PROYECTO ML : CREDIT
Autor :Kino Galvez

User applies via website

User details are sent to the bank

Bank verifies details and credit score before creating a report.

Credit is approved or disapproved

User is informed about status of credit

CONTENIDO:

1.INTRO

2.LIMPIEZA DATOS

3.APROXIMACION Y TECNICAS

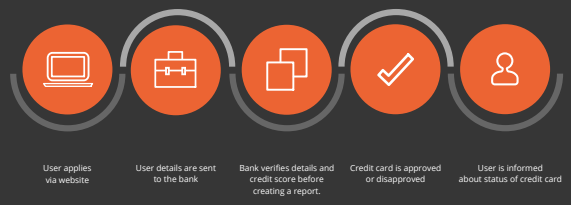
4.MODELOS ENTRENADOS

5.PROCESO ENTRENAMIENTO

6.RESULTADOS-MEJOR PERFOMANCE

7.CONCLUSIONES

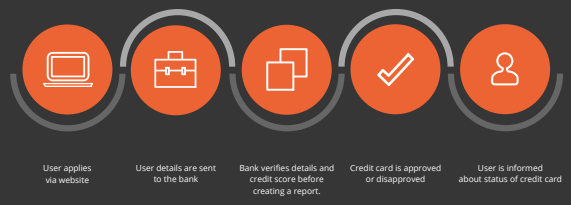
8.ANEXO-STREAMLIT



1. INTRO

Se trata de un proyecto,basado en datos de LendingClub, para predecir si un préstamo según sus características iniciales, sera pagado o NO.

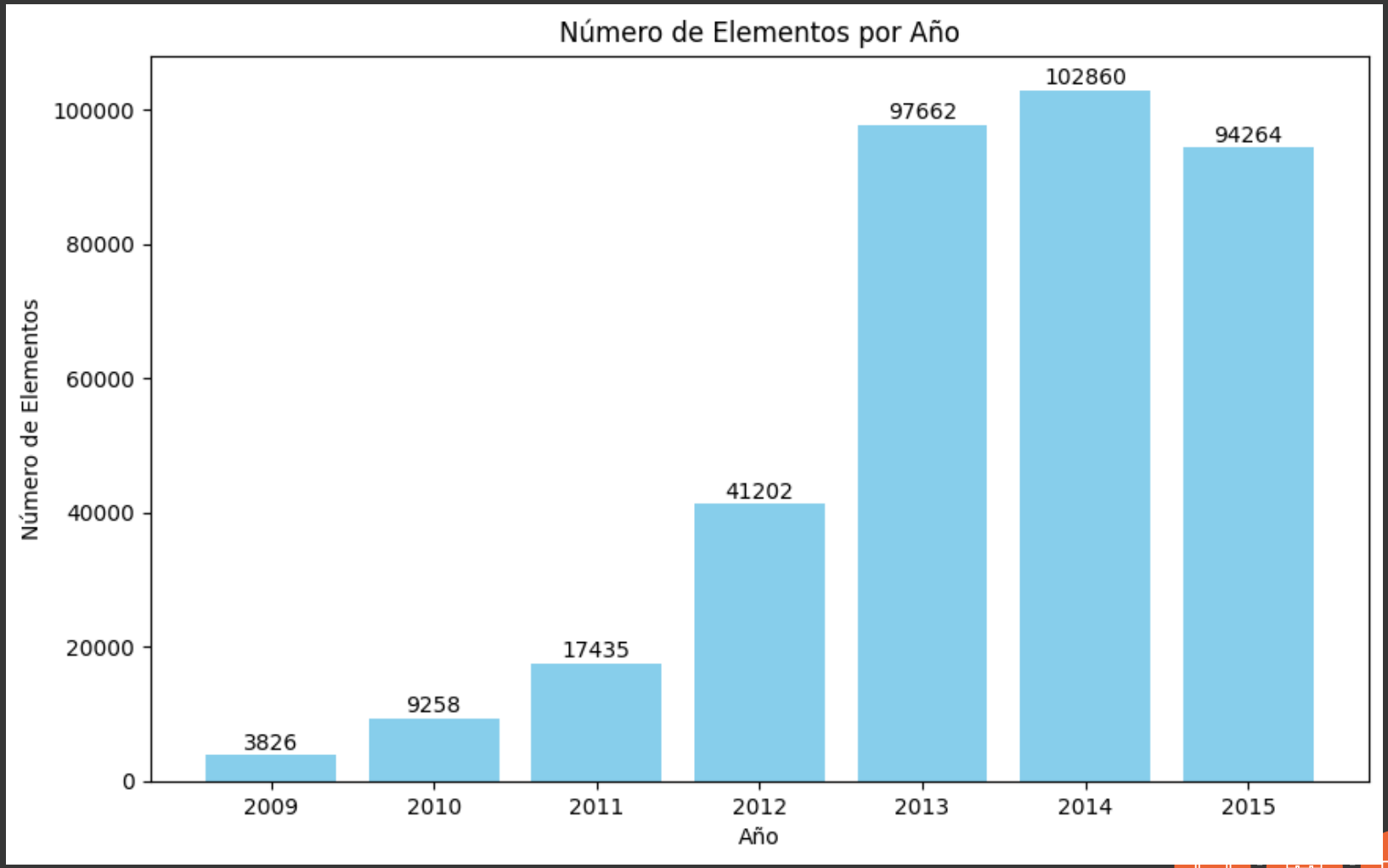
Abordaremos la toma de decisiones de préstamos, minimizando riesgos y maximizando la eficiencia del negocio mediante el análisis de datos pasados y la aplicación de modelos predictivos.



