Sanjukta Nad

# Detection of Autism Spectrum Disorder in Children

## Problem Statement:

Autism Spectrum Disorder (ASD) is a neurological disorder which might have a lifelong impact on the language learning, speech, cognitive, and social skills of an individual. Diagnosing ASD can be difficult because there is no medical test, like a blood test, to diagnose the disorder. Doctors look at the child's developmental history and behavior to make a diagnosis. Monitoring, screening, evaluating, and diagnosing children with ASD as early as possible is important to make sure children receive the services and supports, they need to reach their full potential.

To improve the precision and time required for diagnosis, machine learning techniques are being used to complement the conventional methods. I have applied models such as Logistic Regression (LR), Support Vector Machines (SVM), Random Forest Classifier (RF), Naïve Bayes (NB) and K Nearest Neighbor (KNN) to the dataset and constructed predictive models based on the outcome. The main objective is to determine if the child is prone to ASD in its early stages, which would help streamline the diagnosis process. Based on the results, Logistic Regression gives the highest accuracy for the selected dataset.

## Data Wrangling:

The dataset I have used (https://www.kaggle.com/fabdelja/autism-screening-for-toddlers) contained categorical, continuous and binary attributes. Originally, the dataset had 1054 instances along with 18 attributes (including class variable). Since the dataset contained a few non-contributing attributes, namely 'Case_No', 'Who completed the test', and 'Qchat-10-Score', those were dropped from the data set. The clean dataset is now ready for exploration before training and feeding to model.

## Exploratory Data Analysis:

In this dataset, ten behavioral features (Table 1) plus other individuals characteristics were recorded that have proved to be effective in detecting the ASD cases from controls in behavior science. I plotted several graphs to get different visual perspectives of the dataset.

Table 1: Features mapping with Q-CHAT-10 screening method

| Dataset variable | Description |
|---|---|
| A1 | Child responding to you calling his/her name |
| A2 | Ease of getting eye contact from child |
| A3 | Child pointing to objects he/she wants |
| A4 | Child pointing to draw your attention to his/her interests |
| A5 | If the child shows pretense |
| A6 | Ease of child to follow where you point/look |
| A7 | If the child wants to comfort someone who is upset |
| A8 | Child’s first words |
| A9 | If the child uses basic gestures |
| A10 | If the child daydreams/stares at nothing |

