5650 Final Project

Davis Benson, Morgan Hales, Erik Dickamore

April 2022

1 Introduction

Autism Spectrum Disorder is a neurodevelopment disorder commonly associated with young boys. If detected early, parents and doctors can make the appropriate efforts towards helping to improve the lives of the autistic individual. If not detected early however, autism becomes increasingly difficult to diagnose as patients age up. Knowing this, the purpose of our analysis is to assist in discovering the best process to diagnose adults so that the necessary steps can be taken in helping the individual to work through their disorder. Using the adult autism screening data from Fadi Fayez Thabtah's study with the Department of Digital Technology at the Manukau Institute of Technology, Auckland, New Zealand, we performed a number of analyses to discover which variables are most important in classifying autism.

2 Data

This data set contains 704 observations and 21 attributes (variables). This dataset is used to detect autism in adults. It includes ten behavioral features as well as ten characteristics that have been effective in detecting ASD. The data set contained missing variables, but only 1.29 % of the data is missing. These discrepancies occur in the variables age, ethnicity, and relation with 0.28%, 13.49%, and 13.49% missing, respectively. As our analysis continued the importance of these variables decreased, therefore the missing values has little to no effect on our models.
The following is a table of the attribute information:

Table 1: Features and their descriptions

Attribute	Type	Description
Age	Number	Age in years
Gender	String	Male or Female
Ethnicity	String	List of common ethnicities in text format
Born with jaundice	Boolean (yes or no)	Whether the case was born with jaundice
Family member with PDD	Boolean (yes or no)	Whether any immediate family member has a PDD
Who is completing the test	String	Parent, self, caregiver, medical staff, clinician ,etc.
Country of residence	String	List of countries in text format
Used the screening app before	Boolean (yes or no)	Whether the user has used a screening app
Screening Method Type	Integer (0,1,2,3)	The type of screening methods chosen based on age category (0=toddler, 1=child, 2= adolescent, 3= adult)
Question 1 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 2 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 3 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 4 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 5 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 6 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 7 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 8 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 9 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 10 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Screening Score	Integer	The final score obtained based on the scoring algorithm of the screening method used. This was computed in an automated manner

Initial exploration of our data revealed the following demographic data.

1. Gender: 337 females 367 males
2. Age: Ranges from 17 to 64
3. Ethnicity :

Asian	Black	Hispanic	Latino	Middle Eastern
123	43	13	20	92
Pasifika	South Asian	Turkish	White-European	other
12	36	6	233	31

3 Methods

1. **Linear Regression**
   The first model fit to the data was Linear Regression as a baseline for the rest of our analyses. (Logistic Regression was attempted.) All of the predictor variables were used to classify an individual as having ASD or not having ASD. The following table shows each variable's coefficients and significance.

Variable	Estimate	P-Value
a1_score	0.09614	0
a2_score	0.119	0
a3_score	0.1193	0
a4_score	0.1131	0
a5_score	0.1521	0
a6_score	0.2213	0
a7_score	0.1422	0
a8_score	0.1371	0
a9_score	0.263	0
a10_score	0.0511	0.0229
age	0.0009409	0.111
gender	-0.04031	0.0505
ethnicity	0.004269	0.1048
jaundice	0.04755	0.1708
autism	0.00149	0.9610
Country of Residence	0.00002259	0.9657

It becomes clear from linear regression that the survey response scores are the only variables that have a significant effect on the classification of ASD in an individual. With this in mind, the data was subsetted and the linear regression model was refit. The following table is a confusion matrix of the linear model's predictive ability:

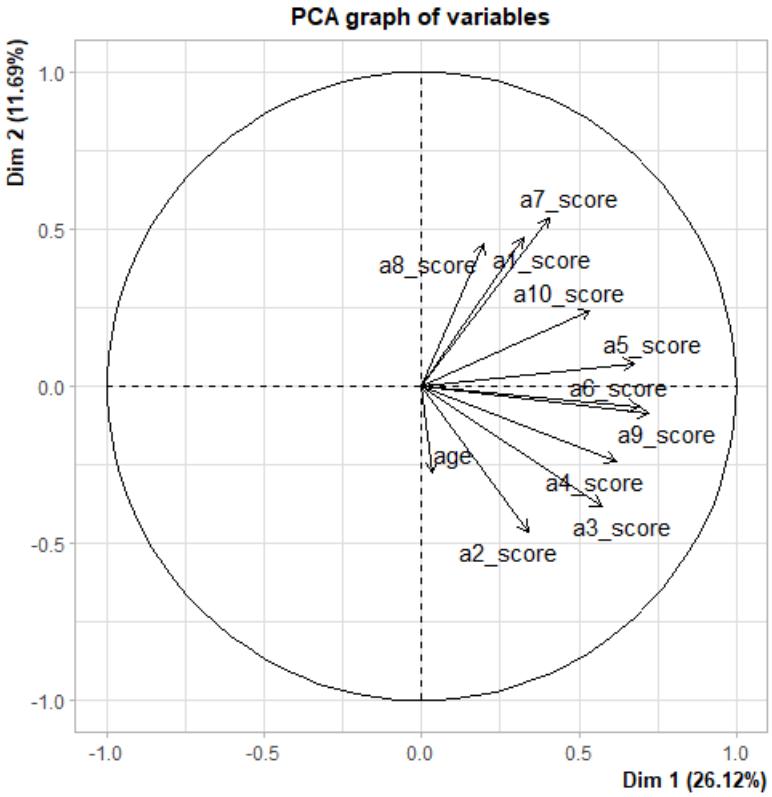
	Not Detected (0)	Detected (1)	error rate
Not Detected (0)	504	11	0.021359
Detected (1)	13	176	0.068783

The subsetted linear model correctly classified 90.9 % of the data.

