Universidad Del Valle de Guatemala
Minería de datos
Sección 30

Ing. Mario Barrientos

Proyecto 2

Análisis y Predicción de Precios de Casas

Emilio Reyes, 22674
Silvia Illescas, 22376
Michelle Mejía, 22596

Guatemala, 23 de marzo de 2025

Contenido

1. Introducción

El presente proyecto tiene como objetivo principal realizar un análisis exploratorio y predictivo sobre un conjunto de datos de viviendas, con el fin de comprender los factores que más influyen en el precio de venta de una casa y construir modelos que permitan estimar dicho precio con la mayor precisión posible.

Para ello, se trabajó con un dataset proveniente de Kaggle (House Prices - Advanced Regression Techniques), el cual contiene información detallada de 1,460 propiedades residenciales, incluyendo características estructurales, de calidad, antigüedad, ubicación, entre otras.

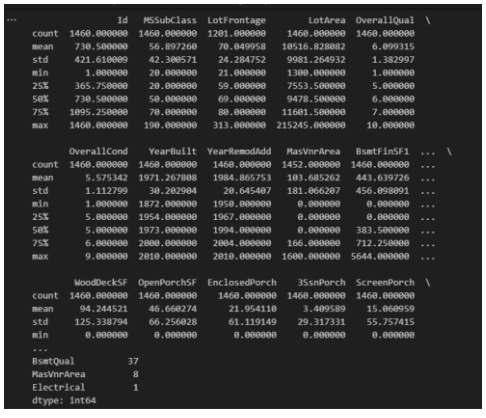
A través de este análisis se buscó:

- ✓ Identificar las variables más relevantes en la predicción del precio (SalePrice).
- ✓ Agrupar las casas por rangos de precio (Económica, Intermedia, Cara).
- ✓ Entrenar diversos modelos de regresión y clasificación.
- ✓ Comparar su desempeño para determinar cuál es más adecuado según el objetivo.

Este informe documenta el proceso completo desde la exploración de los datos, creación de nuevas variables, desarrollo y evaluación de modelos, hasta las conclusiones y recomendaciones para trabajos futuros.

2. Exploración de Datos (EDA)

Se inició con la carga del conjunto de datos y la revisión de su estructura general. El dataset contiene 81 columnas y 1,460 registros, cada uno representando una casa distinta.



```
...                Id   MSSubClass  LotFrontage        LotArea  OverallQual  \
count  1460.000000  1460.000000  1201.000000    1460.000000  1460.000000
mean    730.500000    56.897260    70.049958   10516.828082     6.099315
std     421.610009    42.300571    24.284752    9981.264932     1.382997
min       1.000000    20.000000    21.000000    1300.000000     1.000000
25%     365.750000    20.000000    59.000000    7553.500000     5.000000
50%     730.500000    50.000000    69.000000    9478.500000     6.000000
75%    1095.250000    70.000000    80.000000   11601.500000     7.000000
max    1460.000000   190.000000   313.000000  215245.000000    10.000000

       OverallCond    YearBuilt  YearRemodAdd   MasVnrArea   BsmtFinSF1  ...  \
count  1460.000000  1460.000000   1460.000000  1452.000000  1460.000000  ...
mean      5.575342  1971.267808   1984.865753   103.685262   443.639726  ...
std       1.112799    30.202904     20.645407   181.066207   456.098091  ...
min       1.000000  1872.000000   1950.000000     0.000000     0.000000  ...
25%       5.000000  1954.000000   1967.000000     0.000000     0.000000  ...
50%       5.000000  1973.000000   1994.000000     0.000000   383.500000  ...
75%       6.000000  2000.000000   2004.000000   166.000000   712.250000  ...
max       9.000000  2010.000000   2010.000000  1600.000000  5644.000000  ...

       WoodDeckSF  OpenPorchSF  EnclosedPorch    3SsnPorch  ScreenPorch  \
count  1460.000000  1460.000000    1460.000000  1460.000000  1460.000000
mean     94.244521    46.660274      21.954110     3.409589    15.060959
std     125.338794    66.256028      61.119149    29.317331    55.757415
min       0.000000     0.000000       0.000000     0.000000     0.000000
...
BsmtQual         37
MasVnrArea        8
Electrical        1
dtype: int64
```

Figura 1. Estadísticas Descriptivas

Se realizó un análisis de:

Valores nulos: Se identificaron variables con datos faltantes, como LotFrontage, GarageYrBlt y algunas categóricas como Alley y PoolQC, las cuales fueron tratadas mediante eliminación o imputación según su relevancia.

