"EVALUACIÓN DEL RIESGO CREDITICIO UTILIZANDO APRENDIZAJE AUTOMÁTICO"

PROYECTO FINAL DATA SCIENCE II

MARITZA QUINTERO SIRITT

MOTIVACIÓN OBJETIVO Y AUDIENCIA

El proyecto se centra en analizar datos históricos sobre préstamos, incluyendo información sobre el género del prestatario, el propósito y tipo del préstamo, y características financieras relevantes.

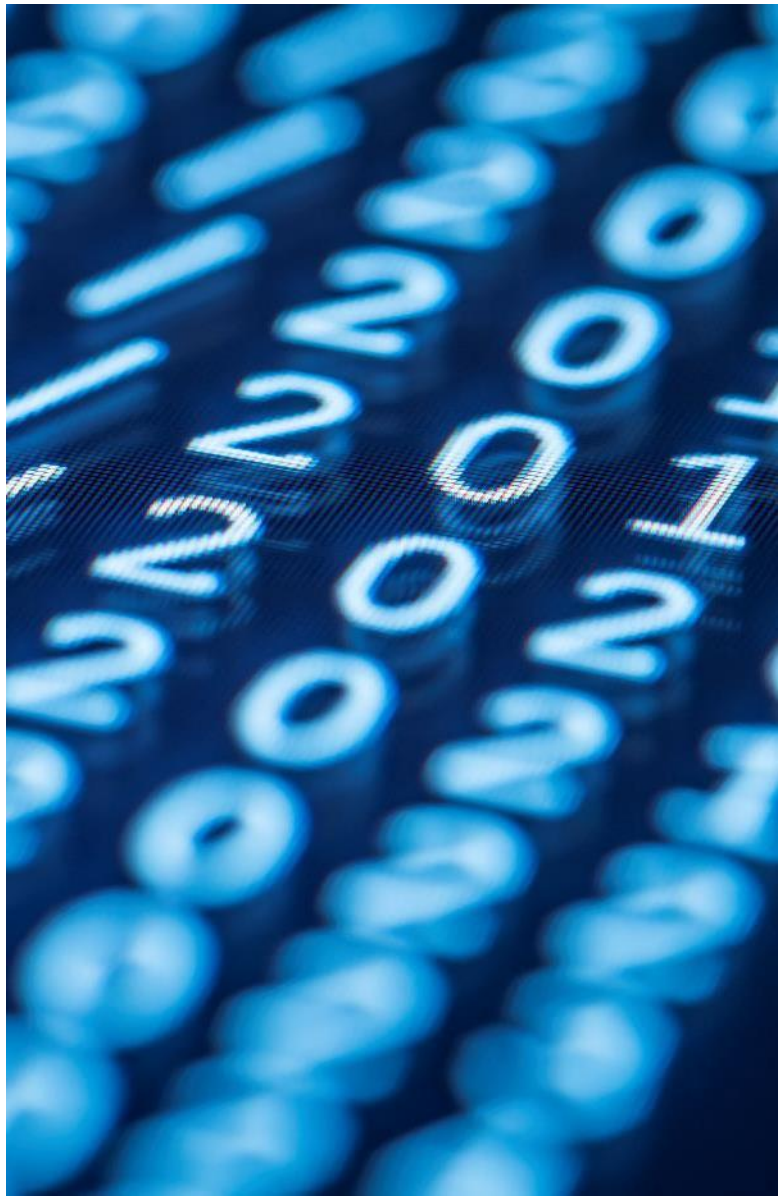
El objetivo principal es desarrollar un modelo de Aprendizaje Automático que pueda predecir con alta precisión si un préstamo será pagado o incumplido.

Este análisis ayudará a las instituciones financieras a mitigar los riesgos asociados con la concesión de préstamos, optimizar su cartera y reducir la tasa de incumplimiento, mejorando la toma de decisiones financieras informadas.

La audiencia objetivo son gerentes de riesgo financiero, analistas de datos y ejecutivos de instituciones financieras.

