Automated ML Approaches to Discriminate the Autism-Categorical Spectrum Disorder

Noora Saleem Jumaah[*], Fouad A. Arif [**], Rasha Hani Salman [***]

[*] Ministry of Higher Education & Scientific Research, Iraq
Email: noorasaleem86@gmail.com
https://orcid.org/0009-0007-3126-4358

[**] College of Engineering, Al-Iraqia University, Iraq
Email: fdshaikley@yahoo.com
https://orcid.org/0009-0000-6193-1456

[***] Department of Studies and Planning, Wasit University, Iraq
Email: rashahany609@gmail.com
https://orcid.org/0009-0009-9615-2079

Abstract

This disorder can be considered as neurodevelopmental which has an affection on an individual's behavior, communicating as well as learning. Its diagnosis needs a lot of time and money. And if it is detected early, it is going to be very helpful via giving the infected ones the suitable treatment in good time. As the development of AI and ML, this disorder is able to be predicted very soon. Thus, the goal here can be the proposing of a system called ASD prediction system by utilizing Naïve bayes classifier for predicting if a person is infected with ASD or not to develop the screening method used to detect the illness. The experimental outcomes showed that the Naïve bayes classifier achieved a great precision, recall and F- measure of about (98. 975%, 98.972 % and 98 .972 %) in kid's dataset. On the other hand, it was noticed that the precision, recall and F- measure in the dataset of the teenagers can be approximately (98.136%, 98.076% and 98.067%) with an estimated error rate that is (0.044).

Keywords- Autism spectrum disorder (ASD), Machine learning (ML), Logistic Regression (LR), K-Nearest Neighbours (KNN).

I. INTRODUCTION

ASD is a frequent, intricate case characterized by sicknesses concerning enhancing the sociability as well as the availability of every usual attitude besides hobbies. It is the fastest improving incapability, and the numbers of commonness execute for rising very greatly. New estimations show that approximately 1.5% of the earth population suffer from autism, and it is thought that several individuals infected by it have not been detected yet. Therefore, the increase in comprehending sickness increased the demand for every service for the purpose of diagnosis.[1] It is quite important since receiving its diagnosing a 'critical milestone', which makes parents understand their children's needs and making them able to get to a decisive backup. ASD can be linked with significant costs connected to health, and early diagnosis decreases them. [2] Unfortunately, the diagnosis consumes a long time since the new procedures of diagnosis consume time and has no production with respect to the costs. All machine learning methods offer automatic effective classifying models for ASD since a mixture of mathematical and lookup methods were taken out of the computer field [3]. The objective was suggesting a system to predict autism depends upon the ML techniques that have the ability to fulfill a predicting related to autism features effectively. In another word, it focuses on developing an app for screening the disorder for predicting ASD features concerning children from 4 to 11 years old and 12 to 17..

II. DATASET

The dataset utilized here that has been taken from UCI Machine Learning Repository [3] is an dataset that is open-source. Two types of datasets have been utilized:

1) "Autism screening Data Set" consists of 104 notes of 21 features
2) "Autism screening child Data Set" consists of 229 notes of 21 features.

III. MACHINE LEARNING METHODS

Several classification algorithms used for the type of the classification task will fulfill here. Three of them (Naive Bayes, Logistic Regression, K-Nearest Neighbors). Using these algorithms will ensure the results to be more depended on and helpful in specifying the one that is very accurate and the error rate is low which, as the classifier in our suggested system, to be selected.

A. Naïve Bayes (NB)

This algorithm's classifier is observed learning and plain probabilistic that calculates for a set of probabilities by computing value and frequency collections. The assumption of NB that is known for us as category conditional independence since NB was based upon the theory of Bayes independently assuming among the features. NB assumes that the illustrating of a characteristic with no dependance on in front of the others, NB is described to be a quick algorithm of learning in different issues of observed classification and it works fine [4].

IV. MODEL EVALUATION METRICS

All Machine learning performances contain certain level of compromising among true positive and true negative rate, and among precision and recall. And besides F-Measure, they are used to retrieve information to measure the performance [5].

A. Cross-Validation

It exemplifies a mechanism used in assessing the method by which any classifier classifies unused any task. Repeating cross-validation contains the division of a data so they become two complementing subsets: the training set and the testing set. In the validation that is k-fold, the real information is firstly divided into k folds or segments. Consequently, k exercising and validating repetitions implemented in all repetitions a different data fold were sustained to validate as the residual folds, which are k-1, utilized for the purpose of learning. In the field of data mining, a cross-validation where k = 10 can be considered very common, and fulfills a serving to be a standard proceeding to fulfill an estimation and model selection [6].

