



Universidad del Valle de Guatemala  
Facultad de Ingeniería  
Minería de Datos

## **HDT 6. K Nearest Neighbors**

Entrega # 4 – Proyecto 2

### **Autor/Carnet:**

María José Villafuerte 22129

Fabiola Contreras 22787

### **Catedrático:**

Mario Barrientos

### **Sección 20**

### **Fecha:**

23/03/2025

## I. Índice

I.	Índice .....	2
II.	Introducción.....	3
III.	Modelo de regresión .....	4
IV.	Comparación con el modelo de regresión lineal y el árbol de regresión.....	5
V.	Modelo de clasificación.....	7
VI.	Análisis de eficiencia.....	7
VII.	Análisis del modelo ¿hay sobreajuste? .....	8
VIII.	Modelo usando validación cruzada .....	9
IX.	Prueba con varios valores de hiperparámetros .....	11
	Modelo de regresión: .....	11
	Modelo de Clasificación.....	12
X.	Comparar la eficiencia del algoritmo con el árbol de decisión, el modelo de random forest y naive bayes .....	12
XI.	Repositorio/ Documento.....	14

## **II. Introducción**

Esta entrega analiza diversos modelos de predicción para el mercado inmobiliario, destacando el rendimiento del algoritmo KNN (K-Nearest Neighbors) tanto en regresión como en clasificación. La investigación compara sistemáticamente el desempeño de KNN con otros modelos como Regresión Lineal, Árbol de Decisión, Random Forest y Naive Bayes, evaluando métricas como RMSE,  $R^2$  y tiempo de ejecución para determinar la mejor opción para predecir precios de propiedades.

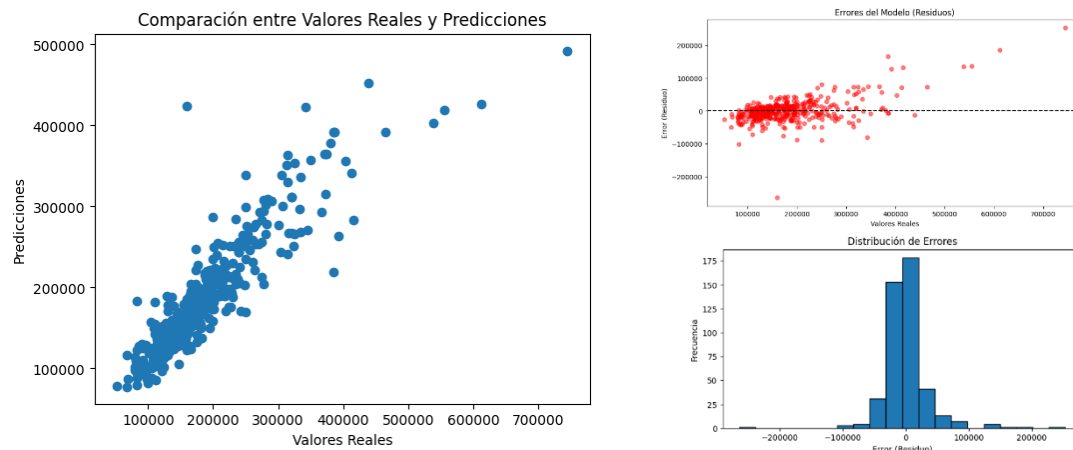
A través de un riguroso proceso de validación cruzada y optimización de hiperparámetros, el documento demuestra cómo KNN ofrece un equilibrio óptimo entre precisión y eficiencia computacional. Los resultados revelan que aunque Random Forest presenta la mayor precisión predictiva, KNN sobresale en velocidad de ejecución, convirtiéndolo en una opción ideal para aplicaciones donde el tiempo de respuesta es crucial. El estudio concluye que el modelo de regresión con validación cruzada supera al modelo de clasificación, proporcionando herramientas más potentes para la valoración inmobiliaria.

