Challenge 1

Nicolas Madella-Mella / k12021216

1 Self Preparing

I started witch rewatching the recordings and reading up some papers, to get a better understanding for the task.

One of those papers was "Machine learning models for classification tasks related to drug safety" published online 10.06.2021 by Rácz et al.(Link provided at the end of the document)

2 Start

First I imported the modules which grew in number the further I proceeded in this project. Then I loaded the csv-files in pandas dataframes and furher inspected the data. To be specific, I checked the training set for NaN values and how balanced the datapoints are. After that i calculated the fingerprints from the molecules of the train and test set and split up the training set in predictions and fingerprints.

In my submission is no further preprocessing done, but for a while i tried to use oversampling such as ADASYN, which I couldn't implement correctly or in a helping way. The idea came from the document "Predicting carcinogens with machine learning and molecular fingerprinting" uploaded 10.07.2021 by Gurkamal Deol (Link provided at the end of the document)

3 Training , Model Setup and Evaluation

For the train and test split I used a 80/20 ratio. In my submission I implemented two Ml-methods, one was Logistic Regression inspired by Gurkamal Deol and the second one was Random Forrest inspired by the lecture. For Random Forrest i chose 120 seeds and 100 estimators and for Logistic regression I used 10-fold cross validation and 1000 max. iterations. The results where slightly better with Logistic Regression than with Random Forrest, but i haven't yet tried out any techniques such as gradient boosting or cross validation on the Random Forrest approach. From one of the readings inspired I also would've liked to try out an approach with a SVM. Unfortunately I lacked time for that, but I'll definitely come back to this task and dig deeper.

4 Links

"Machine learning models for classification tasks related to drug safety",
published online 10.06.2021 by Rácz et al:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8342376/

"Predicting carcinogens with machine learning and molecular fingerprint"ing,
uploaded 10.07.2021 by Gurkamal Deol:
https://medium.com/@gurkamaldeol/predicting-environmental-carcinogens-with-logistic-regression-knn-gradient-boosting-and-7973f88eb8b3