CREDIT SCORING PREDICTION

Data Science Project

Introducción

La importancia de reducir el riesgo crediticio ha llevado a una institución financiera alemana a buscar soluciones innovadoras.

Nuestra tarea es construir un modelo de machine learning preciso y confiable que sea capaz de evaluar con mayor precisión la probabilidad de incumplimiento crediticio de sus clientes.

Metodología

1. Preprocesamiento de Datos: Realizar limpieza de datos, manejar valores faltantes, codificación de variables categóricas y normalización/escalado de datos.

2. Exploración de Datos: Analizar y comprender el conjunto de datos proporcionado, identificar variables llaves y realizar visualizaciones para entender las relaciones entre las variables y seleccionar las características relevantes.

3. Construcción de Modelos: Experimentar con algunos algoritmos de machine learning como Regresión Logística, Árboles de Decisión, Random Forest, Naive Bayes, entre otros.

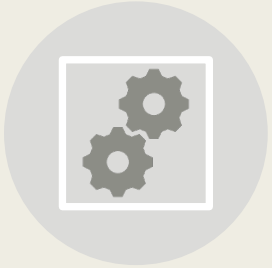
4. Evaluación y Selección del Modelo: Evaluar los modelos utilizando métricas como precisión, recall, área bajo la curva ROC, y F1-score. Seleccionar el modelo con el mejor rendimiento para la predicción de la solvencia crediticia.

Preprocesamiento de Datos

El preprocesamiento de datos es una etapa crucial en el desarrollo de modelos de machine learning y desempeña un papel fundamental en la mejora del rendimiento y la eficacia de dichos modelos. Algunas acciones que se realizaron en esta etapa fueron:

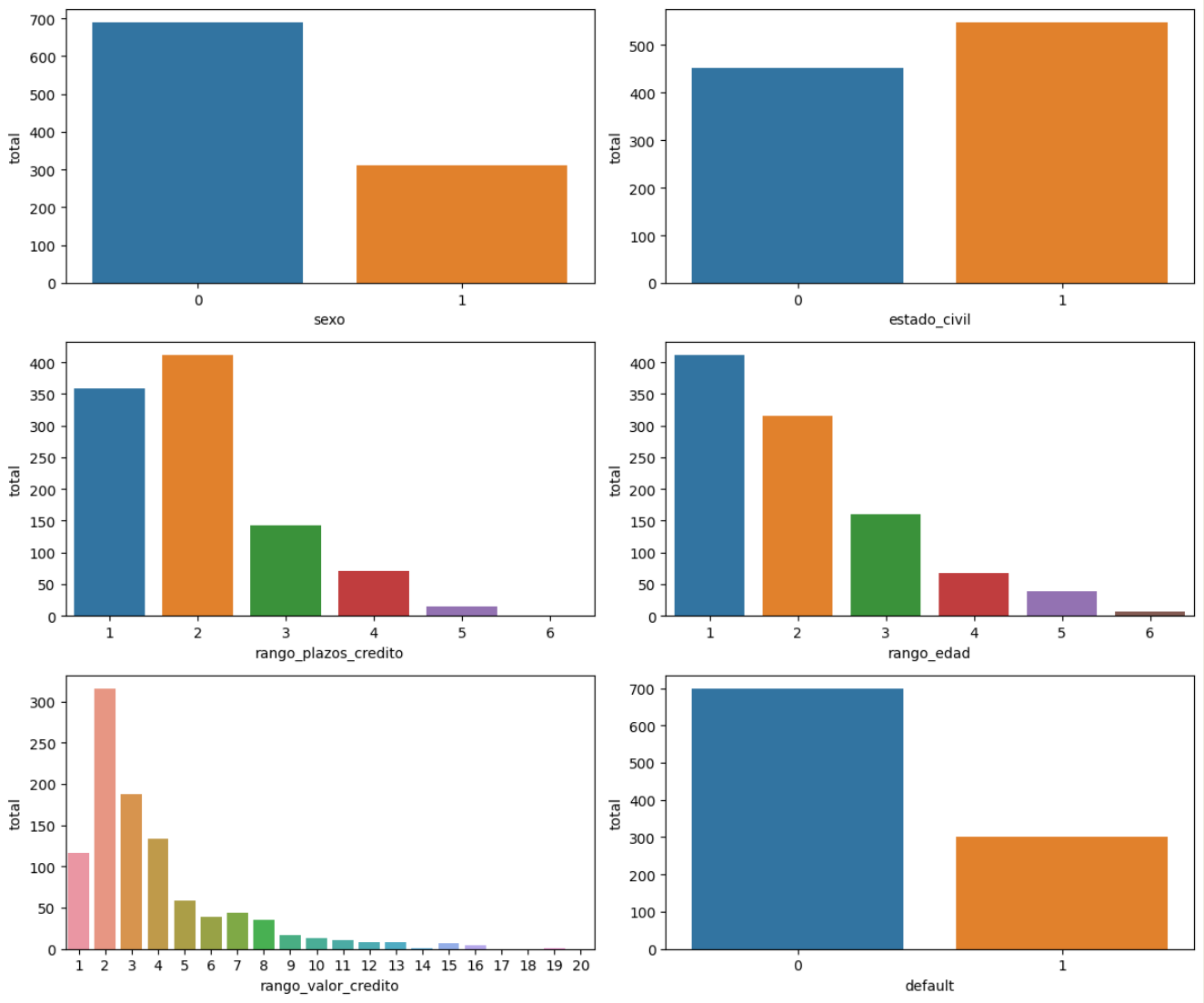
Manejo de datos faltantes: El preprocesamiento permite abordar estos valores nulos mediante la imputación, eliminación o alguna otra estrategia.

