KAUNAS UNIVERSITY of TECHNOLOGY

Faculty of Mathematics and Natural Sciences

Philipp Schlaus

Study module

P160B124 Machine Learning Methods

Laboratory work 1 report

Kaunas, 2024

CONTENTS

1 LAB TASK: 2.1 LINEAR REGRESSION

1.1 TASK 1

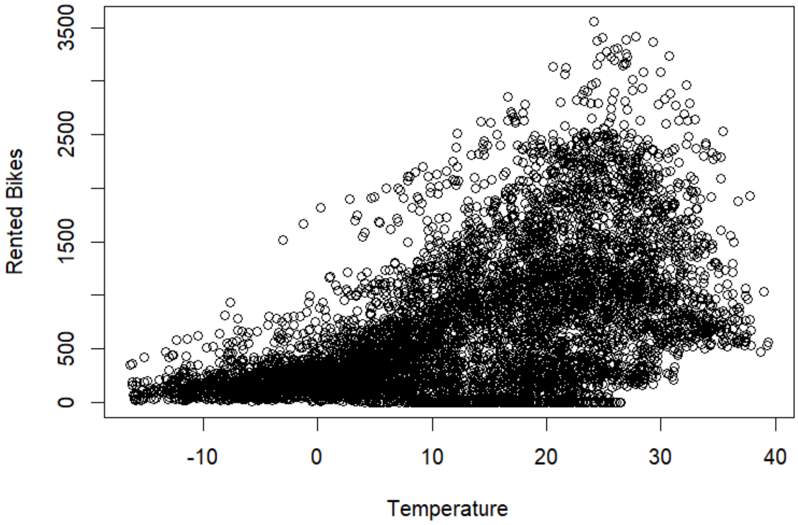
The datasets train_bike.csv and test_bike.csv have 14 features. The features and their types can be seen in the following list:

Feature	**Type**
Date	Character
Rented_Bike_Count	Integer
Hour	Integer
Temperature	Numerical
Humidity	Integer
Wind_speed	Numerical
Visibility	Integer
Dew_point_temperature	Numerical
Solar_Radiation	Numerical
Rainfall	Numerical
Snowfall	Numerical
Seasons	Character
Holiday	Character
Functioning_Day	Character

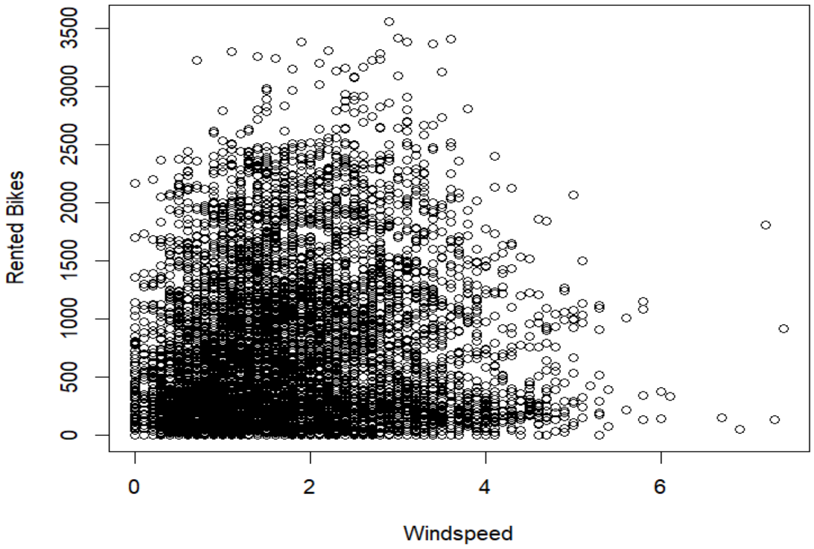
The train dataset has 7200 observations. The test dataset has 1560 observations.

1.2 TASK 2

I chose to compare the features Temperature and Windspeed to the amount of rented bikes. In the first plot, I plotted Temperature against Rented Bike Count.



