INFORME DEL PROYECTO - CREDIT DEFAULT

- **Carrera**: Licenciatura en Ciencias de Datos
- **Materia**: Introducción a la Ciencia de Datos
- **Profesor**: Claudio Gonzalo Pavón
- **Integrantes**:
    - ABRAHAM, Milena Paula
    - ARESE, Juan Cruz
    - CESCHINI, Francisco
    - COGORNO DÍAZ, Lautaro Javier

Repositorio de GitHub: https://github.com/JuanCruzArese/TP-Final-ICD

Fecha de entrega: 27 de junio del 2025

OBJETIVO DEL PROYECTO

Analizando un dataset de clientes de tarjetas de crédito de Taiwán y utilizando modelos de Machine Learning (Random Forest y Decision Tree), predecimos si una persona incumplirá o no el pago de una tarjeta de crédito.

VARIABLES DEL DATASET

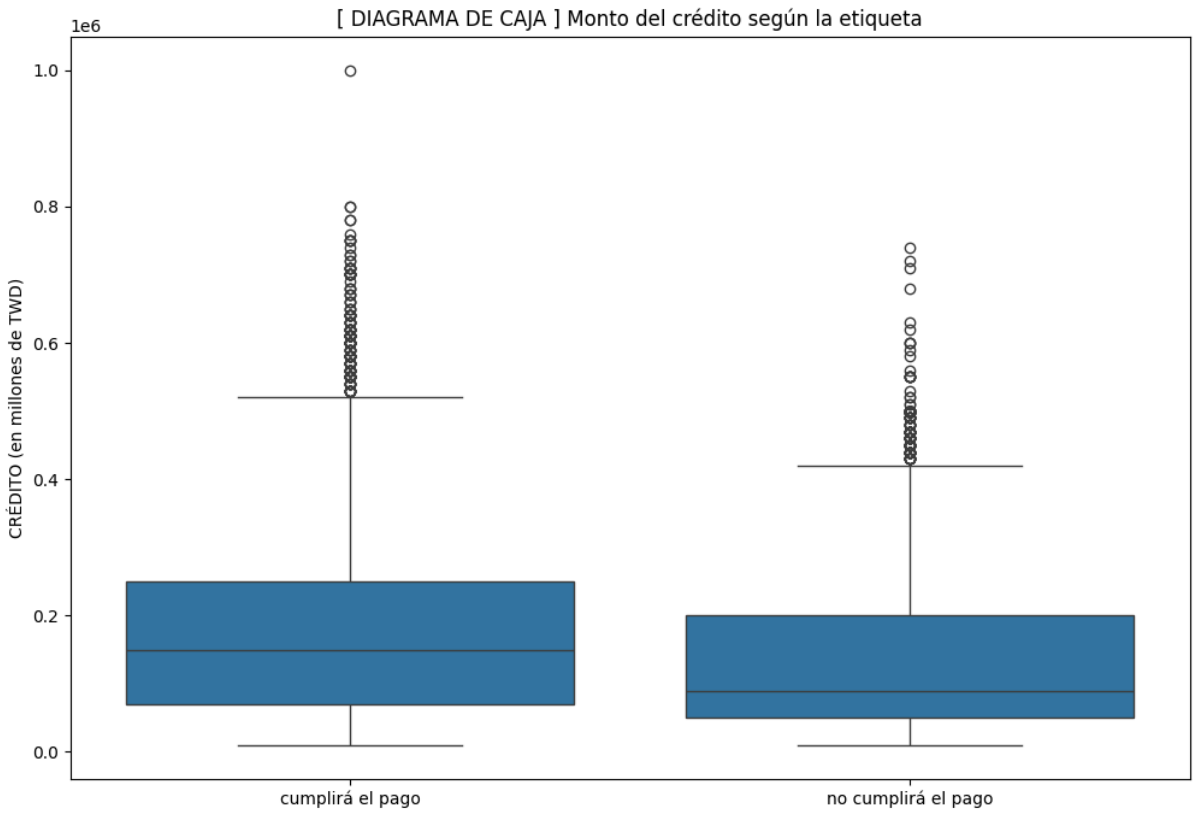
En la pagina del dataset en Kaggle.com, el autor del mismo explica el significado de cada columna:

COLUMNA	DESCRIPCIÓN
ID	ID de cada cliente
LIMIT_BAL	Monto del crédito otorgado en dólares taiwaneses (incluye crédito individual y familiar/complementario)
SEX	Género (1=masculino, 2=femenino)
EDUCATION	1=posgrado, 2=universitario, 3=preparatoria, 4=otros, 5=desconocido, 6=desconocido
MARRIAGE	Estado civil (1=casado, 2=soltero, 3=otros)
AGE	Edad en años
PAY_X	Estado de pago en un determinado mes de 2005 (-1=pago al día, 1=un mes de retraso en el pago, 2=dos meses de retraso en el pago,...9=nueve meses o más de retraso en el pago)

BILL_AMTX	Importe del extracto de factura en un determinado mes de 2005 (dólares taiwaneses)
PAY_AMTX	Importe del pago anterior en un determinado mes de 2005 (dólares taiwaneses)
default.payment.next.month	Pago default (1=sí, 0=no)

ANÁLISIS DE LOS DATOS

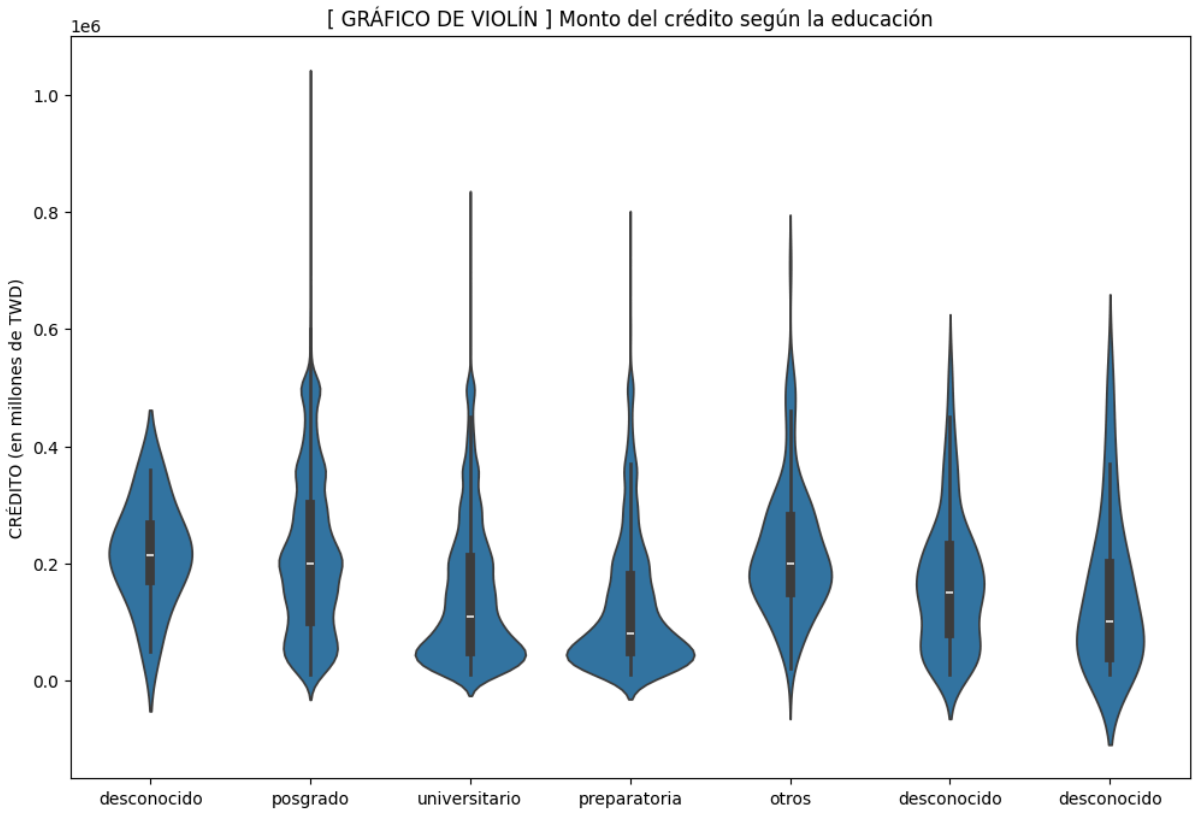
- BOX PLOT



En este box plot, se puede ver que, en general, las personas que no cumplirán con el pago fueron otorgadas un crédito más bajo. El recorrido

intercuartílico (el rectángulo azul), donde se encuentra el 50% de la población, es más compacto en aquellos que no cumplen con el pago, sugiriendo que la mayoría obtuvo un crédito parecido.

- **VIOLIN PLOT**



En este gráfico de violín, podemos ver el crédito que obtuvo cada población según su nivel de educación. Los estudiantes de preparatoria y universitarios, en promedio, recibieron créditos más bajos.

