5-16-2023

Early detection of autism spectrum disorder (ASD) using Machine Learning techniques

Mahra Musabbah Bin Beyat Alfalasi
mma2056@rit.edu

Early detection of autism spectrum disorder (ASD) using Machine Learning techniques

by

Mahra Musabbah Bin Beyat Alfalasi

A Capstone Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Professional Studies:

Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology

RIT Dubai

May 16th 2023

RIT

Master of Science in Professional Studies:

Data Analytics

Graduate Capstone Approval

Student Name: Mahra Musabbah Bin Beyat Alfalasi

Graduate Capstone Title: Early detection of autism spectrum disorder (ASD) using Machine Learning techniques

Graduate Capstone Committee:

Name: **Dr. Sanjay Modak** **Date:**

Chair of committee

Name: **Dr. Ioannis Karamitsos** **Date:**

Member of committee

Acknowledgments

First and foremost, I would like to thank my parents and siblings and the rest of my extended family for their continuous support during this process. I would also like to share my gratitude towards my workplace that enabled me to gain enough flexibility and empowerment to be able to pursue this project. Moreover, I would like to show immense gratitude towards my friends for always believing in me. Finally, I am extremely grateful and honored to be taught by the staff of Rochester Institute of Technology, they gave me the knowledge and skills to be able to execute this in the best manner possible.

Abstract

ASD is a neurological disorder that affects over 1 in 44 chilren and this rate has increased. The diagnosis process can be timely and costly. This will make it difficult for the patients to adhere to the prescribed treatments and will hinder the progress of the patient. This project is focused on increasing the efficiency of this diagnosis process through machine learning techniques. The proposed datasets are; ASD Screening Data for Adult, ASD screening Data for Children and ASD Screening Data for Adolescent. These datasets are a categorical, continuous, and binary data type. They have 21 common attributes and a total of 1,100 instances.

Keywords:

Autism Spectrum Disorder (ASD)

Jaundice

Machine learning

Random Forest Classifier

Support Vector Machine

Logistics Regression

Confusion Matrix

Table of Contents

List of Figures

List of Tables

Chapter 1

1.1 Background

Autism spectrum disorder (ASD) is a serious neurological disorder that has an impact on impairing social skills, learning, speech, social and cognitive abilities of an individual. The severity of its effects can hugely vary however, many common symptoms are difficulties in communication, obsessive interests and mannerisms that occur in a repetitive form (Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques,2021).

Autism is generally a result of genetics as well as environmental factors. The characteristics of an autistic child can be detected in early childhood however autism usually goes undiagnosed until later. Early detection and interventions can optimize the health, wellbeing, and quality of life of those suffering with ASD as well as improve the cognitive development of these individuals. However, the traditional diagnosis methods take a up to 13 months(Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques,2021).

1.2 Statement of problem

In the past few decades, there has been a global increase in the prevalence of autism spectrum disorder (ASD). Furthermore, statistics indicate that 1 in 146 births in the UAE is affected by ASD. (Autism spectrum disorder in the United Arab Emirates: potential environmental links,2020). ASD affects over 1 in 44 children. The prevalence of autism has increased from 1 in 150 in 2002 as high as 1 in 44 by 2018. It is also important to note that boys are four times more likely to have autism than girls.

The problem here lays in the extremely tedious and long process of traditional diagnosis of ASD. By the time the patient is properly diagnosed and prescribed the appropriate treatment and therapy, it could hinder their chance of minimizing the extent of the symptoms they may suffer from. It could also reduce their chances of living a more independent life without the need of special care all the time. Moreover, the costs associated with diagnosis can be a huge burden on families. It could lead some families to make the hard choice to leave their child undiagnosed and could result in additional suffering.

This project aims to solve this issue by applying machine learning techniques to accelerate the rate of diagnosis and make is less costly on the affected families.

1.3 Project goals

The main objective of the project is to propose a method for the early detection of autism spectrum disorder (ASD) using Machine Learning techniques with the use of historical data (to be more specific- attributes). A secondary objective would be that through early detection, the quality of life of those suffering with ASD. This is because caretakers will be able to take the correct decisions and actions to minimize or even prevent the symptoms that the person diagnosed with ASD may suffer from.

The project goals are as follows:

1. To develop an algorithm that can diagnose autism spectrum disorder earlier, faster and with higher accuracy than a health care practitioner
2. To have an algorithm that would not replace, but support, healthcare partitioners and care givers in making the correct and most informed decisions.

1.4 Methodology

The datasets were obtained from the UCI machine learning repository. The three chosen datasets are; ASD Screening Data for Adult, ASD screening Data for Children and ASD Screening Data for Adolescent. These datasets were then loaded into R and combined. The variable that we are trying to predict, the dependent variable, is whether the patient suffers from autism or not. There are 20 variables that would determine the results of this prediction. The next phase would be to prepare the data to fit and run the models efficiently and with accuracy. Dealing with missing data is vital to ensure the data is not skewed and misinterpreted. This process will involve either removing the missing data or imputing it.

Prior to modelling the data to fit the selected predication models, some exploratory data analysis was conducted. This was done with the aim to deepen the understanding of the data and their respective trends. This step is extremely important to assess which models will work best and to also expand the understanding of the topic further in hand.

The data modelling phase involved testing several techniques to analyze and assess their performance to make a well-informed decision on which model would be the best fit. To do so, the dataset will be split into training and testing subsets. The training dataset will contain most of the observations and will feed the data into the model. The testing subset will allow the testing of the models' accuracy and effectiveness in predicting the response.

1.5 Limitations of the Study

The main limitation is that a larger dataset is required to be obtained as it would allow for better training of the dataset as well as better model prediction accuracy.

The resultant models have small margin of error, the proper caution, validation, and regulation are required to be put into place. This is because in this study, the issue being dealt with is highly crucial as it deals with behavioral and cognitive problems that may seem subjective to the healthcare professionals.

Chapter 2 – Literature Review

Autism spectrum disorder (ASD) is a complicated condition which impairs a person's capability to act appropriately, both in verbal and non-verbal communication, and moreover involves recurrent or aberrant conduct. It is indeed a condition that grows in complexity as a person's mind develops. Although autism spectrum disorder (ASD) can eventually become irreversible and many people who have been diagnosed with this condition do go on to lead a normal, full lives, it is essential to treat this condition as young as possible in order to reduce the number of obstacles that it presents, as it only becomes more rigid as the person grows older. Prior on, a significant amount of headway was achieved in this area by applying strategies for machine learning to a wide variety of algorithms that were then used to handle a wide range of datasets. This Literature review was motivated by the need to bring to light a variety of outcomes obtained from various methodologies that have previously been put into practice in order to provide a perspective on probable ways in which they might complement one other. Nevertheless, the idea of computer vision and how successful it may be is something that has received less research attention but is now being examined.

Machine learning has been used in a variety of research in an effort to enhance and expedite the diagnostic process for autism spectrum disorder (ASD). Tests which are standardized with in clinical setting have been the only procedures that are utilized at this moment to identify autism spectrum disorder (ASD). Over this, so will the diagnostic process take much more time, but also the associated expenditures will skyrocket. In addition to the more traditional ways, practitioners are increasingly turning to machine learning strategies in an effort to enhance the diagnostic process in terms of both accuracy and efficiency. (Vakadkar, Purkayastha, and Krishnan, 2021) have created statistical models based on the result of basic concepts to our database including such Support Vector Machines (SVM), Random Forest Classifier (RFC), Naive Bayes (NB), Logistic Regression (LR), and KNN. The primary purpose of the study written by

(Vakandkar, Purkayastha, and Krishnan, 2021) is thus to assess whether the kid is vulnerable to ASD in its embryonic phases, which would assist speed the process of diagnosing autism spectrum disorder. (Vakadkar, Purkayastha, and Krishnan, 2021) have achieved the best accuracy with Logistic regression for their selected dataset.