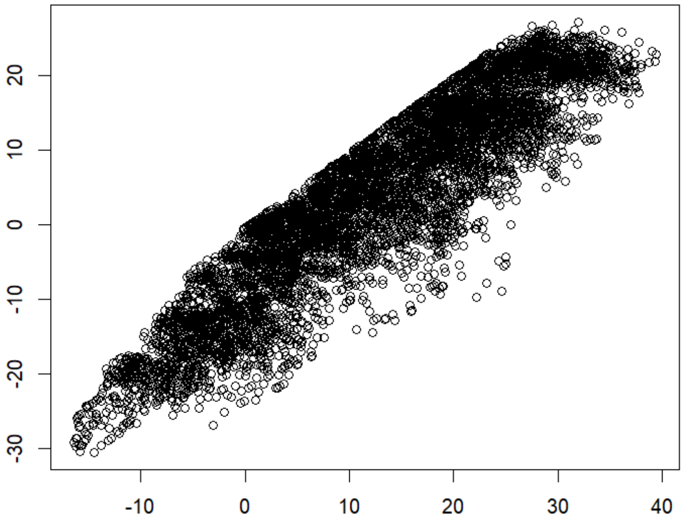
From this imagine, it can be seen, that the temperature has an effect on the amount of rented bikes. When the temperature is low, there are only a few rented bikes. But when the temperature is high, there are observations with very low and very high amounts of rented bikes. Therefore I would say the temperature is importatnt to predict rented bikes. The next plot shows Windspeed plotted against Rented Bike Count:



There are many cases with high and low amounts of rented bikes for any windspeed. This means the windspeed is not useful for predicting the amount of rented bikes.

1.3 TASK 3

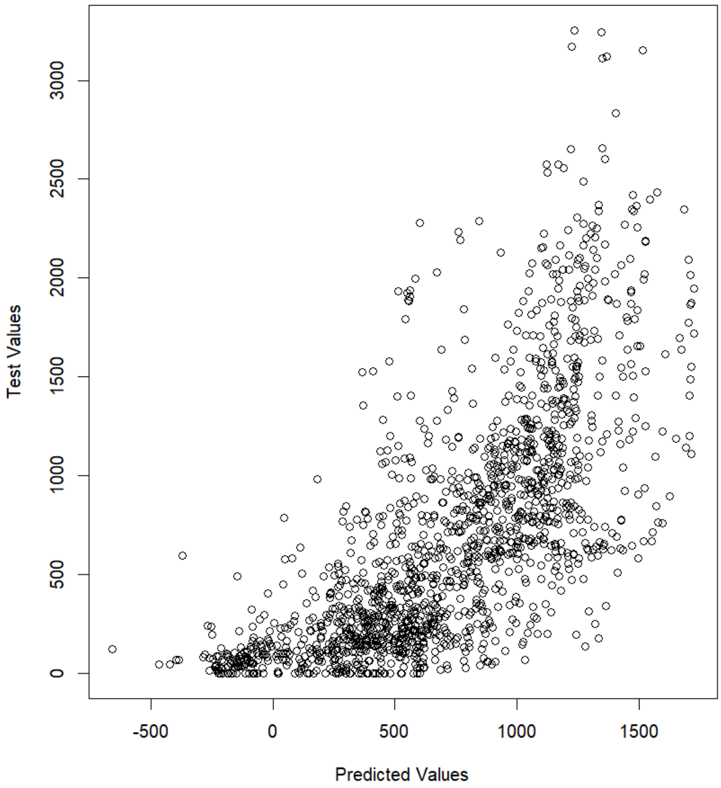
Out of all the variables in the calculation matrix Temperature and Dew Point Temperature have the highest correlation. Their correlation is 0.91. The following image shows the Temperature on the y-axis and the Dew Point Temperature on the x-axis:



The plot looks similar to a straight line. All days with a high temperature have a high dew point temperature and all days with a low temperature have a low dew point temperature and vice versa. Because of this, I would say removing one of these variables from the dataset will barely affect the prediction.

1.4 TASK 4

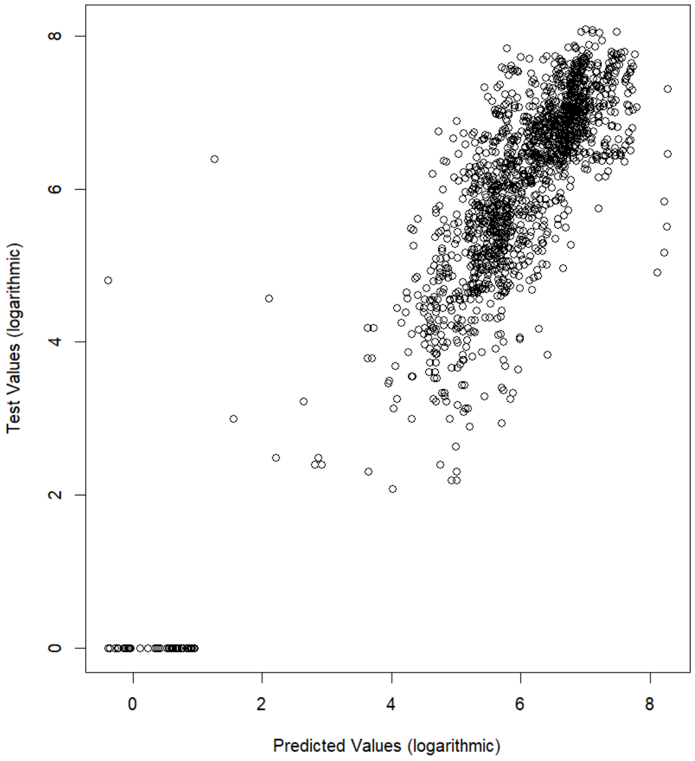
This plot shows the predicted values of the model and the actual values from the testing dataset:



The plot rather resembles an exponential graph instead of a linear graph. If the models predictions were accurate, the plot would look similar to y = x.

1.5 TASK 5

After the transformation using a logarithm, a new model was trained. This is the plot of its predictions:



In comparison to the previous plot, it looks much more like a linear graph and the x-values are similar to the y-values, which means that the second model predicts the results better then the first model. The R^2-values confirm this claim. The first model has an R^2-value of 0.55 and the second model has the R^2-value 0.79.

1.6 TASK 6

Parameters	**R^2-value**
Temperature and Rainfall	0.11
Temperature + Rainfall	0.18
Temperature - Rainfall	0.14
Rainfall - Temperature	0.04
Temperature*Rainfall	0.19

Adding categories increases the R^2-value and subtracting categories decreases it. The temperature has an higher impact than the Rainfall. The multiplication produces the most accurate prediction.

1.7 TASK 7

Transformation	R^2-value
None	0.793
Square of Temperature	0.799
Square root of Rainfall	0.817
Square root of humidity	0.783
Square of Dew Point Temperature	0.796

All the transformations slightly change the R^2-value but not significantly.

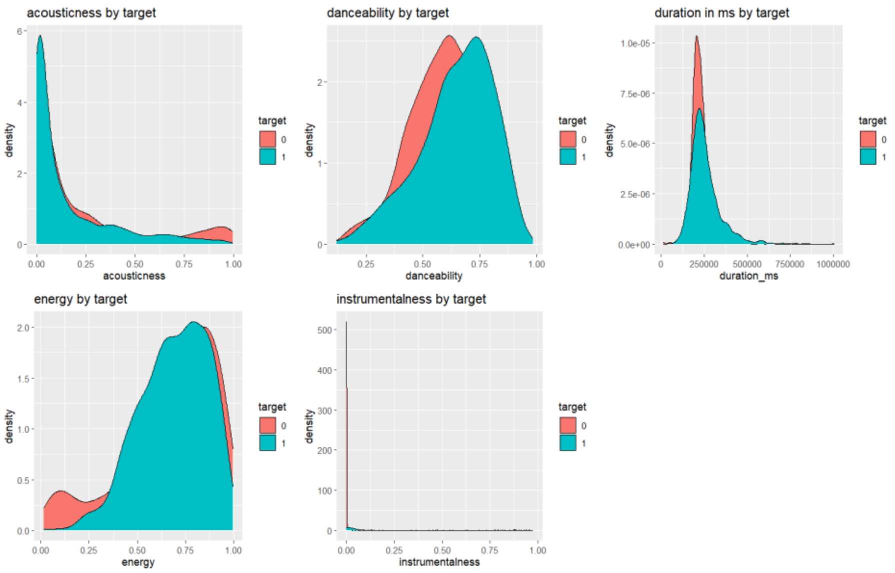
2 LAB TASK: 3.1 LOGISTIC REGRESSION

2.1. TASK 1

The dataset has 2017 observations of these 17 variables X (integer and the ID of the song), accousticness (numerical from 0 to 1), danceability(numerical from 0 to 1), duration_ms(integer), energy(numerical from 0 to 1), instrumentalness(numerical from 0 to 1), key(integer from 0 to 11), liveness(numerical from 0 to 1), loudness(numerical), mode(0 or 1), speechiness(numerical from 0 to 1), tempo (numerical), time_signature(integer), valence(numerical from 0 to 1), target(0 or 1), song_title(characters) and artist(characters).