HIPÓTESIS POR RESPONDER

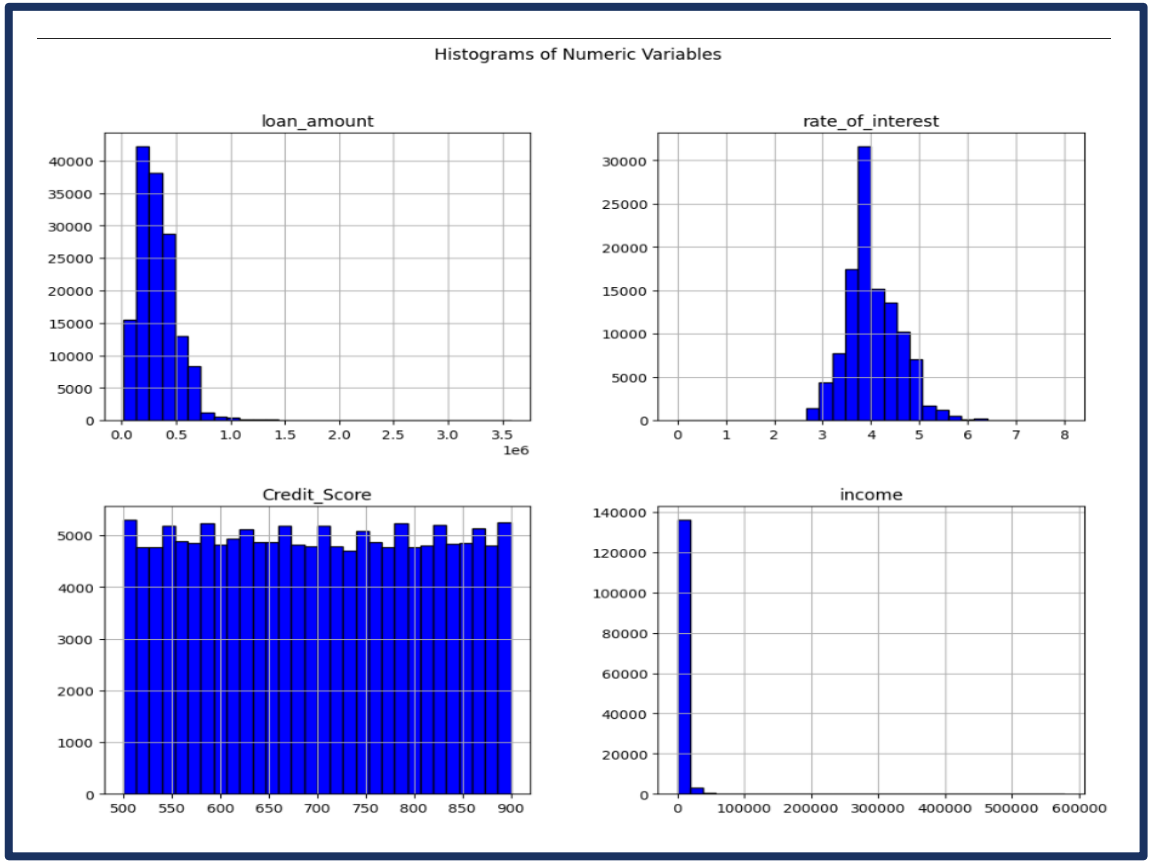
- ¿Podemos predecir si un cliente incumplirá con el pago de un préstamo basado en sus características financieras y del préstamo?
  - Se plantea que, al analizar los atributos de los préstamos, tales como el monto, tipo, propósito del préstamo, y la solvencia crediticia del prestatario, junto con otros factores relevantes, se podrá desarrollar un modelo de Aprendizaje Automático capaz de predecir con alta precisión si un prestatario incumplirá con su préstamo.
  - Este modelo proporcionará información valiosa para la toma de decisiones estratégicas, mejorando la capacidad de la institución financiera para evaluar la rentabilidad y el riesgo crediticio de manera más eficiente.



ANÁLISIS EXPLORATORIO DE LOS DATOS (EDA)

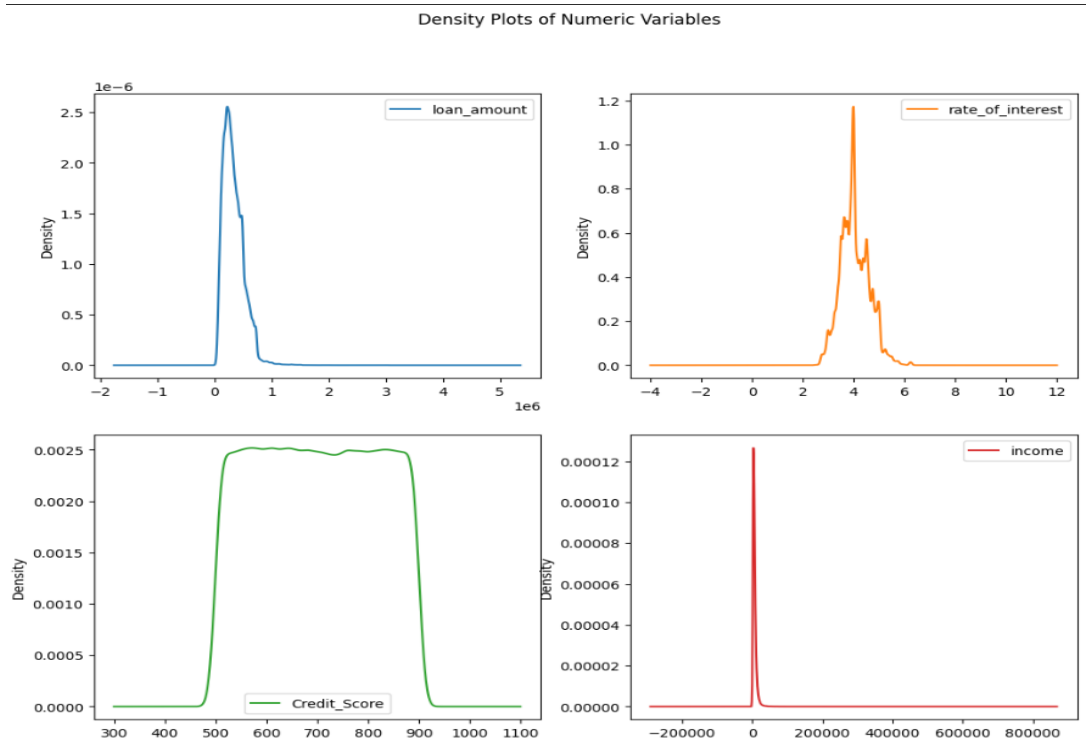
- CONTENIDO:
  - Dataset de préstamos con 34 columnas y 148,670 entradas.
  - Sin valores duplicados.
  - Verificaciones iniciales: valores nulos, duplicados, y tipos de datos.
  - Variables clave: loan_amount, rate_of_interest, credit_score, income.
  - Total de filas con valores nulos: 50.483
  - Porcentaje de filas con valores nulos: 33.96%

