

NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

AN AUTONOMOUS INSTITUTION AFFILIATED TO VTU, BELGAUM, ACCREDITED by NAAC ('A' Grade)

YELAHANKA, BANGALORE-560064



KNOWLEDGE • CHARACTER • UNITY

A Project Report on

DETECTION OF AUTISM SPECTRUM DISORDER USING MACHINE LEARNING

Submitted in partial fulfilment of the requirement for the award of the degree of

BACHELOR OF ENGINEERING

IN

INFORMATION SCIENCE AND ENGINEERING

By

Arya Dev	1NT15IS018
B S Anirudh	1NT15IS021
Milind Kulgod	1NT15IS051
P Vishnu	1NT15IS065

Under the Guidance of

Mrs. Disha D N
Assistant Professor
Information Science and Engineering
Nitte Meenakshi Institute of Technology



DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

(Accredited by NBA Tier-1)

2018-19

NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY
DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

Accredited by NBA Tier-1



CERTIFICATE

This is to certify that the project entitled **DETECTION OF AUTISM SPECTRUM DISORDER USING MACHINE LEARNING** is a bonafide work carried out by **Arya Dev(1NT15IS018)**, **B S Anirudh(1NT15IS021)**, **Milind Kulgod(1NT15IS051)** and **P Vishnu(1NT15IS065)** in fulfilment for the award of the degree Bachelor of Engineering in Information Science and Engineering of Visveswaraya Technological Institute, Belgaum during the year 2018-2019. It is certified that all corrections/suggestions indicated for internal assessment have been incorporated in the report deposited in the departmental library. The project report has been approved as it satisfies the academic requirements of the project work prescribed for the Bachelor of Engineering degree.

Internal Guide

Mrs. Disha D N
Asst. Professor, Dept. of ISE
NMIT, Bangalore

Head Of Department

Dr. Sanjay H A
Dept. of ISE
NMIT, Bangalore

Dr. H. C. Nagaraj

Principal
NMIT, Bangalore

External Viva

Name of the examiners

Signature with date

- 1.
- 2.

NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

Accredited by NBA Tier-1



DECLARATION

We, **Arya Dev (1NT15IS018)**, **B S Anirudh (1NT15IS021)**, **Milind Kulgod (1NT15IS051)** and **P Vishnu(1NT15IS065)**, bonafide students of Nitte Meenakshi Institute of Technology, hereby declare that the project entitled "**DETECTION OF AUTISM SPECTRUM DISORDER USING MACHINE LEARNING**" submitted in fulfilment for the award of Bachelor of Engineering in Information Science and Engineering of the Visvesvaraya Technological University, Belgaum during the year 2018-2019 is our original work and the project has not formed the basis for the award of any other degree, fellowship or any other similar titles.

Name and Signatures of the Students with Date

- 1)
- 2)
- 3)
- 4)

Place: Bangalore

Date:

ACKNOWLEDGEMENT

The satisfaction that one experiences after successfully completing any task only makes sense if everyone involved in the task is credited for their valuable contribution, support, guidance and encouragement throughout the course of the project. Those who have been the very backbone of this project, without whom, it's success would be far from becoming true.

We would like to thank the Management of Nitte Meenakshi Institute of Technology for providing a productive environment that motivated us for the successful completion of the report work.

It gives us immense pleasure to thank the Academic Dean of NMIT, Dr. Sridhar V for his support and encouragement.

We would like to thank the Head of ISE Department, Dr.Sanjay H.A for guiding us and bringing us on the right track throughout, till the very end.

We would like to express my gratitude to the Principal of NMIT Dr. H C Nagaraj for his motivation and making sure we never lost confidence in reaching our goals.

Also, We would like to express our gratitude to Mrs. Disha D N, Assistant Professor, Department of Information Science and Engineering for her constant support and guidance throughout the project, guiding us along the stepping stones towards the completion of the project.

We would also like to thank all other teaching and non-teaching staff of Information Science and Engineering, who has directly or indirectly helped us in the completion of the report.

Thank You

Arya Dev
B S Anirudh
Milind Kulgod
P Vishnu

ABSTRACT

Mental health is something that has to be taken care of with maximum priority. A developmental disorder is a disorder that comes into existence due to an abnormal or a retarded development, either physically, mentally or both, and the prediction of such disorders needs to be carried out very carefully as it is slightly tedious to diagnose it. It requires a careful analysis of numerous observations. Autism is one among them. There are various methods to detect autism, which also include expensive screening methods that are used in well equipped hospitals, among which, the most preferred method has been used, the questionnaire method. The data which is recorded is based on the responses provided by people who have either been diagnosed for Autism, or show behavioural symptoms of Autism. The main intention of this project is to make sure that Autism is diagnosed accurately, and not be mistaken for any other disorder. The treatment for a misdiagnosed disorder will be of no significance and will be rendered useless. In order to accurately achieve results, Machine Learning has been incorporated. With the usage of algorithms, the aim is to achieve an accurate diagnosis based on the answers received. The most accurate algorithm among all the algorithms used, will be selected as the optimum solution to detect autism.

Contents

1	INTRODUCTION	1
1.1	Motivation	1
1.2	Issues and Challenges	1
1.3	Problem Statement	2
1.4	Objective	2
1.5	Organization of Report	2
2	ARCHITECTURE DESIGN	3
3	LITERATURE SURVEY	5
3.1	Machine learning for early detection of autism using a parental questionnaire and home video screening	5
3.2	Efficient autism spectrum disorder prediction with eye movement: A machine learning framework	5
3.3	Development of a Machine Learning Algorithm for the Surveillance of Autism Spectrum Disorder	5
3.4	A machine learning approach for identification and diagnosing features of Neurodevelopmental disorders using speech and spoken sentences	6
3.5	Using Machine Learning for Detection of Autism Spectrum Disorder	6
3.6	A Machine Learning-based Method for Autism Diagnosis Assistance in Children	6
4	DATA PREPROCESSING	8
4.1	System and Software Specifications	8
4.2	Collection of Data	8
4.2.1	Data-set from the Internet	8
4.2.2	Real time Data Collection	9
4.3	Data Cleaning	9
5	IMPLEMENTATION	10
5.1	k-Means Clustering	10
5.2	Logistic Regression	10
5.3	Support Vector Machine	11
5.4	K-Nearest Neighbours	12
5.5	Naive Bayes	13

5.6	Random Forest	13
5.7	Real time Data Collection and Analysis	14
6	UNIT TESTING	15
7	RESULTS	16
7.1	Module 1 Data	16
7.2	Module 2 Data	22
8	CONCLUSION AND FUTURE WORK	29

List of Figures

2.1	Project Workflow	3
5.1	k-Means Clustering	10
5.2	Sigmoid Function	11
5.3	Logistic Regression	11
5.4	Support Vector Machine	12
5.5	Minkowski Distance	12
5.6	K Nearest Neighbours	13
5.7	Baye's Formula	13
5.8	Random Forest	14
7.1	Cleaned and Processed Data	17
7.2	K Means Clustering for Module 1 Data	17
7.3	Autistic and Not Autistic clustering	18
7.4	Accuracy obtained for Logistic Regression	19
7.5	Support Vector Machine for Module 1 Data.	20
7.6	SVM Graph with Hyper plane and intermediate values.	20
7.7	SVM Graph classifying into Autistic and Not Autistic.	21
7.8	SVM function to determine whether the individual is autistic or not. . .	21
7.9	Web page with Autism Questionnaire	22
7.10	Questions to assess Autistic individuals	23
7.11	Real time data-set	23
7.12	K Means Clustering for Module 2 Data	24
7.13	K Means Clustering for Module 2 Data	25
7.14	K Means Graph into Autistic and Not Autistic	25
7.15	K Nearest Neighbours accuracy for Module 2 Data	26
7.16	Logistic Regression accuracy for Module 2 Data	27
7.17	Donation of data-set to UCI	28

List of Tables

7.1	Accuracy of algorithms used for Module 1 and Module 2 Data	27
7.2	Accuracy of algorithms used for Module 2 Data only	27
7.3	Comparison of built-in and coded algorithms	28

Chapter 1

INTRODUCTION

1.1 Motivation

Developmental disorders are the type of disorders that need to be treated with the utmost care and precision has to be maintained when it comes to it's diagnosis. They directly affect the individual's behaviour, the way they interact with the outside world, the way they respond to situations. Autism is one among the several disorders that an individual can be diagnosed with, and can often be confused with Asperger's Syndrome and Attention-deficit/hyperactivity disorder(ADHD). There was an incident that took place in Bangalore, Karnataka, India, where an individual, who was actually autistic, and was exhibiting the symptoms that fell into the spectrum, but the doctor could not diagnose him in a short span of time, and eventually, was misdiagnosed. The treatment towards the wrongly diagnosed disorder, therefore, did not help the individual in any way, and no improvement was seen in the individual. After this observation, another doctor was approached, where the individual was properly diagnosed with Autism through expensive screening techniques. To make sure that this suffering should not be faced by anyone else, and provide a cost efficient, user friendly solution to the issue, this was taken as a motive for the project.

1.2 Issues and Challenges

The issues with this project is that the symptoms of Autism are not specific. ASD stands for Autism Spectrum Disorder. The term "spectrum" is used for the disorder, as the symptoms vary along a very broad range, hence prediction of Autism becomes difficult. Also, as Autism is a mental disorder, the acquiring of the information needed for the data-set to train the algorithm is extremely difficult, as the information is kept confidential by doctors and choose to not disclose the information about the patient as they consider the data as highly sensitive. Selecting of symptoms in order to yield efficient diagnosis is completely analysis based, and hence, is time consuming.

1.3 Problem Statement

Detection and Prediction of Autism Spectrum Disorder using Machine Learning algorithms.

1.4 Objective

- To create a budget friendly, efficient method to detect Autism and differentiate from other disorders with maximum accuracy using Machine Learning.
- To build a simple interface that can be used by anyone to collect data for initial screening purposes.
- To provide a budget friendly, self-assessment method to parents to detect Autism in an individual at initial stages before consulting a neurologist or a paediatrician.
- To donate the data-set to the UCI Repository for other researchers to utilize.

In order to obtain these objectives, the following steps have been taken:

- 1) **Data Preprocessing:** Cleaning of the data-set by removing outliers and junk values, to make the data-set efficient and understandable, thus, making it easy to feed the algorithms with the data and carry out an analysis and obtain the results.
- 2) **Feature Extraction:** Selection of the attributes that could be used to determine whether the person is autistic or not.
- 3) **Algorithm Training:** The data is fed into the algorithm after narrowing down the attributes that need to be utilized.
- 4) **Algorithm Testing:** Once the algorithm is trained, data is obtained from various sources and is fed into the algorithm to get the appropriate diagnosis with the accuracies and graphs.

