Childhood Autistic Spectrum Disorder Screening using Machine Learning

Project Report

Submitted in fulfillment of the
requirement for the degree of
Bachelor of Technology
In
Computer Science and Engineering and Information Technology

By

Avisha Singh (171232)

Under the supervision of Dr.Mrityunjay Singh

To

Department of computer science and Information Technology

Jaypee University of Information Technology Waknaghat, Solan-173234, Himachal Pradesh

Certificate

Candidate's Declaration

I hereby declare that the work presented in this report entitled "Childhood Autistic Spectrum Disorder Screening usingMachine Learning" in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering/Information Technology submitted in the department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology, Waknaghat is an authentic record of my own work carried out over a period from January 2021 to May 2021 under the supervision of Dr. Mrityunjay Singh (Assistant Professor, Senior Grade, Computer Science &Engineering Department).The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Avisha Singh(171232)

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Dr.Mrityunjay Singh
Designation: Assistant Professor (Senior Grade)
Department name: Computer Scienceand Engineering and Information Technology

ACKNOWLEDGEMENT

I want to take this chance to thank almighty for blessing me with his grace and taking my job to a successful culmination. I owe my profound gratitude to my project supervisor, Dr. Mrityunjay Singh who took keen interest and guided me all along in my project work titled - “Childhood Autistic Spectrum Disorder Screening usingMachine Learning” tilltheproject. The project development helped me in research and I got to know a lot of new things in my domain.

I am really thankful to him.

Table of content

LIST OF FIGURES

List of Graphs

LIST OF TABLES

ABSTRACT

The disease known as Autistic Spectrum Disorder (ASD) is known to be a condition which comes under neurodevelopment and that is related severe charges of medication, while the conclusion of this is really helpful in order to diminish this.Unfortunately, the not so friendly situations for the ASD analysis are time consuming and are costly. Its result in terms of mental imbalance and the increase in the number of ASD cases all across the world requires immediate improvement of the non-complex and actually powerfully computing techniques.Therefore, ASD screening which proves to be time efficient and is of appropriate cost is helping the medical experts in informing the individuals whether they should pursue for the formal clinical diagnosis. The development in the number of cases of ASD all over requires the datasets to be verified from their characteristics and traits.

CHAPTER 1: INTRODUCTION

1. Introduction

Autism spectrum disorder (ASD) or Autism will be the solution of the disturbances group for the growth having the origins containing neurological and effects in the communicative, social and challenges dependent on the behavior. ASD now is considered to be the 3rd most common disturbance of development. Most commonly it is noticed in the social interaction that is negative and the presence of most acute behaviors and interests that are constrained with ASD can be effected including their emotional sensitive side which usually means that they can be more or less sensitive to various sensor(like high pitch sounds, various fabric etc.).

Since we already know that there is nothing like single autism rather there are different types of Autism which are dependent on genes mixture and certain factors of environment. Each person suffering from Autism suffer certain situations like power and groups of difficulties as it is a spectrum disorder. The people who usually suffer from autism have ways of solving certain problems, way of learning them and find a way to think about them which starts with huge potential and is a main challenge. People suffering from ASD can require a support in their everyday lives, there are also cases where people do require only a little support and even no support at all and live on their own.

There are certain factors that are causing disease like autism and are mainly regarded to neurological sensitiveness including some medical problem likegastrointestinal disorders (GI), depression attention and anxiety challenges which may also cause rashes and sleeping challenges which affect mentality of a person.

Age of 3 and 4 years is usually considered when the proper symptoms of autism are noticed. But few of the developing delays can come earlier than this age and mostly they have the capability to be diagnosed in the time period of 18 months. It is noticed by professionals that premature interruption can cause good results afterwards in their lives having autism.

There is a guide line that is issued by the Association of American Psychiatric which is used in detecting disorders on mental health, according to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), people with ASD have:

● Interacting problem with other people in the public.

● Not showing any interest in anything and showing absurd behaviors.

● Cause disabilities in the children that impairs their ability to work properly in the school, at their work place and in other aspects of life.

Because of the presence of large dissimilarities in symptoms' effectiveness and seriousness which people are facing there is one more name for autism that is spectrum. All the groups are facing ASD including ethnic, racial and economic groups.

Behavior of social platform and communitative includes:

● Having very little and inconsistent face.

●Scared at looking or hearing from public.

●Memorizing other things by pointing at them and showing it to others.

●Not able to respond when are being called and unable to receive any attention.

● Difficulties in having to move forward.

●Oftentimes talking about things you are disinterested in and not even realizing it by not giving chance to others.

●Not matching their guise, emotions and gestures with what is being said.

● Facing ambiguity in the noise that appears to be singing and resembles robots.

● By not being able to actually get other person's feelings.

Repetitive behavior are:

● Not so normal behavior is repeated e.g. echolalia.

●Not having any kind of interest in deep facts and details.

●Thinking about weird things like parting and moving things.

●Getting annoyed by certain changes in quotidian.

●Getting affected more easily by clothing, light, heat and sound etc.

1.2Problem Statement:

The more effective model based on ML modeling will be promulgated meanwhile an application will also be developed based on ASD which is available for all kinds of people of all ages.

1.3 Motivation:

Nowadays, the problems of people facing health issues are gradually increasing, and mostly cases are actually not available because of the absence of correction and wrongly done diagnosis. Currently no correct solution is available. In order to first begin take a step forward, even if it is not so big. Our project of health care is very much beneficial for our country. So, a system is created with the help of this work where we have to anticipate the increasing cases which will happen according to the previous records at some place.A program will be created for mainly this project which will predict the disease according to the symptoms.

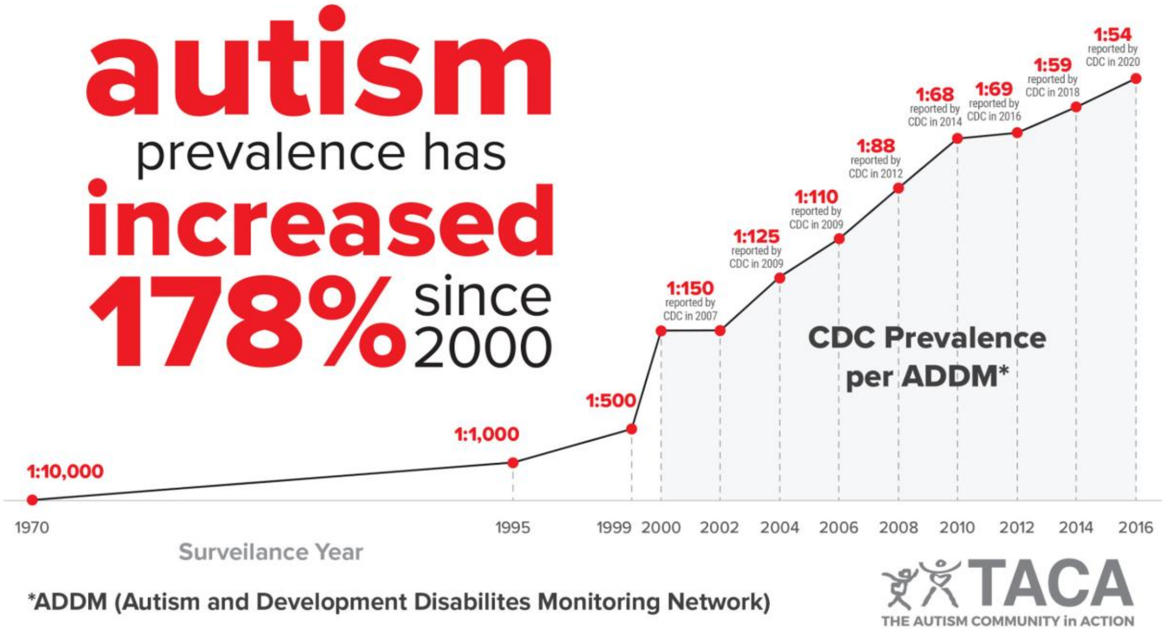


Figure 1.1: Rate of increase of Autism since 2000

1.4 Objectives

The basic motive of this project is:

