DETECTING WEBSITES INVOLVED IN PHISHING

A PROJECT REPORT

for

DATA MINING TECHNIQUES (ITE2006)

in

B.Tech– Information Technology and Engineering

by

SUNAINA AGARWAL (19BIT0118)

AASHI GUPTA (19BIT0121)

TEJASWI PADALA (19BIT0310)

Under the Guidance of

Dr. SENTHILKUMAR N C

AssociateProfessor, SITE

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Information Technology and Engineering

June, 2021

DECLARATION BY THE CANDIDATE

We hereby declare that the project report entitled **"DETECTING WEBSITES INVOLVED IN PHISHING"** submitted by us to Vellore Institute of Technology University, Vellore in partial fulfillment of the requirement for the award of the course **Data Mining Techniques (ITE2006)**is a record of bonafide project work carried out by us under the guidance of **Dr. Senthilkumar N C.**We further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other course.

Place : Vellore Signature

Date : 08/06/2021

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Information Technology & Engineering [SITE]

CERTIFICATE

This is to certify that the project report entitled **“DETECTING WEBSITES INVOLVED IN PHISHING”** submitted by **Sunaina Agarwal(19BIT0118)** to Vellore Institute of Technology University, Vellore in partial fulfillment of the requirement for the award of the course **Data Mining Techniques (ITE2006)** is a record of bonafide work carried out by them under my guidance.

Dr. Senthilkumar N C

GUIDE

AssociateProfessor, SITE

Detecting Websites Involved In Phishing

Abstract

Phishing is an example of a highly effective form of cybercrime that enables criminals to deceive users and steal important data. The user information is hacked by the attacker with the web-based content present in the website. As technology continues to grow, phishing techniques started to progress rapidly and this needs to be prevented by using anti-phishing mechanisms to detect phishing. This paper surveys the features used for detection and detection techniques using machine learning. Here, we explain phishing domain (or Fraudulent Domain) characteristics, the features that distinguish them from legitimate domains, why it is important to detect these domains, and how they can be detected using machine learning and natural language processing techniques. In this paper, we have used two machine learning techniques of data mining namely, Random Forest and SVM (Support Vector Machine) to compare and analyze for the detection of phishing in URLs. We have used 11055 datasets to run the tests. Experimental results show the comparative results of Random Forest and SVM and it is detected and concluded that Random Forest have an accuracy of about 89.942%.

Keywords –Phishing, legitimate domains, machine learning, Random Forest Algorithm.

I. INTRODUCTION

Phishing is a type of social engineering where an attacker sends a fraudulent ("spoofed") message designed to trick a human victim into revealing sensitive information to the attacker or to deploy malicious software on the victim's infrastructure like ransom-ware. Phishing attacks have become increasingly sophisticated and often transparently mirror the site being targeted, allowing the attacker to observe everything while the victim is navigating the site, and transverse any additional security boundaries with the victim.

Phishing is by far the most common attack performed by cyber-criminals, with the FBI's Internet Crime Complaint Centre recording over twice as many incidents of phishing than any other type of computer crime.

We have many types of phishing like-

Filter Evasion

Social Engineering

Link manipulation

We have taken link manipulationas our type of phishing. Most types of phishing use some form of technical deception designed to make a link in an email appear to belong to the organization the attackers are impersonating. Misspelled URLs or the use of sub-domains are common tricks used by phishers. In the following example URL, http://www.yourbank.example.com/, it can appear to the untrained eye as though the URL will take you to the *example* section of the *yourbank* website; actually, this URL points to the "*yourbank*" (i.e., phishing) section of the *example* website. Another common trick is to make the displayed text for a link suggest a reliable destination, when the link actually goes to the phishers' site. Many desktop email clients and web browsers will show a link's target URL in the status bar while hovering the mouse over it. This behavior, however, may in some circumstances be overridden by the phisher. Equivalent mobile apps generally do not have this preview feature.

It has now become important to safeguard online users from becoming victims of online fraud, divulging confidential information to an attacker among other effective uses of phishing as an attacker’s tool, phishing detection tools play a vital role in ensuring a secure online experience for users.

Detecting phishing websites is not easy because of the use of URL obfuscation to shorten the URL, link redirections and manipulating link in such a way that it looks trustable and the list goes on. This necessitated the need to switch from traditional programming methods to machine learning approach.

II. BACKGROUND

The preliminary concepts used for our project are:

a. Basic knowledge of various data mining: The whole project's purpose is to analyze phishing in website URLs and this is done by using data mining. A basic knowledge of the steps involved in data mining like data integration, data preprocessing, data mining, knowledge representation, etc. must be known and studied well for implementation purpose.

b. Basic knowledge of various data mining techniques: There are a number of data mining techniques like Decision Tree, Random Forest, Naïve Bayes, Regression, Support Vector Machines, etc. For a deeper understanding of the qualities of various data mining techniques, andto understand which technique works faster with which kind of data and which technique gives more accuracy as compared to others, a basic knowledge of various techniques and their performance needs to be known so as to choose efficient techniques for comparison.

   The data mining techniques compared in this paper are Random Forest and Support Vector Machines.

c. Programming knowledge: There are programming languages that support data mining efficiently like Python, R, etc. Choosing a programming language and implementing the proposed algorithm needs an understanding of these languages and applications. To carry out the analysis of website URLs, programming is needed. We have implemented the codes using Python language and its various modules like sklearn, etc. Hence, knowledge about the programming language and its various modules is a must for this project.

   The programming language used in our project is Python and it's various libraries required by our code are numpy pandas, sklearn, matplot, etc. A deep knowledge about Python syntax and its libraries is a must for this project.

d. In-depth knowledge of Random Forest classification and Support Vector Machines: These are the two data mining techniques compared in the project and without having an in-depth knowledge about these two techniques, analysis would be impossible. Hence, a deep knowledge of these two techniques is required.

Random Forest Classification uses a process in which n-number of decision trees are generated from the training data and then testing data is applied on these decision trees to conclude the accuracy of the model and its efficiency in new test data. Larger number of decision trees makes the random forest algorithm more accurate.