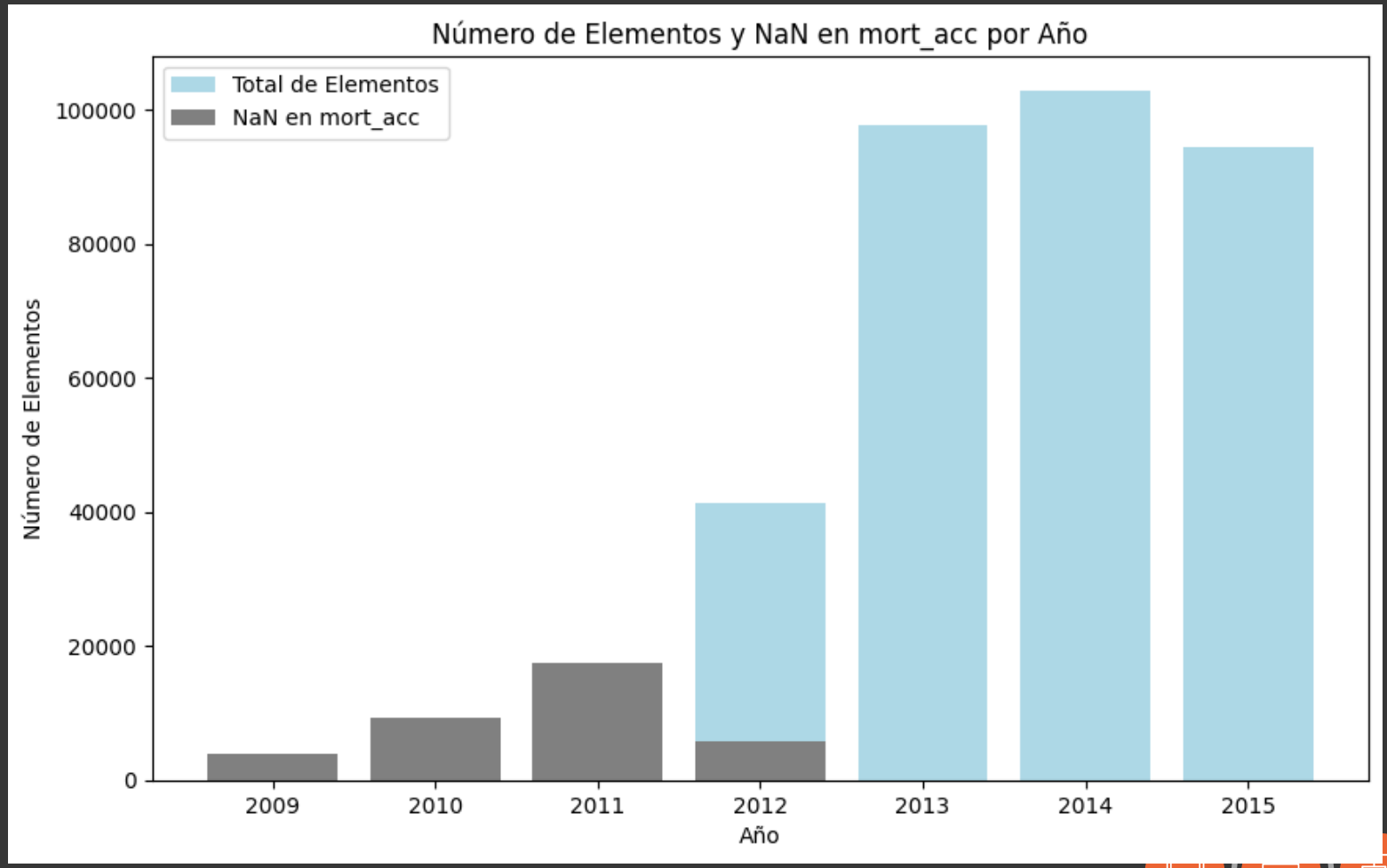
2.LIMPIEZA DE DATOS
MUCHAS COLUMNAS-FEATURES 27

