



RANDOM & FOREST XGBOOST

**PARA LA CLASIFICACIÓN DE
DATOS DESBALANCEADOS**

Elaborado por:
María Isabella Meneses Ospina

Deteccción de fraudes con tarjetas de crédito

Este conjunto de datos contiene transacciones con tarjetas de crédito que han sido anonimizadas y etiquetadas como fraudulentas o genuinas.

Total de transacciones:

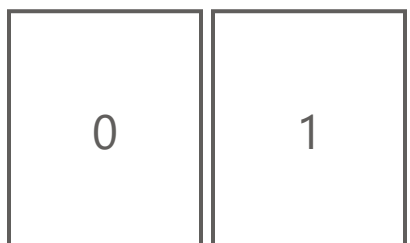
284,807 mil

Tipo de transacción

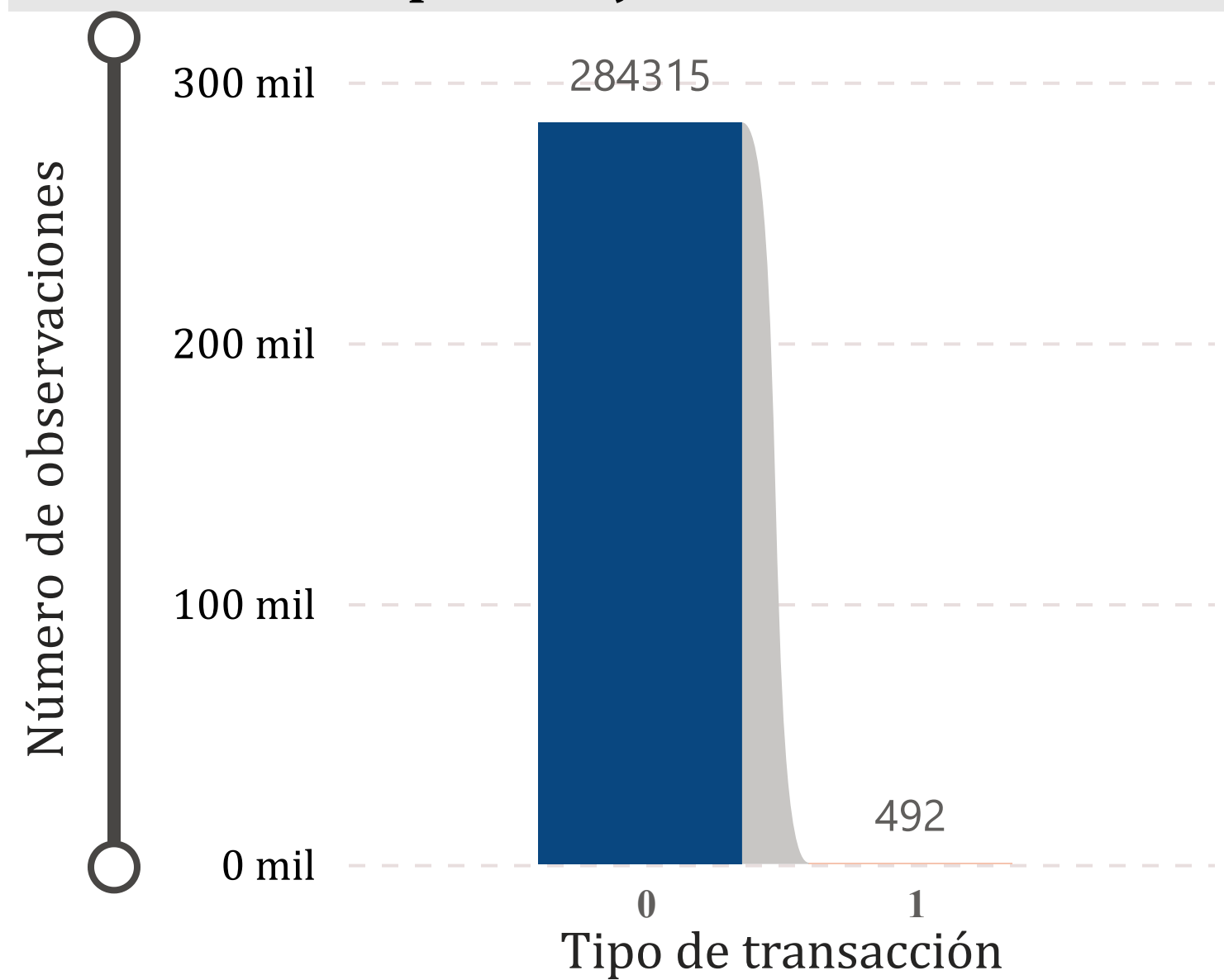
0: Transacción **no fraudulenta**

