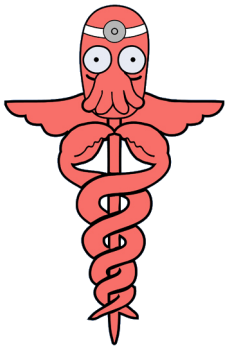
T7 - AI & Big Data

T-DEV-810

zoidberg2.0

Project



2.1.4

{EPITECH.}

zoidberg2.0

binary name: zoidberg2.0_*$AcademicYear*_*$GroupNumber*.zip
language: whatever works

Given some X-ray images, use machine learning to help doctors detecting pneumonia.

Doctors granted you access to 3 datasets if needed.

It's up to you to decide when and how to use the dataset (training, testing, evaluating performance, tuning parameters, …).

You must:

- use a train-validation-test procedure,
- use a cross validation procedure,
- compare your results with a simple train test split,
- use one of the datasets to tune your algorithms.

Specialists insist on the importance that you explore and test various methods, **and to compare results**.

optimization, feature engineering, metrics, PCA

A clear and concise way to present results should always prevail.

You are expected to deliver:

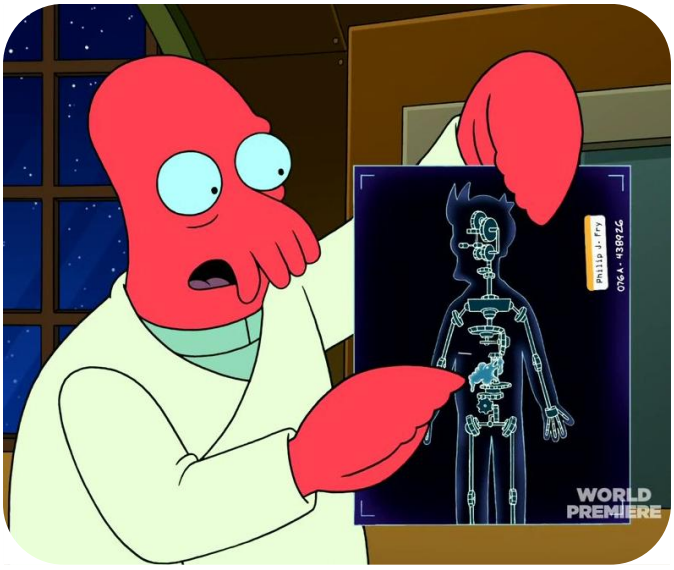
- technical documents
  a **Jupyter notebook-like file**, containing code and text, possibly graphics and an **html-file** to prove your results without rerunning the code
- a synthesis document
  a **pdf file** to sum up your results and figures

There are ways to save a trained algorithm and load it afterwards in order to obtain the same results when you run it again!

Bonus

You can improve your project in many ways, including:

- implementing a self-organizing map to help you visualize your result,
- thorough learning through neural networks,
- predicting on 3 classes: no pneumonia, virus pneumonia, bacteria pneumonia,
- …



Recommandations

- **Time and space**

> *So much effort for such a poor result.*
>
> *P. Sorentino*

Think carefully about needed resources (time and space) and a procedure, **before starting implementation**.
For quicker or better results, you may *-or may not-* want to transform your data.

Some libraries could help you.

- **Bad habits**

Do not let your algorithm(s) make bad habits.
Find the correct balance between **bias** and **variance**.
Cross validation procedures can help you solve this problem.
It's also a good idea when you don't have a lot of data.
Algorithms' parameters can also help you to find a trade-off between bias and variance, this is why you need a good understanding on how they work.

- **Good metrics**

Find an explicit way to show your results, in a readable way.
You must choose among many various metrics and **select the appropriate one(s)**.

Look at more advanced metrics like ROC-AUC score.
It will give you a deeper understanding of your results.
You should also be able to explain what AUC measures and its advantages over other metrics.

First ML project!? You should dig a bit and find answers: the only way out is the way in!