International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 03 | March - 2023 Impact Factor: 7.185 ISSN: 2582-3930

Study and Research on Autism Spectrum Disorder using Supervised Machine Learning Techniques

[1]Asst.Prof.Vinod Babu P,[2]Mr.Kakarla Sanchit, [3]Mr.Nimmadala Nithin,
[4]Mr.P Saketh, [5]Mr.Nekkanti M Sri Rama Chowdary
[1]Asst.Professor,[2345]UG Students,[12345]Computer Engineering Department.
Gandhi Institute of Technology and Management, Visakhapatnam, India.

Abstract-Machine learning is embedded in our life so deeply it is used around the world by various organizations and in various fields,mostly the healthcare industry it brought a drastic difference in the industry .Autism Spectrum Disorder is affecting the people unknowingly well we are predicting the ASD on advance on bases of different aspects mainly the way the people interact/communicate with others ,the way the do there tasks,The way they react to things ,it is very essential to detect ASD in early stage so that we can control it from being worse.In this study we used child,Adult,Adolescent data set and we used Logistic Regression ,KNN,Random Forest,SVMand Naive Bayes to predict weather the person is having ASD .To know the accurate accuracy of the above algorithm we have used 20 percent data for testing and remaining for training the algorithms .As the outcome the Logistic Regression got 95.0% and KNN ended up with 84.72% for optimal k value and Random Forest and SVM,Naive Bayes got 96.36% and 86.36%,95.15%

Keywords-*Autism Spectrum Disorder,Logistic Regression,KNN,Random Forest,SVM,Naive Bayes*

I.INTRODUCTION

Autism Spectrum Disorder is a genetic and Neurological Disorder this is mainly possed by 1 out of every 150 in 2000,In 2020 the Digits has changed and became 1 out of every 100 ,it is very important to detect and find ASD in individual on before had because if we find we cant cure it but we can stop it from being worse many study prove that symptoms are mostly diagnosed in children in 18 months of age but few studies aslo told they will possessed after school age also may be after ther teenage so we got 3 different data sets which are of 3 different age groups

child,Adult,Adolsent Mainly the ASD can be predicted on the bases if the physical test conducted by the concern doctors ,our data set is having 10 attributes which are mainly based on the [1]how the individual reacts to the environment ,[2]how the react to his/her name,[3]How they give eye contact to that person,[4]How they indicate if they want something,[5]How they are looking at someone ,[6]How they are behaving out of their comfort zone,[7]How the person is pretending,[8]How they are using simple gestures ,[9]Would you describe initial word,[10]How they stare ,on the based up on how they react to these 10 question we analyze them and attached there age nationality and few more attributes such as family history and we prepared the data and got ready to feed the machine and we have used 4 different machine learning techniques which are Logistic Regression,K-NearestNeighbor,SVM,Random Forest,Naive Bayes for accurate result and to avoid underfitting and overfitting we had fited our model with 80 percent training data and rest rest of the data is used for testing after vigorous amount of train as an end result we were ended up with 95.0% accuracy for the Logistic Regression and 84.72% for KNN and SVM ,Random Forest,Naive Bayes ended up with 86.36%,96.36% ,95.15%

II.Research Methodology

There are numerous amount of research and they are still under process until till date there is no specific reason for ASD but mostly all the cases are caused due to Neurological and Genetic caused due to the family history of specific person during our research and studied we found that there are few studies which have proved that autism can be only detected in children who are between 18 months old but not in all cases some cases also said that ASD can be detected at their school age and some at after the teenage so we have found a data sent which are divided into 3 different parts which are kid,adult,adolescent the main reason we divided our data set into 3 parts because we want to train our model and predict weather the individual is having ASD or not irrespective of their age.it is very important to detect/predict whether individual is having asd or not they might not be studies which conclude that if we detect the ASD we can prevent them but they are numerous amount of studies which prove that we can stop the ASD in a person

from being more worse.On the other hand we have found and got accurate data so we can used Supervised machine learning techniques to get accurate results and end up with a well fitted models so we choose Machine Learning models such as Linear Regression, KNN, SVM, RandomForest, Naive Bayes for better prediction/accuracy .

