Problema de Riesgo Crediticio

La importancia de reducir el riesgo crediticio ha llevado a una institución financiera alemana a buscar soluciones innovadoras. Como científicos de datos, hemos sido convocados para 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.

Tareas principales:

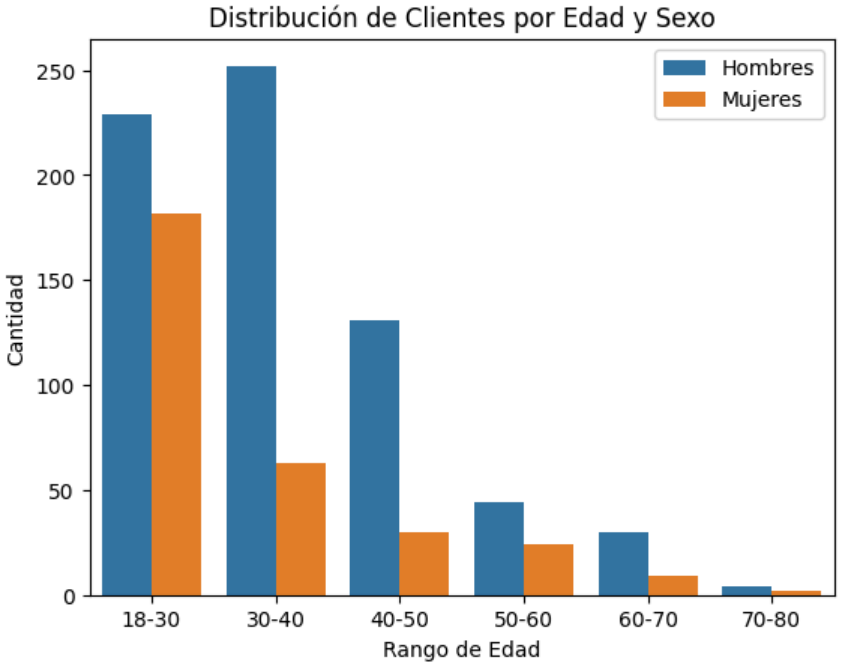
1. Preprocesamiento de Datos: Realizar limpieza de datos, manejar valores faltantes, codificación de variables categóricas y normalización/escalado de datos.

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.

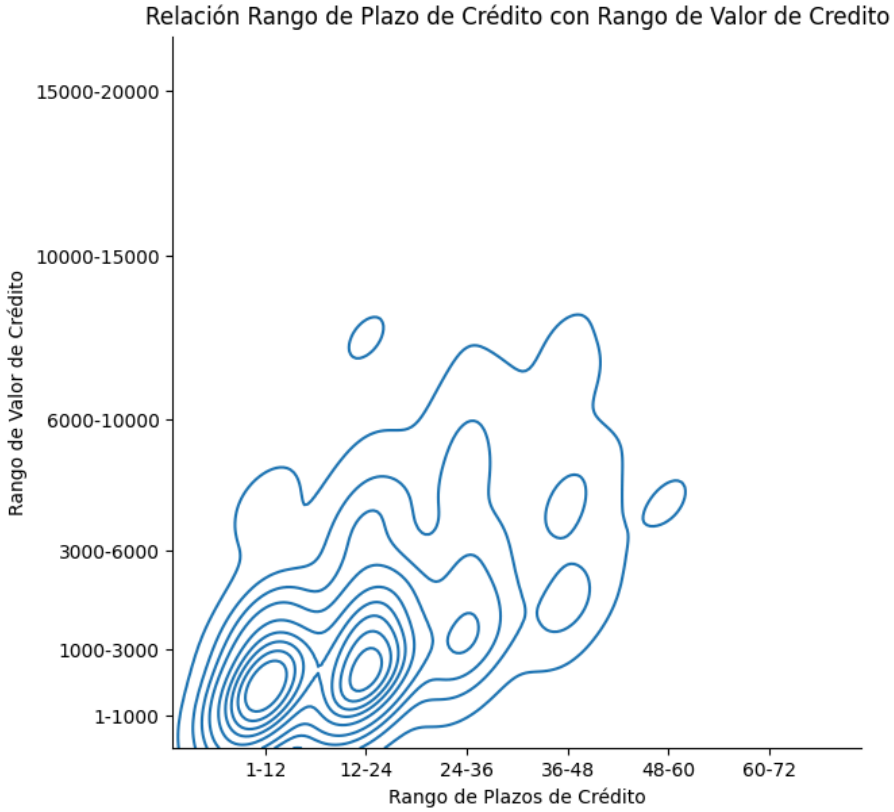
3. Construcción de Modelos: Experimentar con algunos algoritmos de machine learning como Regresión Logística, Árboles de Decisión, Random Forest, Naive Bayes, entre otros.

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.

