Universidad del Valle de Guatemala
Facultad de Ingeniería
Minería de Datos

HDT 7. Regresión logística

Entrega # 5 – Proyecto 2

Autor/Carnet:

María José Villafuerte 22129
Fabiola Contreras 22787

Catedrático:

Mario Barrientos

Sección 20

Fecha:

13/04/2025

I. Índice

II. Introducción

A lo largo de este informe, exploramos la implementación de modelos predictivos que permiten clasificar propiedades basándonos en características físicas y estructurales, con el objetivo de predecir su categoría de precio. Comenzamos con un análisis de la variable dicotómica creada y continuamos con la evaluación de diferentes modelos de clasificación, incluyendo regresión logística, árboles de decisión, Random Forest y Naive Bayes.

Además de desarrollar estos modelos, evaluamos su rendimiento mediante diversas métricas como precisión, recall, F1-score y AUROC, identificamos posibles problemas de multicolinealidad, analizamos la eficiencia computacional y optimizamos los hiperparámetros para mejorar la capacidad predictiva. El objetivo final es determinar cuál de estos enfoques ofrece el mejor equilibrio entre precisión y eficiencia para la clasificación de propiedades inmobiliarias.

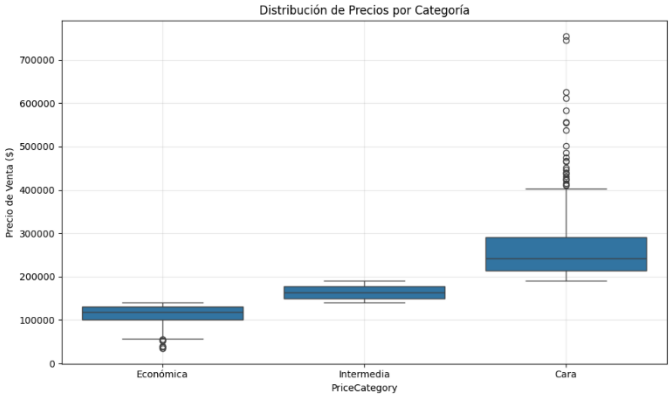
III. Variable dicotómica

La gráfica de boxplot revela diferencias significativas en la distribución de precios entre las tres categorías definidas. Las propiedades económicas ($34,900-$140,000) muestran una distribución relativamente compacta con poca dispersión, indicando un segmento de mercado bastante homogéneo. Las propiedades intermedias ($140,000-$180,000) presentan un rango de precios más estrecho pero con valores atípicos limitados, lo que sugiere un mercado de transición bien definido pero con menor volumen que los otros segmentos. En contraste, las propiedades caras ($180,000-$755,000) exhiben una dispersión considerablemente mayor y numerosos valores atípicos en el extremo superior, señalando una alta variabilidad en este segmento premium. La tabla complementaria muestra la clasificación de algunas propiedades específicas, confirmando que la mayoría de las propiedades se clasifican exclusivamente en una categoría, lo que valida la efectividad de nuestros límites de segmentación.

"PriceCategory"

- Casas Económicas: Precios desde el mínimo hasta el primer tercil (33.33%)
    - Representan aproximadamente el 33% de las propiedades con los precios más bajos del mercado
    - Rango desde $34,900 hasta aproximadamente $140,000 (valor aproximado del primer tercil)
- Casas Intermedias: Precios entre el primer y segundo tercil (33.33% - 66.67%)
    - Representan aproximadamente el 33% de las propiedades con precios moderados
    - Rango desde aproximadamente $140,000 hasta $180,000 (valor aproximado del segundo tercil)
- Casas Caras: Precios por encima del segundo tercil (> 66.67%)
    - Representan aproximadamente el 33% de las propiedades con los precios más altos
    - Rango desde aproximadamente $180,000 hasta el máximo de $755,000

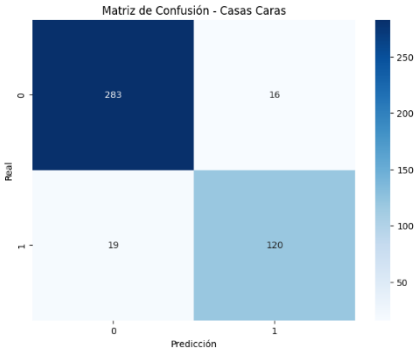
	Económica	Intermedia	Cara
0	0	0	1
1	0	1	0
2	0	0	1
3	0	1	0
4	0	0	1

