Spam Email Classification using different Machine Learning Algorithms

Sadman Sadik	**180104110**
Shaiuf Sadique	**180104111**
Fabliha Nahid	**180104116**

Project Report

Course ID: CSE 4214

Course Name: Pattern Recognition Lab

Semester: Spring 2021

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

September 2022

Spam Email Classification using different Machine Learning Algorithms

Submitted by

Sadman Sadik	**180104110**
Shaiuf Sadique	**180104111**
Fabliha Nahid	**180104116**

Submitted To

Faisal Muhammad Shah

Sajib Kumar Saha Joy,

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

September 2022

ABSTRACT

Messages are distributed from one computer to another through electronic means via network. Nearly 4.26 billion email users are there worldwide. With the rapid increase of using email there occurs also an issue which is Spam Email. Spam email is also known as junk email. So it becomes a necessary part to detect the email which one is spam and which one is ham. We used Email Spam Collection Dataset from kaggle for our project. The dataset contains the data of spam and ham message data. We have applied some machine learning algorithm such as Multinomial Naive Bayes, Support Vector Machine, K Nearest Neighbor, Random Forest and Decision Tree for our project. Random Forest performs best according to the result in our project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In today's world, email is one of the important sources of communication. This communication may vary according to personal, business, corporate to government. In the era of information technology, information sharing has become very easy and fast. There are also many platforms available for users to share information across the world. Among all information sharing mediums, email is the simplest method of information sharing worldwide. But emails are also vulnerable to different kinds of attacks. No one wants to receive emails which are not related to their interest as it wastes receivers' time as well as resources. So, it requires a massive demand for securing of the email system. Spam emails may carry viruses also. Attackers may send spam emails that contain attachments with the multiple-file extension, packed URLs etc. that indicates the user to spiteful and spamming websites and end up with some sort of data.

Chapter 2

Literature Reviews

We have reviewed several papers for this project.

2.1 Improved email spam detection model based on support vector machines

In this paper, Corpus Benchmark Spam Dataset was used which is very popular and widely used. It consists of 57 features and 1 target attribute. The authors of this paper, proposed a model based on support vector machines (SVM) and tried to achieve better accuracy of detecting spam emails while focusing on approximately using exhaustive parameter search techniques to ensure better spam detection accuracy. The accuracy of the proposed model in this paper was 94.06% which was the highest accuracy on this dataset overtaking the NSA-PSO approach which had an accuracy of 91.22%.[1]

2.2 Email Spam Detection Using Machine Learning Algorithms

The authors in this paper implemented different machine learning algorithms like Naïve Bayes, Support Vector Machine, Decision Tree etc to filter the emails. Email datasets from numerous websites like Sklearn, Kaggle is used here. They used a spam email data set from Kaggle to train their model and the other email dataset is used the result. The best two algorithms for their model was Naïve Bayes and Decision Tree and their accuracy was 98% and 95% respectively.[2]

2.3 Comparative Analysis of Classification Algorithms for Email Spam Detection

In this work, the performance of classification algorithms that are used for grouping emails as spam or ham including Bayesian Logistic Regression, RBF Network, Hidden Naïve Bayes, Voted Perceptron is evaluated. A combination of Particle Swarm Optimization and Artificial Neural Network was used for feature selection and Support Vector Machine was classify the spam emails. The Spambase dataset was used for their work and Rotation Forest emerged as the best classifier with 94.2% accuracy. [3]

2.4 Performance Evaluation of Machine Learning Algorithms for Email Spam Detection

In this research, the authors worked on the UCI Machine Learning Repository Spambase Data Set and presented a performance evaluation on different machine learning algorithms. Random Forest algorithm outperformed the other classification algorithms with a accuracy of 99.93%. Though KNN produced the same result but it was more time consuming compared to Random Forest algorithm.[4]

Chapter 3

Data Collection & Processing

In our project, Email Spam Collection dataset [5] is collected from the Kaggle. It contains set of SMS messages in English of 5,574 messages, tagged according to ham or spam. A training sample of email and labels, where label '1' denotes the email is spam and label '0' denotes the email is ham. Preprocessing helps us to transform data so that we can get a better machine learning model to build by providing higher accuracy. The preprocessing performs various functions such as duplicate data, stop word remove, token etc. In the dataset, 4516 samples are classified as ham email, and 653 are classified as spam email.