1.5 Organization of Report

The report is organized in the following way:

Chapter 2: This chapter includes the literature survey of the research papers that have been done for the project and the various methods that could be used to detect Autism.

Chapter 3: This chapter includes the specifications of the system and the requirements, data acquisition and data cleaning.

Chapter 4: This chapter includes the algorithms that have been used for both supervised and unsupervised data and the way they interact with the data provided to them.

Chapter 5: This chapter includes the solution and the accuracy and graphical representation of the algorithm's outcome post training in the form of screenshots.

Chapter 6: This chapter denotes the scope this project has in the future and the way it can be improvised in order to perform more complex predictions

Chapter 2

ARCHITECTURE DESIGN

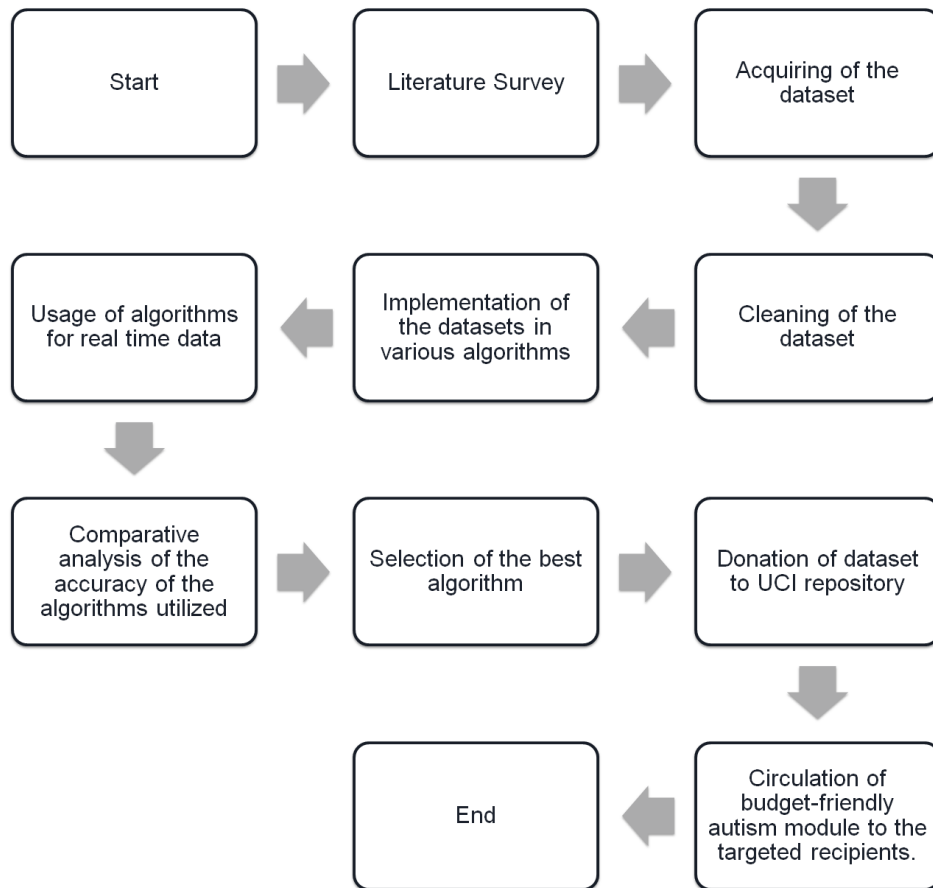


Figure 2.1: Project Workflow

The above flowchart represents the workflow of the project, right from the start to the end. The initial stages of the project was the carrying out of the literature survey, which involved the study of research papers that have already been published in order to gain insight, and the collection of data-set for computational purposes. The data-set had to be processed to make it understandable, and was fed into the algorithms. Post this event, the data was collected in real time and along with the used algorithms, another set of algorithms were implemented to bring about a comparison with the results presented in [6]. After obtaining the results from all the algorithms, the best among them all is selected to carry out the analysis. Due to the scarcity of data-sets online, the data-set which has been curated to collect data in real time, has been donated to the UCI Machine Learning repository. The prediction module will be circulated as a product to autistic schools, hospitals that are in need of the facility to screen patients for autism.

Chapter 3

LITERATURE SURVEY

3.1 Machine learning for early detection of autism using a parental questionnaire and home video screening

This paper was written by Halim Abbas et al. at the 2017 IEEE International Conference on Big Data (BIGDATA)[1]. The paper is about using machine learning to a standardized set of data which has been collected from over a thousand students who either have autism or have a high risk of becoming autistic to create a cost efficient and user-friendly autism screening tool that performs more efficiently than all the expensive high-standard screening equipment.. This new tool combines two screening methods into a single assessment, one based on short, structured questionnaires answered by parents, and the other based on tagging key behaviors from short videos of children. Further discussing about the challenge of extending machine learning algorithms to conditions beyond autism, propose a generalized framework for using machine learning algorithms extensively.

3.2 Efficient autism spectrum disorder prediction with eye movement: A machine learning framework

This paper was written by Wenbo Liu et al. and was published at 2015 International Conference on Effective Computing and Intelligent Interaction (ACII)[3]. The paper is about detection of Autism by observing the eye movement of an individual. This was done with the help of the image level feature extraction process. A machine learning based framework was proposed for facial detection to extract features. K-Means was used to cluster the information retrieved and represent the histogram using Bag of Words concept(BoW), and Support Vector Machine was proposed to maximise the margin of separating the positive and the negative data.

3.3 Development of a Machine Learning Algorithm for the Surveillance of Autism Spectrum Disorder

This paper was written by Matthew J. Maenner et al.[4] It's about how trained clinicians review developmental evaluations collected from multiple health and education sources

to determine whether the child meets the ASD surveillance case criteria or not. Data analysis has to be very careful as the data collected will exist in a vast spectrum. Text processing was used to calculate the frequency of the words in each individual's record and an analysis is carried out on that. There is a mention about using random forests method to accomplish two tasks. First task is to identify the subset of words and phrases that are important for classifying ASD. Second task is to build an Algorithm from useful words to perform the classification. Random Forests generate many independently-grown decision trees and the consensus vote of all trees forms the conclusion.

3.4 A machine learning approach for identification and diagnosing features of Neurodevelopmental disorders using speech and spoken sentences

This paper was written by Anjali Pahwa et al. for the International Conference on Computing, Communication and Automation (ICCCA2016)[5]. It's about identifying autistic traits with the help of interaction with people, and based on the way they communicate, the formation of sentences and fluency of speaking. There are distinctive features which help in detection of ASD between children with ASD and normal functioning children they are mainly pitch estimation, jitter and harmonic to noise ratio(HNR). Pitch estimation can be done using an approach that maximises the energy of the signal by eliminating the noise which might affect the accurate estimation . Jitter referred to as short term periodic variations in the glottal pulses during voice production. Harmonic to noise ratio is used by researchers for evaluating voice disorders. Any noisy disturbance can be subtracted from the original signal of speech by giving reconstructed signal as source.

3.5 Using Machine Learning for Detection of Autism Spectrum Disorder

This paper was written by Bram van den Bekerom[6] has made use of the data that has been retrieved from the National Survey of Children's Health(NSCH). The 10 fold cross validation is being used to determine the prediction of Autism, with the pretext of it being less biased, resulting in the division of the data-set into 10 parts. The algorithms Naive Bayes, Random Forest and Support Vector Machine have been used, to determine the most suitable algorithm for classification. The prediction of Autism was carried out in two ways, by dividing the entire data-set into two classes, and into four classes. The accuracies mentioned in the paper have been taken into consideration for comparison with the results that have been computed with the algorithms and the data-set being used in the current scenario.

3.6 A Machine Learning-based Method for Autism Diagnosis Assistance in Children

This paper was written by Sushama et al. [2] for the International Conference on Information Technology, 2017. This paper includes the the extraction of symptoms using

The Machine learning Association Rule, which is used for figuring out the relationship between two or more variables in a given database, and the Minimum Redundancy-Maximum Relevance method, which is used to select the right features. For the selection of additional symptoms to strengthen the set, Mutual Information, which measures the mutual dependency of the variables, has been used. With the help of data retrieved with the help of body sensors and smartphones, or the smartphone, the symptoms were extracted, additional symptoms were added in order to make the prediction more accurate by using the Highest Information Gain, which is the maximum decrease in the uncertainty in the result obtained, and the prediction of the disease was carried out with the help of machine learning.

Chapter 4

DATA PREPROCESSING

4.1 System and Software Specifications

In order to carry out this project, Windows 10 Operating System is being used. The programming language that is being used is Python.

A front-end web page that is used to take in real time data has been created using HTML5, CSS and JavaScript, integrated with a Google form.

The back-end used is Microsoft Excel as it is easy for data to be read from a CSV(Comma Separated Values) format file.

Jupyter Notebook via the Anaconda Navigator is used as the Python notebook and extensive use of the Python libraries are being made.

Machine Learning related packages are imported via Scikit-learn.

4.2 Collection of Data

4.2.1 Data-set from the Internet

Initially, the search for data in order to establish the data-set was a very tedious process. This is because Autism has a spectrum of symptoms to determine whether the individual is autistic or not. The selection of attributes for the data-set was difficult.

Neurologists, paediatricians and Head of Autistic schools were consulted to help us with curating the data-set. Due to the doctor's code of confidentiality, none of the doctors and medical institutions provided the data. Later, a data-set was found. It has been extensively used by a lot of people for their research. The data-set has been obtained from [textithttps://archive.ics.uci.edu](https://archive.ics.uci.edu), which is a machine learning data repository, and the Autism data-set was created by Fadi Fayez Thabtah, which consisted of around 700 instances, which was sufficient to train the machine learning algorithms used for the project.

4.2.2 Real time Data Collection

After the data was collected from the UCI repository, in order to make the algorithm usable with the data that exists in the real world, data was collected from peers and was circulated to the consulted doctors, and also to the autistic centers. This was made possible by hosting a web page online, which consisted of a questionnaire i.e a set of questions which had exactly two answers, yes and no. The questionnaire was formed in such a way, it included all the attributes that could determine if the individual is autistic or not. The responses were recorded onto an Excel sheet, where the final score was calculated and fed into the algorithm for testing purposes.

