Predicting Autism Capstone

Background

Data comes from https://www.kaggle.com/competitions/autismdiagnosis/data which was a kaggle competition in 2022 with data that was given by the following study: Tabtah, F. (2017). Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment. Proceedings of the 1st International Conference on Medical and Health Informatics 2017, pp.1-6. Taichung City, Taiwan, ACM.

Why would you need to predict autism?

The purpose of this model would be possibly for psychological use where a clinician could have the patient input their information into this model to see if they should go get a formal diagnosis of ASD to help them make a decision using other people who are already diagnosed.

Children or teenagers who are diagnosed can then receive the special help they need as a student to help them know how to succeed in school if that is a difficulty for them.

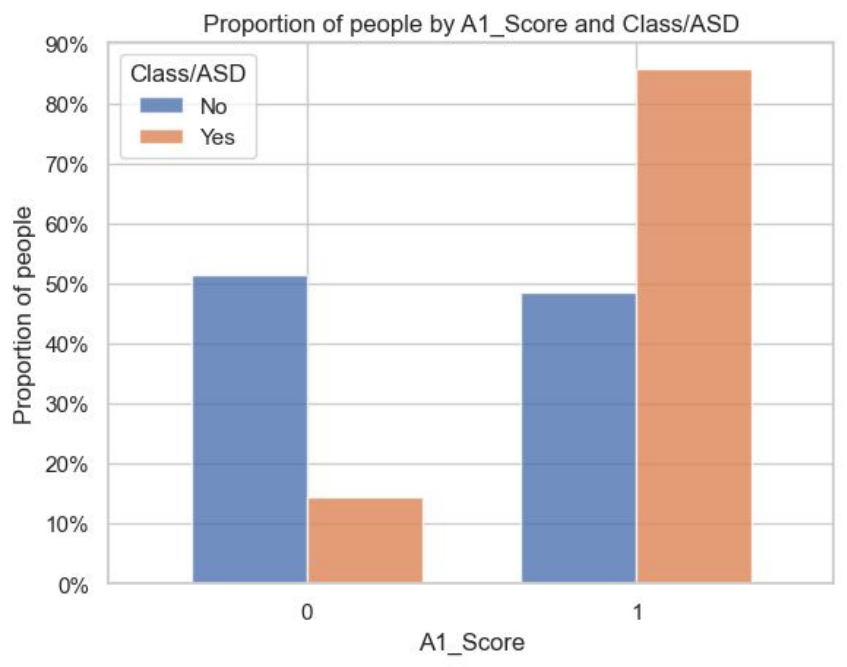
Data Cleaning

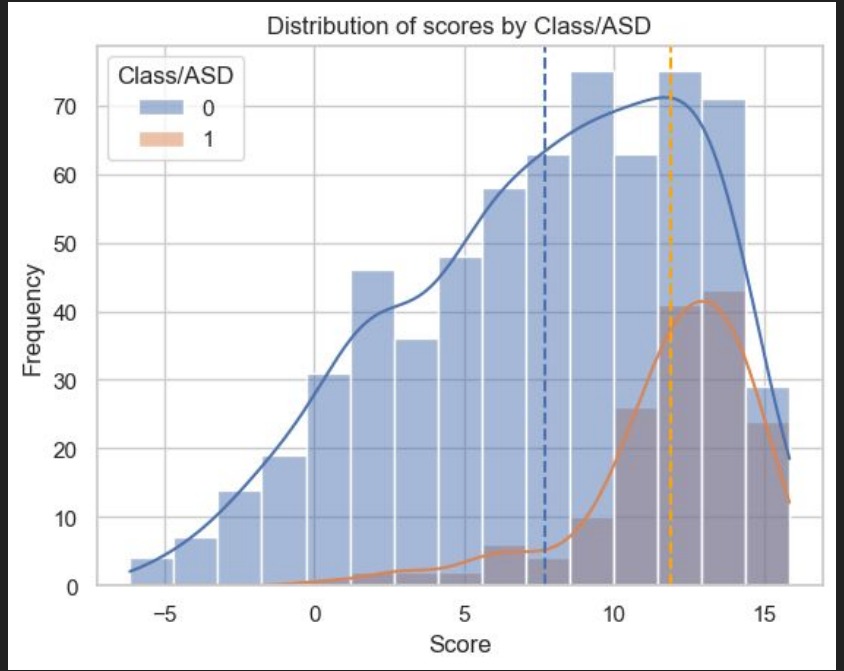
Simple replacing was used to change ‘no’ to 0, and ‘yes’ to 1.

