Proyecto del curso
Modelos y Simulación II

Sebastián Amaya Pérez, Jhon Alejandro García Pareja y David Felipe Tovar Zurita

I. Introducción

El aprendizaje supervisado permite encontrar relaciones entre las variables de un conjunto de datos para hacer predicciones a partir de ejemplos ya conocidos. Según lo visto en el curso *Introduction to Machine Learning*, cuando la salida que se quiere estimar es un valor numérico continuo, la forma adecuada de abordar el problema es mediante técnicas de regresión. Entre ellas, la regresión lineal se convierte en el primer punto de referencia gracias a que es fácil de interpretar y sirve como modelo base para comparar métodos más avanzados más adelante.

En este proyecto, el objetivo es predecir el precio de venta de viviendas utilizando el dataset *House Prices: Advanced Regression Techniques* de Kaggle. Para lograrlo, se realizará un análisis exploratorio de los datos, su limpieza, la imputación de valores faltantes y la codificación de variables categóricas, para posteriormente entrenar un modelo de regresión lineal. Una vez entrenado, su rendimiento será evaluado con métricas comunes en problemas de regresión, con el fin de medir qué tan bien logra ajustar la información disponible.

Finalmente, los resultados obtenidos permitirán identificar qué tanto puede aportar un modelo lineal en este tipo de problema y si es necesario acudir a modelos más complejos en posibles futuras etapas del trabajo.

A. Contexto del Problema

En el mercado inmobiliario, estimar el precio real de una vivienda es un reto debido a la cantidad de factores que influyen en su valor, tales como la ubicación, el estado de la construcción, el área total, las mejoras internas y las condiciones económicas del entorno. A pesar de contar con información histórica, en la práctica muchos cálculos siguen basándose en métodos subjetivos o modelos tradicionales que no logran capturar relaciones complejas entre las variables, lo que genera incertidumbre y decisiones poco acertadas en procesos de compra, venta o inversión.

El uso de técnicas de *Machine Learning* ofrece una alternativa más sólida y eficiente para este tipo de problemas. Estos métodos permiten analizar grandes volúmenes de datos, identificar patrones no lineales y generar modelos predictivos con mayor precisión que los enfoques estadísticos convencionales. Desarrollar una solución basada en *Machine Learning* para la predicción del precio de viviendas no solo mejora la exactitud de las estimaciones, sino que también facilita la toma de decisiones informadas, reduce el margen de error y aporta mayor transparencia al mercado inmobiliario.

Fig. 1. *

B. Composición de la Base de Datos

La base de datos utilizada se obtuvo de la plataforma Kaggle en el desafío *House Prices: Advanced Regression Techniques*. Esta base de datos contiene los siguientes cuatro archivos:

- **train.csv:** 1 460 muestras con 81 columnas, incluyendo la variable objetivo `SalePrice`.
- **test.csv:** aproximadamente 1 459 muestras con 80 columnas (no incluye la variable objetivo).
- **data_description.txt:** documento que detalla el significado de cada variable, sus valores posibles y notas adicionales.
- **sample_submission.csv:** archivo ejemplo que muestra el formato requerido para la presentación de predicciones en el concurso.

Las variables presentes en la base de datos describen distintos aspectos de cada vivienda. Estas pueden agruparse conceptualmente en las siguientes categorías:

- Características físicas del lote: área, forma, frente del terreno, pendiente, entre otros.
- Características estructurales de la vivienda: número de pisos, año de construcción, área habitable, materiales y estado exterior.
- Características internas: número de habitaciones, baños, sótano, chimenea, cocina, acabados y su nivel de calidad.
- Características adicionales: garaje, piscina, cercas, porches y otras mejoras.
- Información del vecindario: ubicación y clasificación residencial del sector.

1) Existencia de Datos Faltantes: Durante la revisión del conjunto de entrenamiento se identificó la presencia de valores representados como `NA`. Estos casos se dividen en dos situaciones:

1) **Valores `NA` que significan "no aplica":** Aparecen principalmente en variables categóricas asociadas a características que no existen en la vivienda. Por ejemplo, si una propiedad no tiene chimenea, la columna que describe su calidad aparece con `NA`, indicando ausencia de la característica y no pérdida de información.
2) **Valores `NA` que representan datos faltantes reales:** Se presentan en columnas numéricas donde el valor sí debería existir. Un ejemplo es la variable `LotFrontage`,

donde el frente del lote no se encuentra registrado para algunas viviendas.

Con el fin de preparar los datos para la construcción del modelo, se adoptará la siguiente estrategia:

- Para variables numéricas con datos faltantes reales: imputación mediante la mediana, con el fin de evitar distorsiones ocasionadas por valores atípicos.
- Para variables categóricas donde `NA` indica ausencia de característica: sustitución por la categoría `"None"`, manteniendo así la coherencia semántica de la información.
- Eliminación de variables con alta proporción de valores faltantes: en los casos donde una variable presente un porcentaje elevado de `NA` y aporte poca información al modelo, se considerará su eliminación.

2) Codificación de Variables: Debido a la presencia de variables de distintos tipos, se emplearán diferentes estrategias de codificación:

- Variables numéricas: se utilizarán sin codificación adicional.
- Variables categóricas nominales: se empleará *One-Hot Encoding* para evitar la introducción de relaciones inexistentes entre categorías.
- Variables categóricas ordinales: se utilizará *Label Encoding*, respetando el orden natural de sus niveles (por ejemplo: *Poor* $<$ *Fair* $<$ *Typical* $<$ *Good* $<$ *Excellent*).