 # loan_amnt: Monto del préstamo solicitado.
 # term: Plazo del préstamo en meses.
 # int_rate: Tasa de interés del préstamo.
 # installment: Pago mensual del préstamo.
 # grade: Grado asignado al préstamo por LendingClub (A, B, C, etc.).
 # sub_grade: Subgrado asignado al préstamo por LendingClub (A1, A2, B1, etc.).
 # emp_title: Título laboral del prestatario.
 # emp_length: Antigüedad laboral del prestatario en años.
 # home_ownership: Estado de propiedad de la vivienda del prestatario (RENT, OWN, MORTGAGE, OTHER).
 # annual_inc: Ingreso anual del prestatario.
 # verification_status: Estado de verificación del ingreso del prestatario.
 # issue_d: Fecha en la que se emitió el préstamo.
 # loan_status: Estado actual del préstamo (Fully Paid, Charged Off, etc.).
 # purpose: Propósito del préstamo.
 # title: Título del préstamo proporcionado por el prestatario.
 # dti: Relación entre las deudas y el ingreso del prestatario.
 # earliest_cr_line: Fecha en que se abrió la primera línea de crédito del prestatario.
 # open_acc: Número de líneas de crédito abiertas en el archivo del prestatario.
 # pub_rec: Número de registros públicos desfavorables.
 # revol_bal: Saldo total de las cuentas de crédito renovable.
 # revol_util: Tasa de utilización de las cuentas de crédito renovable.
 # total_acc: Número total de cuentas de crédito del prestatario.
 # initial_list_status: Estado inicial de la lista del préstamo (W, F).
 # application_type: Tipo de aplicación (INDIVIDUAL, JOINT).
 # mort_acc: Número de cuentas hipotecarias.
 # pub_rec_bankruptcies: Número de quiebras en los registros públicos.
 # address: Dirección del prestatario.

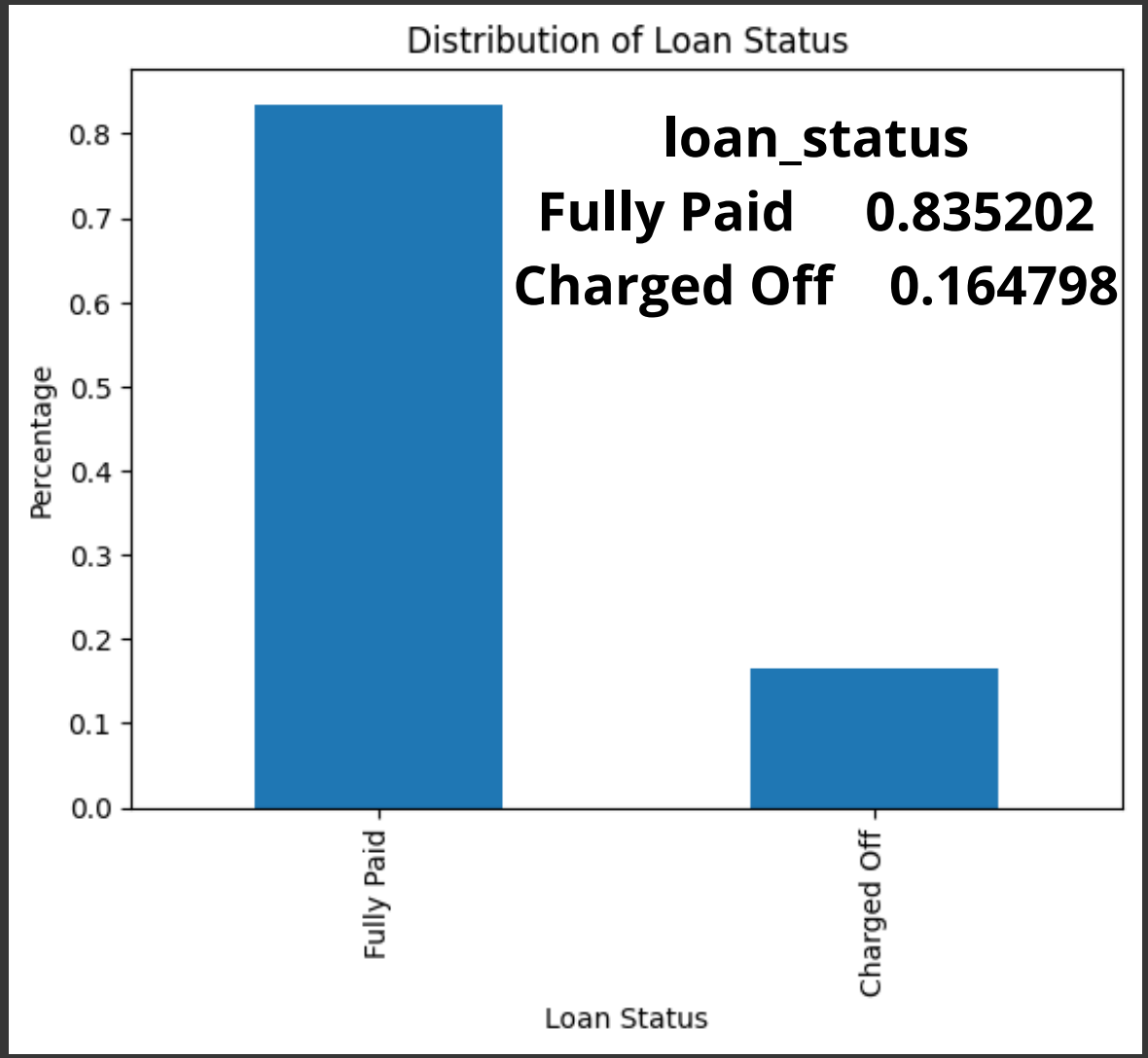
2.LIMPIEZA DE DATOS ELECCION DEL AÑO: 2012