Manejo de duplicados: Los datos duplicados en conjuntos de datos pueden introducir sesgos y afectar negativamente el rendimiento y la validez de los modelos de machine learning.



Normalización y estandarización: La normalización y estandarización son técnicas comunes para ajustar los datos a una escala específica, lo que facilita el entrenamiento y mejora la convergencia de los algoritmos.

Manejo de datos categóricos: Los modelos de machine learning a menudo requieren que todas las variables sean numéricas. El preprocesamiento incluye la conversión de variables categóricas en un formato numérico adecuado.



Exploración de Datos

Se observa como existen más hombres (0) que mujeres (1).

Hay más personas solteras (1).

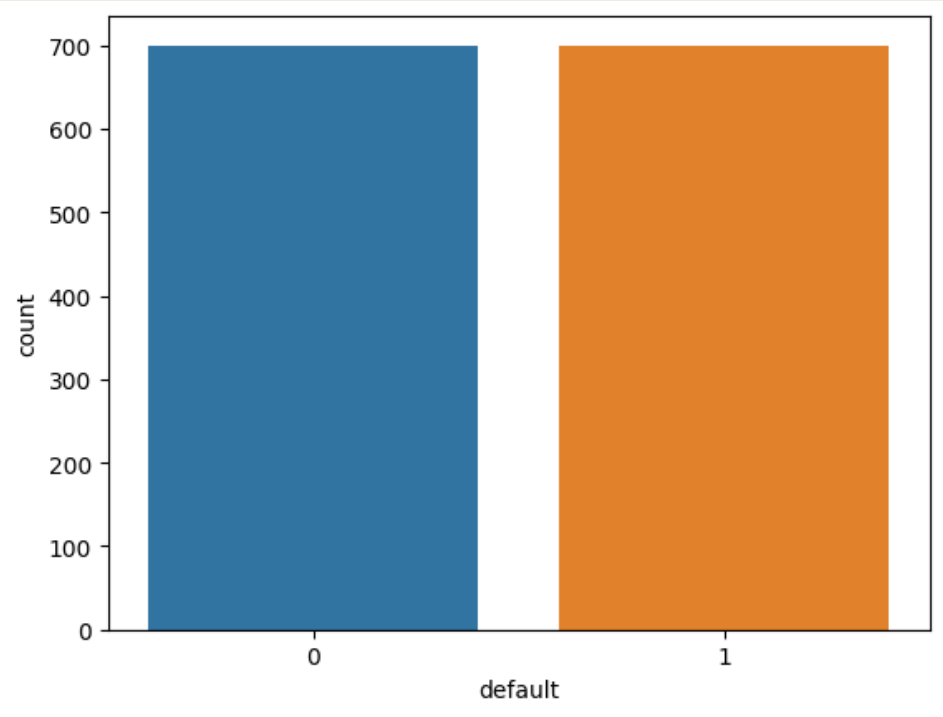
Las personas más jóvenes piden más créditos a un plazo no tan largo.

Los valores de los créditos son relativamente pequeños.