1. ML basics and concepts.

2. To come up with an idea of autism..

3. To detect the most suitable algorithm in ML for the health datas.

4. Check and run those algorithms that are not in use.

5. To test the algorithm of the basic machine that is used in the dataset.

1.5 Methodology

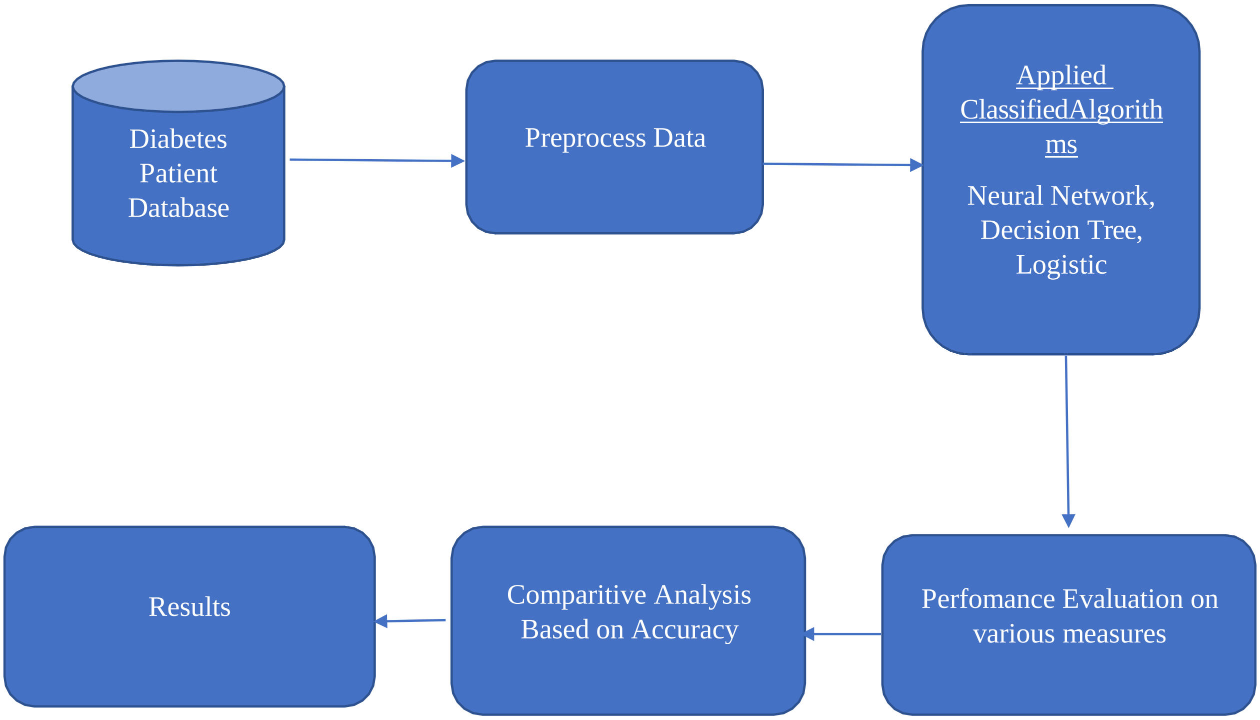

Diabetes Patient Database
Preprocess Data
Applied ClassifiedAlgorithms
Neural Network, Decision Tree, Logistic
Perfomance Evaluation on various measures
Comparitive Analysis Based on Accuracy
Results

Fig 1.5:Methodology

1.5.1 Description of data set

1.5.2 Data Type: Multiple values such as categorical, real and binary

Task: To classify

Column Type: real, categorical and binary.

Some null values are also there

Number of Instances in the dataset: 292

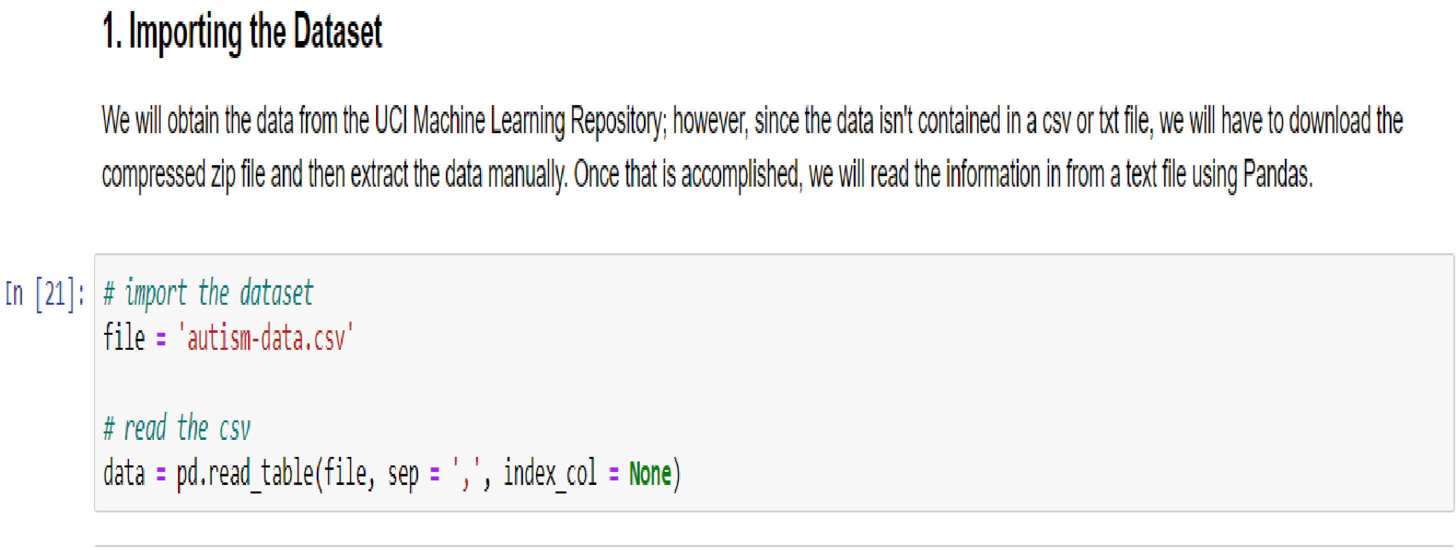
Number of columns in the record: 21

1.5.3 Libraries to be imported

Tools to install

Libraries used:

```
import pandas as pd
import numpy as np
import sys
import sklearn
import keras
from sklearn.model selection import train test split
```



1. Importing the Dataset

We will obtain the data from the UCI Machine Learning Repository; however, since the data isn't contained in a csv or txt file, we will have to download the compressed zip file and then extract the data manually. Once that is accomplished, we will read the information in from a text file using Pandas.

```
In [21]: # import the dataset
         file = 'autism-data.csv'

         # read the csv
         data = pd.read_table(file, sep = ',', index_col = None)
```

Fig 1.2: Reading dataset

1.5.4 Visualizing and selecting the dataset

1.5.5 CSV File: also known as value file of comma separation, where each row has several fields which are separated by comma. We can place each and every row as row and all the fields as a separate column. Reading of a csv file is as follows:

df=pd.read_csv('autism-data.csv) df.head()

Type of data: Multiple valueslike categorical, real and binary.

Task: Classify

Type of column: Binary, real and categorical

Some null values are there in the data

Number of Instances inthe dataset: 292

Number ofcolumns in each record: 21

1.5.6 Data preprocessing

Ques: Why is it a must to clear the data?

Rarely it happens when datasets are completely filled. In our knowledge it is a fact that every datapoint estimates for all the copyrights. Only several estimates are not laid correctly which when laded into pandas. Mark Nan along with null. There are ample of causes of the data loses. It may also happen that data collecting person could forget or can even started to collect these variatetion in middle of process of data collection.

We have to first maintain the dataset before working with it.For eg,supposedly while performing analysis on the data and while we are in the middle of the process and gain few precious info regarding the dataset from a varying feature like var F. Soon we'll come to know about the thing that 95% of var F values in the data are Nan : We'll not be in a state to make stiff decisions about the data set from only a variable that is representing only 7% of it.At the time of training of machine learning model, Nan is assumed to be zero or infinity in the process, by abandoning the training process.

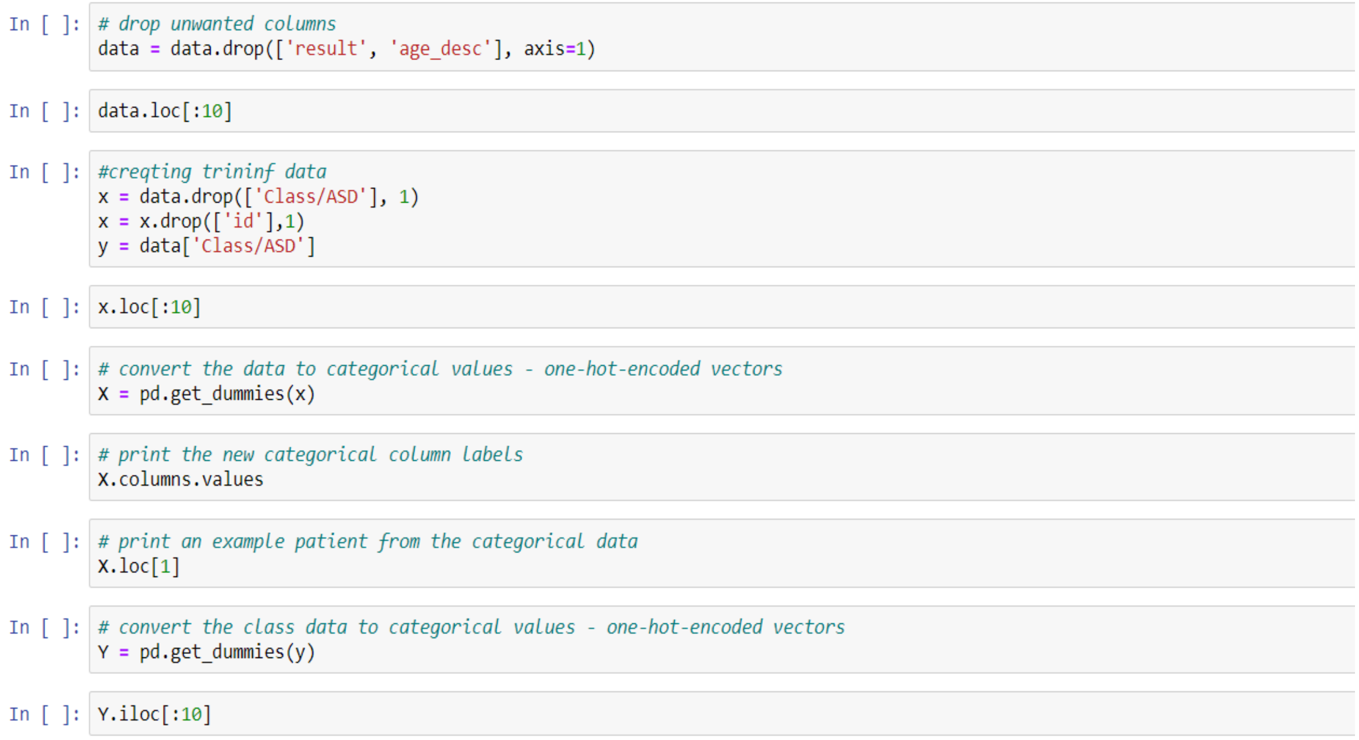
In order to find the lost info in pandas:

- Look at NAN's : pd.is null which finds the absent value from the data set giving both NAN and none.

