Universidad del Valle de Guatemala
Facultad de Ingeniería
Minería de Datos

HDT 5. Bayes Ingenuo

Entrega # 3 – Proyecto 2

Autor/Carnet:

María José Villafuerte 22129
Fabiola Contreras 22787

Catedrático:

Mario Barrientos

Sección 20

Fecha:

16/03/2025

I. Índice

II. Introducción

El presente trabajo tiene como objetivo analizar y evaluar el rendimiento del algoritmo de Naive Bayes en problemas de regresión y clasificación aplicados al mercado inmobiliario. Como parte del curso de Minería de Datos, esta investigación busca determinar la eficacia de este modelo probabilístico para predecir y categorizar precios de viviendas, comparándolo con técnicas previamente estudiadas como la regresión lineal, árboles de decisión y Random Forest.

El algoritmo de Naive Bayes, fundamentado en el teorema de Bayes, presenta una característica distintiva al asumir independencia entre las variables predictoras, lo que simplifica significativamente los cálculos pero puede afectar su rendimiento en datasets donde existen fuertes correlaciones entre variables. A lo largo de este informe, exploraremos si esta limitación teórica impacta negativamente en nuestro caso de uso específico o si, por el contrario, el modelo logra capturar eficientemente los patrones presentes en los datos inmobiliarios.

Para una evaluación integral, implementaremos tanto un modelo de regresión para predecir valores exactos de las propiedades, como un modelo de clasificación que categoriza las viviendas en segmentos de precio (Económica, Intermedia, Cara). Adicionalmente, exploraremos técnicas de optimización como validación cruzada y ajuste de hiperparámetros para mejorar el rendimiento del algoritmo, analizando en cada caso su precisión, eficiencia computacional y capacidad de generalización.

Los resultados de este análisis proporcionan información valiosa para determinar si es apropiado utilizar Naive Bayes en el contexto de valoración inmobiliaria, contribuyendo así a la toma de decisiones informadas en el análisis de los datos.

III. Modelo de regresión

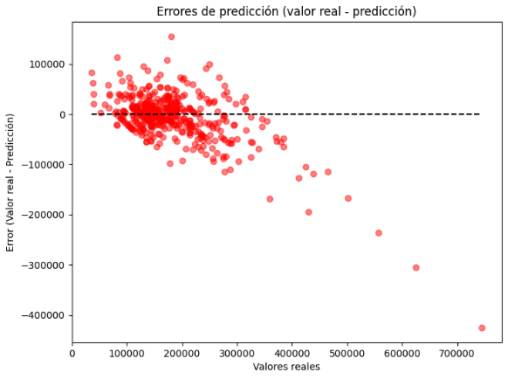
Para este modelo se emplearon las mismas variables de modelos anteriores (de regresión lineal y árbol de decisión). Según el análisis realizado al inicio de este proyecto se determinó que esas eran las variables más relevantes y sus resultados lo comprobaron. Estos son los resultados del modelo con Naive Bayes:

R^2: 0.6356

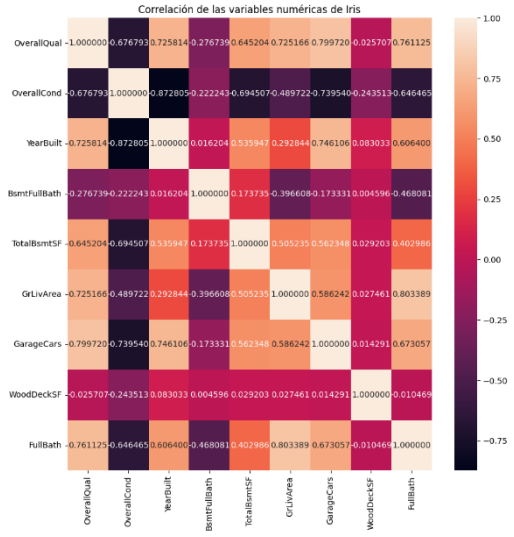
MAE: 31814.5548

RMSE: 49301.6193

El modelo de Naive Bayes aplicado a regresión muestra un desempeño considerablemente deficiente. Con un R^2 de 0.6356, la capacidad explicativa del modelo es limitada, indicando que solo logra explicar alrededor del 63.56% de la varianza en los datos.



Además, el MAE de 31,814 y el RMSE de 49,301 evidencian errores de predicción elevados, lo que indica que el modelo tiene dificultades para hacer predicciones precisas. El gráfico de errores refuerza esto, mostrando una dispersión notable de puntos a medida que los valores reales aumentan, lo que sugiere un problema de sesgo. La presencia de valores negativos y positivos a lo largo del eje y = 0 (línea de error cero) también sugiere una posible falta de capacidad del modelo para capturar patrones en los datos de manera precisa.



