Team: Agata Kaczmarek, Mateusz Stączek
***Name of the chosen project:* A - Feature selection**
Progress: We chose methods and datasets, which we will use during the project:

- methods using information theory measures
  - MIM
  - JMI
  - mini-max
- arbitrary methods
  - random forest
  - lasso regression
- datasets
  - **Artificial example**, for which the set of significant features is known
    - generate x from normal distribution, y from normal with other parameters, z also
    - target=x+y
    - so x and y should be important features and z not
  - **Artificial example** in which one of the MI-based methods doesn't work
    - make columns correlated, then MIM will fail
  - **Real world datasets x5 - classification**
    - Heart Disease (https://archive.ics.uci.edu/dataset/45/heart+disease)
      - 13 features
      - target: integer (0,1,2,3,4) but can be treated as True (1,2,3,4) and False (0)
    - Wine (https://archive.ics.uci.edu/dataset/109/wine)
      - 13 features
      - target: categorical, three types of wine
    - Predict students dropout and academic success (https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success)
      - 36 features
      - target: categorical, three statuses of students: dropout, enrolled, graduate
    - CDC Diabetes Health Indicators (https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators)
      - data is huge (253680 records), we will probably take only some subset of it
      - 21 features
      - target: binary
    - Phishing Websites (https://archive.ics.uci.edu/dataset/327/phishing+websites)
      - 30 features
      - target: binary