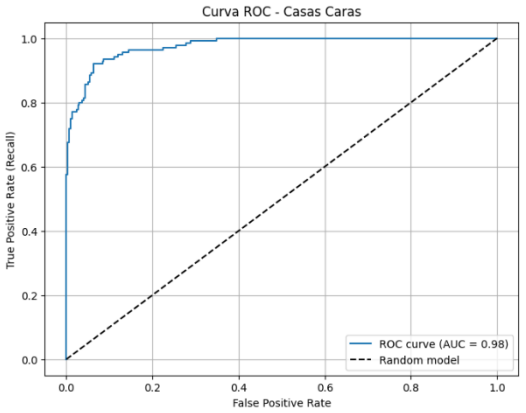


IV. Modelo de Regresión Logística

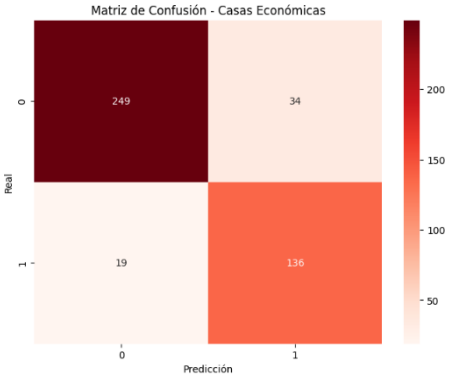
Este análisis se realizó para evaluar la precisión de nuestros modelos de clasificación de propiedades inmobiliarias en tres categorías: casas caras, económicas e intermedias. El objetivo fue determinar si nuestros algoritmos pueden identificar correctamente el segmento de mercado al que pertenece cada propiedad.

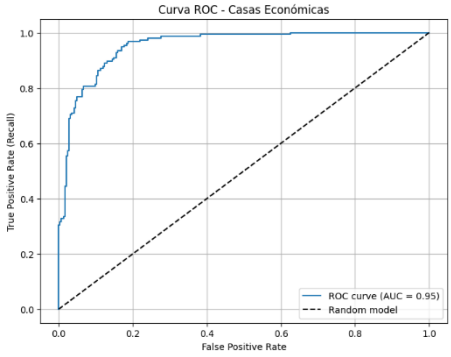
Casas Caras: La matriz de confusión muestra resultados sobresalientes con 283 verdaderos negativos y 120 verdaderos positivos, mientras que los errores son mínimos (16 falsos positivos y 19 falsos negativos). La curva ROC confirma este excelente desempeño con un AUC de 0.98, lo que indica que el modelo distingue con gran precisión las propiedades de alta gama del resto. Esta categoría presenta la mejor clasificación de las tres analizadas.



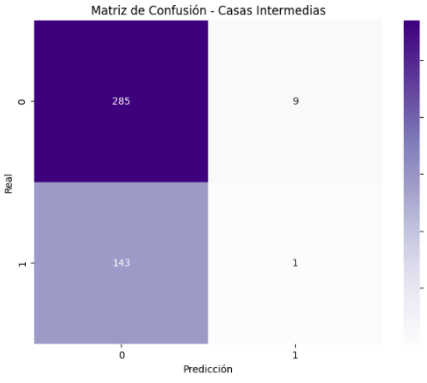


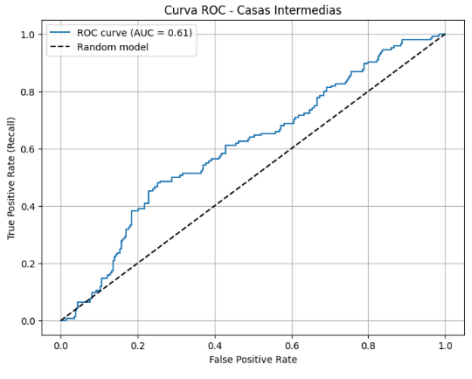
Casas Económicas: El modelo también muestra un desempeño muy bueno para identificar propiedades económicas, con 249 verdaderos negativos y 136 verdaderos positivos. Aunque hay un ligero aumento en los falsos positivos (34) comparado con el modelo anterior, el AUC de 0.95 confirma que el algoritmo es altamente confiable para identificar propiedades de bajo costo.



