Sistema de Detección de Fraudes con Técnicas de Machine Learning

Integrantes:

Cristóbal Camarena Hernández – A01642653
Victor Jaziel Coronado Flores – A01644090
Rodrigo Yael Morales Luna – A01637721
Omar Michel Carmona Villalobos – A01644146

Inteligencia Artificial Avanzada para la Ciencia de Datos I

Grupo: 101

19 de septiembre de 2025

1. Introducción

La detección automática de fraude financiero mediante algoritmos de aprendizaje supervisado constituye un área crítica en la ciencia de datos aplicada a sistemas bancarios y aseguradoras. Este documento describe de manera técnica el diseño, entrenamiento y validación de un modelo de clasificación binaria enfocado en la identificación de transacciones fraudulentas. El análisis se centra en métricas robustas como AUC-ROC, curva Precision-Recall, estadístico KS y matriz de confusión normalizada.

2. Contexto

Los sistemas de pago digitales y la alta frecuencia de transacciones incrementan la vulnerabilidad al fraude. Dado el bajo porcentaje de fraudes en el dataset (menor de 1 %), el problema se caracteriza como *highly imbalanced classification*. Esta naturaleza obliga a aplicar técnicas especializadas de re-muestreo, ajuste de umbral y ensambles para evitar un modelo sesgado hacia la clase mayoritaria (transacciones legítimas).

3. Descripción del Dataset

El dataset empleado contiene transacciones financieras históricas con las siguientes características principales:

- **Tamaño:** Aproximadamente 1,000,000 de registros y 32 variables.
- **Variable objetivo:** Indicador binario (0 = No Fraude, 1 = Fraude).
- **Variables numéricas:** Incluyen monto de la transacción, tiempo transcurrido desde la última operación, número de transacciones recientes, entre otras.
- **Variables categóricas:** Tipo de comercio, canal de pago, país de origen, dispositivo utilizado.
- **Distribución de clases:** Menos del 1 % de los registros corresponden a fraudes confirmados, lo que confirma el desbalance extremo.
- **Valores faltantes:** Aproximadamente 3.5 % de los datos contenían nulos en variables de ubicación y dispositivo.

Este conjunto de características permitió la extracción de *features* derivados y la aplicación de técnicas de balanceo que mejoraron la capacidad predictiva del modelo.

4. Objetivos

- Diseñar un pipeline de preprocesamiento robusto que incluya normalización, codificación de variables categóricas y tratamiento de valores atípicos.
- Implementar clasificadores base (Random Forest, Gradient Boosting, Regresión Logística) y combinarlos mediante ensamble (Voting Soft y EasyEnsembleClassifier).
- Evaluar el modelo con métricas sensibles al desbalance de clases (KS, PR-AUC, matriz de confusión balanceada).

5. Retos

- **Desbalance de Clases:** Proporción muy baja de fraudes respecto a no fraudes, lo que requiere técnicas como EasyEnsemble y SMOTE.
- **Sobreajuste:** Riesgo de que modelos complejos (e.g., Gradient Boosting) memoricen patrones específicos.
- **Costo de Error Asimétrico:** En este dominio, los falsos negativos tienen un costo mucho mayor que los falsos positivos. Se ajustó el umbral para priorizar recall.

6. Metodología

El flujo metodológico incluyó:

1. **Preprocesamiento de datos:** imputación de nulos, eliminación de duplicados, codificación de categóricas mediante one-hot encoding y normalización de variables numéricas sensibles a magnitudes.
2. **Selección de modelos base:** Random Forest (bagging), Gradient Boosting (boosting secuencial) y Regresión Logística regularizada.
3. **Ensamble:** Voting Soft combina probabilidades de modelos base, mientras que EasyEnsemble genera subconjuntos balanceados para entrenar múltiples clasificadores.
4. **Optimización de umbral:** Se definió recall objetivo $\geq$0.80. El umbral final (0.517) se seleccionó maximizando TPR-FPR (KS test).
5. **Evaluación:** Se calcularon métricas ROC-AUC, PR-AUC, KS, matriz de confusión y curvas de distribución de probabilidades.

7. Resultados

A continuación, se presentan las gráficas más relevantes del desempeño del modelo:

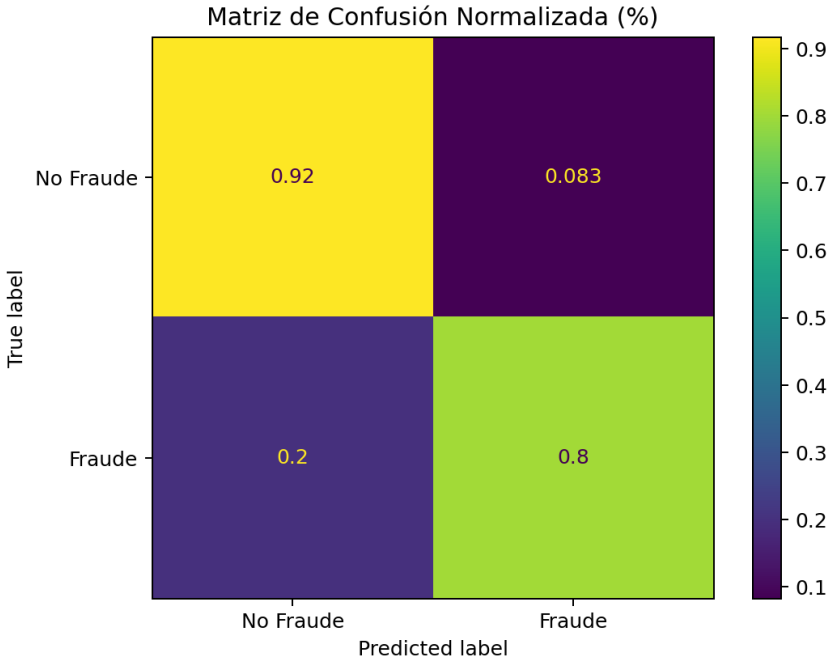

Figura 1: Matriz de Confusión Normalizada: recall=0.80, precisión=0.32.