4.3 Data Cleaning

Data Cleaning is very important as it makes it easier to feed the algorithm with data and achieve the expected results. When any form of a data-set is created or retrieved, the data is not perfect. It contains values that are not essential for the computation of results, and there are values which are most needed for the computation of results, but do not exist in the data-set. The data-set had to be restructured. In the data-set, all the missing values were filled with the mean of all the values in the attribute's column and all the duplicate values were merged together. Outlier values were also eliminated from the data-set, and the attributes that did not make a difference to the computational process were removed, leaving with a limited number of attributes which had complete records, thus making it convenient for the algorithm to train with.

Chapter 5

IMPLEMENTATION

5.1 k-Means Clustering

k-Means clustering is one of the most important algorithms that is used to solve clustering problems. The way the algorithm works is by classifying a given data-set into a certain number of clusters. The number of clusters is denoted by k , which means that the entire data-set is divided into k clusters with k centers. The clusters are being placed far away from each other in order to obtain maximum accuracy based on the data-set fed into the algorithm. Each point is placed to the nearest center. New centers have to be recalculated based on the clusters in the previous iteration. The k centers change their location until no more changes are possible. For the data-set, the attributes 'age', 'gender', 'jaundice' and 'result' have been used, to divide the data-set into two major clusters, namely Autistic and Non Autistic.

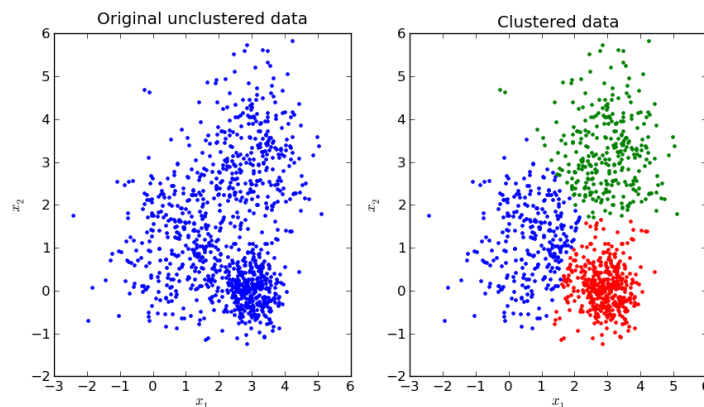


Figure 5.1: k-Means Clustering

5.2 Logistic Regression

Logistic Regression is a technique that has been borrowed by Machine Learning from statistics. It is the most widely used method for problems that involve binary classification. It measures the relationship between the dependent variable and independent

variables by estimating probabilities using a sigmoid function. It computes a weighted sum of the input variables and runs the result through a non-linear function i.e the sigmoid function in order to produce the result of the computation. With the help of logistic regression, the output of the algorithm can be converted into a class variable, i.e 0, which is 'no', and 1, which is 'yes', and for this the sigmoid function is used.

$$h = g(z) = \frac{1}{1 + e^{-z}}$$

Figure 5.2: Sigmoid Function

where e is the base of the natural logarithm and ' z ' is the actual numerical value that has to be transformed. In this scenario, we are taking in parameters from the data-set, in order to distinctively classify the data set into two classes, autistic and non-autistic. For the algorithm, the features selected are jaundice and the collective result of the answers given by the parents.

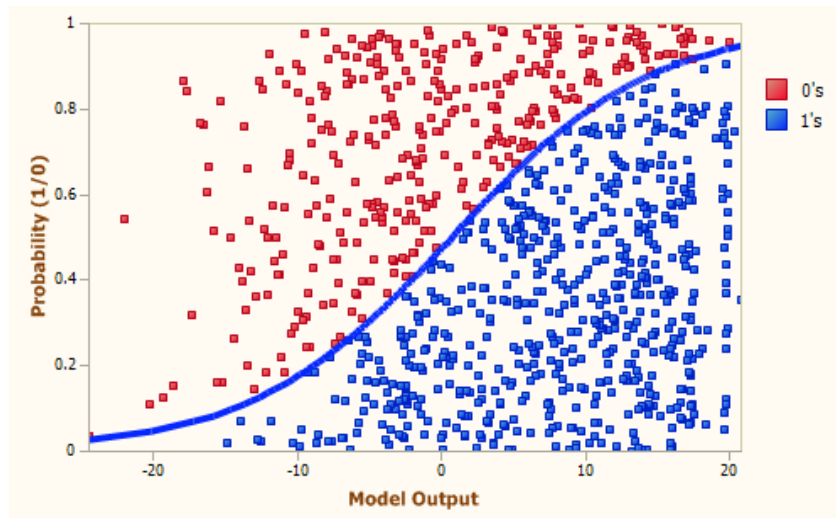


Figure 5.3: Logistic Regression

5.3 Support Vector Machine

The Support Vector Machine(SVM) is an algorithm that is very fast and a dependable algorithm which is used for classification and yields accurate results irrespective of the size of the data-set. Here, there are two major tags, i.e Autistic and Non-autistic. From the data-set, two main features are being considered, i.e 'jaundice' and 'result' attributes. An SVM takes these points and produces a hyper plane, which is a decision boundary that helps in the classification of the points, resulting in the splitting of the data into two classes, which are autistic and non autistic. Here, the output of the linear function is taken into consideration. If the result crosses 1, it defines one class, the rest are identified as the other class, which defines a range from -1 to 1. This is considered as the margin. A margin is the separation of the line to the closest class points. A margin is said to be good when the separation is large for both the classes. The points remain in the respective

classes without merging into the other class. A bad margin is when the separation is small and where there are points, which are supposed to be a part of one class, exists in the other.

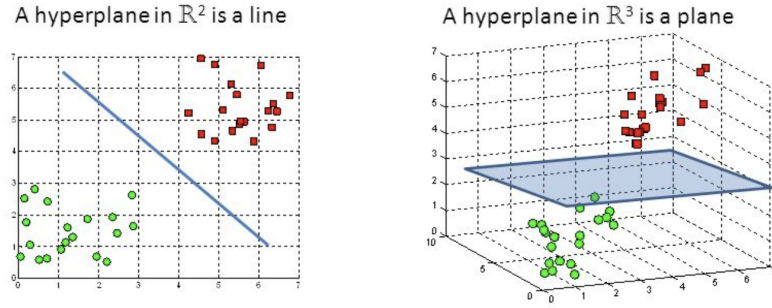


Figure 5.4: Support Vector Machine

5.4 K-Nearest Neighbours

This is an algorithm which is highly considered when it comes to classification in Machine Learning. It is an algorithm that is extensively used in applications like pattern recognition, intrusion detection and data analysis. It is widely flexible as it is non-parametric, which means it does not make any form of assumptions about how the data has been distributed. A part of the data is used to train the algorithm, where the classification of the points take place based on a specific attribute. The distance between two points on a graph defines the similarity among the points. Here, Minkowski Distance formula has been used to calculate the distance between two points.

$$\left(\sum_{i=1}^n |X_i - Y_i|^p\right)^{1/p}$$

Figure 5.5: Minkowski Distance

Where X and Y are the two points, and the value of $p=1$ if it's Manhattan distance and $p=2$ if it's Euclidean distance. For KNN, initially, the data-set has to be loaded, following which k has to be initialized to the number of chosen neighbours. For each example in the data-set, the distance has to be calculated between the query example and the current example that exists in the data-set. Then, the distance and index should be added to an ordered collection and sort it in ascending order. From the collection, the initial k number of entries are considered and the retrieval of labels takes place. To classify, the mode of the k labels is returned as the result. In this scenario, the attributes 'age' and 'result' have been used as input for the algorithm, from the data-set that has been collected through real-time input.

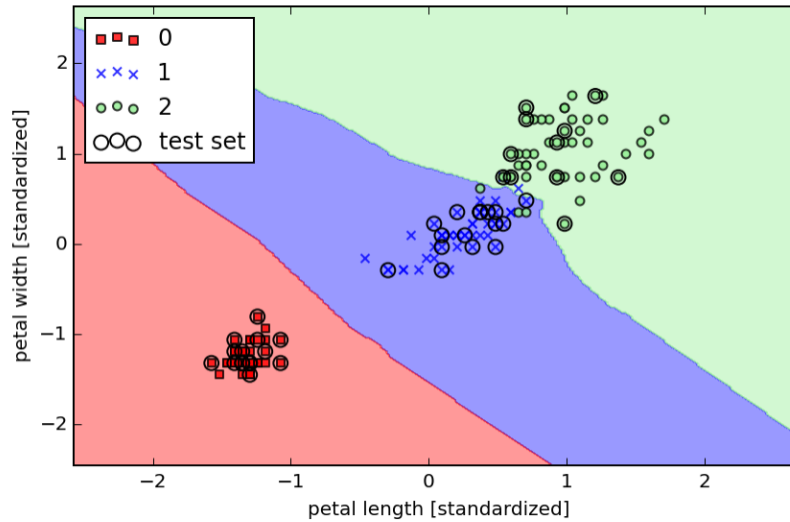


Figure 5.6: K Nearest Neighbours

5.5 Naive Bayes

This is a powerful algorithm which is very much considered for predictive modelling and this is suitable for a huge amount of data. This algorithm is considered for the analysis of text and its classification, filtering of emails and also prediction of diseases. It uses Bayes' Theorem of Probability in order to carry out prediction of classes. Naive Bayes makes the assumption that the influence of a certain feature in a class is purely independent of the other existing features. The computational process is simplified here. This assumption is called as conditional independence.

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$

Figure 5.7: Baye's Formula

Where $P(A)$ is the probability of hypothesis A being true irrespective of the data and $P(B)$ is the probability of the data. $P(A | B)$ is the probability of hypothesis A given the data B, and $P(B | A)$ is the probability of the data B given that the hypothesis of A is true. The classification of the data is purely based on its probability, and this methodology has been utilized for the Module 2 data that has been collected.

5.6 Random Forest

The Random Forest, also known as random decision forests, is an Ensemble Learning method which is used for tasks like classification and regression. It operates by creating decision trees during the time of training and gives out the mode of classes or the mean

prediction of each tree. Multiple trees are brought in together in order to obtain a more stable and accurate prediction.

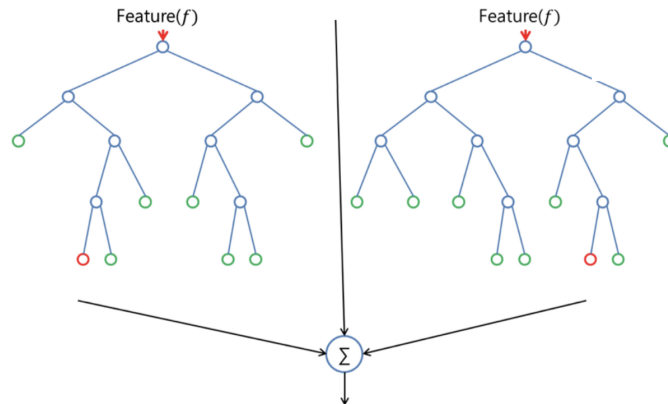


Figure 5.8: Random Forest

Randomness is added to the model, and it searches for the best feature among the random subset of features. A large number of trees are created. The training process might be quick, but the testing process is time consuming, due to the increasing number of trees being created.