El mapa de correlación de las variables utilizadas en el modelo de Naive Bayes revela relaciones significativas entre varias características, lo que puede afectar negativamente su rendimiento. En el gráfico, se observan correlaciones fuertes entre varias variables, como **"OverallQual" y "GrLivArea" (0.72)**, **"YearBuilt" y "OverallCond" (0.87)**, o **"FullBath" y "GrLivArea" (0.80)**. Estas correlaciones indican que las variables están relacionadas entre sí, violando el supuesto de independencia de Naive Bayes. Como resultado, el modelo puede hacer suposiciones incorrectas sobre la distribución de los datos y no capturar correctamente las relaciones entre las variables, lo que explicaría su bajo desempeño en comparación con los otros modelos.

IV. Comparación con el modelo de regresión lineal y el árbol de regresión

El modelo de Regresión Lineal mostró un desempeño moderado, con errores relativamente bajos y una capacidad aceptable para capturar relaciones lineales entre las variables. Aunque es una técnica sencilla y rápida de entrenar, tiene limitaciones cuando los datos presentan relaciones no lineales o interacciones complejas entre variables.

Por otro lado, el modelo de Árbol de Decisión obtuvo un mejor rendimiento que la Regresión Lineal, con errores reducidos y una mayor capacidad para capturar relaciones no lineales en los datos. Sin embargo, los árboles de decisión individuales son susceptibles al sobreajuste, lo que podría afectar su capacidad de generalización.

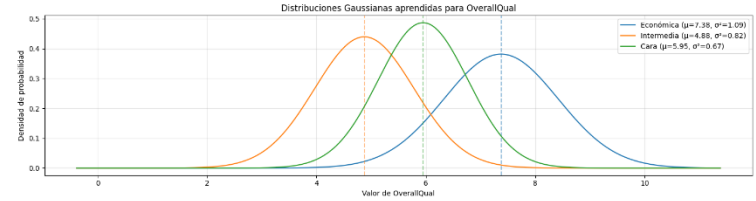
El modelo de Random Forest superó a ambos en términos de precisión, obteniendo menores valores de MAE y RMSE, lo que sugiere que es capaz de capturar de manera más efectiva la compleja relación entre las variables de entrada y la variable objetivo. Esto es esperable, ya que Random Forest combina múltiples árboles de decisión, reduciendo la varianza y logrando un modelo más estable y robusto.

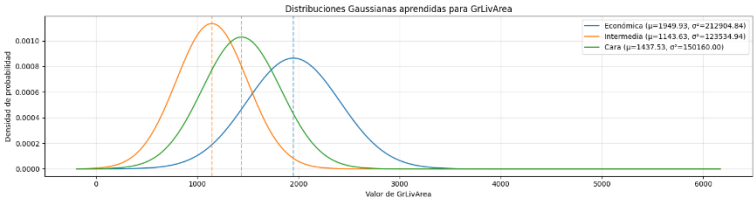
En contraste, el modelo de Naive Bayes tuvo el peor desempeño de los tres. Sus métricas de error más elevadas y la visualización de errores indican que sus predicciones están significativamente desviadas, posiblemente debido a la fuerte suposición de independencia entre las variables predictoras. Como se ve en el heatmap de correlación, hay varias relaciones fuertes entre variables (como TotalBsmtSF con OverallQual y GarageCars), lo que puede violar los supuestos de independencia de Naive Bayes y explicar su menor precisión en esta tarea de regresión.

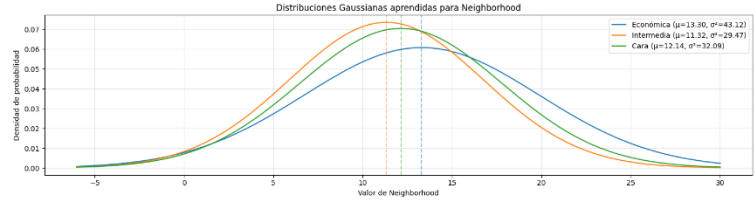
En conclusión, para este conjunto de datos de precios de viviendas, el Random Forest ha demostrado ser la mejor opción, seguido del Árbol de Decisión, mientras que Naive Bayes ha tenido el peor rendimiento. Esto sugiere que los modelos de árboles de decisión son más adecuados para capturar la complejidad de las relaciones entre las variables predictoras y la variable objetivo, especialmente en problemas de regresión con patrones no lineales.