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
0	ham	Go until jurong point, crazy.. Available only ...	NaN	NaN	NaN
1	ham	Ok lar... Joking wif u oni...	NaN	NaN	NaN
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	NaN	NaN	NaN
3	ham	U dun say so early hor... U c already then say...	NaN	NaN	NaN
4	ham	Nah I don't think he goes to usf, he lives aro...	NaN	NaN	NaN
5	spam	FreeMsg Hey there darling it's been 3 week's n...	NaN	NaN	NaN
6	ham	Even my brother is not like to speak with me. ...	NaN	NaN	NaN
7	ham	As per your request 'Melle Melle (Oru Minnamin...	NaN	NaN	NaN
8	spam	WINNER!! As a valued network customer you have...	NaN	NaN	NaN
9	spam	Had your mobile 11 months or more? U R entitle...	NaN	NaN	NaN

Figure 3.1: : Dataset before pre-processing.

3.1 Dataset Preprocessing

3.1.1 Tokenization

It is mainly indicates the removal of URLs, break a string into a list of tokens. It is used to count the occurrence of words in email. In this project email is splits into smaller pieces or tokens from a longer string.

3.1.2 Stemming

The words are replaced by their root words. For example: send, sent and sending. Although these words are different by tense but their root word is send. In this project it is applied for getting to root word to preprocess the data.

3.1.3 Stop words removal

It is the most commonly used to preprocess the data. It removes high occurrence of data which rarely have any influence on the overall email. Basically articles and pronouns are generally classified as stop words.

3.1.4 Remove Pattern

This removes the urls, numbers, special characters and punctuation etc. from email. Through this entire process it cleans the unnecessary strings which impact less value in email.

	result	text	characters	words	sentences	transformed_text
0	0	Go until jurong point, crazy.. Available only ...	111	24	2	go jurong point crazi avail bugi n great world...
1	0	Ok lar... Joking wif u oni...	29	8	2	ok lar joke wif u oni
2	1	Free entry in 2 a wkly comp to win FA Cup fina...	155	37	2	free entri 2 wkli comp win fa cup final tkt 21...
3	0	U dun say so early hor... U c already then say...	49	13	1	u dun say earli hor u c alreadi say
4	0	Nah I don't think he goes to usf, he lives aro...	61	15	1	nah think goe usf live around though

Figure 3.2: : Dataset after pre-processing.

3.2 Spam words in dataset

We have extracted some spam words such as text, mobile, call etc. from the dataset. We have also plot the frequently used spam words in the dataset.

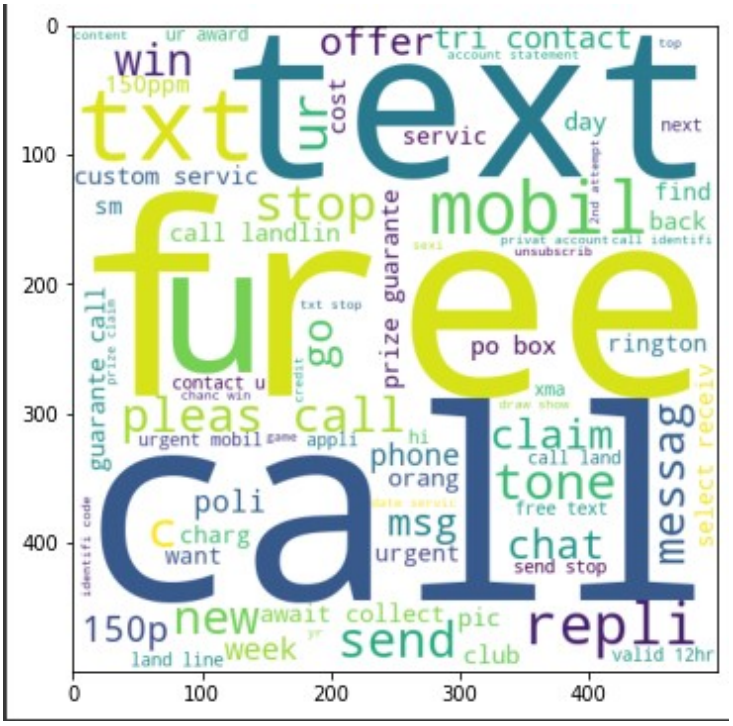

Figure 3.3: : Spam words in dataset.

