National College of Ireland

Detection of Phishing URL using Ensemble Learning Techniques

MSc Internship
Cyber Security

Sharad Rajendra Parmar
Student ID: x18176381

School of Computing
National College of Ireland

Supervisor: Niall Heffernan

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Sharad Rajendra Parmar
Student ID:	x18176381
Programme:	Cyber Security
Year:	2020
Module:	MSc Internship
Supervisor:	Niall Heffernan
Submission Due Date:	17/08/2020
Project Title:	Detection of Phishing URL using Ensemble Learning Techniques
Word Count:	XXX
Page Count:	13

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.

I agree to an electronic copy of my thesis being made publicly available on NORMA the National College of Ireland's Institutional Repository for consultation.

Signature:	Sharad Rajendra Parmar
Date:	26th September 2020

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):	

Detection of Phishing URL using Ensemble Learning Techniques

Sharad Rajendra Parmar
x18176381

Abstract

Phishing is one of the prevailing means of performing cyber-attacks. Spoofed email, social media, development of clone website are the main medium used by various phishers in order to steal the private information of an individual. Uniform Resource Locators (URLs) are the main source for sharing malwares, trojans and false information. Therefore, the accurate classification between legit and phishing url is very much important. Traditional methods of detecting phishing url were mainly rely on the blacklisting and signature based methods. Both of these methods are time consuming process and can not work effectively on new set of URL. Many machine learning classifiers also have been used, to classify the URL as phishing or legit. But, with traditional machine learning approaches, low accurate results have been achieved. Therefore, in this work we propoosed the use of ensemble learning methods. Where we have used the Bagging, AdaBoost, Random Forest and Gradient boosting algorithms. Later, the results were compared with Non-ensemble learning algorithms such as Decision tree, K-nearest neighbour and Logistic regression. After Training the models we have achieved a highest accuracy of 96.15% using random forest classifier, which is an ensemble learning method.

1 Introduction

We all know that the world is emerging into a better place day by day in all terms including the evolution of the internet of things and much more processes that makes use of the internet. As its advantages and developments increases, it also paves the way for the dangerous outcomes that could follow by misusing it. For a sample of speech let's consider the online-based business sites like e-commerce sites, banking firms and much more, these sites could be under threat if hacked by people who are experts in accessing websites. Overall, the role of security plays a vital role as it is the major concern of everyone both by large scale as well as small scale business sites. Email spoofing is one of the techniques used when phishing someone, email spoofing is where you will receive a professional kind of Email from a specific site or user where you can brief into the sender's information. While you take a close look at that, the sender info would be made in the much-hidden format so that it doesn't create a frightening kind of feeling for the person in the receiving end to trust and do the further reply. They can also reach out to you via IM technique, IM is generally referred to as instant messaging where the experience of live chatting could be promoted. It creates a space for both the persons in the opposite end to have a smooth communication between them, it is little similar to text messaging but instant messaging as per the word it differs in a certain aspect possible. All these

