Identification and Classification of Phishing Websites Using Machine Learning – Random Forest

MSc Internship
Cyber Security

David Damilola Ibinaiye
Student ID: x18170455

School of Computing
National College of Ireland

Supervisor: Imran Khan

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	David Damilola Ibinaiye		
Student ID:	X18170455		
Programme:	M.Sc. Cyber Security	**Year:**	2019
Module:	Academic Internship		
Lecturer:	Imran Khan		
Submission Due Date:	Thursday, 12 December 2019		
Project Title:	**Identification and Classification of Phishing Websites Using Machine Learning – Random Forest**		
Word Count:	3555 **Page Count:** 15		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……………………………………………………………………………………………………

Date: 12 Dec. 19

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Identification and Classification of Phishing Websites Using Machine Learning – Random Forest

Ibinaiye David Damilola
Student ID – x18170455

ABSTRACT

Majority of our day to day activities involves sending and receiving information via the Internet. Although the major threat to the acceptability and public embrace of the Internet is the perceived increase the rate of Cybercrimes, the ease of transaction online has encouraged many people to embrace the Internet as a reliable platform. Financial Institutions, Educational Institutions, Medical Services etc. all have platforms where users can visit whenever the need for their service arise. As such, sensitive data are used as means of validation of each user of the Internet. A good example is Internet banking which requires user supplying information such as username, PIN, Password, token etc. One major way in which User sensitive and personal data are being compromised is Phishing. Phishing has really cost some unsuspecting Internet users a lot of their fortune. The term phishing refers to the process whereby Internet fraudsters present a look-alike website just to deceive Internet users into releasing their sensitive data. While some other research papers have sort to use other Techniques and Algorithm, the focus of this research work is to make use of Random Forest Algorithm to Identify and Classify Phishing Websites.

1. INTRODUCTION

Phishing is an Internet crime which is targeted at vital information of the victim. It is a cybercrime technique used in getting personal information through Email links. Many people have been a victim of Phishing. It is a serious Security problem involving the mirroring of legitimate websites to deceive Internet users so as to [1] steal important information from them. Phishing emails do seem to have been sent from a credible Sender which is why it is sometimes difficult to differentiate it from Legitimate emails. While other forms of online data theft such as Spoofing, spamming are easily detected by email filters, the case of Phishing has been different because of the highly sophisticated mode of operations being used by Internet Fraudsters. They are spurious websites created by people to imitate web designs of genuine websites [2]. They then request for your personal data — such as credit card number, username, Private Identity Number (PIN), security number, account number or password. At times, phishing attempts come from websites, service providers and companies with which you have no an account.

Internet fraudsters set to deceive their targets by requesting them to click on a link sent via Email. Internet fraudsters also deceive their victims by creating a replica of a known website. Some make use of software that can grab the code of the website, save it using the same web design programming language and then host it on another web server. Phishing attack majorly makes use of Social Engineering exploits to deceive victims by sending spoofed hyperlinks which redirects the victims to a fake web page [3].

Legitimate organizations would never request for this kind of information from you via email. It could also be referred to as a high-level form of Social Engineering. This fraudulent technique has been used to deceive a lot of Internet users. The Hackers make it look like you are communicating with a trusted party meanwhile the link you are been redirected to is fake. A lot of people have lost their fortune to this act. Phishing websites are also used to attract Internet users into divulging their Credit Card data. Since it is easy to create the replica of any website using Hypertext Mark-up Language (HTML), victims of Phishing attacks will find it hard to believe that they are currently surfing an illegitimate website. There are different strategies being used to make Phishing websites attractive and convincing to victims. Few of such offers are;

- Scintillating Offers and convincing statements to attract your attention immediately you read it.
- Immediate Call-to-Action: This technique will implore you to act fast as the offer has limited expiry time. Some will even inform you of an impending suspension of your Account if you fail to take an immediate action.
- Fake Hyperlinks and Image Links: the link to the redirected page usually looks very much like a legitimate one. For instance, www.bankoflreland.com in this link, letter ‘i' has been replaced with ‘l’. Always ensure you hover on a link before clicking it to view the spelling of the redirection page

2. LITERATURE REVIEW

Quite some numbers of Scholars have proposed different means of classifying and identifying Phishing Websites. This segment seeks to review some past work done on this same Topic. It evaluates the approaches used as well as the result and level of accuracy derived.

Yaokai Yang [13] in his work as an Internet based crime which attempts to trick unsuspecting users to reveal their personal data. Spoofed email is one of the instruments used by Hackers to deceive their victims into giving out their information. In his research, He proposed an Effective Phishing detection using Machine Learning Approach. He made use of feature selections which begins with the study of differences in behavior of phishing and genuine websites. He proposed a detection system that crawls websites and automatically detects malicious websites. His detection system makes use of supervised learning algorithm with rich set of features. His proposed system is only suitable to be applied by top Service providers, it is not effective for all Search Engines. The approach achieved 92% accuracy.

Maher Aburrous et al. [4] in their work, described phishing as a semantic attack aimed Users of Computer. Data Mining classification techniques such as JRip, PART as well as C4.5 algorithms where adopted for learning and comparison of relationships of the different Phishing classification. Phishing is a global issue that involves stealing of online data [1]. Padmaprabha et al. made use of feature selection using rough sets and ant colony optimization for detection of Web Phishing. The feature reduction algorithm was implemented to get features. Raw dataset was supplied to WEKA. The result was compared to other models. Feature Selection was able to detect smaller percentage of attributes in the normalized dataset. Anukool et al. [12] explains that Phishing activities give opportunities to criminals to tamper with Sensitive information of the user. Some features such as popup window, IP address, URL where collected to differentiate between Phishing websites and genuine ones. Ishika in his work proposed a Wrapper Based Phishing Detection. His submission is that Phishing reduces People’s trust in Online transaction which in turn affects the progress, acceptability and development of Electronic Market. He adopted decision trees and rule induction for classification learning. He collected features such as Lengthy URL, URL with an Internet Protocol address, Domain age and made use of wrapper feature selection to conduct a search of all possible feature subsets. The wrapper feature selection carries out search of all possible feature subsets. His method however consumes more time as well as extra computation. Aburrous et al. [4] in their research paper titled Predicting Phishing Websites using Classification mining Techniques with Experimental Case Studies considered two case studies. The first being Phone Phishing while the second was Website Phishing. They got authorized permission to lure some employees into giving away their personal Electronic banking account details through social Engineering calls. Many of them fell for this trick. A replica of a website was also created to deceive some set of people, about 44% also gave out their sensitive bank details. Data Mining classification and association rule approach was then used to compare how related the phishing classification were. Their result showed that factors like page Style and content were not significant. Classification Data Mining Techniques will be suitable for recognizing electronic Banking Phishing websites. The Research shows how feasible Associative Classification usage will be in live Applications involving large data warehouse. Nandhini et al. [1] used Web Mining Technique for Extraction of Features as well as the Classification on Phishing Websites. According to their research work, Phishing is a classification issue in Data Mining. It is a security issue which involves cloning or duplicating authentic websites to deceive visitors and gain access to their information [1]. The different phishing datasets of websites were used to test the classification mode and find out the most efficient.

