



Optimización del ciclo de vida de un proyecto ML

Un caso de uso de Mlflow



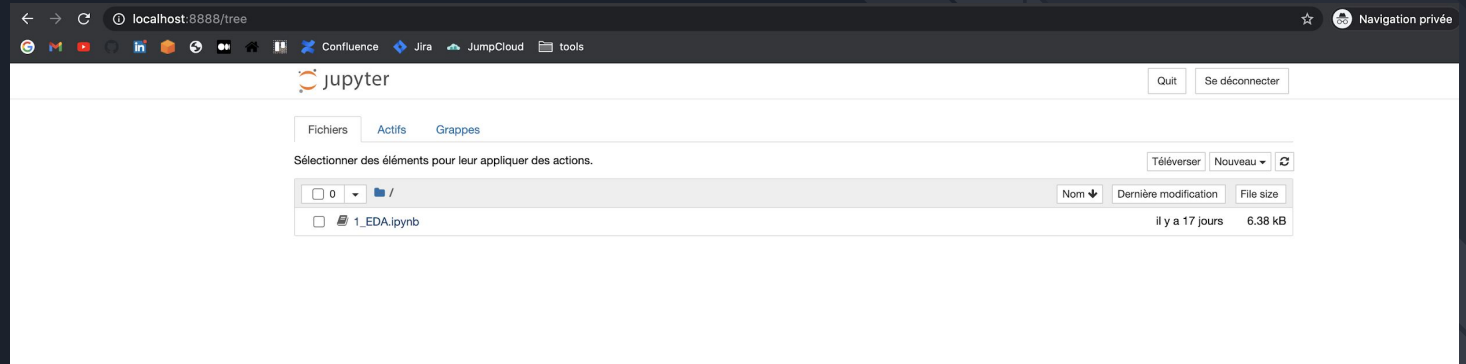
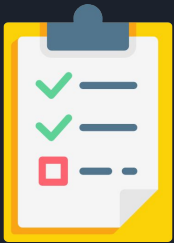
Plan

1. Una situación bastante común
2. Presentación de Mlflow
3. Demo

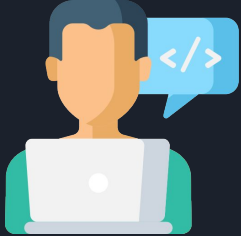


Una situación bastante común ...

Semana 1 - comienzas tu proyecto



Semana 2: tienes más datos



Home Page - Select or create x 1_EDA - Jupyter Notebook x +

localhost:8888/tree

Confluence Jira JumpCloud tools

jupyter Quit Se desconectar

Fichiers Actifs Grappes

Sélectionner des éléments pour leur appliquer des actions. Téléverser Nouveau ↕

| <input type="checkbox"/> | Nom | Dernière modification | File size |
|--------------------------|------------------------------------|--------------------------------|-----------|
| <input type="checkbox"/> | data_v1 | il y a 4 minutes | |
| <input type="checkbox"/> | data_v2 | il y a 4 minutes | |
| <input type="checkbox"/> | 1_EDA.ipynb | Actif il y a 17 jours | 6.38 kB |
| <input type="checkbox"/> | 2_modeling_on_data_v1.ipynb | Actif il y a 3 minutes | 72 B |
| <input type="checkbox"/> | 2_modeling_on_data_v1_and_v2.ipynb | Actif il y a quelques secondes | 72 B |
| <input type="checkbox"/> | 2_modeling_on_data_v2.ipynb | Actif il y a 3 minutes | 72 B |
| <input type="checkbox"/> | feedbacks.sqlite | il y a 4 minutes | 0 B |
| <input type="checkbox"/> | metadata.csv | il y a une minute | 0 B |
| <input type="checkbox"/> | new_data.xlsx | il y a 4 minutes | 0 B |



