

Aprendizaje máquina - Proyecto final

1. Descripción General

El presente proyecto final tiene como objetivo principal la aplicación práctica e integral de los conceptos y técnicas de Aprendizaje Máquina abordados durante la asignatura. Se busca que el alumno demuestre su capacidad para abordar un problema de principio a fin, siguiendo el ciclo de vida estándar de un proyecto de ciencia de datos: desde la selección y comprensión de un conjunto de datos hasta la evaluación y comunicación de los resultados de un modelo predictivo.

Este trabajo representa la oportunidad de consolidar los conocimientos teóricos en un desafío práctico, desarrollando tanto habilidades técnicas como analíticas.

2. Objetivos de Aprendizaje

- Aplicar el flujo de trabajo completo de un proyecto de Machine Learning.
- Realizar un análisis exploratorio de datos (EDA) para extraer insights relevantes.
- Implementar técnicas de preprocesamiento y preparación de datos de manera justificada.
- Entrenar, optimizar y evaluar modelos de aprendizaje supervisado.
- Desarrollar la capacidad de comunicar de forma clara y profesional los métodos y resultados obtenidos.

3. Directrices y Requisitos del Proyecto

3.1. Formación de Grupos

El proyecto se podrá realizar de forma individual o en grupos de hasta un máximo de cuatro (4) integrantes.

3.2. Selección del Dataset

La elección del conjunto de datos es libre. Se recomienda encarecidamente seleccionar un problema de aprendizaje supervisado (clasificación o regresión), ya que se alinea mejor con los contenidos de la asignatura.

Sugerencia: Pueden explorar repositorios como Kaggle, UCI Machine Learning Repository, Google Dataset Search o datos abiertos de gobiernos para encontrar un dataset de su interés.

3.3. Estructura Obligatoria del Proyecto

El trabajo deberá estar estructurado en las siguientes secciones, que deben estar claramente diferenciadas y explicadas:

1. **Análisis Exploratorio de Datos (EDA):** Comprensión inicial del dataset, estadísticas descriptivas, identificación de variables, visualizaciones para entender distribuciones, correlaciones y posibles problemas como valores nulos o atípicos.
2. **Preprocesamiento y Preparación de Datos:** Descripción de las transformaciones aplicadas a los datos para prepararlos para el modelado. Esto incluye, pero no se limita a: limpieza de datos, imputación de valores faltantes, codificación de variables categóricas, escalado de características, etc.
3. **Modelización:** Construcción y entrenamiento de los modelos de aprendizaje automático.
4. **Evaluación de Resultados:** Análisis cuantitativo y cualitativo del rendimiento de los modelos utilizando métricas apropiadas. Se deben interpretar los resultados en el contexto del problema original.

3.4. Enfoque del Análisis

Se proponen dos posibles enfoques para la fase de modelización. El grupo deberá elegir uno:

- **Opción A: Análisis Profundo.** Centrarse en un único tipo de algoritmo (o una familia, ej. modelos basados en árboles) y realizar un análisis exhaustivo. Esto implica una búsqueda y optimización avanzada de hiperparámetros (ej. usando Grid Search, Randomized Search o Búsqueda Bayesiana) y un análisis detallado de su comportamiento.
- **Opción B: Análisis Comparativo.** Implementar y comparar el rendimiento de múltiples algoritmos de diferente naturaleza (ej. Regresión Logística, SVM, Árboles de Decisión, Redes Neuronales simples, etc.) para determinar cuál se adapta mejor al problema, justificando la elección final.

3.5. La Clave del Éxito: Explicación y Justificación

Este es el requisito más importante. No basta con presentar bloques de código y sus salidas. Cada paso del proyecto debe ir acompañado de una explicación detallada en texto. Deben justificar sus decisiones: ¿Por qué eligieron esa técnica de imputación? ¿Qué significa la correlación encontrada? ¿Qué métrica de evaluación es más relevante para este problema y por qué? ¿Qué conclusiones se extraen de los resultados?

Un entregable que contenga únicamente código no será considerado suficiente para superar la asignatura.

4. Formato y Entregable

El proyecto se puede entregar en uno de los siguientes dos formatos:

1. **Cuaderno Computacional:** Un único fichero (.ipynb) de Jupyter Notebook o Google Colab, bien estructurado, con celdas de texto (Markdown) para las explicaciones y celdas de código para la implementación. Debe ser ejecutable y mostrar los resultados finales.
2. **Repositorio de GitHub:** Un repositorio público que contenga el código, los datos (o un enlace a ellos) y un fichero README.md detallado que sirva como el informe principal del proyecto, explicando cada sección de manera clara y profesional.

3. En ambos casos, se valorará positivamente la claridad, la limpieza del código, la calidad de las visualizaciones y el profesionalismo general del entregable.

5. Hitos y Fechas de Entrega

- Hito 1 - Formación de Grupos: Se deberá notificar al profesor la composición de los grupos de trabajo por correo electrónico antes del inicio de la sesión número 6.
- Seguimiento: Se valorará positivamente que los grupos compartan avances parciales o dudas específicas durante las sesiones para recibir feedback y asegurar que el proyecto progresa adecuadamente.
- Entrega Final: El proyecto deberá ser enviado por correo electrónico al profesor antes de las 23:59h del día siguiente a la sesión número 8. El email debe incluir el enlace al repositorio de GitHub o el fichero del cuaderno adjunto.