1: Transacción **fraudulenta**

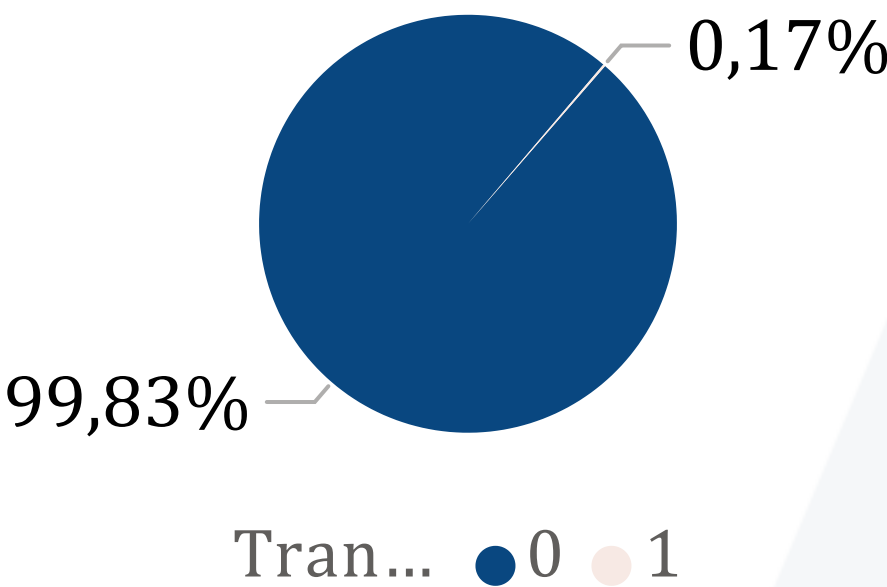
Clase



Distribución de clases detección de fraude por tarjeta de crédito

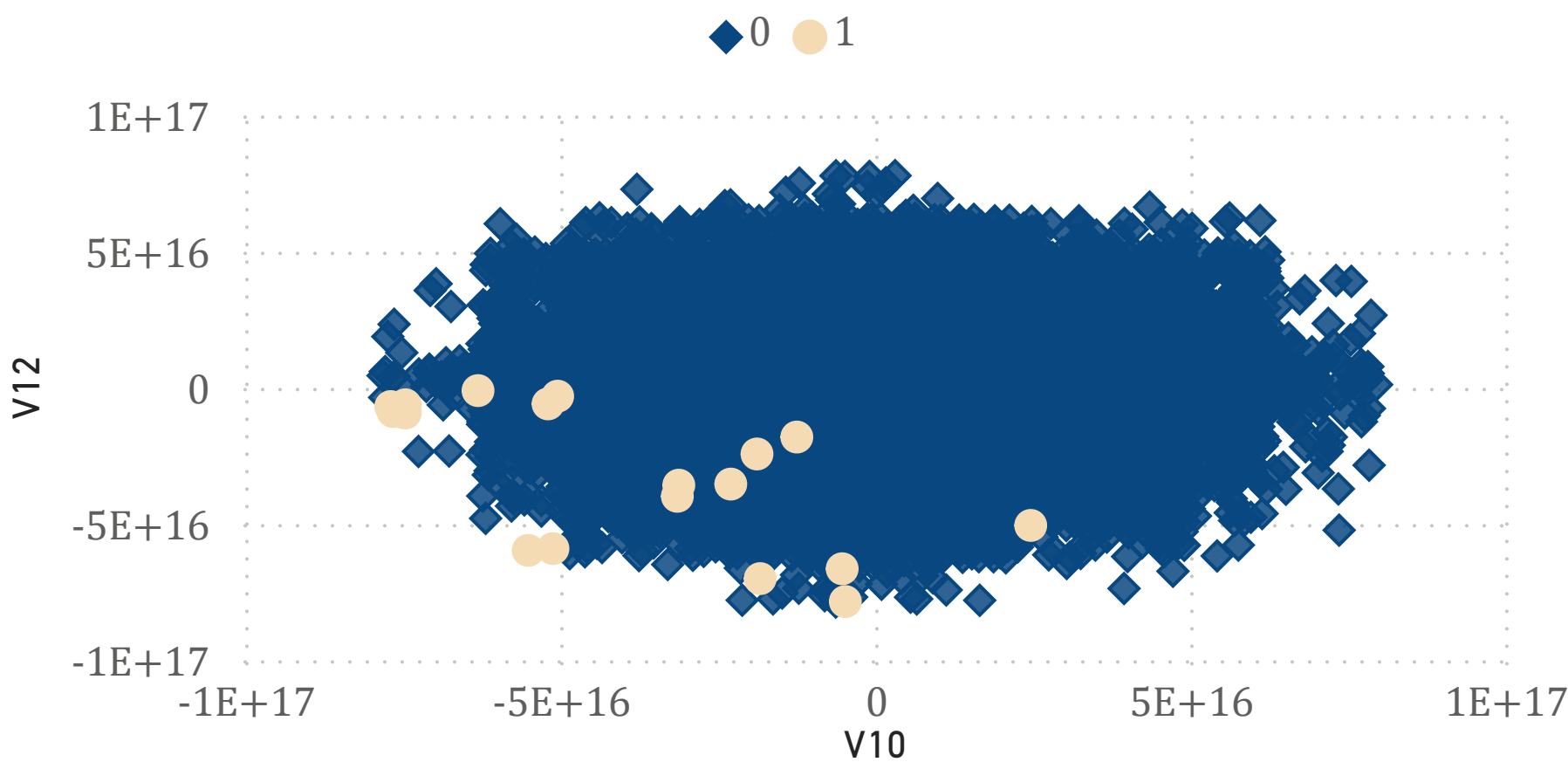


Proporción de Clases: Fraude vs. No Fraude



El dataset presenta un desbalance extremo (Malek, 2023), con la clase minoritaria representando solo el 0,17% del total.

Relación entre Características de Transacciones por Clases



X

V10

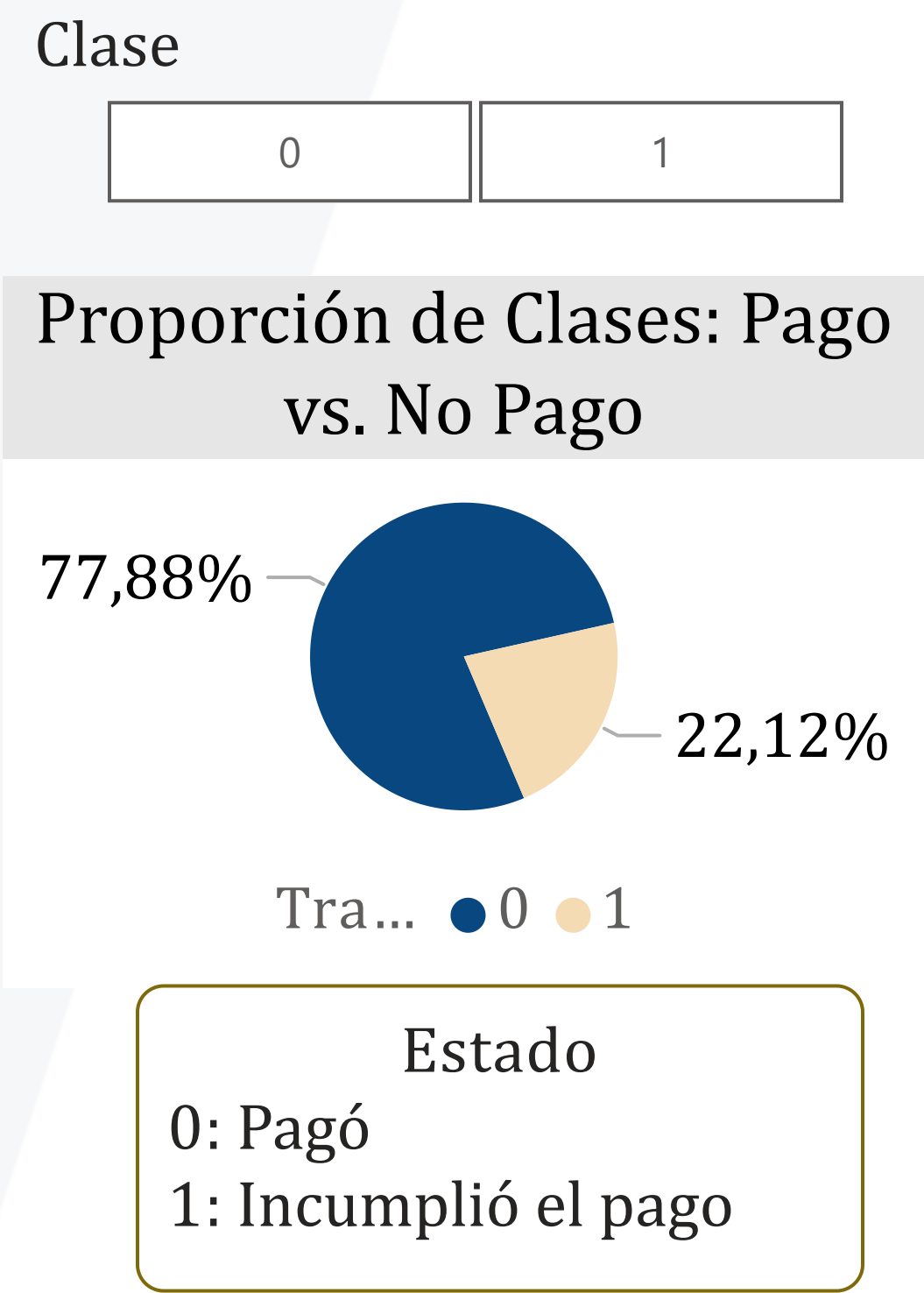
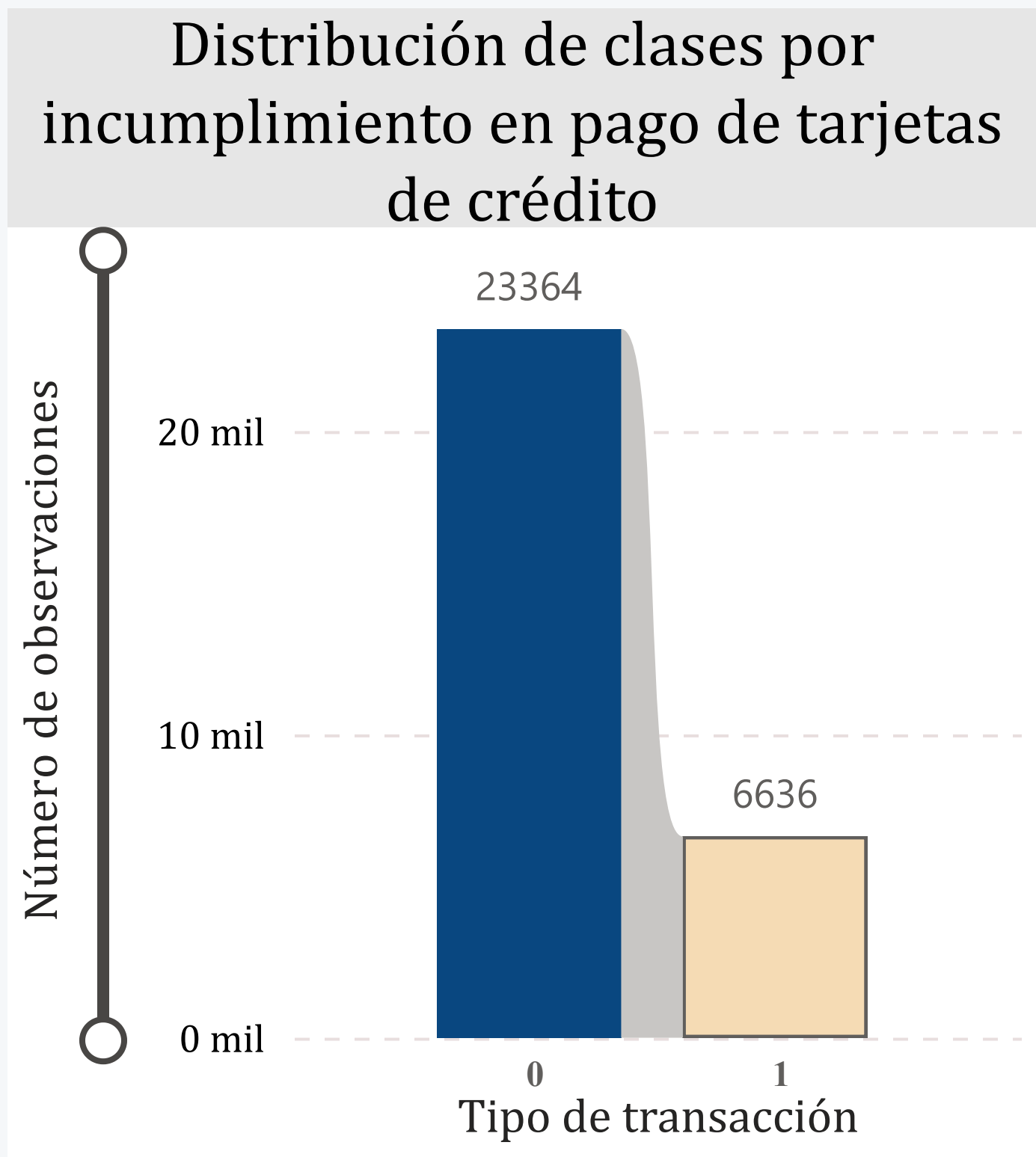
Y