- Arrange the lost data: It will return the frame of data of deleting data points having NAN.

- Exchange the lost data: the values are exchanged to value from to replace. It would be helpful only if you know what are those features to be changed.

- Setting the feature: if the value of the variable has 90% greater value in dataset, then we should neglect that in the database.



```
In [ ]: # drop unwanted columns
        data = data.drop(['result', 'age_desc'], axis=1)

In [ ]: data.loc[:10]

In [ ]: #creqting trininf data
        x = data.drop(['Class/ASD'], 1)
        x = x.drop(['id'],1)
        y = data['Class/ASD']

In [ ]: x.loc[:10]

In [ ]: # convert the data to categorical values - one-hot-encoded vectors
        X = pd.get_dummies(x)

In [ ]: # print the new categorical column labels
        X.columns.values

In [ ]: # print an example patient from the categorical data
        X.loc[1]

In [ ]: # convert the class data to categorical values - one-hot-encoded vectors
        Y = pd.get_dummies(y)

In [ ]: Y.iloc[:10]
```

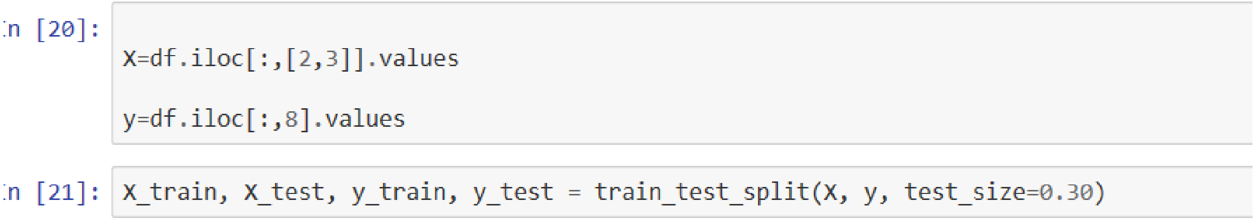
Fig: 1.3 Data Preprocesssing

The columns of the datatable save the results of bi variate like we also have to complete

• Univariate analysis: - it usually gives the analysis results of every field of the green dataset and also for the single and lonely variable. For instancecdf, pdf, violin struct and box struct. (they are mentioned below)

• Bivariate analysis: - it is used to anticipate any kind of connectivity in the variations of database and including the variation that is aimed for the interest by taking two variables and then finding any relation in them. For eg.Bbox struct, and violin struct.

Splitting of the data



```
:n [20]:  X=df.iloc[:,[2,3]].values

          y=df.iloc[:,8].values

:n [21]:  X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30)
```

Fig 1.4 Splitting dataset

1.5.7 Intro to various algorithms in ML

1-Decisiontree

It is the way which uses supervised method of learning and it also finds the solution of class separating situations. It is a kind of mechanism that provides support to data needed for 1 or more training data whose solution is to be found and called overriding mechanism. Its aim is usually to find the result of dependent variable.

It helps in making the model to anticipate anonymous class lables by classifying.

The outputs like binary and continuous are built from this. It finds the root node of the tree on the basis of huge number of entropy cases. It enables the tree to find the most sympatico proposition of the training data. The input value that is given to the tree is known as the data that has the various attributes and output value and also those outputs that eventually become the decision model. I/p: training data set

O/p: decision model

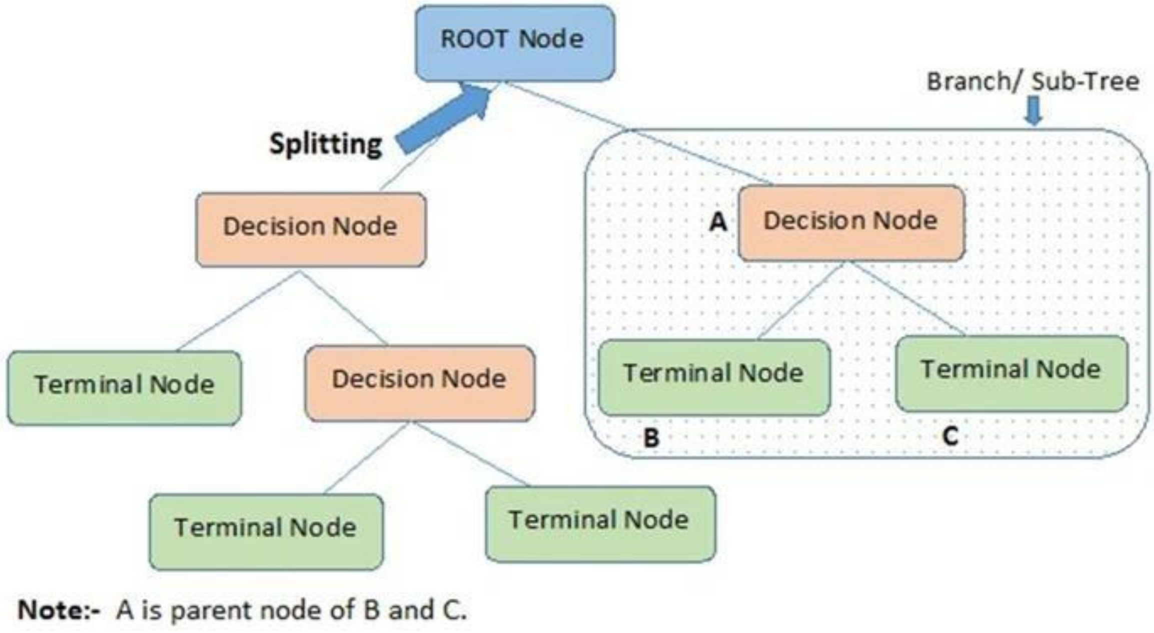


Fig 1.5 DecisionTree

What is the need of decision tree

Since, there are many such algos in ML and to choose the most perfect algo that is given and the question is the main thing of keeping in the mind while we design the ML learning model. Some of the reasons of taking decision tree are mentioned below:

They have the ability to detect how the people think whenever they have to make a decision, in order to make the decision taking process easier.

The main idea of decision tree is quite simple that can be understood by everyone as it gives the structure that looks like a tree.

Some name referring to decision tree:

- root node: It is that node of the tree that is there in the very beginning right from the start, it reflects the datasets, that usually continues to separate from one another in 2 or more identical sets.

- leaf node: They are the output also known as finished node and later on, the tree in incapable of dividing itself in more nodes like this from which a leaf node can be found.

- separating: separate the process of dividing a decision tree / node into its child nodes in accordance of the current situations

- branchingortree following: It is the next tree that comes after dividing a tree.

- shearis: shear is the process in which we remove the meddling branch from the basic tree.

- parent or child node: parent node is the root node and their resulting nodes that are at the end are known as the child nodes.

What is the procedure behind the working of algorithm of decision tree.

In a tree i.e. decision tree the process of predicting the given data's category is the thing, it starts from the root node of the decision tree.It is comparing the value of the root node in comparision to the actual value of the data i.e. the record value and according to that it goes with the branching and later on to the next node.

Along with moving forward, the algorithm double checks the value with all other nodes, along with continuing this process until and unless the edge of the tree if found. The algorithm below helps to understand the process much better:

1: Starting the node that is the root node suppose s which has the overall database.

2: ASM i.e. Attribute selection measure is used to find the perfect attribute.

3: The subsets which have the values that are currently present for the best symbols should be divided into the further subsets.

4: The node having the overall perfect quality is chosen.

5: With the help of the step 3 we have to create the new resolution tree from datasets. This method shall be continued until you reach a point where the further separation of nodes isn’t possible and are able to find the leaf node which is usually the last one

For a better understanding lets suppose a person is having a choice to decide if he has to take the job or not. In order to find the solution to this query, we have to begin with the root node. The differentiation of the root node depends on the upcoming resolution node along with single leaf node depending upon the tables. The following decision thing begins to divide in to a single decision thing and single leaf node also. At last, decision node has been divided in to its two halves i.e. the leaf nodes. Now, lets focus on the below diagram.

The provided steps for selecting:

Whenever we are working on the decision tree, the most prominent problem which comes up is to select the perfect attribute for root along with its sub nodes. Selection named process is there for solving these queries and at this time the most perfect qualities of the tree can be selected , we have two most basic ASM methods for this:

o Info accessing

o Indexes of gin

1. Currently present info is:

o To acquire the info in order to anticipate the entropy change which depends on the merit right after classifying.

o It also detects the amount of details that can be gathered by the single attribute about a certain category.