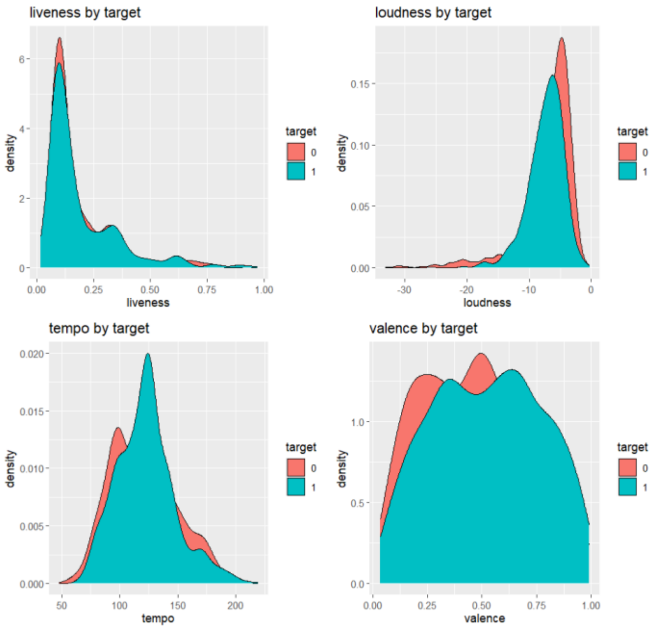
1.2.TASK 2

These are the desnsity plots for the variables:



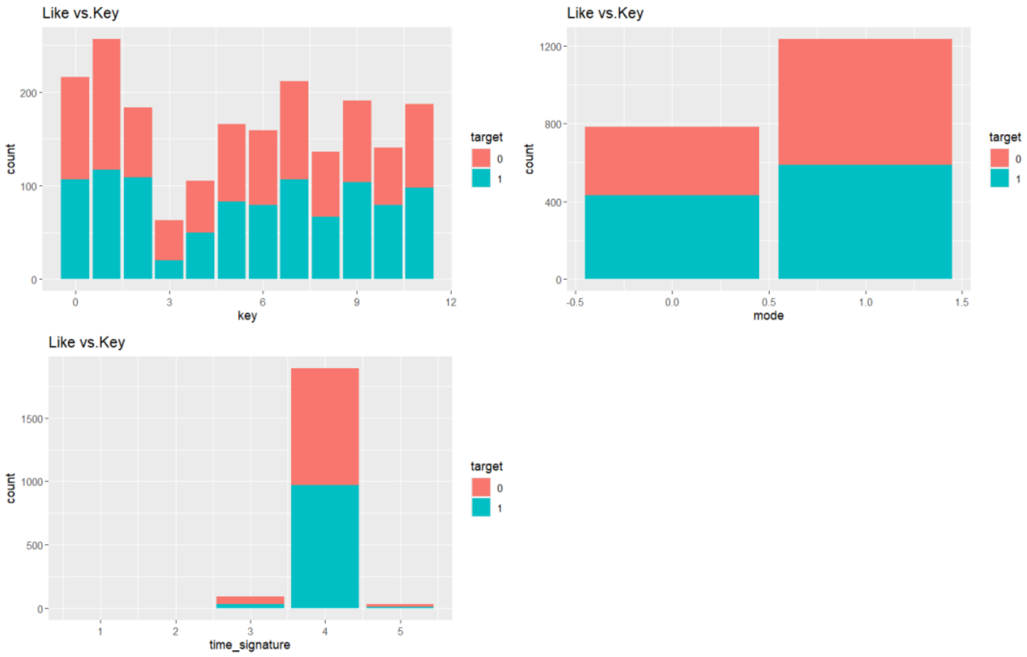
2.3. TASK 3

These are the density plots for the five other variables:



2.4. TASK 4

These are the bar graphs for the three variables:



2.5. TASK 5

The most important variables are the ones, whose graphs for target = 0 and target = 1 have significantly different shapes. Therefore the variables dancebility, energy, loudness, valence and mode have a larger impact on target then the other variables.

2.6. TASK 6

This is the confusion table for the model:

	True	False
Predicted as true	133	71
Predicted as False	65	135

Accuracy for class true: 67.17%

Accuracy for class false: 65.53%

Overall Accuracy: 66.34%

Logistic Regression Equation:

P(x)=$\left(1+e^{-\left(-4.1-1.54\cdot x_1+1.97\cdot x_2+3.25\cdot 10^{-6}\cdot x_3+0.11\cdot x_4+1.20\cdot x_5+0.01\cdot x_6+0.75\cdot x_7-0.1\cdot x_8-0.2\cdot x_9+4.1\cdot x_{10}+0.04\cdot x_{11}+0.02\cdot x_{12}+0.79\cdot x_{13}\right)}\right)^{-1}$

Variable in the equation	Variable in the dataset
X_1	Acousticness
X_2	Danceability
X_3	Duration_ms
X_4	Energy
X_5	Instrumentalness
X_6	Key
X_7	Liveness
X_8	Loudness
X_9	Mode
X_{10}	Speechiness
X_{11}	Tempo
X_{12}	Time_Signature
X_{13}	Valence

