

Group 5

Members:

- Clement
- Engelbert
- Wafaa
- Gbetondji

Approach:

- We have used the graphs of the scatter plots and the correlation matrix to investigate the relationships among features and find out which features are mostly related to the “class”.
- The results of the investigations have shown that features “f1”, “f3”, and “f8” are the ones with the highest correlation value.
- In our model, we created a class for the Rule-based Classifier. The constructor of the class sets the values for the features and threshold values to be used for classification.
- Our rule basically takes threshold values of the three features and sets the class labels accordingly.
- Regarding the training function, because the threshold is not a strict point and we have noticed some grey areas in the graphs, what we did is that we took a range of threshold values, predicted the results for the classes for each threshold value, and then returned the result with the highest accuracy value.
- Regarding the test dataset, we created a function to run the model on the test data and return a list of predicted class labels.
- The accuracy function computes the accuracy by dividing the number of correct samples by the total number of samples in the data provided.
- And finally, for the Bonus question, we implemented a function that takes the id of the sample that it is required to predict the class for using command line arguments, and it returns the predicted class of the given sample. To use this function you should add the flag --predict in the command line, and then pass the sample id, for example: to get the prediction for the sample with ID = 5, run:

```
python3 main.py --predict 5
```