Support Vector Machines uses the concept of plotting the data on hyper-planes based on the number of attribute dimensions and then makes groups of the data based on their relative distances. The whole purpose is to classify the data in several groups so that for test data, a suitable group based on its position in the hyper-plane can be identified and a similar result would follow.

III. Literature Survey

A numerous phishing detection algorithm have been discussed and proposed in papers whose description is as follows:

[1]. This paper aims to use Phishing detection techniques and present the survey of various phishing website detection approaches.The various methods used are CANTINA, Associative classification data mining, SVM, Classification mining techniques, Graph mining techniques, Latent Dirichlet Allocation and Adaboost. CANTINA Accuracy is completely dependent on TF-IDF Algorithm.

[2]. The algorithms used to perceive the satisfactory function sets and assign weights. The sort and quantity of system learning or type algorithms implemented to those features and Using optimization algorithms, and the usage of logic from other anti-phishing strategies.High accuracies are observed.

[3]. Phishing is a duplicitous of sending emails illegally which pretend as it comes from the authorized organization. These emails are the ones that we're receiving day to day unknowingly. This practice leads to loot our personal data including the bank details or other sensitive data.The proposed system helps to detect and prevent the phishing mails using two techniques named as Phishzoo and MLAPT (Machine Learning Anti-Phishing System.Reasonable amount of accuracy found to exist.

[4]. The objective of this paper is to compare different classifier ensemble approaches, i.e. random forest, rotation forest, gradient boosted machine and extreme gradient boosting against single classifiers. Area under ROC curve (AUC) is employed as a performance metric, whilst statistical tests are used as baseline indicator of significance evaluation among classifiers. The experimental results are found using AUC value. This revealed that random forest was superior to other ensembles, i.e. xgboost, rotation forest and GBM and to single classifiers, i.e. C50, C-DT and CART.

[5]. This paper aims to propose and construct a large-scale standard offline dataset and unique requirements are identified by two anti phishing techniques that can be divided into list based (more dynamic and flexible method called PhishList). This method will generate multiple variations URLs based on the existing blacklist and the generated URLs will be served as a predictive blacklist. Heuristic based analyses a query website and extract some discriminative properties as the features for further processing to determine legitimacy. This paper contributes an offline dataset for rapid preliminary method testing and serving as a permanent repository for phishing device.

[6]. Prevention of Phishing of websites using Data Mining. Four different kind of classification algorithms are implemented on the phishing website data set namely Decision tree J48, J Rip rules, PART, Decision table. These algorithms were chosen because of the different strategies they use on datasets.All the algorithms taken for the study show a higher prediction rate and among all of them PART algorithm shows highest accuracy by attaining 96.76%.

[7]. In this paper they are providing the phishing technique to detect the types of websites and emails. In this without the use of internet connection we are going to detect the phishing websites. The pattern mining technique is used for detecting. Phishstorm is used to evaluate Intra-URL related processes which are extracted from words that compose the URL. Detecting spam and phishing emails using SVM and obfuscation URL detection algorithm. They areusing BURL detection algorithm which provides multilayer security. After testing on the data set it is found that Anti-phishing toolbar did a very good job at identifying 94.32%

[8]. This study says that utilizing heuristics, blacklists, visual and machine learning methods, only one single magic bullet cannot eliminate the threats of phishing efficiently. Here, ensemble machine learning techniques are promising methods to classify websites as legitimate or phishing. Along with that classification accuracy, F-measure and area under receiver operating characteristic (ROC) curves (AUC) are utilized to evaluate the performance of the machine learning methods. Experimental results are revealed that Adaboost with SVM has outperformed best among the classification methods by achieving the highest accuracy 97.61%

[9]. Luai Al Shalabi in their paper titled Comparative Study of Data Mining Classification Techniques for Detection and Prediction of Phishing Websites [18th March 2019] aimed to detect or predict the website to either legitimate or phishing class label. They compared nine different classifier techniques based on parameters like accuracy, precision, recall, sensitivity and F-Measure. According to their study, they concluded that the best classification technique was autoMLP whereas the worst was Naïve Bayes. AutoMLP achieved an accuracy of 96.70% which is quite remarkable.

[10]. This paper has a survey of recent literature and methodologies which are conducted to understand and provide an effective solution for the phishing issues. Here, phishtank databases are used where organizations and institutions who are working for cyber security report the phishing cases by some different information. Malicious URL, URL classification, Association rule mining, Classification and clustering are the algorithms used to solve the phishing issues. In conclusion, the paper said that these techniques are not much effective for preventing or accurately recognizing the URL patterns. Though some are useful but a significant amount of computational head are used as classification.

[11]. This paper aimed to perform Extreme Learning Machine based classification for 30 features including Phishing websites data in UC Irvine Machine Learning Repository Database. They used Naïve Bayes algorithm to detect phishing web-pages ELM was compared to Naïve Bayes and ANN methods for result purposes. ELM showed the highest accuracy of 89.3%. ANN showed accuracy of 87.9% and Naïve Bayes showed the least accuracy of 61.3%.

[12]. Utilizing data mining techniques to predict whether a website is phishing or not, relying on a set of factors (URL based features, HTML based features and Domain based features). The proposed system is called "Phishield" which is a merge of "Phish" and “Shield”. Random Forest, Decision Table, C4.5 (J48), SMO and Bayes Net classifiers have been applied on the training dataset and their performance has been evaluated based on their results in predicting the websites status. The results show system effectiveness in predicting phishing websites with 97% as prediction accuracy.

[13].A Data Mining approach to deal with phishing classification URL.The proposed data mining application contributes on the information security. URL classification problem is taken in consideration.The association rule mining-based technique is proposed for solving the URL classification. The a-priory algorithm is implemented for generation and classification of phishing URL. Usage of FP – tree algorithm develops the association rules with less resource requirements. For FP-Tree based technique the accuracy is 97.8% and for Apriori Based Technique, the accuracy is 83.2%

[14]. The main aim of the proposed study is to explore the domain of web security more specifically phishing detection and their identification approaches.The encoding is performed for employing the rule mining algorithm, therefore for mining rules two popular algorithms are implemented in this phase namely a priori and FP-tree. The proposed data model is based on the concept of rule mining and rule-based classification technique.The proposed model can distinguish phishing pages in web keeping money with exactness of 99.14% genuine positive and just 0.86% false negative caution.

