

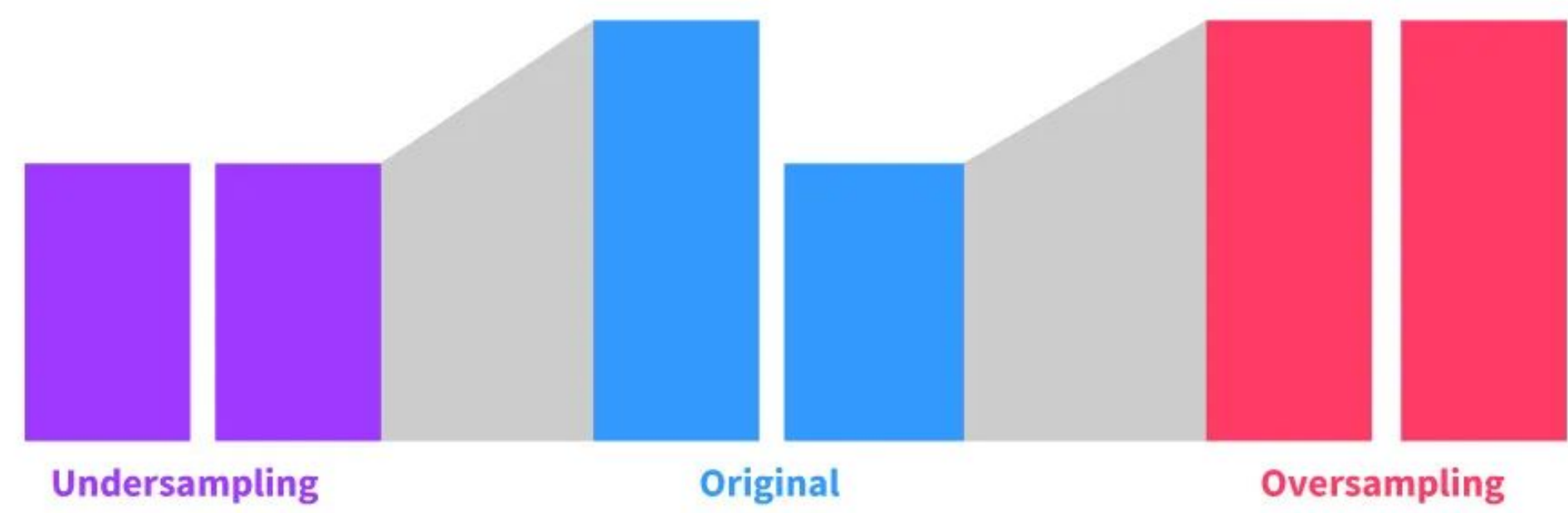
Evaluación de estrategias oversampling vs. subsampling en la detección de intrusos basada en anomalías en hogares inteligentes

Francisco Sánchez, Juan Iturbe (Profesor Guía)
Francisco.sanchez.g@usach.cl, juan.iturbe@usach.cl
Laboratorio / Grupo de investigación - <https://csi.diinf.usach.cl/>



Resumen

El estudio evalúa qué técnica de balanceo de datos (oversampling o subsampling) es más efectiva para mejorar la detección de ciberataques en hogares inteligentes mediante modelos de aprendizaje automático.



Contexto

- IoT en hogares: Automatiza tareas, optimiza recursos, pero genera vulnerabilidades.
- IDS por anomalías: Detecta amenazas nuevas vía comportamientos extraños.
- Impacto del desbalance: Reduce eficacia, provoca errores y falsos positivos.
- Soluciones de balanceo: Oversampling/subsampling mejoran detección y seguridad.

Descripción del problema

- El problema: Desbalance severo (más ataques que normalidad en datos IoT)
- La consecuencia: Métricas infladas crean falsa seguridad; modelos fallan en la práctica.
- El obstáculo: Falta consenso en soluciones y métricas confiables.
- La propuesta: Investigación sistemática para evaluar impacto y mejorar métodos prácticos.

