Predicción del Precio de Viviendas Residenciales mediante Análisis Multivariante y Regularización

Autor: Marco Fernández Pérez

Asignatura: Matemáticas y Estadística para la IA

Fecha: 1 de diciembre de 2025

1. Resumen

El objetivo de este proyecto es desarrollar un modelo predictivo robusto para estimar el precio de venta (*SalePrice*) de viviendas residenciales en Ames, Iowa. Se ha trabajado con un dataset de alta dimensionalidad (79 variables), aplicando técnicas avanzadas de preprocesamiento para evitar *data leakage* y comparando cinco enfoques de modelado: Regresión sobre Componentes Principales (PCR) y regresión regularizada (Lasso y Ridge), tanto sobre variables originales como sobre proyecciones de PCA.

El modelo **Ridge Regression** aplicado sobre las variables originales estandarizadas resultó ser el más eficaz, alcanzando un **RMSE de 0.1197** en validación y un **Error Absoluto Medio (MAE) de $16,001** en el conjunto de test final. Este resultado sugiere que la conservación de todas las características, ponderadas adecuadamente para gestionar la multicolinealidad, es superior a la reducción de dimensionalidad agresiva en este contexto inmobiliario.

2. Introducción y Contextualización

El mercado inmobiliario se caracteriza por su heterogeneidad, donde el precio de un activo depende de una compleja interacción entre características estructurales, de calidad y ubicación. El problema planteado es una tarea de **regresión supervisada** sobre un conjunto de datos que presenta desafíos típicos de datos reales: valores faltantes, mezcla de variables numéricas y categóricas, y una alta multicolinealidad entre predictores (ej. *GarageArea* y *GarageCars*).

El objetivo matemático es estimar una función $Y = f(X) + \epsilon$, donde Y es el logaritmo del precio de venta y X es el vector de características de la vivienda, minimizando el error de predicción en datos no vistos.

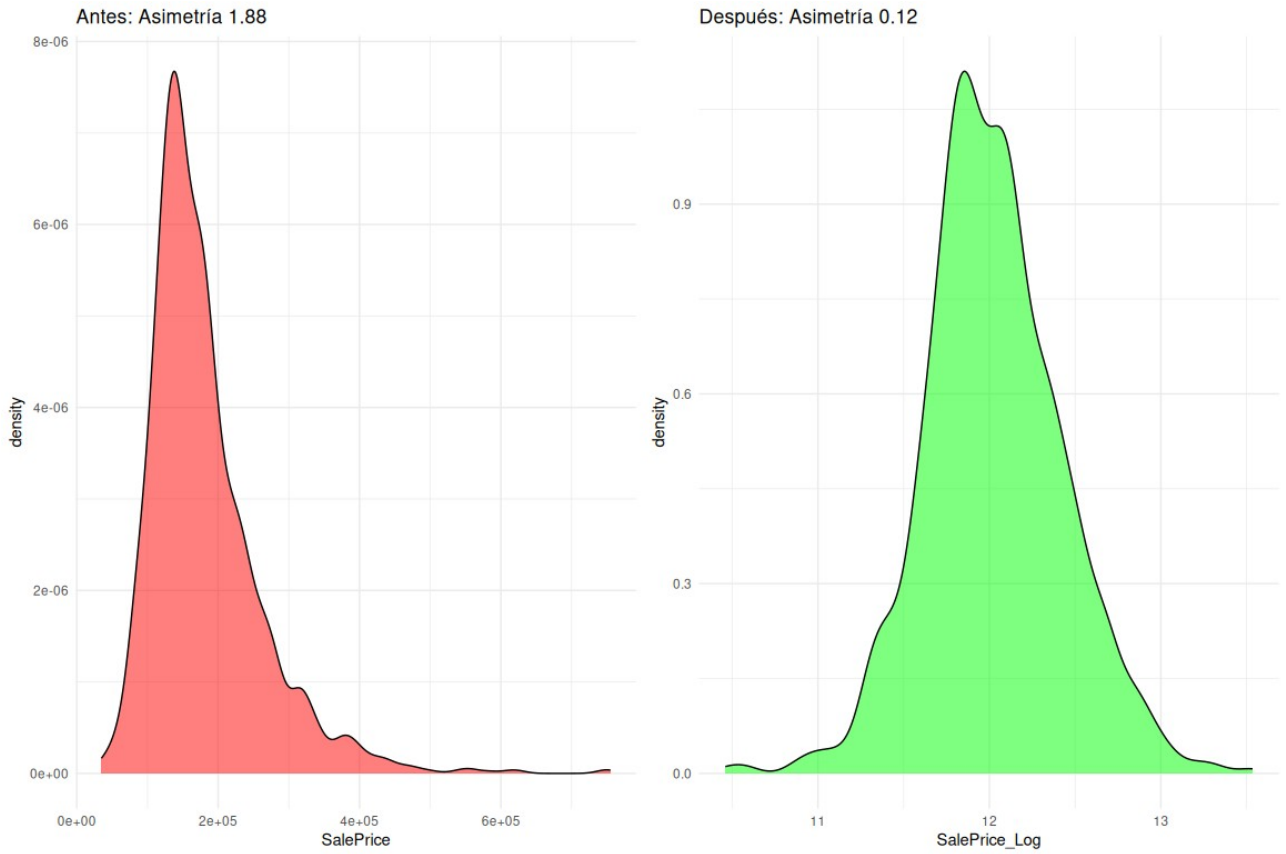
3. Análisis Exploratorio de Datos (EDA)