[15]. This paper aimed to introduce a system to detect phishing websites on Android phones. Their major factors to detect phishing were URL, HTML based features and domain-based features. Their dataset included 496 instances of training dataset and used various classification algorithms namely Random Forest, Decision Table, J48, SMO, and Bayes Net. Results showed that Random Forest algorithm was most accurate out of all the other algorithms even if it wasn't the fastest one in hand. The Random Forest showed an accuracy of 97.7823%.

[16]. There are continuous threats and attacks are carried out over E-mails for various gains. The foremost popular attack over the web is phishing mails. Phishers utilize E-mail services quite expeditiously in spite of different detection and hindrance techniques already in situ.The technique proposed in this research work to classify forged E-mails from the Genuine E-mails also examines the effectiveness of detection of common user's phishing E-mails. The proposed architectural model is tested using the Enron dataset; At first Data Pre-processing is done to eliminates unnecessary words and stop words and also to reduce the size of the data that need to be examined.This paper uses J48 classifier for classification of Fake E-mails from Genuine E-mails and the result shows that the model could able to classify with 99% accuracy.

[17]. The goal of this study is to review the types of phishing attacks and current methods used in preventing those attacks. The machine learning is widely used to prevent phishing attacks. There are several algorithms used in the machine learning method to prevent these attacks like decision tree algorithm, K-means clustering algorithm, Naïve Bayes algorithm, random forest algorithm. This paper discussed 4 features like detection of phishing attacks, which included URL-based, domain based, page based and content-based features. Based on the discussion of ML methods, it is hard to determine which method is the best one because each method has advantages and disadvantages. There is no single method that works best on every problem and can be applied on various problem domains.

[18]. The problem solved in the paper is that the phishing costs Internet users billions of dollars per year and the current solutions of anti-virus, firewall and designated software do not fully prevent the web spoofing method. The detection methods to identify fake URLs are classified in 3 criteria's Lexical based feature are textual properties of URLs itself and is easy to implement. It is better to collect a large dataset to get good result when you develop anti-phishing model.

[19]. This paper aimed at comparing two classification models namely Decision Tree and Naïve Bayes on factors like accuracy, time taken and error rate. Their result showed that Decision tree method has an accuracy of 83.58% with error rate of 16.42% and total time taken as 3.302069 seconds whereas Naïve Bayes showed an accuracy of 78.14% with error rate of 21.86% and total time taken 5.966346 seconds. All the three factors suggest that Decision Tree is a better algorithm for classification as compared to Naïve Bayes.

[20]. This paper aimed to develop a resilient model to predict phishing scam by means of classification algorithms of data mining. They chose five algorithms (Decision table, J48, JRip, OneR, and PART) and compared them on the basis of accuracy, error rate, performance and efficiency. They concluded that PART algorithm had the best results with an accuracy and precision of 97.60% and error rate as less as 0.0277.

[21]. This research work focuses on the classification and analyzing the phishing attacks. Four classification algorithms are used to classify the phishing attacks and performance are evaluated using different performance metrics. The four algorithms used are Random forest, Decision Tree, K-nearest neighbor, Support Vector machine. After performance analysis, the accuracy and misclassification rate of techniques are identified and it’s concluded that random forest algorithm is producing better result.

[22]. All the existing plug-ins send the target URL to an external web server for classification. This project aims to implement the same in browser plug-in removing the need of external web service and improving user privacy.Modern day data mining techniques and use of machine learning is used tocounter such attacks. Dynamic extension to use for easy access and user protection is enabled for user and is highly optimal. The main implementation is porting of Random Forest classifier to java script.94.784% is the accuracy.

[23]. This paper proposes a system which will detect old as well as newly generated phishing URLs that have completely no past behaviors to judge upon, using Data Mining. Based on a comparison of different techniques, the random forest classifier seems to perform better. The model will be based on classification algorithm and will be trained using a training dataset. Random Forest Algorithm can be used to train the proposed model.

[24]. This paper conducted a research on the use of cloud-based model to detect phishing in websites. The model was based on random forest classification algorithm and mainly focused on the URL of the websites to detect phishing. Her dataset consisted of 11055 records of which 4898 were phishing websites. Their main goal was to detect phishing in websites by taking user input as the URL (Uniform Resource Locator) of the website. They achieved high accuracy and claimed that the method was intelligent, flexible and effective in obtaining the desired result.

[25]. This paper proposed a new Association Classification Algorithm as an artificial automated tool to increase the accuracy level of the classification process of any malicious website. Their algorithm IAC outperformed many other prominent algorithms in terms of accuracy, precision, execution time and error rate. IAC showed an accuracy of 85.3659% with a precision of 0.858 and time model of 0.01s which is much better than other models in hand namely FCBA, ECBA, FACA and CMAR.

[26]. This paper conducted research to detect websites involved in phishing using various methods like Decision Tree Classifier, K-nearest neighbors, linear svc classifier, random forest classifier and one class SVM classifier. Their research mainly focused on finding a method with maximum accuracy among the available methods. After completing the research, they concluded that Random Forest Classifier method had the maximum accuracy of 96.87% while One Class SVM was the least accurate with an accuracy of 48.56% only. Hence, they concluded that random forest classifier was the best proposed method to achieve the goal of detecting websites involved in phishing.

[27]. The goal of the paper is to trick the user to voluntarily give his sensitive information such as login credentials. In this paper the publishers have applied knowledge discovery principles from data cleansing, integration, selection, aggregation, data mining to knowledge extraction. False positives, false negatives, mean absolute error, recall, precision and F measure. In the paper it is concluded that Neural Networks is the best classifier for phishing emails detection with overall accuracy of 99.4%. However, neural networks caused a slight but noticeable degradation in classification performance.

