UNIVERSIDAD DEL VALLE DE GUATEMALA

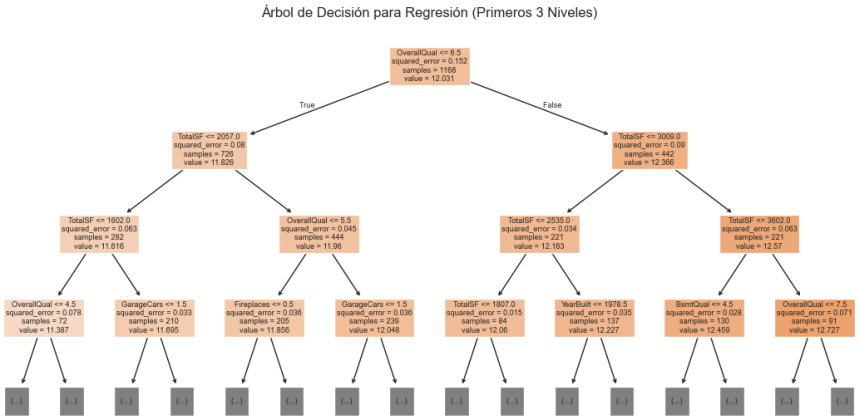
Minería de Datos

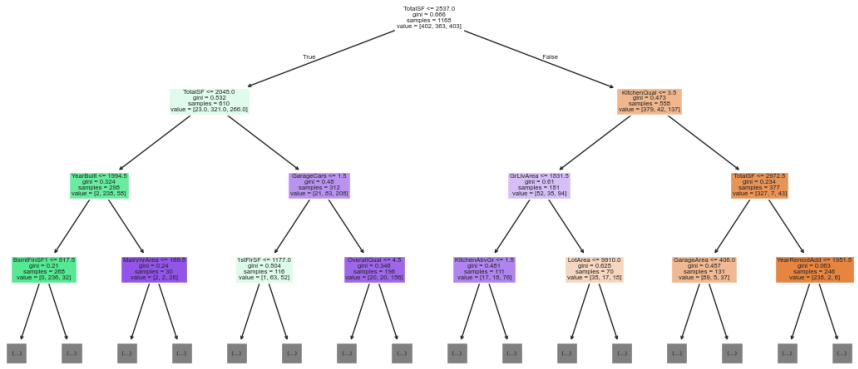
Proyecto No. 2

Isabella Miralles #22293

Guatemala, 2025

Usen los mismos conjuntos de entrenamiento y prueba que usaron para los modelos de regresión lineal en la entrega anterior.

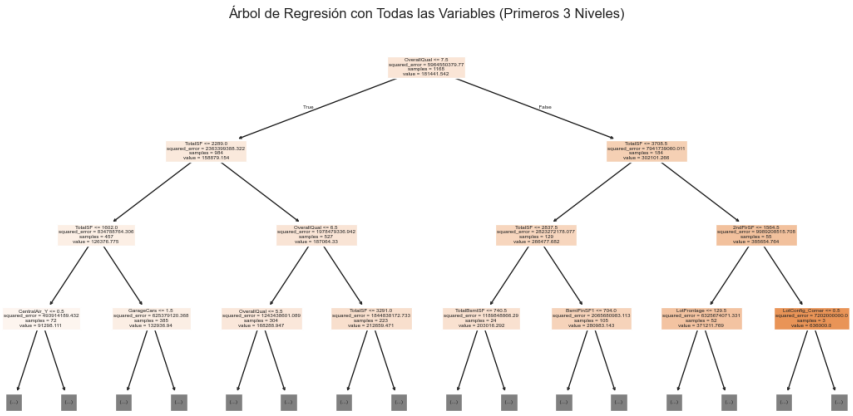




```
Antes de modificar el dataset: (1460, 81)
Después de modificar el dataset: (1460, 123)
Umbrales de clasificación de precios:
  Económicas: < 139000.00
  Intermedias: entre 139000.00 y 189893.00
  Caras: > 189893.00
Filas en entrenamiento: 1168
Filas en prueba: 292

### Evaluación de Regresión Lineal ###
### Evaluación del Modelo en Conjunto de Prueba ###
MSE: 0.0231
RMSE: 0.1521
MAE: 0.1082
R²: 0.8761

### Evaluación del Árbol de Decisión para Regresión ###
MSE: 0.0403
RMSE: 0.2007
R²: 0.7842

### Evaluación del Random Forest para Regresión ###
MSE: 0.0212
RMSE: 0.1456
R²: 0.8865
```

```
### Evaluación del Árbol de Decisión para Clasificación ###
Exactitud: 0.7774
Reporte de Clasificación:
              precision    recall  f1-score   support

       Caras       0.89      0.85      0.87        95
  Económicas       0.78      0.82      0.80       109
 Intermedias       0.66      0.65      0.65        88

    accuracy                           0.78       292
   macro avg       0.78      0.77      0.77       292
weighted avg       0.78      0.78      0.78       292
```