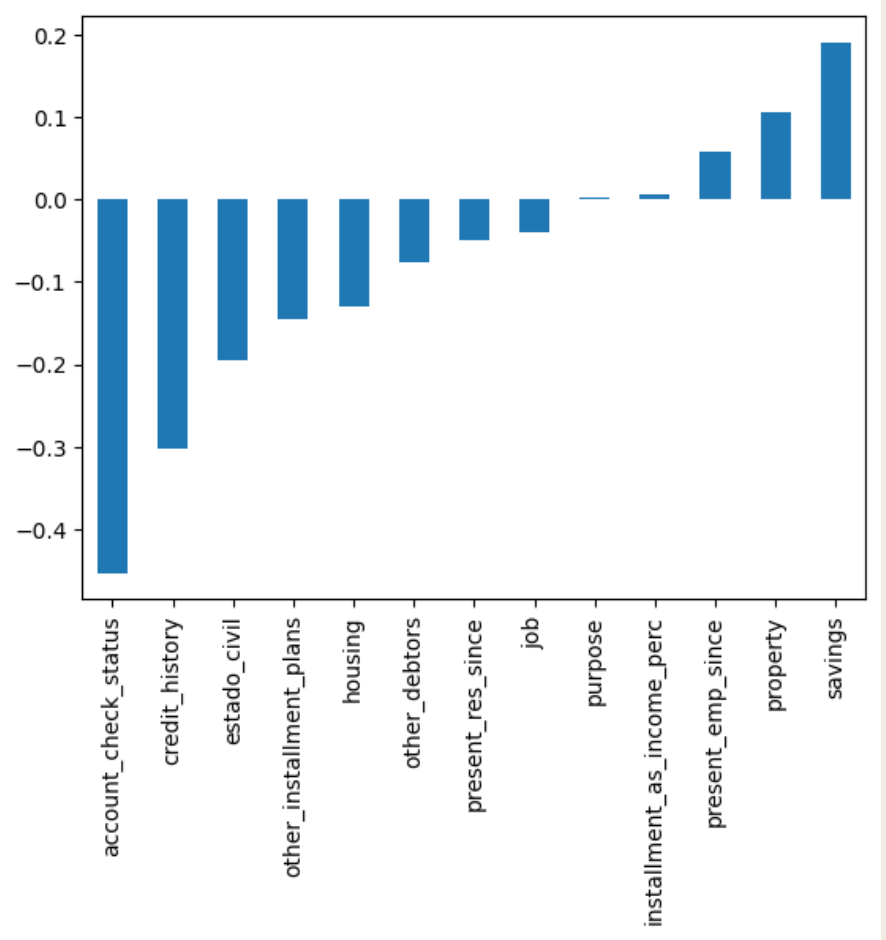
Hay un desbalanceo entre los buenos clients (0) y malos clientes (1).

