Informe Proyecto Final

Daniel Andrés Agudelo García, Paulina García Aristizábal, Emanuel Munera Pérez
Facultad de Ingeniería, Universidad de Antioquia
Medellín, Colombia
Materia: Modelos II
Grupo: 12
daniel.agudelo9@udea.edu.co, paulina.garcia1@udea.edu.co, emanuel.munera@udea.edu.co

I. Introducción

La estimación del valor de arriendo de una vivienda es un desafío relevante en contextos urbanos donde los precios fluctúan según factores estructurales, geográficos y socioeconómicos. Variables como la ubicación, el tamaño, el número de habitaciones o el nivel de amueblamiento influyen de manera conjunta en el precio final, haciendo que los métodos tradicionales de valoración resulten limitados.

En este escenario, las técnicas de *Machine Learning* ofrecen una alternativa eficiente para modelar relaciones y descubrir patrones en grandes volúmenes de datos. El presente proyecto aplica este enfoque al *House Rent Prediction Dataset* —disponible en Kaggle— que recopila información de propiedades en distintas ciudades de India con el fin de predecir el valor mensual de arriendo (`Rent`).

Mediante un enfoque de aprendizaje supervisado, el modelo busca aprender la relación entre las características del inmueble y su valor de renta, permitiendo estimaciones más precisas y objetivas. Este tipo de herramienta puede servir de apoyo a propietarios, arrendadores y plataformas inmobiliarias, contribuyendo a una mayor transparencia y eficiencia en el mercado del alquiler.

II. Descripción del problema

El acceso a la vivienda es uno de los factores más determinantes en la calidad de vida urbana. En ciudades con una alta densidad poblacional y dinámicas inmobiliarias cambiantes, predecir el valor de arriendo de una vivienda puede ser una tarea compleja, influenciada por variables como la ubicación, el tamaño, las comodidades del inmueble, la infraestructura disponible en la zona y las condiciones del mercado en cada momento.

La estimación precisa del precio de arriendo constituye un desafío tanto para arrendadores como para arrendatarios. Por un lado, los propietarios y agentes inmobiliarios buscan fijar precios competitivos que maximicen la ocupación sin sacrificar rentabilidad; por otro, los arrendatarios requieren información confiable que les permita tomar decisiones informadas y evitar sobrecostos injustificados. Sin embargo, en la práctica, los métodos tradicionales de valoración suelen basarse en la experiencia subjetiva o en comparaciones limitadas a propiedades cercanas, lo cual introduce sesgos y dificulta capturar la complejidad real del mercado.

Además, los precios de alquiler están determinados por múltiples factores. Características estructurales de la vivienda, como el número de habitaciones o el área construida, se combinan con factores contextuales como la ciudad, el vecindario o el nivel de amueblamiento. Esta combinación de variables hace que el problema sea ideal para ser abordado mediante técnicas de Machine Learning, las cuales son capaces de modelar relaciones complejas y capturar patrones ocultos en los datos.

En este contexto, una solución basada en aprendizaje automático permite automatizar el proceso de estimación del precio de arriendo, ofreciendo una herramienta útil para agencias inmobiliarias, arrendadores y potenciales inquilinos. Este tipo de sistema podría integrarse fácilmente en plataformas de búsqueda de vivienda, recomendando precios sugeridos en tiempo real o alertando sobre valores atípicos que podrían indicar sobrevaloración o subvaloración del inmueble.

II-A. Aproximación desde Machine Learning

Dado que el objetivo del proyecto es estimar un valor numérico continuo del cual se tienen datos etiquetados —es decir, para cada propiedad se conoce el valor real del arriendo (`Rent`)—, se adoptará un enfoque de aprendizaje supervisado. En este paradigma, el modelo aprende una función de mapeo entre un conjunto de variables de entrada (características del inmueble y su contexto) y una variable objetivo conocida, con el propósito de minimizar el error entre las predicciones y los valores reales.

El uso de aprendizaje supervisado resulta adecuado no solo porque se dispone de datos históricos confiables, sino también porque posibilita la comparación entre distintos algoritmos de regresión, tales como la Regresión Lineal, los Árboles de Decisión, el *Random Forest* o las Redes Neuronales Multicapa (MLP). Esto permitirá analizar el equilibrio entre interpretabilidad y capacidad predictiva, identificando cuál modelo logra capturar mejor las relaciones no lineales entre las variables y el precio de arriendo.

En síntesis, este enfoque proporciona una base sólida para desarrollar un sistema de predicción de rentas interpretable, eficiente y adaptable, que aproveche el potencial del aprendizaje automático para ofrecer estimaciones más precisas y fundamentadas.

II-B. Descripción de la base de datos

El conjunto de datos empleado, *House Rent Prediction Dataset*, contiene información de diferentes propiedades en India, recopiladas a partir de listados reales de arriendo.
Número de muestras: 4,746 registros.
Número de variables: 12 características, incluyendo la variable objetivo.
Variable objetivo: `Rent` — valor del arriendo mensual en rupias.

