Departamento de Informática
Facultad Politécnica
Universidad Nacional de Asunción

ELECTIVA IV: DATA SCIENCE

Laboratorio 2: Clasificación binaria con datos desbalanceados

Actividad

Resuelve un problema de **clasificación binaria** aplicando el ciclo de modelado supervisado con **desbalance de clases** y validación adecuada.

Fuente de datos: datos_clasificación.csv
Notebook: notebook inicial

1. **Objetivos de aprendizaje**

- Diseñar un flujo reproducible de ML con pandas y scikit-learn.
- Construir **pipelines** con preprocesamiento (escalado/One-Hot).
- Comparar al menos **3 modelos** con validación cruzada estratificada.
- Evaluar con métricas adecuadas a desbalance: *precision*, *recall*, *F1*, *ROC AUC*, matriz de confusión y curvas ROC/PR.
  Explorar **umbrales de decisión** y discutir *trade-offs*.

2. **Datos**

- Archivo: datos_clasificacion.csv (incluye variables numéricas, una categórica derivada y objetivo target con desbalance ~70/30).
- Si el CSV no estuviera disponible, la notebook genera un dataset equivalente automáticamente.

3. **Entregables**

- Notebook ejecutada con código, gráficos y conclusiones intermedias.
- **Informe breve** (5–8 párrafos) con: problema, método, resultados y discusión (métricas y umbral), riesgos de *leakage* y *overfitting*, y trabajo futuro.
- CSV con tabla comparativa de modelos (validación cruzada).

4. **Pasos sugeridos (ya guiados en la notebook)**

- EDA y verificación de desbalance.
- Split estratificado train/test.
- Pipeline con ColumnTransformer + modelo base (Reg. Logística).
- Métricas y visualizaciones (confusión, ROC).
- Comparación de modelos (LogReg, RandomForest/AdaBoost/SVC).
- **GridSearchCV** sobre el mejor candidato.
  Ajuste de **umbral** (curva PR y F1 óptimo).
- Evaluación final en test y conclusiones.

Observación: Propósito del conjunto de datos

El dataset fue generado de forma sintética para fines didácticos, con el objetivo de permitir a los estudiantes practicar tareas de clasificación binaria supervisada bajo condiciones realistas: datos mixtos (numéricos y categóricos), presencia de interacciones no lineales, ruido, y desbalance de clases.

Representa un caso genérico de predicción de un evento binario (por ejemplo, *cliente que abandona o no*, *paciente con o sin recaída*, *transacción fraudulenta o legítima*, etc.).