HISTOGRAMAS



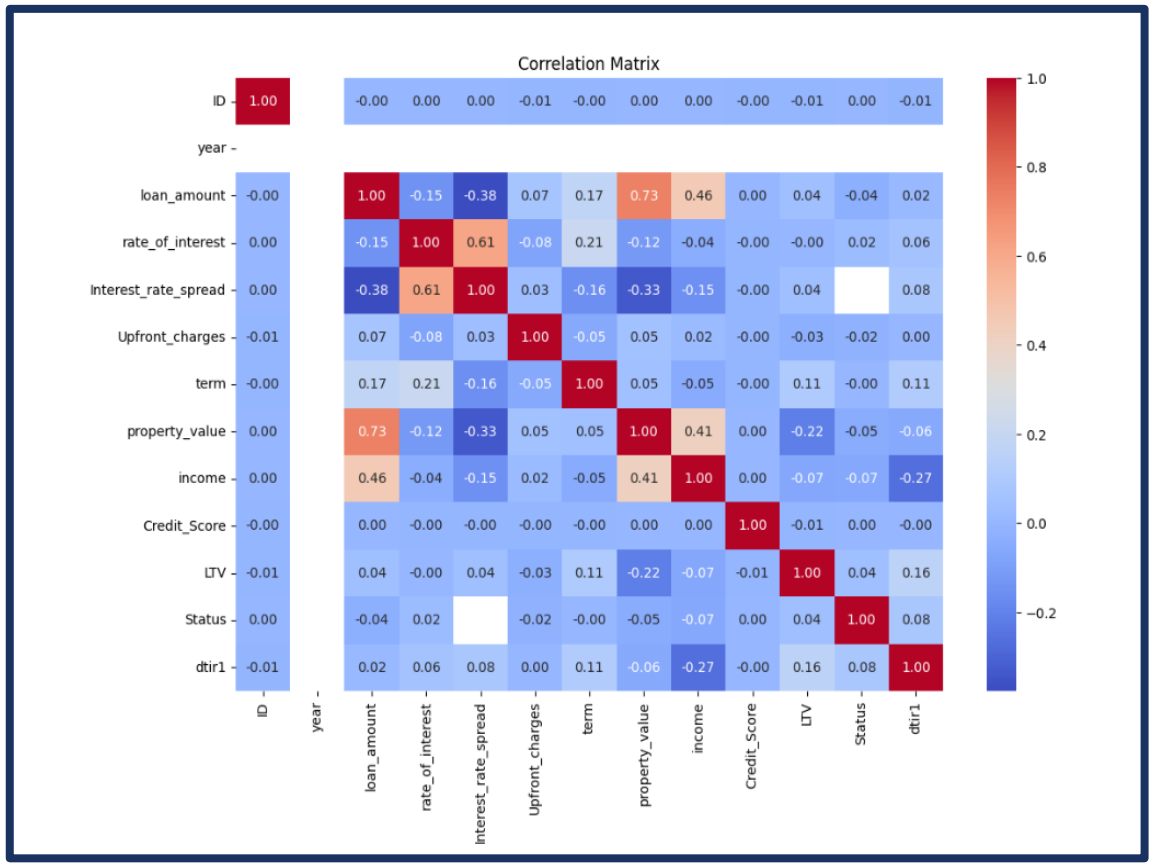
- El análisis de los histogramas muestra lo siguiente:
  - loan_amount: La mayoría de los préstamos están entre 0 y 500,000.
  - rate_of_interest: Las tasas de interés se concentran entre 3 y 5.
  - Credit_Score: Los puntajes de crédito están distribuidos de manera uniforme entre 500 y 900.
  - income: La mayoría de los ingresos son bajos, con una distribución sesgada hacia valores cercanos a 0.

