VISIUM

Crowd Programming

Gaétan Ramet & Adrien Ruault

60+

Engineers and Consultants Throughout Europe

60+

Happy Enterprise Clients

dsm-firmenich
idorsia
MIGROS
Roche
ABB

150+

AI & Data Engagements

VISIUM

Pan-European Company

FT FINANCIAL TIMES

Recognized Leader

72nd Fastest Growing Company in Europe by Financial Times, 2023

Our Partners

ACADEMIA

EPFL
ETH zürich
Stanford University
UNIVERSITAT POLITÈCNICA DE VALÈNCIA

TECHNOLOGY PLATFORMS

databricks
Azure
snowflake
Google Cloud
OpenAI
aws
data iku
Collibra

Context and objectives

Dataset: Student Portuguese Grades

- Tabular data
- ~30 features
- ~600 examples

Objectives

- Process the data to make it usable by a model
- Develop a model that can predict portuguese grades
- Measure model performance
- [Optional]: Investigate visualizations and model explainability.
- **Teach you some stuff!!**

Tech stack

- Python
- Pandas, numpy, sklearn
- DVC
- Pipenv
- Git, pre-commit, ruff

Crowd Programming

What is Crowd Programming?

- Software development approach
- The whole team works on the same thing, at the same time, in the same space, and on the same computer
- Similar to pair-programming, but with more people involved

Workflow

- The team brainstorms together and take decisions (You + Adrien)
- One person writes actual code, implementing other people's idea (Gaétan)

How can you contribute?

- **Option 1:** Just raise your hand and share your ideas :)
- **Option 2**: Send suggestions through *Slido*

All ideas are welcome!!

DVC to structure the ML pipeline

Challenge

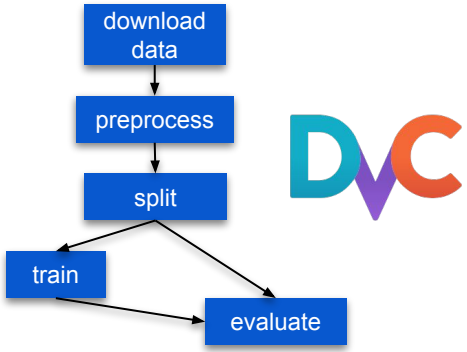
The ML pipeline coded in a notebook or a python script

- The code becomes difficult to maintain as it grows
- Notebooks do not work well with git
- Teamwork is not easy when all the code is entangled in one messy entrypoint.

Solution

Structure your code as a Pipeline

- **Maintainability.** Each step has its own responsibility and is easier to maintain.
- **Collaboration.** Easier to share responsibilities between team members
- **Efficiency.** Possibility to benefit from a cache: run steps only when necessary.
- **DVC** can help you achieve this



Pipenv to manage python environment

Challenge

The python environment is not well managed

- Python packages are difficult to install because package dependencies are not respected
- Reproducing the environment on other machines can be difficult
- Your environment is conflicting with other environment on your machine.

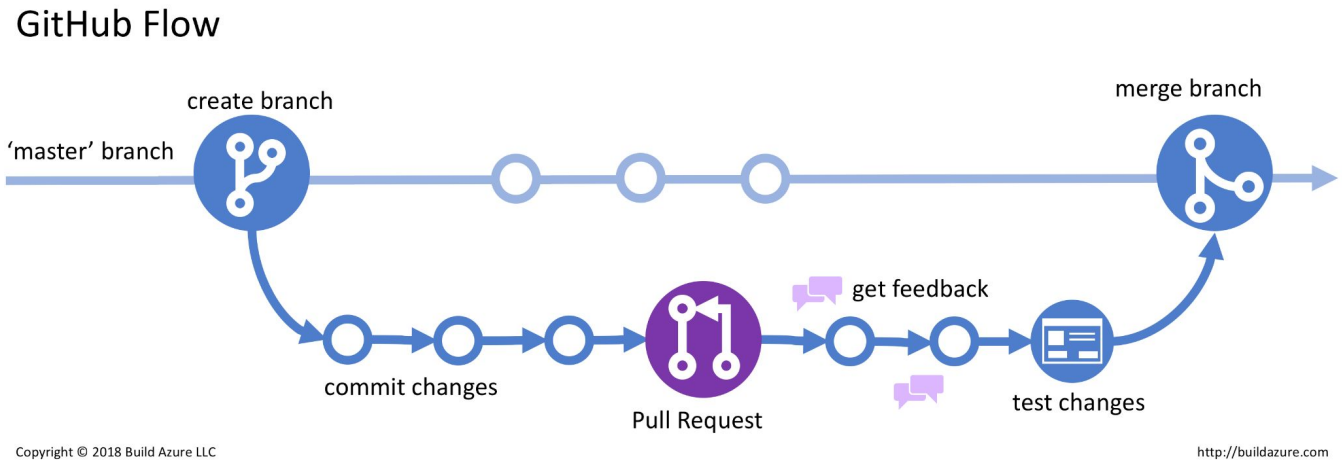
Solution

Use a modern python environment manager.

- **Dependency resolution.** The compatibility between all your packages is ensured.
- **Reproducibility.** Your environment can be reproduced everywhere with the same package versions thanks to locking.
- **Isolation.** Your environment is independent from other python projects on your machine.
- **Pipenv** will be used in this workshop.

```
$ pipenv
```

Git workflow



Tour of the code base

https://github.com/VisiumCH/crowd-programming-workshop-202404

- Where to define the ML pipeline?
- Where to define the logic of the steps?
- Where to store the data?
- Where to define my python environment?

Let's goooo!