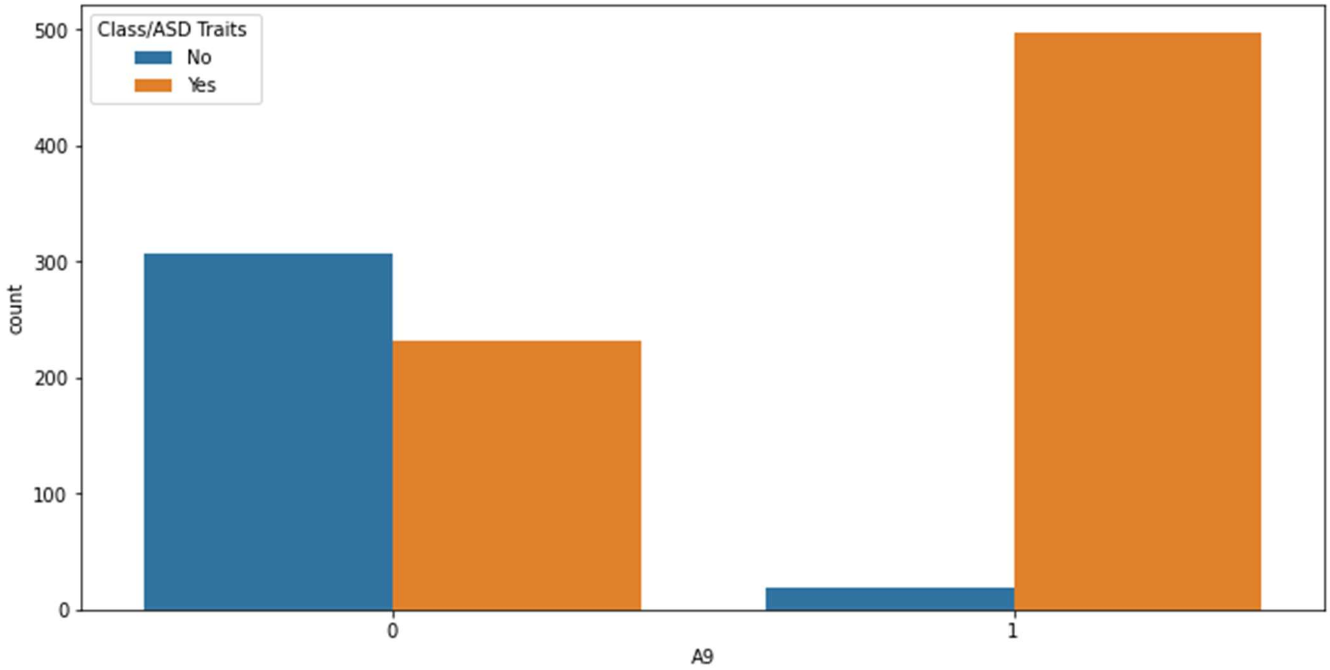


Fig 1: ASD positive and A9 response correlation

From Fig 1 we can see A9 has strong correlation with positive diagnosis.

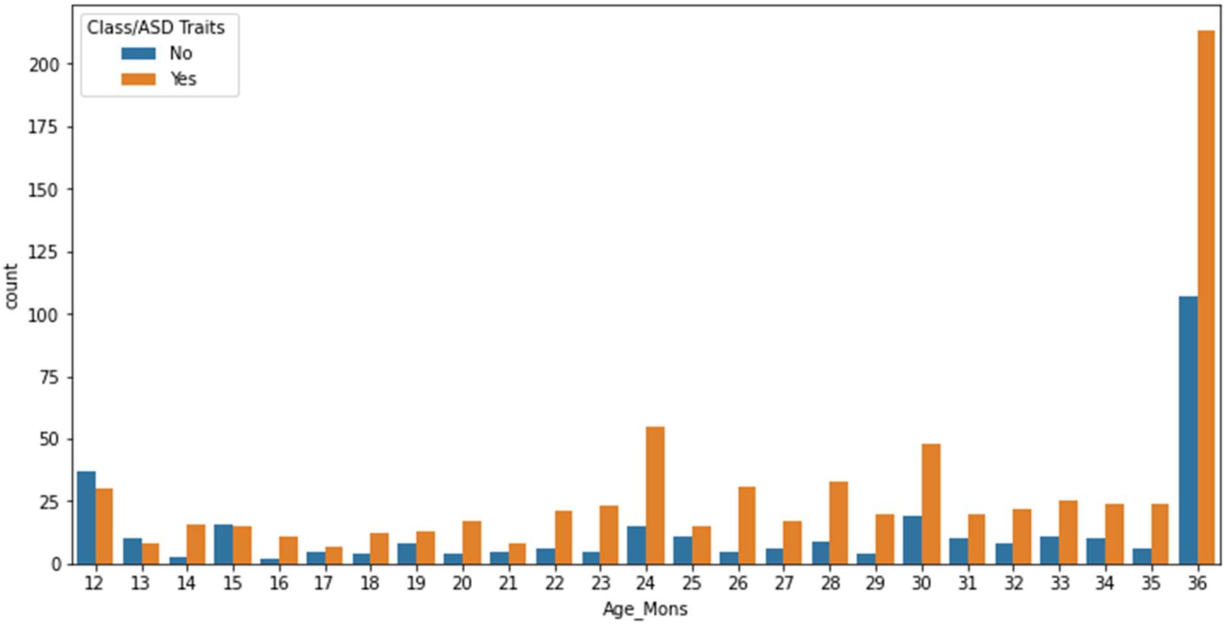


Fig 2: ASD positive and toddler age correlation

For toddlers, most of the ASD positive cases (Fig 2) happen to be at are around 36 months of age. The least number of cases are observed between 15 and 20 months of age. From the graph, it is evident at the age of 3 years substantial signs of autism occur.