Elaboren un árbol de regresión para predecir el precio de las casas usando todas las variables.




```
PS D:\Documentos\Septimo semestre\Mineria de Datos\Proyecto1-MD> python tree_regression_all.py
Antes de modificar el dataset: (1460, 81)
Después de modificar el dataset: (1460, 123)
Filas en entrenamiento: 1168
Filas en prueba: 292
### Evaluación del Árbol de Regresión con Todas las Variables ###
MSE: 1375771161.0137
RMSE: 37091.3893
MAE: 25281.0822
R²: 0.8206
```

Úsenlo para predecir y analicen el resultado. ¿Qué tal lo hizo?

- El árbol explica alrededor del 82% de la variabilidad en el precio de las casas. El error medio de aproximadamente 25000 dólares y la raíz del error cuadrático medio de 37000 dólares, esto quiere decir que el modelo capta la mayor parte de la valoración de los precios. El modelo lo hizo bien al captar la mayoría de los patrones del conjunto de datos, aunque aún hay margen de mejoras si se quiere reducir el error de predicción.

Desarrollen, al menos, 3 modelos más, cambiando el parámetro de la profundidad del árbol. ¿Cuál es el mejor modelo para predecir el precio de las casas?

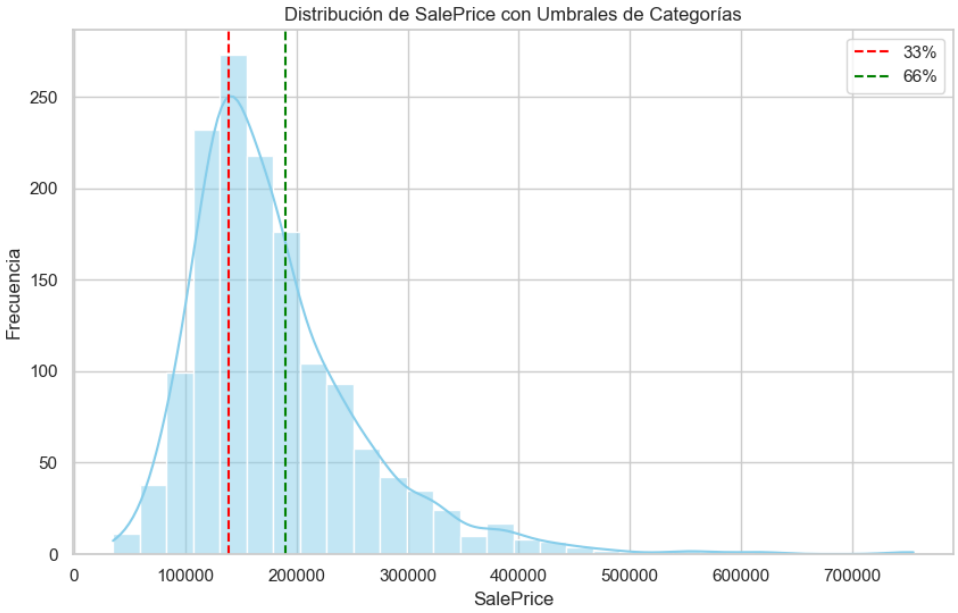
```
Antes de modificar el dataset: (1460, 81)
Después de modificar el dataset: (1460, 123)
Filas en entrenamiento: 1168
Filas en prueba: 292

### Comparación de Árboles de Decisión con Distintas Profundidades ###
max_depth = None
  -> MSE:  1375771161.01
  -> RMSE: 37091.39
  -> MAE:  25281.08
  -> R²:   0.8206

max_depth = 3
  -> MSE:  1628318211.36
  -> RMSE: 40352.43
  -> MAE:  28695.27
  -> R²:   0.7877

max_depth = 5
  -> MSE:  1431851379.67
  -> RMSE: 37839.81
  -> MAE:  24232.16
  -> R²:   0.8133

max_depth = 7
  -> MSE:  1222087753.46
  -> RMSE: 34958.37
  -> MAE:  22551.94
  -> R²:   0.8407
```

```
### Mejor Modelo según R² ###
max_depth = 7 con R² = 0.8407
```