V12

- [V1, V2, ..., V28] : Características obtenidas a través de transformación PCA.
- 'Time': (s) entre cada transacción y la primera
- 'Amount': Dinero retirado en cada transacción.
- 'Class': Variable respuesta

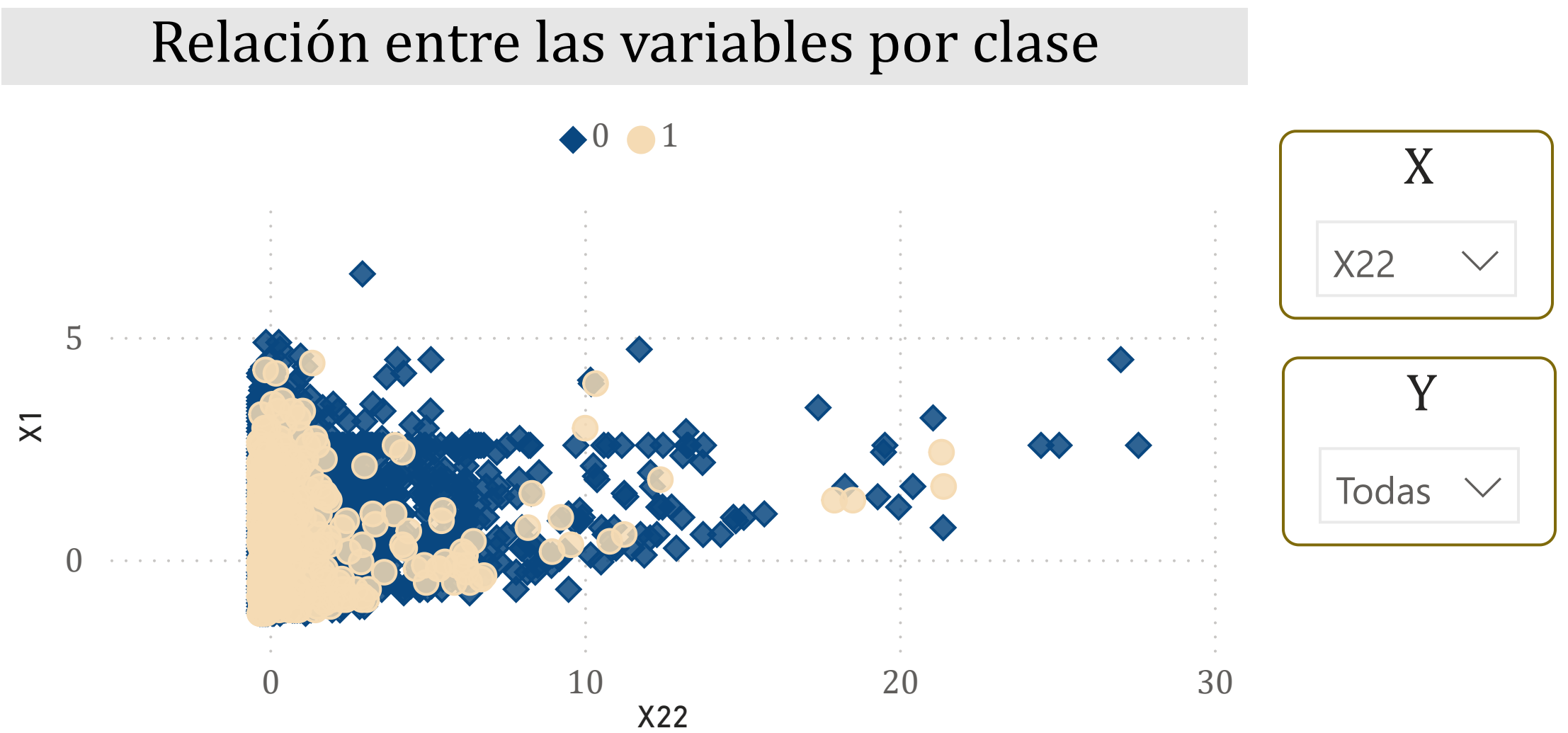
Impago de los clientes de tarjetas de crédito

Este conjunto de datos contiene información de clientes de tarjetas de crédito, incluyendo variables demográficas, financieras e históricas, con una variable de respuesta que indica si el cliente incumplió con el pago.



Total de registros:

30,000 mil



- X1: Monto total del crédito otorgado.
- X2: Género del cliente
- X3: Nivel educativo alcanzado.
- X4: Estado civil.
- X5: Edad del cliente.
- X6-X11 (PAY_0 a PAY_6): Historial de pagos mensuales desde abril a septiembre de 2005. Los valores oscilan de -1 (pagado puntualmente) a 9 (retraso de nueve meses o más).
- X12-X17 (BILL_AMT1 a BILL_AMT6): Monto mensual del estado de cuenta en dólares taiwaneses, desde septiembre (X12) hasta abril (X17) de 2005.
- X18-X23 (PAY_AMT1 a PAY_AMT6): Monto pagado mensualmente en

El dataset presenta un desbalance moderado (Malek, 2023), con la clase minoritaria representando 22,12% del total.

Random Forest

¡El poder colectivo de los árboles de decisión!

0,17%

Clase minoritaria

Base de
datos 1

22,12%

Clase minoritaria

Base de
datos 2

Parámetros configurados:

"n_estimators": [10, 12, 14, 16, 18, 20, 50, 100, 200],

"criterion": ["gini", "entropy"],

"max_features": ["sqrt", "log2"],

"max_depth": [10, 20, 30],

"min_samples_split": [2, 5, 10],

"min_samples_leaf": [1, 2, 4]}



Efecto de los parámetros y el Tamaño del conjunto de entrenamiento en el Recall

Selecciona la base de datos con el control "Base de datos", elige el parámetro de interés desde el menú desplegable "Parámetros RF" y ajusta la "técnica de re-muestreo" con las opciones disponibles (1: sin técnica de re-muestreo, 2: RUS, 3:ROS, 4: SMOTE) para explorar los resultados correspondientes para cada tamaño de conjunto de entrenamiento.

