Q10-Test Analysis

An analysis of the Q10-Test. Is it as good as it seems? And could it be even better..

Wouter Zeevat

BFV3

11-11-2022

Bart Barnard & Dave Langers

Q10-Test Analysis

Wouter Zeevat

Bio-informatica

ILST

Bart Barnard & Dave Langers

11-11-2022

Samenvatting

In order to figure out if the Q10-test is trustworthy, a Weka model was created. This model does not only look at the end score. It also looks at the answers themselves. This is because some questions may be more relevant than others. A lot of Weka classifiers were tested, and an end option was chosen. The Data was manipulated so that it was now balanced. This was done in order to make the model have less false negatives. Which is better because if someone's result is positive (Saying they have autism), they could still go to a doctor to later figure out they do not have ASD. Then to have a false negative. To later never figure out they had ASD all along. The model was implemented in a java wrapper and can now be used to conclude the results of anyone's Q10-test. In conclusion the Q10-test's previous outcome had some flaws but are fixed by balancing the data and looking at more than just the end score. The wrapper is published here: https://github.com/wouterzeevat/thema09/tree/main/Software

Contents

Introduction

The AQ-10 test is a test with 10 questions that will try and predict if you have the autism spectrum disorder. Each question has 4 answers. Slightly agree, agree, slightly disagree and disagree. Every question gives either 1 or 0 points based on the answer. Slightly agree gives the same points as agree and the same for disagree. They made it like this to make it feel like slightly agree/disagree feels like a better answer when in doubt. The more points the people have, the more chance there is of them having ADS according to the test (autism research quotient, 2012). This project uses a dataset that consists of 700 people who made this test (https://www.kaggle.com/datasets/faizunnabi/autism-screening). how accurate can the AQ-10 test predict whether someone has the autism spectrum disorder? The goal of the project is to find this out and create a model that will predict this more accurate than the test itself. The model will look not only look at the end score, but to correlations as well. These correlations will be looked at in R. After finding some significant correlations a model will be made. This will be done by using a Weka model (Machine Learning at Waikato University, n.d.). Which will later be in a java wrapper for other people to use it. Now we will look at those correlations.

Results

First let's look at the correlation of people having ASD and their end scores. This is important to know to see how accurate the test itself is. If the end score does not correlate at all the test would have no point.

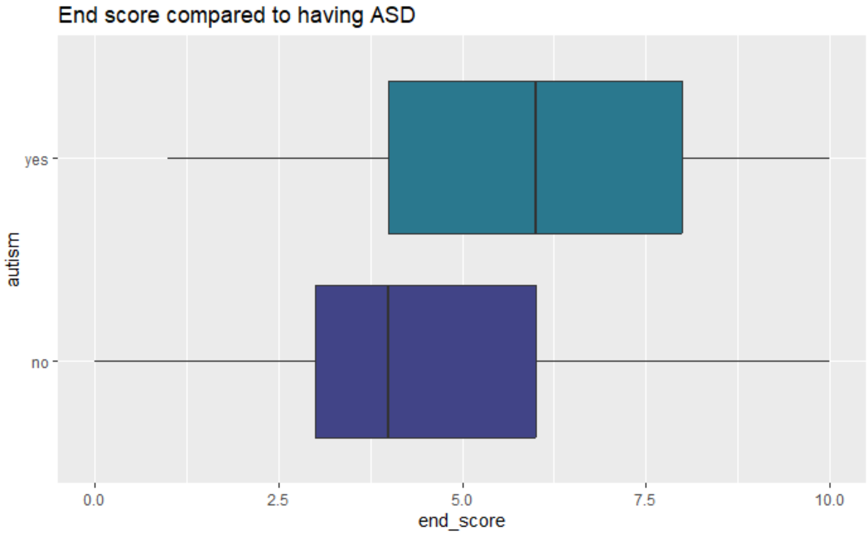
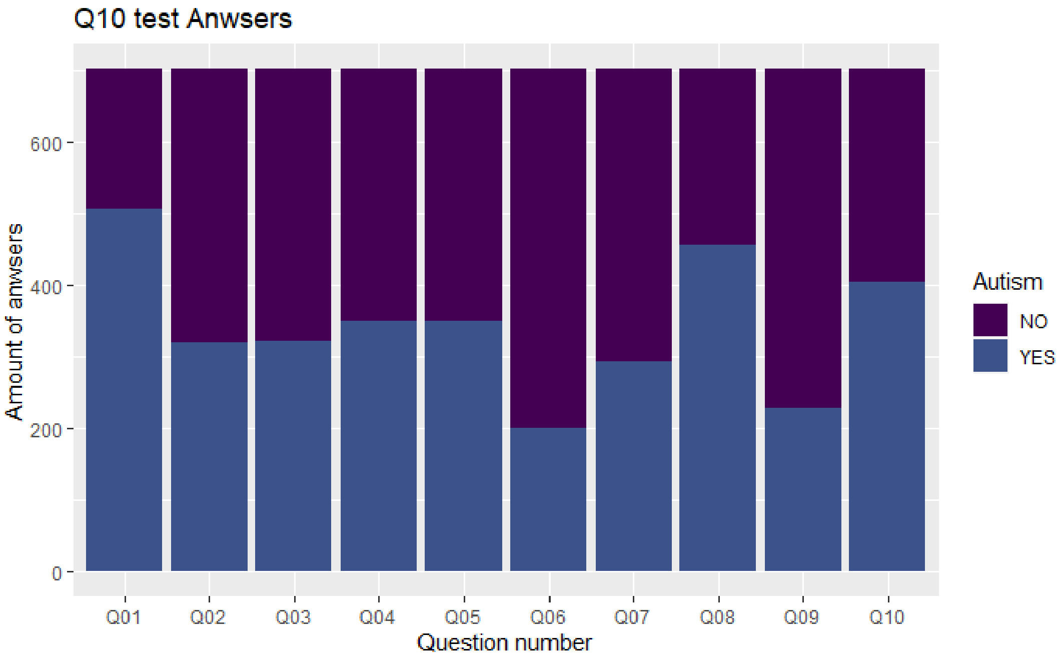


