Pautas y Rúbrica del Proyecto Final

Curso de Machine Learning

Prof. Cristian López Del Alamo

29 de octubre de 2025

1. Objetivo General

El objetivo de este proyecto final es que los estudiantes demuestren su dominio de las técnicas de Machine Learning aprendidas durante el curso, aplicándolas de manera integral a un problema de su elección. El proyecto debe cubrir el ciclo de vida completo de un proyecto de ciencia de datos, desde la definición del problema y la recolección de datos hasta el análisis de resultados y las conclusiones.

Se evaluará con especial atención el rigor metodológico, la originalidad del problema, la profundidad del análisis y la capacidad de comparar y justificar las decisiones tomadas.

2. Tema del Proyecto (Elección Libre)

El tema del proyecto es de elección **libre** por parte del grupo. El problema a resolver puede enmarcarse en una de las siguientes categorías:

- Clasificación (binaria o multiclase)
- Regresión
- Clustering (Segmentación)

3. Requisitos de Datos y Novedad del Problema

Esta es una sección crítica del proyecto. El objetivo es trabajar en problemas que no tengan una solución obvia o fácilmente accesible en la red.

- **Novedad Obligatoria:** Se debe seleccionar un problema y un conjunto de datos que sean **novedosos**.
- **Restricción 1:** No se aceptarán *datasets* çlásicos.º "de juguete"(ej. Iris, Titanic, MNIST, Boston Housing, etc.) ni problemas cuya solución completa se encuentre fácilmente en tutoriales.
- **Restricción 2 (Importante):** Queda estrictamente **prohibido** el uso de data sintética generada en su totalidad por modelos LLM (Large Language Models) para la resolución del problema. Los datos deben provenir de una fuente real.

4. Requisitos de Modelado y Metodología

Cada proyecto debe implementar y comparar rigurosamente un mínimo de **cuatro (4) modelos** distintos:

1. **Un (1) Modelo Base Simple:**
   - Debe ser un modelo simple que sirva como *baseline* o línea base de comparación.
   - *Ejemplos: Regresión Logística, Regresión Lineal (con regularización), K-Means.*
2. **Dos (2) Modelos Adicionales (Vistos en clase):**
   - Deben ser modelos más complejos que el base, seleccionados del material cubierto en el curso.
   - *Ejemplos: Random Forest, Support Vector Machines (SVM), Random Forest, DBSCAN, MLP, CNN, LSTM*
3. **Un (1) Método Moderno (No visto en clase):**
   - El grupo deberá investigar e implementar un método que **no** haya sido cubierto en profundidad durante el sílabo.

5. Evaluación y Métrica

- La selección de las **métricas de evaluación** es fundamental y debe ser justificada.
- Se deben utilizar las métricas correctas según el problema (ej. **AUC-PR** y F1-Score para clasificación desbalanceada, **RMSE/MAE/R^2** para regresión, **Silhouette Score** para clustering).
- La *Accuracy* no será considerada una métrica válida para problemas desbalanceados.

6. Estructura del Informe y Presentación

Se pondrá mucha atención a la forma en que se estructura el análisis. El informe final (o notebook) debe seguir esta estructura:

1. Presentación del Problema
2. Análisis Exploratorio de Datos (EDA)
3. Preprocesamiento y Feature Engineering
4. Metodología y Modelos Seleccionados
5. Análisis de Hiperparámetros (con Validación Cruzada)
6. Presentación de Resultados (de forma compacta y eficiente)
7. Conclusiones y Trabajos Futuros

Rúbrica de Evaluación del Proyecto Final (Base 20 Puntos)

Criterio de Evaluación	Excelente	Bueno	Regular	Deficiente
1. Novedad del Problema y Definición (5 Puntos - 25 %)	**(5 pts)** El problema es **novedoso**, relevante y no trivial. Los datos son reales (no "de juguete"ni 100 % sintéticos de LLM). El objetivo es claro y la métrica de éxito está perfectamente definida.	**(3-4 pts)** El problema es estándar (ej. dataset conocido de Kaggle) pero el objetivo está bien definido. O, el problema es novedoso pero la definición del objetivo es ambigua.	**(1-2 pts)** El problema es estándar y la definición del objetivo es débil. Los datos son problemáticos, su origen no está claro o son "de juguete".	**(0 pts)** El proyecto viola los requisitos: usa un dataset çlásico"(Iris, Titanic) o data sintética de LLM. El problema es trivial o no está definido.
2. Metodología: EDA y Preprocesamiento (5 Puntos - 25 %)	**(5 pts)** El EDA es profundo, visualmente claro y genera hipótesis. El pipeline de preprocesamiento (imputación, *encoding*, escalado) es robusto, está claramente justificado y es reproducible.	**(3-4 pts)** El EDA es correcto pero superficial (gráficos básicos). El preprocesamiento se aplica mecánicamente, sin una justificación clara de por qué se eligió cada técnica.	**(1-2 pts)** El EDA es muy limitado. El preprocesamiento se aplica, pero con errores u omisiones (ej. no escalar datos para un modelo que lo requiere).	**(0 pts)** No hay EDA. Existen errores conceptuales graves en el preprocesamiento (ej. *data leakage*) que invalidan los resultados.
3. Metodología: Modelado y Evaluación (6 Puntos - 30 %)	**(5-6 pts)** Implementa los 4 tipos de modelos (base, 2 adicionales, 1 moderno). El **método moderno** está bien investigado y justificado. Aplica optimización rigurosa con **Validación Cruzada**. Selecciona y justifica las **métricas correctas**.	**(3-4 pts)** Implementa 3 de los 4 modelos (ej. falta el *baseline*). La optimización es superficial o no usa CV. Usa métricas aceptables pero la justificación es débil.	**(1-2 pts)** Implementa 1-2 modelos sin comparación. No hay optimización. El método moderno no se entiende, está mal implementado o no se justifica.	**(0 pts)** No hay comparación. Utiliza métricas demostrablemente **incorrectas** para el problema (ej. *Accuracy* en un dataset desbalanceado).
4. Documentación: Informe y Conclusiones (4 Puntos - 20 %)	**(4 pts)** El informe/notebook es profesional, claro y reproducible. Los resultados se presentan de forma **compacta y eficiente** (tablas, gráficos). Las conclusiones se derivan lógicamente del análisis y se proponen trabajos futuros.	**(3 pts)** El informe es completo pero desordenado (código, texto y gráficos mezclados). Los resultados son correctos pero difíciles de encontrar. Las conclusiones son débiles o genéricas.	**(1-2 pts)** El informe está incompleto. La presentación de resultados es confusa (ej. solo *logs*). Las conclusiones no están respaldadas por los datos o no existen.	**(0 pts)** El informe es solo un volcado de código ilegible. No hay presentación de resultados ni conclusiones. El trabajo no es reproducible.