ISSN: 2582-6433

INTERNATIONAL JOURNAL FOR LEGAL RESEARCH AND ANALYSIS

Open Access, Refereed JournalMulti Disciplinary
Peer Reviewed6th Edition

VOLUME 2 ISSUE 7

www.ijlra.com

 ISSN: 2582-6433

DISCLAIMER

 ISSN: 2582-6433

EDITORIAL TEAM

EDITORS

Megha Middha

Megha Middha, Assistant Professor of Law in Mody University of Science and Technology, Lakshmangarh, Sikar

Megha Middha, is working as an Assistant Professor of Law in Mody University of Science and Technology, Lakshmangarh, Sikar (Rajasthan). She has an experience in the teaching of almost 3 years. She has completed her graduation in BBA LL.B (H) from Amity University, Rajasthan (Gold Medalist) and did her post-graduation (LL.M in Business Laws) from NLSIU, Bengaluru. Currently, she is enrolled in a Ph.D. course in the Department of Law at Mohanlal Sukhadia University, Udaipur (Rajasthan). She wishes to excel in academics and research and contribute as much as she can to society. Through her interactions with the students, she tries to inculcate a sense of deep thinking power in her students and enlighten and guide them to the fact how they can bring a change to the society

Dr. Samrat Datta

Dr. Samrat Datta Seedling School of Law and Governance, Jaipur National University, Jaipur. Dr. Samrat Datta is currently associated with Seedling School of Law and Governance, Jaipur National University, Jaipur. Dr. Datta has completed his graduation i.e., B.A.LL.B. from Law College Dehradun, Hemvati Nandan Bahuguna Garhwal University, Srinagar, Uttarakhand. He is an alumnus of KIIT University, Bhubaneswar where he pursued his post-graduation (LL.M.) in Criminal Law and subsequently completed his Ph.D. in Police Law and Information Technology from the Pacific Academy of Higher Education and Research University, Udaipur in 2020. His area of interest and research is Criminal and Police Law. Dr. Datta has a teaching experience of 7 years in various law schools across North India and has held administrative positions like Academic Coordinator, Centre Superintendent for Examinations, Deputy Controller of Examinations, Member of the Proctorial Board

 ISSN: 2582-6433

Dr. Namita Jain

Head & Associate Professor

School of Law, JECRC University, Jaipur Ph.D. (Commercial Law) LL.M., UGC -NET Post Graduation Diploma in Taxation law and Practice, Bachelor of Commerce.

Teaching Experience: 12 years, AWARDS AND RECOGNITION of Dr. Namita Jain are - ICF Global Excellence Award 2020 in the category of educationalist by I Can Foundation, India.India Women Empowerment Award in the category of "Emerging Excellence in Academics by Prime Time & Utkrisht Bharat Foundation, New Delhi.(2020). Conferred in FL Book of Top 21 Record Holders in the category of education by Fashion Lifestyle Magazine, New Delhi. (2020).Certificate of Appreciation for organizing and managing the Professional Development Training Program on IPR in Collaboration with Trade Innovations Services, Jaipur on March 14th, 2019

Mrs.S.Kalpana

Assistant professor of Law

Mrs.S.Kalpana, presently Assistant professor of Law, VelTech Rangarajan Dr. Sagunthala R & D Institute of Science and Technology, Avadi.Formerly Assistant professor of Law,Vels University in the year 2019 to 2020, Worked as Guest Faculty, Chennai Dr.Ambedkar Law College, Pudupakkam. Published one book. Published 8Articles in various reputed Law Journals. Conducted 1Moot court competition and participated in nearly 80 National and International seminars and webinars conducted on various subjects of Law. Did ML in Criminal Law and Criminal Justice Administration.10 paper presentations in various National and International seminars. Attended more than 10 FDP programs. Ph.D. in Law pursuing.

Avinash Kumar

Avinash Kumar has completed his Ph.D. in International Investment Law from the Dept. of Law & Governance, Central University of South Bihar. His research work is on "International Investment Agreement and State's right to regulate Foreign Investment." He qualified UGC-NET and has been selected for the prestigious ICSSR Doctoral Fellowship.He is an alumnus of the Faculty of Law, University of Delhi. Formerly he has been elected as Students Union President of Law Centre-1, University of Delhi.Moreover, he completed his LL.M. from the University of Delhi (2014-16), dissertation on "Cross-border Merger & Acquisition"; LL.B. from the University of Delhi (2011-14), and B.A. (Hons.) from Maharaja Agrasen College, University of Delhi. He has also obtained P.G. Diploma in IPR from the Indian Society of International Law, New Delhi.He has qualified UGC – NET examination and has been awarded ICSSR – Doctoral Fellowship. He has published six-plus articles and presented 9 plus papers in national and international seminars/conferences. He participated in several workshops on research methodology and teaching and learning.

 ISSN: 2582-6433

ABOUT US

INTERNATIONAL JOURNAL FOR LEGAL RESEARCH & ANLAYSIS ISSN 2582-6433 is an Online Journal is Monthly, Peer Review, Academic Journal, Published online, that seeks to provide an interactive platform for the publication of Short Articles, Long Articles, Book Review, Case Comments, Research Papers, Essay in the field of Law & Multidisciplinary issue. Our aim is to upgrade the level of interaction and discourse about contemporary issues of law. We are eager to become a highly cited academic publication, through quality contributions from students, academics, professionals from the industry, the bar and the bench. INTERNATIONAL JOURNAL FOR LEGAL RESEARCH & ANALYSIS ISSN 2582-6433 welcomes contributions from all legal branches, as long as the work is original, unpublished and is in consonance with the submission guidelines.

