Autism Spectrum Disorder Prediction in Toddlers using Extreme Gradient Boosting

Soorampalli Apoorva[1], Sanjana Gadalay[2], Somu Venkata Sai Susmitha[3], Siddhartha Kolisetti[4]

[1-4]UG student, Dept. of CSE, Mahatma Gandhi Institute of Technology, Telangana, India.

Abstract – *A developmental disability affects a person's life in many ways. People with a developmental disability have trouble speaking, reading, and communicating properly. The number of people being diagnosed with such disabilities has been increasing at a faster pace worldwide. One such disability is Autism Spectrum Disorder (ASD), which is a developmental and behavioral disorder that begins early in childhood and can last through one's life. Raising awareness about such disabilities is very important so that they can be detected at early stages and therapy can be given. The main objective is to detect whether a child has the probability of being diagnosed with ASD using a Machine Learning technique known as Extreme Gradient Boosting as it is fast and efficient.*

Keywords: Autism Spectrum Disorder (ASD), Extreme Gradient Boosting, Extra trees classifier, Accuracy.

1. INTRODUCTION

A mental disorder can be an illness with significant psychological or behavioral issues. Nowadays, with the eradication and treatment of physical illnesses, mental disorders are receiving more attention compared to the past. The centres for disease control recently presented that the prevalence of ASD has risen to approximately 1 in 68 [1]. The families who have a disabled child bear the brunt both socially and economically. There are higher levels of stress and financial burden for treatment and therapies.

Early identification and treatment are vital in reducing symptoms up to a certain extent. There are no actual medical tests available. The diagnosis is performed based on observing the person and noting how he/she talk or act compared to other people. Hence, the identification of disorders like Autism in earlier stages gives a ray of hope that the disabled child can improve to a certain extent. Delay or perversion in language development is an important feature of ASD. About 50% of people with autism can never make a useful speech[2].

Autism Spectrum Disorder is a developmental-behavioral disability in which a person finds difficulty in reading, learning, and communicating with others. It limits the use of communicative, social and cognitive skills as well as abilities of the affected personality whereas its symptoms may vary from person to person.[3] .The problems an autistic person might face are:

- Poor eye contact with others.
- They give abnormal facial expressions.
- They are very sensitive to smell, sound, and touch.
- They get angry when their routine is disturbed.
- They have behavioral disturbances.
- They struggle to understand other people's feelings.

A machine learning model is built to detect Autism using a machine learning algorithm called Extreme Gradient Boosting. It is less expensive and can be used as a pre-screening test by the family members to check whether there is a possibility if a person has Autism or not.

2. EXISTING SYSTEM

Since Autism is a behavioral disability, it is harder to identify it with the help of a blood test or medical tests. It is becoming more challenging to identify, especially in adults. If not identified in the early stage, it may adversely affect the person's everyday life. The existing systems consist of some screening tests, which are very expensive and time-consuming. Autistic people cannot be cooperative for a long time while the screening test is going on.

The main drawbacks of the existing systems include:

- The dataset had a fewer number of instances.
- The focus was more on the comparison.
- Many of them did not focus on the crucial attributes for ASD detection.

3. PROPOSED SYSTEM

A machine learning model is built using extreme gradient boosting. Dr. Fadi Fayez Thabtah developed a dataset based on a mobile app -ASD Tests. The dataset consists of 1054 instances and 19 attributes. This dataset represents the data collected from children of ages between 0-3 years. It is available in UCI Machine. These attributes describe the behavior of an autistic child.

When the ML algorithm is embedded in self- assessment tools, it will provide users with valuable concealed knowledge and guide the process of correct classification selection decision in more efficient manner[4].

The data set is segregated into train and test data. Train data is used to train the model, whereas test data is used to validate the model. The chief goal of the paper is to detect autistic behavior in children with accuracy. Hence the algorithm used is Extreme Gradient Boosting which is fast and efficient. It increases the model's accuracy to detect the probability of a person having Autism is more effective than before.

Advantages of the proposed system:

- It is easy to implement
- All the required software is readily available
- The Accuracy and precision are high.
- The features which are crucial in determining Autism are identified.