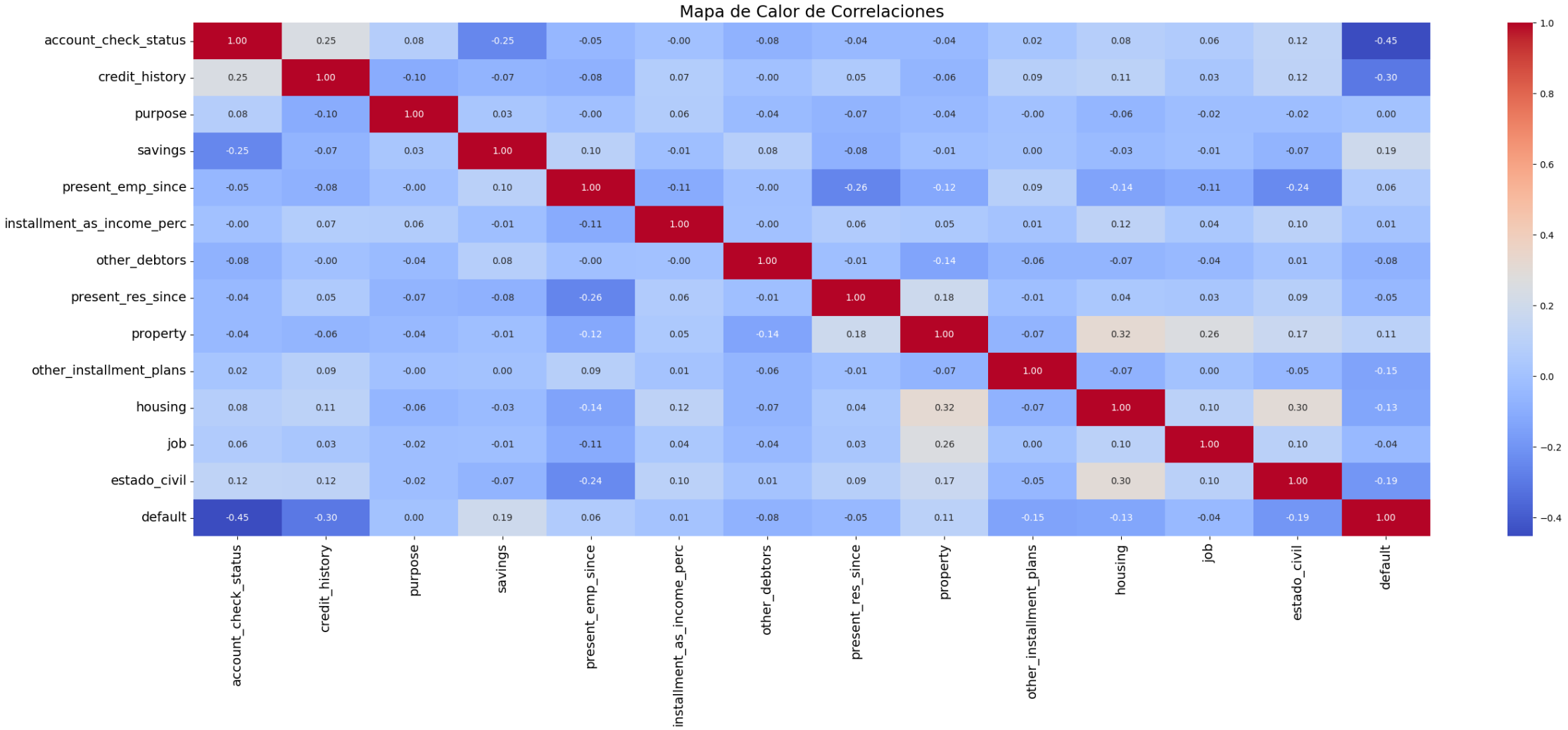
Balanceo de Datos

Se aplicó el balanceo de los datos en la columna ‘default’ con el método SMOTE, así mismo se eliminaron algunas características que se consideraron no relevantes en base al mejor modelo.



Gráfica de correlación de variables respecto a Default




Mapa de Calor de Correlaciones
account_check_status 1.00 0.25 0.08 -0.25 -0.05 -0.00 -0.08 -0.04 -0.04 0.02 0.08 0.06 0.12 -0.45
credit_history 0.25 1.00 -0.10 -0.07 -0.08 0.07 -0.00 0.05 -0.06 0.09 0.11 0.03 0.12 -0.30
purpose 0.08 -0.10 1.00 0.03 -0.00 0.06 -0.04 -0.07 -0.04 -0.00 -0.06 -0.02 -0.02 0.00
savings -0.25 -0.07 0.03 1.00 0.10 -0.01 0.08 -0.08 -0.01 0.00 -0.03 -0.01 -0.07 0.19
present_emp_since -0.05 -0.08 -0.00 0.10 1.00 -0.11 -0.00 -0.26 -0.12 0.09 -0.14 -0.11 -0.24 0.06
installment_as_income_perc -0.00 0.07 0.06 -0.01 -0.11 1.00 -0.00 0.06 0.05 0.01 0.12 0.04 0.10 0.01
other_debtors -0.08 -0.00 -0.04 0.08 -0.00 -0.00 1.00 -0.01 -0.14 -0.06 -0.07 -0.04 0.01 -0.08
present_res_since -0.04 0.05 -0.07 -0.08 -0.26 0.06 -0.01 1.00 0.18 -0.01 0.04 0.03 0.09 -0.05
property -0.04 -0.06 -0.04 -0.01 -0.12 0.05 -0.14 0.18 1.00 -0.07 0.32 0.26 0.17 0.11
other_installment_plans 0.02 0.09 -0.00 0.00 0.09 0.01 -0.06 -0.01 -0.07 1.00 -0.07 0.00 -0.05 -0.15
housing 0.08 0.11 -0.06 -0.03 -0.14 0.12 -0.07 0.04 0.32 -0.07 1.00 0.10 0.30 -0.13
job 0.06 0.03 -0.02 -0.01 -0.11 0.04 -0.04 0.03 0.26 0.00 0.10 1.00 0.10 -0.04
estado_civil 0.12 0.12 -0.02 -0.07 -0.24 0.10 0.01 0.09 0.17 -0.05 0.30 0.10 1.00 -0.19
default -0.45 -0.30 0.00 0.19 0.06 0.01 -0.08 -0.05 0.11 -0.15 -0.13 -0.04 -0.19 1.00
account_check_status
credit_history
purpose
savings
present_emp_since
installment_as_income_perc
other_debtors
present_res_since
property
other_installment_plans
housing
job
estado_civil
default
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Construcción de Modelos