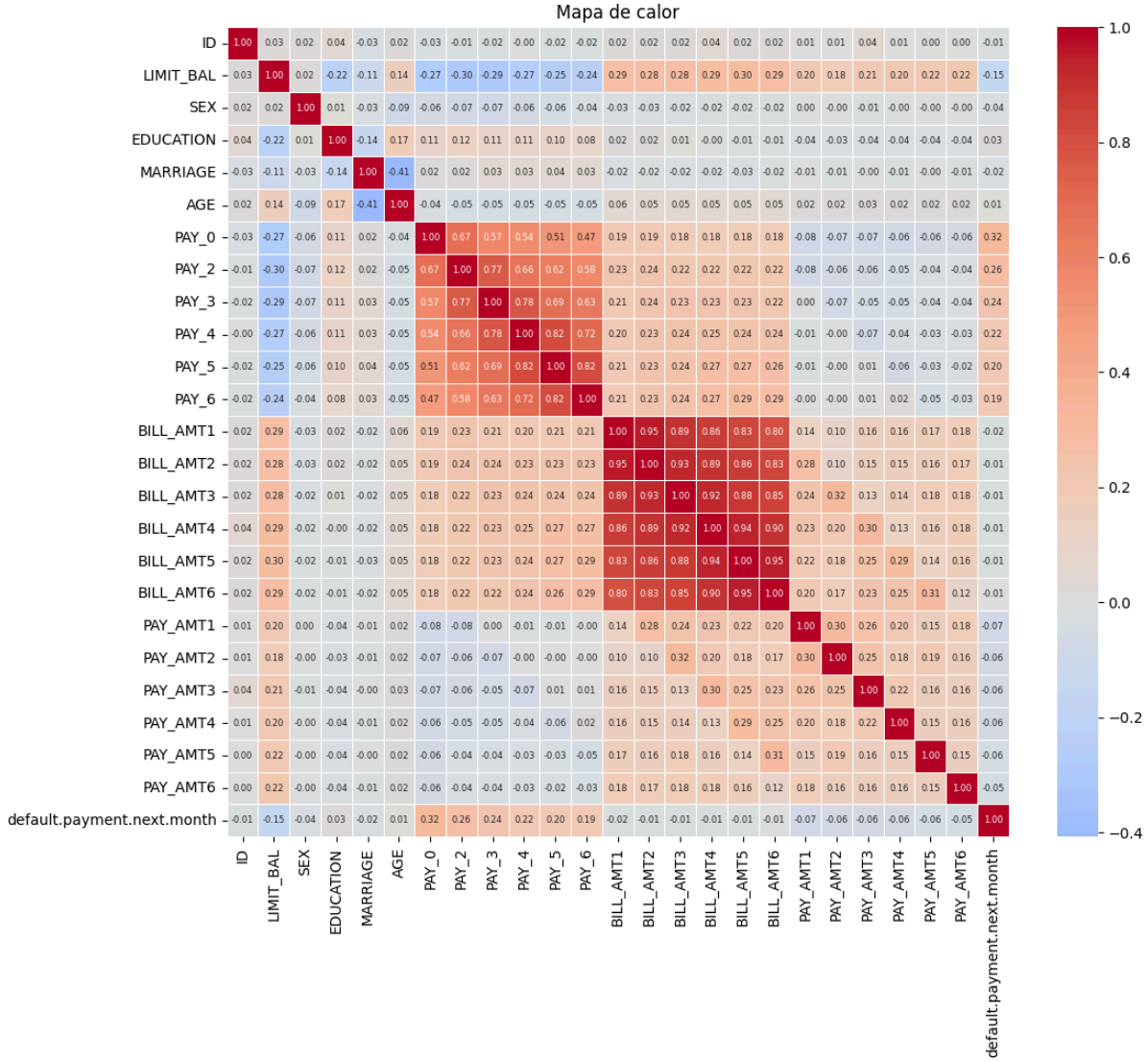
- OUTLIERS

Las 5 columnas con más outliers del dataset:

COLUMNA	OUTLIERS	PORCENTAJE DE OUTLIERS
PAY_2	4410	14.7
PAY_3	4209	14.03
PAY_4	3508	11.69
PAY_0	3130	10.43
PAY_6	3079	10.26

Un outlier es un valor que se sale del intervalo [Q1 - 1.5 * IQR, Q3 + 1.5 * IQR]

- MAPA DE CALOR



En este gráfico se observa una marcada asociación positiva entre los montos de facturación mensual (variables BILL_AMT1 a BILL_AMT6) y los pagos realizados en los meses correspondientes (PAY_AMT1 a PAY_AMT6). Esto indica que, en general, los clientes que acumulan un mayor gasto con la tarjeta de crédito tienden también a efectuar pagos de mayor importe.

Asimismo, se evidencia una correlación negativa, aunque más moderada, entre las variables de historial de pago (PAY_0 a PAY_6) y el límite de crédito (LIMIT_BAL). Este patrón sugiere que los clientes con mayor

historial de mora suelen contar con un límite de crédito más bajo, posiblemente por políticas de riesgo aplicadas por la entidad financiera.

Por otro lado, la variable objetivo default.payment.next.month, que indica si el cliente incumplirá el pago el mes siguiente, no presenta fuertes correlaciones con ninguna variable individual, ya que el fenómeno a predecir no depende de una sola variable, sino de un conjunto de factores combinados.