4. METHODOLOGY

With the advancement of artificial intelligence and machine learning (ML), autism can be predicted at quite early stage[5].We find out whether a toddler is diagnosed with Autism based upon the data used. An analysis is performed on the dataset, which shows relationships between the attributes and the target class-ASD Class/Traits. Label Encoding is used to convert the categorical values into numerical values. To find the essential features that help determine the target class(ASD), we use the ExtraTreesClassifier available in sklearn.ensemble. Then, we perform training on the machine learning model to diagnose Autism for the test data. To accomplish this task, the following process takes place:

- Analyzing the data
- Preprocessing the model
- Model Training
- Model Testing
- Finding the effectiveness of the model

4.1 Data Analysis

Data analysis is one of the critical steps of any machine learning problem. By doing it, we try to find out any insights related to the problem. We represent the relationships between the attributes in the form of bar charts and graphs. The below figure represents the data analysis of the dataset.

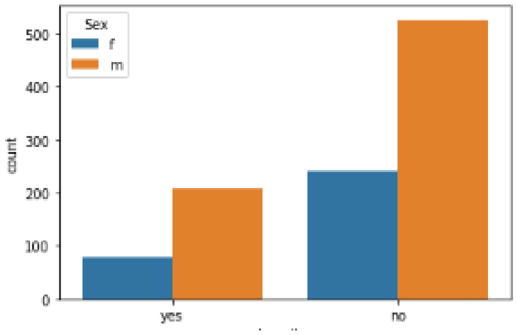


Fig -1: Relationship between gender and autism

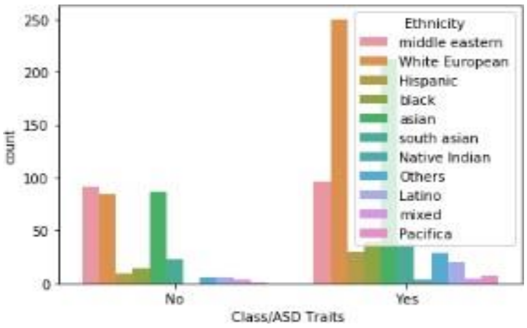


Fig -2: Relationship between ethnicity and autism

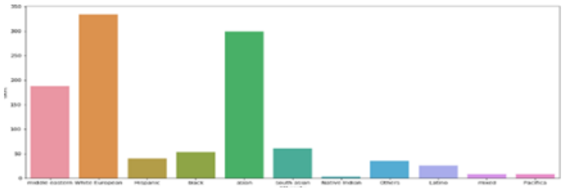


Fig -3: No of people who took the test region wise

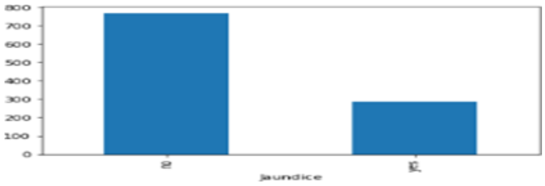


Fig -4: Relationship between jaundice and autism

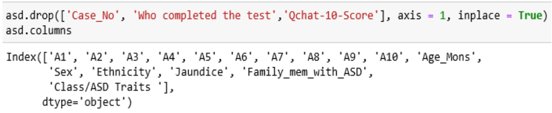
Following are the insights observed in the data:

- Majority of the toddlers who took the test are Europeans, followed by Asians
- Toddlers born without jaundice have a high chance of being ASD positive, and ASD is more common among males than females irrespective of being born with jaundice.
- Majority of the Toddlers diagnosed with Autism are white Europeans.
- Males have a higher chance of getting diagnosed with Autism than females.

4.2 Data Preprocessing

Pre-processing denotes to the transformations applied to the data before using it for training the model. It is important to remove the unnecessary attributes so that they do not affect the machine learning model's accuracy. Various Data pre-processing methods are used to handle incomplete and inconsistent data like as handling missing values, outlier detection, data discretization, data reduction (dimension and numerosity reduction)[6].To remove unnecessary attributes, we use the Series.drop() function available in pandas library.