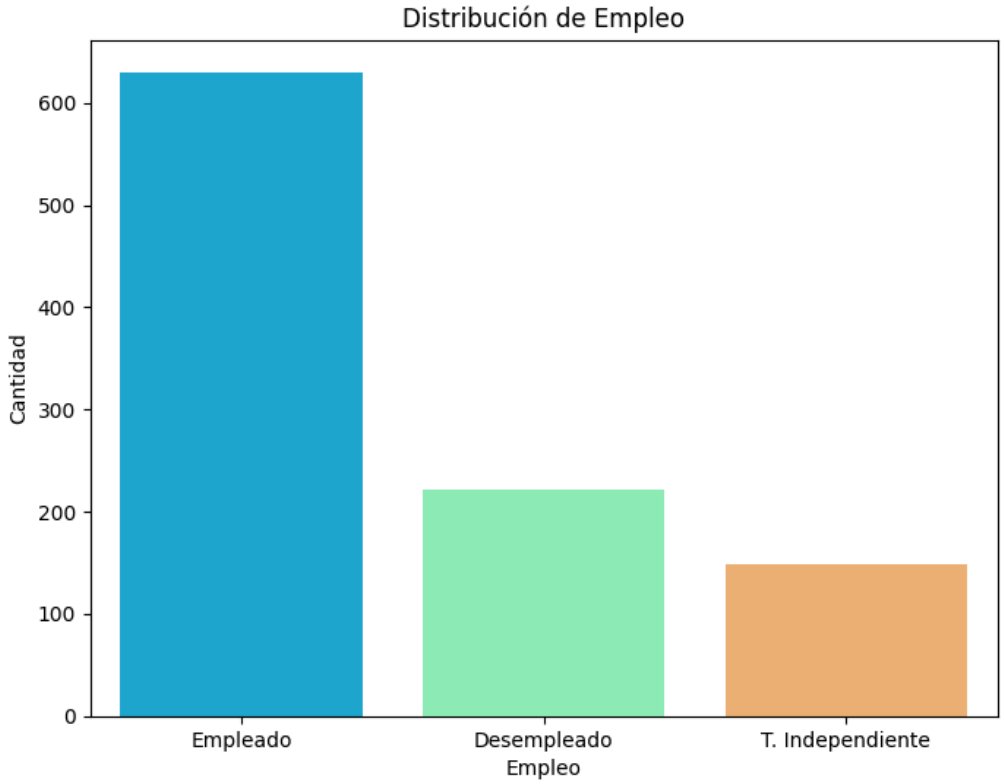
Información de los Clientes



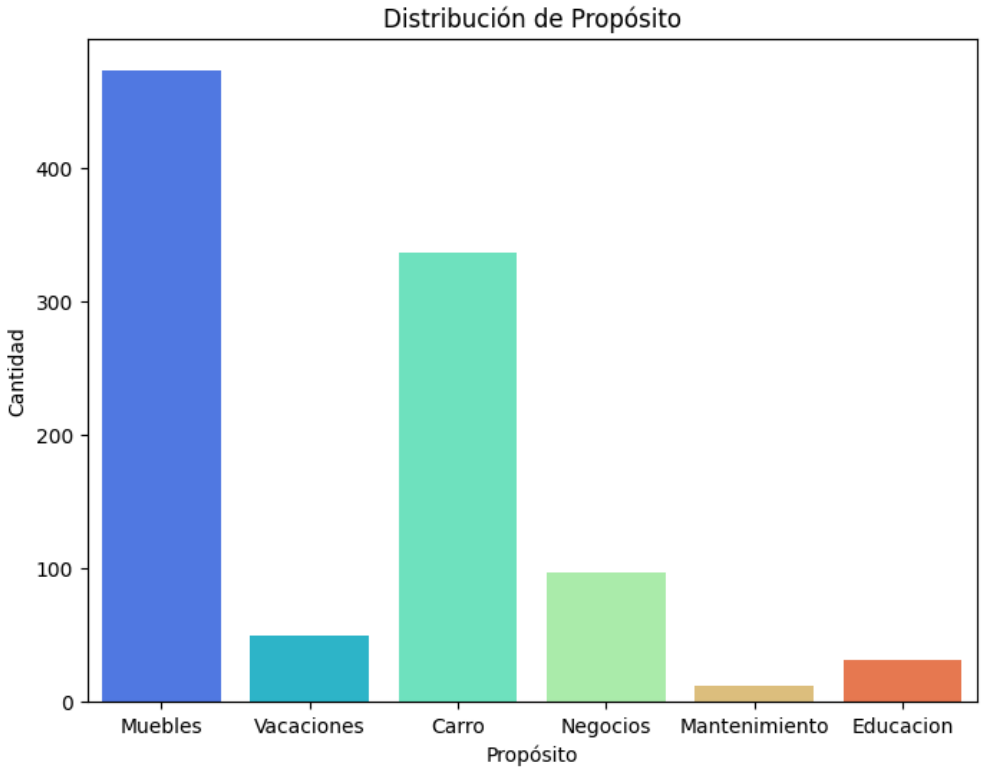
La mayoría de los clientes están en los grupos etarios de hombres entre 18-30 años y de 30-40 años. Así como también las mujeres de entre 18-30 años de edad. Las personas entre 70-80 años no serán consideradas para el análisis posterior porque representan una fracción reducida de los clientes.



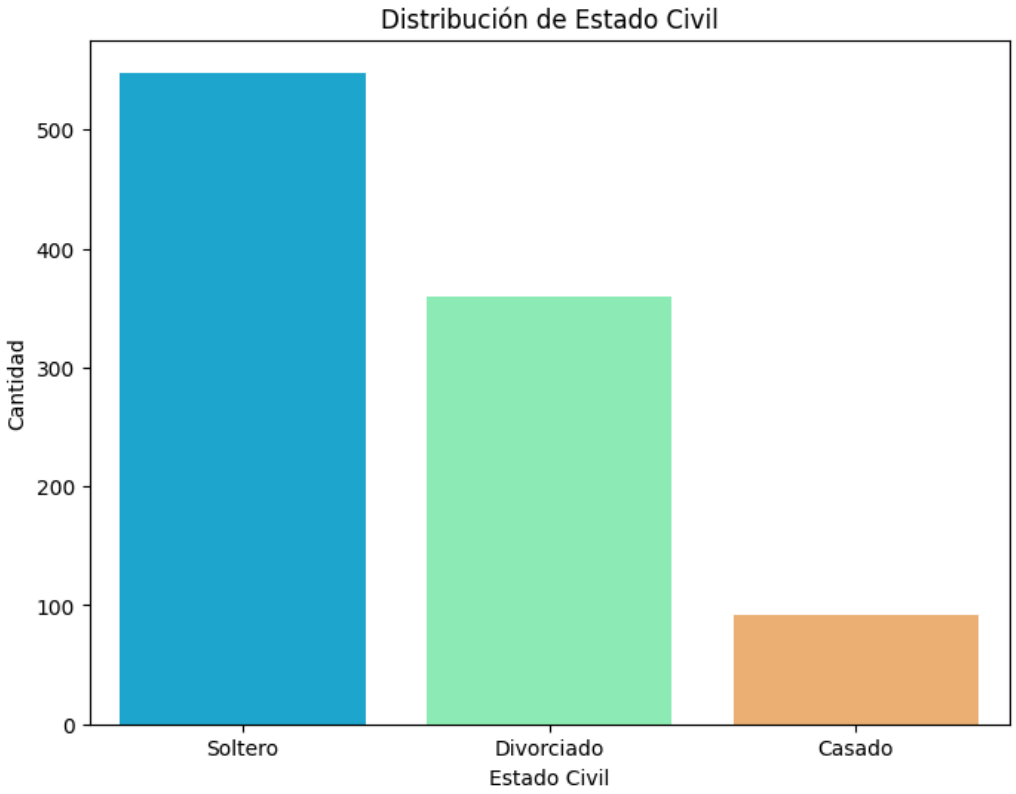
En el gráfico se aprecia una mayor frecuencia en los préstamos a corto plazo con un rango de valor crediticio bajo. Se entiende que la mayoría de los clientes prefieren solicitar montos menores a los 6000, generalmente en plazo de 2 años. Por lo tanto, los préstamos de valor superior a 10000 y con un plazo mayor a los 60 meses, serán excluidos del análisis posterior.