 ISSN: 2582-6433

FEATURE SELECTION ON PHISHING WEBSITES : CYBER CRIME

AUTHORED BY- 1. SAURABH RANKA,
U.G. SCHOLAR
DEPARTMENT OF LEGAL STUDIES
SANGAM UNIVERSITY, BHILWARA
2. CHETAN JAIN
DATA ENGINEER
MAVENWAVE PARTNERS, CHENNAI

Abstract

Internet has led to a drastic change in our lives. It has a wide coverage and reaches which has brought with it a significant rise in malicious attack on digital devices and software systems. Cyber-crimes are carried out in order to steal money or information, sometimes to damage the devices. Phishing attack is one of the common types of cyber-attack. Phisher uses the website which are visually and semantically similar to those real websites attempts to steal user's personal sensitive information such as user name, password, banking details etc. The objective of this project is to provide legal information about the phishing attack and implement the Feature selection techniques using KMO test and to classify the phishing websites using Support Vector Machine (SVM) algorithm in Machine learning. For classifying the phishing websites the dataset have been taken from UCI repository which has been published during the year 2016. The accuracy level because of KMO Test and SVM algorithm leads to 96%. The whole idea is developed with the help of R studio tools.

 ISSN: 2582-6433

I. Introduction

Information security is to protect the sensitive information from the social engineering attack such as phishing attack and money laundering. Social engineering attack is an art of manipulating the people who has less knowledge about these types of attack. Every organization has security issues that have been of great concern to users, site developers, and specialists, in order to defend the confidential data from this type of social engineering attack.

1.1 What are Phishing Attacks?

- Phishing attack is **contour** of social engineering.

  Phishing attack is one of **common type** of cyber-attack.
- The term _**PHISHING'** was coined around 1996 by hackers stealing Online Accounts and Passwords in America.
- Phishing was originally used to **steal of AOL password and corresponding accounts.**
- In recent days, the phisher use variety of vectors such as **Email, Trojan horse key logger and man-in-the-middle attack,** here the phisher use spoofed e-mail to deceive unsuspecting victims to disclose confidential information ;**installation of malware** such as Trojan horse key logger and spyware which may cause data compromises.

  DNS-Based attack in which phisher altered the host name and send to user with aid of fraudulent server.

  Phishing attack can be defined as **PHISHER attempts to fraudulently retrieve user's sensitive information by mimicking electronic communication from trustworthy or public organization;** Figure 1 tells about the Phishing attack.

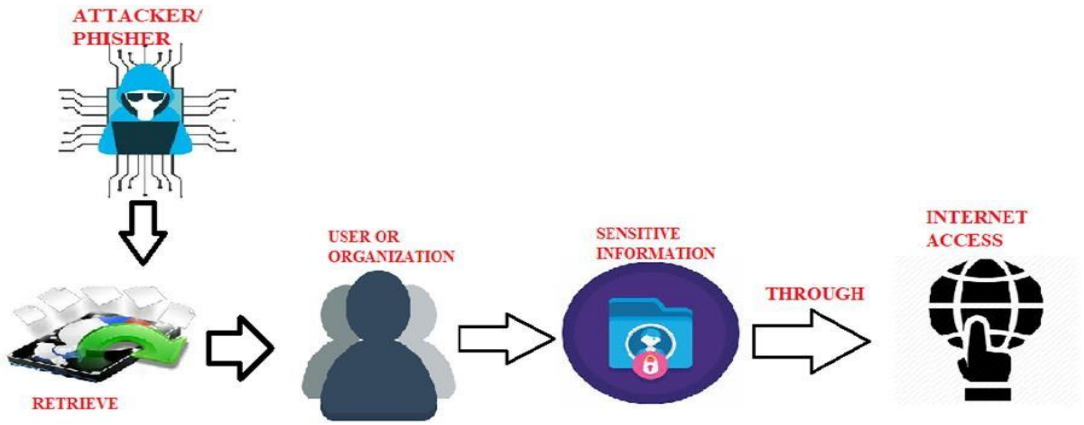


Figure1 Phishing Attack

 ISSN: 2582-6433

1.2 Motivation and Justification

According to Verizon's data breach investigation report for 2021, 85% of all breaches involved the "human element" and 36% of all security breaches were by way of "phishing" up 11% from 2020, reflecting cyber criminals, and continued focus on compromising individual end users.

1.2.1 Phishing Attacks in 2021

- Phishing is a source of 91% of cyber attacks, and costs a business, on average, $4.65 million, according to IBM's 2021 cost of data breach report.
- CISCO's 2021 data suggests that financial services firms are the most likely to be targeted by phishing attacks, having been targeted by 60% more phishing attacks than the next highest sector (which CISCO identifies as higher education).
- Tessian's 2021 research suggests workers in the following industries received a particularly large quantity of malicious emails:
    1. Retail (an average of 49 malicious emails per worker, per year)
    2. Manufacturing (31)
    3. Food and beverage (22)
    4. Research and development (16)
    5. Tech (14)

1.2.2 Sectors affected by Phishing Attacks in 2021

Graph 1 tells about the sectors which are affected by phishing attack during the year 2021, According to the report of of Statista, In 2021, online stores were the most targeted organizations by phishing attacks. Among all the organizations, online stores were targeted by 17.61 percent of all phishing attacks. Global internet portals ranked second, with 17.27 percent of the attacks.