By using forward feature selection in conjunction with it under sampling, (Duda et al., 2016) were able to discern among autism and ADHD with the assistance of a Social Responsiveness Scale that consisted of 65 items. Deshpande et al., (2013) employed measurements caused by brain function to predict autism spectrum disorder. Techniques from the field of computational models, such as probabilistic reasoning, artificial neural networks (ANN), and classifier combination, are also used Pratap et al. (2014). A significant number of the research that were carried out discussed automated machine learning models that just rely on attributes as input features. A couple of the research also relied on information gleaned from neuroimaging of the brain. In the ABIDE database, (Parikh, Li and He, 2019) retrieved 6 personal characteristics from 851 people and carried out the deployment of a cross-validation technique for the training and testing of the ML models. Additionally, they conducted the extraction of these personal characteristics. This was used to categorize patients into those who had ASD and those who did not have the disorder. (Thabtah and Peebles, 2019) introduced a novel machine learning strategy that they dubbed Rules-Machine Learning (RML). This method provides users with a knowledge base of rules that helps them comprehend the fundamental reasoning for the categorization, in addition to recognizing ASD characteristics. (Al Banna et al., 2022) used a tailored AI-based approach that helps with monitoring and support of ASD patients, hence assisting these individuals in coping with the COVID-19 pandemic.

Vaishali R. and Sasikala R., (2018) have come up with a way to diagnose autism using optimal behavior sets. An experiment was carried out in this study using a swarm intelligence based binary firefly feature selection wrapper. The research used a dataset for diagnosis that included features taken from the UI machine learning repository. The alternative hypothesis for the experiment asserts that it is feasible for a

machine learning model to attain greater classification accuracy with fewer feature subsets. This assertion is made in reference to the experiment. It was discovered via the use of swarm intelligence based on a single-objective binary firefly feature selection wrapper that certain characteristics within the features of the dataset are adequate to differentiate between patients and nonpatients. The results that were obtained using this method provide evidence in support of the hypothesis by producing an accuracy rate that is on average somewhere between 92.12 percent and 97.95 percent with optimum feature subsets that are roughly equivalent to the accuracy rate that was generated mostly by entire diagnosis dataset.

Kosmicki et al., (2015) have proposed a strategy for looking for the smallest possible collection of characteristics that may be used to diagnose autism. A technique based on machine learning was used by the authors in this study to evaluate the clinical evaluation of ASD. The ADOS was run on the subset of children's behaviors that were determined to fall somewhere on the autism spectrum. The ADOS system consists of four components. This study used a total of eight distinct machine learning methods, one of which included the stepwise backward identification of features on score sheets derived from 4540 different people. It identifies an autism spectrum disorder risk by using 9 of the 28 behaviors from module 2 and 12 of the 28 behaviors from module 3, and its overall accuracy is 98.27 percent and 97.66 percent, respectively.

Li et al., (2017) have shown that machine learning classifiers are effective in identifying autistic adults via the use of an imitation strategy. Investigating the underlying difficulty relating to discriminative test circumstances and kinematic characteristics was the purpose of this research. There was a total of sixteen ASC individuals included in the dataset, each of whom performed a set of hand motions. With the use of machine learning techniques, forty kinematic constraints were retrieved from eight different imitation situations in this study. This study demonstrates that it is possible to use machine learning approaches to the analysis of high-dimensional data and the diagnostic categorization of autism, even with a small sample size, as shown by the findings of this research. On the AQ-Adolescent dataset, the sensitivity rates

that were attained by RIPPER with the features Va (87.30 percent), CHI (80.95 percent), IG (80.95 percent), Correlation (84.13 percent), CFS (84.13 percent), and "no feature selection" (80.00 percent).

Many other scholars and writers, such as Raj and Masood. (2020) accepted the effectiveness of ML in the early diagnosis of autism. In their article, “Analysis and detection of autism spectrum disorder using machine learning techniques,” They found that early diagnosis of ASD is critical, but for large datasets, it is beyond human capabilities. They stated that machine learning techniques are very useful in medical diagnosis. In the article, they analyze Naïve Bayes, Convolutional Neural Network, Neural Network, KNN, Logistic Regression, and Support Vector Machine, and the use of these techniques for an autism diagnosis. They collected children and adult datasets and tested them with all the techniques. They found 96.88%, 98.30%, and 99.53% accuracy (Raj & Masood, 2020).

In their study, Hossain et al. (2021) stated that ASD (Autism Spectrum Disorder) affects senses, such as touching, smelling, and hearing. Specifically, in children, it is a crucial problem. However, an early diagnosis of the disease helps to improve conditions leading to a better life. In their article, “Detecting autism spectrum disorder using machine learning techniques," they proposed machine learning techniques and compared them with traditional clinical methods. They argued that traditional clinical methods are lengthy and costly. They recommended using classification methods through the automation diagnosis process. They used binary datasets, and evaluation metrics, such as classification errors, F-measures, precision, and recall. They concluded that SVM and SMO (Sequential Minimal Optimization) are the best methods for ASD diagnosis (Hossain, Md & Kabir, Ashad & Anwar, Adnan & Islam, Md. ,2021).

Khan et al. (2019) in their article, “A machine learning approach to predict autism spectrum disorder," stated that in the present world, ASD has become more challenging because of its high occurrence rate and the number of patients. Screening is a highly expensive and time-consuming

diagnosis method. The rapid advancement in technology, algorithms, and machine learning offer a cost-effective and fast solution for ASD early diagnosis. They proposed a prediction model that can help develop a mobile app to predict ASD. The model is a combination of Random Forest-CART, and Iterative Dichotomiser 3 (Random Forest-ID3). They found that their proposed model is effective and accurate in predicting ASD (Islam, Muhammad Nazrul & Omar, Kazi & Mondal, Prodipta & Khan, Nabila & Rizvi, Md., 2019)

Thabtah and David (2020) in their article, “A new machine learning model based on induction of rules for autism detection,” discussed ASD and its impacts on human behavior and senses, such as communication, social interaction, and repetitive behavior. In the clinical environment, only licensed specialists can diagnose autism spectrum disorder. Due to the prevalence and continual growth of ASD, these professionals are fully occupied and charged expensively. In critical, or even in general situations, an early diagnosis of Autism is beneficial both for the patients and families. Therefore, clinical methods and professionals are criticized. They conducted an extensive literature review to support the use of ML in predicting autism (Thabtah, F., & Peebles, D. , 2020).

In the article "Enhancing diagnosis of autism with optimized machine learning models and personal characteristic data" Parikh et al. (2019) claimed that 1% of the global population suffers because of ASD. The early diagnosis and detection of ASD play a pivotal role in controlling and improving ASD conditions. They conducted an experimental and statistical study to verify the effectiveness of ML (Machine Learning) in predicting ASD. They used a cross-validation strategy for training and testing machine learning models to categorize autism disorder patients and non-ASD controls. For assessing classification performance, they tested six personal characteristics. They concluded that machine learning models are far better and more cost effective than traditional clinical methods for predicting ASD (Parikh, M. N., Li, H., & He, L. , 2019).