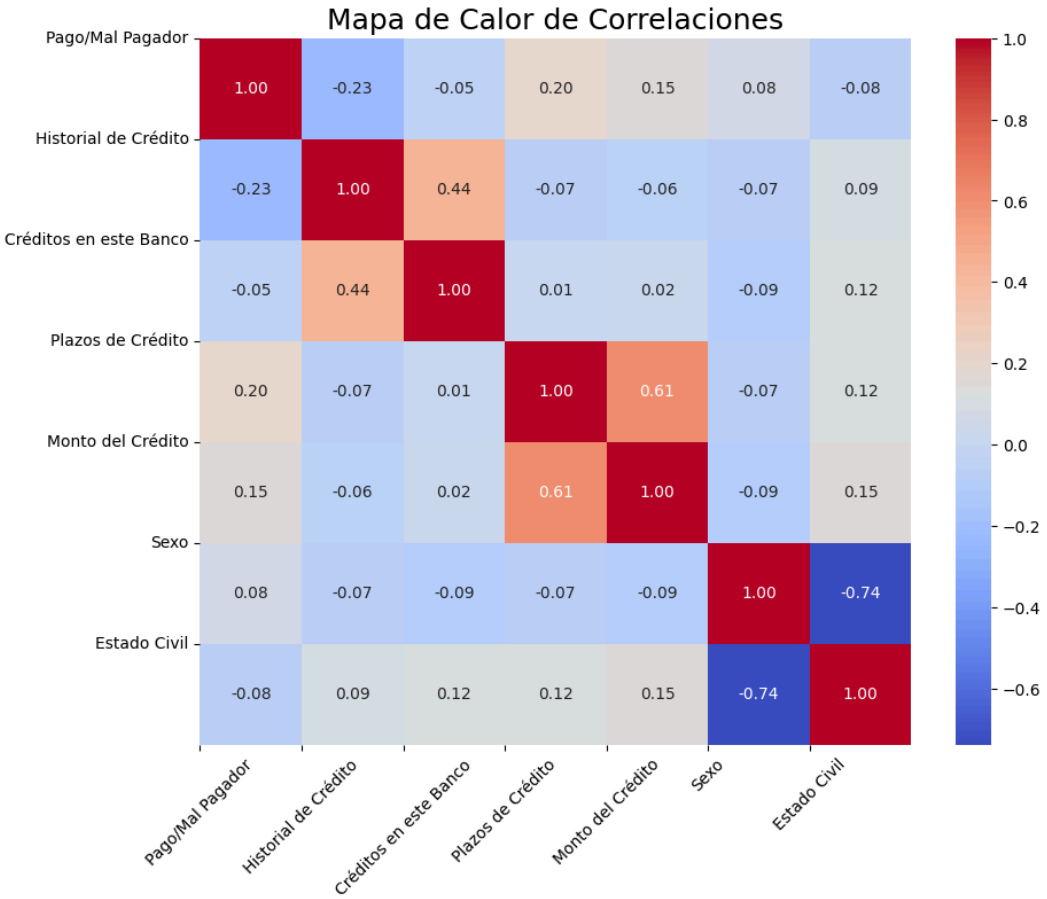
Se puede observar en el gráfico de ‘Distribución de Empleo’ que la mayoría de los clientes son empleados, con más de la mitad de los casos analizados. Los trabajadores independientes son la minoría, no obstante, es sustancial para considerarlos en el análisis con los modelos de ML.