III.Data Analysis

Our data set consists of 3 different age groups namely Child,Adult and Adolescent. These datasets are collected from UCI Repository provided by Fadi Fayez Thabtah.

[Tabtah F].These datasets are suitable for classification and regression models in machine learning.After we examined three datasets for analysis- ASD screening Child Dataset(292,21)/Adolescent Dataset(104,21) Adult Dataset(704,21),The child dataset has 292 records,adolescent and adult dataset has 104 and 704 instances with 21 attributes. Out of 21 attributes, 10 are questionnaire types(AQ1-AQ10).The records or values of this 10 attributes are in binary form ie.either 0 or 1.These questions define the behavior of patients related to Autism and 10 demographic features with one class label.class label also contains binary values ie.0 or 1.The information gives clear picture to ASD diagnosis.Below table will give data description of all 3 datasets

DATASET EXAMINATION:

Attribute name	**Type**
A1_Score A10_ Score	Nominal
age	Numeric
gender	Nominal
ethnicity	Nominal
jaundice	Nominal
autism	Nominal
age_desc	Nominal
country_of_res	Nominal
used_app_before	Nominal
result	Numeric
relation	Nominal
Class/ASD	Nominal

IV.Data Pre-Processing

After studying about the data there are many Diagnosis we needed to do for the present data the first 10 attributes were having string as there data type so we needed to change it to binary so that the machine can understand we used label encoded and standard scalar to convert all the string values into binary and integers within specific range there are many kinds of duplicate values which may lead to retrain the model so we initially dropped all the duplicate data and we had also filled null values by taking out the mean of the specific column and made the data ready for fitting and testing

V.Architecture of the Model

Initially we took the data set from UCI repository as already mentioned in dataset analysis and pre-proceed the specific data set with many different techniques,like standard scalar and we choose specific labels by using label encoder and classified the processed data and gave it to the 5 Supervised Machine Learning algorithms/models that we have used to detect whether the end person is having Autism or not.we use five classification algorithms in machine learning namely Logistic regression,Random Forest, K-Nearest neighbor,Support Vector machine and Naive Bayes.we took best and optimum accuracy of above models we will discuss further .Below Diagram is the flow of Architecture we used for Autism Spectrum Disorder.

Flow of Work :

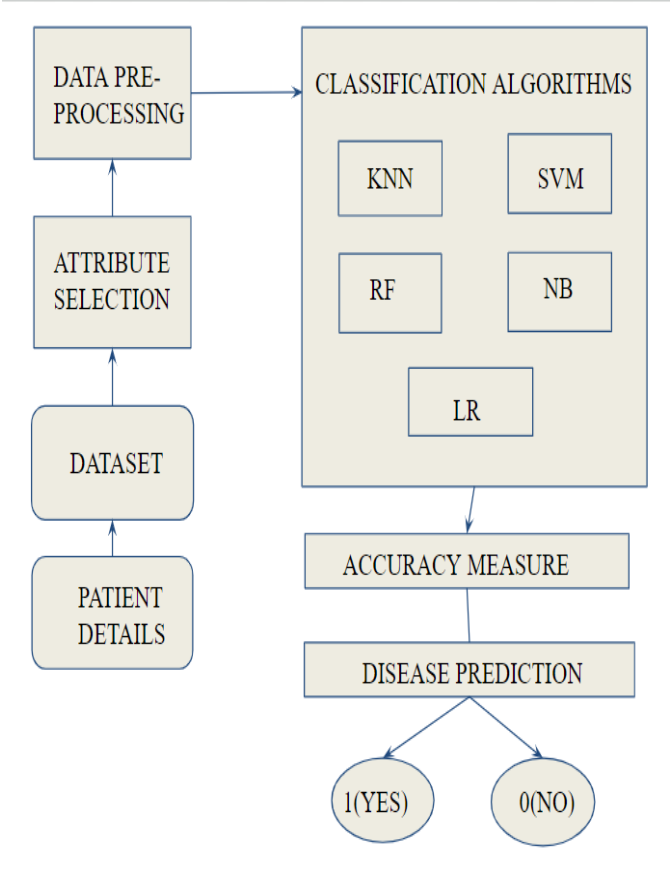

Fig 1:Architecture

VI.Models

a)KNN:

The term KNN stands from K-Nearest Neighbour it is a type of supervised machine learning technique where the whole algorithm is mainly based on the k value it,initially all the distance from uniknow point is calculated and based on the k value the nearest k values points are chosen and the unknown point will be categorized.here in this k value is very crucial for getting the optimal accuracy so we have used seeding point algorithm with the help error we could find the optimal accuracy .in the below Fig we can clearly see than at K=5 the error is minimum so we took the min error k value as input for our KNN model , the reason we did this because to get the optical and better accuracy for the final model .

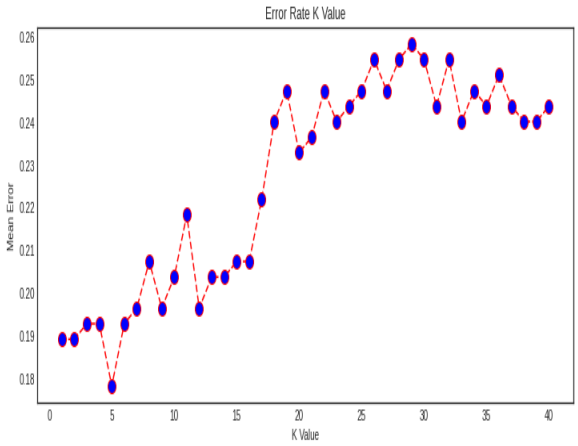

Fig 2

b)Logistic Regression:

Logistic Regression is a regression tool used to evaluate binary dependent variables. Its output number is either 0 or 1. It is used for the continuous value collection. It describes the connection between one dependent binary variable and one nominal or ordinal variable. It can be expressed by the sigmoidal function.

c)Random Forest:

Random Forest Algorithm is widely known for classification and regression, it combines the outcome of several decision trees to obtain a result.

The decision tree consists of the following parameters n_estimators, max_features, max_depth, max_leafnodes, max_samples.Random Forest relies on ensemble learning and uses two techniques Bagging and Boosting.

Measure of Variance

$$Gini\ Index = 1 - \sum_{i=1}^{n}\left(P_i\right)^2$$
$$= 1 - [(P_+)^2 + (P_-)^2]$$

Fig 3

d)Support Vector Machine:

Notable Supervised learning algorithms for classification and regression analysis include Support Vector Machine (SVM).Finding the hyperplane that best divides the data points into different classes is the main goal of SVM.The hyperplane is employed in classification to establish a decision boundary that divides data points into various classes By measuring the distance between the hyperplane and the nearest data points for each class,Support Vectors are the data points that are most closely spaced from the hyperplane.SVM can handle both linear and non linear separable data we are working on binary classification

e)Naive Bayes:

Naive Bayes is a machine learning algorithm for classification and prediction problems that is based on the Bayes theorem.Because it presumes that the features(or variables) used in the model are independent of one another which not always the case in reality,the algorithm is referred to as “naive”.

In a Naive Bayes classifier, the probability of a given output(or class) is determined based on the probabilities of the input features

$$P(Y|X_1,\ldots,X_n) = \frac{P(X_1,\ldots,X_n|Y)P(Y)}{P(X_1,\ldots,X_n)}$$

Fig 4: Bayes Theorem

Here, Y is class or target feature X1,X2,……,Xn are independent features to determine the probability.Naive Bayes is a fast and efficient algorithm

VII.Training and Testing:

We take three datasets called child,adult and adolescent. Each one consists of some instances on the basis of data collected. We combined all three data sets into 1 dataset in .csv format.After making the dataset we have 1100 instances and 21 attributes in which we take 80% for training and 20% for testing.It means 825 instances for training and 275 records for testing 20 attributes are independent ones and the remaining one attribute is class/ASD feature.

