AB FINAL-UNIT 25

ÍNDICE

INTRODUCCIÓN

En este proyecto nos han encomendado la tarea de resolver un problema actual de un ámbito, el mío se basa en el dataset “House Prices: Advanced Regression Techniques” de Kaggle.
Este problema consiste en predecir el precio de venta de viviendas en la ciudad de Ames (EE.UU.) a partir de características estructurales, geográficas y de calidad de construcción.

Este tipo de predicción es altamente relevante en el sector inmobiliario, donde estimar el valor de una vivienda de manera objetiva permite:

- Ayudar a compradores a identificar precios justos.
- Asistir a agencias inmobiliarias en la tasación automática.
- Facilitar estudios de mercado y planificación urbana.

Hipótesis

Las hipótesis iniciales planteadas fueron:

- La superficie habitable (GrLivArea) influye positivamente en el precio de venta.
- La calidad general de la vivienda (OverallQual) es una de las variables más determinantes.
- La ubicación (Neighborhood) impacta significativamente en el valor final.

Se esperaba que el modelo confirmara estas relaciones y permitiera cuantificar su importancia.

Valor para el usuario final

El usuario final puede ser:

- Una agencia inmobiliaria.
- Un comprador particular.
- Un analista de mercado.

El valor del modelo radica en ofrecer una estimación automática y rápida del precio de una vivienda, reduciendo la subjetividad y mejorando la toma de decisiones.

Fundamentación Teórica

Modelo de aprendizaje utilizado

He empleado un modelo de Random Forest Regressor, un algoritmo de aprendizaje supervisado basado en la combinación de múltiples árboles de decisión.

Ventajas principales:

- Captura relaciones no lineales entre variables.
- Tolera bien datos ruidosos y valores faltantes.
- Proporciona medidas de importancia de variables, útiles para storytelling.

Relación con la Red Neuronal en Excel

En la práctica previa de redes neuronales en Excel se estudiaron conceptos como:

- **Valor**: importancia de cada variable.
- **Bias**: ajuste del modelo.
- **Error y función de coste**: diferencia entre predicción y valor real.
- **Iteración y aprendizaje**: ajuste progresivo para minimizar el error.

Estos conceptos son equivalentes en cualquier modelo de Machine Learning; el Random Forest también aprende patrones minimizando errores, aunque mediante árboles en lugar de neuronas.

Ingeniería y Análisis de Datos

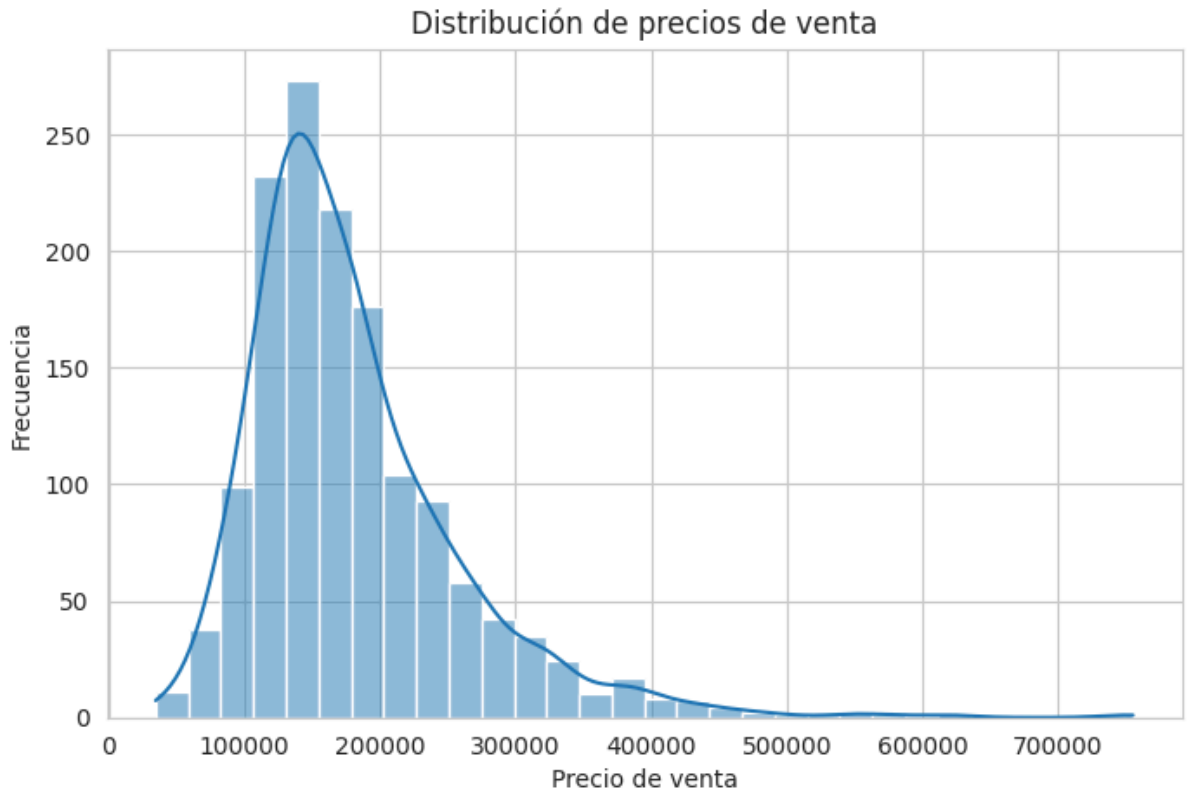
1. Exploración de Datos (EDA)