5.7 Real time Data Collection and Analysis

This is where data is collected in real time by circulating a survey form among parents of autistic children, parents of non-autistic children, autistic individuals, non-autistic individuals, autistic schools within the city and other cities in the country. This step was enforced, with the intention to create a new subset with the attributes taken from the initial data-set, predict autism and diagnose it properly without any faults. The survey form was created using Google Forms, where the responses from the form would get reflected on Google Sheets, an Excel spreadsheet, immediately as it functions on the cloud. The Google Form was integrated into a website that was designed using HTML5, CSS and JavaScript. The responses were imported, cleaned and converted into a Comma Separated Value(CSV) file. The cleaned data was fed into the algorithms that were used for the initial data-set and the results were obtained. The URL for the web page is <https://shfi627krockdqpsxjrj4w-on.driv.tw/HTML%20FORM/FORM.html>. The link for the web page was circulated among the students of the institution, their peers, The Autistic Society of India, Bangalore, Perseverance Special Education School, Bangalore, and Bubbles Center for Autism, Bangalore in order to collect data with their consent. The link was also circulated to the doctors that were consulted for information about Autism and their traits to gather data.

Chapter 6

UNIT TESTING

The data which was initially acquired from the UCI repository was fed into the algorithms in order to identify the features which were insignificant. This was noticeable as there were errors being displayed due to the missing values and values that were not corresponding to the defined data type. The selection of the appropriate attributes for the algorithm were based on a trial and error approach and also with the consultation of doctors as well. This method narrowed down the attributes to be used and the same procedure was carried out for all the algorithms that were utilized. The main goal to be achieved was to get an accuracy more than what other researchers had obtained, and for that, the selection of the right algorithm, the selection of the right attributes, and the composition of the data-set mattered the most, as even the slightest change would bring about a difference in the final result. The final accuracies, which were obtained, were tabulated and compared.

Chapter 7

RESULTS

7.1 Module 1 Data

This section consists of the results that were computed with the data that was obtained from the UCI website, which is an open source data-set repository for Machine Learning, making it available for everyone. The data-set for the algorithm was extracted from <https://archive.ics.uci.edu/ml/data-sets/Autism+Screening+Adult>. This data-set had 704 entries and 21 attributes. Among the 21 attributes, the unnecessary attributes were removed as they only took up space and did not contribute to the computation of the data. The scores of 10 questions, Age, Gender, Ethnicity, Jaundice, Autism history, Country and Result were retained. All the spelling errors were corrected, the missing values were filled with the mean of all the data in that column. All the outliers were discarded. In this manner, the crude data-set was refined and transformed into a usable state. Figure 7.1 shows all the missing data have been substituted and all the irrelevant features or attributes have been removed.

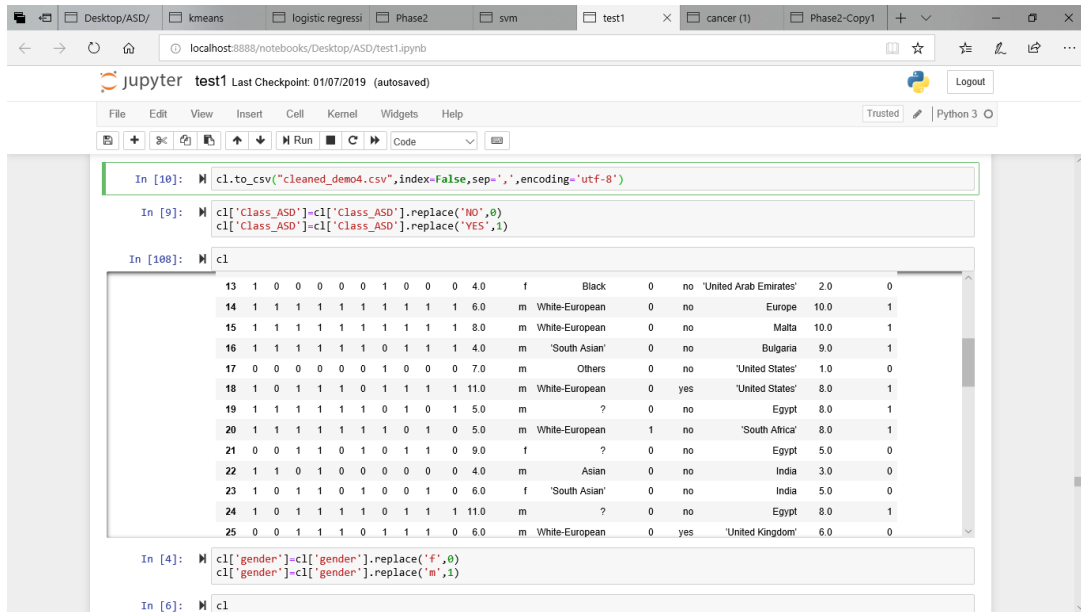


Figure 7.1: Cleaned and Processed Data

After the data is scrutinized and cleaned, k-means clustering was implemented. The features 'age', 'gender', 'jaundice' and 'result' were chosen from the database and was fed into the algorithm.

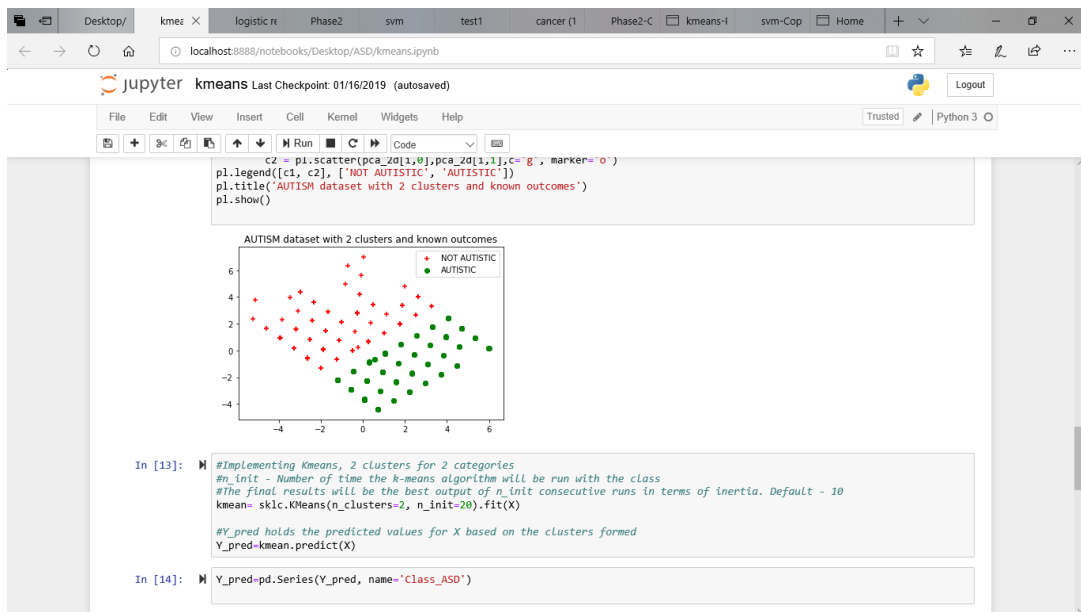


Figure 7.2: K Means Clustering for Module 1 Data

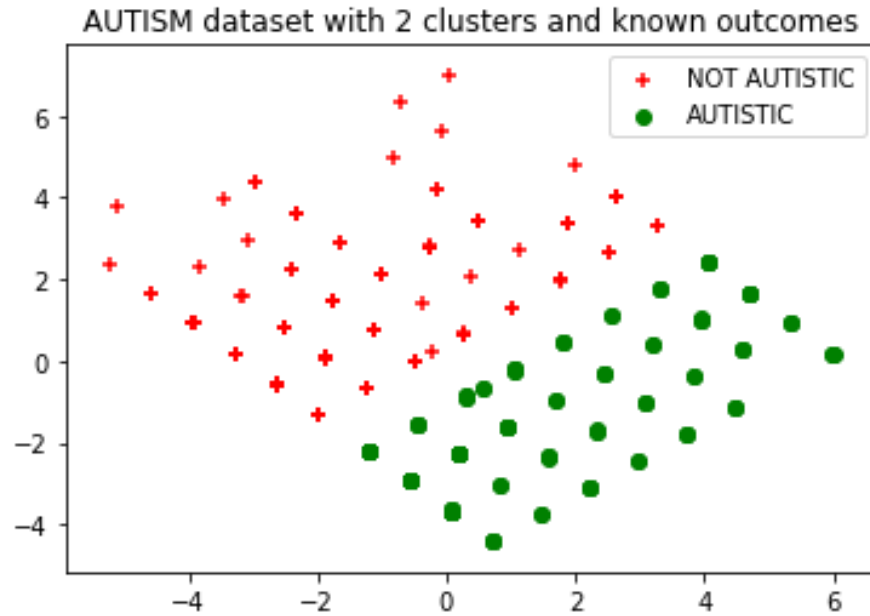


Figure 7.3: Autistic and Not Autistic clustering

The algorithm was successful in classifying into two main classes, namely Autistic and Non Autistic. The red dots represent the Autistic cluster and the green dots represent the Non Autistic cluster. There's a graphical representation of the outcome of the K Means Clustering.

Logistic Regression was used for the classification of the data-set as it was a binary classification problem. The original data-set was split into a training set and a testing set. The features that were used to train the algorithm are 'age', 'gender', 'jaundice' and 'result'.

```

In [3]: feature_cols=['age','result','jaundice','gender']
        x=lg[feature_cols]
        y=lg.Class_ASD

In [4]: #here we split the original dataset into training set and testing set.
        from sklearn.model_selection import train_test_split
        x_train,x_test,y_train,y_test=train_test_split(x,y,random_state=0)

In [5]: #train the model by importing LOGISTICREGRESSION and feed the data.
        from sklearn.linear_model import LogisticRegression
        logreg=LogisticRegression()
        logreg.fit(x_train,y_train)

Out[5]: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
        intercept_scaling=1, max_iter=100, multi_class='warn',
        n_jobs=None, penalty='l2', random_state=None, solver='warn',
        tol=0.0001, verbose=0, warm_start=False)

In [13]: y_pred_class=logreg.predict(x_test)

In [14]: from sklearn import metrics
        print(metrics.accuracy_score(y_test,y_pred_class))

0.9452854794528548

In [15]: y_test.value_counts()

```

Figure 7.4: Accuracy obtained for Logistic Regression

On training the algorithm and using the training the set, and feeding the testing set, an accuracy of 94.52% was obtained.