### III. Modelo de regresión

Para generar este modelo de regresión, utilizamos las mismas variables que se han usado hasta ahora en los otros modelos. Esto nos permite tener una buena base de comparación, además desde el análisis exploratorio inicial fueron esas variables las que se establecieron como de mayor valor para el cálculo de las predicciones. Estos son los resultados:

RMSE en test: 35559.27098445032

$R^2$  en test: 0.8137482181350053

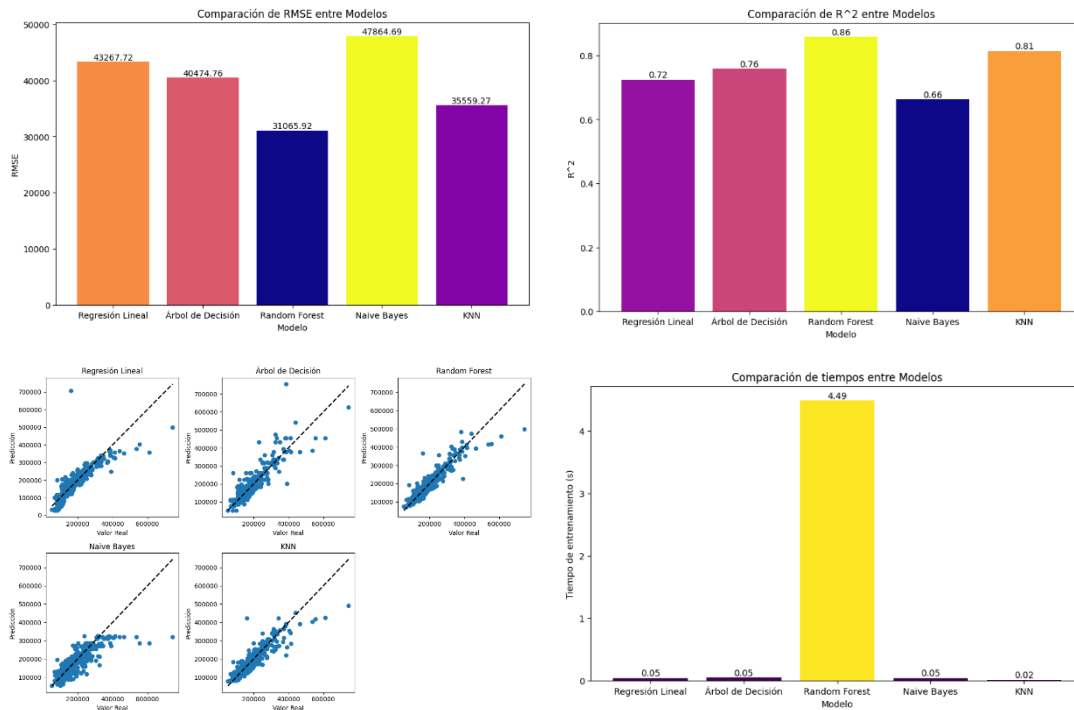


El modelo de regresión KNN obtuvo un RMSE de 35,559.27, lo que indica que, en promedio, las predicciones difieren de los valores reales en aproximadamente 35,559 dólares. Además, el coeficiente de determinación  $R^2$  de 0.814 sugiere que el modelo explica el 81.4% de la variabilidad en los precios de las casas, lo que indica un buen desempeño, aunque con margen de mejora. En el gráfico de dispersión, la mayoría de los puntos siguen la diagonal, lo que demuestra que el modelo hace predicciones razonablemente precisas, aunque se observan algunos valores dispersos que indican errores más grandes en ciertos casos.

El gráfico de distribución de errores muestra una forma aproximadamente normal centrada cerca de cero, lo que es favorable, aunque con algunos valores atípicos significativos en ambos extremos. El gráfico de residuos evidencia un problema de heteroscedasticidad, donde el modelo funciona relativamente bien para propiedades de menor valor (bajo 200,000), pero pierde precisión considerablemente en propiedades más caras, tendiendo a subestimarlas con errores que llegan a superar los 200,000. Para mejorar el rendimiento, se podría ajustar el número de vecinos ( $k$ ), lo cual se trabaja más adelante en este informe.

#### IV. Comparación con el modelo de regresión lineal, el árbol de decisión, el modelo de random forest y naive bayes

En esta ocasión decidimos formar un archivo que cuente con todos los modelos creados para entrenarlos y comparar su rendimiento. Este también servirá para comparación futura. Estos son los resultados de la comparación:



En cuanto a la precisión predictiva, Random Forest destaca con el  $R^2$  más alto (0.8578) y el RMSE más bajo (31,065.92), lo que indica una mayor capacidad para capturar las relaciones complejas entre las variables predictoras y los precios de las viviendas. Observando su gráfico de dispersión, notamos cómo los puntos se distribuyen más uniformemente a lo largo de la línea diagonal ideal, especialmente en comparación con los otros modelos. Este algoritmo mantiene su precisión tanto en propiedades de bajo valor como en las de mayor precio, a diferencia de los demás modelos que muestran mayor dispersión en los extremos superiores.