Thomas et al. [5] collected URLs of some websites and some web browser history. Supplied it into the program. After processing both the phishing and non-phishing URLs were classified. Their experiment showed the characteristics of Phishing URLs were different from other URLs. This is as a result of the difference in the lengths of the Phishing and Non-Phishing URLs. It was also discovered that the Phishing URLs had the name of their target brand attached to their name.

Navin et al. [6] made use of Deep Learning Algorithm for Detecting Phishing. In their work, they focused on Uniform Resource Locator to differentiate between fake and genuine websites. Random Forest Algorithm and Deep learning algorithm was then used to process the URL. Classification of URL using criteria such as Internet Protocol, @ symbol to identify domain names of Phishing Websites. The method could not work for mobile websites. The level of accuracy was not efficient. Islam et.al [7] applied several forms of Machine Learning Classification Techniques ranging from Naïve Bayes, Random Forest, Decision Tree, Neural Net, Random Forest. At the end of their Research, the performance of the Machine Learning Algorithms was measured and compared to check for accuracy. Their result showed that of all the Classification Algorithms used, Random Forest gave the highest accuracy.

Islam et.al [7] focused on detecting of phishing website domain names features. Their research only checks the validity of URLs.

RESEARCH QUESTION

Is it possible to efficiently identify and classify phishing websites using random forest?

Internet is the fastest means of getting Information. Users of the Internet rely basically on the Internet to retrieve information. With the advent of Search Engines, it is easy to redirect a user to a Phishing website especially when such user is not sure of the domain name or link to the website He or She intends to visit. Users at times unconsciously click on links with hints that looks like the URL they intend visiting. In this research, our aim is to implement Machine Learning Technique for the purpose of classification and identification of Phishing Websites. The Algorithm to be used is Random Forest.

3. RESEARCH METHODOLOGY

This research was carried out with the use of a set of data retrieved from a public data platform (Kaggle.com). Kaggle is a free data warehouse for datasets. This dataset contains over Eleven thousand (11,000) labelled data. It contains different Uniform Resource Locators (URLs) of phishing websites. A spreadsheet file containing phishing websites was extracted and converted to the comma separated version (.csv). The dataset is large; therefore, we chose only 1,000 labelled data instances from the whole dataset for quick evaluation. The whole dataset and the small file were both kept in the project folder for referencing. The whole dataset has a size of 825KB.

3.1 Machine Learning

The quest to create computer programs and applications that improve with experience has led to the discovery of Machine Learning. An application is said to learn from Experience E with respect to class tasks *T* and performance *P*, if its performance at tasks in *T*, as measured by value *P*, improves with Experience *E* [8]. Machine Learning is the process of automating the detection of useful data patterns [9]. It is a branch of Artificial Intelligence which focuses on making it easy for machines to carry out their specified tasks with the use of application software that are intuitive. It involves a data to be learnt, pre-processed and trained. It came about as a result of the intersection of Computer Science and Statistical Analysis [10]. For this research, we adopted Machine Learning Concept to achieve our result.

3.2 Random Forests Algorithm

Random forest is a supervised Learning algorithm which comprises of a large set of individual decision trees. The strength of Random Forest Algorithm is in the fact that the higher the number of trees present in the forest the robust the forest looks like. This also applies to the random forest classifier, because the higher the number of trees in the forest the higher the probability of result accuracy. Let’s assume you would like to predict whether your partner will like the new brand of Car you bought for her or not. To model the decision tree, you will use the training dataset like the features of the vehicles your wife has used in the past. So, once you pass the dataset with the target as your partner will like the car or not to the decision tree classifier. The decision tree will start building the rules with the features your partner like as the nodes while the like or not as the leaf nodes. By examining the path from the root node to the leaf node. You can deduce the rules. Each decision tree relies on the value of a random vector. It is a learning method for Data Classification and Regression. It is a model built on decision trees for making predictions whereby the largest set of trees forms the final decision. A decision tree is pictorially represented by a tree for predicting a course of action. Random Forest Algorithm is one of the most effective models for Machine Learning because of its reduced training time and high accuracy level. The advantage of Random Forest Algorithm over other Machine Learning Algorithm is that you can use it for both Classification and regression task. It doesn't overfit and can as well handle missing values

3.3 Phishing

This is a practice of using fraudulence electronic means to obtain information from unsuspecting users. It is usually done via emails; fake website URLs etc. unsuspecting victims reveal their vital financial and personal data thinking they are on a safe and secure platform. A lot of people all their money due to this act. There is need to prevent this occurrence. This research will try to improve on existing prevention methods.

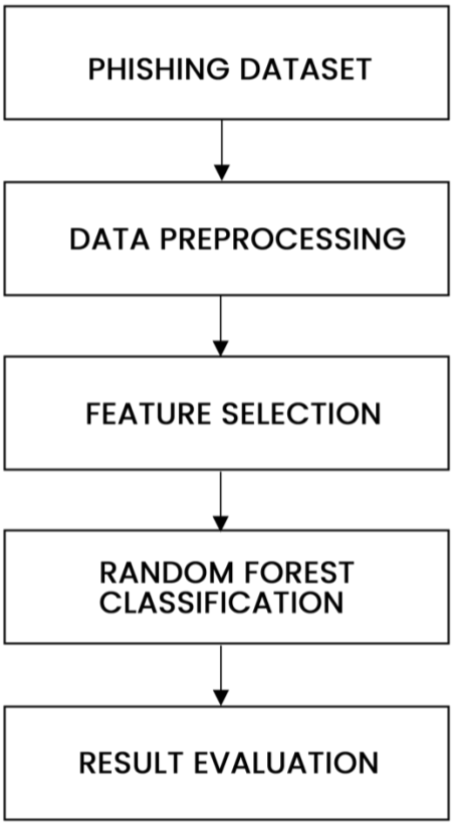


Fig. 1.

