Data Science I: Fundamentos para la Ciencia de Datos

CoderHouse: Comisión N°61.750

Alumno: Agüero García Josias

Documentación DataSet:

Informe Prestamos Default

CODER HOUSE

Introducción

Descripción:

Este proyecto aborda un análisis detallado de una base de datos de préstamos, con el objetivo de identificar patrones relacionados con el incumplimiento de pagos. Utilizando herramientas de ciencia de datos, se exploran las relaciones entre varias variables clave como la duración de los préstamos, el propósito de estos y su impacto en la tasa de incumplimiento.

Objetivo:

Se llevará a cabo un análisis exploratorio para identificar las características y patrones predominantes, y se diseñarán y evaluarán diferentes modelos de clasificación para determinar su efectividad.

Lo que podría ayudar a las entidades financieras a optimizar estrategias de evaluación de riesgos.

Fuente:

https://www.kaggle.com/datasets/yasserh/loan-default-dataset

CODER HOUSE

Importación de Librerías y Carga de Datos:

El análisis se llevó a cabo en un entorno de Google Colab.

Utilizamos las siguientes librerías:

- Pandas para la manipulación de datos.
- Numpy para operaciones numéricas.
- Seaborn y matplotlib para la visualización de datos.
- Missingno para analizar datos faltantes.

Exploración de Datos:

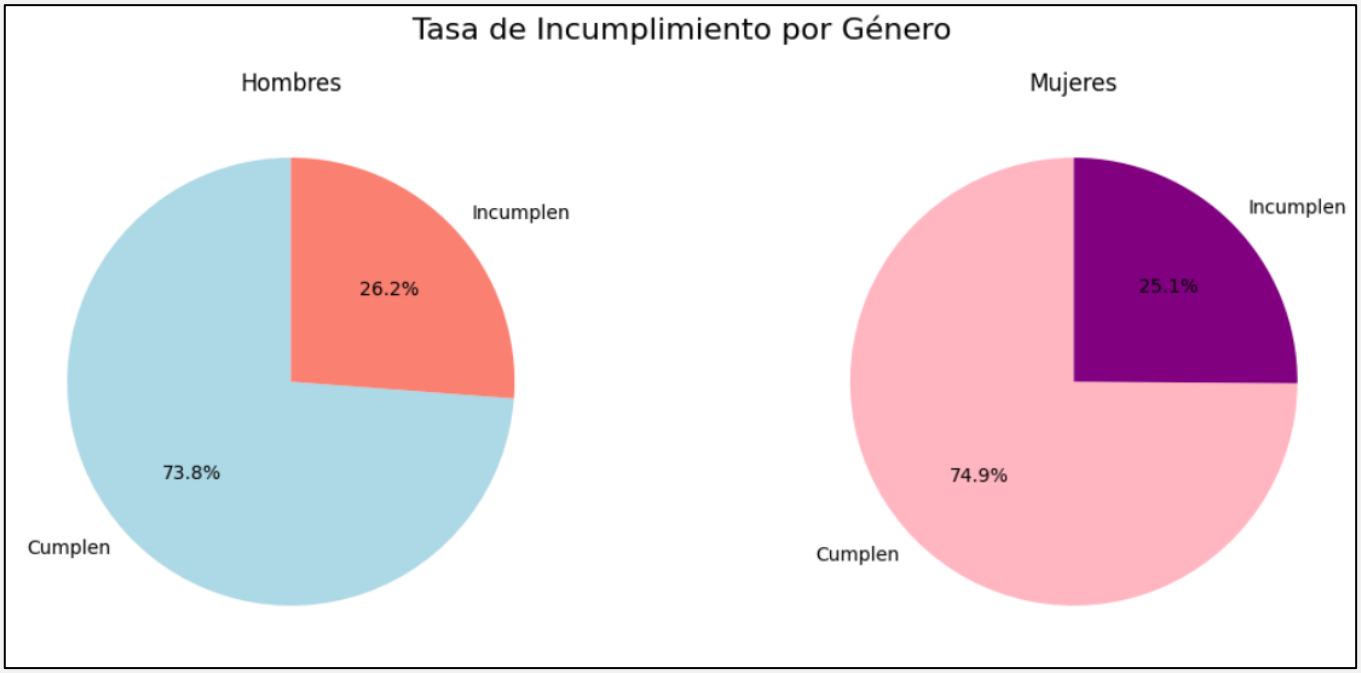
Exploramos el contenido de la base de datos para evaluar su estructura:

- El siguiente código nos muestra que el conjunto de datos contiene un total de 148.670 filas (registros) y 34 columnas (variables).