GRÁFICOS DE DENSIDAD



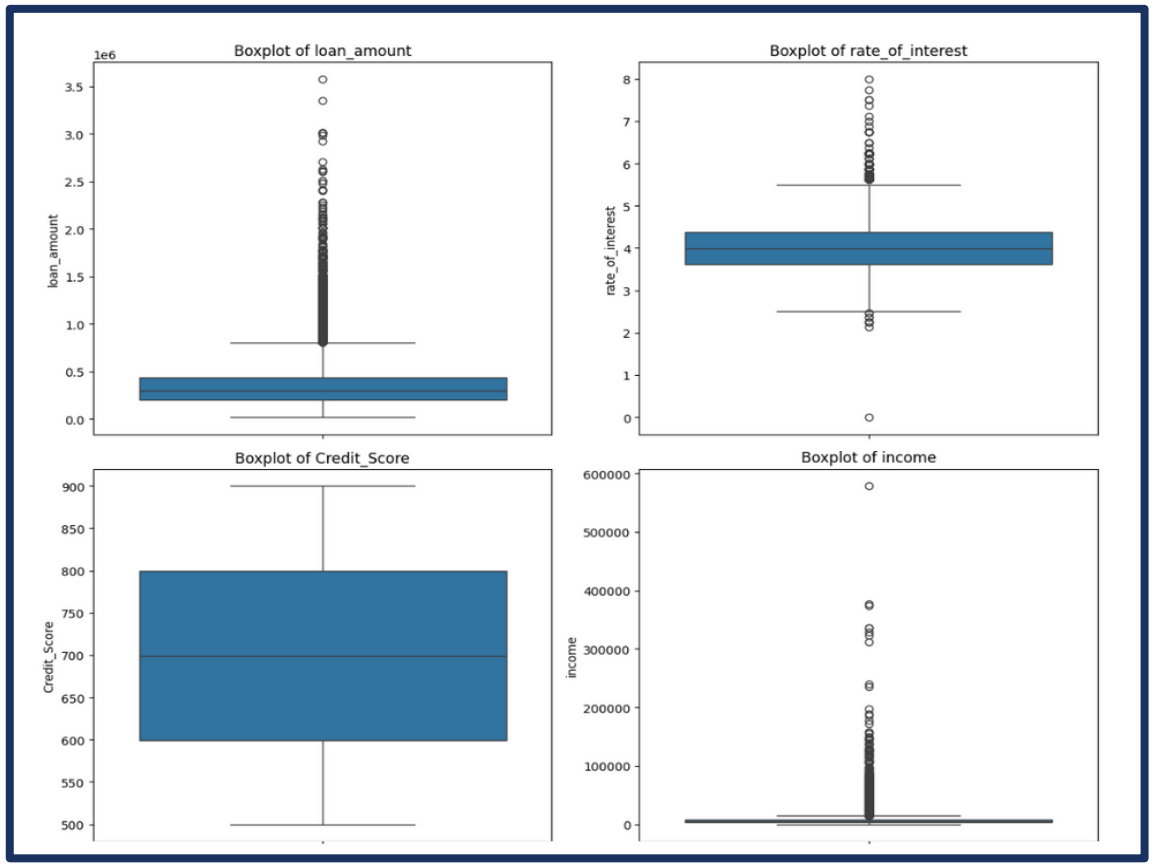
- Este gráfico muestra las curvas de densidad donde:
  - loan_amount: La densidad está fuertemente concentrada en préstamos bajos, con un pico alrededor de los 500,000.
  - rate_of_interest: La distribución tiene una forma normal con un pico en torno a 4.
  - Credit_Score: La densidad muestra una distribución casi uniforme entre 500 y 900, lo que respalda la observación anterior de una distribución balanceada de puntajes de crédito.
  - income: La densidad está altamente sesgada hacia 0, lo que refuerza que la mayoría de los ingresos reportados son muy bajos.

