Proyecto Final (Avanzado): Simulación de un Pipeline de ML de Producción en un Entorno de Notebook

Apertura: martes, 8 de julio de 2025, 00:00
Cierre: domingo, 3 de agosto de 2025, 23:59

Proyecto Final (Avanzado): Simulación de un Pipeline de ML de Producción en un Entorno de Notebook

Curso: Paradigmas de Programación para Inteligencia Artificial y Análisis de Datos
Programa: Maestría en Inteligencia Artificial y Ciencia de Datos
Entrega: Individual
Puntaje Total: 10 puntos

Objetivo General:

Simular un flujo de trabajo de Machine Learning de extremo a extremo (end-to-end), desde el preprocesamiento de datos complejos hasta la evaluación de un modelo y el empaquetado de la lógica para una "predicción como servicio". El proyecto se enfoca en aplicar principios de ingeniería de software (modularidad, configuración, testing) dentro de un notebook de Colab, demostrando una comprensión profunda de cómo se estructuran los proyectos de ML en el mundo real.

Instrucciones Generales:

1.
Selección del Dataset (Enfoque en Complejidad):

- Elige un dataset de **Kaggle** que presente desafíos de preprocesamiento complejos.
- **Requisitos del Dataset:**
  - Debe contener una mezcla de datos numéricos, categóricos, y preferiblemente **texto no estructurado** (ej. descripciones, reseñas) o una alta cardinalidad en las variables categóricas.
  - Debe tener problemas realistas como **clases desbalanceadas** o la necesidad de un **feature engineering significativo**.
- **Sugerencias:** "Toxic Comment Classification", "IMDb Movie Reviews", "Customer Churn con variables de texto", "Credit Card Fraud Detection".
- **IMPORTANTE:** La justificación de por qué el dataset es un buen candidato para un pipeline complejo es una parte crucial de la evaluación.

2.
Entregables:

- Un **único Notebook de Jupyter/Colab (.ipynb)**.
- El notebook debe estar estructurado lógicamente para simular un proyecto modular. Utilizarás celdas de Markdown para delinear las diferentes "partes" del proyecto (configuración, procesamiento, entrenamiento, etc.) y celdas de código que utilizan %%writefile para crear archivos .py virtuales.

Desglose del Proyecto y Tareas:

El notebook debe seguir esta estructura, demostrando una clara separación de responsabilidades.

Parte 1: Configuración y Diseño del Pipeline (2 puntos)

- **1.1. Simulación de un Archivo de Configuración:**
  - En una celda, usa %%writefile config.py para crear un archivo de configuración.
  - Este archivo debe contener todas las variables importantes: la URL del dataset, el nombre de la variable objetivo, listas de columnas por tipo (numéricas, categóricas, texto), parámetros para train_test_split y los hiperparámetros del modelo elegido. **No debe haber "números mágicos" en el resto del código.**
- **1.2. Diseño y Preprocesamiento Avanzado:**
  - En otra celda, usa %%writefile processing.py para crear un módulo de preprocesamiento.
  - Dentro de este archivo, implementa una función o clase que construya un Pipeline de Scikit-learn complejo usando