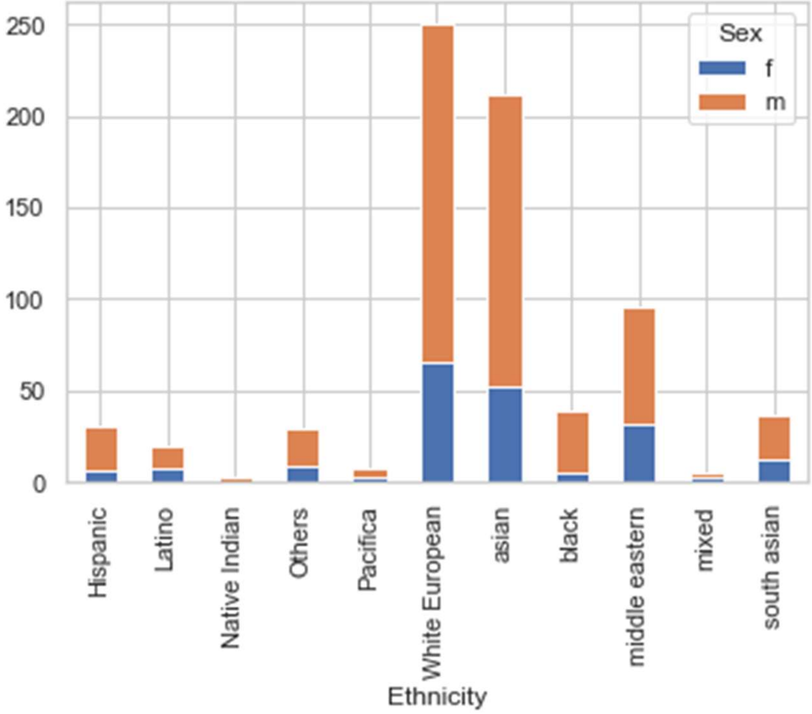
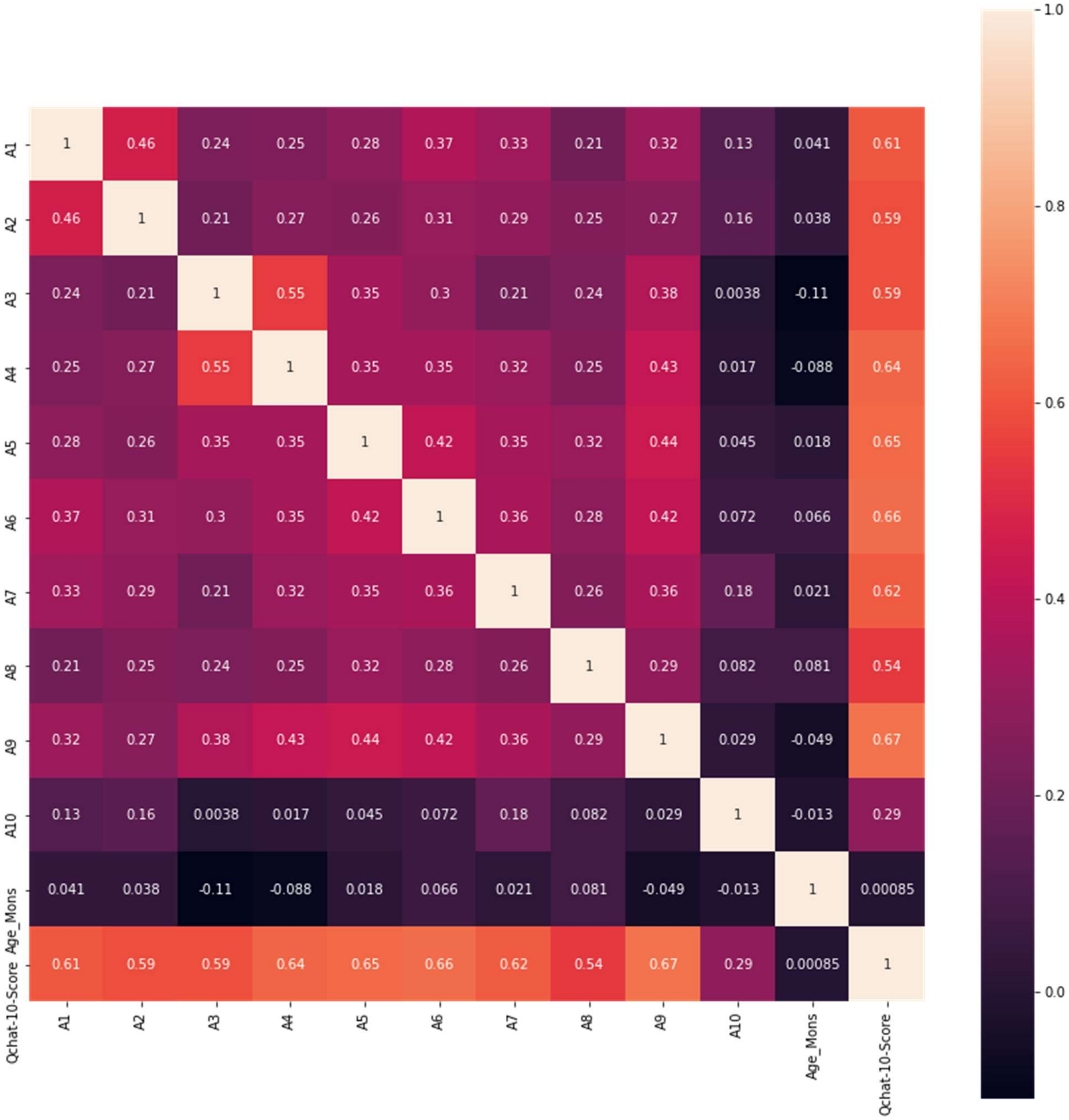