3.1. Análisis de la Variable Objetivo

La variable *SalePrice* original presentaba una distribución fuertemente asimétrica positiva (*skewness* = 1.88), lo cual viola la asunción de normalidad de los residuos en modelos lineales. Se aplicó una **transformación logarítmica**:

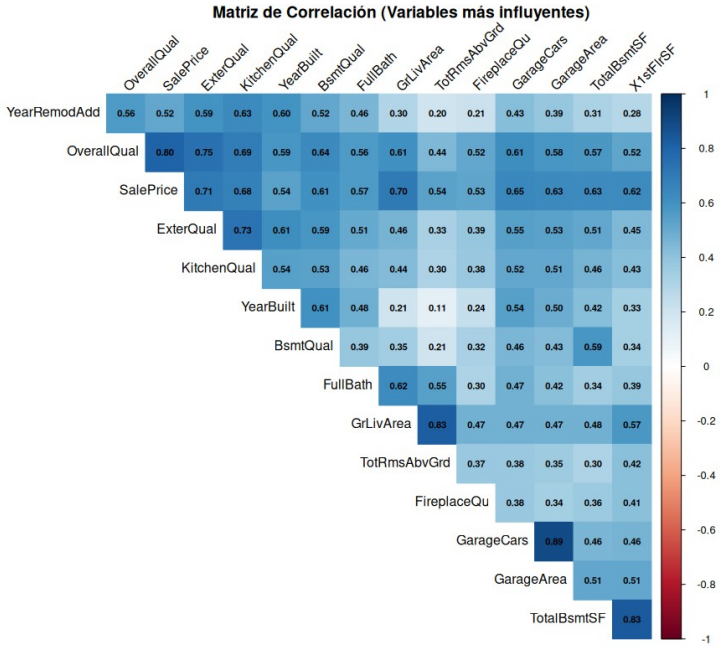
$$Y' = \ln(Y)$$

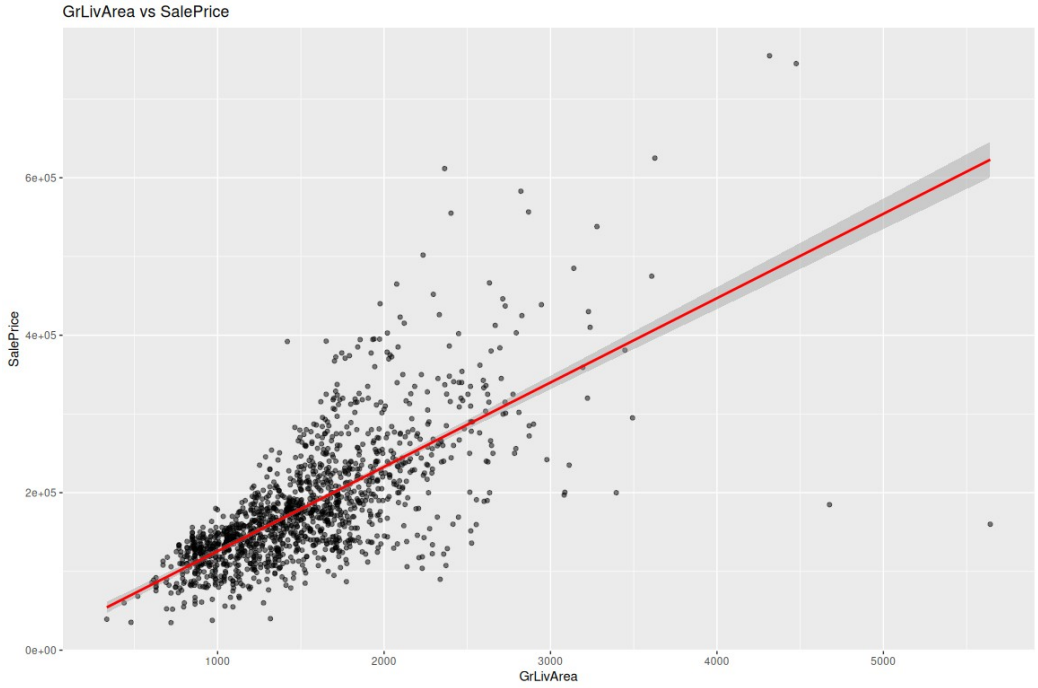
Esta transformación redujo la asimetría a **0.12**, estabilizando la varianza y mejorando la idoneidad para el modelado lineal.