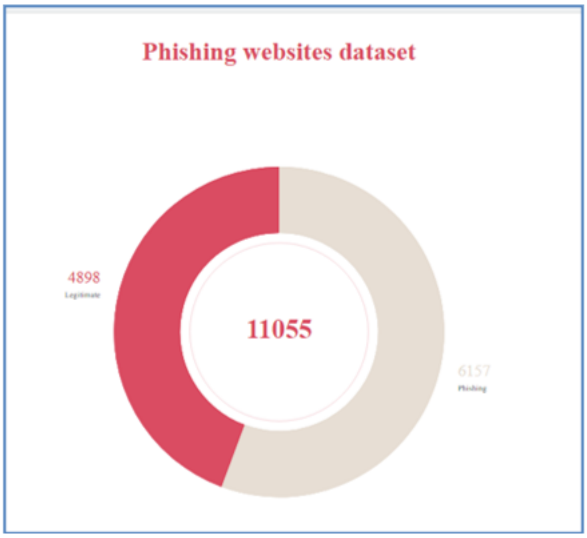
[28]. The goal of the paper is to overcome the problem when there is an attack and sabotage on sensitive data by people trying to steal and destroy information by creating fake sites similar to the original sites of the bank. The most important of this information is the IBAN number of account owner in the bank or password. Here, C4.5 algorithm is used as one of the classification techniques where used this algorithm in WEKA is a tool for analyzing data and categorizing fake websites and reducing phishing in electronic banking websites. The algorithm contains a dataset with 32 attributes, and the accuracy rate was 98.11%.

[29]. This paper aimed to study the effect of classification algorithms, feature selection, or dataset preparation methods and feature extraction. They implemented classification algorithms namely Decision Tree, Random Forest, Support Vector Machines, Logistic Regression, Lazy K-Star, Naïve Bayes, and J48,etc. They concluded that the random forest classifier's performance was better in terms of F1-score, accuracy

[30]. This paper made use of a newly arisen method named Phishing Multi-Class based on Association Rule to tackle the issue of predicting phishing websites. Among the popular algorithms, SVM outperformed all the other algorithms whereas among the AC algorithms, PMCAR outperformed all the other AC algorithms. The accuracy of SVM was 84.4% and the accuracy of PMCAR was 83.8%.

IV. DATASET DESCRIPTION & SAMPLE DATA

The collection we took has 11055 records which will comprise of boththe Legitimate andIllegitimate websites. Every tuple of these records in thedataset willpossess 30 different characteristics (also called as attributes) that a website's URL has. These characteristics will beconsidered as the independent variables for training the model.



The legitimate websites were a total of 4898 and the phishing websites were a total of 6157.

All the attributes have values as -1 or 0 or 1. The various attributes in our dataset are as follows:

1. having_IP_Address: If a URL contains IP address in its domain name, then it is considered to be related to phishing, otherwise not.
2. URL_Length: If the length of URL is below 54, then the website is considered as legitimate otherwise if the length is from 54 to 75, then it is considered as suspicious or else phishing.
3. Shortining_Service: this attribute checks whether a third-party website has been involved to shorten the long URL length. If a URL is very short, then it is considered as phishing.

4. having_At_Symbol: this attribute checks whether there is a presence of @ symbol in a URL or not. Existence of @ symbol denotes phishing website.

5. double_slash_redirecting: this attribute marks the position of two forward slash in a URL. These two forward slashes are used to redirect a page to another webpage. If the position of last // is less than 7, then it is considered as phishing otherwise legitimate.

6. Prefix_Suffix: URL names can be generated using prefix or suffix values. This behavior of a URL is marked in this attribute. Existence of '- 'character in domain name denotes phishing.

7. having_Sub_Domain: A URL domain can have a sub domain included in its URL. This sub domain's presence is checked in this attribute. If there is a single. in the domain, i.e., only two sub domains, then it is considered as legitimate. Otherwise, if the number of dots in domain are more than one then its phishing.

8. SSLfinal_State: SSL stands for Secure Sockets Layer. This attribute stores the SSL final state of a website. If the URL has https from trusted providers and certificate age is greater than a year, then it is considered legitimate else phishing.

9. Domain_registration_length: Domain registration length is the amount of time for which a particular domain name is allotted to a user. If the domain is registered for less than or equal to an year, then it is considered as phishing.

10. Favicon: Favicon is a small icon that is displayed in the tabs alongwith the title of website. Favicons are usually added to improve usability and provides an identity to a website. If the favicon is retrieved from external source, then it is phishing.

11. port: Presence of port number is recorded under this attribute. If the port number has a preferred status, then it is considered as phishing.

12. HTTPS_token: presence of HTTPS in a website URL makes the website secure for any user. Presence of HTTP in place of HTTPS may denote that the website is not secure.

13. Request_URL: Whenever a person requests for a particular service in a website, then the URL of website changes. This changes URL has a link to denote whether the website is a legitimate one or not. If the percentage of request URL is less than 22% in the complete website, then it is considered legitimate else phishing.

14. URL_of_Anchor: Anchor tags are used to add a link in a webpage. The URLs used in these anchor tags might be phishing related. If the percentage of total URLs in a webpage is less than 31%, then it is legitimate otherwise phishing.

15. Links_in_tags: Links contained in various html tags in the source code of the webpage. If the percentage of meta, link and script tags in the whole webpage is less than 17%, then it is legitimate otherwise phishing.

16. SFH: SFH stands for Server Form Handler. If there is a presence of "Is Empty" or "about:Blank" in SFH code, then it is phishing.

17. Submitting_to_email: If mail() services are used in the webpage, it might be considered as phishing.

18. Abnormal_URL: URL without a hostname is considered as abnormal. Hence it is called as phishing URL.

19. Redirect: If the page redirects are less than or equal to one, then it is considered as phishing. Else legitimate.

20. on_mouseover: If the status bar changes on mouse over, then it is probably related to phishing.

21. RightClick: If the webpage has disabled the user to right click on the page, then it is considered to be phishing URL.