Bone et al. (2015) penned the article "Applying machine learning to facilitate autism diagnostics: pitfalls and promises." According to them, the potential of ML (Machine Learning) is undeniable for fast prediction of ASD leading to improving patients' conditions. They stated that criticizing clinical methods and environment for autism detection is unjustified because the clinical domain is an essential requirement to apply machine learning. They quoted Wall et al (2012a; 2012b) to demonstrate how machine learning reduces time and costs (Bone, D., Bishop, S. L., Black, M. P., Goodwin, M. S., Lord, C., & Narayanan, S. S. ,2016). Wall et al. used ADOS (Autism Diagnostic Observation Schedule) for detecting ASD (Wall et al., 2012). The researchers found that ADOS's effectiveness is related to complete administration; otherwise, it does not produce desired results.

Wall et al. (2012) in their famous article, "Use of machine learning to shorten observation-based screening and diagnosis of autism" discussed the significance of ADOS (The Autism Diagnostic Observation Schedule-Generic). ADOS, for autism's behavioral evaluation, is one of the most tested models. It consisted of four modules each designed for particular patients or groups according to communication and developmental levels. The duration of a module is approximately 30 to 60 minutes. They tested their proposed model along with ML algorithms on 612 patients. They found that machine learning improves results and produces 100% results in predicting autism (Wall et al., 2012).

2.2 Main outcomes and learnings from the literature review:

- Autism spectrum disorder (ASD) is a complicated condition which impairs a person's capability to act appropriately, both in verbal and non-verbal communication.
- To reduce the number of obstacles an ASD diagnosed individual will face, it is vital to diagnose and treat the condition as early as possible.

- Several research studies show and effort in enhancing and expediting the ASD diagnostic process using machine learning.
- The traditional standardized method of testing for ASD takes longer amount of time and at a higher cost.
- Healthcare practitioners are increasingly turning to machine learning strategies in an effort to enhance the accuracy and efficiency of the diagnostic process.
- Some statistical models were developed such as Support Vector Machines (SVM), Random Forest Classifier (RFC), Naive Bayes (NB), Logistic Regression (LR), and KNN.
- Some researchers also used techniques from the field of computational models, such as probabilistic reasoning, artificial neural networks (ANN), and classifier combination.
- There is an alternative hypothesis that for a machine learning model to have higher classification accuracy with fewer feature subsets.
- A Study used a total of eight different machine learning techniques, some of which include the stepwise backward identification of features on score sheets from 4540 different people. It identifies an autism spectrum disorder risk by using 9 of the 28 behaviors it resulted in 98.27 percent and 97.66 percent, respectively.
- Other researchers also used machine learning techniques; forty kinematic constraints were retrieved from eight different imitation situations in their study. This study demonstrates that it is possible to use machine learning approaches to the analysis of high-dimensional data and the diagnostic categorization of autism in adults.
- Another study also shows that early diagnosis of ASD is critical. They stated that machine learning techniques are very useful in medical diagnosis. They analyzed the use of Naïve Bayes, Convolutional Neural Network, Neural Network, KNN, Logistic Regression, and Support Vector

Machine for autism diagnosis. They collected children and adult datasets and tested them with all the techniques. They found 96.88%, 98.30%, and 99.53% accuracy.

- A study concluded that SVM and SMO (Sequential Minimal Optimization) are the best methods for ASD diagnosis.
- The advancement of machine learning can make ASD diagnosis more cost-effective and less time consuming. It was proposed that a prediction model that can help develop a mobile app to predict ASD. The model is a combination of Random Forest-CART, and Iterative Dichotomiser 3 (Random Forest-ID3). They found that their proposed model is effective and accurate in predicting ASD.
- Due to the prevalence and continual growth of ASD, these professionals are fully occupied and charged expensively. In critical, or even in general situations, an early diagnosis of Autism is beneficial both for the patients and families
- The article "Enhancing diagnosis of autism with optimized machine learning models and personal characteristic data" concluded that machine learning models are far better and more cost effective than traditional clinical methods for predicting ASD .

Chapter 3- Project Description

3.1 Data Pre-processing

To combine the Datasets using WEKA, which is a open-source programming language often used in data analytics projects.

As shown in figure 3.1 The function Rbind() was used to make up the dataset named Autism.diagnosis1. Then to validate that all the rows were combined, the number of rows for each dataset was checked using the nrow() function. The total ended up being 1,103 rows which is correct. Now that the datasets are combined, the processes of data cleaning will begin. As a further step to validate this combination, figure 3.2 will show a summary of the Autism.diagnosis1 dataset.

In this dataset, there is a total of 21 dimensions. It is vital to avoid the curse of dimensionality. The curse of dimensionality is a phenomenon in the data analytics and machine learning field. Although more data is better, it also will increase redundancy and errors during analysis. On the same note, according to Hughes the performance of a predictive model will increase as the dimensions increase, however there is a limit to this (Karanam, 2021) This phenomenon is clearly depicted in the image below:

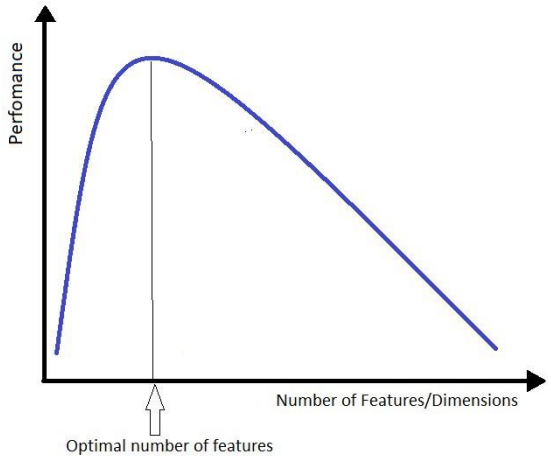

Figure 3.1.1: **Graphical Representation of Hughes Principle**

Figure 3.1.2

```
>
> Autism.diagnosis1 <- rbind(Autism.Child.Data.arff,Autism.Adolescent.Data.arff)
> Autism.diagnosis1 <- rbind(Autism.diagnosis1,Autism.Adult.Data.arff)
> nrow(Autism.Child.Data.arff)
[1] 293
> nrow(Autism.Adolescent.Data.arff)
[1] 105
> nrow(Autism.Adult.Data.arff)
[1] 705
> nrow(Autism.diagnosis1)
[1] 1103
```

Figure 3.1.3

```
> summary(Autism.diagnosis1)
      V1                 V2                 V3                 V4                 V5                 V6                 V7
 Length:1103        Length:1103        Length:1103        Length:1103        Length:1103        Length:1103        Length:1103
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
      V8                 V9                V10                V11                V12                V13                V14
 Length:1103        Length:1103        Length:1103        Length:1103        Length:1103        Length:1103        Length:1103
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
     V15                V16                V17                V18                V19                V20                V21
 Length:1103        Length:1103        Length:1103        Length:1103        Length:1103        Length:1103        Length:1103
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

3.2 Data Cleaning

To begin the process of data cleansing, the function is.na () was used in order to show the missing values. Upon viewing the data, it has been noticed that fields with missing values have the symbol “?” in them. In order to allow the is.na() function to work and show TRUE for missing values, the “?” needed to be replaced with NA. This was done by using the following line of code: Autism.diagnosis1[Autism.diagnosis1 == "?"] <- NA. This resulted in a total of 288 missing data points from 2 attributes as shown in figure 3.3. The attributes with missing data points are ethnicity and who is filling the questionnaire. The missing values were omitted as shown in figure 3.4. It was decided to omit these missing values as it is difficult to impute any personal information. According to the medical article Imputing Race/Ethnicity: Part 1 it is common for people to leave ethnicity-related questions blank. There are methods of imputing race or ethnicity related values, however they do have short comings in terms of validity. It could also result in bias due to misclassifications. It was concluded in this article that it is not advised to impute this sort of data and it could result in poor outcomes (Lines & Humphrey, 2021). Based on this, it was decided, for accuracy and quality’s sake, to omit the missing values rather than attempt to impute them

Figure 3.2.1