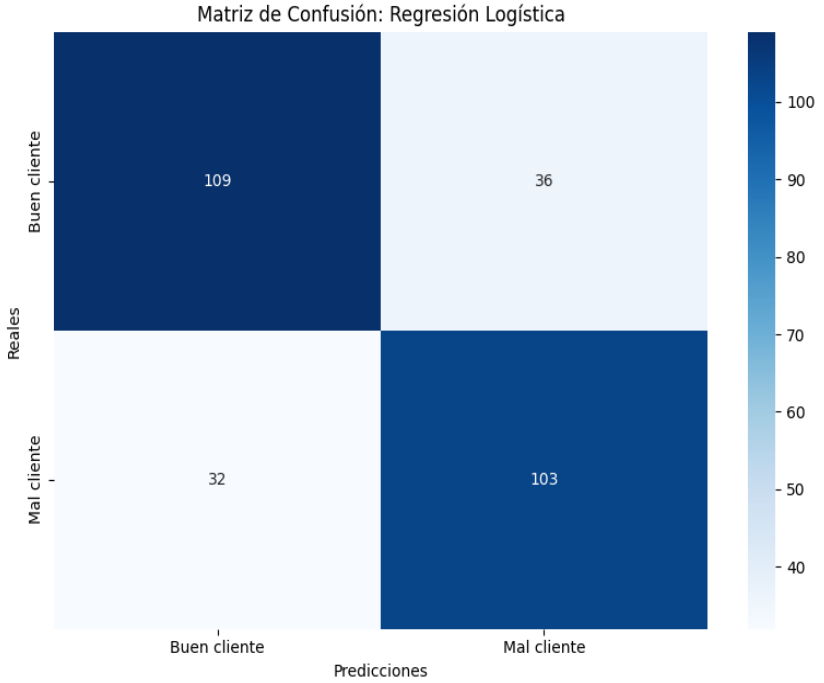
Se probaron diferentes modelos de Machine Learning para este proyecto, con los siguientes resultados en cuanto a sus métricas.

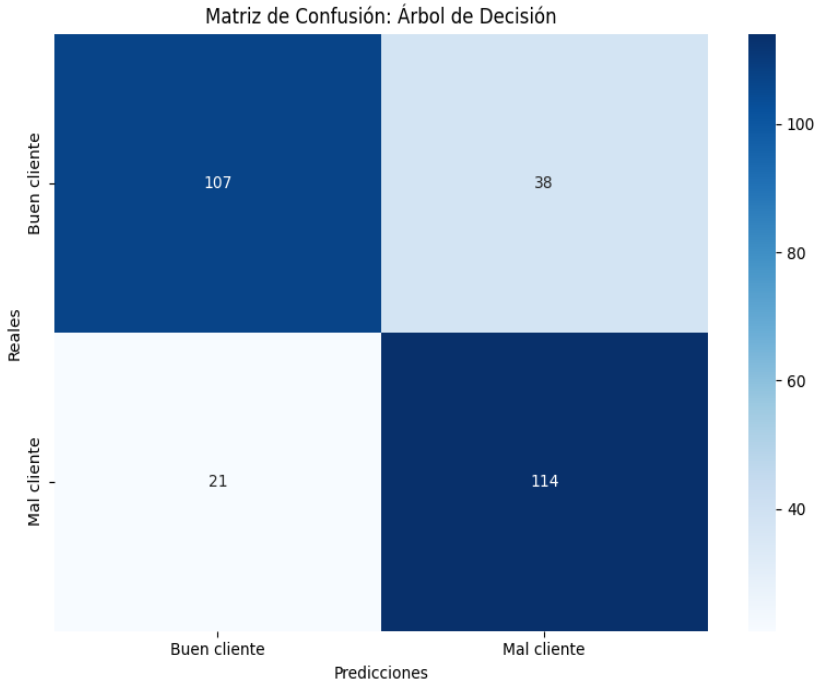
En este caso algunas métricas de interés son Accuracy, Precisión, Recall y F1 para una mejor predicción de solvencia crediticia.

