ASD Cause Prediction

Chang Cao
Matrikelnummer 6405371
chang.cao@student.uni-tuebingen.de

Nianzi Yi
Matrikelnummer 6280976
nianzi.yi@student.uni-tuebingen.de

Zhoutao Zhang
Matrikelnummer 6455587
zhoutao.zhang@student.uni-tuebingen.de

Abstract

Autism spectrum disorder (ASD) is a neurodevelopmental-related mental disorder. With the improvement and application of machine learning techniques in classification problems, the problem of maintaining a high accuracy of ASD diagnosis could be solved through that. However, studies that focus on using machine learning methods to diagnose ASD suffer from data issues, such as trust in using models and feature selection. This present project examined 298 children and 108 adolescents samples from two datasets by using the Random Forest Algorithm. The two models we obtained both have 100% accuracy and the "result" feature contributes the most. The models with either only "result" or only "A1-A10" show strong predictive validity, whereas the model with only features unrelated to AQ-10 underperformed.[1]

1 Introduction

Autism Spectrum Disorder (ASD) is a psychiatric disorder with communication deficit, social deficit, and repetitive stereotyped behavior [Bolton et al., 1994]. Besides the clinical criteria, several non-clinical facts could also affect ASD, e.g. gender and heredity. To diagnose correctly, clinicians must have exceptional expertise and extensive clinical experience. However, the diagnosis difficulty, as a typical classification problem (ASD, Non-ASD), could be overcome by effective machine learning methods, such as random forest, support vector machine, and logistic regression, which are now commonly used to increase the efficiency of screening and optimize the related features, ultimately reduce the rate of misdiagnosis and alleviate severe stress on the healthcare system. Therefore, the primary aim of this project is to use Random Forest (RF) algorithm to make predictions based on two datasets. And we also focus on using feature ranking to find which features contribute to the prediction most and test if fewer features are sufficient to predict ASD.

2 Materials and Methods

This project considers two datasets from the UCI machine learning repository which are "Autistic Spectrum Disorder Screening Data for Children" and "Autistic Spectrum Disorder Screening Data for Adolescent"[Thabtah, 2017a,b] that both contain the scores of "A1-10", and 10 individual characteristics: "age", "gender", "ethnicity", "jaundice" (if born with jaundice), "austim" (family member with PDD, "country_of_residence", "used_app_ before", "result" (total score of AQ-10 test), "age_description", "relation" (who is completing the test). The children's dataset comprised 292 participants (mean age: 6.36 $\pm$ 2.37 *SD*; male: 208, female: 84; Non-ASD: 151, ASD: 141), while

[1] Git repository: https://github.com/NianziYi/ASD-Prediction

the adolescents' dataset contained 108 participants (mean age: 14.13 $\pm$ 1.58 *SD*; male: 54, female: 50; Non-ASD: 41, ASD: 63). Figure 1 shows that all data showed a great combination except for a slight imbalance between male and female children participants.