MATRIZ DE CORRELACIÓN

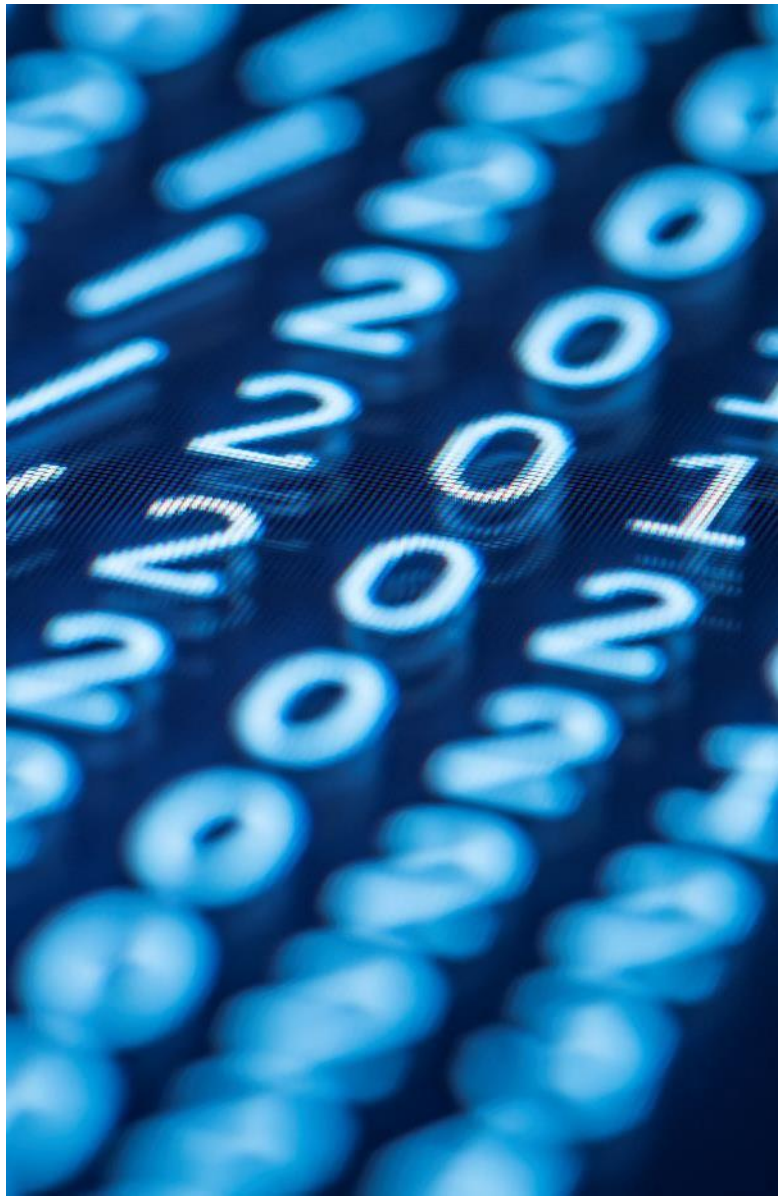


- Este gráfico muestra las relaciones entre diferentes variables numéricas del conjunto de datos.
  - loan_amount tiene una fuerte correlación positiva con property_value (0.73), lo que indica que, a mayor valor de propiedad, mayor es el monto del préstamo solicitado.
  - rate_of_interest tiene una correlación positiva moderada con Interest_rate_spread (0.61), lo que sugiere que las variaciones en la tasa de interés están relacionadas con los márgenes de interés.
  - income muestra una correlación positiva con loan_amount (0.46) y property_value (0.41), lo que indica que personas con mayores ingresos tienden a solicitar préstamos más grandes y poseer propiedades de mayor valor.
  - LTV (Loan-to-Value) tiene correlaciones menores, pero parece influir en variables como loan_amount y rate_of_interest, aunque no de manera significativa.

VALORES ATÍPICOS



- Este gráfico muestra (boxplots) para cuatro variables numéricas:
  - loan_amount: Se observan muchos valores atípicos por encima de 1 millón, mientras que la mayoría de los préstamos están por debajo de los 500,000.
  - rate_of_interest: Hay varios valores atípicos tanto en el rango alto como en el rango bajo (tasa de interés fuera del rango de 3 a 5).
  - Credit_Score: No hay muchos valores atípicos, lo que sugiere que los puntajes de crédito están distribuidos de manera más homogénea entre 550 y 850.
  - income: Existen varios valores atípicos con ingresos significativamente mayores a la media, aunque la mayoría de los datos están concentrados en ingresos bajos.

