Classification Of Autism Data Using Different Classification Techniques

Priyal Sobti
Computer Science Department
Bharati Vidyapeeth's College Of Engineering
Delhi,India
priyalsobti@gmail.com

Niharika
Computer Science Department
Bharati Vidyapeeth's College Of Engineering
Delhi,India
niharika12oct@gmail.com

Pulkit Gupta
Computer Science Department
Bharati Vidyapeeth's College Of Engineering
Delhi,India
pulkitgupta078@gmail.com

Ishita Agarwal
Computer Science Department
Bharati Vidyapeeth's College Of Engineering
Delhi,India
ishita.nishu1999@gmail.com

Nikita Jain
Assistant professor and Research Scholar
Computer Science Department
Bharati Vidyapeeth's College Of Engineering
Delhi,India
nikitajain1210@gmail.com

Abstract: **The research paper focuses on the use of fuzzy approaches to help classify the medical data collected on Autism.The Autism data was first classified using other models namely Logistic Regression wherein the results collected could be compared with the other models.The results from other models can help us identify how classification is done better by a specific model under different circumstances and how data also plays a major role.A number of parameters were evaluated in the dataset like the patients have had any medical ailments like Jaundice or whether autism is hereditary in their families or not.The patients were asked to answer certain questions where the answers can help to identify whether the patients have the disease or not. The classification further helps in survival analysis i.e. the prediction of the survival of the sufferers.**

Keywords: Fuzzy, Autism Spectrum Disorder, ABIDE, Classification,Logistic Regression,Binary

Classification, Jaundice, Autism, Kaggle, Artificial Intelligence, Machine Learning.

I. INTRODUCTION

Autism is neurodevelopmental disorder consisting of problems in social interaction and communication.It is generally identified in infants and children who are two to three. Unfortunately, waiting times for an ASD diagnosis are lengthy and procedures are not cost effective. The economic impact of autism and the increase in the number of ASD cases across the world reveals an urgent need for the development of easily implemented and effective screening methods. Therefore, a time-efficient and accessible ASD screening is imminent to help health professionals and inform individuals whether they should pursue formal clinical diagnosis. The rapid growth in the number of ASD cases worldwide necessitates datasets related to behaviour traits. However, such datasets are rare making it difficult to perform thorough analyses to improve the efficiency, sensitivity, specificity and predictive accuracy of the ASD screening process. Presently, very limited autism datasets associated with clinical or screening are available and most of them are genetic in nature. Hence, we used a dataset related to autism screening of toddlers that contained influential features to be utilised for further analysis especially in determining autistic traits and improving the classification of ASD cases. In this dataset, we utilized ten behavioural features (Q-Chat-10) plus other individuals characteristics that have proved to be effective in detecting the ASD cases from controls in behaviour science years of age.1 in 45 children, ages 3 through 17, have been diagnosed with autism spectrum disorder (ASD).
The data was collected from **Kaggle.**

II.PROBLEM BACKGROUND AND PREVIOUS WORKS

Autism has been studied previously with the help of artificial intelligence.
Data collected in the study have been able to predict autism with 60% accuracy.
The data studied had been collected from the Kaggle wherein the patients have answered certain questions that help to identify symptoms for the same.
The study had been performed on multisite data i.e data collected from different labs.

III.SCOPE OF THE PROJECT

The research aims to diagnose the disease using answers given by the patients and other factors like jaundice or the presence of Autism in the family.

Machine learning algorithms shall be used to classify the data used and to tell whether the particular patient has autism or not. Certain other datasets containing the study of genomes in such diseases will also be used.
Survival analysis is defined as the methods used to analyse the data that produces an output as time until the occurrence of an event of interest.The event can be death, occurrence of a disease, marriage, divorce, etc.
The focus of this study is on autism spectrum disorder and the prediction of survival of its sufferers.

The symptoms for the disease are visible from an early age. The improvement or worsening of symptoms needs to be tracked in patients (termed prognosis) which helps in survival analysis.

The project helps to understand the current scenario of AI in the health based industry.
The project aims at studying the patient's record and understand the development of the disease.
The project also aims at diagnosing the disease by studying data of the patients obtained using answers to certain questions asked by a family member or medical practitioner .

The diagnosis of such diseases in time can help the medical practitioner to help the patient effectively by taking the necessary steps.
The answers to such questions help to know about the response of the patient.

IV OBJECTIVES OF THE STUDY

The study aims to take the studies further in this domain and bring about a change in the diagnosis of some untreatable diseases.
The large volume of medical data should be studied using machine learning algorithms and deep learning approaches.
Hence, the approaches being used for the study are artificial intelligence based.
Such a step in the medical industry helps to reduce the diagnostic and therapeutic errors that are inevitable in the human clinical practice.

V. IMPLEMENTATION

The implementation was carried out using a few classification techniques namely logistic regression, random tree classifier and decision tree.
The dataset were simultaneously analysed to evaluate and compare the datasets.
Both the dataset had similar features and the respective features were studied to make deductions.

```
Toddlersfor dataset1:  69.07020872865274
Toddlers for dataset2:  26.84659090909091
```

Fig 1 : The percentage of toddlers suffering from Autism as per the dataset in different parts of the world.

