Autism Prediction using ML Algorithms

1st Kango Sai Tejaswi
Computer Science and Engineering
Sri Sai Ram Engineering College
Chennai, India
teju19.kango@gmail.com

2nd Meghavarshini K
Computer Science and Engineering
Sri Sai Ram Engineering College
Chennai, India
meghavarshinikanagaraj@gmail.com

3rd. P. Nivedhitha
Computer Science and Engineering
Sri Sai Ram Engineering College
Chennai, India
nivedhithap96@gmail.com

2022 1st International Conference on Computational Science and Technology (ICCST) | 978-1-6654-7655-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCST55948.2022.10040300

Abstract-ASD is gaining ground now more quickly than ever before. Screening testing for autistic features cost a lot of money and take a lot of time. The development of ML and AI has made it possible to predict autism relatively early. Despite the fact that several research have been conducted using various approaches, these investigations have not led to any conclusive findings about the ability to predict autism features in terms of several age factors. The objectives are to provide a web-based program or a website that can identify ASD in persons for any age. as well as to offer a useful prediction model based on ML approach. As a result of this study, using Logistic Regression, AdaBoost and Random Forest Algorithms, autism prediction model was created.

Keywords- ASD-Autism Spectrum Disorder, ML-Machine Learning, AI-Artificial Intelligence

I. INTRODUCTION

Autism is a broad term for a number of illnesses that are characterized by challenges with social interaction, persistent behavior, speech, and nonverbal interaction. An individual's ability to interact, communicate, and learn is affected by autism spectrum disorder, a neurodevelopmental condition. Despite the fact that Autism can be identified at any age. Patients with autism deal with a variety of difficulties, including attention problems physical and sensory impairments, intellectual challenges, mental health conditions like anxiety and sadness. Autism diagnosis takes a long time and costs a lot of money. By giving patients the right medication at an early stage, earlier detection of autism can be very beneficial. In order to forecast an individual's autism features and determine whether they are autistic, a quick, reliable, and simple screening test instrument is desperately required.

This project seeks to provide a machine learning (ML)-based model for predicting autistic traits as well as a web application that can accurately predict autism traits in people of any age. To put it another way, the goal of this effort is to create a tool for detecting autism spectrum disorders(ASDs) in humans.

II. BLOCK DIAGRAM

The UCI Repository was used to get the data on the autism spectrum. After that, the data is processed, by a process known as data preprocessing. Data preparation is the process of preparing (cleaning and organizing) raw information to be suitable for building and training ML models. Data preprocessing in ML is, to put it simply, a data mining approach that converts raw information into a format that is readable and intelligible. A small portion of the original data is employed to train the model, whereas testing data is used to validate the model's accuracy.

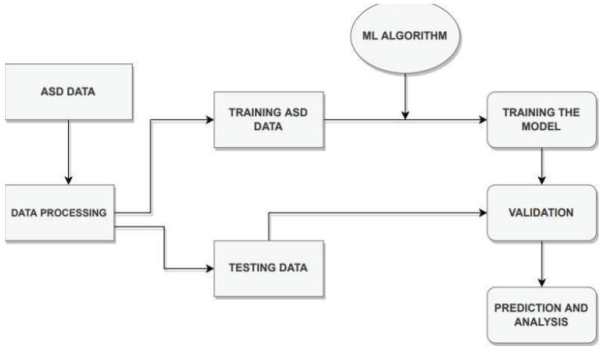


Fig 1. Classification of ASD using ML algorithms

Giving the ML algorithm—the supervised learning where the data are used as a learning resource is the method used to train a ML model. The term "ML model" denotes to the model object produced during training. The proper response, also known as target attribute, must be incorporated into the training phase. The learning algorithm develops an ML model that captures these patterns by looking for similarities in the training phase that relate the parameters of the input data to the target. A crucial phase of predictive analysis is validation. In this step, we conduct a variety of tests to

determine how effective our model is. Here, we give sample input sets so that you can evaluate the accuracy of our model. At this point, the model's accuracy has to be evaluated. Data exploration is a part of the data analysis. To find some patterns or new results from the data collection, we thoroughly examine and analyze the data. During this phase, we gather relevant data, draw a conclusion by spotting some patterns or trends, and then create a predictive model. Using a variety of algorithms, we create prediction models based on the observed patterns in this step of predictive analysis. Standard statistical models are also used to test our hypothesis before the model is actually put into use. By deploying our model, we enable it to operate in a real situation, support regular conversation, and make it usable.