Hipótesis Planteadas:

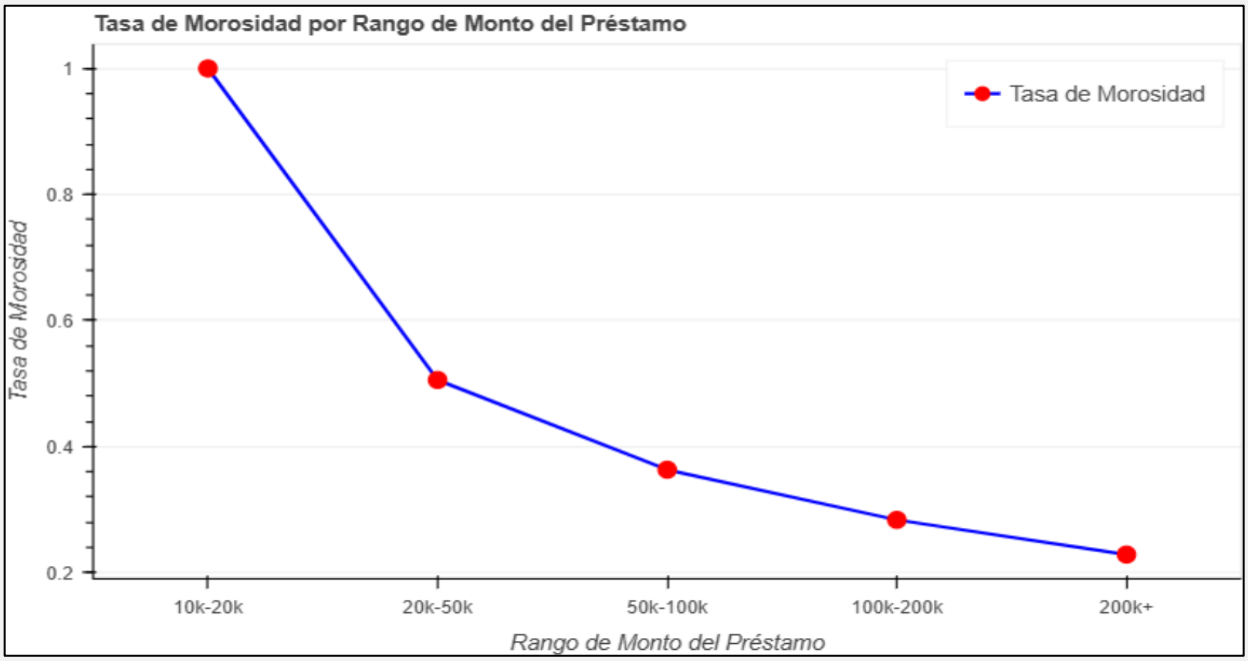
1. Existe una diferencia de tasa de incumplimiento en los préstamos otorgados a personas de género masculino en comparación con las de género femenino?