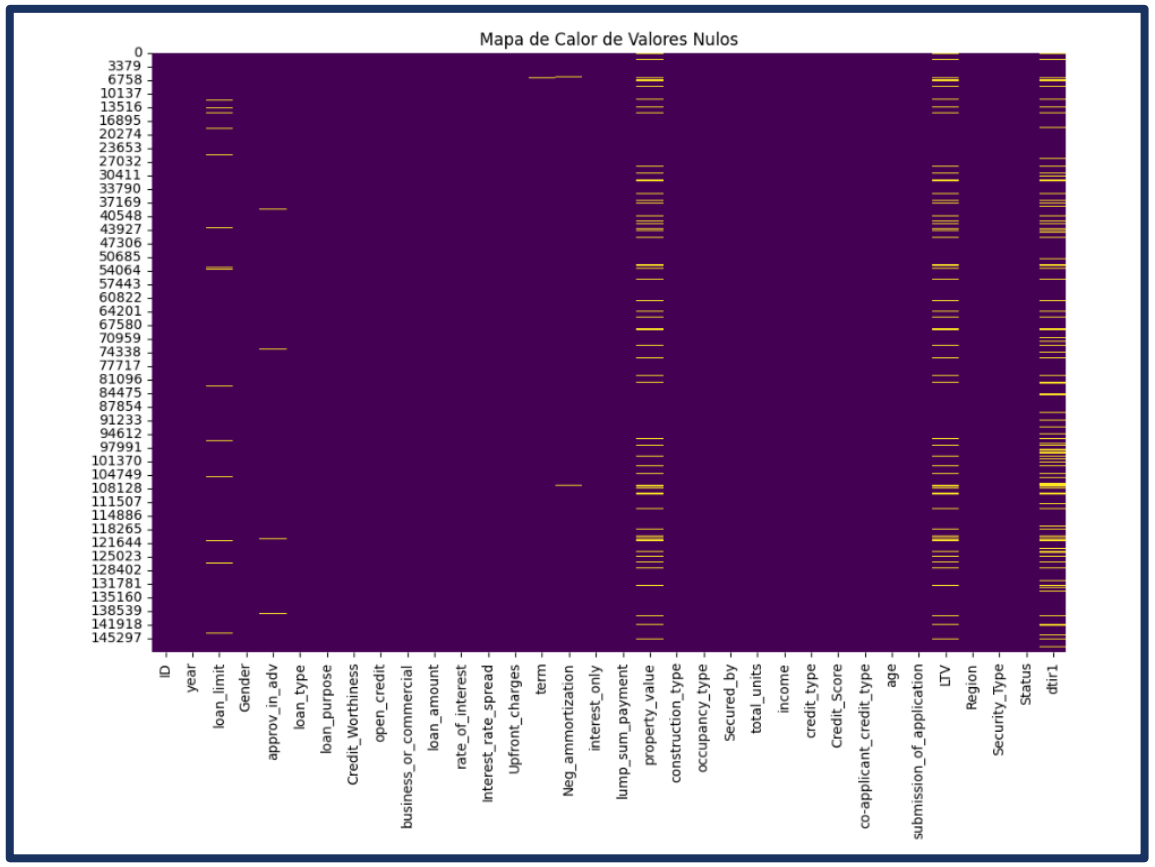


MANEJO DE DATOS FALTANTES

Las columnas como rate_of_interest, interest_rate_spread, income y upfront_charges contienen un alto porcentaje de valores nulos, eliminar todas las filas con datos faltantes podría reducir significativamente el tamaño del dataset.

Además, estas columnas están relacionadas con el status, por lo que, para preservar la integridad del análisis, se opta por reemplazar los valores nulos con la media correspondiente de cada columna.

MANEJO DE DATOS FALTANTES



- Este gráfico es un mapa de calor de valores nulos en el dataset.
  - Las barras amarillas indican la presencia de valores nulos en varias columnas.
  - Las columnas como rate_of_interest, Interest_rate_spread, income, y Status parecen tener una mayor concentración de valores nulos, confirmando que estas variables tienen datos faltantes en varias filas.
  - Algunas otras columnas también tienen valores nulos, pero en menor proporción.

INGENIERÍA DE CARACTERÍSTICAS

- Nuevas variables creadas:
  - Loan_to_property_value: ratio entre el monto del préstamo y el valor de la propiedad.
  - Income_to_loan_amount: proporción de ingresos en relación al préstamo.
  - Se categorizaron variables como rate_of_interest y se aplicaron transformaciones logarítmicas a loan_amount e income.
- Transformaciones aplicadas: logaritmos, categorizaciones binarias.
- Eliminación de valores atípicos usando iqr.
- Verificación de valores nulos.
- Reemplazo de nulos por la moda (eliminar los valores nulos no es el procedimiento mas optimo para este set de datos, debido a que al eliminar las filas con valores nulos eliminamos información critica para los datos como el status).