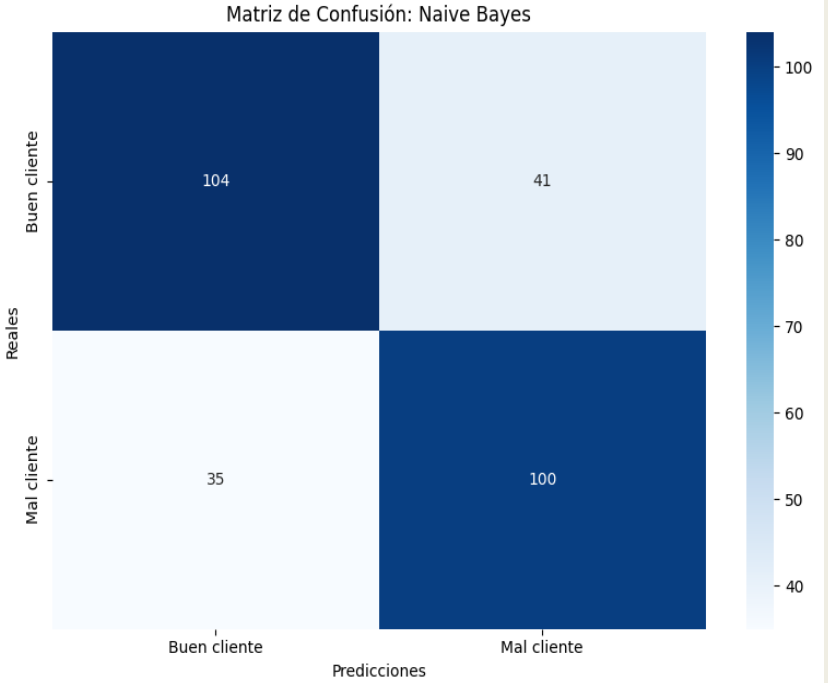
```
Modelo: Regresión Logística
Accuracy: 0.757
Precisión: 0.741
Recall: 0.763
F1_Score: 0.752
AUC-ROC: 0.757
Features importances: None

Modelo: Árbol de Decisión
Accuracy: 0.789
Precisión: 0.75
Recall: 0.844
F1_Score: 0.794
AUC-ROC: 0.791
Features importances: {'account
```

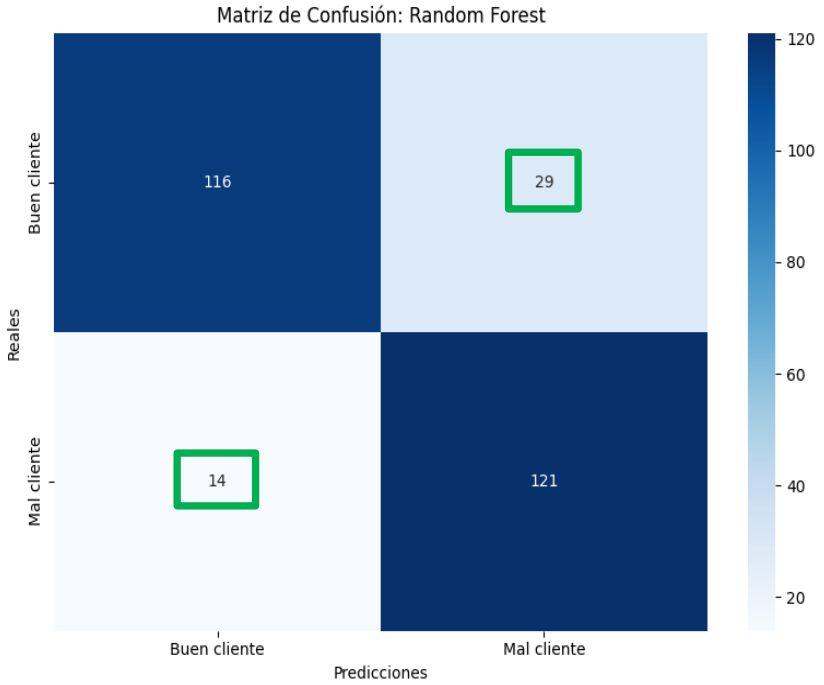
```
Modelo: Random Forest
Accuracy: 0.846
Precisión: 0.807
Recall: 0.896
F1_Score: 0.849
AUC-ROC: 0.848
Features importances: {'account

Modelo: Naive Bayes
Accuracy: 0.729
Precisión: 0.709
Recall: 0.741
F1_Score: 0.725
AUC-ROC: 0.729
Features importances: None
```