Semana 3: otro científico de datos se une al proyecto



Home Page - Select or create x 1_EDA - Jupyter Notebook x +

localhost:8888/tree

Confluence Jira JumpCloud tools

Jupyter Quit Se desconectar

Fichiers Actifs Grappes

Sélectionner des éléments pour leur appliquer des actions. Téléverser Nouveau ↻

| <input type="checkbox"/> | 0 | | | Nom | Dernière modification | File size |
|--------------------------|---|--|--|--|-------------------------|-----------|
| <input type="checkbox"/> | | | | / | | |
| <input type="checkbox"/> | | | | data_v1 | il y a 13 minutes | |
| <input type="checkbox"/> | | | | data_v2 | il y a 13 minutes | |
| <input type="checkbox"/> | | | | 1_EDA.ipynb | Actif il y a 17 jours | 6.38 kB |
| <input type="checkbox"/> | | | | 1_EDA_Alice.ipynb | il y a une minute | 6.38 kB |
| <input type="checkbox"/> | | | | 2_modeling_on_data_v1.ipynb | Actif il y a 12 minutes | 72 B |
| <input type="checkbox"/> | | | | 2_modeling_on_data_v1_Alice.ipynb | il y a une minute | 72 B |
| <input type="checkbox"/> | | | | 2_modeling_on_data_v1_and_v2.ipynb | Actif il y a 9 minutes | 72 B |
| <input type="checkbox"/> | | | | 2_modeling_on_data_v1_and_v2_Alice.ipynb | il y a une minute | 72 B |
| <input type="checkbox"/> | | | | 2_modeling_on_data_v2.ipynb | Actif il y a 12 minutes | 72 B |
| <input type="checkbox"/> | | | | 2_modeling_on_data_v2_Alice.ipynb | il y a une minute | 72 B |
| <input type="checkbox"/> | | | | feedbacks.sqlite | il y a 13 minutes | 0 B |
| <input type="checkbox"/> | | | | metadata.csv | il y a 10 minutes | 0 B |
| <input type="checkbox"/> | | | | new_data.xlsx | il y a 13 minutes | 0 B |

Semana 4: empezas a ver resultados



Fichiers Actifs Grappes

Sélectionner des éléments pour leur appliquer des actions.

☐ 0 ▾

/ exports

| | |
|--------------------------|---------------------------------|
| <input type="checkbox"/> | .. |
| <input type="checkbox"/> | model_v11_alice_best.pkl |
| <input type="checkbox"/> | model_v1_alice_all_data.pkl |
| <input type="checkbox"/> | model_v1_bob.pkl |
| <input type="checkbox"/> | model_v1_bob_auc_0-Copy1.76.pkl |
| <input type="checkbox"/> | model_v1_bob_auc_0.76.pkl |
| <input type="checkbox"/> | model_v2_bob.pkl |
| <input type="checkbox"/> | model_v3_alice.pkl |

Semana 5: se le pide información



- ❑ ¿Cuáles son las métricas de rendimiento del mejor modelo?
- ❑ ¿En qué datos y variables se entrenó este modelo?
- ❑ ¿Qué diferentes experimentos ha realizado cada científico de datos?
- ❑ ¿Es posible implementar los dos mejores modelos para equipos comerciales?





mi**flow**TM



Una plataforma para organizar el ciclo de vida de sus modelos de AA

“MLflow es una plataforma de código abierto para administrar el ciclo de vida de ML, incluida la experimentación, la reproducibilidad, la implementación y un registro de modelo central. MLflow ofrece actualmente cuatro componentes ”

MLflow Tracking

Record and query experiments: code, data, config, and results

[Read more](#)

MLflow Projects

Package data science code in a format to reproduce runs on any platform

[Read more](#)

MLflow Models

Deploy machine learning models in diverse serving environments

[Read more](#)

Model Registry

Store, annotate, discover, and manage models in a central repository

[Read more](#)

Integrado en todo el ecosistema de ciencia de datos

Integrations with:



PyTorch

K Keras



RAPIDS



python



ONNX



XGBoost

LightGBM

spaCy



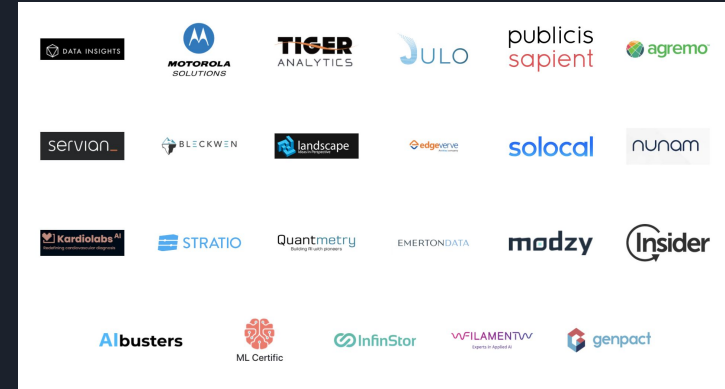
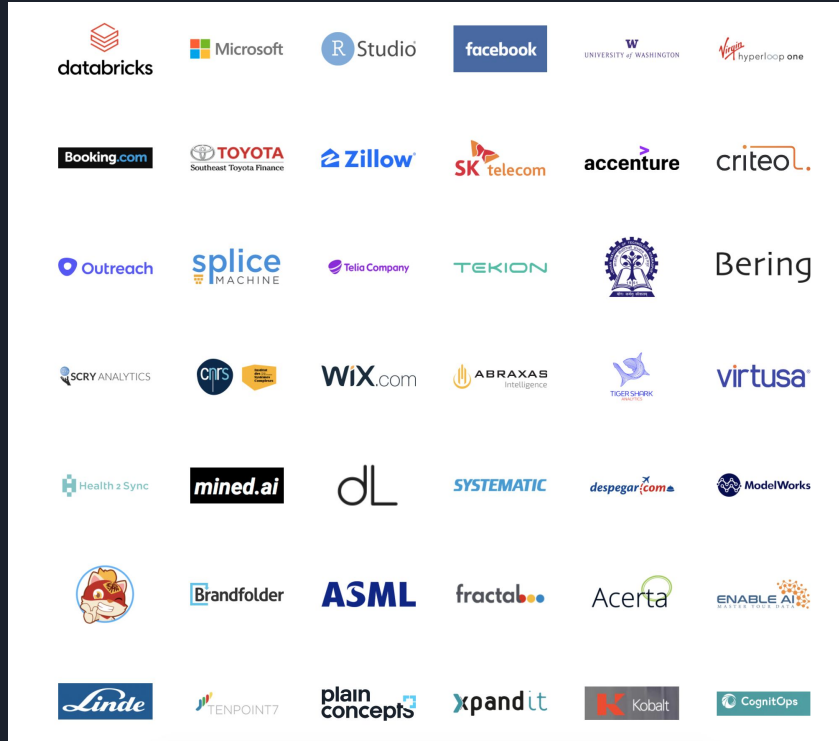
OPTUNA

RAY

CONDA



Y adoptado por varias empresas



Soluciones **alternativas**



TensorFlow Extended



Michelangelo

Jeremy Hermann, Machine Learning Platform @ Uber

Uber




TensorBoard



El ciclo de vida de un proyecto de AA





MLflow nos impulsa a aplicar las mejores prácticas de MLOps

MLOps = Aprendizaje automático + DEV + OP

- **DEV:** empaquetado, implementación, prueba, lanzamiento
 - **OPS:** configuración, monitoreo
-
1. Una cultura que unifica desarrollos y operaciones
 2. Aboga por la automatización y el monitoreo en todas las etapas de la construcción (integración, prueba, lanzamiento e implementación)
 3. Permite acelerar la transición a la producción.
 4. Permite la detección rápida de errores y fallos