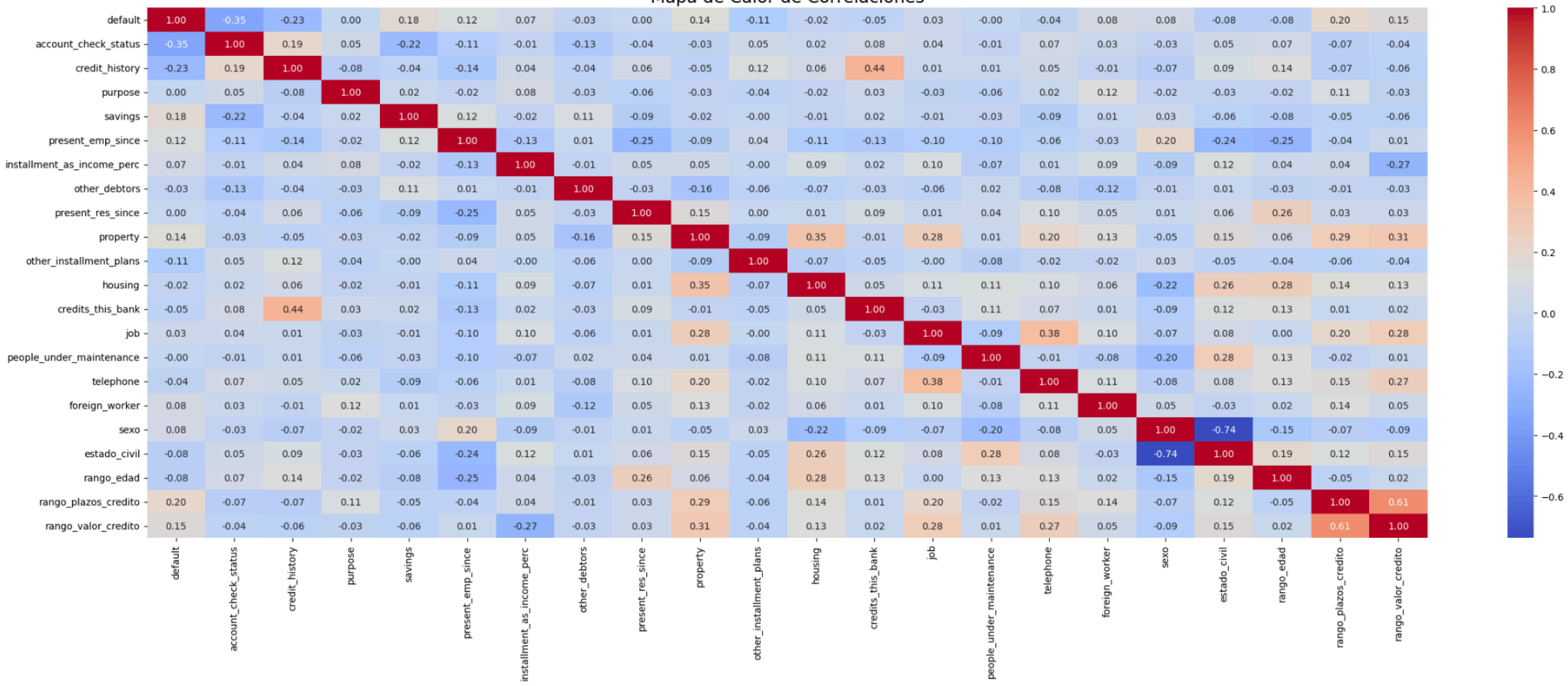
Con respecto al gráfico de ‘Distribución de Propósito’, se puede ver que la gran mayoría de los préstamos se realizan para bienes inmuebles y carros. Por otra parte, el mantenimiento es el propósito que tiene la menor relevancia en los datos, por lo que se puede descartar para el análisis.



En cuánto a la 'Distribución de Estado Civil', es claro que el mayor porcentaje de clientes están solteros o divorciados. Sin embargo, la población casada que solicita préstamos, aún es considerable, por lo que se tomara en cuenta para el análisis posterior.



En este mapa de calor, podemos se destaca la correlación entre los datos de 'Historial de Crédito' - 'Créditos en este Banco' con un 0.41 de relación, entre 'Plazos de Crédito'-'Monto del Crédito', con una correlación positiva de '0.61 y la negativa entre 'Sexo' y 'Estado Civil'.



En este mapa de calor podemos ver la relación entre todas las variables de los clientes.

Modelos de Machine Learning

En el actual panorama financiero, la evaluación del riesgo crediticio es una parte fundamental de la operativa de cualquier institución bancaria. La capacidad de predecir con precisión si un cliente será un "mal pagador" o no, es crucial para mantener la estabilidad y la rentabilidad del negocio. En este documento, se presenta un análisis detallado sobre el uso de técnicas de Machine Learning para abordar este desafío en el contexto de un banco alemán.

Utilizando un conjunto diverso de modelos de clasificación, tales como Logistic Regression, Decision Tree Classifier, Random Forest Classifier y Gaussian Naive Bayes, hemos llevado a cabo una exhaustiva evaluación del riesgo crediticio. Estos modelos ofrecen un enfoque sistemático y cuantitativo para identificar patrones y tendencias en los datos históricos de los clientes, permitiendo así la construcción de sistemas predictivos robustos.

A lo largo de este documento, se analizan en detalle cada uno de estos modelos, incluyendo su implementación, rendimiento y capacidad predictiva. Además, se discuten las ventajas y limitaciones de cada enfoque, así como las consideraciones prácticas para su implementación en un entorno bancario real.

En última instancia, este análisis proporciona una visión integral sobre la aplicación de técnicas de Machine Learning en la evaluación del riesgo crediticio, destacando su potencial para mejorar la eficiencia y la precisión en la toma de decisiones financieras en el sector bancario alemán.

Métricas Utilizadas

Accuracy (Precisión):

La precisión es una medida de la proporción de predicciones correctas realizadas por el modelo. Se calcula dividiendo el número total de predicciones correctas (verdaderos positivos y verdaderos negativos) por el número total de predicciones realizadas. Una alta precisión indica que el modelo está haciendo un buen trabajo en clasificar correctamente tanto los casos positivos como los negativos.

Precision (Precisión):

La precisión es una medida de la proporción de predicciones positivas realizadas por el modelo que fueron correctas. Se calcula dividiendo el número de verdaderos positivos por la suma de verdaderos positivos y falsos positivos. Una alta precisión indica que el modelo tiene una baja tasa de falsos positivos, es decir, que cuando predice que un cliente será un "mal pagador", es probable que realmente lo sea.

Recall (Recuperación o Sensibilidad):

El recall es una medida de la proporción de casos positivos reales que fueron correctamente identificados por el modelo. Se calcula dividiendo el número de verdaderos positivos por la suma de verdaderos positivos y falsos negativos. Una alta sensibilidad indica que el modelo tiene una baja tasa de falsos negativos, es decir, que es capaz de identificar la mayoría de los casos de "mal pagador" en el conjunto de datos.

