International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-4, April 2019

Prognostication of Autism Spectrum Disorder (ASD) using Supervised Machine Learning Models

N V Ganapathi Raju, Karanam Madhavi, G Sravan Kumar, G Vijendar Reddy, Kunaparaju Latha, K Lakshmi Sushma

Abstract*: autism spectrum disorder (ASD) screening is a psychiatric disorder which leads to neurological and developmental growth of a person which begins early in childhood and lasts throughout a person's life. This disorder is caused by differences in the brain, genetics and environmental conditions. The disorder also includes limited and repetitive patterns of behaviour. It affects how a person interacts with others, communicates, and learns. The main areas of functioning affected in people with ASD qualitative impairments in social interaction and qualitative impairments in communication. Among the affected, it was observed that more men were affected with this disorder when compared to women. The cases related to these disorders are increasing progressively. The centres of disease control and prevention (CDC) currently estimate that one in 59 children is diagnosed with ASD disorder. Unfortunately, waiting time to diagnosis these disorders are lengthy and procedures are not cost effective. To overcome the time complexity for identifying the disorder, computational intelligence can be used by making use of advanced technologies such as machine learning to improve precision, accuracy and quality of the diagnosis process. Machine learning helps us by providing intelligent techniques to discover useful hidden (or) concealed patterns, which can be utilized in prediction and to improve decision making.

***Index** autism spectrum disorder, disease control and prevention, healthcare, supervised classification*

I. INTRODUCTION

As in the recent times this autistic spectrum disorder is increasing progressively and waiting time to diagnosis these types of disorders are lengthy and procedures are not cost effective. The reported prevalence rates of autism and its related disorders have been increasing worldwide over the past decades, from approximately 4 per 10,000 to 6 per 1,000 children. The reasons for this increase include wider public awareness of these disorders, broadening of the diagnostic concepts, reclassifications of disorders and improved detection. The possibility that the increase in the reported cases is a result of unidentified risk factor(s) cannot be ruled out, and therefore more research is needed to address this. Family studies have demonstrated that autism is both familial and heritable. The recurrence rate in siblings of an autistic child is 2% to 8%, which is higher than that of the general population. In order to reduce the time complexity for identifying the disorder and diagnosing it, we propose to use any of the machine learning algorithms to improve precision, accuracy and quality of the diagnosis process and which may help to reduce the time complexity. In most recent times, the utility of machine picking up information of to move disciplinary points have been exceptionally dynamic and fruitful, particularly in areas of science and neurology. A learned records representation can help visualize actualities to assist people in logical choice making and anticipate a target variable from set of enter highlights.In this classification algorithms which are available in machine learning have been used and which will help us to predict whether any person is having this type of disorder or not and improves the accuracy of prediction.

Manuscript published on 30 April 2019.
***** Correspondence Author (s)
N.V. Ganapathi Raju, Dept. of I.T., GRIET, Hyderabad, India
Karanam Madhavi, Dept. Of CSE, GRIET, Hyderabad, India.
G Sravan Kumar, Dept. of I.T., GRIET, Hyderabad, India
G Vijendar Reddy, Dept. of I.T., GRIET, Hyderabad, India
Kunaparaju Latha, Dept. of BS, GRIET, Hyderabad, India
K L Sushma, Dept. of I.T., GRIET, Hyderabad, India

II. LITERATURE SURVEY

Analysts have more as of late centered in moving forward the information analytics strategies, especially inside the substance of the complex information. As an effective computational instrument machine learning has appeared the potential for classification and acknowledgment errands. Stahl D, Pickles A, Johnson MH, Elsabbagh M [1] analysed event related potential data for diagnosis of autism. In his work he used many computational methods like regularized discriminant functions analysis (DFA), linear discriminant analysis (LDA) and Support vector machine (SVM) and obtained accuracy of 61 %, 56% and 64% respectively. In this he has highlighted need for eliminating irrelevant variables and find important features to improve the accuracy.

Related work by Bone D, Cleric S, Goodwin Ms, Ruler C [2] connected a machine learning classifier to a huge bunch of verbal people with ASD and non –ASD disarranges. A novel numerous - level SVM show was proposed and classification comes about were detailed of having an exactness of 75%-80%.