Now we have to separate the nodes and make a decision tree.

o The value of information is increasing by the algorithm that is used by decision tree at the same time prominent node with info is categorize first. We have to use this formula now:

Acquisition = Entropy (S) - [(Medium Weight) * Entropy (each element)

What is entropy: it is used to measure the pollution in an attribute by a matrix which indicates random datas. Entropy is basically:

Entropy (s) = -P (yes) log2 P (yes) - P (no) log2 P (no)

In which,

o S = the cases of sample

o P (yes) = possibility of yes

o P (no) = possibility of no

2. Gini'sreference:

Gini index is defined as proportion of the cases that is used in developing the medication that is dynamic in cart i.e. regression and classification calculation.

o We have to prioritize gini index that has lesser value in comparison to the higher value.

o it pairs the allotments in the cart in order to make segments parallel

The equation below is determined to find the gini index.

Gini Indicator = 1-∑jPj2

Advantages of decision tree

It is quite simple and understandable because it carries on the process that was done by a person earlier when to make a decision regarding anything in daily life.

It helps us to finds solutions of the situations of decision.

It allows us to think of all possible solutions to the problem and their results.

o It isn’t necessary to purify the data when algorithms are applied.

Pitfalls of decision tree

o There are very dense layers in the decision tree which convolutes the process.

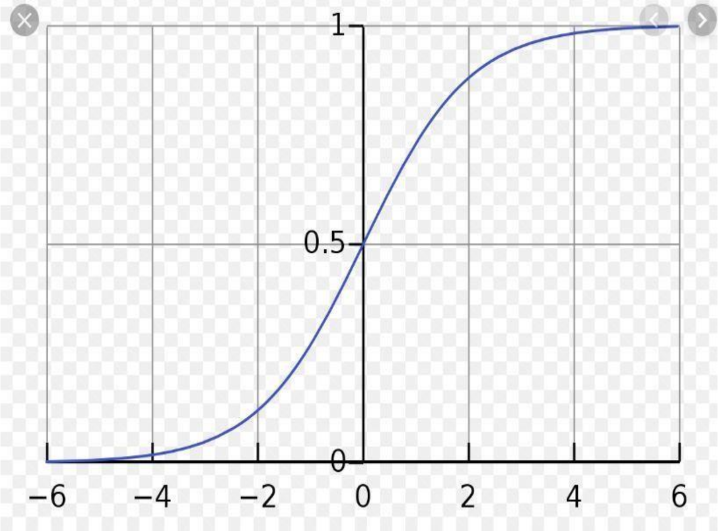
The extremist situation can occur where random forest needs to be applied.

o It can be quite difficult to calculate the decision tree which leads to its increase.

Logistic Regression

The binary split questions which hasatmost 2 values can be solved. There is a way where the relation between the target value and the predictor i.e. the independent value can be investigated which is mainly known as reversal analysis. It can also be known as line algo where the performance of line to the work of sigmoid which is done also done by line algo, where the data is divided between zero and one. After that it makes use of batch base of gradient to advance the model when the reducing operation cost comes in view and also does the comparison of training value with the paid value.

Sigmoid Function:

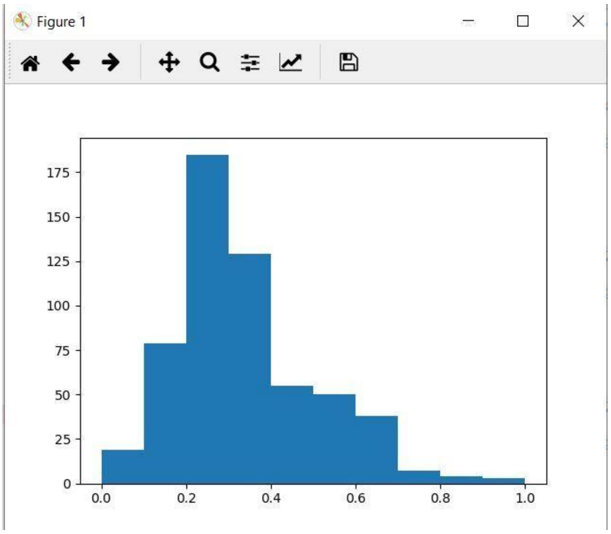


Graph 1.1 : Sigmoid function

We can also say that function called subtraction function creates a s shaped curve. It accepts the prominent numbers and transfers its value to zero or one. If the curve inclines to the positive sidethen the y’s value will become one, and if it inclines to the negative side or confidence then y’s value will be turned to zero.

Then when it has happened, the limit will be 0.5.

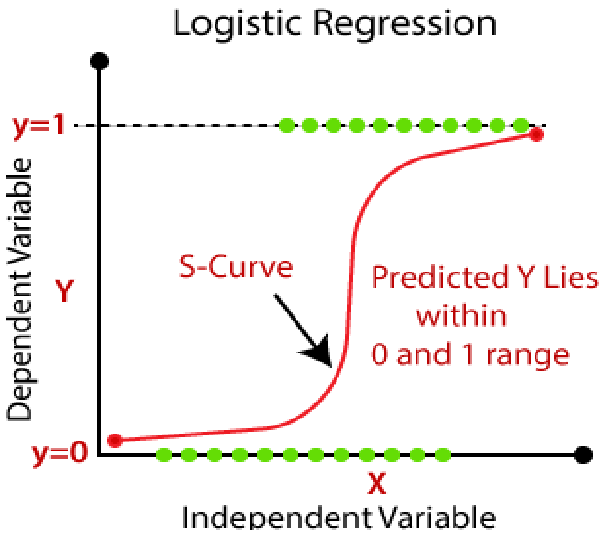
Or if pristine value is less than 0.5 then divide it as zero.



Graph 1.2: Data representation

In this fig, y axis indicates the no. of the data of every feature.

And the x axis indicates the data’s value when the feat is calibrated every time.



Graph: 1.3 Logistic Regression

At the time of 20th century an algorithm was actually used in biology and that is called logistic regression. Along with that it came into use in various sst programs. We generally use logistic regression when our target variable i.e. dependent is categorized. And linear regression when target variable is continuous.

Talking of some examples:

- to anticipate the spam mails

- to detect if a given plant is poisonous or safe.

Think of a situation where we have to find out the junk mails or spam mails and in that situation suppose we are using a line deflection method, so it requires a limit to be set in this based on the segmentations that have to be made. Tell if a given and primitive category is -ve, in order to anticipate value i.e. 0.4 (continuous) or the limit's value to be 0.5, then in that case data's point would be not so correct which gives us the bad or -ve results in realtime situation.

In the eg mentioned below, it is stated that reversing a line is not good and comfortable for separation problem. With infinite being the rotation of a line. Which also takes order of the things in to the picture or into the frame. Along with values ranging between zero to one.

Backbone type

1. Regression i.e. binary logistic

There are 2 possible answers which are spamming and no spamming.

2.Other is imultinomial logistic regression

Now we have 2 or more categories without any order like forecasting the preferable food e.g non-veg, veg and vegan.

3. and standard operationprocesses

More than 3 categories woth order are rating the movies starting from one till five.

The organization containing neurons is called neural network or circuit and with point of view of cutting edge, an organization consisting of counterfeit neurons is called a fake nneural organization. We can say that this institution can become a natural neural organization which consists of actual neurons that are general and genuine and even a neural organization that is actually fake to actually detect artificial consciousness or man made (AI) issues. Loads are said to be the natural neuron associations. As well as a positive association also tells us about the weight of it, whereas the negative weight means that it is an obstructing association . All of the sources are shortened and also weighed which is known to be clear blending. Now the work of entactment takes control of the yield size like adequate scope of the yield is between zero and one and also -1 to 1.

This fake organization is used in making the ephemeral display, apps and dynamic controls in which they are made on the info bases. Inside the organization a self made for a fact is conceived which are reachable for determination from a bundle of convoluted and also inconsequential data.

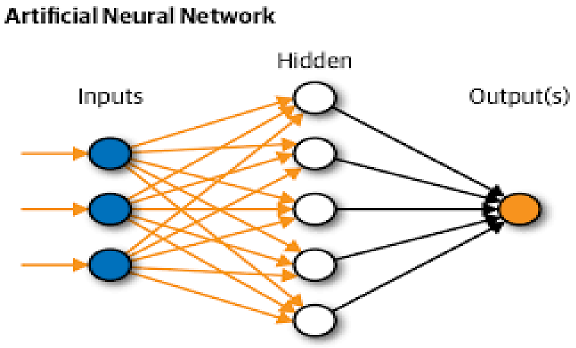


Fig: 1.6 Neural Network

The steps in order to create a neural network are:

- To define a structure model. [no. of inputs and output features]
- To initialize the parameter of model.

    - Looping.
        - Find out the current loss i.e. forward propagation.
        - Find out the current gradient i.e. backward propagation.
        - Updating parameter i.e. gradient descent.
- To preprocess the data is imp.
- To tune the cleaning rate, its example is hyper parameter which can actually change the actual algorithm.
- At first we will show the perceptron which is an artificial neuron. For eg. Suppose that perceptron has 3 inputs X1, X2, and X3. It has basically either more or less inputs. The result is usually a calculated by a simple trick introduced by Rosenblatt. Bells are as follows W1, W2….., which are real no’s. indicating values of input suitable for exit. Output being 00/11 which is determined by wjwjxj∑jwjxj weight is less or moreas compared to the limit. Just like the weights limit is also a real number defined as neuron parameter. algorithm:

o/p = {01if ∑jwjxj≤ thresholdif ∑jwjxj> limit (1) (1) output = {0if ∑jwjxj≤ threshold1if ∑jwjxj> limit

Finally perceptron is ready to work.