KNN emerge como el segundo mejor en precisión ( $R^2$  de 0.8137, RMSE de 35,559.27). Su gráfico refleja un patrón similar al de Random Forest, aunque con algo más de dispersión en los valores altos. Sin embargo, el gráfico muestra cierta tendencia a subestimar los valores más elevados, algo que ya había sido identificado en los gráficos de residuos analizados anteriormente.

El Árbol de Decisión ocupa el tercer lugar en precisión ( $R^2$  de 0.7587, RMSE de 40,474.76). Su gráfico de dispersión muestra mayor variabilidad, con algunos casos de sobrestimación notables en el rango medio y alto. Se aprecia la formación de "bandas" horizontales en las predicciones, lo que puede atribuirse a la naturaleza discreta de este tipo de modelo.

La Regresión Lineal, con un  $R^2$  de 0.7242 y RMSE de 43,267.72, muestra un comportamiento interesante. Mientras mantiene buena precisión en el rango inferior y medio, tiende a subestimar significativamente los precios de las propiedades más costosas, como se evidencia en la clara desviación de la línea diagonal en el extremo superior del gráfico. Finalmente, Naive Bayes presenta el peor desempeño predictivo ( $R^2$  de 0.6625, RMSE de 47,864.69). Su gráfico muestra una notable tendencia a comprimir el rango de predicciones, limitándolas aproximadamente entre 50,000 y 320,000, lo que explica su incapacidad para modelar adecuadamente los extremos de la distribución de precios.

En términos de eficiencia computacional, KNN sobresale con un tiempo de ejecución extremadamente rápido (0.0159s), seguido de Naive Bayes (0.0474s), Regresión Lineal (0.0481s) y Árbol de Decisión (0.0542s). En marcado contraste, Random Forest requiere un tiempo de procesamiento significativamente mayor (4.486s), aproximadamente 282 veces más lento que KNN, reflejando la complejidad inherente al entrenamiento de múltiples árboles.

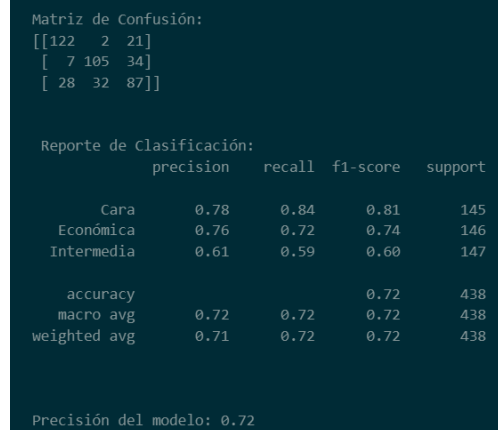
Considerando todo esto, podemos concluir que Random Forest ofrece la mayor precisión predictiva, pero a costa de un tiempo de ejecución considerablemente mayor. KNN representa un excelente equilibrio entre precisión y eficiencia, haciéndolo particularmente atractivo para aplicaciones donde el tiempo de respuesta es crucial. Para casos donde se priorice la interpretabilidad del modelo junto con un buen nivel de precisión, el Árbol de Decisión podría ser una opción razonable. La Regresión Lineal ofrece simplicidad y rapidez, pero con limitaciones claras en la modelación de relaciones no lineales, especialmente en el extremo superior del rango de precios. Naive Bayes, aunque computacionalmente eficiente, presenta limitaciones significativas en su capacidad predictiva para este conjunto de datos específico (como lo vimos en la entrega y análisis anterior).

## V. Modelo de clasificación

El análisis mediante el algoritmo K-Nearest Neighbors (KNN) se realizó para desarrollar un modelo de clasificación capaz de categorizar propiedades inmobiliarias según su rango de precio, alineándose con los objetivos estratégicos de InmoValor S.A. Utilizamos una segmentación en terciles que clasifica las viviendas como "Económica", "Intermedia" y "Cara", estableciendo umbrales en \$139,699 y \$190,000 respectivamente. Este enfoque permite a la empresa identificar patrones de valoración basados en características fundamentales de las propiedades, como la calidad general, área habitable, capacidad del garaje, año de construcción, entre otras variables clave. La implementación de KNN resulta particularmente valiosa por su capacidad de capturar relaciones no lineales entre las características y los precios, así como su adaptabilidad para clasificar nuevas propiedades sin necesidad de reentrenar completamente el modelo.