Finally, the Support Vector Machine(SVM) algorithm was implemented, which is the fastest, most efficient algorithm used for classification. A hyper plane is produced, which separates both the classes. The features that were selected for this was 'Jaundice' and 'result'. The hyper plane was produced and there was a separation of classes. Any score that ranged from 0 to 6 are classified as Non Autistic individuals and any score that exists beyond 6 are classified as Autistic. A function was created, that allows an individual to enter the values for the features 'jaundice' and 'result' in order to check if the algorithm is classifying it accurately or not, ending up with an accuracy of 83.3%.

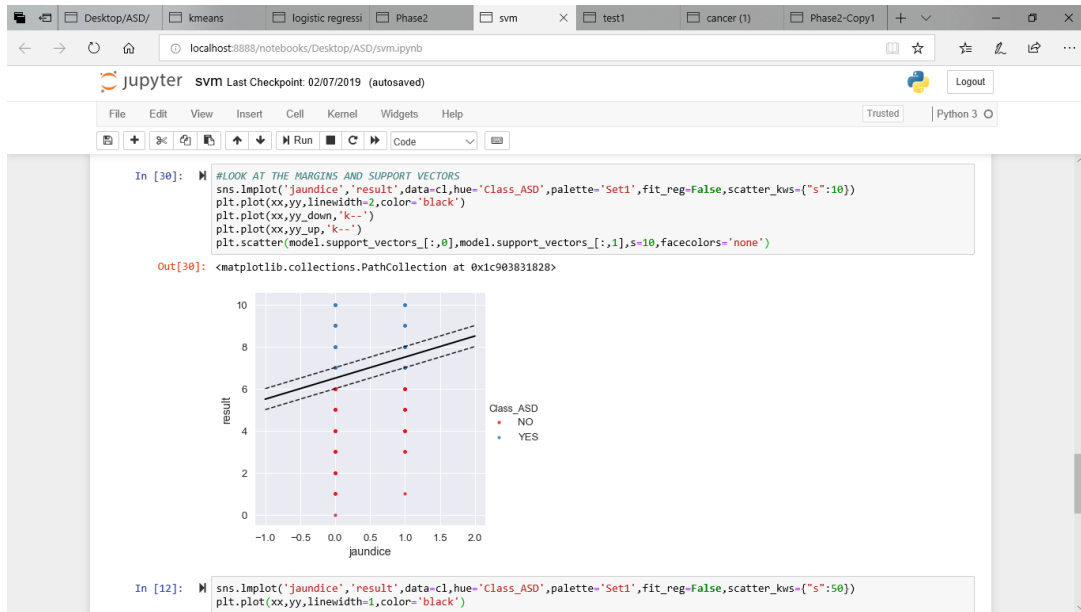


Figure 7.5: Support Vector Machine for Module 1 Data.

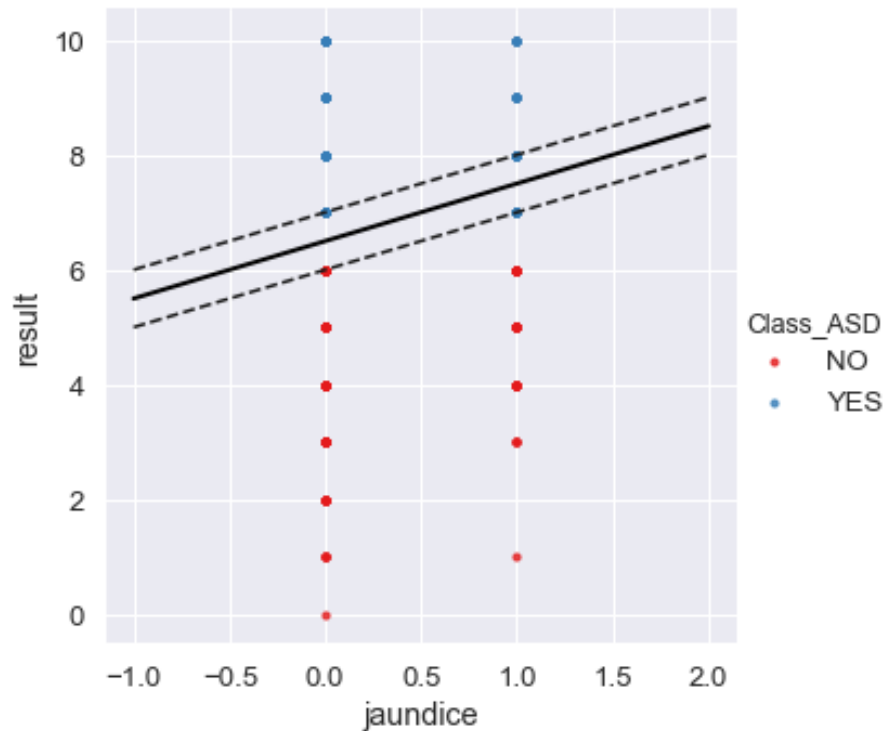


Figure 7.6: SVM Graph with Hyper plane and intermediate values.

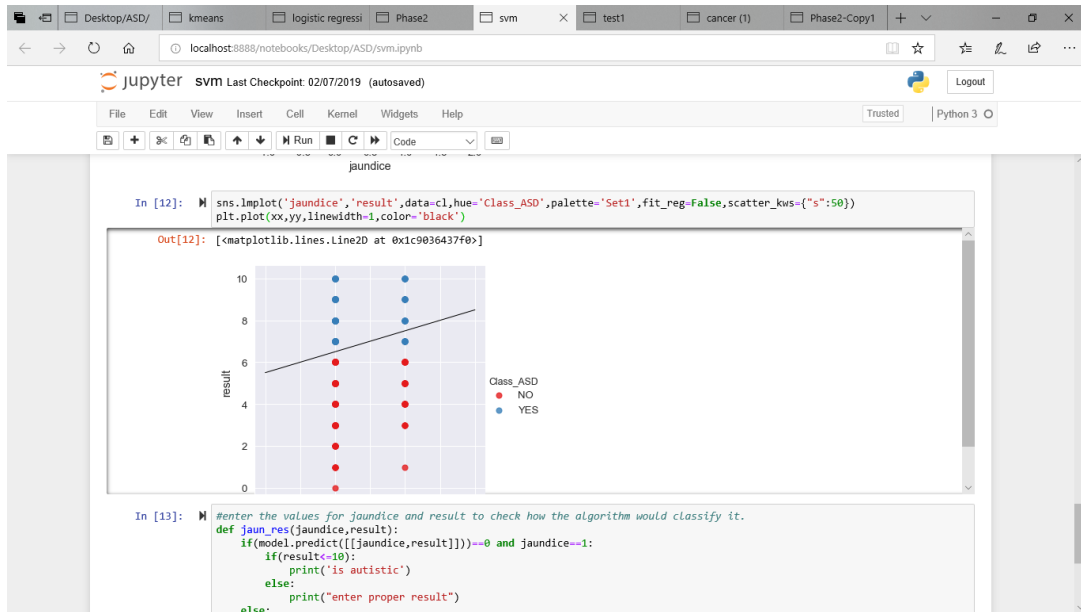


Figure 7.7: SVM Graph classifying into Autistic and Not Autistic.

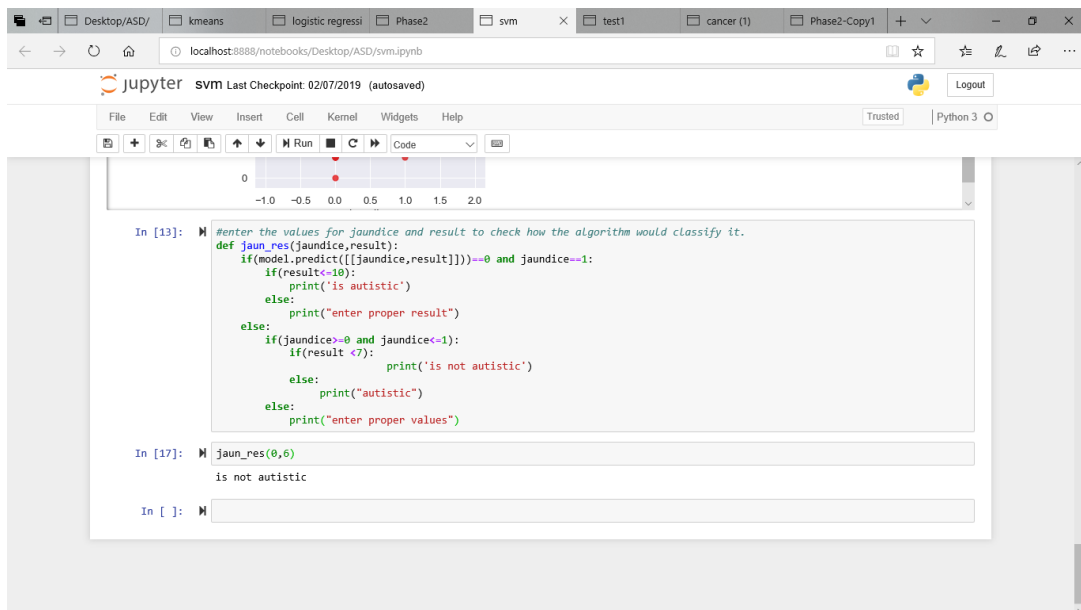
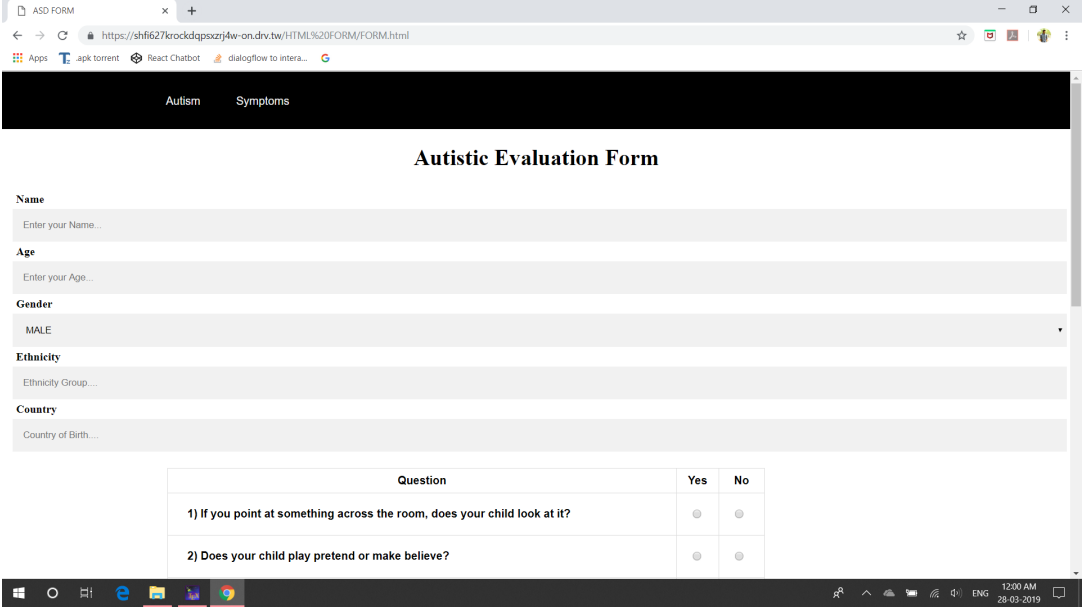


Figure 7.8: SVM function to determine whether the individual is autistic or not.