Base de datos

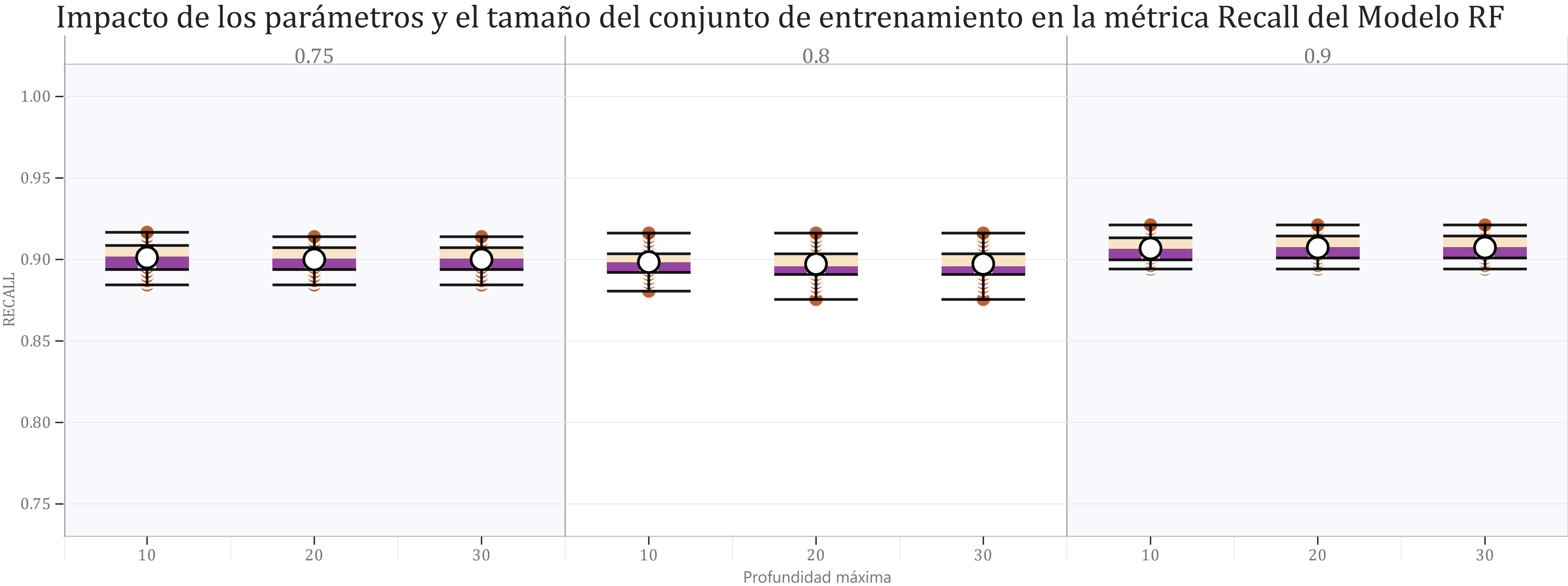
- 1
- 2

Parámetros RF

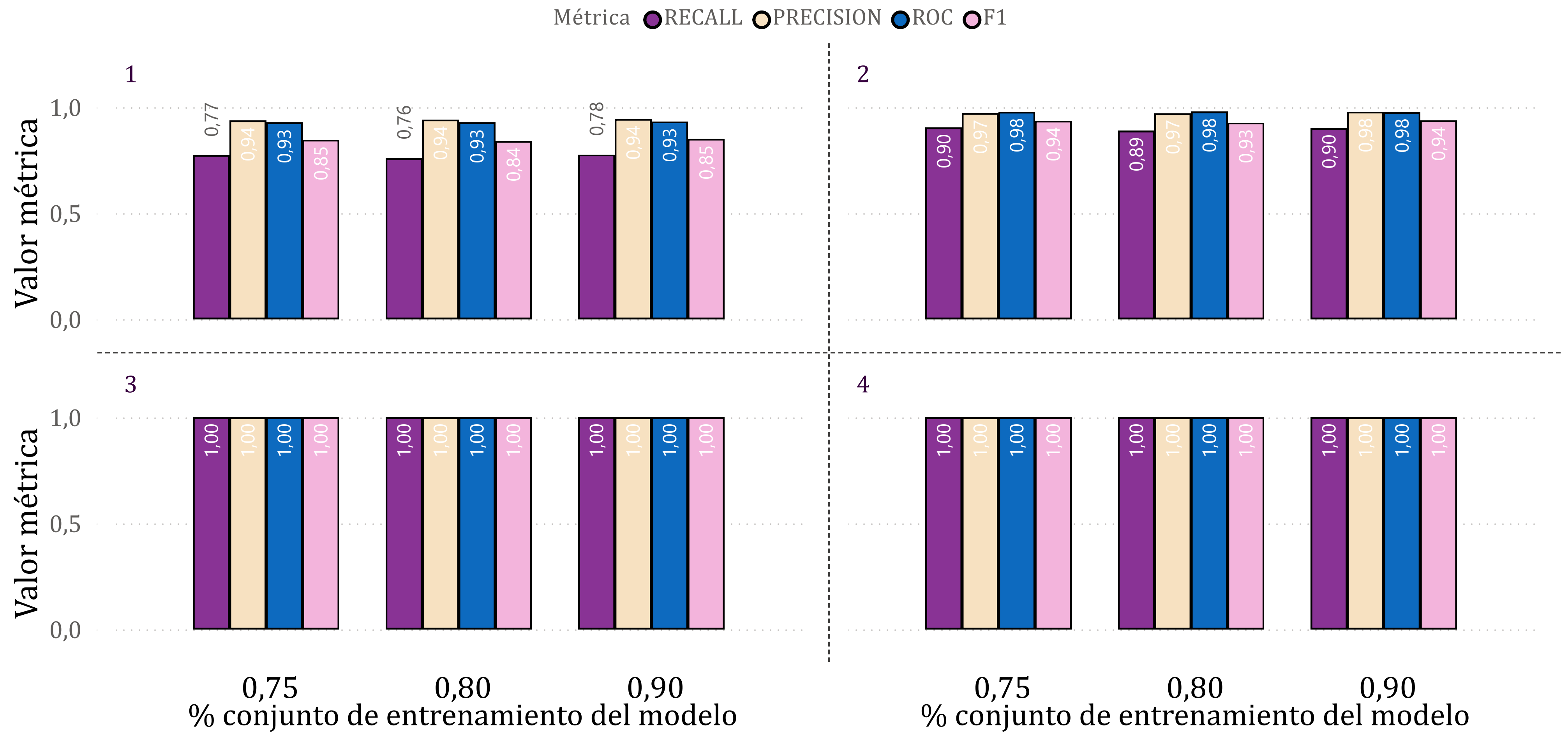
Profundidad máxima

Técnica de re-muestreo

1	3
2	4



COMPARACIÓN DE MÉTRICAS SEGÚN EL TAMAÑO DE ENTRENAMIENTO Y TÉCNICA APLICADA PARA EL MODELO RANDOM FOREST



BASE DE DATOS

1

2

Ajusta los filtros para modificar parámetros como el número de árboles, la profundidad máxima y el criterio de división según la base de datos seleccionada, y analiza cómo estas configuraciones impactan las métricas en cada escenario.

Criterio de división

☒ entropy

☐ gini

Mínimo de muestras p...

☐ 2

☒ 5

☐ 10

Características máximas

☐ log2

☒ sqrt

Mínimo de muestras d...

☐ 1

☒ 2

☐ 4

Número de árboles

☐ 10

☒ 12

☐ 14

☐ 16

☐ 18

Profundidad máxima

☐ 10

☒ 20

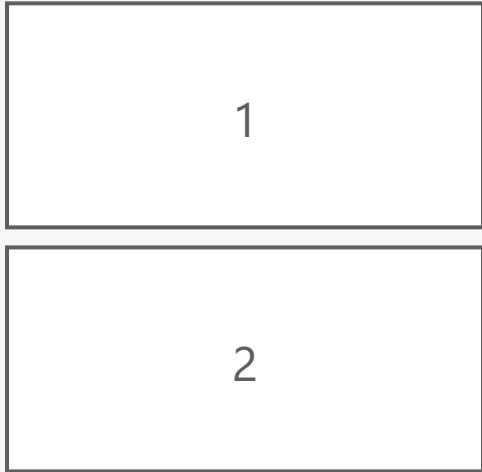
☐ 30

Elección de los mejores modelos a partir de la métrica Recall para el modelo Random Forest.

