International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
ISSN: 2457-0362
www.ijarst.in

EXPERT SYSTEM APPLICATION FOR AUTISM SPECTRUM DISORDER PREDICTION USING XGBOOST MODEL

K. Venkatakrishna[1], Thumula Tejaswi[2], Vadlakonda Venkatesh[2], Choula Yashwanth[2], Gourav Kumar Singh[2]

[1]Assistant Professor, [2]UG Scholar, [1,2]Department of CSE – Data Science
[1,2]Malla Reddy College of Engineering and Management Sciences, Medchal, 501401, Hyderabad

Abstract

Autism Spectrum Disorder (ASD), also known as "autism," is a psychiatric disorder that impacts an individual's language, thinking, and social skills. Autism Spectrum Disorder (ASD) is a widespread condition, affecting around 1 in every 54 births and around 1% of the world's population. Regrettably, despite its widespread occurrence, the etiology and remedy for autism remain elusive, presenting substantial obstacles for parents who fear their kid may have ASD. Prompt identification of autism is vital for a child's progress, but it can be exceedingly challenging when symptoms emerge during the child's maturation. Diagnostic tests administered to children aged 2 to 3 years are generally less dependable compared to tests conducted on children aged 4 to 5 years. This situation is concerning as timely diagnosis is crucial for autistic individuals to achieve their developmental objectives effectively. Autism is frequently distinguished by challenges in social interaction and communication, which can complicate proper diagnosis, even when using sophisticated tools such as the ADOS and ADI. This study aims to solve the issues related to the diagnosis of autism by specifically targeting the enhancement of the diagnostic process. The process entails training and evaluating machine learning models, specifically Random Forest with Standard scaler, using a dataset on autism spectrum disorder. The objective is to determine the most influential factors in identifying autism in toddlers. The objective is to provide a quantitative methodology to assist in the early detection and subsequent management of autism, as prompt intervention can alleviate long-term symptoms associated with the condition. This study seeks to utilize machine learning techniques to offer valuable insights into the accurate diagnosis of autism and to enhance assistance for individuals with ASD and their families.

Keywords: Autism spectrum disorder, Predictive analytics, Data analysis, Supervised learning.

1. Introduction

Autism spectrum disorder (ASD) is a developmental disorder that involves persistent challenges in social interaction, speech and nonverbal communication, and restricted and repetitive behaviours. In the USA, the prevalence of ASD has increased substantially in the past two decades, with an estimate of every 1 in 44 children to be identified with ASD by age 8 in 2016 [1]. Although there exist evidence-based interventions which improve core symptoms in children with ASD, many children with ASD still experience long-term challenges with daily life, education and employment [2]. Early diagnosis is the key to early intervention for improving the long-term outcomes of children with ASD. However, despite the growing evidence shows that accurate and stable diagnoses can be made by 2 years, in real-world settings, the median age of ASD diagnosis is 50 months. To improve early diagnosis, the American Academy of Paediatrics (AAP) has recommended universal screening among all children at 18-month and 24-month well-child visits in the primary care settings using the Modified Checklist for Autism in Toddlers (M-CHAT) [3], a questionnaire that assesses children’s behaviour for toddlers. However, growing evidence has shown that using M-CHAT alone may not yield sufficient accuracy in detecting ASD cases, with a sensitivity below 40% and a positive predictive value (PPV) under 20% [4, 5]. In addition to ASD-specific behavioural questionnaires, general clinical and healthcare records may also contain meaningful signals to differentiate the ASD risks among very young children. Studies have found that children with ASD are oftentimes accompanied by certain symptoms and medical issues such as gastrointestinal problems, infections and feeding problems. This implies that past diagnosis and healthcare encounter information, commonly available from health insurance claims or Electronic Healthcare Record (EHR), could potentially be used for ASD risk prediction. In fact, medical claims and EHR data have been widely used in the health informatics literature for identifying disease-specific early phenotypes even before the hallmark symptoms start to manifest, such as for chronic diseases like heart failures, diabetes and Alzheimer's disease [6].

The main motive is to analyze and find some limitations to propose a new, better, and improved machine-learning based approach for autism spectrum disorder prediction. Automated algorithms for disease detection are being deeply studied for usage in healthcare. Graph theory and machine learning algorithms were used. For each age range being examined, the pipeline automatically selected 10 biomarkers. In discriminating between ASD and HC, measures of centrality are the most operational [8]. The study [9] used a neural network-based feature selection method from teacher-student which was suggested to have the most discriminating features and applied different classification algorithms. The results are compared with the already presented methods at the overall and site level. The authors in [10] also utilize the neural network to acquire the distributions of PCD for the classification of ASD as it has far more hyper parameters that make the model extra versatile. Payabvash et al. [11] used computer leaning algorithms to classify children with autism based on tissue connectivity metrics, hence, observed decreased connectome edge density in the longitudinal white matter tracts. It illustrated the viability of it in identifying children with ASD, connectome-based machine-learning algorithms.