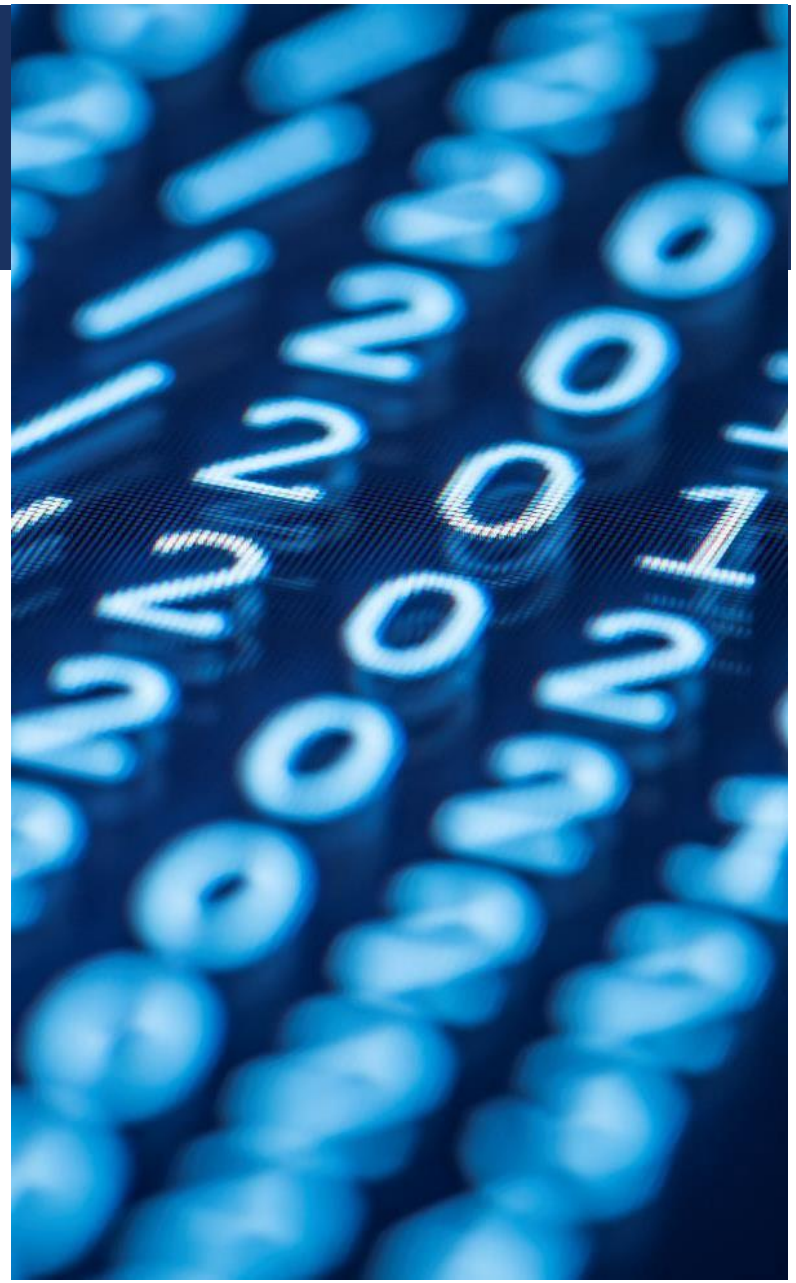
MODELOS DE CLASIFICACIÓN

Random Forest Classifier

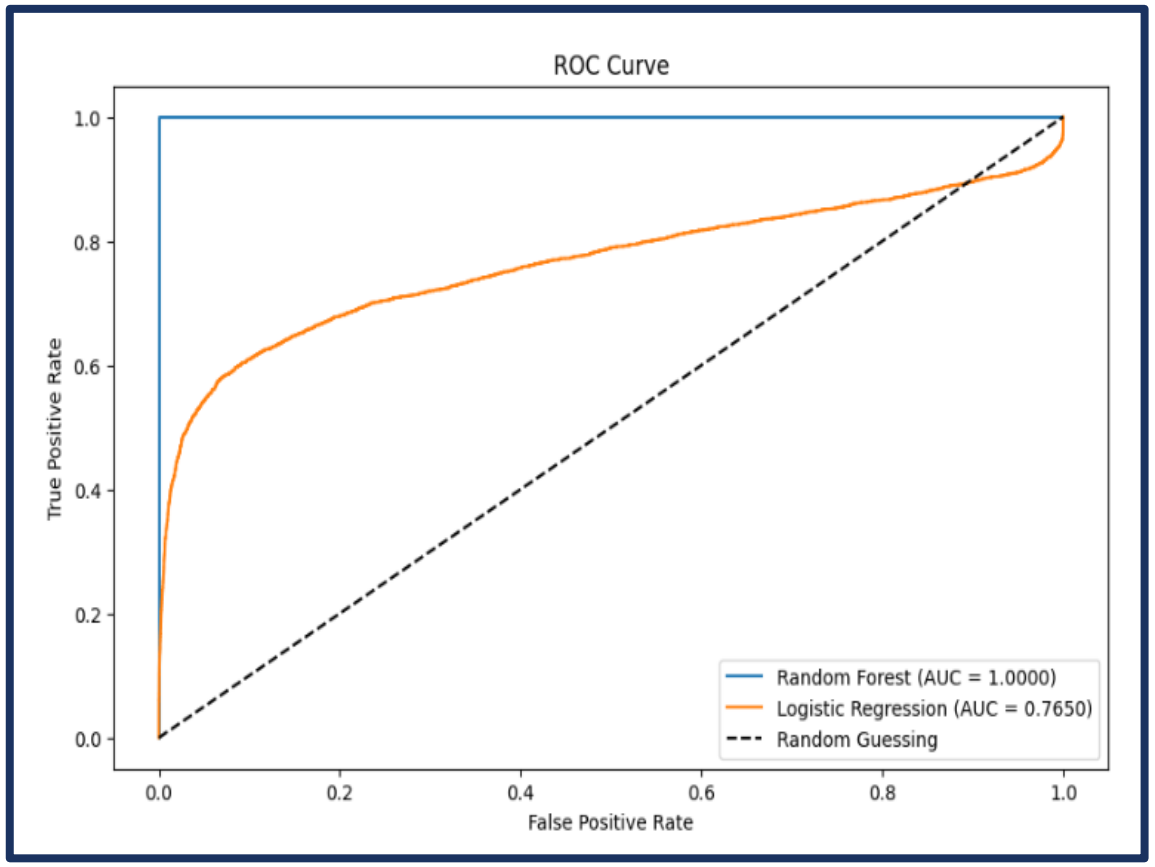
- Optimización: Se aplicaron técnicas de Grid Search para encontrar los mejores hiperparámetros (n_estimators, max_depth, min_samples_split).
- Resultados:
  - AUC = 1.000
  - Accuracy = 1.000