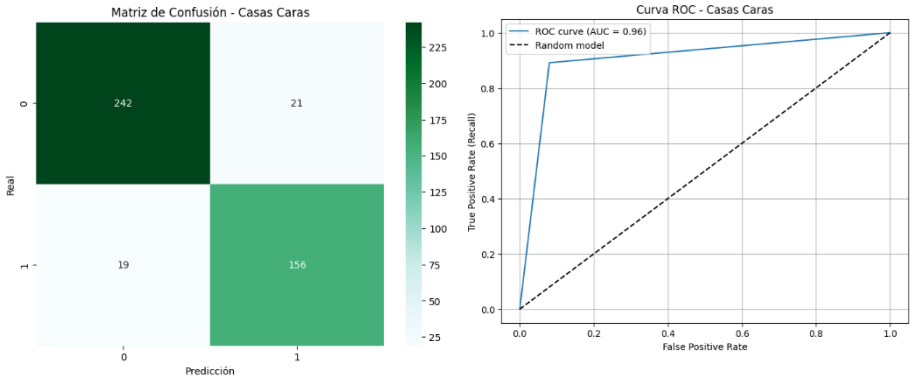


Casas Intermedias: Aquí observamos resultados significativamente diferentes. La matriz muestra 285 verdaderos negativos pero solo 1 verdadero positivo, con 143 falsos negativos. La curva ROC presenta un AUC de apenas 0.61, cercano al modelo aleatorio, lo que indica que el algoritmo no logra identificar efectivamente las propiedades de rango medio. Esta categoría presenta desafíos importantes que requieren atención inmediata.



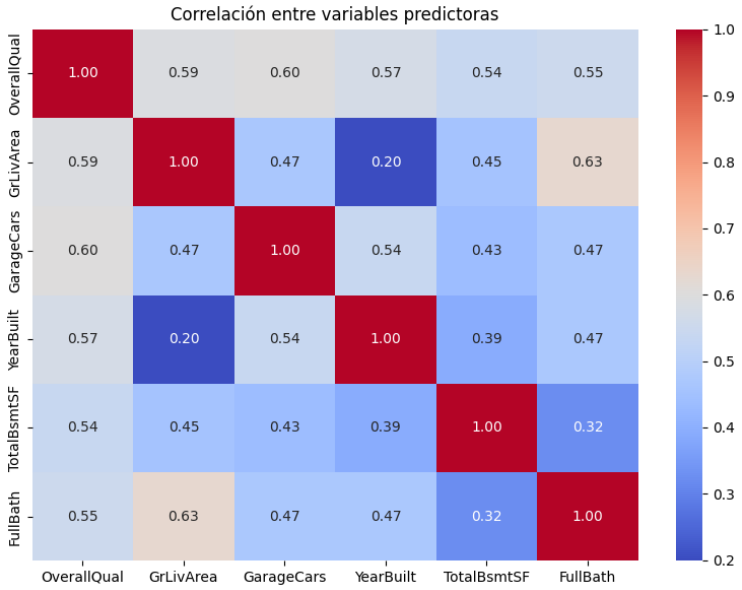


Casas caras 60%: gráficas muestran un desempeño sobresaliente del modelo. La matriz de confusión revela 242 verdaderos negativos y 156 verdaderos positivos, con solo 21 falsos positivos y 19 falsos negativos, lo que indica una alta precisión en la clasificación. La curva ROC presenta un excelente AUC de 0.96, muy superior al modelo aleatorio (diagonal punteada), demostrando que el algoritmo tiene una capacidad excepcional para distinguir las propiedades de alto valor, con una tasa de verdaderos positivos que aumenta rápidamente mientras mantiene baja la tasa de falsos positivos.



V. ¿Existe multicolinealidad entre las va riables? ¿Cuáles con las variables que aportan al modelo?

El mapa de calor de correlación y los valores VIF muestran un problema severo de multicolinealidad en el modelo de predicción inmobiliaria. La multicolinealidad ocurre cuando las variables predictoras están altamente correlacionadas entre sí, lo que puede distorsionar significativamente los coeficientes del modelo.