LIMPIEZA DE DATOS
ELECCION DEL AÑO: 2012



2.LIMPIEZA DE DATOS
DESBALANCEO DE LA MUESTRA



2.LIMPIEZA DE DATOS DISCRETIZACION CATEGORICAS o DROP

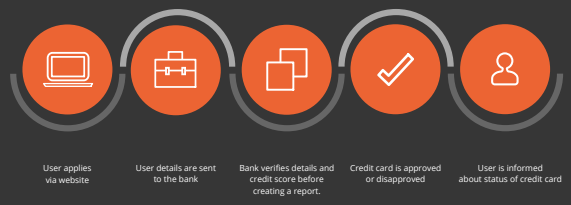
	emp_title	emp_length	home_ownership	purpose	pub_rec	total_acc	initial_list_status	mort_acc	pub_rec_bankruptcies
	30985	12	5	13	6	71	2	23	5

2.LIMPIEZA DE DATOS
DISCRETIZACION CATEGORICAS o DROP

 # Address

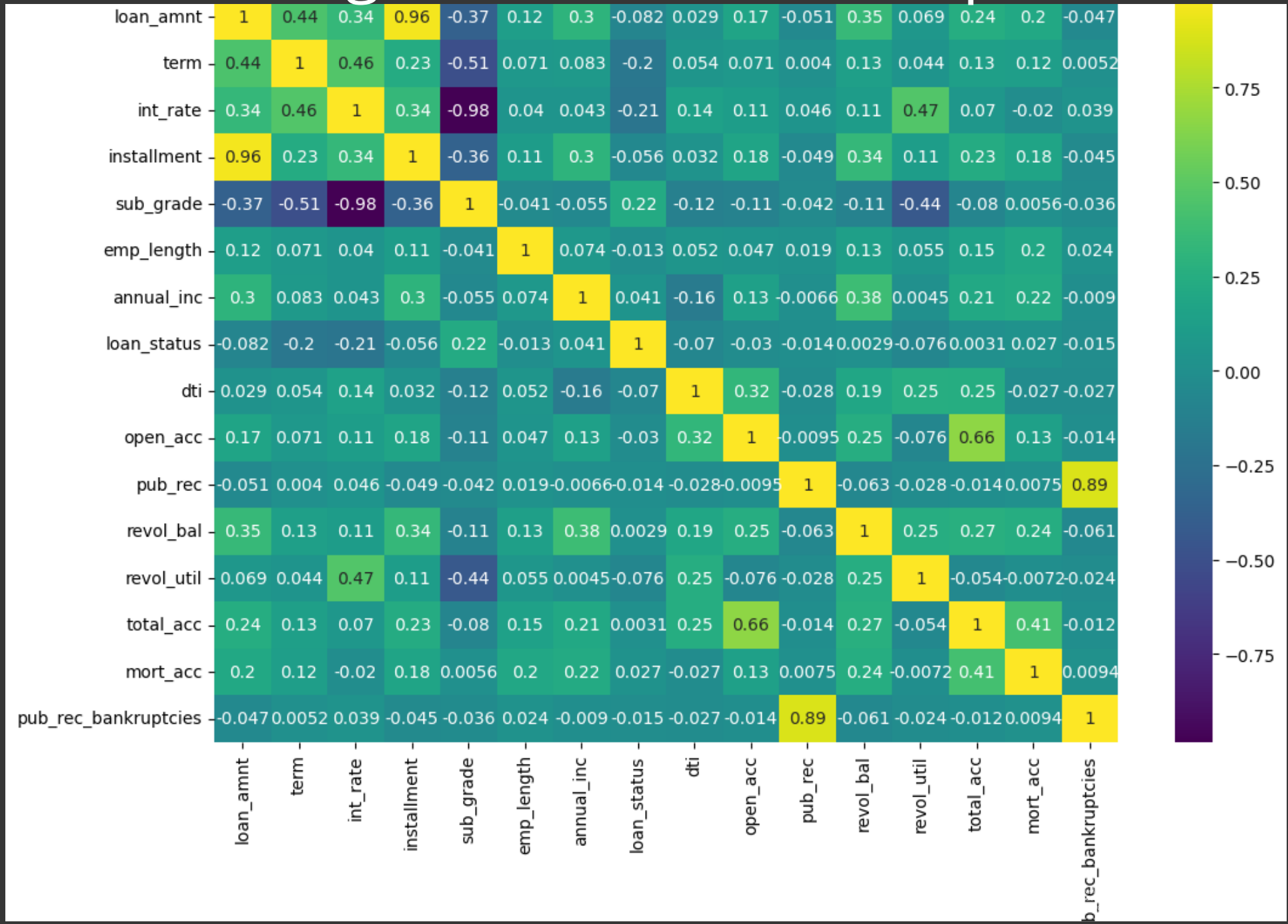
 # La linea mas temprana en el historial crediticio del usuario

 # issue_d fecha de emision



2.LIMPIEZA DE DATOS

'sub_grade','loan_amnt','open_acc'



2.LIMPIEZA DE DATOS

traspasar a cada categoría su predisposición a ser 1 o 0, mediante las medias de la columna target, de cada una de esas categorías.

2.LIMPIEZA DE DATOS

traspasar a cada categoría su predisposición a ser 1 o 0, mediante las medias de la columna target, de cada una de esas categorías.

3.APROXIMACION Y TECNICAS

Resolución como un problema de clasificación.