Comparen los resultados con el modelo de regresión lineal de la hoja anterior, ¿cuál lo hizo mejor?

- El modelo de regresión lineal es mejor en términos de capacidad explicativa y estabilidad. Aunque el árbol de decisión además de tener un desempeño cercano tiene la ventaja de ser interpretado visualmente y poder capturar relaciones no lineales e interacciones que el modelo lineal puede llegar a pasar por alto.

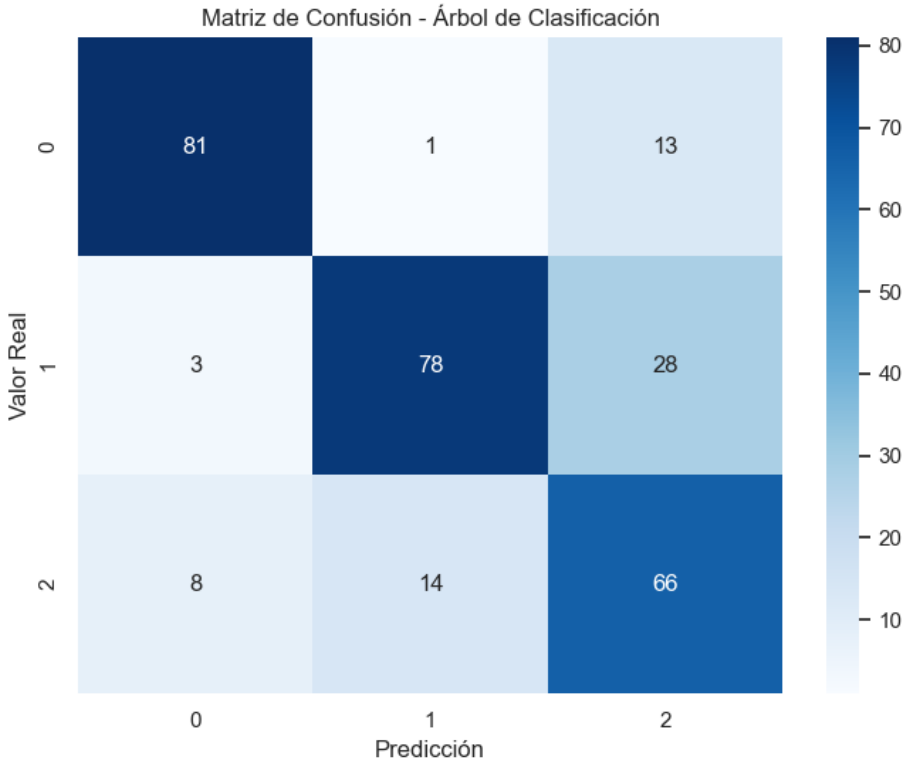
Dependiendo del análisis exploratorio elaborado creen una variable respuesta que les permita clasificar las casas en Económicas, Intermedias o Caras. Los límites de estas clases deben tener un fundamento en la distribución de los datos de precios, y estar bien explicados