```
> colSums(is.na(Autism.diagnosis1))
 V1  V2  V3  V4  V5  V6  V7  V8  V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20 V21
  0   0   0   0   0   0   0   0   0   0   6   0 144   0   0   0   0   0   0 144   0
```

Figure 3.2.2

```
> Autism.diagnosis2 <- na.omit(Autism.diagnosis1)
> View(Autism.diagnosis2)
> colSums(is.na(Autism.diagnosis2))
 V1  V2  V3  V4  V5  V6  V7  V8  V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20 V21
  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
>
```

Another step of data pre-processing and cleansing would be to convert the results for the Final variable, v21 (The result of ASD occurrence) from YES and NO to 1 and 0. The justification for this is that some ML techniques require the output to be numeric. The line of code below was used for this:

Figure 3.2.3

```
> require(dplyr)
> Autism.diagnosis2 <- Autism.diagnosis2 %>%
+     mutate(V21 = ifelse(V21 == "NO",0,1))
```

Then it was noticed that the attributes are labeled as V1,V2,V3 ….. V21 which is not practical for analysis purposes. To overcome this, the following code was used, the output labeled the columns with their actual meaning rather than just v1,v2 and so on. The code below shows the replacement of the column name for the 21st column to “Class ASD”.

Example*: colnames(Autism.diagnosis1) [21] = "Class ASD"*

	A2 Score	A3 Score	A4 Score	A5 Score	A6 Score	A7 Score	A8 Score	A9 Score	A10 Score	Age	Gender	Ethnicity	Jaundice	Autism	Country of residence	Used application	R
2	1	0	0	1	1	0	1	0	0	6	m	Others	no	no	Jordan	no	5
3	1	0	0	1	1	0	1	0	0	6	m	'Middle Eastern '	no	no	Jordan	no	5
6	1	1	1	1	1	1	1	1	1	5	m	Others	yes	no	'United States'	no	1
8	0	1	1	1	1	0	1	0	1	5	m	White-European	no	no	'United Kingdom'	no	7
9	1	1	1	1	1	1	1	0	0	5	f	'Middle Eastern '	no	no	Bahrain	no	8

Figure 3.3.4

Additionally, some fields within the Ethnicity and Country of Residence attributes are inconsistent. For example, some countries or ethnicities’ have speech marks around the value such as ‘Middle Eastern’ and some don’t. Due to this, these values require to be altered to be consistent.

3.3 Data Modelling

The3dataset has been split into two parts, training and testing, where 75% is training and 15% testing. This split is depicted in the diagram below.

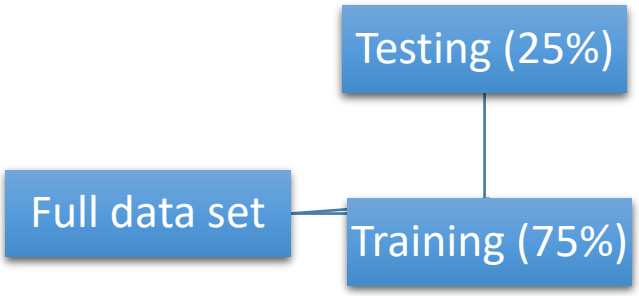

Figure 3.3.1

The chunk below shows the code used to split the dataset into training and testing subsets. Set seed was used to ensure the testing and training split does not change every time the code is run.

```
```{r}
# make the partion of the dataset in training 75% and testing 25%

set.seed(100)

partition <- sample(2,nrow(Autism.data2),prob = c(0.75,0.25),replace = T)
train <- Autism.data2[partition==1, ]
test  <- Autism.data2[partition==2, ]

# Dimensions of the dataset
dim(train)
dim(test)