Las principales variables incluidas en el dataset son las siguientes:

Variable	Descripción
`BHK (Num)`	Número de habitaciones.
`Rent(Num)`	Valor mensual del arriendo (variable objetivo).
`Size(Num)`	Área total de la propiedad (en pies cuadrados).
`Floor (Cat)`	Piso o nivel en el que se encuentra la vivienda.
`Area Type (Cat)`	Tipo de medición del área (Super built-up, Built-up, Carpet).
`Area Locality (Cat)`	Nombre del vecindario o zona
`City (Cat)`	Ciudad donde se ubica la propiedad.
`Furnishing Status (Cat)`	Nivel de amueblamiento (Unfurnished, Semi-Furnished, Furnished).
`Tenant Preferred (Cat)`	Tipo de arrendatario preferido (Family, Bachelors, Company).
`Bathroom (Num)`	Número de baños.
`Point of Contact (Cat)`	Medio de contacto del anunciante.

Durante la inspección inicial de los datos no se identificaron valores faltantes en las variables; sin embargo una vez se hizo una codificación de la variable `Floor`, surgieron 4 datos faltantes a los cuales se les asignó el valor de la columna nueva `Current Floor`. Para la limpieza de datos fue necesario estandarizar los formatos de texto y convertir algunas variables categóricas en representaciones numéricas mediante codificación *Label Encoding*.

III. Estado del arte

Artículo I: Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times [1]

El estudio de Mora-García *et al.* (2022), publicado en la revista *Land*, analiza la predicción del precio de vivienda durante el periodo de pandemia en la provincia de Alicante (España), aplicando algoritmos de machine learning a un extenso conjunto de datos inmobiliarios.

- **Configuración del problema:** Se plantea una tarea de regresión supervisada, donde la variable objetivo es el precio de venta de vivienda. Se emplearon más de 47 000 registros georreferenciados obtenidos del portal Idealista, combinando periodos previos y posteriores al confinamiento por COVID-19.
- **Técnicas y modelos utilizados:** Se evaluaron Regresión Lineal Múltiple, Random Forest, Extra Trees, Gradient Boosting, XGBoost y LightGBM, comparando el desempeño entre modelos lineales y métodos de ensamble basados en *bagging* y *boosting*.
- **Metodología de validación:** Se aplicó una división entre conjuntos de entrenamiento, validación y prueba bajo un esquema *pooled cross-sectional*, controlando efectos temporales y espaciales y analizando el sobreajuste.
- **Métricas empleadas:** MAE, MSE, RMSE y R^2. La **RMSE**, no vista en clase, representa la raíz cuadrada del MSE y mide el error medio en las mismas unidades de la variable objetivo:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

- **Resultados y conclusiones:** Los modelos de boosting (XGBoost, LightGBM, GBR) obtuvieron el mejor desempeño, con menores errores respecto a los modelos lineales y de bagging. La incorporación de variables espaciales y socioeconómicas derivadas —como NDVI, distancias a zonas de interés o renta media— mejoró significativamente la precisión. Los autores concluyen que los métodos de boosting logran un balance adecuado entre precisión y complejidad, destacando la importancia de la calidad del dataset y de las variables.

Referencia: Mora-García, R. T., et al. (2022). *Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times*. Land, 11(10), 2100. Disponible en: https://www.mdpi.com/2073-445X/11/10/2100

Artículo II: Machine Learning Techniques for Predicting Home Rental Prices in India (Jayadharshini et al., 2023)

El estudio de Jayadharshini *et al.* (2023), publicado en la revista *Applied and Computational Engineering*, se enfoca en la predicción de precios de arriendo en India, empleando un conjunto de datos similar al utilizado en este proyecto.

- **Configuración del problema:** Regresión supervisada para estimar el valor de arriendo en ciudades indias como Bangalore y Chennai, considerando variables estructurales (BHK, tamaño, baños) y contextuales (ciudad, tipo de área, amueblamiento, tipo de inquilino).
- **Técnicas y modelos utilizados:** Se emplearon Regresión Lineal, Random Forest, Gradient Boosting y XGBoost, destacando la superioridad de los métodos de *boosting* al manejar interacciones no lineales entre variables.
- **Metodología de validación:** Se usó validación cruzada *k-fold* y conjuntos de prueba independientes. Se aplicaron

técnicas de re-muestreo para equilibrar localidades poco representadas.

- **Métricas empleadas:** MAE, RMSE, R^2 y MAPE (*Mean Absolute Percentage Error*), definida como:

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

- **Resultados y conclusiones:** Los modelos XGBoost y Gradient Boosting obtuvieron los menores errores, mostrando mayor capacidad predictiva. Los autores concluyen que la correcta codificación de variables categóricas y el tratamiento de valores atípicos son esenciales para mejorar el desempeño del modelo.

Referencia: Jayadharshini, P., Santhiya, S., et al. (2023). *Machine Learning Techniques for Predicting Home Rental Prices in India*. Applied and Computational Engineering. Disponible en: https://www.sciencedirect.com/

IV. Entrenamiento y evaluación de los modelos

V. Reducción de dimensión

VI. Evaluación