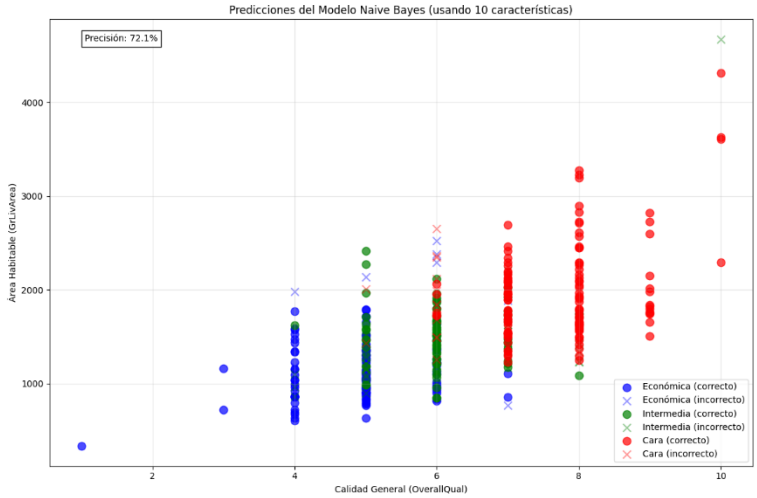
V. Modelo de clasificación

Implementamos un modelo de Naive Bayes Gaussiano para clasificar propiedades inmobiliarias en categorías de precio (Económica, Intermedia, Cara) usando un enfoque probabilístico. Este modelo complementa nuestros análisis anteriores al cuantificar la probabilidad de pertenencia a cada categoría basándose en la distribución normal de características clave, permitiendo evaluar la incertidumbre en las predicciones.









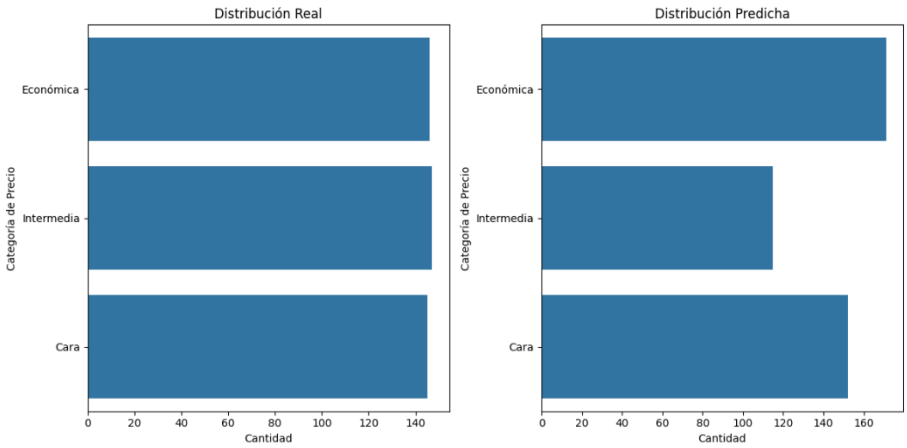
Las gráficas revelan patrones distintivos en la distribución de características por categoría. Para OverallQual, existe una clara progresión entre categorías (medias: 3.38, 4.88, 5.95), confirmándola como fuerte discriminador de precio. En GrLivArea, sorprendentemente, las viviendas económicas muestran mayor superficie media (1949.93 pies²), sugiriendo que el área sola no determina el precio sin considerar otros factores. La visualización de predicciones muestra que propiedades con alta calidad (>7) y mayor superficie tienden a clasificarse como "Caras", mientras la categoría "Intermedia" presenta mayor solapamiento, explicando parte del error del modelo.