B. Precision, Recall, and F-measure

A number of significant measures is used such as recall, F Measure and precision [7]. And Equation (3) is used to calculate F Measure which is a useful stand-by in case of misclassifying rate measurement [8].

Precision = $\frac{\text{TP}}{\text{TP+FP}}$ (1)

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

F1-measure = 2* $\frac{precision * recall}{precision + recall}$ (3)

Where:

TP, True Positives, rows of data in the sets' test having and predicted for having a target that is positive.

TN, True Negatives, rows of data in the sets' test having and predicted for having a target that is negative.

FP, False Positives, rows of data in the tests' set having a target that is negative but predicted for having a positive one.

FN, False Negative, rows of data in the tests' set having a target that is positive but predicted for having a negative one [9].

C. Error Costs and Estimation

There are a few error alternatives within the highest machine learning tools' number [10].

And Mean Absolute Error: used for being a statistical metric to measure errors happened by model performance [11].

MAE = $\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}|\boldsymbol{e_i}|$ (4)

b. Root Mean Square Error: a useful measure used to estimate every error of model evaluation.

RMSE=$\sqrt{\frac{1}{n}\sum_{i}^{n} e_i^2}$ (5)

V. PROPOSAL SYSTEM

In machine learning, a common task is creating algorithms which learn from and predict on data. Algorithms such as these ones function by data-driven predicting and deciding, yet constructing an arithmetical model from the input data. The suggested system comprises the first step in constructing ASD predicting system has assessed three machine learning algorithms, which are (NB, LR, KNN) and then choose one which accomplishes high accuracy and low error as the classifier for suggested system illustrated in Figure (1).

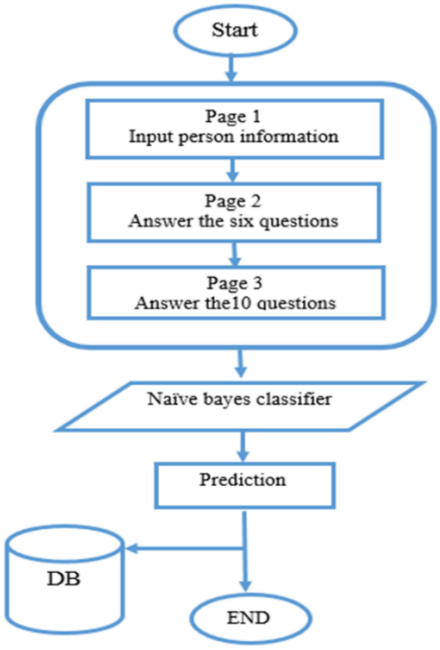

Figure1: Flowchart illustrated Block diagram of ASDPS data flow

This system comprises many steps, which are:
A-Input data: this step consists of the input person's information and answers many questions related of individual's behaviors. The system comprises 3 pages. On the first page, the user inputs personal information (name, age…etc.). The second page includes 6 questions to be answered by the user since it has a role in case diagnosis and the third page includes 10 questions related to ASD patients' behavior.
B-Naïve bayes classifier: After the user, it completes input data, then this data input to NB classifier which is in charge of the predicating process through calculating the probability the person is with ASD or not.
C-predication: Here, the predicating appears as a message says that the patient "has ASD" or "does not have ASD".
D-Database: Here, the predication saves the patient's information and predicting results in system database.

VI. THE RESULTS

Here, the performance results of the classification algorithms are discussed on the kids and teenager dataset completely for all evaluation measures individually.

A-Precision: Table (4.1) illustrates the precision values in kids and teenager dataset which shows the Naïve Bayes algorithm that scored the highest precision (98.975%) if we compare it with the LR algorithm which accomplishes (95.216%) and the KNN algorithm which accomplishes (88.822%) in case of the kids dataset. And the NB algorithm scored the highest precision of (98.136%) if we compare it with the LR algorithm which is (95.229%) and the KNN algorithm is (94.413%) in case of teensger dataset.

Table (4.1) Precision results

Algorithms	**Result of children dataset (kids dataset)**	**Result of adolescent dataset (teenager dataset)**
NB	98.975%	98.136%
LR	95.216%	95.229%
KNN	88.822%	94.413%

B-Recall: Table (4.2) shows the values of recall in kids and teenager dataset. The outcomes illustrate the NB algorithm that scored the highest recall which equals (98.972%) if we compare it with the LR algorithm which is (95.205%) and the KNN algorithm which is (88.698%) in case of kids dataset. And the NB algorithm scored the highest recall which equals (98.076%) if we compare it with the LR algorithm which equals (95.192%) and the KNN algorithm which is (94.230%) in case of teenager dataset.