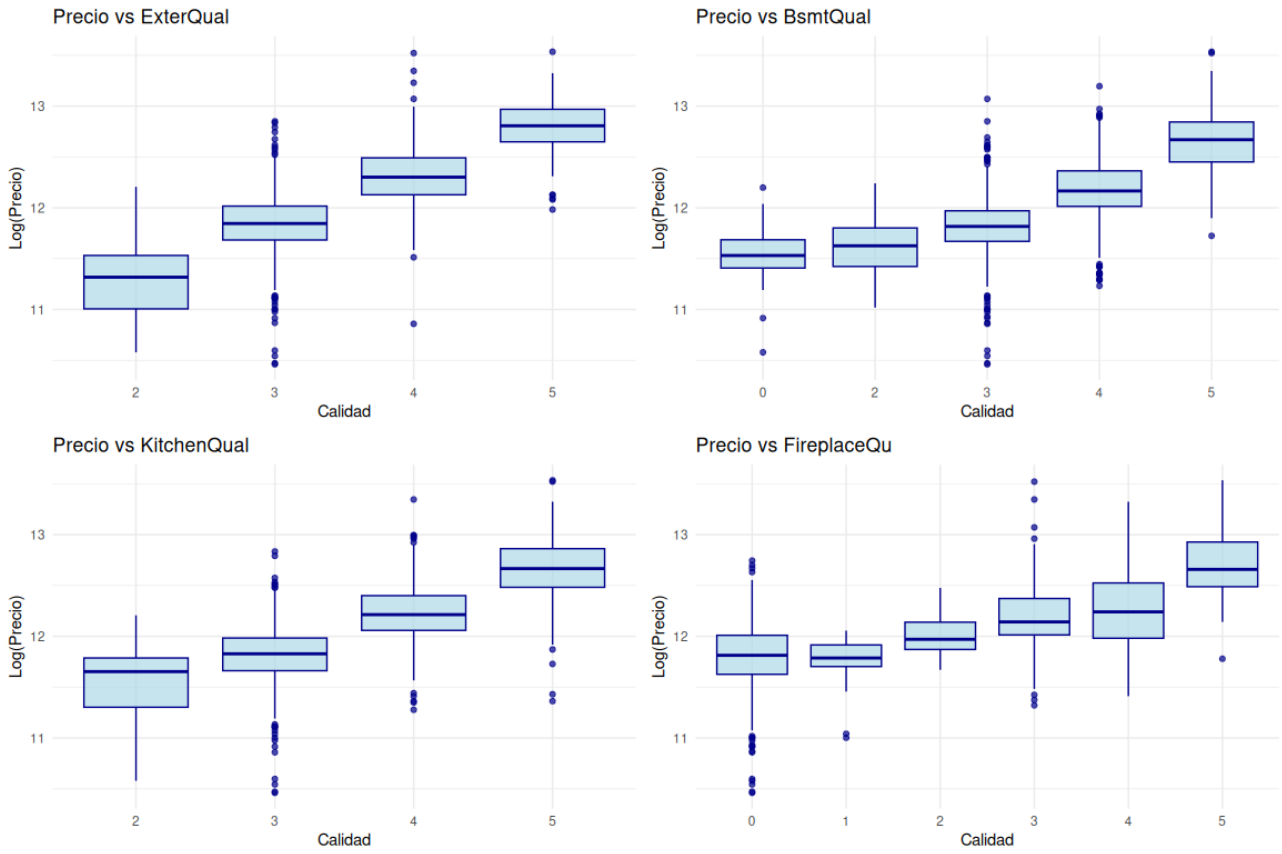
3.2. Correlaciones y Outliers

Mediante la matriz de correlación, se identificaron variables con fuerte relación lineal con el precio, destacando *OverallQual* (Calidad General) y *GrLivArea* (Área habitable sobre rasante). Se detectaron anomalías visuales en el gráfico de dispersión de *GrLivArea* vs *SalePrice*: dos propiedades con más de 4,000 pies cuadrados pero con precios inusualmente bajos (< $300,000). **Decisión:** Estos dos puntos fueron eliminados del dataset antes del entrenamiento para evitar que actuaran como puntos de palanca (*leverage points*) que sesgaran la regresión.





Respecto a las variables categóricas ordinales, que fueron transformadas a numéricas, se realizaron boxplots para observar su correlación con la variable objetivo *SalePrice.*



Entre estas variables, las correlaciones superiores a 0.5 respecto a la variable objetivo pertenecían a las variables *ExterQual*, *BsmtQual*, *KitchenQual* y *FireplaceQu*.

4. Metodología y Preprocesamiento

Siguiendo las mejores prácticas para evitar la fuga de datos (*data leakage*), se estableció el siguiente protocolo estricto:

1. **División de Datos:** Se realizó una partición aleatoria (semilla 77) del dataset limpio: **60% Entrenamiento**, **20% Validación** y **20% Test**.

2. **Imputación y Limpieza:**

- **Variables Categóricas (NA estructural):** Los valores NA en variables como *PoolQC* o *GarageType* no indican datos perdidos, sino la ausencia de la característica. Se imputaron como "None" o 0.