Similarly, Zhang F [3] sought to identify children with ASD from typically developed children. Utilizing the information of 70 children analyzed with ASD and with the assistance of SVM and 10-fold cross approval and by extricating highlights from each cluster they were able to classify the ASD children and their demonstrate came up with an precision of 78.33%.

Brian McNamara, Camila Lora, Donyoung Yang, Fablana Flores [4] applied supervised machine learning algorithms Random Forest and Decision Tree to a dataset derived from ASD. Initially they cleansed the dataset by removing the observations containing missing values and dropping unnecessary variables, and later they used repeated K-Fold cross validation t aid in a selecting model as well as partition of testing and training data and the accuracy they scored is 61 % for Decision Tree and 79 % for Random Forest.

Related work by Vaishali R and Sasikala R [5] experimented with swarm intelligence based binary firefly feature selection wrapper to achieve a better classification with minimum feature selection and by using this swarm intelligence based single objective binary firefly they have selected 10 features among 21 features of ASD dataset to distinguish between ASD and non- ASD persons. Initially they thought of using exhaustive search based feature for selection of features but due to increase in time complexity they shifted to use stochastic search algorithms and finally proposed use of this swarm intelligence based feature selection wrappers as better alternatives in feature selection and they achieved accuracy in the range of 92.18% - 94.26%.

Recently Fadi Thabtah [6] made her studies on Autistic Spectrum Disorder and builds models using machine learning algorithms and used few screening tools like ADOS (Autism Diagnostic Observation Schedule) and ADI (Autism Diagnostic Interview) which is used to evaluate an individual's behaviour to assess Autism traits. Majority of her studies are based on DSM-IV rather than new DSM 5 and results have showed that cases with ASD are least likely to be classified with DSM5. The author has taken a dataset with 612 cases and applied decision tree-based algorithms and results revealed that the best classifier in accuracy contains only eight features among the 28 features present in the dataset.

To date, there have been few ponders that have connected machine learning to ASD dataset. Although a few machine learning strategies have been connected in ASD dataset, the reasonableness of machine learning and guideline of choosing distinctive calculations got to be examined with regard to conduct inspected and amount of the information display in dataset. In the current work, we tested the accuracy scores of all the classification algorithms available in the machine learning and found that the boosting algorithms like XGBoost and Gradient boosting algorithms have performed well when compared to all the other algorithms with an accuracy of 97.17%.

III. METHODOLOGY

The present paper consists of the following steps for the classification of Autism Spectrum Disorder using supervised classification algorithms such as XGBoost, Logistic Regression and Stochastic Gradient Descent.

3.1 Data Collection: The dataset that we have used in this paper has been taken from Kaggle dataset 2018 Autistic Spectrum Disorder dataset. The dataset that we are using is a predictive type of data and it consists of 17 attributes including the class variable. Total number of instances in the dataset is 1345 and attributes type can be nominal data or binary data. There is only one variable present with numerical data with the remaining variables being categorical and binary in nature. The dataset describes ASD screening results.

***Nominal Data*:** A nominal data is a type of data that is used to label variables without providing quantitative value. These kinds of data cannot be measured or evaluated, in order to use these kinds of data we have to convert these types of data into numerical data. An example of nominal data is "Male" or "Female" in gender of a person and "pass" or "fail" for a student.

Binary Data: Binary data is data whose unit can take on only two possible values, traditionally labelled as "0" or "1".

3.2 Confusion Matrix:

The confusion matrix is a pre-processing of the facts that describes how well a class version is predicting its magnificence variables. A confusion matrix is a tabular content material material fabric which indicates total wide form of correct and wrong predictions made through way of the elegance algorithm in contrast to the actual consequences in the dataset. the general overall performance is regularly evaluated the usage of the dataset within the matrix, permitting the visualization of the algorithms. The matrix is N thru N where N is form of commands with anticipated training and actual instructions. As predictive models don't make assumptions, so it is vital that the overall performance is measured.