Situation is non-linear separation.

Nonlinearity meaning the label can’t be predicted with model b + w1x1 + w2x2 i.e. , "decision area" is not a line. Earlier, the feature cross process was the solution in order to measure the offline things.

Now take a look at this:

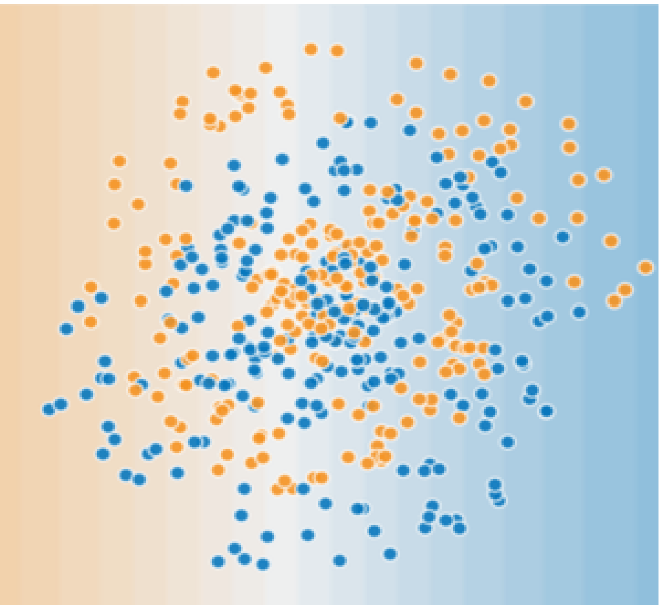


Fig 1.8 Non Linear separations

Non- linearity problem is hard to solve.

The collection of information that is there is actually unable to predict the instant model.

In order to see whether the neural organization can assist the non- straight issue first of all start it by speaking to the model as chart:

Direct model as diagram.

Every blue colored circle tells the info component and the green colored circle tells us about absolute load associated with the information.

What is the way to increase the capacity of managing the non-straight issue ?

Concealed Lists are:

As in the model which consists info by accompanying chart, an extra layer called shrouded layer is added of the intercession esteems. Every hub that is yellow in this layer is weighted as an incentive for hub that is blue esteems Yield is a fully completed load of yellow hubs.

2-layered model:

Now the question arises whether the model is precise or not?

The answer is yes, yield is a blend of info lines.

As said by the accompanying chart in a certain model that we should add a 2nd shrouded layer of weighing figures.

3 layered chart of model.

Now the question arises will the model be in line now? The answer is yes. Some time when it is shown that yield is information work and should be improved Is this model still in line? Obviously it is. At the point when you show the yield as an info work and should be improved then another information measure is added.

Starting activity

To make a non-straight model then in that order we have to bring the irregularities in it. By adding a hub in every concealed layer with non straight model.

As told by the accompanying diagram for the model, the estimated thing in each hub of hidden layer1 is changed in look by the non connecting capacity for moving the counts with the weight of the layer . And the disconnecting capacity is known as the enactment work.

3 layered diagram model along with actuation

As the addition of actuation work begins then the layers are added as major effect. Among the inputs and also the detected results, the no straight overlays of exact setting allows to show the perplexed connection. Then every layer learns highly troublesome and great work . If in case you want to add up the extra feeling of this function’s working and also have a look at amazing blog entry of Chris Olah’s.

Starting activities

The present capacity converts into the estimation who range varies from zero and one from the sigmoid actuation .

$F (x) = 11 + e - x$

Following is the ambiguity:

Sigmoid’s starting works:

To ensue the fixed work unit also known as ReLu function for various works in the past time that is usually good like a sigmoid as compared to the smooth work, meanwhile it is also quite easy for computation.

$F (x) = \text{plural} (0, x)$

The extension work of the ReLu function is dependent on the solid disclosure, which is recognizable within a reaction range. It’s response also dwindles along the two sides.

Beginning work of ReLu

To be truthful, in the place of element of actuation a numerical value can take place. For a while σ acts as the actuation like ReLu, and sigmoid. Thus, the number of hubs is shown in the following recipe of the organization.

$\cdot (w \cdot x + b)$

The out-of-container upholds are given by the tensorflow for the initiation assignments. We can also detect the initiation capacity within the tensorflow for a primitive organization of neural activity for the covering. To suggest the starting of the ReLu we have to take care of everything.

Splitting of news in the training and test.

We divide the info into the training and testing model depending on the 20, 60, 40 or 80%. For preparing the data the help of the training model is taken which consists of the relized field and our info is getting ready from the earlier data and while testing the model we have to test the model in the prepared model.

2 ways of preparing the test info:

(I) By taking work in library:

This thing divides the data set into the 80% training data and 20% of the testing data.

```
from sklearn.model_selection import train test_split xTrain, xTest, yTest = train_test_split (x, y, test_size = 0.2, random_state = 0)
```

The arbitrary model works like the non regular generator of the part of the info.

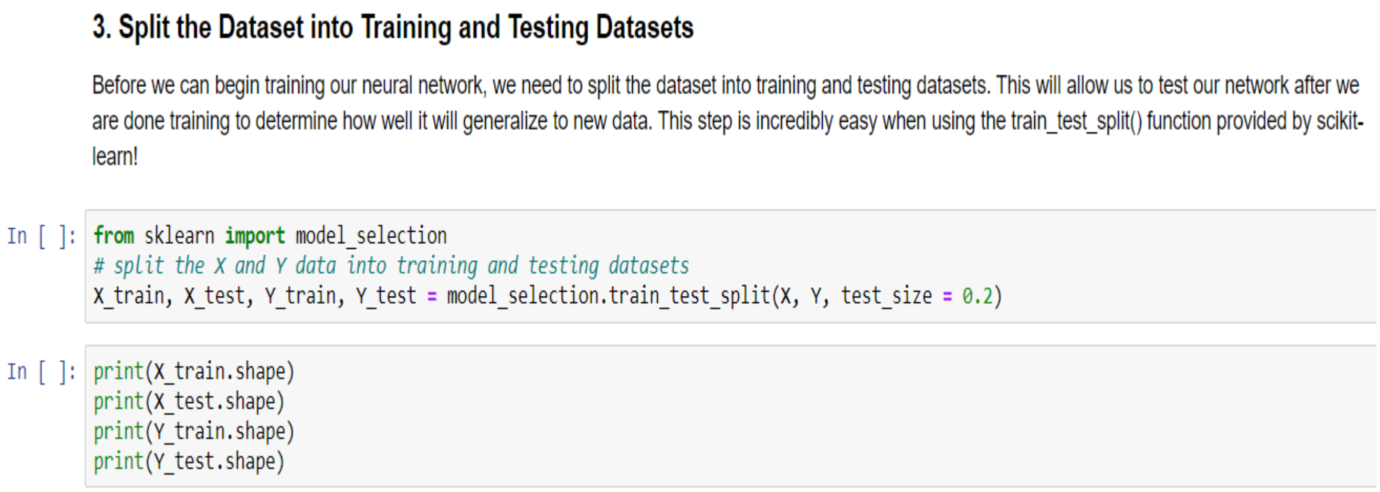
(ii)Use of random and the grant of work:

```
shuffle_indices = np.random.permutation (input.shape [0]) test_size = int (input.shape [0] * proportion)
```

```
train_indices = shuffle_indices [: test_size] test_indices = shuffle_indices [test_size:]
```

Our info is made in a way as indicated by our reaction rate which is 0.2 and training testing will become 80-20% which states that our 80% of the data will be used up in the preparation of the information data and 20% will be used up in the testing data info.

Soon after the division of the train and test data, we can arrange our information index in this way:



3. Split the Dataset into Training and Testing Datasets

Before we can begin training our neural network, we need to split the dataset into training and testing datasets. This will allow us to test our network after we are done training to determine how well it will generalize to new data. This step is incredibly easy when using the train_test_split() function provided by scikit-learn!

```
In [ ]: from sklearn import model_selection
        # split the X and Y data into training and testing datasets
        X_train, X_test, Y_train, Y_test = model_selection.train_test_split(X, Y, test_size = 0.2)
```

```
In [ ]: print(X_train.shape)
        print(X_test.shape)
        print(Y_train.shape)
        print(Y_test.shape)
```

Fig:1.7 Splitting in training and testing

Bias and Variance

- Bias – it is the combination among the general forecast of the model and the right worth where we have to forecast and try to detect.

•The model consisting of the highest inclination gives almost no recognition which inclines the model for preparation info.

- It persuades the large mistake for the preparation of the testing data.

- Variance – it is the variation of forecasting model for certain info which tells us the spreading of the data.

- Model giving the highest changes offers a good deal in order to prepare the data and does not pile up upon the dataset where the data is not seen earlier.

- So, these types of models actually work well on the preparing dataset and also has great gaffe rates on the testing dataset.

When our data is unable to detect the basic example then underfitting takes place

- Such types of models have usually lesser difference and higher pre disposition.