techniques are mainly used for manipulating the people and make them believe in the authenticity to collect the information that is needed by them. It could also be stopped by creating awareness among the user and to build the security system much stronger and spam-free. Many types of phishing or engineered techniques that are used currently are as follows, Spear phishing is something where a specific individual or a specific company is targeted. They try to collect personal pieces of information from them and try to work it out with it. Once they have a firm hold on the personal data, they try to track the account linked to it or the accounts that made a connection to it, all sorts of accounts that are anyway linked with them are also hacked for their motive. Whaling is one such other way where a specific individual is targeted than the entire company. He may be targeted for his higher position or other reasons. Cat phishing or Catfishing is one of the ways where they try to be close with the person they need to do the phishing on, they seem to be a friend kind of person and without their knowledge, they collect all the information they need and use them as a puppet for their wishes. Clone phishing is so common where the target mail information is collected and a clone of some sites mail ID visited by him is created where all the malicious links are replaced with the normal links and are sent to the target claiming that this is an updated mail from them (the site's he have accessed before). Voice phishing also plays a major part where the target receives a call from the bank saying that if they have any problem please dial the number provided by them as customer care number. Both the call and the number are made by these fraudulent people to collect their data. In the customer care number, the voice is made formatted as per assisted voice of the default system voice used in banks and much more, the target believes the formatted voice to be real and gives the account number along with the pin as asked by the formatted voice. SMS phishing or smishing is so common that the mobile phone is filled with many spam messages every day, they try to send text messages to many people which has links if clicked on it could steal all your data without your knowledge, it also has a number to contact for further information if that number is made a call it directs to send a message of your data to them without your knowledge. The emergence of technology on one side serves people with numerous benefits and at the same time on the other side, the chance of it being misused is very high. Few people who have much more knowledge into dark web than the other users of sites such technocrat tends to mislead the people by performing some fraudulent activities to cheat the common people, one of the techniques or methods used in phishing attacks, they contact the user in terms of offering genuine authority which works to help them from all web-based issues, then they try to steal bank security codes, secret accessing numbers and much more. There aren't many ways to try to avoid these kinds of cheating bestowed. Akila and many other authors have used machine learning as a strategy to detect Phishing tools and mechanism. Several other algorithms like classifiers and decision tree are used in this process. The application of a machine learning method alone doesn't seem to quite well perform the task of detection accurately. Therefore, an enhanced model is required to improve the rate of accuracy. Ensemble learning methods are the kind of techniques where the multiple models are combined in order to solve a computation problem. From the previous studies for other applications, it has been observed that the results of ensemble learning methods are quite efficient as compared to the machine learning models [1]. The main aim of this research work is to detect such phishing URLs with greater accuracy using ensemble machine learning techniques. A comparative analysis is done between the ensemble learning and machine learning techniques for the detection of Phishing URLs. The algorithms like Bagging Classifiers, Random forest, Adaboost and Gradient boosting are used under the

ensemble learning protocols and for traditional machine learning process algorithms like Decision tree, KNeigbors and Logistic Regression are used. The comparison is performed using various metrics of classification such as precision, rate of accuracy, the training time, recall and F1 score.

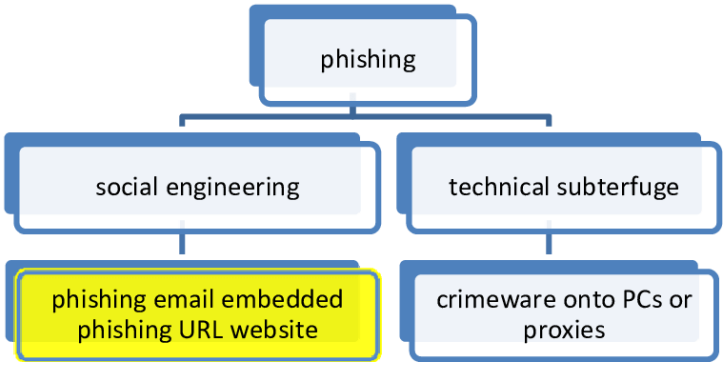

Figure 1: Types of Phishing attacks

1.1 Research question

Here are some of the research questions that this research paper will provide detailed insight.

- Does ensemble learning algorithms performs better than traditional machine learning approaches for detection of phishing URLs?
- Which algorithm accurately classifies the phishing URL and What is the highest accuracy have been achieved by the algorithm ?

The ultimate goal of this research paper is to brief more about the process that is involved in detecting the presence of Phishing attacks using a comparative model of machine learning and ensemble learning methodologies.

2 Literature Review

This section of the research paper consists of the descriptions of the past works existing on the same particular domain. It shows the gist of others work using the sampling technique to increase the effectiveness and efficiency and also in producing an effective model. Fadi in this research paper compares huge numbers of the dataset that are used for various Phishing techniques. He took for his research various kinds of ML methods that is the chief technique of phishing. The analysis is also done on different kind of metrics. After the considerate comparison, the advantage and the negative aspects are revealed and analysed. It is performed to establish an effective working and the performance of the machine learning models when analysed for all phishing attacks. The outcomes of such an analysis showing the potential for the arise of an enhanced model for the detection of the phishing mechanism [2]. Baykara in his research model keep forth a theory in which he proposes a new kind of model that doesn't allow any of the Phishing mechanism to take

place called as an Anti-Phishing Simulator. This model or the simulator gives detailed information about the presence of al kind of Phishing traps and the complete process on how to detect and identify them. It also focuses more on the process of the detection process of Phishing emails. A database model is taken as an input for the detection process which consists of all the samples of spam email alerts using the process of the Bayes algorithm. To change the original URL that is mostly displayed on the Brower's top bar the JavaScript is the thing that is mostly used by the attackers who try to Phish out the information. The method of identification involves the process of making and carrying out the complex kind of word processing where the keywords that are gotten from the sample emails and texts of Phishing mechanism. That is why the Anti Phishing simulator was developed. It checks all the contents that are received and run a total checking analysis to determine whether they contain any particulars elements regarding Phishing [3].