La tasa de incumplimiento es mínimamente mayor en los hombres con 1.1%



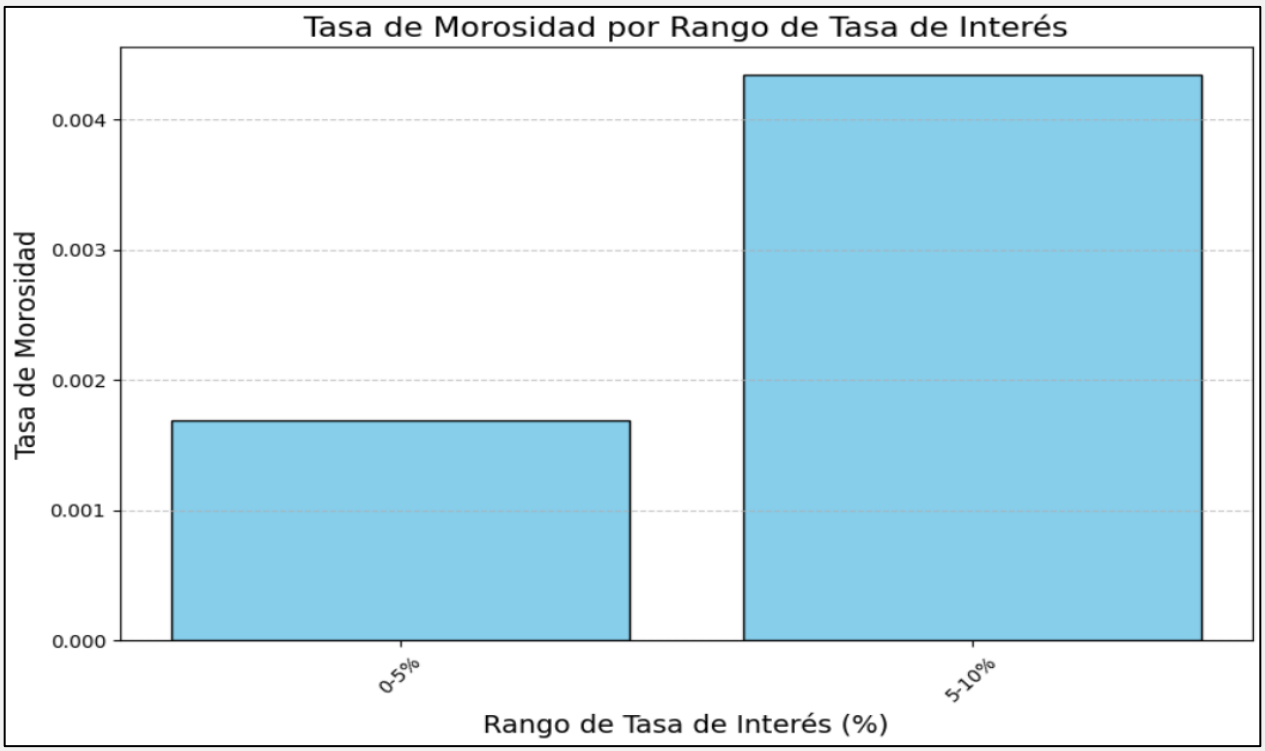
2. La morosidad varía de acuerdo con el monto del préstamo?

En conclusión, la tasa de morosidad disminuye a medida que el monto del préstamo es mayor.



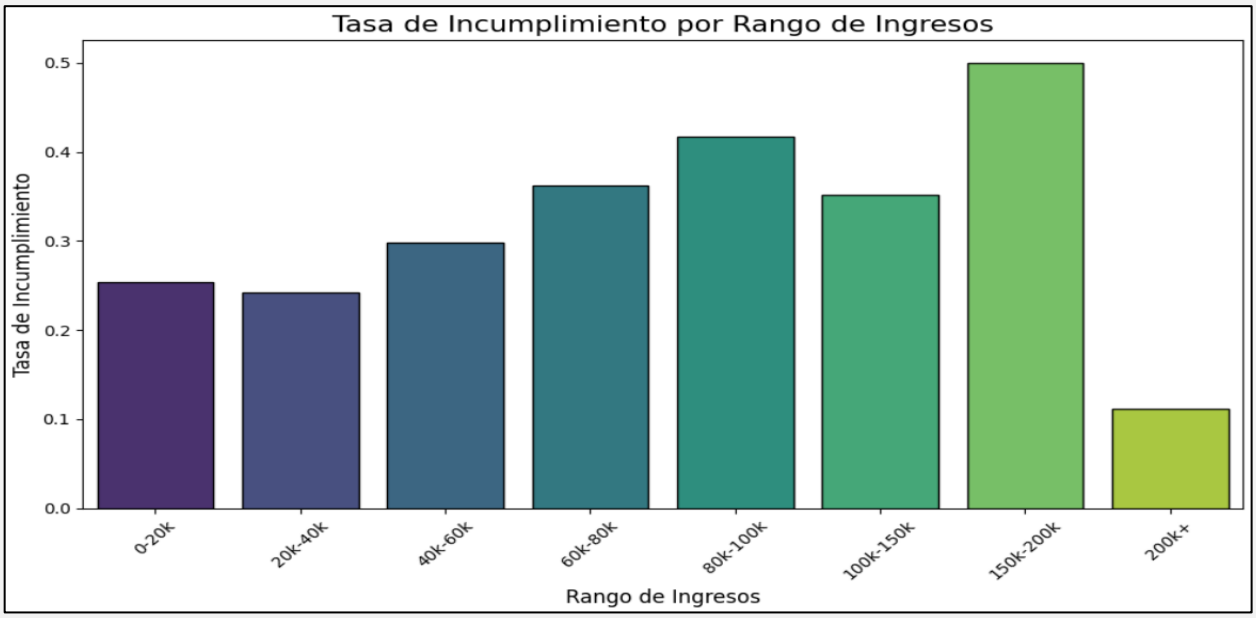
3. La morosidad aumenta con la tasa de interés?

