Supervised Learning Classification of ecoli.csv dataset

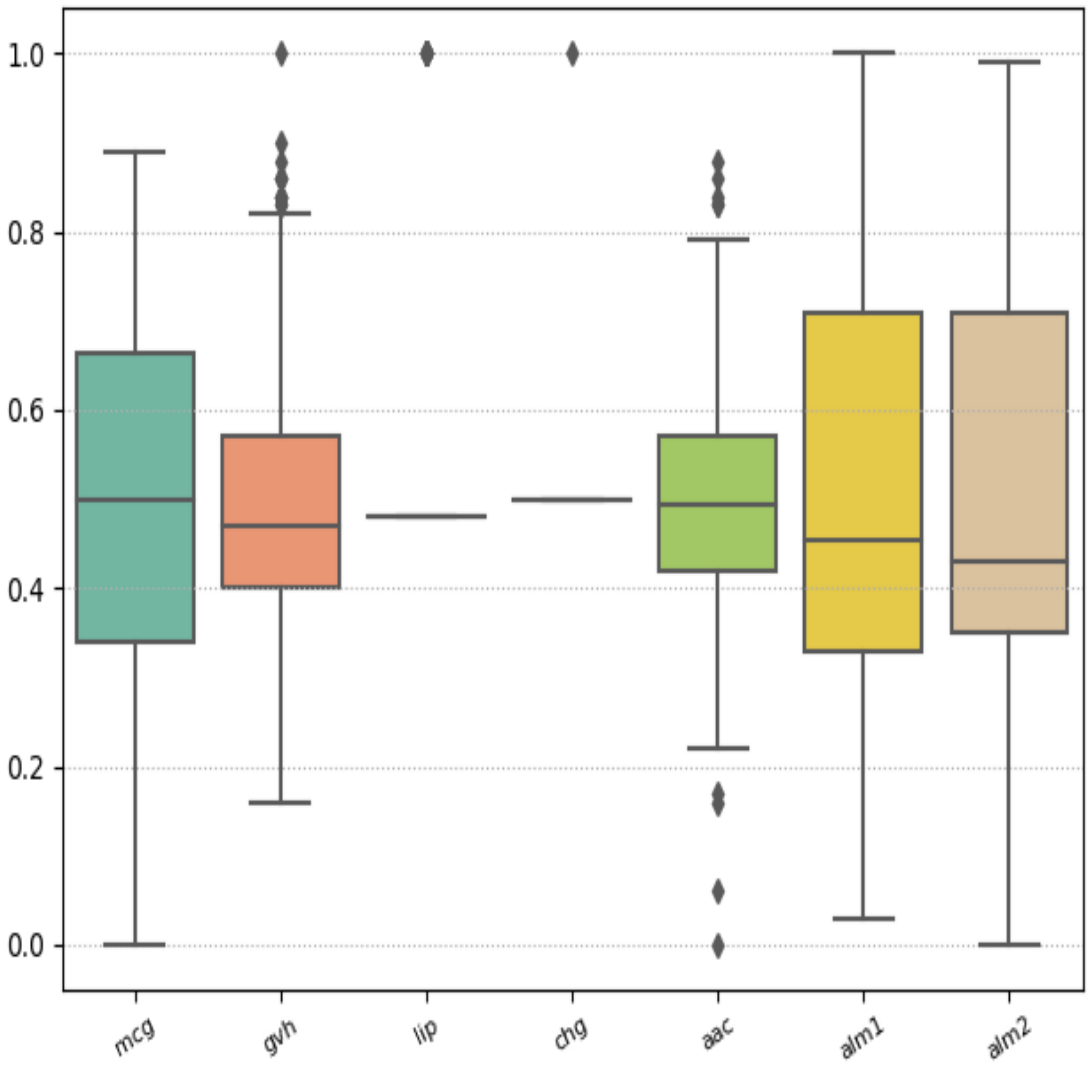
Content:

1) Get an overview of dataset
2) Compare Classification algorithms

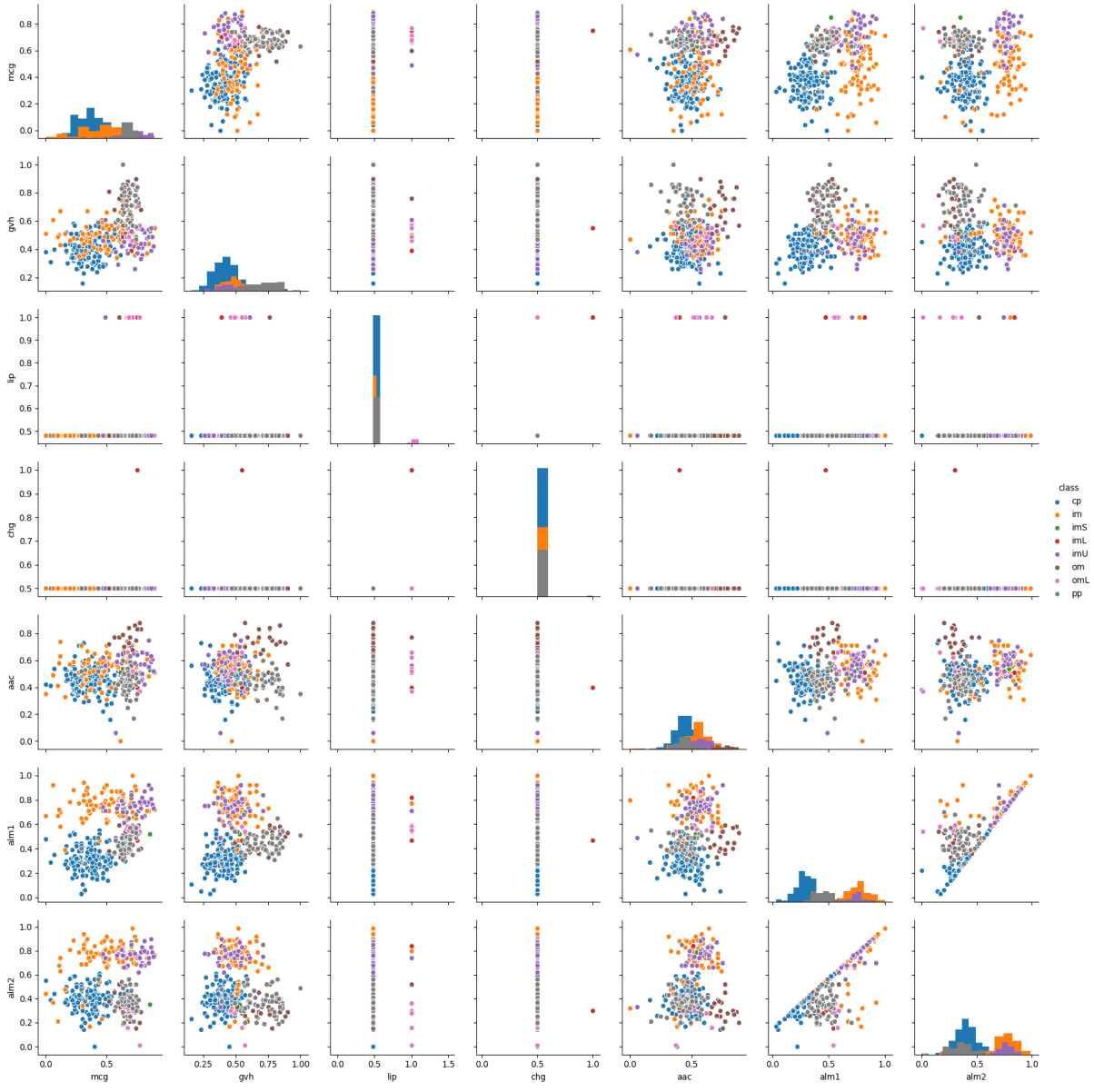
The following report shows a comparison of different supervised learning algorithms that classify the data of the ecoli.csv dataset.

1) Get an overview of the dataset:

The boxplot diagram shows an overview of the different features within the dataset and its attributes. On the X-Axis the different features of theecoli.csvdataset are represented. The Y-Axis shows how the attributes of the different attributes of the features. The boxplot is based on the minimum, first quartile, median, third quartile, and maximum.

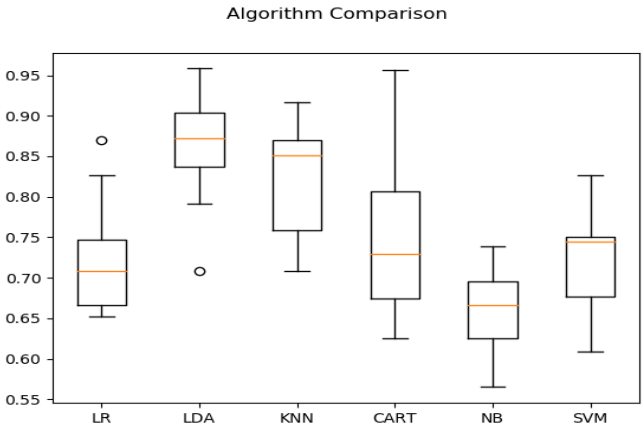


The second graph is a scatterplot that shows how the feature attributes are distributed for the different classes. Classification algorithms should show better results if clusters can be recognized within the scatterplot.



2) Compare classification algorithms

Now that we got a feeling of how the dataset is distributed, a random split for training and evaluation data has been conducted. The following boxplot diagram shows the comparison of the different supervised learning algorithms:



The accuracies of the different algorithms are as follows:

LR: 0.723913 (0.070639)

LDA: 0.864312 (0.071463)

KNN: 0.826268 (0.069874)

CART: 0.745652 (0.095326)

NB: 0.660145 (0.056979)

SVM: 0.727717 (0.067044)

Finally, the confusion matrix shows a comparison of the predictions by the algorithms with the test data (Y-Axis) and the validation through the actual class labels (X-Axis). In the optimum all data is distributed on a diagonal line from the top left corner to the bottom right corner. This means that the predictions match the validation.