Peng in his research methodology keeps forth the idea of a new and enhanced technique for the identification of all kinds of phishing attacks. Here he proposes a developed model called as a SEAHound. This method searches all the documents line by line to analyse and verify the presence of any phishing technique. If any of the tools found to be present it returns a notification to alert the user of such engineered attacks. The main aim of the methodology is to focus on the nature of the language that is used in the terms which have the phishing attack-oriented to it. To detect and identify the presence of the malicious threat of phishing tools a detailed process called semantic analysis is used [4]. The working of the process called semantic analysis is done in a way that it performs the technique on each line of all the files which contains the text that is sent from the attacker to check the authenticity of the text that is sent. Li has come forward with a new mechanism for the identification of Phishing attacks. He formulated a model in this research paper called a PhishBox. This PhishBox works in a very simple way by collecting all the information or data sent by the attacker. Then determines the legitimacy of the content that is being sent or received. The detection of the presence of Phishing tools is done using this same method. It collects all the information about various kind of phishing tools and websites as the valuable input data for the process of detection and validation of the samples that are taken for the analysis [5]. It is a tool that keeps an eye out for such phishing tools and if found or detected any it gives out an alert and marks the email or the website as an illegal product that is used for the phishing mechanism. It also performs the analysis and frames the outcome in real-time. In another research paper Moradpoor proposes another model for the identification of Phishing tools presence using machine learning techniques. He keeps forth the concept of using neural networks that are used for the identification and detection of the presence of any threats regarding the Phishing emails. It uses real-time spam samples as an input in the detection process. For that specific usage a dataset, called as SpamAssassin is used [6]. That dataset consists of all the essential data that is needed for carrying out the analysis and the detection process of phishing emails and the tools used for it. All the sample spam emails are collected in a separate grouping known as the Phishcorpus dataset which is another essential part of the analysis. Certain features of the analysis are measured using the factors of the results. Such factors that are taken for consideration are rate of accuracy, true positive percentage, false positive percentile, some manual errors and the performance of the network. All these analyses are done with the help of MATLAB tool and Python programming. Radha in their research methodology proposes a new way in the detection process of phishing tools detection. He keeps forth a long list of various kinds of Anti-phishing techniques that are currently in the

use and these techniques come under two extensive types and they are known as blacklist and the white list. The blacklist that is taken for his research model consists of all types of Phishing models and techniques which are mostly associated with spam websites. On the other hand, the white list which is taken into consideration contains the details of all other authentic websites where no phishing mechanism or tools are found [7]. Then these two kinds are listed are taken for the analysis that involves the detection process. In many of the techniques that are based on the model of heuristics for the determination of anti-phishing models, the true nature and the behaviour of the websites are first displayed according to the process. Then with all the procured information, the machine learning techniques are applied for the detection of the presence of phishing tools in the emails. Aydin in his research paper tells another sustainable way of a model that is used in the process of detecting malicious phishing attacks. This paper formulates a model for the identification process by collecting all the samples from the available websites. Then by analysing all the URLs and its subsets the process is carried on. The process of feature selection is done to analyse and estimate this methodology [8]. It also uses some other methods like feature extraction and the selecting methods along with the protocols of machine learning algorithms to detect the presence of phishing tools. The URLs that are obtained from the websites and the spam folders are then categorised into some of the groups and then implemented to be used in the process. The main features are analysed using some of the main processes. Alphanumeric character analysis, Keyword analysis, domain and rank based analysis are some of the common techniques that are used for examining the factors. Security analysis is also additionally performed. Many of the obtained links of the URLs are based on the properties of the originally procured websites itself. The third-party services are also used in this process.

