

# Phishing Website Detection System Using Machine Learning

**Md. Arif Khan**

*Department of Computer Science  
And Engineering, City University,  
Bangladesh*

**Md Sazzad Hossen**

*Department of Computer Science  
And Engineering, City University,  
Bangladesh*

**Md Ataullah Bhuiyan**

*Department of Computer Science  
and Engineering, City University,  
Bangladesh*

**Mahir Labib Hossain**

*Department of EEE  
Ahsanullah University Science  
and Technology*

**Abstract:** Phishing, categorized as a Social Engineering Attack, poses a prevalent security threat by deceptively extracting private and confidential information from users without their awareness. This sensitive data encompasses usernames, passwords, account numbers, and more. To encounter this, we proposed a website equipped with a Machine Learning (ML) Algorithm to empower users in identifying phishing websites. The effectiveness of this approach is particularly notable when applied to extensive datasets, addressing limitations present in current methodologies and enabling the detection of zero-day attacks. While ML-based classifiers generally demonstrate optimal accuracy and performance, their efficacy is contingent upon factors such as the scale of training data, feature set, and classifier type. Notably, these classifiers may fall short in detecting instances where attackers employ compromised domains for hosting their sites. Despite the absence of a foolproof system for identifying all phishing websites, adopting these methods promises an efficient means of detection.

**Keywords:** Phishing Website Detection System; Machine Learning Algorithm; Dataset; Feature Selection; Domain; Email

## Nomenclature

Abbreviation	Expansion
ML	Machine Learning
MCAC	Multi-label Classifier based Associative Classification
DT	Decision Tree
NB	Naïve Bayesian
SVM	Support Vector Machine
NN	Neural Network
RF	Random Forest
HS	Harmony Search
ELM	Extreme Learning Machine
URL	Uniform Resource Locator
SLS	Second Level Space
VoIP	Voice over Internet Protocol
SMS	Short Message/Messaging Service
KNN	K-Nearest Neighbours
IP	Internet Protocol
HTTP	Hypertext Transfer Protocol
DNS	Domain Name System
ID3	Iterative Dichotomised 3
MAP	Maximum A Posteriori
ER	Entity Relationship
API	Application Programming Interface
GUI	Graphical User Interface
CSS	Cascading Style Sheets

## 1. Introduction

Phishing is one of the most dangerous criminal activities in cyberspace [19]. It is estimated that 3.4 billion spam emails are sent every day [20]. There are numerous individuals are mindful of utilizing the web to perform different exercises like Online shopping, Online charge installment, Online versatile revives, and keeping money exchange. Due to the wide utilization of these clients confront different security dangers like cybercrime. There's much cybercrime that's broadly performed for illustration spam, extortion, cyber terrorism, and phishing. Among these phishing is modern cybercrime

exceptionally well-known these days. Phishers make a fake URL but web pages imitate genuine websites. A non-technical individual can effectively accept that's genuine. Ordinarily, phishing assaults misuse social design to bait the casualty by sending a spoofed interface by diverting the casualty to a fake web page. The spoofed interface is set on the prevalent web pages or sent through mail to the casualty. The fake web page is compared to the trueblue webpage. In this way, instead of coordinating the casualty ask to the real web server, it'll be coordinated to the aggressor server.

Phishing exercises can happen by a person or a gathering. Within the current scenario, when the end-user needs to get to his private data online (within the frame of cash exchange or installment door) by logging into his bank account or secure mail account, the individual enters data like Username, Watchword, Credit card no. etc. on the login page. But very regularly, this data can be captured by aggressors using phishing strategies (For occasion, a phishing site can collect the login data the client enters and redirect him to the initial location). There's no such data that cannot be straightforwardly gotten from the user at the time of his login input. Phishing assaults can help organizations keep an eye on a tremendous number of cash per assault in fraud-related hardships and staff time. Distant more loathsome, costs associated with the corruption of brand picture and shopper certainty can keep running into a tremendous number of dollars. Phishing assaults may show up in numerous sorts of communication shapes such as informing, SMS, VOIP, and fraudster emails. Clients commonly have numerous client accounts on different websites including social systems, mail, and conjointly accounts for keeping money. Subsequently, blameless web users are the foremost helpless targets for this attack since the truth is that most individuals are unaware of their profitable data, which helps to make this assault effective.

We have developed a framework for identifying and predicting phishing websites using classification information mining algorithms. Our approach involves implementing classification algorithms to extract criteria from phishing datasets, aiding in the authentication classification process. Key characteristics such as URL structure, domain characteristics, and security and encryption criteria are utilized to effectively identify phishing sites. This framework enhances the accuracy of distinguishing phishing websites by considering critical features within the last phishing location rate. Then the main objective of the proposed method is

- To create a secured Internet transaction facility for Online payment with higher accuracy.
- To implement confidential online goods purchases with the aid of technology.

The phishing website detection system achieved high accuracy using ML algorithms. The paper is organized as Section 2 represents the Literature review, Section 3 explains the methodology, Section 4 covers the Experimental Analysis & Result, Section 5 mentions the System Analysis & Design, Section 6 demonstrates the technology, Section 7 covers Web Implementation, Section 8 covers advantages and disadvantages and Section 9 concludes the paper with future scope.

## 2. Background and Related Study

### 2.1 Machine Learning

ML is an application of AI that gives outlines the capability to concurrently learn and progress from encounters without being expressly modified. ML focuses on the enhancement of computer programs to get information and utilize it to memorize for themselves. There are many ML methods.

**Supervised ML Algorithms** can forecast future events by leveraging labeled examples and past knowledge applied to new data.

**Unsupervised ML Algorithms** are employed when training data lacks both classification and labeling.

**Semi-supervised ML Algorithms** provide both unlabelled and labeled data for training and a small quantity of labeled data and a significant amount of unlabelled data were placed in between supervised and unsupervised learning.

**Reinforcement ML Algorithms** area kind of learning that responds to its surroundings by acting and identifying mistakes or rewards.

### 2.2 Existing Work

In 2024, Abdelhamid *et al.* [1] used MCAC for detecting the phishing website with high accuracy. The method implemented the "If-Then" rule with a high degree of predictive accuracy. The result showed higher accuracy in real-time data collection from various sources. However, MCAC generated a new

hidden knowledge (rules) that other algorithms were unable to find and this had to be improved its classifier's predictive performance.

In 2016, Arun Kulkarni *et al.*, [2] developed methods of defence utilizing various approaches to categorize websites. They had developed a system that uses ML techniques to classify websites based on their URL. It used four classifiers: the DT, NB classifier, SVM, and NN. Then, The classifiers were tested with a data set and categorized as a Suspicious site, Legitimate site, or Phishing site. The results showed that the classifiers were successful in distinguishing real websites from fake ones.

In 2020, Nur Sholihah Zaini *et al.*, [3] investigated and evaluated the effectiveness of the ML approach in the classification of attacks. It was a heuristic approach through an ML classifier to find the phishing attacks in the website applications. Then they were compared with five classifiers and found RF can achieve high detection accuracy.

In 2015, Bhagya shree E *et al.*, [4] implemented a feature-based approach to classify URLs into phishing or non-phishing categories. The usage of a variety of URL features was carried by the anatomy of URLs. To classify URLs, two different algorithms were used. RF-ML algorithm was an efficient classifier to decide whether the URL was phishing or not. Moreover, a scheme was used to detect phishing URLs by mining the publicly available content on the URLs.

In 2019, Mehdi Babagoli *et al.*, [5] executed a meta-heuristic-based nonlinear regression algorithm along with the feature selection approach. The dataset was used to validate the legitimate web pages and extract features from the websites. DT and Wrapper were used as feature selection methods. After that, the prediction and detection of fraudulent websites are performed through HS and SVM. The result showed that the nonlinear regression-based HS results in better performance compared to SVM.

In 2019, Sandeep Kumar Satapathy *et al.*, [17] devised an ELM-based classification for features including Phishing website data in the UC Irvine ML Repository database. They used various features based on web pages. NB was used to detect phishing web pages. Finally, ELM was compared with other ML methods such as NB and ANN for the highest accuracy.

In 2019, Sophiya Shikalgar, *et al.*, [18] used URL features to identify features that phishing site URLs contain. This method employed features for phishing detection for maximum accuracy. Various ML algorithm was combined to increase the accuracy in the prediction of phishing attacks.

## 2.3 Comparison with Existing Works

Table 1 portrays the Algorithm, Accuracy, Real-world application, and feature used. We considered seven papers that used a different algorithm for the detection of Phishing Websites. Each method has certain accuracies and features that were explained in detail.

**Table 1:** Comparison with Existing Work

Name	Algorithm	Accuracy	Real-World Application	Feature used
Abdelhamid, <i>et al.</i> [1]	AC	Not Specified	No	16
Arun Kulkarni, <i>et al.</i> , [2]	DT and NB	90%, 86%	No	9
Nur SholihahZaini, <i>et al.</i> , [3]	RF and KNN	92.79%, 90%	No	30
Bhagyashree E, <i>et al.</i> , [4]	RF and Content-based Algorithm	Not Specified	No	Not Specified
Mehdi Babagoli, <i>et al.</i> , [5]	HS and SVM	92.80%, 91.83%	No	20
Sandeep Kumar Satapathy, <i>et al.</i> , [17]	ELM, NN, and NB	89.3%, 87.9%, 61.3%	No	30
Sophiya Shikalgar, <i>et al.</i> , [18]	XGBoost, SVM, and NB	85.5%, 86.3%, 80.2%	No	9
Proposed System	DT, RF, SVM, XGBoost, KNN, Naive Bayes	95.4, 95.6, 93.8, 72.5, 92.8, 75.2	Yes	19

## 2.4. Research Gap

- Existing approaches in phishing website detection systems cannot detect when attackers use compromised domains for hosting their sites.
- The complexity and time-consuming nature of random forest algorithms pose challenges in comparison to other ML algorithms.
- There is a need for efficient feature selection and dataset analysis to improve the accuracy of phishing website detection systems.

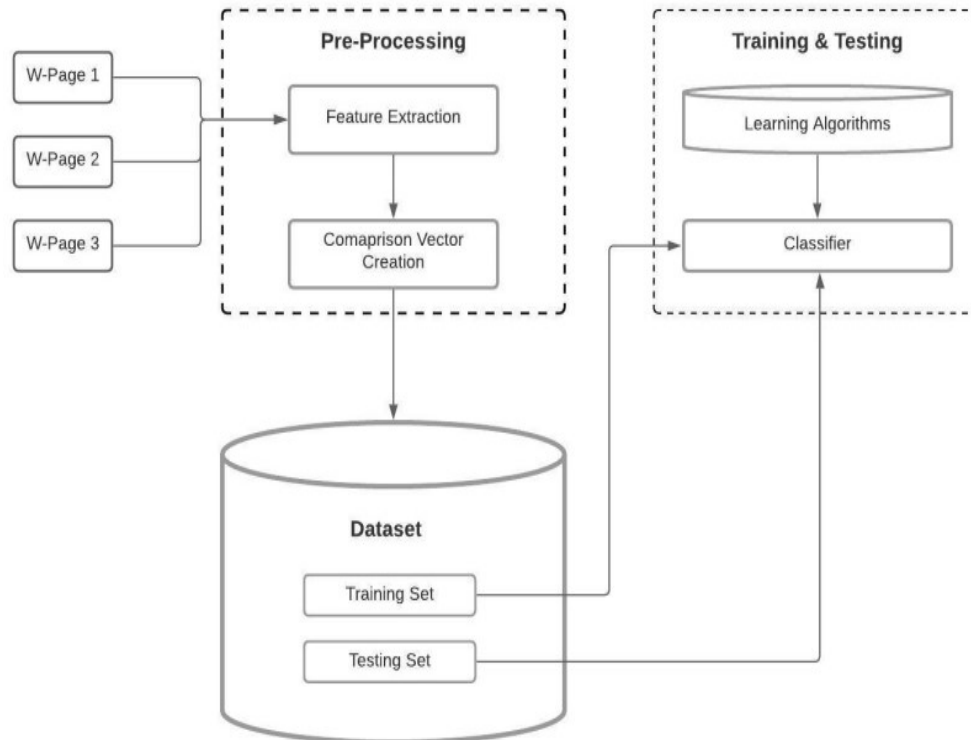
## 3. Methodology

### 3.1 Proposed System Architecture

We aim to create techniques that can identify the phishing page layouts by taking layout features. ML is used, where it is used to create similarity models based on the statistical properties obtained from the training data sets. First, we get a classifier for determining page similarity based on layout features. This phase consists of a pre-processing phase and a training phase. In the pre-processing phase, the feature is extracted and sent to comparison vector creation. The comparison vectors that summarise key similarity features of each web page pair accordingly. In the classifier training phase, we take into account a set of labeled comparison vectors. The similar page classifiers obtained in this phase can then be used to decide whether 2 web pages are comparable based on their comparison vectors. Finally, In the detection stage, the classifier directly identifies the phishing website or the URL is compared with the Blacklist URL stored in the database from which the phishing website is identified.

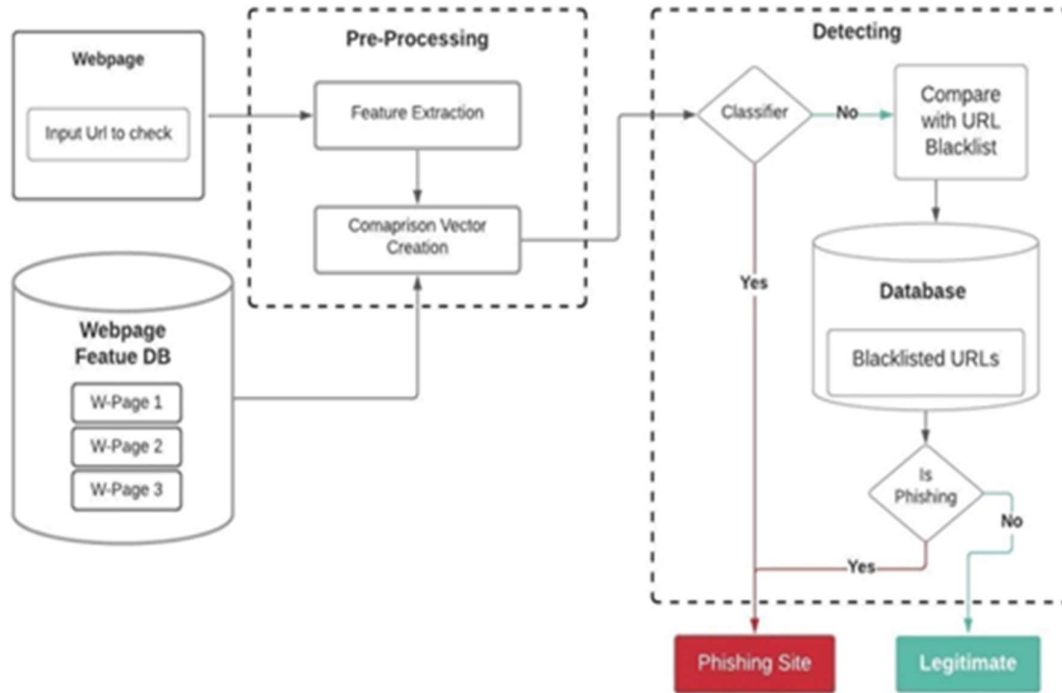
The main target users are persons, educationalists, Online shoppers, new users, and other browsers. Around 2,00,000 BDT (Twolakh) is needed to implement this process professionally. This experiment was carried out from 07-06-2020 to 10-12-2020 approximately six months.