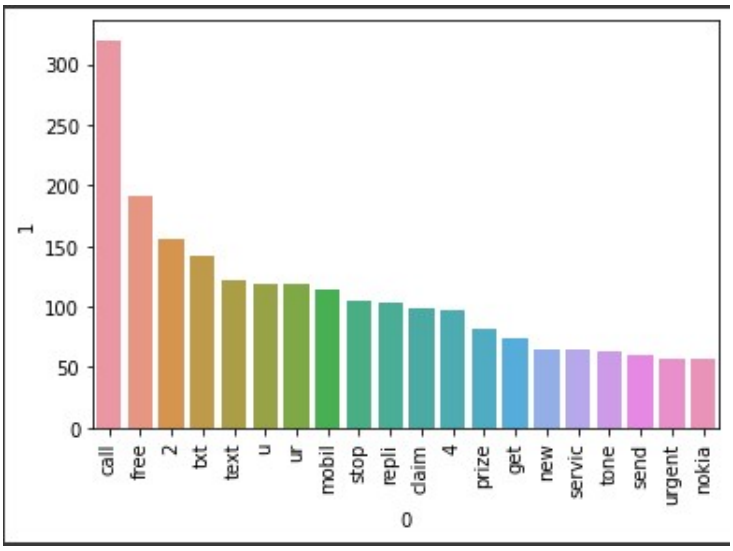

Figure 3.4: : Spam word frequency.

3.3 Ham words in dataset

We have extracted some nonspam words such as u, go, com etc. from the dataset. We have also plot the frequently used spam words in the dataset.

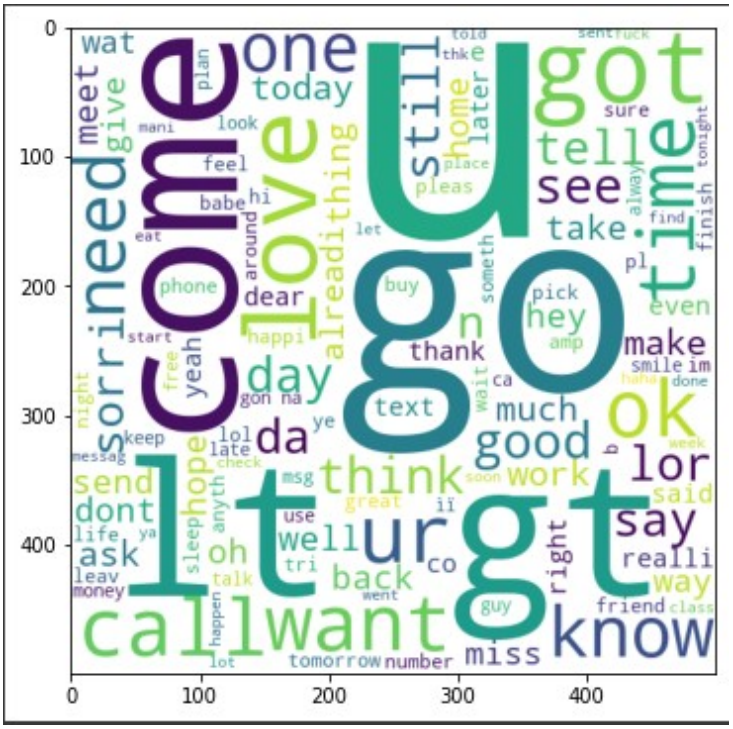

Figure 3.5: : Ham words in dataset.