3.4 Attributes

The data used for the research contained numerous attributes selected for detection and assessing of possible Phishing feature. The dataset contained over two thousand (2,000) records, we however reduced to three hundred for testing.

- Request_URL
- URL_of_Anchor
- Links_in_tags

- Submitting_to_email
- Abnormal_URL
- Redirect
- on_mouseover
- RightClick
- popUpWidnow
- Iframe
- age_of_domain
- DNSRecord
- web_traffic
- Page_Rank
- Google_Index
- Links_pointing_to_page
- Statistical_report

4. IMPLEMENTATION

This section describes the process used for the actualization of the Project aim. The steps described in this section was adopted and applied for both identification and classification of Phishing Websites.

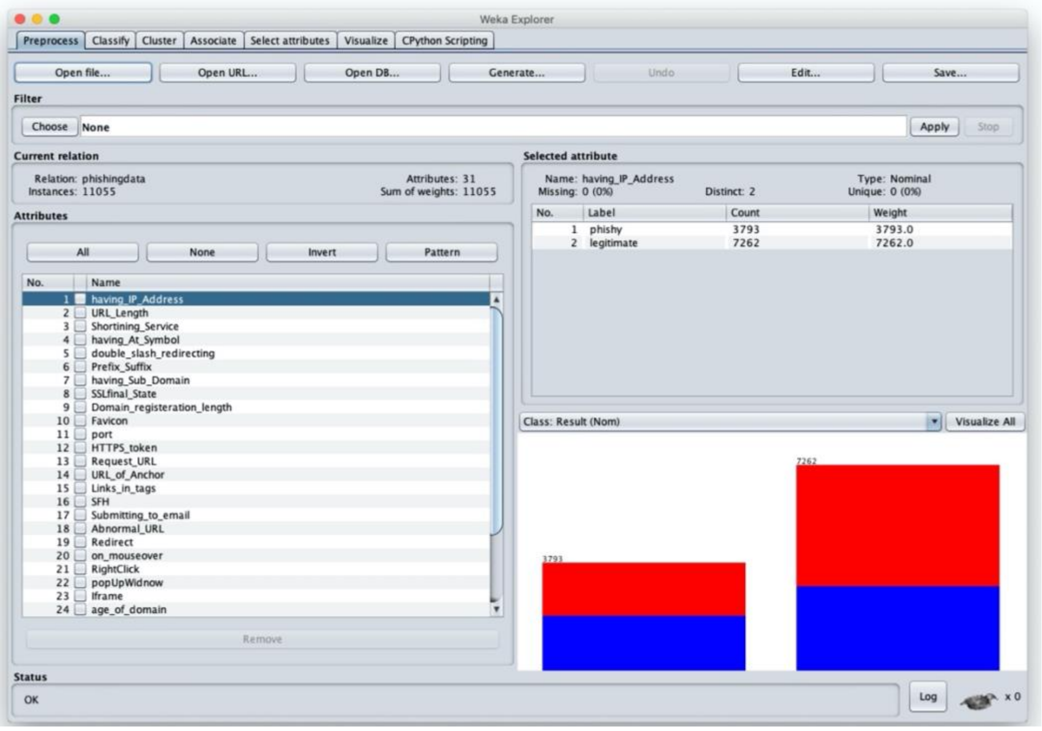
4.1 Data Gathering

We sourced for the data used for this Research from www.kaggle.com.

4.2 Data Pre-processing

Data pre-processing is a major aspect of any Artificial Intelligence implementation. Pre-processing steps involves transforming the sourced dataset into a format that can be readable and processed by our Machine Learning Tool. This is essential because raw data is always unrefined and could lead to misleading conclusions. Raw data could also have missing records hence the need for Pre-Processing. In this stage, the dataset is extracted, transformed into Comma Separated Version, standardized into readable ARFF (Attribute Relation File Format) before feature Selection is applied.

4.3 Dataset loading



Filter: None

Current relation

Relation: phishingdata
Instances: 11055
Attributes: 31
Sum of weights: 11055

Selected attribute

Name: having_IP_Address
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	phishy	3793	3793.0
2	legitimate	7262	7262.0

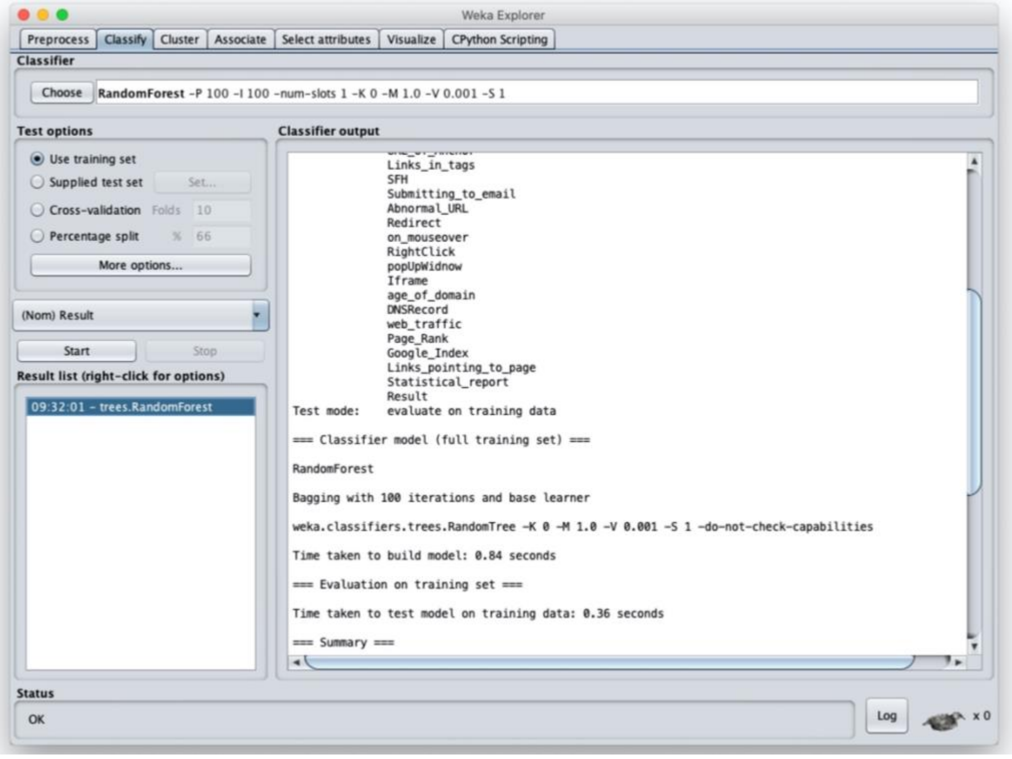
Class: Result (Nom)