RESULTADOS

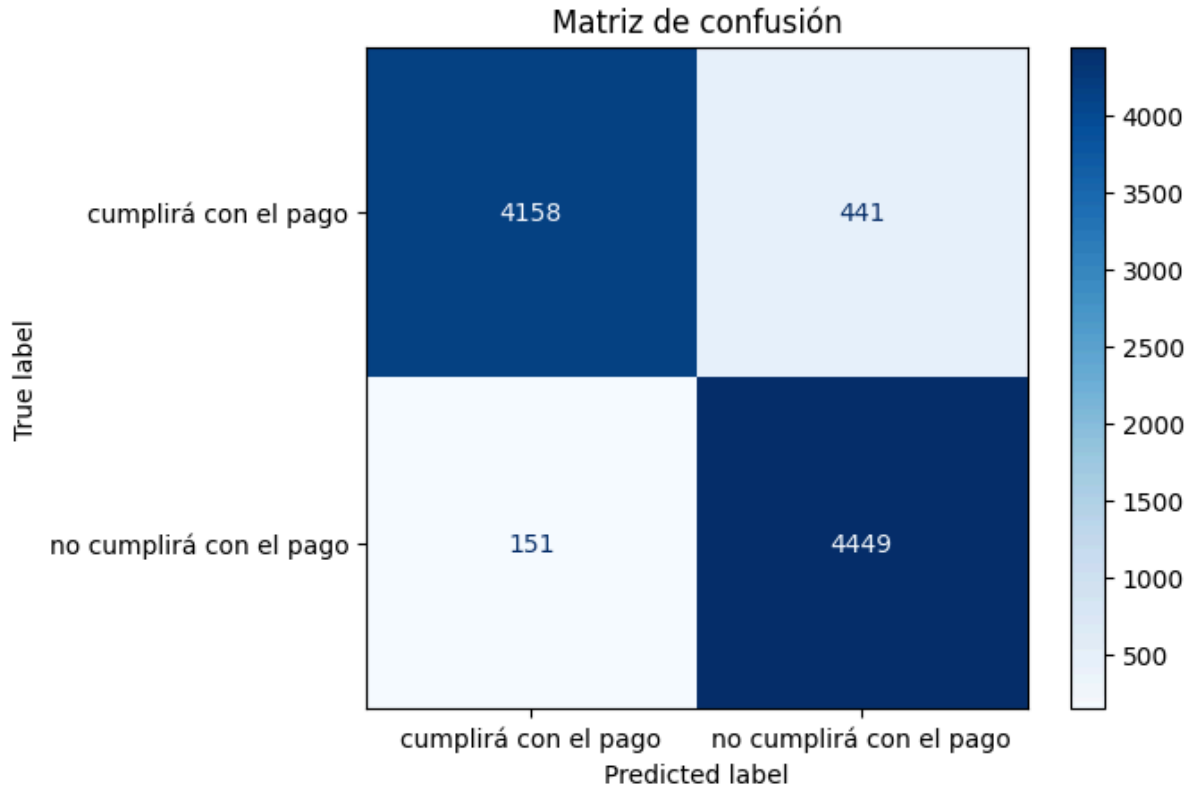
RANDOM FOREST

* Reporte de clasificación:

	precision	recall	f1-score	support
0	0.96	0.90	0.93	4599
1	0.91	0.97	0.94	4600
accuracy			0.94	9199
macro avg	0.94	0.94	0.94	9199
weighted avg	0.94	0.94	0.94	9199

(*) Resultados de una ejecución, podrían no ser siempre los mismos.

MATRIZ DE CONFUSIÓN



La matriz de confusión del Random Forest nos dice que de las 4599 veces en las que el cliente cumplirá con el pago, 4158 veces la predicción del modelo fue correcta (verdaderos positivos) y 441 veces incorrecta (falsos positivos). Por el otro lado, de las 4600 veces en las que el cliente no pagará, 4449 veces su predicción fue correcta (verdaderos negativos) mientras que 151 veces su predicción fue incorrecta (falsos negativos).

