

Project 2018-19

Choose a dataset from publicly available datasets such as:

UCI - Machine learning Repository:

<https://archive.ics.uci.edu/ml/datasets.html>

DATA.GOV.UK

<https://data.gov.uk/data/search>

US.GOV

<https://www.data.gov/>

KAGGLE

<https://www.kaggle.com/datasets>

WORLD BANK

<https://data.worldbank.org/data-catalog>

The selected dataset must be at least order of 10,000 instances.

Your task is to rank/filter/process/adjust features in the dataset **and** run several ML algorithms to predict a class (the class depends upon the dataset).

For this project, **max. 2 students each team.**

It is preferable not to use WEKA, scikit-learn or other tools more suitable for large datasets.

A good project should include:

- description of the dataset and of the ML task
- (non trivial) analysis, filtering, engineering of features,
- description of the algorithms used
- comparison of experiments,
- careful evaluation with robustness test
- final comments on performance metrics.

See examples of good submitted projects on the course web site.

Free choice projects

Students **are allowed** to propose their own projects in which they build their own dataset, however, they **MUST** contact the instructor in order to evaluate the complexity of the task (required background knowledge, availability of a sufficiently large dataset or clear ideas on how to create a dataset, clear ideas on what is the problem, what is the predictive task to be considered, how can it be evaluated).

Projects must be delivered BEFORE September 2019 (before the new 2019-20 course)