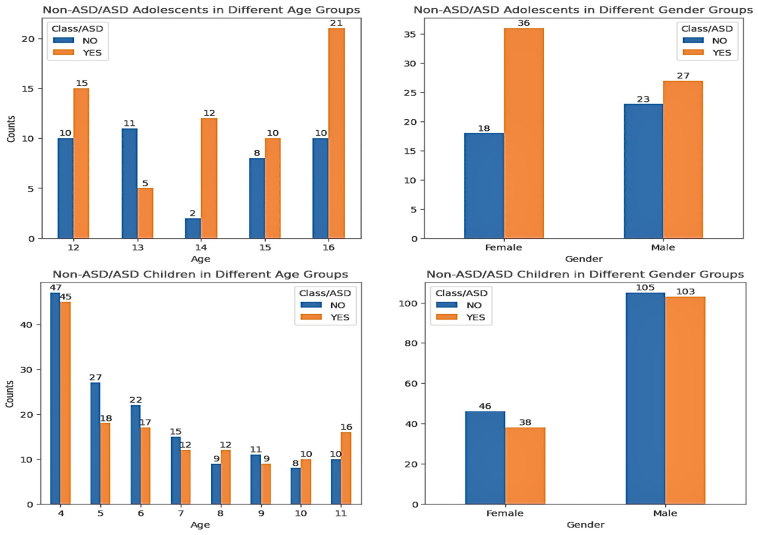
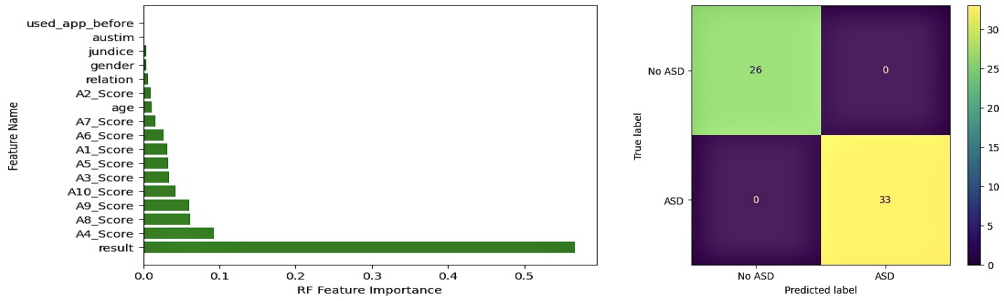


Figure 1: Illustration of ASD and Non-ASD in Age and Gender of Children and Adolescents Datasets

Before training our prediction model, we removed 3 features that were irrelevant for the prediction purpose in this project: "ethnicity", "country of residence", and "age_description". The one-hot encoding technique was used to convert all variables to numerical types. In terms of processing missing values, because the AQ-10 questionnaire is conducted with parents by clinicians, only in this experiment it was changed to parent and caregiver to avoid some parents spending less time with their children and cannot observe their children's daily behavior well, nevertheless according to our data, in most cases, the questions were still answered by the parents. So we imputed missing values in the "relationship" column with "parent", and in the "age" column with the mean age. Since the class label is binary, RF Classifier Model was chosen to classify samples. 20% of data from the two pre-processed datasets were used for testing while the rest 80% was used for training. The number of estimators was set to 500 in our RF model. And the prediction result was analyzed by using statistical tools like accuracy, F1 score, and confusion matrix. At the same time, to find which feature contributes to the model's prediction most, the built-in feature ranking tool of the RF model was used to rank the features.



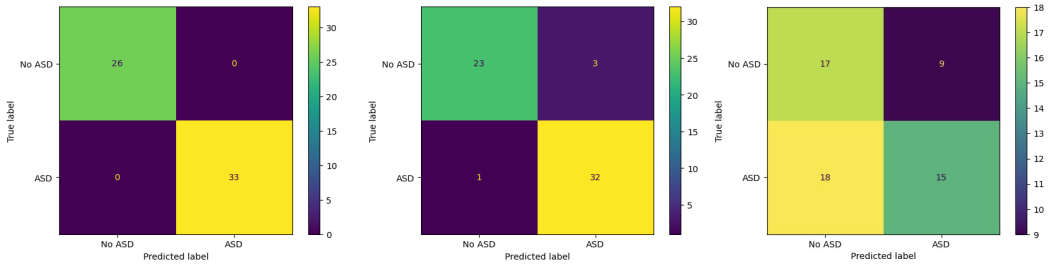
(a) (b)

Figure 2: Model Data based on Children Dataset: (a) Feature Ranking; (b) Confusion Matrix