Se analizó el archivo train.csv, que contiene 1460 viviendas con 79 variables explicativas y la variable objetivo SalePrice.

Distribución del precio de venta

El histograma de SalePrice muestra:

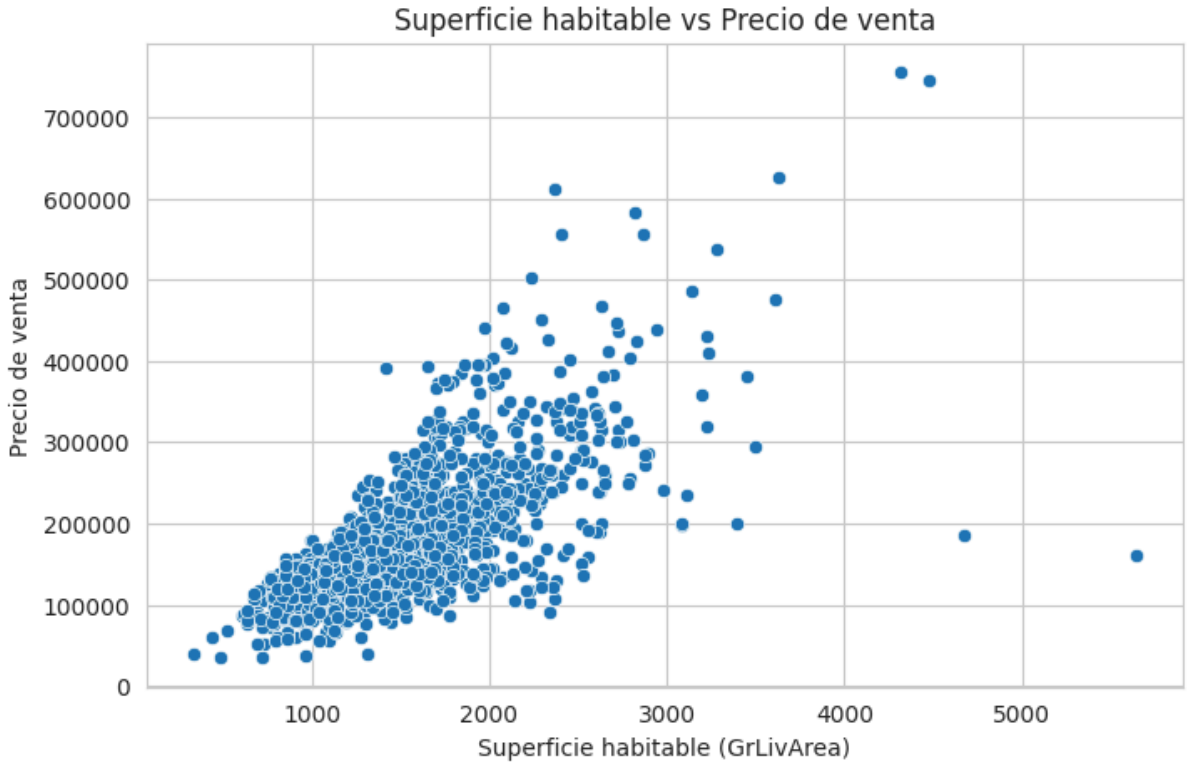
- Concentración de precios en rangos medios.
- Ligera asimetría hacia precios altos.
- Presencia de outliers de viviendas de lujo.



Superficie habitable vs Precio

El scatterplot entre GrLivArea y SalePrice evidencia:

- Fuerte correlación positiva.
- Algunas viviendas muy grandes con precios más bajos de lo esperado (outliers).