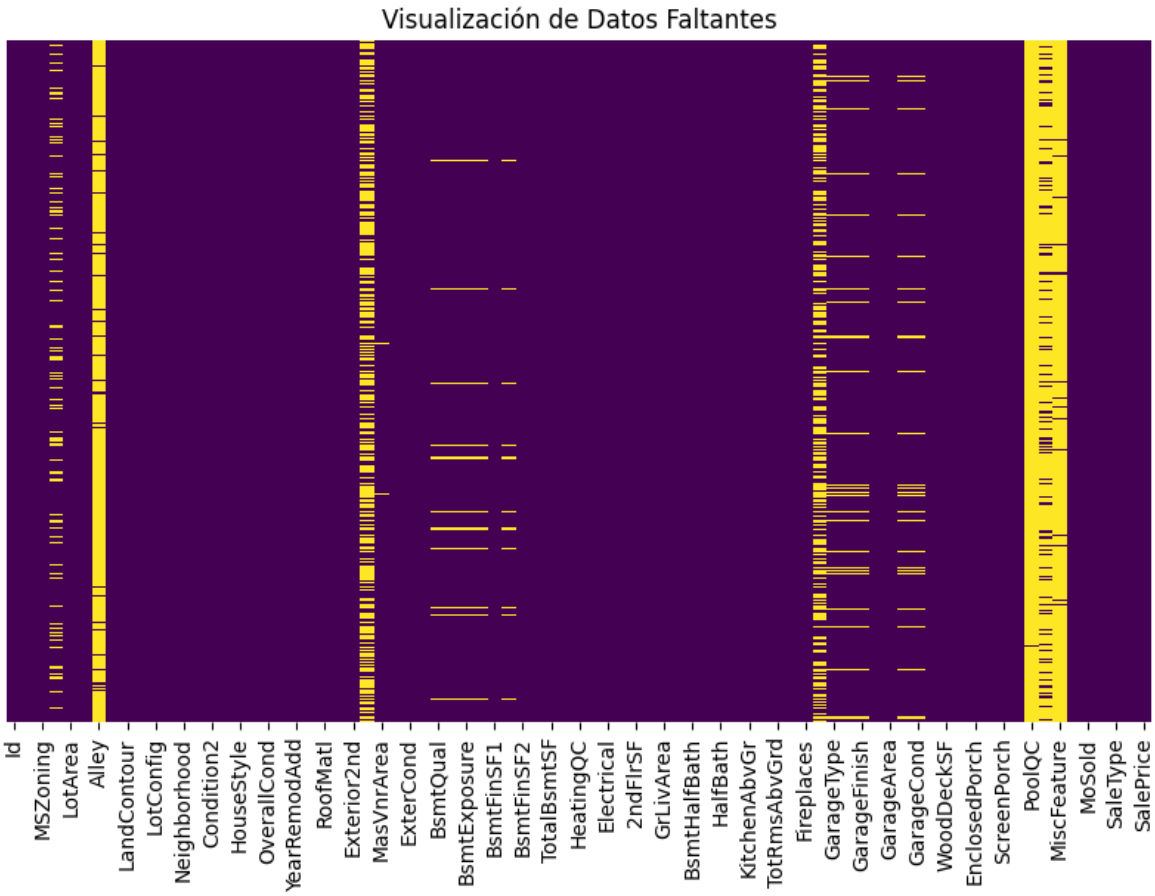


Figura 2. Valores Faltantes

Variables numéricas: Se analizaron estadísticas como la media, mediana, desviación estándar y presencia de outliers mediante diagramas de cajas. Variables como GrLivArea, SalePrice y TotalBsmtSF mostraron distribuciones sesgadas hacia la derecha.