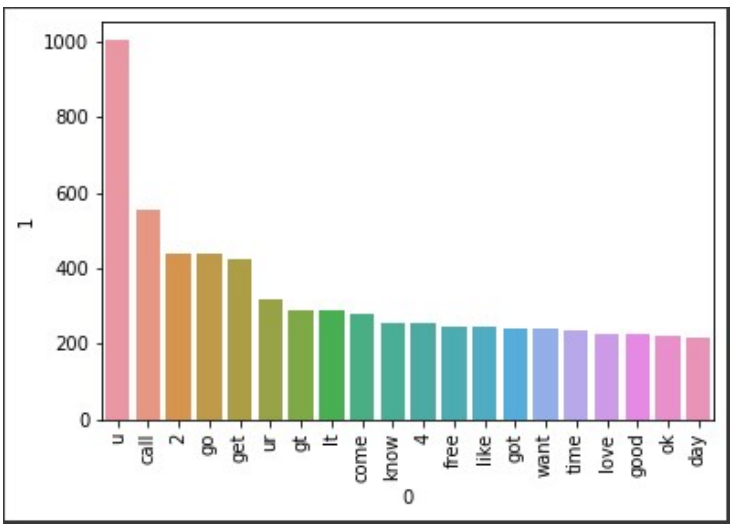

Figure 3.6: : Spam word frequency.

Chapter 4

Methodology

4.1 Proposed Model

For our project we used five machine learning algorithms like Support Vector Machine (SVM), K Nearest Neighbour (KNN), Decision Tree (DT), Random Forest (RF), Multinomial Naive Bayes. For this we pre- processed the data and split the data for training and testing the model. For training we used 80% of the data 20% for testing purpose. After training the models, we evaluated the models with testing data.

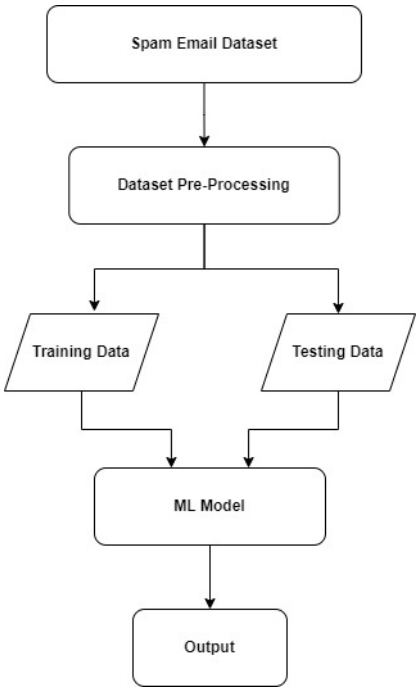

Figure 4.1: Methodology.

4.2 Feature Extraction

Here TF IDF is used in feature extraction. TF is term frequency and IDF is inverse document frequency.
TF is the measurement of how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (the total number of terms in the document) as a way of normalization:
TF(t) = Number of times term t appears in a document / Total number of terms in the document
IDF is the measurement of how important a term is. While computing TF, all terms are considered equally important. However, it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus, it is needed to weigh down the frequent terms while scaling up the rare ones by computing the Inverse Document Frequency.
IDF(t) = log (Total number of documents / Number of documents with term t in it)

4.3 Model Description

4.3.0.1 Support Vector Machine

Support vector machine is a supervised machine learning technique. It seeks to establish the best cut-off between the potential outputs. Depending on the chosen kernel function, SVM performs complicated data transformations. Based on these transformations, SVM attempts to optimize the separation boundaries between the data points according to the labels or classes established. To generalize, the objective of SVM is to find a hyperplane that maximizes the separation of the data points or margin to their potential classes in an n-dimensional space. The line closest to positive and negative points of the decision boundary is called gutter which is also parallel to the decision boundary. Distance between two gutters is called margin. The points that decides the decision boundary is called support vector.

4.3.1 K Nearest Neighbour

A powerful classification algorithm used in pattern recognition. K nearest neighbors stores all available cases and classifies new cases based on a similarity measure(e.g distance function). In this algorithm an object (a new instance) is classified by a majority votes for its neighbor classes. The object is assigned to the most common class amongst its K nearest

neighbors.(measured by a distant function).

4.3.2 Decision Tree

Decision tree may be an n-ary, n2 tree. Decision trees can be used for both classification and regression task. The top, or first node, is called the root node. The last level of nodes are the leaf nodes and contain the target label values. The intermediate nodes are the descendant or "hidden" layers. Nodes can contain one or, more questions. In a binary tree, by convention if the answer to a question is "yes", the left branch is selected. In a decision tree, the non-terminal nodes are used to make local decisions based on local information they process. Terminal nodes make the final decisions. Attribute selection for nodes is needed when splitting the data with the selected attribute, the split groups can have maximum purity. Gini impurity is used to calculate how much pure the split groups are. The lower the Gini impurity value, the better.