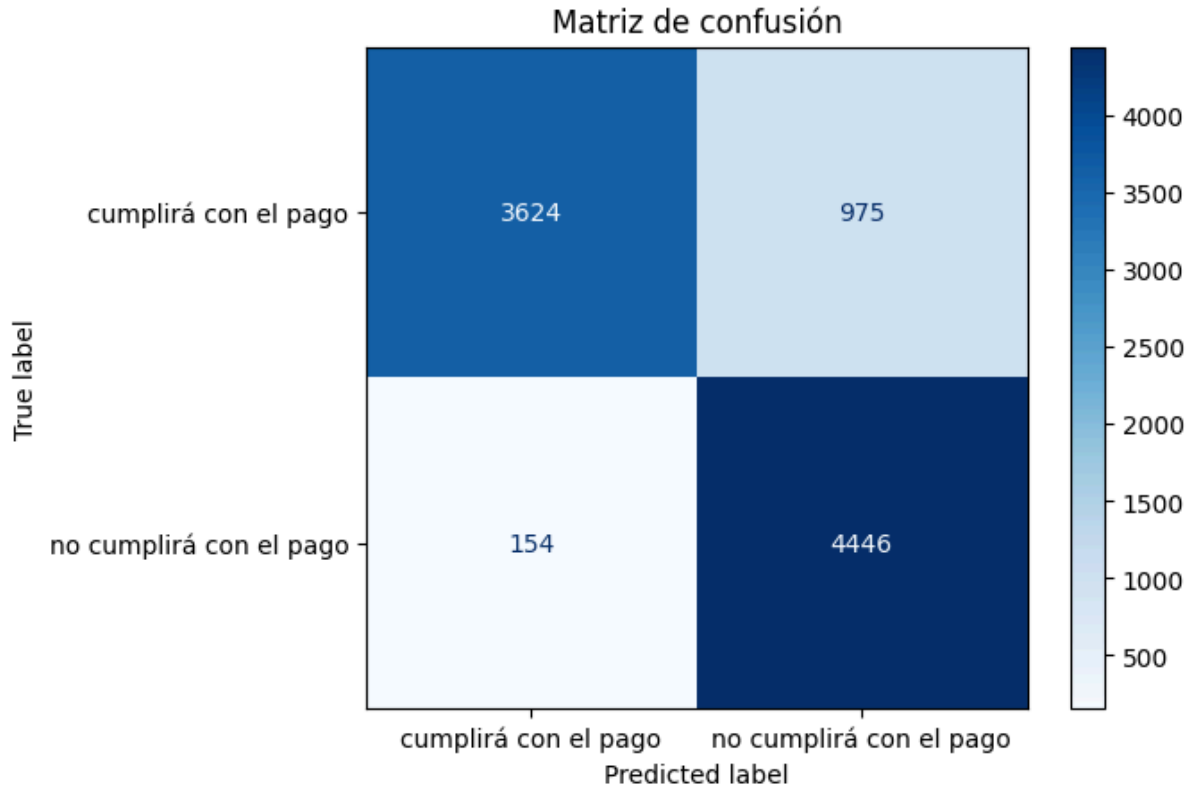
DECISION TREE

* Reporte de clasificación:

	precision	recall	f1-score	support
0	0.96	0.79	0.87	4599
1	0.82	0.97	0.89	4600
accuracy			0.88	9199
macro avg	0.89	0.88	0.88	9199
weighted avg	0.89	0.88	0.88	9199

(*) Resultados de una ejecución, podrían no ser siempre los mismos.