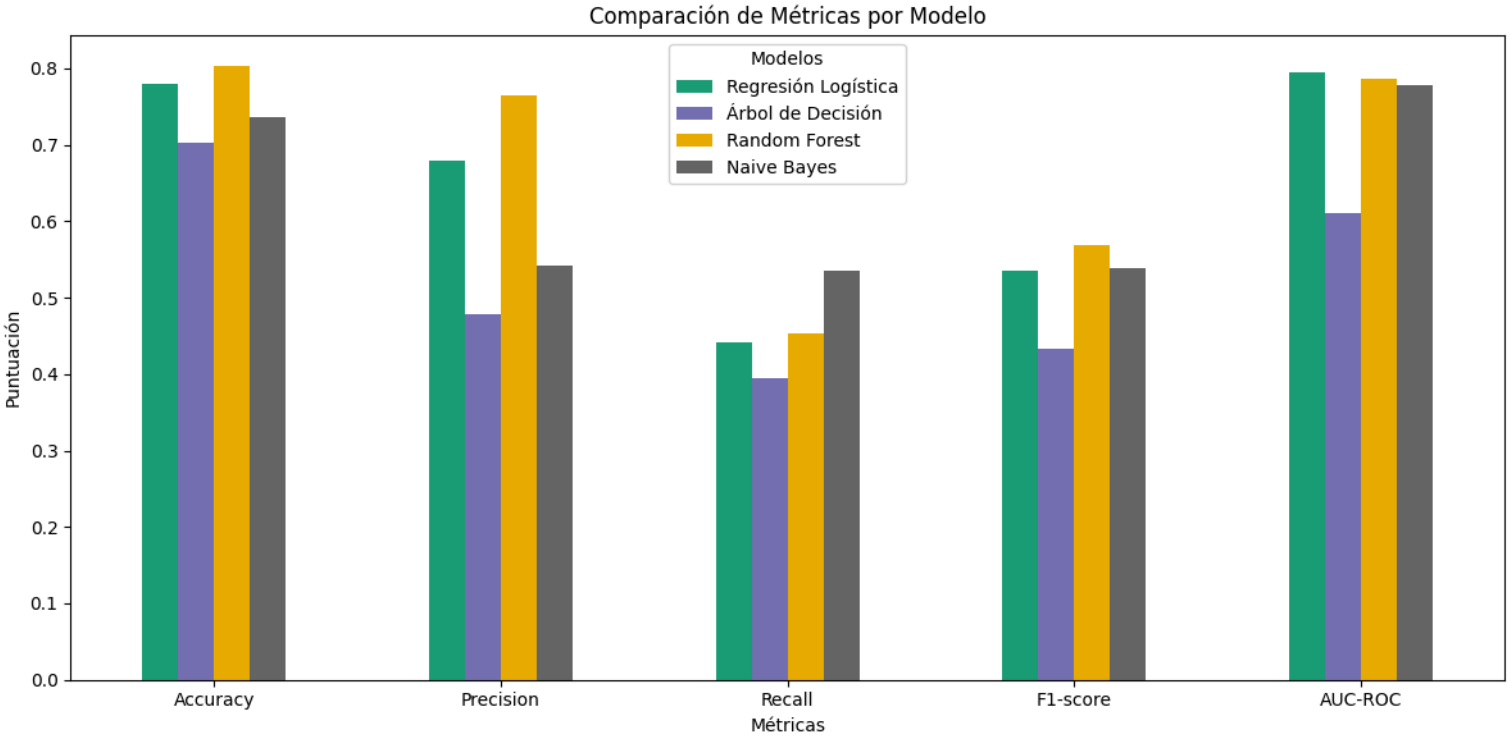
F1-score (Puntuación F1):

La puntuación F1 es una medida de la media armónica entre precisión y recuperación. Se calcula como 2 veces el producto de precisión y recuperación, dividido por la suma de precisión y recuperación. La puntuación F1 es útil cuando hay un desequilibrio entre las clases de predicción, ya que tiene en cuenta tanto los falsos positivos como los falsos negativos.

AUC-ROC (Área bajo la curva ROC):

El área bajo la curva ROC es una medida de la capacidad discriminativa del modelo. Representa el área bajo la curva ROC (Receiver Operating Characteristic), que muestra la tasa de verdaderos positivos frente a la tasa de falsos positivos para diferentes umbrales de clasificación. Un valor de AUC-ROC cercano a 1 indica un modelo con una buena capacidad para distinguir entre clases positivas y negativas.

Resultados



Considerando que el objetivo de los modelos es identificar a los clientes que serán malos pagadores de préstamos de crédito, debemos priorizar las métricas que se centran en la capacidad del modelo para identificar correctamente estos casos positivos (malos pagadores). Por lo tanto, las métricas más relevantes en este contexto son Recall y AUC-ROC.

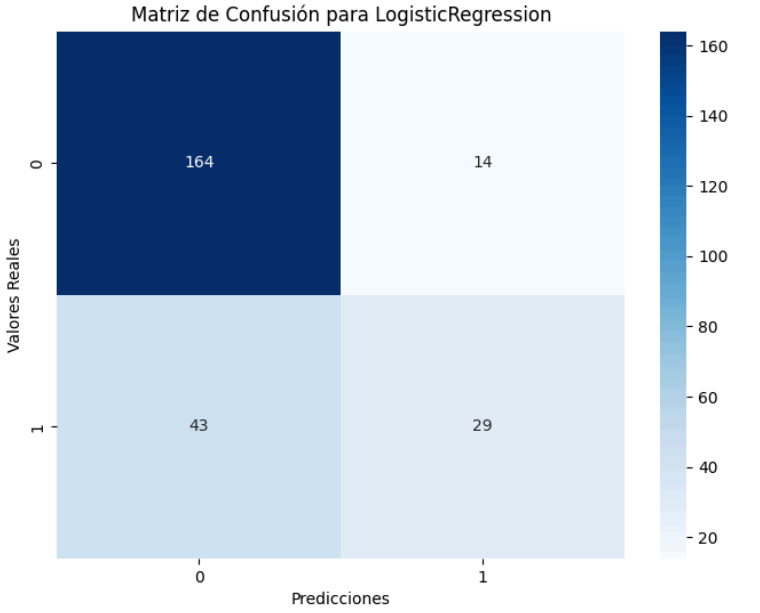
En base a esta consideración, observamos que el modelo de Naive Bayes tiene el Recall más alto y una AUC-ROC considerablemente alta, lo que sugiere que es mejor para identificar a los malos pagadores en comparación con los otros modelos. Aunque Random Forest tiene una precisión y Accuracy generalmente más altas, lo que indica que es mejor para clasificar correctamente todos los casos, incluidos los buenos pagadores, no se desempeña tan bien como Naive Bayes en la identificación específica de los malos pagadores.