La tasa de morosidad aumenta mínimamente con una mayor tasa de interés.



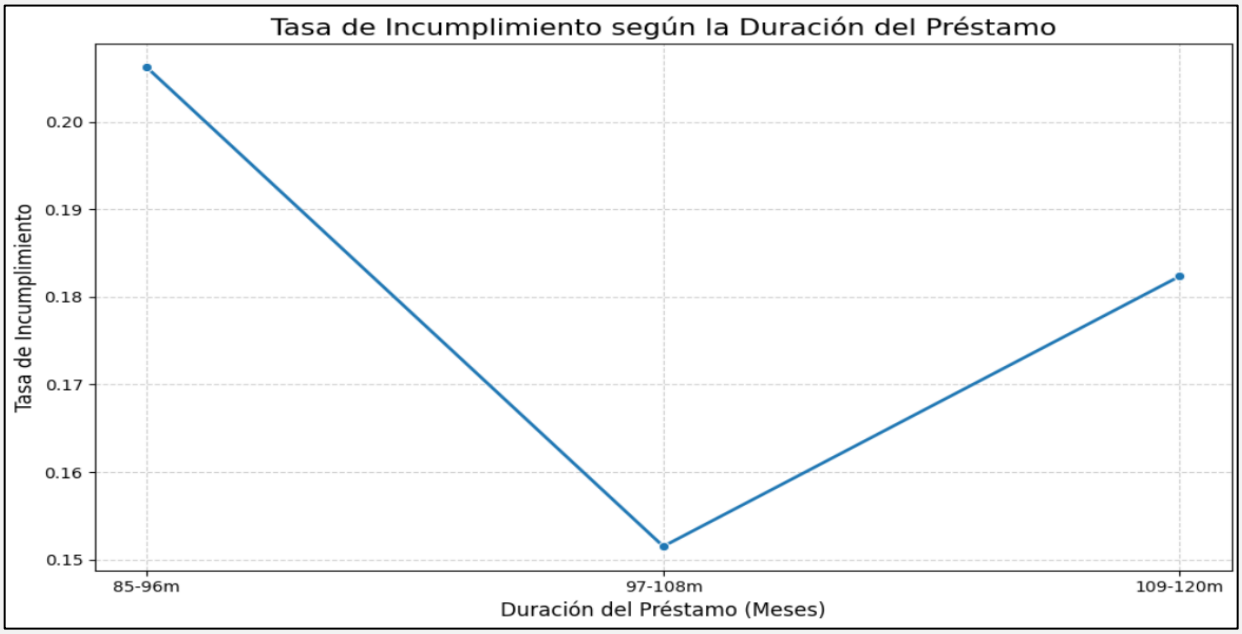
4. Los solicitantes con ingresos más bajos presentan una mayor tasa de incumplimiento en comparación con aquellos con ingresos más altos?

Podemos observar que existe una mayor tasa de incumplimiento en comparación de solicitantes de menores ingresos con los de mayores ingresos.