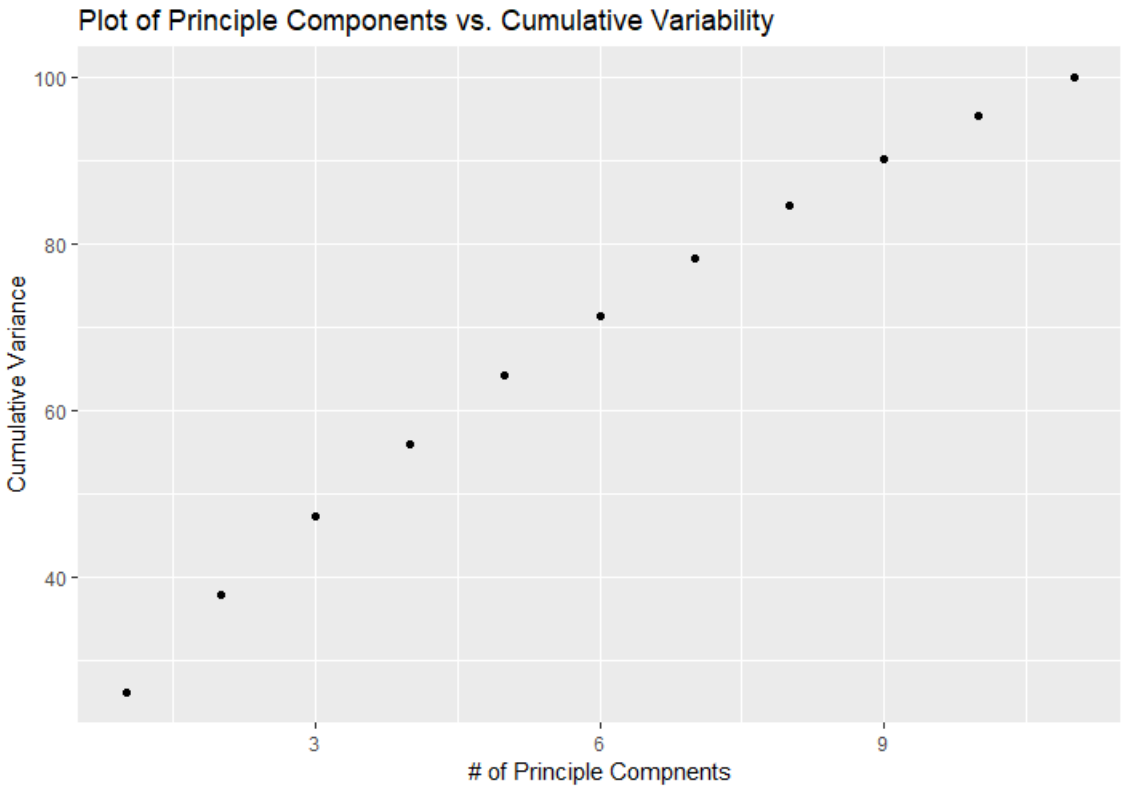
2. **Principle Component Analysis**

PCA was applied to the data. Imputation was necessary, so the *imputePCA()* function from the **missMDA** package was used. Then, the *PCA()* function from the **FactoMineR** package was used to attain principle components and their corresponding eigenvectors. It took 9 principle components to reach 90% cumulative variability explained and 11 to explain 100%. The following table provides information about the principle components. With 21 predictor variables, it is reasonable to expect 11 principle components to explain the observed variance. The first 3 components explained nearly half of the observed variance.

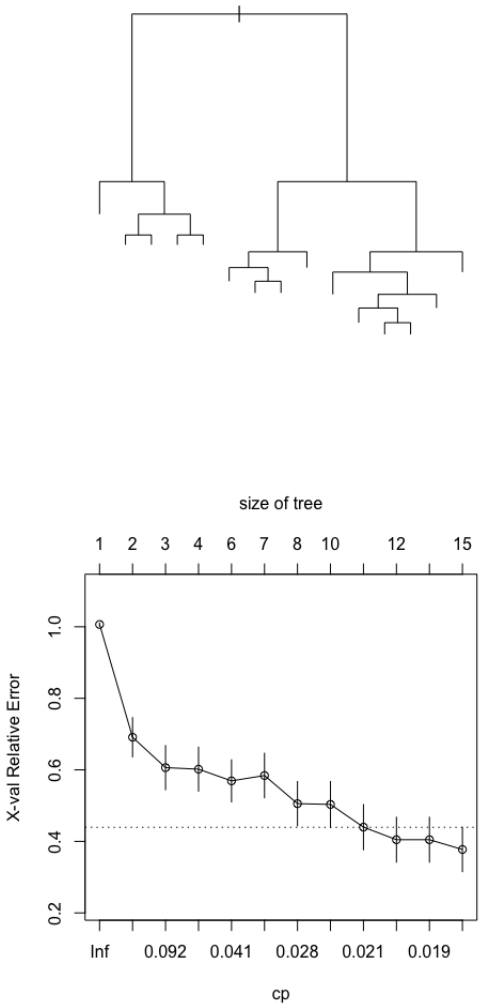
	PCA1	**PCA2**	**PCA3**	**PCA4**	**PCA5**	**PCA6**
% of Variance	26.117	11.686	9.452	8.745	8.161	7.199
Cumulative % of Variance	26.117	37.801	47.255	55.999	64.160	71.359
	PCA7	**PCA8**	**PCA9**	**PCA10**	**PCA11**	
% of Variance	6.871	6.322	5.551	5.218	4.679	
Cumulative % of Variance	78.230	84.553	90.104	95.321	100.00	



To visualize the cumulative variability explained with each additional principle component, a plot of cumulative variability vs. the number of principle components was obtained (Figure 2). The cumulative variability quickly increases to nearly 60% by the fourth principle component.


Plot of Principle Components vs. Cumulative Variability
Cumulative Variance
100
80
60
40
3
6
9
# of Principle Compnents

3. **Classification Tree**



Looking at Figure 3 We can see in our cp plot that a tree with 13 nodes and a cp of about 0.019 would be a good fit four our data. The resultant tree is displayed in Figure 3 The following Table shows the miss-classification rate for our classification tree. The resulting success rate for our classification tree model was 91.55 %

	Not Detected (0)	Detected (1)	error rate
Not Detected (0)	509	6	0.01165049
Detected (1)	58	131	0.3068783