22. popUpWindow: If there are browser popups with text boxes in a website, then it is considered as phishing.

23. Iframe: iframe tag is used to embed another html document in the present document. This can keep the user in illusion that he/she is accessing other webpage because its content is displayed there but in reality, it is a phishing website extracting user information.

24. age_of_domain: If the domain age is greater than 6 months, then it is considered as legitimate else phishing.

25. DNSRecord: Domain without a DNS record is considered as phishing.

26. web_traffic: If the webpage rank is less than 10,000 then it is considered legitimate. This is stated by the SEO (Search Engine Optimization) techniques. These techniques consider various factors of a webpage and provide it a specific rank.
27. Page_Rank: If the page rank is less than 0.2, then it is considered phishing.
28. Google_Index: A webpage must have a google index to be considered legitimate.
29. Links_pointing_to_page: Number of links pointing to a webpage is 0, then it's a phishing website.
30. Statistical_report: If the host is having topmost phishing IP address, then it is considered a phishing website. Otherwise, a legitimate website.

V. PROPOSED ALGORITHM WITH FLOWCHART

Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of over-fitting.

Let’s suppose the training set has m number of columns and n number of rows.

Then we create a training data which has m’ number of columns and n’ number of rows here

M’<m and n’<n.

Similarly the data in the training set is divided to many number of training data1,..training data n.

The training data then makes the decision tree.

Assumptions for Random Forest

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result. As we have divided our whole data into number of data sets and then the predictions from these values come. Now we can either use the Regression models (taking out the maximum of all the values) or the continuous values (which includes calculating mean and standard deviation of the values)

- The predictions from each tree must have very low correlations. When initially the data is there in the voting(averaging) then this will have high correlations but after calculating the regression and continuous values the low correlations are formed.

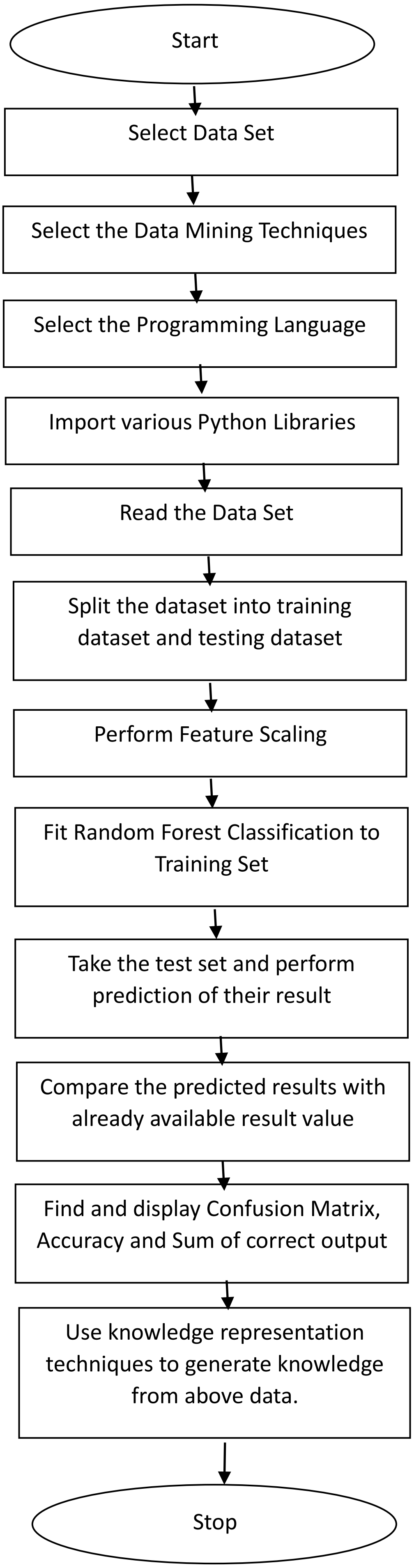
How does Random Forest algorithm work?

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:
Step-1: Select random K data points from the training set.
Step-2: Build the decision trees associated with the selected data points (Subsets).
Step-3: Choose the number N for decision trees that you want to build.
Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

Flowchart:


Start
Select Data Set
Select the Data Mining Techniques
Select the Programming Language
Import various Python Libraries
Read the Data Set
Split the dataset into training dataset and testing dataset
Perform Feature Scaling
Fit Random Forest Classification to Training Set
Take the test set and perform prediction of their result
Compare the predicted results with already available result value
Find and display Confusion Matrix, Accuracy and Sum of correct output
Use knowledge representation techniques to generate knowledge from above data.
Stop

VI. EXPERIMENTS RESULTS

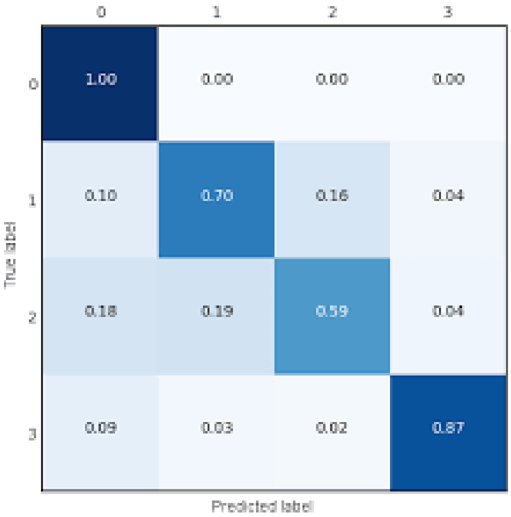
Accuracy

It is the ratio of correctly predicted target values to total number of target values.

The accuracy of Random Forest is **89.9%** and SVM is **89.0%**

Confusion Matrix

A confusion matrix is a table that is often to describe the performance of classification model (Or classifier) on a set of test data. Each row of matrix represents the instances in the actual target class while each column represents the instances in the predicted target class.



	0	1	2	3
0	1.00	0.00	0.00	0.00
1	0.10	0.70	0.16	0.04
2	0.18	0.19	0.59	0.04
3	0.09	0.03	0.02	0.87

Sum of correct Output

Sum of on diagonal elements in the confusion matrix is the sum of correct output.