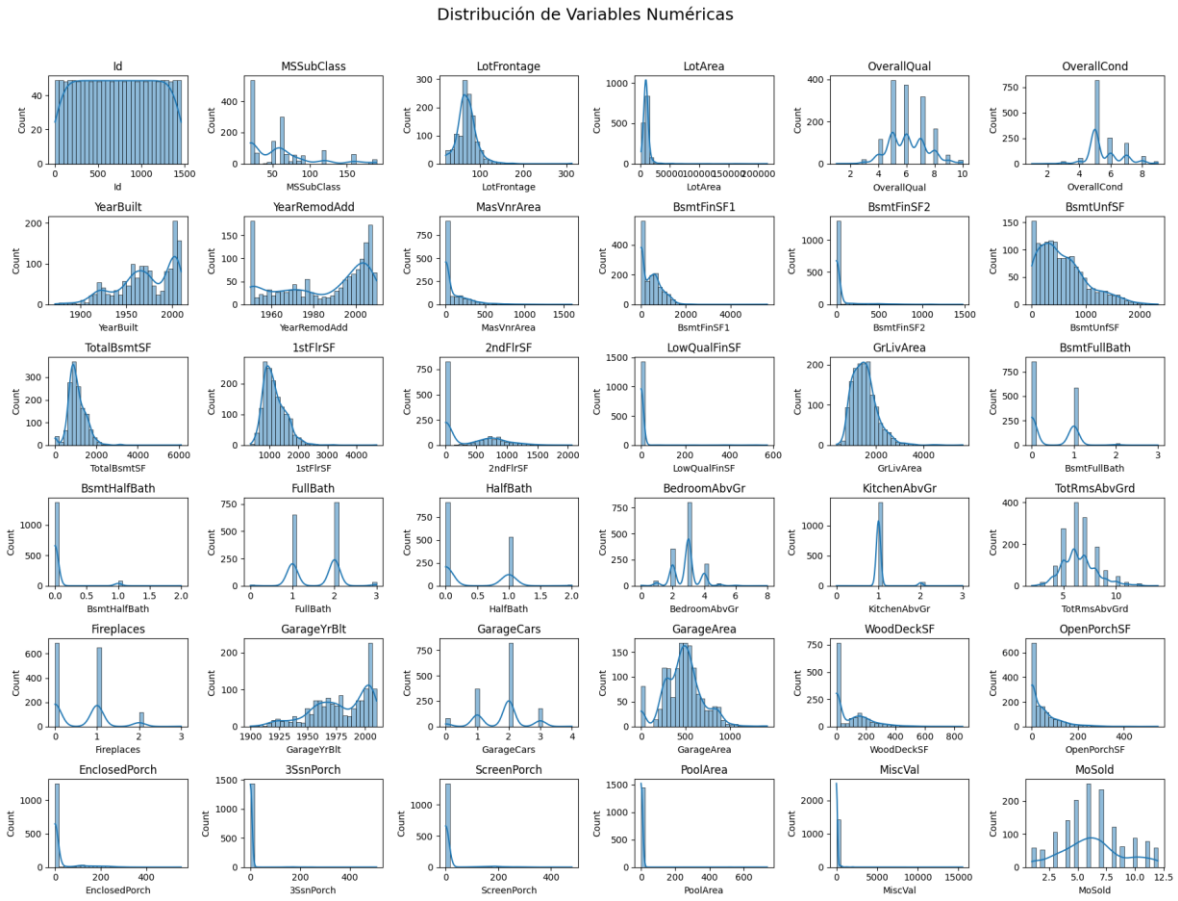


Figura 4. Distribución de Variables Numéricas

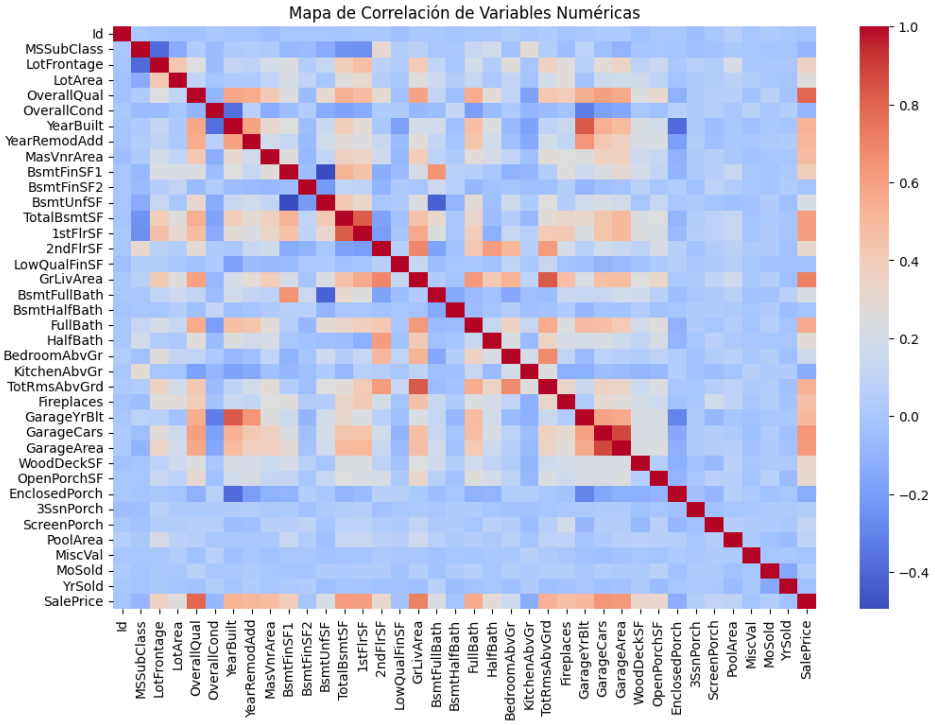
Variables categóricas: Se exploraron frecuencias con gráficos de barras para columnas como Neighborhood, HouseStyle y ExterQual. Se identificaron categorías dominantes y se evaluó su relación con el precio.

También se generaron:

Histogramas para evaluar la distribución de precios y variables numéricas.

Boxplots para visualizar la relación entre la calidad general (OverallQual) y el precio.

Matriz de correlación (heatmap): que mostró que OverallQual, GrLivArea y GarageCars están fuertemente correlacionadas con el precio de venta.




```
Top Variables Correlacionadas con SalePrice:
SalePrice       1.000000
OverallQual     0.790982
GrLivArea       0.708624
GarageCars      0.640409
GarageArea      0.623431
TotalBsmtSF     0.613581
1stFlrSF        0.605852
FullBath        0.560664
TotRmsAbvGrd    0.533723
YearBuilt       0.522897
Name: SalePrice, dtype: float64
```

Figura 4. Mapa e índice de correlación entre variables con SalesPrice

Por último, se realizó un análisis de grupos utilizando KMeans, donde se detectaron tres segmentos principales de viviendas, los cuales ayudaron a respaldar la creación de la variable categórica PriceCategory.

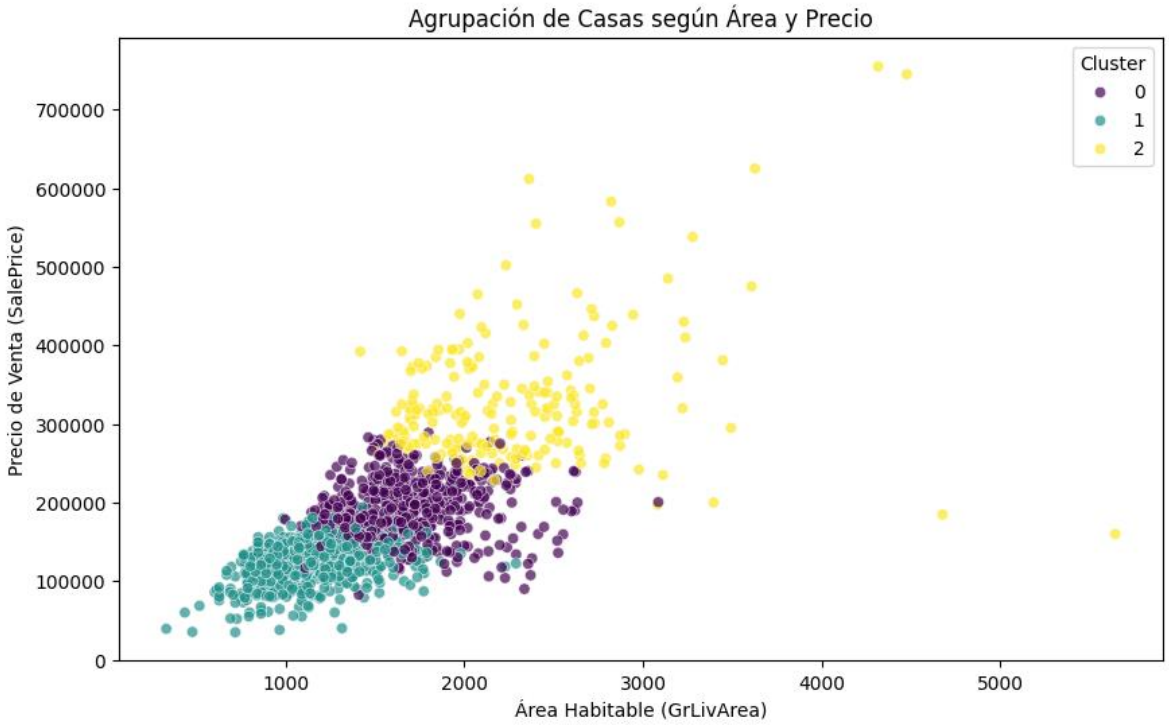


Figura 5. Agrupación de casas según precio (Clústers)

