This project data comes from a Kaggle competition held in 2022 and can be found here https://www.kaggle.com/competitions/autismdiagnosis/overview The data comes from a survey that had questions asked relating to certain attributes that people with Autism commonly display, then they were asked some demographic information about themselves. The purpose of this data is to create a model that could be used in theory with other participant data to get an idea to whether they should seek an ASD diagnosis or not.

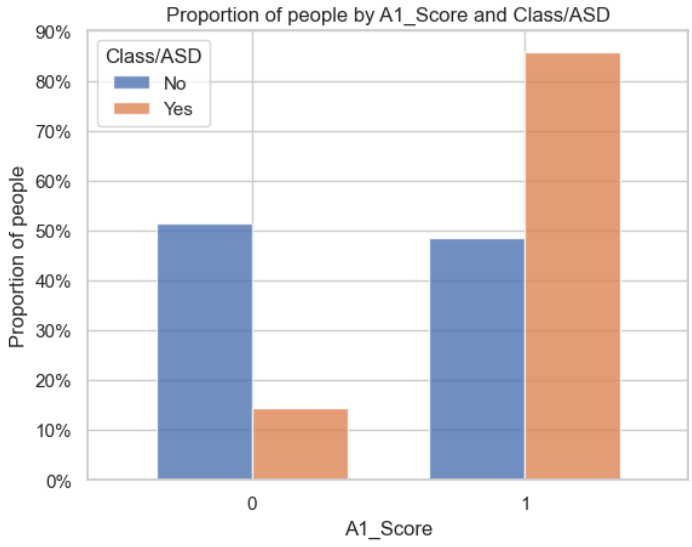
The data features are labeled as the following and came straight from the kaggle dataset:

- A1_Score - A10_Score: Questions asked in a survey, 1 being 'yes' and 0 being 'no'
- Age: Age in years
- gender: 0 for the person being a female and 1 if they are male
- Jaundice: 1 for if the person was diagnosed with Jaundice at birth and 0 if they weren't
- ASD_relative: 1 if they know of someone in their close relatives that was diagnosed with ASD during their lifetime
- result: Score for AQ1-10 screening test
- Class/ASD: Classified result as 0 or 1. 0 represents No and 1 represents Yes. This is the target column.

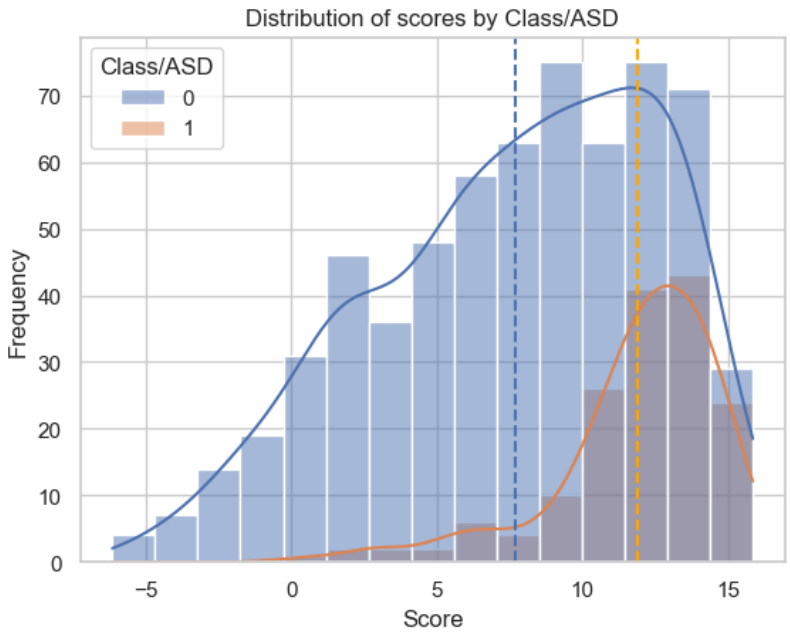
The majority of the data was already in the proper numerical format for the modeling process. There were no categorical columns in the original data. Most of the cleaning was converting 'no' and 'yes' to 0 and 1 respectively. There were a couple columns that were dropped due to there being too many missing values, only one unique value, or just not enough 'yes' values for the column to be of importance to us. After the dropping of the columns and renaming no and yes values, the data was ready to be explored.

The columns that display little importance are dropped below. Age_desc are all the same category of '18 and more', the relation and used_app_before columns didn't have much varriance in the answers either, and the ethnicity column had too many missing values for it to be important.

