Description:

Part 1 - Data Governance:

- Make an initial setup using `Data version control` tool. Add a dataset, so another could obtain it via dvc pull after cloning your repository
- Define a DVC pipeline that will:
    - preprocess data
    - train a model
    - evaluate the model
    - generate a feature importance plot with a model agnostic method
- The pipeline should be reproducible using dvc repro
- Run experiments (for example, different Scaling, models) using development environment from previous step and save metrics using dvc metrics

Part 2 - CICD, testing:

- Create unit tests for python code from Part 1
- Create a github action which at least performs:
    - code quality check:
        - auto-formatting with black
        - linting with pylint - fail if less than a threshold example
    - run unit tests

Criteria:

- DVC pipeline defined in a simple, reproducible manner
- There is an existing remote from which one could pull data (use free tier of AWS/GCP, Google Drive, or any other that would be easy to share)
- Code style / code quality tools are used
- Use github actions for CICD

Materials:

Data governance & CICD:

DVC

- Intro to CI with DVC, CML and GithubActions
- DVC for data versioning
- DVC Pipelines+githubActions

Github Actions

- GA docs

Testing & monitoring:

- intro

- DS testing & monitoring (some examples from Google):
    - concept of drift
    - behavioral model test

Libraries for testing:

- code testing:
  pytest | unittest
- code quality (auto-formatting + testing):
  black & pylint

Explainable ML:

model agnostic methods for feature importance