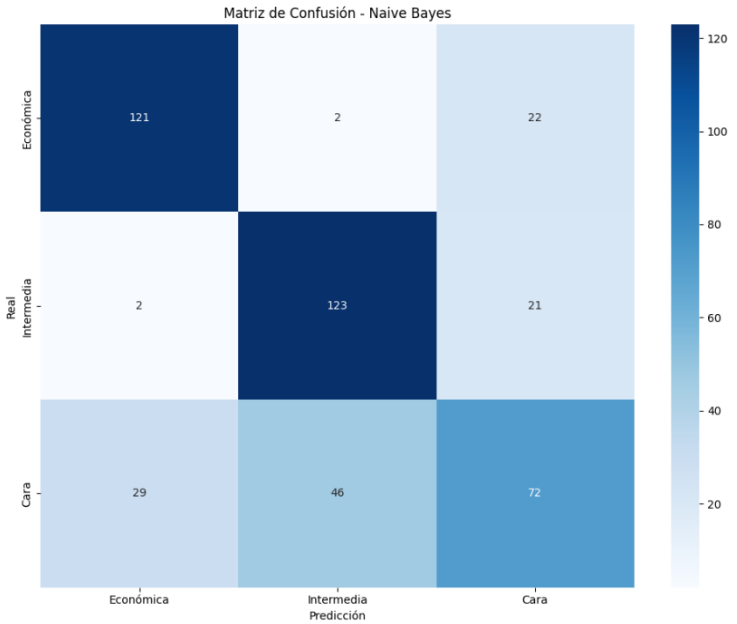
VI. Análisis de eficiencia

Implementamos un modelo de clasificación Naive Bayes para evaluar su eficacia en la predicción de categorías de precio inmobiliario (Económica, Intermedia, Cara) usando datos de prueba.

```
Accuracy: 0.7214611872146118
Precision: 0.7214611872146118
Recall: 0.7214611872146118
F1-score: 0.7214611872146118
```

Las métricas de evaluación (Imagen 1) muestran un rendimiento consistente del modelo Naive Bayes con valores de 72.1% para accuracy, precision, recall y F1-score, indicando un equilibrio entre su capacidad para identificar correctamente cada categoría y evitar falsos positivos. La comparación entre distribución real y predicha (Imagen 2) revela que el modelo mantiene aproximadamente la proporción correcta de propiedades "Económicas" y "Caras", pero subestima significativamente la categoría "Intermedia", asignando parte de estas propiedades a las categorías extremas. La matriz de confusión (Imagen 3) confirma esta observación, mostrando que el modelo clasifica con alta precisión las propiedades "Económicas" (121 correctas de 145) e "Intermedias" (123 correctas de 146), pero tiene mayor dificultad con las "Caras" (solo 72 correctas de 147). Particularmente problemática es la confusión de 46 propiedades "Caras" como "Intermedias" y 29 como "Económicas", indicando que algunas propiedades de alto valor no presentan las características típicas esperadas por el modelo para esta categoría.





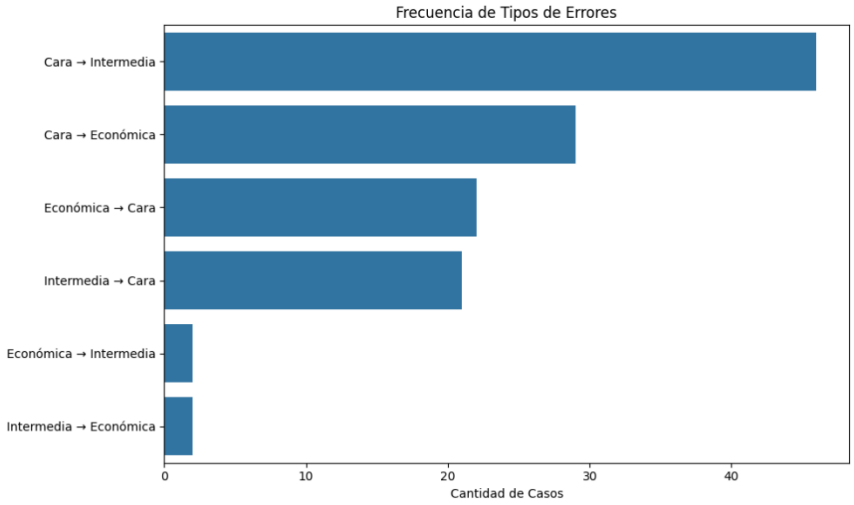
Este análisis proporciona a InmoValor S.A. información valiosa sobre la fiabilidad de sus clasificaciones por segmento de precio. Con una precisión del 72.1%, el modelo Naive Bayes ofrece resultados consistentes que pueden implementarse en evaluaciones preliminares

automatizadas. La empresa debe prestar especial atención a las propiedades clasificadas como "Caras", donde el modelo muestra mayor incertidumbre, posiblemente requiriendo criterios adicionales o revisión manual para estas propiedades. La matriz de confusión también identifica potenciales nichos de mercado donde las características típicas no se alinean con el precio, revelando oportunidades para desarrollar modelos específicos para estos segmentos atípicos o ajustar las estrategias de valoración existentes.