Columns containing large amounts of missing data such as ethnicity and country of origin were dropped.

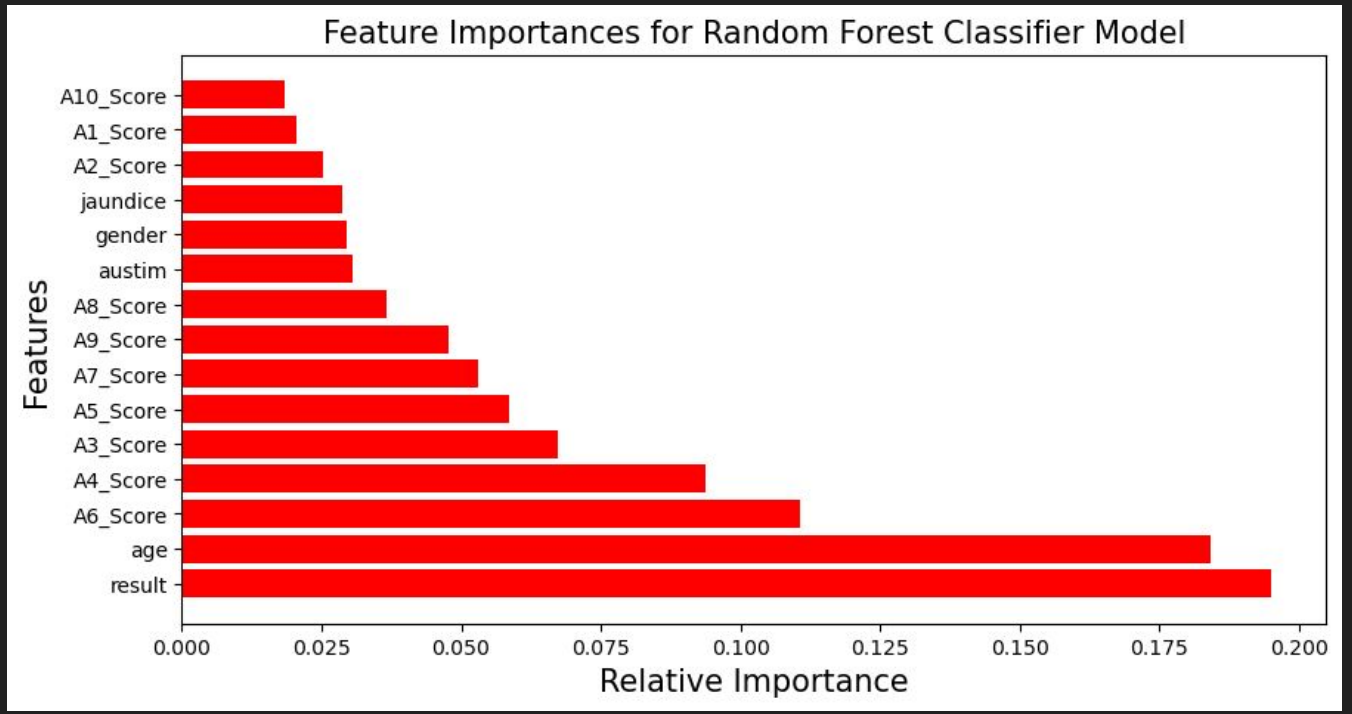
There were no categorical columns that needed dummy columns.

EDA


Proportion of people by A1_Score and Class/ASD
Class/ASD
No
Yes
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Proportion of people
0
1
A1_Score