A



**Fig. 1.** Train layout similarity classifier

B



**Fig. 2.** Detect suspicious web pages.

### 3.2 Data Collection

The information set utilized in this paper was downloaded from Kaggle. It contains highlights from 11,054 URLs. Out of these, 4,897 are genuine and 6,157 are phishing. The information set moreover contains an add-up of 32 columns and 31 highlights that were extricated from each URL. 30 input highlights and one yield highlight. The traits give data such as the Long URL, Short URL, Using IP, Diverting//, Image@, Sub Domains, Prefix Suffix-, Domain Reg Len, HTTPS, Non-Std Port, Favicon, Request URL, HTTPS Domain URL, Links In Script Tags, Anchor URL, Server Form Handler, Abnormal URL, Info Email, Status Bar Cust, Website Forwarding, Using Popup Window, Disable Right Click, Age of Domain, Iframe Redirection, website traffic, DNS Recording, google index, Page Rank, Stats Report, Links Pointing To Page, course. Each includes esteem that holds parallel values. Parallel values show that the presence or the need of the presence of the included inside the URL decides the esteem relegated to that include.

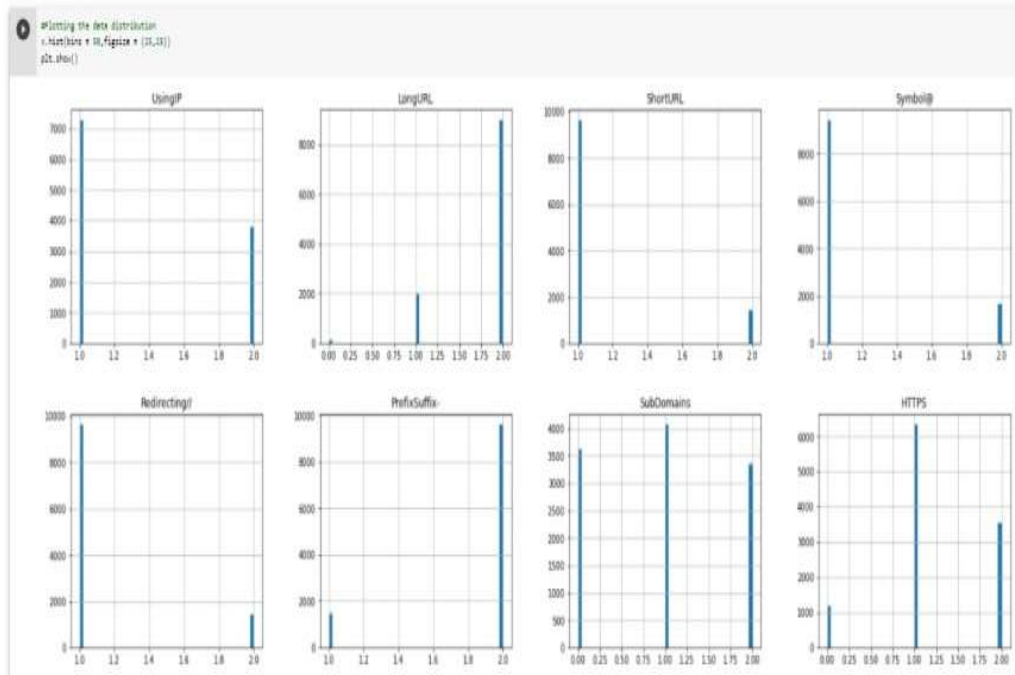
### 3.3 Data Pre-Processing

ML calculations do not work so well with handling crude information. so that it must pre-process the crude information to bolster various ML calculations. In the data pre-processing, crude information is removed to form a clean information set.

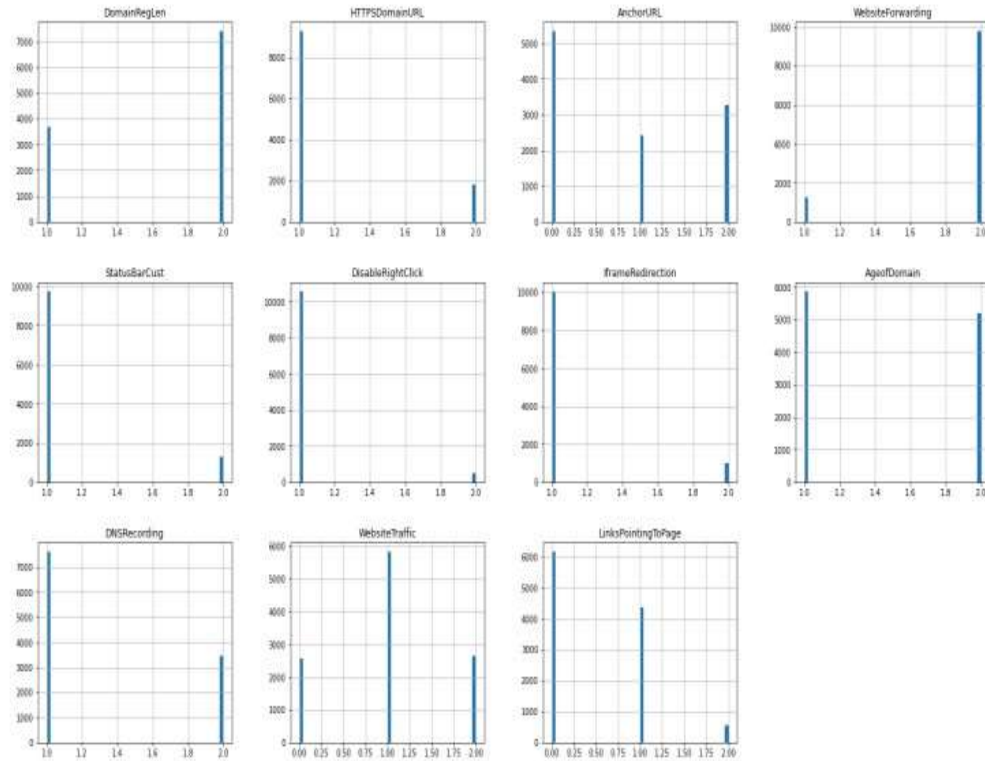
Fig 3 and 4 represent the data before getting started on the cleaning procedure.

#### 4. Visualizing the data

Few plots and graphs are displayed to find how the data is distributed and the how features are related to each other.



**Fig. 3. Data Visualization**



**Fig. 4. Data Visualization**

### 3.3.1 Data Cleaning

The new complete guide in Python.

- Missing Data
- Irregular Data (Outliers)
- Unnecessary Data — Repetitive Data, Duplicates, and more.

- Inconsistent Data — Capitalization, Addresses, and more.

### 3.3.2 Missing Data Checking

Information cleaning or cleansing is the method of identifying and adjusting (or evacuating) degenerate or wrong records from a record set, table, or database and refers to recognizing inadequate, inaccurate, wrong, or unessential parts of the information and after that supplanting, adjusting, or erasing the grimy or coarse information. No models make meaningful results with chaotic information.

To find the missing value, we used this code which is shown in Fig. 5.

```
[13] #checking the data for null or missing values
      data.isnull().sum()
```

```
UsingIP          0
LongURL          0
ShortURL         0
Symbol@         0
Redirecting//    0
PrefixSuffix-   0
SubDomains      0
HTTPS           0
DomainRegLen    0
HTTPSDomainURL  0
AnchorURL       0
WebsiteForwarding 0
StatusBarCust   0
DisableRightClick 0
IframeRedirection 0
AgeofDomain     0
DNSRecording    0
WebsiteTraffic  0
LinksPointingToPage 0
class           0
dtype: int64
```

*Fig. 5: List of Missing value*

In our dataset, we didn't get any missing values in a single column.

### 3.3.3 Missing Data Filled Up

We don't have lost values, so we do not require this step. But in case we have missing esteem, ready to drop either columns or columns with lost information. We will utilize dropna () to evacuate all columns with lost information.

### 3.3.4 Data Integration

Information integration frameworks are progressively looking to utilize ML-based approaches for finding and highlighting the islands of valuable information within the tremendous sea of dull information (and in this way move forward analytics). Metadata is picking up a more grounded accentuation and is being captured expressly or gathered with offer assistance of ML. A few cases are the use of ML within the Deduction of patterns, information conveyance, and common esteem designs.

### 3.3.5 Split Dataset

The performance of ML algorithms when they are applied to prediction on non-training data is estimated using the train-test split approach. We have 25% of the data for testing and 75% for training. To find the split dataset, we used this code which is shown in Fig. 6.

```
[15] # Separating & assigning features and target columns to X & y

x = data.drop('class',axis=1)
y = data['class']
x.shape, y.shape

((11054, 19), (11054,))

[16] x.shape, y.shape

((11054, 19), (11054,))

[17] from sklearn.feature_selection import SelectKBest

[34] # Splitting the dataset into train and test sets: 75-25 split
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(x, y, test_size = 0.25, random_state = 4)
```

*Fig. 6. Split Dataset*

### 3.3.6 Training Data

For training, we have 8290 data from 11,054 data. The code for training data is mentioned in Fig. 7.

```
[37] X_train.shape, y_train.shape

((8290, 19), (8290,))
```

*Fig. 7. Training Data*

### 3.3.7 Testing Data

For testing, we have 2764 data from 11,054 data. The code for testing data is mentioned in Fig. 8.

```
[38] X_test.shape,y_test.shape

((2764, 19), (2764,))
```

*Fig. 8. Testing Data*

## 3.4 Phishing Websites Feature Analysis

The main challenge confronted by our investigation was the inaccessibility of dependable preparing datasets. In truth, this challenge faces any analyst within the field. Be that as it may, even though a bounty of articles approximately foreseeing phishing websites utilizing information mining procedures have been spread these days, no solid preparing dataset has been published publicly, possibly since there's no assentation within the writing on the conclusive highlights that characterize phishing websites. Subsequently, it is troublesome to shape a dataset that covers all conceivable highlights.

### 3.5 Feature Selections

Based on the significance of highlights, we chose 19 input highlights from 30 input highlights on a given dataset. These highlights are Using IP, Long URL, Short URL, Image@, Diverting//, Prefix Suffix-, Sub Domains, HTTPS, Domain Reg Len, HTTPS Domain URL, Anchor URL, Website Forwarding, Status Bar Cust, Disable Right Click, Iframe Redirection, Age of Domain, DNS Recording, Website Traffic, Links Pointing To Page.

In this venture, we shed light on the critical highlights that have been demonstrated to be sound and viable in anticipating phishing websites. The highlights that we utilized in this investigative work are depicted in the following passages (Fig.9):

Importance of Column		
10	AnchorURL	3172.030980
7	HTTPS	215.591225
5	PrefixSuffix-	82.691979
8	DomainRegLen	74.931140
18	LinksPointingToPage	47.112919
6	SubDomains	34.201724
15	AgeofDomain	27.615439
0	UsingIP	16.401287
16	DNSRecording	10.324318
2	ShortURL	5.123713
3	Symbol@	3.437305
17	WebsiteTraffic	2.546147
9	HTTPSDomainURL	2.079493
12	StatusBarCust	1.815869
4	Redirecting//	1.665146
1	LongURL	1.524063
11	WebsiteForwarding	0.243742
13	DisableRightClick	0.070163
14	IframeRedirection	0.009517

*Fig. 9. Important Features*

#### 3.5.1 Address Bar-based Features

##### 3.5.1.1 Using the IP Address

On the off chance that an IP address is utilized as an elective of the space title within the URL, such as "http://125.98.3.123/fake.html", clients can be sure that somebody is attempting to take their data. Sometimes, the IP address is even changed into hexadecimal code as appeared within the taking after interface "http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html".

##### 3.5.1.2 Long URL to Hide the Suspicious Part

Phishers can utilize a long URL to stow away the far-fetched portion within the address bar. For case: [http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/?cmd=\\_home&amp;dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@phis hing.website.html](http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/?cmd=_home&amp;dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@phis hing.website.html). To guarantee the exactness of our consideration, we calculated the length of URLs within the dataset and created a normal URL length. The comes about appears that if the length of the URL is more noteworthy than or rises to 54 characters at that point the URL is classified as phishing.

##### 3.5.1.3 Using URL Shortening Services "Tiny URL"

URL shortening could be a strategy on the "World Wide Web" in which a URL may be made impressively smaller in length and still lead to the desired webpage. This can be fulfilled using an "HTTP Redirect" on a brief space title, which joins to the webpage that incorporates a long URL. For illustration, the URL "http://portal.hud.ac.uk/" can be abbreviated to "bit.ly/19DXSk4".

##### 3.5.1.4 URLs having "@" Symbol

When you use the "@" symbol in a URL, the browser will ignore everything that comes before it. The actual address often comes after the "@" symbol.

### 3.5.1.5 Redirecting using “//”

The presence of “//” inside the URL way implies that the client will be diverted to another site. A case of such a URL is “http://www.legitimate.com/http://www.phishing.com”. We look at the area where the “//” shows up. We discover that on the off chance that the URL begins with “HTTP”, meaning the “//” ought to show up within the 6th position. Be that as it may, on the off chance that the URL employs “HTTPS” at that point the “//” ought to show up within the seventh position.

### 3.5.1.6 Adding Prefix or Suffix Separated by (-) to the Domain

The sprint image is seldom utilized in authentic URLs. Phishers tend to include prefixes or additions isolated by (-) to the space title so that clients feel that they are managing an authentic site. For illustration <http://www.Confirme-paypal.com/>.