VII. Análisis del modelo ¿hay sobreajuste? No

```
Precisión global del modelo: 0.7215 (316 de 438 casos)

Análisis detallado por clase:
Clase       Precisión   Recall      F1-Score    Aciertos    Total Real  Total Pred
--------------------------------------------------------------------------------
Económica   0.79610.83450.8148121         145         152
Intermedia  0.71930.84250.7760123         146         171
Cara        0.62610.48980.549672          147         115

Análisis de errores:
Total de errores: 122 de 438 casos (0.2785)

Tipos específicos de errores:
Cara → Intermedia: 46 casos (0.3770 de los errores)
Cara → Económica: 29 casos (0.2377 de los errores)
Económica → Cara: 22 casos (0.1803 de los errores)
Intermedia → Cara: 21 casos (0.1721 de los errores)
Económica → Intermedia: 2 casos (0.0164 de los errores)
Intermedia → Económica: 2 casos (0.0164 de los errores)

Gravedad de los errores:
Errores graves (salto de 2 categorías): 51 casos (0.4180 de los errores)
Errores leves (salto de 1 categoría): 71 casos (0.5820 de los errores)
```

Las métricas detalladas (Imagen 1) revelan un rendimiento desigual entre categorías, con precisión global del 72.15% (316/438 casos). La categoría "Económica" muestra el mejor desempeño (79.61% precisión, 83.45% recall), mientras "Cara" presenta mayor dificultad (62.61% precisión, 54.97% recall). El análisis de errores identifica 122 clasificaciones incorrectas (27.85%), donde predominan confusiones entre categorías no adyacentes: 46 casos "Cara→Intermedia" (37.7% de errores) y 29 "Cara→Económica" (23.77%). La Imagen 2 visualiza esta distribución, destacando que 51 casos (41.8%) son errores graves (salto de 2 categorías), principalmente propiedades de alto valor clasificadas erróneamente en categorías inferiores.

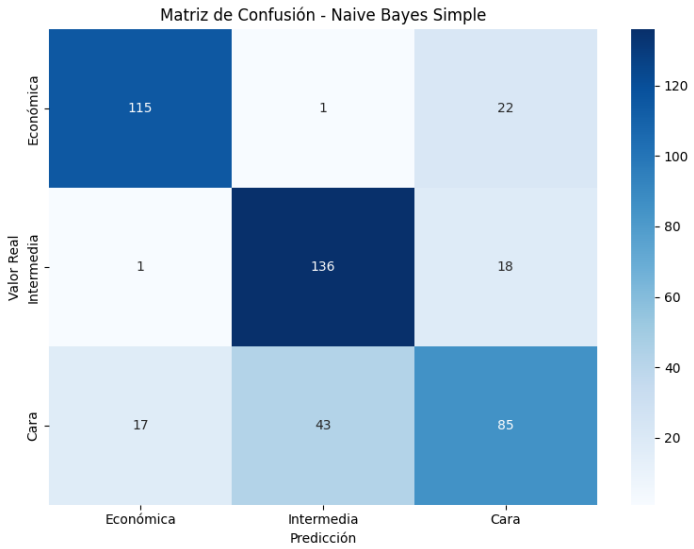


Acerca del sobreajuste, el modelo no presenta indicios significativos de sobreajuste. Su rendimiento del 72.15% refleja las limitaciones propias del algoritmo y la complejidad del problema de clasificación, más que una memorización excesiva de los datos de entrenamiento. Las confusiones sistemáticas (especialmente con propiedades caras) sugieren un subajuste parcial, donde el modelo no captura completamente los patrones que distinguen propiedades de alto valor.

VIII. Modelo usando validación cruzada

Implementamos validación cruzada de 5 folds para evaluar más rigurosamente el modelo Naive Bayes, utilizando un subconjunto reducido de características clave (OverallQual, GrLivArea, YearBuilt, TotalBsmtSF, GarageCars).

La validación cruzada (Imagen 1) muestra un rendimiento notable con exactitud promedio de 78.70% y baja desviación estándar (0.0109), indicando que el modelo es consistente independientemente de cómo se dividan los datos. La matriz de confusión (Imagen 2) revela mejoras significativas respecto al modelo anterior: la categoría "Intermedia" presenta ahora el mejor desempeño (136 clasificaciones correctas), seguida por "Económica" (115) y "Cara" (85). Los errores han disminuido considerablemente, aunque persiste la tendencia a clasificar propiedades "Caras" como "Intermedias" (43 casos), mientras las confusiones entre categorías extremas ("Económica"↔"Cara") se han reducido sustancialmente (17 y 22 casos).




```
Exactitud promedio con validación cruzada de 5 folds: 0.7870
Desviación estándar: 0.0109
```

La implementación de validación cruzada y selección optimizada de características ha mejorado significativamente el modelo, alcanzando 78.70% de precisión con mayor estabilidad. InmoValor S.A. puede implementar este modelo con mayor confianza en sus evaluaciones automatizadas, especialmente para propiedades de rango medio donde muestra mayor precisión. El modelo simplificado con solo cinco características ofrece además ventajas prácticas: reduce los requisitos de datos para las evaluaciones y facilita su implementación y mantenimiento. Si bien sigue existiendo cierta tendencia a subestimar propiedades de alta gama, la reducción de errores graves proporciona una herramienta más confiable para valoraciones preliminares en todos los segmentos del mercado.