		Predicted	
		Positive	Negative
Observed	Positive	TP (# of TPs)	FN (# of FNs)
	Negative	FP (# of FPs)	TN (# of TNs)

Fig 1: Confusion Matrix

3.3 Heat Map: A heat map is a two-dimensional representation of the data in the form of map or diagram which values are represented by colours. It provides an immediate visual summary of the information present in the given dataset. Correlation states how the features are related to each other or to the target variable. Correlation can be positive or negative. Heat map makes easy to identify which features are most related to the target variable and helps us to increase the accuracy for predicting the output.

3.4 Model Selection: The present paper utilizes classification algorithms available in machine learning and boosting algorithms like XGBoost and Gradient Boosting classifiers which has shown more accuracy than remaining classification algorithms.

XGBoost classifier is one of the foremost well-known and effective execution of Angle Boosted Trees algorithm, it could be a administered learning demonstrate which is based on work guess by optimizing particular misfortune capacities as well as applying regularization procedures. XGBoost classifier is well known to supply superior arrangements when compared to other machine learning calculations. In reality, since of its inception, it has gotten to be the "state-of-the-art" machine learning calculation to bargain with organized information. Boosting method, which works on the rule of an outfit. It combines a set of powerless learners and gives progressed expectation exactness. At any moment t, the show results are weighed based on the results of past moment t-1.

3.5 Model Evaluation and Selection

Following are the measures used to evaluate the performance of machine learning models in comparison to human judgements are precision, recall and F-measure.

Precision and Recall are useful measures of success of prediction when machine learning algorithms are used. Precision refers to the percentage of your results which are relevant, and recall refers to the percentage of total relevant results correctly classified by your algorithm.

Precision: Precision is defined as the number of true positives (Tp) over the number of true positives plus the number of false positives (Fp).

$$P = \frac{T_p}{T_p + F_p}$$

Recall: Recall is defined as the number of true positives (Tp) over the number of true positives plus the number of false negatives (Fn).

$$R = \frac{T_p}{T_p + F_n}$$

F-measure: F-measure is defined as the harmonic average of precision and recall

$$\text{F-measure} = (2 \times \text{precision} \times \text{recall}) / (\text{precision} + \text{recall})$$

IV. RESULTS AND OBSERVATIONS

The present paper is implemented with Python 3.6 with Anaconda 3.x distribution using Scikit-learn machine learning package for the implementation of various unsupervised machine learning classifiers. The Fig 2 shows the sample corpus of autism spectrum disorder data set.

Case_No	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Age	Sex	Ethnicity	Jaundice	with ASD	Who completed the test	Class
1	1	1	0	0	1	1	0	1	0	0	6 m		Others	no	no	family member	No
2	1	1	0	0	1	1	0	1	0	0	6 m		'Middle E	no	no	family member	No
3	1	1	0	0	0	1	1	1	0	0	6 m		?	no	no	?	No
4	0	1	0	0	1	1	0	0	0	1	5 f		?	yes	no	?	No
5	1	1	1	1	1	1	1	1	1	1	5 m		Others	yes	no	family member	Yes
6	0	0	1	0	1	1	0	1	0	1	4 m		?	no	yes	?	No
7	1	0	1	1	1	1	0	1	0	1	5 m		White-Eu	no	no	family member	Yes
8	1	1	1	1	1	1	1	1	0	0	5 f		'Middle E	no	no	family member	Yes
9	1	1	1	1	1	1	1	0	0	0	11 f		'Middle E	no	no	family member	Yes
10	0	0	1	1	1	0	1	1	0	0	11 f		?	no	yes	?	No
11	1	0	0	0	1	1	1	1	1	1	10 m		White-Eu	yes	no	Self	Yes
12	0	1	0	0	1	0	0	0	0	1	5 f		?	no	no	?	No
13	0	1	1	1	1	1	1	1	1	1	4 m		White-Eu	yes	no	family member	Yes
14	1	0	0	0	0	0	1	0	0	0	4 f		Black	no	no	family member	No
15	1	1	1	1	1	1	1	1	1	1	6 m		White-Eu	no	no	family member	Yes
16	1	1	1	1	1	1	1	1	1	1	8 m		White-Eu	no	no	family member	Yes
17	1	1	1	1	1	1	0	1	1	1	4 m		'South As	no	no	family member	Yes
18	0	0	0	0	0	0	1	0	0	0	7 m		Others	no	no	family member	No
19	1	0	1	1	1	0	1	1	1	1	11 m		White-Eu	no	yes	family member	Yes
20	1	1	1	1	1	1	0	1	0	1	5 m		?	no	no	?	Yes
21	1	1	1	1	1	1	1	0	1	0	5 m		White-Eu	yes	no	family member	Yes
22	0	0	1	1	0	1	0	1	1	0	9 f		?	no	no	?	No
23	1	1	0	1	0	0	0	0	0	0	4 m		Asian	no	no	family member	No
24	1	0	1	1	0	1	0	0	1	0	6 f		'South As	no	no	family member	No

