Proyecto de aula: Aprendizaje de Máquina aplicado

Marco Teran
EAFIT
mtteranl@eafit.edu.co

Resumen

Este proyecto integrador pone en práctica el ciclo completo de un caso real de *Machine Learning* siguiendo **CRISP-DM**: comprensión del problema y la métrica, EDA, preparación de datos, modelado, validación honesta, interpretación y comunicación. Cada estudiante o equipo seleccionará un conjunto de datos, construirá *baselines* y comparará familias de modelos (lineales, SVM, ensambles y, cuando aporte valor, *deep learning*). La implementación será en **Python** con `pandas`/`scikit-learn` y cuadernos `IPython`, procurando reproducibilidad y reportes claros orientados a la toma de decisiones. El proyecto se estructura en **tres entregas** alineadas con el cronograma del curso.

1. Introducción

Este proyecto busca consolidar los aprendizajes del curso mediante la ejecución guiada de un problema aplicado de *Machine Learning*. Se espera una implementación rigurosa, con control de calidad de datos, métricas adecuadas al objetivo, comparación justa de modelos, análisis de umbrales en clasificación desbalanceada cuando aplique, e interpretación de resultados para soporte de decisiones.

2. Objetivos

2.1. Objetivo general

Ejecutar un proyecto de *Machine Learning* de principio a fin, con **CRISP-DM** como marco, entregando evidencia reproducible y conclusiones útiles para un caso real.

2.2. Objetivos específicos

- Seleccionar y caracterizar el conjunto de datos; definir problema y métrica(s) de éxito.
- Realizar EDA enfocada en calidad de datos, sesgos, fugas potenciales y variables clave.
- Construir *pipelines* de preparación y *baselines* reproducibles.
- Entrenar y comparar familias de modelos (lineales, SVM, árboles/ensambles; *DL* cuando agregue valor).
- Evaluar con validación honesta (*CV* adecuada; validación temporal en series), calibración y análisis de umbrales.
- Interpretar (SHAP/*permutation*) y comunicar resultados con recomendaciones claras.

3. Descripción del proyecto

Se espera que utilice la metodología de trabajo propuesta en el curso y las herramientas de modelamiento para llevar a cabo la planeación y ejecución de un proyecto aplicado. El conjunto de datos sobre el que trabajará puede ser seleccionado por ustedes (entre los conjuntos de datos propuestos) de acuerdo con sus intereses. El

objetivo es que a través de un proceso de extensiva experimentación con modelos de *Deep learning* poder llegar a obtener conclusiones con información valiosa que aporte en procesos de toma de decisiones en un dominio de aplicación particular.

El proyecto se desarrollará utilizando el lenguaje de programación *Python* y su entorno de herramientas para la computación científica, en forma de *Notebook* en el formato `iPynb`. Se debe presentar el proyecto tomando como referencia las etapas previas al despliegue de la metodología *CRISP-DM* para análisis de datos (IBM, 2012).

Se trabajará sobre un conjunto de datos elegido por el estudiante o equipo (ver Sección 6). La implementación será en **Python** con el ecosistema científico habitual y entrega en cuaderno `.ipynb`. El desarrollo seguirá las etapas previas al despliegue de **CRISP-DM**.

Herramientas sugeridas: `pandas`, `numpy`, `scikit-learn`, `matplotlib`. Para *deep learning* (si procede): `TensorFlow/Keras`.

4. Cronograma y entregas del proyecto

Las sesiones del curso se dictan los **jueves** de 18:00 a 21:00, con cierre el **miércoles 05/11/2025**. A continuación se detallan las **fechas clave** del proyecto:

Sesión	Fecha	Hito del proyecto
2	18/09/2025	Asignación de **Entrega 1** (semanal)
3	25/09/2025	**Entrega 1**: EDA + *data card* + *baseline*
5	09/10/2025	**Entrega 2** (mitad del curso)
9	05/11/2025	**Entrega 3** (final) + Presentación + Examen