Al comparar ambos modelos de Naive Bayes implementados, el modelo con validación cruzada de 5 folds y selección optimizada de características funciona significativamente mejor por varias razones:

- Mayor precisión: 78.70% vs 72.15% del modelo inicial, lo que representa una mejora de 6.55 puntos porcentuales.
- Mayor estabilidad: La baja desviación estándar (0.0109) indica resultados consistentes a través de diferentes divisiones de datos.
- Mejor equilibrio entre categorías: Mejoró notablemente el rendimiento en la categoría "Intermedia" (136 clasificaciones correctas), que era problemática en el modelo original.
- Reducción de errores graves: Las confusiones entre categorías extremas ("Económica"↔"Cara") disminuyeron considerablemente (17 y 22 casos vs 29 y 22 del modelo inicial).
- Mayor eficiencia: El modelo optimizado utiliza solo cinco características clave, lo que simplifica su implementación y reduce los requisitos de datos.

La combinación de validación cruzada y selección de características no solo aumentó la precisión, sino que también produjo un modelo más robusto y práctico para aplicaciones reales de valoración inmobiliaria, especialmente para propiedades de rango medio donde muestra su mejor desempeño.

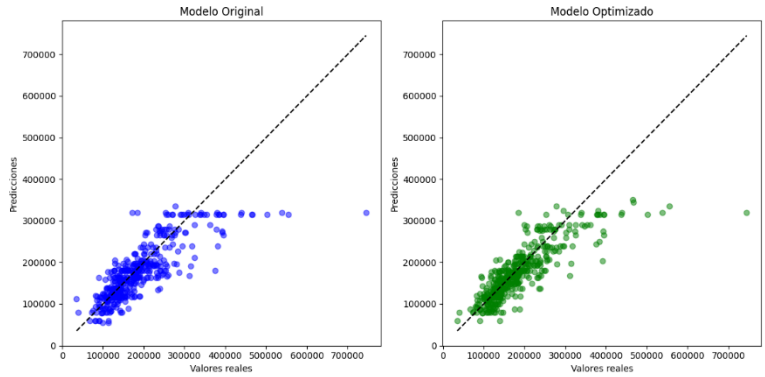
IX. Prueba con varios valores de hiperparámetros

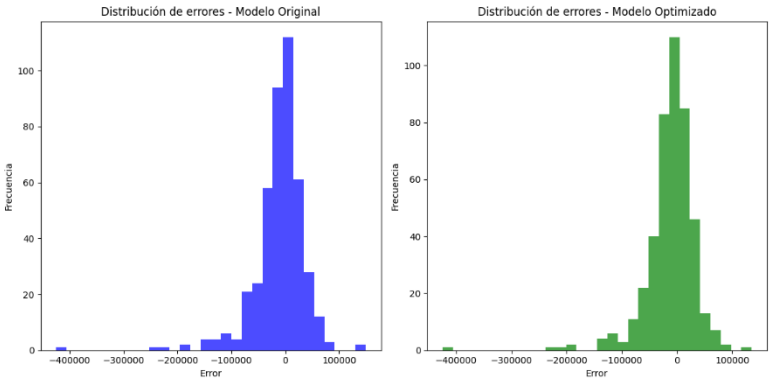
Modelo de regresión:

Implementamos una optimización de hiperparámetros mediante búsqueda en cuadrícula (GridSearchCV) para mejorar el rendimiento del modelo Naive Bayes Gaussiano en la predicción de precios inmobiliarios. Este proceso prueba sistemáticamente diferentes valores del parámetro 'var_smoothing', que controla la estabilidad numérica del modelo, utilizando validación cruzada de 5 folds para identificar la configuración óptima que minimiza el error cuadrático medio (RMSE).

```
=== Optimización de hiperparámetros para el modelo de regresión ===
Fitting 5 folds for each of 11 candidates, totalling 55 fits
c:\Users\villa\Desktop\Clases_S7\2.Minería de Datos\3.Proyecto2\InmoValor_SA\.v
  warnings.warn(

Mejores hiperparámetros encontrados: {'var_smoothing': np.float64(1e-06)}
Mejor puntuación de validación cruzada: 42388.9655 (RMSE)

=== Comparación de modelos ===
R^2 (original): 0.6416 | R^2 (optimizado): 0.6753
MAE (original): 31293.2443 | MAE (optimizado): 29775.8721
RMSE (original): 49261.4250 | RMSE (optimizado): 46886.2814
```