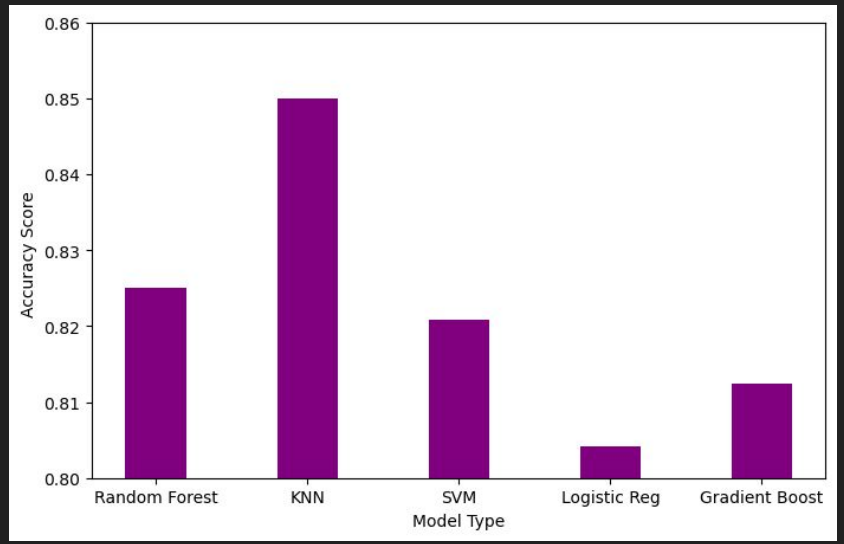

Distribution of scores by Class/ASD
Class/ASD
0
1
70
60
50
40
30
20
10
0
Frequency
−5
0
5
10
15
Score

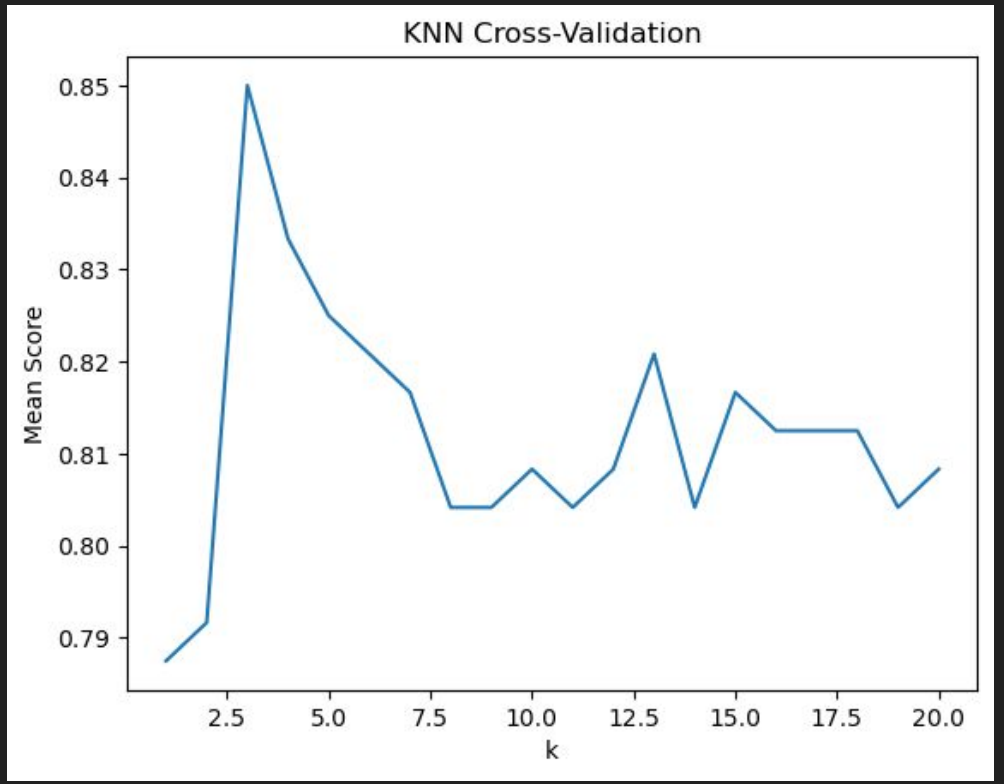
Modeling


Feature Importances for Random Forest Classifier Model
A10_Score
A1_Score
A2_Score
jaundice
gender
austim
A8_Score
A9_Score
A7_Score
A5_Score
A3_Score
A4_Score
A6_Score
age
result
Features
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Relative Importance

Modeling

KNN modeling had the highest accuracy score at 85%. This was achieved using k = 3.





Conclusion

The ROC curve shows that true positives can be detected a majority of the time. Improvements to the model could be done using more complex ensembling techniques.