Uso de técnicas de aprendizaje supervisado Y no supervisado.

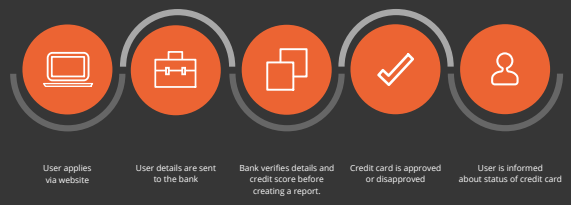
4.MODELOS ENTRENADOS

Regresión Logística.
Árbol de Decisión.
Random Forest Classifier.
Gradient Boosting Classifier.
SVM.
Kmeans-CLUSTER(no supervisado)

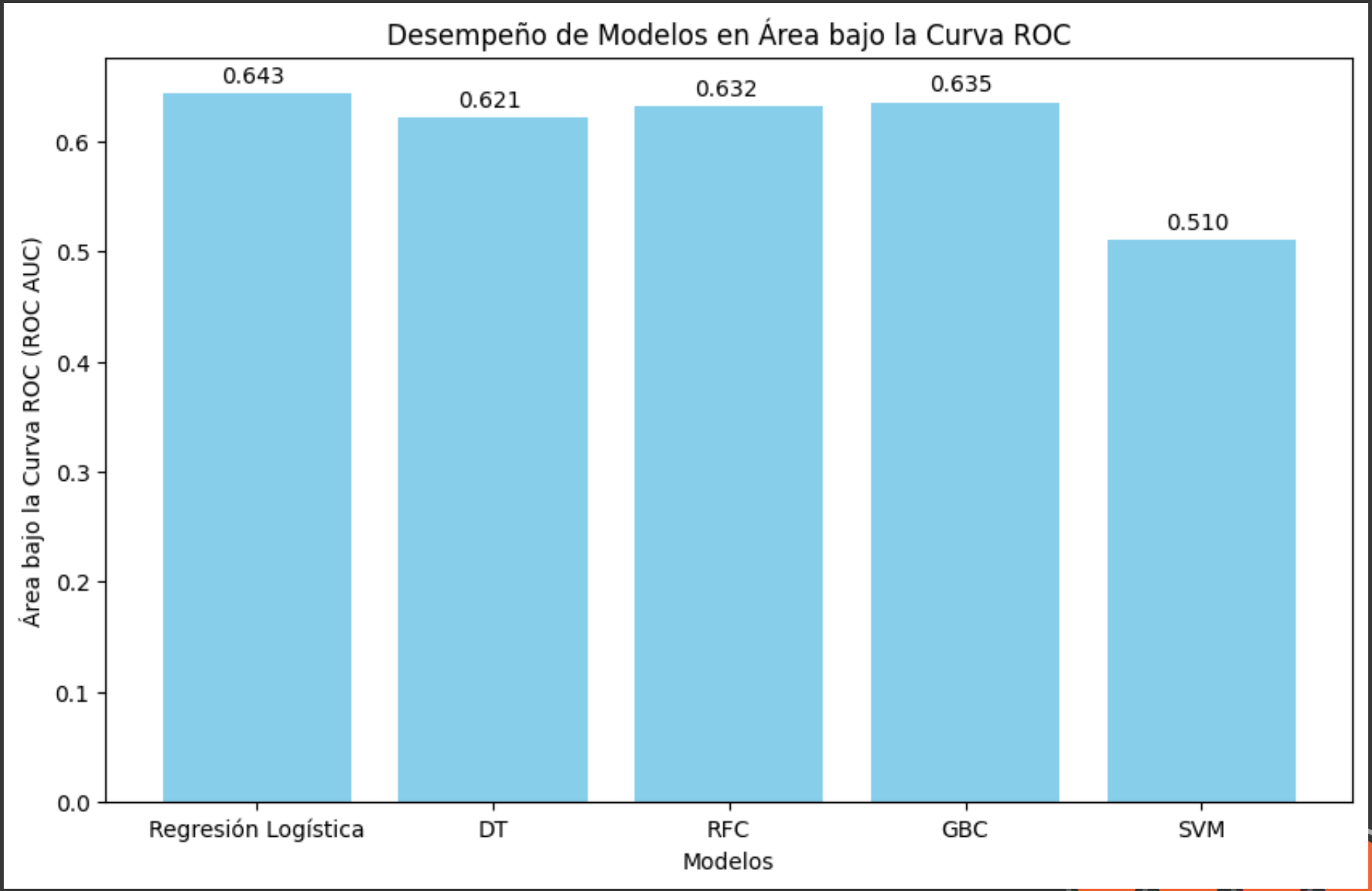
5.PROCESO ENTRENAMIENTO

Optimización de hiperparámetros con GridSearchCV.

Validación cruzada para evaluar el rendimiento.