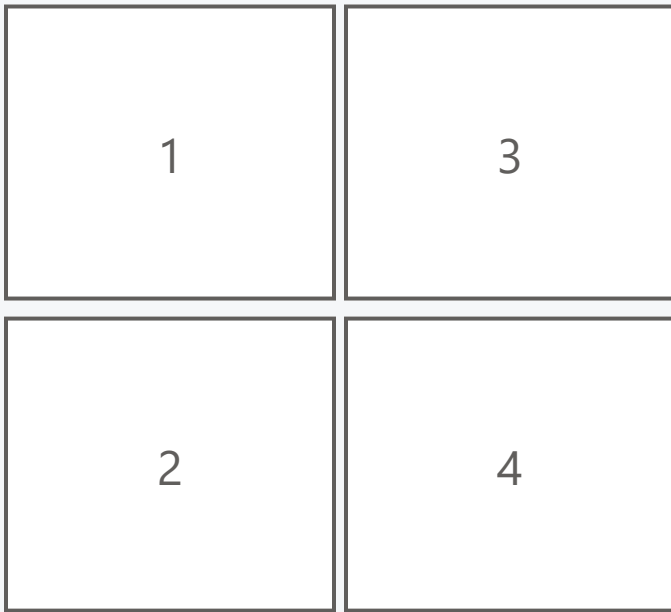
Los resultados se organizaron en función de las estrategias empleadas: 1: sin re-muestreo, submuestreo (2: RUS) y sobremuestreo (3: ROS y 4: SMOTE). Se evaluó el modelo Random Forest mediante métricas clave como **Recall**, **Precision** y **AUPRC** , considerando diferentes tamaños de conjuntos de entrenamiento (**75%**, **80%** y **90%**) y se seleccionó la configuración recomendable de hiperparámetros al maximizar el rendimiento en la detección de la clase minoritaria a través del **Recall** para cada conjunto de entrenamiento y técnica.

train_size	n_estimators	min_samples_split	min_samples_leaf	criterion	max_features	max_depth	RECALL	PRECISION	PRECISION-RECALL
0,90	50	10	1	gini	sqrt	30	0,3736	0,6596	0,5464
0,80	50	5	1	gini	sqrt	30	0,3770	0,6571	0,5416
0,75	50	5	1	gini	sqrt	30	0,3799	0,6531	0,5505
0,90	10	10	1	gini	log2	10	0,6594	0,7137	0,7652
0,75	14	2	2	gini	sqrt	20	0,6614	0,7067	0,7577
0,80	10	10	2	entropy	sqrt	30	0,6680	0,7006	0,7602
0,80	50	2	1	entropy	sqrt	10	0,7792	0,9512	0,8440
0,75	14	2	2	gini	sqrt	20	0,7811	0,9209	0,8210

DATA SET



TÉCNICA



XGBoost

¡La optimización inteligente de los árboles de decisión!

0,17%

Base de
datos 1

Clase minoritaria

22,12%

Base de
datos 2

Clase minoritaria

Parámetros configurados:

"***n_estimators***": [10, 20, 50, 100, 200, 500],

"***max_depth***": [3, 4, 7, 9],

"***learning_rate***": [0.02, 0.15, 0.2, 0.4, 0.5],

"***gamma***": [0, 0.25, 0.5],

"***lambda***": [0, 0.25, 0.5],

"***alpha***": [0, 0.25, 0.5],

"***scale_pos_weight***": [1, "sum_wneg/sum_wpos"],

"***colsample_bytree***": [0.2, 1]



Efecto de los parámetros y el Tamaño del conjunto de entrenamiento en el Recall

Selecciona la base de datos con el control "Base de datos", elige el parámetro de interés desde el menú desplegable "Parámetros XGB" y ajusta la "técnica de re-muestreo" con las opciones disponibles (1: sin técnica de re-muestreo, 2: RUS, 3:ROS, 4: SMOTE) para explorar los resultados correspondientes para cada tamaño de conjunto de entrenamiento.