Characteristics	**Percentage of Phishing attacks**
Other	**22.71%**
Online Stores	**17.61%**
Global Internet Portals	**17.27%**
Payment Systems	**13.11%**
Banks	**11.11%**
Social networks and Blogs	**6.34%**
IMS	**4.36%**
Telecommunication companies	**2.09%**

 ISSN: 2582-6433

IT Companies	**2%**
Financial Services	**1.9%**
Delivery companies	**1.51%**

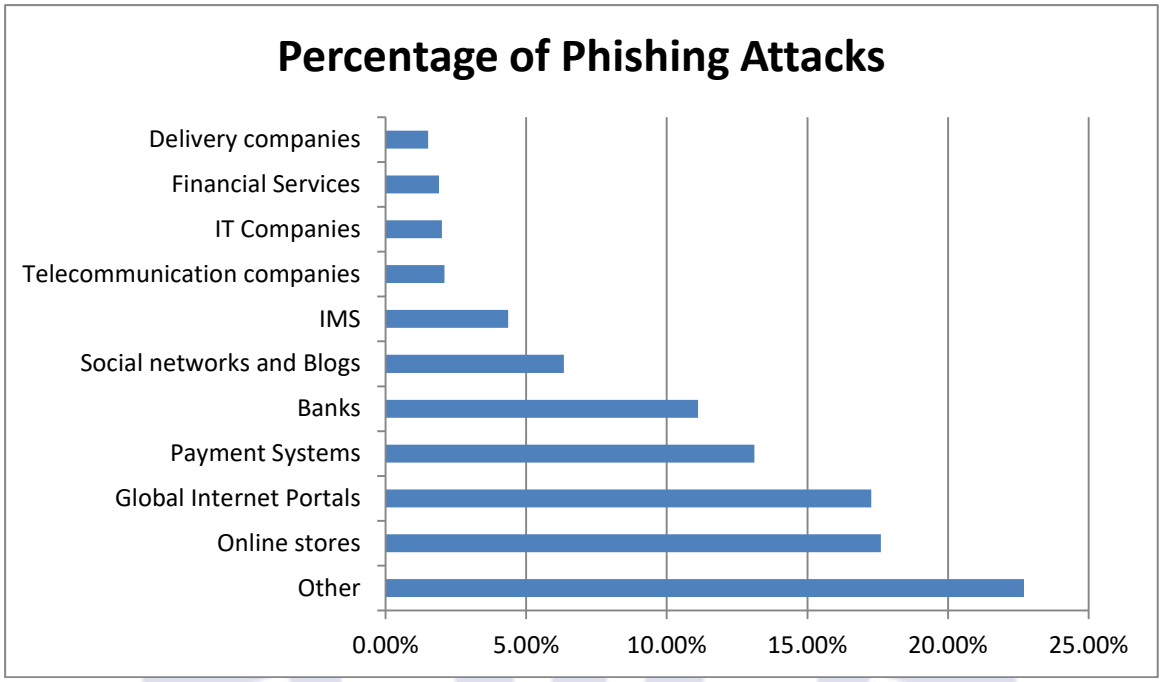


Figure 3. Sectors affected by Phishing attack

1.3 How Phishing Attack works?

Attackers create a cloned website or page in the form of phishing kit and upload it on the site.

The phisher lures its victim to his **fake login page by sending emails.**

The **victim** visits the phishing page, inputs **her/his login and password**, confidential or sensitive information have been **proceeded site.**

- Then the **site** sent the **information to the email** which is controlled by the phisher.

Phisher retrieves the **stolen credentials and hijacks the victim account;** Figure 4 displays the working of phishing attack.

 ISSN: 2582-6433

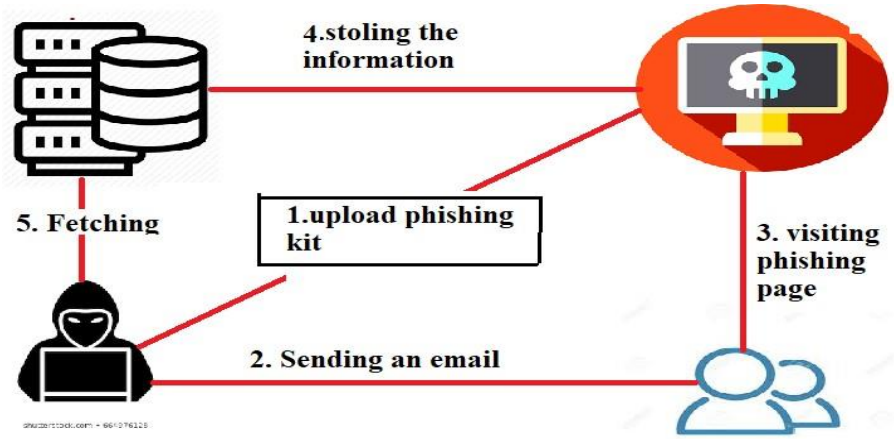


Figure 2.Working of Phishing Attacks

1.4 Execution of Phishing Attacks

Phishing attacks can be executed in the form of:

- E-mails
- ***Websites***
- Phone calls
- SMS

In this project Phishing Website have been implemented.

1.5 Phishing Detection and Prevention Tools

The phishing attack can be detected and prevented with the aid of prevention tools, they are

Anti-Phishing Plug-In [Browser Extension] Anti-Phishing Toolbar
Phi-Stank Spam Filters
Malware Software.

1.6 Handling Phishing Attacks

Before being trapped into phishing attack we can work on its avoidance. Different types of avoidance are given as follows:

Before Responding: User gets aware to respond such phishing e-mails which demand for our personal information.

Suspicious Website: If user found any suspicious website then the user can check for its authenticity. By checking its https in the beginning of URL, padlock icon in the browser any sign which makes it different form original site.

Use of Secure Browser: User must use the browser with latest security against phishing attack;

 ISSN: 2582-6433

we can use latest versions of browser with updates phishing filters.

Fantastic Offer: Don‘t believe such offers that are not easy to believe check for the all necessary details of the website and ask too many questions before sharing any personal details over the internet.

1.7 Laws governing phishing in India

Phishing is subject to the provisions of the Information and Technology Act, 2000 (('IT Act). The provisions dealing with the crime were incorporated via the 2008 amendment. The provisions that have been incorporated and regulate the crime of phishing are :

Section 43 – extracting or accessing data without consent

Section 43 stipulates that if an individual accesses another person's computer system or network for the purposes of downloading, accessing, disrupting, denying or corrupting the data contained therein, without the consent of the owner – then that person may be held liable under this provision.