6.RESULTADOS-MEJOR PERFOMANCE



6.RESULTADOS-MEJOR PERFOMANCE

GBC-PARAMs

learning_rate: 0.1
max_depth: 2
max_features: 3
n_estimators: 150

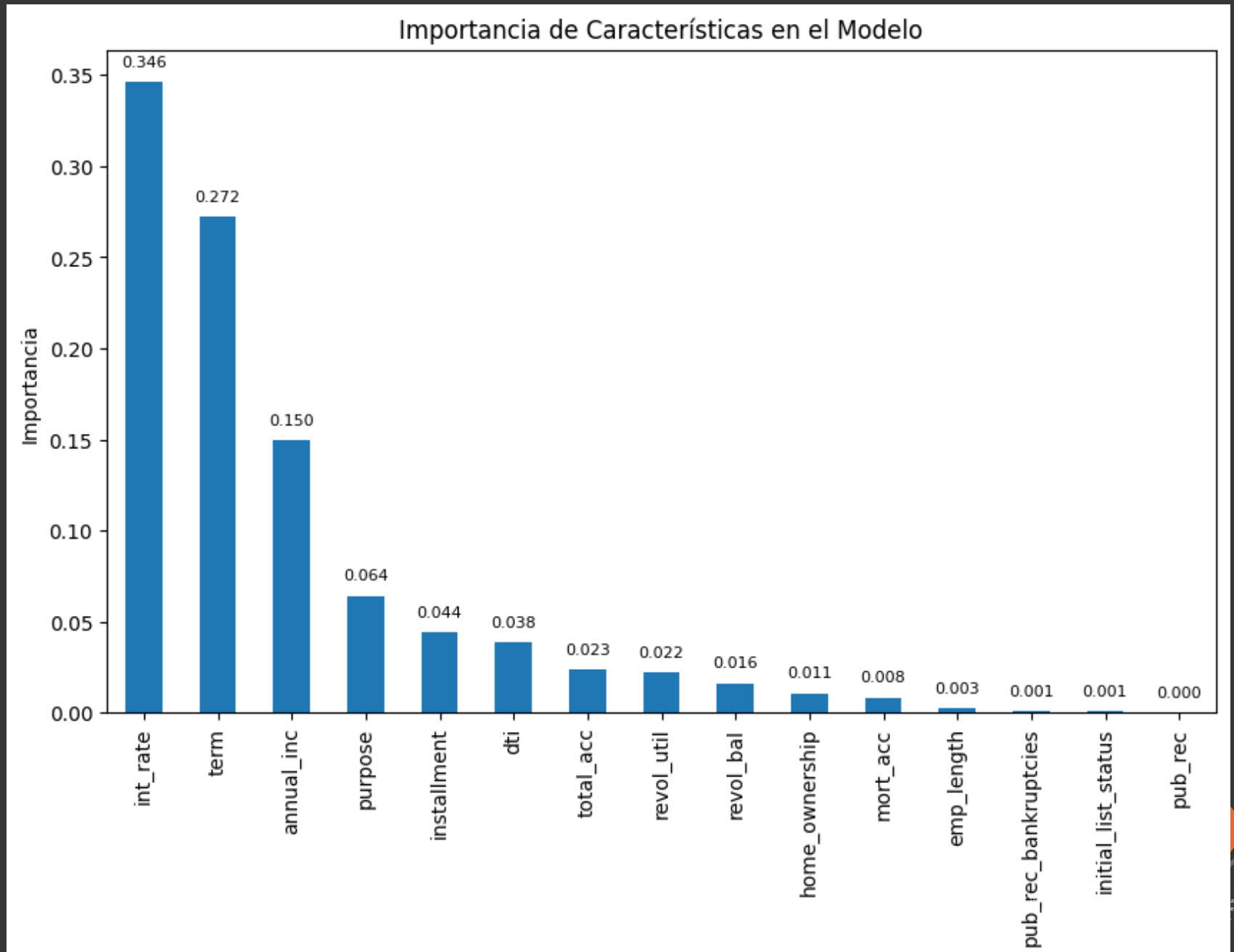
Accuracy: 0.635 (63.5% de predicciones correctas).
Precision: 0.635 (63.5% de verdaderos positivos entre los positivos predichos).
Recall: 0.643 (64.3% de positivos reales identificados).
F1 Score: 0.639 (63.9% de equilibrio entre precision y recall).
ROC AUC Score: 0.635 (Área bajo la curva ROC).

[0.70031627, 0.6918349, 0.6872482, 0.70729893, 0.7042142, 0.67790482,
0.68390895, 0.69112586, 0.6837893, 0.66525156]