4.3.3 Random Forest

In machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. Random forest is a popular ensemble supervised machine learning technique for classification and regression. On various samples, it constructs decision trees and uses their majority decision for classification and average decision for regression. A group of models rather than a single model is employed to create predictions. In random forest, n records are randomly selected from a set of k records. For each sample, a different decision tree is built. An output will be produced by each decision tree. For classification and regression, the final result is evaluated using a majority vote or an average.

4.3.4 Multinomial Naive Bayes

Naïve Bayes algorithm is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Multinomial Naïve Bayes consider a feature vector where a given term represents the number of times it appears or very often i.e. frequency.
P (X = x1, x2, x3, ..., xk) = (n! * P1^{x}1 $*$ $P1^x2 * P1^x3$.....Pn^xn)/($x1! * x2! * x3!$.....$xk!$)

Chapter 5

Experiments and Results

In our project we calculated accuracy and precision for each of the models and also the confusion matrix for each of them. We mainly focused the precision value because the higher the precision the lower the False positive value.

5.1 Support Vector Machine

The model achieved 97.29% accuracy and a gave 97.41% precision on the testing data.

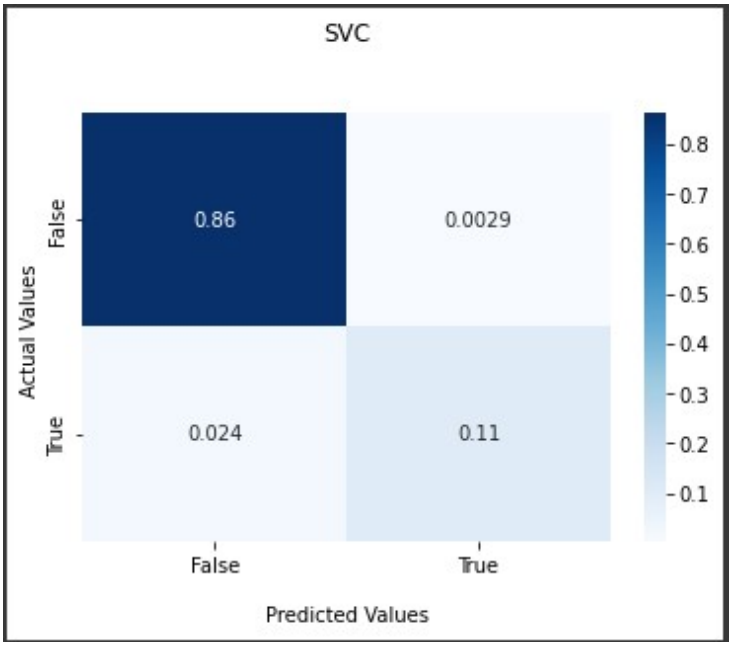

Figure 5.1: Confusion Matrix of SVM.

5.2 K Nearest Neighbour

The model achieved 90.03% accuracy and a gave 100% precision on the testing data.