Fig 2: Snapshot of the Autism Spectrum Disorder dataset

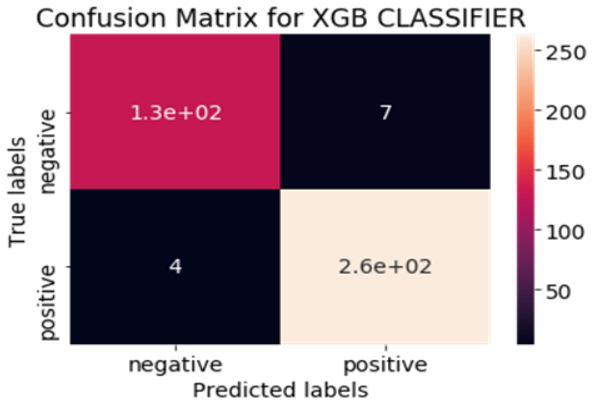

Fig 3: Confusion matrix for XGBoost Classifier

By plotting the confusion matrix, it is observed that model correctly predicted target variable Yes (True Positive) 262 times and target variable No (True Negative) 132 times and incorrectly predicted target variable Yes (False Positive) 4 times and target variable No (False Negative) 7 times. Accuracy is: 97.28%.

```
XGB CLASSIFIER Accuracy: 97.28%
XGB CLASSIFIER AUC: 96.70%
XGB CLASSIFIER Classification report:

              precision    recall  f1-score   support

           0       0.97      0.95      0.96       137
           1       0.97      0.99      0.98       267

 avg / total       0.97      0.97      0.97       404
```

Fig 4: Accuracy, Precision, Recall and F-Score for XGBoost Classifier

```
Logistic Regression Accuracy: 94.55%
Logistic Regression AUC: 93.75%
Logistic Regression Classification report:

              precision    recall  f1-score   support

           0       0.93      0.91      0.92       137
           1       0.96      0.96      0.96       267

 avg / total       0.95      0.95      0.95       404
```

Fig 5: Accuracy, Precision, Recall and F-Score for Logistic Regression

```
STOCHASTIC GRADIENT DESCENT Accuracy: 72.52%
STOCHASTIC GRADIENT DESCENT AUC: 78.15%
STOCHASTIC GRADIENT DESCENT Classification report:

              precision    recall  f1-score   support

           0       0.56      0.96      0.70       137
           1       0.96      0.61      0.74       267

 avg / total       0.83      0.73      0.73       404
```

Fig 6: Accuracy, precision, Recall and F-scores for Stochastic Gradient Descent

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

K Nearest Neighbors Accuracy: 87.13%
KNN AUC: 85.11%
KNN Classification report:

	precision	recall	f1-score	support
0	0.82	0.79	0.81	137
1	0.89	0.91	0.90	267
avg / total	0.87	0.87	0.87	404

Fig 7: Accuracy, Precision, Recall and F-Scores for K Nearest Neighbours

Random Forest Accuracy: 93.32%
Random Forest AUC: 93.34%
Random Forest Classification report:

	precision	recall	f1-score	support
0	0.88	0.93	0.90	137
1	0.97	0.93	0.95	267
avg / total	0.94	0.93	0.93	404

Fig 8: Accuracy, Precision, Recall and F-Score for Random Forest

SVM Accuracy: 96.29%
SVM AUC: 96.48%
SVM Classification report:

	precision	recall	f1-score	support
0	0.92	0.97	0.95	137
1	0.98	0.96	0.97	267
avg / total	0.96	0.96	0.96	404