## VI. Análisis de eficiencia

Los resultados del modelo KNN reflejados en la matriz de confusión y métricas de evaluación muestran un rendimiento notable con una precisión global del 72%. El análisis detallado de la matriz de confusión [122, 2, 21; 7, 105, 34; 28, 32, 87] revela que el modelo es particularmente efectivo para identificar propiedades en los extremos del espectro de precios, con 122 predicciones correctas para viviendas económicas y una alta precisión (0.78) para las propiedades caras. Las métricas por categoría indican un desempeño superior en la clasificación de propiedades "Caras" (f1-score: 0.81) y "Económicas" (f1-score: 0.74), mientras que la categoría "Intermedia" presenta mayor dificultad (f1-score: 0.60). Esta disminución en el rendimiento para la categoría intermedia es esperada, ya que representa una zona de transición donde las características distintivas son menos evidentes. El resultado más preocupante se encuentra en la confusión de 34 propiedades intermedias clasificadas como caras, indicando una tendencia del modelo a sobrevalorar ciertas propiedades, lo que podría requerir ajustes adicionales.



```
Matriz de Confusión:
[[122  2  21]
 [  7 105  34]
 [ 28  32  87]]

Reporte de Clasificación:
              precision    recall  f1-score   support

    Cara         0.78        0.84        0.81        145
  Económica      0.76        0.72        0.74        146
  Intermedia     0.61        0.59        0.60        147

 accuracy         0.72
 macro avg        0.72
weighted avg        0.71

Precisión del modelo: 0.72
```

El modelo se equivocó más significativamente en dos áreas principales:

1. La clasificación errónea de 34 propiedades "Intermedias" como "Caras" representa el error más frecuente. Este tipo de error es particularmente importante porque implica una sobrevaloración de propiedades, lo que podría llevar a InmoValor S.A. a establecer expectativas de precio demasiado altas, resultando potencialmente en propiedades que permanecen más tiempo en el mercado o negociaciones fallidas con clientes.

2. El segundo error más común es la clasificación de 32 propiedades "Caras" como "Intermedias", seguido por 28 propiedades "Caras" clasificadas erróneamente como "Económicas". Estos errores de subvaloración son igualmente problemáticos, ya que podrían ocasionar pérdidas financieras significativas al vender propiedades por debajo de su valor real de mercado.

## **VII. Análisis del modelo ¿hay sobreajuste? No**

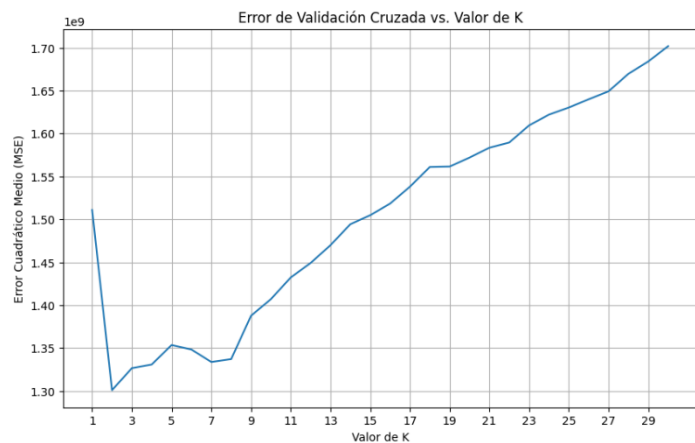
Analizando el modelo KNN implementado para la clasificación de propiedades inmobiliarias, no se observan signos claros de sobreajuste. La precisión general del 72% representa un equilibrio razonable entre el ajuste a los datos de entrenamiento y la capacidad de generalización. El rendimiento relativamente uniforme a través de las diferentes métricas (precision, recall y f1-score) para cada categoría, así como la distribución bastante equilibrada de errores en la matriz de confusión, sugieren que el modelo no está memorizando los datos de entrenamiento. La categoría "Intermedia" presenta un rendimiento más bajo (f1-score de 0.60), pero esto se debe probablemente a la inherente dificultad de clasificar propiedades en este rango transitorio más que a problemas de sobreajuste.



