Introducción

El mercado inmobiliario es un sector donde la estimación del valor de una propiedad es fundamental para la toma de decisiones de compra, venta e inversión.
En este proyecto se utiliza el **Ames Housing Dataset**, un conjunto de datos muy popular en competiciones de Machine Learning (Kaggle), con el objetivo de **predecir el precio de venta de viviendas** a partir de múltiples características estructurales y de ubicación.

El propósito principal es **comparar diferentes algoritmos de regresión** para encontrar el modelo que ofrezca un mejor ajuste y precisión en la predicción de precios.

Dataset

- **Fuente:** Ames Housing Dataset – Kaggle
- **Tamaño:** más de 1.400 observaciones y 80 variables.
- **Características principales:**
  - ✓ Variables numéricas: superficie construida, año de construcción, número de baños, área del sótano, etc.
  - ✓ Variables categóricas: tipo de zona, calidad de materiales, estilo de la vivienda, etc.

Tratamiento de datos

- **Nulos:**
  - ✓ Columnas con más del 45% de nulos fueron eliminadas.
  - ✓ En variables numéricas, los nulos se imputaron con la mediana.
  - ✓ En variables categóricas, los nulos se imputaron con la moda.
- **Outliers:**
  - ✓ Identificados en SalePrice y otras variables mediante boxplots.
  - ✓ Se aplicó winsorización para reducir el efecto de valores extremos.
- **Codificación:**
  - ✓ Variables categóricas convertidas a numéricas mediante **One-Hot Encoding**.
- **Escalado:**
  - ✓ Se aplicó escalado estándar en modelos lineales para mejorar la estabilidad.

Andres Vallejo
2025

Metodología

El flujo del proyecto se desarrolló en varias etapas:

1. **Análisis Exploratorio (EDA):**
    - Histogramas y distribuciones de precios.
    - Correlaciones entre variables numéricas y el precio de venta.
    - Visualización de variables categóricas frente al precio.

2. **Preprocesamiento:**
    - Manejo de nulos y outliers.
    - Codificación de variables categóricas.
    - Escalado en modelos lineales.

3. **Modelado:**
   Se probaron múltiples algoritmos de regresión:
    - RandomForestRegressor
    - CatBoostRegressor
    - XGBRegressor
    - LGBMRegressor
    - MLPRegressor
    - ElasticNet
    - HuberRegressor

4. **Evaluación:**
    - Métricas utilizadas: **MSE, RMSE, MAE y R^2**.
    - Comparación de desempeño entre modelos.

5. **Conclusión de resultados:**
    - **CatBoostRegressor y LightGBM** fueron los mejores modelos ($R^2 \approx 0.80$).
    - Los modelos lineales tuvieron un desempeño inferior.
    - El **MLPRegressor** presentó el peor rendimiento en este dataset.

Conclusiones

- El proyecto permitió explorar un caso práctico de predicción de precios de vivienda, muy cercano a aplicaciones reales en el sector inmobiliario.
- Los modelos basados en **árboles de decisión** (CatBoost, LightGBM) fueron los más efectivos.
- Los modelos lineales mostraron limitaciones debido a la complejidad de las relaciones no lineales en el dataset.

Andres Vallejo
2025