The authors in [12] conclude that the data may be used to establish diagnostic biomarkers for the progression of autism spectrum disorders and to distinguish those with the condition in the general population. Wang et al. [13] proposed an ASD identification approach which focuses on multi-atlas deep feature representation and ensemble learning technique. In study [14], the multimodal automated disease classification system uses two types of activation maps to predict whether the person is healthy or has autism. It was able to achieve 74% accuracy. Rakić et al. [15] suggested a technique which is based on a system composed of autoencoders and multilayer perceptron. Because of a multimodal approach that included a set of structural and functional data classification classifiers, the highest classification precision was 85.06%.

2. Existing model

Naive Bayes algorithm

Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output. Naive Bayes classifier is a collection of many algorithms where all the algorithms share one common principle, and that is each feature being classified is not related to any other feature. The presence or absence of a feature does not affect the presence or absence of the other feature.

Naive Bayes is a powerful algorithm that is used for text data analysis and with problems with multiple classes. To understand Naive Bayes theorem's working, it is important to understand the Bayes theorem concept first as it is based on the latter. Bayes theorem, formulated by Thomas Bayes, calculates the probability of an event occurring based on the prior knowledge of conditions related to an event. It is based on the following formula:

$$P(A|B) = P(A) * P(B|A)/P(B)$$

Where we are calculating the probability of class A when predictor B is already provided.

P(B) = prior probability of B

P(A) = prior probability of class A

P(B|A) = occurrence of predictor B given class A probability

Drawbacks

The Naive Bayes algorithm has the following disadvantages:

- The prediction accuracy of this algorithm is lower than the other probability algorithms.
- It is not suitable for regression. Naive Bayes algorithm is only used for textual data classification and cannot be used to predict numeric values.

3. Proposed methodology

The machine learning model for autism prediction in toddlers is a data-driven approach aimed at identifying whether a toddler is at risk of having ASD based on certain features and characteristics. Here's an overview of the key aspects of this predictive model:

1. Data Collection and Preparation

- The first step involves collecting a dataset that includes information about toddlers. This dataset should ideally contain both toddlers who have been diagnosed with ASD and those who have not.
- Data preparation includes cleaning the dataset, handling missing values, and ensuring it is well-structured for analysis.

2. Feature Selection and Engineering

- Identifying the most relevant features (variables) is crucial. These features may include demographic information (e.g., age, gender), family history, behavioral traits, and responses to specific tests or questionnaires designed to assess ASD risk.
- Feature engineering may involve creating new features or transforming existing ones to improve the model's predictive performance.

3. Data Encoding: Categorical variables, such as gender or family history, may need to be encoded into numerical values to be used by machine learning algorithms. Label encoding or one-hot encoding can be applied for this purpose.

4. Model Selection:

- Choosing an appropriate machine learning algorithm is a critical decision. Common choices include logistic regression, decision trees, random forests, support vector machines, neural networks, and ensemble methods like XGBoost.
- The choice of the model depends on factors like the dataset's size, complexity, and the desired level of interpretability.

5. Data Splitting: The dataset is typically divided into two subsets: a training set and a testing set. The training set is used to train the model, while the testing set is used to evaluate its performance.

6. Model Training: The selected machine learning model is trained on the training data, learning the patterns and relationships between the features and the target variable (ASD diagnosis).

7. Model Evaluation: The model's performance is assessed using various evaluation metrics, including accuracy, precision, recall, F1-score, and the confusion matrix. These metrics provide insights into how well the model is making predictions.

3.1 XGBoost Classifier

XGBoost is an optimized distributed gradient boosting library designed for efficient and scalable training of machine learning models. It is an ensemble learning method that combines the predictions of multiple weak models to produce a stronger prediction. XGBoost stands for “Extreme Gradient Boosting” and it has become one of the most popular and widely used machine learning algorithms due to its ability to handle large datasets and its ability to achieve state-of-the-art performance in many machine learning tasks such as classification and regression. One of the key features of XGBoost is its efficient handling of missing values, which allows it to handle real-world data with missing values without requiring significant pre-processing. Additionally, XGBoost has built-in support for parallel processing, making it possible to train models on large datasets in a reasonable amount of time. XGBoost can be used in a variety of applications, including recommendation systems, and click-through rate prediction, among others. It is also highly customizable and allows for fine-tuning of various model parameters to optimize performance.