- This situation take over only when the info is of lesser measure to cover up the model and when we try to assemble a direct model with the non linear info.

- And such models are very easy to catch up the certain examples in info just like the linear relapse.

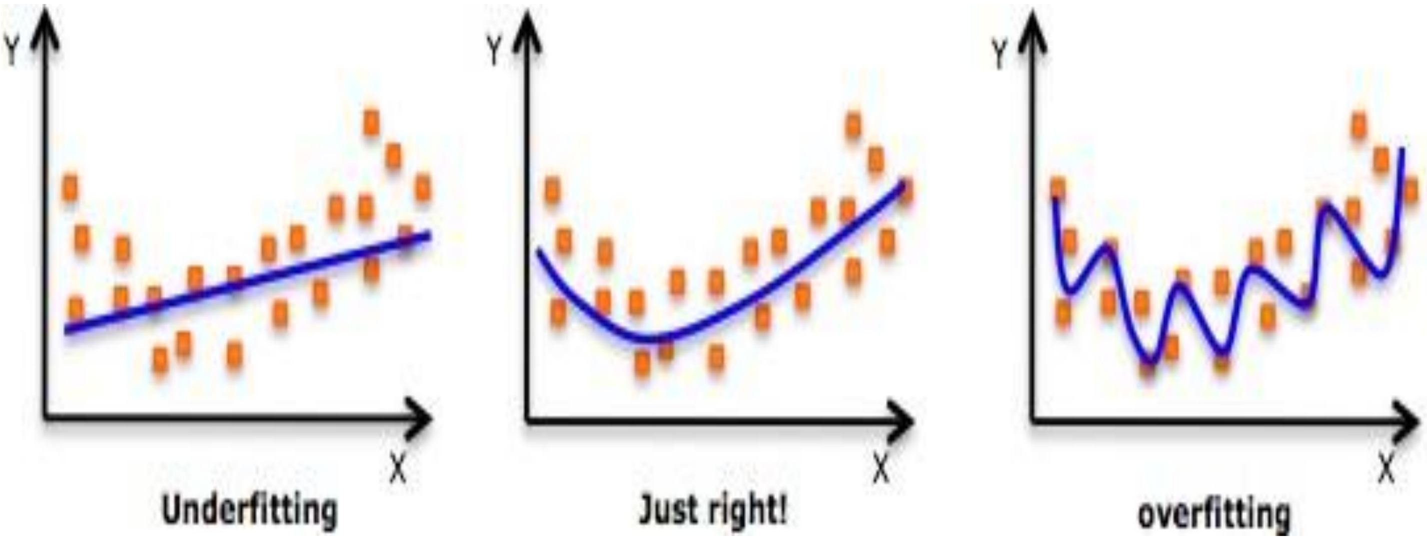
When the model catch up on the commotion while the hidden eg is there in the dataset then overfitting comes in the picture.

- It takes place only when we train the model dataset(uproarious)

.

- Those models have the greatest changes and also less inclination.

- These types of models are very amazing just like decision tree which inclines towards the overfitting.



Graph: 1.4: Model

2.5.1 Results validation

There are mainly 2 methods of checking if the algorithm is accurate:

(i) MSE-MEAN SQUAREDERROR:

In this method if the algorithm is efficient then its value is stored in the cost function which has earlier been calculated. The lesser the value of the cost function of both the training and testing data the more the value of the linear model in order to predict the target value.

(ii) CONFUSION MATRIX – Accuracy calculation:

When we get the prediction of our target values then we get a 2 by 2 matrix.

When the 4 values of binary classification are calculated i.e. either one or zero, then accuracy is calculated when we add the left elements of the diagonal and after we have to divide that sum of all the values. The more the rate of our accuracy the more is the value of the classification model.

By using the earlier method of Linear regression and after that using the other one in classification algo like Logistic regression, naïve bayes and KNN.

2.6 Organization

In the first chapter we have studied about the autism about its type, all the factors that contribute to the autism along with its signs, symptoms and causes.

Then in the next chapter, we will discuss the literature review of the lesson which consists of the key terms and the functioning and value of algorithms with its type. We have already studied about the several algorithms and their advantages and disadvantages so that we can find out the proper algorithms for our data base.

Then in another chapter, we’ll be focused on improving the architecture of our system and mentioning all the important details of the program in order to use the algorithm and also test them in environment in order to visualize the results for better answers.

After in the next chapter, all the remaining algos are discussed along with their statistics of the algos for better understanding and their working also. We have to consider the parameters in the algo.

In the end we’ll conclude the report and add the result in the reading list of the project. The future scope and outcome of the project will also be discussed.

Chapter-2 LiteratureReview

(2.1) Terminologies

2.1.1 Machine Learning: It explores the techniques that generally build the models from data of history which is used for various tasks like it is used in predicting and decision making. When the algorithm takes the values as input then it finds the similarity in the pattern and after the learning it usually helps algo to detect the o/p.

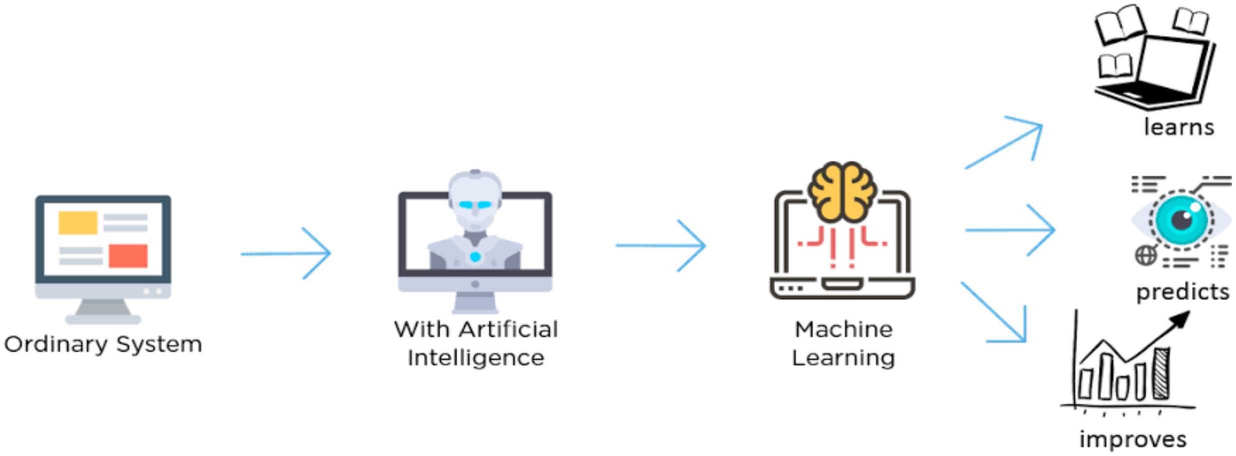


Fig 2.1: Machine Learning

2.1.2 ML working

This process usually starts by the collection of the data from dataset of the type of table. Apart from that there are algorithms that exist which are usable. After that we have to select the algorithm that is to be applied. Then the database is divided into the 2 dimensional scale: training and testing data. 70% being the training data and 30% being the testing data. Then our algo will be trained. Finally we have to test out algo and the answer will be put into the sequence.

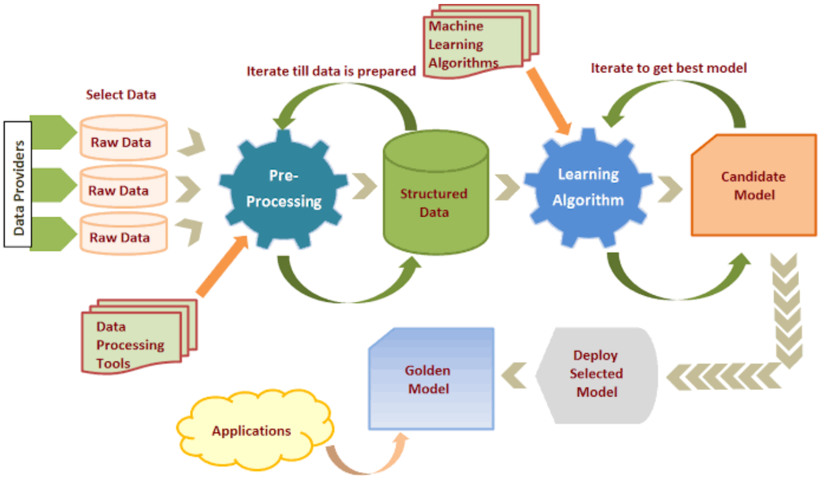


Fig 2.2: Working of machine learning

2.1.3 Classifying the machine learning

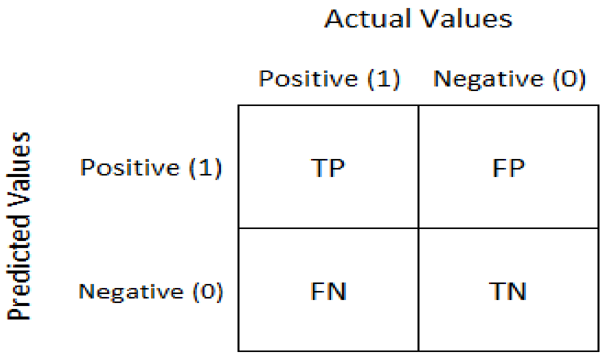
1.Supervised Machine Learning: This kind of classification technique is applied to a dataset that prepares from the earlier outcomes and also using the info to detect the results of those data who are recent. Then the realized data is broken into parts then it actually articulates which is used in detecting the outcomes of newly prepared data. It can separate the dataset from the result and also the separation of spared dataset to identify the wrongdoings and allowing the change to take place and the preparation of the model.