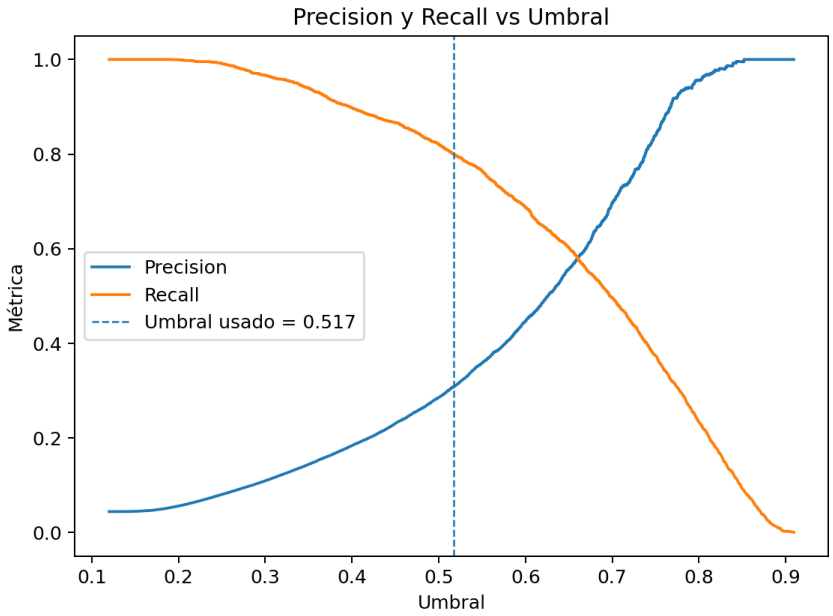

Figura 2: Precisión y Recall vs Umbral: selección óptima en 0.517.

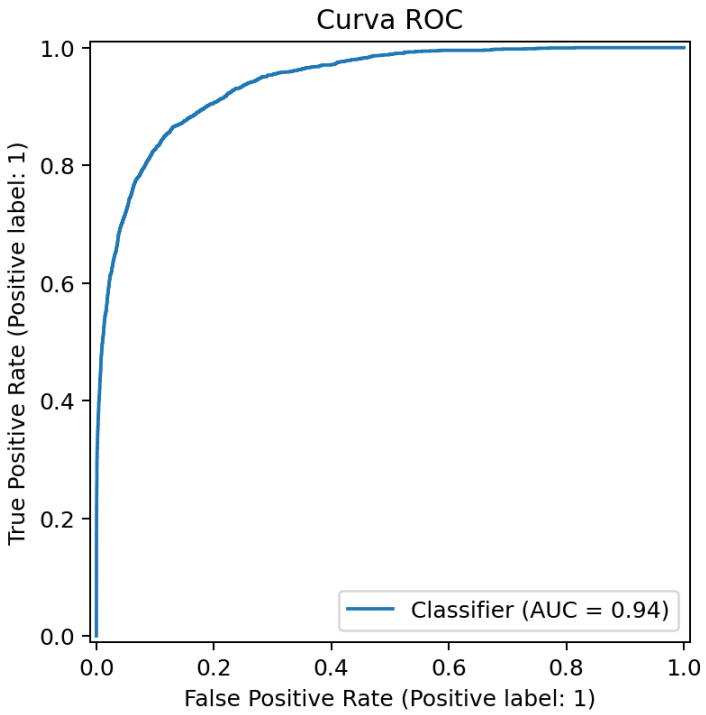

Figura 3: Curva ROC: AUC=0.94, lo que refleja un modelo altamente discriminativo.

8. Análisis de Resultados

El modelo final cumple con el recall objetivo (80 %) a costa de una reducción en precisión (32 %). Este compromiso es aceptable dado el costo elevado de los falsos negativos. El AUC-ROC de 0.94 valida la capacidad del modelo para discriminar entre clases, mientras que el KS=0.735 confirma una segmentación efectiva. La distribución de probabilidades revela un solapamiento moderado en la zona de decisión (0.4-0.6), lo cual explica parte de los falsos positivos.

9. Posibles Mejoras

- Probar algoritmos como XGBoost, LightGBM o CatBoost.
- Experimentar con SMOTE-ENN, ADASYN u otras técnicas híbridas que combinan sobremuestreo y limpieza de ruido.
- Cost-sensitive learning, ajustar la función de pérdida para penalizar más los falsos negativos (fraudes no detectados).
- Usar validación cruzada estratificada en lugar de un único train/test. Evaluar estabilidad en diferentes subconjuntos de datos.

10. Conclusiones

El sistema desarrollado logra recall=0.80 y AUC-ROC=0.94, métricas que validan su efectividad en entornos financieros con alto desbalance de clases. Si bien la precisión es

limitada, este trade-off es estratégico: detectar la mayoría de fraudes es prioritario. El pipeline propuesto es escalable y puede ser extendido con técnicas más sofisticadas de feature engineering y algoritmos de boosting de última generación.