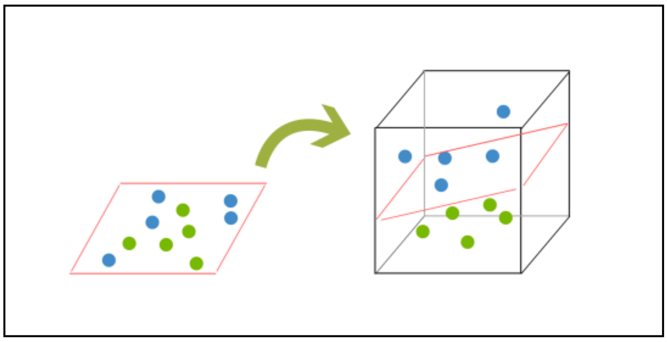
The sum of correct output is Random Forest is **2486** and SVM is **2462**.

VII. COMPARATIVE STUDY / RESULTS AND DISCUSSION

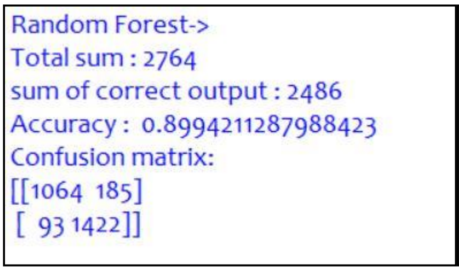
How we came to the conclusion that we will use Random Forest as our proposed algorithm?

We derived on this conclusion by reading research papers and making a comparative study of Random Forest with Support Vector Machine (SVM).

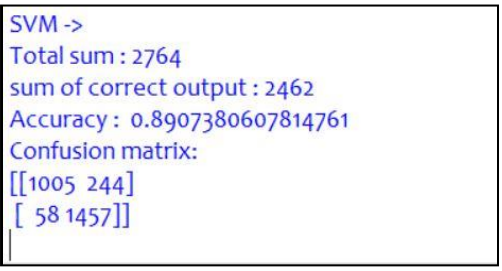
Support vector machine algorithm is apowerful yet flexible supervised machine learning algorithms. It is used both for classification and regression. The goal is to divide the datasets into classes to find a maximum marginal hyperplane (MMH) and can perform data mining on infinite dimensions.



After executing our code, we can recognize that from the dataset accuracy scores and sum of correct output values are different.



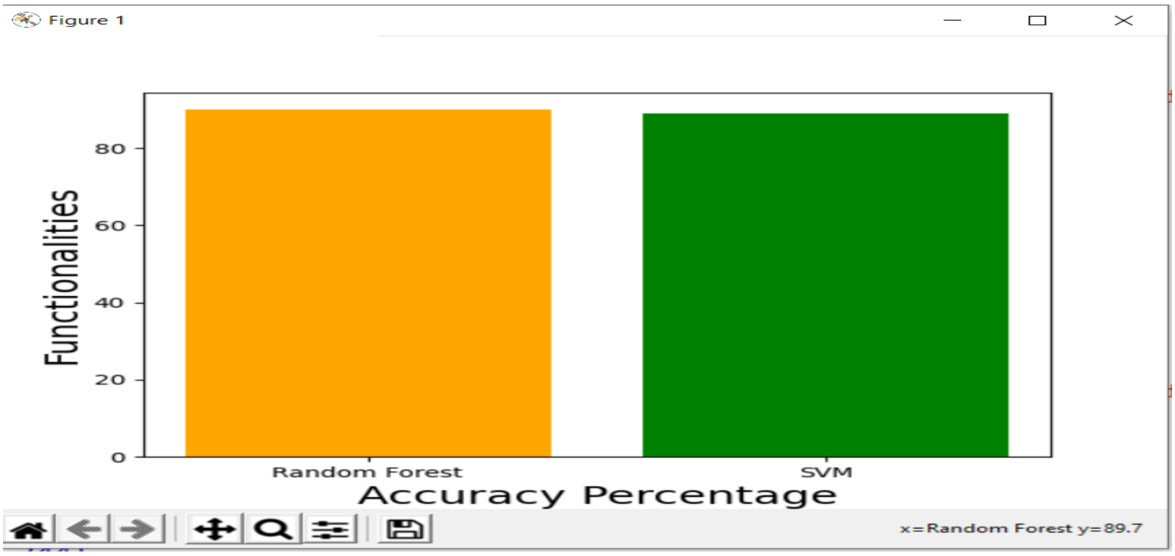
```
Random Forest->
Total sum : 2764
sum of correct output : 2486
Accuracy :  0.8994211287988423
Confusion matrix:
[[1064  185]
 [  93 1422]]
```

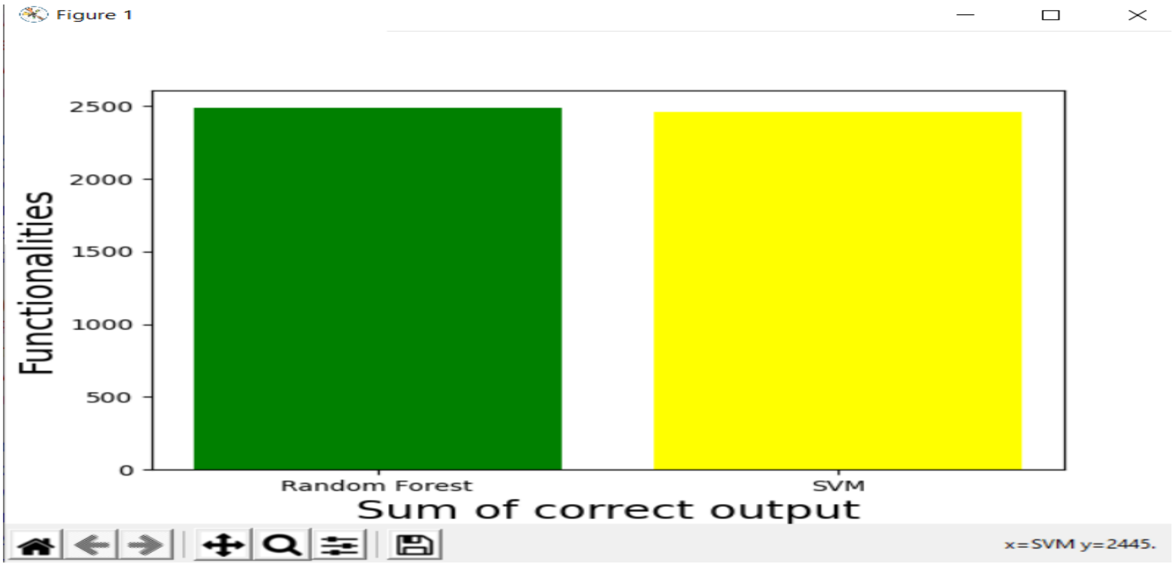


```
SVM ->
Total sum : 2764
sum of correct output : 2462
Accuracy :  0.8907380607814761
Confusion matrix:
[[1005  244]
 [  58 1457]]
```

