OVERVIEW OF MACHINE LEARNING IN CYBERSECURITY COMPARATIVE ANALYSIS OF CLASSIFIERS USING WEKA[1]

Shweta Sharma
Guru Gobind Singh Indraprastha University

Abstract: *Technologies have made a drastic change over years from mainframe computers to laptops, from telephone to cellular phone everything is changing and becoming digital. The online platform is the new way of working whether it is related to education, social gathering or business everything is going online which is easy, comfortable and consumes less time. Smart tv smartphones smartwatches that come under the category of IoT has been deployed all over the world nowadays, features like voice recognition system face detection system have become a crucial part of the most of the smart device. Nowadays it has become an essential part of our daily life but with the benefits, there is also a major concern that is increasing day by day that is cyber-attack. Security over cyberspace is a most crucial thing what user seeks for When security & machine learning both come into one picture it makes a huge impact on user's safety. This research paper deals with the overview of machine learning and the need for machine learning in cybersecurity. I have also performed a comparison between two classifiers Naïve Bayes and decision tree by feeding the spam email dataset in the WEKA tool. The motive behind doing this classification is to check which classifier can interpret the result more accurately.*

Keyword: **Decision Tree, Machine learning, Naïve Bayes, Spam email classification, WEKA**

1. Introduction

National airlines of the country – Air India claimed to have cyberattack on its data servers which has affected around a 4.5millon people around the world [1]. Rehoboth McKinley Christian Health Care Services (RMCHCS), a non-profit healthcare provider operating in Arizona and New Mexico, has reported a data breach claiming to impact of around 200,000 patients and employees, in the investigation they had found an unauthorized party was able to access certain systems that contained patient information and remove some data between January 21 and February 5, 2021[2].Attack on cyberspace is becoming a serious threat these days according to the report published in security magazine, the world experiences 2,200 per day that turns out to be nearly 1 attack every 39 seconds in other to deal with it we have to go in-depth of the problem [3]. Pegasus software that is used for hacking affecting millions of users all over the world.

Cyber-attack is an attack when somebody intrudes into any person's or group or institution's privacy and steal confidential data from it, there can be various reasons for attacks like gaining business or customers financial details like credit card or account details, personal details about the individual, access of credential detail of employees or patient Even countries sensitive and confidential data is also under threat due to cyber-attacks according to the report published on 27 February by Subex, a Bengaluru-based firm providing analytics to telecom and communication service stated that India faced the most cyber-attacks in the world, while the US was the most cyber-targeted nation in 2019, India held the top spot in April, May and June [4]. To protect and prevent our sensitive data, hardware, software from cyberattacks we need security. In order to deal with it we use Cybersecurity to protect against unauthorized access to data and other computerized systems. It is the branch of technologies, processes and controls that deals with the security of systems, networks, programs, devices and data from cyberattacks. The objective of cybersecurity is to minimize the risk of cyberattacks and protect against the unauthorised exploitation of systems, networks and technologies. Data can be secured by various methods like network security, data loss prevention, cloud security, intrusion prevention or detection encryption or by installing antivirus in the systems.

This paper is sponsored by Mamta Sharma and Prashant Sharma.

2. Literature Review

Our world is surrounded by the internet everything is becoming smart and digital along with the deployment of internet cyber threats is also increasing at large scale which can be very harmful to mankind as unwanted malicious software can find a way to steal your confidentiality to prevent them from harming or security cybersecurity come into the light. We can use machine learning and data mining for the security of our devices as these two plays an important role in cybersecurity. Data mining and Machine learning are interrelated as Data Mining utilizes techniques created by machine learning for predicting the results while machine learning is the capability of the computer to learn from a minded data set.

One of the widely used Data mining techniques for the implementation of classification procedures is the Decision Tree by Dr Archana Saxena. It is very easy to implement and generate Classification rules that a layman can practice to generate a class label for the unknown dataset. The decision tree is also used in the Trust prediction of the cloud provider [5].