XGBoost is an implementation of Gradient Boosted decision trees. In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

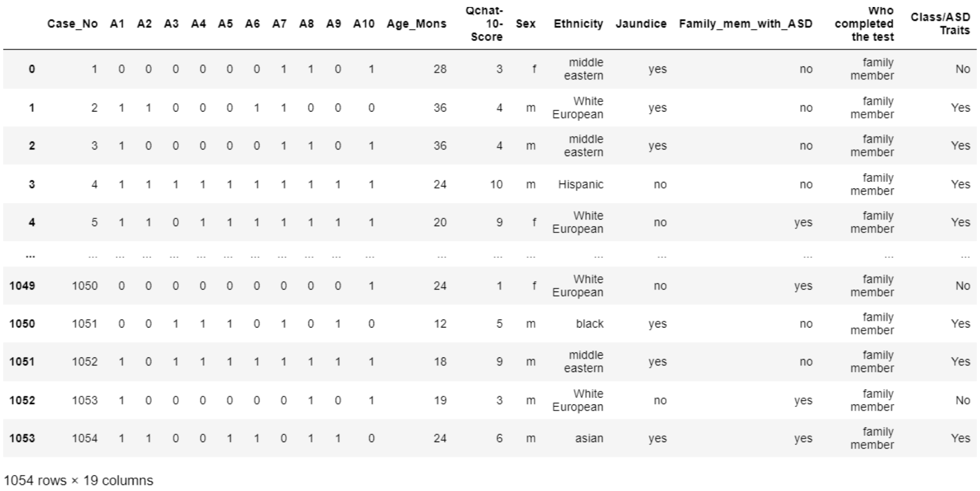
Advantages of XGBoost

- Performance: XGBoost has a strong track record of producing high-quality results in various machine learning tasks, especially in Kaggle competitions, where it has been a popular choice for winning solutions.
- Scalability: XGBoost is designed for efficient and scalable training of machine learning models, making it suitable for large datasets.
- Customizability: XGBoost has a wide range of hyperparameters that can be adjusted to optimize performance, making it highly customizable.
- Handling of Missing Values: XGBoost has built-in support for handling missing values, making it easy to work with real-world data that often has missing values.
- Interpretability: Unlike some machine learning algorithms that can be difficult to interpret, XGBoost provides feature importances, allowing for a better understanding of which variables are most important in making predictions.

4. Results and Discussion

This dataset contains information about various attributes of individuals, including demographic information (age, gender, ethnicity), medical history (jaundice), family history of ASD, and assessment scores (Q-CHAT-10) used to predict the presence of ASD traits. The "Class/ASD Traits" column appears to be the target variable used for classification purposes, indicating whether the individual exhibits ASD traits or not.

Figure 1 displays a representation of the sample dataset used for the classification of autism spectrum disorder (ASD) in toddlers. The dataset likely includes various features that are used as input for the machine learning models to predict whether a toddler has autism or not. Figure 2 presents a count plot that illustrates the distribution of classes within the dataset. It shows the number of instances labeled as "Autism" and the number labeled as "Normal." This visualization gives an idea of the class imbalance or balance in the dataset. Figure 3 showcases the dataset after the label encoding operation has been applied. Label encoding is a preprocessing step that converts categorical labels into numerical values, making the data suitable for machine learning algorithms that require numerical input. Figure 4 displays the accuracy of the Gaussian Naive Bayes (GNB) model along with a classification report. The classification report likely includes metrics such as precision, recall, and F1-score for each class (Autism and Normal). Similar to Figure 4, the Figure5 shows the accuracy and classification report, but for the XGBoost model. It presents the model's performance metrics for both classes (Autism and Normal).



	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1049	1050	0	0	0	0	0	0	0	0	0	1	24	1	f	White European	no	yes	family member	No
1050	1051	0	0	1	1	1	0	1	0	1	0	12	5	m	black	yes	no	family member	Yes
1051	1052	1	0	1	1	1	1	1	1	1	1	18	9	m	middle eastern	yes	no	family member	Yes
1052	1053	1	0	0	0	0	0	0	1	0	1	19	3	m	White European	no	yes	family member	No
1053	1054	1	1	0	0	1	1	0	1	1	0	24	6	m	asian	yes	yes	family member	Yes

1054 rows × 19 columns

Figure 2: Sample dataset used for classification of ASD.

