Machine Learning

# Submission Assignment 03

von
Benjamin Ellmer
(S2210455012)

Mobile Computing Master
FH Hagenberg

December 10, 2023

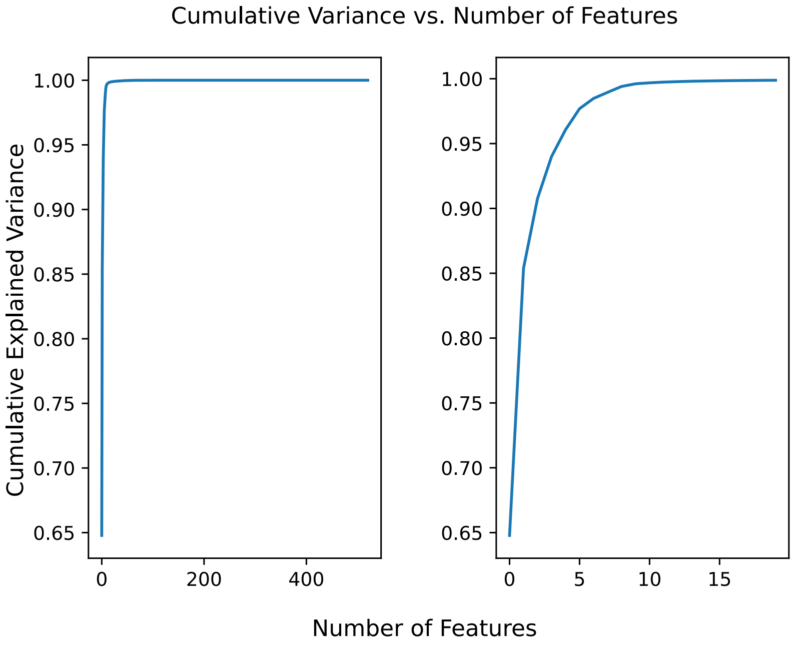
# 1. Feature Selection - (Recap from last assignment)

The Feature Extraction was done in the last exercise and the code of the last exercise was modified to write all features into a csv file. This gives me the advantage, that I do not have to recalculate all features for each run, just read the created csv.

Additionally, the targets were appended to the features csv file, but they were immediately dropped. This was just done because I need the targets for the model selection and I did not want to read and write 2 csv files.

## 1.1. PCA

According to the PCA in Figure 1, only 10 features have 99% variance. Therefore I chose to continue with only 10 principal components and find the best model.



# 2. Model Selection

## 2.1. Pre-Model Selection

First of all I wanted to try the configuration from the lecturer code examples, but I did rewrite the provided code a little bit:

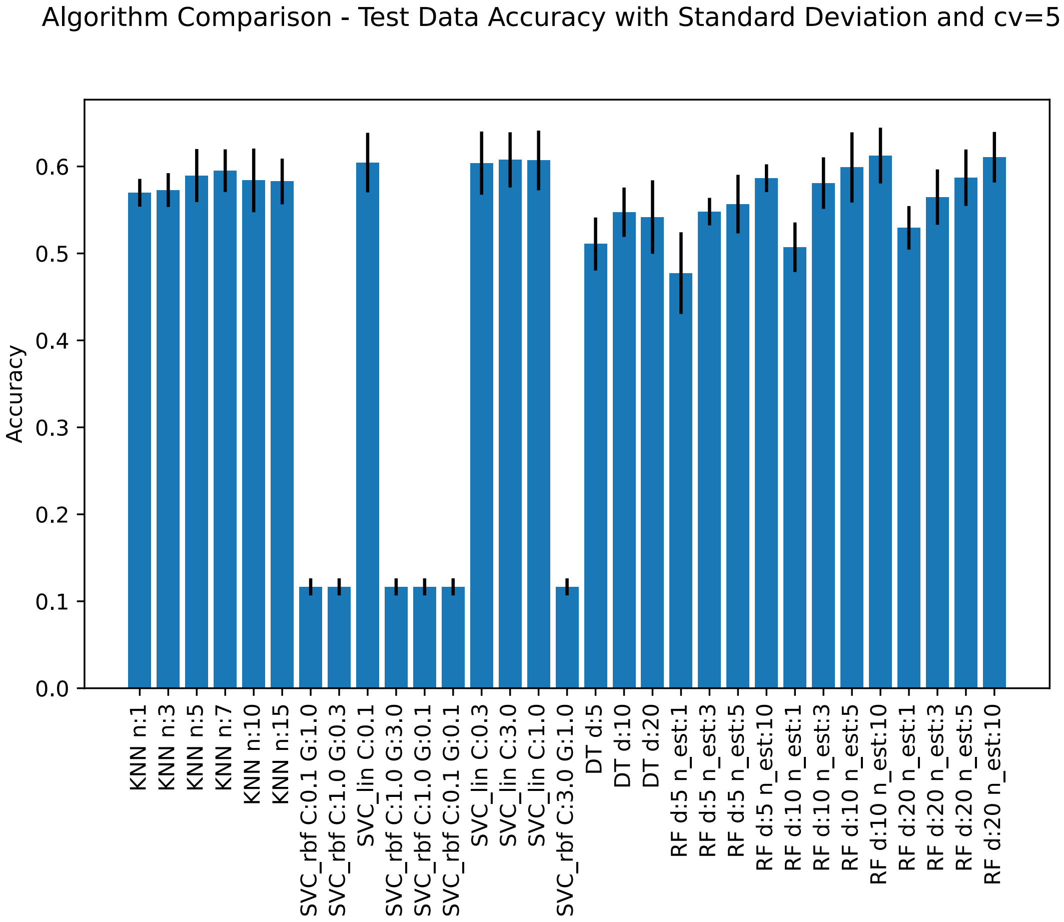
- I changed all range parameters for SVC from (-5, 5) to (-2, 2) to reduce the number of fits
- I did not store all models at once, but I stored each model it its own file appending the run id (e.g. `trained_NearestNeighbors_1.sav`)
- I used `RandomizedSearchCV` for SVC, because I read it is faster
- I set the `n_jobs` parameter of the searches to `-1`, so it uses all available CPUs
- I tried to shorten the labels, so we can see the parameters in the plot

Saving the models with their run id makes it easier for me to jump around between the runs without recalculating all models. Additionally, splitting up the sav file for each algorithm makes it possible to change the parameters of one single algorithm without recalculating the model for all algorithms.

## 2.2. Run 1

In Figure 2 we can see the results with 5-fold cv, my findings are:

- KNN gets worse wit >= 10 neighbors -> use 7, 8, 9
- SVC_rbfs all achieve only 10% accuracy -> remove rbf
- SVC_linear have very similar results -> increase gamma
- Decision Tree gets worse >= 10 -> use 10 - 20
- Random Forest gets better with more estimates -> increase and start with 10


Algorithm Comparison - Test Data Accuracy with Standard Deviation and cv=5
Accuracy
KNN n:1
KNN n:3
KNN n:5
KNN n:7
KNN n:10
KNN n:15
SVC_rbf C:0.1 G:1.0
SVC_rbf C:1.0 G:0.3
SVC_lin C:0.1
SVC_rbf C:1.0 G:3.0
SVC_rbf C:1.0 G:0.1
SVC_rbf C:0.1 G:0.1
SVC_lin C:0.3
SVC_lin C:3.0
SVC_lin C:1.0
SVC_rbf C:3.0 G:1.0
DT d:5
DT d:10
DT d:20
RF d:5 n_est:1
RF d:5 n_est:3
RF d:5 n_est:5
RF d:5 n_est:10
RF d:10 n_est:1
RF d:10 n_est:3
RF d:10 n_est:5
RF d:10 n_est:10
RF d:20 n_est:1
RF d:20 n_est:3
RF d:20 n_est:5
RF d:20 n_est:10