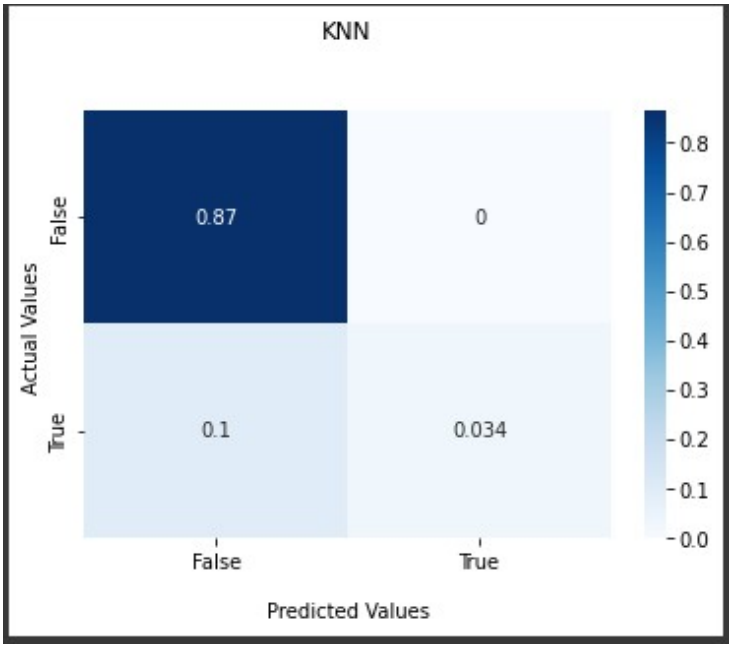

Figure 5.2: Confusion Matrix of KNN.

5.3 Multinomial Naive Bayes

The model achieved 95.93% accuracy and a gave 100% precision on the testing data.

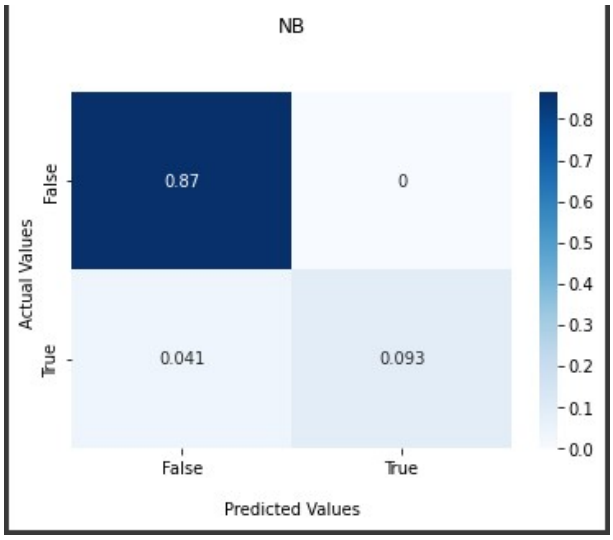

Figure 5.3: Confusion Matrix of MNB.

5.4 Decision Tree

The model achieved 93.61% accuracy and a gave 84.61% precision on the testing data.

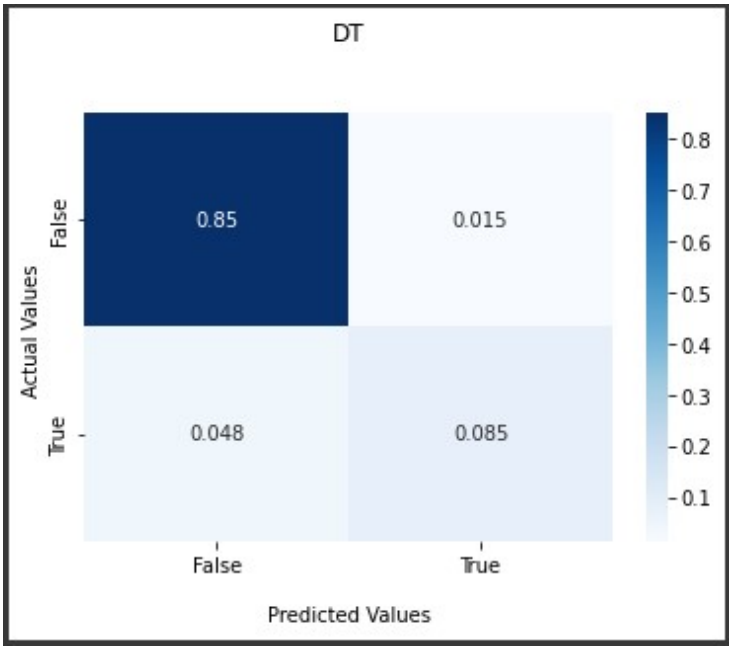

Figure 5.4: Confusion Matrix of Decision Tree.

5.5 Random Forest

The model achieved 97.38% accuracy and a gave 100% precision on the testing data.