Another research work by Marchel gives us a detailed idea on an improved version of making the detection mechanism to find out the presence of phishing techniques in the emails and so on. He gives us the term PhishStorm, which idea is formulated and designed by his team to automatically detect the phishing tools and identify them in real-time. To carry out such a technique the sample URLs are collected from many phishing websites as the input. This improves the accuracy in the detection process more than other models. The model of PhishStorm is designed it such a way that it works effective in analysing the real time datasets. It also finds its way in accurate detection and estimation of phishing websites. It also has an inbuilt rating system to rate the sites as an illegal or black one so that the future potential users can be aware of the duplicity of that particular website. This model also alerts the user if detects or identifies any malicious phishing links or emails [9]. The model of PhishStorm also has an enhanced feature of providing the Phishing score for every URL that is tested. The score ranges from least threat to major threat mechanism, indicating the level of threat content in it. It also acts as an alternative as a website reputation rating system. Research work carried out by Almomani and team has focused on building a superior model for the detection of Phishing threats that are surfing around the web through the passing of emails. These Phishing sites which generate the URL on the emails have a very short lifetime. They tend to expire very quickly. Many of the features are used in the process of transferring and sending emails that have the phishing tool. The name that is displayed on the domain is mostly termed to be known as the message transfer agent (MGA). This MGA is the sole entity that is responsible for the sending and receiving messages from the end to end system. This technique uses the SMTP protocol and sends the message to any other system. Then those messages are reviewed by the MDA who is known as the

message delivery agents. The main work of these agents is to allocate the message to the mailbox of various peoples. Some programs like MUA are also involved in these kinds of processes [10]. Their responsibility is to check and access the working status of the program and to read and deliver the mentioned information. Various methods of phishing mechanism are mentioned and classified in this research. The filtering mechanism is done with the help of automatic classification method and they are dividing the phishing emails from the legitimate ones. Ahmed in his research paper has listed out the new techniques in the classification process of filtering the Phishing emails. It can group them to use or to classify them separately. There is also a filter which serves as the main component in the process of detection. Such a filter helps in analysing the collections that are taken as the sample of phishing data. This machine learning-based filter also labels the gotten data when it is queued for the training process. Some of the main and important parameters are measured and considered to be taken for the analysis. Those parameters are the rate of positive and negative detection rates that are performed. The true positive, False positive, true negative and false negative etc. These features are classified as a typical matter in the detecting of phishing tools present in the email. Some of the basic features are also analysed. They are the featuring structures, link and element features, spam features and the word list [11]. Some of the secondary protocols are also observed in this process and they are the Dynamic Markov chain feature and latent topic model features.

Jain in this research that is based on using the machine learning algorithms for the detection mechanism of phishing emails presents all the extracts that are completely done on the client-side of this process. This anti-phishing approach proves to be a useful and effective model. It also sees that this model enhances the privacy of the user without letting out any valuable information. This research model processes and analyses all the basic features and the factors that are associated with spam and phishing websites. This model also explains the user the interconnected relationship between the webpage, URL and the malicious email received by them. The usage of algorithms like pattern matching is done to relate the page which displays the domain name with the elements that are present on the website. All the working propaganda are based on the URL that is linked to the web page. Some of the main five categories are taken for the sample working and analysis. They are termed as fake web usage, fake login format, Forgery that is taking place in the URL formation, copying the CSS format and the information taken from the hyperlink. These are some of the listed factors that are taken for the analysis [12]. The dataset used for the process of detecting the phishing tools and mechanism are mainly procured from two main sources. The sources are called as the Openphish and the Phishtank. The URLs that are verified of its authenticity are found here. For sample analysis, the dataset for phishing detection consists of 2400 spam and 2000 authentic websites. James in this proposal shares where the users can identify the information of the phishing samples and tools. He refers to the word called as web links. They are the primary means by which these phishing attacks take place according to his research paper. The main purpose of the paper was to group and classify the existing models that are used to determine whether the URL which is hosted by the website is lexical or host-based. The methods for tested the algorithms are MATLAB and WEKA. The techniques of machine learning that are widely used in the analysis is Naïve Bayes, decision tree and vector machine models [13]. The factors which were measure for the output classification are false and true positive rate, the rate of accuracy and matrix etc. In another paper of selection, the features using the machine learning methods for the detection of Phishing websites the author Abbas documents the clustering problem

of the models using the machine learning algorithms. The techniques that were used are support vector machines, random forest, decision tree and KNN. Other methods like spatial clustering and noise scan are also done with the other methods. The detection model involving machine learning protocols are mostly list based different from normal methods [14].

3 Methodology

Phishing is a kind of fraud, where the attacker attempts to get the sensitive information of user such as login credential or account information. The Uniform Resource Locator (URL) is the main source of spreading the phishing attacks. Phisher has the full control over the sub-domains of the URL. As the URL can have the file components and paths, that can be changed by the phisher. Various antivirus companies struggles to solve such issues. In this work we are proposing ensemble learning methods in order to accurately classify the phishing URL. The proposed framework is shown in the Figure 2.