Further deductions were made to find the distribution of patients suffering from Jaundice and the age distribution of the total number of subjects in the dataset who are mainly toddlers.

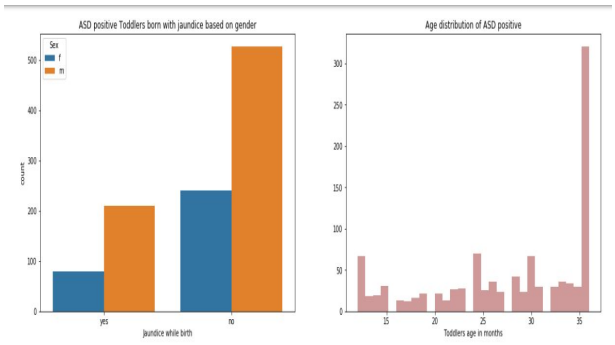

Fig 2 : Distribution of Toddlers with Jaundice on the basis of gender and age distribution of ASD Positive Children for one dataset

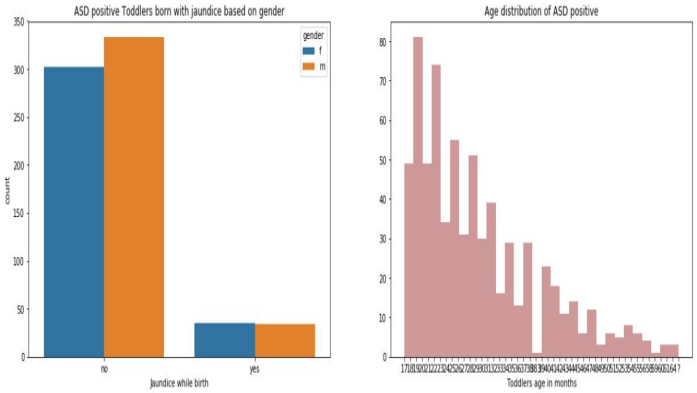

Fig 3 : Distribution of Toddlers with Jaundice on the basis of gender and age distribution of ASD Positive Children for other dataset

Various models have been used for implementation namely Logistic Regression, Gaussian Naive bayes,Stochastic Gradient Descent,Random Forest and Decision Tree

Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.
An explanation of logistic regression can begin with an explanation of the standard logistic function. The logistic function is a Sigmoid function, which takes any real value between zero and one.
It is defined as

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

Fig 4: Mathematical representation for Logistic Regression

Let's consider ***t*** as linear function in a univariate regression model.

$$t = \beta_0 + \beta_1 x$$

Fig 5: Mathematical Representation for Logistic Regression

Flowchart for logistic regression algorithm:

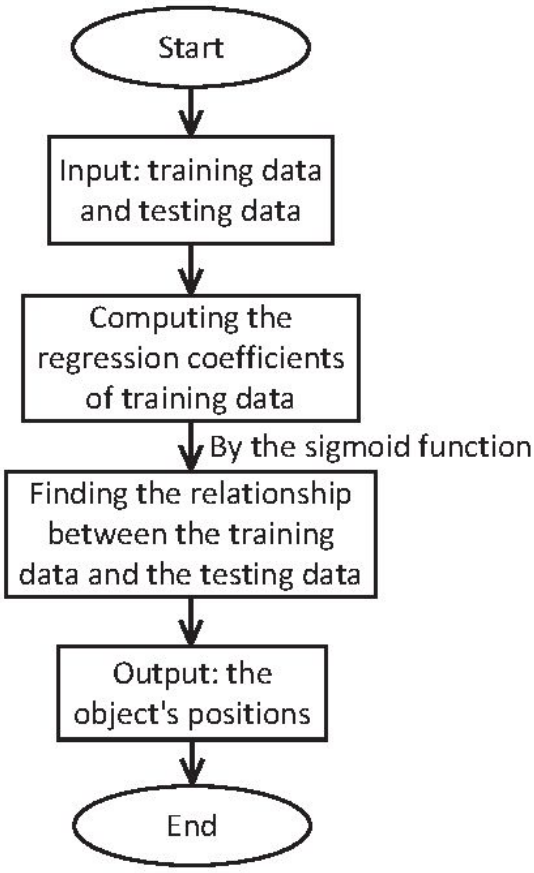

Fig 6: Algorithmic depiction of the model used

Mathematically, logistic regression estimates a multiple linear regression function defined as:

$$= \log\left(\frac{p(y=1)}{1-(p=1)}\right) = \beta_0 + \beta_1 x_{i2} + \beta_2 x_{i2} + \_ + \beta_p x_{in}$$

Fig 7 : The mathematical multiple regression function used.

For binary classification the data was converted into binary using general techniques and one-hot encoding.

The features namely gender, Jaundice and Age were mapped with 1 and 0.

Mean for the age was calculated and the mapping was done on the basis of the mean obtained.

If the age was greater was 27 months that was mapped to 1 and age less 27 months was mapped to 0.This was done on both the datasets.

Similar mapping for gender was implemented.