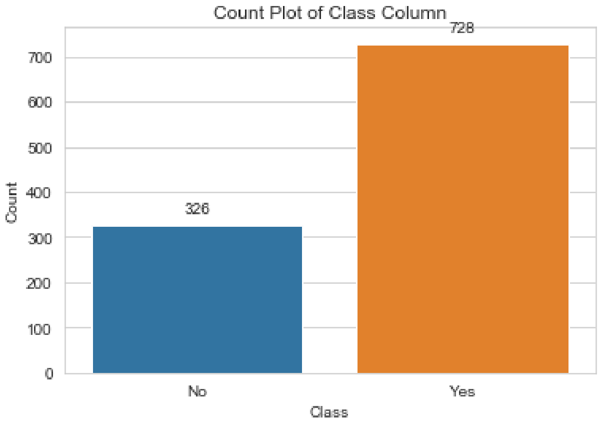
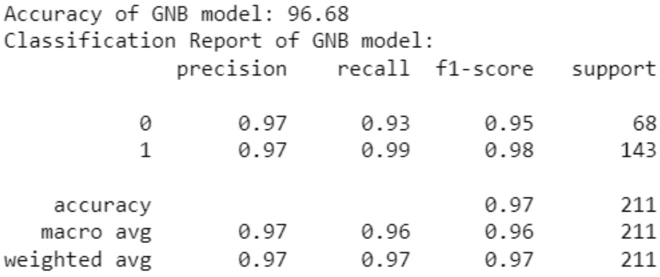


Figure 3: Count plot of class column i.e., Autism or Normal.

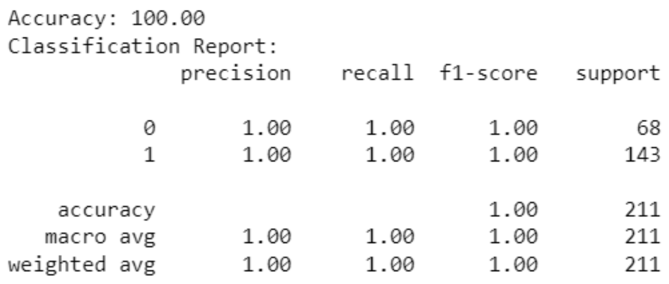


```
Accuracy of GNB model: 96.68
Classification Report of GNB model:
              precision    recall  f1-score   support

           0       0.97      0.93      0.95        68
           1       0.97      0.99      0.98       143

    accuracy                           0.97       211
   macro avg       0.97      0.96      0.96       211
weighted avg       0.97      0.97      0.97       211
```

Figure 4: Obtained accuracy and classification report using GNB model.



```
Accuracy: 100.00
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        68
           1       1.00      1.00      1.00       143

    accuracy                           1.00       211
   macro avg       1.00      1.00      1.00       211
weighted avg       1.00      1.00      1.00       211
```

Figure 5: Obtained accuracy and classification report using XGBoost model.

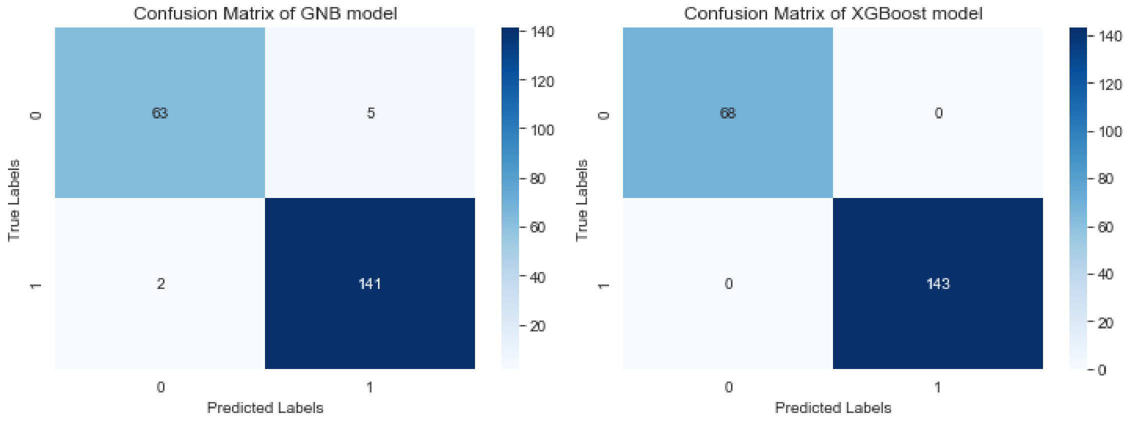


Figure 6: Confusion matrix of GNB model (left). Confusion matrix of XGBoost model (right).