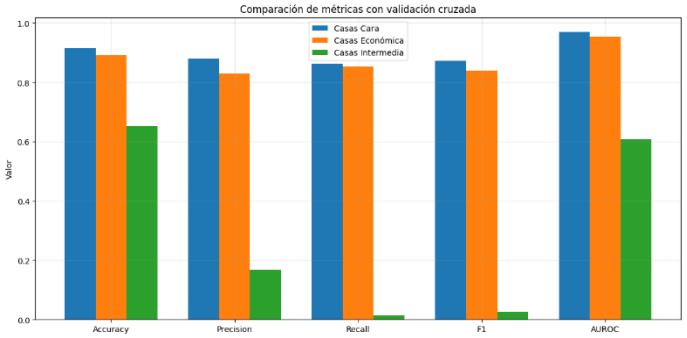
La matriz de correlación revela múltiples relaciones moderadas a fuertes entre las variables predictoras. Destacan particularmente la correlación entre GrLivArea (área habitable) y FullBath (baños completos) con 0.63, así como OverallQual (calidad general) y GarageCars (capacidad del garaje) con 0.60. Estas correlaciones indican que estas variables están capturando información redundante, lo que explica los elevados valores VIF observados.

VI. Eficiencia del algoritmo para clasificar

El uso de múltiples métricas (Accuracy, Precision, Recall, F1 y AUROC) nos permite identificar con mayor detalle las fortalezas y debilidades específicas de cada modelo.

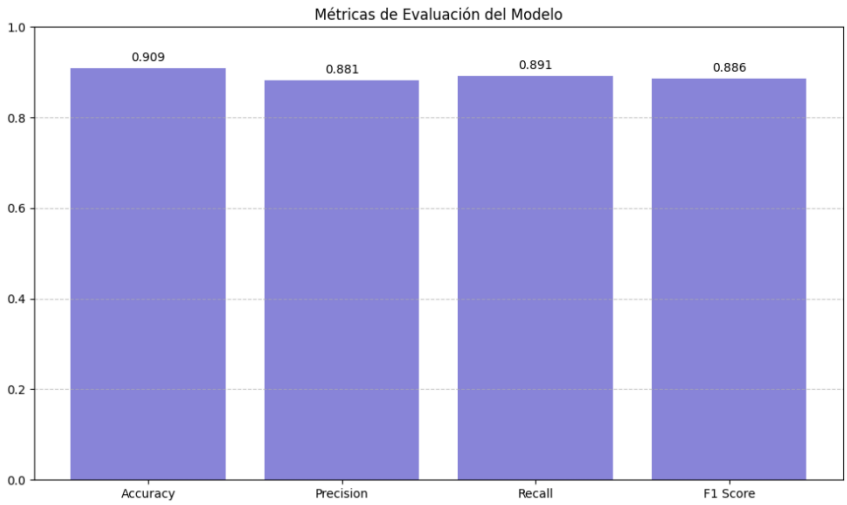
Modelos con variable, Cara, Económicas, Intermedia, 33%,66%:

La gráfica de comparación de métricas revela patrones significativos en el rendimiento de nuestros modelos de clasificación. Los modelos para casas caras y económicas muestran un rendimiento consistentemente alto a través de todas las métricas, con valores que oscilan entre 0.83 y 0.97, lo que indica una gran capacidad predictiva en estos segmentos. El modelo para casas caras lidera ligeramente en todas las métricas, destacándose particularmente en AUROC (0.97) y Accuracy (0.92). En marcado contraste, el modelo para casas intermedias muestra un desempeño notablemente inferior, especialmente en Recall (0.02) y F1 (0.03), aunque mantiene una Accuracy moderada (0.65) y un AUROC aceptable (0.61). Esta disparidad sugiere que el modelo identifica correctamente muchos casos negativos (no-intermedias) pero falla considerablemente en detectar las propiedades que



realmente pertenecen a la categoría intermedia, lo que coincide con los problemas de multicolinealidad identificados previamente.