Syntax:Series.drop(labels=None,axis=0,index=None,columns=None,level=None, inplace=False, errors=raise)



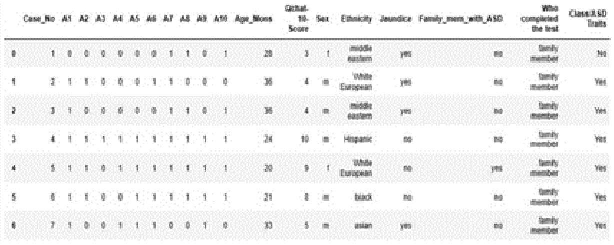
```
asd.drop(['Case_No', 'Who completed the test','Qchat-10-Score'], axis = 1, inplace = True)
asd.columns
```

```
Index(['A1', 'A2', 'A3', 'A4', 'A5', 'A6', 'A7', 'A8', 'A9', 'A10', 'Age_Mons',
       'Sex', 'Ethnicity', 'Jaundice', 'Family_mem_with_ASD',
       'Class/ASD Traits '],
      dtype='object')
```

Fig -5: Removing unnecessary data

Machines work well with numerical attributes as compared to categorical attributes. Hence categorical attributes need to be converted to numerical attributes for better efficiency. One such method we can use is Label Encoding.

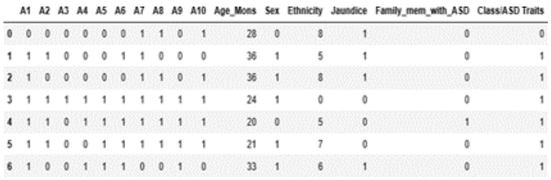
Label encoding refers to changing the labels into numeric types, therefore converting them into the machine-readable type. Machine learning algorithms will then decide in an enhanced way on how those labels should be operated. It is an essential preprocessing step for the structured dataset in supervised learning. We use the LabelEncoder, which is available in sklearn. Preprocessing package to perform label encoding and to transform we use the fit_transform() function. There are many factors affecting the outcome of autism, overfitting occurs particularly when using small data sets[7].



	Case_No	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Age_Mons	Qchat-10-Score	Sex	Ethnicity	Jaundice	Family_mem_with_ASD	Who completed the test	Class/ASD Traits
0	1	0	0	0	0	0	0	1	1	0	1	28	3	f	middle eastern	yes	no	family member	No
1	2	1	1	0	0	0	1	1	0	0	0	36	4	m	White European	yes	no	family member	Yes
2	3	1	0	0	0	0	0	1	1	0	1	36	4	m	middle eastern	yes	no	family member	Yes
3	4	1	1	1	1	1	1	1	1	1	1	24	10	m	Hispanic	no	no	family member	Yes
4	5	1	1	0	1	1	1	1	1	1	1	20	9	f	White European	no	yes	family member	Yes
5	6	1	1	0	0	1	1	1	1	1	1	21	8	m	black	no	no	family member	Yes
6	7	1	0	0	1	1	1	0	0	1	0	33	5	m	asian	yes	no	family member	Yes

Fig -6: Data before label encoding

We use the LabelEncoder, which is available in sklearn. preprocessing package to perform label encoding and to transform we use the fit_transform() function.



	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Age_Mons	Sex	Ethnicity	Jaundice	Family_mem_with_ASD	Class/ASD Traits
0	0	0	0	0	0	0	1	1	0	1	28	0	8	1	0	0
1	1	1	0	0	0	1	1	0	0	0	36	1	5	1	0	1
2	1	0	0	0	0	0	1	1	0	1	36	1	8	1	0	1
3	1	1	1	1	1	1	1	1	1	1	24	1	0	0	0	1
4	1	1	0	1	1	1	1	1	1	1	20	0	5	0	1	1
5	1	1	0	0	1	1	1	1	1	1	21	1	7	0	0	1
6	1	0	0	1	1	1	0	0	1	0	33	1	6	1	0	1

Fig -7: Data after Label Encoding

4.3 Feature Importance

It is essential to know what attributes in the dataset are crucial for determining the target variable, known as Feature Importance. For feature importance, we use the ExtraTreesClassifier available in sklearn.ensemble. Extremely Randomized Trees Classifier or Extra trees classifier is an ensemble learning technique that combines many decorrelated decision tree results in a forest and gives the final classification result. The importance of each feature is calculated using a measure called as Gini index. A score is assigned to each feature based on its importance. A graph is plotted by describing the top features which determine the target class label.

The feature_importance attribute of tree-based classifiers is used to assign each attribute score, showing each attribute which is vital to determine the target variable. Figure 3.10 shows that A9 is an essential attribute in finding out whether the child has autistic traits or not, followed by A6 and A7. Moreover, some recently developed machine learning algorithms, such as XGBoost, have shown better performance over traditional algorithms on many tasks beyond biomedical domain[8].