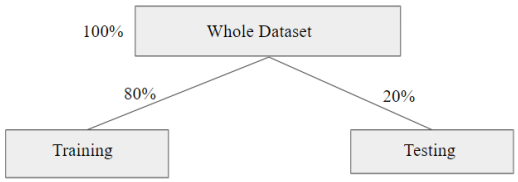

Fig 5:image for training and testing

VIII.Data Visualization:

So as per data we intially made age frequency using voiln plot in fig 6 we clearly observe Adult data has maximum instances and followed by adolescent and child datasets.Child data has least instances compared to adolescent.We observe that age 10 to 20 has maximum instances

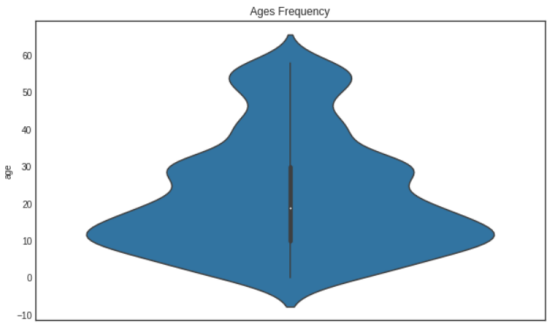

Fig 6:Violin plot

To check male and female comparison from our dataset we plot barplot in that we observe male instances are more compared to females as shown in fig 7.ie around 600 Males and 500 Females roundly present in the dataset

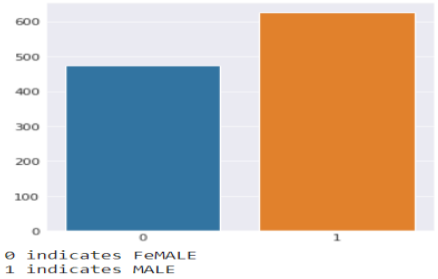

Fig 7:Bar plot

Here Fig 8 depicts how A1 to A10 have answers for questions in binary values, that is if a patient has that problem mentioned in the question is true then it is 1 else it is 0. Below image consist of 10 histogram plot

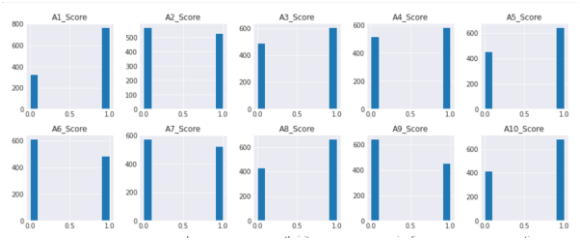

Fig 8:Histogram plot

The Fig 9 shows the count of Traits that have autism or not.ie. the number of people affected by autism or not

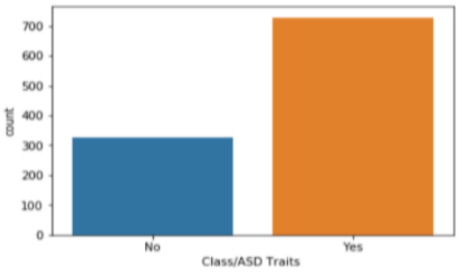

Fig 9:count of traits

IX.Displaying Results:

Accuracy score: 0.9515

	precision	recall	f1-score	support
0	0.98	0.95	0.97	183
1	0.91	0.97	0.94	92
accuracy			0.96	275
macro avg	0.95	0.96	0.95	275
weighted avg	0.96	0.96	0.96	275