3 Result

3.1 Children Dataset

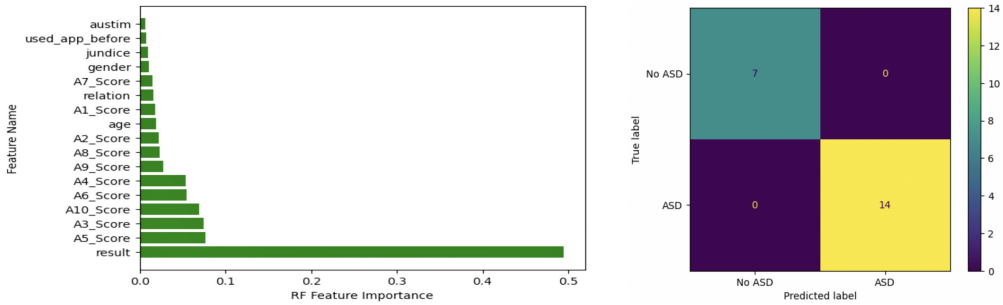
After we tested the trained model on the children's test set, the confusion matrix shows that, out of 59 samples in the test set, the model classified 26 Non-ASD samples and 33 ASD samples correctly,



(a) (b) (c)

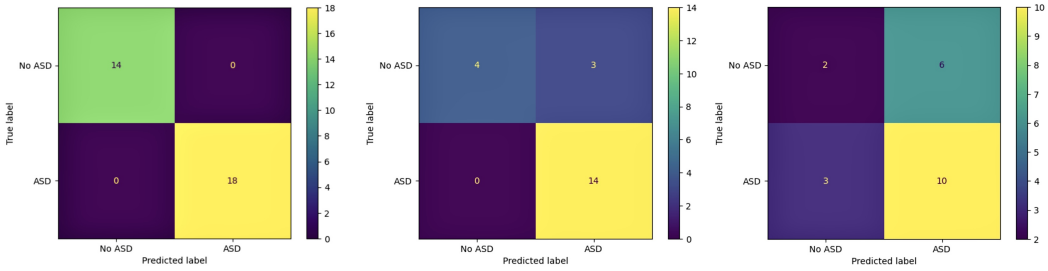
Figure 3: The Confusion Matrices of Limited Features of Children Dataset: (a) Only Using "Result"; (b) Only Using "A1-A10"; (c) Only Using Features Unrelated to AQ-10

which means both the accuracy and the F1 score are 1. Our trained RF model did a perfect job of predicting if a children's sample has ASD or not. Figure 2 shows that from the feature ranking results, the AQ-10 questionnaire helped the model classify the samples mostly, especially the "result" feature, which is the total score of A1 to A10 questions, contributed most to the model's prediction. However, the "used_app_before" feature was the least important feature among the 17 features which are used to train. To find if we could use fewer features to do the prediction, we trained 3 RF models based on 3 different subsets of features. As illustrated in Figure 3, the models still could achieve 100% accuracy using only the "result" feature to train (a). The accuracy would reduce to 93.2% if the model was trained based on the A1 to A10 scores (b). Also, the accuracy would reduce to 54.2% when using only the 6 features that are unrelated to the AQ-10 questionnaire (c). In this case, the model classified 9 Non-ASD samples into ASD samples and 18 ASD samples into Non-ASD samples.



(a) (b)

Figure 4: Model Data based on Adolescents Dataset: (a) Feature Ranking; (b) Confusion Matrix



(a) (b) (c)

Figure 5: The Confusion Matrices of Limited Features of Adolescents Dataset: (a) Only Using "Result"; (b) Only Using "A1-A10"; (c) Only Using Features Unrelated to AQ-10

3.2 Adolescents Dataset

On the adolescents' dataset, the RF model achieved 100% accuracy and a F1-Score of 1.0 on the test set. As illustrated in the Figure 4, the model classified all 9 Non-ASD samples and 12 ASD samples

correctly. This test and an extra sanity check we performed show that the model can predict ASD diagnosis based on all the features given by the data set mentioned above with very high accuracy. The "result" of the AQ-10 questionnaire contributes the most to the classification. Following by the answers A1 to A10 of the questionnaire. Features like "austim", "used_app_before" ranked last among all features, which suggests they are not strongly related to ASD. To make sure the importance of features given by the model is accurate, we trained another 3 RF models on 3 subsets of features. The results are shown in Figure 5. With only the "result", the model still achieved 100% accuracy (a). The model trained on answers "A1" to "A10" got only 85.7% accuracy (b). When being trained on features unrelated to AQ-10 questionnaire, accuracy reduced to 57.1% (c). Both ASD and Non-ASD samples were classified incorrectly by this model.