## VIII. Modelo usando validación cruzada

El análisis mediante validación cruzada se implementó para evaluar y optimizar el rendimiento del modelo KNN de regresión aplicado a la predicción de precios inmobiliarios. Esta técnica divide sistemáticamente los datos en múltiples subconjuntos (5 folds en este caso), permitiendo entrenar y validar el modelo en diferentes combinaciones de datos, lo que proporciona una evaluación más robusta y realista de su capacidad predictiva. El objetivo fundamental de aplicar validación cruzada fue identificar la configuración óptima de hiperparámetros que minimizara el error de predicción, evitando tanto el sobreajuste como el subajuste del modelo. Adicionalmente, se implementó GridSearchCV para explorar exhaustivamente diferentes combinaciones de parámetros clave: número de vecinos, sistema de ponderación y métrica de distancia. Este enfoque metodológico riguroso garantiza que el modelo desarrollado no solo tenga buen rendimiento en los datos de entrenamiento, sino que también generalice efectivamente a nuevas propiedades inmobiliarias no vistas previamente.

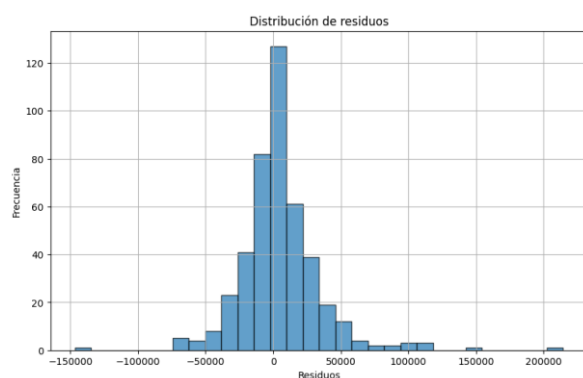
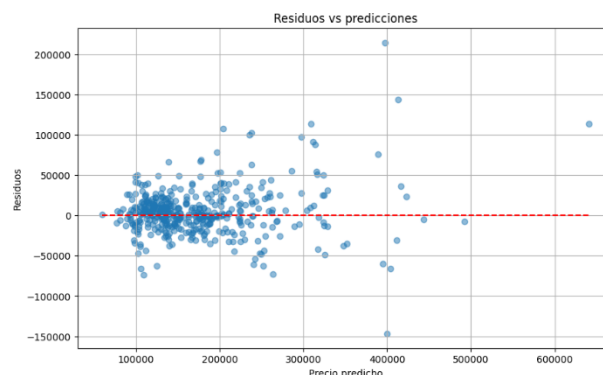
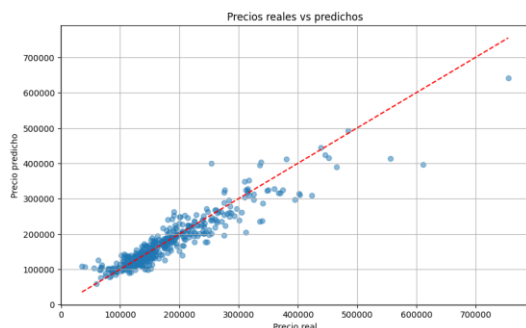
Los resultados de la validación cruzada revelan patrones significativos sobre el comportamiento del modelo KNN. La gráfica de error vs. valor de K muestra claramente que el error de predicción se minimiza con valores pequeños de K (entre 2 y 4), con un incremento progresivo y constante a medida que aumenta el número de vecinos considerados. Esta tendencia indica que la información más relevante para la predicción de precios proviene de propiedades muy similares, sugiriendo una alta especificidad en los factores que determinan el valor inmobiliario. La búsqueda mediante GridSearchCV identificó como configuración óptima: 4 vecinos, ponderación por distancia (asignando mayor importancia a propiedades más similares) y métrica Manhattan ( $p=1$ ), confirmando la importancia de las propiedades cercanas en el espacio de características. Los resultados de la validación cruzada muestran un RMSE promedio de \$32,747 con una desviación estándar relativamente baja de \$2,426, lo que indica consistencia



```
Mejores parámetros encontrados:  
{ 'n_neighbors': 4, 'p': 1, 'weights': 'distance' }
```

```
Resultados de la validación cruzada (RMSE):  
Fold 1: $30,445  
Fold 2: $35,179  
Fold 3: $34,596  
Fold 4: $29,223  
Fold 5: $34,290  
RMSE promedio: $32,747  
Desviación estándar: $2,426  
  
Resultados en el conjunto de prueba:  
Error cuadrático medio (MSE): $932,138,290  
Raíz del error cuadrático medio (RMSE): $30,530  
Error absoluto medio (MAE): $19,854  
Coeficiente de determinación (R²): 0.8664
```

del modelo a través de diferentes subconjuntos de datos. Esta estabilidad se refleja también en los valores de RMSE para cada fold (entre \$29,223 y \$35,179), demostrando que el modelo mantiene un rendimiento similar independientemente de la partición de datos utilizada para su evaluación.