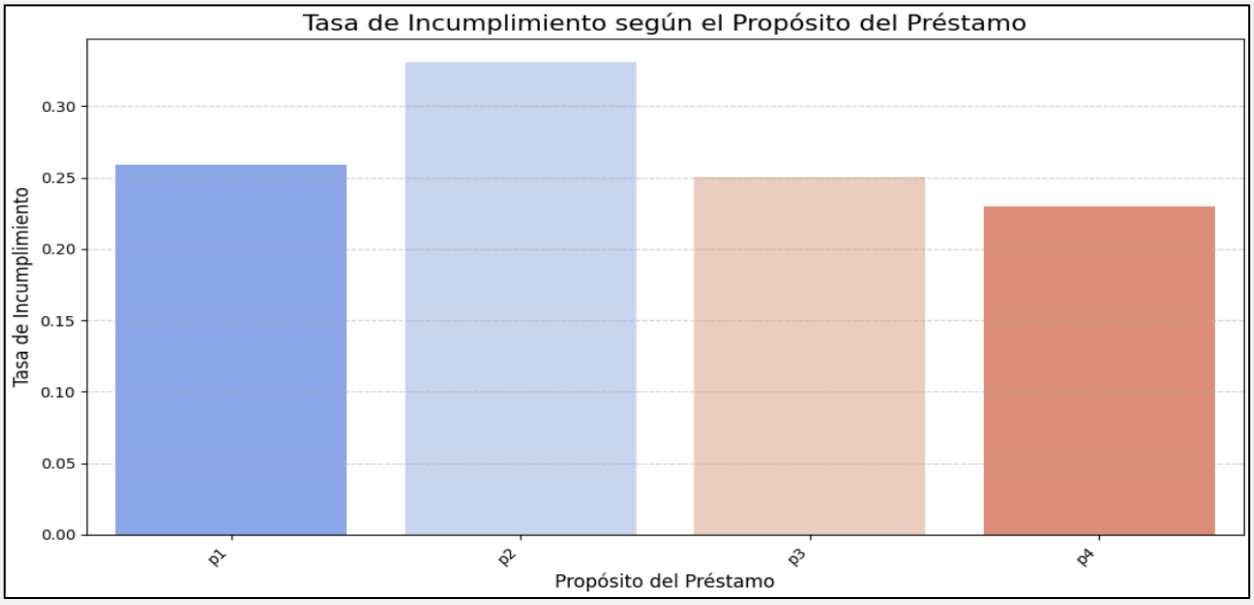
5. La tasa de incumplimiento varía según la duración del préstamo (plazo en meses o años).

La tasa de incumplimiento aparece recién entre los meses 85 y 96 con un casi 22%, luego cae un 7% en los meses 97 y 108. Y entre los últimos 109 a 120 meses se acomoda en un término medio de un 18%.



6. Los préstamos destinados a diferentes propósitos tienen tasas de incumplimiento distintas?

Podemos ver que la tasa de incumplimiento varia mínimamente entre los distintos propósitos de préstamos, situada en un promedio del 25%.



Construcción de Modelos:

A continuación, creamos la función **matriz de confusión** que visualiza el desempeño del modelo en términos de clasificación correcta e incorrecta. Genera dos representaciones:

1. **Sin Normalizar:** Indica el número absoluto de predicciones correctas e incorrectas.
2. **Normalizada:** Muestra las proporciones relativas, lo que facilita la interpretación en DataSet desbalanceados.

Esta herramienta permite identificar posibles problemas del modelo, como tendencias a equivocarse más en una clase específica (falsos positivos o falsos negativos).

1 - Selección de Algoritmos: Elegimos algoritmos de clasificación adecuados (Regresión Logística, Random Forest, Árbol de Decisión).

Regresión Logística:

Utilizamos el modelo de **Regresión Logística** para predecir la variable objetivo.

- El modelo se entrena y evalúa de manera eficiente, proporcionando métricas clave y visualizaciones que permiten analizar su rendimiento.
- La matriz de confusión permite analizar los errores específicos del modelo en términos de falsos positivos y falsos negativos.

Al evaluar el modelo de **Regresión Logística**, obtuvimos un **accuracy del 86.5%**, lo que indica que, en general, el modelo clasifica correctamente la mayoría de las observaciones.

Sin embargo, al analizar métricas más específicas, como el **recall del 48.42%**, identificamos que el modelo tiene dificultades para detectar todos los casos de incumplimiento (Default), lo que resulta en un alto número de falsos negativos.