Attributes

No.	Name
1	having_IP_Address
2	URL_Length
3	Shortining_Service
4	having_At_Symbol
5	double_slash_redirecting
6	Prefix_Suffix
7	having_Sub_Domain
8	SSLfinal_State
9	Domain_registeration_length
10	Favicon
11	port
12	HTTPS_token
13	Request_URL
14	URL_of_Anchor
15	Links_in_tags
16	SFH
17	Submitting_to_email
18	Abnormal_URL
19	Redirect
20	on_mouseover
21	RightClick
22	popUpWidnow
23	Iframe
24	age_of_domain

Status: OK

Fig. 2. WEKA showing the list of the file attributes. After Loading the .csv file in the WEKA application, it brings out the list of attributes and number of instances the file contains. The above image contains 31 attributes and 11055 instances.



Classifier: RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Test options

● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) Result

Result list (right-click for options)

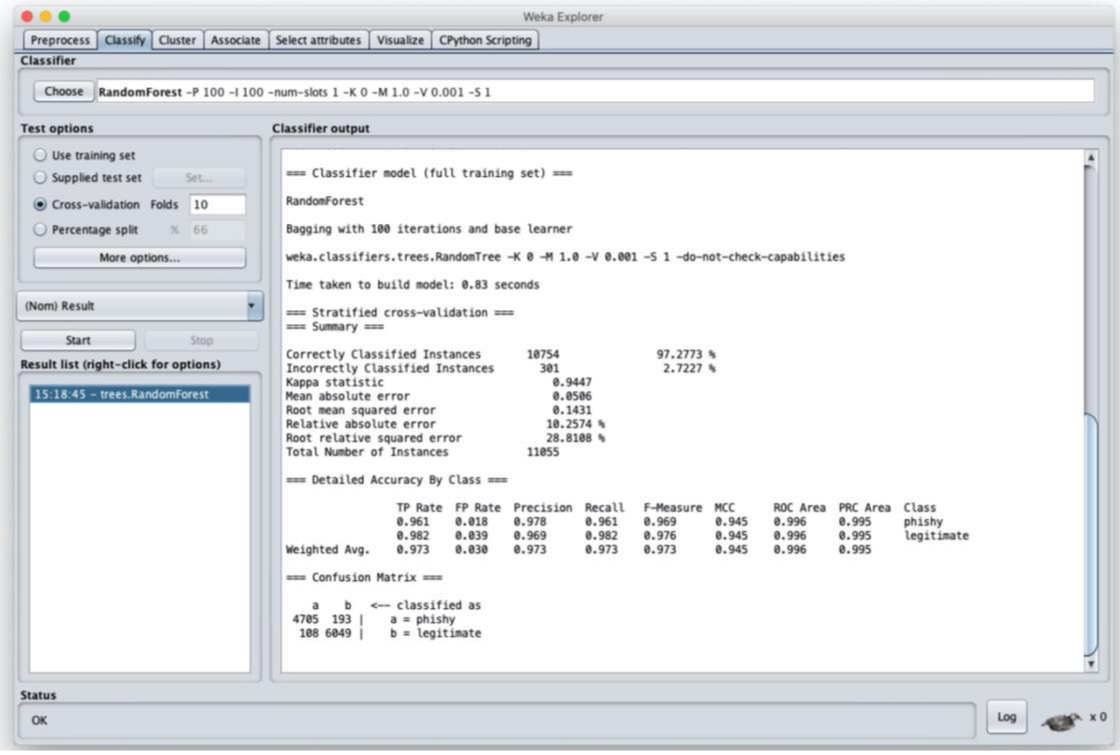
09:32:01 - trees.RandomForest

Classifier output

```
              Links_in_tags
              SFH
              Submitting_to_email
              Abnormal_URL
              Redirect
              on_mouseover
              RightClick
              popUpWidnow
              Iframe
              age_of_domain
              DNSRecord
              web_traffic
              Page_Rank
              Google_Index
              Links_pointing_to_page
              Statistical_report
              Result
Test mode:    evaluate on training data

=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.84 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.36 seconds

=== Summary ===
```

Status: OK

Fig. 3. First Weka Classification using Random Forest showing classifier output. This image above shows the test mode result evaluating on training data. It shows some part of the list of attributes and the chosen classifier, Random Forest.



```
RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
```

Cross-validation Folds 10

Percentage split % 66

(Nom) Result

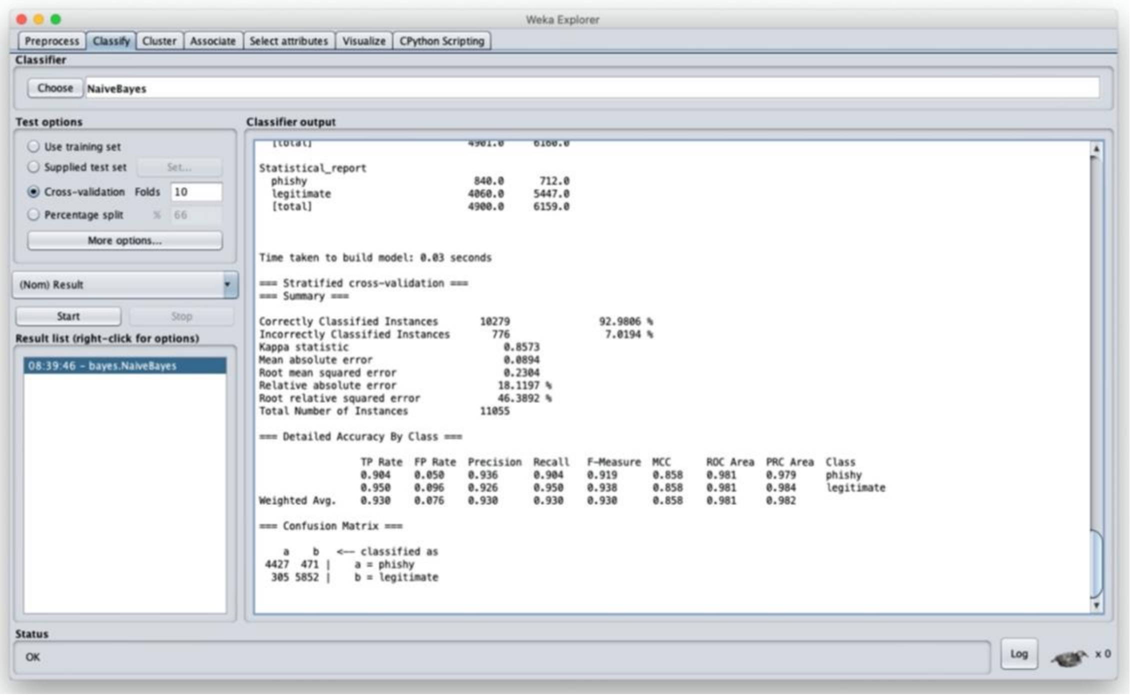
15:18:45 - trees.RandomForest

```
=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.83 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       10754               97.2773 %
Incorrectly Classified Instances       301                2.7227 %
Kappa statistic                          0.9447
Mean absolute error                      0.0506
Root mean squared error                  0.1431
Relative absolute error                 10.2574 %
Root relative squared error             28.8108 %
Total Number of Instances            11055

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.961    0.018    0.978      0.961    0.969      0.945    0.996     0.995     phishy
                 0.982    0.039    0.969      0.982    0.976      0.945    0.996     0.995     legitimate
Weighted Avg.    0.973    0.030    0.973      0.973    0.973      0.945    0.996     0.995

=== Confusion Matrix ===

    a    b   <-- classified as
 4705  193 |    a = phishy
  108 6049 |    b = legitimate
```

