Assignment 1: Classification

1. Study the data set carefully and answer the questions below

a) Report the class distribution. Is this a balanced or unbalanced data set?

The data set is not totally balanced because the label is "No" 156 times (53,3%) and "Yes" 136 times (46,7%), meaning the labels do not occur in the same number of times.

Nonetheless, it can be approximately said that this is a balanced dataset. This assumption can be made up to a 60:40 distribution.

b) Please select and justify a suitable metric to evaluate the performance of your classification model.

It should be avoided as much as possible that someone who suffers from depression is recognized as healthy, because the clinical picture can worsen greatly by a lack of therapy. In contrast, it would not do too much harm if someone who is healthy starts treatment. Therefore, increased attention is paid to achieving a high recall value for the class "True", i.e. that depression is present. Of course, the value for precision should also be as high as possible. But a high recall value has priority.

During my experiment, however, there were frequent difficulties with this, such as the recall value of both classes was optimized, or the accuracy was very poor. So, I decided to use the f1 score. This is usually a good choice, since it considers the two opposing parameters recall and precision.

c) Given the size of the data set, which validation option (e.g., percentage split, k-fold cross validation) do you think is suitable to be used in your machine learning experiments. Specify the validation option you are selecting for your machine learning experiments. Briefly explain the reason for using the validation option.

In my opinion, k-fold cross-validation would be the best option, as it offers the advantage that each instance is tested once. It also reduces the bias. The disadvantage of the computationally intensive load can be neglected due to the small amount of data in this data set.

2. Features extracted from the emotion data are represented in two forms

a) Which feature representation produces a better model?

In the experiments carried out, the different representations were mostly equal. In KNN the normalized representation performed a lot better, the decision tree also performed better with the normalized data. In linear SVM, although, the non-normalized data set performed better.

b) Explain how you determine the best performing model based on the performance metric you have selected. Can you explain why one feature representation is better than the other?

The best performing models were identified based on their f1 score. This was calculated using the classification_report method from the sklearn library. The best model with the non-normalized dataset was the linear SVM. The best model with normalized data set was the decision tree. It should be noted that the decision tree with non-normalized data set should also be taken into account, as it

has a recall value of 0.74 for the label "depression", which is the highest value of all models in this respect.

That the normalized representation performs better more often can be attributed to the fact that with the non-normalized form, distortions can arise if a person reports his or her mood several times a day. For example, if a person has a bad day and reports sadness five times on that day, the pure number seems high compared to a person who reports only once on that day.

4. Is your overall best performing model able to achieve very promising results (reach at least 0.8 of your selected performance metric)? Provide two suggestions on future strategies that can be used to improve the performance of your best performing model.

Unfortunately, despite several different metrics and many tried parameters, no model could be found that has a good performance.

One strategy to improve performance would be to collect more data. The more data the model has for training, the better it can learn the hypothesis function. It could also make sense to tune more parameters or to tune them more precisely. It might also make sense to ban some features from the dataset that have no influence on the target but cause noise.