

Proyecto Curso II – Especialización Machine Learning Engineering

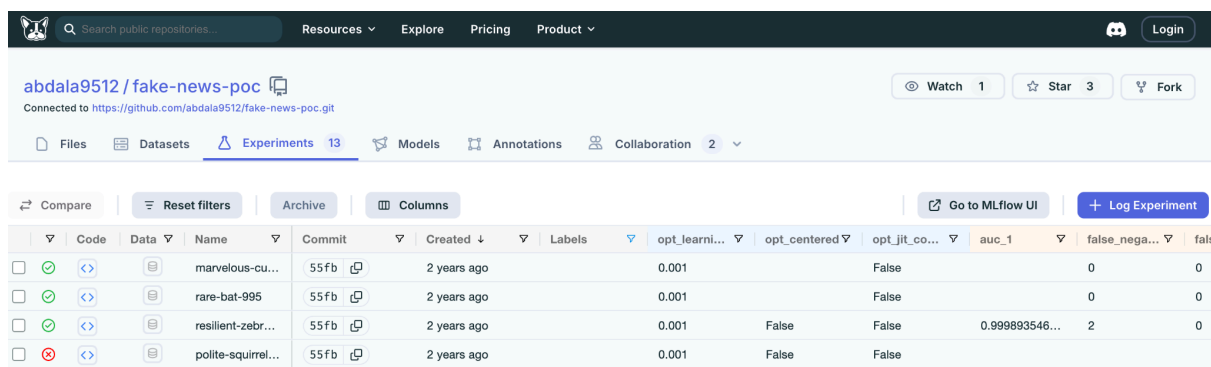
El proyecto del curso II de la especialización de Machine Learning Engineering tiene como objetivo enfrentar al estudiante a:

- Conocimiento teórico de algoritmos populares y avanzados de Machine Learning en la industria
- Administración de ciclo de vida de modelos de Machine Learning

Fecha máxima de entrega: Domingo 24 de Agosto de 2025

Entregable: Repositorio de GitHub en versión 1.0.0** con las siguientes secciones:

1. README.md
 - a. Problema de ML
 - b. Diagrama de flujo del proyecto
 - c. Descripción del dataset con su respectivo diccionario de datos
 - d. Model Card <https://www.kaggle.com/code/var0101/model-cards>
 - e. Resultados con Métricas de evaluación offline y online
 - f. Conclusiones
2. Estructura del repositorio de código con:
 - a. Carpeta de notebooks
 - i. Notebook preprocesamiento de datos
 - ii. Notebook de Machine Learning
 - b. Carpeta de datos (.csv, .txt, .parquet)
 - c. Módulo de código reusable
 - d. Scripts de ejecución (Preprocesamiento, entrenamiento y predicción)
3. Link con evidencia de experimentos realizados en MLflow con sus respectivo artefactos y un modelo productivo (e.g. <https://dagshub.com/abdala9512/fake-news-poc/experiments>)
 - a. Se espera de los experimentos tener **métricas, parámetros y artefactos** en ellos.



The screenshot shows the MLflow DAGshub interface for the 'fake-news-poc' project. The table lists four experiments, each with a status icon, code link, name, commit hash, creation time, and various metrics.

	Code	Data	Name	Commit	Created	Labels	opt_learni...	opt_centered	opt_jit_co...	auc_1	false_nega...	fals
<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	marvelous-cu...	55fb	2 years ago		0.001		False		0	0
<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	rare-bat-995	55fb	2 years ago		0.001		False		0	0
<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	resilient-zebr...	55fb	2 years ago		0.001	False	False	0.999893546...	2	0
<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	polite-squirrel...	55fb	2 years ago		0.001	False	False			

4. Release de la versión 1.0.0** con sus respectivas notas <https://docs.github.com/es/repositories/releasing-projects-on-github/managing-releases-in-a-repository>
5. Ramas Main y Development (al menos una pull request cerrada exitosamente)
6. Documentación sobre la estrategia de git utilizada
7. [OPCIONAL] Otros (.gitignore, requirements.txt, instrucciones de ejecución)

Indicaciones Generales del proyecto

1. El conjunto de datos es de libre elección. Se sugiere utilizar las siguientes fuentes para seleccionar un conjunto de datos:
 - a. Kaggle <https://www.kaggle.com/datasets>
 - b. Google Datasets <https://datasetsearch.research.google.com/>
 - c. UCI datasets <https://archive.ics.uci.edu/datasets>
2. Definir el problema de Machine Learning que se quiere resolver (Supervisado, no supervisado) y el subconjunto de problemas (Regresión, clasificación, Clustering, Reducción de dimensiones).

Sugerencias para el proyecto

- Para los experimentos de MLflow, se recomienda revisar esta excelente guía de métricas de ML <https://arize.com/blog-course/model-evaluation-metrics/>
- Usar datasets tabulares de tamaño no mayor a 100MB, además se recomienda no usar datos que requieran procesos de ingeniería de datos complejos, pues esto no será evaluado.
- Crear estructura del proyecto con cookiecutter <https://www.cookiecutter.io/>
- Revisar la guía básica de Markdown para preparar el README.md con mejor estructura. <https://markdown.es/>
- Agregar excepciones adicionales para archivos que estemos manejando. <https://docs.github.com/es/get-started/git-basics/ignoring-files>
- Usar gitkeep en carpetas provisionales del proyecto (data, tmp, etc).
- Para los Pull Request, documentarlos acorde a los cambios y usar **git and merge** para fusionarlo con la rama main. (Se recomienda github Flow <https://docs.github.com/es/get-started/using-github/github-flow>)

Evaluación

Componente	Tipo	Descripción	Porcentaje en la evaluación
Repositorio de Github	Obligatorio	El estudiante presenta su proyecto de acuerdo con la estructura definida por el profesor.	25%
Modelo de Machine Learning	Obligatorio	El estudiante documenta y desarrolla un modelo de machine learning alineado con sus hipótesis, además de evaluarlo con métricas que tengan sentido para el problema que previamente escogió y usando mlflow como herramienta de administración de modelos y experimentos.	50%
Buenas prácticas de desarrollo	Obligatorio	El estudiante sigue el proceso estándar de desarrollo basado en control de versiones, usando commits, pull requests y releases en su proyecto.	15%
Documentación	Obligatorio	El estudiante documenta el repositorio con archivos tipo markdown, o jupyter notebooks, así como también el código generado con docstrings, naming conventions, etc.	10%
Reto ML 1	Opcional	El estudiante demuestra dominio de los algoritmos vistos en clase realizando experimentos con más de uno y obteniendo resultados excepcionales (estos dependerán e tipo de problema que se busque resolver)	10% [Acumulable para todos los cursos de la especialización]
Reto ML2	Opcional	El estudiante hace uso del feature store para almacenar y servir carateristicas para modelos de machine learning. (Recomendado usar Feast , que fue la herramienta vista en clase)	10% [Acumulable para todos los cursos de la especialización]

Instrucciones de Envío

En el classroom del curso se habilitará una tarea donde se debe colocar el enlace del repositorio de github con la estructura previamente definida (Para la fecha de entrega final no debe haber ningún commit adicional y el repositorio debe estar en la versión 1.0.0**).

NOTA: El enlace del repositorio lo pueden compartir desde su creación, **El entregable se evaluará con el último commit del 24 de Agosto de 2025.**

****En caso de querer usar el mismo proyecto y repositorio del curso 1 de la especialización, crear una versión 2.0.0 en GitHub**

En caso de dudas adicionales estas serán atendidas por los canales de comunicación habituales.