### 3.5.1.7 Sub Domain and Multi-Sub Domains

Let us accept we have the taking after connect: <http://www.hud.ac.uk/students/>. A space title might incorporate the country-code top-level spaces (ccTLD), which in our illustration is “uk”. The “ac” portion is shorthand for “academic”, the combined “ac.uk” is called an SLD, and “hud” is the real title of the space. To deliver a run the show for extricating this highlight, we to begin with ought to exclude the (www.) from the URL which could be a sub domain in itself. At that point, we have to evacuate the (ccTLD) if it exists. At long last, we number the remaining specks. On the off chance that the number of dabs is more noteworthy than one, then the URL is classified as “Suspicious” since it has one sub domain. However, on the off chance that the dabs are more prominent than two, it is classified as “Phishing” since it'll have numerous sub domains. Something else, on the off chance that the URL has no sub domains, we'll relegate “Legitimate” to the include.

### 3.5.1.8 HTTPS with Secure Sockets Layer

The presence of HTTPS is exceptionally imperative in giving the impression of site authenticity, but typically it is not sufficient. The creators [7] recommend checking the certificate allotted with HTTPS counting the degree of the believed certificate backer, and the certificate age. Certificate Specialists that are reliably recorded among the beat dependable names incorporate: “Geo Trust, Go Daddy, Arrange Arrangements, Thawte, Comodo, Doster and VeriSign”. Besides, by testing out our datasets, we discovered that the least age of a legitimate certificate is two a long time.

### 3.5.1.9 Domain Registration Length

Based on the truth that a phishing site lives for a brief period of time, we accept that reliable spaces are routinely paid for a few a long time in progress. In our dataset, we discover that the longest false spaces have been utilized for one year as it were.

### 3.5.1.10 The Existence of “HTTPS” Token in the Domain Part of the URL

The phishers may add the “HTTPS” token to the domain part of a URL to trick users. For example, <http://https-www-paypal-it-webapps-mpp-home.soft-hair.com/>.

## 3.5.2 Abnormal Based Features

### 3.5.2.1 URL of Anchor

An anchor is an element defined by the <a> tag. This feature is treated exactly as a “Request URL”. However, for this feature, we examine:

1. If the <a> tags and the website have different domain names. This is similar to a request URL feature.
2. If the anchor does not link to any webpage, e.g.:
  - i <a href="#">
  - ii <a href="#content">
  - iii <a href="#skip">
  - iv <a href="JavaScript::void(0)">

## 3.5.3 HTML and Java Script-based Features

### 3.5.3.1 Website Forwarding

The fine line that distinguishes phishing websites from true blue ones is how numerous times the website has been diverted. In our dataset, we discover that true blue websites have been diverted one-time max. On the other hand, phishing websites containing this highlight have been diverted at slightest 4 times.

### 3.5.3.2 Status Bar Customization

Phishers may utilize JavaScript to show fake URLs within the status bar to clients. To extricate this, we must digout the webpage source code, especially the “on Mouse Over” occasion, and check on the off chance that it makes any changes on the status bar.

### 3.5.3.3 Disabling Right Click

Phishers utilize JavaScript to debilitate the right-click work so that clients cannot see and spare the webpage source code. This includes is treated precisely as “Using on Mouse Over to cover up the Link”. In any case, for this, we are going to rummage around for the occasion “event. Button==2” within the webpage source code and check on the off chance that the right-click is impaired.

### 3.5.3.4 Iframe Redirection

The Iframe is an HTML tag utilized to show an extra webpage into one that's right now appearing. Phishers can make utilize of the “iframe” tag and make it undetectable i.e. without outline borders. In this respect, phishers make utilize of the “frame border” trait which causes the browser to render a visual depiction.

## 3.5.4 Domain-based Features

### 3.5.4.1 Age of Domain

This include can be extricated from the WHOIS database (Who is 2005). Most phishing websites live for a brief period of time. By checking on our dataset, we discovered that the least age of the true-blue space is 6 months.

### 3.5.4.2 DNS Record

For phishing websites, either the claimed character isn't recognized by the WHOIS database (Who is 2005) or no records are found for the hostname (Skillet and Ding 2006). In case the DNS record is purged or not found at that point the site is classified as “Phishing”, something else it is classified as “Legitimate”.

### 3.5.4.3 Website Traffic

This includes measuring the ubiquity of the site by deciding the number of guests and the number of pages they visit. Be that as it may, since phishing websites live for a brief period of time, they may not be recognized by the Alexa database (Alexa the Net Data Company., 1996). By looking into our dataset, we discover that in most noticeably awful scenarios, true blue websites are positioned among the beat 100,000. Moreover, on the off chance that the domain has no activity or isn't recognized by the Alexa database, it is classified as “Phishing”. Something else is classified as “Suspicious”.

### 3.5.4.4 Number of Links Pointing to Page

The number of joins indicated on the webpage demonstrates its authenticity level, indeed if a few joins are of the same space. In our datasets and due to their brief life span, we discover that 98% of phishing datasets things have no joins indicating to them. On the other hand, authentic websites have at slightest 2 outside joins indicating to them.

## 3.6 Registration

To use the website, a user must register for access.

## 3.7 Login

Following a successful registration, the user can log in to the system by entering his login credentials. The administrator can log into the system using the admin panel.

## 3.8 Add to Blacklist

Here, the framework director includes the malevolent site in the boycott. Even though there does not exist a framework that can identify all the phishing websites, utilizing this includes making a proficient way to distinguish the phishing sites.

## 3.9 Check Website

Here, the user enters the URL to check for a phishing website.

### 3.10 View Profile

The client can see his profile points of interest from the client side. An admin can see the client's list and subtle elements from the admin board.

### 3.11 Update Profile

Clients can overhaul their profile (name, mail, profile picture, password) from the client side. An admin moreover can upgrade his data from the admin board.

### 3.12 Feedback

A user will receive feedback based on the conditions of the website.

### 3.13 Select the Right ML Algorithms

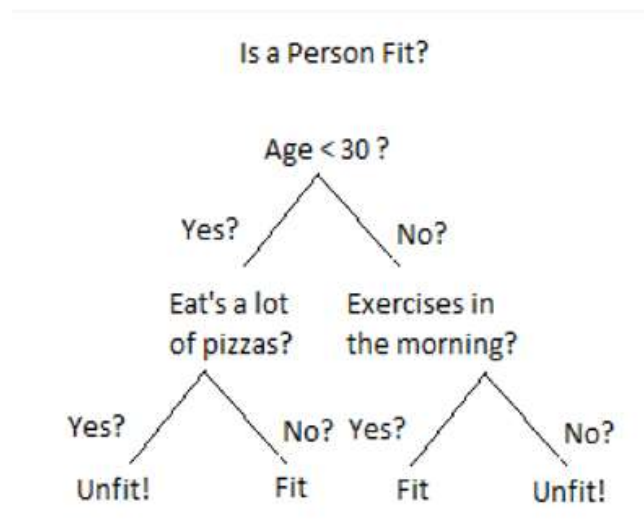
- DT
- RF
- XG Boost
- SVM
- KNN
- NB Classifiers

### 3.14 Classifier Algorithms

This work utilized the over-information set to compare the execution of Six classifiers. Particularly, we utilized the choice tree, Arbitrary Woodland classifier, Back Vector Machine, SVM, KNN, Credulous Bayes, and XG Boost to classify the URLs within the data set, and after that, we compared the comes about utilizing perplexity networks.

#### 3.14.1 Decision Tree Algorithm

Choice trees are non-parametric classifiers. As its title shows, a choice tree may be a tree structure, where each non-terminal hub indicates a test on a trait, each department speaks to the result of the test, and the leaf hubs signify classes. The essential calculation for choice tree acceptance could be an eager calculation that builds the choice tree in a top-down recursive divide-and-conquer way. Choice Trees are a sort of Administered ML (that clarifies what the input is and what the comparing yield is within the preparing information) where the information is persistently part agreeing to a certain parameter. The tree can be clarified by two substances, to be specific choice hubs and takes off. The clears out are the choices or results. And the choice hubs are where the information is part. It is represented in Fig 10.



**Fig. 10.** Decision tree algorithm

A case of a choice tree can be clarified utilizing the over-the-parallel tree. Let's say you need to foresee whether an individual is fit given their data like age, eating propensities, physical movement, etc. The choice hubs here are questions like 'What's the age?', 'Does he exercise?', 'Does he eat a parcel of

pizzas'? And the takes off, which are results like either 'fit', or 'unfit'. In this case, this was a double classification issue (a yes-no sort issue).

### 3.14.1.1 Decision Tree Working Procedure

Presently that we know what a Choice Tree is, we'll see how it works inside. There are numerous calculations out there that build Choice Trees, but one of the finest is called an ID3 Calculation.

**Entropy:** Entropy, moreover called Shannon Entropy is signified by  $H(S)$  for a limited set  $S$ , which is the degree of the sum of instability or haphazardness in information.

$$H(S) = -\sum p(x) \log_2 p(x)$$

Naturally, it tells us almost the consistency of a certain occasion. For the case, consider a coin hurl whose likelihood of heads is 0.5 and the likelihood of tails is 0.5. Here the entropy is the most noteworthy conceivable since there's no way of deciding what the result can be. On the other hand, consider a coin that has heads on both sides, the entropy of such an occasion can be anticipated impeccably since we know in advance that it'll continuously be executed. In other words, this occasion has no haphazardness consequently its entropy is zero. In specific, lower values infer less vulnerability whereas higher values infer greater instability.

### 3.14.1.2 Information Gain

Information gain is also called Kullback-Leibler divergence denoted by  $IG(S, A)$  for a set  $S$  is the effective change in entropy after deciding on a particular attribute  $A$ . It measures the relative change in entropy concerning the independent variables.

$$IG(S, A) = H(S) - H(S|A)$$

Alternatively,

$$IG(S, A) = H(S) - \sum_{x \in S} P(x) H(x|A)$$

where  $IG(S, A)$  is the information gain by applying feature  $A$ .  $H(S)$  is the Entropy of the entire set, while the second term calculates the Entropy after applying the feature  $A$ , where  $P(x)$  is the probability of event  $x$ .

ID3 Algorithm will perform the following tasks recursively.

1. It begins with the original set  $S$  as the root node.
2. On each iteration of the algorithm, it iterates through the very unused attribute of the set  $S$  and calculates the Entropy ( $H$ ) and Information gain ( $IG$ ) of this attribute.
3. It then selects the attribute that has the smallest Entropy or Largest Information gain.
4. The set  $S$  is then split by the selected attribute to produce a subset of the data.
5. The algorithm continues to recur on each subset, considering only attributes never selected before.

### 3.14.1.3 Advantages

- Compared to other algorithms, decision trees require less effort for data preparation during pre-processing.
- A DT does not require the normalization of data.
- A DT does not require scaling of data as well.
- Missing values in the data also do NOT affect the process of building a decision tree to any considerable extent.
- A Decision tree model is very intuitive and easy to explain to technical teams as well as stakeholders.

### 3.14.1.4 Disadvantage

- A small change in the data can cause a large change in the structure of the DT causing instability.
- For a DT sometimes calculation can be far more complex compared to other algorithms.
- DT often involves more time to train the model.
- DT training is relatively expensive as the complexity and time taken are greater.
- The DT algorithm is inadequate for applying regression and predicting continuous values.

### 3.14.2 Random Forest

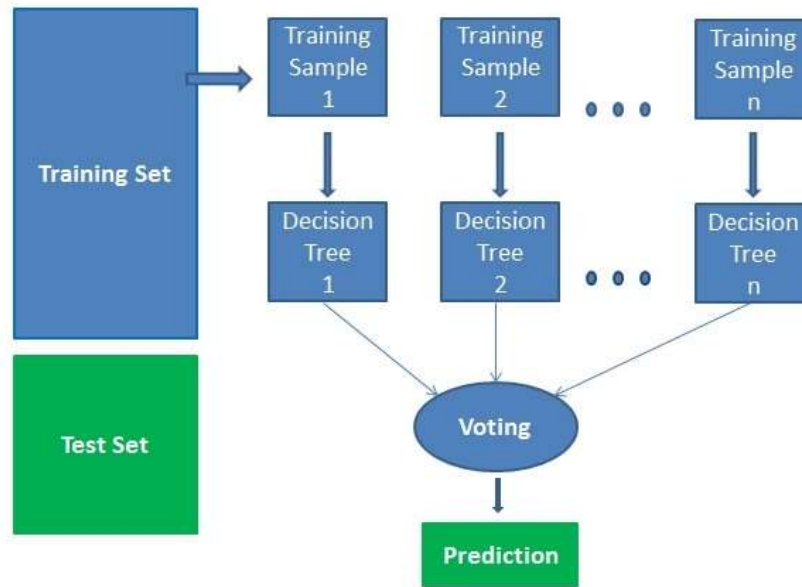
RF is a supervised learning algorithm that is used for both classifications as well as regression. However, it is mainly used for classification problems. As we know a forest is made up of trees and more trees mean more robust forests. Similarly, the RF algorithm creates decision trees on data samples then gets the prediction from each of them, and finally selects the best solution using voting. It is an ensemble

method that is better than a single decision tree because it reduces the over-fitting by averaging the result.

### 3.14.2.1 Working of Random Forest Algorithm

We can understand the workings of the RF algorithm with the help of the following steps. Fig 11 represents the block diagram of RF.

- Step 1 – First, start with the selection of random samples from a given dataset.
- Step 2 – Next, this algorithm will construct a DT for every sample. Then it will get the prediction result from every DT.
- Step 3 – In this step, voting will be performed for every predicted result.
- Step 4 – At last, select the most voted prediction result as the final prediction result.



*Fig. 11. Block diagram of RF*

### 3.14.2.2 Advantages of Random Forest

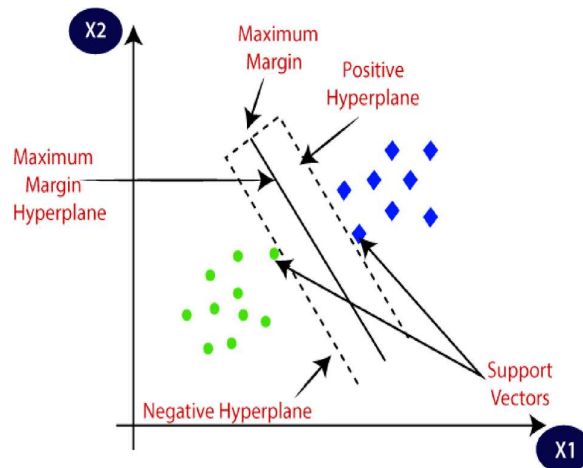
- It overcomes the problem of overfitting by averaging or combining the results of different DTs.
- RF works well for a larger range of data items than a single DT does.
- The RF has less variance than a single DT.
- RF is very flexible and possesses very high accuracy.
- Scaling of data does not require an RF algorithm. It maintains good accuracy even after providing data without scaling.
- RF algorithms maintain good accuracy even if a large proportion of the data is missing.