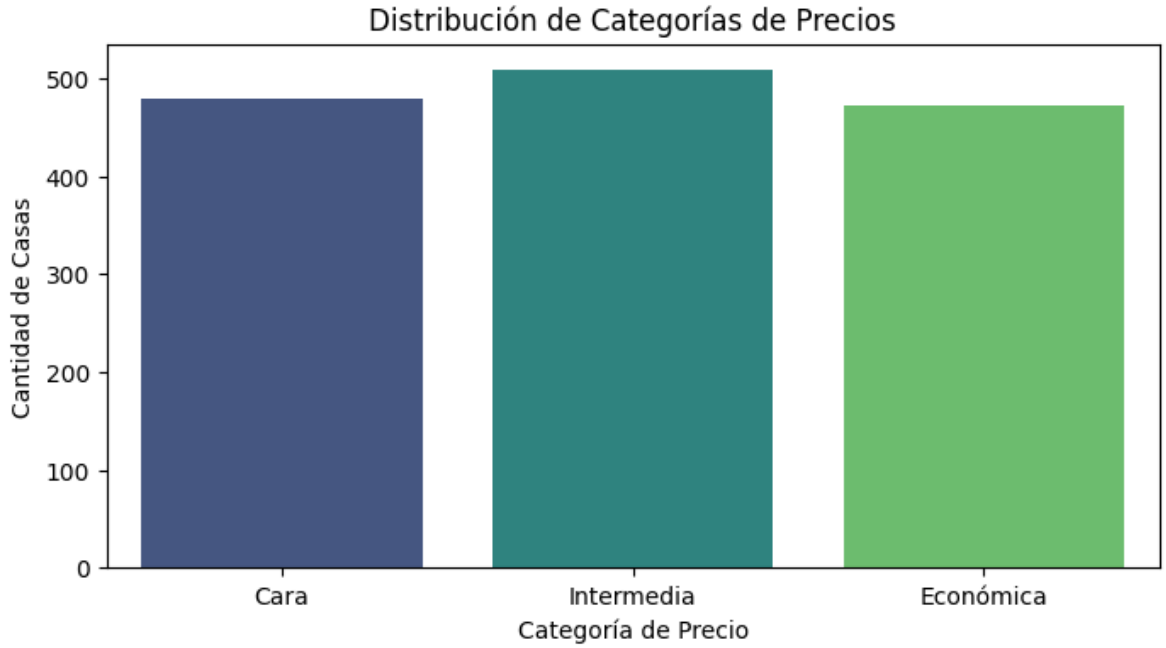


Figura 6. Cantidad de casas según precio (Dividido según percentiles 33 y 67)

El EDA permitió definir una base sólida para la ingeniería de características y selección de variables predictoras relevantes.

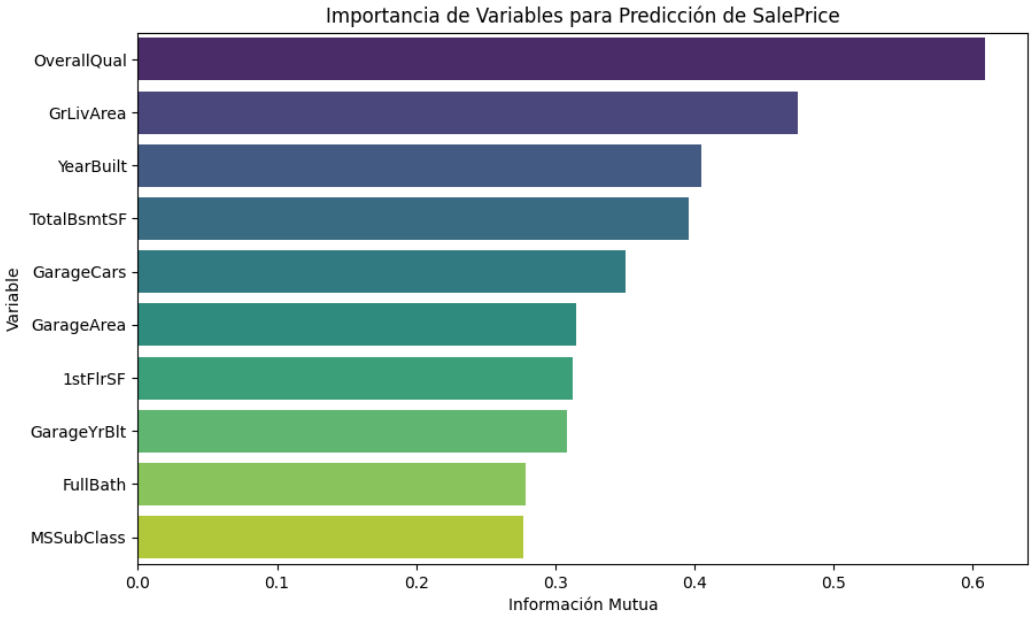


Figura 7. Selección de variables de importancia

3. Ingeniería de Características

Con base en el análisis exploratorio, se procedió a la creación de nuevas variables y selección de aquellas que más influencia tenían en el precio de las casas.

Se creó la variable categórica PriceCategory, que clasifica las casas en tres rangos: Económica, Intermedia y Cara, según los percentiles 33 y 67 de la distribución de SalePrice.

Se seleccionaron como variables predictoras principales aquellas que mostraron mayor correlación con el precio, tales como: OverallQual, GrLivArea, GarageCars, TotalBsmtSF, YearBuilt y FullBath.

Se aplicó codificación numérica a algunas variables categóricas si se usaban en los modelos (como Neighborhood o ExterQual, en pruebas específicas).

No se realizó normalización o estandarización para modelos basados en árboles, pero sí para algoritmos sensibles a la escala como KNN.

Estas transformaciones permitieron entrenar modelos con información más estructurada y relevante, mejorando su capacidad predictiva y reduciendo el ruido de variables poco informativas.

4. Modelos de Regresión

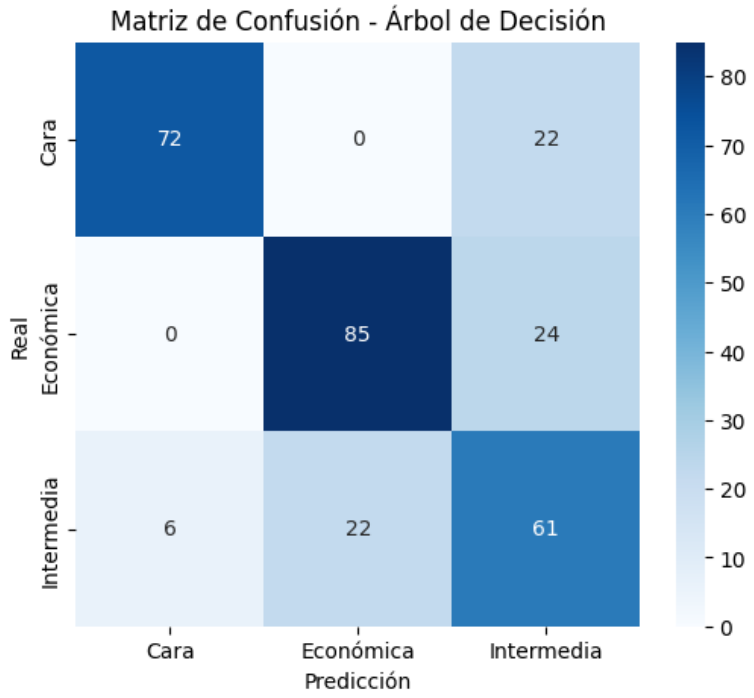
Se entrenaron distintos modelos de regresión para predecir el precio de venta (SalePrice) de las casas a partir de las variables seleccionadas. Los modelos implementados fueron:

Regresión Lineal

Modelo base que mostró buen rendimiento general (RMSE: 39,710.99 – R^2: 0.79).

