TDDE01. Lab1. Group B24 report.

# **Statement of contribution**

Firstly, general analysis was performed teamwise. Approaches and strategies of solving the tasks were elaborated teamwise as well.

Student Elham was responsible for code writing and problem solving for the Assignment 1. Student Anton was responsible for code writing and problem solving for the Assignment 2. Student Elena was responsible for code writing and problem solving for the Assignment 3.

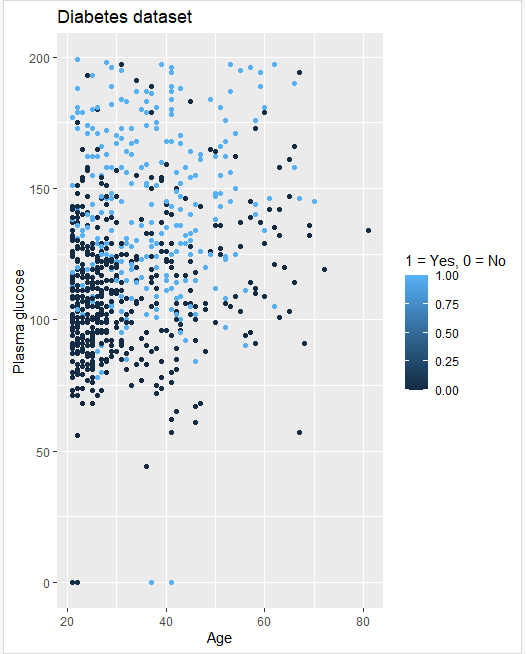
After completion of coding stage group analyzed the results together and peer reviewed the results of each other’s work. Finally, each student corrected their solutions according to the received reviews from groupmates.

# **Assignment 1. Handwritten digit recognition with Knearest neighbors.**

# **Assignment 2. Linear regression and ridge regression**

# **Assignment 3. Logistic regression and basis function expansion**

During this assignment raw data from "pima-indians-diabetes.csv" file was loaded and this full dataset was used in further analysis. The initial dataset was put on the plot where x axis is patients age and y axis is glucose levels in plasma. Data was colored by actual diabetes levels.



We assume that diabetes is relatively easy to classify, as there are only two possible outputs (whether a person has diabetes or not) and there is a certain interconnection between actually having diabetes and a pair of parameters (Age+Blood sugar levels).

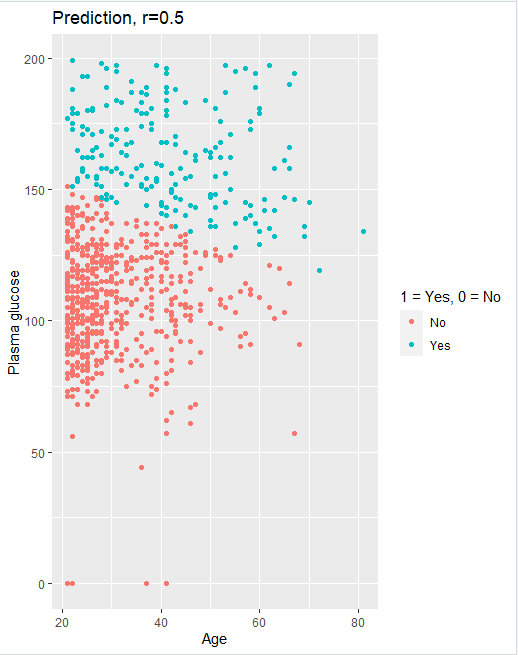
A logistic regression model was trained with the respect to 𝑥1 =Plasma glucose concentration and 𝑥2 =Age as features and 𝑦 =Diabetes as target. Prediction was made for the whole dataset by using r=0.5 as the classification threshold.

The probabilistic equation of the estimated model looks like

p(y|X) = 1 / (1+exp(-(-5.91+0.035x1+0.024x2))).

The training misclassification error computed MSE = 58.29971.

Acquired prediction was put on the plot:



Furthermore, the confusion matrix was calculated:

|  |  |  |
| --- | --- | --- |
| Prediction/True data | 0 | 1 |
| No | TN = 436 | FN = 138 |
| Yes | FT = 64 | TP = 130 |

As can be seen from the data and the graph above, the quality is pretty mediocre because as it is seen on the plot, there is no concrete boundary between two domains of decisions so there is a high amount of errors (more than 25% errors, i.e., out of 768 objects 202 were false positive or false negative).