Table (4.2) Recall results

Algorithms	Result of children dataset (kids dataset)	Result of adolescent dataset (teenager dataset)
NB	98.972%	98.076%
LR	95.205%	95.192%
KNN	88.698%	94.230%

C-F-measure: Table (4.3) shows the values of F-measure in kids and teenager dataset, where the NB algorithm that scored the highest F-measure which equals (98.972%) if we compare it with the LR algorithm which equals (95.206%) and KNN algorithm which is (88.675%) in case of kids dataset. And the NB algorithm scored the highest F-measure which equals (98.067%) if we compare it with the LR algorithm which is (95.202%) and KNN algorithm which is (88.171%) in case of teenager dataset.

Table (4.3) F-measure results

Algorithms	Result of children dataset (kids dataset)	Result of adolescent dataset (teenager dataset)
NB	**98.972%**	**98.067%**
LR	**95.206%**	**95.202%**
KNN	**88.675%**	**94.171%**

D-Root Mean Squared Error (RMSE): Table (4.5) shows the values of RMSE in kids and teenager dataset, where the NB algorithm that scored the lowest RMSE which is (0.109) if we compare it with the value of RMSE in the LR algorithm which is (0.217) and KNN algorithm which equals 0.304 in case of kids dataset. And the NB algorithm scored the lowest MAE which equals (0.126) is we compare it with the value of MAE in the LR algorithm which equals (0.219) and KNN algorithm which is (0.284) in case of teenager dataset.

Table (4.5) Root Mean Squared Error results

Algorithms	Result of children dataset (kids dataset)	Result of adolescent dataset (teenager dataset)
NB	**0.109**	**0.126**
LR	**0.217**	**0.219**
KNN	**0.304**	**0.284**

VII. CONCLUSION

The work here gives the following result: the suggested system has the ability for predicting if an individual owns ASD or does not have ASD through the Naive bayes classifier. The outcomes illustrate that the NB algorithm accomplishes high accuracy that equals (98.975%), and (98.136%) accuracy in the case of kids, teenager in order and lower error than LR and KNN algorithms. Utilized K-fold cross validation mechanism to split dataset which helps for achieving high performing. The suggested system developed the detecting process for autism traits. It illustrates the method that every correlation among autistic features helps physicians, family and caregivers to improve how ASD screening performs which will soon have the ability to substitute the scoring functions and handcrafted rules recently in progress within the tools of classic screening. Through the ASDPS database, a novel ASD dataset is suggested for adding to the ASD dataset.

REFERENCES

[1] American Psychiatric Association (APA). Diagnostic and statistical manual of mental disorders. Fifth ed. Arlington, VA: APA.

[2] Vishal, V., Singh, A., Jinila, Y.B., Shyry, S.P. and Jabez, J., 2022, April. A comparative analysis of prediction of autism spectrum disorder (ASD) using machine learning. In 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 1355-1358). IEEE.

[3] Ruzich E, Allison C, Smith P, et al. measuring autistic traits in the general population: a systematic review of the Autism-Spectrum Quotient (AQ) in a nonclinical population sample of 6,900 typical adult males and females. Mol Autism 2015; 6: 2.

[4] Centers for Disease Control and Prevention. Prevalence of autism spectrum disorder among children aged 8 years – autism and developmental disabilities monitoring network, 11 sites, United States, 2010. MMWR Surveill Summ 2014; 63(2): 1–21.

[5] Brugha T, McManus S, Bankart J, et al. Epidemiology of autism spectrum disorders in adults in the community in England. Arch Gen Psychiatry 2011; 68(5): 459–466.

[6] Fitzgerald M. The clinical gestalts of autism: over 40 years of clinical experience with autism. In: Fitzgerald M and Yip J (eds) Autism – paradigms, recent research and clinical applications. London: InTech, 2017, pp. 1–13.

[7] Russell AJ, Murphy CM, Wilson E, et al. The mental health of individuals referred for assessment of autism spectrum disorder in adulthood: a clinic report. Autism 2016; 20(5): 623–627.

[8] Thabtah F. Machine learning in autistic spectrum disorder behavioural research: a review and ways forward. Inform Health Soc Care 2018; 2: 1–21.

[9] Crane L, Chester J, Goddard L, et al. Experiences of autism diagnosis: a survey of over 1000 parents in the United Kingdom. Autism 2016; 20(2): 153–162.

[10] Wang, H., Chi, L., Yang, H., Li, L., & Zhao, Z. (2021). A deep learning predictive classifier for autism screening and diagnosis. In Neural Engineering Techniques for Autism Spectrum Disorder (pp. 215-236). Academic Press.

[11] Magrelli, S., Jermann, P., Noris, B., Ansermet, F., Hentsch, F., Nadel, J., & Billard, A. (2013). Social orienting of children with autism to facial expressions and speech: a study with a wearable eye-tracker in naturalistic settings. Frontiers in psychology, 4, 840.