

**A Methodical Machine Learning Approach for Autism Stage
Classification**

BY

**Md Nazrul Islam
ID: 191-15-2593**

AND

**Md Mohaimanul Islam
ID: 191-15-2424**

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

Supervised By

Aliza Ahmed Khan
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised By
Al Amin Biswas
Lecturer
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

This Project/internship titled “**A Methodical Machine Learning Approach for Autism Stage Classification**”, submitted by Md Nazrul Islam ID:191-15-2593 and Md Mohaimanul Islam ID:191-15-2424 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfilment 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 **25/01/2023**.

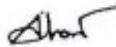
BOARD OF EXAMINERS


29/1/23

Dr. Touhid Bhuiyan
Professor and Head

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

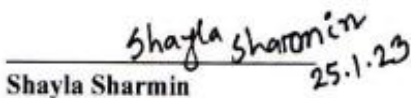
Chairman



Dr. Md. Atiqur Rahman
Associate Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

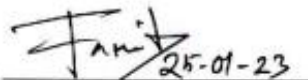
Internal Examiner


25.1.23

Shayla Sharmin
Senior Lecturer

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner


25-01-23

Dr. Dewan Md Farid
Professor

Department of Computer Science and Engineering
United International University

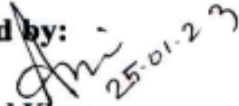
External Examiner

DECLARATION

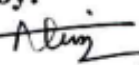
We hereby declare that this project has been done by us under the supervision of **Aliza Ahmed Khan, Senior Lecturer, Department of CSE** Daffodil International University.

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:

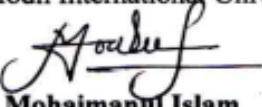

Aliza Ahmed Khan
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:


Al Amin Biswas
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Submitted by:


Md Nazrul Islam
ID: 19-15-2593
Department of CSE
Daffodil International University


Md Mohaimanul Islam
ID: 191-15-2424
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

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

We are really grateful and wish our profound indebtedness to **Aliza Ahmed Khan, Senior Lecturer**, 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 drafts 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 members 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 discussion while completing the course work.

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

ABSTRACT

Since autism is thought to present differently in each person, it is commonly viewed as a spectrum condition. Some, but not all, of these features are shared by people with autism, and an individual's level of autism-related symptoms may also vary. While some autistic persons have no spoken language at all, others are able to communicate normally. There is a wide range in how much assistance an individual needs, and even the same individual may appear in quite different ways at various times. Recent versions of the leading diagnostic manuals identify ASD as a single diagnosis, despite the fact that autism was formerly separated into subtypes that have since been questioned for their validity. Overall and by subgroups, this study examined whether the rate of new autism diagnoses has been rising over the past two decades. Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), Gradient Boosting (GB), X-Gaussian Boosting (XGB), k-Nearest Neighbors (KNN), Adaboost Classifier (ADB), and Support Vector Machine (SVM) classifier were used to achieve a maximum classification accuracy of 99.68% in this study. The data was cleaned up and prepared for analysis, and then the ML algorithm was used. The results of each algorithm were compared, and an accurate result was found. The F1 score, together with precision, recall, and the AUC score, is used to evaluate performance. Analyses show that Svm classifier and Gradient Boosting are the most effective methods for reaching an overall accuracy of 99.75%.

Keyword

Logistic Regression Classifier (LRC), Random Forest Classifier (RFC), Decision Tree, Gradient Boosting, Support Vector machines, K-Nearest Neighbors Classifier(KNNC), Adaboost classification, Gaussian training, XGB classifier.

TABLE OF CONTENTS

CONTENTS	PAGE
Approval Page	i
Declaration	ii
Acknowledgement	iii
Abstract	iv
List of Tables	vi
List of Figures	vii
Chapter 1: Introduction	1-3
1.1 Overview	1-2
1.2 Motivation	3
1.3 Research Objectives	4
1.4 Research Questions	4
1.5 Research Layout	4
Chapter 2: Background Study	5-7
2.1 Related Works	5-7
Chapter 3: Research Methodology	8-22
3.1 Overview	8
3.2 Pre-processing	10
3.3 Classification Algorithm	12-18
3.5 Ensemble Classifier	19-20
3.6 Dataset Description	21-22
Chapter 4: Experimental Results and Discussion	23-27
4.1 Experimental Result Analysis	23-27
Chapter 5: Conclusion and Future Work	28
5.1 Conclusion and Future Work	28
References	29-30

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.3.1: Logistic Regression Classified	12
Figure 3.3. 2: Support Vector Machine (SVM) Classification	13
Figure 3.3.3: Decision Tree Classification	14
Figure 3.3.4: Random Forest Classification	15
Figure 3.3.5: Extreme Gradient Boosting	16
Figure 3.3.6: K-Nearest Neighbors (KNN) Classification	17
Figure 3.3.7: Gradient Boosting Classification	18
Figure 3.3.8: AdaBoost Classification	19
Figure 3.3.9: Bagging and Boosting	19
Figure 3.5: Performance evaluation measures formula	22
Figure 4.1: ROC Curve Analysis of Introduced Algorithm	24

LIST OF TABLES

TABLES	PAGE NO
Table 01: Details of dataset	20-21
Table 02:Result of the introduced Algorithm	23
Table 3: Experimental Results of Bagging Classifiers	25
Table 4: Experimental Results of Boosting classifier	26

CHAPTER 1

Introduction

1.1 Overview

Excessive and restricted behaviors, as well as varying degrees of difficulty with social interaction and language expression, define autistic (or autism spectrum disorders, ASD) (Lord et al., 2018). By 18–24 months of age, it is possible to distinguish the classic symptoms of autism from those of typical growth, other delays, as well as other disorders of early brain development. New insights on autism have had a far-reaching impact on public policy on a worldwide scale. With the major policy shifts that have occurred. Improvements in domains including such psychological health, maternal and child health and human rights have aided in the spread of autism awareness and advocacy across the world (WHO 2013a,b, 2014, 2015, 2018; WHO & UNICEF, 2018). The United Nations Convention on the Rights of Persons with Disabilities (UNCRPD) has been the foundation and the driving force behind these advancements. The UNCRPD outlines key principles such as regard for dignity, liberty of choice and autonomy, quasi, full participation and inclusion in society, and acceptance of people with disabilities as part of human diversity. There has been a push for autism research to pivot from a focus on raising awareness to one that emphasizes strengthening existing institutions and infrastructure (WHO, 2013a). While there has been a significant increase in autism awareness in many parts of the world, this hasn't necessarily led to a corresponding increase in assistance. The World Health Organisation has approved the World Health Organization's (WHO) Integrated Mental Action Plan (2013-2020). (WHO, 2013b). The policy supports "high-quality, culturally significant health and social care," which is available when needed. All individuals affected by mental health issues should be able to fully exercise their human rights and have ready to receive high-quality, culturally sensitive social and medical care in a culture that values and promotes mental health. The plan is guided by a number of principles, including universal health care, scientific proof treatment, multisectoral methods, a life-course approach, and the independence of persons with mental illnesses. Success has also been shown with country- and geographical area policies for autism that were developed or enhanced in response to local needs (e.g., the UK's Autism Act, 2009; Ireland's Schon & Started by researching, 2018; Que, Canada's Zeidan et al., 2019; Laudon and laudon et al., 2018). Increases in autism awareness related public health responses throughout the world have gone hand in hand with epidemiological studies that give objective indications of the effect of autism, such as estimates of instances and their accompanying social and economic implications. To be more specific, epidemiological estimates can show policymakers how to better assist the affected population through better identification, services, and supports (Fombonne, 2019; Franz et al., 2017; Raina et al., 2017). Studies in epidemiology are increasingly being used, albeit covertly, to derive conclusions about the

etiology of autism. For instance, if the rate of occurrence has been rising, that is interpreted as a result of increased exposure to risky environmental conditions. It is thought that differences in frequency by significant sociodemographic variables (such as location, ethnicity, socioeconomic status, and so on) reflect real differences in biology and/or environmental causation (e.g., Hewitt et al., 2016). Health inequalities, where stigma and/or structural obstacles marginalize ethnic or socioeconomic minority, modify their access to care, and so impact prevalence, provide one alternative explanation for these connections (e.g., Durkin et al., 2017; Elsabbagh, 2020). The purpose of this updated analysis is to provide more accurate estimates of autism's global prevalence. The most recent systematic assessment of global prevalence, published in 2012, used 70 estimations to determine that the prevalence of ASD was, on average, 62 per 10,000 children, with males consistently having a greater prevalence (Elsabbagh et al., 2012). There was a lot of variation in estimates both within and between geographical regions, and some numbers were either scarce or nonexistent in many parts of the world, like Eastern Europe and Africa. Additionally, the estimations that were available mostly applied to youngsters, with relatively few research including individuals older than 18.

- ◆ We have pre-processed the dataset using a variety of techniques, including feature scaling, addressing imbalanced data, and categorical data encoding.
- ◆ We have put together a comparison of the algorithms using eight distinct feature scaling techniques.
- ◆ Using common machine learning techniques (DT, RF, KNN, XGB, GB, ADB, GT, GDB) we have been able to make accurate predictions about the course of the disease with better accuracy.
- ◆ We've utilized cross-validation to check our work and eliminate any bias by running it 10 times.
- ◆ In order to assign the optimum settings and increase accuracy, we have incorporated hyperparameter adjustment.
- ◆ Using ensemble learning classifiers, we have enhanced and guaranteed real-world performances (Bagging and Boosting).
- ◆ In order to evaluate our performance to that of others, we have determined the outcomes of numerous evaluation measures.

1.2 Motivation

Differences in brain structure and function are at the root of autism spectrum disorder (ASD), a developmental impairment. Characteristics of those on the autism spectrum include difficulties in social interaction and communication, as well as narrow or repetitive hobbies or habits. Asperger's persons may also behave, learn, and pay attention in unusual ways. It's worth noting that some folks who don't have autism spectrum disorder could also exhibit these signs. However, these traits can make daily living difficult for those with autism spectrum disorder. People on the autism spectrum may struggle with social communication and interpersonal skills. Few research papers on early prediction are available. So, we do numerous studies to predict diabetes. The majority of them do not attain higher accuracy. As a result, we were driven to find the best accuracy outside of our method in order to attain the best results. Another typical worry that people may have when thinking about ABA therapy is the use of extrinsic rewards. But those in charge of ABA programs prepare ahead of time and work to gradually switch from extrinsic to more intrinsic reinforces. Extrinsic rewards, also known as transitional positive reinforcements, are a crucial component of a child's program that fosters greater positive associations with new behaviors. It can be a good idea to get in touch with a qualified expert if you are having trouble figuring out what inspires your autistic child or you would like to see changes in what motivates your child. You can get assistance determining how your child's motivation and autism are related.

1.3 Research Objectives

To make classification smooth and efficient.

To generate better outcome comparing to the other related works.

To raise public awareness to the autistic people

To categorize autistic people to provide special care to them

1.4 Research Questions

From where the dataset was collected?

What are the key features of the dataset?

What feature selection techniques were employed in this study?

What approach had taken to generate highest accuracy outscoring previous researches?

1.5 Report Layout

Background

Research Methodology

Experimental Results & Discussion

Conclusion & Future Work

CHAPTER 2

Background Study

2.1 Related Works

Numerous studies have examined autism and utilized various methods to gauge the performance of various machine learning algorithms. [2] Researchers in the fields of primary care, clinical research research datalinks, temporal patterns, and autism/autism spectrum disorder/diagnosis utilized a total of five classification models in their work. The number of people receiving a diagnosis of autism has climbed exponentially over the past two decades, rising by 787 percent between 1998 and 2018 ($R^2 = 0.98$, exponentiated coefficient = 1.07, 95% CI [1.06, 1.08], $p < .001$). Increases in diagnoses were more pronounced in females than males, and they were moderated by age group, with increases in adult diagnoses being the highest (exponentiated interaction coefficient = 1.06, 95% CI [1.04, 1.07], $p < .001$) and being more noticeable in women than in men (exponentiated interaction coefficient = 1.02, 95% CI [1.01, 1.03], $p < .001$). Regularly collected data from 738 GP surgeries in England and Northern Ireland (10% of all GP surgeries in England and NI) is stored in a database named CPRD Aurum. Over 19 million patients' diagnoses, symptoms, medications, referral, and test results are all included. Wolf et al. (2019) report that in 2018, seven million patients, or 13% of England's inhabitants, were actively contributing data. In 2019, we incorporated data from Northern Ireland. To locate research published in peer-reviewed journals between 2008 and 2018, the authors searched the electronic databases PubMed and Ebsco by using terms (music therapy OR music OR songs OR sound treatment) AND autism. He provides information from 36 research on the effectiveness of music therapy for children with autism spectrum disorder. The current study discovered problems with the methodology of some of these studies. As a result of these caveats, it is now difficult to determine the precise effects of music therapy on autism. Consequently, he reasoned, further research is required using sufficiently large samples, standard experimental methods, and objective evaluations of therapeutic success. advocate for further research on the effects of music therapy to incorporate neuroimaging techniques. [5] This study categorizes cases of ASD according to four criteria: prevalence (how often ASD cases are), incidence (how often ASD occurs), and epidemiology. Six of the 17 elements used to diagnose ASD on the Autism Treatment Assessment and Classification System (ATAC) are related to either language or communication, social interaction, or limited and repetitive behavior. correlation coefficients in the primary model. The model using the function gee derived from degree has cluster-resistant standard errors. By using ASD symptom score and birth cohorts, a model of the ASD impairments score was developed, complete with 95% confidence intervals. The data was gathered from the Child and Teenage Sibling Research in Sweden (27,240 respondents), in which parents rated their child's ASD symptoms and disability. Impairment caused by ASD symptoms was reversed using symptom ratings. There were five groups of people who all had the same birth year

(1995–1997, 1998–2000, 2001–2003, 2004–2006, and 2007–2009). [8] During the years 2007–2013, information was collected from 18 Autism

Oregon Health & Science University, where the authors are based, is part of the American Cancer Society's Cancer Treatment Network (ATN). Participants who suspected they had Autism Spectrum Disorder were referred to local Autism Treatment Network (ATN) locations for evaluation (ASD). The ATN has designed a standardized registration procedure for use by all network locations. The site's registry included children and adolescents between the ages of 2 and 18, provided that they had been diagnosed with ASD in accordance with DSM-IV-TR criteria, and had received the ADOS (Lord et al. 1999). Using African-Americans, Autism, Race, Adaptive Behavior, Intellectual Functioning, Behavioral Problems, and Emotional Issues. Included in this tally are all 245 of the study's non-Hispanic Black participants who had previously been tested using Module 1 of the Autism Diagnostic Observation Schedule (ADOS). Early Learning Measures by Mullen (MSEL; Mullen, 1995). The Mullen Scales of Early Learning were used to evaluate 275 individuals (mean age, 3.4 years; 82.5% males; 92 Black, 183 White). One of the most often used tests with samples of young children with ASD is the MSEL Early Learning Composite (ELC) Standard Score (Duvall et al., 2020). Standardized to a average of 100 as well as a standard deviation of 15, it was used to evaluate general cognitive development. In pilot trials, the ELC scores deviated noticeably from the typical distribution. An examination of the numbers uncovered a statistically significant positive skew, with a percentage of the sample scoring below the manual's minimum threshold of 49. In this study, children with ASD were split into two groups and given either an inflated wrap (group I) or a manual pull (group II) for 20 minutes, twice a week, for three weeks. GSR, or skin reaction, was employed to measure the sympathetic stress response, and the Galvanic Parent Rating Scale-48 (CPRS-48) and the Conners' Parent Rating Scale (CPRS) were utilized to grade behavior. Counting all of them, there were 20 children with autism spectrum disorder (14 boys and 6 girls, 7 to 13 years old). Disruptive behaviors were more common in the inflatable group compared to the manual pull group ($p = 0.007$), suggesting that CPRS-48 had conduct problems. In groups, however, the GSR showed a statistically significant reduction in sympathetic response ($p = 0.01$). [11] Two-feature decision tree classifier performed best (92.11% accuracy). His findings suggest that there may be measurable biomarkers of ASD in the kinematics of head motions. According to the findings, a DT classifier using just two features was able to attain a 92.11% rate of success in classification. Because of advances in technology that allow for the gathering of objective characteristics, it is possible that other new ASD biomarkers may be identified. Features gathered with neuroimaging, eye tracking, EEG, and motion capture methods have showed potential in detecting ASD, as evidenced by research by Crippa et al. (2015), Grossi et al. (2020), Liu et al. (2016), Plitt et al. (2015), and Zhao et al. (2019). OpenFace 2.0, a head posture tracking system, was used to compute six features, which were then fed

into four ML classifiers. [10] There are still gaps in the knowledge base that prevent us from drawing definitive conclusions from epidemiological data on underlying causative processes evidence.

Meanwhile, worldwide estimates of the prevalence of disease have been growing, especially in regions like Africa and the Middle East that have been under-represented in the past (al-Mamari et al., 2019; Alshaban et al., 2019; Chinawa et al., 2016).

In order to obtain a more accurate worldwide estimate of ASD prevalence, this study increased the previous estimate of 62/10,000 to a median prevalence of 65/10,000. Prevalence increases over time have been seen in studies focusing on countries and populations as diverse as the U. S. (Christensen et al., 2019; Jariwala-Parikh et al., 2019), South Korea (Hong et al., 2020), and Taiwan (Liu et al., 2019). (Lai et al., 2012). In a similar vein, researchers in Australia (May et al., 2020; May et al., 2017; Randall et al., 2016) and France have observed an increase in assessed prevalence in consecutive birth cohorts (van Bakel et al., 2015). [3] The purpose of this study was to systematically review previous studies that have looked into GM in ASD children. We were able to narrow the field down to 18 investigations. Our data showed that both *Streptococcus* (SMD+ = 0.999; 95% CI 1.549, 0.449) and *Bifidobacterium* (SMD = 0.513; 95% CI 0.953, 0.073) relative abundances were lower in children with ASD. Overall. The EndNote X7 application was used to compare the 2391 studies located in the datasets to check for any repetition. After reading the brief synopses of the remaining pieces. In his research, he found no evidence of a significant GM-ASD link in the phyla Bacteroidetes, Firmicutes, Proteobacteria, or Actinobacteria.

CHAPTER 3

Research Methodology

3.1 Overview

Data for this analysis was retrieved from the UCI machine learning repository. Our research was based on information gathered through a survey that was part of a larger project called Special studies addressing relevant topics at the Sylhet diabetes clinic. Data collection and preprocessing are just two of the many components of the current approach. The current research employs five distinct algorithms: Decision Tree (DT), the Random Forest (RF), the Logistic Regression (LR), the k-Nearest Neighbors (KNN), the Support Vector Machine (SVM), the XGBclassifier (XGB), and the Gradient Boosting (GDB). Where XGB's 99.67% accuracy is tops. After that, we employ hyper parameter tweaking based on 10 iterations of cross-validation. Then, we employed a technique called bagging and boosting to improve accuracy, and the resulting SVM model had a 99.68% success rate.

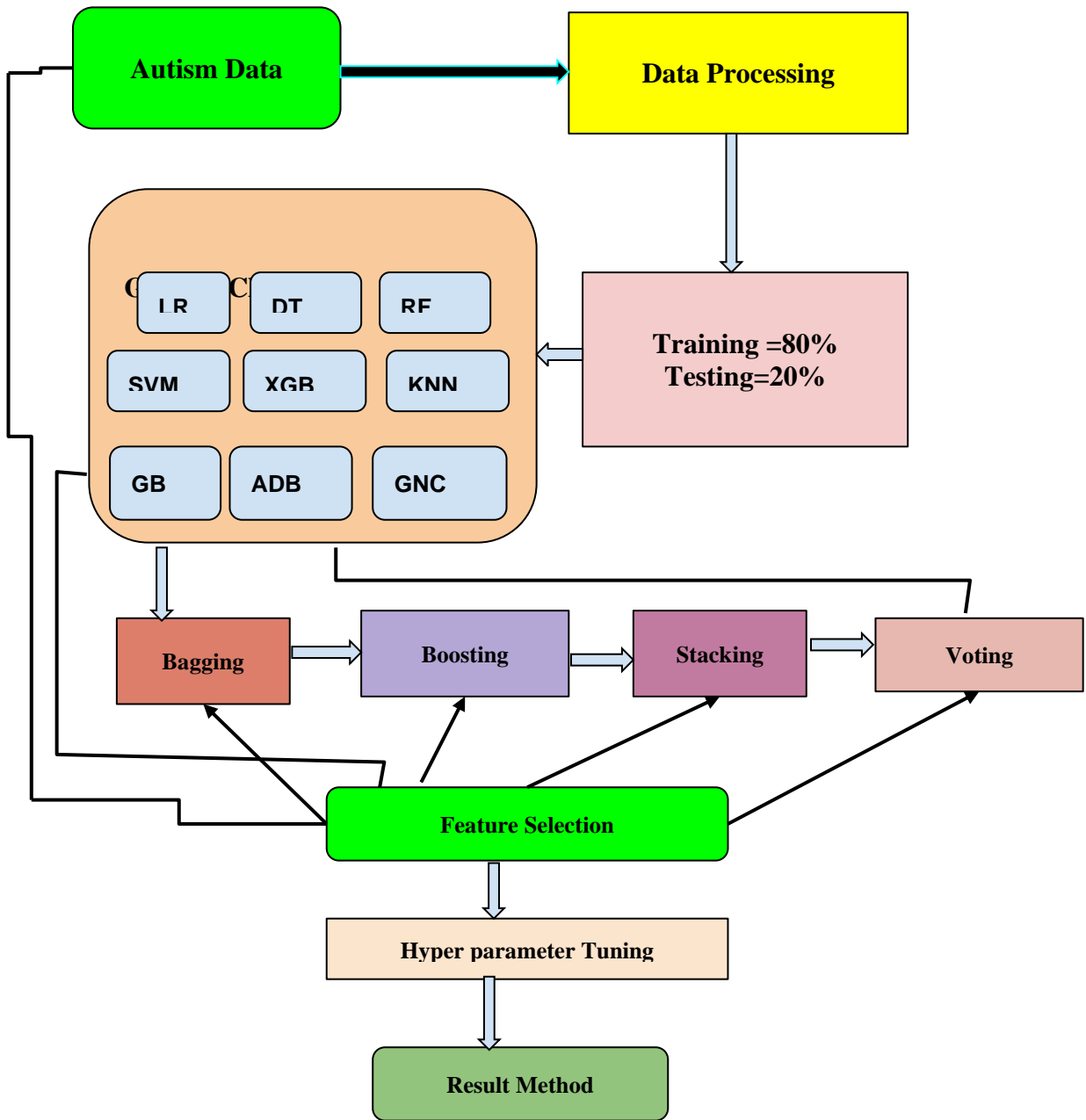


Figure 3.1: Overall Process Diagram

3.2 Pre-processing

3.2.1 Categorical to Numerical Conversion:

Categorical data may be encoded by translating each piece of information into a number value. In order for machine learning models to function, the variables used as inputs and results must, as far as we can tell, in order to fit a model, categorical data must be transformed to integers if the dataset contains them. In our diabetes dataset, we have both categorical and numeric information; with the exception of the age column, all other columns include nominal data. The dataset is now encoded.

3.2.2 Missing Value Imputation:

Missing value imputation is indeed the method of filling in gaps in data using estimates derived from other sources of information. Fortunately, none of our data sets include any blanks.

3.2.3 Handling Imbalanced Data:

In order to deal with unbalanced data, it is necessary to modify the classification performance of the dataset. The Synthetic Minority Oversampling Method (smote) has been useful for dealing with our skewed data. It manages information by uniformly expanding the sample size of a dataset. It utilizes the whole dataset while simultaneously boosting the data of the underrepresented group.

3.2.4 Feature Scaling:

The feature scaling technique is used to standardize the distribution of data's unrelated attributes (FS). The MinMax data scaler uses the interval [0, 1] when no negative values are present in the input data, and the interval [-1, 1] otherwise.

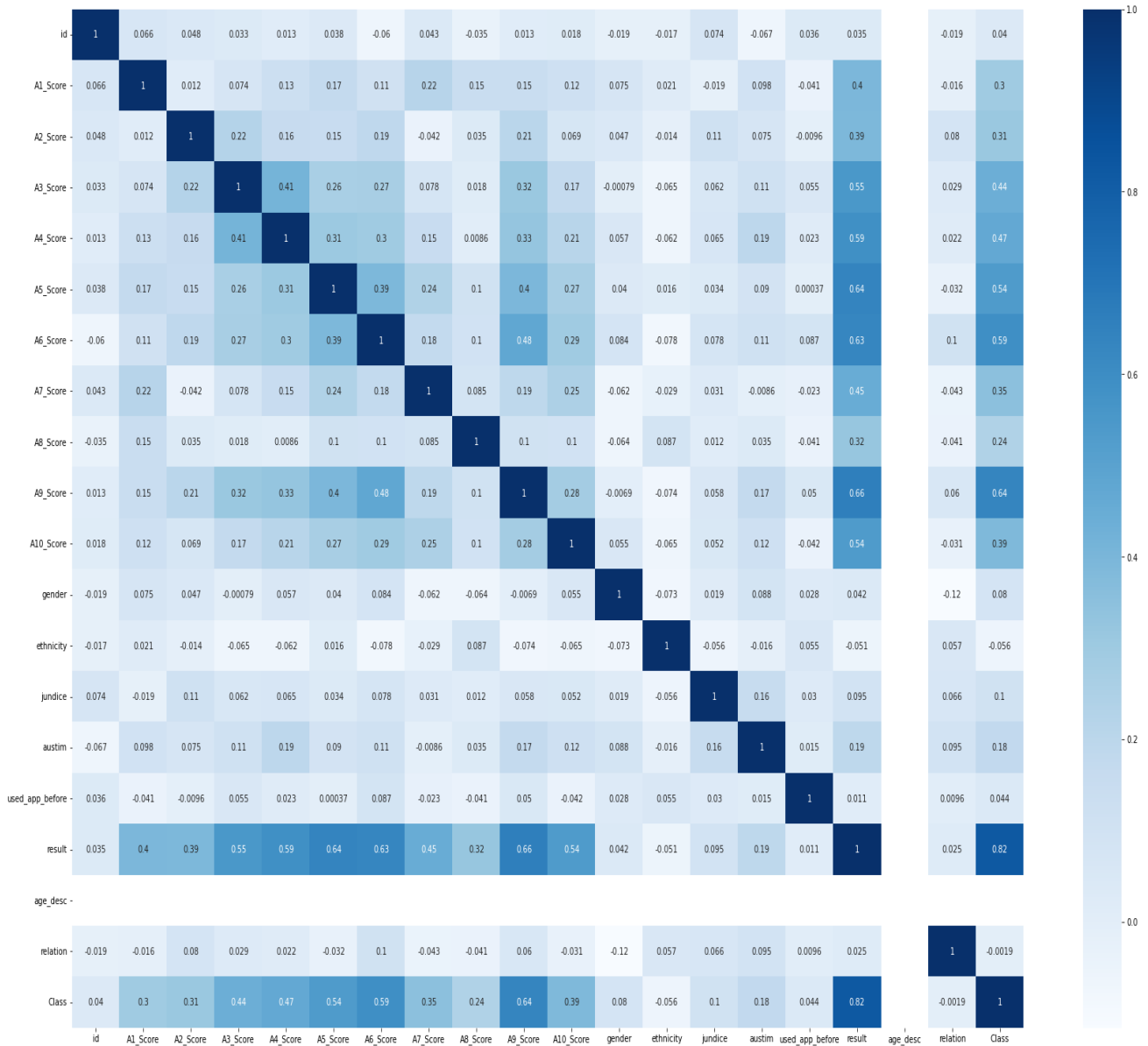


Figure 3.2: Correlation Between the Attributes

3.3 Classification Algorithm

The datasets have been classified in this study using ML-based classifiers including Logistic Regression (LR), Support Vector Machine (SVM), Gradient Boosting (GB), Naive Bayes (NB) and Decision Tree (DT).

3.3.1 Logistic Regression(LR)

The approach is ML-based, and the classifier consists of a simple yes/no pair (yes/no = binary 0/1) that may be represented as yes or no. Logistic regression assumes the seers are insufficient to fix the response variable, even though they confront a scenario in which a logistic action of a linear aggregation of them might be achievable. Some situations benefit greatly from LR's strengths, such as when the seer is not content and provides extra probabilistic assessment of the feedback.

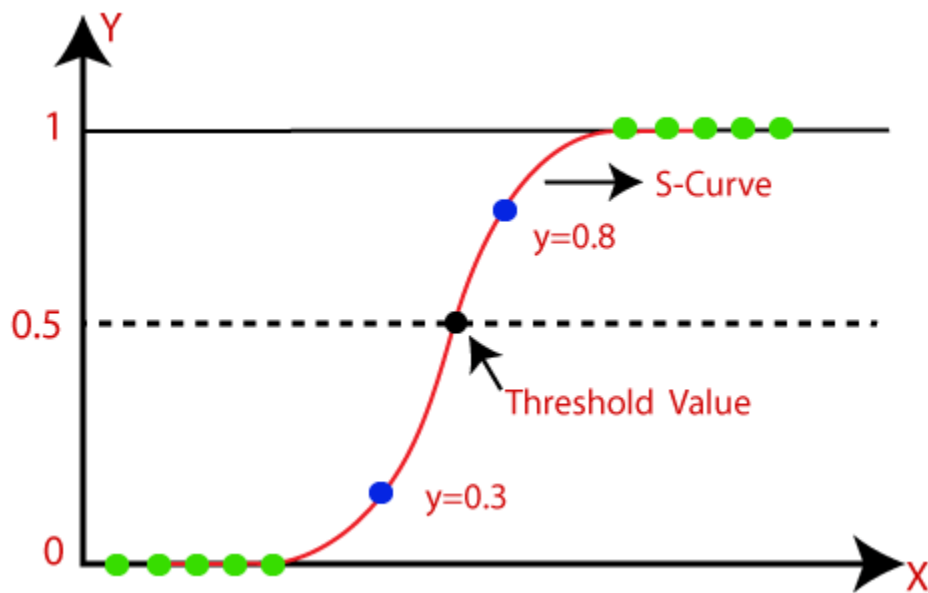


Figure 3.3.1: Logistic Regression Classifier

3.3.2 Support Vector Classifier (SVM):

Used for both linear and nonlinear data, SVM is a supervised machine learning model. It creates an N-dimensional hyperplane, converts the original data onto it, and then ideally creates a split in the data that may be utilized for further analysis, such as classification or regression. [7] When compared to other computational methodologies, SVM can provide superior accuracy for long-term predictions in many functional applications.

SVM adopts the outstanding point that aids in creating the hyperplane. The computer is known as a Support Vector Machine because of these egregious situations, which are called support vectors. Recognize the covered chart, which is organized with a desired hyper plane using two specific classes.

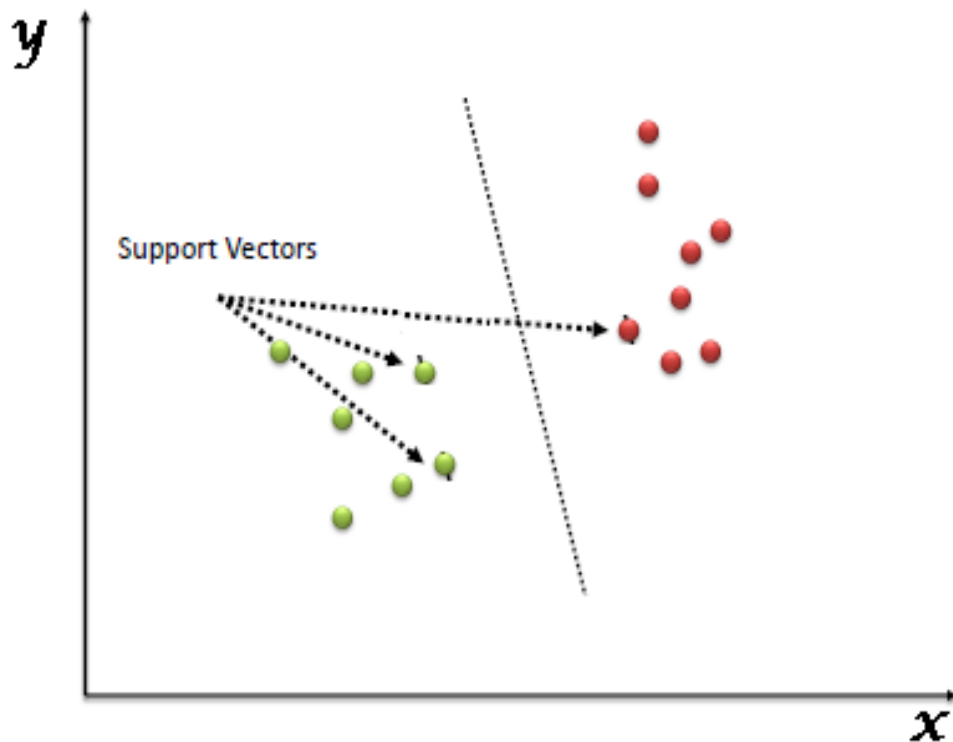


Figure 3.3.2: Support Vector Machine (SVM) Classification

3.3.3 Decision Tree (DT):

Decision trees are used to make predictions about the category of data by working backwards from the highest-level node in the tree. It categorizes information based on attributes' values as shown in the training set. In this approach, we are looking for a correlation in between values of the actual property and the attribute in the record (the real dataset). As soon as the comparison is finished, the process proceeds to the following node, moving down the appropriate branch. In order to maximize the value of the information gain, a decision tree method is employed, and an attribute is first split in order to get the largest information gain [].

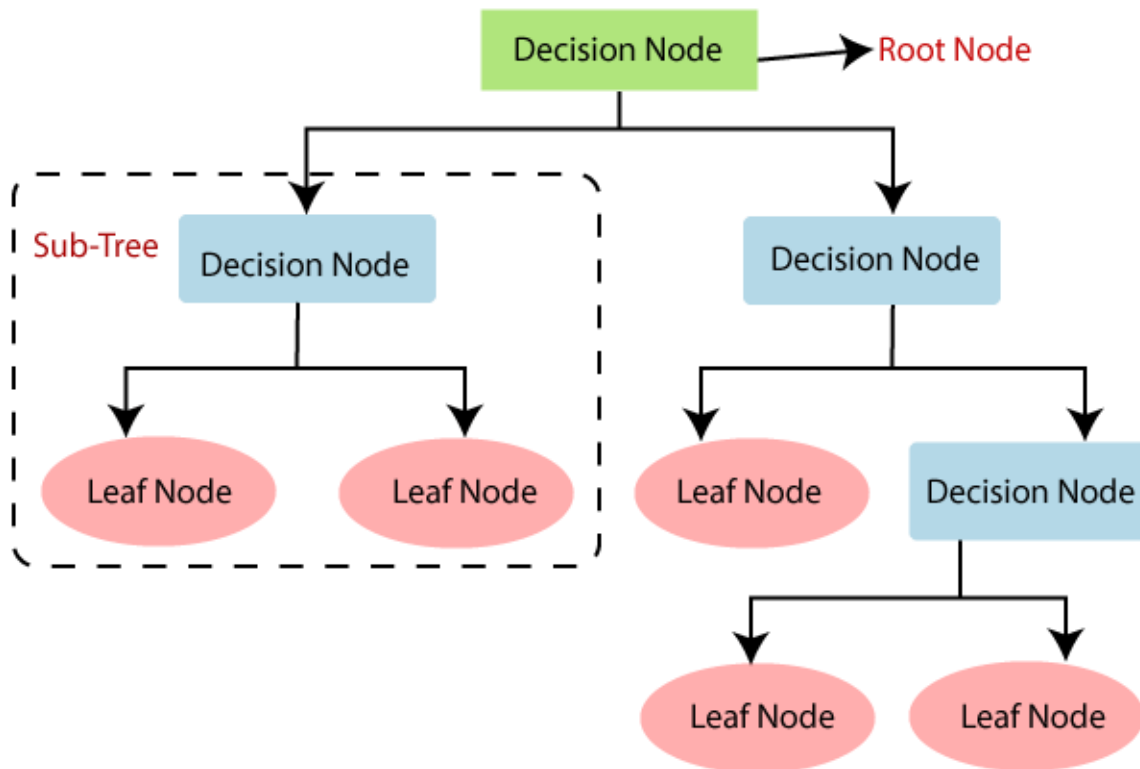


Figure 3.3.3: Decision Tree Classification

3.3.4 Random Forest:

Image processing techniques, recommendation engines, feature extraction, etc. are just some of the many areas where random forest has found usage. It takes random samples of data and generates decision trees. Then, it takes each tree's forecast and uses voting to choose the best option. One of random forest's strengths is its ability to provide a reasonably accurate measure of the significance of features. The fact that it is so simple to determine how much weight each feature should be given when making a forecast is the most impressive aspect.

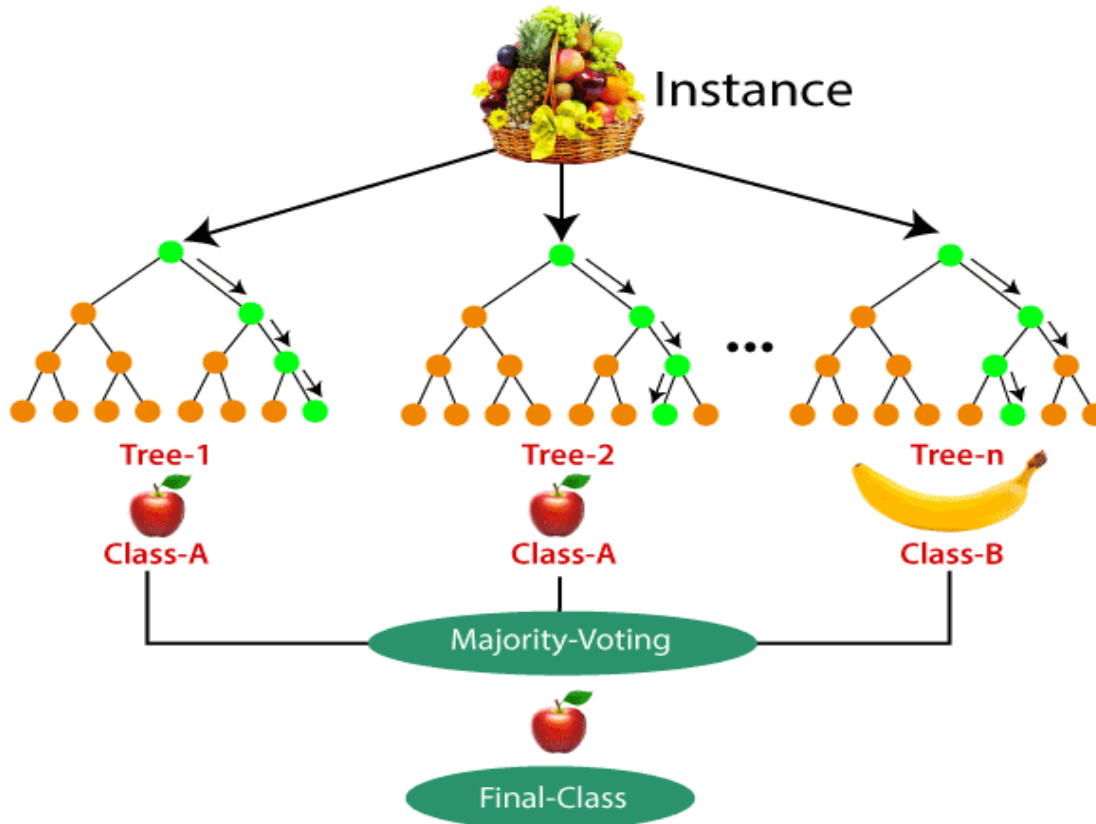


Figure 3.3.4: Random Forest Classification

3.3.5 Extreme Gradient Boosting:

With its sequence-following and ensemble-like performance, XGBoost is a model that deserves attention. It brings together groups of ineffective learners to improve prediction []. This approach is built as a distributed library that is optimized for performance and adaptability. As with other tree-building algorithms, Portable XGB uses Matching Score and Gain to determine which node splits are optimal.

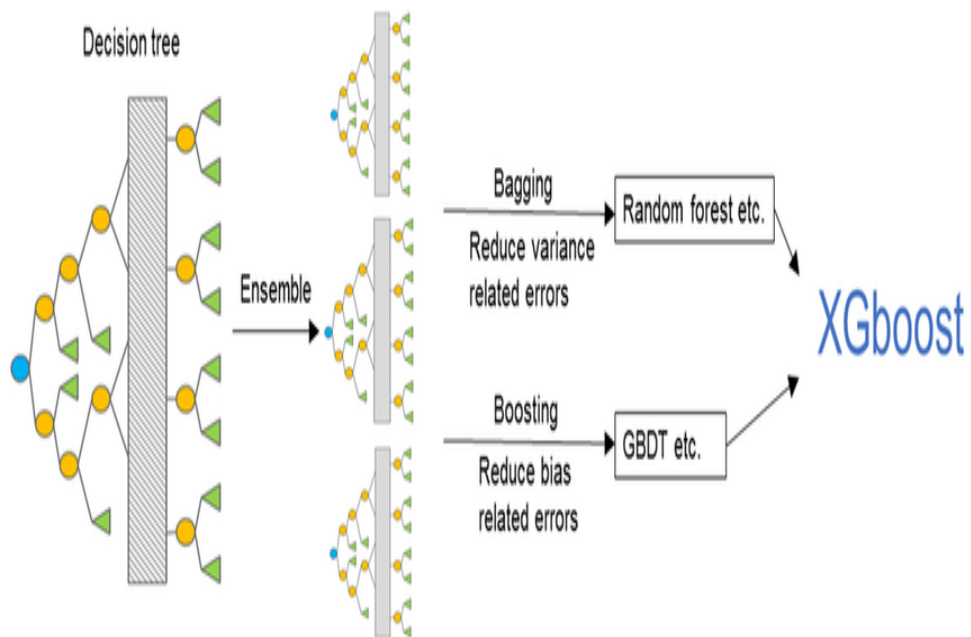


Figure 3.3.5: Extreme Gradient Boosting

3.3.6 K-Nearest Neighbors:

Because it is a non-parametric method, K-NN makes no assumptions about the data it is using. An example of a pattern recognition method is k-nearest neighbors (k-NN), which takes a set of training data points and stores their information based on how those points compare to others in n-dimensional space. With the goal of classifying unknown instances by locating the nearest data in patterns space forecast generated from Euclidean distance, K-NN seeks to identify the k nearest associated data points from future data that has not yet been observed. Afterwards, the distance between two points in the pattern space is determined by using the Euclidean distance $d(x, y)$. The sample's unknown population class is decided by a simple majority vote. After calculating the prediction distance, the Euclidean Formula is used to isolate the test data from the training data and the query point from the instances.

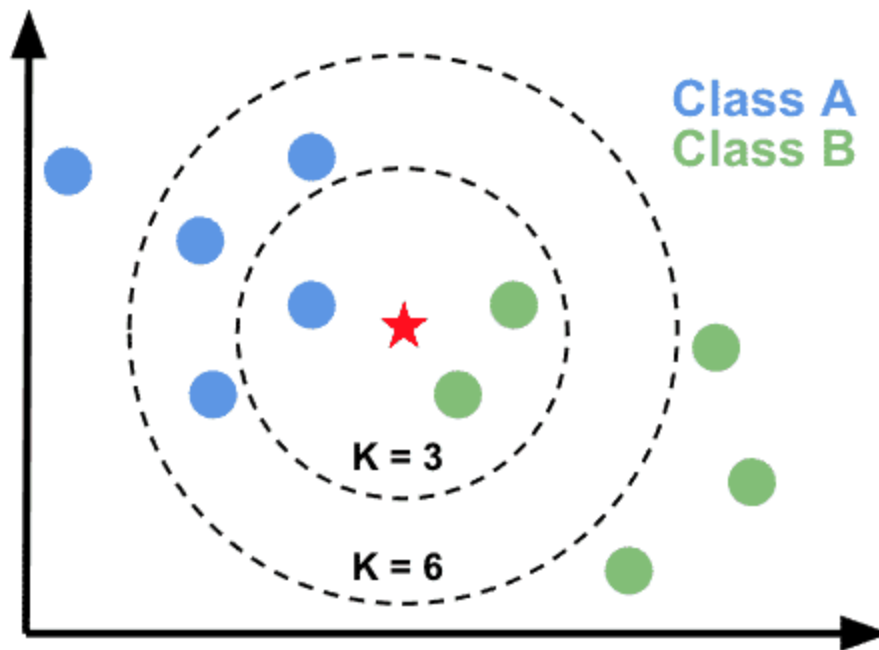


Figure3.3.6: K-Nearest Neighbors(KNN) Classification

3.3.7 Gradient Boosting:

For classification, regression, and other tasks, a prediction model is created using machine learning. Boosting Gradient Classifier deals with somewhat flimsy learning processes or flimsy ideas. It adds a number of weeks to it, improving the learner's or the hypotheses' strength. Gradient boosting (GB) is an ensemble boosting method that starts with "weak learners" in a "regression tree." Overall, because the GB model employs a sequential sampling strategy, it adds a second system that lowers the loss function [. The difference between the expected value and the actual value is determined by the loss function. GB improves accuracy by reducing bias and variation. The setting for least squares regression, which is straightforward to explain, is another fantastic aspect of this technique.

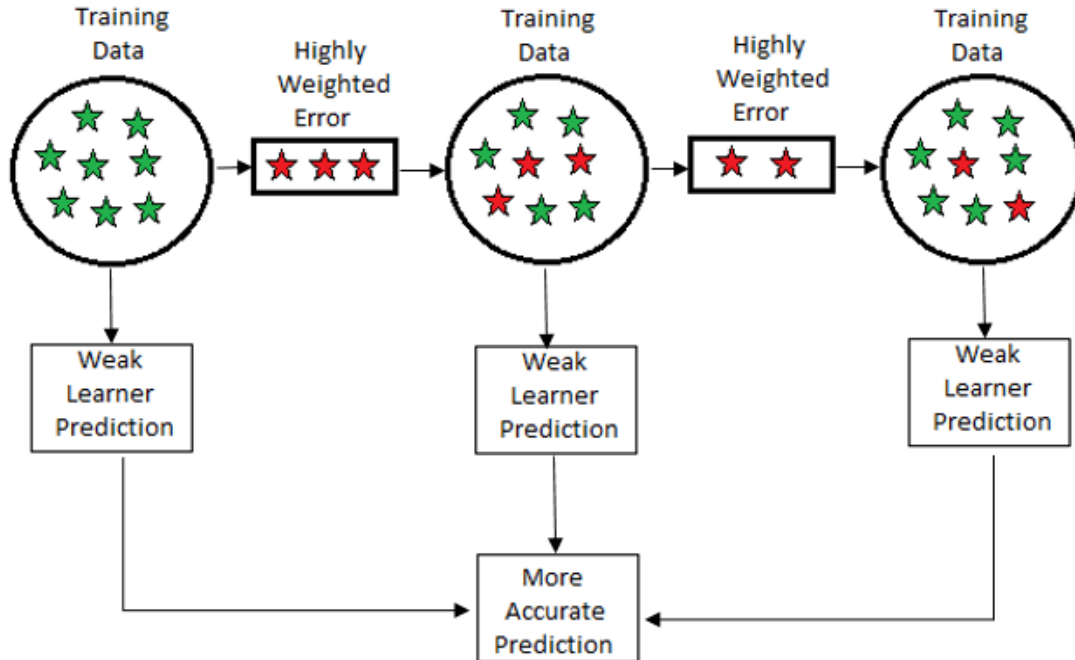


Figure 3.37: Gradient Boosting Classification

3.3.8 AdaBoost(ADB):

The Ada-Boost classifier is a potent technique that combines a number of classifiers with subpar performance to increase the resultant classifier's accuracy. As a result, it can offer a reliable categorization with a high accuracy rate. Iteration is a part of the ensemble method called AdaBoost[19]. ADB is used to ensure accurate predictions of aberrant observations by fixing the classifier weights [20] and training the data sample in each cycle.

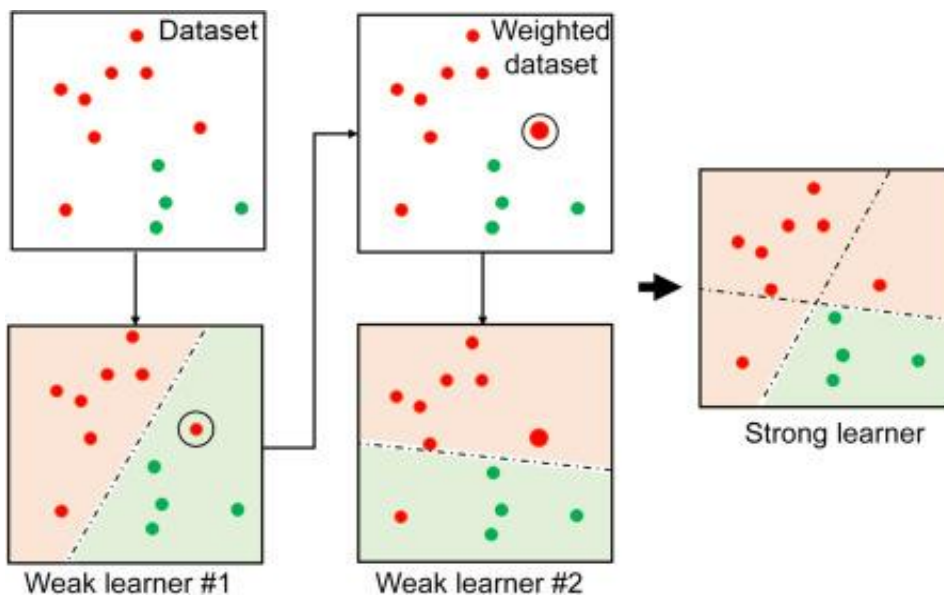


Figure3.3.8 Adaboost Classification

3.3.9 Bagging:

From the entire dataset, this bootstrap aggregating method generates several random bootstrap samples. A model is utilized as a base estimator to determine the evaluation report, and simple voting is employed to determine the total evaluation report. When you want to minimize differences of opinion while being biased, bagging is typically used. This occurs when predictions are equated over different regions of the given feature space. Bagging is helpful as long as you are improving a model's accuracy by using several versions of it that are proficient with different sources of data. Bagging is not advised for models with significant bias. Adaboost is employed in similar situations to propel onward motion and prevent the attainment of a high.

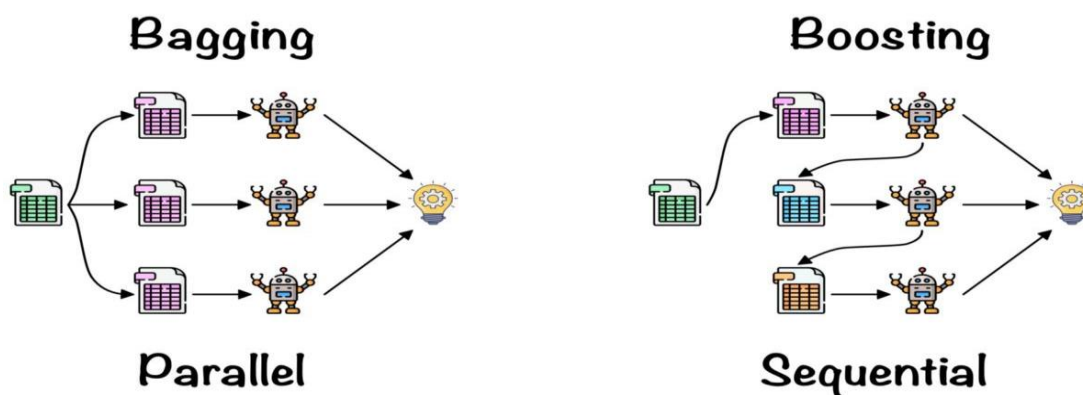


Figure3.3. 9: Bagging and Boosting Classifiers

3.3.10 Boosting:

Boosting uses several different types of basis learners, all of which made an error in classifying the original instance. Ultimately, it was a comparison of options that led to our conclusion. Adaboost is a more versatile algorithm. Actual forecasts are made by a boosting model using the most recent data. When a new inspection is obtained with its own set of characteristics, this data is processed by all of the models, each of which then makes its own forecast.

3.3.11 Research Subject and Instrumentation

All of the studies on the Google Collaborator platform are written in Python. You may use it to write and run Python scripts directly in your browser. Colab is a hosted Jupyter notepad that needs no initial configuration. All evaluations were carried out on a PC outfitted with just an Intel(R) Core(TM) i5-8250U processor running at 2.70GHz and 8GB of RAM, and operating under Windows 10 Pro 64-bit.

3.4 Dataset Description

The research makes use of the Preliminary phase diabetes prediction models dataset from the UCI Machine-Learning Repository [], which is widely regarded as the best such dataset available to the general public at large. Intended Publication: Journal of Informatics in Healthcare and Social Care. This month of December 2017 (in press). Implementing Machine Learning for DSM-5 Compliance in Autism Spectrum Disorder Screening. Medical and Healthcare Informatics 2017: Proceedings of the First Global Forum. Pages 1–6. ACM City of Taichung, Taiwan. There are 704 samples totaling 21 characteristics in this collection.

Table 01: Details of the dataset

Number of Attribute =21	Total number of Instances =704	Number of missing value = 0
Attribute	Description	Types of value
Id	such as a card number	integer value(1,2..)
A1_Score	integer	Nominal
A2_Score	integer	Nominal
A3_Score	integer	Nominal
A4_Score	integer	Nominal
A5_Score	integer	Nominal
A6_Score	integer	Nominal
A7_Score	integer	Nominal
A8_Score	integer	Nominal
A9_Score	integer	Nominal
Gender	Gender of the person(m/f)	Nominal
ethnicity	Randomly country	Nominal
jaundice	No/Yes	Nominal
autism	No/Yes	Nominal
contry_of_res	Randomly country	Nominal
used_app_before	No/Yes	Nominal
result	integer	Nominal
age_desc	Age of the person(M/F)	Nominal
relation	Self/parent	Nominal
Class/ASD	No/Yes	Nominal

3.5 Performance Evaluation Measure

To gauge how well the present model is doing, we may utilize a number of performance evaluation measures. These methods assess overall performance using hidden information.

Accuracy: That's the percentage of the test data that can be predicted with that much precision. A situation where precise measurements outperform easy access. It depends on just one variable. Inaccuracy refers to errors that occur repeatedly.

Precision:

It measures the extent to which a set of positive predictions matches actual data. Accuracy reveals the actual true share of all the times that they would have expected true.

Recall: It is the rate at which true positive observations are expected.

F1 Score: This is a middle ground between accuracy and memory recall.

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

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

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

$$F1-score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Figure3.5: Model for Calculating Performance Measures

CHAPTER 4

Experimental Results & Discussion

4.1 Experimental Result Analysis

The dataset for this study consists of 704 instances and 21 attributes, including 11 numerical features and 14 nominal features. This dataset's label encoding is Data that was categorized and was converted to a number. After testing and training, we used the test data to calculate accuracy and cross-validation scores for the classifier, and we compared them. Following that, we voted classifiers and 10-Fold cross validation were used to teach, evaluate, and test the classifiers. We calculated measures like as reliability, accuracy, f1-score, recall, and area under the curve (AUC) to assess the efficacy of the models. Only when someone exhibits autistic characteristics can they be classified positively. In contrast, if an individual does not acquire autism, they will be classified negatively.

Table 2: The Outcome of the New Algorithm

Model	Accuracy	precision	Recall	F1 score
RF	99.01%	99.12%	99.24%	99.01%
DT	99.22%	99.34%	99.54%	99.55%
SVM	99.60%	97.30%	99.20%	99.05%
XGB	99.75%	99.20%	99.30%	99.22%
ADB	98.0%	98.0%	99.0%	99.0%
GB	99.11%	99.21%	99.54%	99.23%
KNN	99.33%	99.23%	99.67%	99.21%
LR	99.69%	99.21%	99.50%	99.31%
GNB	99.43%	99.0%	99.34%	99.43%
GCV	99.34%	99.11%	99.22%	99.45%

First of all, utilizing the SVM classifier, the best accuracy of 99.68% has been achieved when comparing the results of traditional classifiers. Applying LR resulted in the highest

score of 99.69% when precision values were taken into account. Regarding the recall section, SVM again provided the highest score, which is 99.30%. SVM classifier has consistently performed at its best level of 99.60% in the case of the f1-score.

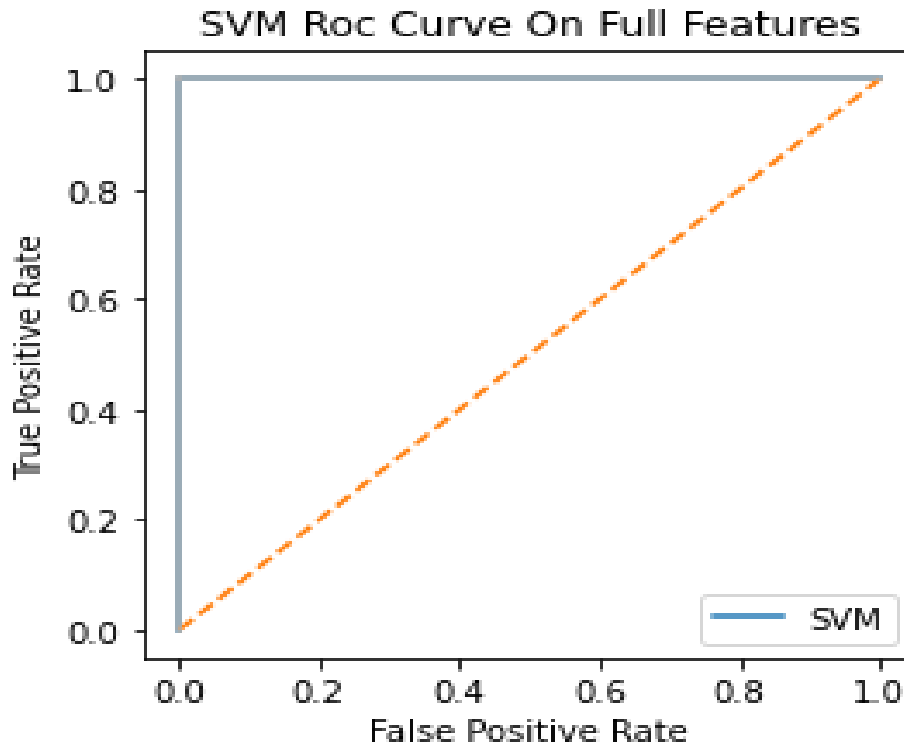


Figure 4.1: ROC Curve Analysis of Introduced Algorithm

The identical scenario where SVM obtained the greatest score of 1 is also shown by the ROC Curve in Figure 3. As a result, the SVM classifier may be declared to be the best traditional classifier based on the analysis presented aforementioned, and Fig. 9's extensive results and visual display.

Table 3: Experimental Results of Bagging Classifiers

Model	accuracy	precision	Recall	F1 score
RF	99.12%	99.20%	99.22%	99.23%
DT	99.60%	99.63%	99.67%	99.63%
SVM	99.68%	99.72%	99.31%	99.46%
XGB	99.60%	99.53%	99.41%	99.40%
ADB	99.42%	99.48%	99.51%	99.57%
GB	99.56%	99.50%	99.42%	99.45%
KNN	99.60%	99.71%	99.65%	99.67%
LR	99.65%	99.40%	99.38%	99.45%
GNB	99.63%	99.47%	99.61%	99.56%
GCV	99.37%	99.43%	99.41%	99.45%

Second, it is important to discuss the results of bagging classifiers. With precision at the forefront, the explanation demonstrates that the highest precision was achieved using the SVM classifier and stands at 99.68%. Following that, using KNN yielded the best score of 99.72 when precise performances were taken into account. SVM has regained its position in the recall segment, obtaining the best score of 99.68%. The similar situation has occurred once more in the f1-score case. observed that SVM has the best 99.72% performance. The ROC Curve's illustration has illustrated a distinct scenario in which The optimal score, 1, was agreed upon by SVM and DT. It is also clear that bagging has altered the classifiers' results in this case. While the accuracy of SVM and RF has declined and that of GB has remained constant, that of LR and DT has improved. Although the outcomes of NB and GB have reduced, the precision of LR, SVM, and DT has grown. The recall values have also changed, with XGB and DT increasing while LR, SVM, and GB being unable to do so. F1 rating for LR, and SVM have given the equivalent. Therefore, the SVM classifier

may be regarded as the best bagging classifier in light of the foregoing disagreement as including the entire findings shown graphically in Fig 3. The output from boosting classifiers ought to be the last consideration. First, regarding accuracy, the DT classification was utilized to get the highest possible rate of 99.6 percent. DT's precise ability is the highest, at 99.67%. efficiency of 99.6 percent Not everyone anticipated the ROC Curve to yield a perfect score of 1 for SVM, DT, and GB. The effects of Adaboost on classifier performance have also been studied. Starting with precision, DT and GB scores have risen while LR, SVM, and NB scores have dropped.

Table 4: Experimental Results of Boosting classifier

Model	accuracy	precision	Recall	F1 Score
RF	99.25%	99.34%	99.34%	99.45%
DT	99.60%	99.56%	99.32%	99.33%
SVM	99.67%	99.63%	99.36%	99.54%
XGB	99.61%	99.50%	99.43%	99.43%
ADB	99.42%	99.48%	99.57%	99.44%
GB	99.51%	99.32%	99.43%	99.30%
KNN	99.34%	99.45%	99.54%	99.20%
LR	99.02%	99.33%	99.43%	99.62%
GNB	99.56%	99.54%	99.12%	99.34%
GCV	99.34%	99.00%	99.25%	99.23%

While LR, XGB, and GB have not been able to increase precision, SVM and DT have. In the recall example, the scores of DT and RF have increased while the values of LR and SVM have declined, but ADB has remained constant. It has been noticed that the results from DT and GBAs have improved, whereas the results from LR, SVM, and KNN have

declined. While the ROC scores of GB have increased and those of DT, SVM, and LR have declined, respectively, while ADB has the opposite trend, Consequently, according Regarding the aforementioned disagreement and the detailed findings with visual representation shown in Table 4, the DT classifier may be considered the best boosting classifier. Our analysis shows that SVM (SVM) is the best performing model once boosting is applied, with an accuracy of 99.67%. In terms of processing time, SVM is the most efficient classifier.

CHAPTER 5

5.1 Conclusion and Future Work

One of the most difficult issues in autism informatics is the classification of autistic data. However, it is one of the oldest pieces of work in the field of research. As a result, different approaches have been put out by others. There are still many opportunities for improvement. So, using 10-fold cross-validation and a variety of pre-processing techniques, we demonstrated a correlative Measurement of the efficacy of a number of traditional classifiers, as well as modern variants like bagging and boosting on the diabetes dataset. By assigning the best parameters through hyperparameter tuning, the models perform better than previous models. We have carefully gone over the dataset. As a result, various performances employing various algorithms have been noted. They claim that we suggested the SVM classifier for this dataset. Overall, Because we additionally compared employing measures of performance assessment – an area where no other studies have made any headway despite their best efforts examined as many measures as we have—the exploration yields superior results than the existing approaches carried out by others. By including additional algorithms and using different datasets as our next step, we can finish our experiment.

Our future goals should include experimenting with various pre-processing methods and applying deep learning approaches.

Reference

- [1] Russell, G., Stapley, S., Newlove-Delgado, T., Salmon, A., White, R., Warren, F., ... & Ford, T. (2022). Time trends in autism diagnosis over 20 years: a UK population-based cohort study. *Journal of Child Psychology and Psychiatry*, 63(6), 674-682..
- [2] Marquez-Garcia, A. V., Magnuson, J., Morris, J., Iarocci, G., Doesburg, S., & Moreno, S. (2022). Music therapy in autism spectrum disorder: A systematic review. *Review Journal of Autism and Developmental Disorders*, 9(1), 91-107.
- [3] Andreo-Martínez, P., Rubio-Aparicio, M., Sánchez-Meca, J., Veas, A. and Martínez-González, A.E., 2022. A meta-analysis of gut microbiota in children with autism. *Journal of autism and developmental disorders*, 52(3), pp.1374-1387.
- [4] Lundström, S., Taylor, M., Larsson, H., Lichtenstein, P., Kuja-Halkola, R. and Gillberg, C., 2022. Perceived child impairment and the 'autism epidemic'. *Journal of Child Psychology and Psychiatry*, 63(5), pp.591-598.
- [5] Afif, I. Y., Farkhan, M., Kurdi, O., Maula, M. I., Ammarullah, M. I., Setiyana, B., Jamari, J. and Winarni, T. I., 2022. Effect of short-term deep-pressure portable seat on behavioral and biological stress in children with autism spectrum disorders: A pilot study. *Bioengineering*, 9(2), p.48.
- [6] Zeidan, J., Fombonne, E., Scolah, J., Ibrahim, A., Durkin, M. S., Saxena, S., Yusuf, A., Shih, A. and Elsabbagh, M., 2022. Global prevalence of autism: a systematic review update. *Autism Research*, 15(5), pp.778-790.
- [7] Zhao, Z., Zhu, Z., Zhang, X., Tang, H., Xing, J., Hu, X., ... & Qu, X. (2022). Identifying autism with head movement features by implementing machine learning algorithms. *Journal of Autism and Developmental Disorders*, 52(7), 3038-3049.
- [8] Ramani, R. Geetha, and G. Sivagami. "Parkinson disease classification using data mining algorithms." *International journal of computer applications* 32.9 (2011): 17-22.
- [9] P. Ghosh, S. Azam, A. Karim et al., "Use of Efficient Machine Learning Techniques in the Identification of Patients with Heart Diseases", In 5th ACM International Conference on Information System and Data Mining (ICISDM2021), pp. 14 –20, May 2021, <https://doi.org/10.1145/3471287.3471297>
- [10] Shamrat, F. J. M., Chakraborty, S., Billah, M. M., Das, P., Muna, J. N., & Ranjan, R. (2021, June). A comprehensive study on pre-pruning and post-pruning methods of decision tree classification algorithm. In 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 1339-1345). IEEE.
- [11] Asra, Taufik, Ahmad Setiadi, Mahmud Safudin, Endah Wiji Lestari, Nila Hardi, and Doni Purnama Alamsyah. "Implementation of AdaBoost Algorithm in Prediction of Chronic Kidney Disease." In 2021 7th International Conference on Engineering, Applied Sciences and Technology (ICEAST), pp. 264-268. IEEE, 2021.

- [12]Almasoud, M., & Ward, T. E. (2019). Detection of chronic kidney disease using machine learning algorithms with the least number of predictors. *International Journal of Soft Computing and Its Applications*, 10(8)
- [13]Zhao, Z., Zhu, Z., Zhang, X., Tang, H., Xing, J., Hu, X., ... & Qu, X. (2022). Identifying autism with head movement features by implementing machine learning algorithms. *Journal of Autism and Developmental Disorders*, 52(7), 3038-3049.
- [14]P. Ghosh, S. Azam, A. Karim, M. Hassan, K. Roy, M. Jonkman, "A Comparative Study of Different Machine Learning Tools in Detecting Diabetes", *Procedia Computer Science*, Vol. 192, PP. 467-477, 2021, <https://doi.org/10.1016/j.procs.2021.08.048>.
- [15]NehaVerma, Dr BP. "Performance and Comparison of Classification Algorithms of MLwith Comparative Mean for Heart Disease Prediction." *Annals of the Romanian Society for Cell Biology* 25, no. 6 (2021): 12791-12813.
- [16]P. Ghosh et al., "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques," in *IEEE Access*, vol. 9, pp. 19304-19326, 2021, doi: 10.1109/ACCESS.2021.3053759.
- [17]S. Zobaed, F. Rabby, I. Hossain, E. Hossain, S. Hasan, A. Karim, and K. Md. Hasib, "Deepfakes: Detecting forged and synthetic media content using machine learning," *Advanced Sciences and Technologies for Security Applications*, pp. 177–201, 2021.
- [18]Ploner, Tina, Steffen Heß, Marcus Grum, Philipp Drewe-Boss, and Jochen Walker. "Using gradient boosting with stability selection on health insurance claims data to identify disease trajectories in chronic obstructive pulmonary disease." *Statistical Methods in Medical Research* 29, no. 12 (2020): 3684-3694.
- [19][1]Zeidan, J., Fombonne, E., Scolah, J., Ibrahim, A., Durkin, M. S., Saxena, S., ... & Elsabbagh, M. (2022). Global prevalence of autism: a systematic review update. *Autism Research*, 15(5), 778-790
- [20]Pellicano, E., & den Houting, J. (2022). Annual Research Review: Shifting from ‘normal science’ to neurodiversity in autism science. *Journal of Child Psychology and Psychiatry*, 63(4), 381-396.
- [21]Silverman, J.L., Thurm, A., Ethridge, S.B., Soller, M.M., Petkova, S.P., Abel, T., Bauman, M.D., Brodtkin, E.S., Harony-Nicolas, H., Wöhr, M. and Halladay, A., 2022. Reconsidering animal models used to study autism spectrum disorder: current state and optimizing future. *Genes, brain and behavior*, p.e12803.
- [22]Fombonne, E., & Zuckerman, K. E. (2022). Clinical profiles of Black and White children referred for autism diagnosis. *Journal of autism and developmental disorders*, 52(3), 1120-1130.

Nazrul_Plagiarism_Report

ORIGINALITY REPORT

20%

SIMILARITY INDEX

19%

INTERNET SOURCES

9%

PUBLICATIONS

11%

STUDENT PAPERS

PRIMARY SOURCES

1

Submitted to Daffodil International University

Student Paper

5%

2

dspace.daffodilvarsity.edu.bd:8080

Internet Source

5%

3

www.researchgate.net

Internet Source

2%

4

link.springer.com

Internet Source

1%

5

doctorpenguin.com

Internet Source

1%

6

Amparo V. Marquez-Garcia, Justine Magnuson, James Morris, Grace Iarocci, Sam Doesburg, Sylvain Moreno. "Music Therapy in Autism Spectrum Disorder: a Systematic Review", Review Journal of Autism and Developmental Disorders, 2021

Publication

1%

7

doi.org

Internet Source

1%