Árbol de Decisión

Modelo no lineal que logró un mejor ajuste con menor error (RMSE: 37,056.55 – R^2: 0.82).



Reporte de Clasificación (Árbol de Decisión):

	precision	recall	f1-score	support
Cara	0.92	0.77	0.84	94
Económica	0.79	0.78	0.79	109
Intermedia	0.57	0.69	0.62	89
accuracy			0.75	292
macro avg	0.76	0.74	0.75	292
weighted avg	0.77	0.75	0.75	292

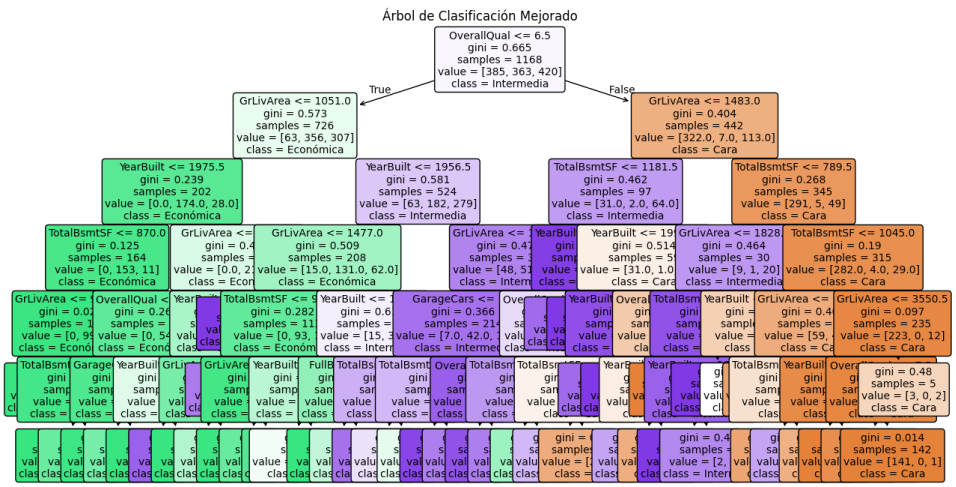


Figura 8. Árbol de decisión y resultados obtenidos

Random Forest

Modelo ensamblado que fue el más preciso (RMSE: 30,287.63 – R^2: 0.88).

Naive Bayes

No funcionó bien para regresión (RMSE: 53,380.67 – R^2: 0.63), debido a suposiciones de normalidad poco realistas.

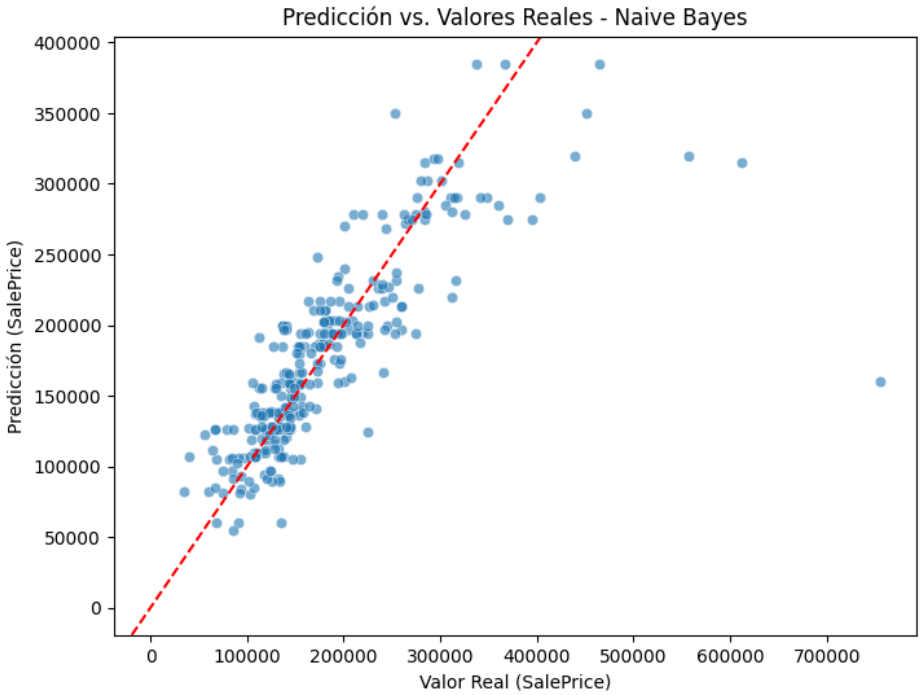


Figura 9. Resultados obtenidos del modelo Naive Bayes

KNN (K-Nearest Neighbors)

Con k=5, el rendimiento fue aceptable (RMSE: 46,838.75 – R^2: 0.71). Al optimizar hiperparámetros, se mejoró a (RMSE: 43,166.03 – R^2: 0.76).