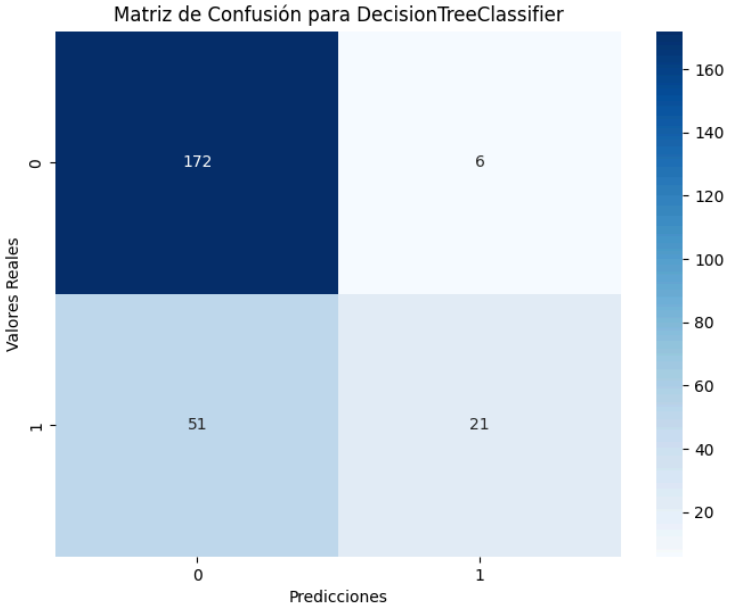
Por lo tanto, en este contexto, el modelo de Naive Bayes sería la mejor opción para predecir qué clientes serán malos pagadores de préstamos de crédito.

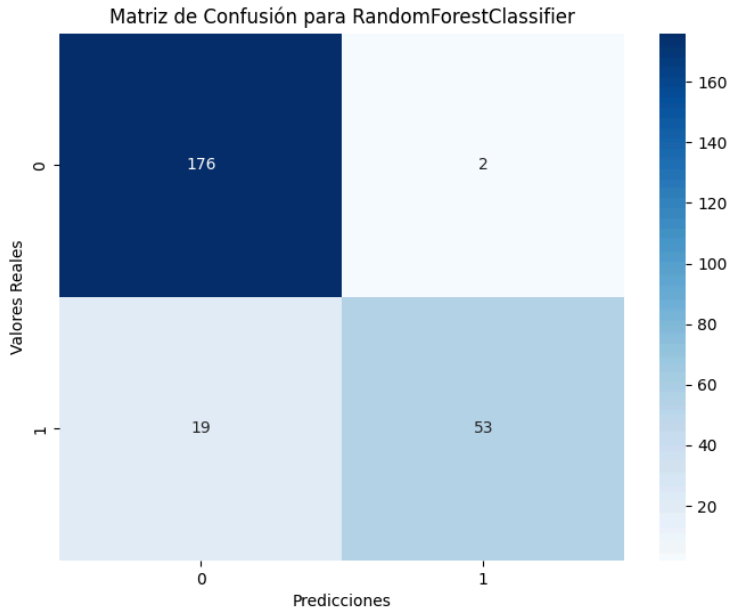
Matriz de Confusión

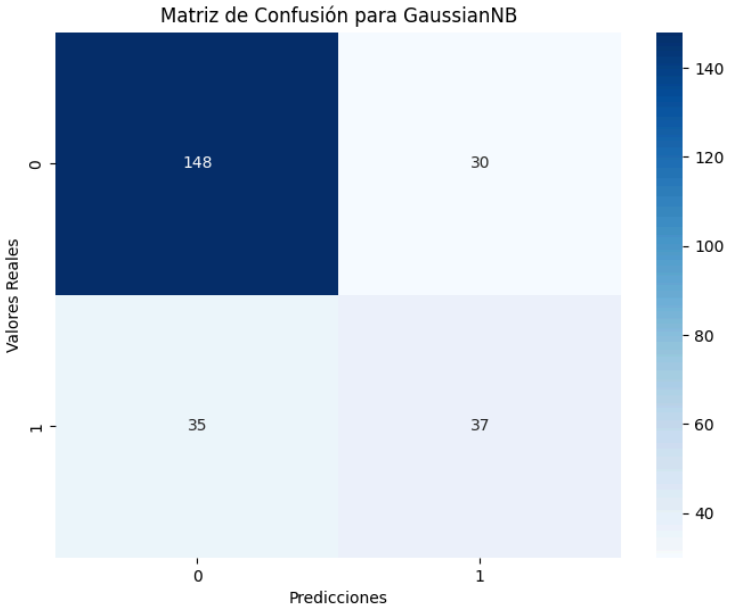
La matriz de confusión es una herramienta que visualiza el rendimiento de un modelo de clasificación comparando las predicciones del modelo con las clases reales en un conjunto de datos. Se organiza en una matriz bidimensional con cuatro cuadrantes:

- Verdaderos Positivos (TP): Casos en los que el modelo predijo correctamente la clase positiva.
- Falsos Positivos (FP): Casos en los que el modelo predijo incorrectamente la clase positiva cuando en realidad era negativa.
- Verdaderos Negativos (TN): Casos en los que el modelo predijo correctamente la clase negativa.
- Falsos Negativos (FN): Casos en los que el modelo predijo incorrectamente la clase negativa cuando en realidad era positiva.









Al considerar las matrices de confusión, podemos observar cómo se distribuyen los verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos en cada modelo. Dado que el objetivo es identificar a los malos pagadores de préstamos de crédito, nos enfocaremos en el número de falsos negativos (FN), es decir, los casos en los que el modelo predijo incorrectamente que un cliente sería un buen pagador cuando en realidad fue un mal pagador.