Figure 1, The end scores of people with ASD are significantly higher than the other people. This is confirmed by doing a T test that has an outcome of p-value < 2.2e-16. This is a good thing. Because the Q10-test is doing it's work. Now the other correlations can be looked at to find out if there's more things the model could look at instead of the end score. Things like gender and ethnicity can not be used because those would make the model too overfitting due to the big amount of options.

Something that could be looked at is the questions itself though. Some questions could be more relevant than others. The model could use this and give certain questions a higher weight. This will be checked in the following plot.

Figure 2, As excepted some questions are more relevant than others. This is because apparently people with ASD have the answers to those questions in common. This is a huge thing for the model. This means it can implement this by attaching weight to the questions itself and use it to predict the test outcome.



A little concern is the influence of age on the test. A possible correlation could be age. Older people are sometimes more likely to have certain autism symptoms. Therefore, this correlation will be looked at.

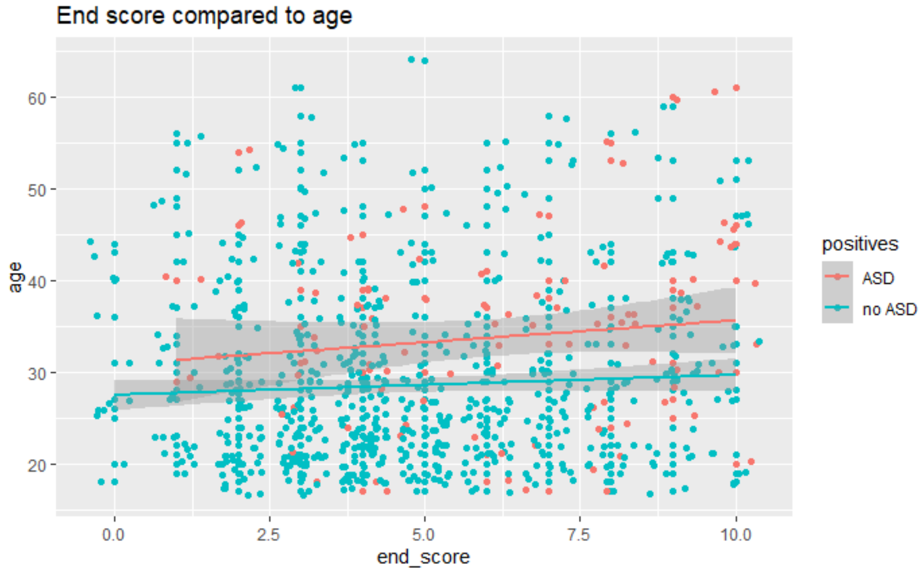
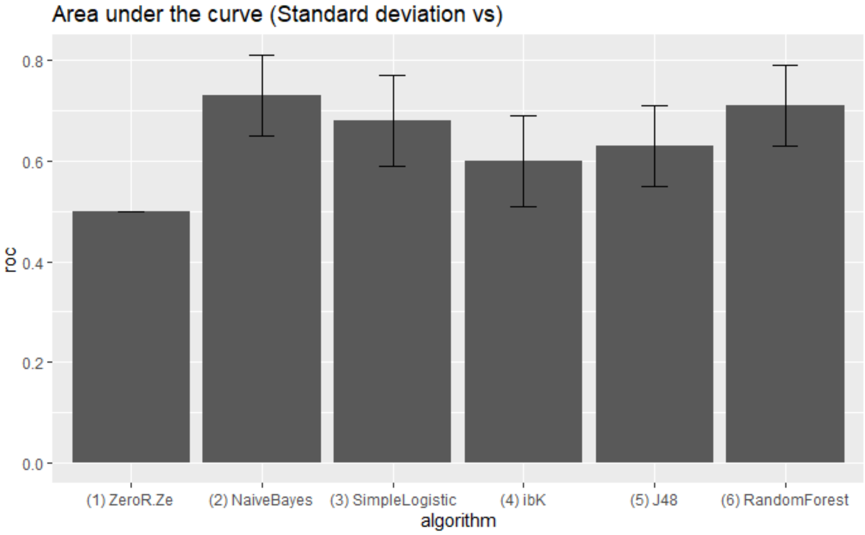