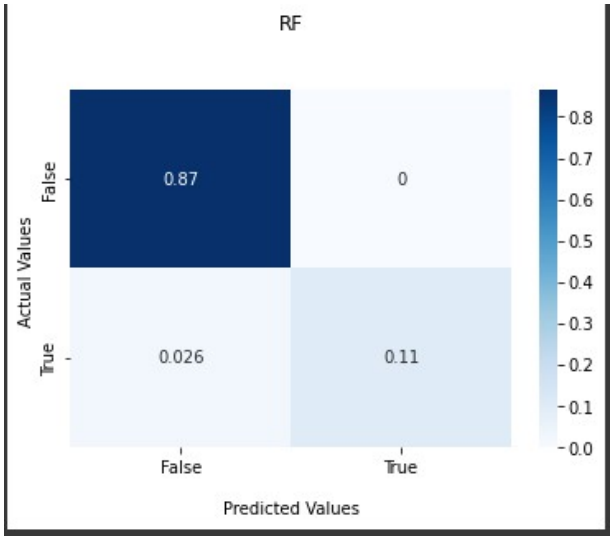

Figure 5.5: Confusion Matrix of Random Forest.

5.6 Performance of models

Table 5.1: Model Results

Model	Accuracy	Precision
SVM	97.29%	97.41%
KNN	90.03%	100%
Multinomial Naive Bayes	95.93%	100%
Decision Tree	93.61%	84.61%
Random Forest	97.38%	100%

We achieved the best result for Random forest after that Multinomial Naïve Bayes gave the second highest result.

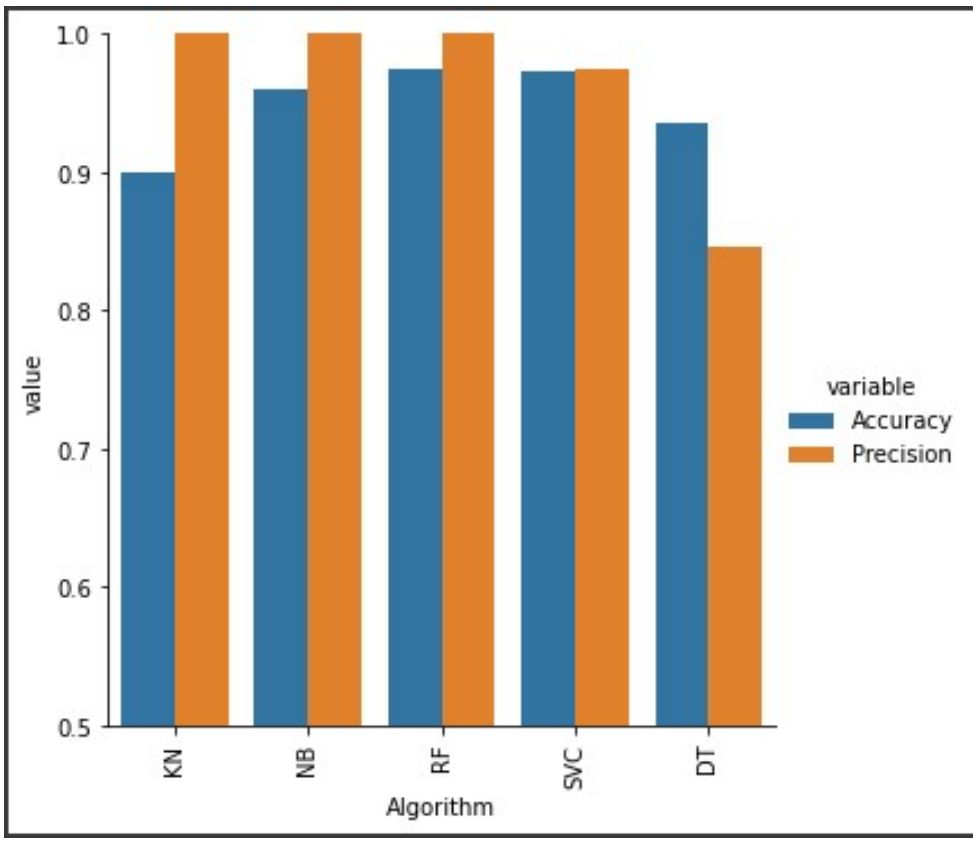

Figure 5.6: : Accuracy and Precision of each model.

