

**PREDICTION OF AGORAPHOBIA DISEASE BASED ON MACHINE LEARNING  
TECHNIQUES**

**BY**

**NAME: Israt Jahan Intia**

**ID: 181-15-10751**

**NAME: Khondoker Sangida Ferdous**

**ID: 181-15-10516**

**NAME: Nahid Hasan Hridoy**

**ID: 181-15-10663**

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

Supervised By

**Md. Zahid Hasan**

Associate Professor

Department of Computer Science and Engineering

Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**

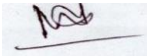
**DHAKA, BANGLADESH**

**4 JANUARY 2022**

## APPROVAL

This Project titled “**PREDICTION OF AGORAPHOBIA DISEASE BASED ON MACHINE LEARNING TECHNIQUES**”, submitted by **Israt Jahan Intia , Khondoker Sangida Ferdous** and **Nahid Hasan Hridoy** 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 (BSc) and approved as to its style and contents. The presentation has been held on 4 January 2022.

### BOARD OF EXAMINERS

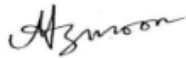


---

**Dr. Md. Ismail Jabiullah**  
**Professor**

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

**Chairman**



---

**Nazmun Nessa Moon (NNM)**  
**Assistant Professor**

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

**Internal Examiner**



---

**Aniruddha Rakshit (AR)**  
**Senior Lecturer**

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

**Internal Examiner**



---

**Dr. Md Arshad Ali**  
**Associate Professor**

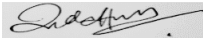
Department of Computer Science and Engineering  
Hajee Mohammad Danesh Science and Technology  
University

**External Examiner**

## DECLARATION

We hereby declare that, this project has been done by us under the supervision of Name, Designation, 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. Zahid Hasan**  
Associate Professor  
Department of CSE  
Daffodil International University

### Submitted by:



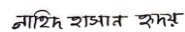
---

**Israt Jahan Intia**  
ID: 181-15-10751  
Department of CSE  
Daffodil International University



---

**Khondoker Sangida Ferdous**  
ID: 181-15-10516  
Department of CSE  
Daffodil International University



---

**Nahid Hasan Hridoy**  
ID: 181-15-10663  
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 successfully.

We really grateful and wish our profound our indebtedness to Supervisor Name, Designation, 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 **Prof. Dr. Touhid Bhuiyan**, and 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**

In today's world, agoraphobia has become a very common disorder. Agoraphobia disease is a group of mental illnesses marked by intense emotions of fear and anxiety. Majority of people are unaware of the condition. It is essential to recognize it early on so that doctors can give better treatment and prevent it from progressing into a significant problem. Recently machine learning algorithms can be used to assess a patient's history and find abnormalities by simulating human thinking or drawing logical inferences. This study reviews the basic ideas and applications of machine learning algorithms in predicting anxiety disorder types. We try to detect agoraphobia in the primary stage in this research. We primarily used three feature selection strategies, as well as a variety of classification algorithms, to accomplish this. We use some classification methods include the Naive Bayes, Random Forest ,Decision Tree, KNN, and Support Vector Machine (SVM). After testing random forest classification method in achieving higher accuracy with 98.02% accuracy than all other classification methods currently in use.

## TABLE OF CONTENTS

<b>CONTENTS</b>	<b>PAGE</b>
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
<b>CHAPTER</b>	
<b>CHAPTER 1: INTRODUCTION</b>	<b>1-4</b>
1.1 Introduction	01
1.2 Motivation	02
1.3 Problem Definition	02
1.4 Research Question	03
1.5 Research Methodology	03
1.6 Research Objective	03
<b>CHAPTER 2: BACKGROUND</b>	<b>5-7</b>
2.1 Introduction	05
2.2 Related work	04
<b>CHAPTER 3: RESEARCH METHODOLOGY</b>	<b>8-15</b>
3.1 Introduction	08
3.2 Data Collection	09
3.3 Preprocessing	09
3.4 Dataset	10
3.5 Algorithm Implementation	10
3.5 Statistical Analysis	14
3.6 Evaluation	15

<b>CHAPTER 4: RESULT COMPARISON AND DISCUSSION</b>	<b>16-20</b>
4.1 Experimental Setup	16
4.2 Experimental Result and Analysis	16
4.2 Experimental Result and Analysis	17
4.4 Confusion Matrix	19
<b>CHAPTER 5: Conclusion and Future Work</b>	<b>21</b>
5.1 Introduction	21
5.2 Conclusion	21
5.3 Implication for Future Study	21
<b>REFERENCES</b>	<b>22</b>
<b>APPENDIX</b>	<b>24</b>

## LIST OF FIGURES

<b>FIGURES</b>	<b>PAGE NO.</b>
Figure 3.1: Methodology diagram	08
Figure 3.2: Agora phobia affected and not affected percentage	10
Figure 3.3: functionality of knn algorithm	11
Figure 3.4: working procedure of decision tree algorithm	12
Figure 3.5: Random Forest algorithm	13
Figure 3.6: All collected ages	14
Figure 3.7: Affected age ratio.	15
Figure 3.8: Male and female ratio	15
Figure 4.1: Roc comparison between svm, knn, random forest	18
Figure 4.2: Roc comparison between svm, knn, random forest	18
Figure 4.3: Comparison between real and predicted class	19
Figure 4.4: Confusion Matrix	20



## LIST OF TABLES

<b>TABLE NO.</b>	<b>PAGE NO.</b>
Table 3.1: Parameter usage	11
Table 4.1: Accuracy table	16

# CHAPTER 1

## INTRODUCTION

### 1.1 Introduction

Agoraphobia is an anxiety condition characterized by a strong dread of being overpowered or unable to flee or seek help. When people with agoraphobia are in a stressful situation, they often experience signs of a panic attack, such as a racing heart and nausea. They may also exhibit similar signs prior to entering the feared circumstance. In other situations, the disease is so bad that people avoid going to the bank or the grocery store and instead stay at home for the majority of the day. Agoraphobia symptoms are comparable to those of a panic attack. Like symptoms: chest pain or a fast heart rate, fear or a shaky sensation, breathing difficulties or hyperventilation dizziness or lightheadedness, chills that come on suddenly or flushing (red, hot face), upset stomach.

In Bangladesh, Agoraphobia disease affects 6.5-31% people in Bangladesh [1], according to specialists. Agoraphobia disease affects women more than it does males. In one study, those over the age of 60 were shown to have a higher prevalence. Between the ages of 18 and 35, it usually begins. 85 percent of those with agoraphobia will also have another psychiatric condition at some point in their lives, such as panic disorder, social anxiety disorder, particular phobia, generalized anxiety disorder (GAD), substance use disorder.

Machine learning has become a very widespread tool for diagnosing many diseases in recent years. It is very simple and powerful to predict illnesses using machine learning algorithms [2]. We utilized feature selection and classification algorithms to forecast Agoraphobia in the initial stage. We collect data from online survey. We use several classification techniques include the Naive Bayes, Random Forest, Decision Tree, KNN, and Support Vector Machine. The three major feature selection methodologies we use are principal component analysis (PCA), univariate feature selection (UFS), and recursive feature elimination (RFE) (NB). Finally, we discovered that the RFE feature selection strategy improved accuracy regardless of the classification technique used.

### 1.2 Motivation

Agoraphobia disease affects around 3 percent people in Bangladesh. Day by day this occurs are increasing. The majority of people are unaware that they have this disease. Agoraphobia can become more severe and difficult to treat if it is not treated early. However, good treatment can help function if suffer from agoraphobia. But in our country most of people can't understand that they have Agoraphobia. As a result, they do not receive proper therapy, putting their lives being at risk.

Machine learning technologies are now widely used in the medical industry, and the application of machine learning techniques in early stage disease identification is becoming increasingly widespread. The selection of appropriate attributes is critical for making an accurate prediction. Early disease identification is challenging due to the difficulty of extracting attribute selection from data. It's possible that if we don't pick the right attribute we'll get a negative result. Our main goal is to employ feature selection techniques to select acceptable qualities from data to aid in the early detection of disease with high accuracy.

### **1.3 Problem Definition**

Machine learning is a very significant topic in the medical area in current modern day. Machine learning will aid in the improvement of the medical industry by detecting diseases early. Firstly we need to find out the main problem for this disease and then we need to discover the best solution. Without find out the exact reason it can cause harmful. When it comes to detecting disease, we must assess whether the method we're using is reliable. Take a short look at hospitals and diagnostic centers to discover more about the issues surrounding thyroid disease (Agoraphobia) and how it develops.

### **1.4 Research Questions**

The following are some questions on which this thesis is focused:

- What is the present situation of Agoraphobia disease in Bangladesh?
- Which classification is best for predicting disease in its early stages?
- How to labeling the dataset?
- How can this work benefit people?
- What is the most appropriate method for categorizing the disease?

## **1.5 Research Methodology**

The manner we plan to conduct your study is referred to as your research technique. This chapter discusses how we plan to handle concerns including data collection methods, statistical analysis, participant observations etc .We collected the Agoraphobia Data Set, data cleaning and picking attributes, and feature selection procedures in this section of our study report Train the model, then use a classification method (SVM, DT, RF, KNN, NB) to determine the outcome.

## **1.6 Research Objectives**

There are some advantages of using Machine learning techniques in disease detection. There are some objectives of using Machine learning techniques.

The following are some of the goals of machine learning techniques:

- Predict the beginning of Agoraphobia disease using data from Bangladesh.
- Which feature selection strategies get the best results?
- Also, figure out the best classification algorithm.
- Increase the accuracy of disease detection by using an artificial intelligence-based system.

## **1.7 Report Layout**

Chapter 1 will cover the following topics: introduction, motivation, problem definition, research question, research methodology, and our program's predicted conclusion. In this part, we also explain why we decided to conduct this study.

Chapter 2 will explore the history of this study, as well as related studies and present state from the perspective of Bangladesh. It contains both a context analysis and a brief summary of the work.

Chapter 3: will describe research methodology. This chapter goes through the technique or workflow in great detail. This section will explain how the study was performed.

Chapter 4: will discuss performance of the proposed model.

Chapter 5: The goal of this chapter is to look at the results. It includes a graph as well as the research findings.

Chapter 6: This chapter is included in the report's conclusion. This section summarizes the model's performance. A comparison of accuracy is also included in this section. This section also covers

the model's web implementation and output. A review of the work's limitations finishes the chapter. It was also encoded with information about future work.

Chapter 7 contains all of the references we utilized during our study.

## **CHAPTER 2**

### **BACKGROUND**

#### **2.1 Introduction**

There is no work or science in our country that can accurately predict disease and have a solution. As a result, the context is the current state of infant mortality and the use of Machine Learning in Bangladesh's medical field.

#### **2.2 Related Works**

Machine learning is extensively applied to resolving problems that are subject to forecasting. For taking actions toward the death of children, a lot of research done with machine learning. ML has made this strategy extremely convenient.

Depression is a severe danger to one's personal and society's health these days. Tens of millions of people suffer from depression each and every year. But Just a small minority receives medical help. They [3] looked at the possibilities of detecting and diagnosing major depression in people on social media. They started with a list of people on social media who claimed to have been diagnosed with clinical depression using a traditional psychometric assessment. They analyzed behavioral variables relating to social emotion, ego, network language and linguistic types, and antidepressant medicine references in their social media activity for a year before they suffered depression. . According to their data, those who are depressed had less social engagement, more negative feelings, more self-attentional focus, more relational and health problems, and a higher expression of religious ideas. They developed an SVM classifier that may forecast the chance of a person acquiring depressed before it occurs. The classifier produced good results, with a classification accuracy of 70%.

These days, Psychological health concerns such as anxiety, sadness, and stress have become quite frequent. In this paper the author [4], described Machine learning algorithms were used to predict anxiety, depression, and stress. In order to use these algorithms, they collected data from working and jobless people from diverse cultures and groupings (DASS 21). Then five different categorization approaches were used: K-Nearest Neighbour (KNN) ,Random Forest Tree (RFT), Nave Bayes, Decision Tree (DT) and Support Vector Machine (SVM). Despite the fact that

Random Forest was shown to be the best model, the accuracy of naive Bayes was found to be the highest.

Anxiety disorders are a group of mental illnesses marked by intense emotions of dread and anxiety. Recently, advances in machine learning techniques have aided in the development of systems that help clinicians detect mental diseases and provide patient care. Using machine learning techniques, a comparative literature search was undertaken on studies for the prediction of certain forms of anxiety disorders described by Arif M. et.al. [5]. Medical specialists have found clinical decision support systems based on data-driven classifier design to be advantageous in a variety of ways. The rise of social media in the recent decade, as well as wearable sensor technology, has opened up new avenues for bettering clinical decision-making. Still, there is scope for change in diagnostic quality, and new treatment scenarios can be used to enhance overall population mental health and reduce mental disease tendencies that lead to serious outcomes such as suicides. Knowing the incidence of depression, anxiety, and mood disorders in the community might also help governments make better decisions.

Artificial intelligence is one of the most significant upcoming disruptions in healthcare [6]. AI aids in the prediction of disease patients for medical procedures. The use of AI in healthcare is vast, and it is used by not just doctors, but also patients, the pharmaceutical industry, health services, insurance companies, and medical institutes. AI systems will evolve to the point where they will be able to do a greater range of tasks without the need for human intervention. While inspiring and driving innovation in the field, AI is created and applied in a transparent and public-interest-friendly manner. AI assists in patient monitoring, screening, and clinical and medical investigations.

Using Feature Selection And Classification Techniques, Thyroid Disease Can Be Forecast Early. One of the most prevalent diseases among Bangladeshi women is thyroid disease. Because the majority of people are unaware of the condition, it is fast spreading. a life-threatening illness It is critical to discover it in the early stages. so that doctors may better treat patients in the basic stage. The author of this paper [7] used three feature selection techniques, as well as a variety of classification methods. They used feature selection techniques like Recursive Feature Selection Univariate Feature Selection and Principal Component Analysis (PCA), as well as classification algorithms like Decision Tree(DT), Naive Bayes, Support Vector Machine(SVM), Logistic

Regression(LR), and Random Forest(RF). According to the statistics, the RFE feature selection approach helps us achieve constant 99.35 percent accuracy for all four classification algorithms.

Many people use online communities to talk about mental health difficulties, which provides opportunity for fresh understanding of these communities. The author T. Nguyen et. al. [8] looked into online depression communities and looked at how they differed from other online communities. Then They used machine learning and statistical techniques to discriminate online messaging between depression and control groups using mood, psycholinguistic processes, and content subjects extracted from postings generated by members of these communities. the written material, and other factors are all taken into consideration. The writing styles of two people are discovered to be drastically diverse. Various forms of communities These two groups were found to be considerably different in all respects, including affect, textual substance, and writing style. Furthermore, latent subjects were found to be a better predictor of depressed communities than linguistic characteristics. Data mining of online blogs, according to this study, can be used to identify useful data for depression research. The discovery highlights machine learning's potential in psychiatric therapy and research.

Machine learning techniques are commonly employed to diagnose defects in order to ensure that systems operate safely and reliably described by R Razavi-Far et. al. [9]. With partially labeled data, when only a small number of labeled observations are recorded alongside a large number of unlabeled observations, semi-supervised learning, among other approaches, can help in the detection of problematic states and decision making. The author of this paper[ref] investigates a variety of semi-supervised dimensionality reduction algorithms in order to find the optimal combination of classification and dimensionality reduction strategies for induction motor bearing defect diagnosis. Based on the testing results, FME produces the best characteristics for the decision-making phase. The ASSEMBLE algorithm, on the other hand, can best match FME with its k-nearest neighbors base learner. Their combination produces exceptional bearing problem diagnostic accuracy in induction motors.



## CHAPTER 3

### RESEARCH METHODOLOGY

#### 3.1 Introduction

The forms or methodologies utilized to discover, select, handle, and analyze data on a subject are alluded to as inquire about technique. The technique parcel of a inquire about article makes a difference the peruser to equitably look at the by and large legitimacy and unwavering quality of the consider. The methodology covers an absolute of five steps which conclude our research that is displayed in 3.1 The steps are the following:

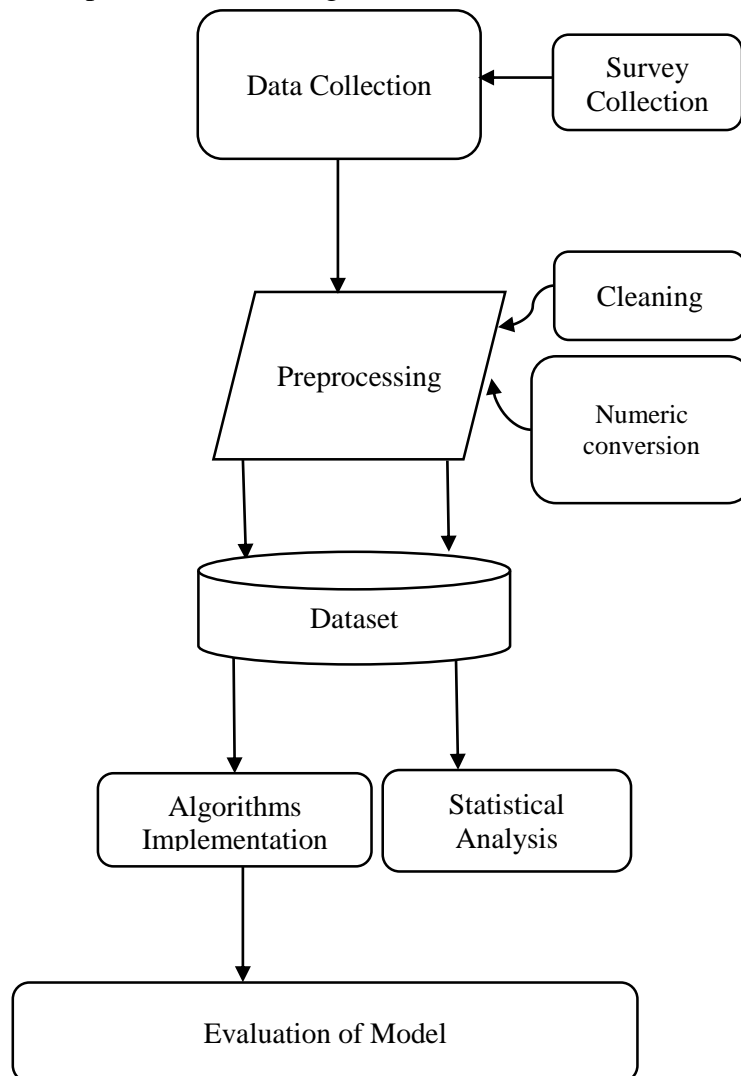


Figure 3.1: Methodology diagram

## 3.2 Data Collection

We surveyed and gathered almost 1685 pieces of information. This poll was conducted on persons of various socioeconomic backgrounds. We chose EnclosedSpace, MeetingRoom, SnadingInCrowd, and HomeAlone as common for all because these four parameters are extremely significant for everyone because if any individual has either of the two problems, he or she is likely to have agoraphobia disease. Our study includes not only disease prediction but also a statistical examination of the characteristics that cause these types of diseases. As a result, each of our chosen characteristics is critical to our job. As a result, we carefully gathered these parameters using the following questionnaire:

1. Are you Male or Female?
2. What is your age?
3. Is your father die when you are child?
4. Are you sexually abused by anyone?
5. Do you have any Bereavement?
6. Are you divorced?
7. Are you lost your job?
8. Do you have any chest pain?
9. Do you have shortness of breath problem?
10. Are you stress when you are leaving from house?
11. Do you rely other's when you are going to shopping?
12. Do you afraid to see open space?
13. Do you afraid to see enclosed space?
14. Do you afraid to see meeting room?
15. Do you afraid when you standing in crowd?
16. Do you afraid when you stay in home alone?

## 3.3 Preprocessing

1. Data cleaning: Name and location are included in our preliminary data gathering, however they are not required for training our model. As a consequence, we removed all of the primary dataset's names and locations.

2. Transformation: Data transformation is an essential component of every dataset for obtaining high-quality information. [1] Because machine learning algorithms are incapable of comprehending strings, all strings must be translated to numbers. In this situation, we used 1 for yes and 0 for no, and 2 for unemployed.

### 3.4 Dataset

Our dataset contains total 17 features. 16 of them are independent features and 1 target or dependent feature. Each of entity of our dataset are divided in to two classes one is agoraphobia positive and agora phobia negative. The figure 3.4.1 represents our dataset representation. 34.1% people of our dataset affected by agora phobia disease. and 65.9% people are not affected by this disease.

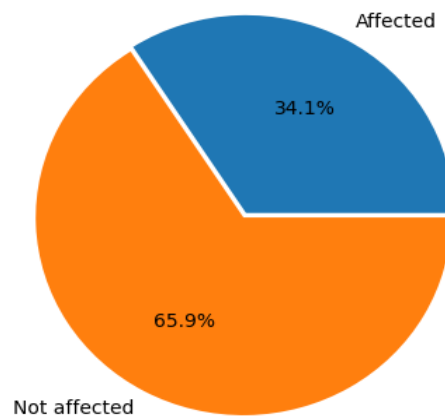


Figure 3.2: Agora phobia affected and not affected percentage

### 3.4 Algorithm Implementation

Using algorithms, we achieved the greatest 98.02 percent accuracy using Random Forest by using 30 percent data usage rate. The other three algorithms performed admirably as well. Because Random Forest was the most efficient. We have decided to use this algorithm for final prediction of agoraphobia. The parameter usages of each algorithm are shown in Table 3.4.1. We use the hyper parameter tuning approach to develop the algorithm and select the parameters. And then select the parameter that allows the algorithm to deliver the most accurate results.

TABLE 3.1. PARAMETER USAGE

Algorithms	Details
KNN	n_neighbors = 3 , metric = 'minkowski'
Naïve Bayes Classifier	Random_state = 1, classifier = GaussianNB
Decision Tree	Criterion = 'gini', min_samples_split = 2, random_state = 42
SVM	C=1.0, degree=3, probability=False, coef0=0.0, shrinking=True, kernel='rbf', gamma='scale'
Random Forest	n_estimators=100, min_samples_leaf = 1

### 3.4.1 K-Nearest Neighbor

K-Nearest Neighbor could be a fundamental Machine Learning strategy that uses the Administered Learning approach. It may be a separate learning calculation. most of the time it works with Euclidean remove between any two point. The K-NN strategy assumes similarity between the new case/data and existing cases and places the unused case within the category that's most comparative to the existing categories. The K-NN calculation keeps up all existing information and classifies modern information focuses based on similitudes. This suggests that when new information is produced, it may be rapidly categorized into a well-suited category utilizing the K- NN strategy. Figure 3.3 represents the functionality of knn algorithm [11].

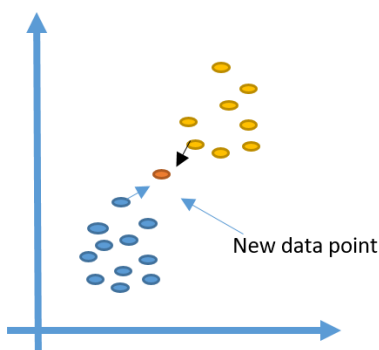


Figure 3.3: Functionality of knn algorithm

### 3.4.2 Naïve Bayes Classifier

It is a classification strategy based on Bayes' Theorem and the assumption of predictor independence. A Naive Bayes classifier, in basic words, posits that the existence of one feature in a class is independent to the presence of any other feature. Although Naive Bayes is a simple methodology, it has the potential to outperform very complicated classification systems. Equation 1 represents the formulae of naïve bayes classifier algorithm [12].

$$P(c|x) = \frac{P(x|c)p(c)}{P(x)} \quad (1)$$

### 3.4.2 Decision Tree

The choice tree Strategy may be a directed machine learning calculation. It may be utilized for both classification and relapse issues. The reason of this approach is to create a demonstrate that predicts the esteem of a target variable, for which the choice tree utilizes the tree representation to illuminate the issue, where the leaf hub compares to a lesson name and characteristics are spoken to on the inner hub of the tree [13].

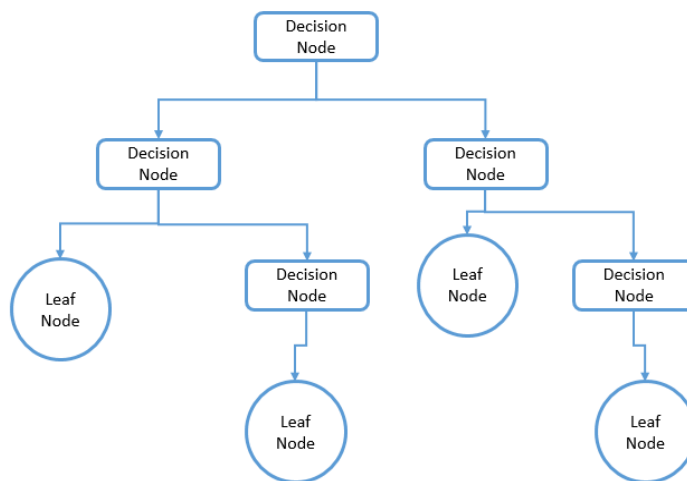


Figure 3.4: working procedure of decision tree algorithm

The root hub symbolizes the whole populace or test, which is at that point partitioned into two or more homogenous sets. The method of isolating a hub into two or more sub-nodes is known as part. Decision Node: When a sub-node divides into other sub-nodes, it is referred to as a decision node. Leaf/Terminal Node: Nodes that do not divide are referred to as Leaf or Terminal nodes.

Pruning: When we eliminate sub-nodes from a decision node, this is referred to as pruning. We may call it the inverse process of splitting.

Branch / Sub-Tree: A branch or sub-tree is a segment of a tree.

### 3.4.3 Random Forest algorithm

Random Forest is a supervised machine learning technique that is commonly used to solve classification issues. It also supports both categorical and continuous input and output variables, similar to a decision tree. In the CART model, a single tree grows in the random forest, whereas several trees grow in the random forest. To categorize a new object based on characteristics, each tree provides a classification, which we refer to as the tree's 'vote' for that class. In the case of regression, the forest selects the classification with the most votes, and it takes the average of the outputs from several trees[14].

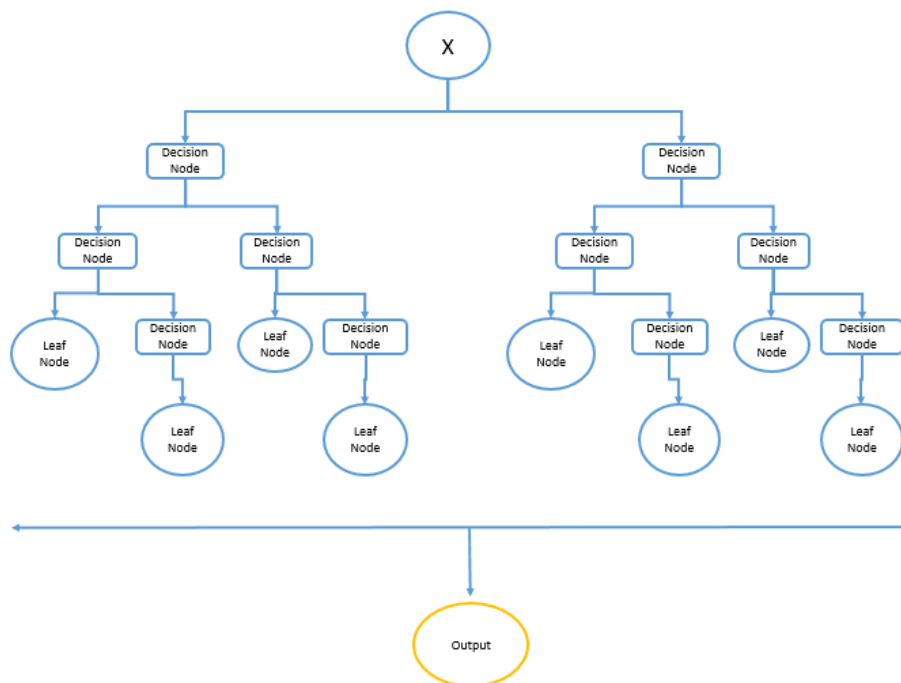


Figure 3.5: Random Forest algorithm.

### 3.5 Statistical Analysis

Statistical analysis is the use of quantitative data to investigate trends, patterns, and correlations. [15] Scientists, governments, corporations, and other organizations rely on it for research. After gathering data from our sample, we may use descriptive statistics to arrange and summarize the information. Then, we may utilize inferential statistics to formally test hypotheses and create population estimates. Finally, our findings may be interpreted and generalized.

#### 3.5.1 Age Analysis

We collected our data from 11 different ages of people. Figure x represents the age distribution. From our survey we can see that the people whose age is 16 less attended to survey. the percentage is about 5.7% among 1685 people. The people whose age is 18 are highest position to interest our research. And the percentage rate is about 15.9%. Another things is in our dataset we have seen that the age above 20 are more interested that the bellow 20. Figure 3.6 represents the all age analysis.

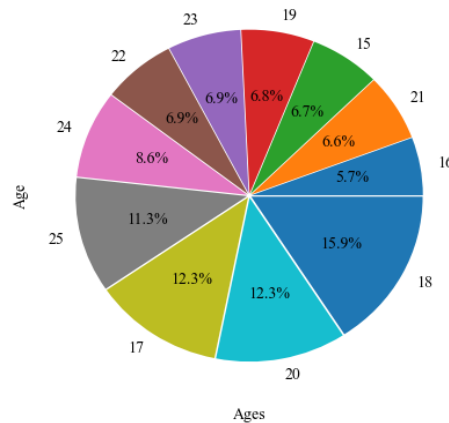


Figure. 3.6: All collected ages

#### 3.5.2 Affected age ratio

Figure 3.7 represents the affected age ratio of agoraphobia disease. From this graph we can see that

The highest affected age is 18. That means people at the age of 18 mostly suffer from this disease. and the second position is 17. The people whose age is 21 are less affected this disease. This incident gives an important information to us that is the ratio of bellow 20 is very less by the compare of above 20 but the affected rate is very high below 20 than above 20 years old.

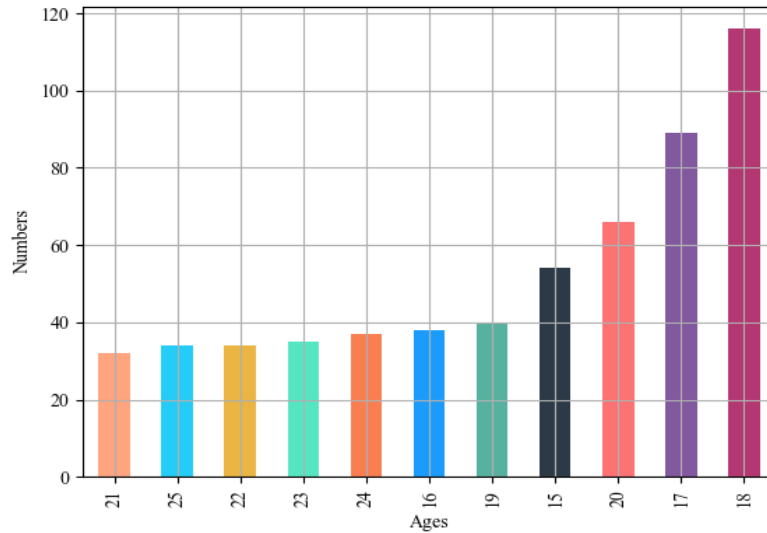


Fig. 3.7. Affected age ratio.

### 3.5.3 Male and Female Ratio

Figure 3.8 represents the male and female attendant ratio in our survey. the male rate is higher than female. Though we have seen from different article most of the time female are higher than male in term of agoraphobia disease.

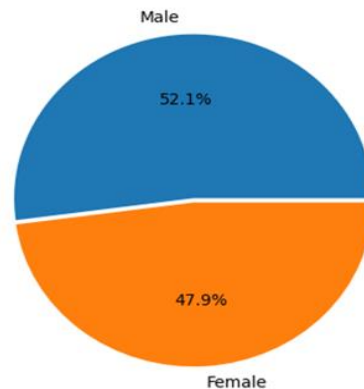


Fig. 3.8. Male and female ratio.



## CHAPTER 4

### Experimental Results and Discussion

#### 4.1 Experimental Setup

This chapter focuses on empirical evidence and descriptive research. When we examine it, we first consider what the findings are. The Implications section should be constructed in such a way that the findings are delivered without the need for awareness or examination. Suggestions are included under the section about research papers. This chapter will provide documentation of the findings and the test.

#### 4.2 Experimental Result and Analysis

We have used KNN(k Nearest Neighbor) Naive Bayes classifier, Support Vector Machine (SVM), Decision Tree, and Random Forest for prediction. On the basis of accuracy and accuracy score, we compared these algorithms. From the accuracy table, for 30% data usage rate, we got 98.02% by Random Forest and Decision Tree algorithm.

TABLE 4.1. ACCURACY TABLE

Data usage rate	Algorithms				
	<i>KNN</i>	<i>Naive Bayes</i>	<i>SVM</i>	<i>Decision Tree</i>	<i>Random Forest</i>
30%	95.85%	97.43%	97.83%	97.58%	98.02%
40%	97.48%	96.44%	97.03%	97.48%	97.48%
50%	96.56%	96.32%	97.03%	96.56%	97.51%
60%	96.74%	95.45%	96.05%	96.74%	96.71%
70%	97.20%	92.96%	92.20%	97.20%	96.54%

For training and testing data splitting we used a technique. That is we used 30 to 70% test data usages rate. That means when test size is 30 % then validation automatically 70%. Same way when test data is 70% then train data is 30%. We used this technique because we tried to see which percentage is better for our algorithms. And data quality, now the question is how can we check data quality? And the answer is if dataset contains more vulnerability then it never performs very well with less data training. Now the discussion is given below about our algorithm performance. Table 4.1.1 describes the accuracy score of our algorithm. yellow color background represes that the highest accuracy among 30-70% test data usage rate for each algorithm. Most of the algorithm perform very well by using 30% data usages rate except knn algorithm. Knn algorithm achieved 97.48% accuracy by using 40% test data. Other 4 algorithms like Naïve Bayes, Decision Tree, Support Vector Machine(SVM) achieved highest accuracy by using 30% test data. And their accuracy rate is 97.43%, 98.02%, 97.83%, 98.02% correspondingly. From this table we can see that random forest generates highest accuracy. So we decided to use this algorithm for prediction. But we have to test it more different way. The roc curve comparison is given bellow.

### **4.3 Roc Curve**

The ROC curve is a graphical representation of the connection between sensitivity and specificity that aids in the selection of the appropriate model by calculating the best diagnostic test threshold. [16] when we discuss about roc curve then we need to know two terms of machine learning one is sensitivity and specificity.

The division of real positive occasions that were anticipated as positive is known as affectability (or genuine positive). Review is another equivalent word for affectability. And The division of genuine negatives that were expected as negatives is known as specificity (or genuine negative). This implies that a portion of genuine negatives will be forecasted as positives, which may well be referred to as untrue positives. This rate is additionally known as the wrong positive rate.

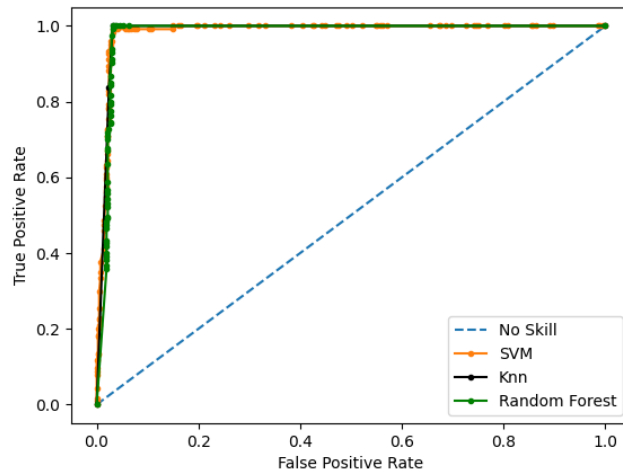


Fig. 4.1: Roc comparison between svm, knn, random forest

Figure 4.1 represents the roc curve of svm knn and random forest. Orange color represents svm. Black color represents knn and random forest represented by green color line graph.

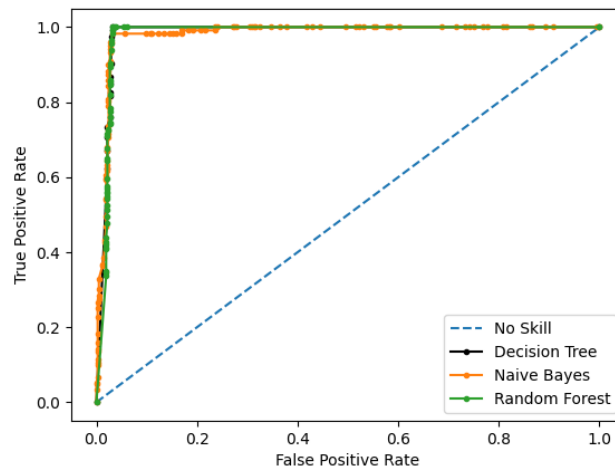


Fig. 4.2: Roc comparison between svm, knn, random forest

Figure 4.2 represents the roc curve of Decision Tree, Naïve Bayes, and random forest. Black color represents Decision Tree. Orange color represents Naïve bayes ad green color represents random forest algorithm.

Now what idea we get from roc curve? Answer is A higher X-axis value suggests more False positives than True negatives in a ROC curve. A higher Y-axis value implies a greater number of

True positives than False negatives, whereas a lower Y-axis value suggests a lower number of True positives.[17] The capacity to balance False positives and False negatives is thus a factor in determining the threshold. That means our system is better for true positive.

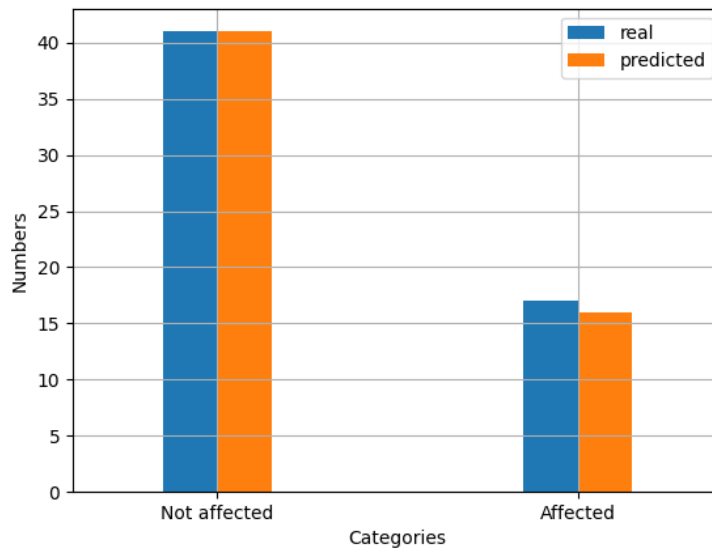


Fig. 4.3: Comparison between real and predicted class

Fig. 4.3 Illustrates, the real value is shown in orange color, while the predicted value is shown in blue color. In the overwhelming majority of cases, our expected value was similar to the actual value. For evaluation of our model we used 58 sample that is never seen by our model. We predicted 41 not affected and our system predicted 41 accurately and there is no error. And 17 for affected. In this case our system generates only one error.

From above experiments we can say that our system can generates very good results for real life data.

#### 4.4 Confusion Matrix

A confusion matrix is a table that shows how well a classification model performs on a set of test data for which the real values are known. The confusion matrix is straightforward in and of itself, but the associated language can be challenging [2]. So we used confusion matrix for test data for measuring the efficiency of our research.

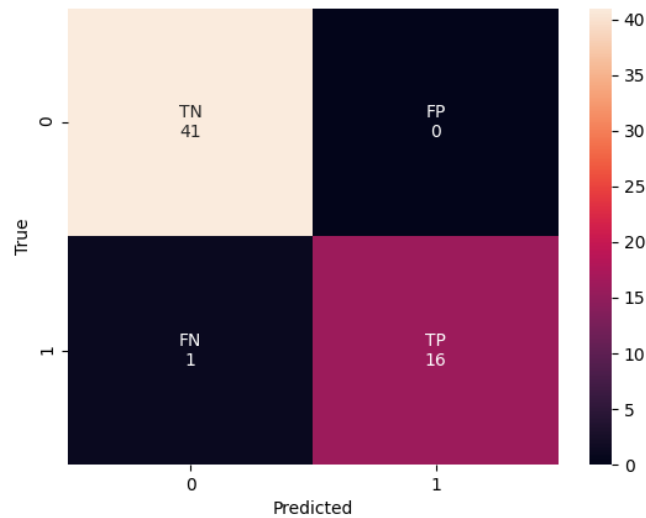


Fig. 4.4: Confusion Matrix.

$$\text{Accuracy} = \frac{41 + 16}{41 + 16 + 1 + 0} = 0.982 * 100$$

$$= 98.2\%$$

$$\text{Error} = 1 - 0.982 = 0.01 * 100 = 1\%$$

Recall rate for positive:

$$\frac{16}{16+1} = .94 * 100 = 94\%$$

$$\text{Recall rate for Negative: } \frac{17}{17+0} = 1 * 100 = 100\%$$

To detect overall outcomes, we employed the Confusion Matrix. The uncertainty matrix for the validation dataset is shown in Figure. In the evaluation procedure, we have a 99.2 percent accuracy rate. This also means that our model works with both visible and hidden data. The percentage of positive memory is 94%, whereas the rate of negative recall is 100%. It's a nice example for our agoraphobia negative detection prediction model.

## CHAPTER 5

### Conclusion and Future Work

#### 5.1 Introduction

The primary objective of our research is to detect agoraphobia disease by machine learning approach. In our case, we saw random forest algorithm perform very well with validation data as well as test data. We achieved this accuracy by maintaining all rules of data preprocessing and hyper parameter tuning.

#### 5.2 Conclusion

In this paper, five machine learning algorithms were used to detect Agoraphobia. We collect data from online survey. We use several classification techniques include the Decision Tree, Support Vector Machine (SVM), KNN, Random Forest (RF), and Naive Bayes, KNN. We trained 70% data and tested 30% data. After testing we get best accuracy by random forest with 98.02% accuracy. This was our test accuracy. For validation accuracy it performed very well. It generates only one error among for 58 unseen data. Our system is very good for agora phobia negative detection. We found this by confusion matrix approach.

#### 5.3 Implication for Future Study

1. we'd like to work with larger dataset, in the future,
2. In future, we want to use a real-time dataset. That mean dataset will be continuously updating and train by our model in online.
3. We will produce a web & android app.
4. We will also make a api, that can be integrated by anywhere like telemedicine website, health care application.
5. Our work is machine learning based, In future we will use deep learning algorithm to increase the efficiency of our research.

## REFERENCE

- [1] Hossain, Mohammad Didar, et al. "Mental disorders in Bangladesh: a systematic review." *BMC psychiatry* 14.1 (2014): 1-8.
- [2] B. Nithya and V. Ilango, "Predictive analytics in health care using machine learning tools and techniques," 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), 2017, pp. 492-499
- [3] Choudhury, Munmun De et al. "Predicting Depression via Social Media." ICWSM (2013).
- [4] Anu Priya, Shruti Garg, Neha Prerna Tigga, Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms, *Procedia Computer Science*, Volume 167, 2020, Pages 1258-1267
- [5] Arif M, Basri A, Melibari G, et al. (2020) Classification of Anxiety Disorders using Machine Learning Methods: A Literature Review. *Insights Biomed Res* 4(1):95-110.
- [6] Vijai, C; Wisetsri, Worakamol. Rise of Artificial Intelligence in Healthcare Startups in India Vol. 14, Iss. 1, (Mar 2021): 48-52.
- [7] M. Riajuliislam, K. Z. Rahim and A. Mahmud, "Prediction of Thyroid Disease(Hypothyroid) in Early Stage Using Feature Selection and Classification Techniques," 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), 2021, pp. 60-64
- [8] T. Nguyen, D. Phung, B. Dao, S. Venkatesh and M. Berk, "Affective and Content Analysis of Online Depression Communities," in *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 217-226, 1 July-Sept. 2014.
- [9] R. Razavi-Far, E. Hallaji, M. Saif and L. Rueda, "A Hybrid Scheme for Fault Diagnosis with Partially Labeled Sets of Observations," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 2017, pp. 61-67
- [10] K-Nearest Neighbor(KNN) Algorithm for Machine Learning, available at << <https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning> >> last accessed on 08-09-2021 at 8PM
- [11] K. Taunk, S. De, S. Verma and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 1255-1260
- [12] Sri Krishnan, 6 - Machine learning for biomedical signal analysis, Editor(s): Sri Krishnan, *Biomedical Signal Analysis for Connected Healthcare*, Academic Press, 2021, Pages 223-264
- [13] JOUR , Lv, Zhihan, Zhang, Zhifei, Zhao, Zijian, Yeom, Doo-Seoung, Decision Tree Algorithm-Based Model and Computer Simulation for Evaluating the Effectiveness of Physical Education in Universities , 2020, SN - 1076-2787

[14] Sarica A, Cerasa A and Quattrone A (2017) Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. *Front. Aging Neurosci.* 9:329.

[15] M. Mehedi Hasan, M. Omar Faruk, B. Biswas Biki, M. Riajuliislam, K. Alam and S. Farjana Shetu, "Prediction of Pneumonia Disease of Newborn Baby Based on Statistical Analysis of Maternal Condition Using Machine Learning Approach," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2021, pp. 919-924

[16] Sarang Narkhede, Understanding AUC - ROC Curve , available at << <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5> >> , last accessed on 05-07-2021 at 1 PM.

[17] Simple guide to confusion matrix terminology, available at << <https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/> >> March 25, 2014. last accessed on 06-07-2021 at 8AM.



## **APPENDIX**

The first problem we had when doing the analysis was establishing the analytical technique for our investigation. It wasn't standard job, and little had been done in this subject previously. As a result, we weren't able to get much help from any source. We also started gathering data by hand. After a lengthy time of hard labor, we might be able to achieve it.

# PLAGIARISM REPORT

Final Thesis_Report_V2			
ORIGINALITY REPORT			
29%	21%	14%	20%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS
PRIMARY SOURCES			
1	Submitted to Daffodil International University Student Paper	5%	
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%	
3	Submitted to Columbia High School Student Paper	2%	
4	Submitted to Coventry University Student Paper	1%	
5	Submitted to Athlone Institute of Technology Student Paper	1%	
6	Submitted to University of Essex Student Paper	1%	
7	scholars.direct Internet Source	1%	
8	"Proceedings of International Joint Conference on Advances in Computational Intelligence", Springer Science and Business Media LLC, 2021 Publication	1%	