Section 66 – Punishment for phishing

The provision under Section 66 of the IT Act prescribes the punishment that can be inflicted for the act of stealing a victims account by a phisher. The punishment includes either imprisonment for a term that can exceed up to three years or a fine that can exceed up to five lakh rupees, or both, depending on the severity of the crime.

Section 66A – spreading false information

The provision stipulates that the act of spreading information knowing that it is false, with the intent of causing some form of damage to the victim would be punishable. The provision additionally, outlines the offences that attract the punishment prescribed under the provision.

Section 66C

The provisions under this Section forbids the use of passwords, electronic signatures, or any other feature which is a unique identification of any person. Phishers commit fraudulent actions while disguising themselves as the legitimate owner of the account and carrying out fraudulent acts.

Section 66D – Impersonation

Cheating by impersonating another person while utilising communication devices or computer sources is covered under the provisions outlined under this section. Fraudsters commit fraud by impersonating banks and other organisations by using URLs that take customers to phoney versions of the official websites, giving the impression that they are part of the same organisation.

Additionally, the provision under Section 81 of the Act is an obstante clause whereby the provisions of the IT act would take precedent and override all other provisions within the exiting framework. However, it is also important to note that, pursuant to Section 77B of the IT Act (Amendments 2008), phishing scams are bailable. This is based on the inability to determine with certainty, who is the perpetrator behind the crime. The mode of the crime creates a translucent screen before the phisher, which hides their identity and results in situations wherein an innocent person can get convicted for a crime that they have never committed; this creates the need to make provisions for bail under Section 81.

Additionally, the Indian Penal Code contains the following provisions under which an individual can be held liable for the crime of phishing:

Theft under Section 378 and 379

Criminal breach of trust under Section 405 and 406

Cheating under Section 415 to 419

 ISSN: 2582-6433

Mischief under Section 425 and 426, and
Forgery under Sections 463 – 465, and Sections 467-477.

1.8 Objective and Scope of the Project

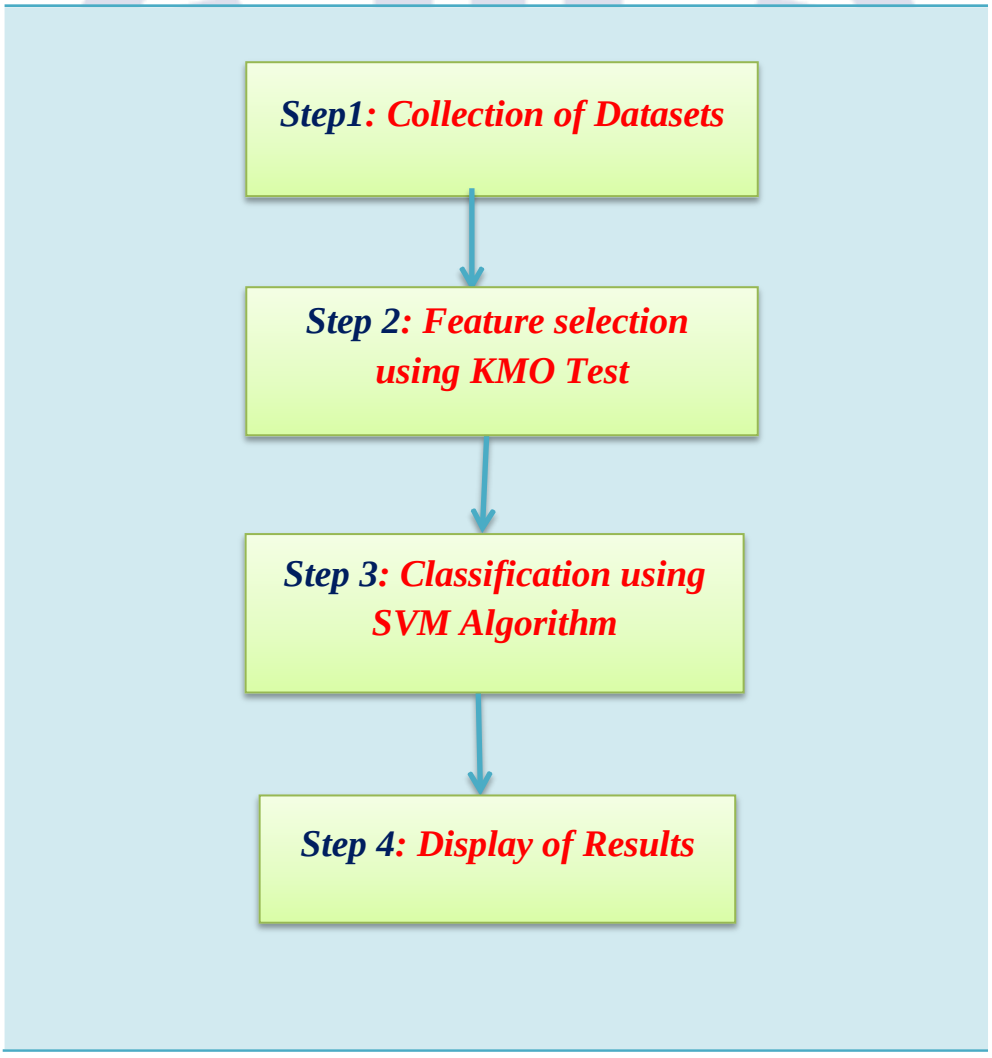
The objective of this project is to classify the phishing attack with the aid of machine learning algorithms. The scope of the project is to select the feature using (KMO test) and applying the machine learning algorithms (SVM).

1.9 Organization of the Report

Chapter 1 discussed the introduction, motivation and justification, objective and scope of mini project. Chapter 2 discusses the overall flow of the project. Chapter 3 gives the result and discussion. Chapter 4 discusses the conclusion and future scope. Chapter 5 gives the references and Chapter 6 presents the annexures.

II METHODOLOGY

Chapter II gives overall design for classification of Phishing Websites. There are four steps involved in this design. Figure 3 gives the overall flow of the project.

