The two main challenges in the Task 3 were handling the imbalanced dataset and finding a suitable classifier. To deal with the imbalanced dataset, we wrote a resampling method which applies down- and up-sampling to the training dataset. Down-sampling was done by just taking only a fraction of all the possible inactive (majority) datapoints. Up-sampling was done by repeating the active (minority) datapoints multiple times to increase their occurrence. The rate of the down- and up-sampling could be set with parameters. The optimal parameters were determined by a grid search, where the most relevant factors were 30 for down-sampling and 15 for up-sampling in our function.

We used the OneHotEncoder from sklearn for python to encode the characters. This encoder was chosen to prevent dependencies from the other values, which would have happened with an ordinal encoder.

Finding a suitable classifier has also been done using a stratified 3-Fold cross validation from the sklearn library for pyhton. We tried multiple classifiers such as RandomForest, GaussianNaiveBayes, SVC and MultiLayerPerceptronClassifier (MLPClassifier). All these classifiers were obtained from the sklearn library for python. We found out that the best classifier was the MLPClassifier with the parameter alpha set to 0.00001.