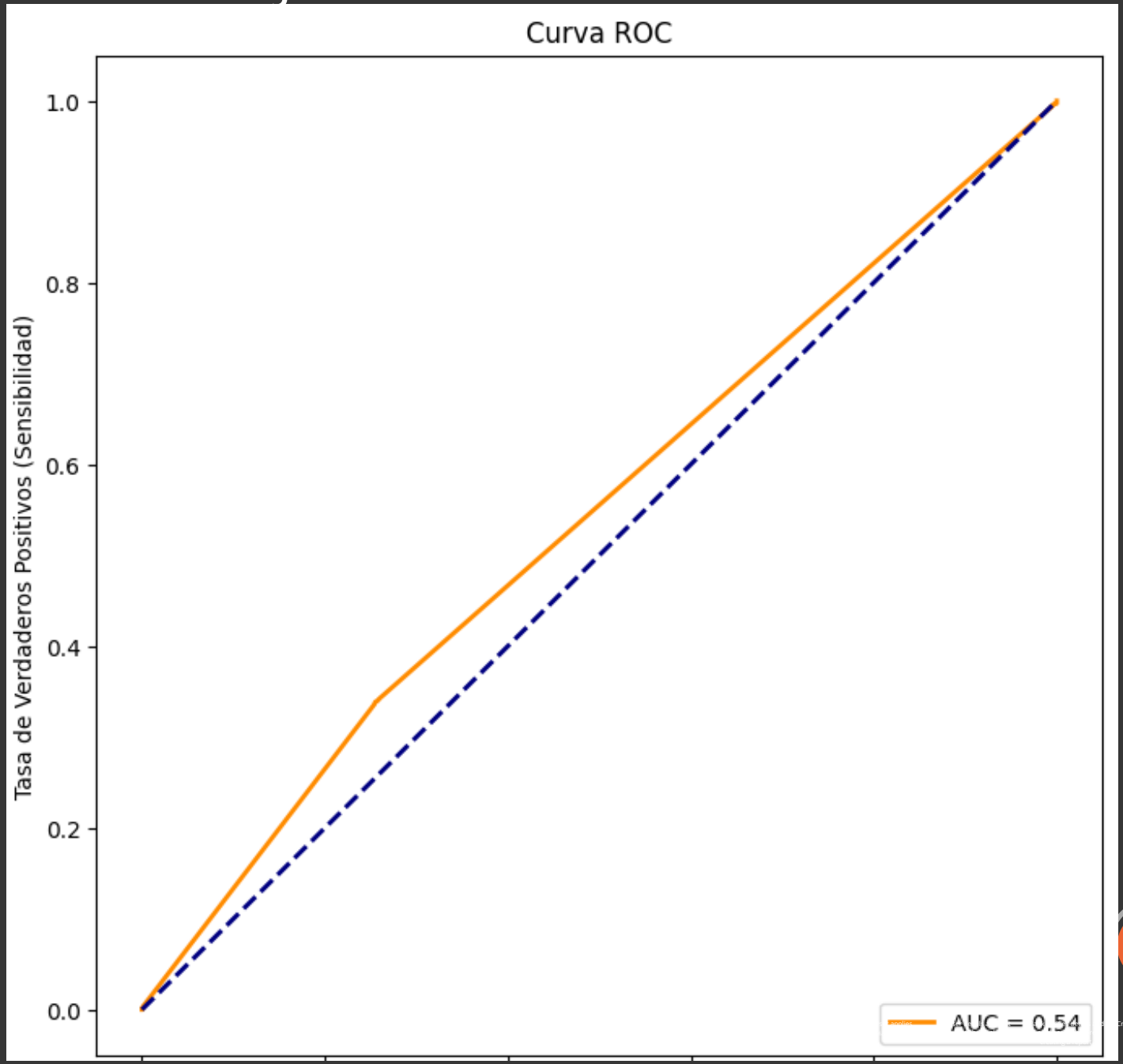
Promedio de ROC AUC Scores en Validación Cruzada: 0.689