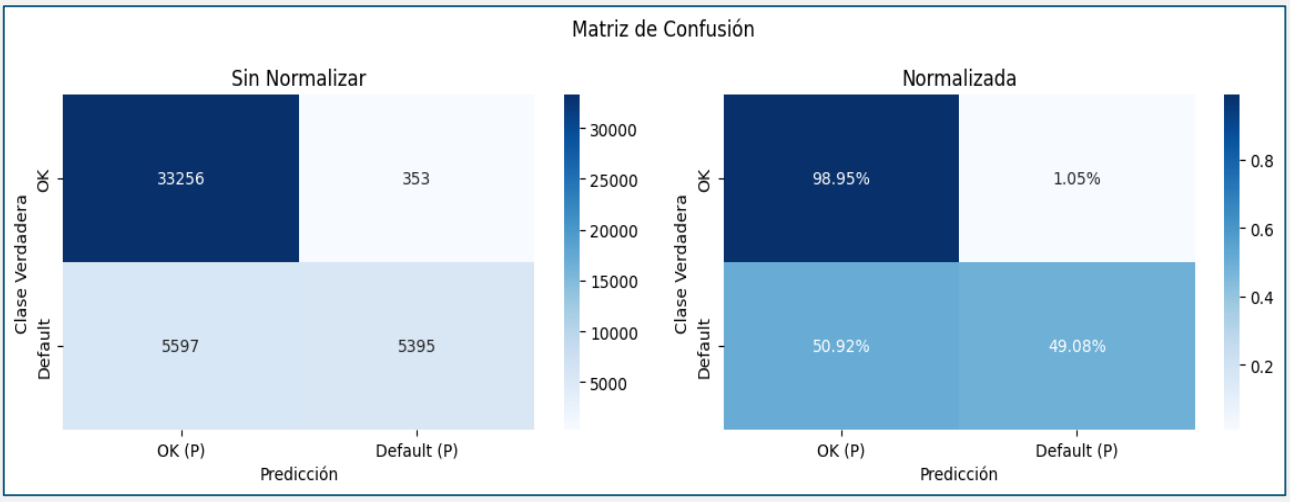
Esto se refleja claramente en la matriz de confusión, donde 5670 incumplimientos no fueron detectados.

Por otro lado, el **F1-Score del 63.86%** nos muestra un equilibrio moderado entre precisión y recall, aunque con una mayor inclinación hacia la precisión (93.78%), lo que significa que el modelo es confiable al predecir incumplimientos pero tiene un costo significativo en términos de casos omitidos.

En resumen, el modelo tiene un desempeño aceptable para casos generales, pero requiere ajustes para minimizar falsos negativos, como es el caso en aplicaciones críticas de riesgo crediticio.

Resultados del modelo: Regresión Logística

Accuracy: 0.8666
Precision: 0.9386
Recall: 0.4908
F1-Score: 0.6446
ROC-AUC: 0.7402



Random Forest:

Utilizamos el modelo de **Random Forest** para predecir la variable objetivo. Al igual que Regresión Logística.

- El modelo se entrena y evalúa, proporcionando métricas clave y visualizaciones que permiten analizar su rendimiento.
- La matriz de confusión permite analizar los errores específicos del modelo en términos de falsos positivos y falsos negativos.

Al analizar el desempeño del modelo **Random Forest**, observamos resultados perfectos, con métricas clave como **accuracy, precisión, recall, F1-Score y ROC-AUC** alcanzando el 100%. Esto indica que el modelo clasificó correctamente todos los casos, sin cometer errores.

La matriz de confusión confirma este rendimiento ideal. En ella, se puede observar que no hay falsos positivos ni falsos negativos: todos los casos de 'Default' y 'OK' fueron correctamente clasificados.