Observando las matrices de confusión:

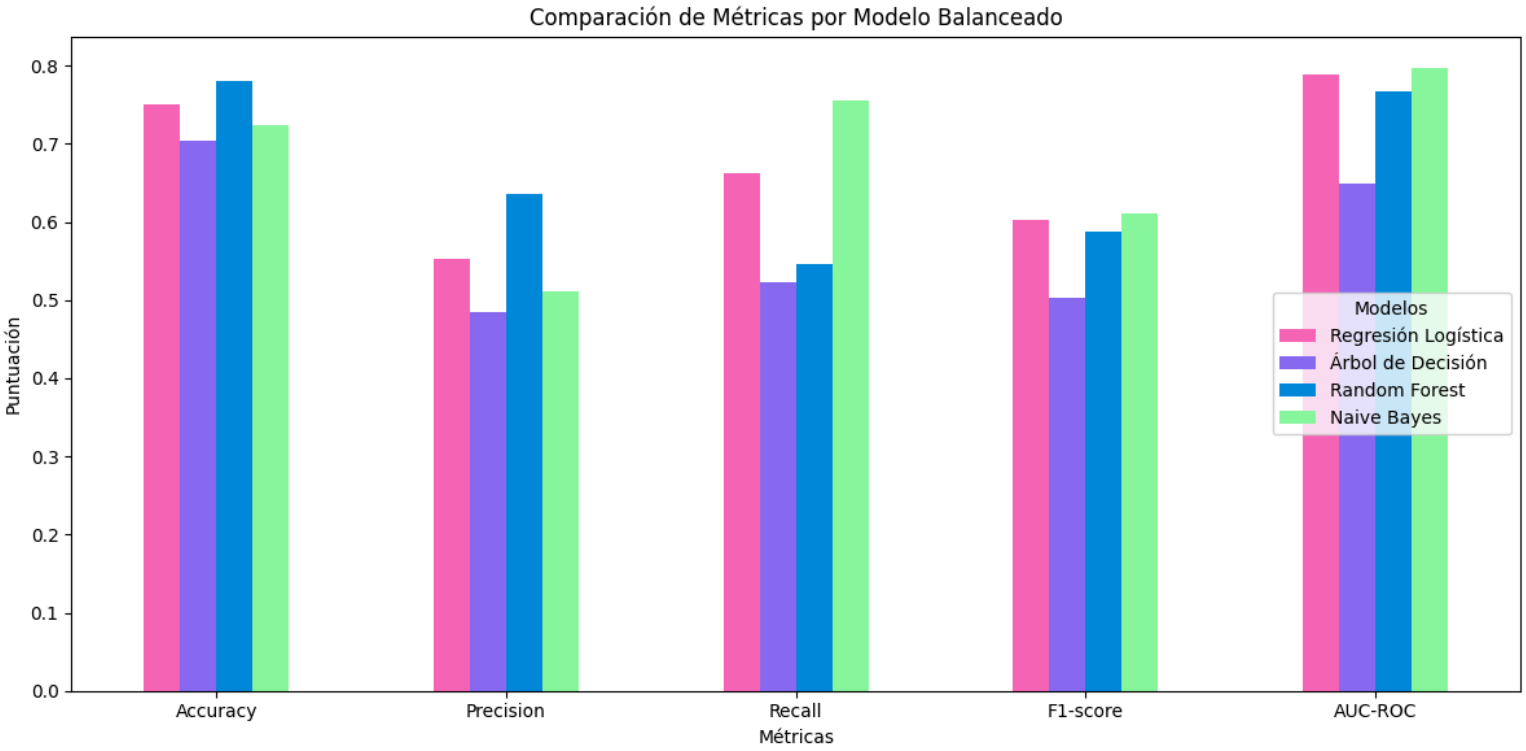
- Regresión Logística: Tiene 43 falsos negativos.
- Árbol de Decisión: Tiene 51 falsos negativos.
- Random Forest: Tiene 19 falsos negativos.
- Naive Bayes: Tiene 35 falsos negativos.

Es notable que el modelo de Random Forest sigue teniendo el menor número de falsos negativos, seguido por el modelo de Naive Bayes. Por lo tanto, en este contexto, el modelo de Random Forest sigue siendo la mejor opción para predecir qué clientes serán malos pagadores de préstamos de crédito, seguido de cerca por el modelo de Naive Bayes.

Balanceo de Variables

El balanceo de variables, se refiere a la situación en la que las diferentes categorías o clases de una variable (también conocida como feature o característica) están representadas en cantidades desproporcionadas dentro de un conjunto de datos.

Cuando se trabaja con conjuntos de datos desbalanceados, es decir, cuando hay una gran diferencia en el número de muestras entre las diferentes clases, puede surgir un problema de sesgo en los modelos de machine learning. Esto se debe a que los algoritmos de aprendizaje automático tienden a favorecer las clases mayoritarias sobre las clases minoritarias, lo que puede llevar a un rendimiento deficiente del modelo en la