Base de datos

- 1
- 2

ParámetroXGB

Profundidad máxima

Técnicas de re-muestreo

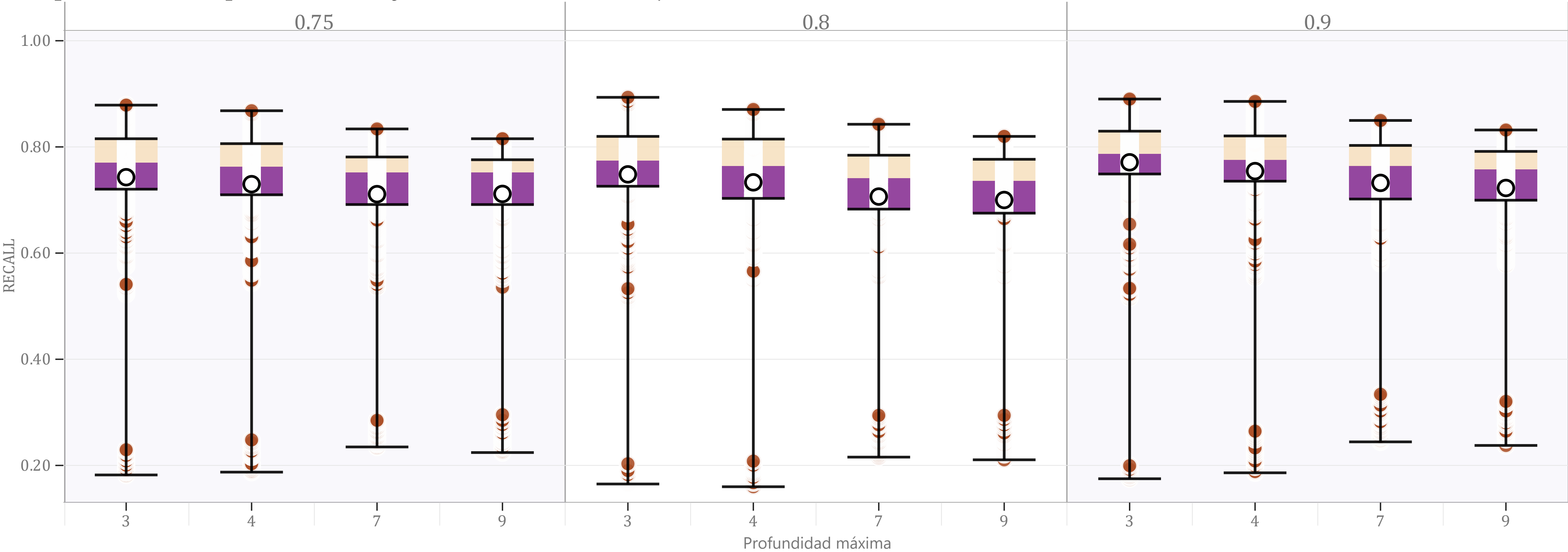
1

3

2

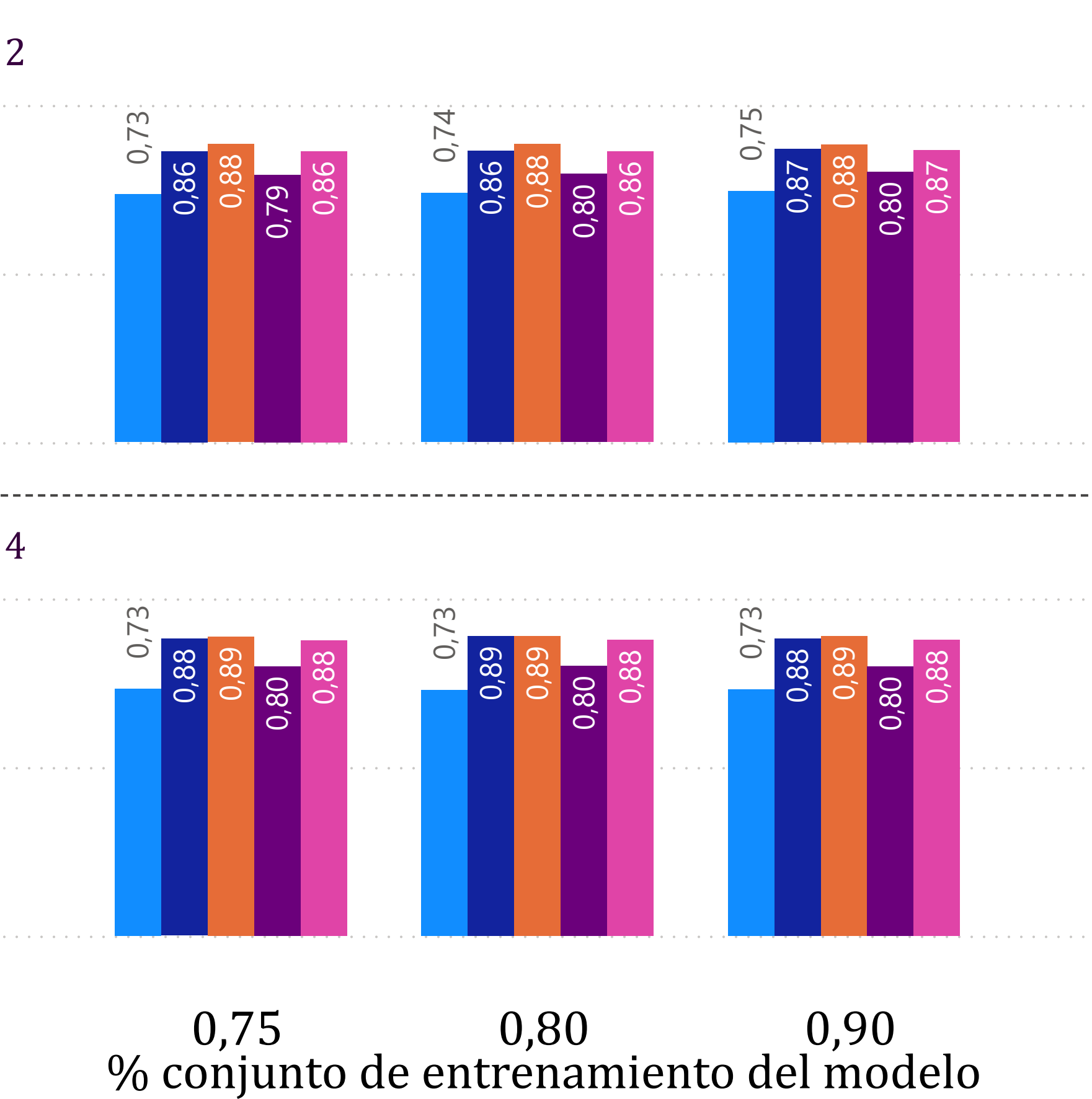
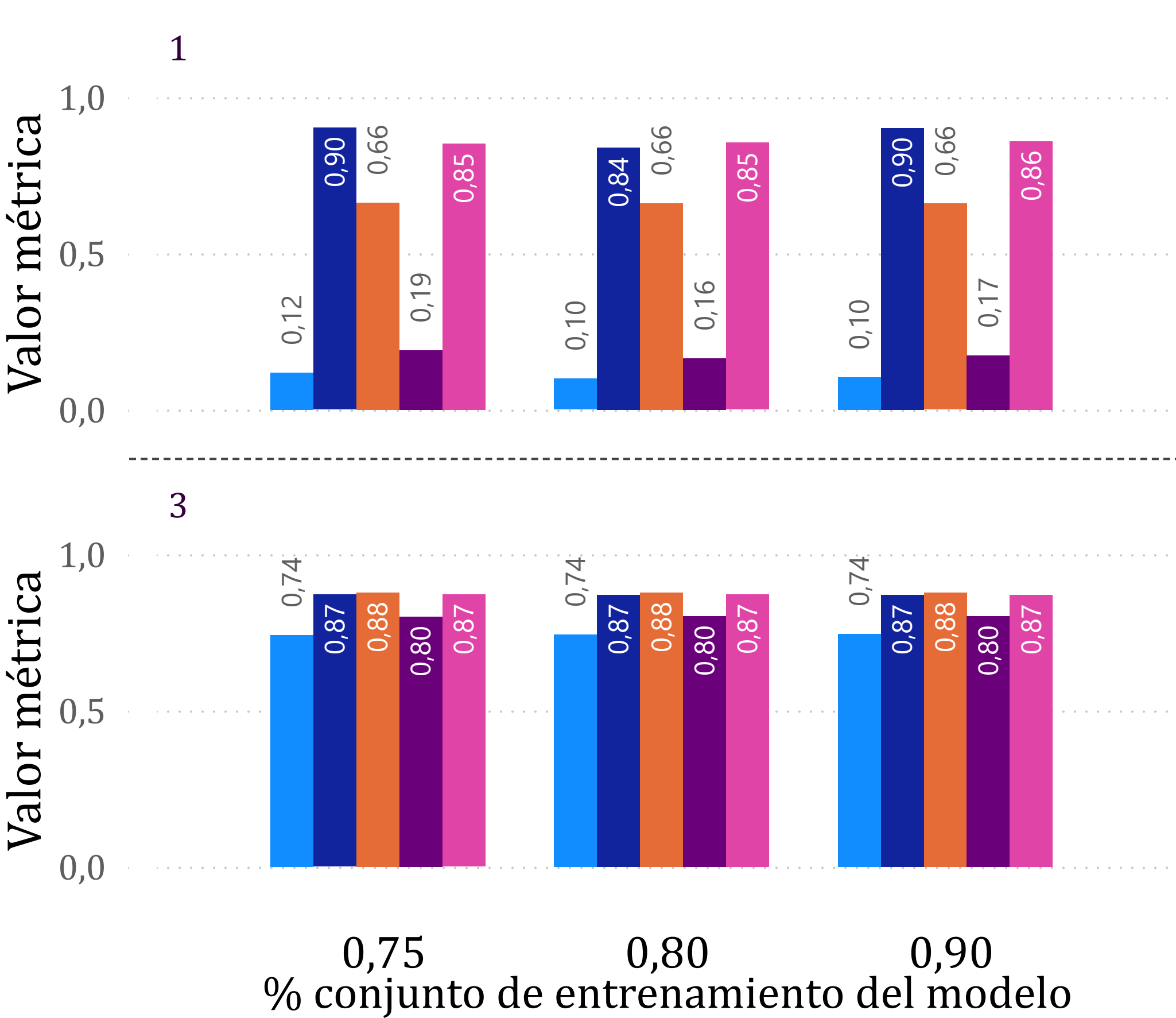
4

Impacto de los parámetros y el tamaño del conjunto de entrenamiento en la métrica Recall del Modelo XGBoost



COMPARACIÓN DE MÉTRICAS SEGÚN EL TAMAÑO DE ENTRENAMIENTO Y TÉCNICA APLICADA PARA EL MODELO XGBOOST

Métrica ● RECALL ● PRECISION ● PRECISION-RECALL ● F1 ● ROC



BASE DE DATOS

1
2

Ajusta los filtros para modificar parámetros como el número de árboles, la profundidad máxima y el tasa de aprendizaje según la base de datos seleccionada, y analiza cómo estas configuraciones impactan las métricas en cada escenario.

Número de árboles

☒ 10

☐ 20

☐ 50

Profundidad máxima

☒ 3

☐ 4

Tasa de aprendizaje

☒ 0,02

☐ 0,15

gamma (reducción mí...

☐ 0,00

☒ 0,25

lambda (regularizació...

☒ 0,00

☐ 0,25

alpha (regularización ...

☒ 0,00

☐ 0,25

Balance de clases desb...

☒ 1,00

☐ 3,49

Características selecci...

☒ 0,20

☐ 1,00

Elección de los mejores modelos a partir de la métrica Recall para el modelo XGBoost.

Los resultados se organizaron en función de las estrategias empleadas: 1: sin re-muestreo, submuestreo (2: RUS) y sobremuestreo (3: ROS y 4: SMOTE). Se evaluó el modelo XGBoost mediante métricas clave como **Recall**, **Precision** y **AUPRC** , considerando diferentes tamaños de conjuntos de entrenamiento (**75%**, **80%** y **90%**) y se seleccionó la configuración recomendable de hiperparámetros al maximizar el rendimiento en la detección de la clase minoritaria a través del **Recall** para cada conjunto de entrenamiento y técnica.

