# Jak zrobić model?

LINK kaggle: https://www.kaggle.com/kemical/kickstarter-projects/version/7#ks-projects-201801.csv

1. Preprocessing

a. Ogólne spojrzenie na dane, co się w nich znajduje (research)

i. kwantyle, odstające wartości,

ii. shapy,

iii. missing values,

iv. pomyłki w danych (powielone rekordy)

b. Data cleaning – doprowadzenie do dobrej formy

i. Normalizacja i standardyzacja danych (m.in. klasyfikowanie ‘strings’ w decision trees jest latwiejsze niż w linear regression, gdzie należy to zakodować)

ii. Kodowanie danych tekstowych jeśli należy

iii. Usuwanie outliers

c. Podział datasetu na trening + validacja + test (metoda crossvalidation + decision)

2. Wybór modelu

a. Linear regression

b. Logistic regression

c. KNN vs. K-means

d. Decision tree

e. Random forest

f. SVN

g. Bayes

h. Xgboost

3. Wybor:

a. Ktore zmienne wziac pod uwage

b. Jakie parametry modelu nalezy uwzglednic

c. Zastosowanie parametrów które umozliwią wybór i porównanie modeli (MSE itd.)

4. Final model + performance na datasecie testowym.