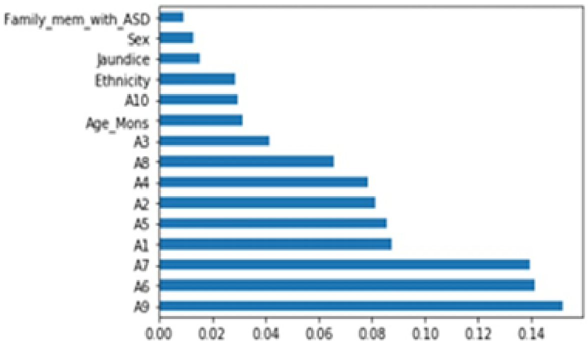


Fig -8: Feature Importance

4.4 Modeling

We are constructing and training a machine learning model using XGBoosting.

Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting is a useful technique to implement Gradient Boosted decision trees. C++ programming language can help implement the XGBoost library. The Extreme Gradient Boosting library is a form of software library delineated to advance model performance and speed. Recently, the XGBoost method has been showing advanced progress and is leading in applied machine learning. In this algorithm, sequential decision trees are created. Here, weights demonstrate a vital role in XGBoost. First, weights are allocated to all the independent variables. Then, those weights are passed into a decision tree resulting in a prediction of the required results. All the weights of the variables predicted incorrectly by the tree are then incremented, and these variables are fed to the second decision tree. This process goes on, and all the individual classifiers or predictors are accumulated to give a powerful and more accurate model. This technique can be implemented on classification, user-defined prediction, regression, and ranking problems. This model supports three primary forms of gradient boosting: Gradient Boosting, Regularized Gradient Boosting, Stochastic Gradient Boosting.

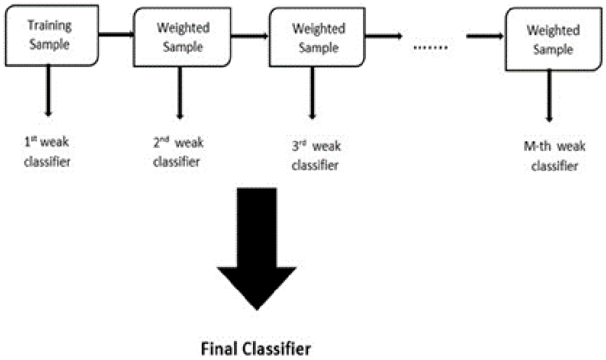
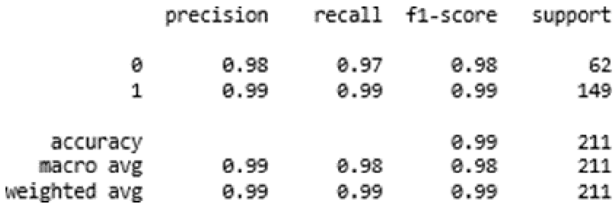


Fig -9: Working of Extreme Gradient Boosting

5. RESULTS

The accuracy of the XGBoost classifier is 95%.The classification report shows the precision,recall and f1-score for ASD negative(0) is 98,97 and 98 whereas precision,recall and f1-score for ASD positive(1) is 99,99 and 99.



```
              precision    recall  f1-score   support

           0       0.98      0.97      0.98        62
           1       0.99      0.99      0.99       149

    accuracy                           0.99       211
   macro avg       0.99      0.98      0.98       211
weighted avg       0.99      0.99      0.99       211
```

Fig -10: Accuracy score and classification report of XGBoost classifier

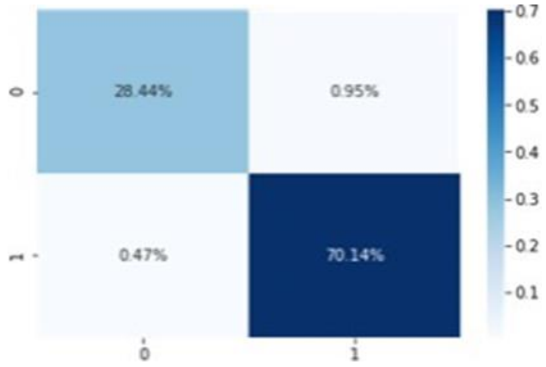


Fig 11: Confusion Matrix of XGBoost Model

The XGBoost classifier model classifies 70.14% of the samples which are ASD positive correctly and 28% of the samples that are ASD negative correctly. It classifies 0.95% of the samples which are actual ASD negative as positive, and 0.47% of the samples which are actually ASD positive as negative.