Status: OK

Fig. 4. WEKA Classification using Random Forest showing classifier output.

The image above shows that Random Forest classifier gave an output of 97.2% correctly classified instances. It produced a Root mean squared error of 0.1431. It produced a True Negative (TN) value of 4705 and a True Positive (TP) value of 6049.



```
NaiveBayes
```

Cross-validation Folds 10

Percentage split % 66

(Nom) Result

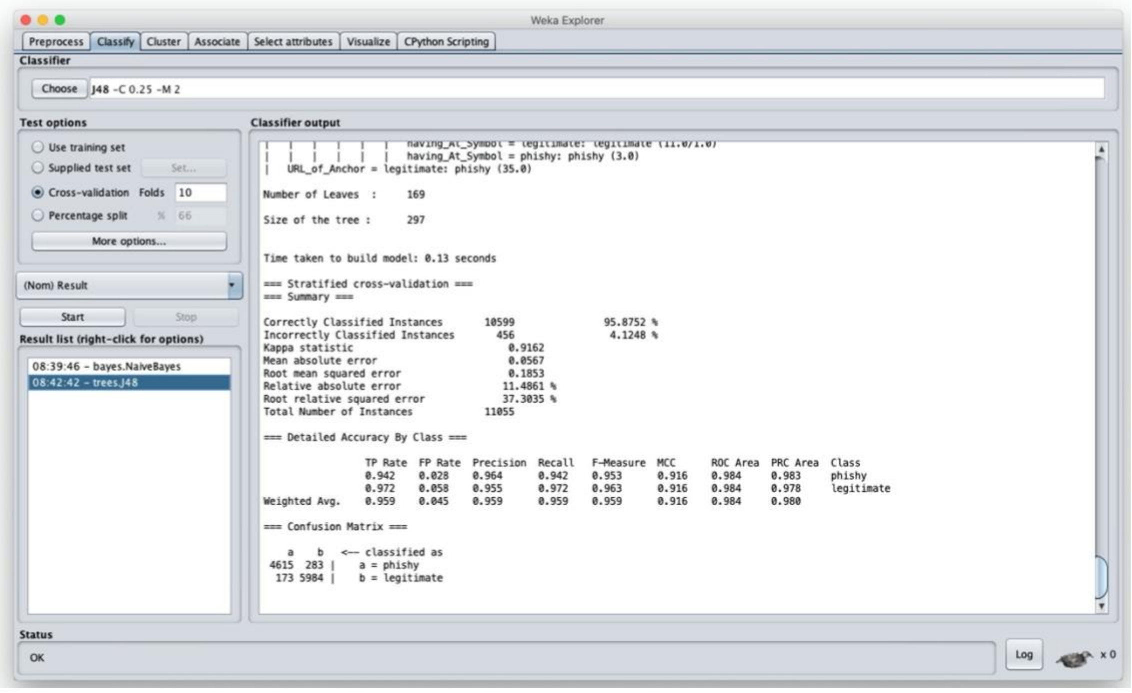
08:39:46 - bayes.NaiveBayes

```
  [total]                          4901.0     6160.0

Statistical_report
  phishy                            840.0      712.0
  legitimate                       4060.0     5447.0
  [total]                          4900.0     6159.0



Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       10279               92.9806 %
Incorrectly Classified Instances       776                7.0194 %
Kappa statistic                          0.8573
Mean absolute error                      0.0894
Root mean squared error                  0.2304
Relative absolute error                 18.1197 %
Root relative squared error             46.3892 %
Total Number of Instances            11055

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.904    0.050    0.936      0.904    0.919      0.858    0.981     0.979     phishy
                 0.950    0.096    0.926      0.950    0.938      0.858    0.981     0.984     legitimate
Weighted Avg.    0.930    0.076    0.930      0.930    0.930      0.858    0.981     0.982

=== Confusion Matrix ===

    a    b   <-- classified as
 4427  471 |    a = phishy
  305 5852 |    b = legitimate
```

Status: OK

Fig. 5. Naïve Bayes Classification showing classifier output.

The image above shows that Naïve Bayes classifier gave an output of 92.9% correctly classified instances. It produced a Root mean squared error of 0.2304. It produced a True Negative (TN) value of 4427 and a True Positive (TP) value of 5852.