### 3.14.2.3 Disadvantages of Random Forest

- Complexity is the main disadvantage of RF algorithms.
- Construction of RF is much harder and time-consuming than DT.
- More computational resources are required to implement the RF algorithm.
- It is less intuitive in the case when we have a large collection of DT.
- The prediction process using RF is very time-consuming in comparison with other algorithms.

### 3.14.3 Support Vector Machine (SVM)

Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane. These are the points that help us build our SVM. Fig 12 represents the SVM.

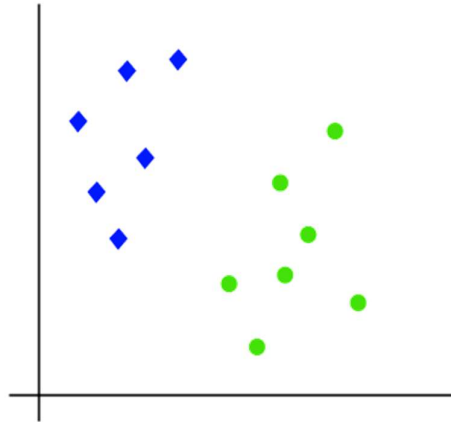


*Fig. 12. Support vector machine*

### 3.14.3.1 SVM Working Procedure

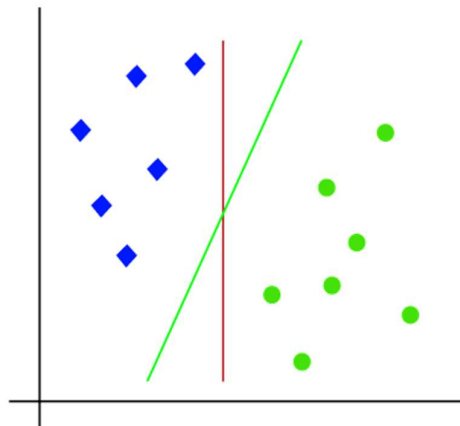
#### Linear SVM

The working of the SVM algorithm can be understood by using an example. Suppose we have a dataset that has two tags (green and blue), and the dataset has two features  $x_1$  and  $x_2$ . We want a classifier that can classify the pair( $x_1$ ,  $x_2$ ) of coordinates in either green or blue. Consider the below image (Fig 13):



*Fig. 13. Linear SVM*

So, as it is 2-d space by just using a straight line, we can easily separate these two classes. But there can be multiple lines that can separate these classes. Consider the below image (Fig 14):



*Fig. 14. Linear SVM*

Hence, the SVM algorithm helps to find the best line or decision boundary; this best boundary or region is called a hyper plane. The SVM algorithm finds the closest point of the lines from both classes. These points are called support vectors. The distance between the vectors and the hyperplane is called a

margin. The goal of SVM is to maximize this margin. The hyperplane with the maximum margin is called the optimal hyperplane.

### 3.14.3.2 Advantages

1. SVM works relatively well when there is a clear margin of separation between classes.
2. SVM is more effective in high-dimensional spaces.
3. SVM is effective in cases where the number of dimensions is greater than the number of samples.
4. SVM is relatively memory efficient.

### 3.14.3.3 Disadvantages

1. The SVM algorithm is not suitable for large data sets.
2. SVM does not perform very well when the data set has more noise i.e., target classes are overlapping.
3. In cases where the number of features for each data point exceeds the number of training data samples, the SVM will underperform.
4. As the support vector classifier works by putting data points, above and below the classifying hyperplane there is no probabilistic explanation for the classification.

### 3.14.4 XG Boost

Gradient boosting is an ML technique for regression and classification problems. That produces a prediction model in the form of an ensemble of weak prediction models.

The accuracy of a predictive model can be boosted in two ways:

1. Either by embracing feature engineering.
2. By applying boosting algorithms straight away.

There are many boosting algorithms like

- Gradient Boosting
- XG Boost
- Ada Boost
- Gentle Boost etc.

Every boosting algorithm has its underlying mathematics. Also, a slight variation is observed while applying them.

XG Boost or Extreme Gradient Boosting is a perfect combination of software and hardware optimization techniques to generate superior results using fewer computing resources in the shortest amount of time.

In short, we can use XG Boost to achieve two goals of the project:

1. Execution Speed
2. Model Performance

### 3.14.5 K-Nearest Neighbours (KNN)

- K-Nearest Neighbour is one of the simplest ML algorithms based on the Supervised Learning technique.
- KNN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories.
- The KNN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a good suite category by using the KNN algorithm.
- KNN algorithm can be used for Regression as well as for Classification but mostly it is used for Classification problems.
- KNN is a non-parametric algorithm, which means it does not make any assumptions on underlying data.
- It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.
- Example: Suppose, we have an image of a creature that looks similar to a cat and a dog, but we want to know whether it is a cat or a dog. So, for this identification, we can use the KNN

algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dog's images and based on the most similar features it will put it in either the cat or dog category.

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point  $x_1$ , so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset. Consider the below diagram (Fig 15):

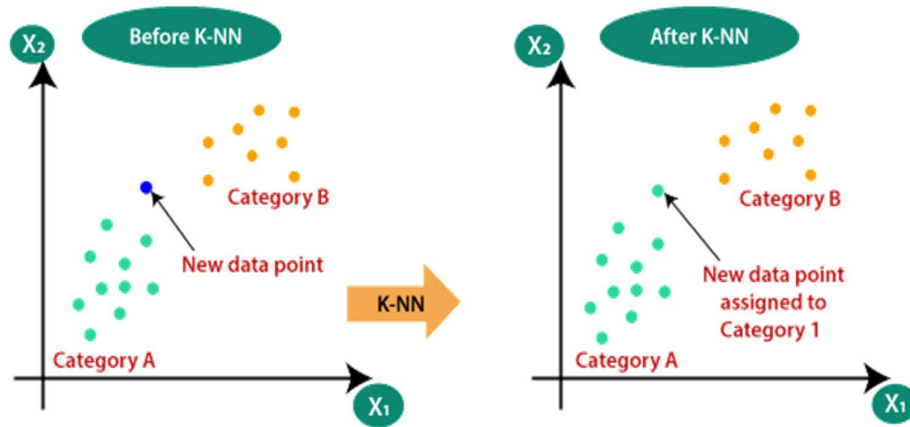


Fig. 15. K-Nearest Neighbour

### 3.14.5.1 KNN Working Procedure

The K-NN working can be explained based on the below algorithm:

- Step1: Select the number K of the neighbors.
- Step2: Calculate the Euclidean distance of the K number of neighbors.
- Step3: Take the K nearest neighbors as per the calculated Euclidean distance.
- Step4: Among these k neighbors, count the number of the data points in each category.
- Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.
- Step6: Our model is ready.

Suppose we have a new data point and we need to put it in the required category. Consider the below image (Fig 16):

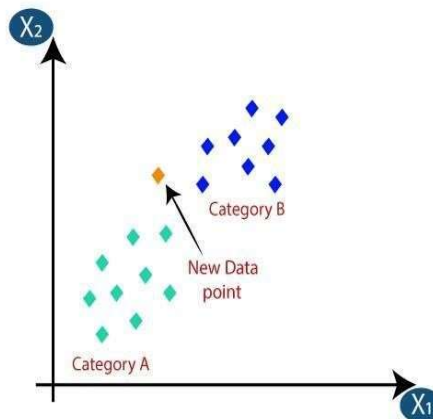
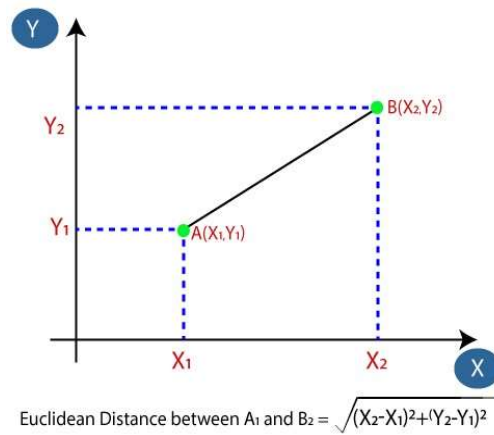


Fig. 16. K-Nearest Neighbour

- Firstly, we will choose the number of neighbors, so we will choose  $k=5$ .
- Next, we will calculate the Euclidean distance between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. It can be calculated as (Fig 17):

**Fig. 17. K-Nearest Neighbour**

- By calculating the Euclidean distance we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B. Consider the below image (Fig 18):

**Fig. 18. K-Nearest Neighbour**

- As we can see the 3 nearest neighbours are from category A, hence this new data point must belong to category A.

### 3.14.5.2 The Value of K in the K-NN Algorithm

Below are some points to remember while selecting the value of K in the K-NN algorithm:

- There is no particular way to determine the best value for "K", so we need to try some values to find the best out of them. The most preferred value for K is 5.
- A very low value for K such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model.
- Large values for K are good, but they may find some difficulties.

### 3.14.5.3 Advantages

- The algorithm is simple and easy to implement.
- There's no need to build a model, tune several parameters, or make additional assumptions.
- The algorithm is versatile. It can be used for classification, regression, and search (as we will see in the next section).

### 3.14.5.4 Disadvantages

- Always needs to determine the value of K which may be complex sometimes.
- The computation cost is high because of calculating the distance between the data points for all the training samples.

### 3.14.6 Naive Bayes (NB)

Naive Bayes is an ML model that is used for large volumes of data, even if you are working with data that has millions of data records the recommended approach is NB. It gives very good results when it comes to NLP tasks such as sentimental analysis. It is a fast and uncomplicated classification algorithm.

- The NB algorithm is a supervised learning algorithm, which is based on the Bayes theorem and used for solving classification problems.
- It is mainly used in text classification that includes a high-dimensional training dataset.
- NB Classifier is one of the simplest and most effective Classification algorithms that help in building fast ML models that can make quick predictions.
- It is a probabilistic classifier, which means it predicts based on the probability of an object.
- Some popular examples of NB Algorithms are spam filtration, Sentiment analysis, and classifying articles.

The NB algorithm is comprised of two words Naïve and Bayes, which can be described as:

**Naïve:** It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified based on color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identifying that it is an apple without depending on each other.

**Bayes:** It is called Bayes because it depends on the principle of **Bayes' Theorem**.

To understand the naive Bayes classifier, we need to understand the Bayes theorem. So let's first discuss the Bayes Theorem.

#### 3.14.6.1 Bayes Theorem

It is a theorem that works on conditional probability. Conditional probability is the probability that something will happen, given that something else has already occurred. The conditional probability can give us the probability of an event using its prior knowledge. The formula for Bayes' theorem is given as:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

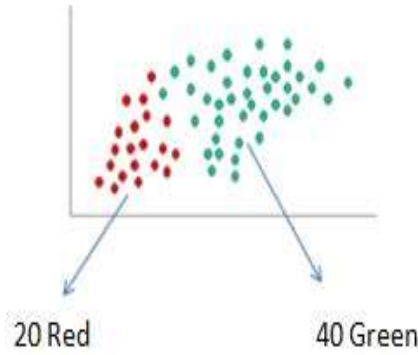
Where,

- $P(c|x)$  is the posterior probability of class (c, target) given predictor (x, attributes).
- $P(c)$  is the prior probability of class.
- $P(x|c)$  is the likelihood which is the probability of a predictor given class.
- $P(x)$  is the prior probability of the predictor.

Therefore, the above equation can be rewritten as:

$$\text{Posterior} = \frac{\text{Prior} \times \text{Likelihood}}{\text{Evidence}}$$

Below is one simple way to explain the Bayes rule. The task is to identify the color of a newly-observed dot.



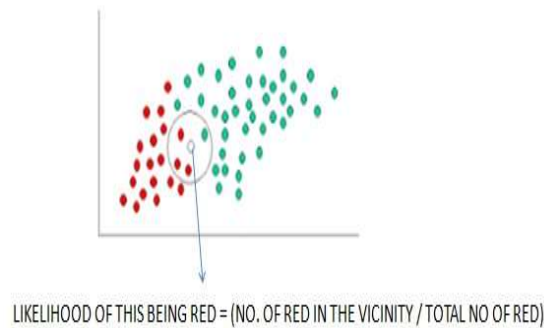
Since there are twice as many GREEN objects as RED, it is reasonable to believe that a new case (which hasn't been observed yet) is twice as likely to have a membership with GREEN rather than RED. In the Bayesian analysis, this belief is known as the prior probability. Prior probabilities are based on previous experience, in this case, the percentage of GREEN and RED objects, and are often used to predict outcomes before they happen.

Since there is a total of 60 objects, 40 of which are GREEN and 20 RED, our prior probabilities for class members can be written as below:

$$\text{Prior Probability of GREEN} = \frac{\text{Number of GREEN objects}}{\text{Total number of objects}} = \frac{40}{60}$$

$$\text{Prior Probability of RED} = \frac{\text{Number of RED objects}}{\text{Total number of objects}} = \frac{20}{60}$$

Having formulated our prior probability, we are now ready to classify a new object (WHITE circle in the diagram below). Since the objects are well clustered, it is reasonable to assume that the GREENER (or RED) objects in the vicinity of X, the more likely that the new cases belong to that particular color. To measure this likelihood, we draw a circle around X which encompasses a number (to be chosen a priori) of points irrespective of their class labels. Then we calculate the number of points in the circle belonging to each class label. From this, we calculate the likelihood:



From the illustration above, it is clear that the likelihood of X given GREEN is smaller than the Likelihood of X given RED since the circle encompasses 1 GREEN object and 3 RED ones.

Although the prior probabilities indicate that X may belong to GREEN (given that there are twice as many GREENER compared to RED) the likelihood indicates otherwise; that the class membership of X is RED (given that there are more RED objects in the vicinity of X than GREEN). In the Bayesian analysis, the final classification is produced by combining both sources of information, i.e., the prior and the likelihood, to form a posterior probability using Bayes' rule.

$$\text{Posterior Probability of GREEN} = \text{Prior Probability of GREEN} \times \text{Likelihood of GREEN} = \frac{40}{60} \times \frac{1}{40}$$

$$\text{Posterior Probability of RED} = \text{Prior Probability of RED} \times \text{Likelihood of RED} = \frac{20}{60} \times \frac{3}{20}$$

Finally, we classify X as RED since its class membership achieves the largest posterior probability.

### 3.14.6.2 Naive Bayes Classifier

- It is a kind of classifier that works on Bayes theorem.