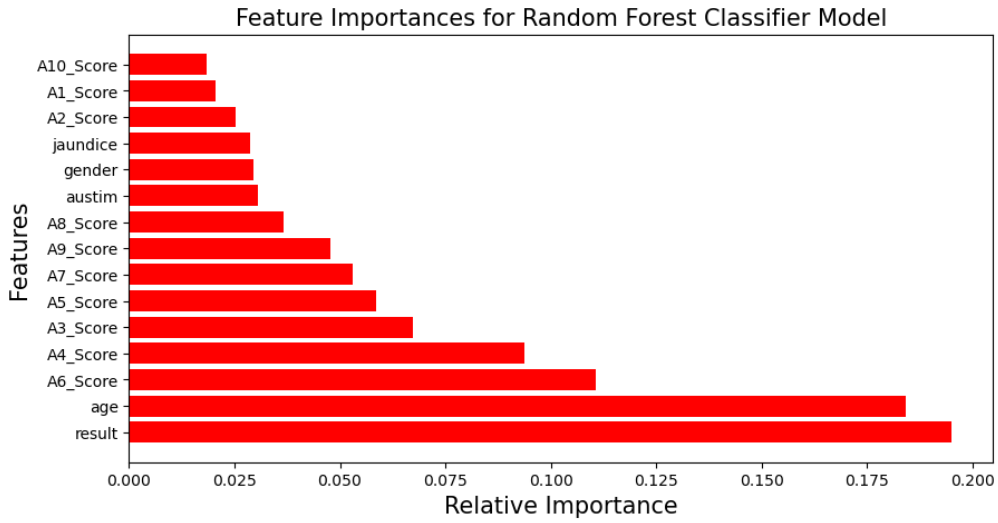
For the A1 - A10 questions, most of them followed this similar trend which shows that people with autism tend to answer yes to the questions rather than no. This strengthens the validity of the questions that were created since this was their purpose.



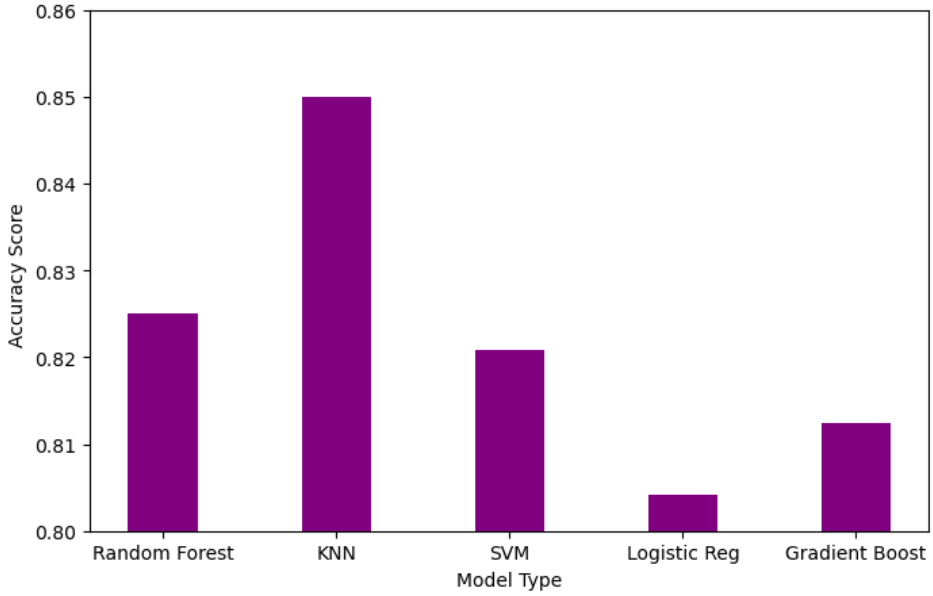
The figure below is made from the results column which is the score that the researchers gave the participants based on their answers to the questions given. It is apparent that pepole with autism in orange tended to have higher scores than those who didn't in blue. It is significant to point out that there were some participants who had really high scores but didn't report having autism. This could be because of misclassification on their own part, or they weren't properly diagnosed and the misclassification was from the psychologist instead.



When starting up the modeling part of this project, it appeared the main two features that the model is using is result and age. Result makes more logical sense since it is a score given to them based on their answers to the questionnaire. The age column is interesting but it might be that way the questionnaire was given might have biases that attract a certain age group.



All five of the models were within less than five percent from the highest score to the lowest. The KNN model scored the highest at 85% which might be due to it being easier to test all the different levels of k using our loop. The lowest was logistic regression at 80.4% which might just be because of the nature of our data. It is interesting how far the KNN model is ahead of the other models despite doing a gridsearch on them.



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.

Things that could be used to create a better model could include:

- Going through more parameters on the gridsearch for the models
- Experimenting with different kernels for the SVM model