7.2 Module 2 Data

This phase involves in the collection of data from the public, which includes parents of autistic and non autistic children, students and faculty members of autistic schools belonging to various parts of the country, and also doctors of some hospitals. The data was collected through a Google Form questionnaire, that was integrated into a web page and was hosted online. All the entries were recorded onto an Excel sheet, and on cleaning the data, they were fed into the algorithms for computation. Along with the existing features, more features were added. These features feed in data related to the history of Autism in an individual's family, data regarding physical disabilities of the individual, and the individual's reaction in a social environment, which is classified as 'socially uncomfortable', 'normal' and 'extreme' behaviours.



The screenshot shows a web browser window with a tab titled 'ASD FORM'. The address bar displays the URL: <https://shf627krockdapszj4w-on.drv.tw/HTML%20FORM/FORM.html>. The page has a black header with 'Autism' and 'Symptoms' links. The main title is 'Autistic Evaluation Form'. Below the title are input fields for 'Name' (with placeholder 'Enter your Name...'), 'Age' (with placeholder 'Enter your Age...'), 'Gender' (with a dropdown menu showing 'MALE'), 'Ethnicity' (with placeholder 'Ethnicity Group...'), and 'Country' (with placeholder 'Country of Birth...'). Below these fields is a table with two columns: 'Question' and two sub-columns 'Yes' and 'No'. The table contains two rows of questions, each with radio button options for 'Yes' and 'No'.

Question	Yes	No
1) If you point at something across the room, does your child look at it?	<input type="radio"/>	<input type="radio"/>
2) Does your child play pretend or make believe?	<input type="radio"/>	<input type="radio"/>

Figure 7.9: Web page with Autism Questionnaire

DETECTION OF AUTISM SPECTRUM DISORDER USING MACHINE LEARNING

The screenshot shows a web browser window with the URL <https://shf627krockdpsxzf4w-on.drv.tw/HTML%20FORM/FORM.html>. The page displays a questionnaire titled "ASD FORM" with a table of questions and two columns for "Yes" and "No" responses. The questions are numbered 1 through 10, followed by "Jaundice", "Anyone in the family has Autism?", and "Physically Disabled". Below the table, there are radio buttons for "Behaviour in Social Environment" with options: "Socially uncomfortable", "Normal Behaviour", and "Extreme Behaviour". A "Submit" button is at the bottom right of the form.

Question	Yes	No
1) If you point at something across the room, does your child look at it?	<input type="radio"/>	<input type="radio"/>
2) Does your child play pretend or make believe?	<input type="radio"/>	<input type="radio"/>
3) Does your child make unusual finger movements near his or her eyes?	<input type="radio"/>	<input type="radio"/>
4) Is your child interested in other children?	<input type="radio"/>	<input type="radio"/>
5) Does your child respond when you call his or her name?	<input type="radio"/>	<input type="radio"/>
6) Does your child get upset by everyday noises?	<input type="radio"/>	<input type="radio"/>
7) Does your child understand when you tell him or her to do something?	<input type="radio"/>	<input type="radio"/>
8) If something happens, does your child look at your face to see how you feel about it?	<input type="radio"/>	<input type="radio"/>
9) Does your child understand what other people say?	<input type="radio"/>	<input type="radio"/>
10) Does your child have extreme reactions to uncomfortable situations?	<input type="radio"/>	<input type="radio"/>
Jaundice	<input type="radio"/>	<input type="radio"/>
Anyone in the family has Autism?	<input type="radio"/>	<input type="radio"/>
Physically Disabled	<input type="radio"/>	<input type="radio"/>

Behaviour in Social Environment : ☐ Socially uncomfortable ☐ Normal Behaviour ☐ Extreme Behaviour

Submit

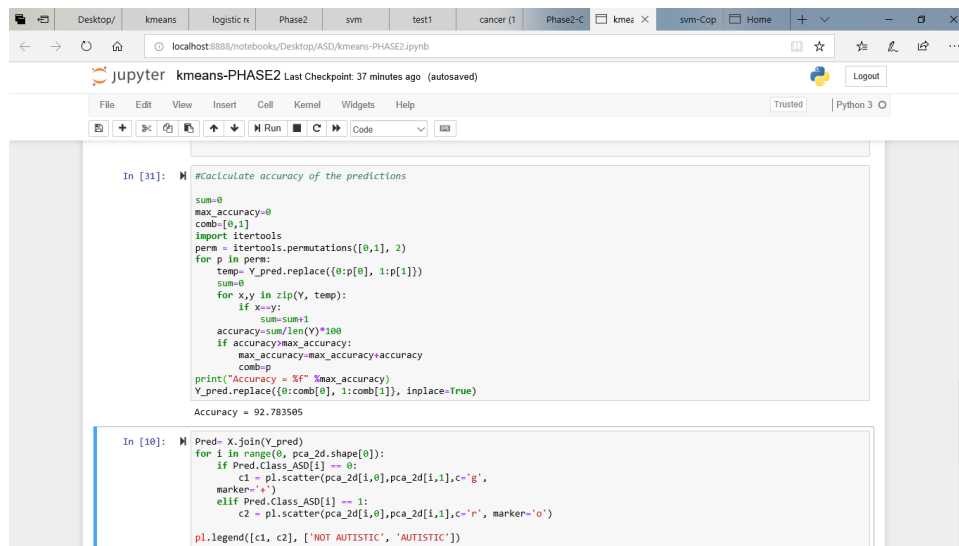
Figure 7.10: Questions to assess Autistic individuals

The screenshot shows a Google Sheet titled "ASD-FORM" with a real-time dataset of responses. The sheet has columns for Timestamp, Name, Age, Gender, Ethnicity, Country, and four columns for the questionnaire results (A1, A2, A3, A4). The data is organized into rows, with each row representing a response from a different individual. The sheet is titled "ASD-FORM" and has a "Share" button in the top right corner. The data is organized into columns: Timestamp, Name, Age, Gender, Ethnicity, Country, A1, A2, A3, A4. The data is organized into rows, with each row representing a response from a different individual.

Timestamp	Name	Age	Gender	Ethnicity	Country	A1	A2	A3	A4
2/28/2019 19:36:03	Jathin Chako	48	1	Indian	India	1	1	1	0
3/1/2019 18:14:45	Akshay S	20	1	Indian	India	1	1	1	0
3/1/2019 18:16:02	Akshay Satish	20	1	Indian	India	1	1	1	0
3/1/2019 18:17:08	Sandhya Jayaraman	20	0	Indian	India	1	0	0	0
3/1/2019 18:18:23	Shanaya	22	0	Hindu	Indian	1	1	1	0
3/1/2019 18:18:39	Adithya Sagar	20	1	Indian	India	1	0	0	0
3/1/2019 18:23:18	Supriya	21	0	Indian	India	1	1	1	0
3/1/2019 18:24:00	Hansha	21	0	Indian	India	1	0	0	0
3/1/2019 18:26:11	Kalyan Venkatesh V	23	1	Bangalore	India	1	1	1	0
3/1/2019 18:36:38	Rishabh D	22	1	Jain	India	1	0	0	0
3/1/2019 18:47:36	Avani Bhardwaj	18	0	Karnataka	India	1	1	1	0
3/1/2019 18:56:26	Sujith	21	1	Dravidian	India	1	1	1	1
3/1/2019 19:01:40	Ayush saraf	20	1	Indian	India	1	1	1	1
3/1/2019 19:21:39	Mohana	19	0	Indian	India	1	1	1	0
3/1/2019 19:31:02	PRATEEK SONTEKE	21	1	Maharashtrian	India	0	1	1	0
3/1/2019 19:42:45	Vallish Prabhu	20	1	Bangalore	India	1	1	1	0
3/1/2019 21:16:49	Smitha	25	0	Indian	India	1	1	1	0
3/1/2019 21:24:24	ABHISHEK SAINI	19	1	Rajasthan	India	1	1	1	0
3/1/2019 21:36:20	Rohil	23	1	Asian	India	1	0	0	0
3/1/2019 21:45:44	Srinjan S	18	1	Kannadiga	India	1	1	1	0
3/2/2019 4:44:15	Ria Borthakur	22	1	Indian	India	1	1	1	0
3/2/2019 8:04:40	Swati Devella	20	0	Indian	India	1	1	1	0

Figure 7.11: Real time data-set

The algorithms that were used for the Module 1 data are being used for the Module 2 data as well. Fabrication of data will lead to the yielding of results that might be too good to be true, hence, a waiting period was kept, to hear back from the people who have received the questionnaire. More than 100 entries were received and was used to train the algorithm, which was further donated to UCI. This methodology can be used by an individual, or an autistic clinic/school, or by hospitals to diagnose or detect Autism accurately in a cost efficient manner.



```

In [31]: #Calculate accuracy of the predictions

sum=0
max_accuracy=0
comb=[0,1]
import itertools
perm = itertools.permutations([0,1], 2)
for p in perm:
    temp= Y_pred.replace([0:p[0], 1:p[1]])
    sum=0
    for x,y in zip(Y, temp):
        if x==y:
            sum=sum+1
    accuracy=sum/len(Y)*100
    if accuracy>max_accuracy:
        max_accuracy=max_accuracy+accuracy
    comb=p
print("Accuracy = %f" %max_accuracy)
Y_pred.replace([0:comb[0], 1:comb[1]], inplace=True)
Accuracy = 92.783505

In [10]: # Pred= X.join(Y_pred)
for i in range(0, pca_2d.shape[0]):
    if Pred.Class ASD[i] == 0:
        c1 = plt.scatter(pca_2d[i,0],pca_2d[i,1],c='g',
            marker='x')
    elif Pred.Class ASD[i] == 1:
        c2 = plt.scatter(pca_2d[i,0],pca_2d[i,1],c='r', marker='o')
plt.legend([c1, c2], ['NOT AUTISTIC', 'AUTISTIC'])