### ¿Cuál funcionó mejor? Validación Cruzada

Comparando los resultados de ambos modelos, el modelo de regresión con validación cruzada demuestra un rendimiento superior al modelo de clasificación para la predicción de precios inmobiliarios. Mientras que el modelo de clasificación alcanzó una precisión general de 72%, categorizando correctamente las propiedades en tres grupos de precios, el modelo de regresión con validación cruzada logró un coeficiente de determinación ( $R^2$ ) de 0.8664, explicando aproximadamente el 87% de la variabilidad en los precios exactos. Esta diferencia es significativa porque el modelo de regresión no solo identifica la categoría de precio, sino que proporciona estimaciones numéricas precisas con un error absoluto medio de solo \$19,854, lo que representa un nivel de granularidad y precisión considerablemente mayor. Además, la metodología de validación cruzada implementada en el segundo modelo garantiza su robustez y capacidad de generalización, como lo demuestra la baja desviación estándar (\$2,426) entre los diferentes folds, asegurando que las predicciones sean confiables independientemente del subconjunto de datos utilizado. Para InmoValor S.A., aunque ambos modelos ofrecen valor, el modelo de regresión con validación cruzada proporciona herramientas de decisión significativamente más potentes y precisas para la valoración inmobiliaria, representando la opción óptima para implementación en entornos de producción.

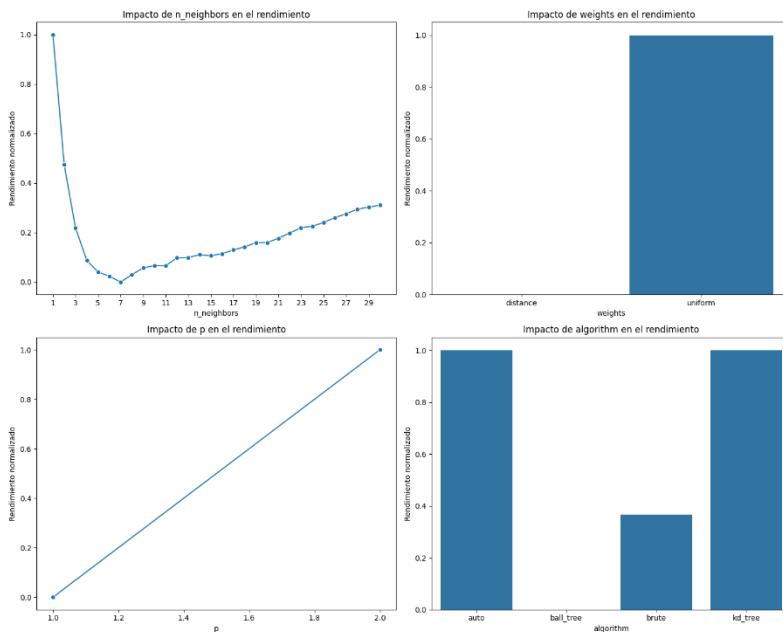
## IX. Prueba con varios valores de hiperparámetros

Los principales hiperparámetros que pueden ajustarse en un KNN son:

- `n_neighbors`: El número de vecinos ( $k$ ) a considerar
- `weights`: Cómo se ponderan los vecinos ('uniform' o 'distance')
- `p`: La potencia del parámetro para la métrica de Minkowski ( $p=1$  para Manhattan,  $p=2$  para Euclidiana)
- `metric`: Métrica de distancia usada
- `algorithm`: Algoritmo usado para calcular vecinos más cercanos
- `leaf_size`: Tamaño de hoja para árboles BallTree o KDTree

### Modelo de regresión:

La optimización de hiperparámetros del modelo KNN de regresión se realizó para mejorar la precisión predictiva en la estimación de precios de viviendas. Mediante el ajuste de los parámetros del modelo (número de vecinos, métrica de distancia, esquema de ponderación y algoritmo de implementación).



