

- 1) We used pandas to read the data from the csv-files into data frames and prepared the feature matrix X and response vector y.
- 2) For the feature selection we have chosen to use SelectKBest from sklearn.feature\_selection with the f\_classif scoring function. It selects the features according to the k highest scores. K was selected with cross-validation.
- 3) The classifier was build using Support Vector Machichines (sklearn.svm.SVC) with a polynomial kernel. The hyperparameters kernel, gamma and C were chosen using a cross-validated grid-search from sklearn.model\_selection.
- 4) To combine the feature selection and the classifier, we put them into a pipeline.
- 5) We trained the classifier using the training data.
- 6) We printed out the cv mean score and the accuracy on the training data to be able to value the quality of the classifier.
- 7) We predicted the labels on the test data and have written them to a csv-file.

remarks:

For all cross-validation tasks we used accuracy as a scoring function, since this is how our predictions will be graded. Cross-validation was mostly performed on the Euler computer. We also tested kNN, Gradient Boost classifier, Logistic Regression classifier, but SVM has yield the best cv-score.