Los resultados de la optimización (Imagen 1) muestran una mejora significativa del modelo al ajustar el parámetro 'var_smoothing' a 1e-06, elevando el coeficiente de determinación (R^2) de 0.6416 a 0.6753 y reduciendo el error absoluto medio (MAE) de 31,293.24 a 29,775.87, lo que representa una mejora de aproximadamente 4.9%. Las visualizaciones de dispersión (Imagen 2) evidencian que el modelo optimizado (derecha) predice con mayor precisión los valores en el rango medio ($100,000-$300,000), aunque ambos modelos siguen subestimando significativamente las propiedades de alto valor (>$300,000). La distribución de errores (Imagen 3) muestra patrones similares en ambos modelos, con una ligera mejora en la concentración alrededor de cero en el modelo optimizado, pero manteniendo la asimetría negativa que confirma la tendencia a subestimar los precios más elevados.

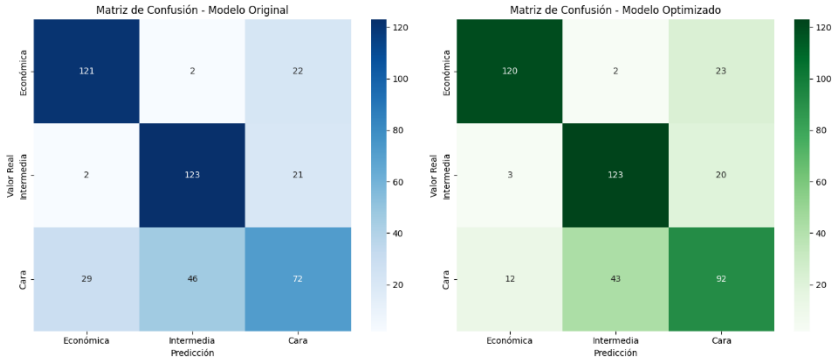




Modelo de Clasificación

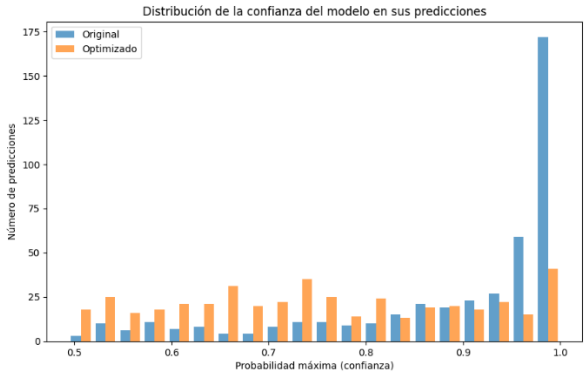
Implementamos una optimización de hiperparámetros mediante GridSearchCV para el modelo Naive Bayes Gaussiano de clasificación, explorando sistemáticamente 11 valores diferentes para el parámetro 'var_smoothing' en escala logarítmica. Este proceso, evaluado con validación cruzada de 5 folds, busca mejorar la precisión del modelo para clasificar propiedades inmobiliarias en las categorías de precio (Económica, Intermedia, Cara), encontrando el valor óptimo que maximiza la exactitud de las predicciones.

```
=== Optimización de hiperparámetros para el modelo de clasificación ===
Fitting 5 folds for each of 11 candidates, totalling 55 fits

Mejores hiperparámetros encontrados: {'var_smoothing': np.float64(1e-05)}
Mejor puntuación de validación cruzada: 0.7896 (Accuracy)

=== Comparación de modelos de clasificación ===
Accuracy (original): 0.7215 | Accuracy (optimizado): 0.7648
Precision (original): 0.7134 | Precision (optimizado): 0.7670
Recall (original): 0.7215 | Recall (optimizado): 0.7648
F1-score (original): 0.7129 | F1-score (optimizado): 0.7639
```