Las visualizaciones de impacto de hiperparámetros revelan patrones significativos para la optimización del modelo KNN. La gráfica de `n_neighbors` muestra un rendimiento óptimo con muy pocos vecinos (1-3) seguido de una caída dramática hasta aproximadamente 7 vecinos, para luego mejorar gradualmente, sugiriendo que las propiedades inmediatas proporcionan la señal más fuerte, pero incorporar más propiedades eventualmente captura tendencias más amplias del

vecindario. Aunque las gráficas individuales sugieren que la ponderación 'uniform' y la distancia Euclidiana ( $p=2$ ) podrían funcionar mejor aisladamente, la búsqueda en cuadrícula determinó que la combinación óptima utiliza ponderación por 'distance' con distancia Manhattan ( $p=1$ ) cuando se combina con el algoritmo 'ball\_tree' y 7 vecinos. Esto destaca la compleja interacción entre parámetros, donde la combinación óptima no necesariamente consiste en configuraciones individualmente óptimas. La comparación de algoritmos demuestra que estructuras de datos especializadas como 'ball\_tree' y 'kd\_tree' superan

significativamente al enfoque de fuerza bruta ('brute') para búsquedas de vecinos.

```
Mejores parámetros encontrados:
{'algorithm': 'ball_tree', 'n_neighbors': 7, 'p': 1, 'weights': 'distance'}

Resultados con el mejor modelo:
RMSE optimizado: 33534.09399051837
R2 optimizado: 0.834358981073425
```

## Modelo de Clasificación:

La optimización del modelo KNN para clasificación de propiedades inmobiliarias con los parámetros algorithm: 'ball\_tree', n\_neighbors: 12, p: 1 (distancia Manhattan) y weights: 'distance' resultó en una precisión global de 0.75, representando una disminución de 0.3 puntos respecto al modelo anterior. Sin embargo, este cambio refleja un equilibrio más

equitativo entre las categorías de precio, donde las propiedades Caras alcanzaron un recall de 0.84 y precisión de 0.78, las Económicas mostraron la mayor precisión (0.82) con recall de 0.75, mientras que la categoría Intermedia, naturalmente más difícil de distinguir, obtuvo métricas de 0.65 (precisión) y 0.66 (recall). La matriz de confusión evidencia una distribución más equilibrada de errores, indicando que el modelo actual, aunque ligeramente menos preciso en términos globales, ofrece predicciones más confiables en todo el espectro de precios, lo que resulta más valioso para decisiones empresariales que abarcan diferentes segmentos del mercado inmobiliario.

```
Mejores parámetros encontrados:
{'algorithm': 'ball_tree', 'n_neighbors': 12, 'p': 1, 'weights': 'distance'}

Resultados con el mejor modelo:
Matriz de Confusión:
[[122  1  22]
 [ 7 109  30]
 [ 27  23  97]]

Reporte de Clasificación:
      precision    recall  f1-score   support

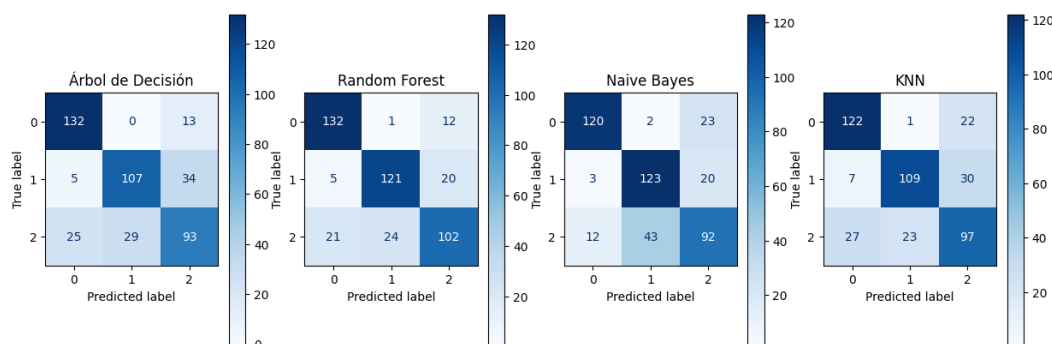
    Cara      0.78      0.84      0.81      145
  Económica    0.82      0.75      0.78      146
  Intermedia    0.65      0.66      0.66      147

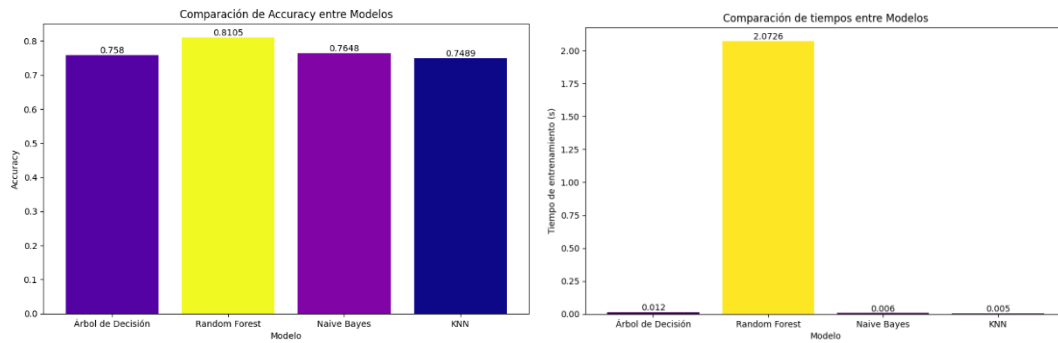
 accuracy      0.75      0.75      0.75      438
  macro avg      0.75      0.75      0.75      438
  weighted avg      0.75      0.75      0.75      438

Precisión del modelo optimizado: 0.75
```

