cercana se extrajo de OpenStreetMap, lo que proporciona información crucial sobre la accesibilidad al transporte público en la ciudad.
3. **UPL (Unidad Administrativa):** Se incluyó la unidad administrativa a la que pertenece la propiedad, utilizando datos disponibles en el portal de datos abiertos de Bogotá.
4. **Distancia al Centro Comercial más cercano:** La distancia a los centros comerciales más cercanos se obtuvo de OpenStreetMap, lo que puede ser relevante para aquellos que valoran la proximidad a centros comerciales.
5. **Estrato:** El estrato de la vivienda, información también obtenida del portal de datos abiertos de Bogotá, desempeña un papel importante en la valoración del precio por metro cuadrado de la propiedad.
6. **Distancia al Parque más cercano:** La distancia al parque más cercano se extrajo de OpenStreetMap, lo que proporciona información sobre la accesibilidad a áreas verdes.
7. **Distancia al CAI más cercano:** La distancia al CAI (Comando de Atención Inmediata), que puede ser indicativa de la seguridad en la zona, se obtuvo de OpenStreetMap.

## Proceso de imputación de datos

La corrección de los valores faltantes en la base de datos se llevó a cabo de manera exhaustiva. Inicialmente, se intentó imputar los valores faltantes a partir del texto descriptivo utilizando expresiones regulares para extraer información sobre la superficie de la propiedad, el número de alcobas y baños. Sin embargo, persistieron numerosos valores faltantes.

Para abordar los valores faltantes en la variable "superficie" (mts2), se implementó un enfoque en el conjunto de entrenamiento (train). Se calculó el precio por metro cuadrado de la propiedad y se buscó el valor más cercano sin faltantes. Luego, se utilizó este valor para estimar la superficie de las propiedades con valores faltantes.

En el caso de las variables de baños y habitaciones, se realizaron divisiones de las propiedades en bandas de metros cuadrados: "<100", "100-200", "200-300", "300+" y, a partir de estas categorías, se imputaron los valores faltantes utilizando la moda en cada banda.

La variable de "estrato" también presentaba un alto número de valores faltantes. Se imputaron estos valores utilizando la propiedad más cercana a 700 metros en el conjunto de entrenamiento, mientras que en el conjunto de prueba (test) se imputaron los valores faltantes con "estrato 5", ya que es común en propiedades en la zona de Chapinero.

Finalmente, se realizó una limpieza de la base de datos de entrenamiento (train) para eliminar observaciones atípicas. Se excluyeron propiedades con precios por metro cuadrado superiores a 12 millones y propiedades con más de 500 metros cuadrados. Como resultado, el conjunto de datos de entrenamiento pasó de 38,644 observaciones a 36,629, lo que contribuye a la calidad y confiabilidad de los análisis y modelos subsiguientes.

## Análisis de la variables finales.

De acuerdo con la Tabla 3, datos facilitados ofrecen información valiosa sobre seis variables numéricas con 36.621 observaciones cada una. La variable "Precio" muestra una amplia gama de precios inmobiliarios, con una media de 648.639.006,0 y una importante desviación típica de 307.081.200,0, que oscila entre 300.000.000 y 1.650.000.000. Esta información es esencial para comprender los valores inmobiliarios de la zona.

Las variables de distancia ("Parada de autobús", "Parque", "Centro comercial" y "CAI") proporcionan información sobre la accesibilidad a los servicios. Por ejemplo, la "Distancia al parque" tiene una media de 160,4 y una desviación típica moderada de 99,0, lo que indica que los parques son generalmente accesibles, con cierta variabilidad en su proximidad. Estas estadísticas ayudan a urbanistas, residentes y responsables políticos a comprender la dinámica espacial.

La variable "Superficie total", con una media de 136,7 y una desviación típica de 79,7, muestra la distribución del tamaño de las propiedades. Este dato es crucial para que compradores y promotores evalúen el tamaño de los inmuebles disponibles. En resumen, estas estadísticas ofrecen una visión global del precio, la accesibilidad y el tamaño de los inmuebles, lo que ayuda a tomar decisiones informadas en el mercado inmobiliario.