Bhatia in 2015 proposed secure group communication techniques using steganography for communication of secret messages on the internet. To maintain the security of secret messages Bhatia in 2014 proposed an image steganography method using the spread spectrum approach. In the proposed technique author uses the properties of orthogonal image planes and the secret message is modulated using pixels of one image plane of the cover image. The modulated message is then replaced by the pixels of another image plane [12].

Bhatia in 2019, proposed a message hiding technique, in this technique author used solutions of Knight tour and 8-Queen's problem in an 8*8 chessboard. The proposed technique applied solutions of moving knight tour and of placing 8-Queen's in a non-attacking manner in 8*8 chessboard to select pixels for embedding secret message bits.[13][14]

Nurul Fitriah Rusland, Norfaradilla Wahid, Shahreen Kasim and Hanayanti Hafit have proposed their work on Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets where they have taken multiple datasets and performed analysis among them [10].

A comparative study for Spam Classifications in Email Using Naïve Bayes and SVM algorithms has proposed by the author in which they worked on two datasets in order to produce classification using two classifiers and compared their outcomes with each other [11].

Chugh A., Sharma V.K., Bhatia M.K., Jain. C (2021) A Big Data Query Optimization Framework for Telecom Customer Churn Analysis. In: 4TH International Conference on Innovative Computing and Communication, Advances in Intelligent Systems and Computing. Springer, Singapore.
The researcher proposed a big data query optimization system for customer sentiment analysis of telecom tweets. The suggested hybrid system leverages the recurrent neural network, a deep learning technique for efficient handling of big data and spider monkey optimization, a metaheuristic technique that helps to train the network faster and enables fast query handling. The comparison is made with deep convolutional networks and model optimized the efficiency and performance for predicting customer churn rate.
The author propounded a Spider Monkey Crow Optimization (SMCO), a hybrid model for sentiment classification and information retrieval of big data. The model was compared with famous sentiment classification techniques and outputs showed an accuracy of 97.7%, precision of 95.5%, recall of 94.6%, and F1-score 96.7% respectively [15].

3. Background Analysis

3.1 How does attack happen?

Communication in devices happens through IP addresses and port numbers if the source device wants to send data to the destination machine it will need the destination's IP address and port number but in the case of attack the attacker lists into that traffic between both the devices, once the data send from sender's side attackers steal it in between, it is very difficult to find the intruder as they hide their activities from the intrusion detection system. Attackers hide from various methods like covering their tracks by editing system logs

3.2 Types Of Cyber Attacks

Pharming- is a type of social engineering cyberattack in which criminals redirect internet users trying to reach a specific website to a different, fake site [8]. These spoofed site's objective is to capture a victim's personally identifiable information (PII) and log-in credentials, such as passwords, social security numbers, account numbers, and so on so that they can attempt to install pharming malware on their computer.

Malware - malicious software like spyware, scareware ransomware, virus and worms come under this category where malware breaches into the system through various vulnerabilities like clicking into some link or an email link can do a lot of harm to a computer like obtaining information from the system, can make the component in operatable.

SQL injection – it occurs when an attacker inserts malicious code into a server that uses SQL to gain confidential data from the server.

Man In The Middle Attack in which attacker inserts itself into a two-party transaction to steal data these attacks generally happens on the public WIFIs where the security of the network is low or doesn't exist anywhere.

Denial-Of-Service where attacker floods systems, networks which exhaust bandwidth. As a result, the system becomes insufficient to handle requests. Take the advantage of this attacker use multiple compromised devices to launch this attack.

A **zero-day** exploit hits after the network having a vulnerability concern before its solution is implemented attacker target the disclosed vulnerability during this window of the daytime.

3.3 Phases of Attack

Attacks generally occur in 5 phases. **Reconnaissance** in this step attacker gathers information about the target that has to be attacked once the information gathered it will check for the vulnerabilities or the weakness in the **scanning phase** like various ports that are open or can easily get attacked so that he can easily penetrate into the network or system any vulnerability found the attacker try to **gain access** to get into those ports once he gains access **he keeps access** it by ensuring the way back to target machine for further procedures it and last he gets into the computer he **covers his tracks by hiding or deleting or modify logs** so that no one can follow him back [8].