2.Unsupervised AI: The calculation transfers from the AI that is regulated because they are engaged due to the model not being prepared and before that neither it is arranged and nor it is checked upon. Solo calculation prepares the framework a shrouded structure inside of the data that isn’t labeled and even detect the potential results along with using the examples meanwhile the anomalies are also removed.

3.Reinforcement Machine Learning: It is one of the learning which is based on prize preparing while the model is connecting with climate by performing the activities and pointing out the faults and rewards. The qualities of the support learning are usually experimenting and postponing the awards. Meanwhile the models subsumes the slip ups and even the new model is arranged for doing the communication with machines for deciding the results and to fortify its working and enhancing its execution.

2.1.4 Measuring the performance

Here are several techniques to measure the performance of algorithms. The notion used is called confusion matrix. It is a table which is used to identify the performance of the classification model in the database where we know the true values beforehand. And while it is quite simple to understand on the other side the items that are related to each other are distracting.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Table 2.1: Confusion matrix

This above table is known as the confusion matrix. It includes the four sections. Namely true positive, true negative are well calculated observations. Other measures being false negatives and false positives which are the wrong predictions and are diminished whenever possible.

1.**True Positives:**These are quantities that are truly detected and they are qualities which are positive and are analyzed as the true prediction of true class and positive prediction of the anticipated class. It is denoted by TP.

2.True negative: It is calculated by the extent of negatives that are a collection of them. It is denoted by TN.

3.False Positive: They are unfortunately believed to be the values and after that they work by considering the other issues. Denoted by FP.

4.False negative: They are the wrongly predicted values and then work incorrectly in primitive section. Denoted by FN.

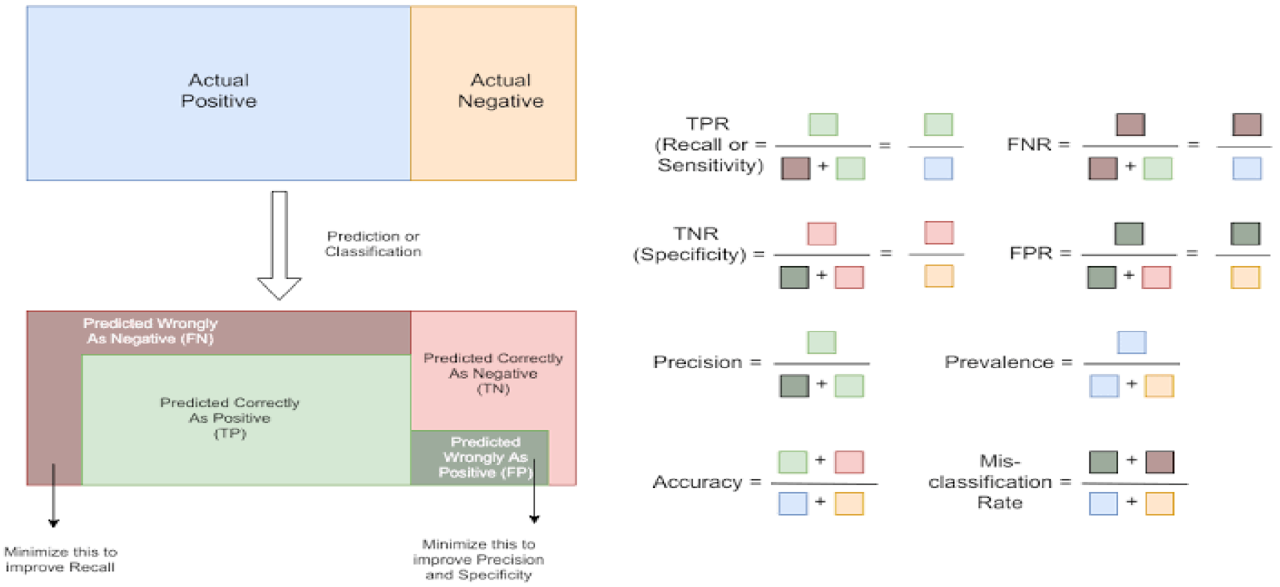

Actual Positive
Actual Negative
Prediction or Classification
Predicted Wrongly As Negative (FN)
Predicted Correctly As Positive (TP)
Predicted Correctly As Negative (TN)
Predicted Wrongly As Positive (FP)
Minimize this to improve Recall
Minimize this to improve Precision and Specificity
TPR (Recall or Sensitivity) =
FNR =
TNR (Specificity) =
FPR =
Precision =
Prevalence =
Accuracy =
Mis-classification Rate =

Fig 2.4: confusion matrix

Accuracy- It is the most used process which actually covers the most of the things in it for exactly and precisely predicting and visualizing the model. In most of the cases when there are large consistencies, the model works appropriately. It is that measure which can't be seen in case of quantitative dataset where our predictions of false negatives and false positives are nearly same. For eg there are in total 1000 cases, negative cases are 995 and 5 cases are positive among them. And if our system predicts them all negatives, then the accuracy would be 99%, despite missing all of the positive cases from the data.

$$\text{Accuracy} = TP + TN / TP + FP + FN + TN$$

Precision- It is that measure which has the ability to anticipate and have the best understanding of this experience. The rating response which is the request, it is known to be for all the testers who are suffering, telling the persistence cases. If the accuracy is large then the false positive rate is also high.

Accuracy = TP / TP + FP

Recall-It is also called the ability to detect feeling which is a component and part of the total number of cases which are related which are found in for real.

Recall = TP / TP + FN

Chapter-3 System Development

(3.1)Requirements of the system

The algorithms utilized in this project require some standard programming because itrequires the processing of algorithms.

- Windows 10 (64-bit)
- JupyterNotebook
- Python
- 4 GBRAM
- Intel (R) Core (TM) i3processor

(3.2)Why Python?

It is a language to do the programing consisting of various factors and is very much accurate in it. It also provides the segment of groups making the most exceedingly awful of errands or undertakings troublesome. Python has libraries in the records utilized for instance - working with pictures, which works with the recordings of the sound. Regardless, when working with another OS, python is awesome. Python is an incredible organization that makes it simple to search for help and tips and deceives.

(3.3) Scikit Learn

Scikit read Python library is frequently used for machine learning and is in a position to include various returns, classifications and compilationalgorithms.

(3.4) The Pandas

Pandas is an open-source, easy-to-use datastructures anddata analysis toolsfor the Python programing language . Python with Pandas is employed during a wide selection of fields including academic and commercialdomains.

(3.5) Numpy

It stands for numerical calculations including fourier transform, linear algebra etc. It is that library if python which consists of an array n- dimensional article, enabling the devices to work in C and C++ etc. Likewise helpful in arithmetical request, arbitrary number force and so forth is a compartment of different sizes of standard information.

(3.6) Anaconda

It is a platform in data science which leads to the machine learning environment of open source.

(3.7) Matplotlib

Matplotlib isaPythonprogramminglibrarythatprovidesanAPI-basedapplicationforembedding sites into applications. It's very similar to MATLAB embedded in Python programming language. Histogram, bar plots, streaming plots, pie plot area, Matplotlib can show a wide range of observations. With a little effort and a tint of visual skills, with Matplotlib, we can create any observation

(3.8) Jupyter Notebook

Jupyter Notebook is a natural computational environment for building Jupyter scripts. It is a document that follows a modified structure, and contains an ordered list of input / output cells that can contain code, text, statistics, sites and rich media. It usually ends with the ".ipynb" extension.

(3.9) Tensorflow It is the library in python that was introduced by the google company in order to execute deep learning models including machine learning models as well as soon as possible. It helps in simple calculations of various mathematical problems by merging the algebra of optimization methods.

(3.10) Keras is a profound learning API written in Python, running on top of the AI stage TensorFlow. For empowering quick experimentation keras was created.

