Detection of Autism Spectrum Disorder Using Machine Learning

BY

Md. Shaimum Hasan Sagor
ID: 191-15-2708

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering.

Supervised By

Md. Abbas Ali Khan
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Shah Md. Tanvir Siddiquee
Assistant professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JULY 2023

APPROVAL

This Project/internship titled **"Detection of Autism Spectrum Disorder Using Machine Learning"**, submitted by Md. Shaimum Hasan Sagor, ID: 191-15-2708 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 30/07/2023.

BOARD OF EXAMINERS

Chairman

Dr. S.M Aminul Haque (SMAH)
Associate Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner 1

Nazmun Nessa Moon (NNM)
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner 2

Mr. Md. Ali Hossain (MAH)
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

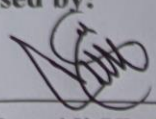
External Examiner 1

Dr. Md. Arshad Ali (DAA)
Professor
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science & Technology University

DECLARATION

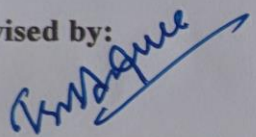
We hereby declare that, this project has been done by us under the supervision of **Md. Abbas Ali Khan, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



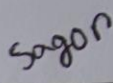
Md. Abbas Ali Khan
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:



Shah Md. Tanvir Siddiquee
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:



Md. Shaimum Hasan Sagor
ID: 191-15-2708
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Md. Abbas Ali Khan Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Field name*" to carry out this project. His endless patience ,scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Dr. Touhid Bhuiyan, Head**,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Autism Spectrum disorder (ASD) is a group of neurological disorders. Autism is often considered a spectrum condition due to its variability. Autistic people have some of these traits, but their symptoms vary. This study processed an ASD dataset and compared machine learning methods. Feature impact was used to identify ASD-causing features in the dataset. The study found that calibrated machine learning algorithms can detect early ASD. Autism is a spectrum of varying symptoms and assistance needs. Though subtypes are debated, diagnostic manuals now list ASD as a single diagnosis. This study used various ML algorithms. Considering hyper parameter tuning and 3-fold cross validation in this study Logistic Regression were the most accurate methods with accuracy 100%.

Keyword

Autistic, Logistic Regression, Random Forest, Decision Tree, XGBoost, Naive Bayes, K-Nearest Neighbors, GirdsearchCV.

TABLE OF CONTENTS

CONTENTS	PAGES
Board of Examiners Bookmark not defined.	Error!
DECLARATION Bookmark not defined.	Error!
ACKNOWLEDGEMENT	**iii**
ABSTRACT	**iv**
CHAPTER 1	**1-3**
1.1 Overview	1
1.2 Motivation	2
1.3 Objectives	3
1.4 Report Format	3
CHAPTER 2	**4-5**
2.1 Related Works	4
CHAPTER 3	**6**
3.1 Overview	6
3.2 Implementation Analysis and Design of Our Proposed System	6
3.3 Dataset Description	7
3.4 Pre-Processing	9
3.4.1 Numerical Conversion from Categorical Data	9
3.4.2 Data Cleaning	9
3.4.3 Handling Imbalanced Data	9
3.4.4 Feature Scaling	10
3.4.5 Feature Impact Analysis	10
3.4.6 Pearson Correlation	11
3.5 Supervised Machine Learning Algorithms	13
3.5.1 Logistic Regression	13
3.5.2 Random Forest	13
3.5.3 XGBoost Classifier	13
3.5.4 K Nearest Neighbor	14
3.5.5 Decision Tree	14
3.5.6 Gaussian Naive Bayes	14

3.6 Cross Validation 15
3.7 Hyper Parameter Tuning 15
3.8 Performance Evaluation 16
3.8.1 Confusion Matrix 16
3.8.2 AUC Score & ROC CURVE 18
CHAPTER 4 19-30
4.1 Experimental Results 19
CHAPTER 5 31
5.1 Impact on Society 31
5.2 Ethical Aspects 31
CHAPTER 6 32
6.1 Conclusion 32
REFERENCE 33
APPENDIX 34

LIST OF TABLES

TABLE	PAGE NO
Table 3.1 Detailed overview of dataset	8
Table 3.8.1: Confusion matrix visualization	16
Table 4.1.1 Accuracy, F-1 Score, Precision & Recall Score	19
Table 4.1.2: performance measurement of ML Algorithms	20-21
Table 4.1.3: AUC and Cross Validation Score	21
Table 4.1.4: Hyper Parameter Tuning (Best parameter & Score)	30

LIST OF FIGURES

FIGURE	PAGE NO
Figure 3.1 Workflow Diagram of the research	7
Figure 3.4.5: Feature Importance Chart	10
Figure 3.4.6: Pearson Correlation Heatmap	12
Figure 4.1.1: Roc Curve Analysis of Logistic Regression	20
Figure 4.1.2 : Confusion Matrix and Classification report Logistic Regression	22
Figure 4.1.3: Confusion Matrix and Classification of Random Forest	23
Figure 4.1.4: Confusion Matrix and Classification of XGBoost	24
Figure 4.1.5: Confusion Matrix and Classification of K-Nearest Neighbor	25
Figure 4.1.6: Confusion Matrix and Classification of Decision Ttree	26
Figure 4.1.7: Confusion Matrix and Classification of Naive Bayes	27
Figure 4.1.8: ROC Curve Diagram	28
Figure 4.1.9: AUC Curve Graph	29

CHAPTER 1

INTRODUCTION

1.1 Overview

Autism is a complicated Neurodevelopmental Condition. Differences in social skills, communication, and behavior are some of the signs of this condition[1]. People on the autism spectrum have difficulty in social situations and communicating their thoughts and feelings. ASD is a subtype of Autism. The distinctive features of autism can be recognized within the age of 18 and 24 months. The World Health Organization reports that ASD affects 0.63 per 1,000 children worldwide [2]. One in 36 American children will be considered to have ASD by 2023, according to data collected in 2020 by the CDC[3]. Autism spectrum disorder diagnoses occur in boys at a rate four times higher than in girls. Although autism can be precisely identified at the age of 2, most children weren't treated until after age 4. Children on the autism spectrum tend to have lower IQs. The average IQ of children with ASD is below 85 in 44% of cases [4]. It does not discriminate in terms of race or socioeconomic status. In general, minority groups are less likely to receive timely diagnoses [5]. An early detection can help reduce the patient's suffering, even though there is currently no known treatment for this illness. Scientists have assumed that human genes are to blame for ASD, but they have yet to pinpoint the disorder's precise origins. The human genome plays a role in shaping the environment in which a person grows. Low birth weight, having an autistic sibling, being the child of elderly parents, etc., are all risk factors for autism spectrum disorder. Four types of autism and their syndrome details:

A. Asperger’s syndrome

In the spectrum of neurodevelopmental disorders that includes autism, one can find Asperger's syndrome or Asperger's disorder. It is characterized by restricted and repetitive patterns of behavior and desires, also have difficulties in communication. which can make it hard for them to form and maintain relationships [6]. They tend to have narrow, focused interests and may engage in ritualistic or habitual actions as a result. However, many people with Asperger's syndrome are bright and even exceptional in some areas, like math or science. Individuals with Asperger's syndrome have the potential to live satisfying lives and make significant contributions to society with the right level of support and acceptance [7].

B. Childhood disintegrative disorder

It is a development disorder, which affects speaking and late reaction. At a certain period of typical maturation, it usually manifests between the ages of three and ten. On a rate ratio boys developed more CDD from girls; the male-to-female ratio is 9:1. Regression in toilet training, language, social skills, and motor abilities are just some of the difficulties faced by children with CDD [8].

C. Pervasive development disorder

Syndromes with compromised relationships, challenges with communication, and restrictive patterns of behavior are all included under the umbrella term pervasive developmental disorder (PDD). Asperger's syndrome and autism spectrum disorder (ASD) are two examples of PDD [9]. These conditions typically manifest in early childhood and have far-reaching consequences for the individual at every stage of growth. People with PDD frequently have issues in relationships, such as misunderstanding and late response to action. They are likely born with an attraction for habit and an attraction toward area of expertise. Individuals with PDD can develop symptoms from mild to severe. With proper help at the right time, people with PDD can realize all that they can do and live better lives. [10].

D. Rett syndrome

Rett Syndrome is a form of female-predominant rare genetic disorder. Normal infant growth and development is followed by a regression and the emergence of diagnostic symptoms [11]. These signs and symptoms include the inability to perform tasks that once required the use of both hands, the development of repetitive hand movements, severe motor impairments, and the inability to express oneself verbally. Abnormal breathing, seizures, and mental impairments are additional symptoms. In most cases of Rett Syndrome, mutations in the MECP2 gene are to blame because of the central role it plays in brain development [12].

1.2 Motivation

Changes in the functioning of the brain are at the core of autism spectrum disorder (ASD), an intellectual disability. People on the autism spectrum tend to have limited interests and routines and may have trouble communicating with others. Individuals with Asperger's may also exhibit peculiar patterns of behavior, learning, and focus. It should be noted that these symptoms may also be displayed by people who do not have autism. However, these behaviors can be difficult for those have autism. Relationships

and interacting with others are the areas where people with the autism may struggle. There are scant academic papers devoted to preemptive forecasting. At that time, reducing the effects of ASD is extremely challenging. The disorder has no known treatment yet, but it can be maintained better if recognized early. However, most of our people are unable to afford a clinical diagnosis due to its costly nature. From here, I hope to develop a ML model which will recognize ASD at an early age. It will be cost effective for all sort of people.

1.3 Objectives

- To promote public awareness of autistic individuals.
- To produce improved outcomes compared to other related works.
- The system allows patients to detect ASD at a lower cost.
- Identifying autistic individuals to provide them with special care.
- Early ASD detection is possible with the proposed system. A machine learning method will maximize efficiency and accuracy.

1.4 Report Format

Background Study

Research Methodology

Experimental Results & Discussion

Impact on Society & Ethical Aspects

Conclusion & Future Work

CHAPTER 2

BACKGROUND STUDY

2.1 Related Works

Many studies have used machine learning to find better and faster methods of ASD diagnosis. In a study (Raj & Masood, 2020), This paper makes an effort to investigate the applicability of Svm, K-NN, LR, NB in the prediction and analysis of ASD issues across the lifespan, from infancy to old age. The methods proposed here are tested on three publicly available, non-clinical ASD datasets. The first dataset on children with autism disorder screening consists of 292 cases and 21 pieces of information. There are 650 unique adults and 21 distinct characteristics. The adolescent autism spectrum disorder (ASD) screening dataset has 104 instances and 21 attributes.