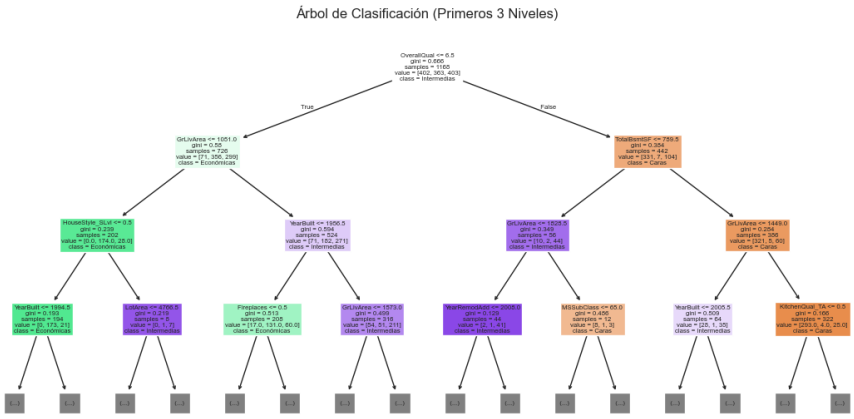



```
PS D:\Documentos\Septimo semestre\Mineria de Datos\Proyecto1-MD> python price_categ
Umbral inferior (33%): 139000.00
Umbral superior (66%): 189893.00

Conteo de casas por categoría:
PriceCategory
Caras          497
Intermedias    491
Económicas     472
Name: count, dtype: int64
```

- Selección de umbrales
  El percentil 33 y 66 son los que se utiliza, esto quiere decir que la división se basa en la distribución real de los precios.
- Grafica
  Muestra la distribución de SalePrice y muestra como se ubican los umbrales, esto respalda la elección de los límites.
- Interpretación
  Económicas, son las casas con un valor inferior al 33 % de la distribución.
  Intermedias, son las casas con los precios en el rango medio.
  Caras, son las casas con el tercio superior de la distribución.

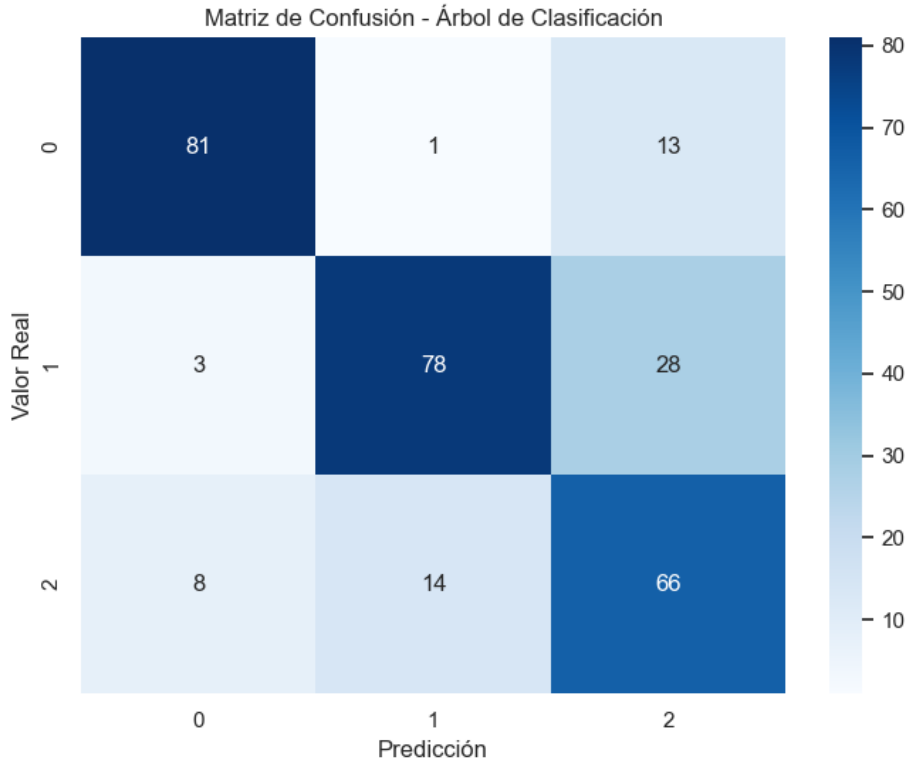
Elaboren un árbol de clasificación utilizando la variable respuesta que crearon en el punto anterior. Expliquen los resultados a los que llegaron. Muestren el modelo gráficamente. Recuerden que la nueva variable respuesta es categórica, pero se generó a partir de los precios de las casas, no incluyan el precio de venta para entrenar el modelo.