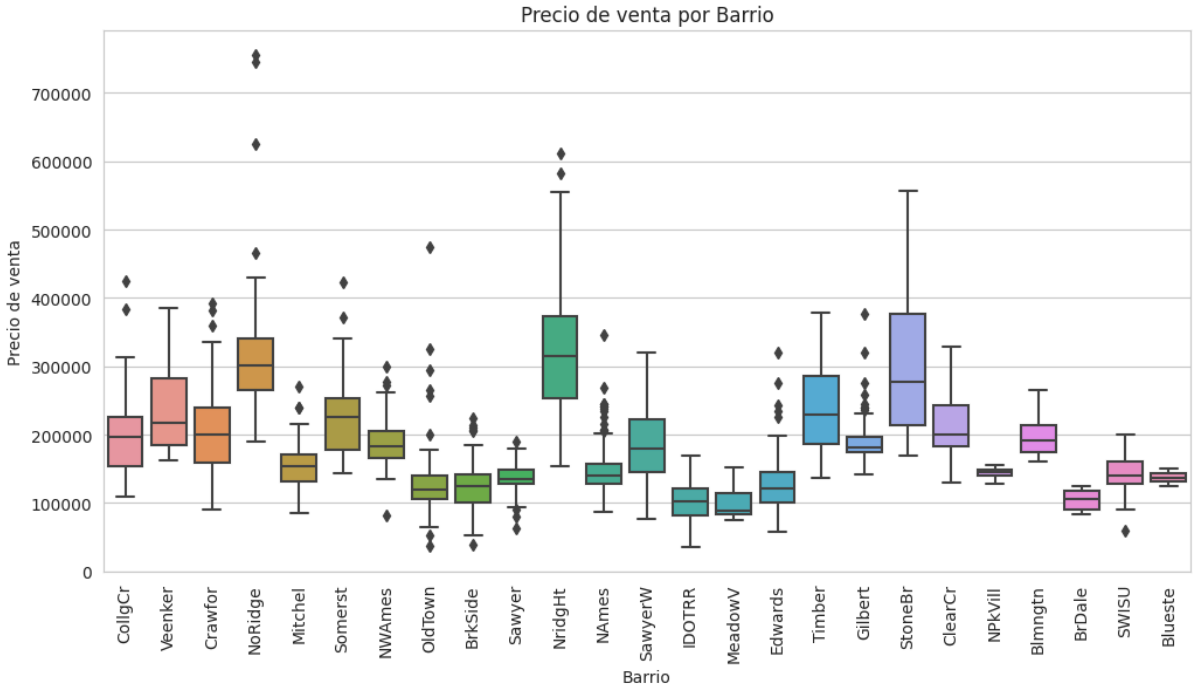


Barrio vs Precio

El boxplot de Neighborhood vs SalePrice muestra:

- Barrios con precios claramente diferenciados.

- Confirmación de la hipótesis sobre la importancia de la ubicación.



2. Preprocesamiento

En el proyecto se aplicaron las siguientes transformaciones:

- **Valores numéricos:**
    - Imputación por mediana.
    - Estandarización (StandardScaler).
- **Valores categóricos:**
    - Imputación por moda.

    - Codificación One-Hot Encoding.

Y se separaron:

- Variables explicativas: X
- Variable objetivo: y = SalePrice

3. Entrenamiento del modelo

- División entrenamiento / validación: 80% / 20%.
- Modelo: RandomForest Regressor con 200 árboles.
- Métrica: RMSE (Root Mean Squared Error).

Resultado obtenido:

RMSE en validación = 28512.645677520482

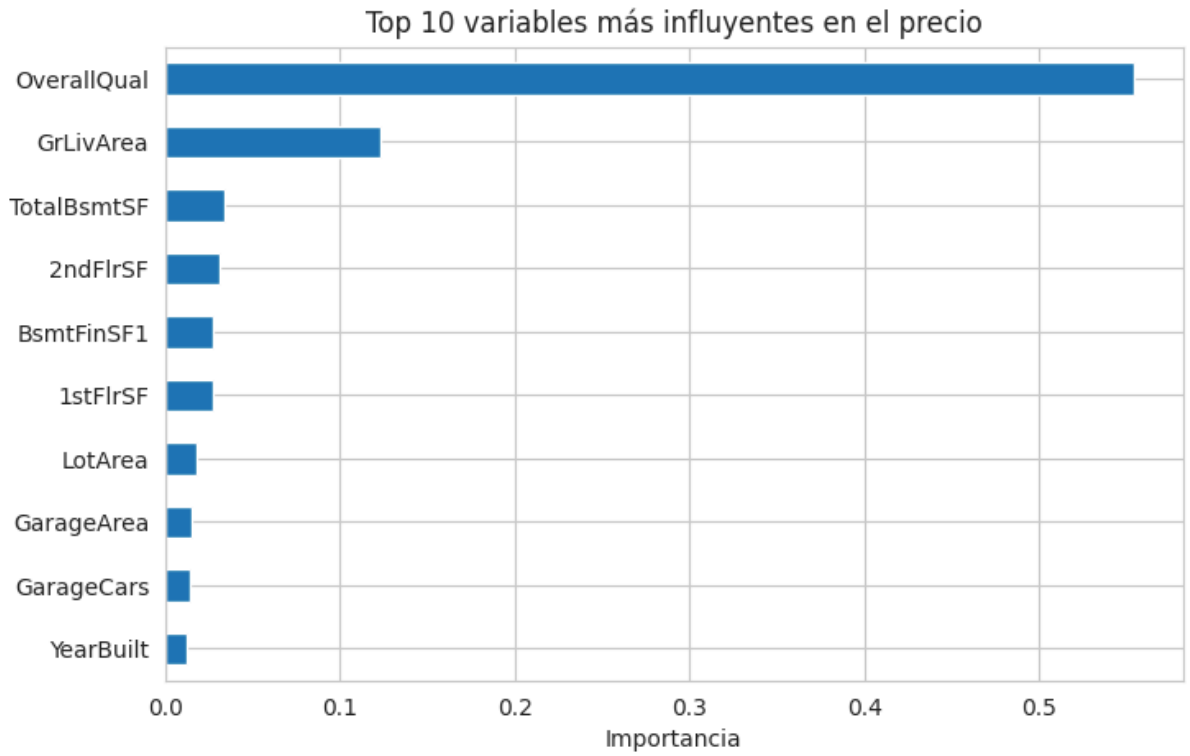
Este es el valor que indica un error promedio razonable considerando la escala de precios de mi dataset.