```
```

Figure 3.3.2

3.4 Data Sources

The analysis will be based on a combination of 3 datasets from the UCI machine learning repository. The three datasets are; ASD Screening Data for Adult, ASD screening Data for Children and ASD Screening Data for Adolescent. These datasets are a categorical,

continuous, and binary data type. They have 21 common attributes and a total of 1,100 instances.

The attributes are as follows:

TABLE 1: Dataset Metadata

#	Attribute Name	Data type	Description/unit
1	Patient age	Number	Years
2	Gender	String	Male or Female
3	Ethnicity	String	A list of ethnicities
4	Born with Jaundice	Boolean	Yes or no
5	Any family member suffered from pervious development disorders	Boolean	Yes or no For immediate family members with PDD
6	Who is completing the test	String	Parent ,self, caregiver etc.
7	The country in which the user lives	String	A list of countries
8	Screening Application used by the user before or not?	Boolean	0 or 1
9	Screening test type	Integer (0,1,2,3)	0- Toddler 1- Child

			2- Adolescent 3- Adult
10-20	Based on the screening method answers of 10 questions	Binary (0, 1)	Based on answers to 10 questions
21	Screening score	Integer	Score

3.5 Data Visualization

This portion of the paper will begin a deeper discovery of this dataset. We will visualize general trends that occur in the dataset. This a vital stage as it will result in a better understanding that will enable us to model the data in the most accurate way possible. To visualize and analyze the data properly, we need to first normalize the categorical data into numeric data. This will be done using an *ifelse* function. For example, male and female will be normalized from m and f to 1 and 0.

Summary of ASD Cases

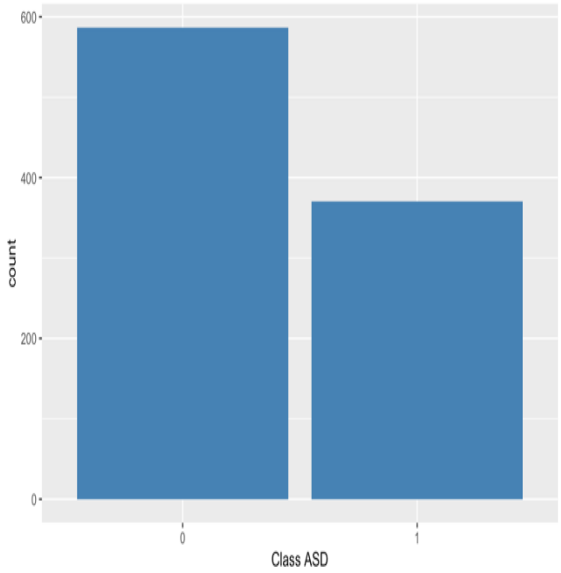

Figure 3.5.1

To start off with better understanding the data, as shown in the bar plot above, it is a summary of the count of cases of Autism in the dataset. The cases that went through this diagnosis process with a positive ASD (less than 400c cases) result is far lower than those that weren't diagnosed with Autism(around 600 cases) specifically, over 200 cases lower than those not diagnosed with ASD. This could be a result of several reasons. Some of which show that a significant portion of people that go through the ASD screening do not get diagnosed with ASD.

Count of diagnosis screening per Ethnicity

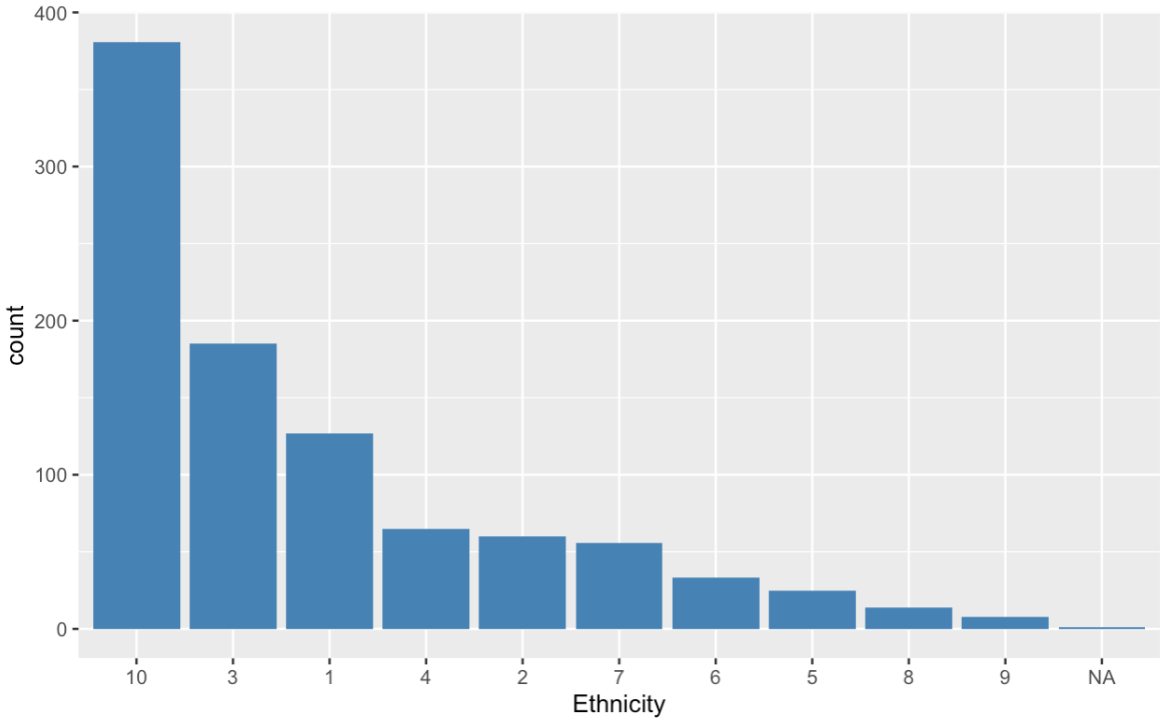

Figure 3.5.2

The Bar chart above shows the number of those that tested for Autism by ethnicity. Clearly, hose that have a White-European(10) ethnicity have the highest rate of ASD screening. This could be related to the accessibility and availability of these forms of screening for this ethnicity. It could also be an indicator of the level of awareness and education on the topic. As a result of these two components, the screening rate tends to be more frequent for White-Europeans(10). On the other side of the spectrum, those with a Hispanic(5), Paskifa(8) and Turkish(9) ethnicity show a lower count. This may be due to a lower wide spread of ASD and the importance of diagnosing it. According to a medical article, those that have a white ethnicity are more likely to

be identifies with autism whereas non-white and lower income children are less able to have access to early autism interventions at an early age where is it vital that is occurs. This is due is a distortion in the health equity because of socio-economic status (Aylward et al., 2021)

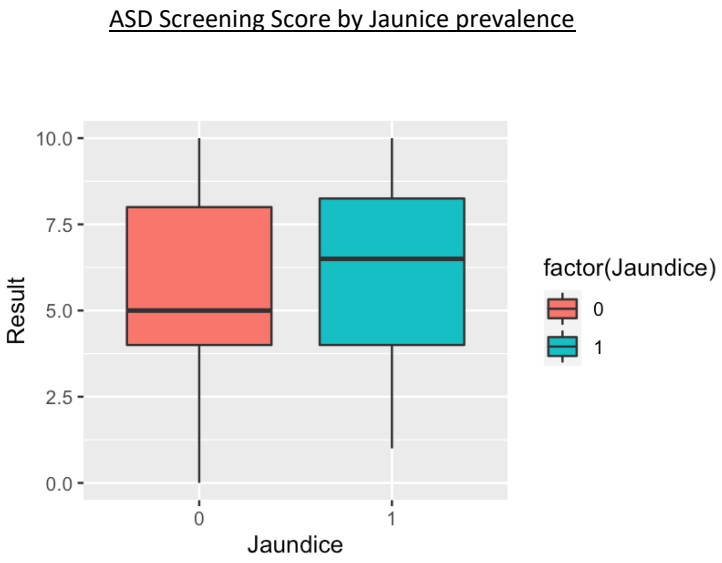

Figure 3.5.3

This boxplot shows the ASD screening score and the prevalence of Jaundice. The results show those that suffered with Jaundice in their earlier years have a higher screen score. The 1st, 2nd, and 3rd quartile of those that suffered from Jaundice in higher than those that did not. The higher the screening score the more likely they are to be diagnosed with Autism. A study by the name of “Neonatal jaundice and autism: Precautionary principal invocation overdue” depicts that the insufficient treatment of neonatal jaundice can lead to a higher risk of developing neurological development disorders (Wilde, 2022).

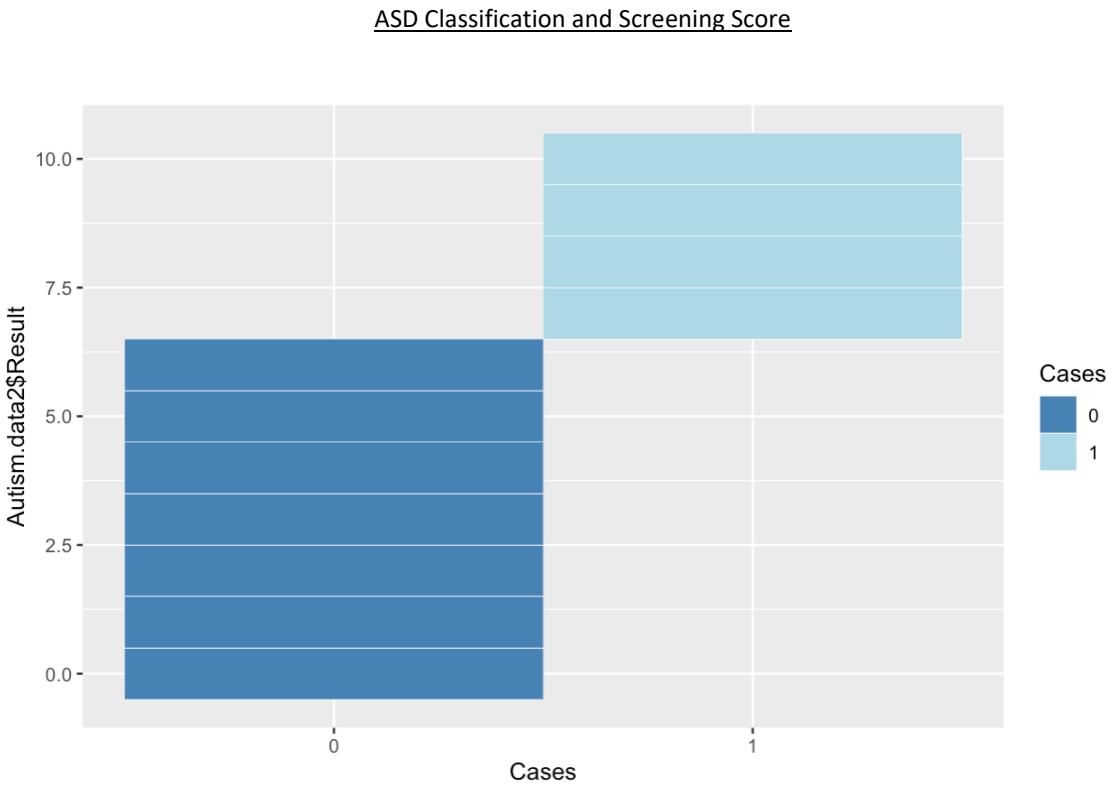

Figure 3.5.4

The visualization above shows the screening scores split into those that were diagnosed with autism (1) and those that weren't (0). The graph portrays and supports the fact that those that score higher in the screening test will more likely be diagnosed with autism then those that have a lower score. The threshold for being diagnosed with autism seems to be those with a score higher than 6. Whereas those that had a score of 6 or lower are not diagnosed with autism.

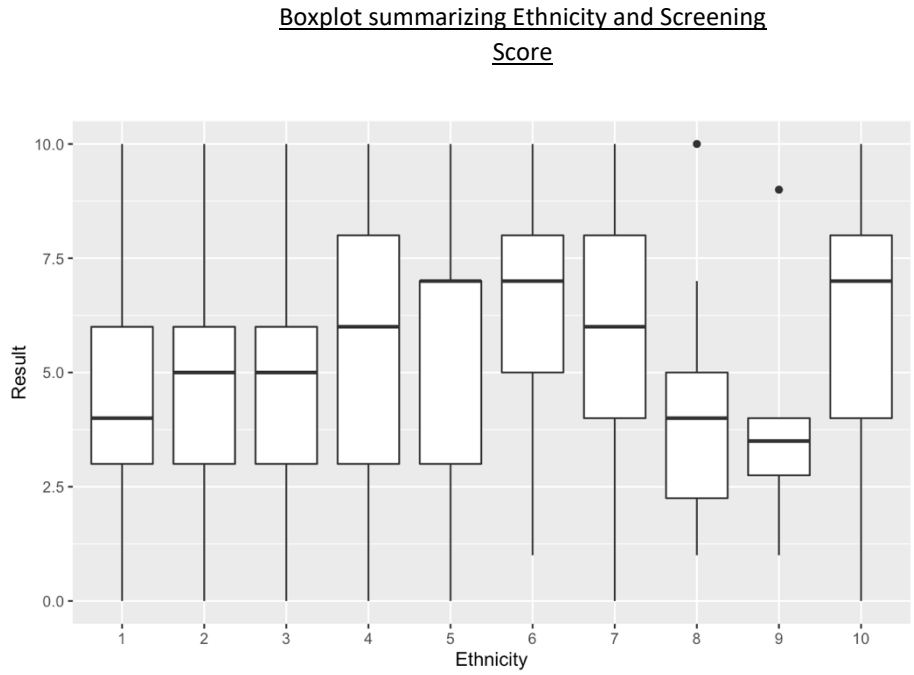

Figure 3.5.6

The boxplot show above summarizes the screening scores per ethnicity. The ethnicity with the lowest median at around 3.6 (Q2) is number 9 which represents those that have a Turkish ethnicity. On the other hand, the ethnicity with the highest median at around 6.6 is 6 and 10 which are Latinos and White-Europeans. Those with a Middle Eastern ethnicity have a high third quartile,6.3`, in comparison to their median which is around 4. Most of the Ethnicities have the same maximum, which is 10, the highest score possible. Ethnicity 8(Pasifika) and 9(Turkish) have some outliers that scored significantly higher in comparison to their respective third quartiles.

3.6 Data processing

The main prediction models that are used in this project are Logistics regression, Random Forest Classifier, Support Vector Machine and Naïve Bayes.

Logistics regression is used in cases where the dependent variable is considered categorical, an example could be whether a child has Autism or not or whether an email is spam or not. The resultant values for Logistics regression are strictly 0 or 1.

$$y = \frac{e^{(b_0 + b_1 X)}}{1 + e^{(b_0 + b_1 X)}}$$

Figure 3.6.1

The elements of the equation are:

- The symbol x = the input value
- The symbol Y= predicted output
- The symbol b0 = bias or interception point
- The symbol b1 = the coefficient for input x

There are several types of logistics regression. They are binary logistics regression which results in only two outcomes. On the other hand, Multinomial logistics regression can consist of three or more categories that are not ordered. Whereas ordinal logistics regression will result in three or more categories that are ordered, for example restaurant ratings (Swaminathan, 2018).

For the case of this study, we will be using binary logistics regression to predict whether a patient suffers from autism or not.

The Random Forest classifier is a combination of individual decision trees. This is the main reason why random forest classifiers work extremely well. Decision trees are the building blocks of random forest classifiers. Each decision tree within the random forest classifier comes up with the result of the dependent variable, then the result with the most occurrence then becomes the final prediction of the model (Yiu, 2019).

The Support vector machine (SVM) algorithm aims to find a hyperplane in an N-dimensional space, where N refers to the number of attributes). The hyperplane is required to clearly classify the data points. Hyperplanes are considered as boundaries for decision making to enable the model to predict the appropriate and accurate classes. The support vectors are those data points which are closer to the hyperplane. The size of the margin of the classifier depends on these support vectors (Gandhi, 2018).

The main objective of the Naïve Bayes classifiers is that it is a probabilistic method for machine learning used in classification. This classification model is built upon the Bayes theorem. The theorem states that we can find the probability of outcome A happening given that outcome B has already occurred. This model assumes that the attributes are independent from one another and do not affect each other. Therefore, it is called naïve. There are several types of Naïve Bayes classifiers. These include, multinominal, Bernoulli and Gaussian. For this particular project we will be using the Bernoulli naïve bayes as it results in predictions with two values (yes and no) (Gandhi, 2018).

Chapter 4- Project Analysis

4.1 Model 1: Logistics Regression

Fitting the training subset on the logistics regression model is shown in the code below:

```
```{r}
# Apply the Logistic Regression

LR = glm(as.factor(`Class ASD`)~., data=train%>%select(-`Country of res`), family='binomial')

summary(LR)

```
```

Figure 4.1.1

The fitted model was then applied to the testing subset and a confusion matrix was conducted to analyse the accuracy of the fitted model. The results of the confusion matrix are presented below:

The logistics regression model has an accuracy of 98.77%, an F1-score of 0.9 against the testing subset. Out of 244 observations, there were 156 true positives, 85 true negatives, 2 false negatives and 1 false positive. This shows that the model is seemingly accurate considering the fairy small size of the dataset.

```
```{r}

# Prediciton on the test data
pred1 <- predict(object = LR, newdata = test,type = 'response')

LOGpredict = ifelse(pred1 > 0.5, 1, 0)
LOGpredict<-as.factor(LOGpredict)

# Plot the confusion metrics and detailed results of the model
performance_model1 <- confusionMatrix(data = LOGpredict,as.factor(test$`Class ASD`),
                                      positive = "1",mode = "everything")
performance_model1


```
```

Figure 4.1.2

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 156   1
         1   2  85

               Accuracy : 0.9877
                 95% CI : (0.9645, 0.9975)
    No Information Rate : 0.6475
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9731

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9884
            Specificity : 0.9873
         Pos Pred Value : 0.9770
         Neg Pred Value : 0.9936
              Precision : 0.9770
                 Recall : 0.9884
                     F1 : 0.9827
             Prevalence : 0.3525
         Detection Rate : 0.3484
   Detection Prevalence : 0.3566
      Balanced Accuracy : 0.9879

       'Positive' Class : 1
```

4.2 Model 2: Random Forest Classifier

Fitting the training subset on the Random Forest Classifier is shown in the code below, the Country of residence variable had to be removed due to it causing errors in fitting the model:

```r
RF<-train(
  `Class ASD`~.,

  method = "rf",
  data=train%>%mutate(`Class ASD`=as.factor(`Class ASD`))%>%
                        select(-`Country of res`),
  tuneLength=10)

RF

plot(RF)
```

Figure 4.1.3

To fine tune this model, we set the tune length to 10 since the default was 3. The accuracy presented in the graph below to reach its peak after the 2nd tune. Therefore, there was no real need to increase the tune length further.