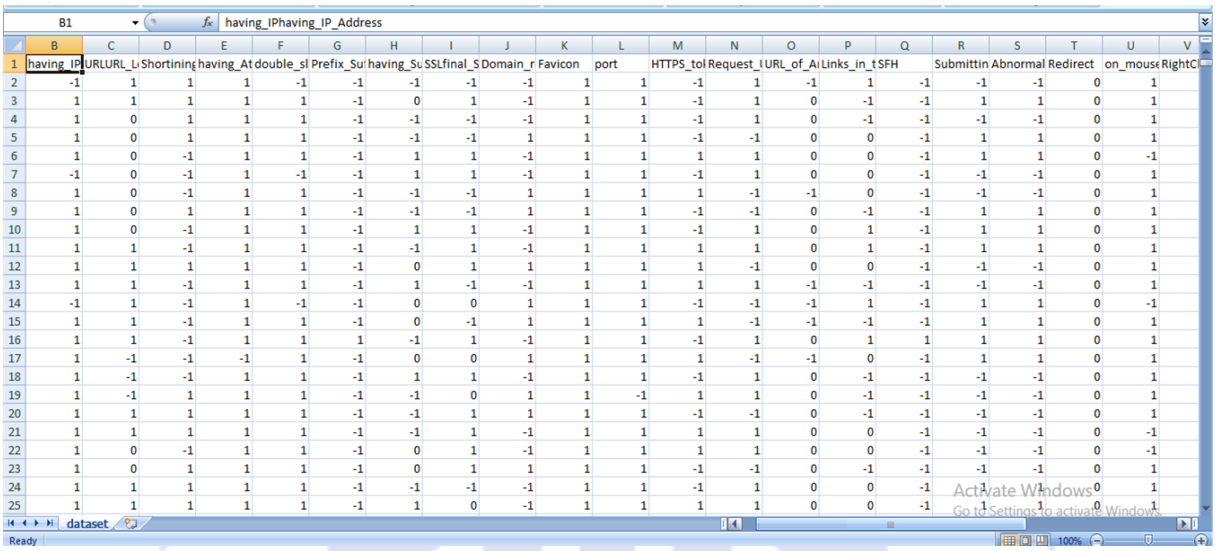


 ISSN: 2582-6433

Figure 3 Overall design of the project.

2.1 Collection of Datasets

- In this project the phishing websites data sets has been available in UCI repository which has been published during the year 2016.
- The Phishing website containing 28 fields with four classes attributes. This dataset also consist of 11055 instances.
- Figure 4 displays the sample dataset which have used for this project.



having_IP	URLURL_L	Shortining	having_At	double_sl	Prefix_Su	having_Su	SSLfinal_S	Domain_r	Favicon	port	HTTPS_tol	Request_L	URL_of_A	Links_in_t	SFH	Submittin	Abnormal	Redirect	on_mouse	RightC
-1	1	1	1	-1	-1	-1	-1	-1	1	1	-1	1	-1	1	-1	-1	-1	0	1	
1	1	1	1	1	-1	0	1	-1	1	1	-1	1	0	-1	-1	1	1	0	1	
1	0	1	1	1	-1	-1	-1	-1	1	1	-1	1	0	-1	-1	-1	-1	0	1	
1	0	1	1	1	-1	-1	-1	1	1	1	-1	-1	0	0	-1	1	1	0	1	
1	0	-1	1	1	-1	1	1	-1	1	1	1	1	0	0	-1	1	1	0	-1	
-1	0	-1	1	-1	-1	1	1	-1	1	1	-1	1	0	0	-1	-1	-1	0	1	
1	0	-1	1	1	-1	-1	-1	1	1	1	1	-1	-1	0	-1	-1	-1	0	1	
1	0	1	1	1	-1	-1	-1	1	1	1	-1	-1	0	-1	-1	1	1	0	1	
1	0	-1	1	1	-1	1	1	-1	1	1	-1	1	0	1	-1	1	1	0	1	
1	1	-1	1	1	-1	-1	1	-1	1	1	1	1	0	1	-1	1	1	0	1	
1	1	1	1	1	-1	0	1	1	1	1	1	-1	0	0	-1	-1	-1	0	1	
1	1	-1	1	1	-1	1	-1	-1	1	1	1	1	-1	-1	-1	-1	-1	0	1	
-1	1	-1	1	-1	-1	0	0	1	1	1	-1	-1	-1	1	-1	1	1	0	-1	
1	1	-1	1	1	-1	0	-1	1	1	1	1	-1	-1	-1	-1	1	1	0	1	
1	1	-1	1	1	1	-1	1	-1	1	1	-1	1	0	1	1	1	1	0	1	
1	-1	-1	-1	1	-1	0	0	1	1	1	1	-1	-1	0	-1	1	1	0	1	
1	-1	-1	1	1	-1	1	1	-1	1	1	-1	1	0	-1	-1	-1	-1	0	1	
1	-1	1	1	1	-1	-1	0	1	1	-1	1	1	0	-1	-1	-1	-1	0	1	
1	1	1	1	1	-1	-1	1	1	1	1	-1	-1	0	-1	-1	-1	-1	0	1	
1	1	1	1	1	-1	-1	1	-1	1	1	1	1	0	0	-1	-1	-1	0	-1	
1	0	-1	1	1	-1	0	1	-1	1	1	1	1	0	0	-1	-1	-1	0	-1	
1	0	1	1	1	-1	0	1	1	1	1	-1	-1	0	-1	-1	-1	-1	0	1	
1	1	1	1	1	-1	-1	-1	-1	1	1	-1	1	0	0	-1	-1	-1	0	1	
1	1	1	1	1	-1	1	0	-1	1	1	1	1	0	0	-1	-1	-1	0	1	

Figure 4 Sample dataset

- The 28 fields are classified into 4 types .Figure 5 gives the classification of field.