(Erkan, Thanh, 2019) Erkan, U. and Thanh, D.N propose classifying ASD data to aid early diagnosis. Three ASD datasets cover children, adolescents, and adults. ASD data was categorized using KNN, SVM, and RF. Training and test sets were randomly selected. 100 random data selections were made to test classification techniques. Average values determined results. SVM and RF classify ASD well. RF classified all datasets with 100% accuracy. Recent research has focused on data collection, clinical diagnosis, and brain imaging, according to the study. This study may not add to ASD detection knowledge.

(Rahman et al., 2020), This research recommended establishing a rapid diagnostic procedure for ASD as a means of facilitating early identification. Dataset that was collected from UCI will be preprocessed and classified for this research. Both demographic data and ASD screening questions were collected in these databases. Information will be converted to a numerical format for easier processing. The researchers will employ a gap-filling strategy to address the absence of certain information. Implementing classification methods can help clinicians make more accurate, timely, and straightforward diagnoses of ASD based on the findings.

(Hossain et al., 2021). The goal of this work is to streamline the diagnostic process by identifying the most important characteristics and automating them using existing classification methods. The best classifier and feature set were determined by comparing them along with powerful classifier. The multilayer MLP classifier beats all standard classification algorithms and achieves a remarkable score with a minimum total of attributes for the four datasets, following experimental results. In addition, they

find that among all four ASD datasets, the "relief F" approach to figure out which features the most crucial ranks highest.

(Vakadkar et al., 2021). This research seeks to improve medical diagnosis efficiency and precision. Machine learning is supplementing traditional methods. Dr. Fadi Thabtah's dataset had various attributes. Predictive models were created using different algorithms. Other symptoms add to ASD evaluations. The logistic regression model was most precise. This research suffers from a lack of publicly available, comprehensive autism spectrum disorder datasets. (Abdullah, 2019) The most important features for three supervised ML algorithms KNN, LR, RF cross validation were selected using Chi-square and LASSO. Logistic Regression, which used a model with 13 Chi-square-selected features, was found to be effective.

(Thabtah 2019) This work provides a critical review of recent studies on autism and makes suggestions for enhancing the application of machine learning to ASD, both conceptually and in terms of implementation and data. Metrics like FPN rates, TN rates, time to process, and accuracy are used for evaluation. These methods for evaluation are frequently seen in machine learning programs. Prediction algorithms for autism datasets have been applied using a variety of statistical methods. The research ignores data limitations and biases. The publication mentions parental or caretaker responses, but it does not discuss bias or measurement error from subjective reports.

(Thangaiyan, 2020) This study aims to improve Autism prediction. Adult and Child Autism Spectrum Disorder Screening Data from UCI was utilized for this study. The detection and management of ASD was sped up and made better by using different algorithms. From the data, DT and standard LR execute the best. Data errors and constraints are not taken into account. It does not address the potential of bias or errors in measurement due to subjective reports, despite acknowledging the use of the parent responses.

(Chowdhury and Iraj 2020) This paper seeks the best early ASD detection method. This paper illustrates classifier measures. SVM generates the best results, but it requires specific kernels. GRK result in the best results. The paper doesn't discuss dataset contributors' demographics. The study's findings won't apply to all age, sex, socioeconomic, and cultural groups.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Overview

The dataset which was used in this study was collected from the UCI database. The current approach involves several components, including data collection and preprocessing. The current approach encompasses various components, including data collection and preprocessing. This research applies a total of six algorithms Logistic Regression, Random Forest, XGBoost, K-Nearest Neighbor, Decision Tree, Gaussian Naive Bayes. Following the initial step, a hyperparameter tweaking approach is implemented also 3-fold cross-validation was used. Accuracy was improved using hyperparameter tuning. Logistic Regression performed well and achieved 100% accuracy. Figure 3.1 outlines all of the phases of the study.

3.2 Implementation Analysis and Design of Our Proposed System

An elaborate diagram depicting the usual procedure to be followed is drawn at the outset. The diagram clearly shows that the first step in our methodology is to gather information. UCI Machine Learning repository dataset. After obtaining the dataset, we preprocessed it to clean it up and make it work better for our purposes. The next stage was selecting features. Feature selection is a method for narrowing down the input variables in a model to those that are most useful. Then, we used a wide variety of machine learning classification methods until we found one with the highest accuracy. Tuning hyper-parameters is the next step. The final result is reached after cross validation and performance evaluation.

In The figure 3.1 provided illustrates the comprehensive workflow and individual steps involved in the process. The dataset is loaded, and the data is preprocessed as the initial step. Machine learning algorithms were then applied. The objective of this analysis is to conduct a performance evaluation and determine the feature importance of a given dataset. Additionally, we will perform hyperparameter tuning and cross-validation to optimize the model's performance. Finally, we have the outcome.

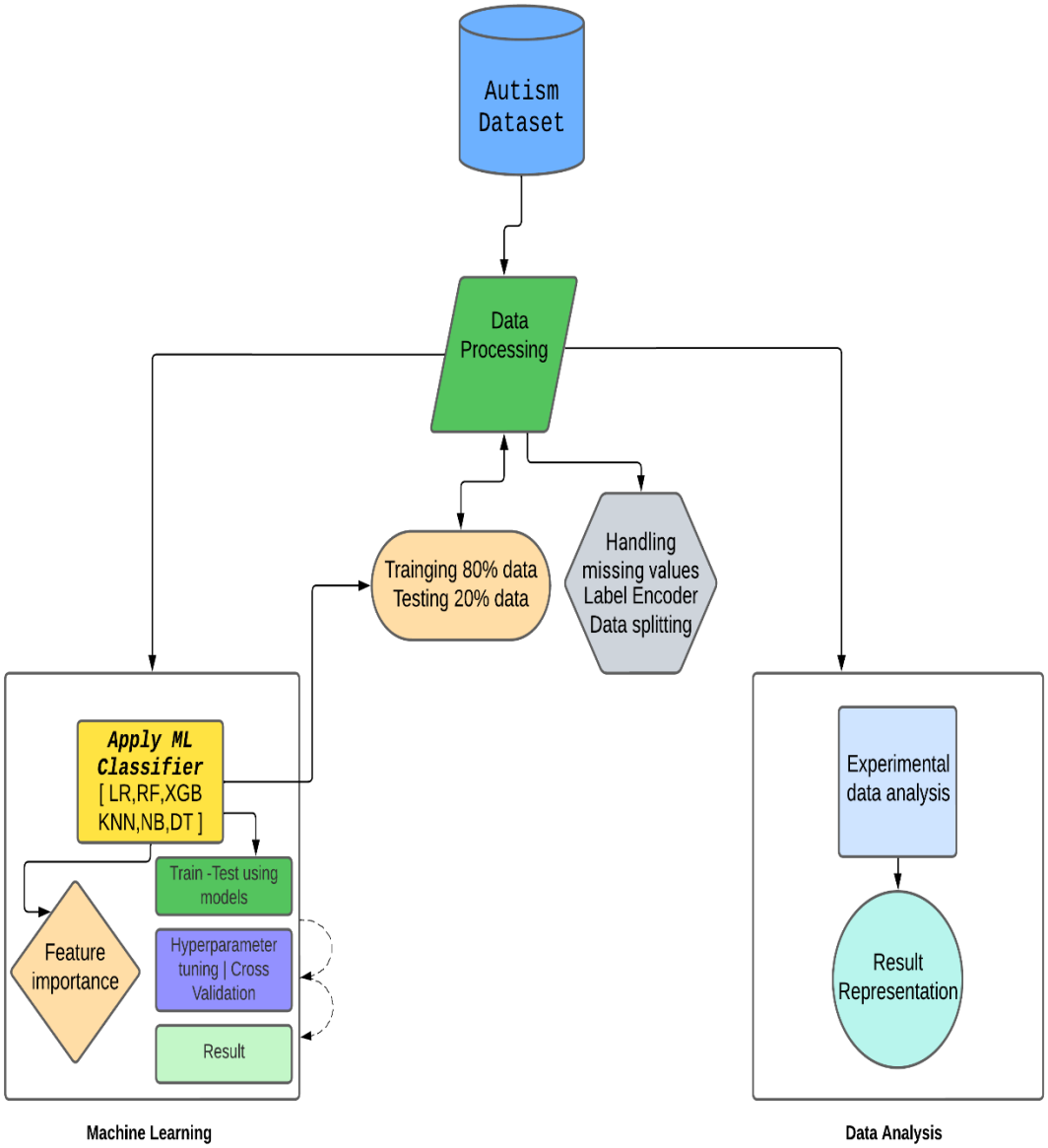


Figure 3.1 Process Diagram of the research

3.3 Dataset Description

The study employs the UCI Machine-Learning Repository's Preliminary phase Autism prediction dataset. This dataset contains a total of 704 samples representing 21 distinct characteristics. The following properties were assigned to each of the 21 characters that were included in this dataset: id, bornwithjundice, familymemberaustim, age, race, A1 to A10 score, usedscreeningbefore etc.

Table 3.1: Detailed overview of dataset

Symbol	Description	Type
Id	sequence	integer
gender	male/female	nominal
race	List of common races in texts	nominal
Bornwithjaundice	Whether people born with jaundice or not	nominal
familymemberwithautism	Any member of family has already autism or not	nominal
countryofresident	List of countries in text format	nominal
usedscreeningbefore	If the user has used screening before	nominal
whoistakingthetest	Self, parent, relative, others, health care professionals	nominal
A1	integer	nominal
A2	integer	nominal
A3	integer	nominal
A4	integer	nominal
A5	integer	nominal
A6	integer	nominal
A7	integer	nominal
A8	integer	nominal
A9	integer	nominal
A10	integer	nominal
HasAutism	People with autism (Yes/NO)	nominal

People's characteristics that may be relevant to autism and autism screening are listed in table 3.1 . It consists of nominal categories representing demographic information such as gender, race, and country of residence. Information about the person's medical background, such as whether or not they were born with jaundice or have a history of autism in their family, is also included.

3.4 Pre-Processing

Here we used various methods including handling missing values, encoding categorical values using label encoding, feature scaling, train-test split, modeling with logistic regression, random forest, XGBoost, Knn, decision tree and naive bayes.

3.4.1 Numerical Conversion from Categorical Data

Label encoding was used to convert the categorical data into a numerical format. To make categorical data more usable as input for machine learning algorithms, label encoding is a common practice. LabelEncoder, found in the sklearn.preprocessing library, was specifically used in the code for the purpose of label encoding. For each distinct category in the categorical feature, the LabelEncoder generates a separate integer (label). If there are three distinct categories for a categorical feature, say A, B, and C, label encoding would convert these to the integers 0, 1, and 2. After all the necessary libraries have been imported, a new instance of a LabelEncoder has been created. Then, adjust the encoder so it best fits the categorized feature used in the training data. Finally Apply the fitted encoder to both the training data and the test data. After this procedure is complete, numerical labels will replace the categorical values in the designated column of the training and testing datasets. Machine learning algorithms can now utilize the data, as they prefer numeric to categorical input.

3.4.2 Data Cleaning

The dataset was checked for missing values using the isnull() function.sum(). The removal of duplicate rows is achieved by utilizing the drop_duplicates() function. Similarly, the elimination of rows containing missing values is accomplished by employing the dropna() function.

3.4.3 Handling Imbalanced Data

The Synthetic Minority Over-sampling Technique (SMOTE) handled imbalanced data. The dataset's class distribution (target values) is imbalanced. In this case, one class may have significantly fewer samples than the other, which can bias model performance, especially if the model cannot capture minority class patterns. SMOTE is used on the training set. SMOTE from imblearn.over_sampling is used for this. It creates synthetic samples for the minority class (HasAutism = 'Yes') to match the number of samples in the majority class (HasAutism = 'No').

3.4.4 Feature Scaling

After the data has been cleaned and prepared, feature scaling is done before the machine learning models are trained. Feature scaling is a method for normalizing the variation in a dataset's standalone features or variables. This is done to prevent some features from dominating the training process of the model because of their larger magnitude, and to ensure that all features contribute equally to the training process. To scale features to a uniform interval (typically 0–1), the provided code employs Min-Max scaling, which considers the minimum and maximum values of each feature.

3.4.5 Feature Impact Analysis

Feature impact is a concept used to determine which features of a dataset have the greatest bearing on the results produced by a machine learning model. To further enhance model precision, feature influence is also applied to the process of selecting features.

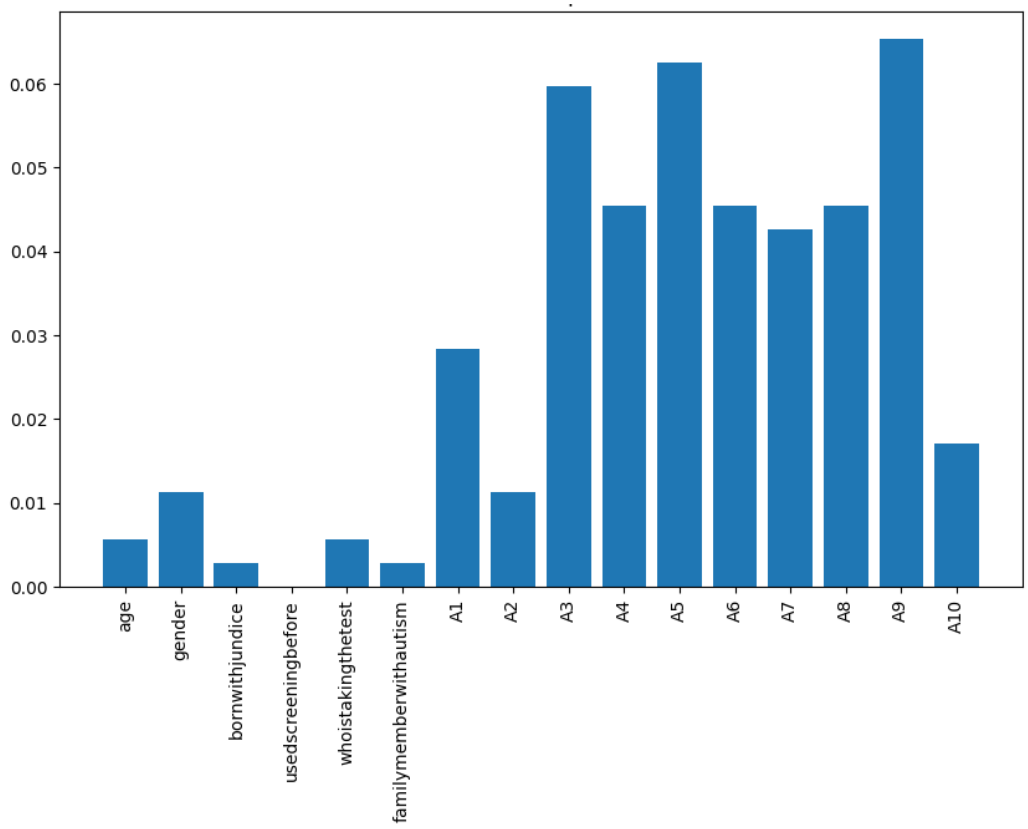


Figure 3.4.5: Feature Importance Chart

Figure 3.4.5 illustrates relative importance of features. How a model prioritizes its input features is described by their feature importance. The importance of each feature is only

partially captured by its score. An increased score indicates a more significant impact on the variable prediction model from this attribute.

3.4.6 Pearson Correlation

The Pearson correlation coefficient is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. The purpose of the Pearson correlation is to establish a line that accurately represents the observed relationship between two variables. According to the user's statement, the Pearson correlation coefficient (r) is used to measure the distance between each data point and the line of best fit. According to the analysis, when the correlation coefficient between two variables is -1, it indicates the presence of a perfect negative linear relationship between the variables. According to the observed relationship, when one variable experience an increase, it is expected that the other variable will undergo a corresponding decrease. According to the analysis, it has been determined that the variables under consideration do not exhibit a linear relationship. This conclusion is based on the fact that the correlation coefficient between the variables is found to be 0. The analysis suggests that a straightforward linear relationship between the two variables is not evident. According to the analysis, a correlation coefficient of 1 indicates a perfect linear relationship between two variables. Based on the analysis, it has been determined that there exists a proportional relationship between the two variables in question.

$$\mathbf{r} = \frac{\mathbf{n}(\sum \mathbf{xy}) - (\sum \mathbf{x})(\sum \mathbf{y})}{\sqrt{[\mathbf{n}\sum \mathbf{x^2} - (\sum \mathbf{x})^2][\mathbf{n}\sum \mathbf{y^2} - (\sum \mathbf{y})^2]}}$$

Where, r is Correlation Coefficient, n holds total of pairs of the stock, $\sum xy$ is total of stock pair-products, $\sum x$ is the sum score of x, $\sum y$ is total of Y, $\sum x^2$ is a tally of the x-score squared, $\sum y^2$ is y-squared scores.

As shown in figure 3.4.6, a positive correlation exists between a number of these risk factors and an individual being given a diagnosis of autism. This is illustrated by the graph. The graph demonstrates, for example, that the likelihood of receiving a diagnosis of autism rises with age, being male, and having a first-degree relative who also has the disorder. As the graph demonstrates, the likelihood of an individual being diagnosed

with autism is significantly reduced when they have previously participated in an autism screening.

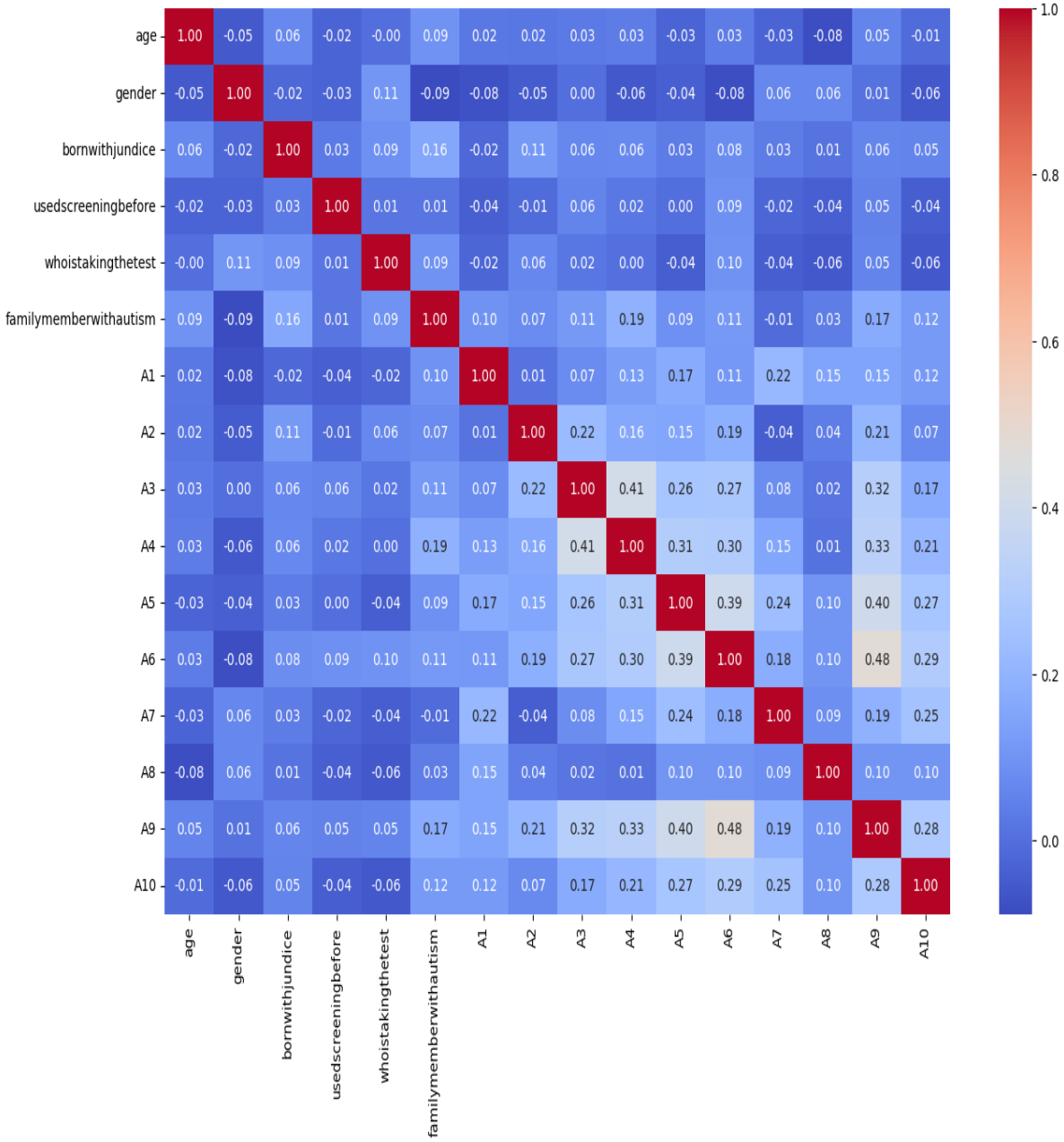


Figure 3.4.6: Pearson Correlation Heatmap

Figure 3.4.6 portraits the correlation between various risk factors and an autism diagnosis. The child's age, gender, whether or not they were born with jaundice, the presence of a family history of autism, prior screening for autism, and the presence of a whistle are all relevant factors.

3.5 Supervised Machine Learning Algorithms

The following classifiers were used in the analysis Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), Decision Tree (DT), K-Nearest Neighbors Classifier (KNNC), and XGBoost. The classifiers were configured using standard machine learning classifier parameters.

3.5.1 Logistic Regression

The Logistic Regression algorithm is a linear classification technique commonly employed for binary classification tasks. In such tasks, the desired outcome is a binary label, typically represented as either 0 or 1. The algorithm is instantiated using the LogisticRegression() function from the scikit-learn library. The model is trained on the training data using the fit method. In this process, the input features are represented by X_train, while the corresponding output labels are represented by Y_train. Following the completion of the training phase, the model is subsequently employed to generate predictions on the test data (X_test) by utilising the predict method. The resulting predicted labels are then stored in the variable Y_pred_lr.

3.5.2 Random Forest

The Random Forest method of ensemble learning is one that makes predictions through the utilisation of multiple decision trees. The Random Forest classifier is produced with the help of scikit-learn's RandomForestClassifier () function. Next, it is trained on the training data using the fit method, where X_train represents the input features and y_train represents the corresponding output labels. This is done so that the model can correctly classify new data.Following the completion of training, the model is put to use to make predictions on the test data (X_test) by employing the predict method, and the results of these predictions are saved in the y_pred_rf variable.

3.5.3 XGBoost Classifier

By iteratively combining multiple weak learners (decision trees), the XGBoost boosting algorithm produces a robust predictive model. Gradient descent optimization is used by XGB to strengthen an ensemble of decision trees, which are relatively poor predictive models. Through the use of parallel processing, tree pruning, and regularization, performance can be enhanced and overfitting can be mitigated. The high predictive accuracy and scalability of XGB have led to its widespread adoption. It is widely used

for regression and classification due to its adaptability, speed, and feature importance analysis. XGBClassifier() from scikit-learn is used to make the XGBoost classifier. Then, the fit method is used to train the model on the training data, where X_train is the set of input features and Y_train is the set of labels. Predictions are made on X_test using the trained model and the predict method the results are saved as Y_pred_xgb.

3.5.4 K Nearest Neighbor

The K-nearest neighbours (KNN) algorithm is an intuitive classification method that assigns data points to a specific class based on the majority class of their K nearest neighbours.The KNN classifier is instantiated using the KNeighborsClassifier() function from the scikit-learn library. The model is then trained on the training data using the fit method. In this process, the input features are represented by X_train, while the corresponding output labels are represented by y_train.

Following the completion of the training process, the model is subsequently employed to generate predictions on the test data, denoted as X_test. These predicted labels are then stored in the variable y_pred_knn. This is achieved by utilizing the predict method, which yields the predicted labels.

3.5.5 Decision Tree

The Decision Tree algorithm is a classification method that utilises a tree structure to make predictions by recursively dividing the data according to its features. The formation of a hierarchical structure of decisions is observed. The Decision Tree classifier is instantiated using the DecisionTreeClassifier() function from the scikit-learn library. The model is then trained on the provided training data using the fit method. In this process, the input features are represented by X_train, while the corresponding output labels are represented by y_train. Following the completion of the training phase, the model is subsequently employed to generate predictions on the test data (X_test) by utilising the predict method. The resulting predicted labels are then stored in the variable y_pred_dt.

3.5.6 Gaussian Naive Bayes

The Gaussian Naive Bayes algorithm is a Bayesian probability tool. It is widely used for classification tasks and works under the assumption that features are uncorrelated and normally distributed. Use the scikit-learn function GaussianNB() to make a Naive Bayes classifier with a Gaussian distribution. Then, the fit method is used to train the

model on the training data, where X_train is the set of input features and y_train is the set of labels. Predictions are made using the trained model and the predict method on the test data (X_test), with the results saved as labels in the y_pred_nb variable.

3.6 Cross Validation

Cross-validation is a commonly used technique to evaluate the performance of machine learning models in making predictions. The analysis reveals potential problems such as overfitting and selection bias. Additionally, insights can be obtained regarding the model's capacity to effectively apply its learnings to unseen data. The importance of cross-validation after model development cannot be overstated. Stratified K-fold cross-validation was employed in this study. The dataset was consistently divided into 3 equal folds for each iteration. The model was constructed by employing a total of nine convolutional layers. Subsequently, its performance was assessed by employing a single evaluation process. The accuracy of the model was evaluated by comparing it to its mean value after 3 iterations. The use of a stratified K-fold technique in this study ensures that the attribute classes of interest are evenly distributed across all folds.

3.7 Hyper Parameter Tuning

An integral part of controlling the behavior of machine learning models is adjusting their hyperparameters. Definition of the hyperparameter grid precedes tuning. To accomplish this, we employ a dictionary in which each key stands for a hyperparameter and each value is an array of candidates for that hyperparameter to evaluate. Initialization of the GridSearchCV object involves setting the machine learning model (SVC), the hyperparameter grid (param_grid), and the cross-validation technique (kfold). The accuracy metric will be used to score the model's performance, as accuracy is the value set for the scoring parameter. Data (in X and Y coordinates) are used to adjust the GridSearchCV object (grid_search). Fitting will be performed with cross-validation using the kfold provided, trying all possible combinations of the hyperparameters defined in the param_grid. The accuracy of each possible combination is considered the most accurate hyperparameters. Once the grid search is complete, you will have access to the optimal set of hyperparameters that yielded the highest accuracy. Rebuild the model using the preferred hyperparameters. Finally, you can use the model's best hyperparameters to re-create it. When using GridSearchCV for hyperparameter tuning, the code will iteratively explore all possible permutations of the

input hyperparameters before settling on the optimal one. This improves the model's ability to generalize to new data and thus optimizes its performance.

3.8 Performance Evaluation

All of the performance evaluation criteria and the corresponding mathematical equation are shown below. Both their definition and their functions are laid out for the reader. These factors were used as the basis for this study's evaluation of classification algorithms in order to find the optimal approach for detection of ASD.

3.8.1 Confusion Matrix

Confusion matrices are helpful for assessing a classification model's efficacy because they reveal more information about how well the model is classifying the positive and negative classes. Table 3.8.1 describes an example of confusion matrix in a two-way classification.

Table 3.8.1: Confusion matrix visualization

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

The confusion matrix shows the number of times the actual values were counted in comparison to the expected values. The output of "True Negatives," or "TN," is the total number of samples that were accurately identified as negative. The percentage of correct predictions is displayed. "FP" stands for "false positive," which indicates an incorrect diagnosis. H. The rate at which false-negative results were actually correct. The term "false negative" (or "FN") refers to the number of actually positive cases that were mislabeled as negative.

Accuracy

Accuracy refers to the measurement of the proportion of cases that have been correctly categorized. The calculation of accuracy is determined by employing the following equation:

$$Acc = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Where,

- True Positive (TP): The predicted outcome is positive while the main value is also positive.
- True Negative (TN): The predicted outcome is negative while the main value is negative.
- False Positive (FP): The predicted outcome is positive, but the actual value is negative.
- False Negative (FN): The predicted outcome is negative, but the actual value is positive.

Precision

The proportion of accurate predictions relative to the sum of accurate and incorrect ones is the main focus. When the model predicts a positive class, it is more likely to be correct if the precision value is high. Conversely, a low precision value indicates that the model incorrectly predicts positive instances, leading to a higher rate of false positives. The calculation of precision is determined by employing the following equation:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Where TP, FP represents (True Positive, False Positive).

Recall

Recall is a performance metric that evaluates how well a classification model does at correctly identifying positive examples. This means that if the dataset contains positive examples, the model is more likely to correctly identify those examples. However, if the model has a low recall value, it likely has a high rate of false negatives and is missing some important positive instances. The calculation of recall is determined by employing the following equation:

$$Recall = \frac{TP}{\text{TP} + \text{FN}}$$

Where TP, FN represents (True Positive, False Negative).

F-Measure

The F1 Score is determined by taking the weighted average of Precision and Recall. The F-measure is utilized to calculate the mean score of Recall and Precision. The score provided in the previous analysis accounts for both false positives and false negatives. The F-measure is often considered more valuable than accuracy, particularly in cases where the class distribution is challenging. The equation used for calculating precision is as follows:

$$F = \frac{2 * Precision * Recall}{\text{Precision} + \text{Recall}}$$

While accuracy is easier to understand intuitively, the F-measure provides a more comprehensive evaluation. According to the analysis, when the costs associated with false positives and false negatives are considered to be equal, the accuracy of the system is deemed to be excellent. In order to make an informed decision, it is crucial to consider both Precision and Recall metrics, particularly when the costs associated with false positives and false negatives differ significantly.

3.8.2 AUC Score & ROC CURVE

AUC Score

The AUC score is useful for assessing the efficacy of a binary classification model because it is insensitive to the relative abundance of the positive class. As a result, the AUC score can be used to evaluate models, regardless of how uncommon the positive class may be. By comparing the TPR and FPR across a selection of cutoffs, the AUC value can be determined. TPR measures how often a positive example is correctly classified, while FPR measures how often a negative example is misclassified.

ROC Curve

The ROC curve provides a visual representation of the accuracy of a binary classifier across a range of decision thresholds, allowing you to fine-tune the threshold to your application's needs. To evaluate the efficacy of a binary classification model, a receiver operating characteristic (ROC) curve is plotted. The false positive rate (FPR) and the true positive rate (TPR) are graphed against each other across a range of classification cutoffs. The (ROC) curve helps evaluate a classifier's recall and specificity.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Results

In total, there are 704 cases and 21 characteristics in the dataset used for this investigation. The label encoder was used to transform categorical information into numeric form. After the classifier was trained and tested, we compared its accuracy and cross-validation scores using the test data. Next, we used voted classifiers to train, assess, and test the classifiers with 3-Fold cross validation. For this purpose, we computed metrics such as reliability, accuracy, f1- score, recall, and area under the curve (AUC) to evaluate the models. The only way to positively label someone as autistic is if they display autistic characteristics. The opposite is true for those who avoid autism's symptoms throughout their lives. In this study we used 6 classifier algorithms. In the Table 4.1 we have displayed accuracy score along with

Table 4.1.1 Accuracy, F-1 Score, Precision & Recall Score

Serial	Model Classifier	Accuracy	F-1 score	Precision	Recall
1	Logistic Regression	100	1.00	1.00	1.00
2	Random Forest	99.43	0.99	1.00	0.99
3	K-Nearest Neighbor	94.86	0.94	0.94	0.93
4	XGBoost	99.43	0.99	0.99	0.99
5	Decision Tree	97.44	0.97	0.97	0.97
6	Naive Bayes	96.87	0.96	0.97	0.96

Precision, recall and F-1 score of the model used in this study. Considering the outcome we get the best results using Logistic Regression with 100% accuracy, precision, recall and f1 is 1.0.

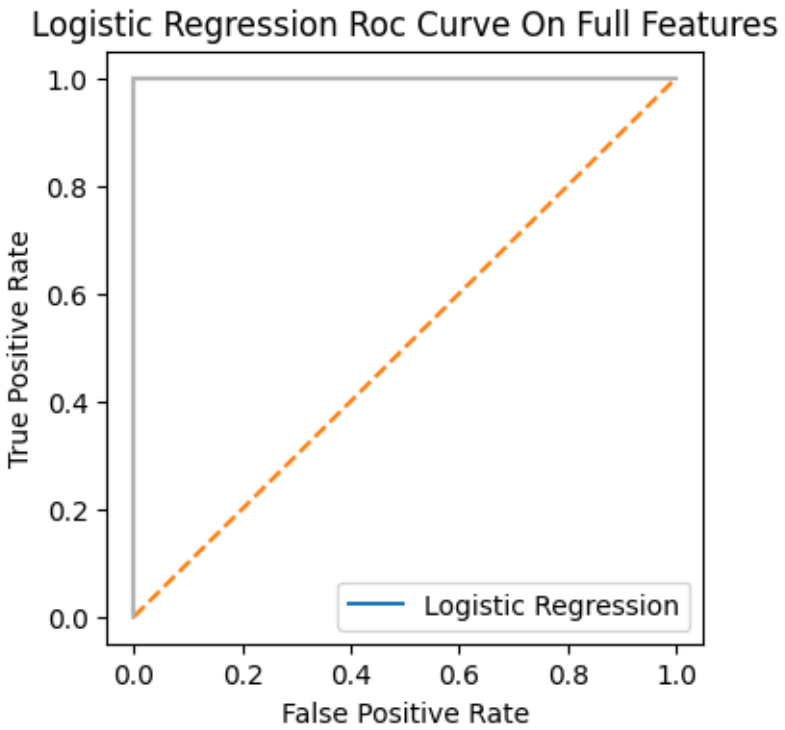


Figure 4.1.1: Roc Curve Analysis of Logistic Regression

Table 4.1.2 Displays some of the other important numbers regarding performance measurements of the classifiers used in this study. The table represents information like

Table 4.1.2: performance measurement of ML Algorithms

SN	Classifier	Testing Accuracy	Sensitivity	Specificity	False positive rate	False negative rate	Negative Predictive value	False discovery rate	Mean absolute error	Mean squared error	Roc Accu	Log loss	Cohen kappa score
1	Logistic Regression	100	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.09	0.99
2	Random Forest	99.43	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.14	0.89
3	XGBoost	99.43	0.972	0.971	0.028	0.027	0.990	0.078	0.028	0.028	0.971	0.09	0.89

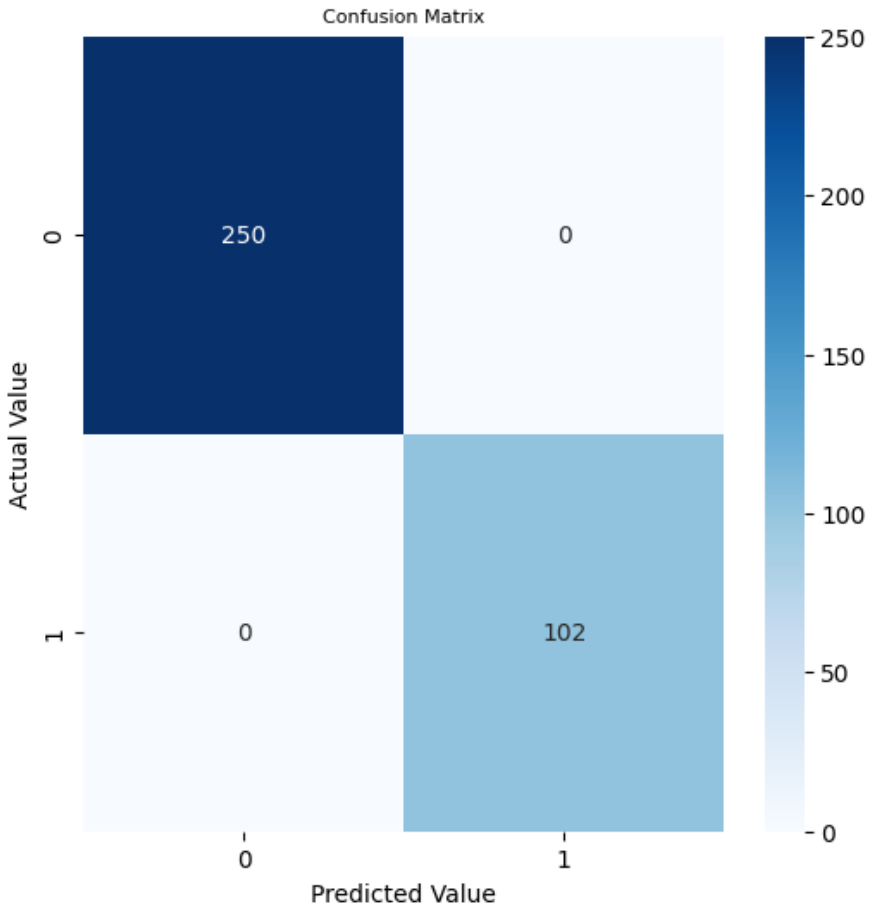
4	K Nearest Neighbor	94.86	0.722	0.971	0.028	0.277	0.910	0.103	0.092	0.091	0.846	0.54	0.66
5	Decision Tree	97.44	0.916	0.914	0.085	0.083	0.969	0.214	0.851	0.085	0.916	3.78	0.73
6	Naive Bayes	96.87	0.972	0.990	0.009	0.027	0.990	0.027	0.014	0.014	0.981	0.10	0.88

testing Accuracy, sensitivity, Specificity, False Positive Rate, False Negative, Negative Predictive Value, False Discovery rate, Mean Absolute Error, Mean Squared Error, Error, Log Loss, Cohen Kappa Scorer.

Table 4.1.3: AUC and Cross Validation Score

Serial	Classifier	AUC Score	Cross Validation
1	Logistic Regression	1.0	0.991
2	Random Forest	0.993	0.948
3	XGBoost	0.994	0.957
4	K Nearest Neighbor	0.933	0.897
5	Decision Tree	0.875	0.893
6	Naive Bayes	0.993	0.931

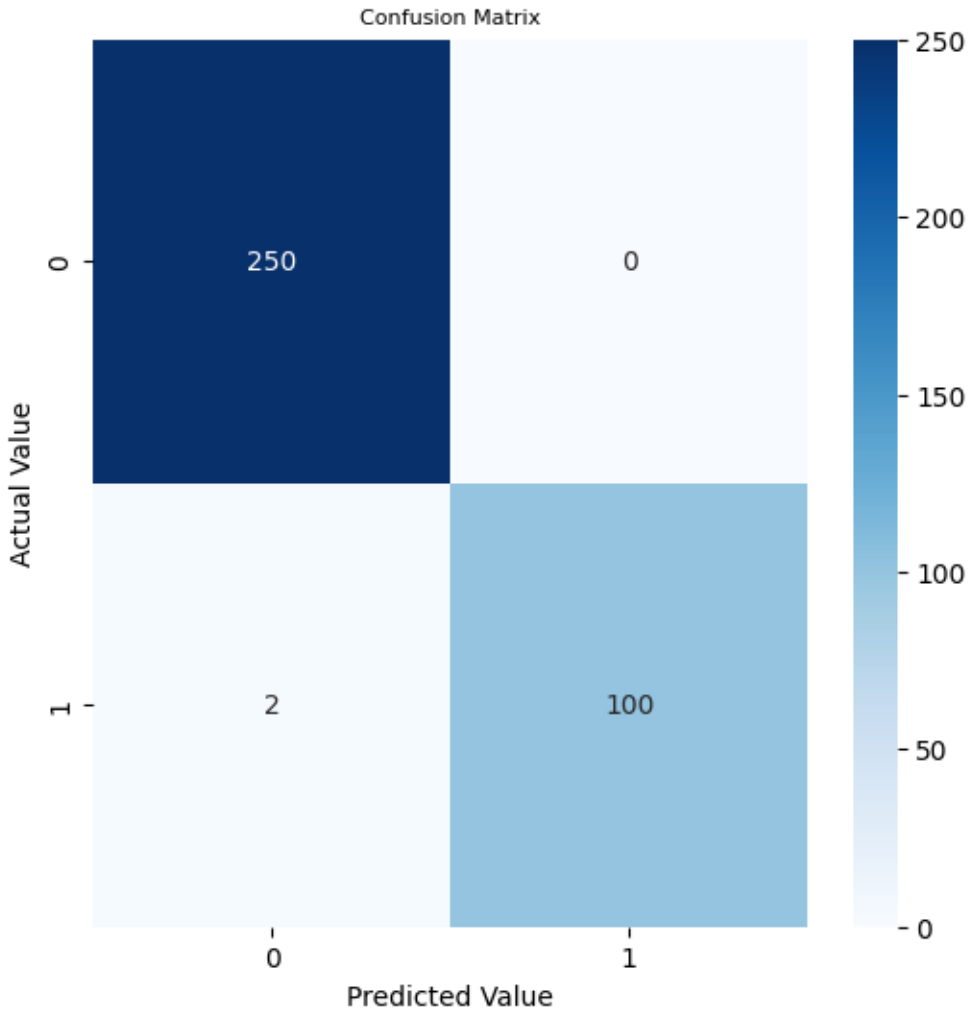
The AUC and the Cross Validation Scores for each classifier are shown in Table 4.1.3. AUC measures how different the results are from each other. Cross-validation is a way to figure out how well a machine learning model can predict new data. The AUC score for Logistic Regression is 1.0, and the Cross the score for validation is 0.991, which is the best. The decision tree with an AUC score of 0.875 and a Cross Validation score of 0.893 has the lowest number.



	precision	recall	f1-score	support
0	1.00	1.00	1.00	250
1	1.00	1.00	1.00	102
accuracy			1.00	352
macro avg	1.00	1.00	1.00	352
weighted avg	1.00	1.00	1.00	352

Figure 4.1.2: Confusion Matrix and Classification report Logistic Regression

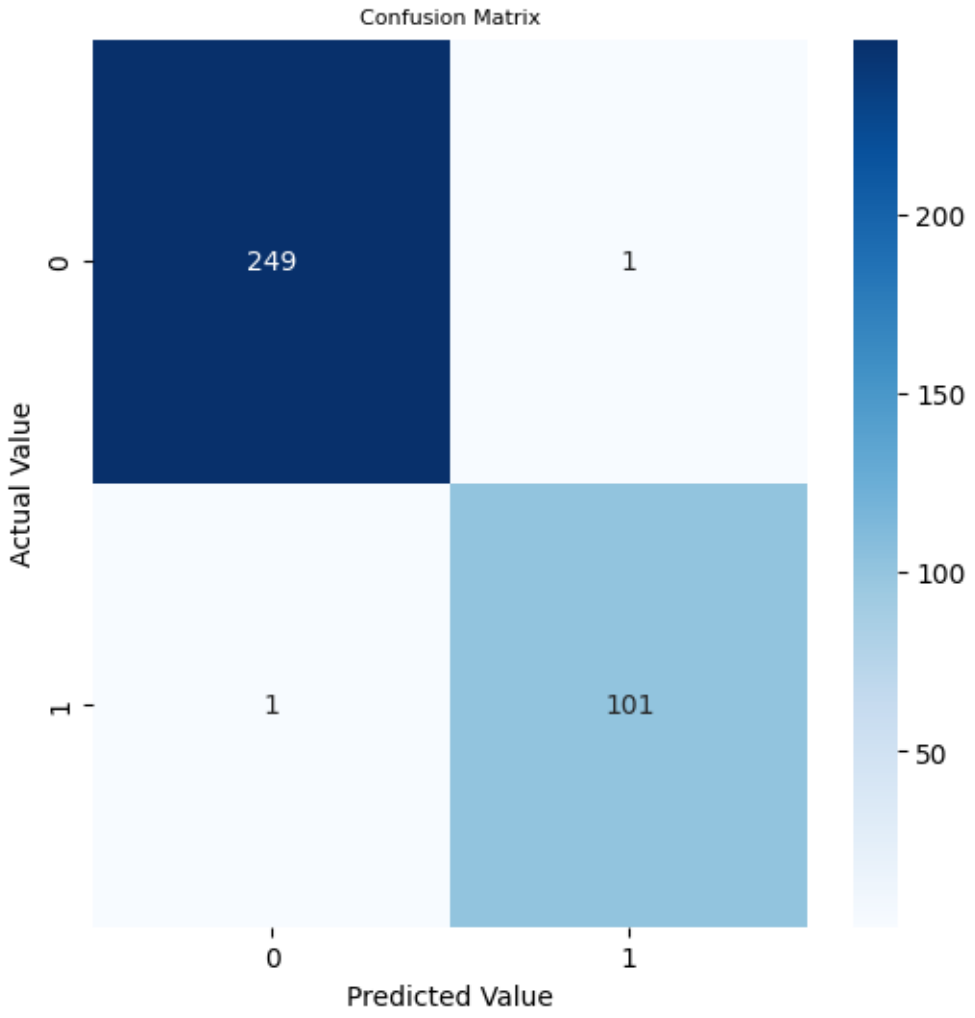
According to the Logistic Regression model's confusion matrix and classification report, the model's accuracy is excellent at 99%. The best performing model for this study has a f1-score close to 1.0, indicating its superiority.



	precision	recall	f1-score	support
0	0.99	1.00	1.00	250
1	1.00	0.98	0.99	102
accuracy			0.99	352
macro avg	1.00	0.99	0.99	352
weighted avg	0.99	0.99	0.99	352

Figure 4.1.3: Confusion Matrix and Classification of Random Forest

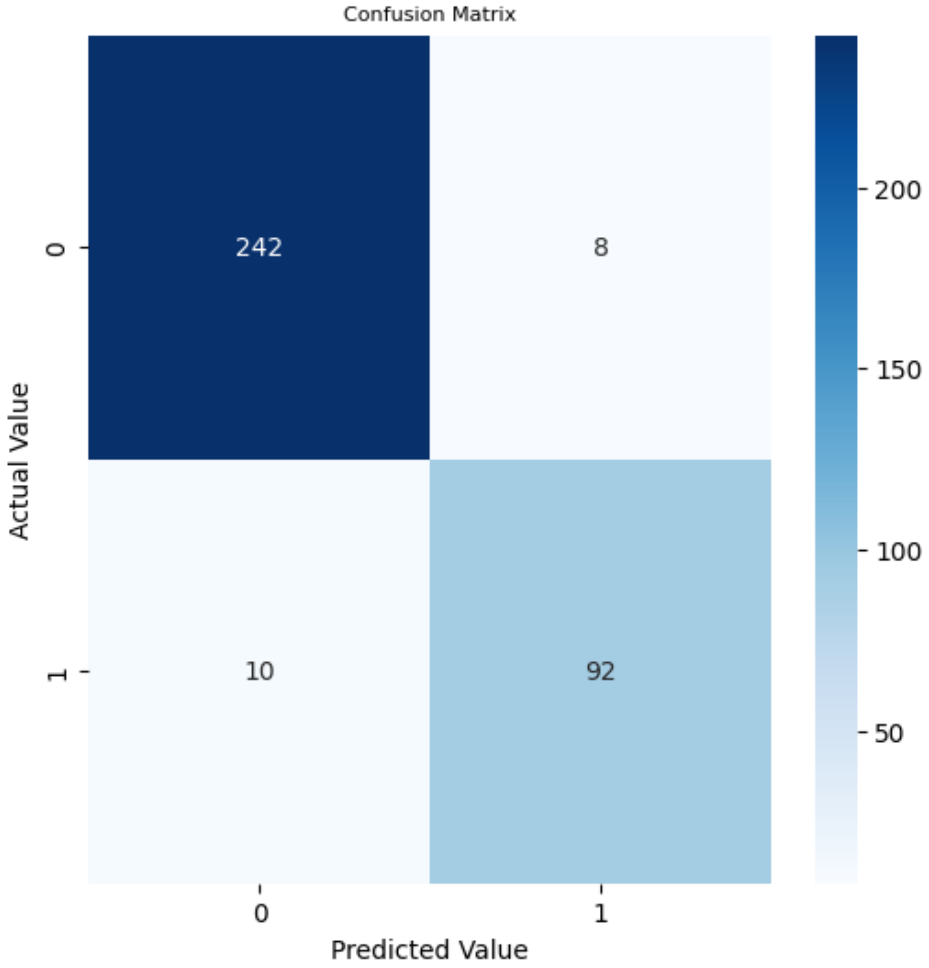
As for the Random Forest the classification report and confusion matrix shows that the accuracy is 99% and f1 score is around 1.0 which is good.



	precision	recall	f1-score	support
0	1.00	1.00	1.00	250
1	0.99	0.99	0.99	102
accuracy			0.99	352
macro avg	0.99	0.99	0.99	352
weighted avg	0.99	0.99	0.99	352

Figure 4.1.4: Confusion Matrix and Classification of XGBoost

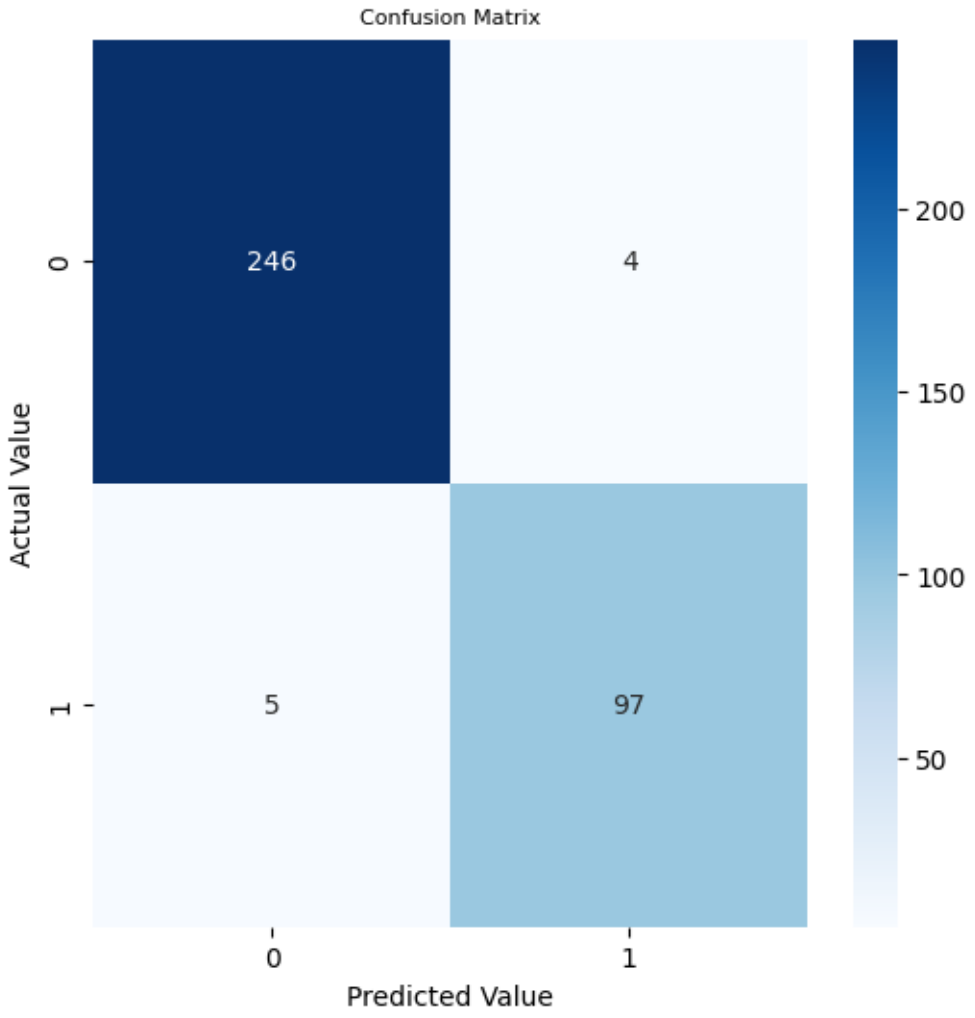
According to the XGBoost model's confusion matrix and classification report, the model's accuracy and other scores of precision, recall and f1 score.



	precision	recall	f1-score	support
0	0.96	0.97	0.96	250
1	0.92	0.90	0.91	102
accuracy			0.95	352
macro avg	0.94	0.93	0.94	352
weighted avg	0.95	0.95	0.95	352

Figure 4.1.5: Confusion Matrix and Classification of K-Nearest Neighbor

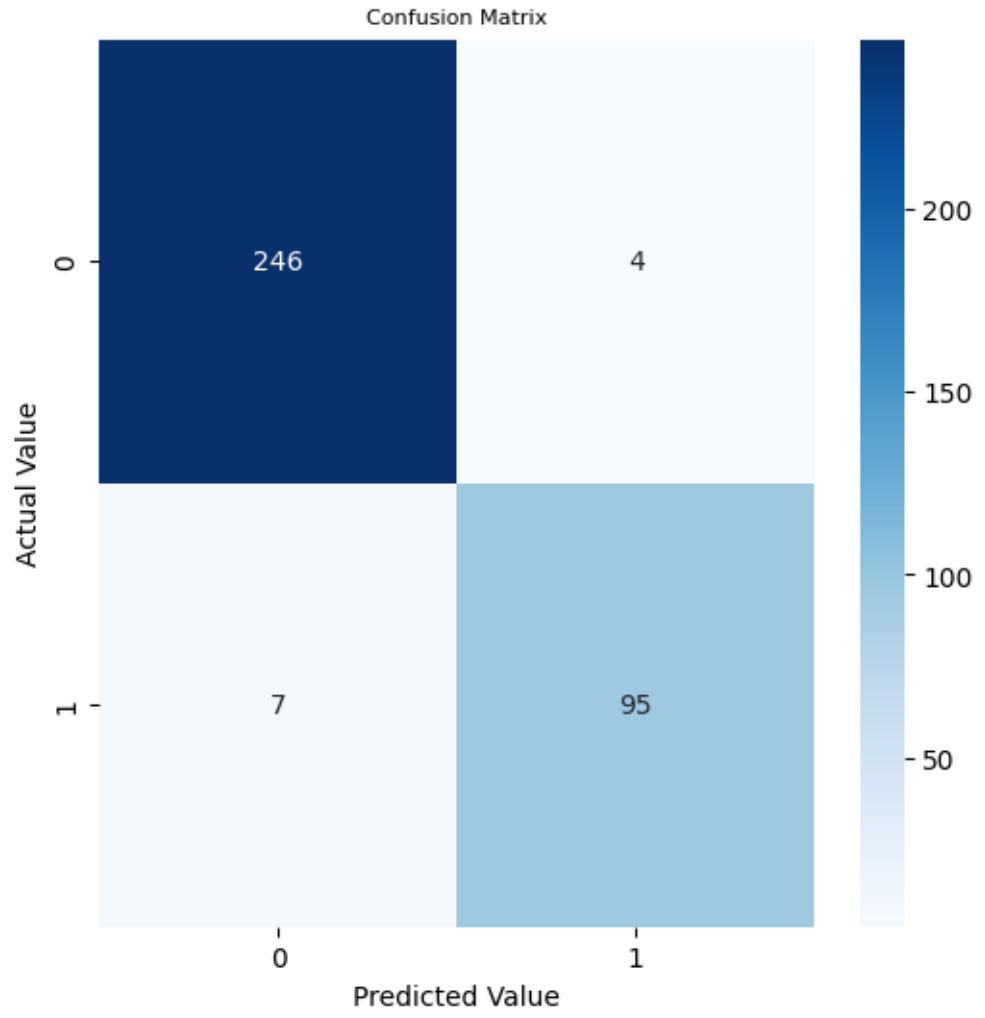
Considering the classification report and confusion matrix of KNN we can see that the accuracy of KNN has decreased which is 95% and the f1-score is satisfactory, indicating an approximate value of 0.96.



	precision	recall	f1-score	support
0	0.98	0.98	0.98	250
1	0.96	0.95	0.96	102
accuracy			0.97	352
macro avg	0.97	0.97	0.97	352
weighted avg	0.97	0.97	0.97	352

Figure 4.1.6: Confusion Matrix and Classification of Decision Ttree

As for the Decision Tree the classification report and confusion matrix shows that the accuracy is 97% and f1 score is around 0.98 which is good.



	precision	recall	f1-score	support
0	0.97	0.98	0.98	250
1	0.96	0.93	0.95	102
accuracy			0.97	352
macro avg	0.97	0.96	0.96	352
weighted avg	0.97	0.97	0.97	352

Figure 4.1.7: Confusion Matrix and Classification of Naive Bayes

Considering the classification report and confusion matrix of Naive Bayes we can see that the accuracy of NB is 97% and The f1-score is satisfactory, indicating an approximate value of 0.98.

Figure 4.1.8 and 4.1.9 represent ROC and AUC curve diagrams for each classifier. ROC curve represents the trade-off between the true positive rate (TPR) and the false positive rate (FPR) at varying thresholds graphically. To evaluate how well a binary classifier performs, we can look at its AUC. It is a numerical summary of the ROC curve, with the best possible value being 1.

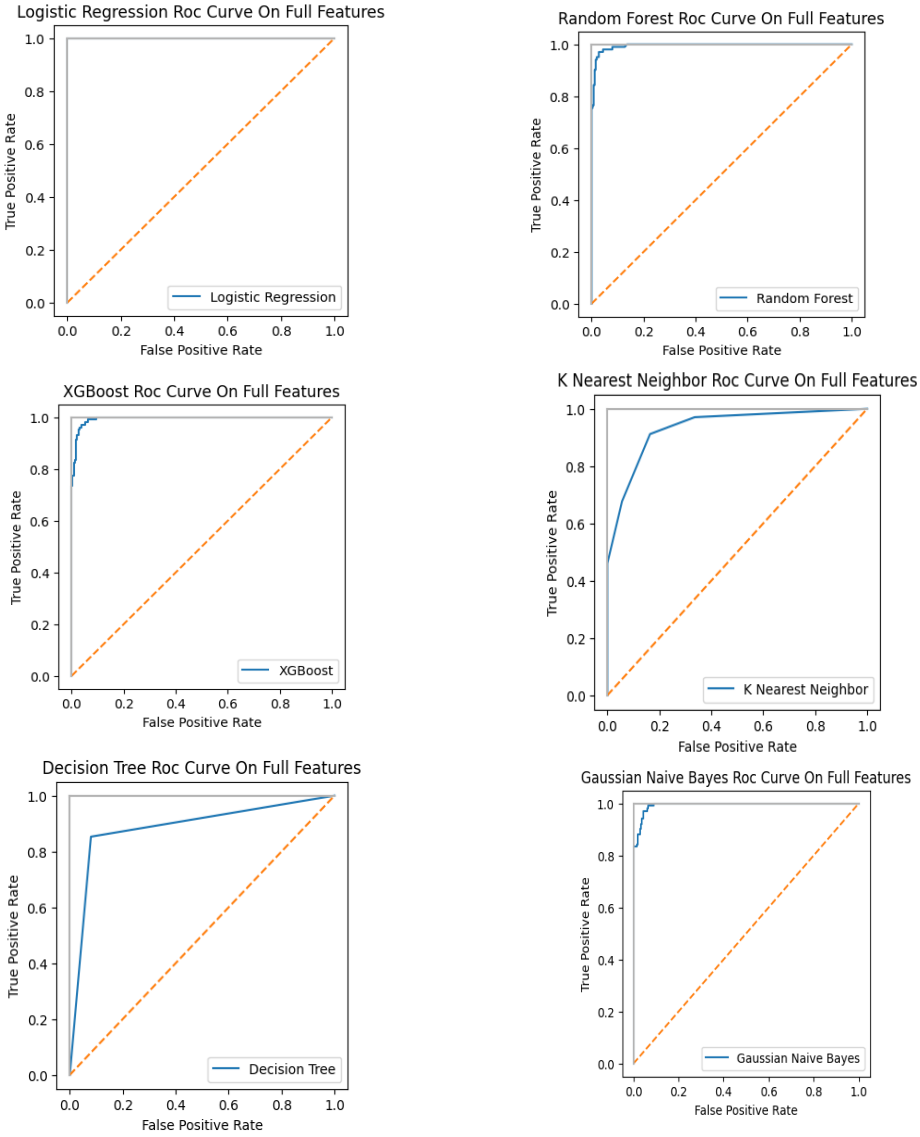


Figure 4.1.8: ROC Curve Diagram

In figure 4.1.8 all six classifier models are represented virtually by each individual graph. True Positive Rate is shown on the vertical axis, and False Positive rate is shown on the horizontal.

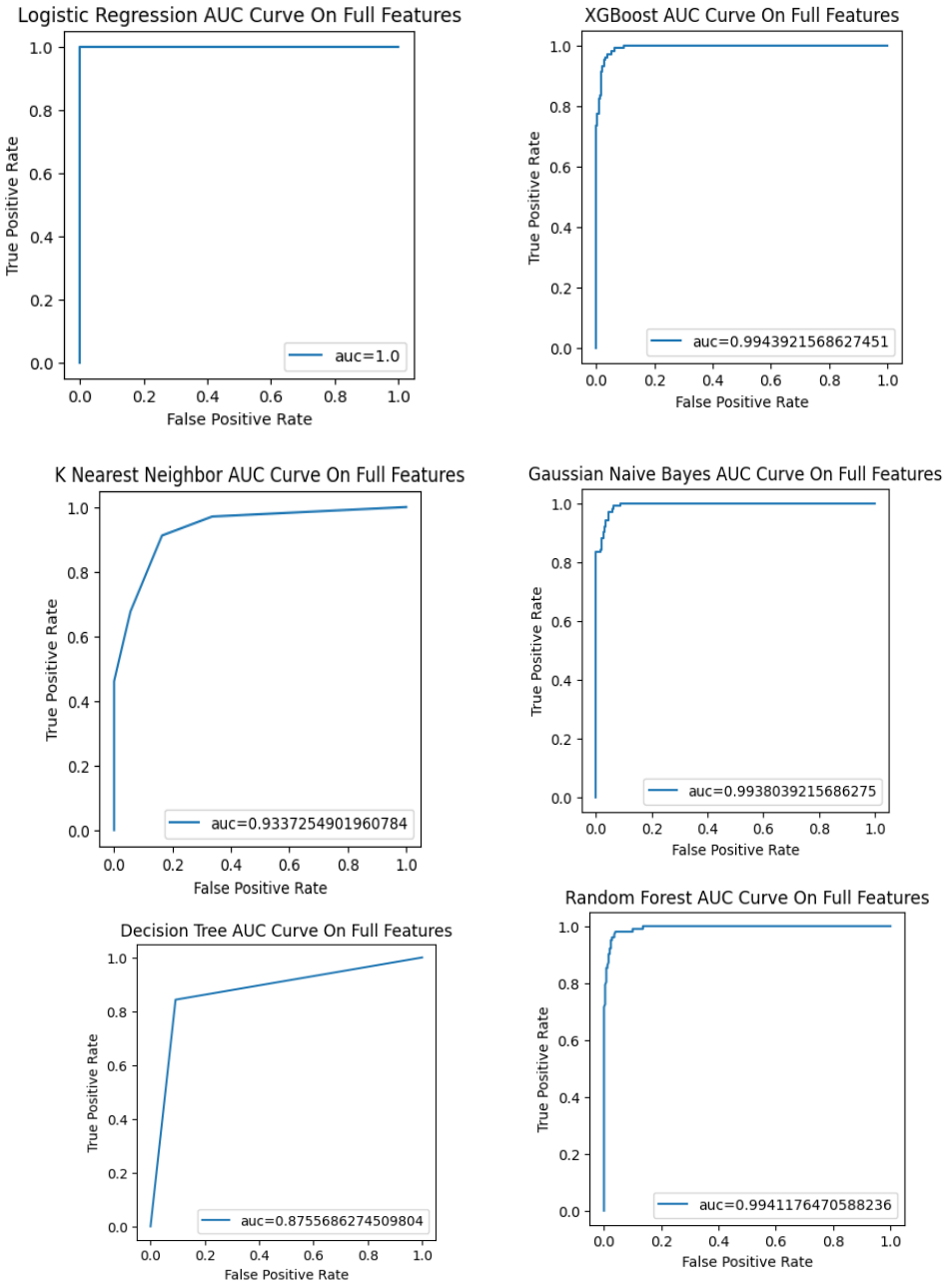


Figure 4.1.9: AUC Curve Graph

Figure 4.1.9 also depicts the AUC curve. The six classifiers are represented by proxies within each graph. The horizontal axis represents false positives, while the vertical axis represents true positives.

Table 4.1.4 displays the optimal settings and scores for Hyper Parameter Tuning. Multiple parameters were used for each classifier in this study. To get the highest possible score, we used the parameters provided.

Table 4.1.4: Hyper Parameter Tuning (Best parameter & Score)

Serial Number	Model Classifier	Used Parameter	Best Parameter	score of Hyper parameter tuning
1	LogisticRegression	{'C': [0.001, 0.01, 0.1, 1, 10],'solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']}	'C': 10, 'solver': 'newton-cg'	1.0
2	Random forest	{'n_estimators': [10, 40,80,100], 'max_features':['auto','sqrt','log2'], 'max_depth': [None, 5,10, 20], 'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2, 4] }	None,'auto',1,5, 80	0.951
3	XGBoost	{'n_estimators': [100, 200, 400, 500], 'learning_rate': [0.01, 0.1, 0.2, 0.3], 'max_depth': [3, 5, 7, 9], 'subsample': [0.5, 0.7, 0.9]}	'learning_rate':0.3, 'max_depth':5, 'n_estimators':500, 'subsample': 0.5	0.974
4	K-Nearest Neighbor	{'n_neighbors': [3, 5, 7, 10], 'weights': ['uniform', 'distance'], 'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'] }	algorithm"ball_tree', 'n_neighbors':3, 'weights': 'distance	0.869
5	Decision Tree	'criterion':['gini', 'entropy'] splitter': ['best', 'random'] max_depth': [None, 5, 10, 20] 'min_samples_split': [2, 5, 10] 'min_samples_leaf': [1, 2, 4]	criterion': 'entropy', 'max_: 20,1, 2,'random	0.911
6	Naïve Bayes	{'var_smoothing': [10, 15, 20]}	'var_smoothing': 10	0.710

CHAPTER 5

IMPACT ON SOCIETY & ETHICAL ASPECTS

5.1 Impact on Society

Autism has the highest prevalence and is also the least treatable developmental disorder. Autism is largely unknown in Bangladesh. It's possible this could happen again if doctors fail to properly diagnose a patient's condition. Patients are more likely to avoid detection if this is the case. Using machine learning methods, we created a model that accurately identified patients with Autism 99.20% of the time. Our method will help medical professionals improve their diagnosis of Autism and treatment options for those with the disorder. By taking these measures, we can hopefully contribute to reducing the mortality rate of those on the Autism Spectrum.

5.2 Ethical Aspects

Autism spectrum disorders (ASD) raise complex ethical questions that call for an empathetic and welcoming response. Individuals with autism spectrum disorder (ASD) should be treated with dignity and respect, and their right to make their own decisions and choices should be upheld. It's crucial to promote diversity and acceptance so that people with autism can take advantage of the same opportunities as everyone else. Individuals with ASD should be given top priority by professionals and caregivers, who should work to improve their quality of life and protect them from harm. It is crucial to provide effective and safe support by employing interventions and treatments that are backed by evidence. Individuals' identities and health records should be treated with the utmost confidentiality. Research involving people with autism requires strict ethical standards, including informed consent and protection of participant rights. When communicating ethically with someone who has autism, it is important to take into account their unique communication needs and preferences. Ethical duty also includes speaking up for the rights of people with autism and educating the public about the gifts and challenges they face. Ethical considerations in this area center on the importance of providing autistic people with a welcoming and safe space that respects their individuality and dignity.

CHAPTER 6

CONCLUSION & FUTURE WORK

6.1 Conclusion

The classification of autistic data stands out as a challenging issue in the field of autism informatics. The work holds the distinction of being one of the most ancient contributions within the realm of research. Based on the current analysis, it has been identified that there are numerous areas that present potential for enhancement. In this study, we employed a cross validation of 3 folds also applied various few methods of preprocessing to assess the correlation between the efficacy of several traditional classifiers and modern variants, such as bagging and boosting, on the ASD dataset. In general, the model seems to be extremely competent when it comes to diagnosing ASD.

Future Scope

Experimenting with a variety of preprocessing techniques and making use of deep learning strategies ought to be among our long-term objectives.

REFERENCE

[1] Raj, S. and Masood, S., 2020. Analysis and detection of autism spectrum disorder using machine learning techniques. *Procedia Computer Science*, *167*, pp.994-1004.

[2] Erkan, U. and Thanh, D.N., 2019. Autism spectrum disorder detection with machine learning methods. *Current Psychiatry Research and Reviews Formerly: Current Psychiatry Reviews*, *15*(4), pp.297-308.

[3] Rahman, M.M., Usman, O.L., Muniyandi, R.C., Sahran, S., Mohamed, S. and Razak, R.A., 2020. A review of machine learning methods of feature selection and classification for autism spectrum disorder. *Brain sciences*, *10*(12), p.949.

[4] Hossain, M.D., Kabir, M.A., Anwar, A. and Islam, M.Z., 2021. Detecting autism spectrum disorder using machine learning techniques: An experimental analysis on toddler, child, adolescent and adult datasets. *Health Information Science and Systems*, *9*, pp.1-13.

[5] Vakadkar, K., Purkayastha, D. and Krishnan, D., 2021. Detection of autism spectrum disorder in children using machine learning techniques. *SN Computer Science*, *2*, pp.1-9.

[6] Abdullah, A.A., Rijal, S. and Dash, S.R., 2019, November. Evaluation on machine learning algorithms for classification of autism spectrum disorder (ASD). In *Journal of Physics: Conference Series* (Vol. 1372, No. 1, p. 012052). IOP Publishing.

[7] Thabtah, F., 2019. Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. *Informatics for Health and Social Care*, *44*(3), pp.278-297.

[8] Elavarasi, S.A., Jayanthi, J. and Jayasankar, N.B.T., 2020. Methods for improving the predictive accuracy of autism spectrum disorder screening using machine learning algorithms. *Methods*, *29*(03), pp.9255-9262.

[9] Chowdhury, K. and Iraj, M.A., 2020, November. Predicting autism spectrum disorder using machine learning classifiers. In *2020 International conference on recent trends on electronics, information, communication & technology (RTEICT)* (pp. 324-327).

[10] Duda, M., Ma, R., Haber, N. and Wall, D.P., 2016. Use of machine learning for behavioral distinction of autism and ADHD. *Translational psychiatry*, *6*(2), pp.e732-e732.

[11] https://www.medicalnewstoday.com/articles/types-of-autism. [Online]

[12] https://www.autismspeaks.org/autism-statistics-asd [Online]

APPENDIX

Variable in dataset	Corresponding related features
A1	Delayed speech and language abilities
A2	Delayed motor skills
A3	Delayed cognitive development or learning abilities
A4	Hyperactivity, impulsivity, and inattention
A5	Epileptic seizures or seizure disorder
A6	Atypical eating and sleeping patterns
A7	Gastrointestinal problems (such as constipation)
A8	Unusual emotional or mood responses
A9	Anxiety, heightened stress, or excessive worry
A10	Abnormal fear responses or heightened fear levels

191-15-2708_Check

ORIGINALITY REPORT

18%	15%	5%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	6%
2	Submitted to Daffodil International University Student Paper	2%
3	Submitted to Michigan Technological University Student Paper	1%
4	Submitted to University of Edinburgh Student Paper	1%
5	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	1%
6	tudr.thapar.edu:8080 Internet Source	1%
7	www.mdpi.com Internet Source	<1%
8	link.springer.com Internet Source	<1%
9	Submitted to University of Westminster	