MLflow Tracking



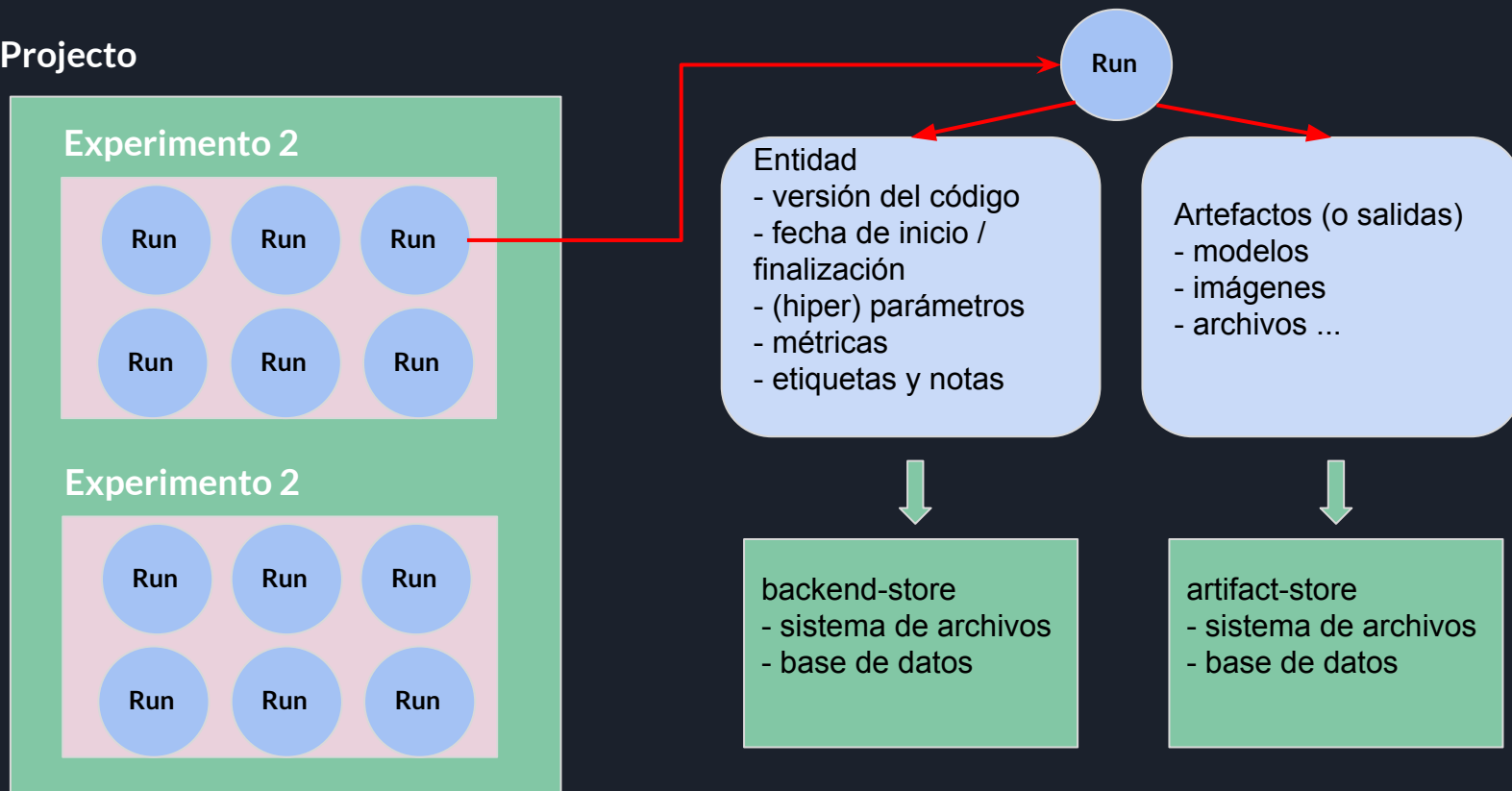


Qué permite MLflow **tracker**?

- ❑ **parámetros:** `n_estimators`, `max_depth`, `epochs`, `kernel_size`, `dropout`, `batch_size`?
- ❑ **métricas:** AUC, MAE, MSE; Puntuación F1, accuracy, R cuadrado (r^2)
- ❑ **datos:** ¿en qué versión de los datos (es decir, volumen y número de características) se entrenó dicho modelo?
- ❑ **artefactos:**
 - ❑ **los modelos:** guardados en disco en formato binario (pickle, h5, joblib, etc.)
 - ❑ **salidas** (que no sean plantillas, por ejemplo, imágenes, csv, texto, html, etc.)
- ❑ **fuentes:** ¿qué script / cuaderno inició este experimento?
- ❑ **etiquetas y comentarios:** información y anotaciones (colaborativas y / o individuales) sobre una ejecución

Terminología de MLflow: ejecuciones y experimentos

Proyecto





```
pip install mlflow
```

Experiment

localhost:5000/#/experiments/1/s?orderByKey=tags.%60mlflow.source.git.commit%60

mlflow Experiments Models

Experiments + -

Search Experiments

Default

training experiment

training experiment

Track machine learning training runs in an experiment. [Learn more](#)