If the toddler was male, it was mapped to 1 and if the toddler was female it was mapped to 0.

After the mapping and conversion into binary classification, the model for logistic regression was trained and tested.

The features namely "ASD_Traits" for dataset1 and "asd_Traits" for dataset2 were used as the target variable and the rest of the features were used as input variable.

The accuracies for the models were hence compared.

The same concept was applied on different algorithms and the performance has been judged using various metrics.

The comparison for the same is shown in a tabular form:

S no	Name	Accuracy	F1 Score
1.	Logistic Regression	95%	0.73
2.	Stochastic Gradient	100%	1.0
3.	Random Forest	100%	0.96
4.	Gaussian Naive Bayes	94%	0.96
5.	Perceptron	99%	0.98
6.	Linear Support Vector Classifier	100%	1.0
7.	Decision Tree	100%	0.93

As per the table , it is clear that Stochastic Gradient and Linear Support Vector Classifier have both accuracy and f-1 score as 100% and 1 respectively.

This study helps to prove that these algorithms can be further used for analysis in this field.

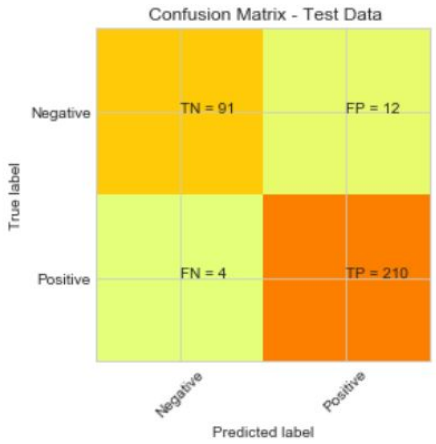

Fig 8 : This is the confusion matrix for Stochastic Gradient

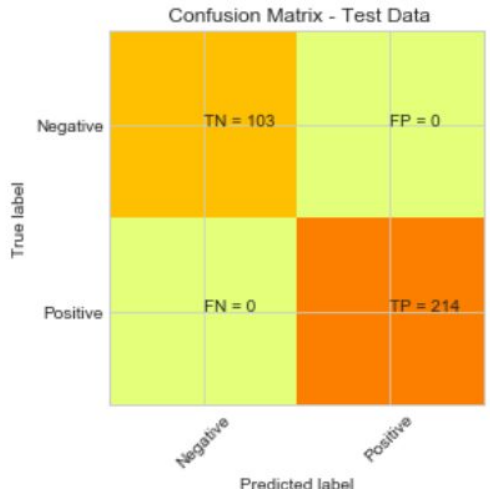

Fig 9 : Confusion Matrix for Linear Support Vector Classifier

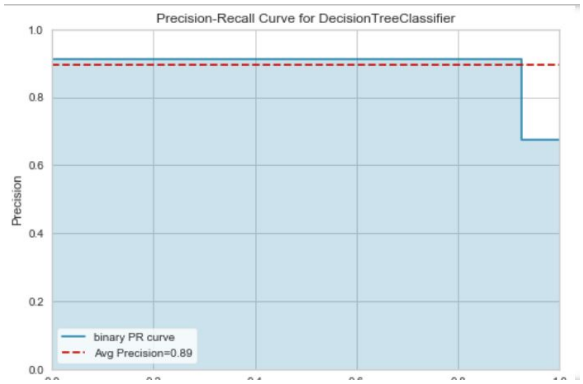

Fig 10 : Precision Recall Curve for Decision Tree Classifier

V. REFERENCES

[1]Yangyang Li1, Guoli Yang1,Haiyang He1, Licheng Jiao1 and Ronghua Shang ," A study of large-scale data clustering based on fuzzy clustering".

[2] Yan Li,Mohammed Diykh ," Fuzzy and Non-Fuzzy approaches for digital image Clustering".

[3] Sofia Visa,et al.,"Fuzzy Classifier for Classification Of Medical Data"

[4]Ashutosh Malviyel ,Liliane Peters,"Fuzzy Handwriting Classification : FOHDEL".

[5]Alessandro Crippa,Christian Salvatore,et al.,"Use of Machine Learning to Identify Children with Autism and Their Motor Abnormalities"

[6]Bram van den Bekerom ,"Using Machine Learning for Detection of Autism Spectrum Disorder"

[7]AbbasH,Garberson F,Glover E,Wall DP,"Machine learning approach for early detection of autism by combining questionnaire and home video screening".

[8] P. K, S. Srinath, S. Seshadri, S. Girimaji, and J. Kommu. Lost time: Need for more awareness in early intervention of autism spectrum disorder. Asian Journal of Psychiatry, 25:13–15, 2017.

[9] A. Keil, C. Breunig, S. Fleischfresser, and E. Oftedahl. Promoting routine use of developmental and autism-specific screening tools by pediatric primary care clinicians. Wisconsin Medical Journal, 113(6), 2014.

[10] I. Kononenko. Inductive and Bayesian Learning in Medical Diagnosis. 1993.

[11] I. Koprinska, M. Rana, and V. G. Agelidis. Correlation and instance based feature selection for electricity load forecasting.