Desviación Estándar de ROC AUC Scores en Validación Cruzada: 0.012

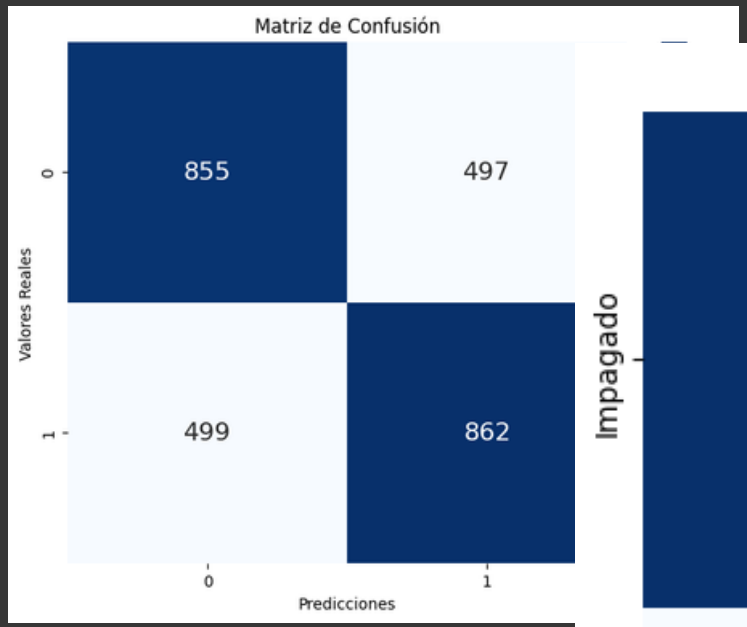
6.RESULTADOS-MEJOR PERFOMANCE

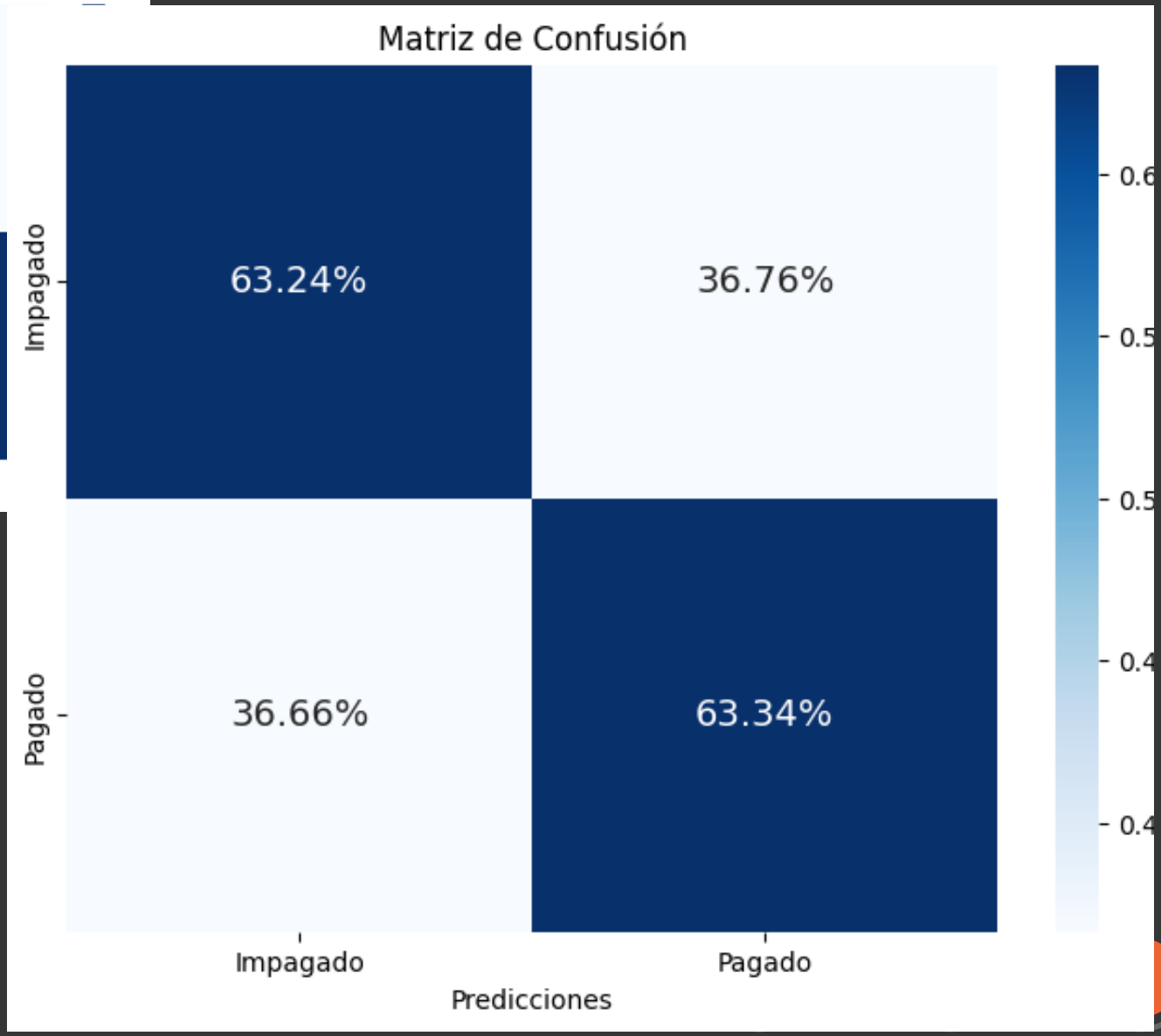


6.RESULTADOS-MEJOR PERFOMANCE



6.RESULTADOS-MEJOR PERFOMANCE





7.CONCLUSIONES

Conclusiones:

GBC optimizado supera a otros modelos.
Importancia de la optimización de hiperparámetros.
Buen equilibrio entre precision y recall.

7.1.RECOMENDACIONES PARA PROX ITERs

Explorar Nuevas Características:

Investigar la inclusión de nuevas características o la ingeniería
de características podría mejorar aún más el rendimiento del modelo.

Evaluar Modelos Ensemble:

Considerar la implementación de modelos ensemble más complejos
o la combinación de varios modelos para explorar sinergias y mejorar la robustez del sistema.

Monitoreo Continuo:

Establecer un proceso de monitoreo continuo del modelo en un entorno de producción
para evaluar su rendimiento a medida que se reciben nuevos datos.

Interpretación de Resultados:

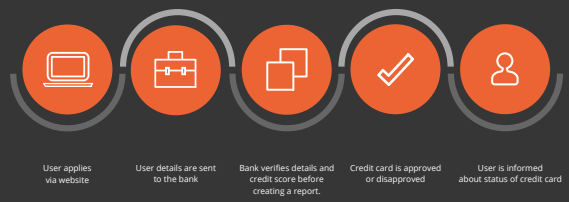
Profundizar en la interpretación de las características más relevantes identificadas
por el GBC para obtener una comprensión más detallada de los factores que influyen en las predicciones.

Considerar Datos Adicionales:

Evaluar la posibilidad de incorporar datos adicionales o mejorar la calidad
de los datos existentes para proporcionar al modelo una información más completa y precisa.

Optimización Continua de Hiperparámetros:

Realizar ajustes adicionales en la búsqueda de hiperparámetros para asegurar
que el modelo esté completamente optimizado.



ANEXO

STREAMLIT