Challenge 1 Report

Name: Lukas Nöbauer
ID: k11702439

For the first Challenge in AILS I first revisited the slides and the provided notebooks for the challenge in order to get familiar with the topic and especially with molecule data. After I did some short explorative analysis on the whole dataset, especially on the labels.

Preprocessing:
For preprocessing I first split the labels and the smiles in two different sets. With the un-preprocessed feature set only containing the smile strings and the label dataset containing the labels for the bioassays 1-11. Up to then I left the labels as they were, also keeping the zeros (no information), as I would get rid of them during training afterwards.

Feature Engineering:
I tried adding several features, such as Morgan Fingerprints, descriptor data and also one-hot encoding the periodic table with the number of atoms in the molecule. The two latter turned out to be not too beneficial for the performance of my models, so I only kept the Morgan Fingerprints dataframe with fp-length of 2024. Furthermore, I also tried to set up an automatic feature engineering pipeline, by applying a task-wise linear regression on the data and keeping only the features with a p-value < 0.05. This also did not turn out to be very successful, probably due to the nature of the data and the problem (classification).

Training:
For training the models I set up a pipeline loop in which for each task I queried the data for non-zero labels (unknowns) and then used this data for training. The resulting 11 models where then trained on between 500 – 4000 datapoints. To store the models from the loop I used a python dictionary to use them later on for predicting the submission set.

Models:
Considering my models, I used a wide variety of different models. I tried a simple Feed Forward Network, a self-regularizing Neural network (SNN) and Convolutional Neural Network. What was kind of strange was that these models seemed to yield a pretty good AUC_ROC performance on my local test set, but not on the public test test (ROC of 0.81 vs ROC of 0.64). Besides Neural Nets I also tried several linear models, primarily SVM, Random Forest and Gradient Boosted Trees. Finally, SVMs, especially SVRs (oddly not SVCs) yielded the best performance on the public test set, with around 0.73 ROC_AUC. To get the best model here I was using sklearns GridSearchCrossValidation algorithm to get the best parameters and afterwards I used a BaggingRegressor to further improve the generalization performance of the model. To fit the respective model I tried 48 different parameter sets in the grid approach. Furthermore, I used around 50 estimators for the BaggingRegressor to generalize the model. The whole training and prediction procedure took around 2 hours to execute.
Although I am not fully happy with my final score, I have learned a lot about processing molecule data and using it for Machine Learning. Looking forward to the next topics!