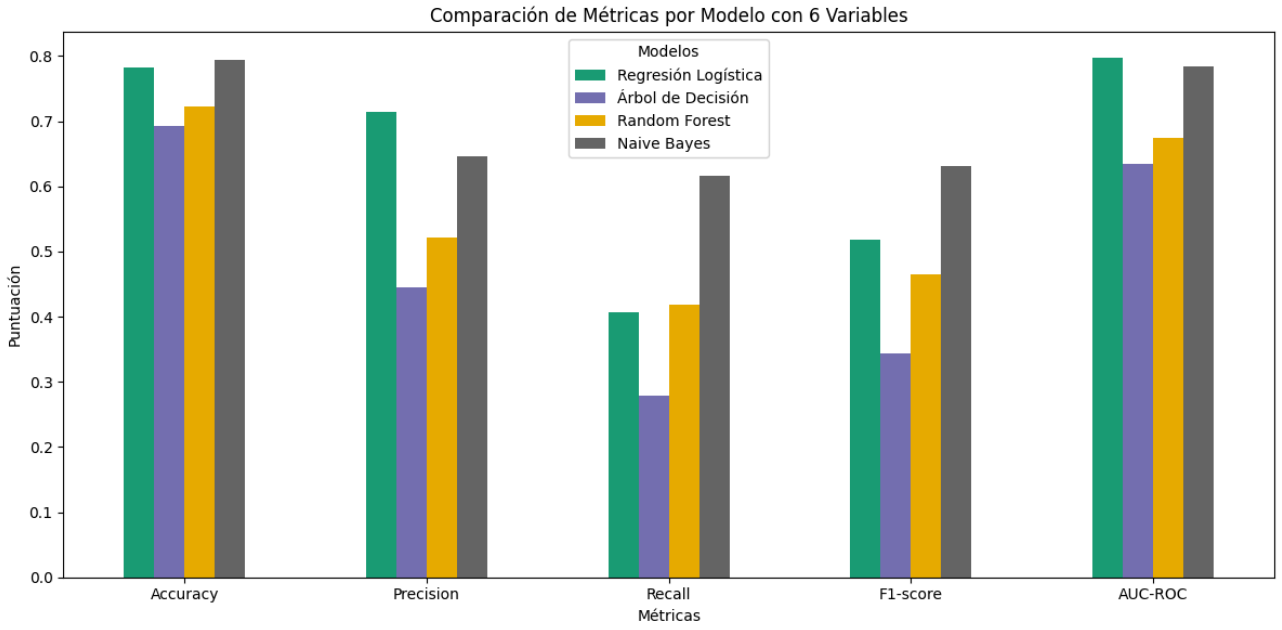
Después de balancear la variable "Default", observamos algunos cambios significativos en las métricas de los modelos:

- **Regresión Logística**: La precisión disminuyó, pero el recall aumentó, lo que indica que el modelo es mejor para identificar casos positivos verdaderos. El F1-score y AUC-ROC también mejoraron ligeramente.
- **Árbol de Decisión**: Las métricas permanecieron prácticamente iguales después del balanceo.
- **Random Forest**: La precisión y el recall aumentaron, lo que indica una mejora en la capacidad del modelo para identificar tanto casos positivos como negativos verdaderos. El F1-score y AUC-ROC también mejoraron.

- **Naive Bayes**: La precisión disminuyó, pero el recall aumentó significativamente, lo que indica una mejor capacidad para identificar casos positivos verdaderos. El F1-score y AUC-ROC también mejoraron.

En general, el balanceo de la variable "Default" condujo a mejoras en la capacidad de los modelos para identificar casos positivos verdaderos (malos pagadores de préstamos de crédito), como se refleja en el aumento del recall en la mayoría de los modelos. Esto sugiere que el balanceo de variables puede ser una estrategia efectiva para mejorar el rendimiento de los modelos en conjuntos de datos desbalanceados.

Prueba con Menos Variables (6)



En general, parece que Naive Bayes es el modelo con el mejor rendimiento en términos de precisión, recall, F1-score y AUC-ROC en este conjunto de datos con solo 6 columnas. Sin embargo, es importante considerar el contexto específico del problema y las necesidades del negocio al seleccionar el modelo final.

Conclusiones

- **Impacto de las variables en los modelos de ML**: La selección cuidadosa de las variables puede tener un impacto significativo en el rendimiento de los modelos de machine learning. Al reducir el conjunto de datos a solo 6 columnas, observamos cambios en las métricas de rendimiento de los modelos, lo que destaca la importancia de comprender y seleccionar las variables más relevantes para el problema en cuestión.

- **Balanceo de datos y su efecto en el rendimiento**: El balanceo de la variable objetivo ("Default") puede mejorar el rendimiento de los modelos, especialmente en términos de recall y AUC-ROC. Los resultados muestran que después de balancear la variable, algunos modelos experimentaron mejoras significativas en la capacidad para identificar casos positivos verdaderos, lo que sugiere que el balanceo de datos puede ser una estrategia efectiva para abordar conjuntos de datos desbalanceados.

- **Comparación de modelos**: Entre los modelos evaluados (Regresión Logística, Árbol de Decisión, Random Forest y Naive Bayes), Naive Bayes demostró ser consistentemente sólido en términos de precisión, recall, F1-score y AUC-ROC. Esto sugiere que Naive Bayes puede ser una opción viable para predecir el riesgo crediticio en este contexto específico.

- **Importancia de métricas de rendimiento múltiples**: Es importante considerar múltiples métricas de rendimiento al evaluar modelos de machine learning. Mientras que algunos modelos pueden tener un rendimiento sólido en ciertas métricas, puede haber compensaciones en otras métricas. Por lo tanto, una evaluación integral que tenga en cuenta todas las métricas relevantes es crucial para seleccionar el modelo más adecuado.

- **Contextualización de los resultados**: Los resultados del análisis de riesgo crediticio deben interpretarse en el contexto específico del negocio y las necesidades del cliente. Si bien los modelos pueden proporcionar predicciones precisas, la implementación exitosa de estrategias basadas en estos resultados requiere una comprensión profunda del entorno operativo y regulatorio, así como una consideración cuidadosa de los posibles impactos en los clientes y la reputación de la empresa.

Recomendación del mejor modelo

Después de evaluar los resultados de los diferentes modelos, se recomienda utilizar el modelo de Naive Bayes para predecir el riesgo crediticio en este contexto. Tanto con datos balanceados como desbalanceados, Naive Bayes demostró un rendimiento sólido en términos de precisión, recall, F1-score y AUC-ROC. Además, Naive Bayes mostró una capacidad destacada para manejar conjuntos de datos desbalanceados, lo que sugiere que es robusto incluso cuando las clases están desproporcionadamente representadas. Por lo tanto, Naive Bayes parece ser el modelo más adecuado para predecir el riesgo crediticio en este escenario específico, ofreciendo un equilibrio entre rendimiento y capacidad para manejar datos desbalanceados. Sin embargo, es importante realizar más análisis y pruebas en un entorno de producción para validar estas conclusiones y asegurar una implementación efectiva en la práctica.