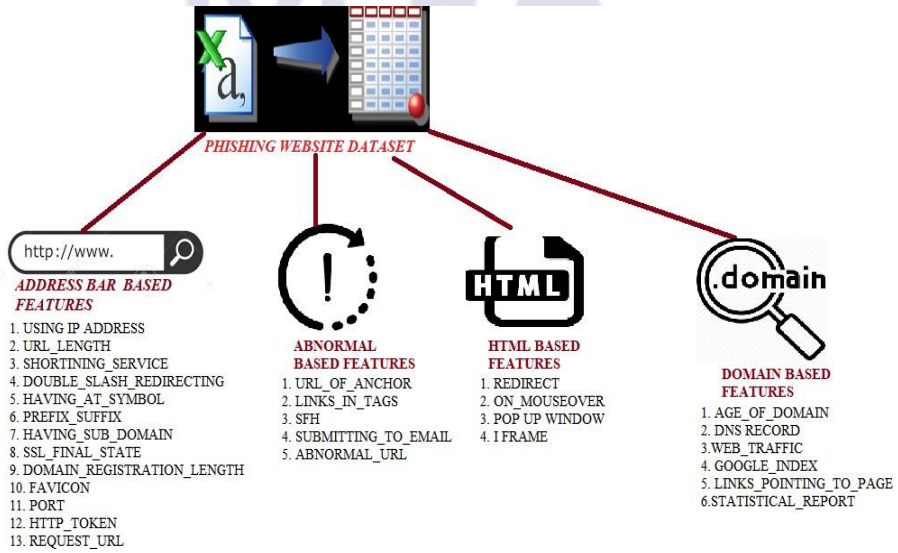


Figure 5 Classifications of Fields

 ISSN: 2582-6433

After the collection of dataset the next step follows the Feature Selection techniques.

2.2 Feature Selection using KMO Test

2.2.1 What is Feature Selection?

In data mining and machine learning, real-world problems often involve a large number of features. Feature selection aims to solve this problem by selecting only a small subset of relevant features from the original large set of features. By removing irrelevant and redundant features, feature selection can reduce the dimensional of the data, speed up the learning process, simplify the learned model, and/or increase the performance.

2.2.2 Methods involved in Feature Selection techniques

Method involved in feature selection technique is classified into three categories.

Table 1 tells about the methods involved in Feature Selection techniques.

Table 1 Methods involved in Feature Selection.

Filter Method	**Wrapper Method**	**Embedded Method**
Filter methods are generally used as a pre-processing step. The selection of features is independent of any machine learning algorithms. Instead, features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable. Pearson‘s correlation: **KMO test** Lda Anova chi-square	In wrapper methods, we try to use a subset of features and train a model using them. The problem is essentially reduced to a search problem. these methods are usually computationally very expensive forward selection backward elimination recursive feature elimination:	Embedded methods combine the qualities‘ of filter and wrapper methods. It is implemented by algorithms that have their own built-in feature selection methods. lasso regression ridge regression

 ISSN: 2582-6433

2.2.3 What is Correlation Matrix?

A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table. A correlation matrix consists of rows and columns that show the variables. Each cell in a table contains the correlation coefficient. KMO takes value between 0 which indicate that the sum of the partial correlation is large compared to the sum of correlation, 1which indicate the variables are clustering among a few variables.

2.2.4 What is KMO Test?

- KMO stands for ‒Kaiser –Meyer-Olkin‖ test is a measure of how suited our data is ***factor analysis***
- This test measures the **sampling adequacy** for each variable in the model and for the complete model.
- The statistic is a measure of the **propagation of variance** among variables.
- The variables in any dataset are more or less correlated but the correlation between 2 variables can be influenced.
- For this reason, we are calculating the partial correlation matrix by calculating the inverse matrix as $[R^{-1}=V_{ij}]$ let we assume the partial correlation matrix as $A=A_{(Ij)}$

$$A_{ij} = - \frac{V_{ij}}{\sqrt{V_{ii} * V_{jj}}} \quad \text{----------[1]}$$

- We compute the KMO index per variable in to detect which feature are not to other

$$KMO = \frac{\sum_{i!=j} r^2_{ij}}{\sum_{i!=j} r^2_{ij} + \sum_{i!=j} a^2_{ij}} \quad \text{----------[2]}$$

- To find the correlation matrix in $[r^{-1}=v_{ij}]$ and partial variance matrix of $a=a_{(ij)}$, finally, we compute the overall KMO index As Equation 3

$$KMO = \frac{\sum_i \sum_{i!=j} r^2_{ij}}{\sum_i \sum_{j!=i} r^2_{ij} + \sum_i \sum_{j!=ij} a^2_{ij}} \quad \text{----------[3]}$$

ISSN: 2582-6433

2.2.4. i. Interpretation of KMO Measure

- ❖ A **rule of thumb** for interpreting the statistic:
- ➢ KMO values between 0.8 and 1 indicate the sampling is adequate.
- ➢ KMO values less than 0.6 indicate the sampling is not adequate
- ➢ KMO values close to zero means that there are large partial correlations compared to the sum of correlations. Table 2 gives the Interpretation of KMO Test.

Table 2 Interpretation of KMO Test

KMO	Interpretation
0.9 And Above	Marvellous
0.8-0.9	Meritorious
0.7-0.8	Midding
0.6-0.7	Mediocre
0.5-0.6	Miserable
Under 0.5	Unacceptable

2.2.4. ii. Pseudo Code for KMO Test

Table 3 displays the Pseudo cod for KMO Test.

Table 3 Pseudo code for KMO Test

i.	*Create a function KMO(a)*
ii.	*Find the correlation matrix of 'a' using cor() function*
iii.	*Apply KMO index per variable formula by calculating the*
	✚ *sum of square of cor(a) divided by overall sum of cor(a) plus the sum of partial correlation of 'a'*
iv.	*Repeat the step (iii) for calculating overall KMO index with the colsum() function.*
v.	*Call the KMO (a) function.*

The above KMO test has been applied as a feature selection. As per the methodology discussed above after feature selection the next step follows classification techniques.