III. LITERATURE SURVEY

The works pertaining to the ASD prediction approaches are briefly presented in this section. It was employed by Wall et al [5]. Using a decision tree (AD Tree) to speed up screening and quicker ASD characteristics identification. Autism Diagnostic was utilized. Interview. However, the test was confined to those between the ages of 5 and 17, and it was unable to predict ASD for various age groups (children, adolescent and adults). Support vector machine (SVM) was utilized by Bone et al[6] to apply ML for the same goal and achieve 89.2% sensitivity and 59% specificity.

In their study, 1264 people with ASD and 462 people without ASD features were involved. However, because of the vast age range (4-55 years), their research was not approved as a screening method for all age groups. Hauck and Kliewer[7] attempted to pinpoint substantially more significant screening questions for ADOS, whereas Bekerom [10] employed a variety of machine learning approaches, such as naive bayes, SVM, and the random forest algorithm, to identify ASD symptoms in children, such as developmental delay, obesity, and a lack of physical activity. The findings were then compared.

IV RESEARCH METHODOLOGY

A. COLLECTION OF DATA
B. DATA SYNTHESIZATION
C. CLEANING AND PREPROCESSING OF DATA
D. DATA VISUALIZATION
E. DEVELOPING AND EVALUATING THE PREDICTION MODEL
F. DEVELOPING A WEB APPLICATION

A. COLLECTION OF DATA:

The UCI Machine Learning Repository was used to choose the Autism Screening Toddler Dataset. Data file and data description file were downloaded in a zip folder from the data folder. We changed the data file from . arff to.csv format because it was previously in that format. There are 1056 parameters in 18 columns in the dataset. A successful prediction model was developed using the AQ-10 dataset, which is made up of three separate datasets based on the questions from the AQ-10 screening instrument. Data from the age categories of children, adolescents, and adults (aged 18 or older) are included in these three datasets (adult). The AQ-10, also known as the Autism Spectrum Quotient tool, is used to determine if a person needs to be referred for a thorough autism evaluation. Each of the 10 questions can only be scored with 1 point, according to the scoring methodology. Depending on their response, the user might receive 0 or 1 points for each question. Age, gender, ethnicity, and other qualitative and numerical data are among the eighteen features that each of the three datasets comprises. Jaundice at birth, ASD in the family, who is taking the test, questions 1 through 10, results, and class.

B.DATA SYNTHESIZATION

In order to eliminate unnecessary characteristics, the gathered data were synthesized. List wise deletion was used to manage null values. Then, the algorithms were employed to remove extraneous features from the dataset. According on the results, removing the “age in terms of months”, "members who used the application previously," and "age" columns would lead to a more precise classification.

C.CLEANING AND PREPROCESSING OF DATA

Data preparation is the process of transforming the original content into a format that can be understood. To maximize the usefulness of the data, data preparation is a vital stage in data mining. Any analytical algorithm's results are directly impacted by the data processing techniques.

Steps Involved Data Pre-processing:

1. **Acquiring the data**
2. **Adding the Dataset and Libraries**
3. **Sorting the dataset into dependent and independent variable**
4. **Dealing with missing values**
5. **Training the data split**

1.ACQUIRING THE DATA

Data is unstructured information that represents both human and automated observations of the outside environment. Depending on the kind of issue you wish to resolve, the dataset will vary greatly. Machine learning problems are approached differently for each one. One of the earliest online sources for the dataset is the UCI Machine Learning Repository. used to choose the required dataset.

2.ADDING THE DATASET &LIBRARIES

Importing the necessary libraries into the program is typically the first step. In essence, a library is a group of callable and usable modules and the 'import' keyword can be used to import libraries into Python code.

→import numpy as np

→import pandas as pd

→import matplotlib.pyplot as plt

→import seaborn as sns

Using the Pandas libraries we can load the data

→import pandas as pd

→df2=pd.read_csv("Toddler autism dataset july 2018.csv")

3.DEALING WITH MISSING VALUES

We may occasionally discover that the dataset contains some missing data. If there are any, the rows will be removed; otherwise, the feature's mean, mode, or median can be calculated and the missing values will be substituted.