Metodología

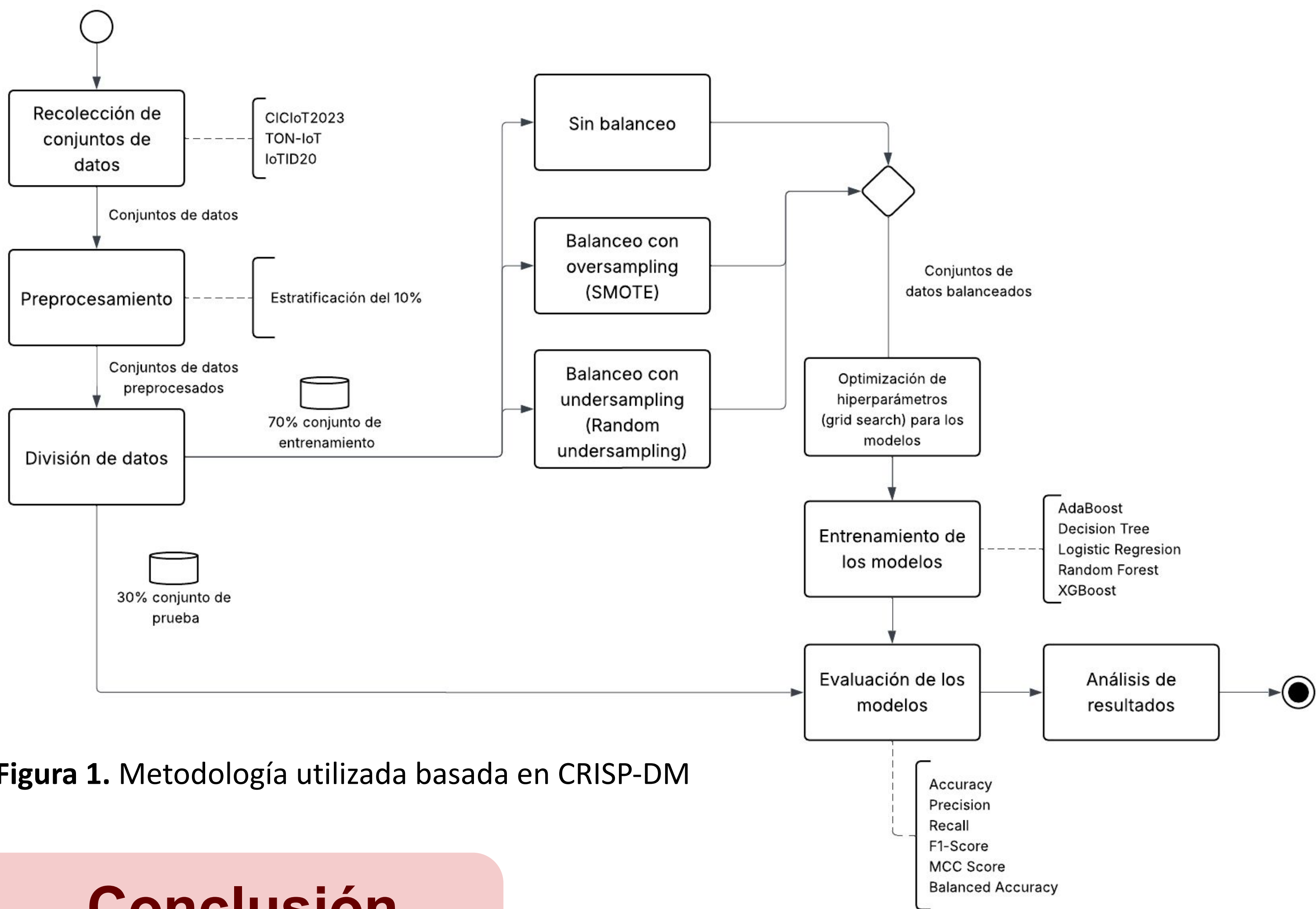


Figura 1. Metodología utilizada basada en CRISP-DM

Resultados

- Sin balanceo domina: 80% de casos superiores (12/15); balanceo poco efectivo en IDS-IoT.
- AdaBoost líder: Mejor sin balanceo en todos los datasets por su robustez con minorías.
- Logistic y DT sólidos: Prefieren no balancear; SMOTE sube falsos negativos y solo empata DT en CICIoT-2023 (costoso).
- RF y XGBoost estables: Mejores sin balanceo, SMOTE ayuda mínimamente en desbalances bajos.

Tabla 1. Comparación de métricas para CICIoT-2023 con diferentes técnicas de balanceo para el modelo Decision Tree

Métrica	Sin Balanceo	SMOTE	Under Sampling
Tiempo (s)	6.29	167.63	166.92
Accuracy	0.9832	0.9835	0.9835
Precision	0.9832	0.9848	0.9848
Recall	0.9832	0.9835	0.9835
F1 Score	0.9832	0.9841	0.9841
MCC Score	0.7760	0.7962	0.7962
Balanced Accuracy	0.9237	0.9529	0.9237

Conclusión

- Sin balanceo domina: Mejor en 12 de 15 escenarios; balanceo (SMOTE, subsampling) rara vez ayuda.
- Modelos destacados: AdaBoost y Logistic Regression brillan sin balanceo; SMOTE sube falsos negativos en Logistic.
- Casos específicos: Decision Tree gana con SMOTE solo en CICIoT-2023 (costoso); RF/XGBoost mejoran levemente en TON-IoT.
- Recomendación: Evaluar por modelo/datos con métricas sólidas, no aplicar balanceo automáticamente.