Figure 4.1.4

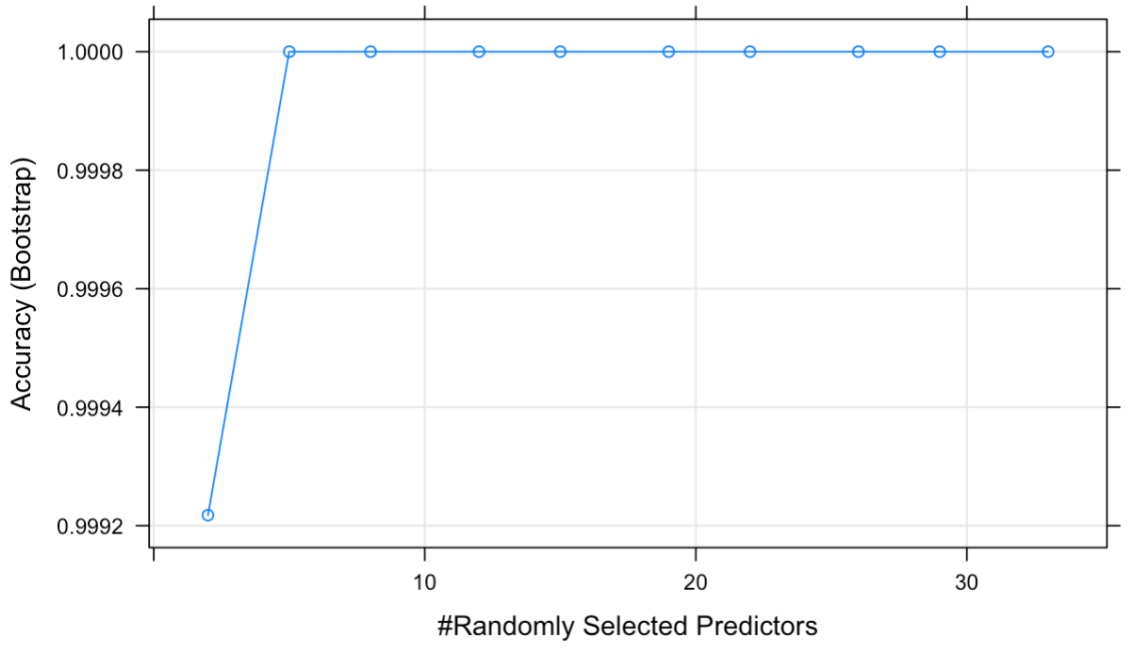

Figure 4.1.6

The fitted model was then applied to the testing subset and a confusion matrix was conducted to analyse the accuracy of the fitted model. The results of the confusion matrix are presented below:

```
```{r}
PredRF <- predict(object = RF, newdata = test%>%filter(complete.cases(.)))
TT<-test%>%filter(complete.cases(.))
performance_model2 <- confusionMatrix(data = PredRF,
                                      as.factor(TT$`Class ASD`),positive = "1",mode = "everything")
performance_model2
plot(PredRF)

```
```

Figure 4.1.7

The Random Forest Classifier resulted in an accuracy of 100% and an F1-score of 1 against the testing subset. Out of 244 testing observations, there were 158 true positives, 86 true negatives, 0 false negatives and 0 false positive. This indicates that the model is completely accurate. However, this may bring up the argument that this model could be over fitted. To further validate the accuracy, a larger testing sample needs to be acquired.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 158   0
         1   0  86

               Accuracy : 1
                 95% CI : (0.985, 1)
    No Information Rate : 0.6475
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
              Precision : 1.0000
                 Recall : 1.0000
                     F1 : 1.0000
             Prevalence : 0.3525
         Detection Rate : 0.3525
   Detection Prevalence : 0.3525
      Balanced Accuracy : 1.0000

       'Positive' Class : 1
```

Figure 4.1.8

4.3 Model 3: Support Vector Machine

Fitting the training subset on the support vector machine is shown in the code below, the Country of residence variable had to be removed due to it causing errors in fitting the model, the gride size was also required to be specified for the code to run smoothly:

```
gridsvm <-expand.grid(sigma = c(0.0577), C = c(2.21049))
SVMModel<-train(
  `Class ASD`~.,

  method = "svmLinear",
  data=train%>%mutate(`Class ASD`=as.factor(`Class ASD`))%>%
                      select(-`Country of res`),tuneGrid = expand.grid(C = seq(0, 2, length = 20)))
```

Figure 4.1.9

The fitted model was then applied to the testing subset and a confusion matrix was conducted to analyse the accuracy of the fitted model. The results of the confusion matrix are presented below:

```
```{r}
SVMPred <- predict(object = SVMModel, newdata = test%>%filter(complete.cases(.)))
TT<-test%>%filter(complete.cases(.))
performance_model2 <- confusionMatrix(data = SVMPred,
                                      as.factor(TT$`Class ASD`),positive = "1",mode = "everything")
performance_model2

```
```

4.1.10

The support vector machine model resulted in an accuracy of 100% and an F1-score of 1 against the testing subset. Out of 244 testing observations, there were 158 true positives, 86 true negatives, 0 false negatives and 0 false positive. This indicates that the model is completely accurate. However, it may also mean that over fitting might be playing a role here. To further validate the accuracy, a larger testing sample needs to be acquired.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 158   0
         1   0  86

               Accuracy : 1
                 95% CI : (0.985, 1)
    No Information Rate : 0.6475
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
              Precision : 1.0000
                 Recall : 1.0000
                     F1 : 1.0000
             Prevalence : 0.3525
         Detection Rate : 0.3525
   Detection Prevalence : 0.3525
      Balanced Accuracy : 1.0000

       'Positive' Class : 1
```

Figure 4.1.11

4.4 Model 4: Naïve Bayes

Fitting the training subset on the Naïve Bayed model using the e1071 library is shown in the code below:

```r
set.seed(42)
library(e1071)
#apply the Naive Bayes model
Naive_Bayes_Model=naiveBayes(`Class ASD` ~., data=train)
#Print the model outcomes
Naive_Bayes_Model

```

Figure 4.1.12

The fitted model was then applied to the testing subset and a confusion matrix was conducted to analyse the accuracy of the fitted model. The results of the confusion matrix are presented below:

```r
# Prediciton on the test data
pred <- predict(object = Naive_Bayes_Model, newdata = test)

# Plot the confusion metrics and detailed results of the model
performance_model <- confusionMatrix(data = pred,as.factor(test$`Class ASD`),
                                     positive = "1",mode = "everything")
performance_model

```

Figure 4.1.13

The Naïve Bayes model resulted in an accuracy of 95.51% and an F1-score of 0.9264 against the testing subset. Out of 244 testing observations, there were 153 true positives, 81 true negatives, 6 false negatives and 5 false positive. Despite the accuracy being high, when considering it against the rest of the prediction models used, it has the worst performance out of the four.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 153   5
         1   6  81

               Accuracy : 0.9551
                 95% CI : (0.9211, 0.9774)
    No Information Rate : 0.649
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9017

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9419
            Specificity : 0.9623
         Pos Pred Value : 0.9310
         Neg Pred Value : 0.9684
              Precision : 0.9310
                 Recall : 0.9419
                     F1 : 0.9364
             Prevalence : 0.3510
         Detection Rate : 0.3306
   Detection Prevalence : 0.3551
      Balanced Accuracy : 0.9521