Fig 9: Accuracy, Precision, Recall and F-Score for Support Vector Machine

NAIVE BAYES Accuracy: 93.32%
NAIVE BAYES AUC: 93.34%
NAIVE BAYES Classification report:

	precision	recall	f1-score	support
0	0.88	0.93	0.90	137
1	0.97	0.93	0.95	267
avg / total	0.94	0.93	0.93	404

Fig 10: Accuracy, Precision, Recall and F-Score for Naive Bayes Classification

DECISION TREE CLASSIFIER Accuracy: 85.89%
DECISION TREE CLASSIFIER AUC: 84.35%
DECISION TREE CLASSIFIER Classification report:

	precision	recall	f1-score	support
0	0.79	0.80	0.79	137
1	0.89	0.89	0.89	267
avg / total	0.86	0.86	0.86	404

Fig 11: Accuracy, Precision, Recall and F-Score for Decision Tree Classifier

Linear Discriminant Analysis Accuracy: 94.31%
Linear Discriminant Analysis AUC: 94.27%
Linear Discriminant Analysis Classification report:

	precision	recall	f1-score	support
0	0.90	0.94	0.92	137
1	0.97	0.94	0.96	267
avg / total	0.94	0.94	0.94	404

Fig 12: Accuracy, Precision, Recall and F-Score for Linear Discriminant Analysis

GRADIENT BOOSTING CLASSIFIER Accuracy: 97.03%
GRADIENT BOOSTING CLASSIFIE AUC: 96.51%
GRADIENT BOOSTING CLASSIFIER Classification report:

	precision	recall	f1-score	support
0	0.96	0.95	0.96	137
1	0.97	0.98	0.98	267
avg / total	0.97	0.97	0.97	404

