

# mle-template

Classic MLE template with CI/CD pipelines

Using technologies:

- Analytics and model training
    - Python 3.x
    - Pandas, NumPy, SkLearn
  - Testing
    - unittest + coverage
  - Data / Model versioning
    - DVC
  - CI/CD
    - GitHub Actions
- 

## Links:

- Docker Image: [ml-pipe-twitter-sentiment \(https://hub.docker.com/repository/docker/danielto1404/ml-pipe-twitter-sentiment/general\)](https://hub.docker.com/repository/docker/danielto1404/ml-pipe-twitter-sentiment/general)
- 

## Dataset

Twitter Sentiment Analysis Dataset from [Kaggle \(https://www.kaggle.com/c/twitter-sentiment-analysis2\)](https://www.kaggle.com/c/twitter-sentiment-analysis2). Sentiment analysis is a common task in the field of Natural Language Processing (NLP). It is used to determine whether a piece of text is positive, negative, or neutral. In this dataset, the task is to classify the sentiment of tweets from Twitter.

---

## Workflow

1. Download dataset from [Kaggle \(https://www.kaggle.com/c/twitter-sentiment-analysis2\)](https://www.kaggle.com/c/twitter-sentiment-analysis2)
  2. Analyze dataset and create simple baseline model in this [notebook \(./notebooks/twitter-sentiment-analysis.ipynb\)](#)
  3. Transform notebook to python scripts in [src \(./src\)](#) folder
  4. Put dataset into S3 bucket using DVC
  5. Created Dockerfile and [docker-compose.yml \(./docker-compose.yml\)](#)
  6. Created CI / CD pipelines using GitHub Actions:
    - [CI \(./github/workflows/ci.yaml\)](#)
    - [CD \(./github/workflows/cd.yaml\)](#)
  7. Added logging with [Greenplum \(https://greenplum.org/\)](https://greenplum.org/) database during functional testing
-

## Run tests

Run data preprocessing tests:

```
python -m unittest src/unit_tests/test_preprocess.py
```

Run model training tests:

```
python -m unittest src/unit_tests/test_training.py
```

## Logs from CD pipeline

```
twitter-sentiment_1 | INFO:root:Fitting model
twitter-sentiment_1 | INFO:root:Train F1 0.8117694303924563 | Valid F1 0.7406303833044623
twitter-sentiment_1 | INFO:root:Predicting on test data
twitter-sentiment_1 | INFO:root:Saving test predictions
twitter-sentiment_1 | .....
twitter-sentiment_1 | -----
twitter-sentiment_1 | Ran 6 tests in 0.679s
twitter-sentiment_1 |
twitter-sentiment_1 | OK
twitter-sentiment_1 | ....
twitter-sentiment_1 | -----
twitter-sentiment_1 | Ran 4 tests in 21.795s
twitter-sentiment_1 |
twitter-sentiment_1 | OK
twitter-sentiment_1 | Name                               Stmts  Miss  Cover   Missing
twitter-sentiment_1 | -----
twitter-sentiment_1 | src/constants.py                   3      0  100%
twitter-sentiment_1 | src/preprocess.py                  49      3   94%  23-25
twitter-sentiment_1 | src/train.py                       75     23   69%  90-91, 95-96, 121-143, 147
twitter-sentiment_1 | src/unit_tests/test_preprocess.py  43      0  100%
twitter-sentiment_1 | src/unit_tests/test_training.py    26      0  100%
twitter-sentiment_1 | -----
twitter-sentiment_1 | TOTAL                             196     26   87%
bigdata-course-01_twitter-sentiment_1 exited with code 0
```