Proyecto 2 — Predicción de Enfermedad Cardíaca usando Técnicas de Aprendizaje Supervisado

INFO1185: Inteligencia Artificial

Prof. Dr. Ricardo Soto Catalán

10 de octubre de 2025

1. Descripción general

La detección automática de transacciones fraudulentas es un problema real y crítico en el ámbito financiero. Las organizaciones deben identificar actividades sospechosas entre millones de transacciones legítimas, en contextos donde los casos de fraude son extremadamente poco frecuentes.

En este proyecto, la tarea es **entrenar y evaluar diferentes modelos de clasificación supervisada** para detectar transacciones fraudulentas en un conjunto de datos reales.

Este problema presenta desafíos importantes:

- El dataset está **altamente desbalanceado** (menos del 1 % de las transacciones son fraudulentas).
- Las variables han sido transformadas mediante PCA, por lo que no existe un significado semántico directo. No tiene sentido aplicar PCA u otro algoritmo de reducción de dimensión a este conjunto de datos.
- Las métricas estándar como la *accuracy* no son adecuadas en este contexto.

Para abordar este problema, los equipos deberán diseñar un **pipeline completo de modelado**, desde el preprocesamiento hasta la evaluación, justificando las decisiones técnicas en cada etapa.

2. Objetivos

1) Comprender y explorar un conjunto de datos reales y desbalanceados.

2) Implementar estrategias adecuadas de preprocesamiento y preparación de datos.

3) Seleccionar un subconjunto de características predictivas usando una técnica de **selección de características basada en modelos o tests estadísticos** (no reducción tipo PCA).

4) Entrenar, ajustar y comparar el desempeño de cuatro clasificadores supervisados:

   - k-Nearest Neighbors (k-NN)
   - Árbol de Decisión
   - Support Vector Machine (SVM)
   - Random Forest

5) Realizar búsqueda y selección de hiperparámetros para cada modelo, documentando el procedimiento y la configuración óptima obtenida.

6) Evaluar el desempeño de los modelos usando métricas adecuadas para problemas desbalanceados y discutir críticamente los resultados.

3. Conjunto de datos

- **Nombre original:** Credit Card Fraud Detection Dataset
- **Fuente:** Kaggle
- **Número de registros:** 284 807
- **Número de fraudes:** 492 (0,172 %)
- **Atributos:**
  - `Time`, `Amount`
  - `V1` – `V28`: componentes principales generadas por PCA
  - `Class`: variable objetivo (0 = legítima, 1 = fraude)

4. Instrucciones técnicas

1) **Preprocesamiento:**

   - Escalado o normalización de variables numéricas.
   - División en conjuntos de entrenamiento (70 %) y prueba (30 %).

2) **Selección de características:**

   - Aplicar al menos una técnica de selección de características, por ejemplo, Sequential Forward Selection.
   - Reporte el vector de características seleccionado.

3) **Entrenamiento y ajuste de modelos:**

   - Entrenar los cuatro clasificadores especificados, considerando minimizar la tasa de falsos negativos (evitar fraudes clasificados como caso normal).
   - Realizar selección de hiperparámetros mediante validación cruzada (por ejemplo, Grid Search o Random Search).
   - Documentar el espacio de búsqueda y la configuración óptima obtenida para cada modelo.

4) **Evaluación:**

   - Reportar las siguientes métricas:
     - Sensibilidad, Especificidad, Precisión, Exactitud.
   - Discutir por qué métricas como la *exactitud* no son representativas en este contexto.
   - Analizar falsos positivos y falsos negativos.

5) **Comparación crítica:**

   - Comparar resultados entre modelos y justificar cuál es más adecuado para este problema.
   - Discutir ventajas y limitaciones de cada técnica en el contexto del desbalance.

5. Entregables

Actualizaré esta parte del documento el día 13 de octubre, dando mayor detalle del informe y rúbricas correspondientes.

Entregable	**Descripción**	**Fecha**
Informe técnico (PDF)	6–10 páginas. Debe incluir descripción del problema, selección de características, tuning, resultados, análisis crítico y conclusiones.	22 de octubre
Código	Código completo y bien documentado, que permita reproducir todos los resultados.	22 de octubre
Presentación oral	10–12 minutos por equipo + preguntas. Enfocada en decisiones técnicas y hallazgos.	22 y 23 de octubre