This is an estimate that can introduce variation into the dataset.

→df2.info

With the use of info(), we are able to determine the entries total as well as the number of non-null values for all features.

4.SORTING THE DATASET INTO DEPENDENT &INDEPENDENT VARIABLE

Identification of the predictor variables (X) and the outcome variable should come next once the dataset has been imported (Y)..A dataset may be labelled or unlabeled in general, but in this case I'm evaluating a tiny dataset for easier comprehension while also evaluating a labelled dataset for an ML classification task. Our dataset has 5 columns: Sex, Affected by Jaundice, Who Completed the Test, Class, and Family Members with ASD. Four independent variables (Sex, Affected by Jaundice, Who Completed the Test, and Family Members with ASD) are present in our dataset, along with one dependent variable (Class), which we must predict.

→x=data_set.iloc[:,[2,3]].values

→y=data_set.iloc[:,4].values

5.TRAINING THE DATA SPLIT

Using data that was not used to train the model, the train-test split method can be used to assess how well machine learning algorithms perform when making predictions on unrelated data. You can evaluate the benefits of machine learning techniques for your particular predictive analytic challenge using the procedure' quickness and ease of completion. Although the method is simple to use and comprehend, there are several situations in which it shouldn't be applied, such as when the dataset is small and more setup is required, or when it is used for classification and the dataset is unbalanced.

```
from sklearn.model_selection import
train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,
y, test_size= 0.02,random_state=20)
from sklearn.preprocessing import
StandardScaler
st_x= StandardScaler()
x_train= st_x.fit_transform(x_train)
x_test= st_x.transform(x_test)
```

D. DATA VISUALIZATION

We may obtain a clear representation of our data via data visualization. The human mind examines and comprehends any given data more easily when it is presented with images, maps, and graphs. Both small and big data sets benefit from data visualization, but large data sets are where it really excels since it is difficult to manually view, let alone analyze, and comprehend, all of our data.

```
→plt.figure(figsize=(15,7),dpi=100
sns.counrplot(x='Sex',hue='Class',data=df2,palette= 'Set1')

→plt.figure(figsize=(15,7),dpi=100
sns.counrplot(x='Age/_Mons'
,hue='Class',data=df2,palette='Set1')

→plt.figure(figsize=(15,7),dpi=100
sns.counrplot(x='Family_mem_with_ASD
',hue='Class',data=df2,palette='Set1')
```

Sex, age, and the number of family members with ASD have been plotted on a bar graph against the overall number of individuals with autism.

E. DEVELOPING AND EVALUATING THE PREDCITION MODEL

Algorithms had been developed, and their accuracy had been evaluated, to generate predictions of autism traits. Ada boost and Random Forest were discovered to be highly feasible and to be more accurate than the other algorithms after achieving results from different supervised learning methods, such as Logistic Regression. Therefore, all three methods were suggested for use in the ASD predictive system implementation. To get even better outcomes, the algorithms required more improvements.

1. ADABOOST ALGORITHM:

Ada Boost, often referred to as adaptive boosting, is a machine learning technique applied in an ensemble setting. They are the single-level decision trees or decision trees with only one split. Another pseudonym for these trees is Decision Stumps. This is how Ada Boost functions When the random forest is employed, the method creates a 'n' number of trees. It creates appropriate trees with a start node and many leaf nodes. Although some trees may be larger than others, a random forest has no defined depth. However, AdaBoost's approach only creates the Stump node, which has two leaves.

Decision trees with a single level are the most popular and useful approach to employ with AdaBoost. These trees accomplish the task of making a single category selection since they are too short and only have one level. Because of this, we commonly refer to these trees as decision stumps. Each entity in our training dataset has been appropriately weighted.

We begin these entities' weights by assigning weight(x)=1/n.

We prepared the base learners on the training data using weighted samples. The decision trees in this approach only support binary classification because they have a single level. This implies that the decision stump only decides once about the input variable based on the data that we provide it. This decision tree produces a first-or second-class value output of either +1 or -1.

After each model iteration, we calculate the misclassification rate using the following formula:

M=(correct-N)/N

Here, properly is the quantity of training sessions that the model correctly predicts, and N refers the overall number of training samples. M is the miscalculation rate. Later, the mentioned equation was changed to make it work with the weighted training samples.