Fig 10: Results of NB

Test set score: 0.8636

	precision	recall	f1-score	support
0	0.85	0.95	0.90	142
1	0.89	0.71	0.79	78
accuracy			0.86	220
macro avg	0.87	0.83	0.84	220
weighted avg	0.87	0.86	0.86	220

Fig 11:Results of SVM

Accuracy score: 0.8472727272727273

	precision	recall	f1-score	support
0	0.86	0.94	0.90	191
1	0.82	0.64	0.72	84
accuracy			0.85	275
macro avg	0.84	0.79	0.81	275
weighted avg	0.84	0.85	0.84	275

Fig 12:Results of KNN

```
#Accuracy Score
print('Accuracy score:', accuracy_score(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

Accuracy score: 0.9636363636363636

	precision	recall	f1-score	support
0	0.95	1.00	0.97	191
1	1.00	0.88	0.94	84
accuracy			0.96	275
macro avg	0.98	0.94	0.96	275
weighted avg	0.97	0.96	0.96	275

Fig 13:Results of RF

Accuracy score for Test data: 0.95

	precision	recall	f1-score	support
0	0.97	0.96	0.96	158
1	0.89	0.94	0.91	62
accuracy			0.95	220
macro avg	0.93	0.95	0.94	220
weighted avg	0.95	0.95	0.95	220

Fig 14:Results of LR

Regarding Confusion Matrix:

		Predictive values	
		False	True
Actual values	False	TN=True Negative	FP=False Positive
	True	FN=False Negative	TP=True Positive

Fig 15 confusion matrix

As shown in fig 15 confusion matrix is a matrix table consist 2x2 rows and columns It shows predictive values and actual values as the table axis.It consists True positive,True negative,False Positive and False Negative all these are used to calculate Recall,precision,F-measure and accuracy.

$$\text{Accuracy}=(TN+TP)/(TN+TP+FN+FP) \quad (1)$$

$$\text{Recall}=TP/(TP+FN) \quad (2)$$

$$\text{Precision}=TP/(TP+FP) \quad (3)$$

$$\text{F-measure}=2(\text{Recall}\times\text{Precision})/(\text{Recall}+\text{Precision}) \quad (4)$$

Fig 16:Formulas

Below fig 17 is the bar plot to visualize accuracy of classification models namely LR,NB,SVM,KNN,RF we got LR-95.0%, NB-95.15%,SVM-86.36%,RF-96.36%, KNN-84.72%.

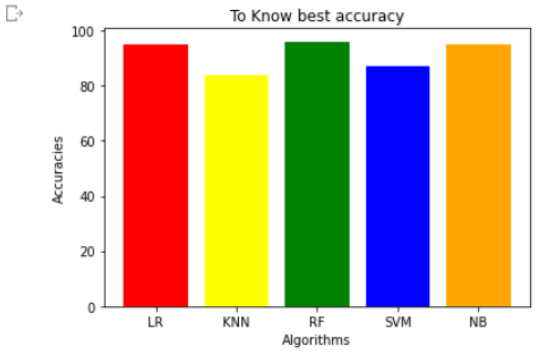

Fig 17 Accuracies

XI.Conclusion

After vigorous amount of processing the data with all the five algorithms which are Random forest,KNN,SVM,Logistic Regression,Navie Bias we finally ended up with Random forest as the most optimal one and it got the most accurate result which is 96.36,followed by that naive bias also got accuracy of 95.15,and remaining SVM,KNN,Linear Regression are ended up with 86.36%,84.72%,95.0%.To conclude with we have finally got Random Forest as our best and most accurate correctly predicted model .

XI.Future Scope

The lack of adequate information to develop the prediction system is the project's main constraint. Our next steps will be to gather more data from various sources and to enhance the efficiency of the proposed machine learning classifier.In the future, we hope to create a flask programme that is more reliable for users.

XII.References

[1]A Comparative Analysis of Prediction of Autism Spectrum Disorder (ASD) using Machine Learning.Vaibhav Vishal,Abhishek Singh,Y.Bevish Jinila,Kavitha.C,S.Prayla Shyry,J.Jabez.ICOEI 2022.