ISSN: 2582-6433

2.3 Classification using SVM Algorithm

2.3.1 What is Machine Learning?

Machine learning is an application of artificial intelligence that provides system the ability to automatically learn and improve experience. The primary aims us to allow the computers to learn automatically without human assistance. The process of learning begins with observations of data's such as in order to look for patters in data and make better decisions in the future. Following are the basic steps in machine learning tool,

- Gather the past data
- Prepare Data
- Develop data model with appropriate algorithm
- Test model accuracy using datasets,
- Check and improve the performance using various datasets.

2.3.2 Types of Machine Learning Algorithm

Machine learning can be grouped into three broad learning tasks. Figure 6 tells about the types of Machine Learning.

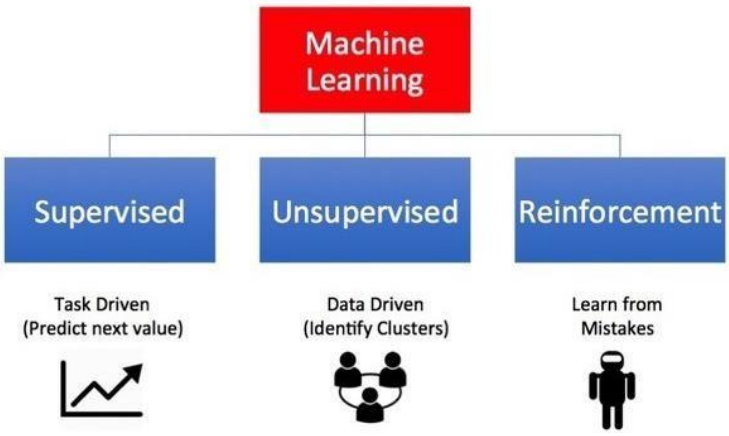


Figure 6 Types of Machine Learning

2.3.2.1 Supervised Learning Algorithm

ISSN: 2582-6433

- Supervised learning, the machine is trained using labelled dataset.
- Labelled dataset is one which have both INPUT and OUTPUT parameters.
- Supervised learning algorithm are classified into
  Classification Regression.

2.3.2.1. i. Classification

Classification is a technique for determine class the dependent belongs to base on the one or more independent variables. A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes. According Ikhlas Abdel-Qader[1], **support vector machine algorithm** works better when compared to Logistic regression, K-NN classifier and back propagation. Figure 7 tells about the types of classification algorithm.

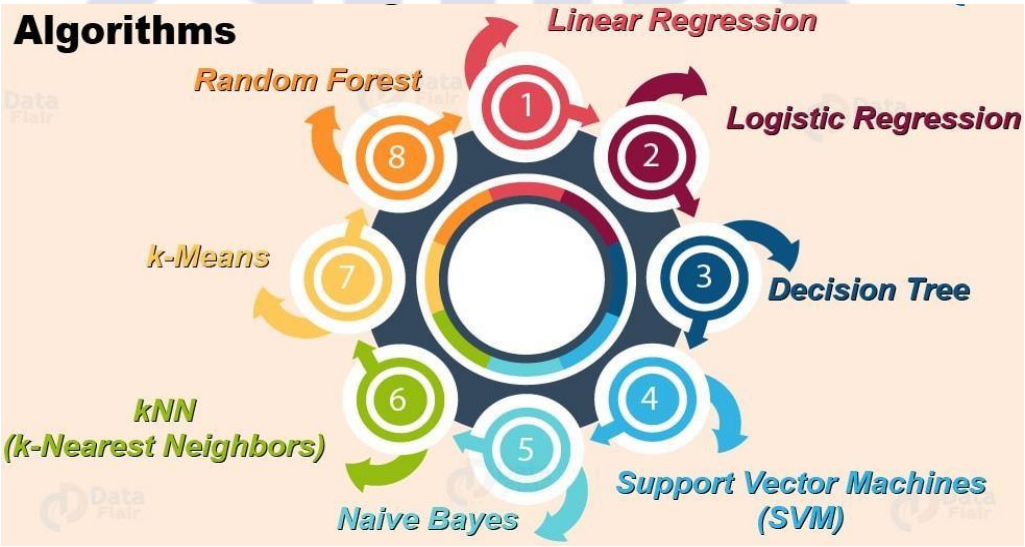


Figure 7 Types of Classification Algorithm

a. Support Vector Machine (SVM)

Support Vector Machines (SVMs) are supervised learning methods used for classification and regression tasks that originated from statistical learning theory. As a classification method, SVM is a global classification model that generates non-overlapping partitions and usually employs all attributes. The entity space is partitioned in a single pass, so that flat and linear partitions are generated. SVMs are based on maximum margin linear discriminant, and are similar

ISSN: 2582-6433

to probabilistic approaches, but do not consider the dependencies among attributes.

➢ SVM Kernel Functions

SVM algorithms use a set of mathematical functions that are defined as the kernel. The function of kernel is to take data as input and transform it into the required form. Different SVM algorithms use different types of kernel functions. These functions can be different types.

Linear kernel Function

Non Linear Kernel Function

Linear Kernel

Linear Kernel is used when the data is linearly separable, that is, it can be separated using a single Line. It is one of the most common kernels to be used. It is mostly used when there are a large number of Features in a particular Data Set. One of the examples where there are a lot of features is Text classification, as each alphabet is a new feature. So it has been most commonly used types. Linear kernel function has been used to classify the Phishing Websites. Table 4 gives the linear classifier formula.

Table 4 Linear classifier formula.

FORMULA:
Linear classifier **f(x) = N. ∑i αi**

Non-Linear Kernel

❖ **Radial basis function kernel (RBF)/ Gaussian Kernel**

Gaussian RBF (Radial Basis Function) is another popular Kernel method used in SVM models for more. RBF kernel is a function whose value depends on the distance from the origin or from some point. Gaussian Kernel is of the following format. Table 5 gives the Non-Linear classifier formula.