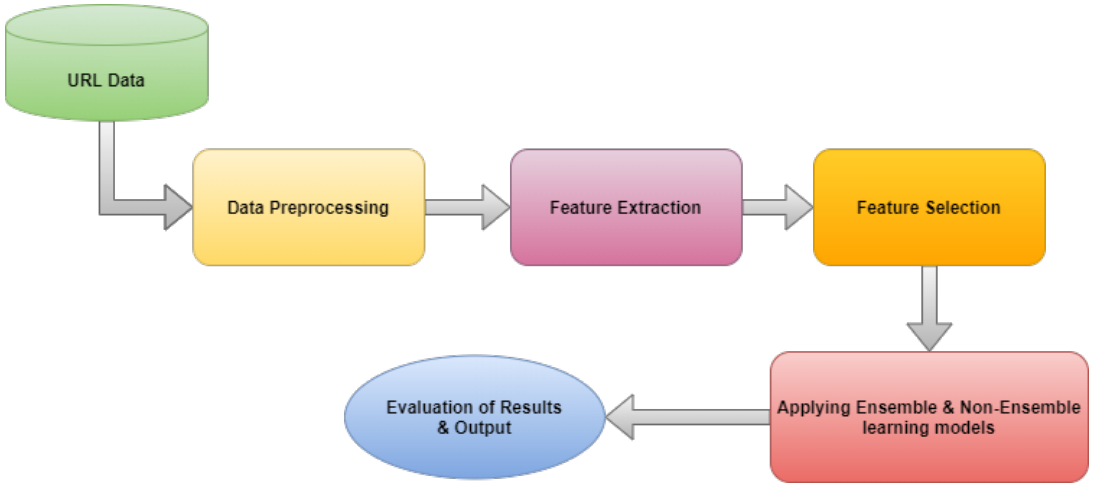

Figure 2: Proposed Framework for detection of Phishing URL

3.1 Data Collection & Dataset Description

The phishing URL data has been collected from [15]. The dataset contains more than 11,000 URLs. The information for each URL is described with the help of 32 features. All the 32 features can be categorized into the 4 types. Address Bar based Features, Abnormal based features, HTML and JavaScript based features and Domain based features. The target feature values are phishing and legitimate. Most of the features in the dataset are either categorical or binary in nature. The frequency count of each feature we have extracted in order to analyze the dataset and perform predictive analysis.

3.2 Data Pre-processing

Data pre-processing includes various operation such as data cleaning, data generalization and data sampling. The dataset contains some null values, we are removing such samples from the dataset. Using min-max scaler we are standardizing the dataset. Some of the data has been transformed into the machine readable format for analysis.

3.3 Feature Extraction & Feature Selection

Every feature in the dataset plays an important role for predictive analysis. There are 32 features which we will use for our analysis. All the features in the dataset are found to be important, therefore no feature selection methods are required for this analysis. The target feature of the dataset contains mainly -1 and +1 values.The -1 value indicates for phishing URL. Whereas, +1 value represents the legit URL. After this we will feed the data into the machine learning and ensemble learning methods.

3.4 Model Training

Ensemble and non-ensemble learning methods has been used in order to perform a comparative analysis between these two methods. The algorithms used for ensemble learning methods are Bagging Classifier, Random forest classifier, Adaboost classifier and Gradient Boosting Classifier. Whereas, the non-ensemble machine learning methods are Decision tree, K-nearest Neighbour method and Logistic Regression algorithm.

3.5 Evaluation of Models

We will use various classification performance metrics for assessment of model over testing data. We will use the 80% of data for training. Whereas, the 20% of the data will be used for testing. The classification performance metrics such as accuracy, training time, precision, recall and f1-score will be used. The higher the score, better the performance of the model.