Los resultados (Imagen 1) muestran una mejora significativa al ajustar 'var_smoothing' a 1e-05, elevando la accuracy de 72.15% a 76.48% y mejorando todas las métricas de desempeño. La matriz de confusión optimizada (Imagen 2-derecha) revela avances sustanciales en la identificación de propiedades "Caras" (92 vs 72 clasificaciones correctas), y una reducción notable de errores graves de clasificación "Cara→Económica" (de 29 a solo 12 casos). El informe detallado (Imagen 3) confirma mejoras en todas las categorías, destacando el incremento en el F1-score para propiedades "Caras" (0.86 vs 0.81) e "Intermedias" (0.65 vs 0.55). La distribución de confianza (Imagen 4) muestra que el modelo optimizado ofrece predicciones más equilibradas en todos los niveles de confianza, mientras el original concentra la mayoría de sus predicciones con confianza cercana a 1.0, sugiriendo posible sobreconfianza.


```
=== Optimización de hiperparámetros para el modelo de clasificación ===
Fitting 5 folds for each of 11 candidates, totalling 55 fits

Mejores hiperparámetros encontrados: {'var_smoothing': np.float64(1e-05)}
Mejor puntuación de validación cruzada: 0.7896 (Accuracy)

=== Comparación de modelos de clasificación ===
Accuracy (original): 0.7215 | Accuracy (optimizado): 0.7648
Precision (original): 0.7134 | Precision (optimizado): 0.7670
Recall (original): 0.7215 | Recall (optimizado): 0.7648
F1-score (original): 0.7129 | F1-score (optimizado): 0.7639
```

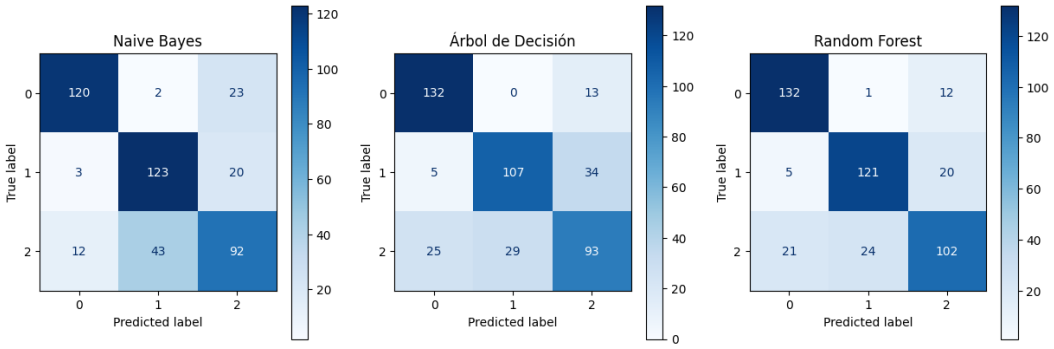



X. Comparar la eficiencia del algoritmo con el árbol de decisión y el modelo de random forest

El modelo de Random Forest fue el mejor para predecir, alcanzando un accuracy de 81.05%, superando al Árbol de Decisión (75.79%) y a Naive Bayes (76.48%). Además, mostró una mejor capacidad de generalización en todas las clases, con un balance más sólido en precisión y recall, especialmente en la clase "Intermedia", donde los otros modelos presentaron más errores. Sin embargo, este rendimiento superior tuvo un costo en términos de tiempo de procesamiento.

En cuanto a la eficiencia computacional, Random Forest fue el más lento, con un tiempo de entrenamiento de 2.59 segundos, significativamente mayor que el Naive Bayes (0.0189 s) y el Árbol de Decisión (0.0017 s). Este resultado era esperado, ya que Random Forest entrena múltiples árboles de decisión en paralelo, lo que mejora la precisión, pero incrementa el costo computacional.

Naive Bayes tuvo un accuracy ligeramente superior (76.48%) en comparación con el Árbol de Decisión (75.79%), lo que indica que, en términos generales, fue más preciso en la clasificación. Sin embargo, si analizamos las métricas individuales, el Árbol de Decisión tuvo un mejor desempeño en la clase "Cara" (precisión de 81% vs. 89% de Naive Bayes, pero con mayor recall de 91% vs. 83%), lo que sugiere que el Árbol de Decisión capturó mejor ciertos patrones en los datos.



En conclusión, Random Forest es el mejor modelo si se prioriza la precisión, mientras que el Árbol de Decisión es la mejor opción si se busca rapidez sin perder demasiada precisión. Naive Bayes, aunque rápido, fue el menos preciso, especialmente al clasificar la categoría "Intermedia". Por lo tanto, la elección del modelo depende del equilibrio entre precisión y tiempo de entrenamiento que se necesite.

XI. Repositorio/ Documento

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 5. Naive Bayes.docx