MATRIZ DE CONFUSIÓN

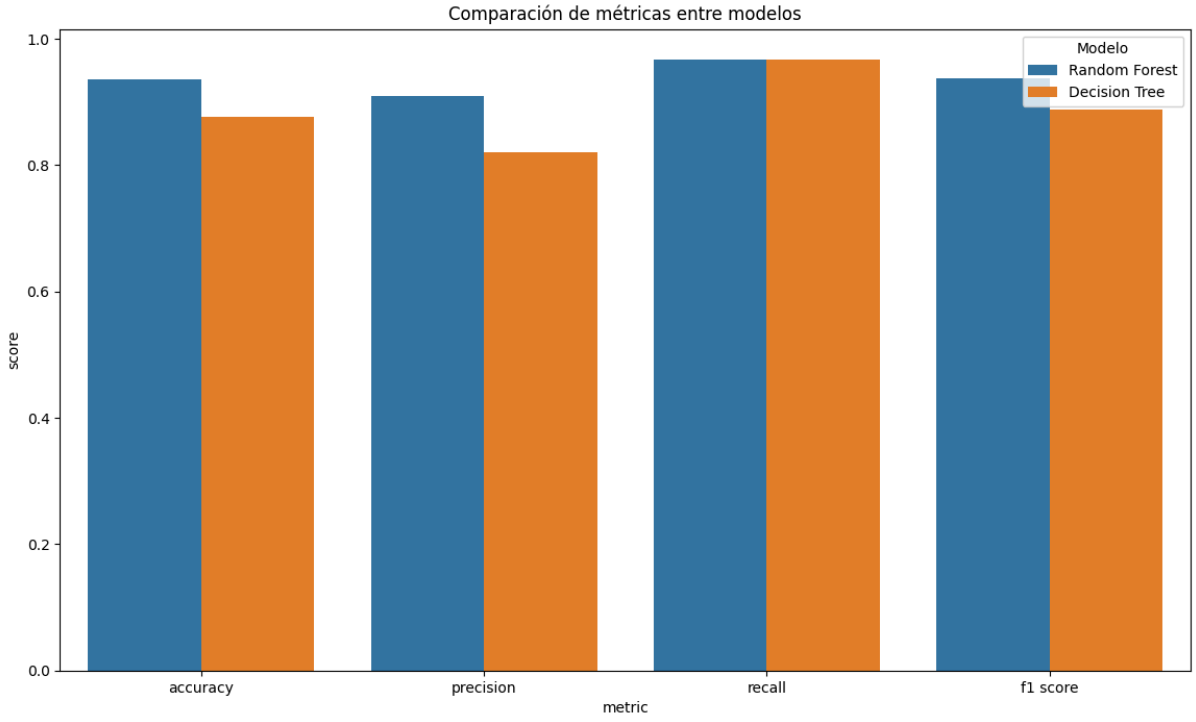


La matriz de confusión del Decision Tree nos dice que de las 4599 veces en las que el cliente cumplirá con el pago, 3624 veces la predicción del modelo fue correcta (verdaderos positivos) y 975 veces incorrecta (falsos positivos). Por el otro lado, de las 4600 veces en las que el cliente no pagará, 4446 veces su predicción fue correcta (verdaderos negativos) mientras que 154 veces su predicción fue incorrecta (falsos negativos).

MÉTRICAS DE EVALUACIÓN

precision	VP / (VP + FP). De las predicciones 'cumplirá el pago', proporción realmente correcta.
recall	VP / (VP + FN). De las predicciones 'cumplirá el pago', la proporción detectada correctamente.
f1-score	2 * (precision * recall)/(precision + recall). Media armónica de precision y recall
support	Número de entradas del dataset con esa etiqueta usadas para la prueba del modelo.

COMPARACIÓN ENTRE MODELOS



Modelo	accuracy	precision	recall	f1-score
Random Forest	0.935	0.909	0.967	0.937
Decision Tree	0.877	0.820	0.966	0.887

El modelo de Random Forest alcanzó una precisión global del 93,5%, mientras que el Árbol de Decisión obtuvo una precisión del 87,7%. Este resultado sugiere que Random Forest fue más efectivo a la hora de predecir correctamente tanto los casos de incumplimiento como los de pago regular.

La ventaja del Random Forest se explica principalmente por su estructura: al ser un conjunto de múltiples árboles de decisión que se entrenan sobre