Figure 3, The correlation of the end score compared to age does not seem to have a correlation. This was tested with a correlation test with values p-value = 0.009571 concluding that it indeed does not have a correlation. This means that the age does not have influence on the test. Which is good because this is no longer something that needs to be taken in account for. After all of this is done a correct model can be chosen. Before doing this there need to be some analysis's done on multiple models. To find the right one. The first one that will be done is the accuracy. The following figure will show how much percentage of the cases it will classify correctly.

Figure 4, This figure shows that SimpleLogistic (Simple logistic regression, n.d.) has the highest accuracy. This kind of makes sense because the test itself uses logic to calculate the result too. This classifier would make most sense to use. But before picking one too carelessly, the area under the curve will be looked at as well. There is still an opportunity that SimpleLogistics' ROC is way lower than other ones. Which would make it less qualifiable to be the model used in the wrapper.



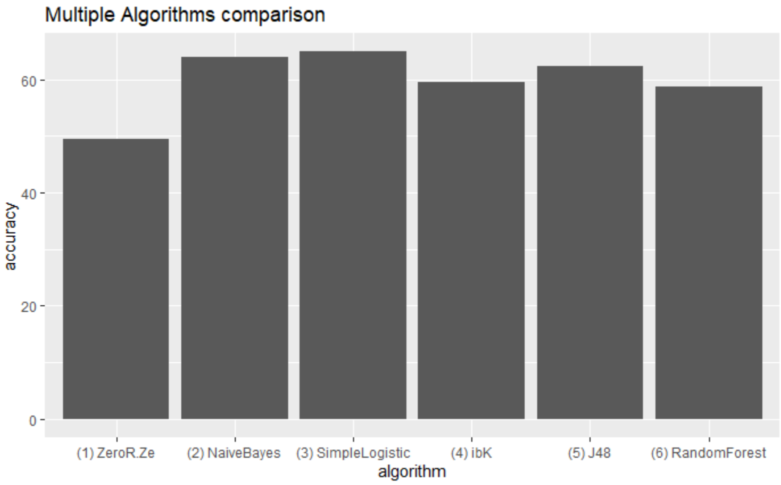


Figure 5, These are all the ROC values together with their error rates. SimpleLogstics is not significantly lower than the highest one (NaiveBayes). It could even be as high or higher than NaiveBayes with the error rates considered. This means that the SimpleLogistics classifier will be used. The data is also balanced because it really wasn't balanced before. The values were changed from 650-50 to 350-350. This reduces the accuracy but lowers the false negatives.

Conclusion & Discussion

All the results conclude that the data is indeed useful for a machine learning algorithm. The plots show that the important variables have a correlation of people having ASD. The first plot shows that the Q10-test indeed actually works which is a good foundation to build on. The second plot marks the questions that are heavily answered with yes by people with ASD. This is a good variable that the classifier can use to predict the result. The third plot shows the end score compared to age. Which didn't have a correlation, meaning it won't be a problem in the classifier. The fourth plot tested the accuracy of multiple classifiers. The outcome of this was that SimpleLogistics was the way to go. The last plot tested the ROC. Which didn't have SimpleLogistics as the best option. But it's not significantly worse when having taken the error rates into account.

The goal was to find out how accurate the Q10-test can predict if someone has ASD. The conclusion of this is that it can never fully know it. But it can predict it sometimes. The weaknesses of the test are that the answers could be based on other things than ASD. But the model indeed is better than the plain test. This is because it takes more variables into account and the data is balanced. Which makes false negatives way less likely to happen. Meaning that any company or doctor who uses this test can use this model instead of the plain test. The model is published on Github. Simple follow the instructions in the readme.MD. https://github.com/wouterzeevat/thema09/tree/main/Software

A great follow-up to this project could be finding out if the questions are good. This could be done by making a way bigger test and finding out what questions are the best. This would have a great influence on the test making the accuracy even higher.

References

autism research quotient. (2012). Retrieved from autism research centre: https://www.autismresearchcentre.com/tests/autism-spectrum-quotient-10-items-aq-10-adult/

Machine Learning at Waikato University. (n.d.). Retrieved from Waikato: https://www.cs.waikato.ac.nz/ml/index.html

Simple logistic regression. (n.d.). Retrieved from Handbook of biological statistics: http://www.biostathandbook.com/simplelogistic.html