M= sum(w(i)* terror(i))/ sum (w)

Here, terror stands for training prediction error, and w is the weight of training sample i. We determine the stage value for the training set to give any predictions the model is capable of making a weight. For the training data, we can determine the stage value as

S= ln((1-M)/M)

Here, ln stands for the natural logarithm, M is the misclassification error, and S is the stage value we used to evaluate model predictions. Using the stage has the benefit of ensuring that the learners with higher weights contribute more to the final forecast The model adjusts the weights of the data points once the initial dataset has been iterated, giving greater weights to instances that were mistakenly predicted and smaller weights to examples that were properly predicted.

By utilizing the formula, we update these weights.

w = w *exp(S*terror)

Here, S stands for the classification error or misclassification rate for base learners, w stands for the weight for a single training sample, exp() is the numerical constant, terror is for the measure of an error made by the model in generating the prediction. fear equals 1 in all other situations and 0 if (y=p). In this case, y is the model's output variable and p is its prediction.

```
fromsklearn.ensemble import AdaBoostClassifier
fromsklearn.metrics import
classification_report,confusion_matrix,accuracy_score
x_train,x_test,y_train,y_test= train_test_split(x, y,test_size=
0.02,random_state=20)
ada=AdaBoostClassifier(n_estimators=60)
ada.fit(x_train,y_train)
prediction_ada=ada.predict(x_test)
cc_ada=accuracy_score(y_true=y_test,y_pred=prediction_ada)
print("Overall accuracy of AdaBoost Classifier using the test-
set is: %f" %(acc_ada*100))
print(classification_report(y_test,prediction_ada))
```

2. RANDOM FOREST ALGORITHM:

ML algorithm to recommend Random Forest is part of the supervised learning methodology. It can be used to solve kernel - based issues in learning algorithms. It is built on the idea of ensemble learning, a technique for merging lots of classifiers to solve complicated issues and enhance model performance. Overfitting is avoided and accuracy is increased due to the larger number of trees in the forest. As instead of focusing exclusively on one decision tree, the random forest makes use of forecasts from all of the trees and bases its prediction of the final outcome on the predictions that garnered the most votes. In order to boost its predicted accuracy, the classifier Random Forest averages data from multiple decision trees applied to various subsets of the input dataset.

As the model grow the trees, random forest introduces more randomness. The best feature within a randomly selected collection of features is sought for when dividing a node as opposed to the most crucial feature. In general, a better model is produced as a result of the great variation this causes.

```
from sklearn.ensemble import
RandomForestClassifier
classifier= RandomForestClassifier(n_estimators=
60, criterion="entropy")
classifier.fit(x_train,y_train)
fromsklearn.metricsimportclassification_report
,confusion_matrix,accuracy_score
predictions_rf= classifier.predict(x_test)
acc_rf=accuracy_score(y_true=y_test,y_
pred=predictions_rf)
print("Overall accuracyofRandom ForestClassifier
using the test-set is : %f" %(acc_rf*100))
print(classification_report(y_test,predictions_rf))
```

3. LOGISTIC REGRESSION:

One predictor binary variable and one or more outcome variables—which may be basic, integer, or ratio level parameters are connected. in a logistic regression. Logistic regression is used to create discrete results. When a dependent variable is continuous, the output can be predicted using logistic regression. Therefore, the result needs to have a discrete or qualitative value. It offers the stochastic numbers between zero and one rather than the precise values between zero and one. It can be either True or False, 0 or 1, or Yes or No.

When classifying observations using various sources of data, logistic regression can be used to quickly identify the factors that will work well.

As a cost function, the "sigmoid function" or "logistic function" is used in logistic regression. The sigmoid function can therefore be used to forecast probability values.

The sigmoid function's mathematical equation, which is shown below.