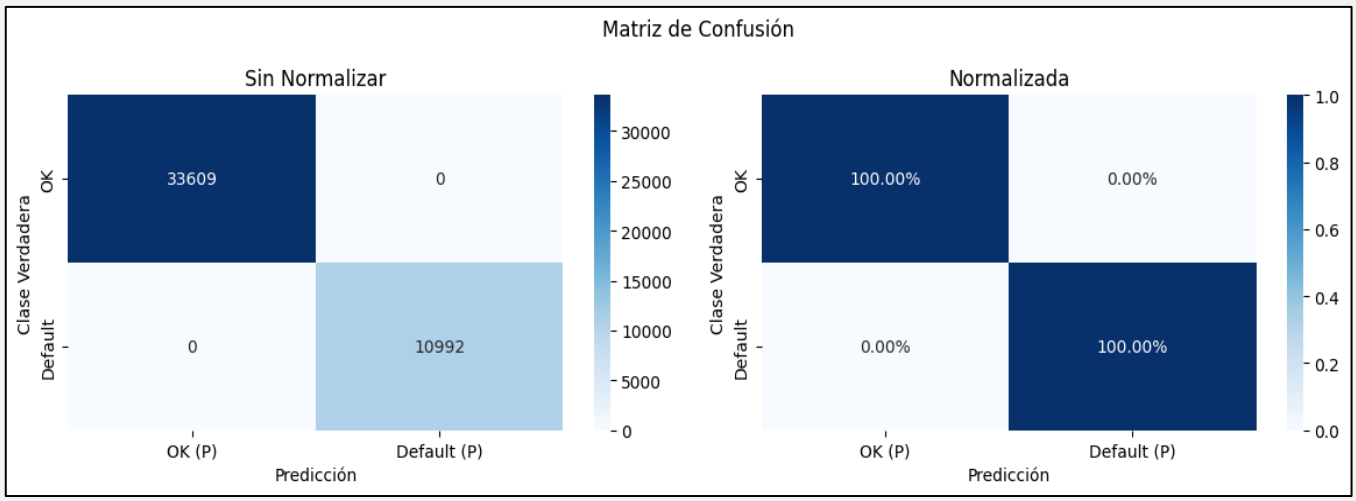
En términos prácticos, esto significa que el modelo tiene una capacidad excepcional para distinguir entre ambas clases.

Es importante señalar que, aunque estos resultados son impresionantes, podrían ser un indicativo de sobreajuste, especialmente si el conjunto de datos de prueba no es suficientemente representativo o si las clases están desbalanceadas.

Por lo tanto, recomendaría realizar validaciones adicionales, como validación cruzada, para confirmar que el modelo generaliza correctamente a nuevos datos.

Resultados del modelo: Random Forest

Accuracy: 1.0000
Precisión: 1.0000
Recall: 1.0000
F1-Score: 1.0000
ROC-AUC: 1.0000



Árbol de Decisión:

El **Árbol de Decisión** es un modelo intuitivo que divide los datos en subgrupos, lo que lo hace útil para entender cómo se toman las decisiones de clasificación.

Al igual que en los modelos anteriores:

- El modelo se entrena y evalúa, proporcionando métricas clave y visualizaciones que permiten analizar su rendimiento.
- La matriz de confusión permite analizar los errores específicos del modelo en términos de falsos positivos y falsos negativos.

El modelo de **Árbol de Decisión** mostró un desempeño casi perfecto en la clasificación, alcanzando un **accuracy, recall**, **F1-Score** y **ROC-AUC del 100%** y métricas como **precisión**, que se encuentran cerca del 100%.

Esto significa que el modelo clasificó correctamente prácticamente todas las observaciones.