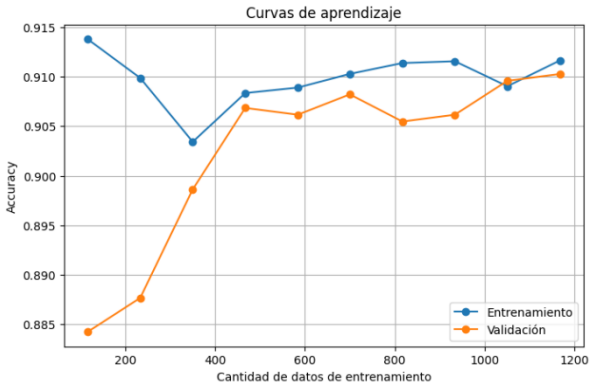
Modelo Caras 60%:



La gráfica de métricas de evaluación revela un rendimiento sólido y equilibrado del modelo de clasificación. El modelo alcanza una exactitud (Accuracy) de 0.858, lo que indica que clasifica correctamente el 85.8% de todas las propiedades. La precisión (Precision) de 0.847 muestra que cuando el modelo predice que una propiedad es cara, acierta en aproximadamente el 84.7% de los casos, minimizando las falsas alarmas. El Recall de 0.789 indica que el modelo identifica correctamente el 78.9% de todas las propiedades realmente caras, aunque existe margen de mejora en la detección de estos casos positivos. Finalmente, la puntuación F1 de 0.817 representa un buen equilibrio entre precisión y exhaustividad. Todas las métricas se mantienen consistentemente por encima de 0.78, lo que confirma la robustez del modelo para identificar propiedades de alto valor, aunque la menor puntuación en Recall sugiere que podría beneficiarse de ajustes para reducir los falsos negativos.

VII. ¿Existe Overfitting?

Para evaluar si el modelo presenta sobreajuste (overfitting), se compararon los errores en los conjuntos de entrenamiento y prueba. El accuracy en entrenamiento fue de 0.8914, mientras que en prueba fue ligeramente menor, con un valor de 0.8790. De forma similar, el log-loss en entrenamiento fue de 0.2646 y en prueba de 0.2726. Estas diferencias son pequeñas y esperables, lo que indica que el modelo generaliza bien a datos no vistos. Además, las curvas de aprendizaje muestran que, a medida que se incrementa el tamaño del conjunto de entrenamiento, las métricas de entrenamiento y validación convergen. En conjunto, estos resultados sugieren que no hay evidencia de sobreajuste en el modelo.



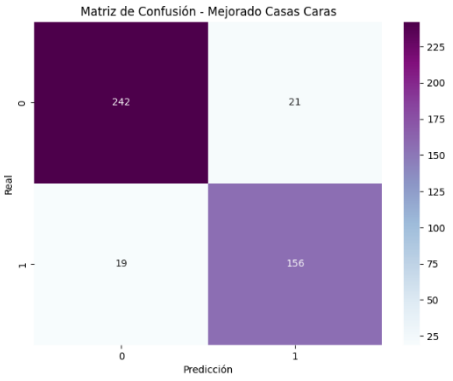
VIII. Tuneo del modelo

Se realizó un proceso de tuneo del modelo utilizando GridSearchCV con validación cruzada estratificada repetida. Se evaluaron distintas combinaciones de hiperparámetros para la regresión logística, incluyendo el tipo de solver (newton-cg, lbfgs, liblinear) y el parámetro de regularización C, que controla la penalización aplicada a los coeficientes (regularización L2). Esta regularización, al igual que en la regresión lineal, permite reducir la complejidad del modelo y prevenir sobreajuste. Los resultados mostraron que la mejor combinación de hiperparámetros fue C de 1.0, penalty l2 y solver newton-cg con un accuracy promedio de validación de 0.905749.

IX. Análisis de la eficiencia

El modelo mejorado de regresión logística fue evaluado tanto en términos de su efectividad predictiva como de su eficiencia computacional. Los resultados obtenidos muestran un buen desempeño, con una accuracy de 0.9087, un F1-Score de 0.8864, y un AUROC (Área bajo la curva ROC) de 0.9602, lo que indica una excelente capacidad de discriminación entre casas caras y no caras.