C. Paradigma de Aprendizaje Seleccionado

El equipo de trabajo decidió abordar el problema bajo el paradigma de aprendizaje supervisado. De acuerdo con el curso, este tipo de aprendizaje se emplea cuando se cuenta con ejemplos donde tanto las entradas como las salidas están claramente identificadas, permitiendo que el modelo aprenda la relación entre ellas con el fin de realizar predicciones sobre nuevos datos.

En este contexto, el proyecto se plantea como un problema de regresión, dado que la variable objetivo corresponde a un valor numérico continuo (el precio de una vivienda). El curso establece que, cuando la salida a predecir es un número real, el enfoque adecuado es la regresión. Por lo tanto, esta configuración se considera apropiada para el tipo de predicción que se desea realizar. A partir de esta definición, el equipo procederá posteriormente a la construcción e implementación del modelo de regresión correspondiente.

II. Estado del arte

Artículo 1

Han, Y. (2023). *Price Prediction of Ames Housing Through Advanced Regression Techniques.* BCP Business & Management EMFRM, 38, 1966–1974.

Referencia: https://www.researchgate.net/publication/369437029_Price_Prediction_of_Ames_Housing_Through_Advanced_Regression_Techniques

Este trabajo aborda el problema de predicción del precio de venta de viviendas utilizando la base de datos *Ames Housing*, la misma empleada en la competencia *House Prices – Advanced Regression Techniques* de Kaggle. El estudio aplica un paradigma de aprendizaje supervisado de tipo regresión, orientado a estimar el valor de la variable continua *SalePrice*.

El autor implementa y compara diferentes técnicas de aprendizaje automático, entre ellas *LASSO*, *Elastic Net*, *Gradient Boosting*, *XGBoost*, *LightGBM* y un modelo ensamblado (*Stacked Model*). El proceso incluyó una exhaustiva ingeniería de características, imputación de valores faltantes y normalización de variables.

Como metodología de validación, se empleó una validación cruzada de 5 particiones (5-fold CV) sobre el conjunto de entrenamiento. La métrica principal fue el *Root Mean Squared Logarithmic Error (RMSLE)*, definida como:

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log(\hat{y}_i + 1) - \log(y_i + 1)\right)^2} \qquad (1)$$

Esta métrica penaliza de forma equilibrada los errores relativos en precios altos y bajos.

En cuanto a resultados, el mejor desempeño se obtuvo con el modelo *Gradient Boosting*, con un RMSLE de 0.046, seguido de *XGBoost* y *LightGBM* con valores cercanos. El autor concluye que los métodos basados en *boosting* ofrecen el mejor equilibrio entre precisión y eficiencia para este conjunto de datos.

Artículo 2

Harsora, H., Ogunleye, B., & Shobayo, O. (2023). *House Price Prediction Using Machine Learning Algorithms. Analytics, 3(1)*, 1–20.

Referencia: https://www.mdpi.com/2813-2203/3/1/3

Este estudio aborda el problema de la predicción del precio de viviendas utilizando el *Ames Housing Dataset*, disponible en Kaggle, el mismo conjunto de datos empleado en la competencia *House Prices – Advanced Regression Techniques*. Los autores aplican un paradigma de aprendizaje supervisado basado en regresión, enfocado en estimar el valor continuo de la variable *SalePrice*.

El trabajo compara el desempeño de cinco técnicas de aprendizaje automático: Regresión Lineal Múltiple, Red Neuronal Multicapa (*MLP*), *Random Forest Regressor*, *Support Vector Regression (SVR)* y *XGBoost*. Tras un proceso de limpieza y codificación de variables, se realizó ingeniería de características para optimizar el rendimiento predictivo de los modelos.

Como metodología de validación, se aplicó validación cruzada (*Cross-Validation*) para garantizar la estabilidad del modelo y reducir el riesgo de sobreajuste. Las métricas utilizadas fueron el coeficiente de determinación (R^2), R^2 ajustado, Error Absoluto Medio (*MAE*), Error Cuadrático Medio (*MSE*) y su raíz (*RMSE*), indicadores que permiten evaluar la precisión y generalización del modelo.

En cuanto a resultados, el algoritmo *XGBoost* alcanzó el mejor desempeño con un R^2 de 0.93, un *MSE* de 0.001, un *MAE* de 0.084 y una precisión promedio del 88.94% en la validación cruzada. Los autores concluyen que *XGBoost* constituye el modelo más estable y preciso para la predicción

de precios inmobiliarios en el conjunto de datos de *Ames*, superando consistentemente a los métodos tradicionales de regresión.

III. Referencias simples

Referencias

[1] A. Cook, "Categorical Variables," *Kaggle*, [Online]. Available: https://www.kaggle.com/code/alexisbcook/categorical-variables

[2] Kaggle, "House Prices – Advanced Regression Techniques," *Kaggle Competition Overview*, [Online]. Available: https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview

[3] J. Darias, "Introducción al Aprendizaje Automático – Unidad 1," *Intro_ML_2025*, [Online]. Available: https://jdariasl.github.io/Intro_ML_2025/titles/U1_description.html

[4] Y. Han, "Price Prediction of Ames Housing Through Advanced Regression Techniques," *BCP Business & Management EMFRM*, vol. 38, pp. 1966–1974, 2023. [Online]. Available: https://www.researchgate.net/publication/369437029_Price_Prediction_of_Ames_Housing_Through_Advanced_Regression_Techniques

[5] H. Harsora, B. Ogunleye, and O. Shobayo, "House Price Prediction Using Machine Learning Algorithms," *Analytics*, vol. 3, no. 1, pp. 1–20, 2023. [Online]. Available: https://www.mdpi.com/2813-2203/3/1/3