Regresión Logística

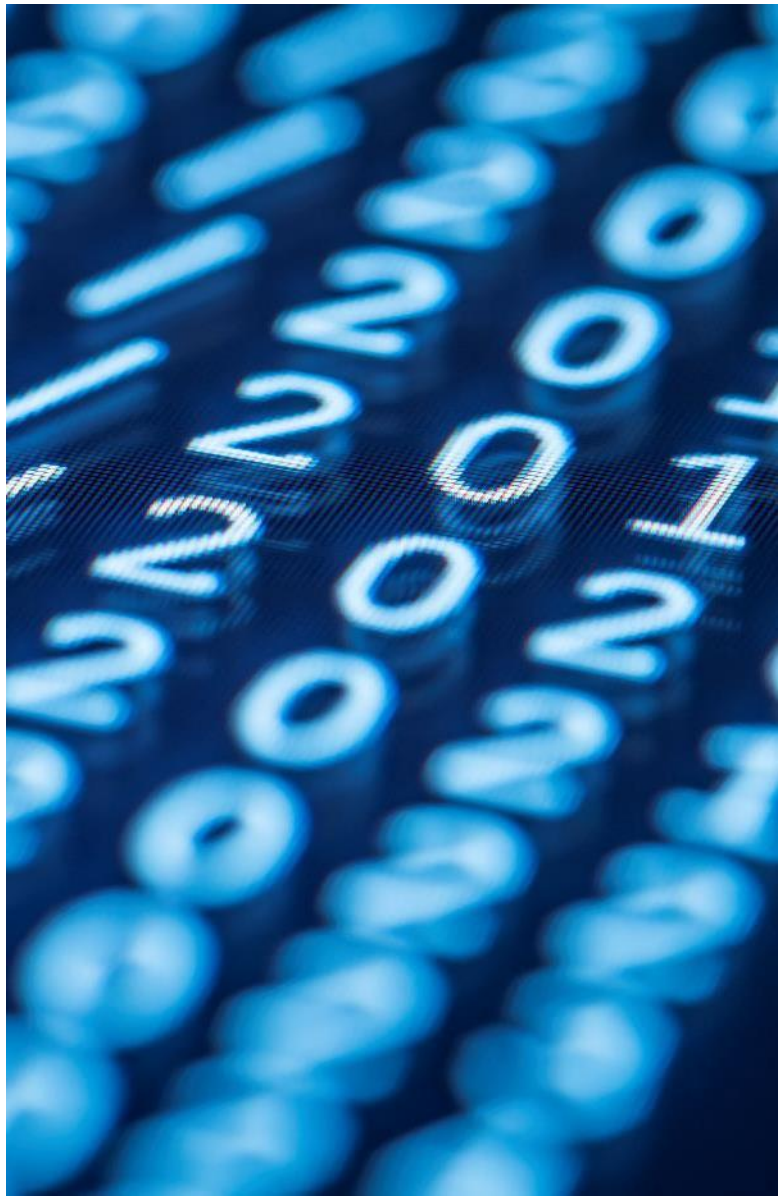
- Optimización: Ajuste de los hiperparámetros utilizando GridSearchCV para mejorar el rendimiento.
- Resultados:
  - AUC = 0.765
  - Accuracy = 0.88



RESULTADOS



- Los resultados sugieren que el Random Forest es un modelo más adecuado para este conjunto de datos, logrando una mayor AUC (Área Bajo la Curva ROC) que la Regresión Logística. Esto implica que el Random Forest tiene una capacidad superior para distinguir entre clientes que incumplirán y los que no.



CONCLUSIONES

Los modelos de machine learning desarrollados, en particular random forest, han demostrado una alta precisión en la predicción de incumplimientos de préstamos. Después de aplicar técnicas de optimización de hiperparámetros, el modelo de random forest alcanzó una precisión del 100% en la clasificación de clientes con riesgo de incumplimiento, mientras que el modelo de regresión logística, aunque efectivo, obtuvo una precisión más baja.

Durante el análisis, se identificó la necesidad de manejar cuidadosamente los valores nulos y atípicos, ya que la eliminación de datos críticos, como el "status", afectaba la capacidad de los modelos para predecir correctamente. El tratamiento adecuado de estos datos fue fundamental para mejorar la calidad de las predicciones.

RESPUESTA A LA PREGUNTA INICIAL

- ¿Podemos predecir si un cliente incumplirá con el pago de un préstamo basado en sus características financieras y del préstamo?
  - Sí, es posible predecir con alta precisión los incumplimientos de préstamos utilizando modelos de Machine Learning. Con los datos adecuados y las técnicas de optimización correctas, las instituciones financieras pueden anticipar posibles incumplimientos y ajustar sus políticas de riesgo crediticio en consecuencia.