4. Overview of Machine Learning

Machine Learning is the branch of artificial intelligence that helps the system to learn and analyze trends from the past data, made the prediction models, whenever a new dataset is received, get a prediction for the output also, the amount of data plays an important role to build a better model which can predict the output more accurately. Generally, the 80:20 rule is used for implementation, in which they train the machine by 80% dataset and the remaining 20% data is used for testing purposes. It also helps in developing various algorithms and making predictions using historical data or information. It enables the machine to automatically train itself from the given dataset, improve performance from experiences, and predict things without being explicitly programmed. For solving a complex problem, we only need to provide data to generic algorithms, and with the help of these algorithms, the machine builds the logic as per the data and prediction of the result. There are 3 types of machine learning. The first one is supervised learning where the machine is taught using labelled data whereas in unsupervised learning machine is trained upon unlabelled data with any supervision or guidance and the last one is reinforcement learning where an agent interacts with its environment by producing actions and discover errors or rewards.

Diagram has illustrated the typical workflow of machine learning:

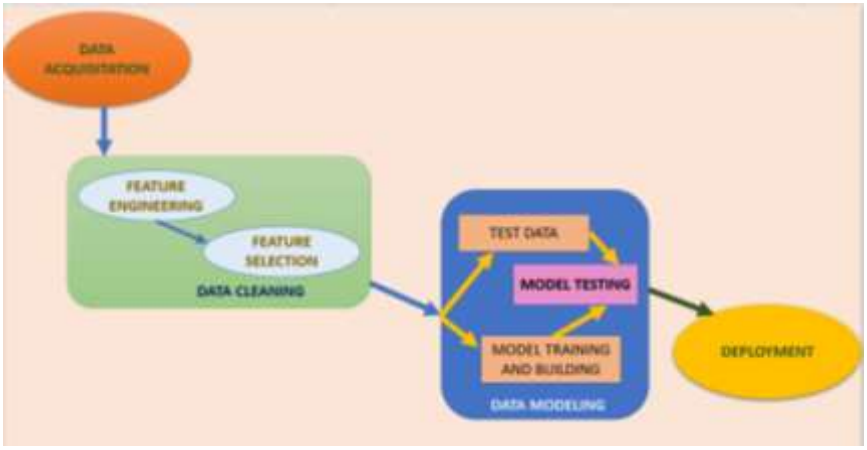

Figure 1: Workflow of machine learning

Machine learning can be used in various cases such as image recognition, speech recognition, Facebook auto-tagging, self-driving care recommender systems. Companies like Netflix, Amazon have built ml models in order to analyse the trend in order to recommend products and services of user's interest.

4.1 Machine Learning Techniques

1) Classification:

A classification is a form of supervised learning where it predicting about the label or class. In cybersecurity, spam detection is used by ML-based classifiers which involve discriminating given email messages as spam or not. The spam filter models are used to separate spam messages from non-spam messages.

2) **Regression:**

This is also a part of supervised learning. It predicts about continuous quantity the value of a dependent feature is estimated based on the values of the independent features by learning from the existing data related to past events and such knowledge is used to handle new events [6].

3) Clustering

Unsupervised learning model, which extracts general patterns from the data when the class label is not defined. [6] Groups of similar events constitute a cluster as they share common features that define a specific(behaviour)pattern.