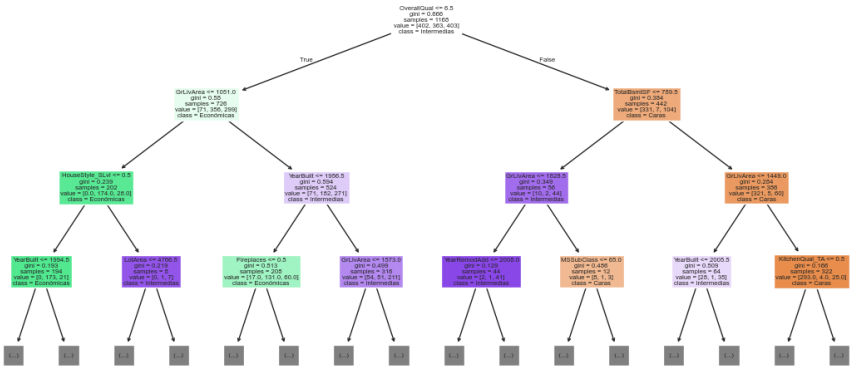


```
PS D:\Documentos\Septimo semestre\Mineria de Datos\Proyecto1-MD> py
Umbral inferior (33%): 139000.00
Umbral superior (66%): 189893.00
Filas en entrenamiento: 1168
Filas en prueba: 292
Exactitud del modelo de clasificación: 0.7705

Reporte de clasificación:
              precision    recall  f1-score   support

       Caras       0.88      0.85      0.87        95
  Económicas       0.84      0.72      0.77       109
 Intermedias       0.62      0.75      0.68        88

    accuracy                           0.77       292
   macro avg       0.78      0.77      0.77       292
weighted avg       0.79      0.77      0.77       292
```

Utilicen el modelo con el conjunto de prueba y determinen la eficiencia del algoritmo para clasificar.