Trabajos futuros

Unificación de conjuntos de datos IoT

Evaluación en streaming real-time

Análisis de estabilidad temporal

Técnicas híbridas avanzadas

Integración Edge computing

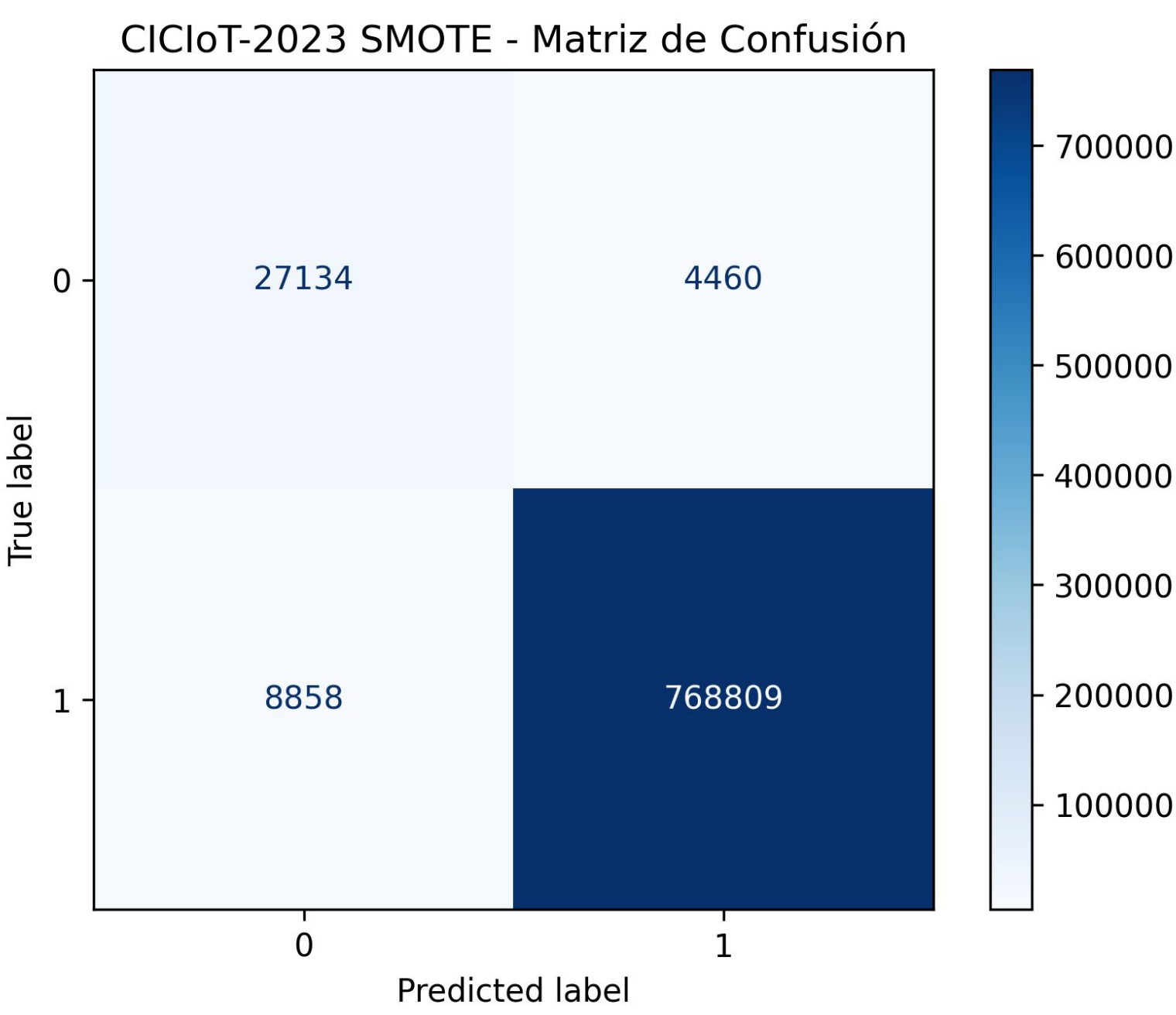


Figura 2. Matriz de confusión del modelo Decision Tree entrenado con el conjunto de datosCICIoT-2023 con balanceo de datos SMOTE