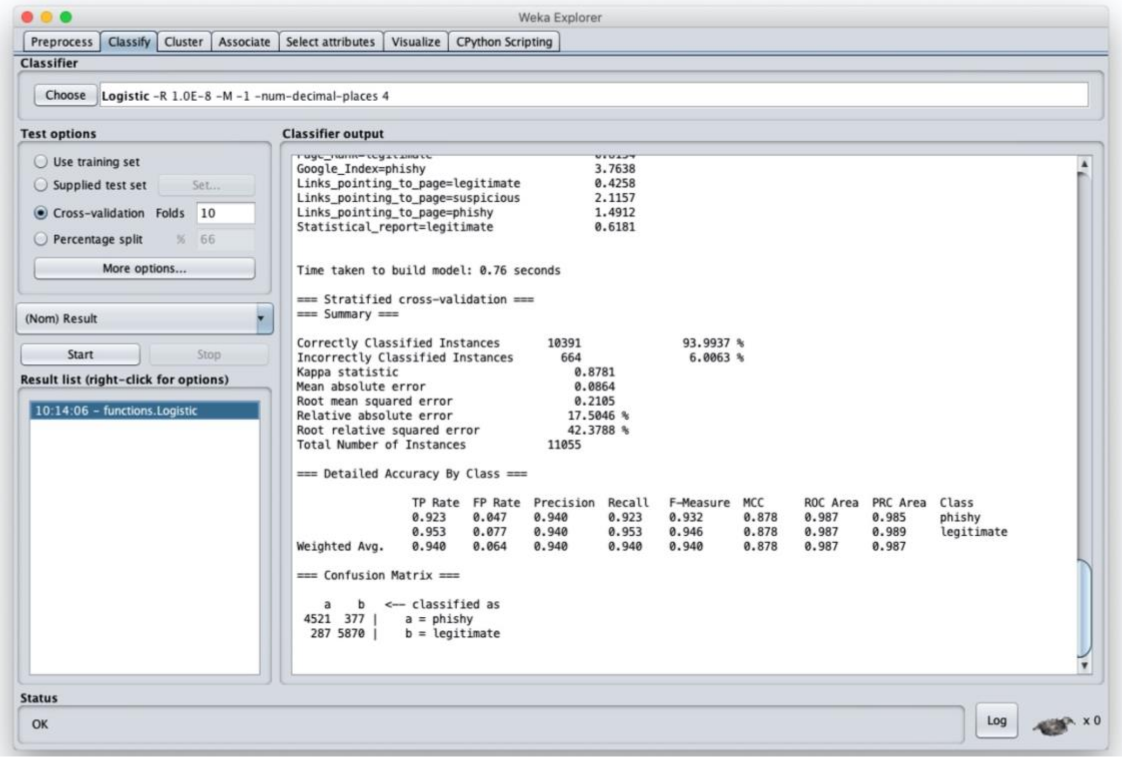
J48 -C 0.25 -M 2

● Cross-validation Folds 10

```
|   |   |   |   |   |   having_At_Symbol = legitimate: legitimate (11.0/1.0)
|   |   |   |   |   |   having_At_Symbol = phishy: phishy (3.0)
|   URL_of_Anchor = legitimate: phishy (35.0)

Number of Leaves  :     169

Size of the tree :      297


Time taken to build model: 0.13 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       10599               95.8752 %
Incorrectly Classified Instances       456                4.1248 %
Kappa statistic                          0.9162
Mean absolute error                      0.0567
Root mean squared error                  0.1853
Relative absolute error                 11.4861 %
Root relative squared error             37.3035 %
Total Number of Instances            11055

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.942    0.028    0.964      0.942    0.953      0.916    0.984     0.983     phishy
                 0.972    0.058    0.955      0.972    0.963      0.916    0.984     0.978     legitimate
Weighted Avg.    0.959    0.045    0.959      0.959    0.959      0.916    0.984     0.980

=== Confusion Matrix ===

    a    b   <-- classified as
 4615  283 |    a = phishy
  173 5984 |    b = legitimate
```

Fig. 6. J48 Classification showing classifier output.

The image above shows that J48 classifier gave an output of 95.8% correctly classified instances. It produced a Root mean squared error of 0.1853. It produced a True Negative (TN) value of 4615 and a True Positive (TP) value of 5984.



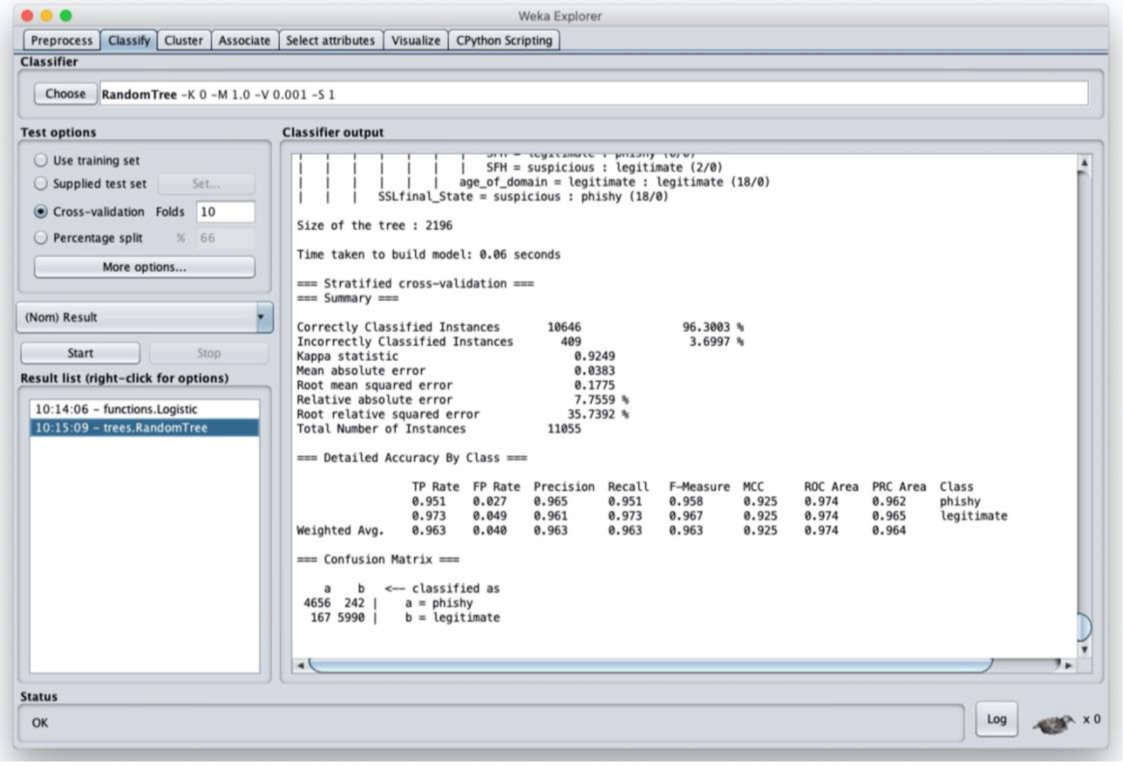
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4