```

Figure 7.12: K Means Clustering for Module 2 Data

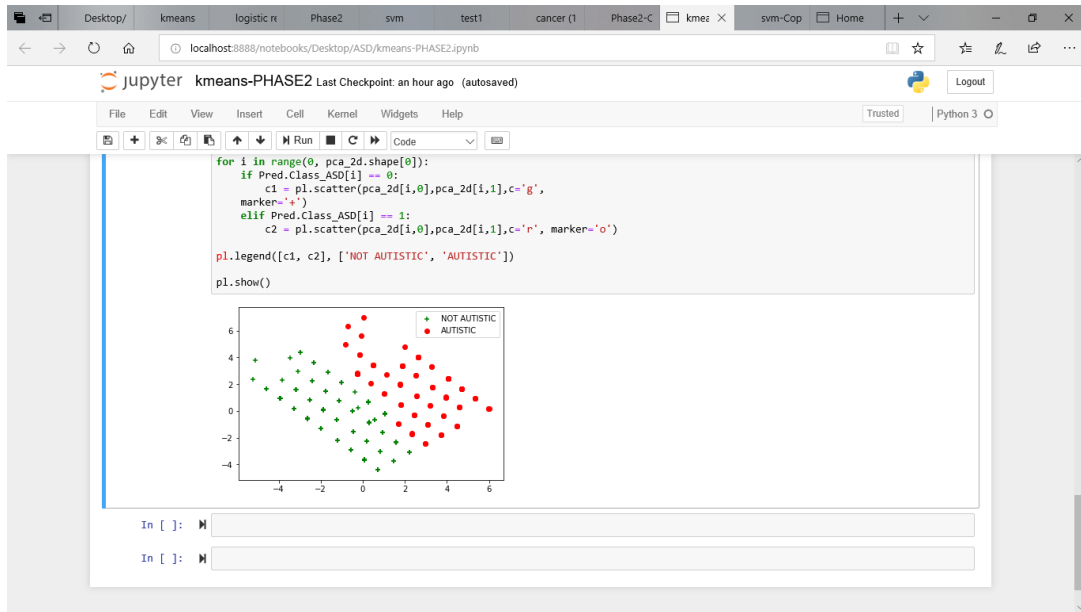


Figure 7.13: K Means Clustering for Module 2 Data

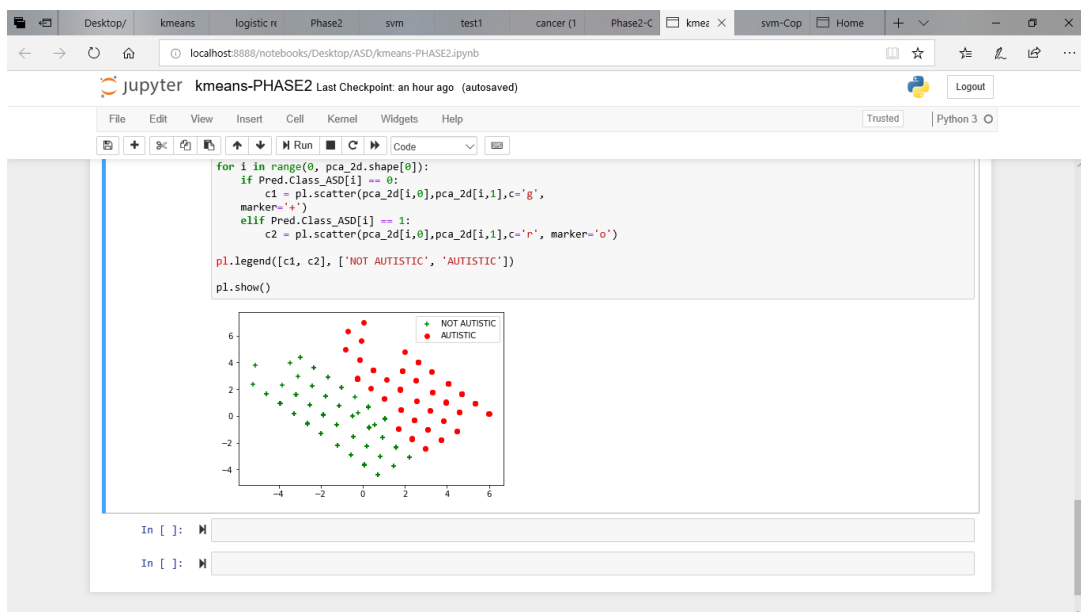
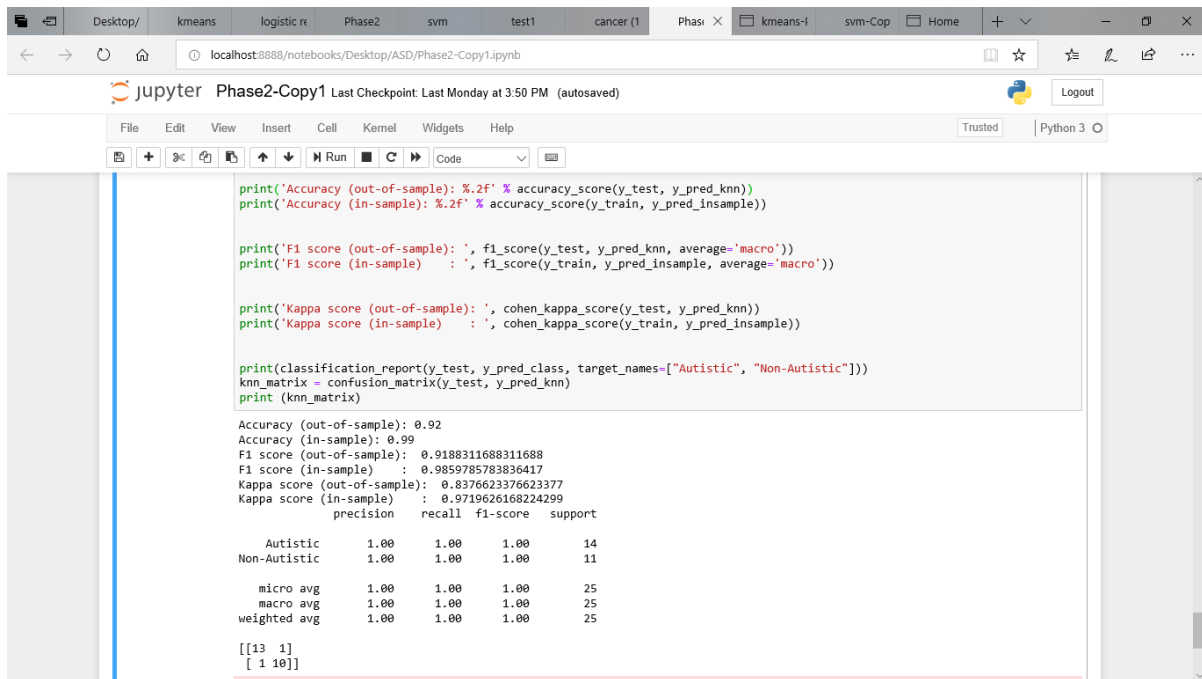


Figure 7.14: K Means Graph into Autistic and Not Autistic

For this data, k-means algorithm has been used, and the features that have been considered are behaviour in social environment (BISE) and result. In Figure 7.15, there is a formation of two different centroids and a cluster of points around them.

After implementing k-means algorithm, the k-nearest neighbors(KNN) algorithm was implemented in order to classify the data into two distinct classes, yielding an accuracy of 92.78%.



```

print('Accuracy (out-of-sample): %.2f' % accuracy_score(y_test, y_pred_knn))
print('Accuracy (in-sample): %.2f' % accuracy_score(y_train, y_pred_insamle))

print('F1 score (out-of-sample): ', f1_score(y_test, y_pred_knn, average='macro'))
print('F1 score (in-sample) : ', f1_score(y_train, y_pred_insamle, average='macro'))

print('Kappa score (out-of-sample): ', cohen_kappa_score(y_test, y_pred_knn))
print('Kappa score (in-sample) : ', cohen_kappa_score(y_train, y_pred_insamle))

print(classification_report(y_test, y_pred_class, target_names=["Autistic", "Non-Autistic"]))
knn_matrix = confusion_matrix(y_test, y_pred_knn)
print(knn_matrix)

```

```

Accuracy (out-of-sample): 0.92
Accuracy (in-sample): 0.99
F1 score (out-of-sample): 0.9188311688311688
F1 score (in-sample) : 0.9859785783836417
Kappa score (out-of-sample): 0.8376623376623377
Kappa score (in-sample) : 0.9719626168224299

```

	precision	recall	f1-score	support
Autistic	1.00	1.00	1.00	14
Non-Autistic	1.00	1.00	1.00	11
micro avg	1.00	1.00	1.00	25
macro avg	1.00	1.00	1.00	25
weighted avg	1.00	1.00	1.00	25

```

[[13 1]
 [ 1 10]]

```

Figure 7.15: K Nearest Neighbours accuracy for Module 2 Data

Logistic Regression was also used for the given data set. Due to the huge number of combinations of the data-set, the accuracy that has been achieved is lower than the accuracy of the data-set retrieved from the internet. Initially the accuracy of the data was below the half way mark, but after the processing of data and clearing out the unwanted information, the algorithm yields an accuracy of 96%. The SVM was also fed with the live collected data, which yielded an accuracy of 88%.

```

In [198]: feature_cols=['Age','PDIS','Res','B1SE','Jaundice']
          x=c[feature_cols]
          y=c1.cluster

In [199]: #here we split the original dataset into training set and testing set.
          from sklearn.model_selection import train_test_split
          x_train,x_test,y_train,y_test=train_test_split(x,y,random_state=0)

In [200]: #train the model by importing LOGISTICREGRESSION and feed the data.
          from sklearn.linear_model import LogisticRegression
          logreg=LogisticRegression()
          logreg.fit(x_train,y_train)

Out[200]: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                             intercept_scaling=1, max_iter=100, multi_class='warn',
                             n_jobs=None, penalty='l2', random_state=None, solver='warn',
                             tol=0.0001, verbose=0, warm_start=False)

In [201]: y_pred_class=logreg.predict(x_test)

In [202]: from sklearn import metrics
          print (metrics.accuracy_score(y_test,y_pred_class) *100)

96.0

