# 1ACC0057 - MACHINE LEARNING

## Enunciado del trabajo Parcial

## 2025 – 10

| | |
|---|---|
| **PROFESORES:** | **LUIS CANAVAL SANCHEZ**<br>**DIEGO ROJAS SIHUAY**<br>**JAIRO PINEDO TAQUIA** |
| **SECCIÓN:** | **TODAS** |
| **FECHA DE EVALUACIÓN:** | **SEMANA 7** |
| **CICLO ACADÉMICOS:** | **2025 - I** |

1. **Título del Proyecto:** Detección de Fraude en Transacciones de Tarjeta de Crédito.

2. **Descripción:**

En la era digital, el fraude en transacciones de tarjeta de crédito representa un desafío significativo para las instituciones financieras. Este proyecto tiene como objetivo desarrollar un sistema de detección de fraude utilizando técnicas de Machine Learning. Los alumnos trabajarán grupalmente con un conjunto de datos que contiene información sobre transacciones de tarjetas de crédito, algunas de las cuales son fraudulentas. El objetivo es construir modelos capaces de clasificar las transacciones como legítimas o fraudulentas, evaluando su rendimiento mediante diversas métricas y analizando la curva ROC.

3. **Objetivos de Aprendizaje:**

- Aplicar técnicas de normalización de datos para mejorar el rendimiento del modelo.
- Implementar y comparar clasificadores Bayesianos (Naive Bayes) para la detección de fraude.
- Diseñar y entrenar un Perceptrón Multicapa (MLP) para la clasificación de transacciones.
- Evaluar el rendimiento de los modelos utilizando diversas métricas de evaluación (Precisión, Recall, F1-score, etc.).
- Analizar y comparar el rendimiento de los modelos mediante la curva ROC.

4. **Conjunto de Datos:**

Se proporcionará un conjunto de datos que contiene las siguientes características:

- TransactionID: Identificador único de la transacción.
- Time: Marca de tiempo de la transacción.

- Amount: Monto de la transacción.
- CardholderName: Nombre del titular de la tarjeta.
- MerchantName: Nombre del comercio.
- MCC: Código de categoría del comercio.
- Location: Ubicación de la transacción.
- Device: Dispositivo utilizado para la transacción.
- IPAddress: Dirección IP de la transacción.
- TransactionSpeed: Velocidad de la transacción.
- Fraud: Variable objetivo binaria (1: Fraude, 0: No Fraude).

**5. Tareas:**

⇒ **Análisis Exploratorio de Datos (AED):**

- Realizar un análisis descriptivo del conjunto de datos, incluyendo estadísticas descriptivas y visualización de la distribución de las variables.
- Identificar y tratar valores atípicos y faltantes, justificando las decisiones tomadas.
- Analizar la distribución de la variable objetivo (Fraud) y determinar si existe desbalance de clases.

⇒ **Preprocesamiento de Datos:**

- Seleccionar las características relevantes para el modelado.
- Codificar variables categóricas utilizando técnicas adecuadas (one-hot encoding, label encoding, etc.).
- Normalizar las variables numéricas utilizando técnicas como MinMaxScaler o StandardScaler.
- Dividir el conjunto de datos en conjuntos de entrenamiento, validación y prueba (70-15-15 o similar).
- Manejar el desbalance de clases si es necesario (sobremuestreo, submuestreo, pesos de clase).

⇒ **Modelado:**

- Implementar un clasificador Naive Bayes Gaussiano.
- Diseñar y entrenar un Perceptrón Multicapa (MLP) con al menos una capa oculta.
- Ajustar los hiperparámetros de los modelos para optimizar su rendimiento.

⇒ **Evaluación:**

- Utilizar el conjunto de prueba para evaluar el rendimiento de los modelos.
- Calcular y presentar las siguientes métricas de evaluación para cada modelo:
  - Precisión (Accuracy)
  - Recall (para la clase positiva: Fraude)
  - Precisión (para la clase positiva: Fraude)
  - F1-score (para la clase positiva: Fraude)
  - Matriz de Confusión
- Generar y comparar las curvas ROC de los modelos.

⇒ **Análisis y Conclusiones:**

- Comparar el rendimiento de los modelos en función de las métricas y la curva ROC.
- Analizar las fortalezas y debilidades de cada enfoque para la detección de fraude.
- Identificar las características más importantes para la predicción del fraude según el MLP.
- Proponer recomendaciones para mejorar el sistema de detección de fraude.

**6. Recursos:**

- Se proporcionará el conjunto de datos.
- Se recomienda utilizar bibliotecas de Python como pandas, scikit-learn, matplotlib y seaborn.

**7. Entregables (FECHA: 11/05/2025 23:59):**

- **Informe detallado:** Un documento que describa el proceso completo del proyecto, incluyendo:
  - El análisis exploratorio de datos realizado.
  - Las decisiones tomadas durante el preprocesamiento de los datos y su justificación.
  - La implementación de los modelos de Machine Learning.
  - Los resultados obtenidos en la evaluación de los modelos.
  - Un análisis comparativo del rendimiento de los modelos.
  - Las conclusiones del proyecto y las recomendaciones propuestas.