TRAIN SIZE	n_estimators	max_depth	learning_rate	gamma	lambda	alpha	colsample_bytree	scale_pos_weight	RECALL	PRECISION	PREC
0,90	500	9	0,50	0,00	0,00	0,50	1,00	1,00	0,38	0,57	
0,80	500	7	0,40	0,00	0,00	0,25	1,00	1,00	0,39	0,57	
0,75	500	7	0,40	0,00	0,00	0,25	1,00	1,00	0,39	0,56	
0,90	20	3	0,50	0,00	0,50	0,50	1,00	3,52	0,64	0,46	
0,75	20	3	0,50	0,00	0,50	0,25	1,00	3,52	0,64	0,46	
0,80	20	3	0,40	0,50	0,25	0,50	0,20	3,52	0,64	0,46	
0,75	50	7	0,40	0,00	0,50	0,25	1,00	1,00	0,67	0,70	
0,80	500	3	0,40	0,00	0,50	0,00	1,00	1,00	0,67	0,69	

DATA SET

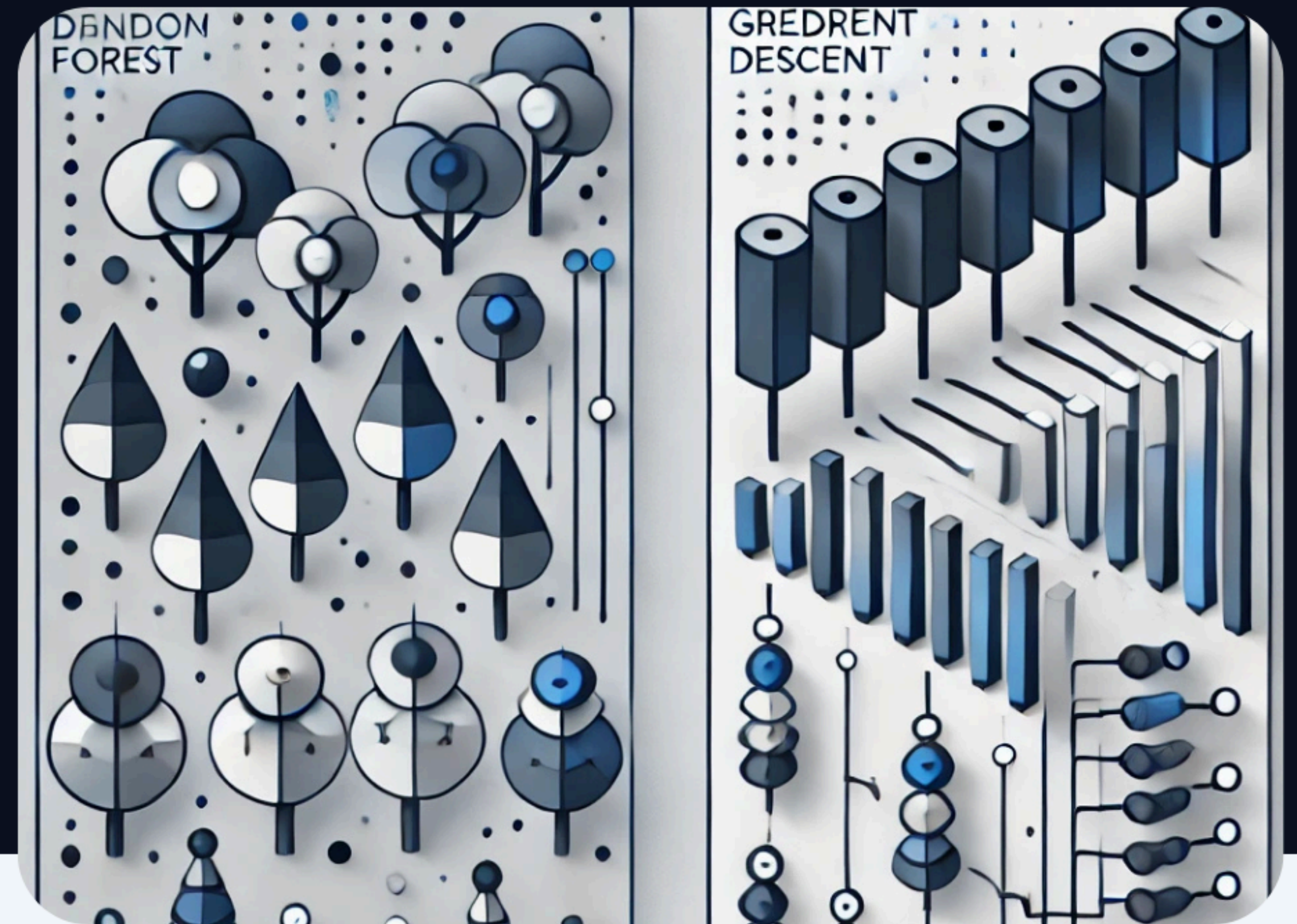
1
2

TÉCNICA

1	3
2	4

COMPARATIVO

RANDOM FOREST **vs** XGBOOST



R
F

- Construcción de Árboles: **PARALELA**
- Optimización: **Promedia árboles independientes**