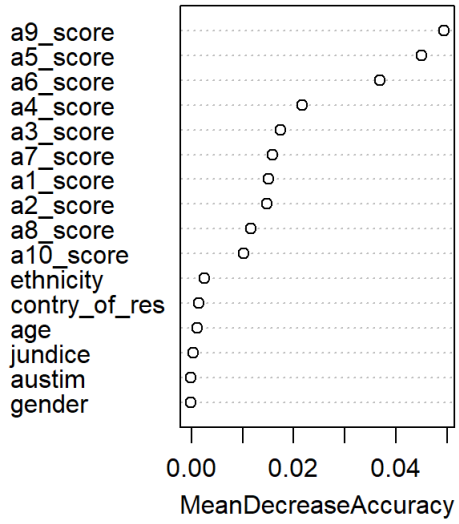
4. **Random Forest**

The Random Forest Model was fit to the data. It was trained on all predictor variables with autism class as the response. The following table is a confusion matrix of the percent correctly classified.

	Not Detected (0)	Detected (1)	error rate
Not Detected (0)	424	5	0.01165501
Detected (1)	17	163	0.09444444

The total percent correctly classified is 89.4% The model did will at classifying individuals without Autism, but had a higher error rate when classifying individuals with Autism.

A plot of variable importance for the random forest model was obtained. Answers to the survey were particularly important; All of the survey scores were ranked as more important than the characteristic variables.



The model was then refit to the data, this time only including the survey response variables (a1_score - a10_score). The following is the confusion matrix on the subsetted data

	Not Detected (0)	Detected (1)	error rate
Not Detected (0)	505	10	0.01941748
Detected (1)	10	179	0.05291005

As expected, this improved the models ability to classify Autism but lowered the models ability to correctly classify individuals without Autism. The percent of individuals correctly classified is 92.76 % a big improvement on the original error rate.

5. **Gradient Boosting Without Tuning**

Another statistical method that was applied to our data was gradient boosting both with and without tuning. For the original, un-tuned analysis, we were able to find the following variable importance:

Variable	Importance
a9 score	36.52
a6 score	16.97
country of res	14.26
a5 score	12.81
a4 score	7.09
a3 score	4.78
a7 score	3.59
a8 score	1.45
a10 score	1.21

Similar to our random forest approach, the results for a9, a6, and a5 all appeared to be very important in the classification of autism. Using 50 trees to avoid overfitting, we were able to get the following accuracies:

	Not Detected (0)	Detected (1)	error rate
Not Detected (0)	504	11	0.0218
Detected (1)	27	162	0.167

From these results, we can see that the gradient machine was experiencing more difficulty in predicting for people who didn't have autism than for those who did. The model correctly classified 81.1% of individuals.

6. **Gradient Boosting With Tuning**
   Tuning the GBM gave the following ideal values for the machine:

Setting	Tuned Value
Interaction Depth	12
Number of Trees	100
Shrinkage	.1
Min Observations Per Node	11

   Utilizing these values boosted the machine up to 100% accuracy. We have some concerns about over fitting of the data, but when testing other tree counts with the original data, we found that 100 wasn't enough to significantly over fit our un-tuned GBM, so the issue may lie somewhere else.

4 Findings and Conclusions

After fitting as many models as possible to the data it was found that Random Forests was the most accurate way to classify an individual as having autism or not having autism. After the data was subsetted to the most important variables the model was able to correctly classify nearly 93% of the individuals. This is incredible accuracy that leads us to believe that random forests will eventually replace logistic/linear regression in the future.

implications:

All of the models found that the survey scores were the most significant/important variables to include in our model. We learned from modeling these data that factors such as age, gender, ethnicity and country of residence have no effect on classifying an individual with autism. This, however, makes a lot of sense. Autism is a behavioral issue that is genetic and doens't depend on outside factors.

5 Data Source

author = "Fadi Fayez, Thabtah"
year = "2017"
title = "UCI Machine Learning Repository"
url = "https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult"
institution = "Department of Digital Technology Manukau Institute of Technology, Auckland, New Zealand"

6 Relevant Papers

1) 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.
2) Thabtah, F. (2017). ASDTests. A mobile app for ASD screening. www.asdtests.com [accessed December 20th, 2017].
3) Thabtah, F. (2017). Machine Learning in Autistic Spectrum Disorder Behavioural Research: A Review. To Appear in Informatics for Health and Social Care Journal. December, 2017 (in press)