RandomForest	SupportVectorMachines
Accuracy= 0.899 Sum of correct output= 2486	Accuracy= 0.890 Sum of correct output= 2462

The bars in the bar graph are of different colours and when we'll hover on these graphs then in the top right corner the values are displayed.





VIII. CONCLUSION AND FUTURE WORK

The purpose of the project has been successfully achieved and a high accuracy data mining technique for detecting phishing is websites has been concluded. Based on the results shown above, the team came to a conclusion that Random Forest algorithm has shown a higher accuracy and has produced a higher number of correct outputs on the test data as compared to Support Vector Machines algorithm. For better accuracy of Random Forest, we will use Grid Search CV. We will remove problems of Random Forest as it doesn't predict beyond the range in the training data, and that they may over fit data sets that are particularly noisy.

The above techniques are suggested in the main code of random forest that is attached in the appendix at the end of this document. However, to make this project available for all and usable by people who are not into programming or technical fields, we propose to work towards converting this project into a browser extension which can be easily downloaded and used by every person using a web-browser. This browser extension will include features like:

a. Detecting a phishing website and provide a warning to the user.
b. If a phishing website stays open for a long duration of time, then the browser will automatically close the tab.
c. Block the detected phishing websites in the browser for future.
d. Provide a warning notification to the user if he/she fills a form field in the phishing website.
e. Prevent the user from copying or pasting the detected phishing websites anywhere on the browser so that these URLs are not transferred to anyone else.

Such browser extensions are quite useful especially for people who do not have the idea of phishing websites. These browser extensions have advantages like:

a. Usable by both non-tech and tech domain people.
b. Does not require efforts of typing a URL and then checking. The URL is automatically checked by the browser and necessary actions are taken.
c. User friendly.

IX. REFERENCES

1. Parasmani M Rathod, Arnab Gajbhiye, Rahul Atri and Anita Sachin Mahajan, *A Survey of Phishing Website Detection and Prevention Techniques*, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 6, Issue 9, September 2017.
2. Rajkumar P, Kogilavani S.V, K. Venkatesh, *A Survey on Data Mining Techniques for Website Phishing Detection*, International Journal of Pure and Applied Mathematics, Volume 119 No. 16 2018, 2127-2133.
3. P. Meena, M. Kavitha, S. Jeyanthi and C.P. Nijitha Mahalakshmi, *Phishing Prevention Using Datamining Techniques,* International Journal of Pure and Applied Mathematics, Volume 119 No. 10 2018, 117-123.
4. Bayu Adhi Tama, Kyung-Hyune Rhee*, A Comparative Study of Phishing Websites Classification Based on Classifier Ensembles,* Journal of Multimedia Information System, January 2018
5. Kang Leng Chiew, Ee Hung Chang, Choon Lin Tan, Johari Abdullah and Kelvin Yong, *Building Standard Offline Anti-Phishing Dataset for Benchmarking*, December 2018
6. M.H.F. Fazliya and H.M.M Naleer*, A Rule Based Prediction of Phishing Websites Using Data Mining Classification Techniques,* Journal of Technology and Value Addition, Vol 1(2)
7. Sreelekha B*, Detecting Phishing Website using Pattern Mining,* 2019
8. Abdulhamit Subasi and Emir Kremic, *Comparison of Adaboost with MultiBoosting for Phishing Website Detection,* 2019
9. Luai Al-Shalabi, *Comparative Study of Data Mining Classification Techniques for Detection and Prediction of Phishing Websites*, Journal of Computer Science, March 2019
10. Rahul Patel and Ananad Rajawat, *Fight Against Phishing: A Data Mining Technique*, International Journal of Recent Scientific Research, Vol 10, May 2019
11. Sandeep Kumar Satapathy, Shruti Mishra, Pradeep Kumar Mallick, Lavanya Badiginchala, Ravali Reddy Gudur and Siri Chandana Guttha, *Classification of Features for detecting Phishing Web Sites based on Machine Learning Techniques,* International Journal of Innovative Technology and Exploring Engineering, Vol 8, June 2019