Fig 3: Proportion of male and female positive cases of Autism with Regards Ethnicity

Middle eastern, Asian and White European are the ethnicities that showed an increase in autism cases (Fig 3).



| | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | A10 | Age_Mons | Qchat-10-Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A1 | 1 | 0.46 | 0.24 | 0.25 | 0.28 | 0.37 | 0.33 | 0.21 | 0.32 | 0.13 | 0.041 | 0.61 |
| A2 | 0.46 | 1 | 0.21 | 0.27 | 0.26 | 0.31 | 0.29 | 0.25 | 0.27 | 0.16 | 0.038 | 0.59 |
| A3 | 0.24 | 0.21 | 1 | 0.55 | 0.35 | 0.3 | 0.21 | 0.24 | 0.38 | 0.0038 | -0.11 | 0.59 |
| A4 | 0.25 | 0.27 | 0.55 | 1 | 0.35 | 0.35 | 0.32 | 0.25 | 0.43 | 0.017 | -0.088 | 0.64 |
| A5 | 0.28 | 0.26 | 0.35 | 0.35 | 1 | 0.42 | 0.35 | 0.32 | 0.44 | 0.045 | 0.018 | 0.65 |
| A6 | 0.37 | 0.31 | 0.3 | 0.35 | 0.42 | 1 | 0.36 | 0.28 | 0.42 | 0.072 | 0.066 | 0.66 |
| A7 | 0.33 | 0.29 | 0.21 | 0.32 | 0.35 | 0.36 | 1 | 0.26 | 0.36 | 0.18 | 0.021 | 0.62 |
| A8 | 0.21 | 0.25 | 0.24 | 0.25 | 0.32 | 0.28 | 0.26 | 1 | 0.29 | 0.082 | 0.081 | 0.54 |
| A9 | 0.32 | 0.27 | 0.38 | 0.43 | 0.44 | 0.42 | 0.36 | 0.29 | 1 | 0.029 | -0.049 | 0.67 |
| A10 | 0.13 | 0.16 | 0.0038 | 0.017 | 0.045 | 0.072 | 0.18 | 0.082 | 0.029 | 1 | -0.013 | 0.29 |
| Age_Mons | 0.041 | 0.038 | -0.11 | -0.088 | 0.018 | 0.066 | 0.021 | 0.081 | -0.049 | -0.013 | 1 | 0.00085 |
| Qchat-10-Score | 0.61 | 0.59 | 0.59 | 0.64 | 0.65 | 0.66 | 0.62 | 0.54 | 0.67 | 0.29 | 0.00085 | 1 |

