## Used Datasets:

- Music
  - Beats per minute
  - Beats per minute statistics
  - Chroma
  - Mel Frequency Cepstral Coefficients
- Heart Failure
- Cover Type (Forest)

The dataset for Cover Type had to be adapted to accommodate the bad scalability behavior of the SVC-Classifier. When trying to process the dataset with all available rows, the processing did not finish in a reasonable amount of time (hours of waiting without any sign of success). Therefore, I reduced the amount of data rows to 200.000. The selection of the rows occurs at random, so there should be a good representation of the entire dataset even after the reduction.

## Used Classifiers:

- K-NN
- Perceptron
- Decision Tree
- Random Forest
- SVC

## Best performing classifier for each dataset:

| Classifier | Parameters | Dataset | Training time (seconds) | Test time (seconds) | Accuracy | Weighted F1 |
|---|---|---|---|---|---|---|
| K-NN | neighbors: 3 | **Forest Cover Type** | 2,68264 | 4,45632 | 0,93456 | 0,93439 |
| Decision Tree | min. samples/split: 50<br>min. samples/leaf: 1 | **Heart Failure** | 0,001 | 0 | 0,76768 | 0,7753 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Random Forest | n estimators: 100 max. feaatures: log2 | **Mel Frequency Cepstral Coefficients** | 0,21655 | 0,01001 | 0,65152 | 0,64978 |
| Random Forest | n estimators: 100 max. feaatures: sqrt | **Chroma** | 0,23396 | 0,01099 | 0,46061 | 0,45999 |
| Decision Tree | min. samples/split: 50 min. samples/leaf: 1 | **Beats per minute statistics** | 0,00301 | 0 | 0,29697 | 0,28009 |
| Decision Tree | min. samples/split: 100 min. samples/leaf: 1 | **Beats per minute** | 0,00099 | 0 | 0,21212 | 0,15518 |

Out of all, K-NN performed best regarding accuracy and the f1 measure. Therefore, I chose this classification for the confusion matrix:

| | **Spruce/Fir** | **Lodgepole Pine** | **Ponderosa Pine** | **Cottonwood/Willow** | **Aspen** | **Douglas-fir** | **Krummholz** |
|---|---|---|---|---|---|---|---|
| **Spruce/Fir** | **0,93755** | 0,00042 | 0 | 0,00253 | 0,05949 | 0 | 0 |
| **Lodgepole Pine** | 0 | **0,78739** | 0,00282 | 0,17121 | 0,02446 | 0,01411 | 0 |
| **Ponderosa Pine** | 0 | 0,00303 | **0,86125** | 0,05701 | 0,00404 | 0,07265 | 0,00202 |
| **Cottonwood/Willow** | 0,00012 | 0,00419 | 0,00214 | **0,94772** | 0,04222 | 0,00357 | 3,00E-05 |
| **Aspen** | 0,00425 | 0,00112 | 4,00E-05 | 0,06168 | **0,93291** | 0 | 0 |
| **Douglas-fir** | 0 | 0,00148 | 0,02964 | 0,03162 | 0,00099 | **0,93207** | 0,0042 |
| **Krummholz** | 0 | 0 | 0,09464 | 0,00315 | 0 | 0,21767 | **0,68454** |

Looking at this result, the model fits each category well except “Krummholz” which shows a lower value (0,68454).

Processing the Sound dataset did not yield as good of a result as the large Cover Type dataset. I assume this is due to the small amount of data available for processing (only 1000 in total and only 100 per category). Compared to the (although reduced) 200.000 datapoints available for the Cover Type, this assumption seems reasonable.

As already mentioned, SVC did not perform satisfactorily when considering the amount of time it takes to train and test the dataset. SVC scales badly (as mentioned on the documentation site for scikit-learn (https://scikit-

learn.org/stable/modules/generated/sklearn.svm.SVC.html): "*The fit time scales at least quadratically with the number of samples and may be impractical beyond tens of thousands of samples. For large datasets consider using* `LinearSVC` *or* `SGDClassifier` *instead, possibly after a* `Nystroem` *transformer.*"

On the Sound dataset, the best working combination for classification is mfcc classified with a Random Forest (n_estimators: 100 and max_features: log2). However, only an accuracy value of 0,652 and an f1 measure value of 0,65 could be achieved. This is considerably worse than what was achieved with a large dataset (Cover Type).

All results can be seen in the also submitted excel sheet (Exercise2_Michael_Goll_se21m003_Detailed_Results.xlsx).