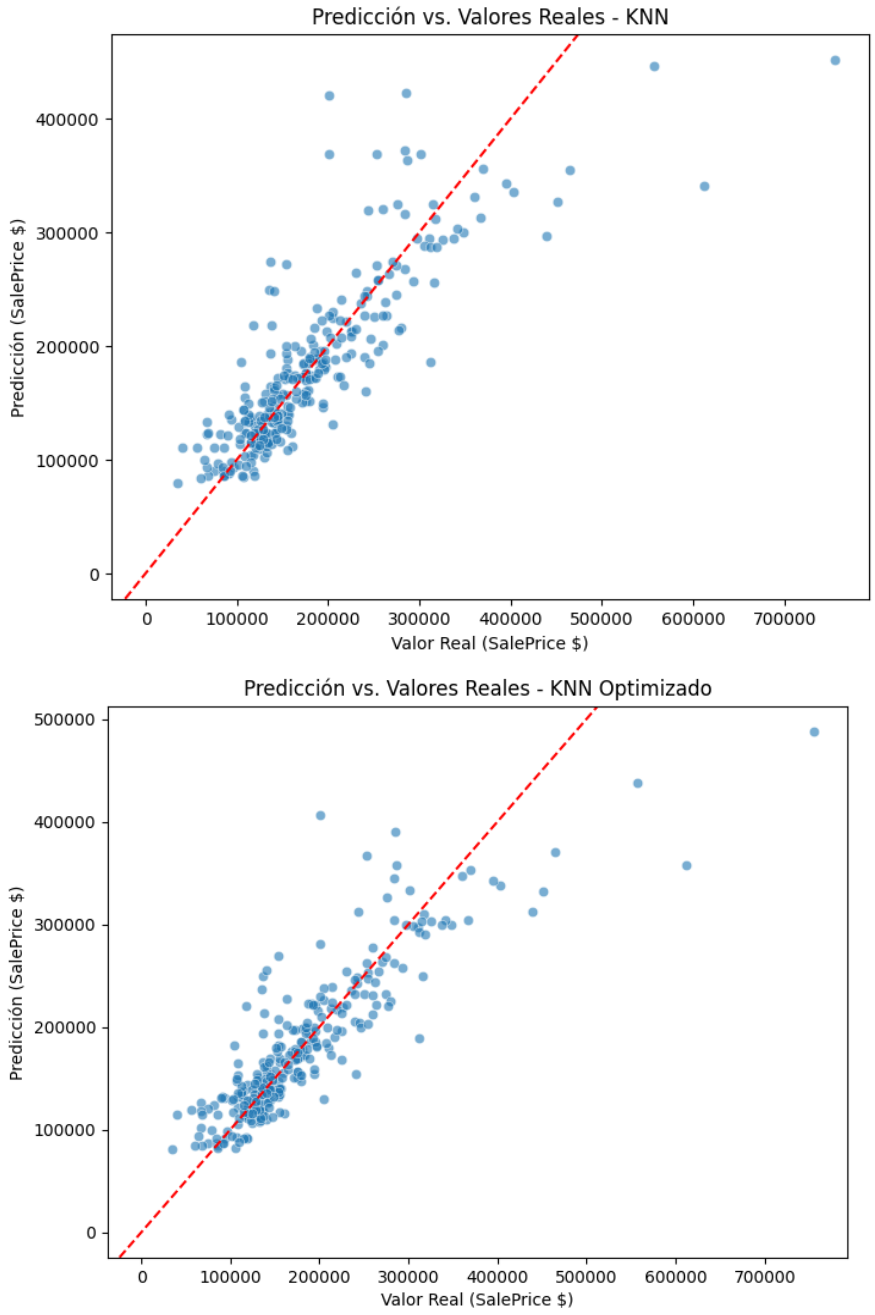


Figura 10. Resultados obtenidos del modelo

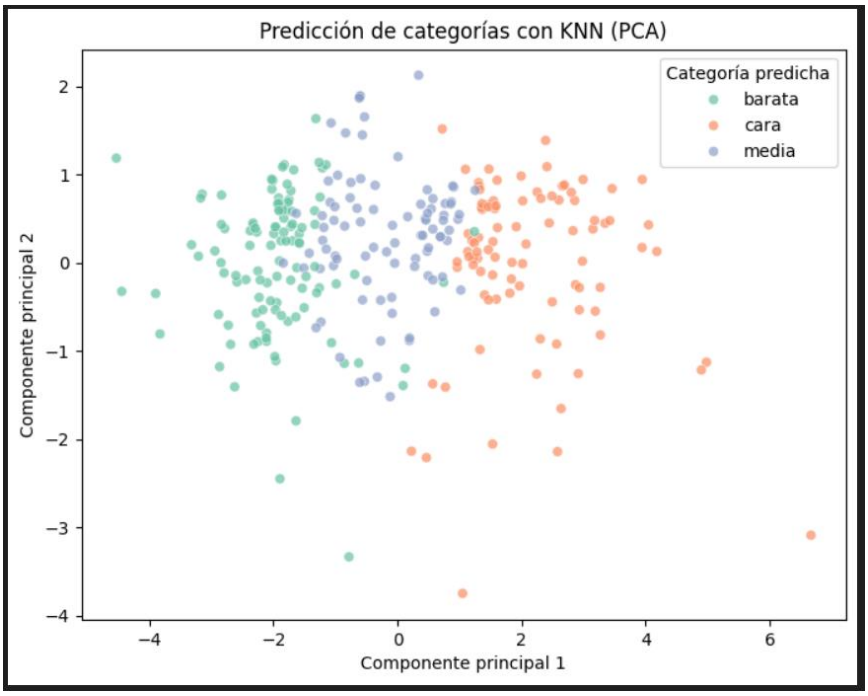


Figura No. 11. Predicción de categorías según KNN.

Métricas de Desempeño

Los modelos fueron evaluados mediante métricas estándar:

RMSE (Root Mean Squared Error)

Mide el error promedio de las predicciones.

R^2 Score

Indica la proporción de la varianza del precio que es explicada por el modelo.

Se utilizaron gráficos de dispersión para comparar valores reales vs. predichos, y gráficos de barras para visualizar comparativamente los resultados de RMSE y R^2 entre modelos.

Comparación entre modelos

Se concluyó que Random Forest fue el modelo más robusto y preciso, superando a los demás tanto en error como en capacidad de generalización.