5. Criterios de evaluación (resumen del curso)

Componente	Porcentaje
Proyecto aplicado (**3 entregas**)	**35 %**
Entrega 1 (semana 3)	5 %
Entrega 2 (mitad del curso)	10 %
Entrega 3 (día final)	20 %

6. Conjuntos de datos sugeridos

El proyecto debe basarse en alguno de los siguientes *datasets* públicos (selecciona uno acorde con tus intereses):

- **Google Play Store Apps:** Datos de 10 mil aplicaciones de la *App Store* obtenidas a través de *web scraping* con el objetivo de analizar el mercado de *Android*. [acceder]
- **Trip Advisor Hotel Reviews:** 20 mil reseñas de hoteles extraídas de *Tripadvisor*. Se puede usar este conjunto de datos para descubrir cómo son los mejores hoteles o usarla en sus propios viajes para *NLP* y análisis de sentimiento. [acceder]
- **Netflix Movies and TV Shows:** Este conjunto de datos consiste en programas de televisión y películas disponibles en Netflix a partir de 2019. El conjunto de datos se recoge de Flixable, que es un motor de búsqueda de Netflix de terceros. En 2018, publicaron un interesante informe que muestra que el número de programas de televisión en Netflix casi se ha triplicado desde 2010. Utilizando este conjunto de datos, se puede averiguar: qué tipo de contenido se produce en qué país, identificar contenido similar a partir de la descripción y muchas más tareas interesantes (vía Flixable); permite *content-based* y análisis descriptivo. [acceder]
- **Avocado Prices:** Datos históricos de los precios del aguacate y volumen de ventas en múltiples mercados de Estados Unidos; ideal para **series de tiempo**. [acceder]
- **Fashion MNIST:** Un conjunto de datos similar a *MNIST* con 70 mil imágenes con tamaño $28x28$ de prendas de ropa; presenta una tarea de **clasificación** supervisada. [acceder]
- **Students Performance in Exams:** Notas obtenidas por estudiantes en varias asignaturas.Estos datos se basan en la demografía de la población. Los datos contienen varias características como el tipo de comida que se le da al estudiante, el nivel de preparación para el examen, el nivel de educación de los padres y el rendimiento de los estudiantes en Matemáticas, Lectura y Escritura. Con los datos se pueden resolver varios tipos de problemas de regresión y clasificación. También se puede utilizar para encontrar qué factores pueden conducir a mejores resultados en los exámenes; CLF/REG. [acceder]
- **Credit Card Fraud Detection:** Este conjunto de datos ayuda a las empresas y equipos a reconocer las transacciones fraudulentas con tarjetas de crédito. El conjunto de datos contiene transacciones realizadas por titulares de tarjetas de crédito europeas en septiembre de 2013. El conjunto de datos presenta detalles de 284807 transacciones, incluidos 492 fraudes, ocurridos durante dos días; **clase desbalanceada**. [acceder]
- **Melbourne Housing Market:** El conjunto de datos del mercado de la vivienda de Melbourne es un recurso de aprendizaje favorito para los principiantes

en la ciencia de los datos. Tiene muchas características: datos numéricos, categóricos e incluso geográficos (latitud y longitud). Por tanto, también puede utilizarse para el análisis geoespacial y otros problemas de agrupación. Del mismo modo, también se pueden realizar tareas de regresión y clasificación con este conjunto de datos. [acceder]

