Contents

1 Feature extraction

For the beginning of the assignment the data from the XL2022.csv file is loaded. The data contains 16 features with each 676 data points. Furthermore the data contains a belonging label to the data points.

1.1 Design matrix and test/train data

In the next part of the assignment the design matrix is extracted from the data frame. Additionally the data is separated into train and test data. The train data set has a size of 572x16, which means 572 data points with 16 features each. The Test data set has only 104 data points with 16 features each, which corresponds to a size of 104x16.

1.2 Covariance matrix

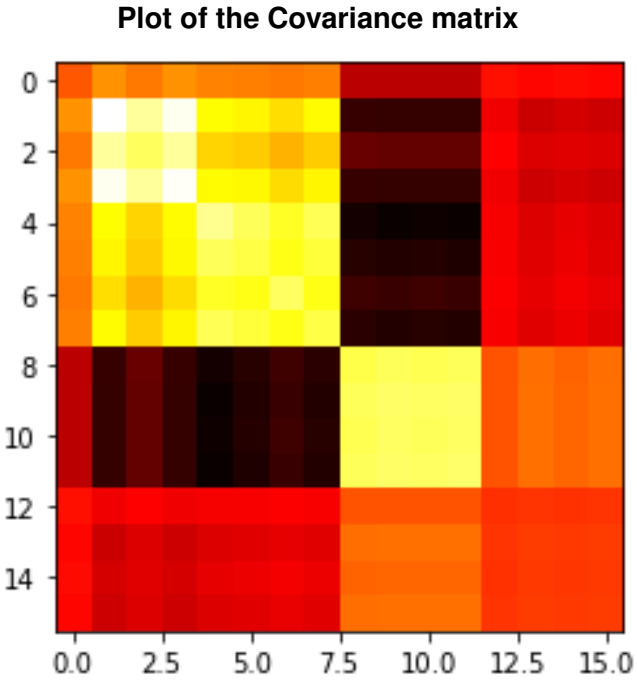

Figure 1.1: covariance matrix of the features

Figure 1.1 shows the result of covariance matrix according to the train data. To determine the reduced set of features, the abs value of a row of the covariance matrix was calculated and compared to a value started with 0. The value was increased till each ffeaturecategory, means contrast, correlation, energy and homogeneity is covered. Features with the index 0, 5, 6, 8, 12 and 15 are selected to reduce the set of features. These indexes relates to Contrast_0, Correlation_45 and 90, Energy_0, Energy_0 and Homogeneity_90 and 135. Afterwards the reduced design matrix X1 is extracted, it has a size of 676x6.

2 Classification

2.1 kNN classifier

The Scikit-learn libry was used to implement the k-nearest neibhbors classifier. Five neighbors were used and predicted classes of the test dataset by the model was compared with the true classes of the test dataset. The accuracy of the classifier when applied to the data containing all features is 0.9230769230769231. When its the reduced data the accuracy is 0.9038461538461539.

2.2 Random Forest

Next task was to train a random forest based on the training data of all features. The results are presented in figure. 2.1.

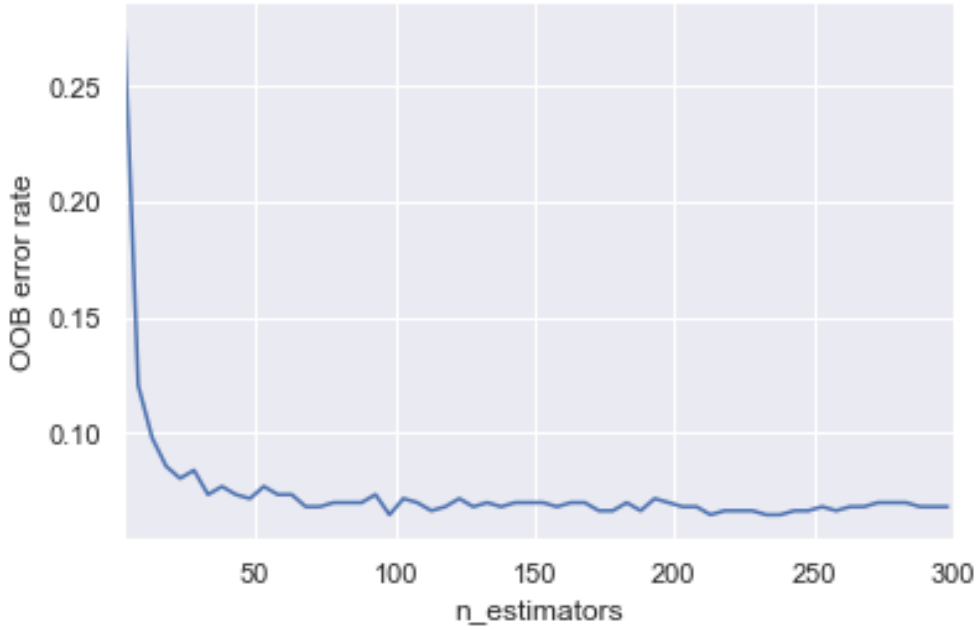

Figure 2.1: Low OOB error rate with more than 60 trees.

2.3 Importance of the various features

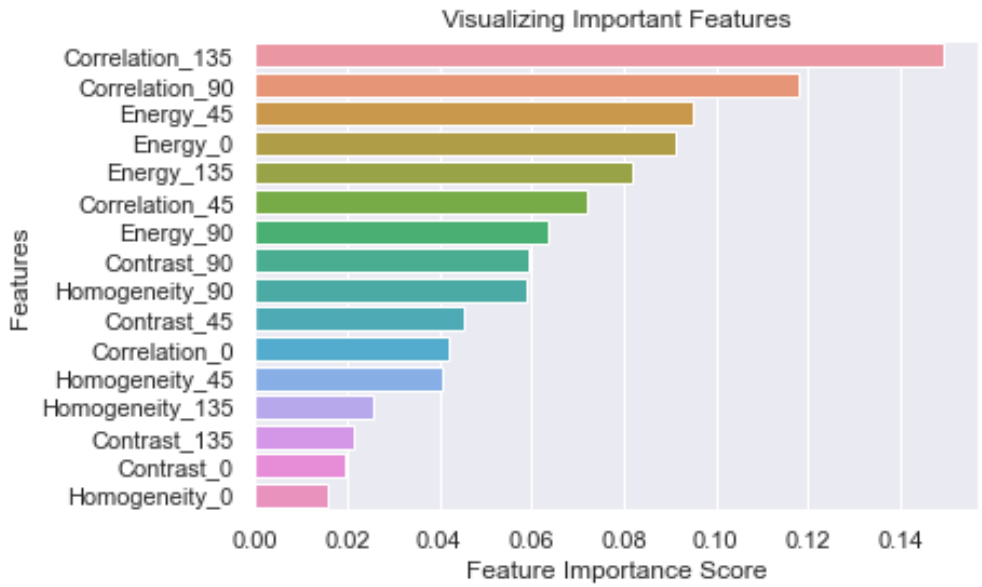

Figure 2.2: Feature importance

Last task of this section was to identify the most valuable feature for the random forest. Fig. 2.2 shows that the correlation and the energy feature categoryäre most valuable for the random forest. The contrast and homogeneity are not so important for the random forest.

List of Figures