distintos subconjuntos de datos y variables, el modelo logra una generalización más robusta y reduce el riesgo de sobre entrenamiento. En cambio, el Árbol de Decisión, al ser un modelo único, tiende a capturar patrones muy específicos del set de entrenamiento, lo que afecta negativamente su desempeño sobre nuevos datos. Esto se refleja en su menor precisión, que indica una mayor cantidad de clasificaciones erróneas entre los clientes que en realidad no incurrirán en incumplimiento.

CONCLUSIÓN

El presente trabajo abordó el desafío de predecir el incumplimiento de pago de clientes de tarjetas de crédito utilizando técnicas de Machine Learning sobre un conjunto de datos reales de Taiwán. A partir de un análisis exploratorio detallado, fue posible identificar comportamientos recurrentes entre los clientes que no cumplen con sus pagos, como menores límites de crédito y patrones distintos según nivel educativo. El mapa de calor evidenció que no existe una única variable fuertemente correlacionada con el default, lo que resalta la necesidad de modelos capaces de captar relaciones complejas y no lineales entre múltiples factores.

En ese sentido, la comparación entre el modelo de Árbol de Decisión y Random Forest resultó clave. Si bien ambos lograron detectar con alto recall a los clientes en riesgo, Random Forest obtuvo mejores métricas generales, con mayor precisión y menor cantidad de errores de

clasificación. Esto se traduce en una herramienta predictiva más confiable para anticipar el comportamiento de pago y reducir el riesgo financiero.