- **Código fuente:** El código desarrollado para el proyecto, que debe estar:

    - Bien comentado, explicando cada sección del código.
    - Organizado, siguiendo una estructura lógica y modular.

- **Presentación:** Una presentación que resuma los hallazgos más importantes del proyecto, incluyendo:

    - El objetivo del proyecto.
    - La metodología utilizada.
    - Los resultados principales.
    - Las conclusiones y recomendaciones.

# ANEXOS:

**RUBRICA**

| Criterio | Excelente (5) | Bueno (3.5) | Suficiente (2) | Insuficiente (1) |
|---|---|---|---|---|
| **Informe Detallado** | El informe es claro, conciso, completo y bien organizado. Demuestra una comprensión profunda del problema y un análisis exhaustivo de los datos. Las decisiones de preprocesamiento están claramente justificadas y la metodología es rigurosa. Los resultados se presentan de manera clara y precisa, y las conclusiones son lógicas y están respaldadas por la evidencia. Las recomendaciones son relevantes y prácticas. | El informe es generalmente claro y completo, pero podría tener algunas áreas de mejora en la organización o la claridad. Demuestra una buena comprensión del problema y un análisis adecuado de los datos. La mayoría de las decisiones de preprocesamiento están justificadas y la metodología es adecuada. Los resultados se presentan de manera comprensible y las conclusiones son razonables. Las recomendaciones son relevantes. | El informe es comprensible pero le falta detalle o claridad en algunas secciones. Demuestra una comprensión básica del problema y un análisis superficial de los datos. Algunas decisiones de preprocesamiento no están justificadas o la metodología es incompleta. Los resultados se presentan de manera confusa o incompleta, y las conclusiones son vagas o no están bien respaldadas. Las recomendaciones son genéricas o poco prácticas. | El informe es confuso, incompleto o mal organizado. Demuestra una falta de comprensión del problema o un análisis inadecuado de los datos. Las decisiones de preprocesamiento no están justificadas o la metodología es incorrecta. Los resultados son incomprensibles o faltan, y las conclusiones no son lógicas o no están presentes. No se presentan recomendaciones o son irrelevantes. |

| | | | | |
|---|---|---|---|---|
| **Código Fuente** | El código es limpio, eficiente, bien comentado y sigue las mejores prácticas de programación. Está organizado en módulos lógicos y es fácil de entender y mantener. La implementación de los modelos de Machine Learning es correcta y eficiente. | El código es funcional y comprensible, pero podría tener algunas áreas de mejora en la limpieza, eficiencia o la calidad de los comentarios. La organización en módulos es adecuada y la implementación de los modelos de Machine Learning es mayormente correcta. | El código es funcional pero desordenado, ineficiente o mal comentado. La organización en módulos es deficiente y la implementación de los modelos de Machine Learning tiene errores menores. | El código es incompleto, no funcional, incomprensible o carece de comentarios. La organización es inexistente y la implementación de los modelos de Machine Learning es incorrecta o faltante. |
| **Presentación** | La presentación es clara, concisa, atractiva y bien organizada. Resume de manera efectiva los hallazgos más importantes del proyecto y utiliza recursos visuales adecuados. El presentador demuestra un dominio del tema y responde a las preguntas de manera clara y precisa. | La presentación es generalmente clara y bien organizada, pero podría ser menos atractiva o concisa en algunas áreas. Resume los hallazgos principales del proyecto y utiliza recursos visuales adecuados. El presentador demuestra un buen conocimiento del tema y responde a la mayoría de las preguntas de manera competente. | La presentación es comprensible pero le falta claridad, concisión o atractivo en algunas secciones. Resume algunos de los hallazgos del proyecto y utiliza recursos visuales limitados o inadecuados. El presentador demuestra un conocimiento básico del tema pero tiene dificultades para responder a algunas preguntas. | La presentación es confusa, desorganizada, aburrida o incompleta. No resume los hallazgos del proyecto o utiliza recursos visuales inapropiados. El presentador no demuestra un conocimiento adecuado del tema o no puede responder a las preguntas. |

| | | | | |
|---|---|---|---|---|
| Análisis del Modelo | El análisis del modelo es exhaustivo y profundo. Compara críticamente el rendimiento de los diferentes modelos, justifica la elección del mejor modelo en función de las métricas de evaluación y la curva ROC, y analiza las fortalezas y debilidades de cada enfoque en el contexto del problema de detección de fraude. | El análisis del modelo es completo y adecuado. Compara el rendimiento de los diferentes modelos, justifica la elección del mejor modelo y analiza las fortalezas y debilidades de cada enfoque. | El análisis del modelo es básico pero incompleto. Compara superficialmente el rendimiento de los modelos y ofrece una justificación limitada para la elección del mejor modelo. El análisis de las fortalezas y debilidades es vago o ausente. | El análisis del modelo es deficiente o ausente. No compara el rendimiento de los modelos o la justificación de la elección del mejor modelo es incorrecta o no está presente. No se analizan las fortalezas y debilidades de los enfoques. |