4. Importancia de variables

El modelo identificó como más influyentes:

- OverallQual
- GrLivArea

- TotalBsmtSF
- Variables de barrio (Neighborhood)



5. Análisis de errores

- El modelo presenta mayor error en viviendas extremadamente caras.
- La hipótesis inicial se cumple:
    - Superficie y calidad tienen alta influencia.
    - Ubicación afecta fuertemente el precio.

- La limitación principal es la escasez de ejemplos de viviendas de lujo.

Este análisis demuestra comprensión del comportamiento del modelo más allá de la métrica numérica.

6. Predicción final

Se aplicó el modelo entrenado sobre test.csv para generar el archivo submission.csv, listo para enviar a Kaggle.

Arquitectura y Despliegue

1. Flujo de datos

Usuario → Interfaz Web → Modelo ML → Predicción → Usuario

↑

Feedback para reentrenamiento

2. Comunicación cliente-servidor

- El usuario introduce características de una vivienda.
- La interfaz web envía los datos al servidor.

- El servidor aplica el pipeline de preprocesamiento + modelo.
- Devuelve la predicción de precio al usuario.

3. Ciclo de feedback

- Los datos introducidos por los usuarios pueden almacenarse.
- Se incorporan periódicamente al dataset.
- El modelo se reentrena para mejorar la precisión futura.

Este flujo representa un sistema básico de MLOps.

4. Tecnologías propuestas

- **Frontend:** HTML / JavaScript.
- **Backend:** Python (Flask o FastAPI).
- **Modelo:** Pipeline de Scikit-Learn.
- **Despliegue:** Vercel.
- **Repositorio:** GitHub.

5. Valor de la arquitectura

- Predicciones en tiempo real.
- Escalabilidad del sistema.
- Posibilidad de mejora continua mediante feedback.

Conclusiones

Con este proyecto hemos aprendido a dominar el uso de datasets y Kagle, una página web con la posibilidad de aprender y practicar cientos de cosas en el ámbito que estudiamos actualmente. Con la posibilidad de programar con un diseño más aseado y cómodo. Respecto a los objetivos del trabajo he demostrado completamente la integración del:

- Análisis exploratorio de datos.
- Preprocesamiento.
- Entrenamiento de modelo predictivo.
- Evaluación crítica de resultados.
- Propuesta de despliegue real.

La hipótesis inicial ha sido validada y hemos conseguido obtener un modelo funcional, interpretable y listo para ser integrado en una aplicación web con muchas ventajas en el sector inmobiliario.

Bibliografía

- Ames Housing Dataset, 2011. *House Prices: Advanced Regression Techniques*. Kaggle. Disponible en: Kaggle competition dataset.
- Breiman, L., 2001. *Random Forests*. Machine Learning, 45(1), pp.5–32.
- Pedregosa, F. et al., 2011. *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, pp.2825–2830.
- Seaborn Library Documentation, 2024. *Statistical Data Visualization*. Disponible en: Seaborn documentation.
- Pandas Development Team, 2024. *Pandas Documentation*. Disponible en: Pandas documentation.
- MyNextEmployee, 2023. *Neural Network in Excel Example*. Disponible en: https://www.mynextemployee.com/post/neural-network-in-excel-example
- Kaggle Learn, 2024. *Intro to Machine Learning*. Kaggle. Disponible en: Kaggle Learn Courses.