12. Karim Hashim Al-saedi, Mustafa Dhiaa Al-Hassani and Huda Yousif, *Mobile Phishing Websites Detection and Prevention Using Data Mining Techniques,* International Journal of Interactive Mobile Technologies, Vol 13, No 10 (2019)
13. Sonam Saxena, Amit Shrivastava and Vijay Birchha, *A Data mining approach to Deal with Phishing URL Classification Problem,* International Journal of Computer Applications, Vol 178 No 41, August 2019
14. Sonam Saxena, Amit Shrivastava and Vijay Birchha, *A Proposal on Phishing URL Classification for Web Security,* International Journal of Computer Applications, Vol 178 No 39, August 2019
15. Karim Hashim Al-saedi, Mustafa Dhiaa Al-Hassani and Huda Yousif, *Mobile Phishing Websites Detection and Prevention Using Data Mining Techniques,* International Journal of Interactive Mobile Technologies, Vol 13, No 10 (2019)
16. Prasanta Kumar Sahoo and Cheguri Rajitha, *Detecting Forged E-mail using Data Mining Techniques*, International Journal of Engineering and Advanced Technology, Vol 9, October 2019
17. John Arthur Jupin, Tole Sutikno, MohdArfian Ismail, MohdSaberi Mohammad, Shahreen Kasim, DerisStiawan, *Review of the machine learning methods in the classification of phishing attack*, Bulletin of Electrical Engineering and Informatics, Vol. 8, No. 4, December 2019
18. Sreelekha B, Harika B,Mrs.L.Sujihelen, *Detectingphishing website using Pattern Mining*, International Conference on Frontiers in Materials and Smart System Technologies, IOP Conf. Series: Materials Science and Engineering 590 (2019) 012024, IOP Publishing, doi:10.1088/1757-899X/590/1/012024
19. Abdulhamit Subasi, Emir Kremic, *Comparison of Adaboost with MultiBoosting for Phishing Website Detection, Procedia Computer Science*, Volume 168, 2020,
20. Abdul Raheem Fathima Shafana, Abdul Raheem Fathima ShihnasFanoon, *Predictive Data Mining for Phishing Websites: A Rule Based Approach,* Journal of Information Systems & Information Technology (JISIT) Vol. 5 No. 2 2020
21. Selvan K, *Prediction of Phishing Websites and Analysis Of Various Classification techniques,* INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 9, ISSUE 02, FEBRUARY 2020
22. M. Kanchana, Prabodhan Chavan and Arjun Johari, *Detecting Banking Phishing websites using Data Mining Classifiers*, EasyChair Preprint, No 2855

23. Mr. Aniket Kote, MrSanketKharche, Mr Pravin Aware, Mr Abhishek Pangavhane, *Detection of Phishing Websites Using Data Mining*, Volume 8, Issue 5 May 2020
24. Ashna Antony M, *Detecting Phishing Websites using Data Mining*, Volume: 07 Issue: 06 | June 2020
25. Mustafa Al-Fayoumi, Jaber Alwidian, and Mohammad Abusaif Computer Science Department, Princess Sumaya University for Technology, Jordan, 2Big Data Department, Intrasoft Middle East, Jordan, *Intelligent Association Classification Technique for Phishing Website Detection*, The International Arab Journal of Information Technology, Vol. 17, No. 4, July 2020
26. Gururaj Harinahalli Lokesh, Goutham BoreGowda, *Phishing Website Detection Based on Effective Machine Learning Approach*, Received 15 Apr 2020, Accepted 14 Aug 2020, Published online: 31 Aug 2020
27. Suhail Palaith, Mohammad Abu Qbeitah, MontherAldwairi, *PhishOut: Effective Phishing technique using selected features*, Zayed UniversityResearch Office, Research Cluster Award # R17079, October 2020.
28. Haya Alhamad, TagreedAlzyadh, Maria AltaibBadawai, *Detecting E-Banking Phishing Website using C4.5 Algorithm*, IJCSNS International Journal of Computer Science and Network Security, VOL.20 No.11, November 2020
29. Shikha Verma, Arun Kumar Gautam, *A Survey on Phishing Detection and The Importance of Feature Selection In Data Mining Classification Algorithms*, Journal of Science and Technology,ISSN: 2456-5660 Volume 5, Issue 6, Nov-December 2020
30. FAISAL ABURUB, WAEL HADI, *A New Association Classification Based Method For Detecting Phishing Websites*, Journal of Theoretical and Applied Information Technology, 15th January 2021. Vol.99. No 1

Appendix

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

# Importing the dataset
dataset = pd.read_csv('Training Dataset1.csv')
#dataset = dataset.drop('id', 1) #removing unwanted column
x = dataset.iloc[: , :-1].values
y = dataset.iloc[:, -1:].values

arr1=[]
arr2=[]
# Splitting the dataset into the Training set and Test set
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.metrics import confusion_matrix
arr = [.1,.2,.3,.4,.5, .60,.65,.70,.75,.80,.85,.90,.95]
for i in arr:
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.25, random_state = 0)
# Feature Scaling
sc = StandardScaler()
x_train = sc.fit_transform(x_train)
x_test = sc.transform(x_test)
# Fitting Random Forest Classification to the Training set
classifier = RandomForestClassifier(n_estimators = 10, criterion = 'entropy', random_state = 0)
classifier.fit(x_train, y_train)
y_pred = classifier.predict(x_test)
cm = confusion_matrix(y_test, y_pred)
```

```
a = cm[0][0]
b=cm[0][1]
c=cm[1][1]
d=cm[1][0]
print('Random Forest->')
print('Total sum :',a+b+c+d)
print('sum of correct output :', a+c)
arr2.append(a+c)
print('Accuracy : ', (a+c)/(a+b+c+d))
arr1.append(((a+c)/(a+b+c+d))*100)
print('Confusion matrix: ')
print(cm)
# Fitting SVM to the Training set
classifier=SVC(kernel = 'rbf', random_state = 0) #classifier which uses SVC from sklearn.svm
classifier.fit(x_train,y_train)
# Predicting the Test set results
y_pred = classifier.predict(x_test)
# Making the Confusion Matrix
cm1 = confusion_matrix(y_test, y_pred)
a = cm1[0][0]
b=cm1[0][1]
c=cm1[1][1]
d=cm1[1][0]
print('SVM -> ')
print('Total sum :',a+b+c+d)
print('sum of correct output :', a+c)
arr2.append(a+c)
print('Accuracy : ', (a+c)/(a+b+c+d))
arr1.append(((a+c)/(a+b+c+d))*100)
print('Confusion matrix: ')
print(cm1)
```

```
arr_values=['Random Forest','SVM']
plt.ylabel('Functionalities',fontsize=18)
plt.xlabel('Accuracy Percentage',fontsize=18)
plt.bar(arr_values,arr1,color=['orange','green'])
plt.show()

plt.ylabel('Functionalities',fontsize=18)
plt.xlabel('Sum of correct output',fontsize=18)
plt.bar(arr_values,arr2,color=['green','yellow'])
plt.show()
```