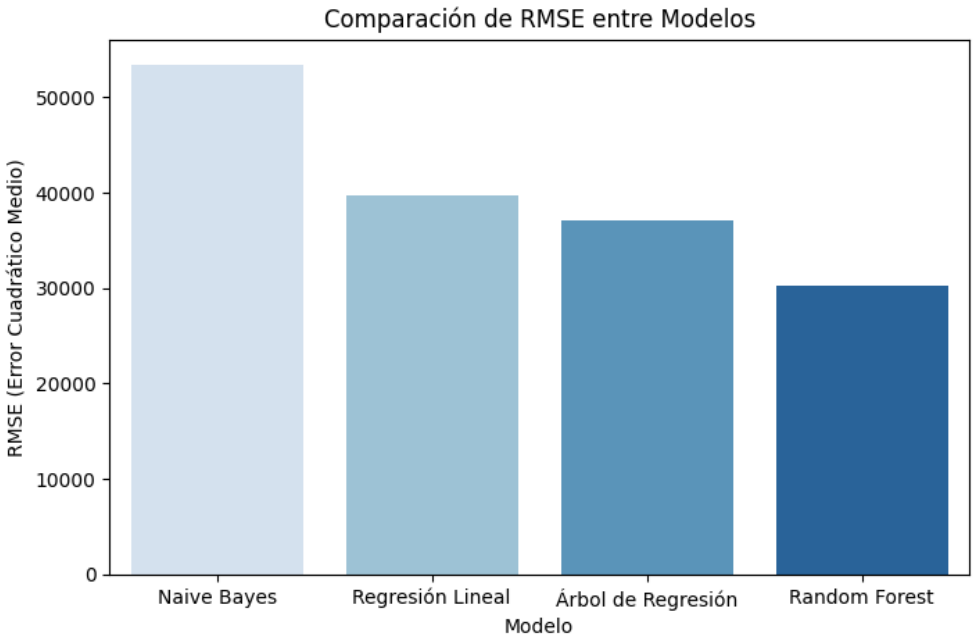


Figura No. 12 Comparación RMSE entre modelos

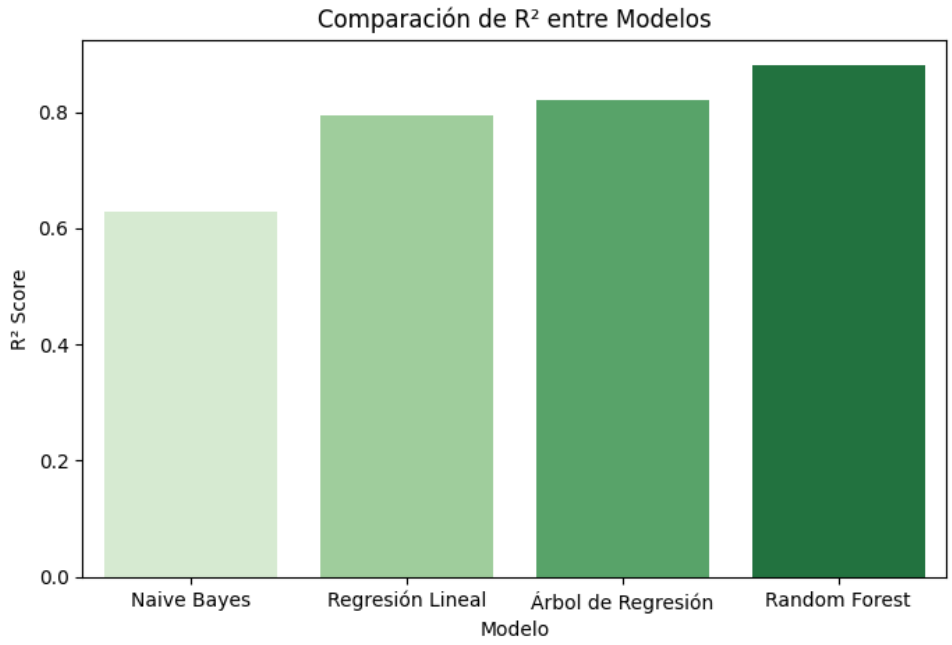


Figura No. 13 Comparación del error al cuadrado.

Comparación de Modelos de Regresión

Modelo	RMSE ↓ (Menor es mejor)	R^2 ↑ (Mayor es mejor)
Naive Bayes	53,380.67	0.63
Regresión Lineal	39,710.99	0.79
Árbol de Regresión	37,056.55	0.82
KNN (k=5)	46,838.75	0.71
Random Forest	30,287.63	0.88

Figura No. 14. Tabla comparativa de los modelos de regresión.

Modelo	Accuracy	F1 Macro	Clase con mayor error
Árbol de decision	0.74	0.73	Intermedia
Random Forest	0.82	0.81	Intermedia
Naive Bayes	0.63	0.60	Todas (más disperso)
KNN	0.82	0.81	Intermedia

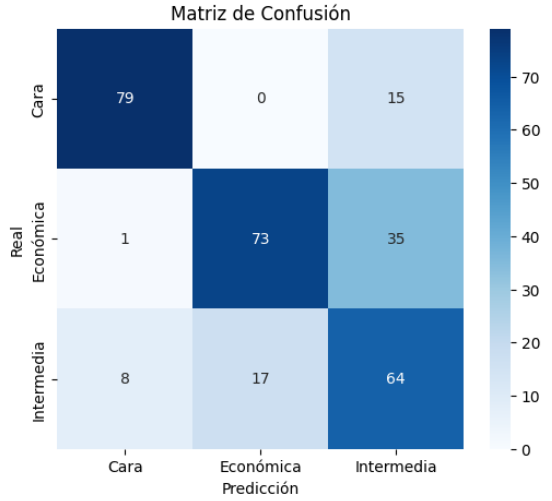
Tabla No. 1. Comparación de desempeño en modelos de clasificación para la variable categórica del precio de vivienda

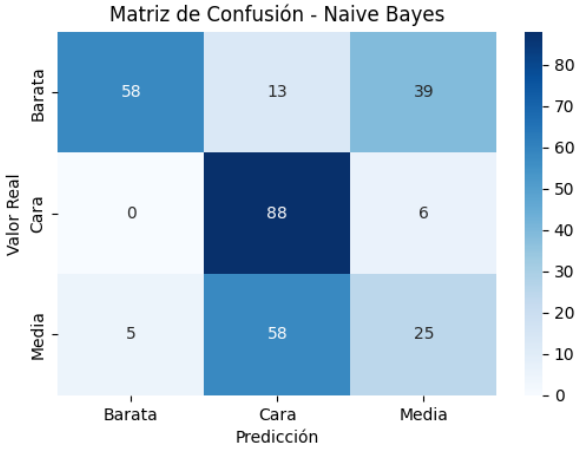
5. Modelos de Clasificación

Además del análisis de regresión, se construyeron modelos de clasificación utilizando la variable PriceCategory (Económica, Intermedia, Cara) como variable objetivo. Se entrenaron los siguientes modelos:

- Árbol de Clasificación: Buen rendimiento general, especialmente en clases extremas (económica y cara). Precisión promedio del 74%.
- Random Forest: Mayor precisión general (82%). Funcionó mejor que todos los demás en todas las clases.
- Naive Bayes: Resultados débiles, con precisión global del 63% y muchas confusiones entre clases.
- KNN (Clasificación): Desempeño aceptable. Con hiperparámetros optimizados (k=11, manhattan, weights='distance'), se alcanzó una precisión del 79%.