Table 5 Non-Linear classifier formula

$$K(X_1, X_2) = exponent(-\gamma \|X_1 - X_2\|^2)$$

||X1 — X2 || = Euclidean distance between X1 & X2

Using the distance in the original space we calculate the dot product (similarity) of X1 & X2.

ISSN: 2582-6433

2.3.2.1. ii. Pseudo Code for SVM Algorithm

Table 6 displays the Pseudo code for SVM Algorithm.

Table 6 Pseudo code for SVM Algorithm

Step1: *Normalize the dataset.*
Step2: *Train and Test the dataset*
Step3: *Compute the SVM linear formula in trained attribute and predict the model*
Step 4: *Repeat the step 3 for calculating the confusion matrix by calling the predication model.*
Step 5: *Call the linear kernel function for calculating the parameters with the cost=1.*
Step 6: *Print and plot the parameters*
Step 7: *End the process.*

SVM algorithm has been used for classifying the phishing websites. The next steps follow the display of results as per the methodology.

III RESULTS AND DISCUSSION

The database for this project has been taken from the **Phishing website dataset** in **UCI** repository. The phishing website dataset contains 11055 instances, including 30 attributes. The features are categorized into four groups: Address Bar, Abnormal based feature, HTML based features, and Domain based features.

With the **help of KMO test 18 features are selected** on the basis of KMO index per variable where the measure lies above 0.78 have been taken where more features are correlated to each other. With the aid of KMO test, 36 percentage features have been reduced. Table 7 tells about the features selected using KMO test.

Table 7 Features selected using KMO Test

S NO	FEATURE	MEASURE
1	ON_MOUSE	0.93
2	HTTP_TOKEN	0.90

ISSN: 2582-6433

3	ABNORMAL_URL	0.87
4	SHORTING_SERVICE	0.87
5	REDIRECT	0.87
6	RIGHT_CLICK	0.86
7	LINKS_IN_TAGS	0.85
8	DOUBLE_SLASH_REDIRECT	0.84
9	PORT	0.84
10	SUBMIITING_TO_EMAIL	0.84
11	PERFIX_SUFFIX	0.81
12	SSL FINAL_STATE	0.81
13	WEB_TRAFFIC	0.81
14	FAVICON	0.80
15	POP_UP_WINDOW	0.80
16	I_FRAME	0.80
17	STATISTICAL_REPORT	0.79
18	HAVING_SUB_DOMAIN	0.78

The SVM classification algorithm is applied to the selected features to evaluate the accuracy.

Table 8 gives the **Confusion Matrix and Statistics**

Table 8 Confusion Matrix and Statistic

	Reference	
Prediction	-1	1
-1	142	7
1	30	876

Table 9 gives the **Sensitivity formula**

Table 9 Sensitivity formula

$$Sensitivity = \frac{TP}{TP+FN}$$

TP - True Positive
FN- False Negative

Table 10 gives the **Specificity formula**

Table 10 Specificity formula

$$Specificity = \frac{TN}{TN+FP}$$

TN- True Negative
FP- False Positive

ISSN: 2582-6433

Table 11 gives the **Accuracy** formula

Table 11 Accuracy formula

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

TP-True Positive
TN- True Negative
FP-False Positive
FN- False Negative

Table 12 displays the Performance metrics of the classification algorithm SVM

Table 12 Performance metrics of the classification SVM

Algorithm	Feature Selection	No of Features	Sensitivity	Specificity	Accuracy	Positive Class
SVM	KMO Test	18	0.8256	0.9921	0.9649	-1(Phishing website)

Table 13 gives the Parameter for SVM Classification.

Table 13 Parameter for SVM Algorithm

SVM-Type: C-classification

SVM-Kernel: linear
cost: 1

Number of Support Vectors: 8

This Chapter briefly discussed the experimental results. Next Chapter discusses the conclusion and future scope of the project.

ISSN: 2582-6433

Iv. Conclusion And Future Scope

4.1 Conclusion

Phishing attack is one of the common types of cyber-attacks where the attackers steal user‘s credential information in the form of websites, e-mails, sms , or through phone calls where the user loses their sensitive information which may leads to cyber-threat. In this project the dataset had been collected for UCI repository and feature selection technique (KMO Test) has been applied based on the correlation matrix was implemented to analyze the dataset and performance was evaluated using Support Vector Machine (SVM) classification algorithm. The accuracy level because of KMO Test and SVM algorithm leads to 96%

4.2 Future Scope

In this project feature selection technique and classification algorithm had been used to classify the phishing website. In future to classify the phishing website many feature selection technique such as Pearson‘s correlation, forward selection method, back propagation method etc. can be applied and many algorithms such as K-NN, Logistic regression algorithm can be applied to improve the performance.

5.1 References

Papers

[1] Wesam Fadheel, Mohamed Abusharkh and Ikhlas Abdel-Qader **–On Feature Selection for the Predication of Phishing Websites‖** IEEE paper published on 3rd International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress (2017).

[2] Minal Chawla, Siddarth Singh Chouhan **–A Survey of Phishing Attack Techniques (2016)‖** IEEE paper published on 7th Annual International Technology. Electronics and Mobile Communication Conference (IEMCON)

[3] A.K. Jain and B.B.Gupta, **‖Comparative analysis of features based machine learning approach for phishing detection(2016)”** IEEE paper published on 3rd International Conference on Computing for Sustainable Global Development(INDIACom), New Delhi 016.pp.2125-2130.

ISSN: 2582-6433

Websites

[4] Phishing Website dataset has been available in http://archive.ics.uci.edu/ml/datasets/Website+Phishing

[5] www.imperva.com

[6] www.statistichowto.datasciencecentral.com

[7] www.torwardsdatascience.com