- **IBM HR Analytics Employee Attrition & Performance:** Prediga el desgaste de sus empleados más valiosos. Descubra los factores que conducen al desgaste de los empleados y explora cuestiones importantes como *La relación entre la distancia de la casa al trabajo por puesto de trabajo y el desgaste* o *La relación entre el ingreso mensual promedio por educación y desgaste*. Este es un conjunto de datos ficticio creado por científicos de datos de *IBM*. [acceder]
- **UJIIndoorLoc:** Muchas aplicaciones del mundo real necesitan conocer la localización de un usuario para ofrecer sus servicios. La localización de usuarios ha sido un tema de investigación de interés en los últimos años. La localización de usuarios consiste en estimar la posición del usuario (latitud, longitud y altitud) mediante un dispositivo electrónico, normalmente un teléfono móvil. El problema de la localización en exteriores puede resolverse con gran precisión gracias a la tecnología GPS en los dispositivos móviles. Sin embargo, la localización en interiores sigue siendo un problema abierto, principalmente debido a la pérdida de la señal GPS en entornos interiores. Aunque existen algunas tecnologías y metodologías de posicionamiento en interiores, esta base de datos se centra en las basadas en huellas digitales WLAN (también conocidas como *WiFi Fingerprinting*). [acceder]
- **COVID19 Global Forecasting (Week 5):** En este desafío, tendrás que predecir el número diario de casos confirmados de COVID19 en varios lugares del mundo, así como el número de víctimas mortales resultantes, para fechas futuras. Este último reto incluye datos de condados del estado de EE.UU. El dataset se refiere a la quinta semana del desafío de pronóstico de COVID-19 en Kaggle, que es una competencia en la que se deben desarrollar pronósticos precisos para casos confirmados y fallecimientos relacionados con COVID-19 en diferentes regiones. El desafío tiene como objetivo identificar factores que parecen influir en la tasa de transmisión del COVID-19, no solo producir pronósticos precisos. La competencia se lanzó en colaboración con el grupo de investigación de la Oficina de Política Científica y Tecnológica de la Casa Blanca y otras organizaciones, y se basa en un conjunto de datos llamado COVID-19 Open Research Dataset (CORD-19), que fue diseñado para abordar preguntas científicas abiertas relacionadas con COVID-19 (competencia Kaggle). [acceder]

6.1. Entrega 1 (2 semana): EDA + *baseline*

Vence: 25/09/2025. **Peso:** 5 %.
Contenido mínimo:

Jupyter Notebook con carga y limpieza, *EDA*, *data card*, *baseline* reproducible; celdas de texto claras.
Reporte (`PDF`) breve estilo artículo: problema y trabajos relacionados (≥3 fuentes), datos y EDA, *baseline* y conclusiones preliminares.
Repositorio Git con estructura limpia (código, datos externos si procede, LaTeX, etc.). Enlace al final del PDF.

6.2. Entrega 2 (mitad del curso): comparación de familias

Vence: 09/10/2025. **Peso:** 10 %.
Contenido mínimo:

Jupyter Notebook con *pipelines* de entrenamiento, *tuning* y validación honesta de 2–3 familias (lineales/SVM/árboles/ensambles); análisis de umbral y PCA/selección de *features* cuando aplique.
Reporte (`PDF`) estilo artículo: métodos, configuración experimental, resultados y discusión.
Repositorio Git actualizado y referenciado en el PDF.

6.3. Entrega 3 (final): modelo, interpretación y recomendaciones

Vence: 05/11/2025. **Peso:** 20 %.
Contenido mínimo:

Jupyter Notebook con modelo final, calibración, interpretación (**SHAP**/*permutation*) y análisis de *trade-offs*.
Reporte (`PDF`) con resumen ejecutivo, metodología final, resultados principales y recomendaciones.
Repositorio Git definitivo (código, LaTeX, video si aplica).
Póster (`PDF`) problema, método y resultados clave de forma visual.

Requisitos de envío (todas las entregas).

- Enviar un único `.zip` con todo lo solicitado. No incluir binarios no requeridos.
- Nombre del archivo: `ml-project-username1-username2.zip`.
- Enviar antes de las 23:59 de la **fecha límite** correspondiente.

7. Recomendaciones

- Planifica las etapas de CRISP-DM con *checklists* y control de versiones.
- Verifica fugas, *data leakage* y sesgos en splits y *pipelines*.
- Usa métricas alineadas con el caso (ROC/PR, costo, calibración).
- Documenta decisiones y supuestos; favorece reproducibilidad.
- Colabora y distribuye tareas según fortalezas del equipo.

8. Bibliografía

IBM. *"Manual CRISP-DM de IBM SPSS Modeler."* CRISP-DM, 2012. [Descargar]