6. CONCLUSION

In the Autism Spectrum Disorder study, toddlers aged 0-3 years were considered. Using feature importance, it was found that A9 was vital in determining autistic trades in toddlers. A Machine learning model has been developed using Extreme Gradient Boosting with an accuracy of 98%. It has very high precision, recall, f1-score, and hence this model can be used as a basis for further research. We observed that an automated approach could predict—with high agreement—whether a child would meet ASD surveillance criteria[9].

7. FUTURE SCOPE

Further research can be done on autism, and a website can be developed to diagnose autism based upon the proposed model and suggest techniques that help improve the state of the autistic child. An autism community app can be developed where the family members can interact with each other. Parents are likely to seek Web- based communities to verify their suspicions of autism spectrum disorder markers in their child [10].

REFERENCES

[1] Oh, Dong Hoon; Kim, Il Bin; Kim, Seok Hyeon; Ahn, Dong Hyun (2017). Predicting Autism Spectrum Disorder Using Blood-based Gene Expression Signatures and Machine Learning. Clinical Psychopharmacology and Neuroscience, 15(1), 47–52. doi:10.9758/cpn.2017.15.1.47 .

[2] Altay, O., & Ulas, M. (2018). Prediction of the autism spectrum disorder diagnosis with linear discriminant analysis classifier and K-nearest neighbor in children. 2018 6th International Symposium on Digital Forensic and Security (ISDFS). doi:10.1109/isdfs.2018.8355354.

[3] Tyagi, Bhawana; Mishra, Rahul; Bajpai, Neha (2018). [IEEE 2018 IEEE Punecon - Pune, India (2018.11.30-2018.12.2)] 2018 IEEE Punecon - Machine Learning Techniques to Predict Autism Spectrum Disorder. , (), 1–5. doi:10.1109/PUNECON.2018.8745405

[4] Thabtah F, Peebles D, "A new machine learning model based on induction of rules for autism detection," Health informatics journal. doi: 10.1177/1460458218824711.

[5] Kazi Shahrukh Oma, Prodipta Mondal, Nabila Shahnaz Khan, Md. Rezaul Karim Rizvi, Md Nazrul Islam(2019).A Machine Learning Approach to Predict Autism Spectrum Disorder. 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE). doi: 10.1109/ECACE.2019.8679454

[6] Raj, S., & Masood, S. (2020). Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques. Procedia Computer Science,167,9941004. doi:10.1016/j.procs.2020.03.399.

[7] Usta, Mirac Baris; Karabekiroglu, Koray; Sahin, Berkan; Aydin, Muazzez; Bozkurt, Abdullah; Karaosman, Tolga; Aral, Armagan; Cobanoglu, Cansu; Kurt, Aysegül Duman; Kesim, Neriman; Sahin, İrem; Ürer, Emre (2018). Use of machine learning methods in prediction of short-term outcome in autism spectrum disorders. Psychiatry and Clinical Psychopharmacology,(),1–6. doi:10.1080/24750573.2018.1545334.

[8] Chen, Q., Qiao, Y., Xu, X., Tao, Y., & You, X. (2019). Urine Organic Acids as Potential Biomarkers for Autism-Spectrum Disorder in Chinese Children. Frontiers in Cellular Neuroscience,13. doi:10.3389/fncel.2019.00150

[9] Maenner, Matthew J.; Yeargin-Allsopp, Marshalyn; Van Naarden Braun, Kim; Christensen, Deborah L.; Schieve, Laura A.; Eapen, Valsamma (2016). Development of a Machine Learning Algorithm for the Surveillance of Autism Spectrum Disorder. PLOS ONE, 11(12), e0168224–. doi:10.1371/journal.pone.0168224 .

[10] Ayelet Ben-Sasson, Diana L Robins, Elad YoM-Tov," Risk Assessment for parents who suspect their child has autism spectrum disorder: Machine Learning approach," Microsoft Research, Israel, Herzelia. doi: 10.2196/jmir.9496.

BIOGRAPHIES

Somu Venkata Sai Susmitha
Student
Computer Science Engineering
Mahatma Gandhi Institute of Technology

Siddhartha Kolisetti
Student
Computer Science Engineering
Mahatma Gandhi Institute of Technology

Sanjana Gadalay

Student
Computer Science Engineering
Mahatma Gandhi Institute of Technology

Soorampalli Apoorva
Student
Computer Science Engineering
Mahatma Gandhi Institute of Technology