5.7 ROC Curve

The receiver operating characteristic curve (ROC curve) is a graph that displays how well a classification model performs across all categorization levels. In the X axis there is false positive rate and Y axis there is true positive rate. The true positive rate (TPR, also called sensitivity) is calculated as TP/TP+FN. TPR is the probability that an actual positive will test positive. The ROC curve demonstrates the trade-off between specificity and sensitivity (or TPR) (1 – FPR).

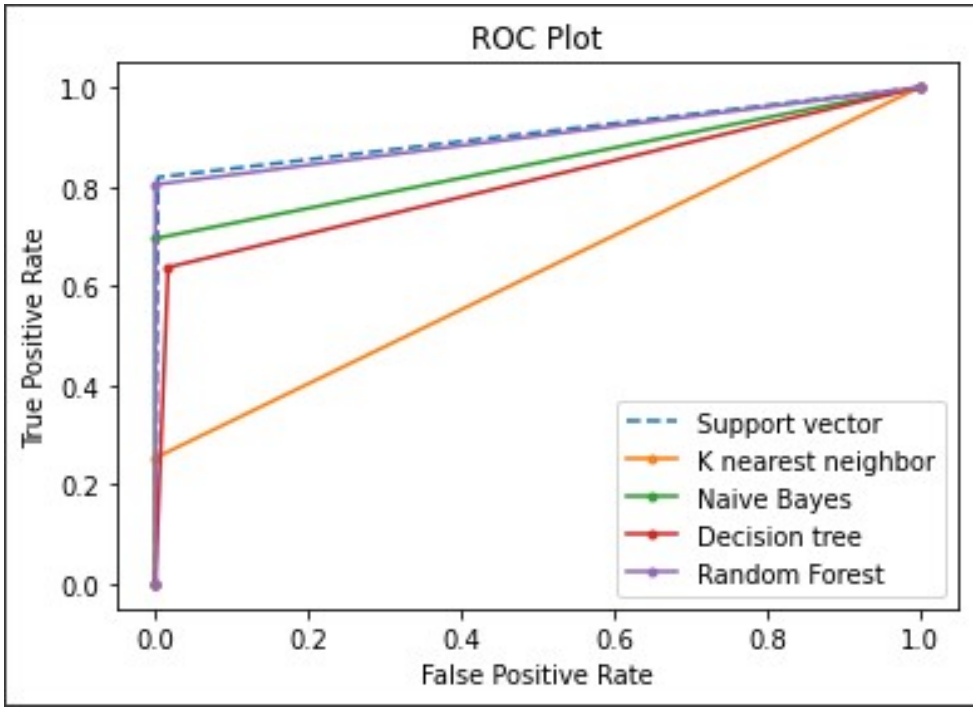

Figure 5.7: : Receiver Operating Characteristic(ROC) Curve .

Chapter 6

Future Work and Conclusion

In this project we have used the spam email dataset and experimented with five machine learning algorithms such as SVM, MultinomialNB, Decision Tree Classifier, K Nearest Neighbor and Random Forest. We trained the data and evaluated the model and found out that Random Forest model performed better than other models to detect spam email from spam sms dataset. By sorting the result according to our model it stands after Random Forest it is MultinomialNB and SVM. Lastly, Decision Tree and KNN. In future, we plan to collect a more enriched dataset which will help us to classify spam email. In addition to this, we also plan to extend this work to evaluate deep learning techniques perform.

References

[1]Improved email spam detection model based on support vector machines , Sunday Olusanya Olatunji
[2]Email Spam Detection Using Machine Learning Algorithms, Nikhil Kumar, Sanket Sonowal, Nishant
[3]Comparative Analysis of Classification Algorithms for Email Spam Detection, Shafi'i Muhammad Abdulhamid, Maryam Shuaib, Oluwafemi Osho, Idris Ismaila and John K. Alhassan
[4]Performance Evaluation of Machine Learning Algorithms for Email Spam Detection, Nandhini.S, Dr.Jeen Marseline.K.S
[5]https://www.kaggle.com/datasets/uciml/email-spam-collection-dataset

Generated using Undegraduate Thesis LaTeX Template, Version 1.4. Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh.

This project report was generated on Wednesday 7th September, 2022 at 4:25am.