5. Conclusion

In the analysis of autism prediction in toddlers, the XGBoost model emerges as the superior choice compared to the Naive Bayes model. This assessment is based on various performance metrics and evaluations. Firstly, in terms of accuracy, the XGBoost model demonstrates a notably higher accuracy score when compared to the Naive Bayes model. This suggests that XGBoost excels at correctly classifying cases, a crucial aspect of any predictive model. Secondly, a closer examination of the classification report reinforces the superiority of the XGBoost model. The classification report provides insights into precision, recall, and F1-score values for both classes—likely autism and not autism. The XGBoost model consistently showcases higher values across these metrics, indicating a more balanced and accurate classification of positive and negative cases. Furthermore, the confusion matrix, which offers a detailed breakdown of true positives, true negatives, false positives, and false negatives, reflects better performance by the XGBoost model.

References

[1] Maenner MJ, Shaw KA, Bakian AV, et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 Years - autism and developmental disabilities monitoring network, 11 Sites, United States, 2018. MMWR Surveill Summ 2021;70:1–16.

[2] McPheeters ML, Weitlauf A, Vehorn A. U.S. preventive services Task force evidence syntheses, formerly systematic evidence reviews. screening for autism spectrum disorder in young children: a systematic evidence review for the US preventive services Task force. Rockville (MD): Agency for Healthcare Research and Quality (US), 2016

[3] Lipkin PH, Macias MM, Council on children with disabilities, section on developmental and behavioral pediatrics. Promoting optimal development: identifying infants and young children

with developmental disorders through developmental surveillance and screening. Pediatrics 2020;145. doi:10.1542/peds.2019-3449.

[4] Guthrie W, Wallis K, Bennett A, et al. Accuracy of autism screening in a large pediatric network. Pediatrics 2019;144.

[5] Carbone PS, Campbell K, Wilkes J, et al. Primary care autism screening and later autism diagnosis. Pediatrics 2020;146. doi:10.1542/peds.2019-2314.

[6] Park JH, Cho HE, Kim JH, et al. Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. NPJ Digit Med 2020;3:46.

[7] Downs J, Velupillai S, George G, et al. Detection of suicidality in adolescents with autism spectrum disorders: developing a natural language processing approach for use in electronic health records. AMIA Annu Symp Proc 2017;2017:641–9

[8] A. Kazeminejad and R. C. Sotero, “Topological properties of resting-state fMRI functional networks improve machine learning-based autism classification,” Frontiers in Neuroscience, vol. 12, p. 1018, 2019.

[9] N. A. Khan, S. A. Waheeb, A. Riaz, and X. Shang, “A three-stage teacher, student neural networks and sequential feed forward selection-based feature selection approach for the classification of autism spectrum disorder,” Brain Sciences, vol. 10, no. 10, p. 754, 2020.

[10] M. N. Parikh, H. Li, and L. He, “Enhancing diagnosis of autism with optimized machine learning models and personal characteristic data,” Frontiers in Computational Neuroscience, vol. 13, no. 9, p. 9, 2019.

[11] S. Payabvash, E. M. Palacios, J. P. Owen et al., “White matter connectome edge density in children with autism spectrum disorders: potential imaging biomarkers using machine-learning models,” Brain Connectivity, vol. 9, no. 2, pp. 209–220, 2019.

[12] R. M. Thomas, S. Gallo, L. Cerliani, P. Zhutovsky, A. El-Gazzar, and G. van Wingen, “Classifying autism spectrum disorder using the temporal statistics of resting-state functional MRI data with 3D convolutional neural networks,” Frontiers in Psychiatry, vol. 11, p. 440, 2020.

[13] Y. Wang, J. Wang, F.-X. Wu, R. Hayrat, and J. Liu, “AIMAFE: autism spectrum disorder identification with multi-atlas deep feature representation and ensemble learning,” Journal of Neuroscience Methods, vol. 343, Article ID 108840, 2020.

[14] M. Tang, P. Kumar, H. Chen, and A. Shrivastava, “Deep multimodal learning for the diagnosis of autism spectrum disorder,” Journal of Imaging, vol. 6, no. 6, p. 47, 2020.

[15] M. Rakić, M. Cabezas, K. Kushibar, A. Oliver, and X. Lladó, “Improving the detection of autism spectrum disorder by combining structural and functional MRI information,” NeuroImage: Clinic, vol. 25, Article ID 102181, 2020.