● Cross-validation Folds 10

```
Google_Index=phishy                           3.7638
Links_pointing_to_page=legitimate             0.4258
Links_pointing_to_page=suspicious             2.1157
Links_pointing_to_page=phishy                 1.4912
Statistical_report=legitimate                 0.6181


Time taken to build model: 0.76 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       10391               93.9937 %
Incorrectly Classified Instances       664                6.0063 %
Kappa statistic                          0.8781
Mean absolute error                      0.0864
Root mean squared error                  0.2105
Relative absolute error                 17.5046 %
Root relative squared error             42.3788 %
Total Number of Instances            11055

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.923    0.047    0.940      0.923    0.932      0.878    0.987     0.985     phishy
                 0.953    0.077    0.940      0.953    0.946      0.878    0.987     0.989     legitimate
Weighted Avg.    0.940    0.064    0.940      0.940    0.940      0.878    0.987     0.987

=== Confusion Matrix ===

    a    b   <-- classified as
 4521  377 |    a = phishy
  287 5870 |    b = legitimate
```

Fig. 7. Logistic Regression Classification showing classifier output.

The image above shows that Logistic classifier gave an output of 93.7% correctly classified instances. It produced a Root mean squared error of 0.2105. It produced a True Negative (TN) value of 4521 and a True Positive (TP) value of 5870.



Classifier: RandomTree -K 0 -M 1.0 -V 0.001 -S 1

Cross-validation Folds 10

Percentage split % 66

Result list:
10:14:06 - functions.Logistic
10:15:09 - trees.RandomTree

```
|   |   |   |   |   |   |   SFH = suspicious : legitimate (2/0)
|   |   |   |   |   |   age_of_domain = legitimate : legitimate (18/0)
|   |   |   SSLfinal_State = suspicious : phishy (18/0)

Size of the tree : 2196

Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       10646               96.3003 %
Incorrectly Classified Instances       409                3.6997 %
Kappa statistic                          0.9249
Mean absolute error                      0.0383
Root mean squared error                  0.1775
Relative absolute error                  7.7559 %
Root relative squared error             35.7392 %
Total Number of Instances            11055

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.951    0.027    0.965      0.951    0.958      0.925    0.974     0.962     phishy
                 0.973    0.049    0.961      0.973    0.967      0.925    0.974     0.965     legitimate
Weighted Avg.    0.963    0.040    0.963      0.963    0.963      0.925    0.974     0.964

=== Confusion Matrix ===

    a    b   <-- classified as
 4656  242 |    a = phishy
  167 5990 |    b = legitimate
```

Status: OK

Fig. 8. Random Tree Classification showing classifier output.

The image above shows that Random Tree classifier gave an output of 95.8% correctly classified instances. It produced a Root mean squared error of 0.1775. It produced a True Negative (TN) value of 4565 and a True Positive (TP) value of 5990.

5. EVALUATION

The evaluation of a new technique is important for the verification of the suitability of such technique. In this section, we evaluate the performance of the proposed system. The machine learning technique was used to identify and classify Phishing Websites. Random Forest (Machine Learning) is suitable for identification and classification.

The output of the classification shows that Random Forest Algorithm is a suitable and reliable Algorithm for Identification and classification.

Table 1 - Classifier Performance – WEKA

OPTIONS	CLASSIFIER	CONFUSION MATRIX	TRUE POSITIVE	FALSE POSITIVE	PRECISION
Percentage Split-60	Random Forest	1851 94 50 2427	0.952	0.02	96.3%
	Naïve Bayes	1747 198 114 2363	0.898	0.046	87%
	J48	1810 135 84 2393	0.931	0.034	86%

Percentage Split-80	Random Forest	962 39 22 1188	0.961	0.018	96.9%
	Naïve Bayes	900 101 57 1153	0.899	0.047	89%
	J48	942 59 27 1183	0.941	0.022	87%