4.2 Need of Machine Learning in Cybersecurity

The need for machine learning is increasing gradually because of the capability of doing complex tasks in a short duration of time. There are different types of attacks that are difficult to detect in order to prevent them from harming the system data we need strong support that is available in the form of Machine learning. Cyber-crime sales are becoming a new source of cyberattack where attackers sell various hacking services in the form of ransomware, malware software that is victimizing people on large scale apart from that there is more network log that makes it difficult to analyse it manually. IPS/IDS are for pattern-based devices follows signature-based which is based on the existing knowledge of the pattern that detects an attack on the basis of supervised learning where device familiar with the type of attack store it in its database so when a data packet comes it will compare it with the previous packet that was of attack it gives the reassembles of previous one it will consider as threat block its entry but if the system is not able to match the pattern it will not be able to detect the attack that going to be a major threat for the system cannot achieve zero-day attack with the technique, this problem can be solved by machine learning using anomaly detection algorithms. It is helpful in the security of big data and can also help in detecting cyber fraud detection. Attackers are playing very smart these days they are hiding their Ip address pretending insight from the user's system by the use of quantum computing, sending scam emails for stealing money can also be solved with the help of withdrawal anomalies email filtering. Data mining and Machine learning are interrelated as Data Mining utilizes techniques created by machine learning for predicting the results while machine learning is the capability of the computer to learn from a minded data set.

5. Spam Classification:

Spam is the latest ongoing threat that can not only harm your device but also manages to trick the victim and get him/her fall into a financial trap so in order to secure our email we need to differentiate which mail is spam and which mail is not. There is a various classifier which can be used for spam detection but in this paper, I will discussing be about Naïve Baye and Decision Tree and based on their outcome I will be performing a comparative analysis of them. I will be using WEKA for the classification. Weka is an assemblage of various machine learning algorithms for data mining tasks. It is an open-source software developed at the University of Waikato New Zealand for knowledge analysis. The algorithms can either be applied directly to the dataset you can upload any file or internet contain format like ARFF, CSV etc. It includes various tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

5.1 Naive Bayes

Naïve Bayes algorithm is a form of supervised learning used for classification, which is based on the mathematics Bayes theorem, as Nurul Fitriah Rusland et has already explained in his work [10]. It is a probabilistic classifier used to predict the outcome based on the occurring of the object. And based on this classifier we will perform spam classification for email. For data, I have taken the SPAMBASE dataset [9]. The dataset contains 4601 E-mail messages and 58 attributes, which is sufficient to detect and predict whether the given dataset is spam or not.

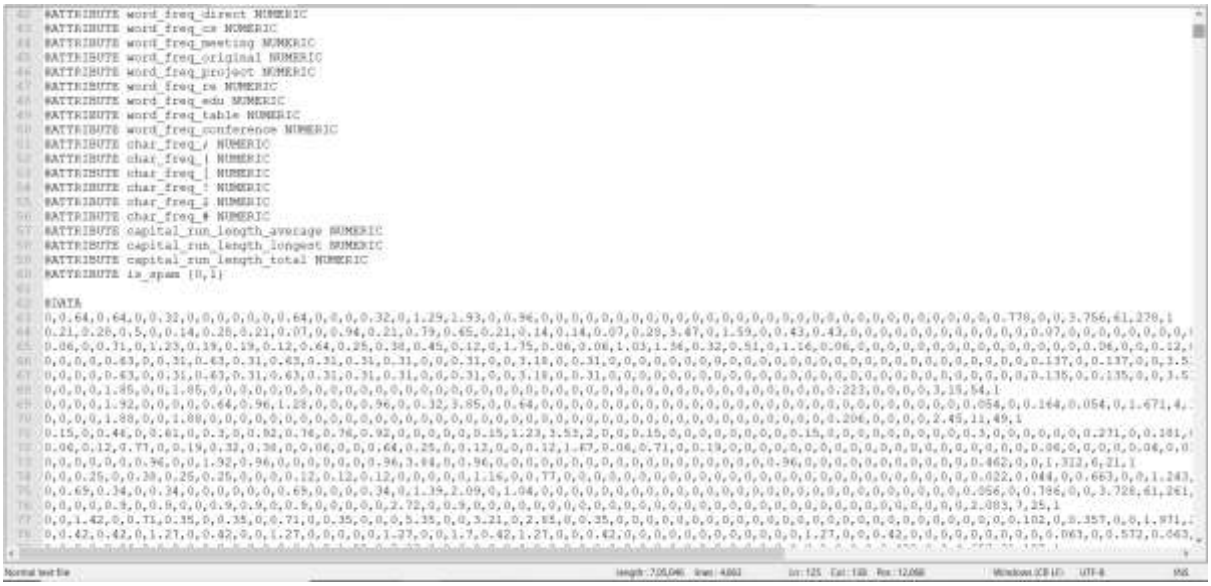

Figure: Representation of SPAMBASE dataset

Now we will be performing the classification technique on the following dataset with the help of Weka. Data needs to be pre-processed first like Attribute's capital run-length average, capital run-length longest and capital run-length total are removed from the list by checking the box to their left and hitting the Remove button, then classification algorithm needs to be applied on the same.