```

Figure 7.16: Logistic Regression accuracy for Module 2 Data

For the Module 2 data, three more algorithms were used to examine whether the algorithms could be trained with the newly collected data and perform accurate predictions. The accuracies of the algorithms exclusively used for the Module 2 data are listed in Table 7.2 and also compared with the accuracies of the same algorithms used in the research paper. The accuracies of the used algorithms are:

Table 7.1: Accuracy of algorithms used for Module 1 and Module 2 Data

Algorithm	Module 1 Data	Module 2 Data
K Means	57.19%	92.7%
SVM	-	88%
Logistic Regression	94.52%	96%

The accuracies of Module 2 Data when compared to the research paper are:

Table 7.2: Accuracy of algorithms used for Module 2 Data only

Algorithm	Module 2 Data	Results in Paper[6]
K Nearest Neighbors	96.00%	-
Random Forest	94.25%	85.1%
Naive Bayes	93.33%	86.5%

A comparative study of the accuracies that were achieved had to be done, and to carry that out, the K-Means Clustering algorithm, the Support Vector Machine algorithm and the Naive Bayes algorithm were programmed from scratch and the accuracies were obtained. Table 7.3 has a comparison of the accuracies.

Table 7.3: Comparison of built-in and coded algorithms

Algorithms	Built in Algorithm	Algorithm from scratch
K Means Classifier	92.70%	92.70%
Support Vector Machine	88%	70%
Naive Bayes	93.33%	84%

As the data-set that was collected in real time was yielding a higher accuracy than the downloaded data-set, the data-set was donated to the online machine learning repository UCI, with around 100 entries and 21 attributes. The data-set was donated with an intention to make it available and make algorithm training easier and more efficient. The donation of the data-set was carried out with a simple procedure where details about the donating entity was made a record of, the number of records and attributes were noted, and the description for each attribute was made a note of. The data-set was successfully donated and on going through the authorities at UCI, the data-set will be made available.

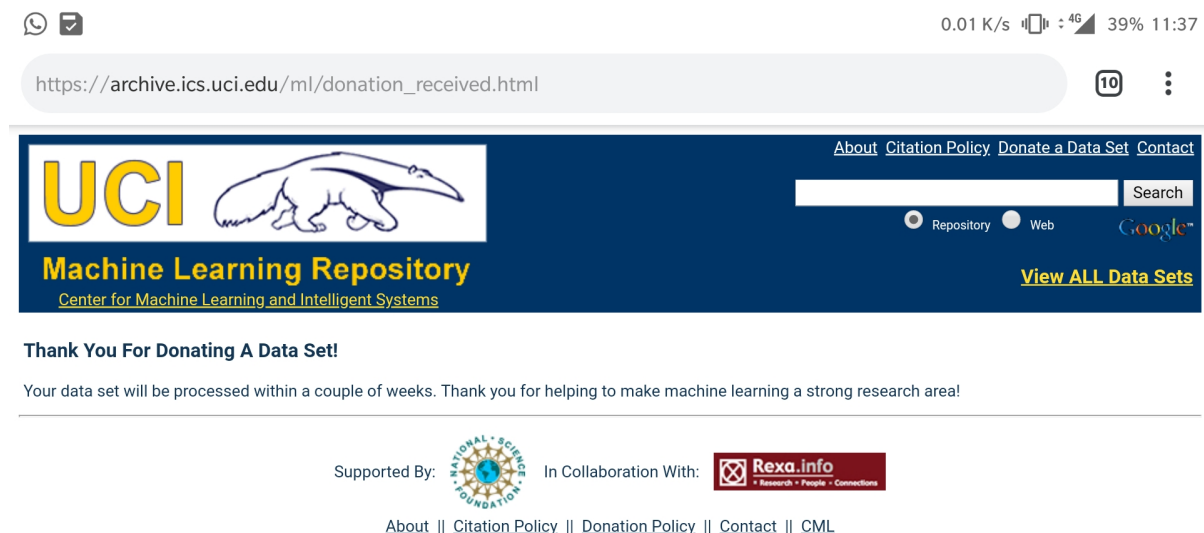


Figure 7.17: Donation of data-set to UCI

Chapter 8

CONCLUSION AND FUTURE WORK

After running through all the different types of data, the algorithms that have been implemented have been trained efficiently and is successfully differentiating the autistic individuals from the non autistic individuals. Also, these algorithms have been yielding a higher accuracy when compared to the accuracy obtained by the authors of the research paper that was referred, thus, improving the efficiency and accuracy of the algorithms. Logistic Regression as well as K Nearest Neighbors algorithms yielded the highest accuracy among the algorithms used for Module 2 data, thus narrowing down the algorithms that can be utilized. This becomes a cost efficient, user friendly tool to help doctors as well as a layman to determine whether someone is autistic or on the verge of being autistic or not. The reach among the public can be easily expanded with the help of the web page to collect the data as computers are available in almost every corner of the world. Smartphones also have not been eliminated as there are smartphones that are affordable by everyone, and hence, working on a mobile application, either on iOS and Android, will make the module portable and user friendly.

There is definitely a scope for future work on this front. For better accuracy of prediction the implementation of Neural Networks and Attention Networks can be incorporated. With this advancement, a huge amount of data can be taken in at once, and collective analysis can be carried out, to produce greater results. As a product, this can be passed on to various medical institutions to help with the diagnosis of Autism. Integration of hardware components in order to obtain data is also possible. The use of retina scanners and facial recognition cameras will help us to obtain information, by processing the data with the help of techniques like image processing.

Bibliography

- [1] Halim Abbas, Ford Garberson, Eric Glover, and Dennis P Wall. Machine learning for early detection of autism (and other conditions) using a parental questionnaire and home video screening. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 3558–3561. IEEE, 2017.
- [2] Sushama Rani Dutta, Soumyajit Giri, Sujoy Datta, and Monideepa Roy. A machine learning-based method for autism diagnosis assistance in children. In *2017 International Conference on Information Technology (ICIT)*, pages 36–41. IEEE, 2017.
- [3] Wenbo Liu, Xhiding Yu, Bhiksha Raj, Li Yi, Xiaobing Zou, and Ming Li. Efficient autism spectrum disorder prediction with eye movement: A machine learning framework. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 649–655. IEEE, 2015.
- [4] Matthew J. Maenner, Kim Van Naarden Braun, Marshalyn Yeargin-Allsopp, Deborah L. Christensen, and Laura A. Schieve. Development of a machine learning algorithm for the surveillance of autism spectrum disorder. In -. PLoS ONE, 2016.
- [5] Anjali Pahwa, Gaurav Aggarwal, and Ashutosh Sharma. A machine learning approach for identification & diagnosing features of neurodevelopmental disorders using speech and spoken sentences. In *2016 International Conference on Computing, Communication and Automation (ICCCA)*, pages 377–382. IEEE, 2016.
- [6] Bram van den Bekerom. Using machine learning for detection of autism spectrum disorder. In -. IEEE, 2017.

ORIGINALITY REPORT

12%

SIMILARITY INDEX

9%

INTERNET SOURCES

6%

PUBLICATIONS

8%

STUDENT PAPERS

PRIMARY SOURCES

1

journals.plos.org

Internet Source

1%

2

Submitted to Engineers Australia

Student Paper

1%

3

Submitted to Visvesvaraya Technological University

Student Paper

1%

4

Anjali Pahwa, Gaurav Aggarwal, Ashutosh Sharma. "A machine learning approach for identification & diagnosing features of Neurodevelopmental disorders using speech and spoken sentences", 2016 International Conference on Computing, Communication and Automation (ICCCA), 2016

Publication

<1%

5

Halim Abbas, Ford Garberson, Eric Glover, Dennis P. Wall. "Machine learning for early detection of autism (and other conditions) using a parental questionnaire and home video screening", 2017 IEEE International Conference

<1%

on Big Data (Big Data), 2017

Publication

6	pidswebs.pids.gov.ph Internet Source	<1 %
7	www.ijmlc.org Internet Source	<1 %
8	arxiv.org Internet Source	<1 %
9	www.msrit.edu Internet Source	<1 %
10	uzspace.uzulu.ac.za Internet Source	<1 %
11	E. Puerto, J. Aguilar, C. López, D. Chávez. "Using Multilayer Fuzzy Cognitive Maps to diagnose Autism Spectrum Disorder", Applied Soft Computing, 2019 Publication	<1 %
12	Submitted to Symbiosis International University Student Paper	<1 %
13	Submitted to CSU, San Jose State University Student Paper	<1 %
14	www.favouriteblog.com Internet Source	<1 %
15	onlinelibrary.wiley.com Internet Source	<1 %

16	Submitted to National Taipei University of Technology Student Paper	<1 %
17	ajaybolar.weebly.com Internet Source	<1 %
18	Submitted to The Hong Kong Polytechnic University Student Paper	<1 %
19	Submitted to Siddaganga Institute of Technology Student Paper	<1 %
20	Submitted to Botswana International University of Science and Technology Student Paper	<1 %
21	vdocuments.site Internet Source	<1 %
22	www.dezyre.com Internet Source	<1 %
23	www.nmit.ac.in Internet Source	<1 %
24	dblp2.uni-trier.de Internet Source	<1 %
25	www.slideshare.net Internet Source	<1 %

26	Submitted to King's College Student Paper	<1 %
27	link.springer.com Internet Source	<1 %
28	data.eufreelance.com Internet Source	<1 %
29	Submitted to Higher Education Commission Pakistan Student Paper	<1 %
30	www.scribd.com Internet Source	<1 %
31	Submitted to Sheffield Hallam University Student Paper	<1 %
32	Lecture Notes in Computer Science, 2013. Publication	<1 %
33	Submitted to University of Bristol Student Paper	<1 %
34	"Methods and Applications of Artificial Intelligence", Springer Nature, 2002 Publication	<1 %
35	www.cse.ust.hk Internet Source	<1 %
36	vision.unipv.it Internet Source	<1 %

37

Murugan Anandarajan, Chelsey Hill, Thomas Nolan. "Practical Text Analytics", Springer Nature America, Inc, 2019

Publication

<1%

Exclude quotes On

Exclude matches < 10 words

Exclude bibliography On