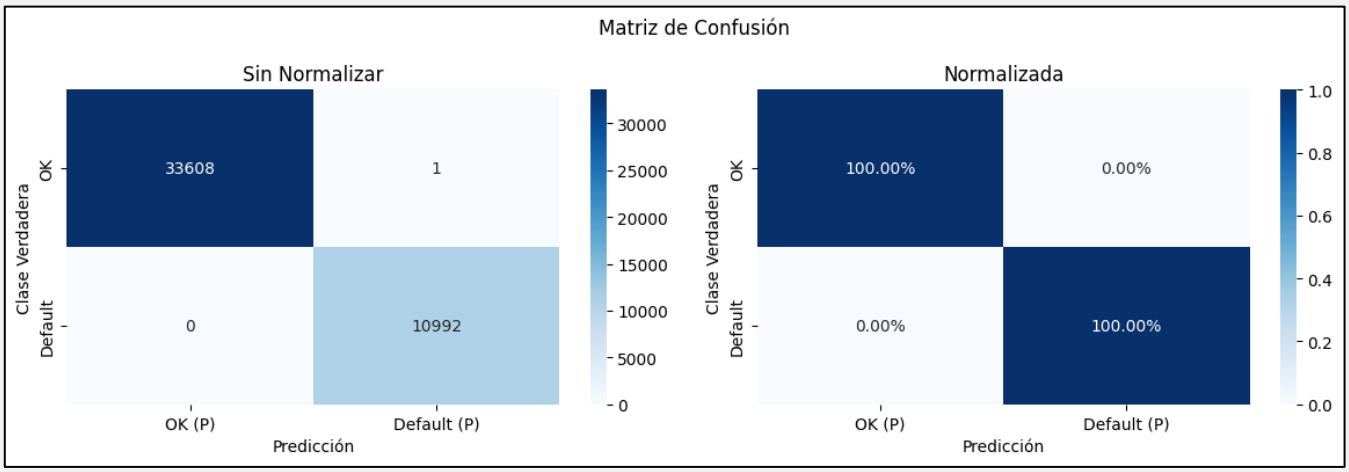
La matriz de confusión refleja este rendimiento. En términos absolutos, el modelo predijo correctamente todos los casos de 'Default' y 'OK', salvo por un único caso erróneo donde se clasificó un cliente como 'Default' cuando en realidad era 'OK'. Esta desviación mínima se traduce en una precisión ligeramente inferior al 100% (99.99%).

En resumen, este modelo demuestra ser altamente confiable y efectivo para distinguir entre las clases 'Default' y 'OK'.

Sin embargo, para garantizar que este rendimiento no se deba a un sobreajuste, recomendaría realizar validaciones adicionales y probarlo en diferentes conjuntos de datos.

Resultados del modelo: Árbol de Decisión

Accuracy: 1.0000
Precision: 0.9999
Recall: 1.0000
F1-Score: 1.0000
ROC-AUC: 1.0000



A continuación, comparamos los resultados de los 3 modelos en el **DataFrem** creado:

```
resultados_df
```

	Modelo	Accuracy	Precision	Recall	ROCAUC	F1-Score	Tiempo
0	LogisticRegression	0.866595	0.938587	0.490811	0.740154	0.644564	4.331839
1	Random Forest	1.000000	1.000000	1.000000	1.000000	1.000000	10.421660
2	Árbol de Decisión	0.999978	0.999909	1.000000	0.999985	0.999955	0.406232

Conclusiones Generales

Resumen de los datos obtenidos:

Rendimiento Global:

Los modelos basados en árboles (**Random Forest** y **Árbol de Decisión**) superan significativamente a la **Regresión Logística** en todas las métricas. Esto los hace más adecuados para tareas donde la clasificación precisa de Default es esencial.

Sobreajuste:

Los resultados perfectos de **Random Forest** y **Árbol de Decisión** podrían ser indicativos de sobreajuste. Se recomienda realizar validaciones cruzadas y pruebas en conjuntos de datos más diversos para confirmar la generalización.

Uso Potencial:

Regresión Logística: Útil para aplicaciones rápidas donde la simplicidad es prioritaria y la interpretabilidad es suficiente.

Random Forest: Ideal para situaciones críticas que requieren alta precisión y sensibilidad, aunque con un costo computacional mayor.

Árbol de Decisión: Excelente opción si buscamos un modelo interpretable con un desempeño casi perfecto.

Recomendaciones sobre los modelos:

Regresión Logística:

Mejoras potenciales: Realizar ajustes en el modelo para mejorar el Recall, como regularización o ajustar el umbral de clasificación.

Aplicar técnicas de manejo de datos desbalanceados, como el sobre muestreó de la clase minoritaria (Default) o el uso de pesos en las clases.

Random Forest:

Mejoras potenciales: Realizar validación cruzada para asegurar que los resultados perfectos no se deban a sobreajuste.

Ajustar hiperparámetros como el número de árboles, la profundidad máxima, y los criterios de división para encontrar un mejor equilibrio entre precisión y generalización.

Árbol de Decisión:

Mejoras potenciales: Limitar la profundidad máxima del árbol para prevenir sobreajuste, especialmente si los datos de prueba no son representativos.

Evaluar la importancia de las características para simplificar aún más el modelo y priorizar interpretabilidad.

Pruebas con modelos adicionales:

Considerar probar otros modelos para comparar su rendimiento y ver si se pueden obtener mejores resultados.