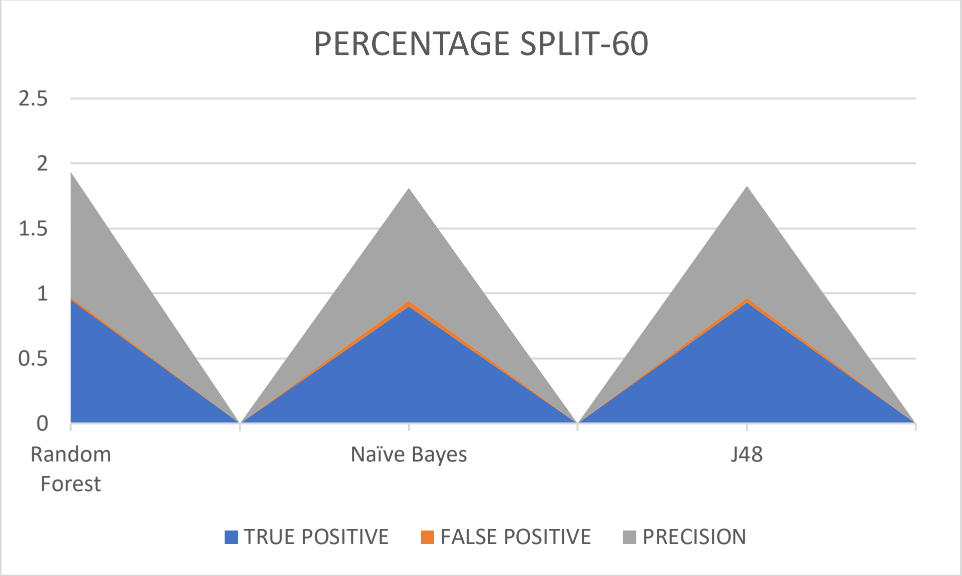


Fig. 9. Graph showing the percentage split between classifiers at 60%

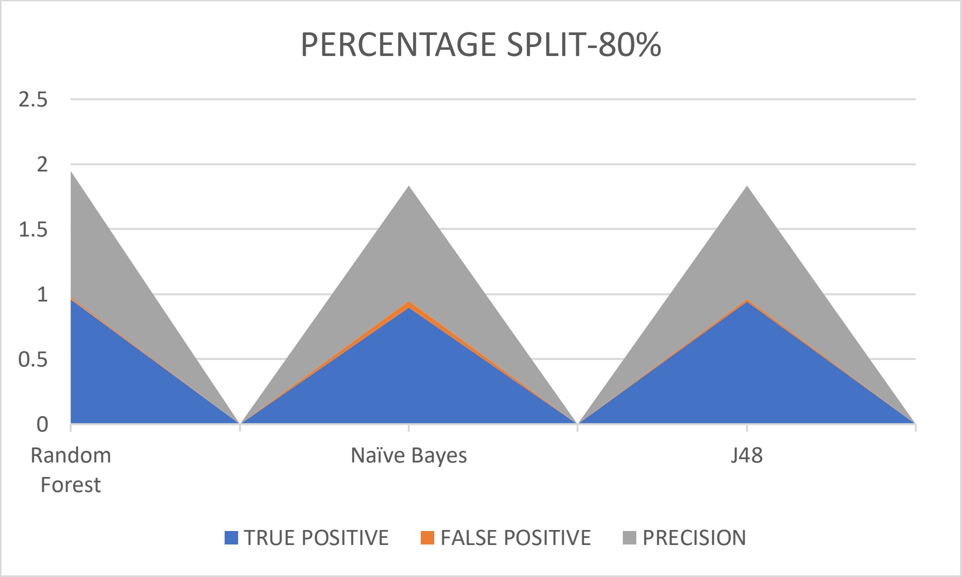


Fig. 10. Graph showing the percentage split between classifiers at 60%

Table 1 shows the result derived from the different data classification models. The Random Forest gave more classification accuracy with a precision accuracy of 96.3% on Percentage split-60 and 96.9% on Percentage Split-80. It is very glaring that in comparison with Naïve Bayes and J48, Random Forest gave the best Accuracy level. Figure 9 and 10 shows a graphical representation of the result of the experiment

6. MODEL COMPARISM

Table 1 shows a comparison of True Positive(TP) Rate, False Positive(FP) Rate, Confusion Matrix as well as precision accuracy of Naïve Bayes, J48 and Random Forest Classifiers with 60% and 90% test split. The three classifiers were used for test and each was able to decide whether it was a Phishing Website or Legitimate Website. The Random Forest Classifier showed the highest accuracy rate.

The previous research done with Naïve Bayes algorithm generates a DAG classifier model with an accuracy of 71.3 [6]. Our model was compared against the Naïve Bayes. The output showed that Random Forest fared better in terms of Prediction Accuracy than Naïve Bayes Theorem.

7. CONCLUSION AND FUTURE WORK

The reduced awareness about the phishing attacks is a major reason why most Phishing attacks succeed. Users should always be sure of the authenticity of a link before opening it. Phishing traps are not easily identified. Another issue is that more techniques are developed almost after new detection and prevention methods are discovered

In this research, we have tried to show that Random Forest Algorithm is an effective tool which can be adopted for the identification and classification of Phishing Website.

The several precious approaches were evaluated and their deficiencies were observed and analysed. We began by downloading a raw dataset from Kaggle which is an open source Online Repository. The dataset was pre-processed and refined for use by the WEKA software. WEKA was adopted for data mining because of its flexibility and ease of use. We tested and trained the dataset before applying the Machine Learning Algorithm. The Algorithm was used to detect features of Phishing which was tested against the testing dataset

In this paper, work was limited to direct application of Random Forest Algorithm to selected dataset. Also, the output will be different when applied to different Phishing Website Data. Future work can focus on capturing and detection of live Phishing Website.

REFERENCE

[1] Nandhini.S , Dr.V.Vasanthi, “Extraction of Features and Classification on Phishing Websites using Web Mining Techniques,” *INTERNATIONAL JOURNAL OF ENGINEERING DEVELOPMENT AND RESEARCH,* vol. 5, no. 4, 2017.

[2] Priya Raj , Meenakshi Mittal, “Detection of Phishing URLs using Bayes Net and Naïve Bayes and evaluating the risk assessment using Attributable Risk,” *International Journal of Computer Sciences and Engineering,* vol. 6, no. 5, 2018.

[3] Hemali Sampat, Manisha Saharkar, Ajay Pandey, Hezal Lopes, “Detection of Phishing Website Using Machine Learning,” *International Research Journal of Engineering and Technology (IRJET),* vol. 5, no. 3, 2018.

[4] Maher Aburrous, Keshav Prasad Dahal, Muhammad Akram Hossain, Fadi Thabtah, “Predicting Phishing Websites Using Classification Mining Techniques with Experimental Case Studies,” www.researchgate.net, 2010.

[5] Joby James, Sandhya L, Ciza Thomas, Detection of Phishing URLs Using Machine Learning Techniques, International Conference on Control Communication and Computing (ICCC), 2013.

[6] Navin R T , Dr Yuvaraj N, “Identification of Phishing Website using Deep Learning Algorithm,” *International Research Journal of Engineering and Technology,* vol. 6, no. 1, 2019.

[7] Mazharul Islam, Nihad Karim Chowdhury, Phishing Websites Detection Using Machine Learning Based Classification Techniques, International Conference on Advanced Information and Communication Technology, 2016.

[8] Shai Shalev-Shwartz, Shai Ben-David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014.

[9] Adele Cutler, D. Richard Cutler, John R. Stevens, Random Forests, 2011.

[10] Mohssen Mohammed, Muhammad Badruddin Khan, Eihab Bashier Mohammed Bashier, Machine Learning Algorithms and Applications, CRC Press, 2016.

[11] P. Kakarlapudi, Web Phishing Detection: Feature selection using rough sets and ant colony optimisation, 2018.

[12] Anukool Shrivastava, Ishika Duggal, Wrapper Based Phishing Detection, 2017.

[13] Y. Yang, Effective phishing detection using Machine Learning approach, 2019.

[14] Ferreira, R.P., Martiniano, A., Napolitano D., Romero M., De Oliveira Gatto D.D., Farias E.B.P, Sassi R.J, Artificial Neural Network for Websites Classification with Phishing Characteristics. Social Networking, 2018.

[15] Purvi Pujara, M. B.Chaudhari, “Phishing Website Detection using Machine Learning : A Review,” *International Journal of Scientific Research in Computer Science, Engineering and Information Technology,* vol. 3, no. 7, 2018.

[16] L. Breiman, Random Forests, Kluwer Academic Publishers, 2001.

[17] T. M. Mitchell, Machine Learning, McGraw-Hill Science/Engineering/Math, 1997.

[18] R. Kiruthiga, D. Akila, “Phishing Websites Detection Using Machine Learning,” *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 2, 2019.