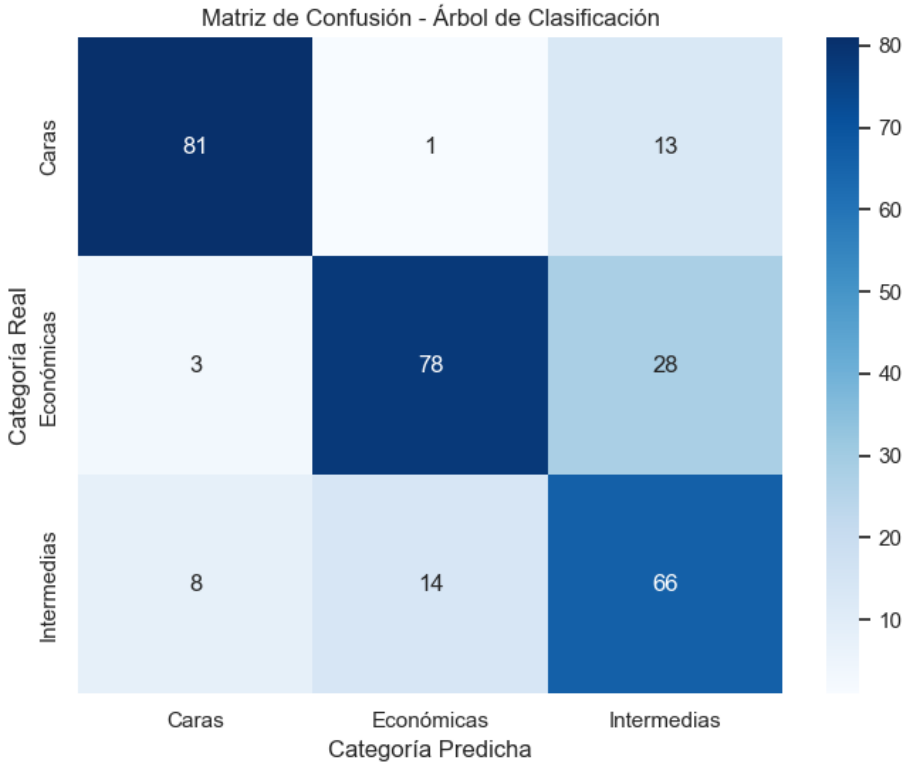


Árbol de Clasificación (Primeros 3 Niveles)




```
Umbral inferior (33%): 139000.00
Umbral superior (66%): 189893.00
Filas en entrenamiento: 1168
Filas en prueba: 292
Exactitud del modelo de clasificación: 0.7705

Reporte de clasificación:
              precision    recall  f1-score   support

       Caras       0.88      0.85      0.87        95
  Económicas       0.84      0.72      0.77       109
 Intermedias       0.62      0.75      0.68        88

    accuracy                           0.77       292
   macro avg       0.78      0.77      0.77       292
weighted avg       0.79      0.77      0.77       292
```

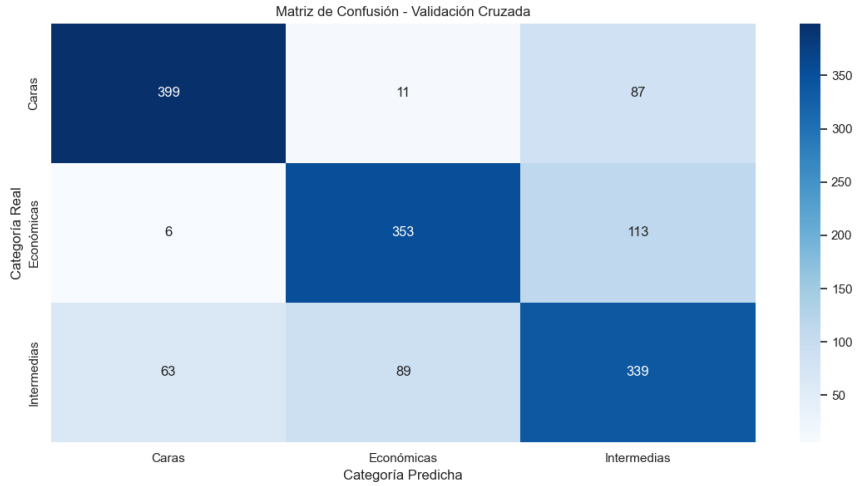
Realicen un análisis de la eficiencia del algoritmo usando una matriz de confusión para el árbol de clasificación. Tengan en cuenta la efectividad, dónde el algoritmo se equivocó más, dónde se equivocó menos y la importancia que tienen los errores.



```
PS D:\Documentos\Septimo semestre\Mineria de Datos\Proyecto
Umbral inferior (33%): 139000.00
Umbral superior (66%): 189893.00
Filas en entrenamiento: 1168
Filas en prueba: 292
Matriz de Confusión:
[[81  1 13]
 [ 3 78 28]
 [ 8 14 66]]
Reporte de Clasificación:
              precision    recall  f1-score   support

       Caras       0.88      0.85      0.87        95
  Económicas       0.84      0.72      0.77       109
 Intermedias       0.62      0.75      0.68        88

    accuracy                           0.77       292
   macro avg       0.78      0.77      0.77       292
weighted avg       0.79      0.77      0.77       292


Total de instancias evaluadas: 292
Instancias correctamente clasificadas: 225
Instancias mal clasificadas: 67
```

```
Análisis por categoría:
  - Caras:
      Total: 95
      Correctos: 81
      Errores: 14 (Tasa de error: 0.15)
  - Económicas:
      Total: 109
      Correctos: 78
      Errores: 31 (Tasa de error: 0.28)
  - Intermedias:
      Total: 88
      Correctos: 66
      Errores: 22 (Tasa de error: 0.25)
Interpretación:
 - Las categorías con mayor número o tasa de error indican dónde el modelo tiene más dificultad para distinguir las clases.
 - Por ejemplo, si 'Intermedias' tiene una tasa de error elevada, puede deberse a que sus características se solapan con las de 'Económicas'
 o 'Caras'.
 - Los errores en ciertas categorías pueden ser más críticos dependiendo del contexto; por ejemplo, confundir una casa 'Cara' con una 'Econó
mica' puede tener implicaciones importantes en la toma de decisiones.
PS D:\Documentos\Septimo semestre\Mineria de Datos\Proyecto1-MD>
```