- Prediction of membership probabilities is made for every class such as the probability of data points associated with a particular class.
- The class having maximum probability is appraised as the most suitable class.
- This is also referred to as MAP.
- The MAP for a hypothesis is:

$$\begin{aligned} MAP(H) &= \max ((H | E)) \\ MAP(H) &= \max ((H | E) * (P(H)) / P(E)) \\ MAP(H) &= \max (P(E | H) * P(H)) \end{aligned}$$

- $P(E)$  is evidence probability, and it is used to normalize the result. The result will not be affected by removing  $(E)$ .
- NB classifiers conclude that all the variables or features are not related to each other.
- The Existence or absence of a variable does not impact the existence or absence of any other variable.
- Example: Fruit may be observed to be an apple if it is red, round, and about 4" in diameter.
- In this case also even if all the features are interrelated to each other, an NB classifier will observe all of these independently contributing to the probability that the fruit is an apple.
- We experiment with the hypothesis in real datasets, given multiple features.
- So, computation becomes complex.

### 3.14.6.3 Types of Naive Bayes Algorithms

1. **Gaussian Naïve Bayes:** When characteristic values are continuous then an assumption is made that the values linked with each class are dispersed according to Gaussian which is Normal Distribution.
2. **Multinomial Naïve Bayes:** Multinomial NB is favored to use on data that is multinomial distributed. It is widely used in text classification in NLP. Each event in text classification constitutes the presence of a word in a document.
3. **Bernoulli Naïve Bayes:** When data is dispensed according to the multivariate Bernoulli distributions then Bernoulli NB is used. That means there exist multiple features but each one is assumed to contain a binary value. So, it requires features to be binary-valued.

### 3.14.6.4 Advantages of Naïve Bayes Classifier

- NB is one of the fastest and easiest ML algorithms to predict a class of datasets.
- It can be used for Binary as well as Multi-class Classifications.
- It performs well in multi-class predictions as compared to the other Algorithms.
- It is the most popular choice for text classification problems.

### 3.14.6.5 Disadvantages of Naïve Bayes Classifier

- NB assumes that all features are independent or unrelated, so it cannot learn the relationship between features.

### 3.14.6.6 Applications of Naïve Bayes Classifier

- **Real-time Prediction:** NB is an eager learning classifier and it is sure fast. Thus, it could be used for making predictions in real time.
- **Multi-class Prediction:** This algorithm is also well known for multi-class prediction features. Here we can predict the probability of multiple classes of a target variable.
- **Text classification/ Spam Filtering/ Sentiment Analysis:** NB classifiers mostly used in text classification (due to better results in multi-class problems and independence rule) have a higher success rate as compared to other algorithms. As a result, it is widely used in Spam filtering (identify spam email) and Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments).
- **Recommendation System:** NB Classifier and Collaborative Filtering together build a recommendation System that uses ML and data mining techniques to filter unseen information and predict whether a user would like a given resource or not.

### 3.15 Google-Colab

We using Google Colab. Collaboratory, or "Colab" for short, allows one to write and execute code in the browser, with

- Zero configuration required
- Free access to GPUs
- Easy sharing

For a student, a data scientist, or an AI researcher, Colab can make work easier.

#### 3.15.1 Advantage

- The best thing is that Google Colab is free.
- To match the industry standards or running time and memory-consuming codes, it is not possible to install costly GPUs on personal machines. To the best of my knowledge, nowadays, using Colab, each notebook can use a free GPU in the Colab environment, independent of what the personal machine is using.
- Google Colab provides an inbuilt version controlling system using Git and it is quite easy to create notes and documentation, including figures, and tables using Markdown.
- As the name suggests, Google Colab comes with collaboration backed by the product. It is a Jupyter Notebook that leverages Google Docs collaboration features.
- It also runs on Google servers using virtual machines and no need to install any packages, which sometimes creates difficulty using different operating systems such as MAC, Windows, and Linux.
- Last but not least, nowadays in industry, the most common requirement in the job packs is knowledge of programming deep learning models using GPUs.
- Link with GitHub profiles. The GitHub links are useful for recruiters. Easily anyone can showcase their work from anywhere around the globe and recruiters can explore the work done by others, which helps them in identifying strong candidates for the jobs.

#### 3.15.2 Disadvantage

- All Collaborative notebooks must be stored in Google Drive — so need to log into a Google account before accessing the tool.
- Long-running background computations may be stopped — “run continuous or long-running computations through Collaboratory’s UI to use a local runtime”.
- Users need to install all specific libraries that do not come with standard Python (and need to repeat this with every session).
- Google Storage is used with the user’s current session, so if users have downloaded a file and want to use it later, they had better save it before closing the session.
- It can be difficult (and potentially costly) to work with bigger datasets as users have to download and store them in Google Drive (only 15GB is free in Google Drive).

## 4. Experimental Analysis & Result

As the phishing URL discovery issue is a parallel classification issue, each URL falls into one of four conceivable categories: genuine positive (TP, accurately classified phishing URL), genuine negative (TN, accurately classified non-phishing URL), wrong positive (FP, non-phishing URL wrongly classified as phishing), and wrong negative (FN, phishing URL wrongly classified as non-phishing). Standard measures such as wrong positive rate (FPR), false-negative rate (FNR), exactness, review, and F-measure were decided utilizing the taking-after conditions:

$$FPR = \frac{|FP|}{Total\ legitimate\ URLs}$$

$$FNR = \frac{|FP|}{Total\ Legitimate\ URLs}$$

$$Precision = \frac{|TP|}{|TP| + |FP|}$$

$$Recall = \frac{|TP|}{|TP| + |FN|}$$

$$F1-Score = \frac{2 * precision * Recall}{Precision + Recall}$$

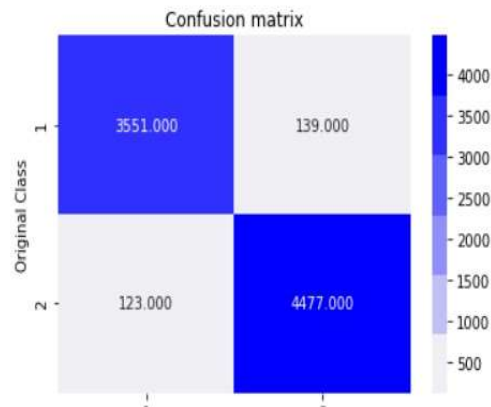
## 4.1 Decision Tree

### 4.1.1 Confusion Matrix Report Generate

For Training Data:

Confusion matrices report on Training data for Decision Tree :

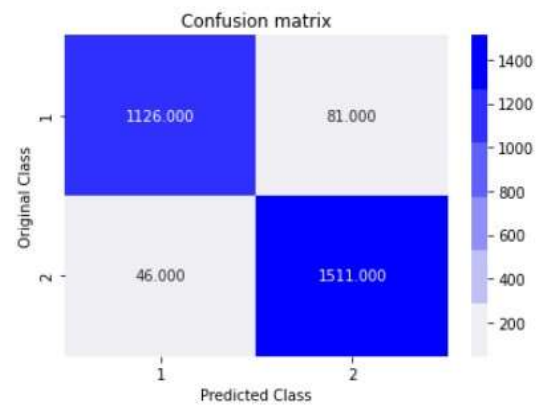
True Positive = 3551  
True Negative = 139  
False Negative = 123  
False Positive = 4477



For Testing data:

Confusion matrices report on Testing data for Decision Tree :

True Positive = 1126  
True Negative = 81  
False Negative = 46  
False Positive = 1511



### 4.1.2 Classification Report Generate

Classification report on training data for Decision Tree :

	precision	recall	f1-score	support
0	0.97	0.96	0.96	3690
1	0.97	0.97	0.97	4600
accuracy			0.97	8290
macro avg	0.97	0.97	0.97	8290
weighted avg	0.97	0.97	0.97	8290

Classification report on testing data for Decision Tree :

	precision	recall	f1-score	support
0	0.96	0.93	0.95	1207
1	0.95	0.97	0.96	1557
accuracy			0.95	2764
macro avg	0.95	0.95	0.95	2764
weighted avg	0.95	0.95	0.95	2764

### Performance Evaluation:

```
[91] #computing the accuracy of the model performance for Decision Tree
      acc_train_tree = accuracy_score(y_train,y_train_tree)
      acc_test_tree = accuracy_score(y_test,y_test_tree)

      print("Decision Tree(ID3): Accuracy on training Data: {:.3f}".format(acc_train_tree))
      print("Decision Tree(ID3): Accuracy on test Data: {:.3f}".format(acc_test_tree))
```

Decision Tree(ID3): Accuracy on training Data: 0.968

Decision Tree(ID3): Accuracy on test Data: 0.954

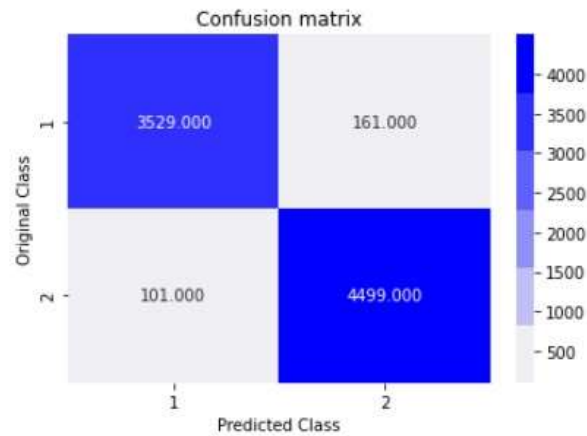
## 4.2 Random Forest

### 4.2.1 Confusion Matrix Report Generate

For Training Data:

Confusion matrices report on Training data For Random forest:

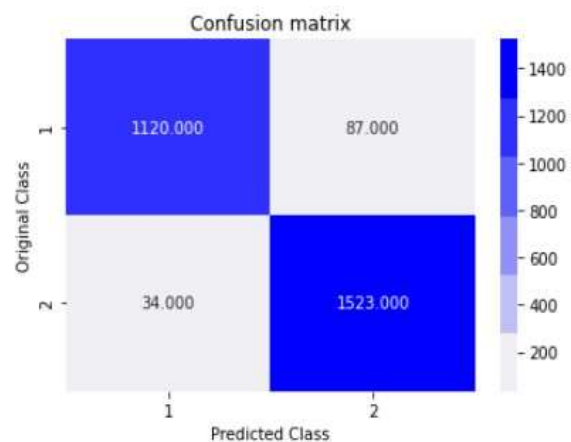
True Positive = 3529  
True Negative = 161  
False Negative = 101  
False Positive = 4499



For Testing data:

Confusion matrices report on Training data For Random forest :

True Positive = 1120  
True Negative = 87  
False Negative = 34  
False Positive = 1523



### 4.2.2 Classification Report Generate

Classification report on training data for Random Forest :

	precision	recall	f1-score	support
0	0.97	0.96	0.96	3690
1	0.97	0.98	0.97	4600
accuracy			0.97	8290
macro avg	0.97	0.97	0.97	8290
weighted avg	0.97	0.97	0.97	8290

Classification report on testing data for Random Forest :

	precision	recall	f1-score	support
0	0.97	0.93	0.95	1207
1	0.95	0.98	0.96	1557
accuracy			0.96	2764
macro avg	0.96	0.95	0.96	2764
weighted avg	0.96	0.96	0.96	2764

### Performance Evaluation:

```
[53] #computing the accuracy of the model performance for Random forest
acc_train_forest = accuracy_score(y_train,y_train_forest)
acc_test_forest = accuracy_score(y_test,y_test_forest)

print("Random forest: Accuracy on training Data: {:.3f}".format(acc_train_forest))
print("Random forest: Accuracy on test Data: {:.3f}".format(acc_test_forest))
```

Random forest: Accuracy on training Data: 0.968

Random forest: Accuracy on test Data: 0.956

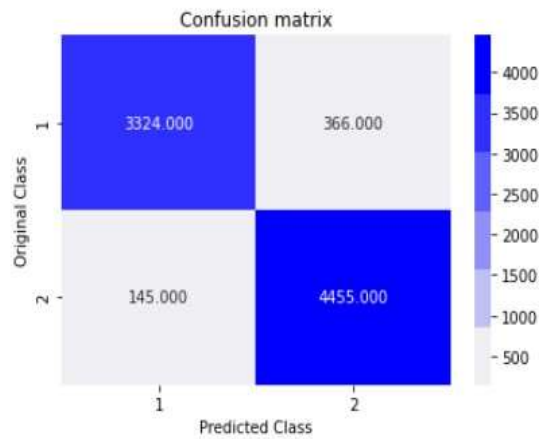
### 4.3 K-Nearest Neighbours

#### 4.3.1 Confusion Matrix Report Generate

For Training Data:

Confusion matrices report on Training data For K-Nearest Neighbors:

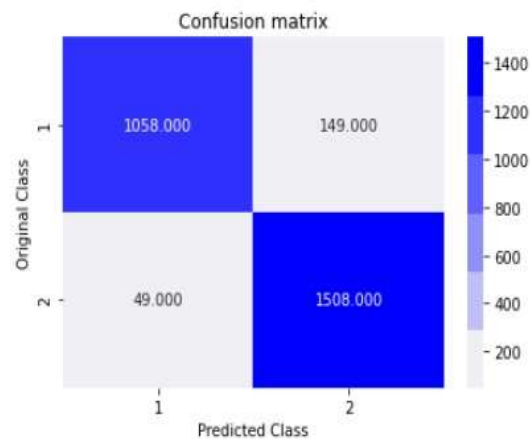
True Positive = 3324  
True Negative = 366  
False Negative = 145  
False Positive = 4455



For Testing data:

Confusion matrices report on Training data for K-Nearest Neighbors :

True Positive = 1058  
True Negative = 149  
False Negative = 49  
False Positive = 1508



### 4.3.2 Classification Report Generate

Classification report on training data for K-Nearest Neighbors :

	precision	recall	f1-score	support
0	0.96	0.90	0.93	3690
1	0.92	0.97	0.95	4600
accuracy			0.94	8290
macro avg	0.94	0.93	0.94	8290
weighted avg	0.94	0.94	0.94	8290

Classification report on testing data for K-Nearest Neighbors :

	precision	recall	f1-score	support
0	0.96	0.88	0.91	1207
1	0.91	0.97	0.94	1557
accuracy			0.93	2764
macro avg	0.93	0.92	0.93	2764
weighted avg	0.93	0.93	0.93	2764

#### Performance Evaluation:

```
[62] #computing the accuracy of the model performance for K-Nearest Neighbors
acc_train_knn = accuracy_score(y_train,y_train_knn)
acc_test_knn = accuracy_score(y_test,y_test_knn)

print("K-Nearest Neighbors: Accuracy on training Data: {:.3f}".format(acc_train_knn))
print("K-Nearest Neighbors: Accuracy on test Data: {:.3f}".format(acc_test_knn))
```

