Machine Learning Approach to Predict Autism Spectrum Disorder

Smit shah
Certification-course(DS - AI),
IIT MADRAS
Surat, India
SMITSHAH0208@gmail.com

***Abstract*—Autism spectrum disorder is one of a group of neuro developmental disorders. ASD's are lifelong developmental disabilities which causes a substantial impairment in an individual's communication, social interactions, range of play and interests, processing and integrating sensory information. It has been noted that symptoms of ASD can range from mild to severe. So, pursuing such a research to find out most significant traits of ASD with the help of the Supervised Classification Learning Algorithms we will construct a model which will predict if a person has autism or not based on several scientific aetiology, symptoms and factors affecting ASD. Hence, we are analysing our dataset using libraries which includes, pandas, sklearn [9], numpy, and many more, to check accuracy, sensitivity, specificity and precision using different models and finding suitable among them.**

I. Introduction

Autism spectrum disorder is a developmental disability that affects how a person interacts and behaves in social situations [8]. The general consensus is that a neurodevelopmental disorder affects brain development and impairs the development of social and communication skills. The term spectrum refers to the wide range of symptoms and behaviours a person might have. According to Centers for Disease Control and Prevention(CDC) about 1% of the world's population has ASD (i.e. 7,50,00,000 people). The number of people diagnosed with ASD in the United States and other countries has increased since the 1970s and particularly since the late 1990s but, it is not clear if the increase is related to the changes in the criteria used to diagnose ASD or if the condition has truly become more common over time.

II. Problem Statement

Our objective is to detect whether "an individual has ASD" or "an individual does not have ASD" with the help of predetermined binary classified dataset.

To pursue our objective we have used 4 classifier models,

1. k-Nearest Neighbors (kNN)
2. Decision Trees
3. Logistic Regression
4. Support Vector Machines (SVM)

III. Dataset Description

We have taken the dataset [1] from UCI Machine learning Repository which was released on 24th of December, 2017 [5]. Dataset consists of different age group, that is, Child(4 to 11), Adolescence(12 to 16) and Adult(Above 16) with 20 features which are:-

1. age: age of an individual in years.
2. gender: gender of an individual.
3. ethnicity: belonging of an individual to a social group.
4. jaundice: whether the individual was having jaundice at birth or not.
5. autism: an immediate family member has a pervasive developmental disorder.
6. relation: relation with the suspected individual.
7. contry_of_res: country of residence.
8. used_app_before [7]: any prior screening or test for ASD.
9. age_desc: Age category.

A1-A10 scores: 1(YES)/ 0(NO) based on the question asked in screening.

10. A1 score: The answer code of: Does the person speak very little and give unrelated answers to questions?
11. A2 score: The answer code of: Does the person not respond to their name or avoid eye contact?
12. A3 score: The answer code of: Does the person not engage in games of pretend with other children?
13. A4 score: The answer code of: Does the person struggle to understand other people's feelings?
14. A5 score: The answer code of: Is the person easily upset by small changes?
15. A6 score: The answer code of: Does the person have obsessive interests?
16. A7 score: The answer code of: Is the person over or under-sensitive to smells, tastes, or touch?
17. A8 score: The answer code of: Does the person struggle to socialize with other children?
18. A9 score: The answer code of: Does the person avoid physical contact?
19. A10 score: The answer code of: Does the person show little awareness of dangerous situations?
20. Result: Count of 1(YES) for the questions.

21. Class: Whether an individual has ASD or not (YES/ NO).

IV. Data Pre-Processing

Data preprocessing is a technique used to transform RAW data into a useful and efficient format. Using "dropna" from pandas we have removed "NULL" value rows from our dataset. In our columns, we have string values but they are categorical so we have applied One-Hot-Encoding using the "get dummies" method from pandas to convert them into binary values after which our dataset consists of 955 rows and 11 columns [4]. Then we have divided our whole dataset into training and testing sets we have kept 70% for training and 30% for testing, also we have set random state to 42 so it chooses the same values from the dataset in training and testing every time.

V. Evaluation Strategy

To evaluate the accuracy of our model we have used 4 metrics [6].

1. Sensitivity: Sensitivity is the metric that evaluates a model's ability to predict the true positives of each available category.

$$\text{Sensitivity} = \frac{TruePositives}{TruePositives + FalseNegatives}$$

2. Specificity: Specificity is the metric that evaluates a model's ability to predict the true negatives of each available category.

$$\text{Specificity} = \frac{TrueNegatives}{TrueNegatives + FalsePositives}$$

3. Precision: Precision is the measure of positive predictive value.

$$\text{Precision} = \frac{TruePositives}{TruePositives + FalsePositives}$$

4. Accuracy: Accuracy is the fraction of predictions our model got right.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP}$$

VI. Model Implementation

A. k-Nearest Neighbors (kNN)

kNN is one of the simplest machine learning model based on supervised learning widely used for classification as well as regression problems. It works on the concept of the assumption that similar things exist in the same proximity. It calculates the distance between the query example and the current example from the data by using Euclidian distance. It performs majority voting among k nearest data points and defines the label of the given point. In classification, k should be an odd value for better results. It is also called a lazy learner Algorithm because it does not learn from the training set immediately hence its computation time is fast and accuracy is low.

Before kNN:

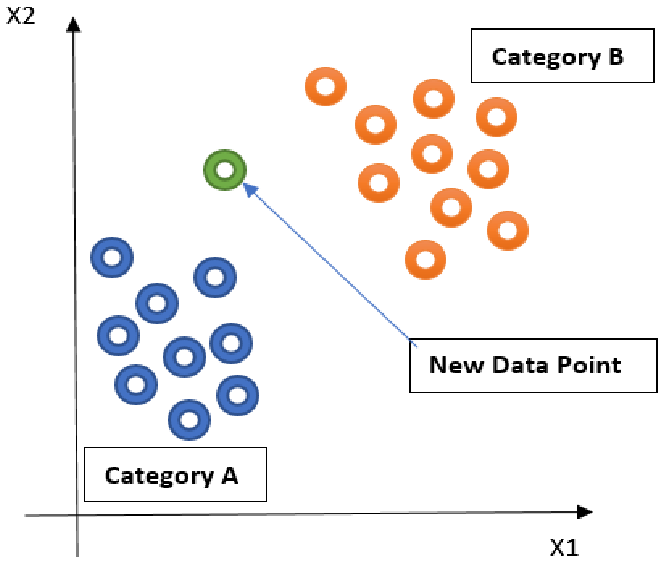

Fig. 1. Before k-NN

After kNN:

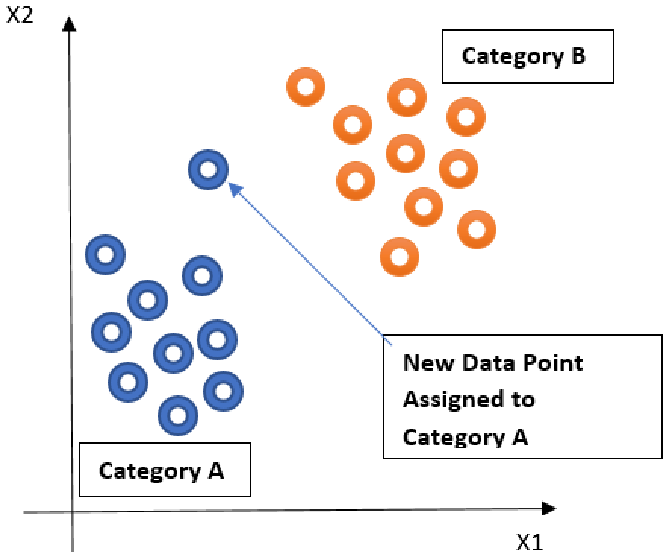

Fig. 2. After k-NN

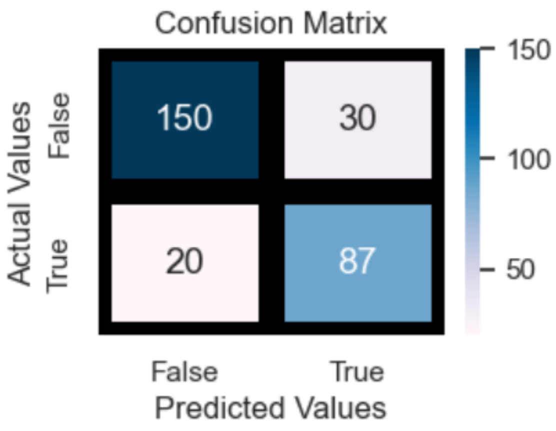

Fig. 3. Confusion Matrix of kNN

B. Decision Trees

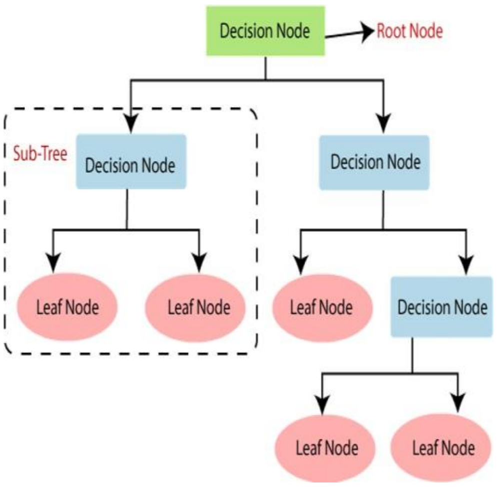

Fig. 4. Decision Tree

Decision Tree - A supervised machine learning model and is preferable to solve classification problems. Internal nodes are the features of the dataset, branches of the tree defines decision rules, and leaf node represents the outcome. In this model, it selects the best attribute in the dataset using the Attribute Selection Measure(ASM). It continues to create nodes until a point where no further classification can be done and the final leaf node is reached. An optimal Decision tree is where it has most amount of data and also the level of questions are minimized. It is linearized to some decision rules, which decides the outcome in terms of leaf node. The accuracy of decision tree can be changed by increasing the depth of it and making more decision rules to classify the data points.

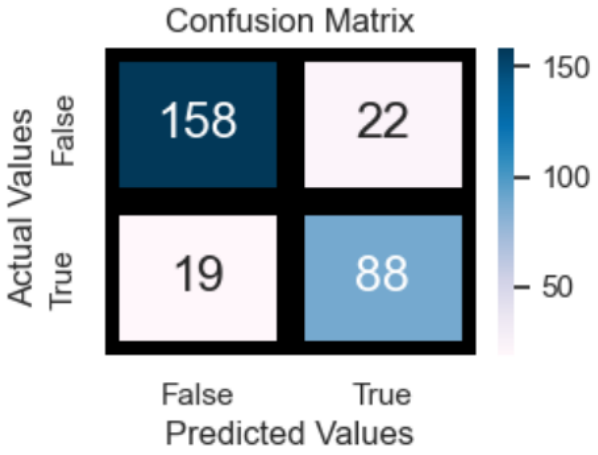

Fig. 5. Confusion Matrix for Decision Tree

C. Logistic Regression

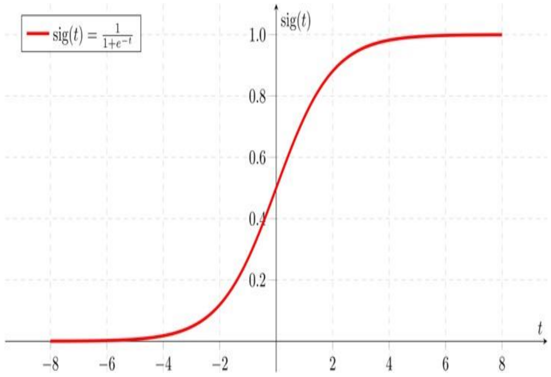

Fig. 6. Logistic Regression

Logistic Regression is a supervised learning model used for classification problems [2]. It predicts the output of a categorical dependent variable in terms of either Yes or No, 0 or 1, True or False, etc. It gives probabilistic values lying between 0 to 1 instead of giving the exact value as 0 or 1. The graph is obtained using the sigmoid function which can be depicted as an ‘S-shaped logistic curve. The sigmoid function helps to map the predicted values to probabilities and convert any real value within a range from 0 to 1. A threshold value is defined and all the points above that are classified as 1 and below that are classified as 0. The dependent variable must be categorical to apply the logistic function. Furthermore, Binomial is used when there are only 2 possible results, Multinomial is used when there are 3 or more unordered possible results, and Ordinal is used when there are 3 or more ordered possible results.

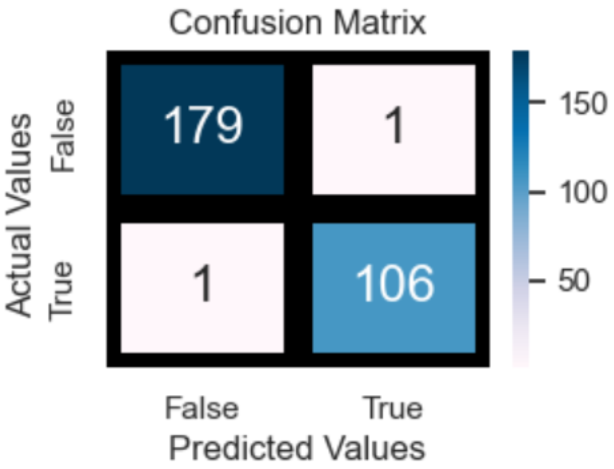

Fig. 7. Confusion Matrix for Logistic Regression

D. Support Vector Machines (SVM)

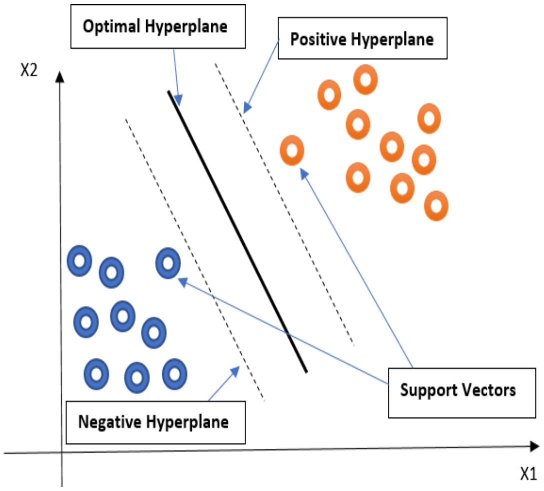

Fig. 8. Linear Support Vector Machine

SVM is a supervised learning model widely used for classification problems and can also be used for regression. It finds the hyperplane (middle line) also called as decision boundary which will help us separate two classes and also the maximum margin for the problem. It chooses extreme points or vectors which helps to create a hyperplane. Linear SVM is used for linearly separable data and non-linear SVM for non-linearly separated data. A straight line is formed to classify both classes in linear SVM and non-linear in the case of non-linear SVM. It is also known as a maximal margin classifier

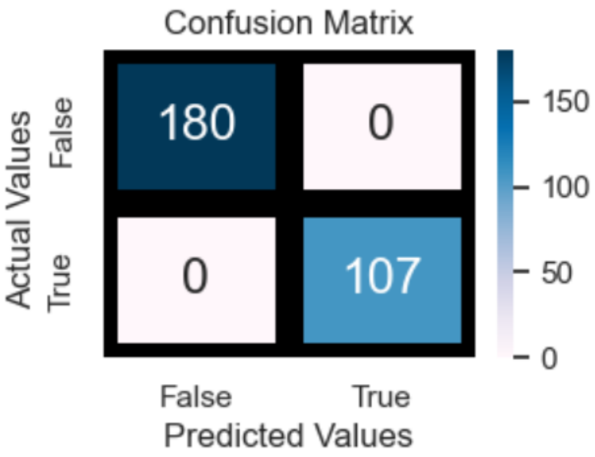

Fig. 9. Confusion Matrix for Linear SVM

VII. Results

A. For Evaluation Strategies

In the below table we have calculated all the evaluation strategies which we have mentioned above:-

Model	Sensitivity	Specificity	Precision	Accuracy
K-Neighbors Classifier	0.813084	0.83333	0.743590	0.825784
Decision Tree Classifier	0.822430	0.87778	0.800000	0.857143
Logistic Regression	0.990654	0.994444	0.993031	0.993031
SVM (linear)	1.000	1.0	1.0	1.000000

B. ROC Curves

ROC curve,which is also known as ”Receiver Operating Characteristics” Curve, is a metric used to measure the performance of a classifier model. The ROC curve shows us the rate of true positive(TPR) with respect to the rate of false positive(FPR), which means the sensitivity of the classifier model. It can be plotted with different thresholds settings. The TPR can be also considered as sensitivity, probability or recall of detection. The FPR can be termed as probability of False Alarm and can be calculated by computing (1-Specificity). Thus, it can be also said that ROC Curve is recall or sensitivity as a function of fall-out.

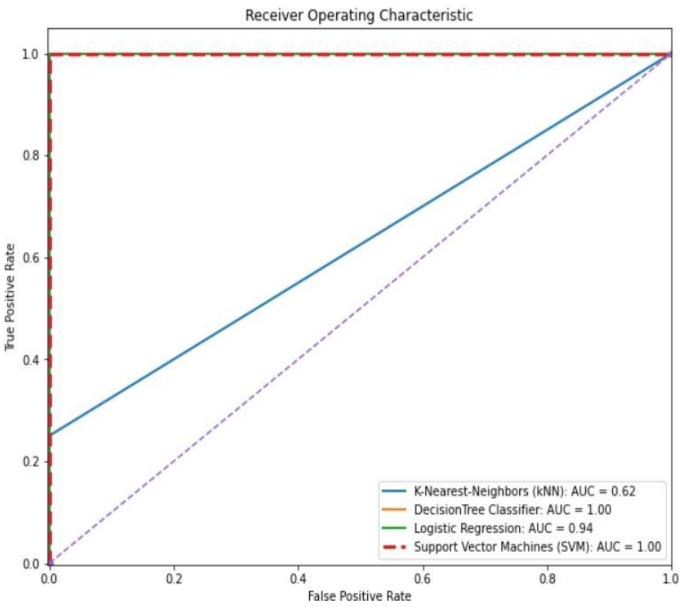

Fig. 10. AUC - ROC Curve

VIII. Conclusion

Screening of Autism at early age should be a fundamental step towards understanding traits of Autism. Although there is no cure for Autism in terms of Medicine, but one can be benefitted by certain therapies and behavioral changes. Talking about our project, we analyzed datasets of Child, Adolescent and Adult of Autism screening. From results, it can be derived that for this dataset Decision Trees and SVM(Linear) are predicting with 100% Accuracy. Decision Trees can easily predict continuous and discrete variables, also it is non-parametric. SVM(Linear) can create the decision boundary(hyperplane) with maximum margin because our data is linearly separable. Also, classes of all testing data are predicted accurately because of the proper classification of classes linearly. We also derived that k-means is predicting with 97.9% accuracy because it is not efficient towards outliers. Logistic is also predicting with 99.65% accuracy because there are some multicollinearity between some features and also it is giving highest running time among all 4 algorithms. Further, it can be concluded that Decision Trees is the best model for screening of Autism as it is giving 100% Accuracy and it is taking least time to run the algorithm and significantly very less than SVM(Linear).

IX. Future Work

The only limitation of this dataset is that it has very less rows after data cleaning. So in the future to build a robust and optimized model, one will require a very large number of datasets [3]. After learning the whole dataset and applying the models, it can be derived that all the models are working at their best and there is no further improvement required in this part. Also, we observed that there is very less content available about ASD screening in terms of questions that can accurately define ASD. So, more research on symptoms can help significantly for Autism screening and detection of ASD. Our research can benchmark accuracy as well as learning for others who want to further explore this dataset and who want to work further in the field of Autism Screening and detecting ASD.

References

[1] Thabtah, Fadi & Abdelhamid, Neda & Peebles, David. (2019). A machine learning autism classification based on logistic regression analysis. Health Information Science and Systems. 7. 10.1007/s13755-019-0073-5.

[2] Y. Zheng, T. Deng and Y. Wang, ”Autism Classification Based On Logistic Regression Model,” 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), 2021, pp. 579-582, doi: 10.1109/ICBAIE52039.2021.9389914.

[3] Hossain, Md & Kabir, Ashad & Anwar, Adnan & Islam, Md. (2020). Detecting Autism Spectrum Disorder using Machine Learning.

[4] Autism-Detection-in-Adults/report.pdf at master · kbasu2016/Autism-Detection-in-Adults · GitHub. URL **https://github.com/kbasu2016/Autism-Detection-in-Adults/blob/master/report. pdf**.

[5] B. Tyagi, R. Mishra and N. Bajpai, ”Machine Learning Techniques to Predict Autism Spectrum Disorder,” 2018 IEEE Punecon, 2018, pp. 1-5, doi: 10.1109/PUNECON.2018.8745405.

[6] S. B. Shuvo, J. Ghosh and A. S. Oyshi, ”A Data Mining Based Approach to Predict Autism Spectrum Disorder Considering Behavioral Attributes,” 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019, pp. 1-5, doi: 10.1109/ICCCNT45670.2019.8944905.

[7] Thabtah, F. (2017). ASDTests. A mobile app for ASD screening. **www.asdtests.com** [accessed April 03rd, 2022].

[8] Tabtah, F. (2017). Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment. Proceedings of the 1st International Conference on Medical and Health Informatics 2017, pp.1-6. Taichung City, Taiwan, ACM.

[9] A. G´eron, Hands-On Machine Learning with Scikit-Learn & Tensor Flow, O’Reilly ISBN: 9781491962299.