4 Design & Implementation

The machine learning models such as logistic regression, Decision Tree and K-nearest neighbour algorithm will be used for predictive analysis. All these algorithms are non-ensemble learning algorithms Logistic regression is the most popular algorithm used for the binary classification. It is mainly used to predict the probability of dependent categorical variable. Whereas, decision tree algorithm can be used for classification as well as for regression analysis. The decision tree still is mainly famous for classification task, therefore, it is also called as classification tree. K-nearest neighbour classifies the new cases based on the majority vote by its neighbour. It calculates the hamming distance to perform the classification. The ensemble learning techniques such as Bagging, AdaBoost, Random forest and gradient boosting are used for predictive analysis purposes. Bagging classifier build the multiple models by choosing random training subset and aggregates the different learners to build an optimal learner. Whereas, AdaBoost classifier is mainly used to convert the weak classifier into strong one, it can combined with any of the classifier to boost it's classifier performance. Random forest is the supervised algorithm mainly used for classification. It combines the properties of decision tree and bagging algorithm. The proposed methods and algorithm has been implemented in a single machine. Therefore, no additional hardware is required to implement the project. To get insight about the dataset pandas framework has been utilized. In order to provide the visualization the seaborn and matplotlib libraries have been used. The sklearn libraries has been utilized to implement all the ensemble and non-ensemble based machine learning algorithms. While implementing 80% of the data will be utilized for training purposes and remaining

20% of the data will be utilized for testing purposes. Programming language is used as python and jupyter notebook is utilized for real time visualization. The above described models have been implemented with the following the configuration and specification :

- RAM : 4GB
- CPU : 2 cores @2.20 GHz
- Hard disk : 20GB
- Operating system : Windows
- Programming Language : Python
- User Interface : Jupyter
- Python Libraries : pandas, numpy, scipy, sklearn, seaborn, matplotlib

5 Evaluation

We will use 3 different cases for evaluation of the different algorithms. In the first case, we will evaluate the accuracy for both the ensemble and non-ensemble learning methods. Whereas, in the second case we will evaluate both kind of methods over precision, recall and f1-score. In the final case, the training time of different algorithms will be compared. In all the cases the training and validation (testing)set are divided in the ratio of 80:20.

5.1 Experiment 1/ Accuracy Comparison

The accuracy describes about the overall performance of the model. We have calculated the accuracy for both ensemble and non-ensemble learning models. By analysing the overall graph we have found that accuracy obtained by all the ensemble learning models such as bagging, Adaboost, random forest and Gradient boosting is high as compared to non-ensemble learning methods. In Non-ensemble learning method the highest accuracy has been achieved by logistic regression algorithm of 92.4%. In case of ensemble learning methods random forest algorithm achieved the highest accuracy of 96.15%. The K-nearest neighbour classification algorithm provided the lowest accuracy of 54.31%. The accuracy graph for every model is shown i Figure 3.

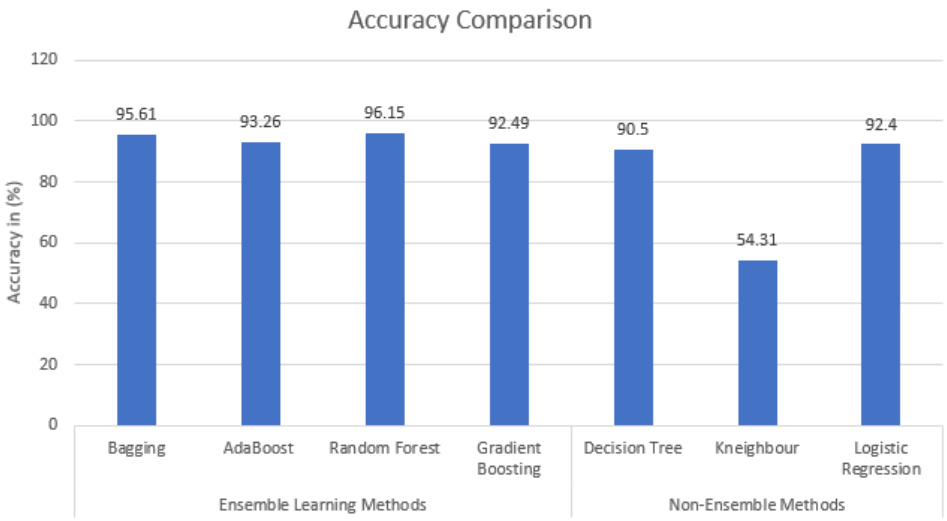

Figure 3: Accuracy Comparison

5.2 Experiment 2/ PRF Score Comparison

The precision, recall and f1-scores are mainly used to check the false-positive and false-negative rate in model performance. The high number of false-positive and false-negative values in model can reduce the f1-score to a significant rate. The PRF value of ensemble learning methods for every model seems to be exactly same. Still, the PRF Score has been achieved using Random forest algorithm. In non-ensemble techniques the highest PRF score has been obtained using logistic regression algorithm and lowest PRF score on comparing both the methods is provided by K-nearest neighbour algorithm. Overall based on PRF score we can conclude that Ensemble learning methods for all models has high PRF score. Graphical analysis of PRF score for all models is shown in Figure 4