Outcome:

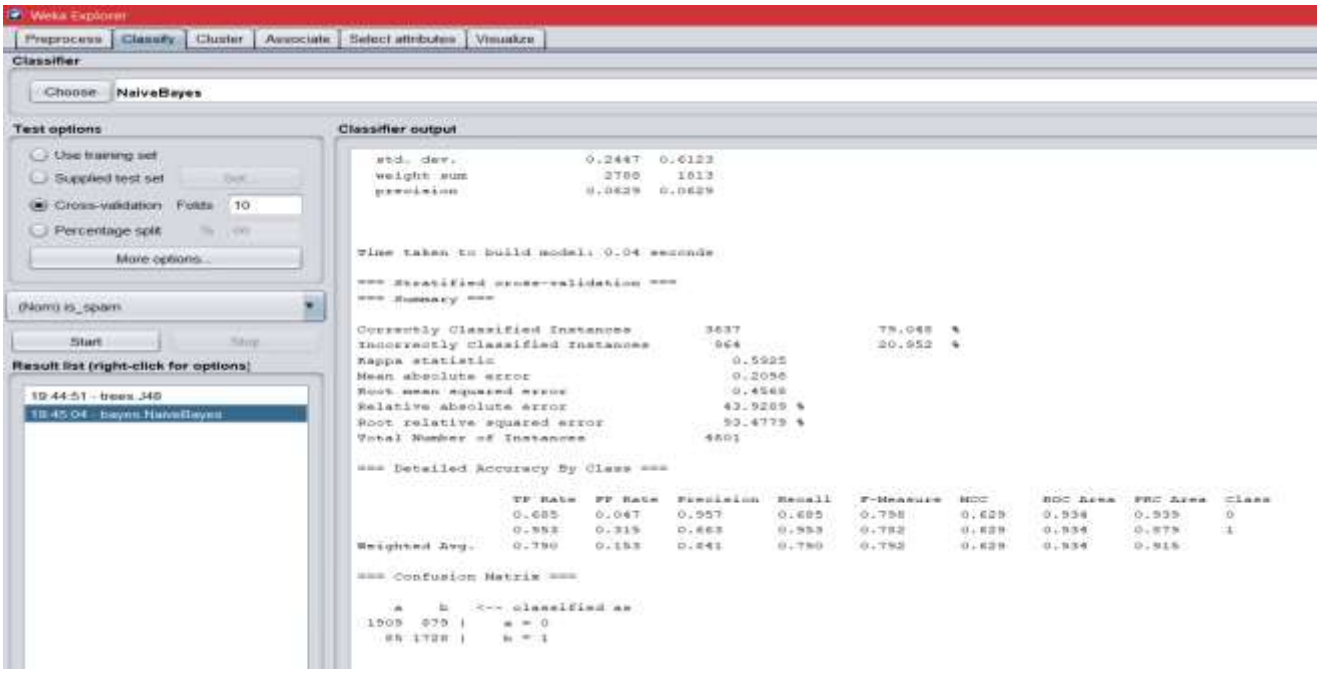

Figure2 : Classification outcome for Naïve Baye

5.2 Decision Tree:

It is also a classification technique for decision support. Decision Tree uses a tree-like structure model used for making decisions where test on an attribute denoted by each internal node, and an outcome of the test is represented by the branches, and each leaf node (terminal node) holds a class label. Let's understand with an example:

The following dataset is for the prediction of whether a child can play outside or not. The dataset for the following attributes and data.

```
@relation weather.symbolic

@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,mild,high,FALSE,yes
rainy,cool,normal,FALSE,yes
rainy,cool,normal,TRUE,no
overcast,cool,normal,TRUE,yes
sunny,mild,high,FALSE,no
sunny,cool,normal,FALSE,yes
rainy,mild,normal,FALSE,yes
sunny,mild,normal,TRUE,yes
overcast,mild,high,TRUE,yes
overcast,hot,normal,FALSE,yes
rainy,mild,high,TRUE,no
```

Figure 3: Inbuilt Weak Weather dataset interpretation

On the basis of the data, we got given classification and prediction from Weka.

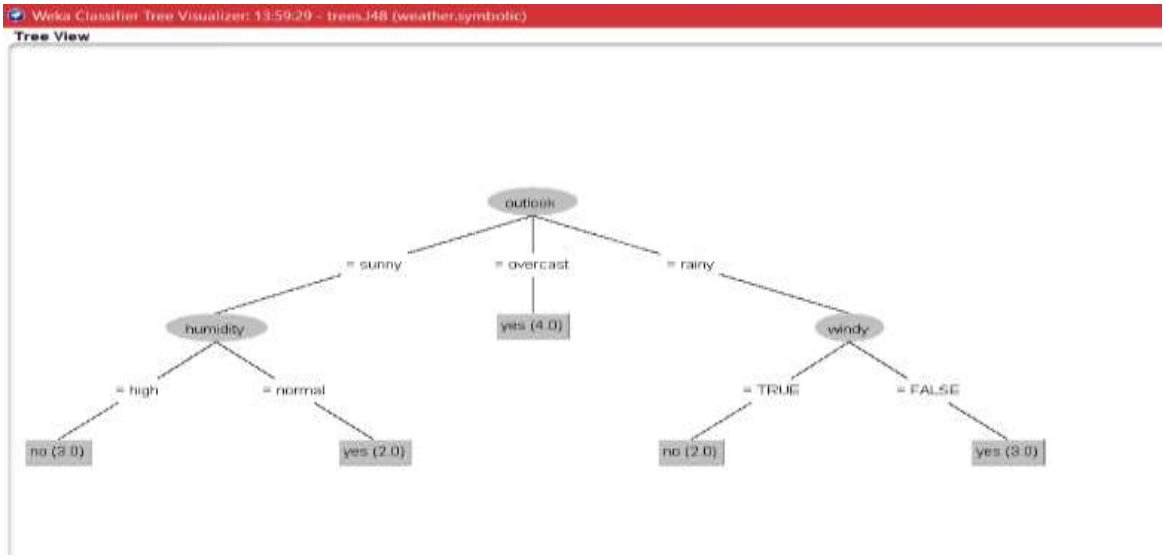

Figure 4: Classification of the above dataset

You can see from the example If the outlook is sunny then it will further get classified on the basis of high and normal if the humidity is high then the child cannot play if it is normal then he/she can play.

Based on the decision tree classifier we will try to implement spam classification on the same SPAMBASE dataset [9]. I have used the j48 classifier for the classification as it is considered a machine-learning algorithm to examine the data categorically and continuously.

The outcome for the spam email classification using a decision tree is:

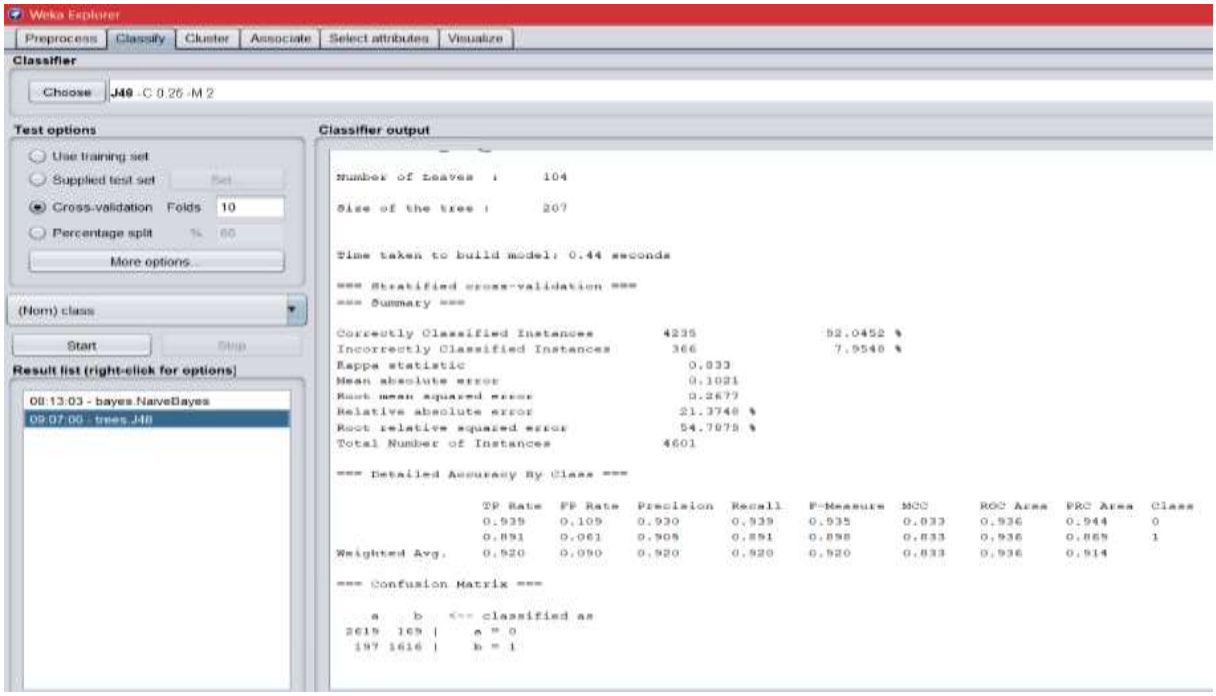

Figure 5: Classification outcome for Decision Tree

COMPARISON OF BOTH CLASSIFIERS:

Algorithms	Correctly Classified Instances	Incorrectly Classified Instances
Naïve Baye	79. 048 %	20. 952 %
Decision Tree	92. 0452 %	7. 9548 %

Table 1: Comparison of both classifiers on the bases of correctly, incorrectly classified instances

Classified as	A	B
a=0	2619	169
b=1	197	1616

Table 2: Confusion Matrix

Based on the above confusion matrix, accuracy, precision, recall can be calculated as:

$$ACCURACY = \frac{TN+TP}{TN+TP+FN+FP} \qquad PRECISION = \frac{TP}{TP+FP} \qquad RECALL = \frac{TP}{TP+FN}$$

Algorithms	Accuracy	Precision	Recall
Naïve Baye	79.04%	84.1%	79%
Decision Tree	92.04%	92%	92%

Table 3: Comparison of both classifiers in terms of accuracy, recall, precision

6. Conclusion

E-mail spam filtering is an important concern in the network security and machine learning techniques, from the above experiment we can conclude that the Decision tree has a high classification rate for the SPAMBASE dataset. But we cannot rely on one algorithm as there is no algorithm that can classify both spam and ham (non-spam) always without any oversight. So, the Decision Tree algorithm is recommended as per my observation to classify the spam and not-spam emails as it exceeds the naïve Bayes algorithm in terms of accuracy, precision, recall, correctly incorrectly classifies instances.

Acknowledgments

I would like to sincerely thank my mentor my guide **Dr Archana Saxsena**[i]**,** who help me in carrying out the research**. Mr Manoj Joshi**[ii] for inspiring me. Also, I would like to show gratitude to my parents who have supported and motivated me for writing the paper.

REFERENCES

[1] Air India cyber-attack: data of millions of customers compromised. *Available: https://www.bbc.com/news/world-asia-india-57210118.*

[2] US healthcare non-profit reports data breach impacting 200,000 patients, employees' 25 May 2021 at 12:58 UTC
Available: https://portswigger.net/daily-swig/us-healthcare-non-profit-reports-data-breach-impacting-200-000-patients-employees.