2.7. TASK 7

I selected the probability threshold values 0.4, 0.5 and 0.6 for this task. This table shows the accuracies for each:

Threshold Value	Accuracy True	Accuracy False	Accuracy Overall
0.4	81.27%	57.77%	69.66%
0.5	67.17%	65.53%	66.34%
0.6	45.59%	85.76%	65.44%

Spotify contains more songs than anyone could ever listen to. A low accuracy in the class True would lead to users missing out a song they like, but they would just listen to another song, because Spotify offers so many songs. On the other hand, the consequence of a low accuracy in the class False would be the user listening to songs they do not like. If the user listens to many songs they do not like, they will become annoyed and quit the app. In conclusion, a higher threshold value is better in this case, because it keeps the users from listening to songs they do not like.

2.8. TASK 8

For this task I trained a model only using the variables dancebility, energy, loudness, valence and mode, because these seem more important to me. The confusion table looks similar to the one in task 6, when I used all variables. I used threshold 0.5 again. This is the confusion table of the new model:

	True	False
Predicted as true	133	78
Predicted as False	72	121

Accuracy for class true: 64.87%

Accuracy for class false: 60.80%

Overall Accuracy: 62.87%

The Accuracy decreased for all categories. Reducing the amount of variables always decreases the accuracy. It is possible to achieve a similar accuracy using only important variables, but it is impossible to reach the same or even higher accuracy.

2.9. TASK 9

I only transform the variables I considered important in the previous tasks, because the other variables will not have significant effects. I will only transform numerical variables.

Results using threshold 0.3:

Transformation	accuracy True	accuracy False	accuracy overall
No Transformation	91.27%	16.65%	54.39%
$(\text{valence})^2$	91.47%	16.65%	54.49%
log(-Loudness)	90.29%	24.07%	57.56%
Sqrt(Energy)	91.67%	18.36%	55.43%
log(Danceability)	91.76%	16.65%	54.64%

Results using threshold 0.5:

Transformation	accuracy True	accuracy False	accuracy overall
No Transformation	67.17%	65.53%	66.34%
$(\text{valence})^2$	65.39%	68.41%	66.88%
log(-Loudness)	67.94%	71.31%	69.61%
Sqrt(Energy)	66.57%	68.61%	67.58%
log(Danceability)	64.80%	68.20%	66.48%

Results using threshold 0.7:

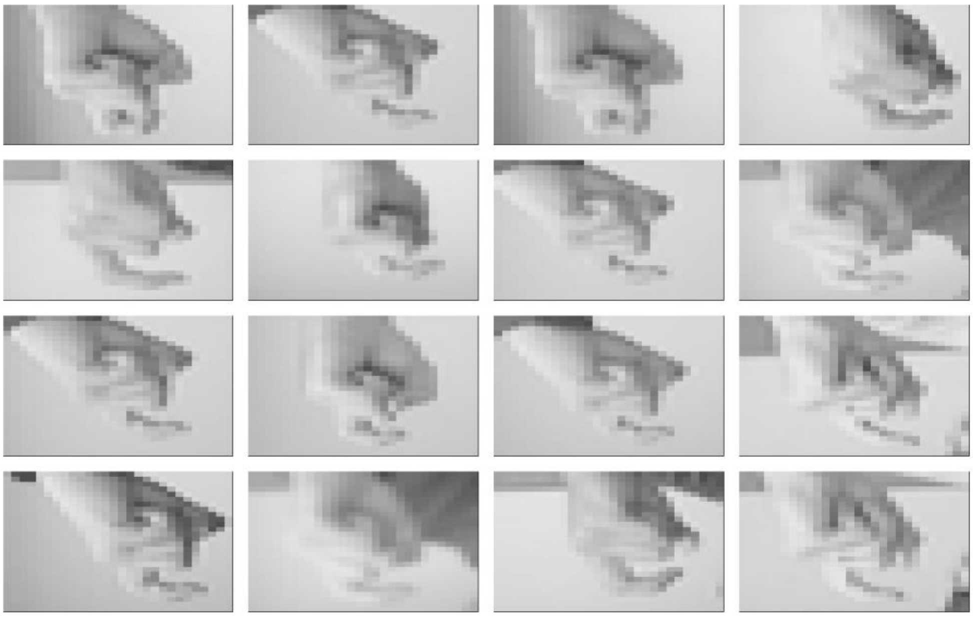
Transformation	accuracy True	accuracy False	accuracy overall
No Transformation	26.27%	94.48%	59.99%
$(\text{valence})^2$	25.98%	93.78%	59.49%
log(-Loudness)	32.84%	93.58%	62.87%
Sqrt(Energy)	28.33%	93.38%	60.49%
log(Danceability)	25.88%	94.78%	59.94%

The transformations change the accuracies of the different classes and the overall accuracy. Using multiple transformations at the same time, which increase one of the accuracies, will lead to a more accurate model for this category. Transformations can be used to make more accurate models.