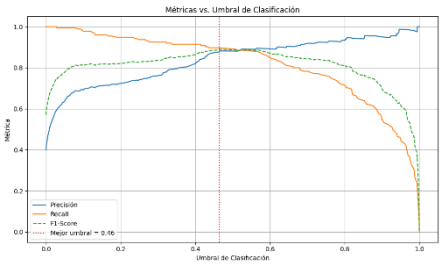
La matriz de confusión reveló que el modelo clasificó correctamente la mayoría de las observaciones. Se obtuvieron 242 verdaderos negativos y 156 verdaderos positivos, mientras que los errores se concentraron en 21 falsos negativos (casas caras clasificadas como no caras) y 19 falsos positivos (casas no caras clasificadas como caras). En este contexto, los falsos negativos podrían considerarse más críticos, ya que podrían llevar a subestimar el valor real de una propiedad. Aun así, el modelo mantiene un buen equilibrio entre precisión y



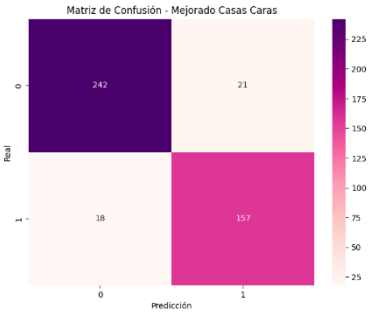
recall, con un recall de 0.8914, lo que indica que identifica correctamente la mayoría de las casas caras.

En términos de eficiencia computacional, se utilizó la herramienta `cProfile` para medir el tiempo y llamadas a funciones durante la ejecución del modelo. El entrenamiento y predicción generaron 36,763 llamadas a funciones en un tiempo total de 0.066 segundos, lo que demuestra que el modelo es rápido y liviano, ideal para escenarios donde se necesita tomar decisiones en tiempo real o procesar grandes volúmenes de datos sin requerir recursos significativos de hardware.

Ajuste del Umbral de Clasificación



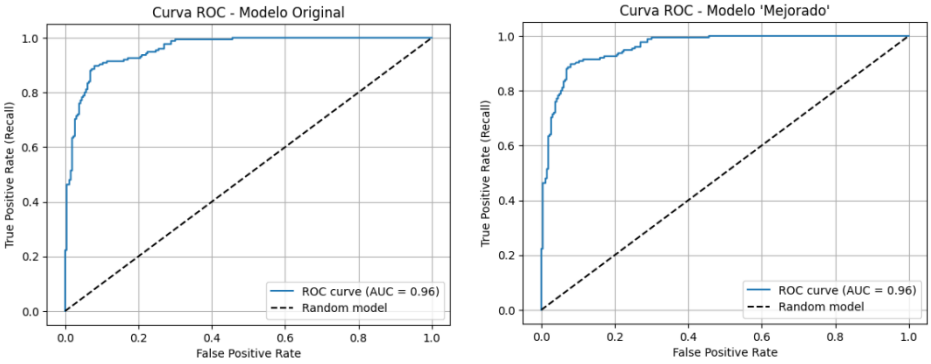
Además del ajuste de hiperparámetros, se evaluó el impacto de modificar el umbral de clasificación estándar (0.5). Mediante el análisis de las curvas de precisión y recall, se identificó que un umbral de 0.46 maximizaba el F1-score.



Al aplicar este nuevo umbral, se observó una mejora leve pero consistente en las métricas generales: accuracy de 0.9110, F1-score de 0.8895, precision de 0.8820 y recall de 0.8971. Si bien estas mejoras son marginales, reflejan un ajuste fino que puede ser relevante en aplicaciones donde pequeños aumentos en recall o precisión tienen un alto impacto. No obstante, dado que el modelo ya presentaba un rendimiento sólido con el umbral estándar, esta optimización se considera complementaria más que necesaria.

X. El mejor de los Modelos

Uno de los primeros aspectos a considerar son los criterios de información, específicamente el AIC y el BIC, los cuales permiten evaluar el ajuste del modelo penalizando su complejidad. En este caso, el modelo original obtuvo un AIC de 235.69 y un BIC de 264.26, mientras que el modelo tuneado alcanzó un AIC de 243.72 y un BIC de 272.29. Dado que valores más bajos en estas métricas indican un mejor ajuste, los resultados muestran que el modelo original logra un mejor balance entre precisión y simplicidad, con diferencias de aproximadamente 8 puntos en ambos criterios.