Se utilizó una matriz de confusión para evaluar los errores de predicción de cada modelo, destacando que la clase más difícil de predecir fue la Intermedia debido a su cercanía con los límites de las otras dos clases.





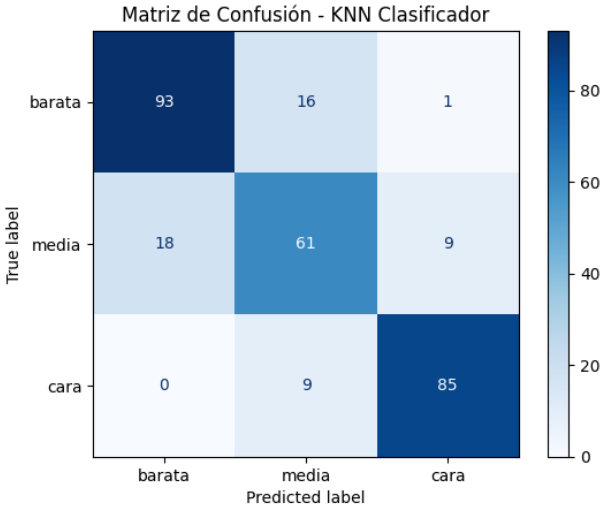


Figura 11: Matrices de confusión

Se emplearon las siguientes métricas para evaluar los modelos:

- Precisión (precision): proporción de aciertos entre todas las predicciones positivas.
- Recall: proporción de verdaderos positivos entre los reales.
- F1-Score: media armónica entre precisión y recall.

El modelo de clasificación con Random Forest fue el que mejor desempeño mostró tanto en métricas individuales como en precisión general. La optimización de hiperparámetros en KNN también mejoró significativamente sus resultados.

6. Optimización de Modelos

Para mejorar el desempeño de algunos modelos, especialmente KNN, se implementó una búsqueda de hiperparámetros mediante GridSearchCV, evaluando combinaciones de:

- Número de vecinos (k)
- Tipo de distancia (euclidean, manhattan, minkowski)
- Peso de los vecinos (uniform, distance)

Resultados destacados:

- KNN Regresión: Mejor configuración con k=7, manhattan, uniform. Mejora del RMSE a 43,166.03 y R^2 a 0.76.
- KNN Clasificación: Mejor configuración con k=11, manhattan, distance. Precisión final del 79%.

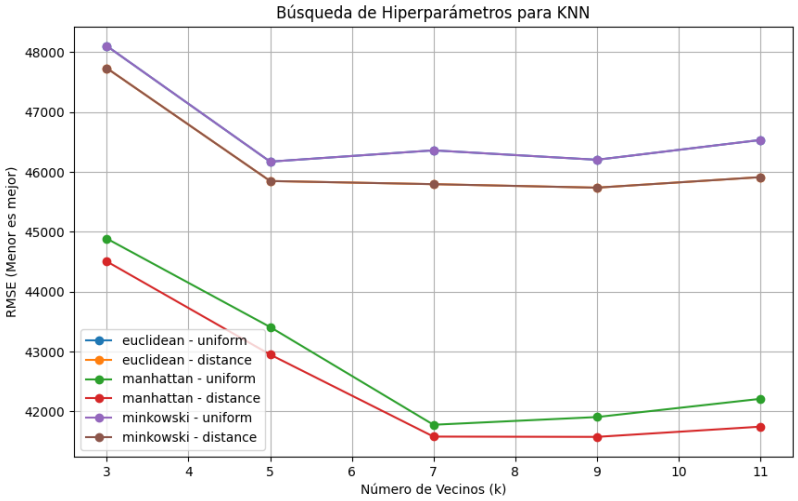


Figura 12: Hiperparámetros para KNN

Estas optimizaciones ayudaron a mejorar el ajuste y desempeño del modelo KNN en ambas tareas.

Otros modelos como Random Forest y Árboles de Decisión también fueron evaluados con distintos parámetros (profundidad máxima, número de árboles), aunque en general su desempeño ya era alto desde la configuración inicial.

Las gráficas de comparación antes y después del ajuste mostraron una reducción en el error y mayor capacidad de generalización para los modelos optimizados.

7. Conclusiones y Recomendaciones

Este proyecto permitió analizar en profundidad los factores que más influyen en el precio de las viviendas, así como evaluar diversos modelos de regresión y clasificación. A través del análisis exploratorio se identificaron variables clave como OverallQual, GrLivArea y GarageCars, las cuales fueron fundamentales para construir modelos predictivos sólidos.

Los resultados mostraron que:

- Random Forest fue el modelo más preciso y robusto tanto para regresión como clasificación.
- KNN mejoró significativamente con optimización, pero aún no superó a Random Forest.
- Naive Bayes no es adecuado para este tipo de tareas debido a suposiciones restrictivas.

En términos de clasificación, la variable PriceCategory permitió categorizar de forma útil los precios y evaluar modelos desde una perspectiva distinta. El mayor reto fue predecir correctamente las casas de clase intermedia.

Se recomienda para futuras versiones:

- Incluir más variables contextuales (ubicación geográfica, tipo de vecindario).
- Probar técnicas de reducción de dimensionalidad o selección de características más avanzadas.
- Implementar validación cruzada más robusta y análisis de errores detallado.
- Este análisis proporciona una base sólida para sistemas de valoración automatizados de propiedades, ayudando a compradores, vendedores y profesionales del sector inmobiliario a tomar decisiones mejor informadas.

8. Referencias

Kaggle. House Prices - Advanced Regression Techniques Dataset. Disponible en: https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques