Predicting the Level of Safety Feeling of Bangladeshi Internet users using Data Mining and Machine Learning

Md. Safiul Alam, Anirban Roy, Partha Protim Majumder, Sharun Akter Khushbu
Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh

***Abstract*—An amazing combination of cutting-edge data mining and machine learning methodologies to predict the level of safety feeling among Bangladeshi internet users, which is a significant departure in this subject. By leveraging cutting-edge algorithms and innovative data sources, this work provides previously unheard-of insights into how this demographic perceives online safety, shedding light on an essential yet underappreciated aspect of their digital lives. This exceptional study's original research increases the body of knowledge of online safety and sets the road for policy recommendations and intervention tactics that will enable Bangladesh to become a global leader in internet security.**

***Keywords*—Bangladesh; data analysis; data mining; important factors; machine learning; prediction; performance evaluation metrics; safety level**

I. Introduction

Every day, more people are using the internet than ever before, all over the world [1]. This rate is increasing in Bangladesh too [2]. Because recently Bangladesh has seen growth in internet usage [3]. At present, the Internet has become a massive part of people's daily lives in Bangladesh [4]. As a result, communication, business, education, banking, service, jobs, etc. are turning online day by day in Bangladesh [5]. In March 2021 Internet users in Bangladesh increased to 116 million whereas the population of Bangladesh at the time was 167 million which means 70% of the population had access to the internet [6]. People want to feel safe, secure, and devoid of any bullying, harassment, and illegal activity when using the internet [7]. According to a UNICEF survey, 32% of Bangladeshi children, aged 10 to 17, are familiar with and have encountered online abuse, harassment, and cyberbullying. 25% of them have access to the internet by the age of 11. Additionally, according to a Telenor Group and Grameenphone report, online bullying is a serious problem for 85% of Bangladeshi youngsters. According to the report, 18% of them experienced worse bullying as a result of the shocking COVID-19 epidemic [8]. So, it is very important to know people's safety feelings at the time of using the internet [9]. The advent of the digital era has created possibilities and challenges never before experienced, altering how people connect, communicate, and access information throughout the globe. As the internet continues to permeate every area of our daily lives, online safety and security have become a huge concern [10]. For that it is essential for the user to feel safe while using the internet [11]. There is a great depiction of an accurate prediction of an individual's safety level at the time of using the internet is indispensable with prior knowledge about the important factors, which have a great impact [12]. Moreover, it is necessary for private and public organizations, industries, banks, and IT companies to find out people's safety level at the time of using the Internet [13]. Because it will make their services more secure and effective [14]. In that case, safety level prediction will act as a guide to making an appropriate safety level which has been fulfilled in this research.

This work closes a huge knowledge gap that has mostly gone unfilled up to this point. Despite the abundance of research on online safety, there are surprisingly few that focus on the unique viewpoints and experiences of internet users in Bangladesh [15]. Because of its geographical emphasis and dedication to illuminating the intricacies of online safety in the context of Bangladesh, this study is a pioneering effort that stands out [16]. This work's significance extends beyond the sphere of academic research; it has a significant impact on Bangladesh's evolving digital ecosystem and tackles urgent problems that have not yet been fully investigated [17]. This groundbreaking study demonstrates its importance in a number of ways. In the age of digital transformation, where internet access is nearly widespread, it is imperative to provide Bangladeshi internet users with the knowledge and tools they need to properly navigate the online world [18]. By predicting people's safety feelings and fostering a sense of control and confidence in the face of online hazards, this study strengthens people's agency in safeguarding their digital experiences [19]. For Bangladeshi internet service providers, regulators, and legislators, the novel approach of this study offers a once-in-a-lifetime chance to tailor safety measures and actions [20]. Knowing the specific factors influencing safety feelings may help them develop more effective strategies and policies that match the local context, which will eventually lead to a safer online environment. The integration of data mining and machine learning in this study has increased the prevalence of data-driven decision-making in the area of internet safety. The efficacy of organizations and authorities may be improved by using the information gathered from this study to assist them in allocating resources and choosing initiatives based on actual evidence. The lack of study on the perspectives of Bangladeshi internet users on online safety highlights the novelty and significance of this endeavor [21]. This study investigates an understudied area, filling a large gap in the literature and setting the stage for future studies that have an emphasis on regional and local variability. As Bangladesh embraces digitization, developing a culture of cybersecurity becomes increasingly important [22]. This work has the potential to promote best practices among internet users, academic institutions, and businesses by igniting dialogues on online safety. Also, this study combines data mining and machine learning, fusing cutting-edge technology with real-world applications. The innovative approaches adopted might serve as a paradigm for future studies on the intersection of data science and cybersecurity in Bangladesh and elsewhere in the world [23]. Even though this study's findings are anchored in the context of Bangladesh, they may still be applicable to other developing nations that are going through rapid digitalization [24]. This research's importance transcends national boundaries since the technique and results developed here may be changed and applied in similar situations. There isn't a single, universal approach to online safety [25]. This unique piece of work is actually innovative since it allows for the customization of safety precautions. By anticipating Bangladeshi users' safety attitudes and allowing interventions and assistance to be personalized to individuals' particular concerns and experiences, online safety is made more pertinent and effective. It also reveals the attitudes and beliefs of Bangladeshi internet users, shedding light on a hitherto unresearched facet of online safety. This highlights the emotional and intangible aspects of cybersecurity that are occasionally overshadowed by technology solutions [26]. By detecting and evaluating these emotions, this approach improves our understanding of the human side of cyber security. The originality of this work opens the door for future research initiatives that focus on the feelings and experiences of internet users in a variety of contexts. It sets a precedent for appreciating the importance of the human element in cybersecurity and might ignite a larger conversation about the psychological aspects of online safety [27]. Along with being creative, it may help the Bangladeshi online community understand online safety by making it more pertinent, relatable, and human.

Additionally, this approach combines the strengths of machine learning and data mining. By exploiting the capabilities of this cutting-edge technology, the research proposes a creative way of predicting safety feelings that are tailored to the Bangladeshi environment. Combining these methods should result in conclusions that are more accurate, and practical, and represent a novel contribution to the field of internet safety.

Data mining and machine learning, a subfield of artificial intelligence (AI), employ statistical techniques to give computers the capacity to learn from data and improve their performance on certain tasks [28]. Data mining and machine learning are used to enable learning and inference across a heterogeneous mix of devices, including PCs, smartphones, IoT devices, and edge devices [29]. A data mining and machine learning probabilistic system is a complex tool that may be used to evaluate obtained data, provide predictions or judgments based on that data, and then present those findings to the user [30].

As Bangladesh continues its journey toward digital transformation, the findings of this study have the potential to inform governmental decisions, empower internet service providers to enhance user safety, and ultimately create a more secure online environment. By bridging the gap between data-driven insights and the specific problems faced by Bangladeshi internet users, this research provides a groundbreaking contribution to maintaining the online experiences of an expanding online community. Here, emphasis has been given to the analysis of some empirical factors of an individual's data to perform the safety level prediction.

In this research, safety level predictions have been done and several factors behind this have been analyzed. Moreover, extensive research and analysis have been conducted. Here, several data mining techniques have been applied for experimentation, and several performance evaluation metrics to evaluate this work. Twelve popular data mining classifiers, including Logistic Regression, MLP, KNN, Decision Tree, Naive Bayes, Search Vector Machine, Gradient Boosting, Linear Discriminant Analysis, Stochastic Gradient Descent, Ada Boosting, Bagging, and Random Forest, have been experimented with on a survey dataset. Several performance evaluation metrics have been calculated to determine the best classifier in the working context, and a result comparison is presented here. From the analysis of the obtained result, it is confirmed that the Decision Tree classifier achieves the best result in terms of metrics.

These are the order of this paper: Section II gives an exhaustive overview of relevant studies. The study methodology is presented in Section III along with a brief overview of the dataset, data analysis, implementation process, classifiers, the outcome of the experiment, and additional findings while the conclusion is given in Section IV. Finally, Section V provides future work.

II. Literature Review

The ultimate purpose of this research work is to the safety level of the user. After going through several articles, it is discovered that there has been no existing work like this done before. However, it has been unable to locate a compass in the large ocean of scholarly works that might direct us through the uncharted area of understanding how Bangladeshi internet users view their online safety [31]. This absence emphasizes the originality and importance of this research, which aims to address this important gap and improve not only the scholarly community but also the daily lives of countless Bangladeshi internet users [32]. The awareness that addressing the safety concerns of internet users goes beyond academic study and constitutes a necessary first step in establishing a more secure and safe online environment for everyone has steered this research down an innovative route [33]. To implement this unique model, some papers have been studied. All of them are described below as per the research paper's theme:

Syeda et al. [34] applied seven approaches of data mining i.e. KNN, Decision Tree, SVM, NN, Naive Bayes, Logistic Regression, and Random Forest to predict user satisfaction and dissatisfaction. They have taken different parameters which are produced with high accuracy. The accuracy for KNN, Decision Tree, SVM, NN, Naive Bayes, Logistic Regression and, Random Forest were 96%, 93.33%, 93.3%, 86%, 89.3% and, 96%. Though the highest accuracy is achieved by three algorithms, they have chosen Random Forest because it shows better precision, recall, and f1 score rather than others.

In order to identify phishing websites, Kaytan et al. [35] suggested a clever model based on extreme learning machines. Website forms differ from one another in terms of how they perform. Therefore, they must make use of special web page features to prevent phishing assaults. Additionally, they proposed a template based on computer training methods for identifying phishing web pages. The model has one output and 30 inputs. In this application, the 10-fold cross-validation test was run. The classification's total accuracy was 95.05 percent.

Salehin et al. [36] Karim advocated combining LSTM and artificial intelligence to produce a straightforward rainfall forecast model. The correctness of the deep learning approach is essential for this manner of application has been established. They used 6 parameters in their article. The accuracy was increased to 76% by looking at all the data integrating LSTM and artificial intelligence to produce a straightforward rainfall forecast model. They used 6 parameters in their article. The accuracy was increased to 76% by looking at all the data.

Salehin et al. [37] recommended utilizing RHMCD as a model to assist machine learning algorithms accomplishes the intended goal. Naive Bayes classifiers, logistic regression, and support vector machines are the algorithms that were tested. The sentiment analysis method was employed to gather information on mental health issues. The amount of depression was assessed using the decision tree method.

Salehin et al. [38] predicted the severity of depression caused by excessive cell phone use. Depression is detected using the Linear Regression technique and two machine learning algorithms, decision trees.

Technologies for agriculture have been created by Salehin et al. [39]. Various viral, fungal, and bacterial illnesses result in a significant loss of agricultural produce. In this research, they categorize crop situations based on various datasets by using the Scale Invariant Transform Feature (SIFT) technique. Finally, the solution was made available through live online portals and SMS services.

Talha et al. [40] draw attention to the significant drawback and its many root causes, including emotional instability, despair, stress, and loneliness. Physical, virtual, and medical reports were the three approaches that were used to collect the data. The detrimental impact of human behavior is demonstrated by the 71% optimistic theorem of Naive Bayes. For measurement purposes in search vector machine (SVM), negative and positive parameters are set. Last but not least, they compare the outcomes of our suggested specialization to those of the three fundamental points of reference.

Syeda et al. [41] used educational data mining to forecast the pupils' success. The entirety of the projection was based on the students' current location and general academic standing.

Yeasin et al. [42] suggested using the data mining approach to forecast students' careers. Only CS grads have had this task completed for them. They used a number of classifiers, and the accuracy varied according on each classifier. Just 506 data records were used in this study, and no distinct training or testing datasets were indicated.

Alonzo et al. [43] provided a thorough analysis of how different machine learning algorithms are used to predict and rate the quality of coconut sugar.

P'erez et al. [44] provided examples of the findings from a case study on educational data analytics that was focused on identifying undergraduate students majoring in systems engineering who had dropped out after six years of attendance. Their experimental findings demonstrated that straightforward algorithms may identify dropout predictors with sustained levels of accuracy. Here, the output of four algorithms—decision trees, logistic regression, naive bayes, and random forest—was examined to suggest the best course of action. The major findings are presented here to lower the dropout rate by identifying probable causes. In addition, they provided some evaluations of the data's quality to help the students refine their data collection techniques.

With the purpose of resolving the dropout prediction problem, Mi et al. [45] developed different temporal models. Specifically, based on substantial research conducted with a few massive open online courses (MOOCs) accessible through edX and Coursera. They claimed that one logical improvement to the model, which would include a max pooling layer before the output layer, would further their work. They anticipated that their model's expansion would increase its robustness.

Aksenova et al. [46] reported an enrollment prediction research that uses support vector machines and rule-based predictive models with the aim of predicting the overall enrollment headcount, which is made up of continuing, returning, and new (freshman and transfer) students. The core prediction findings are generated using a machine learning approach called SVM, which is then applied by a program called Cubist to create simple rule-based predictive models. Lastly, they provided some experimental findings about the forecasting of student enrolment.

III. Methodology

This section is parted into Data Description, Data Collection, Data Preprocessing, Data Analysis, Classifier Description, Implementation Procedure, Result and Discussion, and Evaluation. This section basically presents the approach taken to accomplish this work.

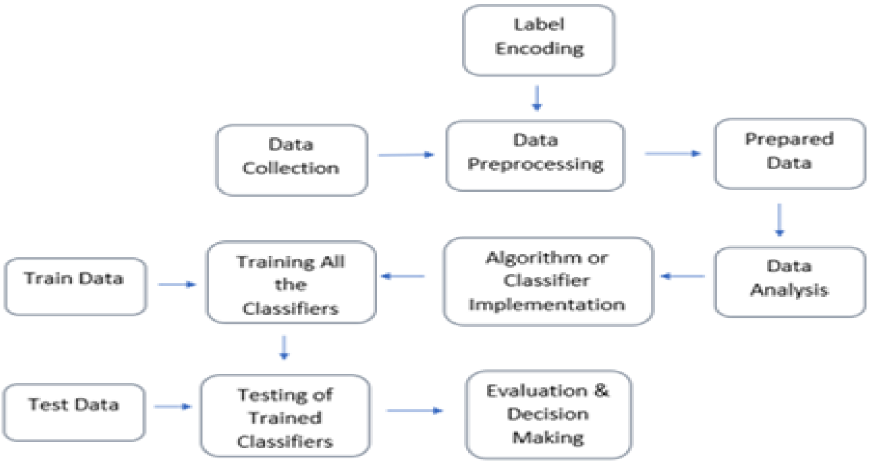
Label Encoding
Data Collection
Data Preprocessing
Prepared Data
Train Data
Training All the Classifiers
Algorithm or Classifier Implementation
Data Analysis
Test Data
Testing of Trained Classifiers
Evaluation & Decision Making

Fig. 1. Methodology diagram.

For this work, several steps have been performed, as presented in Fig. 1. A detailed description of all the subsections is presented below.

A. Data Description

Information that approximates and characterizes is referred to as qualitative data. It is possible to notice and document qualitative data [47]. In statistics, qualitative data is sometimes referred to as categorical data since it can be categorized based on the characteristics and traits of an object or phenomenon [48]. Any information that can be quantified and employed in statistical or mathematical calculations is referred to as quantitative data [49]. Making judgments in real life using mathematical inferences is aided by this type of data [50]. So, in this work, all the data are qualitative before preprocessing, and after preprocessing, they are converted to quantitative data for analysis and to build a machine learning model for prediction. A decision has been made after evaluation.

B. Data Collection

Data is survey-based data. The survey has been performed. Most of the data has been collected by physical survey and some of the data has been collected through an online survey. A total of 5,321 individual records are used here to accomplish this work. The survey mainly consists of 8 questions.

C. Data Preprocessing

After checking for null values, it has been found that there have been no missing values in the dataset as all the answers to the 8 questions have been obtained from the respondents and the information has been carefully compiled to make the dataset. Fig. 2 shows that there are no missing values. The data type information has been checked, and it has been observed that 6 columns have object-type values. The label encoding pre-processing technique has been used to convert these object-type values into numeric. Among the 8 questions, 7 questions

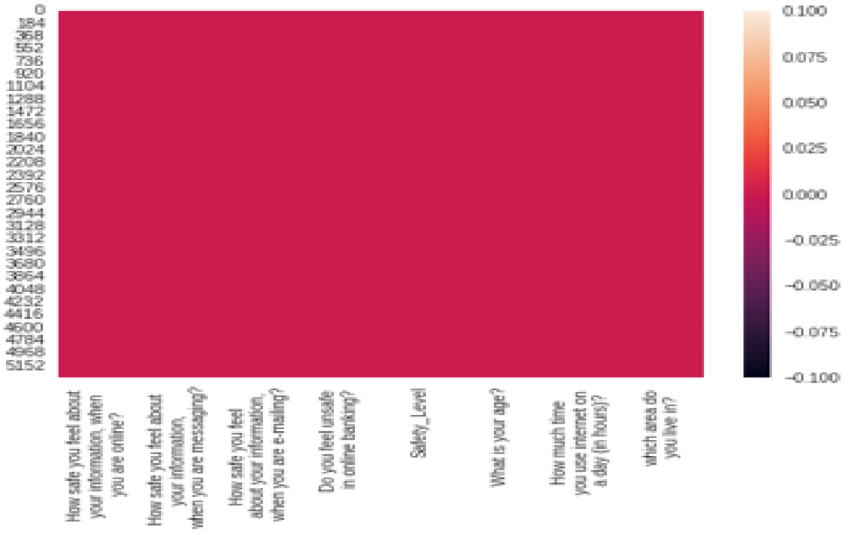

Fig. 2. Heatmap for checking null values.

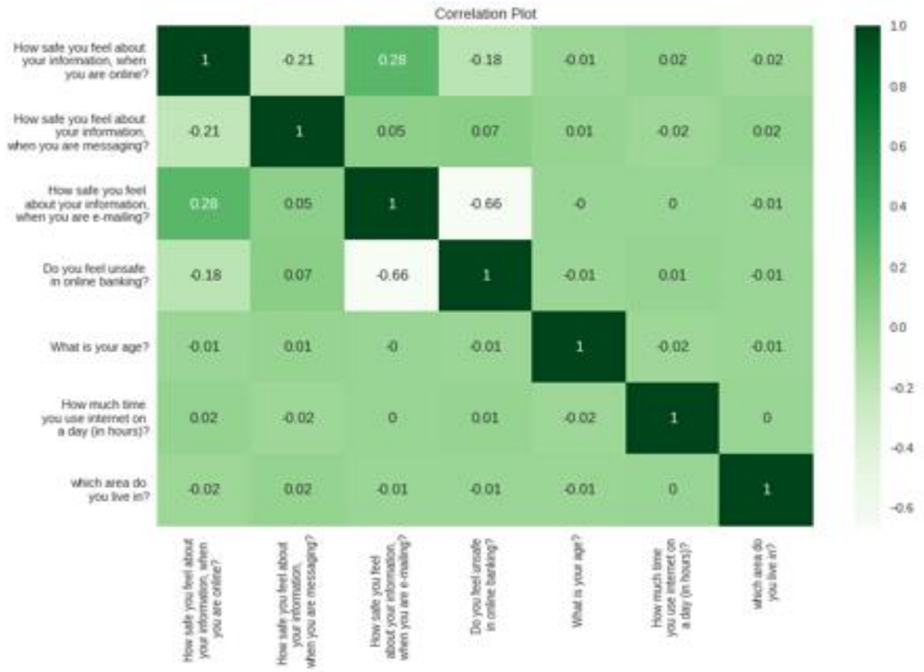

Fig. 3. Correlation matrix.

(How safe you feel about your information, when you are online? How safe you feel about your information, when you are messaging? How safe you feel about your information, when you are e-mailing? Do you feel unsafe in online banking? What is your age? How much time you use internet on a day (in hours)? which area do you live in?) This has been used as the independent variable and only one question (Safety_Level) has been used as the dependent variable.

To prevent overfitting, the dataset has been split into training and testing sets. The correlation of the independent variables in the training dataset has then been determined, as shown in Fig. 3. A total of 73% of the data has been used for the training of the classifier and 27% has been employed for testing purposes. To retrieve appropriate attributes, a threshold value of 0.78 has been set. Using this value, it has surprisingly been found that the 8 features that have been used as the independent variables do not need to be changed.

D. Data Analysis

Data analysis is the process of cleansing, converting, and modeling data to discover useful information for commercial decision-making. The goal of data analysis is to gather useful data and make decisions based on that analysis. When determining what is occurred most recently in real life or how something plays out when making a certain decision, a simple illustration is provided of how the data is interpreted.

In a survey of 5,321 respondents, it is discovered that 31.20% of individuals feel extremely safe about their information when they are online, 32.14% of people feel no safety at all, and 36.65% of people feel poor safety about it. These results are depicted in Fig. 4.

Fig. 5 shows the results of a survey of 5,321 people, which is revealed that 35.38% of respondents feel only moderately safe about their information when messaging, 32.33% of respondents feel no safety at all, and 32.29% of respondents feel extremely safe about their information when messaging.

Fig. 6 illustrates the results of a survey of 5,321 respondents, which is surprisingly revealed that 48.74% of them feel only somewhat secure sending information through e-mail, 26.82% feel no safety at all, and 24.44% feel extremely safe sending information via e-mail.

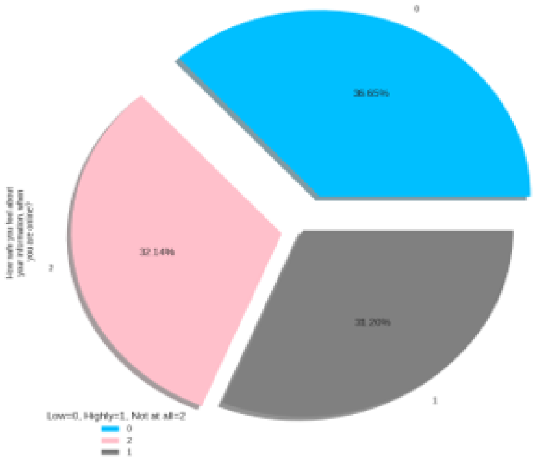

Fig. 4. People’s safety feeling about their information while using the internet.

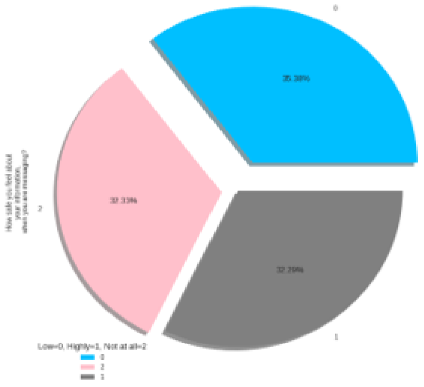

Fig. 5. Safeness feeling while messaging.

The results depicted in Fig. 7 demonstrate that 57.50% of the 5,321 respondents feel unsafe while using internet banking, while the remaining 42.50% feel secure.

The bar in Fig. 8 shows the number of observations for each of the five potential category value combinations. It can be observed that individuals who feel less secure about their information while online are given a lower Safety Level rating than those who feel more secure and those who feel absolutely no security at all. Additionally, it is found that individuals who do not feel secure about their information when online seldom perceive their Safety Level to be high, while those who feel extremely secure about their personal data while online have rated their Safety Level as higher than low.

The bar in Fig. 9 displays the number of observations for each of the five potential category value combinations. It can be observed from the figure that individuals who feel the least safe when texting are assessed to have a lower Safety Level than those who feel the safest and those who feel the least safe while messaging. Additionally, it is surprising that people seldom perceive their Safety Level to be as high when they are texting using the internet when they do not feel safe and feel uncomfortable about their information when they are texting. However, those who feel extremely secure about the privacy of their information when communicating have rated their Safety Level as the highest.

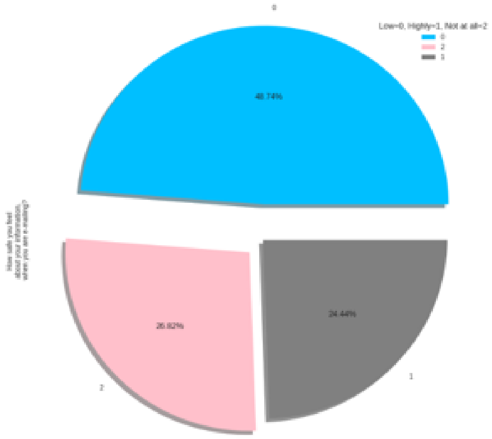

Fig. 6. Sense of safety while e-mailing.

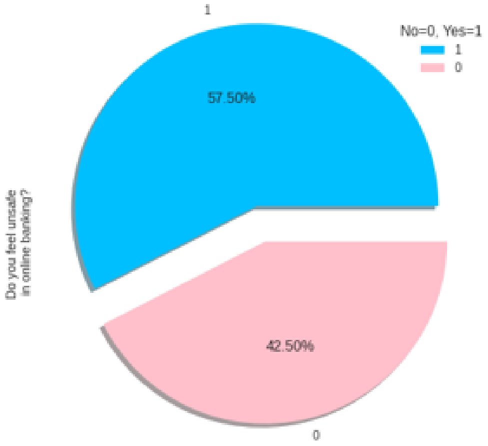

Fig. 7. People's thinking of online banking.

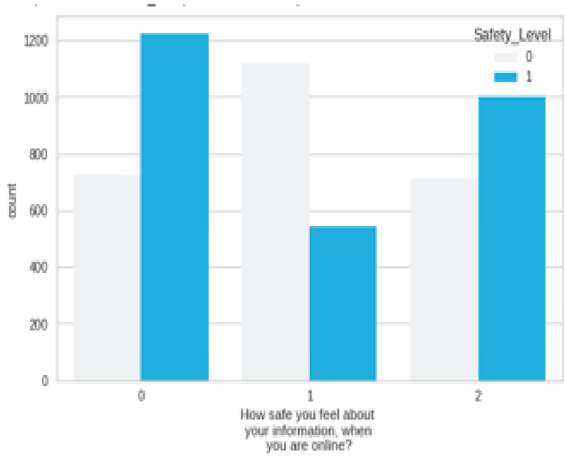

Fig. 8. Impact of First Attribute on Target Variable.

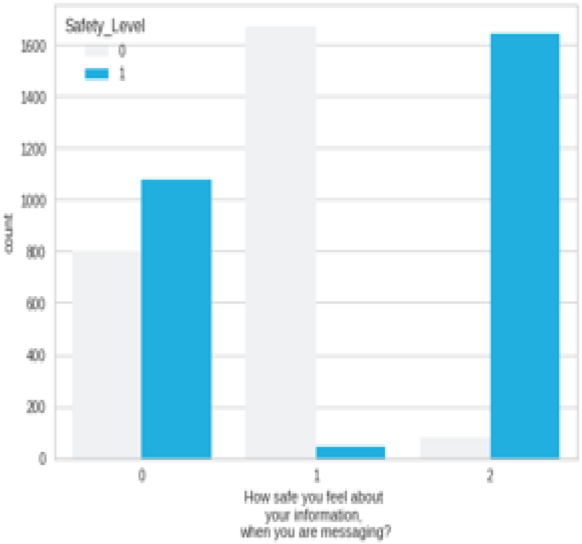

Fig. 9. Second Attribute's Effect on the Target Variable.

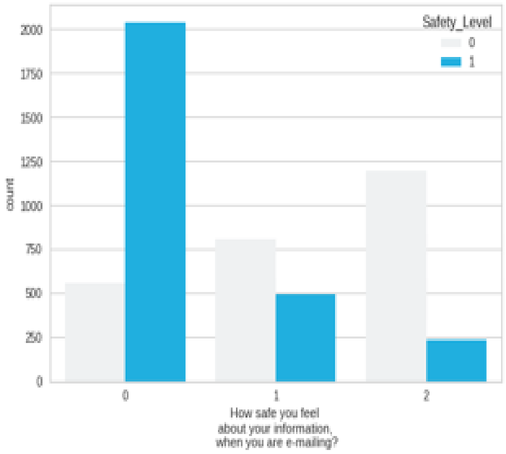

Fig. 10. Influence of the third attribute on the target variable.

Fig. 10 displays the counts of observations for each of the five potential category value combinations. It can be seen that individuals who feel less safe about their information when emailing have a lower Safety_Level rating than those who feel more secure and those who feel no security at all. Moreover, people who feel less safe about their information when emailing rarely perceive their Safety_Level to be high. In contrast, it has found that people who feel unsafe about their information when emailing consider their Safety_Level to be the highest.

The bar chart in Fig. 11 shows the number of observations for each of the four potential category value combinations. It can be observed that individuals who feel unsafe while conducting online banking rated their Safety Level lower than those who feel secure. Interestingly, individuals who feel unsafe when using online banking rarely rated their Safety Level as the highest. However, those who feel secure when using internet banking have rated their Safety Level as the highest.

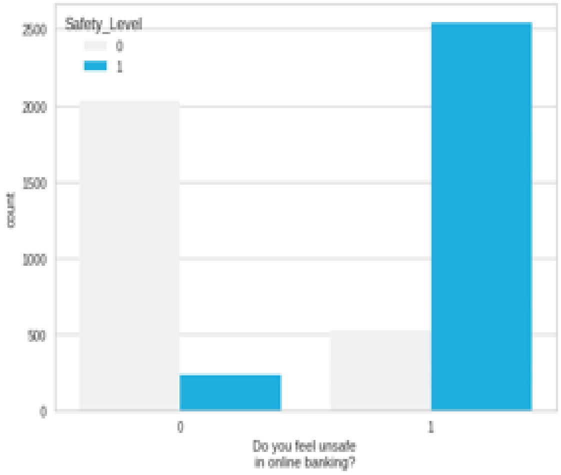

Fig. 11. Significance of the fourth attribute on the target variable.

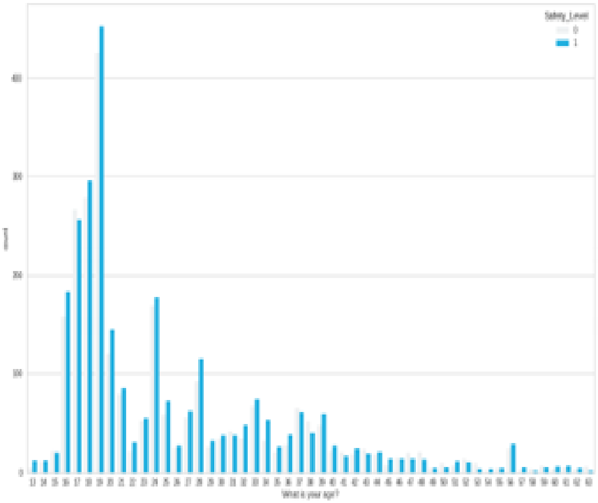

Fig. 12. Role of the sixth attribute on the target variable.

Fig. 12 demonstrates that individuals in the following age groups have believed their safety level to be low: 13 to 14, 16, 18 to 30, 32 to 36, 39 to 40, 42, 44 to 45, 51, 54, 56 to 57, and 60 to 62. On the other hand, those between the ages of 15, 17, 31, 37 to 38, 41 to 43, 46 to 48, 50 to 52, 53 to 58, and 63 thought their safety level is high. Interestingly, respondents between the ages of 49 and 56 are perceived their safety level to be both high and low.

Fig. 13 shows that the people who spend 3 to 4 hours, 7 to 13 hours, and 16 hours a day using the internet have considered their Safety_ Level as low. On the other hand, people who spend 2 hours, 5 to 6 hours, and 14 hours a day using the internet have considered their Safety_ Level as high.

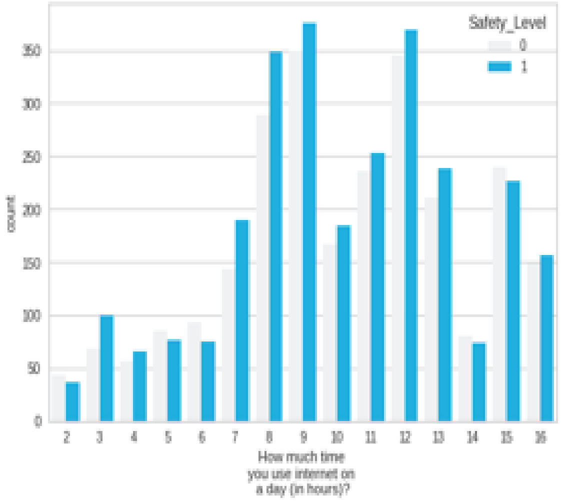

Fig. 13. Importance of the fifth attribute on the target variable.

Among 5,321 respondents, it is found that 51.99% of people consider their Safety_Level as low while they are using the internet and 48.01% of people consider their Safety_Level as high, as shown in Fig. 14.

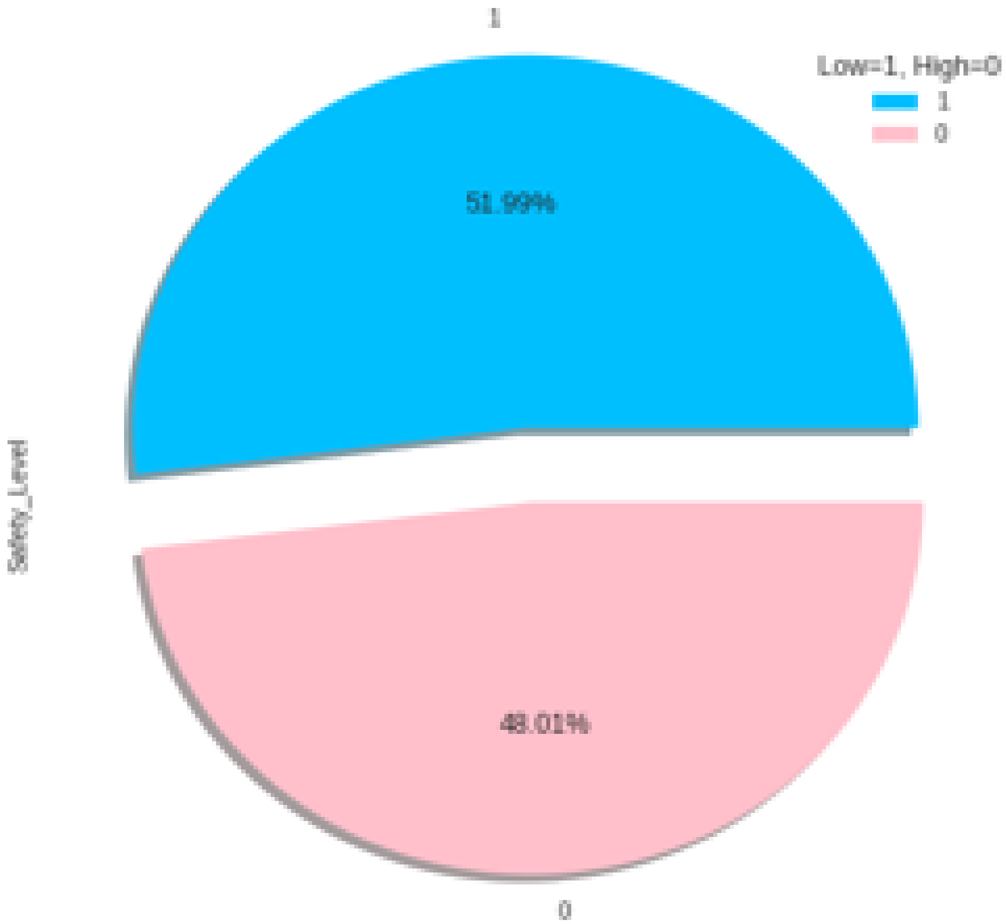

Fig. 14. People’s consideration of safety level while using the internet.

E. Classifier Description

A classifier in machine learning is a tool for forecasting the target characteristic from feature data points. Twelve classifiers have been used to analyze the dataset, and the following theory is pertinent.

In this work, Naive Bayes classifier has been used. This classifier employs Bayesian classification methodologies. It applies Bayes' theorem to class prediction and determines the class-conditional probability by taking into account the fact that the attributes are conditionally independent given the class label. This classifier can handle binary and multiclass classification issues since predictors all make independent assumptions. The work primarily focuses on a binary classification problem.

A Multilayer Perceptron (MLP) consists of an input, an output, and a number of hidden layers (one or more). Single-layer perceptrons can only learn linear functions, but multi-layer perceptrons can learn nonlinear functions. The MLP learning procedure is known as the Backpropagation Algorithm. Once the input layer receives the signal, the output layer anticipates making a decision based on the input [51]. The hidden layers serve as the computational engine for estimating continuous functions [52]. The output of one layer in an MLP serves as the input for the layer that follows.

The optimal division for each node is selected using local knowledge via a greedy method known as a Decision Tree. One conclusion is that a better tree may be produced by altering the divisional components [53]. It is well known that trees are incredibly flexible and exhibit little distortion in their interactions.

The decision tree classifier was designed particularly for the ensemble method known as Random Forest. The main function of the random forest classifier is to integrate the predictions of several trees (decision trees), where each decision tree is constructed from the output of a different dataset of random vectors. Problems with grouping are generally resolved with it. Using data samples, Random Forest algorithms build decision trees, predict those trees, and then let users vote on the best result. The group technique is superior to a single tree since it supports the outcome and reduces over-adjustment.

A statistical approach for analyzing a data set containing one or more independent variables that affect the outcome is logistic regression. To assess the result, a dichotomous variable is employed in this (only two possible outcomes). The goal of this classifier is to choose the model that best depicts, using the logistic function as support, the connection between the outcome variable and the predictor factors.

To address two-group classification issues, supervised machine learning models called support vector machines (SVM) use classification methods. Once provided with a set of labeled training data for each category, an SVM model can classify incoming text. They perform better with fewer samples and are more effective, which are their two main advantages (in the thousands). The method works well for text classification problems since it is customary to only have access to datasets with a small number of tags on each sample.

The k-nearest neighbors method, often known as KNN, is a supervised learning classifier that makes predictions or classifications about how a single data point will be grouped. It is frequently used as a categorization strategy since it is predicated on the notion that similar points could be found adjacent to one another. The k parameter of the k-NN algorithm determines how many neighbors will be looked at in order to categorize a certain query point. If k=1, for example, the case will be put in the identical class as it's only nearest neighbor.

Bagging, often referred to as bootstrap aggression, is an effective collective tactic. A technique for combining the results of different machine-learning algorithms to create predictions that are more accurate is called an ensemble approach. A broad method known as bootstrap aggregation may be used to minimize variation in algorithms with a lot of it. As with hybrid approaches like classification and regression, bagging has a large variance. A high-variance machine learning system, like decision trees, is exposed to the Bootstrap technique during the bagging process.

A quick and effective method for training linear classifiers and regressors under convex loss functions is stochastic gradient descent (SGD). SGD has been present in the machine learning field for a while, but in the context of large-scale learning, it has just lately attracted a lot of interest. Because the update to the coefficients is done for each training instance rather than at the end of examples, it has been successfully used for large-scale datasets. The Stochastic Gradient Descent (SGD) classifier essentially implements a straightforward SGD learning method that supports multiple classification loss functions and penalties.

In 1996, Freund and Schapire proposed AdaBoost. By transforming a number of weak learners into strong learners, these methods increase prediction ability. It creates a classifier by combining a number of subpar classifiers. Each iteration involves training the data and setting the classifier weights.

The combination of gradient descent and boost is known as Gradient Boosting. Each new model in gradient boosting employs the gradient descent method to reduce the loss function from its forerunner. This process is repeated until the target variable's estimation becomes even better. In contrast to previous ensemble approaches, gradient boosting builds a succession of trees, each one attempting to fix the flaws of the one before it.

For supervised classification issues, a dimensionality reduction method called Linear Discriminant Analysis is frequently employed. It is used to represent group distinctions, i.e. to distinguish between two or more classes [54]. In a lower dimension space, it is used to project the characteristics from a higher-dimension space. In order to save money and dimensions, this can be used to project characteristics from higher dimensional space into lower dimensional space.

F. *Implementation Procedure*

The aims of this work are to perform the safety level prediction and to analyze the important factors behind choosing a particular safety level for an individual. Many significant parameters are considered here to ensure an effective prediction.

The work primarily focuses on a binary classification problem. A questionnaire form containing 8 questions was created and data was collected from different professions of people and many random people through this questionnaire. Preprocessing techniques were used to feed this data into the classifier. To label the answer to the particular question, numbers (e.g. 0, 1) were used. The dataset had a variable/attribute named "Safety_Level" with two possible outcomes High (0) and Low (1). After preprocessing, the prepared data was partitioned into the training and testing set. 73% of the data from the total dataset was used for training purposes and the rest 27% of the whole dataset was used for testing purposes. The classifiers were trained with the training data and then used to predict the Safety_Level using both the test data and train data. Metrics were calculated for the performance evaluation and the best classifier was determined based on the confusion matrix generated by the classifier.

G. *Result and Discussion*

In this section, the experimental result and the discussion of the obtained result of the study are presented. The result of the confusion matrix for the test data of twelve classifiers is tabulated in Table I. Since it is a two-class problem, so the classifiers generate a 2*2 matrix.

At the time of implementation, 1,437 respondent instances are put into the testing set where the actual safety level of 667 students is high or positive. On the other hand, the actual safety level of 760 respondents is low or negative. After implementation, it has been found that a confusion matrix for each classifier which is stated in Table I. The experimental result of the confusion matrix in detail for the most competent classifier and the worst classifier has been found from Table I. From Table I, it has been found that the decision tree classifier is correctly able to predict that 607 respondents will be considered their safety level as high among 667 respondents. So, the rest of the 70 respondents among the 667 respondents are incorrectly classified that they will not be considered their safety level as high. On the other hand, this classifier is correctly able to predict that 729 respondents will be considered their safety level as low among 760 respondents. So, the rest of the 31 respondents among the 760 respondents are incorrectly classified that they do not be considered their safety level as low. From Table I, it has been found that the decision tree classifier is correctly able to predict that 530 respondents will be considered their safety level as high among 667 respondents. So, the rest of the 147 respondents among the 667 respondents are incorrectly classified that they will not be considered their safety level as high. On the other hand, this classifier is correctly able to predict that 578 respondents will be considered their safety level as low among 760 respondents.

So, the rest of the 182 respondents among the 760 respondents are incorrectly classified that they do not be considered their safety level as low. From Table I it has been found that the MLP algorithm has the highest specificity which is 0.98. On the other hand, the KNN algorithm has the lowest specificity which is 0.76. Specificity means the true negative rate. In this work, the specificity of a classifier refers to how well a classifier identifies respondents who will be considered their safety level as low. Decision Tree has 0.96 specificity means that it can identify 96% of respondents consider their safety level as low. From the value of the confusion matrix, a classification report, macro average, and weighted average of test data for each of the classifiers has been computed which are presented in Table II and Table III. From Table II it has found that the precision of the MLP classifier for the High class is 0.97 and of the Bagging classifier for the Low class is 0.93 which are the highest, the recall of the MLP classifier for the Low class is 0.98, and of the Bagging classifier for High class is 0.92 which are the highest, and the f1-score of the Decision tree classifier for High and Low class is 0.92, 0.94 which are the highest. From Table III, it has surprisingly found that the Decision Tree classifier has the highest precision, recall, and f1-score. On the other hand, the KNN classifier has the lowest precision, recall, and f1-score.

TABLE I. CONFUSION MATRIX AND SPECIFICITY RESULT OF THE TWELVE WORKING CLASSIFIER

Classifier Name	True Positive	False Negative	False Positive	True Negative	Specificity
Decision Tree	607	70	31	729	0.96
Random Forest	606	71	54	706	0.93
Naive Bayes	562	115	84	676	0.89
Logistic Regression	590	87	89	671	0.88
KNN	530	147	182	578	0.76
SVM	594	83	68	692	0.91
Gradient Boosting	602	75	40	720	0.95
Stochastic Gradient Descent	574	103	73	687	0.90
Linear Discriminant Analysis	587	90	89	671	0.88
MLP	568	109	17	743	0.98
Ada Boost	601	76	43	717	0.94
Bagging	628	49	92	668	0.88

TABLE II. CLASSIFICATION REPORT OF ALL THE TWELVE CLASSIFIERS

Classifier Name	Class Name	Precision	Recall	F1-Score	Support
Decision Tree	Low	0.91	0.96	0.94	760
	High	0.95	0.90	0.92	677
Random Forest	Low	0.91	0.93	0.92	760
	High	0.92	0.90	0.91	677
Naive Bayes	Low	0.86	0.89	0.87	760
	High	0.87	0.83	0.85	677
Logistic Regression	Low	0.89	0.88	0.88	760
	High	0.87	0.87	0.87	677
KNN	Low	0.80	0.76	0.78	760
	High	0.74	0.78	0.76	677
SVM	Low	0.89	0.91	0.90	760
	High	0.90	0.88	0.89	677
Gradient Boosting	Low	0.91	0.95	0.93	760
	High	0.94	0.90	0.91	677
Stochastic Gradient Descent	Low	0.87	0.90	0.89	760
	High	0.89	0.85	0.87	677
Linear Discriminant Analysis	Low	0.89	0.88	0.88	760
	High	0.87	0.87	0.87	677
MLP	Low	0.87	0.98	0.92	760
	High	0.97	0.84	0.90	677
Ada Boost	Low	0.90	0.94	0.92	760
	High	0.93	0.89	0.91	677
Bagging	Low	0.93	0.88	0.91	760
	High	0.87	0.92	0.90	677

TABLE III. MACRO AVERAGE AND WEIGHTED AVERAGE OF ALL THE TWELVE CLASSIFIERS

Classifier Name	**Macro Average**				**Weighted Average**			
	Precision	***Recall***	***F1- Score***	***Support***	***Precision***	***Recall***	***F1- Score***	***Support***
Decision Tree	0.93	0.93	0.93	1437	0.93	0.93	0.93	1437
Random Forest	0.91	0.91	0.91	1437	0.91	0.91	0.91	1437
Naive Bayes	0.86	0.86	0.86	1437	0.86	0.86	0.86	1437
Logistic Regression	0.77	0.77	0.77	1437	0.77	0.77	0.77	1437
KNN	0.92	0.92	0.92	1437	0.92	0.92	0.92	1437
SVM	0.88	0.88	0.88	1437	0.88	0.88	0.88	1437
Gradient Boosting	0.88	0.88	0.88	1437	0.88	0.88	0.88	1437
Stochastic Gradient Descent	0.92	0.91	0.91	1437	0.92	0.91	0.91	1437
Linear Discriminant Analysis	0.88	0.88	0.88	1437	0.88	0.88	0.88	1437
MLP	0.92	0.91	0.91	1437	0.92	0.91	0.91	1437
Ada Boost	0.92	0.92	0.92	1437	0.92	0.92	0.92	1437
Bagging	0.90	0.90	0.90	1437	0.90	0.90	0.90	1437

TABLE IV. AUROC SCORE OF TWELVE CLASSIFIERS

Classifier Name	**AUROC Score**
Decision Tree	0.983
Random Forest	0.914
Naive Bayes	0.913
Logistic Regression	0.950
KNN	0.846
SVM	0.949
Gradient Boosting	0.977
Stochastic Gradient Descent	0.887
Linear Discriminant Analysis	0.942
MLP	0.981
Ada Boost	0.962
Bagging	0.940

TABLE V. USED PARAMETERS AND ACCURACY OF TWELVE CLASSIFIERS

Classifier Name	**Parameter Detail**	**Accuracy (For Test Data)**	**Accuracy (For Train Data)**
Decision Tree	max_depth=6	0.93	0.93
Random Forest	n_estimators=1	0.91	0.96
Naive Bayes	alpha=1.0, fit_prior=True	0.86	0.86
Logistic Regression	random_state=1	0.88	0.88
KNN	n_neighbors=3	0.77	0.89
SVM	probability=True, kernel='linear'	0.89	0.89
Gradient Boosting	n_estimators=88, learning_rate=1.0,max_depth=1, random_state=0	0.92	0.93
Stochastic Gradient Descent	loss="modified_huber"	0.88	0.88
Linear Discriminant Analysis	n_components=1	0.88	0.88
MLP	random_state=1, max_iter=300	0.91	0.88
Ada Boost	n_estimators=105	0.92	0.92
Bagging	n_estimators=2, random_state=0	0.90	0.96

Table IV shows us the AUROC score for each classifier. AUC means the area under the curve which helps to understand the performance of the model [55]. From Table IV it has been found that the Decision Tree classifier has the highest AUROC score which is 0.983. On the other hand, KNN has the lowest AUROC score which is 0.846.

Table V represents the accuracy of all algorithms for both training data and testing data. Also, Table V illustrates the parameters and the different things that are used in this work to implement the algorithms selected. These parameters have been taken for better accuracy. After analyzing Table V in other words after comparing the accuracy of test data and train data for each classifier it can ensure that there is no overfitting and underfitting situation in this model [56]. The highest accuracy for test data is 0.93 which is achieved by Decision Tree. On the other hand, the lowest accuracy for test data is 0.77 which is achieved by KNN.

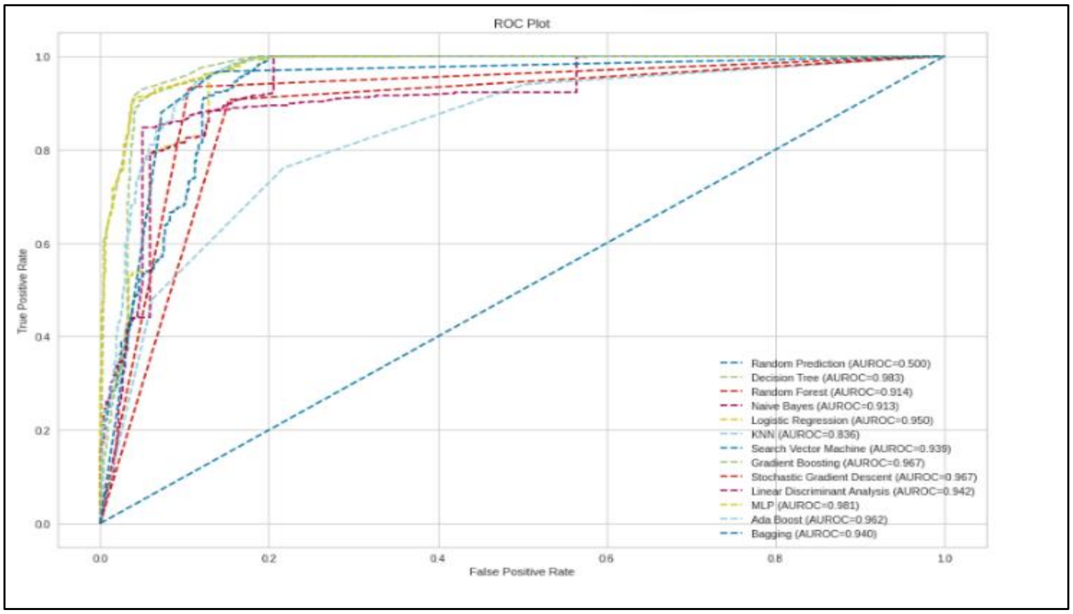

Fig. 15. ROC graph of all the twelve classifiers.

Fig. 15 shows the ROC. ROC means receiver operating characteristic which has been helped to evaluate the performance of diagnostic tests [57]. The blue line actually cuts diagonally across the rectangle here across a call which is actually a random classification that is made not based on any classifier so it simply splits the data into two so it is based on chance [58]. Also, in the blue line, the recall and specificity are equal. Fig. 15 has been made from Table IV where it has been seen that the Decision Tree classifier gives the highest performance than others. It has also been found that the KNN classifier gives the lowest performance than any other classifier.

Table VI provides a list of each algorithm's name, the mean accuracy, and the standard deviation accuracy. From the above table, it has amazingly found that four algorithms that have the highest mean accuracy for train data which are Decision Tree, Gradient Boosting, MLP, and Ada Boost. These four algorithms' mean accuracy is 0.92. On the other hand, the KNN classifier has the lowest mean accuracy for train data which is 0.77. From the above table, it has also been found that the Stochastic Gradient Descent is the highest standard deviation accuracy for train data which is 0.08. Also, it has surprisingly found that Decision Tree, Gradient Boosting, MLP, and Ada Boost have the lowest standard deviation accuracy for train data which is 0.01.

Fig. 16 shows the comparison of different algorithms which have been used to build the model. From these results, it is suggested that Decision Tree, Gradient Boosting, MLP, and Ada Boost are perhaps worthy of further study on this problem.

TABLE VI. MEAN ACCURACY AND STANDARD DEVIATION OF TWELVE ALGORITHMS

Algorithm Name	Mean Accuracy (For train data)	Standard Deviation Accuracy (For Train Data)
Decision Tree	0.92	0.01
Random Forest	0.89	0.02
Naive Bayes	0.85	0.02
Logistic Regression	0.87	0.02
KNN	0.77	0.03
SVM	0.86	0.02
Gradient Boosting	0.92	0.01
Stochastic Gradient Descent	0.81	0.08
Linear Discriminant Analysis	0.87	0.02
MLP	0.92	0.01
Ada Boost	0.92	0.01
Bagging	0.90	0.02

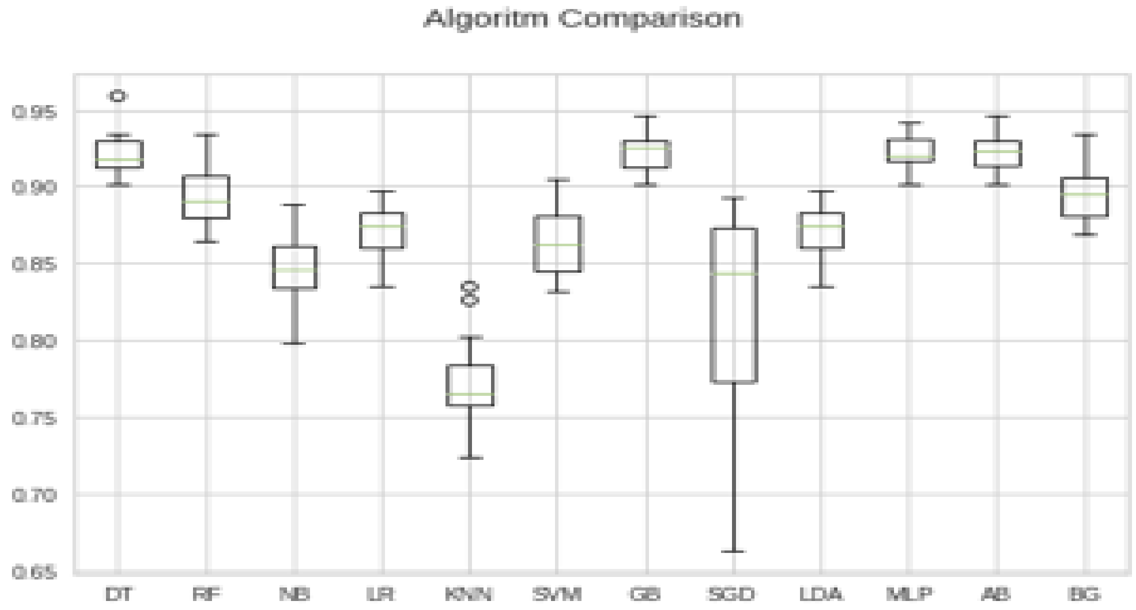

Fig. 16. Comparing all the twelve classifiers by using boxplot.

H. Evaluation

Comparison of training accuracy and testing accuracy is very important to understand the overfitting situation and underfitting situation in a machine learning model [59]. However, most of the previous research works had not shown the comparison of the test accuracy and train accuracy of their model which has been the main reason for being unable to verify their model's performance properly. This problem has been solved in this amazing piece of work and has been shown in Section III (G). The Decision Tree algorithm achieved the highest accuracy of 0.93. Also, based on the results analyzed in Section III (G), this algorithm was chosen as the final algorithm.

IV. CONCLUSION

The major goals of this work are to anticipate a person's level of online safety feeling and to identify the deciding elements that affect that person's decision to select a specific level of internet safety feeling. It is concluded from the analysis of the collected data that 48.01% of individuals feel extremely safe while using the internet, compared to 51.99% who don't, which raises serious concerns for the future growth of the nation. A variety of data mining approaches are used. A total of 73% and 27% of the data are used to train and test the classifier, respectively, in order to complete this task. A number of performance assessment measures are examined to gauge how well the functional classifier performed. The decision tree classifier surpasses conventional data mining algorithms.

V. FUTURE WORK

It is speculated that Decision Tree, Gradient Boosting, MLP, and Ada Boost are probably worthy of additional investigation on this subject based on Fig. 16 in section III (G).

ACKNOWLEDGMENT

Gratitude is expressed to each and every responder who assisted with data gathering by sharing information. Also, many thanks to the reviewers for their insightful and helpful reviews.

REFERENCES

[1] Hine, C. (2015). Ethnography for the Internet: Embedded, Embodied and Everyday (1st ed.). Routledge. https://doi.org/10.4324/9781003085348.

[2] Al Mamun, M. A., and Mark D. Griffiths. "The association between Facebook addiction and depression: A pilot survey study among Bangladeshi students." *Psychiatry research* 271 (2019): 628-633. doi: 10.1016/j.psychres.2018.12.039.

[3] Abdul Aziz (2020) Digital inclusion challenges in Bangladesh: the case of the National ICT Policy, Contemporary South Asia, 28:3, 304-319, doi: 10.1080/09584935.2020.1793912.

[4] Shammi, M., Bodrud-Doza, M., Islam, A.R.M.T. *et al.* Strategic assessment of COVID-19 pandemic in Bangladesh: comparative lockdown scenario analysis, public perception, and management for sustainability. *Environ Dev Sustain* **23**, 6148–6191 (2021). https://doi.org/10.1007/s10668-020-00867-y.

[5] Hoque, Md Rakibul. "The impact of the ICT4D project on sustainable rural development using a capability approach: Evidence from Bangladesh." *Technology in Society* 61 (2020): 101254. https://doi.org/10.1016/j.techsoc.2020.101254.

[6] "Internet in Bangladesh", Available online: https://en.wikipedia.org/wiki/Internet_in_Bangladesh [Last Accessed 30 January 2023].

[7] Hasler, Laura, Ian Ruthven, and Steven Buchanan. "Using internet groups in situations of information poverty: Topics and information needs." *Journal of the Association for Information Science and Technology* 65.1 (2014): 25-36. https://doi.org/10.1002/asi.22962.

[8] "Safe internet and digital security in Bangladesh", Available online: https://www.observerbd.com/news.php?id=373165 [Last Accessed 30 January 2023].

[9] Lavis, Anna, and Rachel Winter. "# Online harms or benefits? An ethnographic analysis of the positives and negatives of peer-support

around self-harm on social media." *Journal of child psychology and psychiatry* 61.8 (2020): 842-854. https://doi.org/10.1111/jcpp.13245.

[10] Djenna, A.; Harous, S.; Saidouni, D.E. Internet of Things Meet Internet of Threats: New Concern Cyber Security Issues of Critical Cyber Infrastructure. *Appl. Sci.* **2021**, *11*, 4580. https://doi.org/10.3390/app11104580.

[11] R. Roman, P. Najera and J. Lopez, "Securing the Internet of Things," in *Computer*, vol. 44, no. 9, pp. 51-58, Sept. 2011, doi: 10.1109/MC.2011.291.

[12] Haight, Michael, Anabel Quan-Haase, and Bradley A. Corbett. "Revisiting the digital divide in Canada: The impact of demographic factors on access to the internet, level of online activity, and social networking site usage." *Current Research on Information Technologies and Society*. Routledge, 2016. 113-129. doi: 10.4324/9781315751474-9.

[13] Tawalbeh, L.; Muheidat, F.; Tawalbeh, M.; Quwaider, M. IoT Privacy and Security: Challenges and Solutions. Appl. Sci. 2020, 10, 4102. https://doi.org/10.3390/app10124102.

[14] Masud, M., Gaba, G.S., Choudhary, K. *et al.* A robust and lightweight secure access scheme for cloud based E-healthcare services. *Peer-to-Peer Netw. Appl.* **14**, 3043–3057 (2021). https://doi.org/10.1007/s12083-021-01162-x.

[15] Kshetri, Nir. "Diffusion and effects of cyber-crime in developing economies." *Third World Quarterly* 31.7 (2010): 1057-1079. https://doi.org/10.1080/01436597.2010.518752.

[16] Uddin, N. (2023). Methodological Issues in Social Research: Experience from the Twenty-First Century. In: Uddin, N., Paul, A. (eds) The Palgrave Handbook of Social Fieldwork. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-031-13615-3_1.

[17] Mannan, Sushmita, Dewan Mohammad Enamul Haque, and Netai Chandra Dey Sarker. "A study on national DRR policy in alignment with the SFDRR: Identifying the scopes of improvement for Bangladesh." *Progress in disaster science* 12 (2021): 100206. https://doi.org/10.1016/j.pdisas.2021.100206.

[18] Mathrani, Anuradha, Tarushikha Sarvesh, and Rahila Umer. "Digital divide framework: online learning in developing countries during the COVID-19 lockdown." *Globalisation, Societies and Education* 20.5 (2022): 625-640. https://doi.org/10.1080/14767724.2021.1981253.

[19] Berson, Ilene R. "Grooming cybervictims: The psychosocial effects of online exploitation for youth." *Journal of School Violence* 2.1 (2003): 5-18. https://doi.org/10.1300/J202v02n01_02.

[20] Howard, H., Knoppers, B., Cornel, M. *et al.* Whole-genome sequencing in newborn screening? A statement on the continued importance of targeted approaches in newborn screening programmes. *Eur J Hum Genet* **23**, 1593–1600 (2015). https://doi.org/10.1038/ejhg.2014.289.

[21] F. M. Awaysheh, M. N. Aladwan, M. Alazab, S. Alawadi, J. C. Cabaleiro and T. F. Pena, "Security by Design for Big Data Frameworks Over Cloud Computing," in *IEEE Transactions on Engineering Management*, vol. 69, no. 6, pp. 3676-3693, Dec. 2022, doi: 10.1109/TEM.2020.3045661.

[22] Tao, Hai, et al. "Economic perspective analysis of protecting big data security and privacy." *Future Generation Computer Systems* 98 (2019): 660-671. https://doi.org/10.1016/j.future.2019.03.042.

[23] Sarker, I.H., Kayes, A.S.M., Badsha, S. *et al.* Cybersecurity data science: an overview from machine learning perspective. *J Big Data* **7**, 41 (2020). https://doi.org/10.1186/s40537-020-00318-5.

[24] Alam, Md Jahangir, Rakibul Hassan, and Keiichi Ogawa. "Digitalization of higher education to achieve sustainability: Investigating students' attitudes toward digitalization in Bangladesh." *International Journal of Educational Research Open* 5 (2023): 100273. https://doi.org/10.1016/j.ijedro.2023.100273.

[25] Shillair, Ruth, et al. "Online safety begins with you and me: Convincing Internet users to protect themselves." *Computers in Human Behavior* 48 (2015): 199-207. https://doi.org/10.1016/j.chb.2015.01.046.

[26] Slupska, J. War, Health and Ecosystem: Generative Metaphors in Cybersecurity Governance. *Philos. Technol.* **34**, 463–482 (2021). https://doi.org/10.1007/s13347-020-00397-5.

[27] Slupska, J. War, Health and Ecosystem: Generative Metaphors in Cybersecurity Governance. *Philos. Technol.* **34**, 463–482 (2021). https://doi.org/10.1007/s13347-020-00397-5.

[28] P. Ongsulee, V. Chotchaung, E. Bamrungsi and T. Rodcheewit, "Big Data, Predictive Analytics and Machine Learning," *2018 16th International Conference on ICT and Knowledge Engineering (ICT&KE)*, Bangkok, Thailand, 2018, pp. 1-6, doi: 10.1109/ICTKE.2018.8612393.

[29] Lavallin, Abigail, and Joni A. Downs. "Machine learning in geography–Past, present, and future." *Geography Compass* 15.5 (2021): e12563. https://doi.org/10.1111/gec3.12563.

[30] Bose, Indranil, and Radha K. Mahapatra. "Business data mining—a machine learning perspective." *Information & management* 39.3 (2001): 211-225. https://doi.org/10.1016/S0378-7206(01)00091-X.

[31] Burns, S., Roberts, L. Applying the Theory of Planned Behaviour to predicting online safety behaviour. *Crime Prev Community Saf* **15**, 48–64 (2013). https://doi.org/10.1057/cpcs.2012.13.

[32] CheshmehSohrabi, M., Mashhadi, A. Using Data Mining, Text Mining, and Bibliometric Techniques to the Research Trends and Gaps in the Field of Language and Linguistics. *J Psycholinguist Res* **52**, 607–630 (2023). https://doi.org/10.1007/s10936-022-09911-6.

[33] Von Schomberg, Rene. "A vision of responsible research and innovation." *Responsible innovation: Managing the responsible emergence of science and innovation in society* (2013): 51-74. https://doi.org/10.1002/9781118551424.ch3.

[34] Syeda Farjana Shetu , Israt Jahan , Mohammad Monirul Islam , Refath Ara Hossain , Nazmun Nessa Moon and Fernaz Narin Nur. Predicting Satisfaction of Online Banking System in Bangladesh by Machine Learning. 2021 International Conference on Artificial Intelligence and Computer Science Technology (ICAICST). Publisher: IEEE, DOI: 10.1109/ICAICST53116.2021.9497796.

[35] Kaytan, M , Hanbay, D . (2017). Effective Classification of Phishing Web Pages Based on New Rules by Using Extreme Learning Machines. Computer Science, 2 (1) , 15-36 . Retrieved from https://dergipark.org.tr/en/pub/bbd/issue/30846/333818.

[36] Salehin, I., Talha, I. M., Hasan, M. M., Dip, S. T., Saifuzzaman, M., & Moon, N. N. (2020, December). An Artificial Intelligence Based Rainfall Prediction Using LSTM and Neural Network. In 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE) (pp. 5-8). IEEE.

[37] Salehin, I., Dip, S. T., Talha, I. M., Rayhan, I., Nammi, K. F. "Impact on Human Mental Behavior after Pass through a Long Time Home Quarantine Using Machine Learning", International Journal of Education and Management Engineering (IJEME), Vol.11, No.1, pp. 41-50, 2021. DOI: 10.5815/ijeme.2021.01.05.

[38] Salehin, I., Talha, I. M., Moon, N. N., Saifuzzaman, M., Nur, F. N. & Akter, M. "Predicting the Depression Level of Excessive Use of Mobile Phone: Decision Tree and Linear Regression Algorithm" on 2nd International Conference on Sustainable Engineering and Creative Computing (ICSECC-2020), 16 - 17 December 2020, President University, Indonesia. Indexing: IEEE Xplore, EI-Compendex, SCOPUS.

[39] Salehin, I., Talha, I. M., Saifuzzaman, M., Moon, N. N., & Nur, F. N. (2020, October). An Advanced Method of Treating Agricultural Crops Using Image Processing Algorithms and Image Data Processing Systems. In 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA) (pp. 720-724). IEEE.

[40] Talha, I. M., Salehin, I., Debnath, S. C., Saifuzzaman, M., Moon, M. N. N., & Nur, F. N. (2020, July). Human Behaviour Impact to Use of Smartphones with the Python Implementation Using Naive Bayesian. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

[41] Shetu, S. F., Saifuzzaman, M., Sultana, S., Yousuf, R., & Moon, N. N. (2020). Students performance prediction through education data mining depending on overall academic status and environment. In 3rd International Conference on Innovative Computing and Communication (ICICC-2020).

[42] M.Y. Arafath, M. Saifuzzaman, S. Ahmed, and S.A. Hossain, "Predicting career using data mining," in Proceedings of the International Conference on Computing, Power and Communication Technologies (GUCON), pp. 889-894, IEEE, 2018.

[43] L. M. B. Alonzo, F. B. Chioson, H. S. Co, N. T. Bugtai and R. G. Baldovino, "A Machine Learning Approach for Coconut Sugar Quality Assessment and Prediction," 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), Baguio City, Philippines, 2018, pp. 1-4, doi: 10.1109/HNICEM.2018.8666315.

[44] B. Pérez, C. Castellanos, and D. Correal, "Predicting student drop-out rates using data mining techniques: A case study," In IEEE Colombian Conference on Applications in Computational Intelligence, pp. 111- 125, Springer, 2018.

[45] F. Mi and D. Yeung, "Temporal models for predicting student dropout in massive open online courses," In 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 256-263, IEEE, 2015.

[46] S. S. Aksenova, D. Zhang, and M. Lu, "Enrollment prediction through data mining," In 2006 IEEE International Conference on Information Reuse & Integration, pp. 510-515, IEEE, 2006.

[47] Kuckartz, Udo, and Stefan Rädiker. *Analyzing qualitative data with MAXQDA*. Cham: Springer International Publishing, 2019.doi: 10.1007/978-3-030-15671-8.

[48] Mertler, Craig A., Rachel A. Vannatta, and Kristina N. LaVenia. *Advanced and multivariate statistical methods: Practical application and interpretation*. Routledge, 2021. https://doi.org/10.4324/9781003047223.

[49] “Qualitative & Quantitative Data”, Available online: https://www.questionpro.com/blog/qualitative-data/ [Last Accessed 9 February 2023].

[50] Rubin, Donald B. "Causal inference using potential outcomes: Design, modeling, decisions." *Journal of the American Statistical Association* 100.469 (2005): 322-331. https://doi.org/10.1198/016214504000001880.

[51] D. Yan *et al*., "Improving Brain Dysfunction Prediction by GAN: A Functional-Connectivity Generator Approach," *2021 IEEE International Conference on Big Data (Big Data)*, Orlando, FL, USA, 2021, pp. 1514-1522, doi: 10.1109/BigData52589.2021.9671402.

[52] M. T. Sami, D. Yan, H. Huang, X. Liang, G. Guo and Z. Jiang, "Drone-Based Tower Survey by Multi-Task Learning," *2021 IEEE International Conference on Big Data (Big Data)*, Orlando, FL, USA, 2021, pp. 6011-6013, doi: 10.1109/BigData52589.2021.9672078.

[53] J. Khalil, D. Yan, G. Guo, M. T. Sami, J. B. Roy and V. P. Sisiopiku, "Traffic Study of Shared Micromobility Services by Transportation Simulation," *2021 IEEE International Conference on Big Data (Big Data)*, Orlando, FL, USA, 2021, pp. 3691-3699, doi: 10.1109/BigData52589.2021.9671455.

[54] Ahad, Md Taimur, et al. "Comparison of CNN-based deep learning architectures for rice diseases classification." *Artificial Intelligence in Agriculture* 9 (2023): 22-35. doi: 10.1016/j.aiia.2023.07.001.

[55] Bowers, Alex J., and Xiaoliang Zhou. "Receiver operating characteristic (ROC) area under the curve (AUC): A diagnostic measure for evaluating the accuracy of predictors of education outcomes." *Journal of Education for Students Placed at Risk (JESPAR)* 24.1 (2019): 20-46. https://doi.org/10.1080/10824669.2018.1523734.

[56] H. Zhang, L. Zhang and Y. Jiang, "Overfitting and Underfitting Analysis for Deep Learning Based End-to-end Communication Systems," *2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*, Xi'an, China, 2019, pp. 1-6, doi: 10.1109/WCSP.2019.8927876.

[57] Lobo, Jorge M., Alberto Jiménez‐Valverde, and Raimundo Real. "AUC: a misleading measure of the performance of predictive distribution models." *Global ecology and Biogeography* 17.2 (2008): 145-151. https://doi.org/10.1111/j.1466-8238.2007.00358.x.

[58] Kumar, R., Indrayan, A. Receiver operating characteristic (ROC) curve for medical researchers. *Indian Pediatr* **48**, 277–287 (2011). https://doi.org/10.1007/s13312-011-0055-4.

[59] Huang, Wenjiang, Pedro Martin, and Houlong L. Zhuang. "Machine-learning phase prediction of high-entropy alloys." *Acta Materialia* 169 (2019): 225-236. https://doi.org/10.1016/j.actamat.2019.03.012.