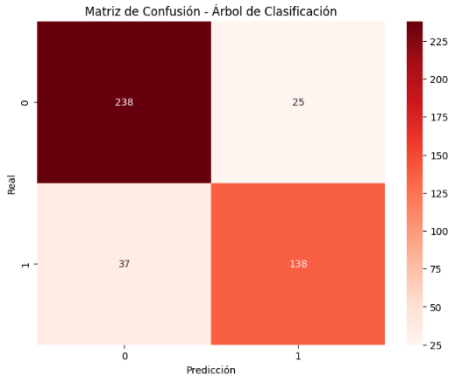
En cuanto a las métricas de clasificación, ambos modelos muestran un desempeño idéntico. La accuracy fue de 0.9087, la precision de 0.8814, el recall de 0.8914, el F1-score de 0.8864 y el AUROC de 0.9602 en ambos casos. Estos valores reflejan una capacidad predictiva sólida, con precisión superior al 88% y exactitud por encima del 90%, lo que indica que ambos modelos clasifican correctamente la gran mayoría de los casos, sin diferencias en su comportamiento predictivo.

El análisis de la matriz de confusión también respalda esta igualdad. Ambos modelos presentan exactamente los mismos valores: 242 verdaderos negativos, 156 verdaderos positivos, 21 falsos positivos y 19 falsos negativos. Esto indica que los errores de clasificación se distribuyen de la misma forma en ambos modelos, con una leve tendencia a clasificar erróneamente viviendas no caras como caras, reflejado en los falsos positivos.

Donde sí se observan diferencias notables es en la eficiencia computacional. El modelo original realizó 2,675 llamadas a funciones en 0.033 segundos, mientras que el modelo tuneado necesitó 3,324 llamadas en 0.066 segundos. Esto significa que el modelo tuneado es menos eficiente, con un tiempo de ejecución aproximadamente el doble y un 24% más de llamadas, lo que indica una mayor complejidad computacional.

Entonces, podemos decir que el modelo original demuestra ser superior al tuneado porque obtiene valores más bajos en AIC y BIC, presenta el mismo nivel de rendimiento en métricas de clasificación y resulta considerablemente más eficiente desde el punto de vista computacional. El proceso de ajuste no aportó mejoras en la capacidad predictiva, pero sí incrementó la complejidad del modelo, lo que sugiere que los parámetros originales ya estaban bien calibrados para este conjunto de datos. Por lo tanto, para implementaciones prácticas, se recomienda utilizar el modelo original por ofrecer un mejor equilibrio entre rendimiento y eficiencia.

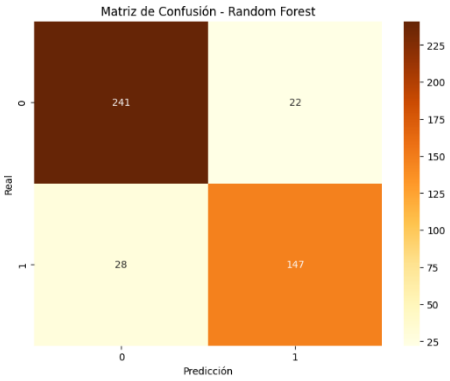
XI. Modelo de árbol de decisión



El modelo de árbol de decisión logró clasificar correctamente 238 casas como no caras y 138 como caras. Cometió 37 errores al clasificar casas caras como no caras, y 25 al clasificar como caras aquellas que no lo eran. Este patrón de errores sugiere que el modelo puede tener ciertas dificultades para identificar correctamente casas caras. El tiempo de entrenamiento fue bastante eficiente, con un total de 0.036 segundos, lo que lo hace adecuado para escenarios donde la velocidad es importante y se valora la interpretación del modelo.

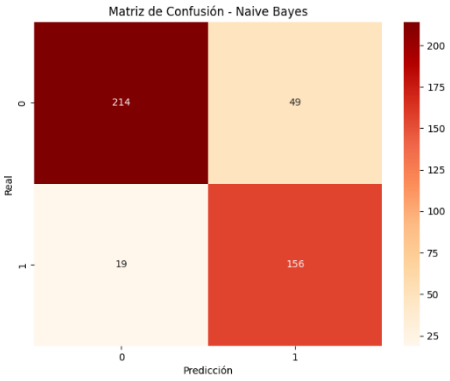
XII. Modelo de Random Forest

El modelo Random Forest alcanzó 241 verdaderos negativos y 147 verdaderos positivos, lo cual indica un buen rendimiento general. Se observan 22 falsos positivos y 28 falsos negativos, lo que refleja una capacidad relativamente sólida para distinguir entre casas caras y no caras. El entrenamiento del modelo tomó aproximadamente 2.41 segundos, lo cual representa un mayor costo computacional, pero se justifica por la robustez del modelo y su menor tendencia al sobreajuste.