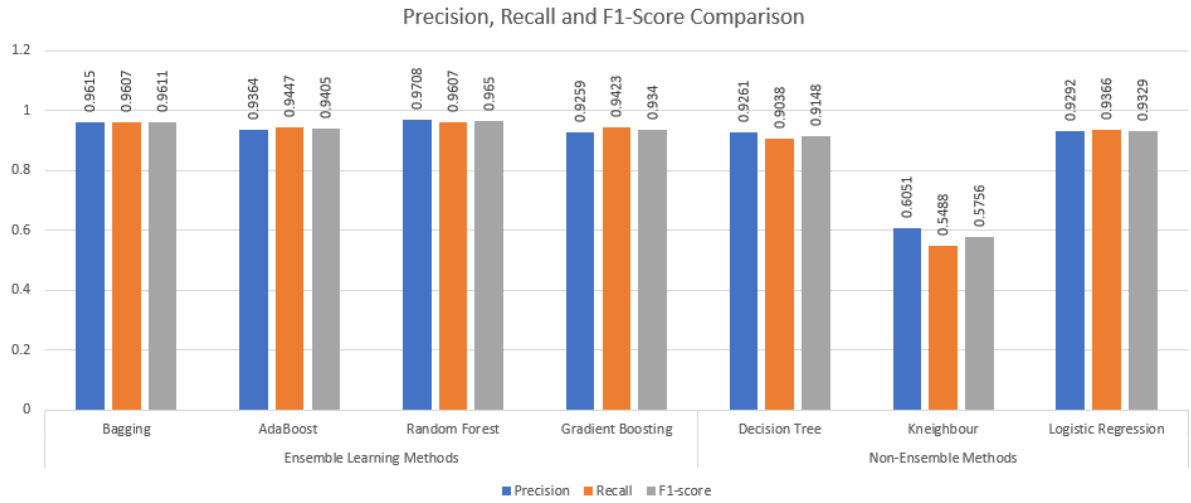

Figure 4: Precision, Recall and F1-Score Comparison

5.3 Experiment 3/ Training Time Comparison

The training time for each model also have been compared with one another. It has been observed that, the logistic regression which will provides the highest accuracy among non-ensemble methods takes the highest training time. Whereas, the decision tree classifier takes the lowest time to train the model. Further on analysing the training time graph for all the models shown in Figure 5, it can be concluded that time taken by non-ensemble techniques is less as compared to ensemble approaches. In ensemble approach, the lowest training time is taken by random forest algorithm.