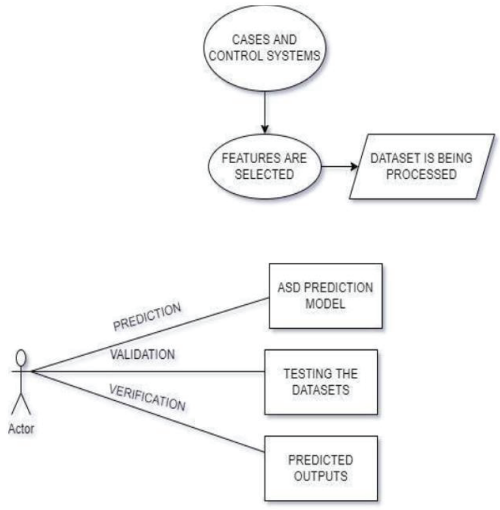
$$\mathbf{F(z)=1/(1-}e^{-z}\mathbf{)}$$

The co-efficient of the model, which is determined via maximum likelihood estimation and the attributes are being used to describe the co-efficient. The calculation of the binary outcome probability is the last phase of the equation, when the probabilities are split into two groups in accordance with the given data point (x). The predictor variable is the variable for the target class that we will make a prediction for. The outcome variables, on the other hand, are the qualities or attributes we'll use to forecast the target class.

```
from sklearn.preprocessing import StandardScaler
st_x= StandardScaler()
x_train= st_x.fit_transform(x_train)
x_test= st_x.transform(x_test)
from sklearn.linear_model import LogisticRegression
logr= LogisticRegression(random_state=60)
logr.fit(x_train,y_train)
x_train, x_test, y_train, y_test= train_test_split(x, y,
test_size= 0.02,random_state=20)
prediction_logr=logr.predict(x_test)
acc_logr=accuracy_score(y_true=y_test,y_pred=prediction
_logr) print("Overall accuracy of Logistic Regression
Classifier using the test-set is : %f" %(acc_logr*100))
print(classification_report(y_test,prediction_logr))
```

F. DEVELOPING THE WEB APPLICATION

Using algorithms like Ada Boost, Random Forest, and Logistic Regression, a web application is developed to detect the presence of autism spectrum disorder based on a variety of factors using the various parameters that the user entered in the front end



V.DISCUSSIONS AND CONCLUSIONS

Three things emerge out of this research: first, a model was created to predict the features associated with autism. The suggested approach can detect autism with 95% accuracy using the AQ-10 dataset. Additionally, the suggested model has a feature that many other existing techniques lack: the ability to predict autism features for distinct age groups The second aspect of this study is performance comparison of several techniques for machine learning.

The Random Forest, Ada Boost and the Logistic Regression Algorithm have accuracy percentages of 91.36,93.68 and 95.89, respectively. According to the facts, the Ada boost algorithm performed better in respect to the random forest algorithm, whereas as compared with other methods, the specified logistic regression algorithm outperforms the Random Forest and the Ada boost algorithm.

This result demonstrated an extension of many other current works because the majority of the previous studies largely focused on developing and assessing the effectiveness of prediction models or approaches rather than investing in creating web applications for end users.

In conclusion, the findings of this study offer a practical method for identifying autism features in individuals of various ages. Since identifying the characteristics of autism is a time-consuming and expensive process, diagnosing autism in children and teenagers is frequently put off Using a screening tool for autism, a person can obtain help early on to stop the issue from getting worse and cut down on the expenditures related to a delayed diagonsis

REFERENCES

[1] L Zwaigenbaum, Susan Bryson, and Nancy Garon, "Early identification of autism spectrum disorders," Behavioral Brain Research, vol. 251, pp. 133–146, Aug. 2013.

[2] Libero, L.E., et al., Multimodal neuro imaging based classification of autism spectrum disorder using anatomical, neurochemical, and white matter correlates. Cortex, 2015. 66: p. 46-59 [7] "autismspeaks.org/what-autism"

[3] Pag nozzi, A.M., et al., A systematic review of structural MRI biomarkers in autism spectrum disorder: A machine learning perspective. International Journal of Developmental Neuroscience, 2018. 71: p. 68-82

[4] Sushma Rani Dutta, Soumya jit Giri, Sujoy Datta, Mani deep Roy - A Machine Learning-based Method for Autism Diagnosis Assistance In children- Machine Learning Algorithm-MID-IEEE CONFERENCE 2017- vol. 2, no. 3, pp 18–22

[5] B.van den BK,"Using machine learning for detection of autism spectrum disorder,"2017

[6] Guo, X., et al., Diagnosing Autism Spectrum Disorder from Brain Resting-State Functional Connectivity Patterns Using a Deep Neural Network with a Novel Feature Selection Method. Front Neuro sci, 2017. 11: p. 460.