Entrenen un modelo usando validación cruzada, predigan con él. ¿le fue mejor que al modelo anterior?




```
PS D:\Documentos\Septimo semestre\Mineria de Datos\Proyecto1-MD> python cross_validation_model.py
Umbral inferior (33%): 139000.00
Umbral superior (66%): 189893.00
Exactitud en cada fold (validación cruzada, 5-fold):
[0.73287671 0.72945205 0.77739726 0.76027397 0.73630137]
Exactitud media (CV): 0.7473

Reporte de clasificación (validación cruzada):
              precision    recall  f1-score   support

       Caras       0.85      0.80      0.83       497
  Económicas       0.78      0.75      0.76       472
 Intermedias       0.63      0.69      0.66       491

    accuracy                           0.75      1460
   macro avg       0.75      0.75      0.75      1460
weighted avg       0.75      0.75      0.75      1460


Análisis comparativo:
El modelo entrenado con simple split tenía una exactitud de aproximadamente 0.7705.
La validación cruzada muestra una exactitud media de CV de 0.7473, que es similar o inferior al modelo anterior.
```

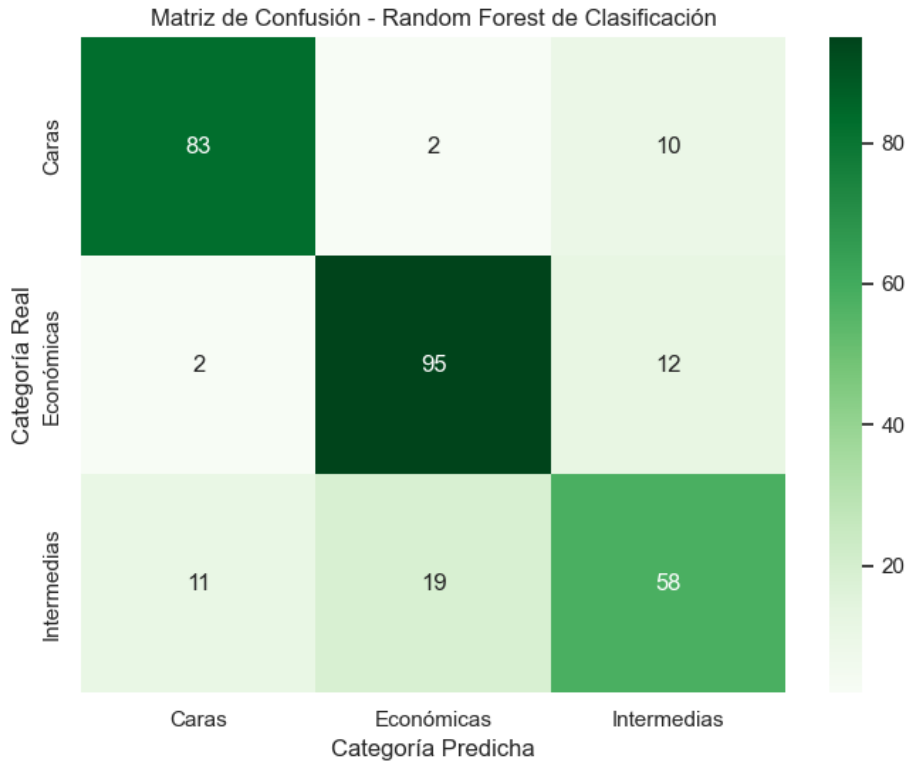
Hagan al menos, 3 modelos más, cambiando la profundidad del árbol. ¿Cuál funcionó mejor?

```
PS D:\Documentos\Septimo semestre\Mineria de Datos\Proyecto1-MD> python compare
Umbral inferior (33%): 139000.00
Umbral superior (66%): 189893.00
Filas en entrenamiento: 1168
Filas en prueba: 292
### Comparación de Árboles de Clasificación con Diferentes Profundidades ###

max_depth = None
  -> Exactitud: 0.7945
  -> Reporte de Clasificación:
              precision    recall  f1-score   support

       Caras       0.92      0.83      0.87        95
  Económicas       0.82      0.83      0.82       109
 Intermedias       0.66      0.72      0.68        88

    accuracy                           0.79       292
   macro avg       0.80      0.79      0.79       292
weighted avg       0.80      0.79      0.80       292
```

```
max_depth = 3
  -> Exactitud: 0.7363
  -> Reporte de Clasificación:
              precision    recall  f1-score   support

       Caras       0.92      0.77      0.84        95
  Económicas       0.76      0.74      0.75       109
 Intermedias       0.58      0.69      0.63        88

    accuracy                           0.74       292
   macro avg       0.75      0.73      0.74       292