Experiment ID: 1 Artifact Location: file:///Users/ahmed.besbes/projects/mlflow/mlruns/1

Notes

None

run

Search Runs: metrics.rmse < 1 and params.model = "tree" and tags.mlflow.source.type = "LOCAL"

Showing 100 matching runs

| | Start Time | Run Name | User | Source | Version | Models | Parameters | Metrics | | | | |
|--|---------------------|----------|--------------|----------|---------|---------|------------|--------------|--------------|----------|-------|-------|
| | | | | | | | max_depth | max_features | n_estimators | accuracy | auc | f1 |
| | 2021-03-05 17:38:55 | - | ahmed.bes... | train.py | 0bbaa7 | - | 21 | None | 275 | 0.785 | 0.691 | 0.538 |
| | 2021-03-05 17:38:54 | - | ahmed.bes... | train.py | 0bbaa7 | sklearn | 21 | sqrt | 275 | 0.787 | 0.688 | 0.535 |
| | 2021-03-05 17:38:53 | - | ahmed.bes... | train.py | 0bbaa7 | sklearn | 19 | log2 | 275 | 0.788 | 0.693 | 0.542 |
| | 2021-03-05 17:38:50 | - | ahmed.bes... | train.py | 0bbaa7 | sklearn | 19 | None | 275 | 0.785 | 0.694 | 0.543 |
| | 2021-03-05 17:38:48 | - | ahmed.bes... | train.py | 0bbaa7 | sklearn | 19 | sqrt | 275 | 0.788 | 0.692 | 0.54 |
| | 2021-03-05 17:38:47 | - | ahmed.bes... | train.py | 0bbaa7 | sklearn | 17 | log2 | 275 | 0.79 | 0.696 | 0.547 |
| | 2021-03-05 17:38:44 | - | ahmed.bes... | train.py | 0bbaa7 | sklearn | 17 | None | 275 | 0.786 | 0.699 | 0.55 |
| | 2021-03-05 17:38:43 | - | ahmed.bes... | train.py | 0bbaa7 | sklearn | 17 | sqrt | 275 | 0.789 | 0.697 | 0.548 |
| | 2021-03-05 17:38:41 | - | ahmed.bes... | train.py | 0bbaa7 | sklearn | 15 | log2 | 275 | 0.791 | 0.702 | 0.555 |
| | 2021-03-05 17:38:39 | - | ahmed.bes... | train.py | 0bbaa7 | sklearn | 15 | None | 275 | 0.788 | 0.704 | 0.559 |
| | 2021-03-05 17:38:38 | - | ahmed.bes... | train.py | 0bbaa7 | sklearn | 15 | sqrt | 275 | 0.793 | 0.704 | 0.56 |
| | 2021-03-05 17:38:37 | - | ahmed.bes... | train.py | 0bbaa7 | sklearn | 13 | log2 | 275 | 0.793 | 0.712 | 0.57 |

mlflow

ExperimentsModels

GitHub

training experiment > Run 2622ec9357dc4bb29104a41e0337e1d4

Date: 2021-03-05 17:38:54

User: ahmed.besbes

Source: train.py

Duration: 1.4s

Git Commit: 0bbaa7a1f9af0f402865b336d737884d0c03f889

Status: FINISHED

Notes

None

Parameters

| Name | Value |
|--------------|-------|
| max_depth | 21 |
| max_features | sqrt |
| n_estimators | 275 |

Metrics

| Name | Value |
|-----------|-------|
| accuracy | 0.787 |
| auc | 0.688 |
| f1 | 0.535 |
| precision | 0.588 |
| recall | 0.491 |

▼ Tags

| Name | Value | Actions |
|------|-------|---------|
|------|-------|---------|

No tags found.

Add Tag

| | | |
|-----------------------------------|------------------------------------|------------------------------------|
| <input type="text" value="Name"/> | <input type="text" value="Value"/> | <input type="button" value="Add"/> |
|-----------------------------------|------------------------------------|------------------------------------|

▼ Artifacts

▼ model

- MLmodel
- conda.yaml
- model.pkl

Full Path: s3://mlflow-artifact-store-demo/1/70c1791605c448f89f01df584597e9b5/artifacts/model
Size: 0B

[Register Model](#)

MLflow Model

The code snippets below demonstrate how to make predictions using the logged model. You can also [register it to the model registry](#).