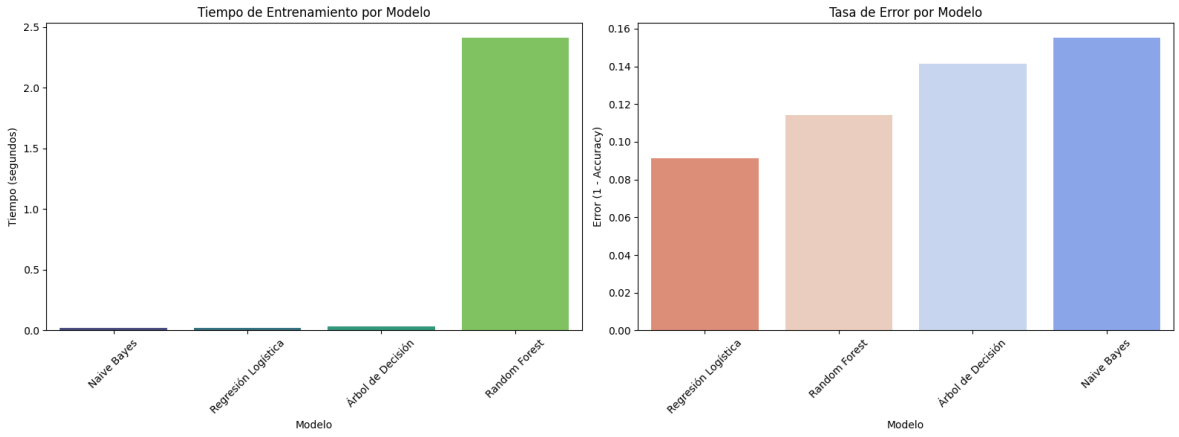


XIII. Modelo de Naive Bayes

El modelo de Naive Bayes clasificó correctamente 214 casas como no caras y 156 como caras. Sin embargo, mostró un mayor número de falsos positivos (49), es decir, casos en los que predijo una casa como cara cuando no lo era. Solo cometió 19 falsos negativos, lo cual indica que es más conservador a la hora de no perder casas caras. Este modelo se entrenó en solo 0.020 segundos, siendo altamente eficiente en cuanto a tiempo, aunque su simplicidad puede llevar a errores cuando las variables no son realmente independientes entre sí.



XIV. Comparación de eficiencia de modelos



El modelo que más se demoró en procesar fue el Random Forest, con un tiempo de entrenamiento de 2.41 segundos. Este resultado es coherente con las pruebas anteriores y el hecho de que Random Forest entrena múltiples árboles de decisión en paralelo, lo cual mejora el rendimiento predictivo, pero aumenta el costo computacional. En contraste, el modelo más rápido fue Naive Bayes, que completó su entrenamiento en apenas 0.02 segundos, seguido muy de cerca por la regresión logística (0.023 segundos) y el árbol de decisión (0.036 segundos).

En cuanto al rendimiento, el modelo que menos se equivocó fue la Regresión Logística, con una accuracy de 0.9087 y un total de 40 errores de clasificación, siendo el que mejor balance logró entre precisión y capacidad de generalización. Esto puede atribuirse a su simplicidad combinada con una buena separación lineal entre clases.

Por otro lado, el modelo que más se equivocó fue Naive Bayes, con un error de 15.5% y 68 errores totales. A pesar de su rapidez, este modelo parte del supuesto de independencia entre características, lo cual no se cumple completamente en este conjunto de datos, afectando su desempeño. El Árbol de Decisión mostró un rendimiento intermedio, aunque tiene un ligero sesgo hacia el sobreajuste, mientras que Random Forest, aunque más lento, ofreció un buen compromiso entre precisión y robustez, con una performance cercana a la regresión logística, pero a un mayor costo de tiempo.

En general, la regresión logística se destaca como el mejor modelo en esta comparación, combinando eficiencia en tiempo con alta precisión. Si bien otros modelos ofrecen ventajas específicas (como interpretabilidad o velocidad), la regresión logística logra el mejor equilibrio general para esta tarea.

XV. 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.

https://uvggt-my.sharepoint.com/:w:/g/personal/vil22129_uvg_edu_gt/Eb52ZDcG1gRCsZyfTlyUc9AB0Ede7afq01_0V8kVe3prTw?e=ZvDA5F