3 LAB TASK: 4.1 LDA AND QDA CLASSIFIERS

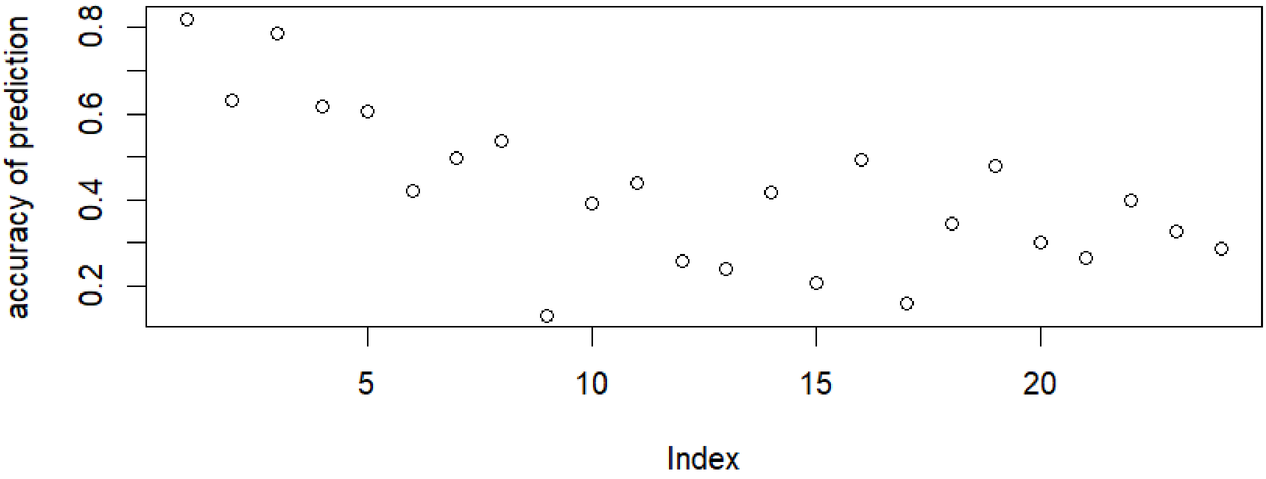
3.1 TASK 1

I chose the number 23 for this excercise. These 16 show the letter X in sign language:



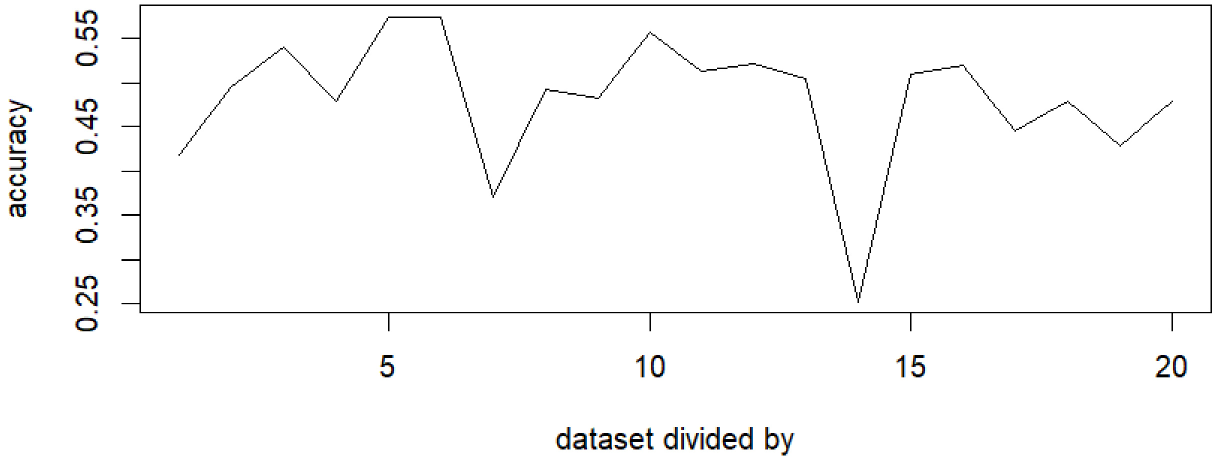
3.2 TASK 2

In this plot, I visualized the accuracy of the LDA for the different classes in the dataset:

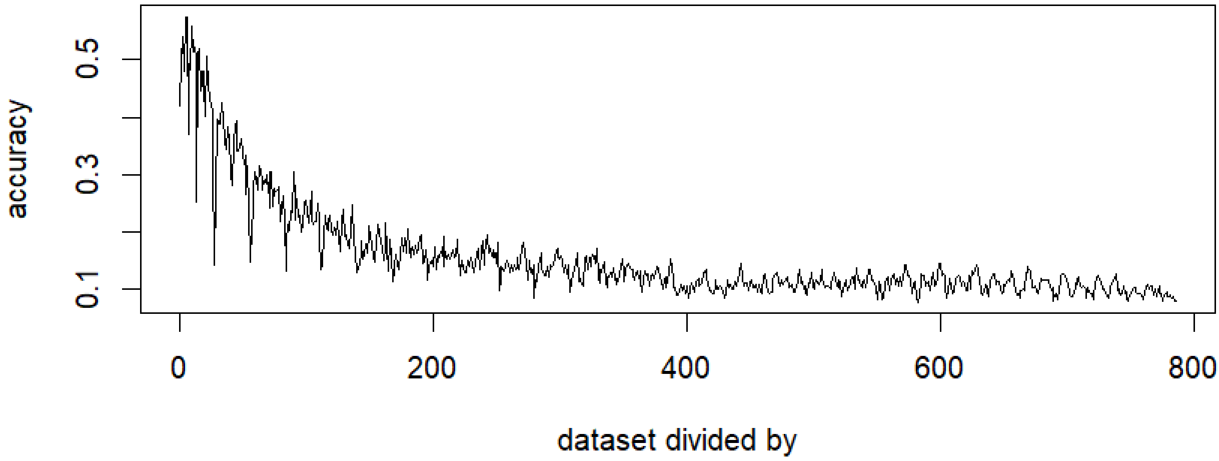


3.3 TASK 3

The overall accuracy of the LDA for the whole data is 0.42. After only using half of the features in the training data, the accuracy increased to 0.49. Reducing the number of features further, increased the accuracy. The following graph shows the accuracy for the values from 1 to 20.

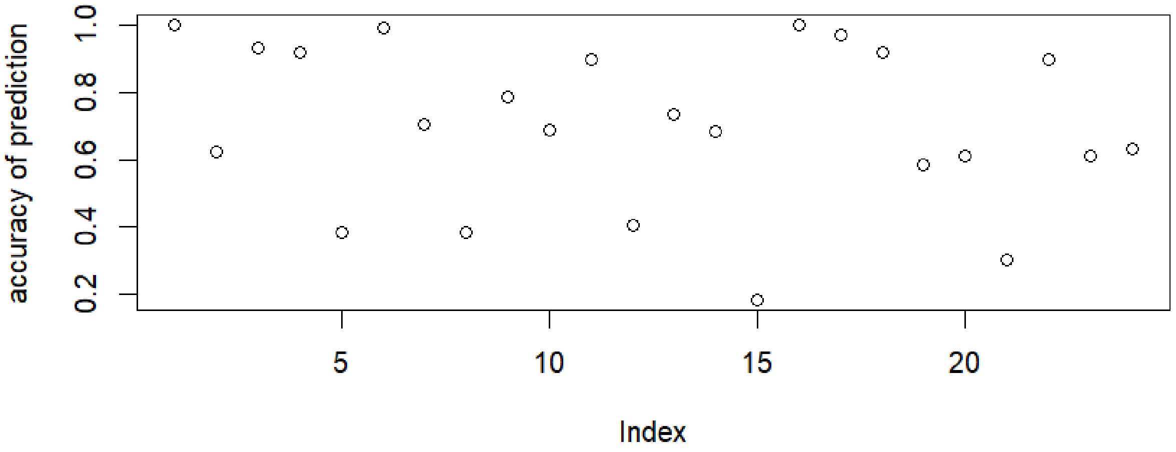


The highest accuracy 0.58 is achieved by using a fifth of the available features. Using less than a fifths of features leads to lower accuracies. Using less variables increases the accuracy, because it filters some of the mistakes by the camera. Sometimes there is a very bright pixel among dark pixels or vice versa. These pixels are often ignored when only using a fraction of the available pixels. Using not enough features leads to losing important information for the model. The following graph includes all the possible reductions of the dataset showing that the accuracy drastically decreases for large reductions:



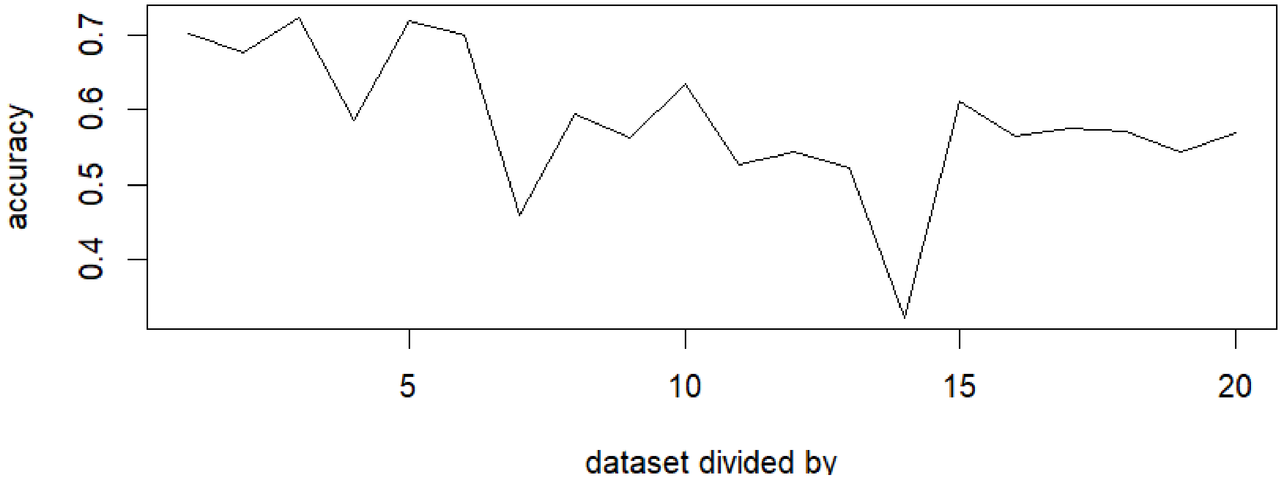
3.4 TASK 4

In this plot, I visualized the accuracy of the QDA for the different classes in the dataset:



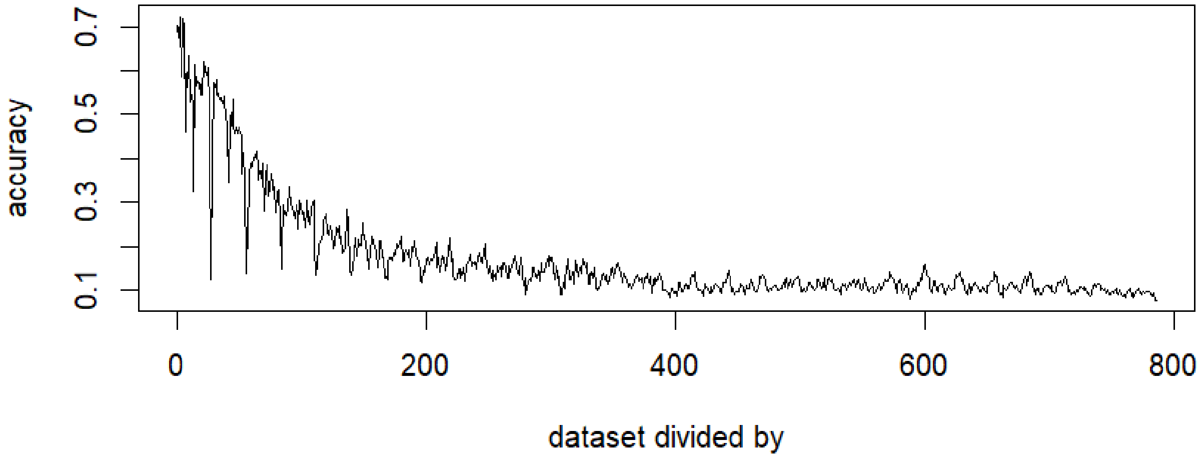
3.5 TASK 5

The overall accuracy of the QDA for the whole data is 0.70. After only using half of the features in the training data, the accuracy decreased to 0.68. Reducing the number of features further, increased the accuracy. The following graph shows the accuracy for the values from 1 to 20.



The highest accuracy 0.72 is achieved by using a third of the available features. Using less than a third of features leads to lower accuracies.This has the same reason as the increase and decrease described in task 3.

The following graph includes all the possible reductions of the dataset showing that the accuracy drastically decreases for large reductions:



For low or no reduction, the QDA is more accurate than the LDA. For large reductions the LDA is more accurate.