Igualmente, las gráficas y mapas presentados en la sección 7 “Anexos” ofrecen una visión detallada de las relaciones entre diversas variables y los precios de las propiedades en Bogotá, brindando valiosas perspectivas para los analistas y posibles compradores de viviendas.

# Modelo y resultados

Kskss

# Conclusiones y Recomendaciones

Jsk

# Bibliografía

Bahia, I. (2013). A Data Mining Model by Using ANN for Predicting Real Estate Market: Comparative Study. *International Journal of Intelligence Science*, 162-169.

Grudnitski, G., Shilling, J., & Quang Do, A. (1995). A Neural Network Analysis of Mortgage Choice. *Intelligent Systems in Accouting, Finance and Management, 4*(2).

Hari Hara Kumar, G., & et al. . (2023). Estimating the Price of House Using Machine Learning. *Journal of Engineering Sciences, 14*(7).

Hromada, E. (2015). Mapping of Real Estate Prices using Data Mining Techniques. *Creative Construction Conference*, 233-240.

Jaen, R. (2002). Data Mining: An Empirical Application in Real Estate Valuation. *The Florida AI Research Society*.

Kaihla, P., Copeland, M., & et al. (2006). The New Rules of Real Estate. *Business 2.0 The New Rules of Real Estate Survival Tips for a Sluggish Market how to Buy how to Sell the 10 Best Places to Invest, 7*(10).

Kontrimas, V., & Verikas, A. (2011). The Mass Appraisal of the Real Estate by Computational Intelligence. *Applied Soft Computing Journal, 11*(1), 443-448.

Mu, J., Wu, F., & Zhang, A. (2014). Housing Value Forecasting Based on Machine Learning Methods. *Abstract and Applied Analysis*. doi:10.1155/2014/648047

Muralidharan, S., Phiri, K., & Sinha, S. (2018). Analysis and Predicton of Real Estate Prices: A Casa of the Boston Housing Market. *Issues in Information Systems, 19*(2), 109-118.

Zhang, D., & Zhou, L. (2004). Discovering golden nuggets: data mining in financial application. *IEEE Transactions on Systems, Man and Cybernetics, 34*(4), 513-522.

# Anexos

|  |  |
| --- | --- |
| **Tabla 1. Variables Originales de la base** | |
| Variable | Descripción |
| property\_id | ID de propiedad |
| city | Ciudad de la propiedad |
| price | Precio de la propiedad |
| month | Mes del anuncio |
| year | Año del anuncio |
| surface\_total | Superficie de la propiedad |
| surface\_covered | Superficie de la propiedad |
| rooms | Número de Habitaciones |
| bedrooms | Número de Alcobas |
| bathrooms | Número de Baños |
| property\_type | Si es casa o apartamento |
| operation\_type | Tipo de operación (Venta en este caso) |
| lat | Latitud |
| lon | Longitud |
| title | Título de la propiedad |
| description | Párrafo descriptivo de la propiedad |

|  |  |
| --- | --- |
| **Tabla 2. Análisis de missing values** | |
| Variables | Missing |
| property\_id | 0 |
| city | 0 |
| price | 0 |
| month | 0 |
| year | 0 |
| surface\_total | 30,79 |
| surface\_covered | 30,079 |
| rooms | 18,26 |
| bedrooms | 0 |
| bathrooms | 10,071 |
| property\_type | 0 |
| operation\_type | 0 |
| lat | 0 |
| lon | 0 |
| title | 22 |
| description | 9 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tabla 3. Estadísticas descriptivas de las variables numéricas.** | | | | | |
| Statistic | N | Mean | St. Dev. | Min | Max |
| price | 36,621 | 648,639,006.0 | 307,081,200.0 | 300,000,000 | 1,650,000,000 |
| distancia\_bus | 36,621 | 948.7 | 681.1 | 3.6 | 5,856.8 |
| distancia\_parque | 36,621 | 160.4 | 99.0 | 1.0 | 1,102.8 |
| distancia\_cc | 36,621 | 658.0 | 382.2 | 0.6 | 4,004.8 |
| distancia\_police | 36,621 | 1,012.4 | 507.6 | 2.4 | 2,236.9 |
| surface\_total | 36,621 | 136.7 | 79.7 | 28.0 | 500.0 |

1. Obtenida de la página <https://www.properati.com.co/> [↑](#footnote-ref-1)