Keras is additionally an exceptionally adaptable system reasonable to emphasize on cutting edge research thoughts. Keras follows the standard of reformist divulgence of intricacy: it makes it simple to begin, yet it makes it conceivable to deal with subjectively progressed use cases, just requiring gradual learning at each step.Keras is utilized by CERN, NASA, NIH, and a lot more logical associations around the globe (and truly, Keras is utilized at the LHC). Keras has the low-level adaptability to execute discretionary exploration thoughts while offering discretionary elevated level comfort highlights to accelerate experimentation cycles

Chapter-4 Performance analysis

Features according to their respective scores

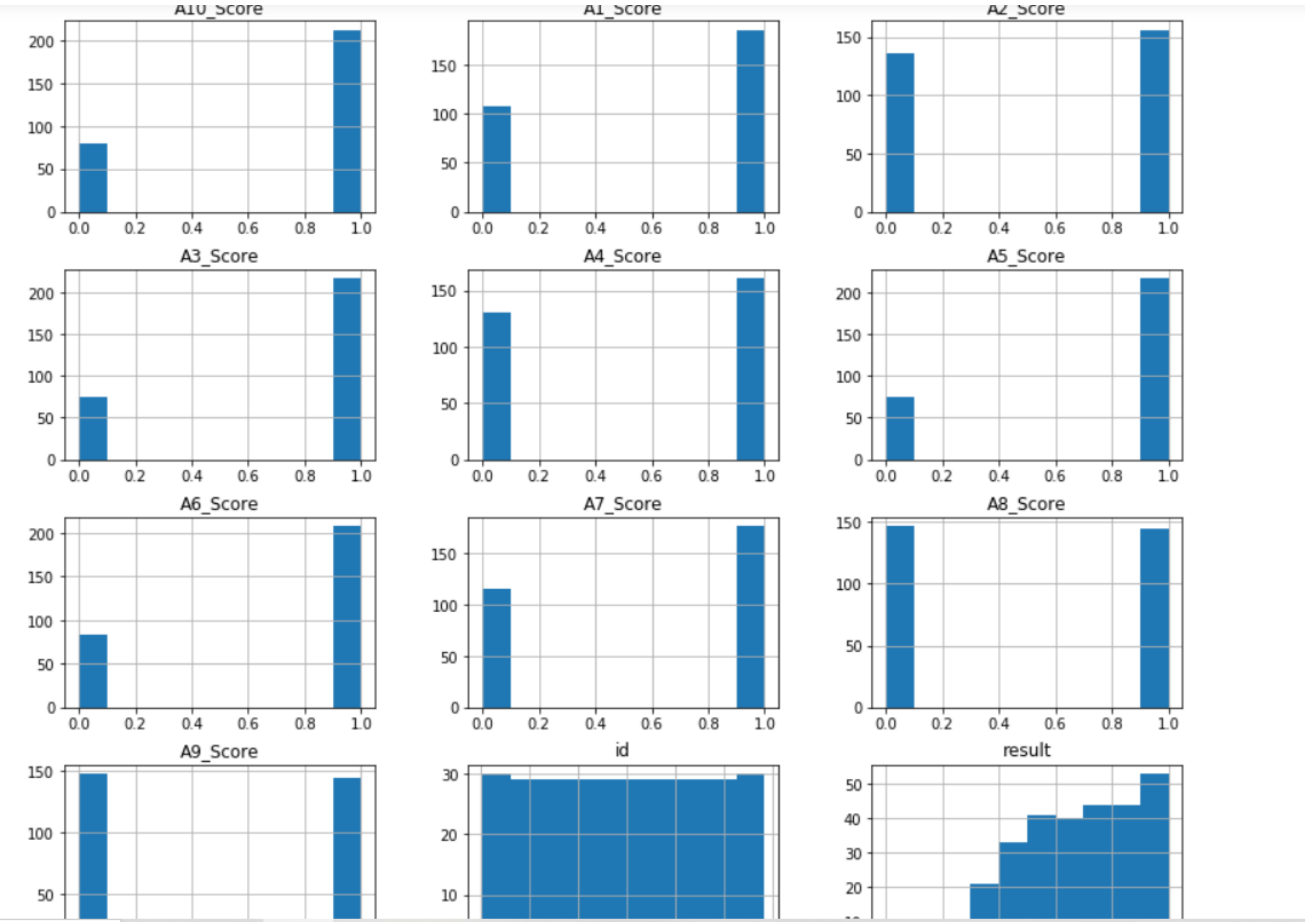


Fig 4.1: correlation between eachcolumn

Decision Tree

	Precision	**Recall**	**F-Measure**
0	0.84	0.33	0.47
1	0.52	0.92	0.67

Table 4.1: Decision Tree

Logistic Regression

	Precision	**Recall**	**F-Measure**
0	0.67	0.34	0.45
1	0.55	0.83	0.66

Table 4.2: Logistic Regression

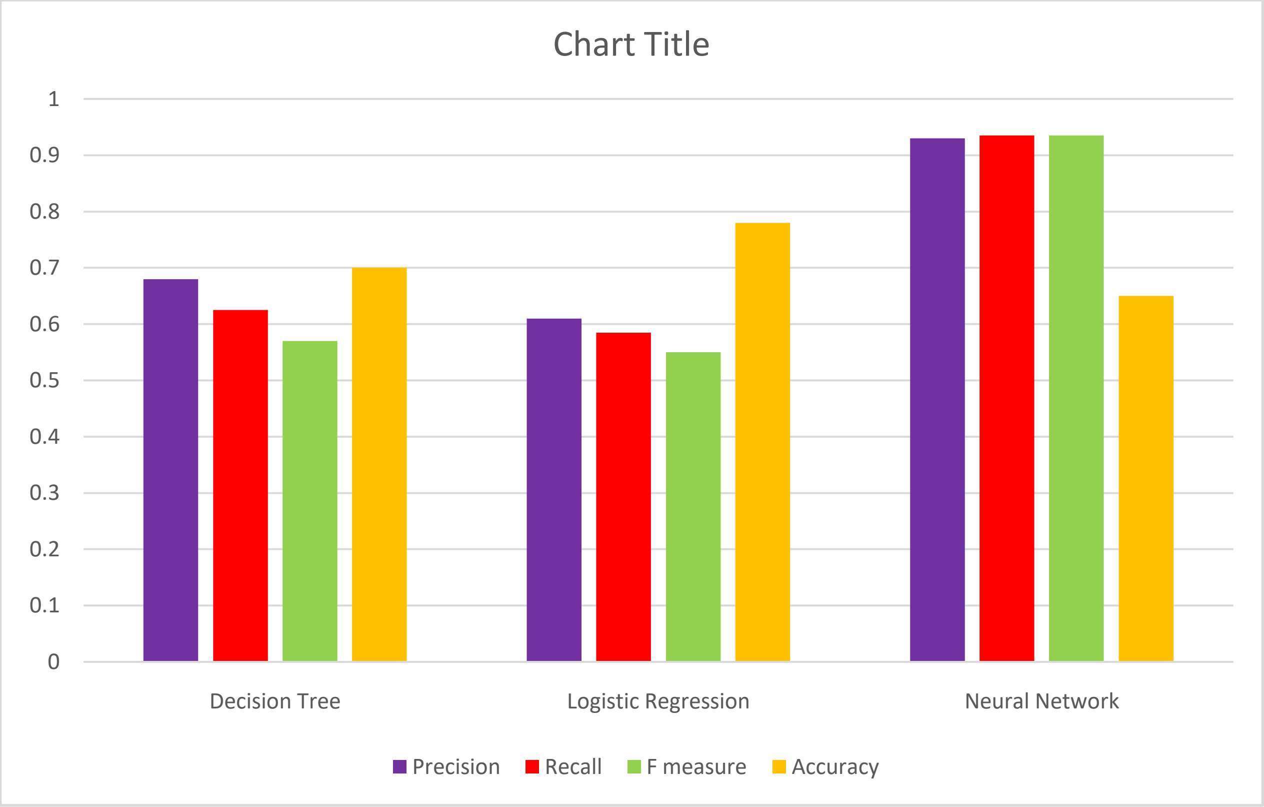
Neural Networks

	Precision	**Recall**	**F-Measure**
0	0.91	0.97	0.94
1	0.96	0.90	0.93

Table 4.3: Neural Networks

Algorithm Names	**Decision Trees**	**Logistic Regression**	**Neural Networks**
Precision	0.68	0.61	0.935
Recall	0.625	0.585	0.935
F measure	0.57	0.55	0.935
Accuracy	0.70	0.78	0.65

Table 4.4: Average of Precision, Recall, F-Measure and Accuracy



Graph 4.1: Comparative analysis of all the algorithms

Corresponding to classifiers performance over Accuracy, Precision F-measure and Recall listed in Table 4 it is analyzed that Neural Networks showing the maximum accuracy. So the Neural Network classifiercanpredict thechancesofdiabeteswithmoreaccuracyascomparedto otherclassifiers.

Chapter 5 Conclusion

One of the foremost important real medical problems is that the early detection of Autism. during this study,accurated effortswere made to develop a system that resulted within the prediction of diseases like Autism. During this work, the three algorithms Decision Tree, Logistic Regression and neural networkare applied and checked by various outcome. The study was conducted on the Autistic Spectrum Disorder Screening Data for Children Data Set.

. Test results determine the efficiency of the system with an accuracy of 93% using the Neural Network algorithm. within the forecast, a program designed with excestingalgorithms for machine learning are often wont to predict or diagnose .Thework are often expanded and improved on the automation function using differentalgorithms.

Future Scope

As theresult of this project, an autism prediction model can be developed by merging Random Forest-CART (Classification and Regression Trees) and also a mobile application can be developed based on the proposed prediction model.

.

References

[1] http://neuralnetworksanddeeplearning.com/

[2] History of machine learning, Online available at: https://www.doc.ic.ac.uk/~jce317/history- machine-learning.html

[3] Regression formula,Online available at:https://www.wallstreetmojo.com/regression-formula/ https://www.geeksforgeeks.org/machine-learning/

[4] Understanding confusion matrix, Online available at: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

[5] Confusion matrix, Online available at:https://devopedia.org/confusion-matrix

[6] https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult

[7] https://towardsai.net/p/programming/decision-trees-explained-with-a-practical-example-fe47872d3b53

[8] https://www.kaggle.com/faressayah/ensemble-ml-algorithms-bagging-boosting- voting/comments#711179

[9] https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861

[10] https://towardsdatascience.com/machine-learning-workflow-on-diabetes-data-part-01- 573864fcc6b8

[11] https://www.dataquest.io/blog/top-10-machine-learning-algorithms-for-beginners/