[3] Cybersecurity Statistics for 2021 Jacob Fox Mar 1, 2021.*Available: https://cobalt.io/blog/cybersecurity-statistics-2021.*

[4] India was the most cyber-attacked country in the world for three months in 2019 by THE PRINT *Available: 3 March 2020 3:25 pm IST https://theprint.in/tech/india-was-the-most-cyber-attacked-country-in-the-world-for-three-months-in-2019/374622/.*

[5] Saxena, A. B., & Dawe, M. (2020). Decision Tree: A Predictive Modelling Tool used in cloud Trust Prediction. *Published in International Journal of Engineering and Advanced Technology (IJEAT), ISSN:2449-8958, Volume -8, Issue-6, August 2019.*

[6] A Survey on Machine Learning for Cyber-Security A. Lakshmanarao M. Shashi. *Published: International Journal Of Scientific & Technology Research Volume 9, Issue 01, January 2020 ISSN 2277-8616.*

[7] A Review on Cyber Security and the Fifth Generation Cyberattacks A. Saravanan1 and S. Sathya Bama published *ISSN: 0974-6471, Vol. 12, No. (2) 2019, Pg. 50-56(www.computerscijournal.org).*

[8] The Attack Types and Phases
Available:http://etutorials.org/Networking/Cisco+Certified+Security+Professional+Certification/Part+V+Intrusion+Detection+Systems+IDS/Chapter+23+Intrusion+Detection+System+Overview/The+Attack+Types+and+Phases/#:~:text=The%20three%20types%20of%20attacks,attack%20on%20the%20network%20resources.

[9] SPAMBASE from the UCI machine learning repositories was created by Jaap Suermondt, George Forman, Erik Reeber, and Mark Hopkins *Available: http://archive.ics.uci.edu/ml/datasets/Spambase.*

[10] Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets Nurul Fitriah Rusland 2017
Published: IOP Conf. Series: Materials Science and Engineering 226 (2017) 012091 doi:10.1088/1757-899X/226/1/012091.

[11] A Comparative Study for Spam Classifications in Email Using Naïve Bayes and SVM Algorithm by Ziyan Mohammed Mohammed Farhaz J A Mohammed Irshad M P Bearys Institute of Technology Bearys Institute of Technology Bearys Institute of Technology Mangalore, India Mangalore, India Mangalore, India Mustafa Basthikodi Dept. of CSE Beary's Institute of Technology Mangalore, India Ahmed Rimaz Faizabadi Dept. of CSE Beary's Institute of Technology. *Published: www.jetir.org (ISSN-2349-5162).*

[12] Bhatia, M. P. S, Muttoo, S. K. and Bhatia, M. K (2014). "An Image Steganography Method Using Spread Spectrum Technique" in Springer sponsored International Conference on Soft Computing for Problem Solving (SocProS 2014) organized by *NIT SILCHAR, Assam, India in 2014, pp:219-236.*

[13] Bhatia, M. K. (2017), "8-Rooks Solutions for Image Steganography Technique*", International Journal of Next-Generation Computing (ISSN: 2229-4678 (Print) and 0976-5034 (Online)), Vol. 8 no. 2, pp. 127-139, July 2017.*

[14] Bhatia, M. K. (2019), "Knight Tour for Image Steganography Technique", International Journal of Engineering and Advanced Technology (IJEAT) (ISSN: 2249 – 8958), Volume-9 Issue-1, pp. 1610-1613, October 2019.

[15] Chugh A., Sharma V.K., Kumar S., Nayyar A., Qureshi B., Bhatia M.K., Jain. C (2021) Spider Monkey Crow Optimization Algorithm with Deep Learning for Sentiment Classification and Information Retrieval. *Published in IEEE Access, vol. 9, pp. 24249-24262, 2021, DOI: 10.1109/ACCESS.2021.3055507.*

i)Assistant Professor at JIMS college.
ii)Assistant Manager at Accenture.