K-Nearest Neighbors: Accuracy on training Data: 0.938

K-Nearest Neighbors: Accuracy on test Data: 0.928

#### Performance Evaluation:

```
[70] #computing the accuracy of the model performance for SVM
acc_train_svm = accuracy_score(y_train,y_train_svm)
acc_test_svm = accuracy_score(y_test,y_test_svm)

print("SVM: Accuracy on training Data: {:.3f}".format(acc_train_svm))
print("SVM : Accuracy on test Data: {:.3f}".format(acc_test_svm))
```

SVM: Accuracy on training Data: 0.949

SVM : Accuracy on test Data: 0.938

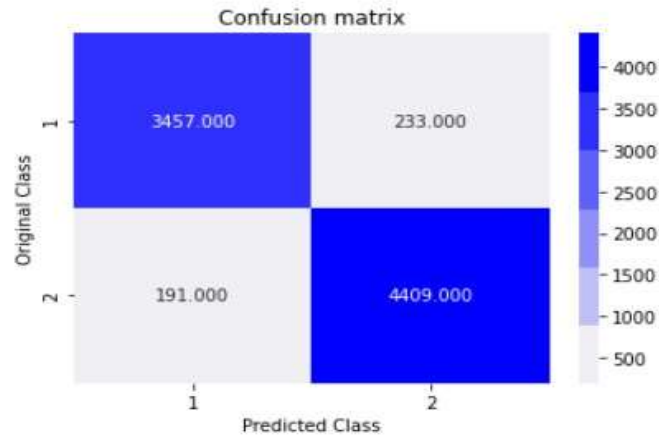
## 4.4 Support Vector Machines

### 4.4.1 Confusion Matrix Report Generate

For Training Data:

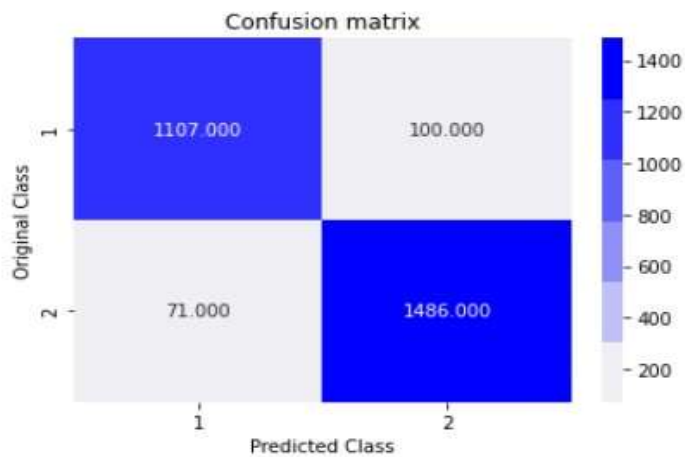
Confusion matrices report on Training data For SVM:

```
True Positive = 3457
True Negative = 233
False Negative = 191
False Positive = 4409
```



Confusion matrices report on Testing data For SVM:

```
True Positive = 1107
True Negative = 100
False Negative = 71
False Positive = 1486
```



**4.4.2 Classification Report Generate**

Classification report on training data for SVM :

	precision	recall	f1-score	support
0	0.95	0.94	0.94	3690
1	0.95	0.96	0.95	4600
accuracy			0.95	8290
macro avg	0.95	0.95	0.95	8290
weighted avg	0.95	0.95	0.95	8290

Classification report on testing data for SVM :

	precision	recall	f1-score	support
0	0.94	0.92	0.93	1207
1	0.94	0.95	0.95	1557
accuracy			0.94	2764
macro avg	0.94	0.94	0.94	2764
weighted avg	0.94	0.94	0.94	2764

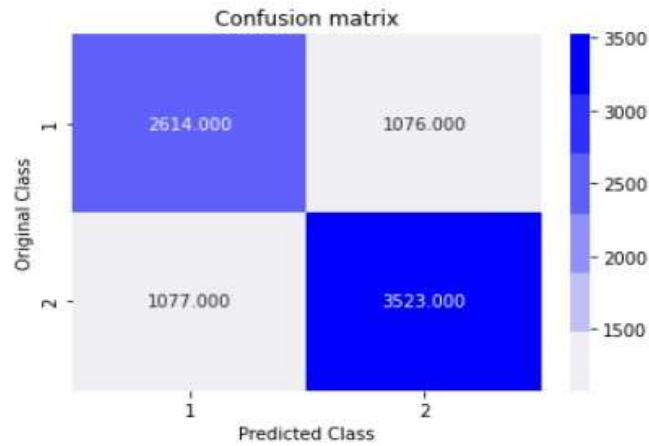
## 4.5 Extreme Gradient Boosting

### 4.5.1 Confusion Matrix Report Generate

For Training Data:

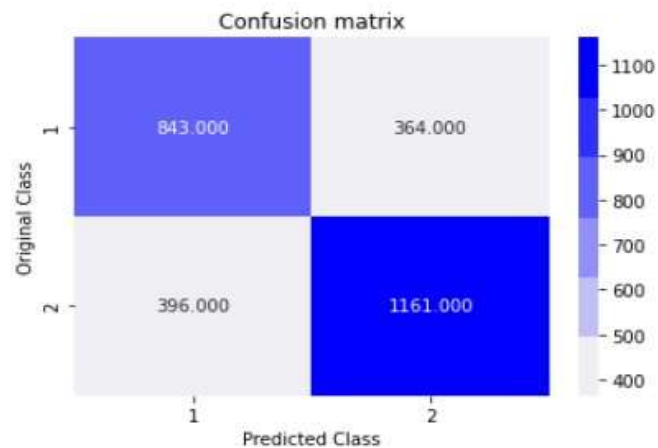
Confusion matrices report on Training data For XGBoost:

```
True Positive = 2614
True Negative = 1076
False Negative = 1077
False Positive = 3523
```



Confusion matrices report on Training data for XGBoost:

```
True Positive = 843
True Negative = 364
False Negative = 396
False Positive = 1161
```



## 4.5.2 Classification Report Generate

### Performance Evaluation:

```
[58] #computing the accuracy of the model performance for XGBoost
acc_train_xgb = accuracy_score(y_train,y_train_xgb)
acc_test_xgb = accuracy_score(y_test,y_test_xgb)

print("XGBoost: Accuracy on training Data: {:.3f}".format(acc_train_xgb))
print("XGBoost : Accuracy on test Data: {:.3f}".format(acc_test_xgb))
```

XGBoost: Accuracy on training Data: 0.740

XGBoost : Accuracy on test Data: 0.725

Classification report on training data for XGBoost :

	precision	recall	f1-score	support
0	0.71	0.71	0.71	3690
1	0.77	0.77	0.77	4600
accuracy			0.74	8290
macro avg	0.74	0.74	0.74	8290
weighted avg	0.74	0.74	0.74	8290

Classification report on testing data for XGBoost :

	precision	recall	f1-score	support
0	0.68	0.70	0.69	1207
1	0.76	0.75	0.75	1557
accuracy			0.73	2764
macro avg	0.72	0.72	0.72	2764
weighted avg	0.73	0.73	0.73	2764

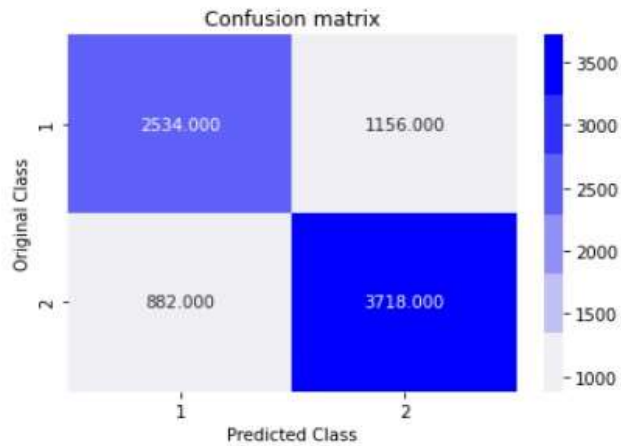
## 4.6 Naive Bayes

### 4.6.1 Confusion Matrix Report Generate

For Training Data:

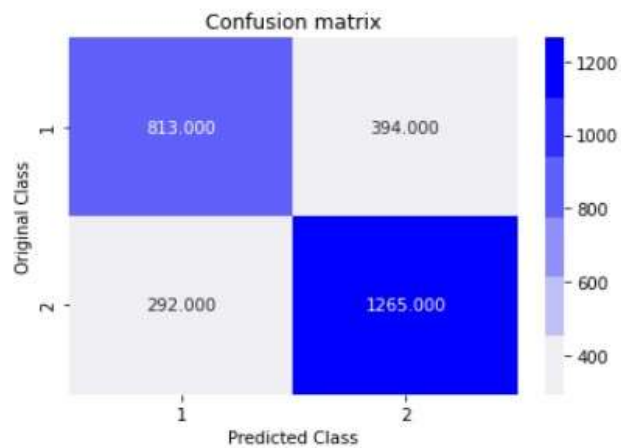
Confusion matrices report on Training data For Naive Bayes:

True Positive = 2534  
 True Negative = 1156  
 False Negative = 882  
 False Positive = 3718



Confusion matrices report on Training data For Naive Bayes

True Positive = 813  
 True Negative = 394  
 False Negative = 292  
 False Positive = 1265



## 4.6.2 Classification Report Generate

### Performance Evaluation:

```
[92] #computing the accuracy of the model performance for MultinomialNB
acc_train_NB = accuracy_score(y_train,y_train_NB)
acc_test_NB = accuracy_score(y_test,y_test_NB)

print("Naive Bayes: Accuracy on training Data: {:.3f}".format(acc_train_NB))
print("Naive Bayes: Accuracy on test Data: {:.3f}".format(acc_test_NB))
```

Naive Bayes: Accuracy on training Data: 0.754

Naive Bayes: Accuracy on test Data: 0.752

Classification report on training data for Naive Bayes :

	precision	recall	f1-score	support
0	0.74	0.69	0.71	3690
1	0.76	0.81	0.78	4600
accuracy			0.75	8290
macro avg	0.75	0.75	0.75	8290
weighted avg	0.75	0.75	0.75	8290

Classification report on testing data for Naive Bayes :

	precision	recall	f1-score	support
0	0.74	0.67	0.70	1207
1	0.76	0.81	0.79	1557
accuracy			0.75	2764
macro avg	0.75	0.74	0.74	2764
weighted avg	0.75	0.75	0.75	2764

## 4.7 Result

	ML Model	Train Accuracy	Test Accuracy
1	Random Forest	0.968	0.956
0	Decision trees	0.968	0.954
5	SVM	0.949	0.938
3	K-Nearest Neighbors	0.938	0.928
4	Naive Bayes:GaussianNB	0.754	0.752
2	XGBoost	0.740	0.725

**Fig. 19. Result**

Agreeing with the result, arbitrary woodland and choice tree have the same preparation precision but Irregular woodland testing precision is superior to choice tree testing precision. Within the perplexity lattice, genuine positive and wrong positive esteem of irregular timberland is superior to the decision tree. So, ready to select the Irregular Timberland Classifier Calculation for our venture.

## 5. System Analysis & Design

### 5.1 Principles of System Analysis

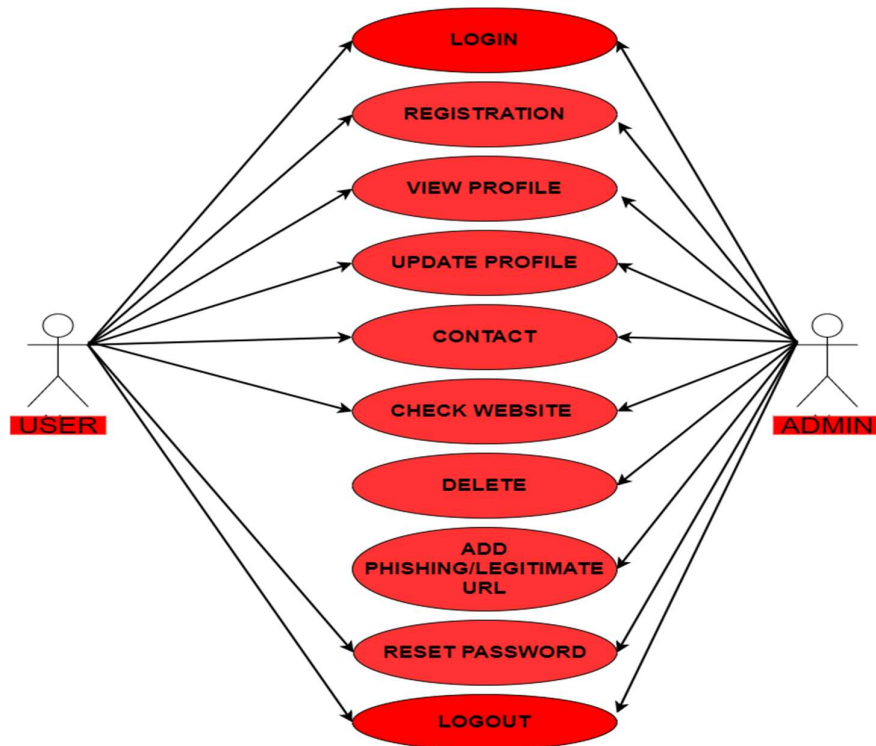
- Get it the issue sometime recently you start to make the investigation demonstrate.
- Create models that empower a client to understand how human-machine interaction will happen.
- Record the beginning of and the reason for each necessity.
- Utilize different sees of necessities like building information, work, and behavioral models.
- Work to dispense with equivocalness.