The equation of the decision boundary between the two classes looks like:

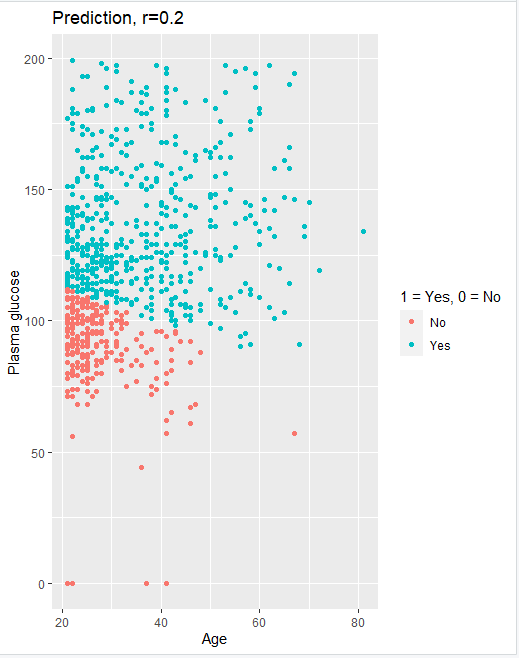
-5.91+0.035\*x1+0.024\*x2 = 0

and it can be seen at the graph below:



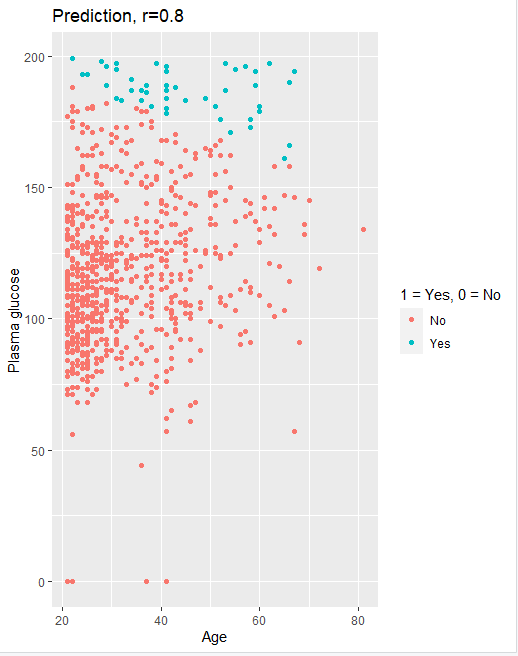
Looking at the plot we cannot say that the decision boundary catches the data distribution well which is approved by the confusion matrix.

Further on we looked at prediction using 𝑟 = 0.2:



It can be seen both from the graph and confusion matrix that with r=0.2 the number of FALSE POSITIVEs increased significantly. There are more than 37% errors, i.e. out of 768 objects 286 were false positive or false negative. Hence low threshold is not the best choice.

Then we looked at prediction using 𝑟 = 0.8:



Although the number of errors slightly decreased, it is still high. There are approx. 32% errors, i.e., out of 768 objects 242 were false positive or false negative. This time there are more FALSE NEGATIVES (which probably is worse than in previous case). High threshold is even worse in this case because it misleads in a detrimental way.

After that we computed new features 𝑧1 = 𝑥1^4, 𝑧2 = 𝑥1^3\*𝑥2, 𝑧3 = 𝑥1^2\*𝑥2^2, 𝑧4 = 𝑥1\*𝑥2^3, 𝑧5 = 𝑥2^ 4 (basis function expansion). New features were added to the dataset and a new logistic regression model was calculated with features {x1, x2, z1, z1, z3, z4, z5}.

Confusion matrix was calculated:

|  |  |  |
| --- | --- | --- |
| Prediction/True data | 0 | 1 |
| No | TN = 433 | FN = 121 |
| Yes | FT = 67 | TP = 147 |

And the graph was plotted:



a

Out of 768 objects 188 were false positive or false negative which makes 24% errors. Also, there can be seen that the number of false negatives decreased a bit. From all the iterations we would say the model after BFE + r=0.5 gave the best results because it gave a reduction in the number of errors and the number of false negatives was the least which is crucial in the public health domain.

# **Appendix. Code.**

