The task 2 was split into 3 subtasks. The first subtask was to predict multiple labels per patient. Therefore, we used a multilabel classification approach. This was done using the OneVsRestClassifier with the estimator RandomForestClassifier from the sklearn library for python. The goal of the second subtask was to classify a binary label, this was done with the classifier RandomForestClassifier from the sklearn library as well. The last subtask was a regression task, this was solved by using the Ridge regression class from the sklearn library.

We couldn’t find much of an improvement from using a cross validation because checking all possible parameter configurations would have taken too long, so we decided on parameters that fitted well (For Ridge: alpha = 50000 and For RandomForest: n\_estimator=500).

The biggest challenge in this task was the imputation of the data and the feature engineering. That is because the provided data had a lot of missing elements and the dataset was quite big which made it difficult to handle.

For data imputation we manipulated the data in such a way that: if there was only one entry for a column per patient, we’d fill the whole column with that value. When there were more than two elements, we’d use a LinearRegression estimator from sklearn to predict the missing values per patient. And if there were no elements in a column per patient, we set all the values to the global mean in that column.

Additional to the imputation, we reduced the feature vector by only using the mean values of certain columns because we didn’t gain additional information over time, for example for the age of a patient.