- **Codificación Ordinal:** Variables cualitativas con orden intrínseco (ej. *ExterQual*: Po, Fa, TA, Gd, Ex) fueron mapeadas a una escala numérica 1 ...5 para preservar la información de jerarquía.

3. **Prevención de Leakage en Estandarización:** Los parámetros de estandarización (media μ y desviación estándar σ) y de imputación por mediana se calcularon **exclusivamente usando el conjunto de entrenamiento**. Posteriormente, estas transformaciones se aplicaron a los conjuntos de validación y test:

$$Z_{val} = \frac{X_{val} - \mu_{train}}{\sigma_{val}}$$

Esto garantiza una evaluación honesta del rendimiento en escenarios reales.

5. Modelado Predictivo: Formulación Matemática

Se entrenaron cinco modelos distintos. A continuación se detalla la formulación matemática de los enfoques principales.

5.1. Análisis de Componentes Principales (PCA)

Para reducir la dimensionalidad y eliminar la multicolinealidad, se aplicó PCA sobre las variables predictoras estandarizadas. PCA busca una transformación ortogonal tal que los nuevos componentes Z_1, Z_2, ... sean combinaciones lineales de las variables originales X que maximicen la varianza explicada. Se seleccionaron los primeros k componentes necesarios para explicar el **95% de la varianza total**.

5.2. Regresión Lineal Regularizada

Dado el alto número de predictores, la regresión lineal simple tiende al sobreajuste (alta varianza). Se utilizaron métodos de regularización que añaden un término de penalización a la función de coste de Mínimos Cuadrados Ordinarios (OLS).