Fig 4:  Correlation matrix

Above Correlation shows A9, A7, A5, A6 has very strong correlation with Qchat-10- score

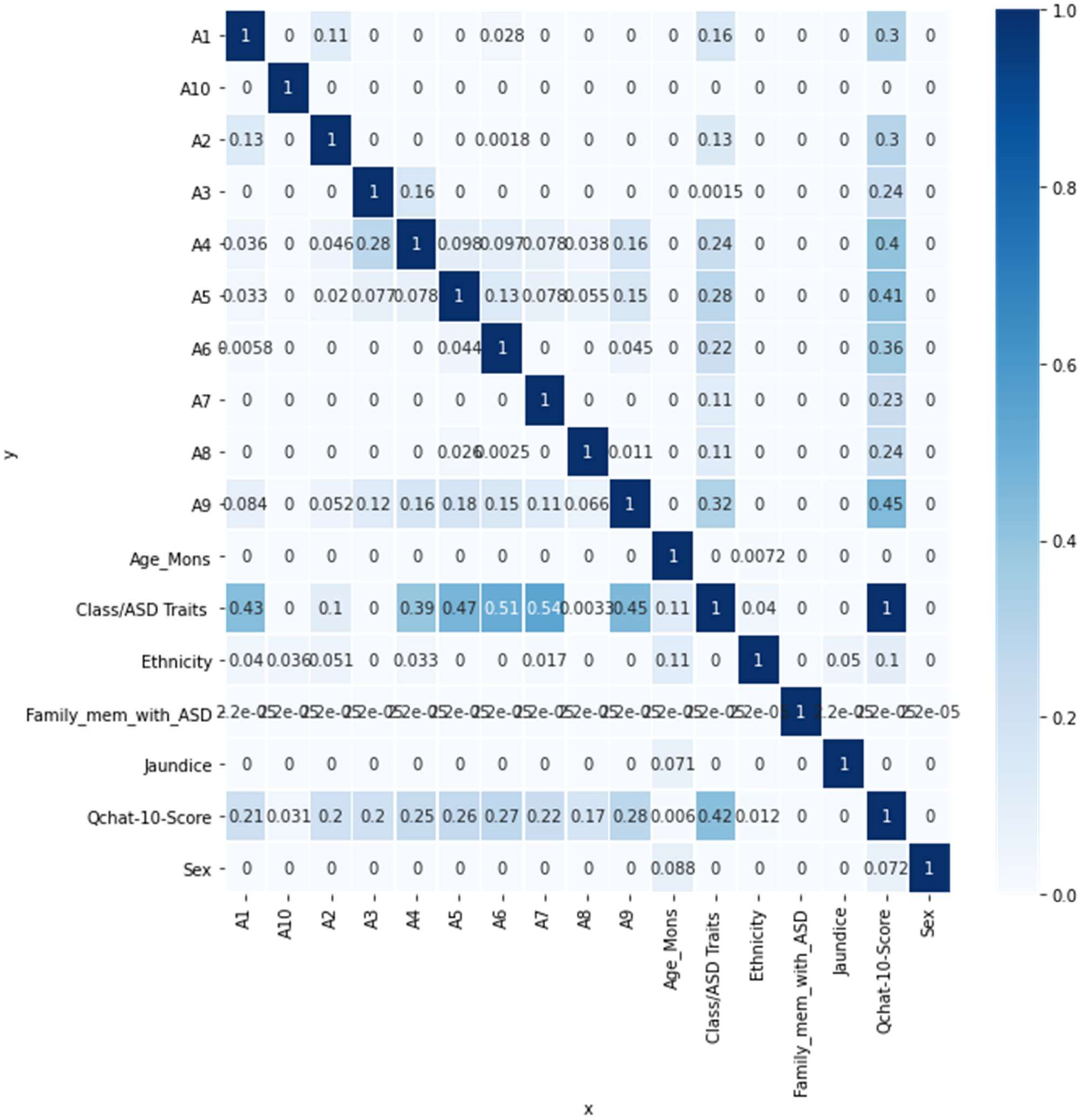


Fig 5:  Predictive Power Score (PPS) matrix

PPS can also handle categorical and nominal data as well, in addition to numerical values. Unlike the correlation matrix, PPS is asymmetric. Fig 5 shows the A7 PPS score on Class/ASD Traits is the highest (0.54), followed by A6 and A5.

## Training and Modelling:

To deal with the categorical values, label encoding was used. Label Encoding converts the labels into numeric form to make it machine-readable. Four features having 2 classes (Sex, Jaundice, Family_mem_with_ASD, and Class/ASD_Traits) binary label encoding is used. Label Encoding proves to be ineffective when there are more than 2 classes. For multiclass feature, ‘Ethnicity’ pandas get_dummies one hot encoder is used. The ‘Ethnicity’ feature has 11 classes.

The dataset was split into two parts—training set and test set. The training set consisting of 70% of the data (737 samples) is used to train the classification model. The remaining 30% of the data (317 samples) was reserved for testing the accuracy and effectiveness of the model on unseen data. I applied five classification models, namely Logistic Regression, Naive Bayes, Support Vector Machine, K-Nearest Neighbors, and Random Forest Classifier.

To evaluate the performance of all these models the accuracy, confusion matrix and F1 score were used. Table 2 shows a comparison of all the classification models I used.

Table 2: A comparison of the applied ML models

| | LR | NB | SVM | KNN | RF |
|---|---|---|---|---|---|
| Accuracy | 100 % | 88.9 % | 81.7 % | 95.2 % | 96.2 % |
| Confusion matrix | `[[ 95   0]`<br>`[ 0 222]]` | `[[ 91   4]`<br>`[31 191]]` | `[[37   58]`<br>`[0 222]]` | `[[ 89   6]`<br>`[ 9 213]]` | `[[ 86   9]`<br>`[ 3 219]]` |
| F1 score | 1.0 | 0.89 | 0.79 | 0.95 | 0.95 |

Logistic Regression:

Logistic Regression model testing accuracy was 100%. Recall, Precision, F1score were 100% for both yes autism and no autism patients, false positive and false negative= Zero. Logistic Regression with default parameters is an ideal model for this dataset.

Naïve Bayes model:

Naïve Bayes model had test accuracy: 0.89 suggesting Naïve Bayes is a not a great model.

KNN model:

KNN model had test accuracy: 0.95 with best k = 20. Weight average Recall was 95% for positive cases, so KNN is good model.