### 5.2 The System Design

The framework plan stage depicts the utilitarian capabilities of the proposed framework. Framework's plan is the method of characterizing the engineering, modules, interfacing, and information for a framework to ful fill specified requirements. The framework's plan might be seen as the application of the framework's hypothesis to item advancement. This is often isolated into the taking-after-plan stages:

#### 5.2.1 Use Case Diagram

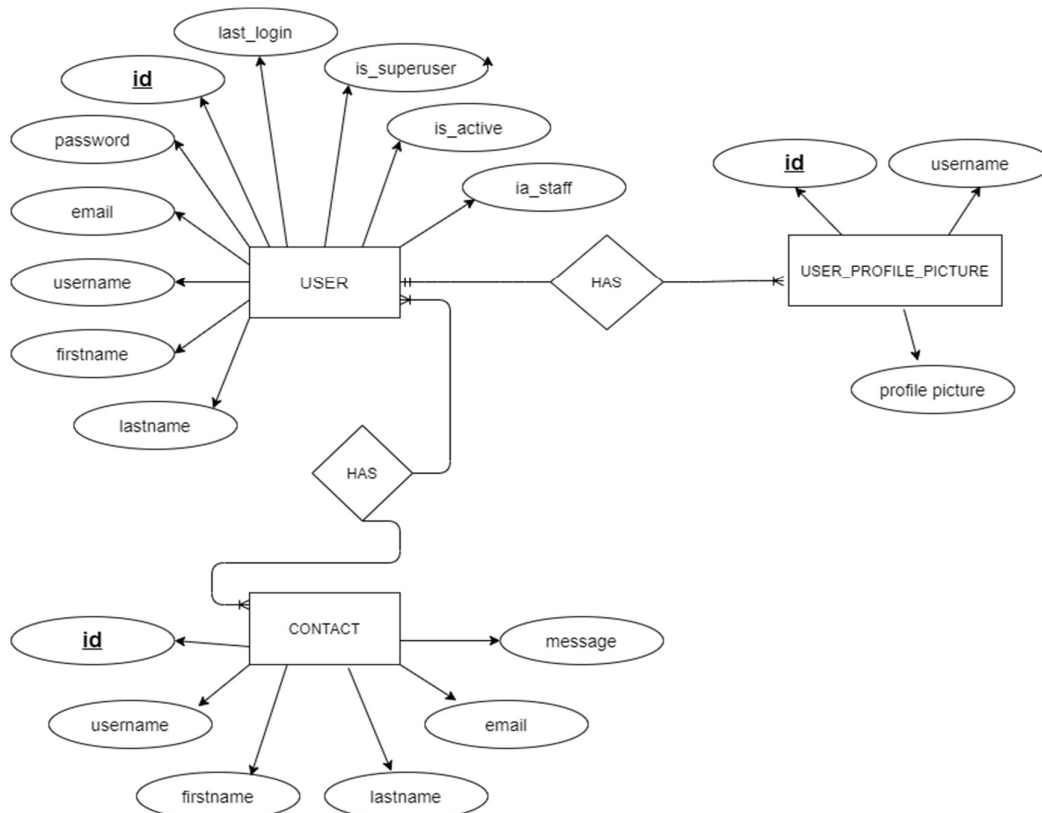
Utilize case graphs to speak to the capacities of a framework from the user's point of see. Fig 20 represents the use-case diagram.



**Fig. 20.** Use Case Diagram

### 5.2.2 Entity Relationship Diagram

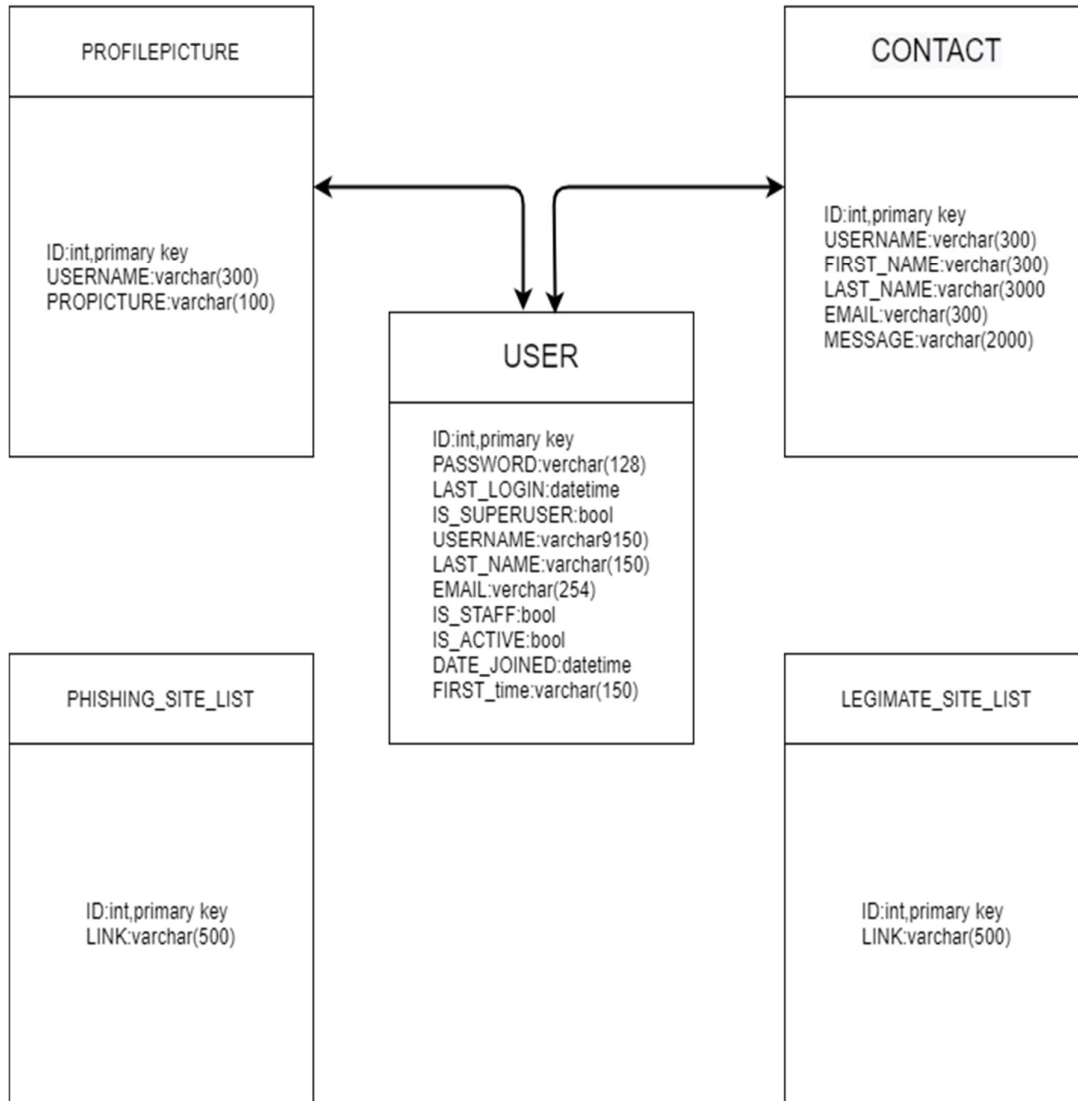
A substance relationship chart (ERD) shows the connections of substance sets put away in a database. ER graphs outline the coherent structure of databases. Fig 21 represents the Entity Relationship Diagram.



**Fig. 21.** Entity Relationship Diagram

### 5.2.3 Database Construction

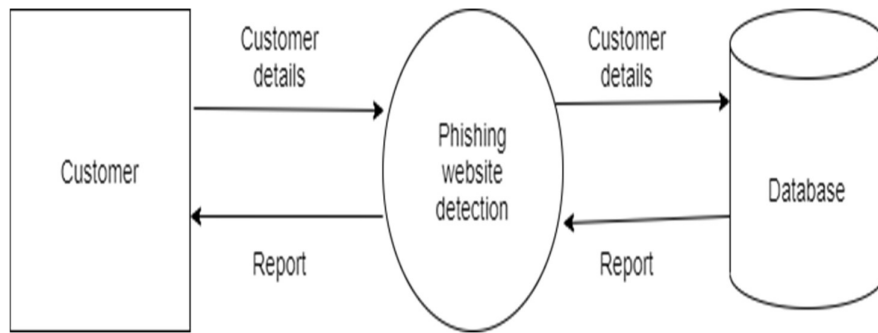
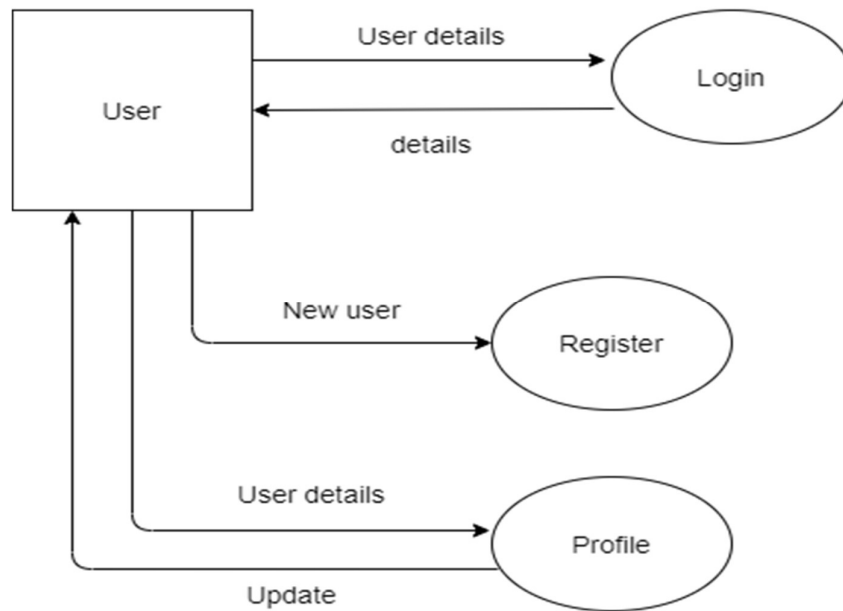
A database Construction is the skeleton structure that speaks to the coherent see of the complete database. It characterizes how the information is organized and how the relations among them are related. It is represented in Fig 22.



**Fig. 22.** Database Construction

### 5.2.4 Customer Data Flow Diagram

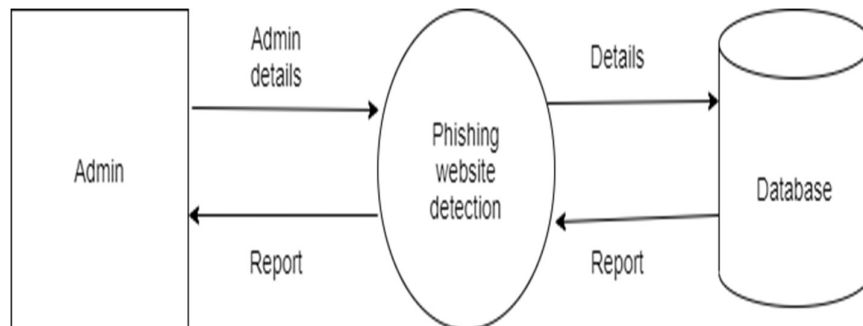
A data-flow diagram is a way of representing the flow of data in a process or a system. It is represented in Fig 23.

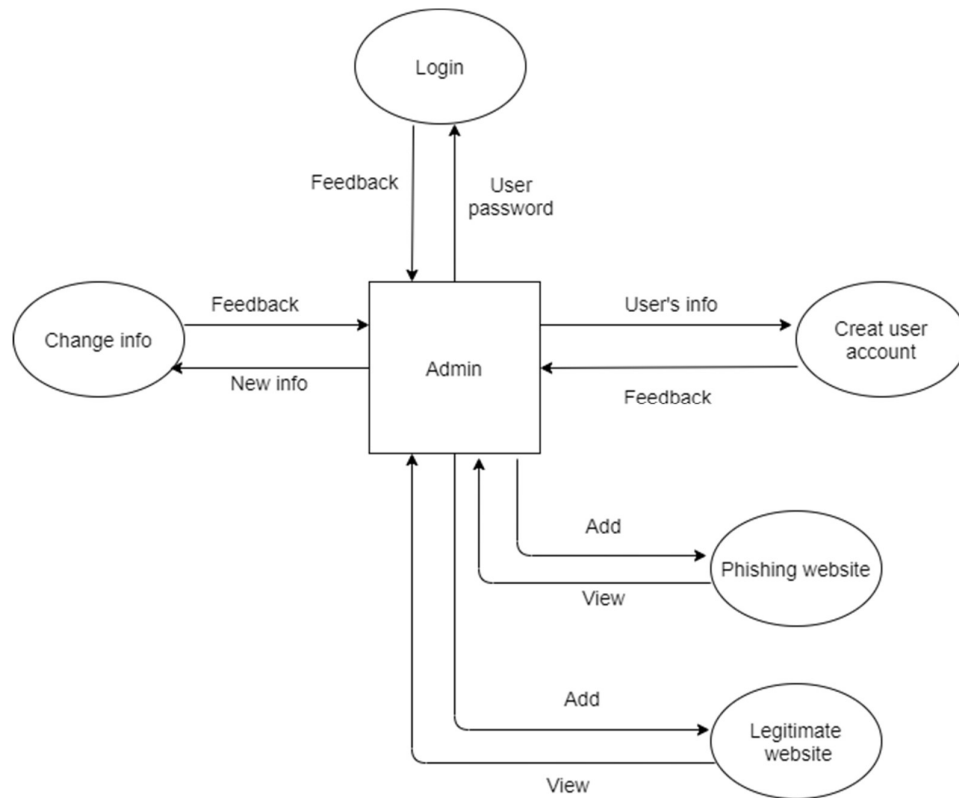
**Level 0 :****Level 1 :**

**Fig. 23.** Customer/Level 0, Level 1 Data Flow Diagram

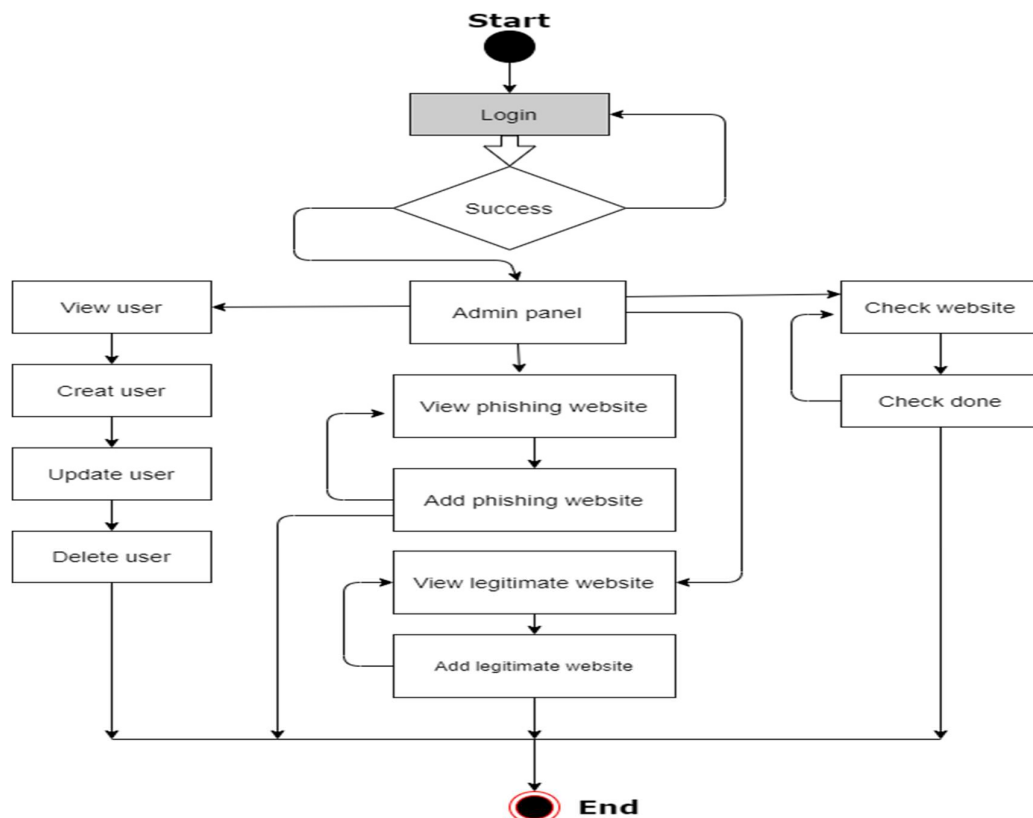
### 5.2.5 Admin Data Flow Diagram

Setting a chart could be a top-level see of the data framework. Context diagram has one handle image speaking to the whole data framework. It is represented in the Fig 24.