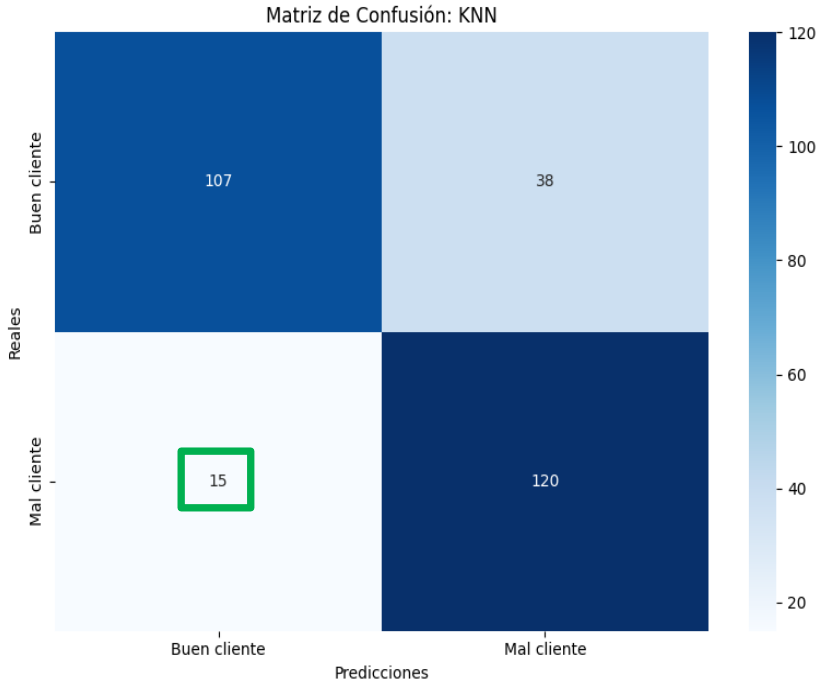
```
Modelo: KNN
Accuracy: 0.811
Precisión: 0.759
Recall: 0.889
F1_Score: 0.819
AUC-ROC: 0.813
Features importances: None

Modelo: SVM
Accuracy: 0.807
Precisión: 0.787
Recall: 0.822
F1_Score: 0.804
AUC-ROC: 0.808
Features importances: None
```

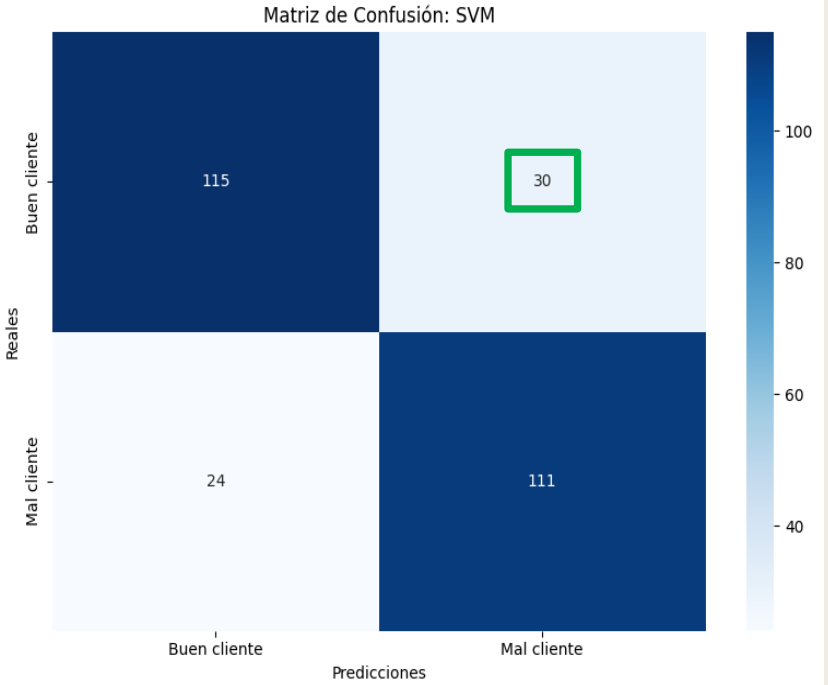

Matriz de Confusión: Regresión Logística
Buen cliente
Mal cliente
Reales
109
36
32
103
Buen cliente
Mal cliente
Predicciones
100
90
80
70
60
50
40


Matriz de Confusión: Árbol de Decisión
Buen cliente
Mal cliente
Reales
107
38
21
114
Buen cliente
Mal cliente
Predicciones
100
80
60
40