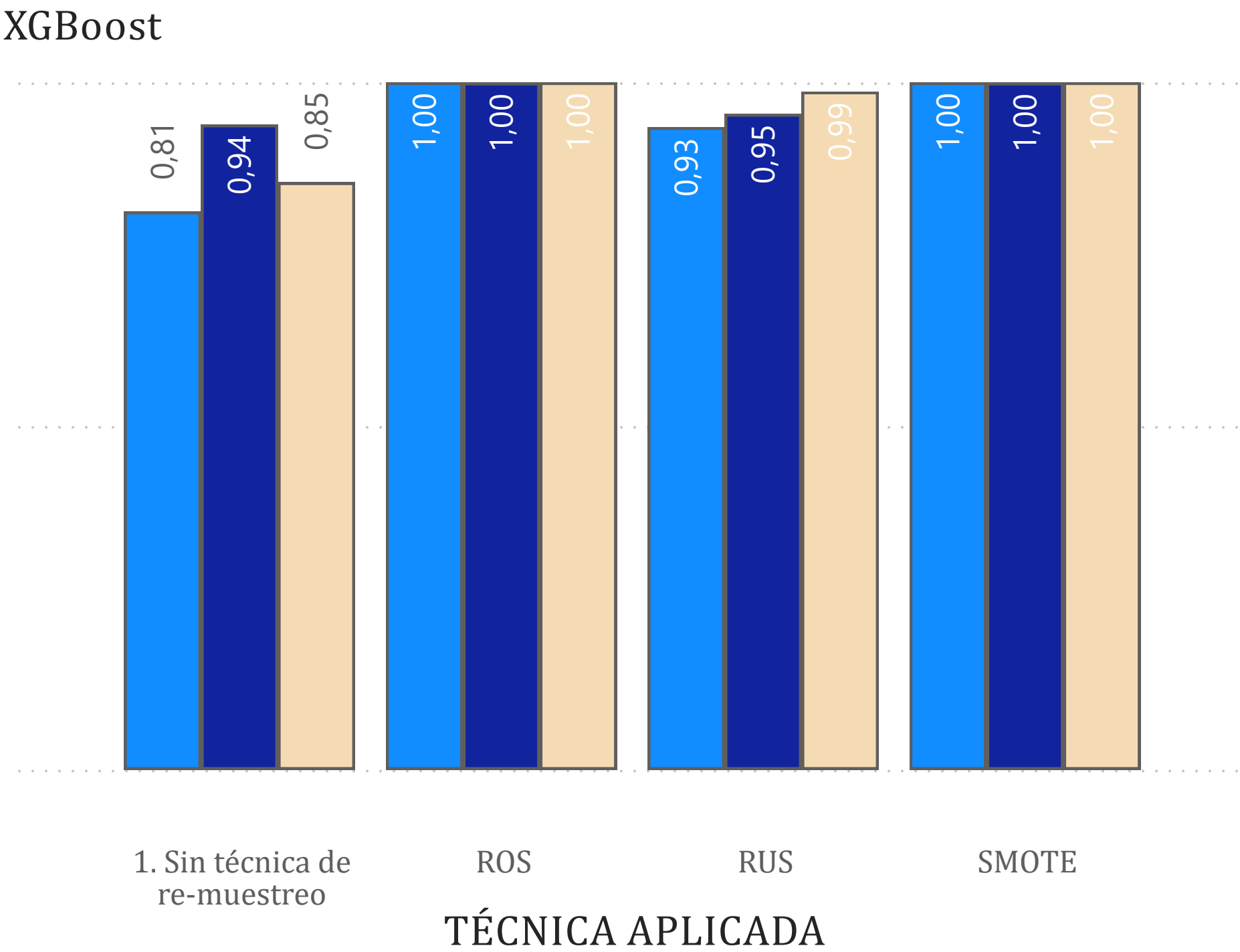
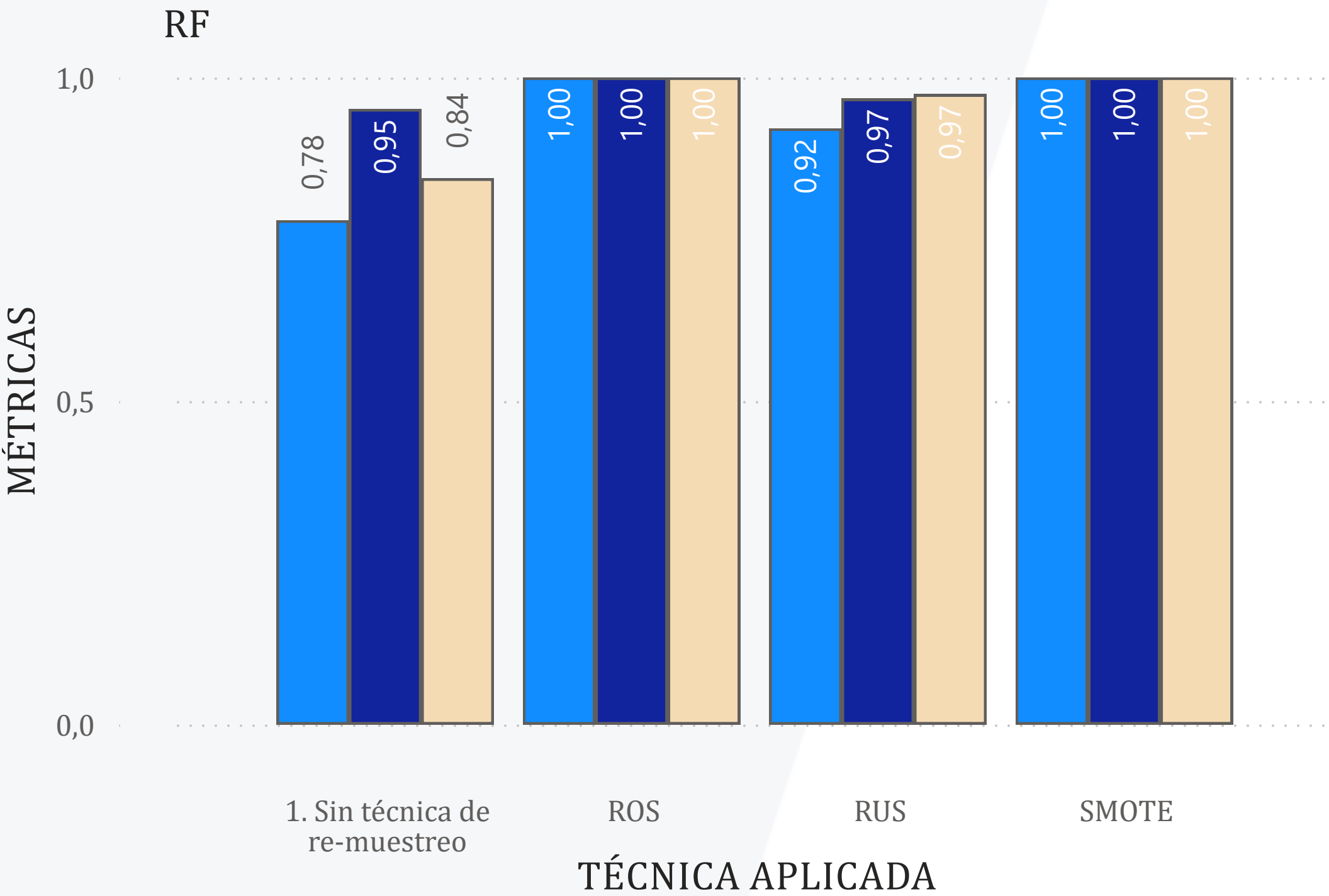
X
G
B

- Construcción de Árboles: **SECUENCIAL**
- Optimización: **Reduce errores gradualmente**



COMPARACIÓN DE MÉTRICAS PARA LOS MODELOS RANDOM FOREST Y XGBOOST

Métrica ●RECALL ●PRECISION ●AUPRC



BASE DE DATOS

credit card fraud
detection

default of credit
card clients

Técnica

- 1. Sin técnica de re-muestreo
- ROS
- RUS
- SMOTE

Modelo	Técnica	train_size	params
RF	ROS	0,80	{'criterion': 'gini', 'max_depth': 20, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}
RF	SMOTE	0,80	{'criterion': 'gini', 'max_depth': 20, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}
RF	RUS	0,90	{'criterion': 'gini', 'max_depth': 10, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 10}
RF	1. Sin técnica de re-muestreo	0,80	{'criterion': 'entropy', 'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50}
XGBoost	1. Sin técnica de re-muestreo	0,90	{'alpha': 0.25, 'booster': 'gbtree', 'colsample_bytree': 1, 'gamma': 0.25, 'lambda': 0.25, 'learning_rate': 0.15, 'max_depth': 3, 'n_estimators': 100, 'scale_pos_weight': 1}
XGBoost	RUS	0,90	{'alpha': 0, 'booster': 'gbtree', 'colsample_bytree': 1, 'gamma': 0, 'lambda': 0.5, 'learning_rate': 0.4, 'max_depth': 4, 'n_estimators': 100, 'scale_pos_weight': 1}
XGBoost	SMOTE	0,80	{'alpha': 0, 'booster': 'gbtree', 'colsample_bytree': 0,2, 'gamma': 0, 'lambda': 0, 'learning_rate': 0.02, 'max_depth': 9, 'n_estimators': 500, 'scale_pos_weight': 1}

CONCLUSIONES

- El estudio recomendó Random Forest y XGBoost combinados con RUS (submuestreo aleatorio) como enfoques efectivos para problemas de clasificación con datos desbalanceados, ajustándose al contexto del problema. Random Forest con RUS destacó en escenarios de desbalance moderado (como BD2), logrando un balance entre Recall y Precisión, reduciendo falsos positivos y siendo ideal para contextos donde la precisión en transacciones legítimas es crítica. Por otro lado, XGBoost con RUS sobresalió en casos de desbalance extremo (como BD1), maximizando el Recall y reduciendo falsos negativos, lo que lo hace idóneo para aplicaciones críticas como detección de fraudes. Estos resultados subrayan la importancia de seleccionar el modelo y la estrategia en función del nivel de desbalance y las características del conjunto de datos.

CONCLUSIONES

- . Técnicas como ROS y SMOTE lograron métricas perfectas, se evidenció que estas introducen sobreajuste significativo, limitando su capacidad de generalización y haciendo que su uso no sea viable en aplicaciones prácticas para estos casos.
- . El análisis comparativo confirmó que no existe una solución universal para problemas de clasificación con datos desbalanceados, pero resaltó la efectividad de combinar modelos ensamblados (Random Forest y XGBoost) con estrategias de re-muestreo adaptadas al contexto del problema, como el submuestreo aleatorio (RUS), para optimizar el balance entre Recall y Precisión según el nivel de desbalance y el tamaño del conjunto de datos.
- . La herramienta de visualización desarrollada no solo facilitó la interpretación de los resultados, sino que también se estableció como un recurso práctico para la toma de decisiones en problemas reales, permitiendo explorar dinámicamente el impacto de distintas estrategias y modelos.