Fig 13: Accuracy, Precision, Recall and F-Score for Gradient Boosting Classifier

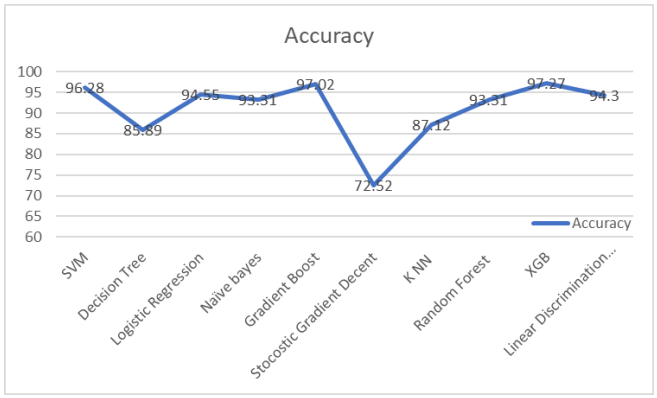

Fig 14: Accuracy scores for various machine learning classification algorithms

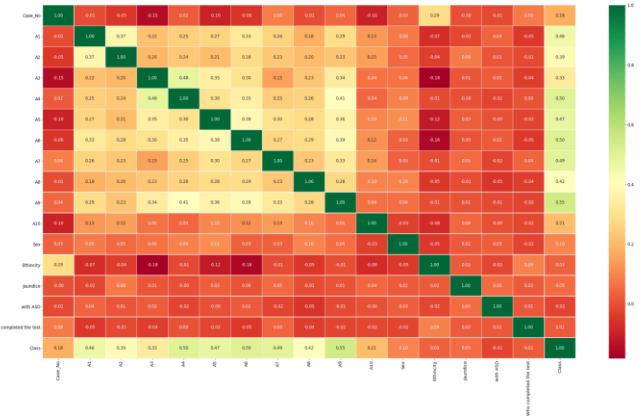

Fig 15: Heat map of the Autism Spectrum Disorder dataset

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

V. CONCLUSION

Computational methodologies, including machine learning are powerful tools for understanding data available in the dataset and to interpret the dataset, but there will be few miss-predictions as these are not completely accurate. Our goal with this study was to apply supervised machine learning algorithms to a dataset derive from Autism Spectrum Disorder. We cleansed the dataset by correcting the outlier values i.e the missing values with the mode of the column as the data present in the dataset is having categorical or binary data and we can't able to use mean or median of the data. After converting this categorical or nominal data into numerical data as well as dividing the dataset into training data and testing data, we were able to build few machine learning models and confirm that these models can predict the target variable of Class ASD with a good level of accuracy. From this study we have observed that by applying the machine learning models and finding the accuracy of each model, XGBoost Classifier and Gradient Boosting Classifiers are performing more accurate the other models with an accuracy of 97.1%. However, the fact is that all the other models also resulted in considerable amounts in predictions but we must accept that these are not reliable predictors of class variable. We conclude that it may be possible to improve the accuracy of predicting the target variable by collecting a lot more data from which a more balanced dataset with equal representation of both class ASD =Yes/No, and by retraining these algorithms on such a dataset would give us a clearer picture of the prediction possibilities of this data.

REFERENCES

1. Stahl D, Pickles A, Elsabbagh M, Johnsn MH, Team B, et al "Novel machine learning methods for ERP analysis: a validation from research on infants at risk fir autism" published on 2012.
2. Bone D, Bishop S, Black MP, Goodwin MS, Lord C, Narayanan SS "Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency and multi-instrument fusion" published in 2015.
3. Zhang F et al "An application to Autism using machine learning" published on 2018.
4. Brian McNamara, Cammila Lora, Donyoung Yang, Fabiana Flores, Paul Daly "Machine
5. Vaishalli R, Sasiala R "A machine learning based approach to classify Autism with optium behavior sets" published as research paper in 2016.
6. Fadi Thabtah in Nelson Marlborough Institute of Technology" Autism Spectrum Disorder Screening: Machine Learning Adaption and DSM-5 Fufillment", published on 2017.
7. Learning Classification of Adults with Autism Spectrum Disorder" published on April 29th, 2018. B van den Bekerom in University of Twente "Using Machine Learning for Detection of Autism Spectrum Disorder" published on 2017.
8. Kayleigh K. Hyde, Marlena N. Novack, Nicholas LaHaye "Applications of supervised machine learning in Autism Spectrum Disorder Research: a review" published on 2019.
9. N V Ganapathi Raju,"ML based HCC Survival Prediction System", IJITEE, Vol.8, Iss.6,2019
10. Sukanya L, "Sentiment Analysis using Legion Kernel Convolutional Neural Network with LSTM", IJITEE, Vol 8, Iss 4, 2019.

AUTHORS PROFILE

Dr. N V Ganapathi Raju working as Professor, in the department of IT, GRIET. Completed Ph.D. in CSE from Jawaharlal Nehru Technological University Kakinada and did Master of Technology in Computer Science and Technology from Andhra University, having 18 years of teaching experience which includes seven years of research experience in the area of Computer Science and Engineering. Research interests include Text Mining, Information Retrieval, NLP, Machine Learning and Data Science. The result research work was published in various national and international journals including Scopus indexed, Free journals and Springer journals.

Dr. K. Madhavi, working as a Professor in Computer Science and Engineering Department, Gokaraju Rangaraju Instittute of Engineering and Technology. She has completed her B.E in 1997, M.Tech from JNTUA in 2003 and awarded Ph.D from JNTUA in 2013. She has 20 years of teaching experience. She has published several papers in reputed international journals and international conferences. Her research interest includes software engineering, Model Driven Engineering, Data Mining, ICT and Education Technology.

G Sravan Kumar, studying currently pursuing his final semester of Bachelor of Technology in Information Technology from Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India.

Mr. G Vijendar Reddy, Associate Professor of Information Technology, pursing Ph.D in Computer Science Engineering from JNTU, Kakinada and M.Tech. in Software Engineering from JNTU, Anantapur. Having over 13 years of academic and research experience.

Ms. Kunaparaju Latha, Assistant Professor of Basic Sciences, completed her M.B.A from Andhra University and has eight years of academic experience. Her area of interest is Human Resources

Ms. Lakshmi Sushma Kolli, Assistant Professor of Information Technology, completed her M.Tech from KL University and has six years of academic experience. She had earned Bachelor of Technology of Engineering in Computer Science Engineering, and Master of Technology of Engineering in Computer Science and Engineering. Ms. Sushma's research interests include data mining, opinion mining and cloud computing in which she has two publications, in various journals.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)