Model schema

Input and output schema for your model. [Learn more](#)

| Name | Type |
|------------|------|
| No Schema. | |

Make Predictions

Predict on a Spark DataFrame:

```
import mlflow
logged_model = 's3://mlflow-artifact-store-demo/1/70c1791605c448f89f01df584597e9b5/artifacts/model'

# Load model as a Spark UDF.
loaded_model = mlflow.pyfunc.spark_udf(logged_model)

# Predict on a Spark DataFrame.
df.withColumn(loaded_model, 'my_predictions')
```

Predict on a Pandas DataFrame:

```
import mlflow
logged_model = 's3://mlflow-artifact-store-demo/1/70c1791605c448f89f01df584597e9b5/artifacts/model'
```

Auto logging - Keras

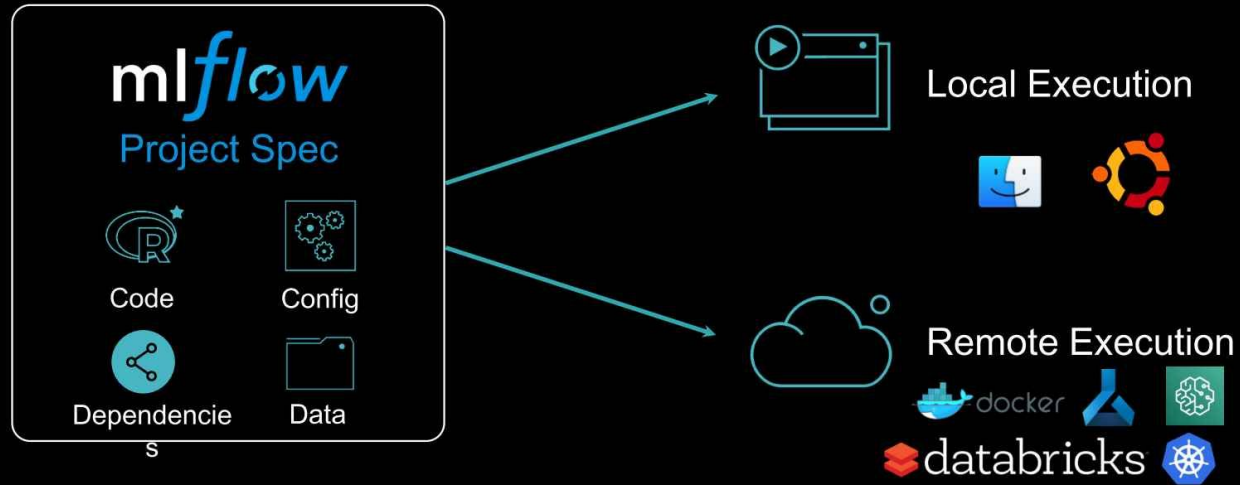
```
import mlflow
import mlflow.keras
# Build, compile, enable autologging, and train your model
keras_model = ...
keras_model.compile(optimizer="rmsprop", loss="mse", metrics=["accuracy"])
# autolog your metrics, parameters, and model
mlflow.keras.autolog()
results = keras_model.fit(
    x_train, y_train, epochs=20, batch_size=128, validation_data=(x_val, y_val))
```

Enables (or disables) and configures autologging from Keras to MLflow. Autologging captures the following information:

Metrics and Parameters

- Training loss; validation loss; user-specified metrics
- Metrics associated with the **EarlyStopping** callbacks: **stopped_epoch**, **restored_epoch**, **restore_best_weight**, **last_epoch**, etc
- **fit()** or **fit_generator()** parameters; optimizer name; learning rate; epsilon
- **fit()** or **fit_generator()** parameters associated with **EarlyStopping**: **min_delta**, **patience**, **baseline**, **restore_best_weights**, etc

MLflow Projects



MLflow Models





Recursos

- <https://kaskada.com/insights/a-guide-to-mlops-for-data-scientists-part-1>
- <https://medium.com/swlh/hyperparameter-tuning-with-mlflow-tracking-b67ec4de18c9>
- <https://www.mlflow.org/docs/latest/tutorials-and-examples/tutorial.html>