       'Positive' Class : 1
```

Figure 4.1.14

4.5 Summary of the results

Although the results of the confusion matrices show that all the applied methods have high accuracy, the most accurate would be the Random Forest Classifier and the Support Vector Machine with an accuracy of 100% for both. Following that would be the logistics regression model which has an accuracy of 98.77%. The least effective model would be the Naïve Bayes model as it has the lowest accuracy which is 95.51%. Despite all the models having considerably high accuracy rates, expanding the dataset and applying it to the real would will only increase the accuracy of the predictions as the models will be fed with larger sets of training data.

Chapter 5 Conclusion

5.1 Conclusion

Considering that the diagnosis of Autism can be timely and costly for those in need. The models applied in this project resulted in high accuracy and low error rates. This gives an indication that the models can successfully diagnose patients with autism in the real-world settings. This would mean in ASD diagnosis procedures that would be more time and cost efficient. With that, this can result in a higher probability of those diagnosed with autism to go ahead with getting the required treatments without any further delay to this process. On the other hand, those that may be misdiagnosed with autism can then run further tests with their healthcare professionals to ensure that they have the correct diagnosis and treatment to their condition.

5.2 Recommendations

The main recommendation would be to expand the dataset in order to better train the models and have more observations to test against the fitted model. Additionally, if real-life cases were able to get obtained from healthcare professionals as testing subsets that will enable better cross-validation of the models as well as increase the confidence of healthcare professionals when used machine learning models to diagnose patients.

5.3 Future Work

Considering the limitations and recommendations that resulted from this study, future work could consist of obtaining a larger dataset to build the prediction tools on. This may require collaboration with health institutions to share data and results for exchange of the results of this study. Moreover, future work will have to include putting these prediction tools through a test on real world cases alongside healthcare professionals and comparing the results produced by the machine learning models against the results produced by the healthcare professional in the traditional methods.

Bibliography

Al Banna, M., Ghosh, T., Taher, K., Kaiser, M. and Mahmud, M., 2022. A Monitoring System for Patients of Autism Spectrum Disorder Using Artificial Intelligence. 13th International Conference on Brain Informatics.

Deshpande, G., Libero, L., Sreenivasan, K., Deshpande, H. and Kana, R., 2013. Identification of neural connectivity signatures of autism using machine learning. Frontiers in Human Neuroscience, 7.

Duda, M., Ma, R., Haber, N. and Wall, D., 2016. Use of machine learning for behavioral distinction of autism and ADHD. Translational Psychiatry, 6(2), pp.e732-e732.

Kosmicki, J., Sochat, V., Duda, M. and Wall, D., 2015. Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. Translational Psychiatry, 5(2), pp.e514-e514.

Li, B., Sharma, A., Meng, J., Purushwalkam, S. and Gowen, E., 2017. Applying machine learning to identify autistic adults using imitation: An exploratory study. PLOS ONE, 12(8), p.e0182652.

Parikh, M., Li, H. and He, L., 2019. Enhancing Diagnosis of Autism With Optimized Machine Learning Models and Personal Characteristic Data. Frontiers in Computational Neuroscience, 13.

Pratap, A., Kanimozhiselvi, C., Vijayakumar, R. and Pramod, K., 2014. Soft Computing Models for the Predictive Grading of Childhood Autism- A Comparative Study. International Journal of Soft Computing and Engineering (IJSCE), 4(3).

Thabtah, F. and Peebles, D., 2019. A new machine learning model based on induction of rules for autism detection. Health Informatics Journal, 26(1), pp.264-286.

Vaishali, R., and R. Sasikala. "A machine learning based approach to classify Autism with optimum behaviour sets. (2018) " International Journal of Engineering & Technology 7(4):

Vakadkar, K., Purkayastha, D. and Krishnan, D., 2021. Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques. SN Computer Science, 2(5).

Virolainen, S., Hussien, W., & Dalibalta, S. (2020). Autism spectrum disorder in the United Arab Emirates: potential environmental links. Reviews on environmental health, 35(4), 359–369. https://doi.org/10.1515/reveh-2020-0025

World Health Organization. (n.d.). Autism. World Health Organization. Retrieved September 3, 2022, from https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders

Aylward, B. S., Gal-Szabo, D. E., & Taraman, S. (2021). Racial, ethnic, and sociodemographic disparities in diagnosis of children with autism spectrum disorder. Journal of developmental and behavioral pediatrics : JDBP. Retrieved October 23, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8500365/

Wilde, V. K. (2022, February 23). Neonatal jaundice and autism: Precautionary principle invocation overdue. Cureus. Retrieved October 23, 2022, from https://www.cureus.com/articles/68849-neonatal-jaundice-and-autism-precautionary-principle-invocation-overdue

Erkan, U., & Thanh, D. N. H. (2019). Autism spectrum disorder detection with machine learning methodsUğur Erkan. Retrieved October 24, 2022, from https://www.researchgate.net/profile/Ugur-Erkan/publication/337195907_Autism_Spectrum_Disorder_Detection_with_Machine_Learning_Methods/links/5e0dea544585159aa4ac2c67/Autism-Spectrum-Disorder-Detection-with-Machine-Learning-Methods.pdf

Raj, S., & Masood, S. (2020, April 16). Analysis and detection of autism spectrum disorder using machine learning techniques. Procedia Computer Science. Retrieved October 24, 2022, from https://www.sciencedirect.com/science/article/pii/S1877050920308656

Hossain, Md & Kabir, Ashad & Anwar, Adnan & Islam, Md. (2021). Detecting autism spectrum disorder using machine learning techniques. Health Information Science and Systems. 9. 10.1007/s13755-021-00145-9.

Islam, Muhammad Nazrul & Omar, Kazi & Mondal, Prodipta & Khan, Nabila & Rizvi, Md. (2019). A Machine Learning Approach to Predict Autism Spectrum Disorder. 10.1109/ECACE.2019.8679454.

Bone, D., Bishop, S. L., Black, M. P., Goodwin, M. S., Lord, C., & Narayanan, S. S. (2016). Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion. Journal of child psychology and psychiatry, and allied disciplines, 57(8), 927–937. https://doi.org/10.1111/jcpp.12559

Thabtah, F., & Peebles, D. (2020). A new machine learning model based on induction of rules for autism detection. Health informatics journal, 26(1), 264–286. https://doi.org/10.1177/1460458218824711

Parikh, M. N., Li, H., & He, L. (2019, January 1). Enhancing diagnosis of autism with optimized machine learning models and personal characteristic data. Frontiers. Retrieved October 24, 2022, from https://www.frontiersin.org/articles/10.3389/fncom.2019.00009/full

Wall, D. P., Kosmicki, J., DeLuca, T. F., Harstad, E., & Fusaro, V. A. (2012, April 10). Use of machine learning to shorten observation-based screening and diagnosis of autism. Nature News. Retrieved October 24, 2022, from https://www.nature.com/articles/tp201210

Swaminathan, S. (2018, March 15). Logistic regression — detailed overview - towards Data Science. towardsdatascience. Retrieved November 19, 2022, from https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

Yiu, T. (2019, July 12). Understanding random forest - towardsdatascience.com. towardsdatascience. Retrieved November 19, 2022, from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Gandhi, R. (2018, June 7). Support Vector Machine — introduction to machine learning algorithms ... towardsdatascience. Retrieved November 19, 2022, from https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Gandhi, R. (2018, May 5). Naive Bayes classifier. what is a classifier? | by Rohith Gandhi ... towardsdatascience. Retrieved November 19, 2022, from https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

Lines, L., & Humphrey , J. (2022, June 27). Imputing race/ethnicity: Part 1. RTI. Retrieved November 19, 2022, from https://www.rti.org/insights/imputing-raceethnicity-part-1