[2]A Hybrid Recommender System using a Multi-Classifier Regression Model for Autism Detection.K. Vijayalakshmi,Dr. M. Vinayakamurthy, Dr. Anuradha.
(ICSTCEE 2020)

[3]Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques Suman Raja ,Sarfaraz Masoodb.
(ICCIDS 2019)

[4]Predicting ASD Using Optimized Machine Learning.Shaikhah Almana, Mustafa Hammad. 2022 International

Conference on Decision Aid Sciences and Applications (DASA)

[5]Prediction of Autism Spectrum Disorder using Random Forest Classifier in Adults.Kanchana,RashmitaKhilar. 2022 IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)

[6]Autism Spectrum Disorder Detection in Toddlers fro early diagnosis using machine learning .shirajul islam,tahmina akter sarah zakir,shareea sabreen muhammad iqbal hossain.(2020 IEEE)

[7]Autism Spectrum Disorder Detection using Machine Learning Approach .nabila zaman,Jannatul ferdus,abdus sattar(2021 IEEE)

[8]Prediction of The Autism Spectrum Disorder Diagnosis With linear Discriminant Analysis Classifier And K-Nearest Neighbor in Children.osman altay,mustafa ulas(2018 IEEE)

[9]Discriminant Analysis And Binary Logistic Regression enable More Accurate Prediction Of Autism Spectrum Disorder Than Principal Component Analysis .wail m.Hassan,Abeer al-dbass,liala al-ayadhi,ramesa shafi bahat,afaf EL-ansary(SCIENTIFIC REPORTS)

[10]Evaluation Of Machine Learning Algorithms For Classification Of Autism Spectrum Disorder (ASD).azian azamimi abdullah(Journal of physics 2019)

[11]Global prevalence of autism: A systematic review update.Jinan zeidan,Eric Fombonne,JulieScorah,alaaibrahim,maureen s.durkin,shekar saxena,afiqah yasuf,andy shih,mayada elsabbagh

[12]An Immersive Computer-Mediated Caregiver-Child Interaction System for Young Children With Autism Spectrum Disorder,Guangtao Nie , Akshith Ullal , Zhi Zheng , Amy R. Swanson , Amy S. Weitlauf, Zachary E. Warren, and Nilanjan Sarkar.2021 IEEE

[13]A Review on Predicting Autism Spectrum Disorder(ASD) meltdown using Machine Learning Algorithms.Sara Karim , Nazina Akter , Muhammed J. A. Patwary, and Md. Rashedul Islam.2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)

[14]Predicting Autism Spectrum Disorder (ASD) meltdown using Fuzzy Semi-Supervised Learning with NNRW. Sara Karim1, Nazina Akter , Muhammed J. A. Patwary.2022 IEEE.

[15]A Machine Learning Framework for Early-Stage Detection of Autism Spectrum Disorders.S. M. MAHEDY HASAN , MD PALASH UDDIN, MD AL MAMUN, MUHAMMAD IMRAN SHARIF, ANWAAR ULHAQ and GOVIND KRISHNAMOORTHY.IEEE 2022.

[16]f-MRI Based Detection of Autism using CNN Algorithm.U.B Mahadevaswamy, Chandini Manjunath.2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)

[17]Theoretical study on supervised learning models was done on javatpoint. https://www.javatpoint.com/machine-learning

[18]A Data Mining Based Approach to Predict Autism Spectrum Disorder Considering Behavioral Attributes Shaon Bhatta Shuvo, Joyoshree Ghosh,Atia Sujana Oyshi

[19]Autism Spectrum Disorder Detection in Toddlers for Early Diagnosis Using Machine Learning
Shirajul Islam,Tahmina Akter,Sarah Zakir Shareea Sabreen,Muhammad Iqbal Hossain