4 Discussion

This project presents two RF models, both of which achieved 100% accuracy, i.e. the accuracy of the chosen features in diagnosing ASD has been demonstrated. However, many co-occurring psychiatric disorders have similar symptoms and are related to ASD, such as affective and anxiety disorders, and hyperactivity and conduct disorders, so a given positive result may be influenced by ASD, a co-occurring disorder, or a combination of both. There is a lack of such data in databases, so the model will likely show high accuracy but still make an incorrect diagnosis if the patient's symptoms of autism are mild. It would be preferable to add features of whether the child or adolescent has an associated psychiatric disorder to the database, and future studies could also try to modify the model by analyzing the differences between ASD and these co-occurring psychiatric disorders to further improve accuracy and increase the reliability of the model.

In order to reduce the number of features without compromising the accuracy of the model, feature ranking was also tested. In both databases, we found that the "result" feature contributed most to the prediction. Therefore, this project also attempted to use some limited features to test how well the models will work. We found that when only the "result" feature was used for prediction, two models could still maintain 100% accuracy. This shows similarity with previous studies[Booth et al., 2013], indicating that the AQ-10 questionnaire is well-designed and valid for overall autism prediction.

Since the "result" feature is crucial for prediction, and it is a simple summation of the answers to "A1-A10", we also tried to build and test the model with A1-A10 alone. Interestingly, the predictive validity of both logically should be the same, but the accuracy of the child model slightly decreased to 93.2% and the adolescent model dropped to 85.7%. The reason could be due to the ensemble learning approach that the RF algorithm uses to classify the question. Unlike if we just use the "result" feature to make a prediction, if we take "A1-A10" as features, the RF algorithm will treat them as separate events and try to find the relationship between them, then use it to make a prediction. But all we need to do is to add up 10 responses. For this reason, if the ASD prediction model is to be used in clinical diagnosis, it is necessary to calculate the total score of the questionnaire at first, i.e. at least include the total score in the feature, rather than only throwing the answers of items from the AQ-10 questionnaire into the model, which will make the model less accurate. Alternatively, models like support vector machine and neural networks may perform better since they will sum up weighted inputs, which enables them to treat "A1-A10" the same as "result".

We also tested the model's performance based on the other features unrelated to the AQ-10 questionnaire in predicting ASD and obtained an accuracy of 54.2% for the child model and 57.1% for the adolescent model. Since the class labels are binary, if we let the model randomly classify each sample into the "Non-ASD" class and the "ASD" class, the randomly predicted results would be binomially distributed and the model would get an accuracy of 50%. This value is very close to the values of the accuracy of the child model and the adolescent model. The model actually learned nothing when it was only given the 6 features unrelated to the AQ-10 questionnaire, which means it is not able to predict ASD only based on these features.

To conclude, the performance of RF learning model shows that it could be an effective tool in diagnosing ASD. Nevertheless, through our analysis of the features, more consideration needs to be given to the selection, addition, and deletion of features on the basis of guaranteeing the validity of the model.

References

P. Bolton, H. Macdonald, A. Pickles, P. Rios, S. Goode, M. Crowson, A. Bailey, and M. Rutter. A case-control family history study of autism. *Journal of child Psychology and Psychiatry*, 35(5): 877–900, 1994. doi: https://doi.org/10.1111/j.1469-7610.1994.tb02300.x.

T. Booth, A. L. Murray, K. McKenzie, R. Kuenssberg, M. O'Donnell, and H. Burnett. Brief report: An evaluation of the aq-10 as a brief screening instrument for asd in adults. *J Autism Dev Disord*, 43, 2013. doi: https://doi.org/10.1007/s10803-013-1844-5.

F. F. Thabtah. Autistic spectrum disorder screening data for adolescent data set, 2017a. URL https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Adolescent+++.

F. F. Thabtah. Autistic spectrum disorder screening data for children data set, 2017b. URL https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Children++.