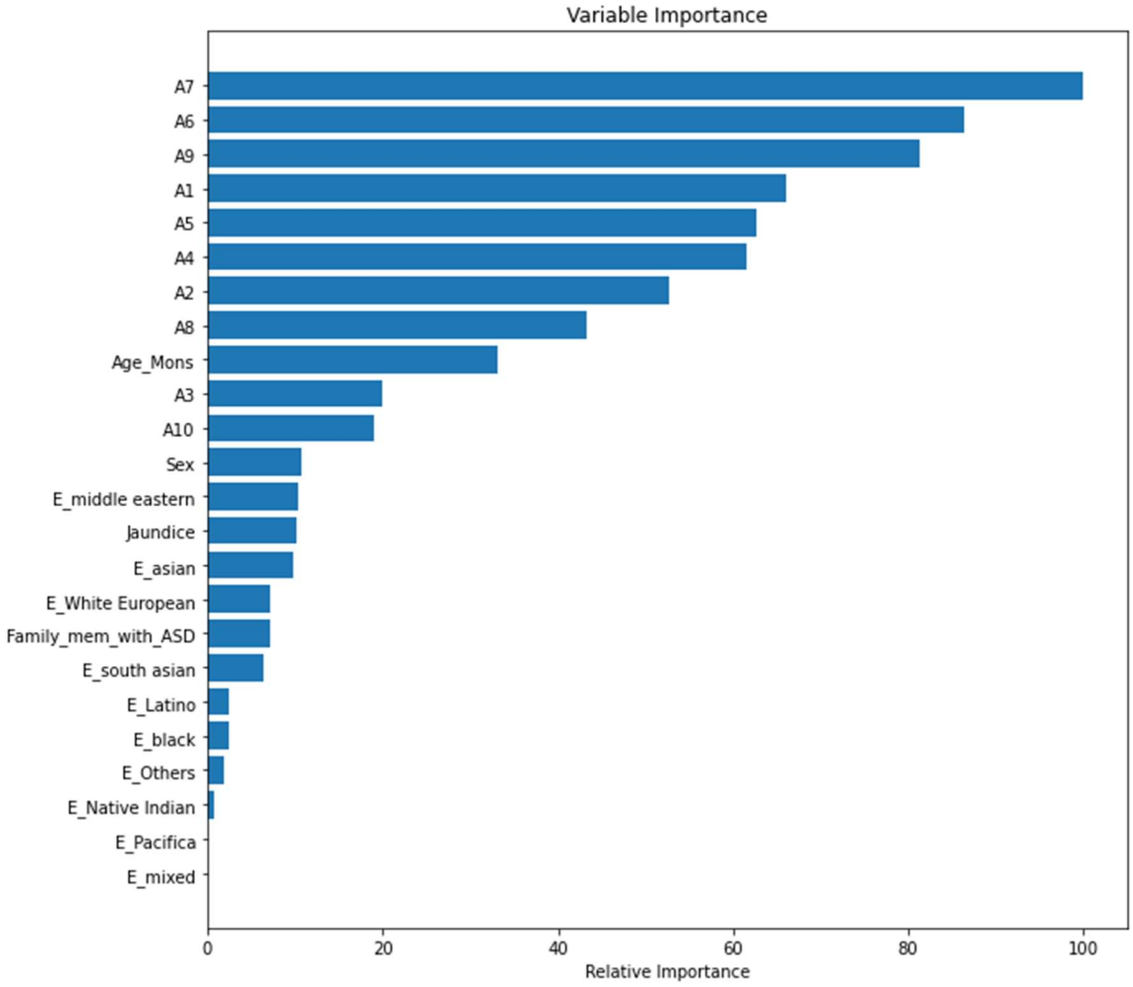
SVM model:

SVM had low test accuracy: 0.81. It is not a good model for this dataset.

RF model:

Random Forest model had test accuracy: 0.95 and weight average Recall = 99% for positive cases. After hypertuning Random Forest had performed even better with accuracy 0.96.

The variable importance plot reveals A7 is the most important features, other important features are A6, A9, and A1. Jaundice and the type of gender 'sex' are less important features. Genes in this study which appears through the column of family member with ASD is not an important feature. Males are more positive to autism than females.



## Conclusion:

There is currently no diagnostic test that can quickly and accurately detect ASD, or an optimized and thorough screening tool that can clearly identify the onset of ASD. This project has provided useful insights in the development of an automated model that can assist medical practitioners in detecting autism in children. Out of the five models that I applied to the dataset; Logistic Regression was observed to give the highest accuracy.

In future I will be considering using a larger dataset to improve generalization. The project analyzes different classification models that can accurately detect ASD in children with given attributes based on the child’s behavioral and medical information. The analysis of these classification models can be used for further exploring this dataset or other autism spectrum disorder data sets.