A. Regresión Ridge (Norma L_2)

Ridge penaliza la suma de los cuadrados de los coeficientes, encogiéndolos hacia cero pero sin anularlos. Es ideal para manejar la multicolinealidad presente en este dataset (ej. variables de área correlacionadas).

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right)$$

B. Regresión Lasso (Norma L_1)

Lasso penaliza la suma de los valores absolutos de los coeficientes. Debido a la geometría de la restricción (romboidal), Lasso tiene la propiedad de forzar coeficientes a ser **exactamente cero**, realizando una selección automática de variables.

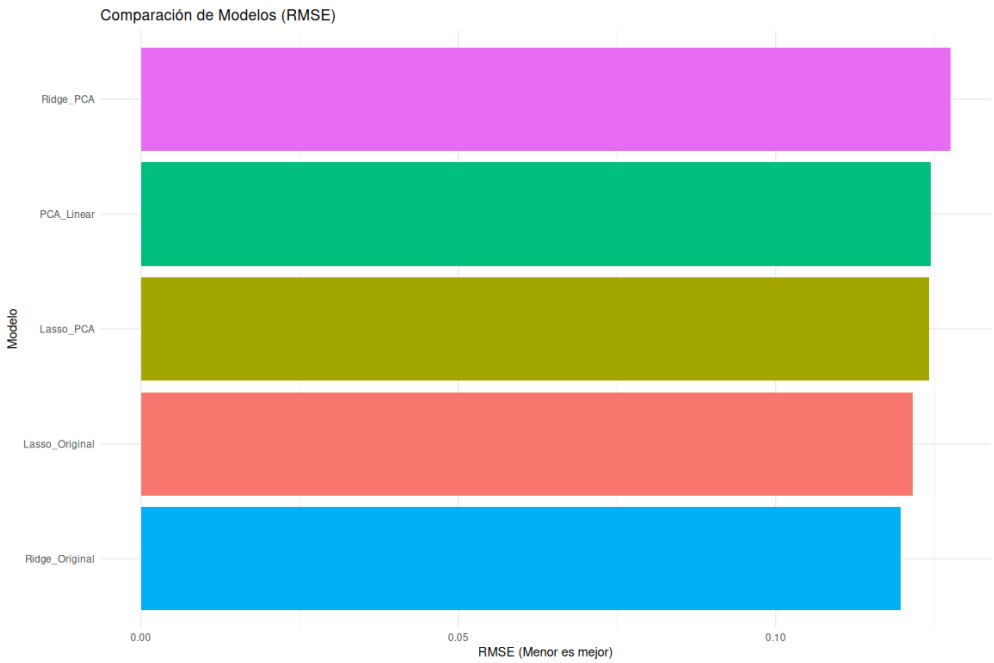
$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

Optimización de Hiperparámetros: El parámetro de penalización λ se optimizó mediante validación cruzada (*5-fold Cross-Validation*) sobre el conjunto de entrenamiento.

6. Resultados y Discusión

6.1. Selección del Modelo (Conjunto de Validación)

Se evaluó el rendimiento de los cinco modelos utilizando la raíz del error cuadrático medio (RMSE) sobre el conjunto de validación (datos no vistos durante el ajuste de parámetros).



Modelo	RMSE (Escala Log)	Interpretación
Ridge Original	**0.1197**	**Mejor rendimiento.** Gestiona bien la multicolinealidad
Lasso Original	0.1216	Buen rendimiento, eliminó variables menos relevantes.
PCA + Linear	0.1244	Pérdida de información al reducir dimensiones.
Lasso sobre PCA	0.1241	No aporta mejora significativa sobre PCA simple.
Ridge sobre PCA	0.1275	Peor rendimiento relativo.