weighted avg       0.76      0.74      0.74       292


max_depth = 5
  -> Exactitud: 0.7705
  -> Reporte de Clasificación:
              precision    recall  f1-score   support

       Caras       0.88      0.85      0.87        95
  Económicas       0.84      0.72      0.77       109
 Intermedias       0.62      0.75      0.68        88

    accuracy                           0.77       292
   macro avg       0.78      0.77      0.77       292
weighted avg       0.79      0.77      0.77       292
```

```
max_depth = 7
  -> Exactitud: 0.7740
  -> Reporte de Clasificación:
              precision    recall  f1-score   support

       Caras       0.91      0.81      0.86        95
  Económicas       0.83      0.76      0.79       109
 Intermedias       0.62      0.75      0.68        88

    accuracy                           0.77       292
   macro avg       0.78      0.77      0.78       292
weighted avg       0.79      0.77      0.78       292

### Mejor Modelo Según Exactitud ###
max_depth = None con exactitud = 0.7945
```

Repitan los análisis usando Random Forest como algoritmo de predicción, expliquen sus resultados comparando ambos algoritmos.




```
PS D:\Documentos\Septimo semestre\Mineria de Datos\Proyecto1-MD> python
Umbral inferior (33%): 139000.00
Umbral superior (66%): 189893.00
Filas en entrenamiento: 1168
Filas en prueba: 292
Exactitud del modelo Random Forest de clasificación: 0.8082

Reporte de clasificación:
              precision    recall  f1-score   support

       Caras       0.86      0.87      0.87        95
  Económicas       0.82      0.87      0.84       109
 Intermedias       0.72      0.66      0.69        88

    accuracy                           0.81       292
   macro avg       0.80      0.80      0.80       292
weighted avg       0.81      0.81      0.81       292
```

En el árbol de decisión se obtuvo una exactitud del 77%. Por otro lado el modelo Random Forest aprovecha la agregación de múltiples arboles para mejorar la robustez y la exactitud en general. La exactitud del Random es mayor a la del árbol de decisión, con eso

podemos concluir que el ensamblaje de arboles dan mayor estabilidad y menor sobreajuste, esto mejora la capacidad predictiva.

Enlace github

https://github.com/Isabella-22293/Proyecto1-MD.git