## X. Comparar la eficiencia del algoritmo con el árbol de decisión, el modelo de random forest y naive bayes

Al igual que en la comparación de modelos de regresión en esta ocasión decidimos formar un archivo que cuente con todos los modelos creados para entrenarlos y comparar su rendimiento. Estos son los resultados de la comparación:





En términos de precisión (accuracy), Random Forest destaca notablemente con un 81.05%, superando a sus competidores por un margen considerable. Al examinar su matriz de confusión, se observa un excelente rendimiento en las tres clases, con 132 aciertos en la clase 0, 121 en la clase 1 y 102 en la clase 2. Se resalta su capacidad para minimizar los falsos positivos y falsos negativos en todas las categorías, logrando un equilibrio óptimo en la clasificación.

Naive Bayes ocupa el segundo lugar con una precisión de 76.48%, mostrando un comportamiento interesante en su matriz de confusión. Si bien clasifica correctamente 120 instancias de la clase 0, 123 de la clase 1 y 92 de la clase 2, presenta cierta confusión entre las clases 0 y 2, con 23 instancias de la clase 0 clasificadas erróneamente como clase 2.

El Árbol de Decisión, con una precisión de 75.80%, presenta un rendimiento sólido, pero con debilidades específicas. Su matriz de confusión revela una tendencia a confundir las clases 1 y 2, con 34 instancias de la clase 1 clasificadas incorrectamente como clase 2, y 29 instancias de la clase 2 clasificadas como clase 1. Sin embargo, muestra un excelente desempeño en la clase 0, con 132 clasificaciones correctas.

KNN presenta el accuracy más bajo (74.89%), aunque la diferencia con los dos modelos anteriores no es dramática. Su matriz de confusión muestra un patrón similar al del Árbol de Decisión, con dificultades en la distinción entre las clases 1 y 2. Identifica correctamente 122 instancias de la clase 0, 109 de la clase 1 y 97 de la clase 2, pero confunde 30 instancias de la clase 1 como clase 2.

En cuanto a tiempos de ejecución, KNN sobresale como el modelo más rápido (0.0050s), seguido muy de cerca por Naive Bayes (0.0060s). El Árbol de Decisión requiere aproximadamente el doble de tiempo (0.0120s), mientras que Random Forest, consistentemente con su mayor complejidad, demanda un tiempo significativamente mayor (2.0726s), siendo aproximadamente 414 veces más lento que KNN.

La comparación muestra que Random Forest ofrece la mejor precisión en todas las clases, pero a costa de un tiempo de entrenamiento sustancialmente mayor. Naive Bayes proporciona un excelente balance entre precisión y eficiencia, haciéndolo particularmente atractivo para aplicaciones donde el tiempo es un factor crítico, pero no se quiere sacrificar demasiada precisión. El Árbol de Decisión y KNN, aunque ligeramente menos precisos, mantienen un rendimiento competitivo con tiempos de ejecución razonables.

Para aplicaciones donde la precisión es primordial y los recursos computacionales no son una limitación, Random Forest se ve como la opción óptima. Sin embargo, para implementaciones en tiempo real o con restricciones de recursos, Naive Bayes o KNN podrían representar alternativas más adecuadas, ofreciendo un buen equilibrio entre velocidad y precisión.

## **XI. Repositorio/ Documento**

[https://github.com/Fabiola-cc/InmoValor\\_SA](https://github.com/Fabiola-cc/InmoValor_SA)

- Hemos dividido el repositorio en branches, cada una de esta tiene una de las entregas, en el repositorio también se encuentran los datos y el pdf con este informe.

[HDT 6. KNN.docx](#)