

Detection of Autism Spectrum Disorder using Machine Learning

Mrs. K. Swarupa Rani¹, Sk. Tabasum Fathima², S. Gayathri³, M. Pavani⁴, Y. Swathi⁵

Assistant Professor, Department of Information Technology¹

B. Tech Students, Department of Information Technology^{2,3,4,5}

Prasad V Potluri Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh, India

Abstract: *AUTISM SPECTRUM DISORDER (ASD) is a neurological disorder caused by differences in brain. It is one among the major disorders by which adults as well as children are suffering. In Autism Spectrum Disorder the word Autism means- a developmental disorder that is characterized by difficulty in social interaction and communication. It is derived from Greek word 'auto' which means 'self'. Spectrum means- which is not limited to specific conditions and Disorder means- a state of confusion/mess. Characteristics related to ASD can include can't able to make eye contact with others, does not respond to name and can't able to show facial expressions like happy, sad, angry, and surprised by 9 months of age, showing no gestures by 12 months of age, does not show interest to play with other children by 36 months of age. Some other characteristics include-delaying language skills, delaying learning skills, unusual eating and sleeping habits, gastrointestinal issues (for example, constipation), unusual mood or emotional reactions, anxiety, stress, lack of fear or more fear than expected [1].*

Keywords: Logistic Regression, Data pre-processing, numpy, pandas, matplotlib, seaborn, sklearn

I. INTRODUCTION

ASD is mainly caused by genetics or by environmental factors. However, its conditions can be improved by detecting and treating it at earlier stages. Or else it leads to great impact on future. In present, to diagnose ASD the clinical standardized methods are the only way. This not only requires more time but also results in increase of medical costs. To improve the analysis and time required for diagnosis, machine learning techniques are being used to detect autism spectrum disorder in early stages. These techniques act as complement to conventional methods which are more helpful and faster in detecting ASD [1].

II. PROPOSED SYSTEM

We consider Logistic Regression algorithm to detect Autism Spectrum Disorder (ASD) [1] in children. Logistic regression [2] is one of the best Classification Algorithms in Machine Learning. Logistic Regression comes under supervised Learning. It consists of a function known as sigmoid function which classifies the output as categorical variable. It will classify the output in binary format as YES for having Disorder and NO for not having disorder or 0 or 1.

$$\text{Sig}(x) = \frac{1}{1 + e^{-x}}$$

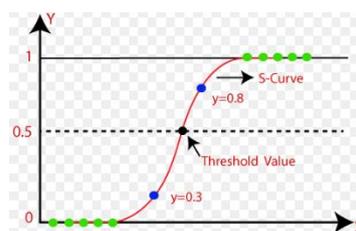
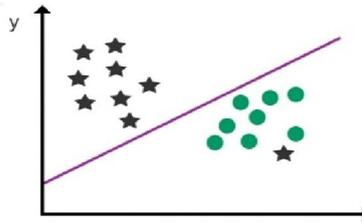


Figure 1.1: Sigmoid curve



The output of this function depends on the x variable, it indicates the attributes that are taken in dataset and Sig(x) will be the output varies between 0 or 1. Here threshold value is 0.5. If the function value comes above threshold value then it predicts 1 or if the function value below the threshold value then it predict as 0. A hyperplane is used as a decision line to separate two categories (as far as possible) after data points have been assigned to a class using the Sigmoid function. The class of future data points can then be predicted using the decision boundary.



Logistic regression helps to predict events that have a binary outcome, such as whether a person will successfully pass a driving test. In order to make predictions in this scenario, you need data from past test results. The model takes this data and predicts the likelihood that the same person will pass the test in the future. The main idea behind logistic regression is to use a model based on the probability of an outcome occurring.

III. TECHNOLOGIES USED

3.1 Jupyter Notebook

Jupyter notebook is a interactive web application used to run code by uploading datasets. It is an open source application which enables users to share documentation include code, multimedia resources etc. Jupyter notebook is easy to use [6].

3.2 Python Libraries

In this application we mainly use these libraries- numpy, pandas, matplotlib, seaborn, sklearn. Numpy library is used to work with array operations. Pandas library is used to manipulate the data. It is built on top of numpy so it can also provide support for multidimensional arrays. Matplotlib library is used for visualizations- to plot graphs and figures. Seaborn is also used for plotting graphs with themes. Sklearn is one of the famous and most important library in python because it provides tools for classification, regression ad clustering [4].

3.3 Dataset

Dataset is taken from kaggle [3]. It is a CSV file of having 1054 rows and 19 columns. The attributes present in the dataset are 'A1', 'A2', 'A3', 'A4', 'A5', 'A6', 'A7', 'A8', 'A9', 'A10', 'Age_Mons', 'Qchat-10-Score', 'Sex', 'Ethnicity', 'Jaundice', 'Family_mem_with_ASD', 'Class/ASD Traits '. 'Case_No', 'Who completed the test' these attributes are not used so we can remove them in order to increase data efficiency and reduce time for processing the data.

IV. RESULTS

4.1 Bar Graphs

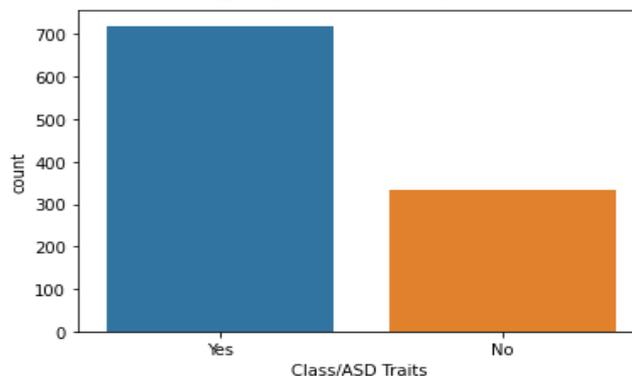


Figure 1.3: ASD Traits analysis



This figure depicts the classification of children based on gender who are having autism disorder and those who are not having disorder. Here blue represents female children and yellow represent male children. For example if we take full count as 700, in those count above 300 male children are affected by autism and 700 children are affected by ASD [1].

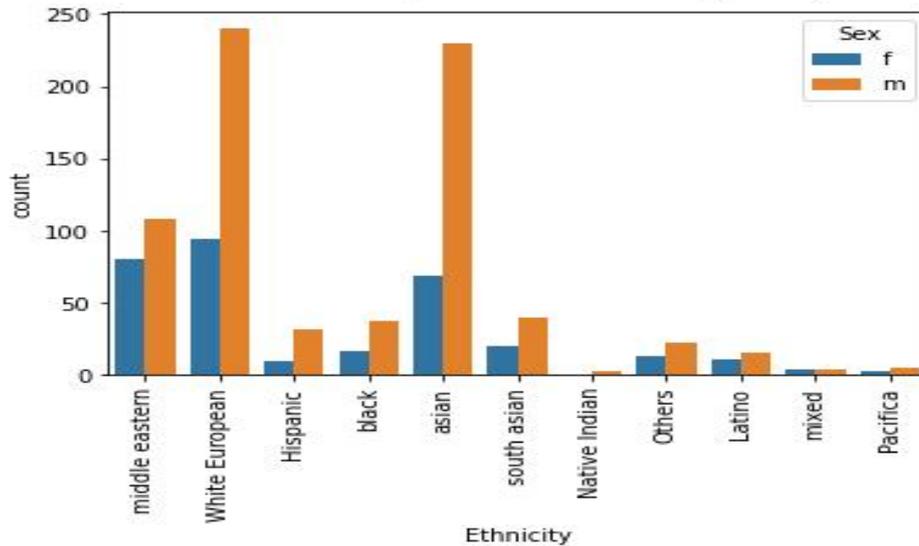


Figure 1.4: Ethnicity analysis based on gender

This figure depicts the classification of children based on Ethnicity. As we can see from above figure the most effected male and female children in countries with ASD are Europe, Middle East and South Asia. And least effected male and female children in countries with ASD are India and Pacifica [1].

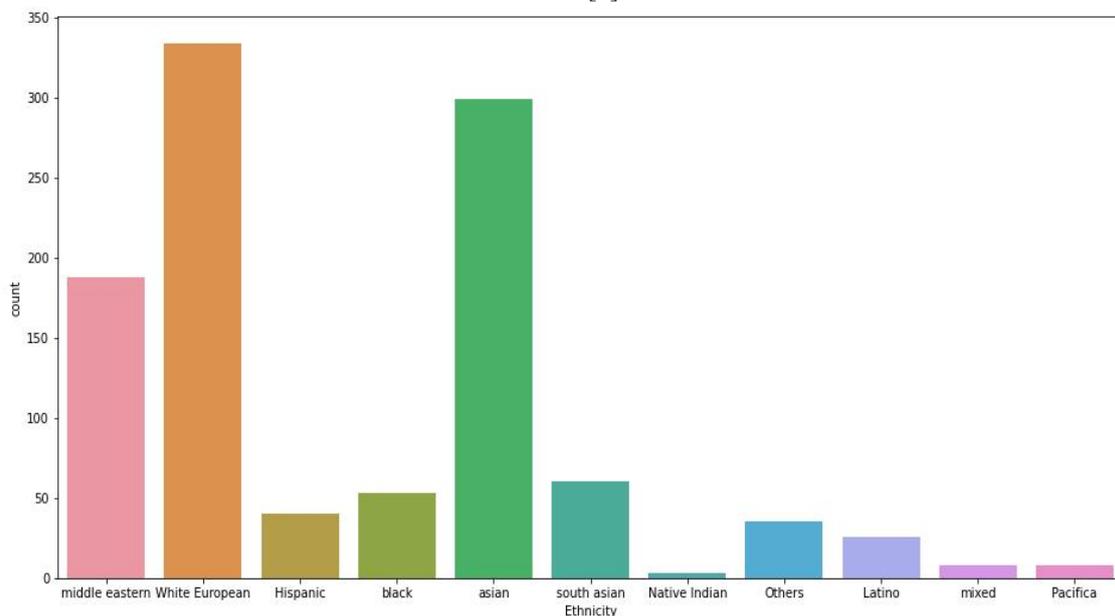


Figure 1.5: ASD count based on Ethnicity

This figure predicts the count of children having ASD [1] in different countries. Countries having highest count with ASD cases are Europe, Middle East and Asia. And countries with least count of ASD cases are India, Pacifica.

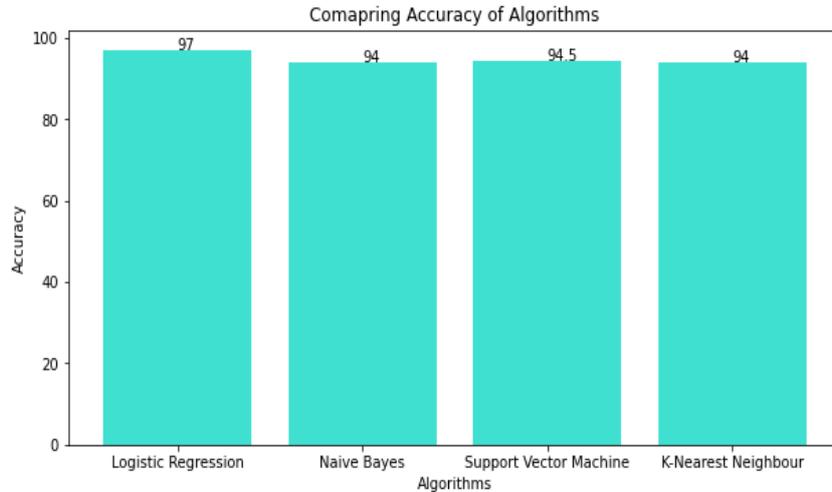


Figure 1.6: Analysis of Algorithms

This figure depicts the accuracy of four classification algorithms in Machine Learning. ML model accuracy is the measurement of best algorithm relationship with variables comparing with other algorithms by taking input as text, images or trained data. From the above bar graph we can understand that Logistic Regression [2] algorithm gives high accuracy compared to Naive Bayes, Support Vector Machine (SVM), K -Nearest Neighbour [9].

The accuracy of Logistic Regression is 97%. The accuracy of other algorithms is less 95%. We can use of seaborn library for themes. Taking this bar graph as an example we can say that Logistic Regression is more efficient than other classification algorithms.

V. SCOPE OF FUTURE USE

The application can be improved further by taking images as input. And the data can be analyzed in a better format by applying efficient software tools. The future use of this application may include in children hospitals or childcare homes or it can also used in anganwadi schools to detect whether a child is suffering from ASD [1] or not in early stages in order to get treated in starting stages only.

VI. CONCLUSION

Therefore, out of all four algorithms- Logistic Regression, Naive Bayes, Support Vector and K-nearest neighbour, the Logistic Regression [2] has given highest accuracy in prediction of Autism Spectrum Disorder [1] in children of age below 3years. Other algorithms have some drawbacks like SVM works well only for unstructured and semi-structured data, Naive Bayes gives good accuracy only with small datasets (doesn't suit for big datasets), K -NN is slow because it is a non-parametric model. Logistic Regression [2] overcome these drawbacks and work efficient and gives high accuracy (97%) compared to other algorithms [9].

REFERENCES

- [1]. <https://www.cdc.gov/ncbddd/autism/signs.html>
- [2]. <https://www.javatpoint.com/logistic-regression-in-machine-learning>
- [3]. <https://www.kaggle.com/fabdelja/autism-screening-for-toddlers>
- [4]. <https://towardsdatascience.com/top-10-python-libraries-for-data-science-cd82294ec266>
- [5]. <https://www.javatpoint.com/data-preprocessing-machine-learning>
- [6]. <https://jupyter.org/>
- [7]. <https://towardsdatascience.com/performance-metrics-confusion-matrix-precision-recall-and-f1-score-a8fe076a2262>
- [8]. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html
- [9]. <https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222>