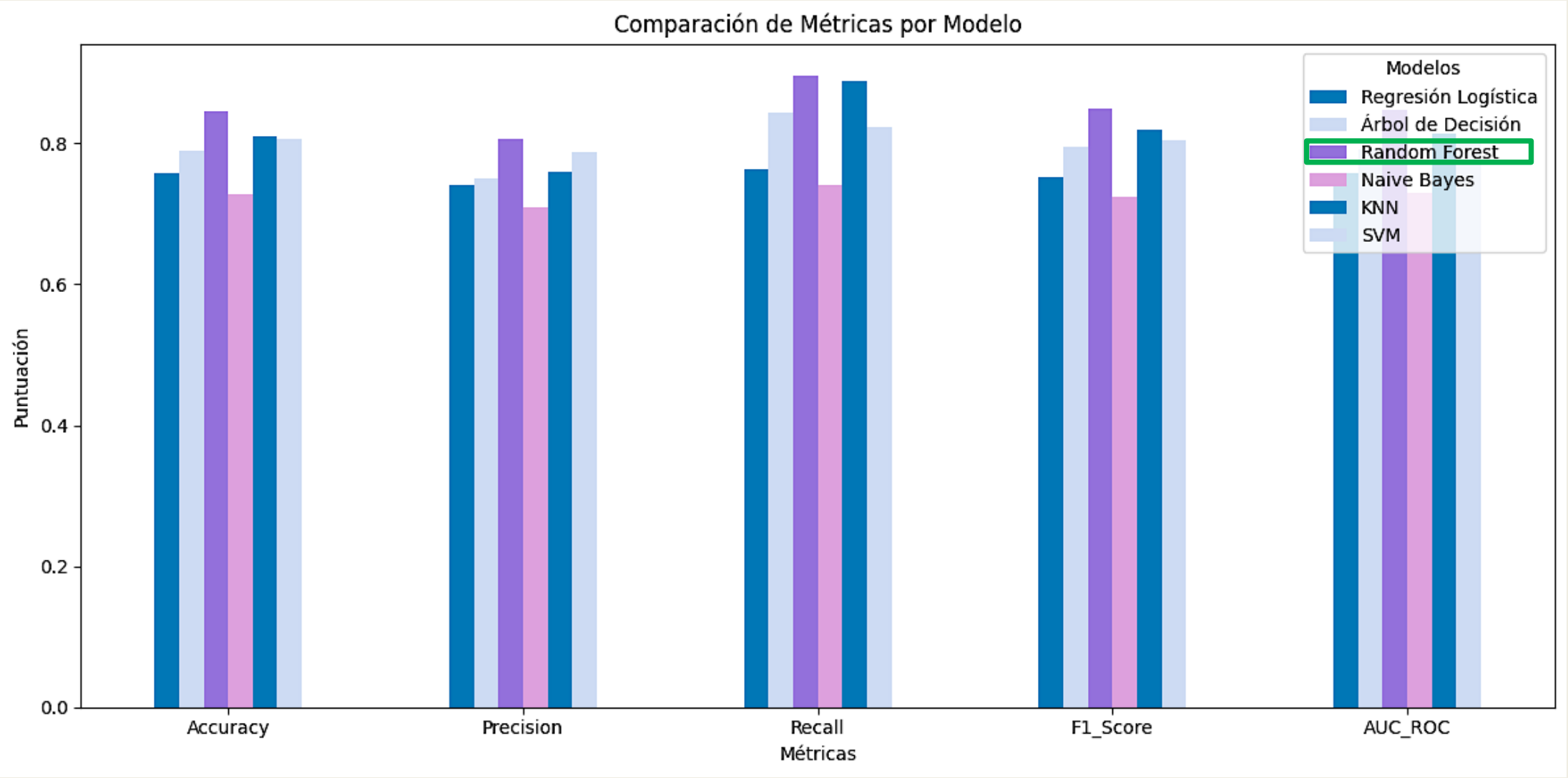

Matriz de Confusión: Naive Bayes
Buen cliente
Mal cliente
Reales
104
41
35
100
Buen cliente
Mal cliente
Predicciones
100
90
80
70
60
50
40


Matriz de Confusión: Random Forest
Buen cliente
Mal cliente
Reales
116
29
14
121
Buen cliente
Mal cliente
Predicciones
120
100
80
60
40
20


Matriz de Confusión: KNN
Buen cliente
Mal cliente
Reales
107
38
15
120
Buen cliente
Mal cliente
Predicciones
120
100
80
60
40
20


Matriz de Confusión: SVM
Buen cliente
Mal cliente
Reales
115
30
24
111
Buen cliente
Mal cliente
Predicciones
100
80
60
40

Evaluación y Selección de Modelo



Conclusiones

En este proyecto de machine learning el modelo con las mejores métricas fue **Random Forest Classifier**, obteniendo un Accuracy: 0.846, Precisión: 0.807 y Recall: 0.89, el cual lo hace el modelo con mejores métricas para este proyecto de clasificación de potenciales buenos y malos clientes. Como se observa en la matriz de confusión es el modelo que menos se equivoca en general al clasificar clientes.