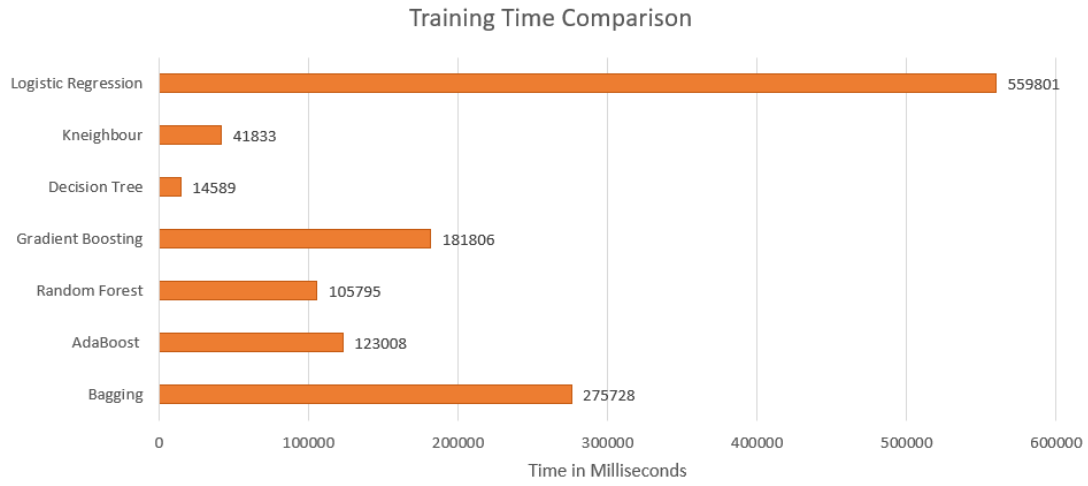

Figure 5: Training Time Comparison

5.4 Discussion

After performing a comparative analysis between multiple algorithms, we can conclude that random forest classifier, which is an ensemble learning methods outperforms as compared to all the other ensemble and non-ensemble models. The highest accuracy achieved by random forest algorithm is 96.15% and seconds highest accuracy has been achieved bagging classifier, which is again an ensemble learning approach. When comparison is performed among the non-ensemble methods logistic regression model outperforms than decision tree and K-nearest neighbour classifier. In terms of performance, random forest wins the race. Whereas, in terms of training time the random forest algorithm requires high time as compared to other models. The lowest performing model is K-nearest neighbour classifier, which provides the lowest accuracy of 54.31%.

6 Conclusion and Future Work

In order to reduce the phishing attacks or malware attacks, ensemble learning methods can be a very effective technique, As it can accurately classifies the phishing URL, with lower false positive and false negative rate. Overall, we can conclude that using ensemble learning a high performing classification output can be used. The reason behind that is ensemble learning combines the best properties of multiple models in order to solve a particular problem. This technique improves the classification to a great extent. Random

forest classifier is an combination of decision tree and bagging approach, using this model we achieved the highest accuracy of 96.15%. Although, the training time of random forest classifier is little high as compared to other classifier but it does not matter much when we are mainly concerned about the performance of the model. In the future work, more combination of different machine learning models can be explored to achieve much better results. From the previous studies on another application, we have also observed that online learning algorithms outperforms than batch processing techniques, which can be another area of research in future.

References

[1] A. D., "Phishing Websites Detection Using Machine Learning," vol. 8, pp. 111–114, Sep. 2019.

[2] N. Abdelhamid, F. Thabtah, and H. Abdel-jaber, "Phishing detection: A recent intelligent machine learning comparison based on models content and features," Jul. 2017, pp. 72–77.

[3] M. Baykara and Z. Gurel, "Detection of phishing attacks," Mar. 2018, pp. 1–5.

[4] T. Peng, I. Harris, and Y. Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning," Jan. 2018, pp. 300–301.

[5] J.-H. Li and S.-D. Wang, "PhishBox: An Approach for Phishing Validation and Detection," Nov. 2017, pp. 557–564.

[6] N. Moradpoor, B. Clavie, and W. Buchanan, "Employing Machine Learning Techniques for Detection and Classification of Phishing Emails," Jul. 2017.

[7] D. RADHA and D. VALARMATHI, "RBL Global Toolbar with Clustering Algorithm for Fake Website Detection," *International Journal of Computer Applications*, vol. 9, Nov. 2010.

[8] M. Aydin and N. Baykal, "Feature extraction and classification phishing websites based on URL," Sep. 2015, pp. 769–770.

[9] S. Marchal, J. Francois, R. State, and T. Engel, "PhishStorm: Detecting Phishing With Streaming Analytics," *IEEE Transactions on Network and Service Management*, vol. 11, pp. 458–471, Dec. 2014.

[10] D. Almomani, B. B. Gupta, S. Atawneh, A. Meulenberg, and E. Almomani, "A Survey of Phishing Email Filtering Techniques," *IEEE Communications Surveys & Tutorials*, vol. 15, pp. 2070–2090, Apr. 2013.

[11] A. Ahmed and N. Abdullah, "Real time detection of phishing websites," Oct. 2016, pp. 1–6.

[12] A. K. Jain and B. B. Gupta, "Towards detection of phishing websites on client-side using machine learning based approach," *Telecommunication Systems*, vol. 68, no. 4, pp. 687–700, Aug. 2018. [Online]. Available: https://doi.org/10.1007/s11235-017-0414-0

[13] J. James, S. L., and C. Thomas, "Detection of phishing URLs using machine learning techniques," Dec. 2013, pp. 304–309.

[14] A. Abbas, S. Singh, and M. Kau, "Detection of Phishing Websites Using Machine Learning," Jan. 2020, pp. 1307–1314.

[15] Akashkr, "Phishing url eda and modelling ," Jun 2020. [Online]. Available: https://www.kaggle.com/akashkr/phishing-url-eda-and-modelling