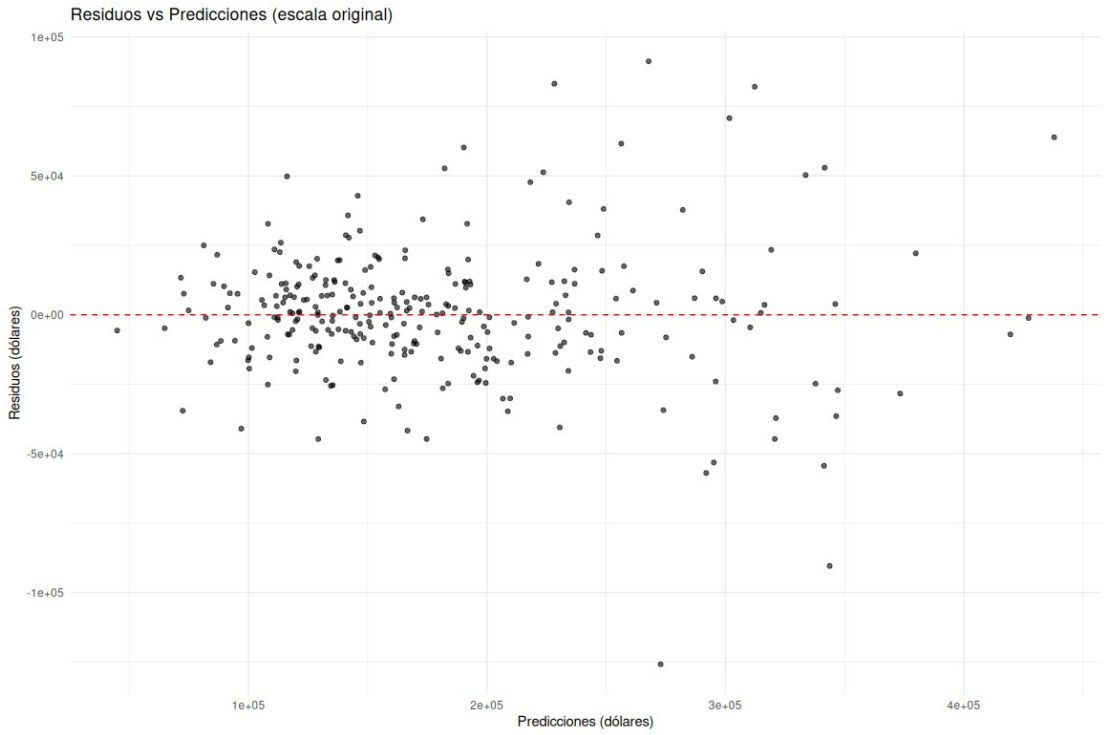
Discusión: El modelo **Ridge sobre variables originales** superó a los modelos basados en PCA y a Lasso. Esto sugiere que, en la valoración de viviendas, la mayoría de las variables aportan información marginal valiosa. Mientras Lasso elimina variables y PCA descarta varianza "pequeña", Ridge mantiene toda la información "amortiguando" el ruido de la multicolinealidad, lo cual resulta óptimo en este escenario denso.

6.2. Evaluación Final (Conjunto de Test)

El modelo seleccionado (Ridge Original) fue evaluado finalmente sobre el conjunto de Test (el 20% de datos reservados hasta el final).

- **RMSE Final (Log):** 0.1306

- **MAE Final (Dólares):** $16,001.42



El análisis de residuos confirma que el modelo cumple razonablemente con los supuestos de homocedasticidad y normalidad, aunque se observa una leve dispersión en las viviendas de precio muy alto, lo cual es habitual en este dominio.

7. Conclusiones

- **Eficacia de la Regularización:** Se ha demostrado matemáticamente y empíricamente que la regularización L_2 (Ridge) es superior a la reducción de dimensionalidad no supervisada (PCA) para este problema específico. PCA, al maximizar solo la varianza de los predictores sin considerar la variable respuesta, puede descartar información predictiva crucial contenida en componentes de baja varianza.

- **Robustez Metodológica:** La aplicación estricta de la separación de datos y el cálculo de parámetros de preprocesamiento exclusivamente en *train* garantiza que el error reportado de $16,001 es una estimación realista y no optimista del desempeño del modelo en producción.

- **Valor de Negocio:** El modelo final permite a una entidad inmobiliaria realizar tasaciones automáticas con un margen de error medio de aproximadamente 16,000 dólares. Esta herramienta es altamente valiosa para el filtrado inicial de carteras de inversión o para ofrecer estimaciones instantáneas en plataformas web, optimizando significativamente los tiempos respecto a la tasación manual.