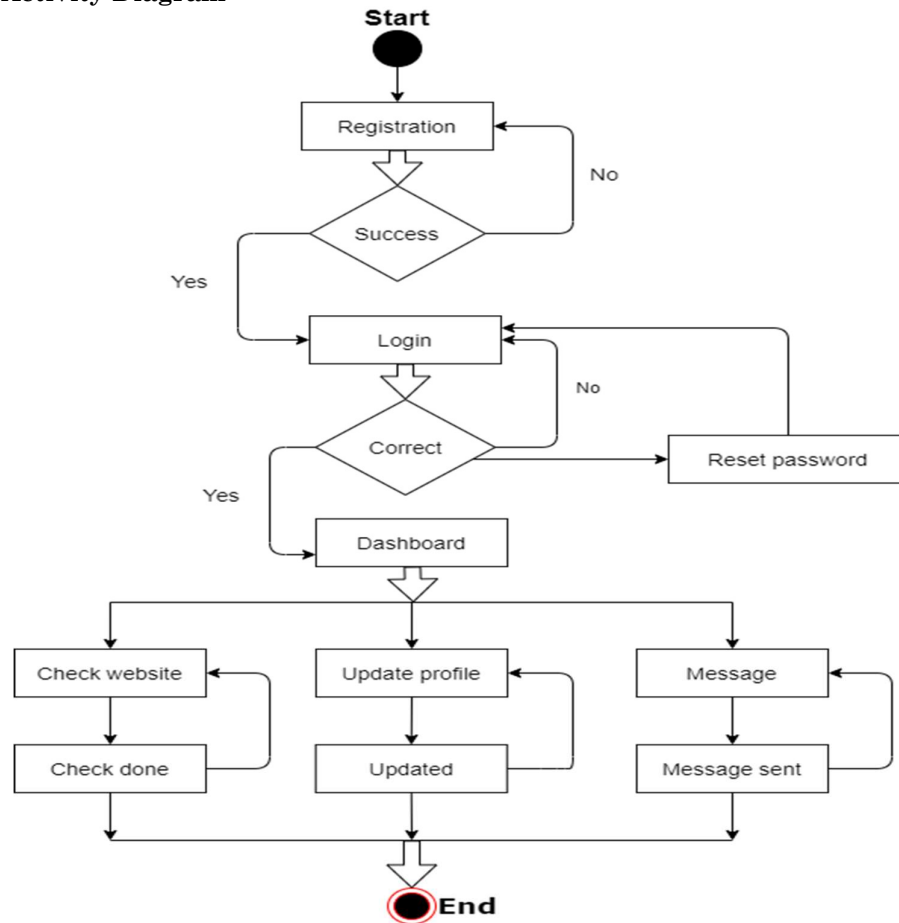
**Level 0:**

**Level 1:****Fig. 24.** Admin/Level 0,Level 1 Data Flow Diagram**5.2.6 Activity Diagram**

Movement charts are to speak to the parallel behavior of an operation as a set of activities.

**5.2.6.1 Admin Activity Diagram****Fig. 25.** Admin Activity Diagram

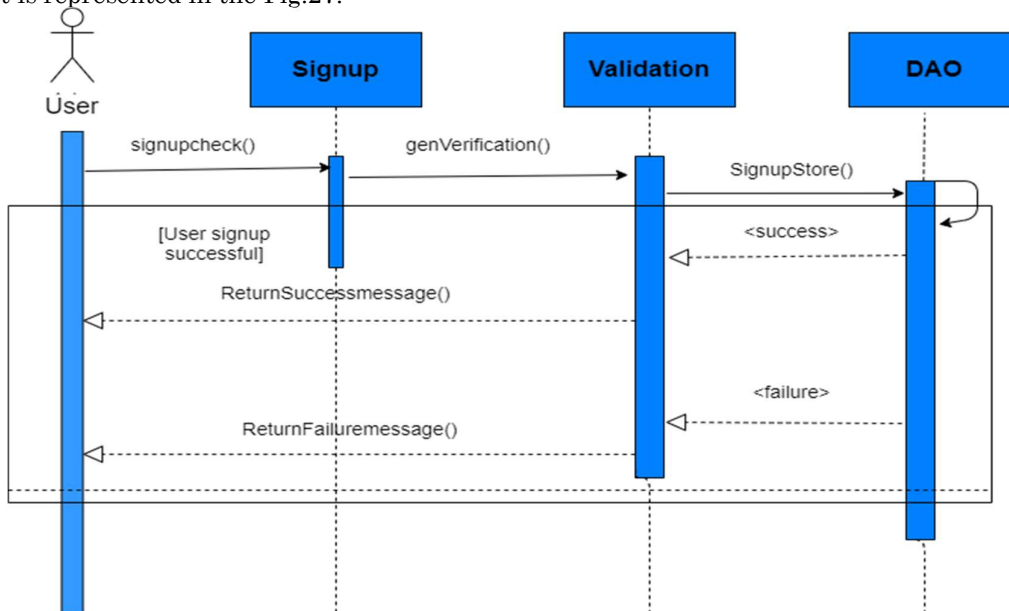
### 5.2.6.2 User Activity Diagram



**Fig 26.** User Activity Diagram

### 5.2.7 Signup Sequence Diagram

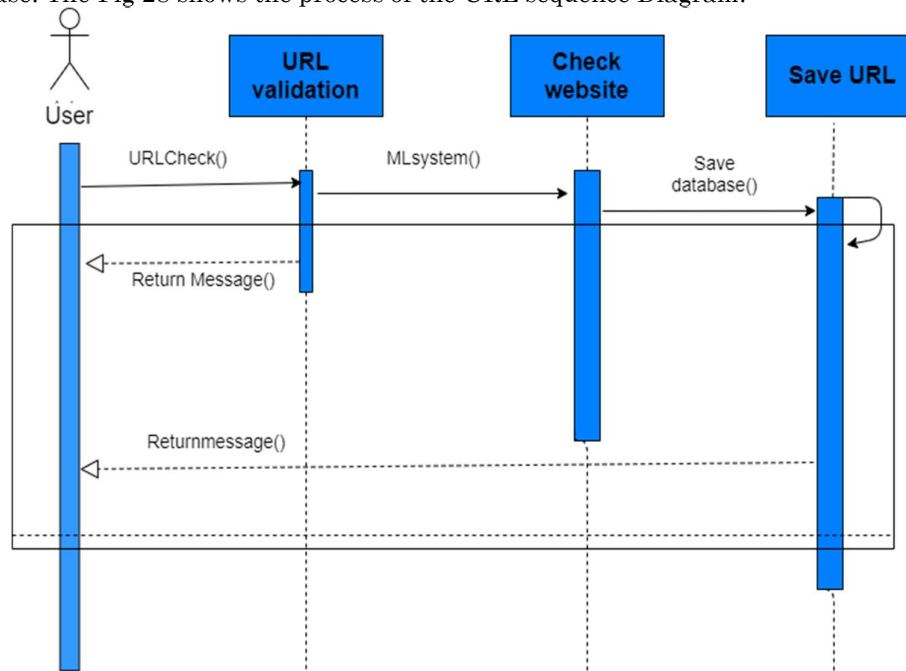
In the sign-in process, the user data is entered and the data is validated through `getVerification()` function the data moves to DAO which has two results. If it is correct then the user receives the message as “User signup successful”. If the data is incorrect then it moves to “failure” and then a failure message is sent. It is represented in the Fig.27.



**Fig. 27.** Signup Sequence Diagram

### 5.2.8 Check URL Sequence Diagram

When the user enters the URL, the link is validated and the website is checked. Then the URL is saved in the database. The Fig 28 shows the process of the URL sequence Diagram.



**Fig 28.** Check URL Sequence Diagram

## 6. Features

- Check whether a website is using Phishing or not by URL
- Add URL to Blacklist (admin)
- View Blacklisted List (admin)
- View User List (admin)
- Feedback
- Registration
- Login
- View Profile
- Update Profile

### 6.1 Technology

#### 6.1.1 Software & Hardware Requirements for Running Software (End-User)

Hardware: Computer, Mobile, Tab, etc with minimum requirements.  
 Software: Operating System: Platform Independent. Browsers: Almost every browser.  
 Network: Must have an internet connection.

#### 6.1.2 Software & Hardware used for Development

##### Hardware used for development

Configuration	Personal Computer (PC)
Processor	Intel Core i7-9750H (12M Cache, 2.60 GHz up to 4.50 GHz) Processor.
RAM	8 GB 3600MHz
HDD + SSD	1TB + 128GB
OS	Windows 10

### 6.1.3 Software used for development Tools

Offline Software	Anaconda, Jupiter Notebook, VS Code, Chrome Browser, DB Browser (SQLite), Python 3.8
Online Editor	Google Colab, Chrome Browser, GitHub Desktop

### 6.1.4 Programming Language used for development

Application Server	Live Server
Front End	HTML, CSS ( Bootstrap ), JavaScript (J-query)
Back End	Python ( Django ), SQLite

### 6.1.5 Software used for design

For prototype design and diagram design

System Design	<a href="https://app.diagrams.net/">https://app.diagrams.net/</a>
Prototype Design	Justinmind
Modelling	<a href="https://lucid.app">https://lucid.app</a>

## 7. Web Implementation

We actualized the front conclusion utilizing Visual Studio. Microsoft Visual Studio is a coordinated improvement environment (IDE) from Microsoft. It is utilized to create computer programs, as well as web locales, web apps, web administrations, and portable apps. Visual Studio employs Microsoft program improvement stages such as Windows API, Windows Shapes, and Windows.

Introduction Establishment, Windows Store, and Microsoft Silver light. It can deliver both local code and overseen code.

Visual Studio incorporates a code editor supporting IntelliSense (the code completion component) as well as code refactoring. The coordinates debugger works both as a source-level debugger and a machine-level debugger. Other built-in instruments incorporate a code profiler, shapes architect for building GUI applications, web architect, course creator, and database construction architect. It acknowledges plug-ins that improve the usefulness at nearly every level—including bolster for source control frameworks (like Subversion) and including modern instrument sets like editors and visual originators for domain-specific dialects or toolsets for other perspectives of the computer program advancement lifecycle (just like the Group Establishment Server client: Group Pioneer).

Visual Studio bolsters 36 diverse programming dialects and permits the code editor and debugger to back (to shifting degrees) about any programming dialect, given a language-specific benefit exists.

#### A . Python

Python could be a broadly utilized high-level, general-purpose, deciphered, energetic programming dialect. Its plan reasoning emphasizes code readability and its language structure permits software engineers to precise concepts in fewer lines of code than would be conceivable in dialects such as Java and C++. The dialect develops expecting to empower clear programs on both little and expansive scales.

Python bolsters multiprogramming ideal models, counting object-oriented basic and useful programming or procedural styles. It highlights an energetic sort framework and programmed memory administration and contains an expansive and comprehensive standard library. Python mediators are accessible for establishment on numerous working frameworks, permitting Python code execution on a wide assortment of frameworks.

#### B. Django

Django could be a high-level Python Web system that empowers quick advancement and clean, business plans. Built by experienced designers, it takes care of much of the hassle of Web advancement, so you'll be able to center on composing your app without requiring to rehash the wheel. It's free and open source.

#### C. Bootstrap

Bootstrap may be a free and open-source CSS system coordinated at responsive, mobile-first front-end web improvement. It contains CSS- and (alternatively) JavaScript-based plan formats for typography, shapes, buttons, routes, and other interface components.

Bootstrap is the seventh-most-starred extend on GitHub, with more than 142,000 stars, behind free Code Camp (nearly 312,000 stars) and possibly behind Vue.js system

## D. Heroku

Heroku may be a cloud stage as a benefit (PaaS) supporting a few programming dialects. One of the primary cloud stages, Heroku has been in advancement since June 2007, when it backed as it were the Ruby programming dialect, but presently underpins Java, Node.js, Scala, Clojure, Python, PHP, and Go. For this reason, Heroku is said to be a multilingual stage because it has highlights for a designer to construct, comparably run, and scale applications over most dialects. Applications that are run on Heroku ordinarily have an interesting space (regularly "applicationname.herokuapp.com") utilized to course HTTP demands to the proper application holder or dyno. Each of the dynos isspread over a "dyno framework" which comprises a few servers. Heroku's Git server handles application store pushes from allowed clients. All Heroku administrations are facilitated on Amazon's EC2 cloud-computing stage.

## 7.1 User side

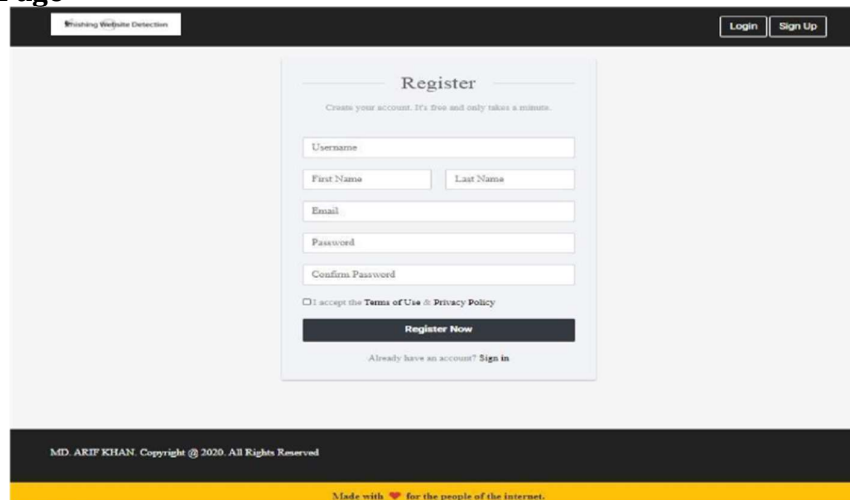
### 7.1.1 Index page

*Fig. 29. Index page*

### 7.1.2 Login Page

*Fig. 30. Login Page*

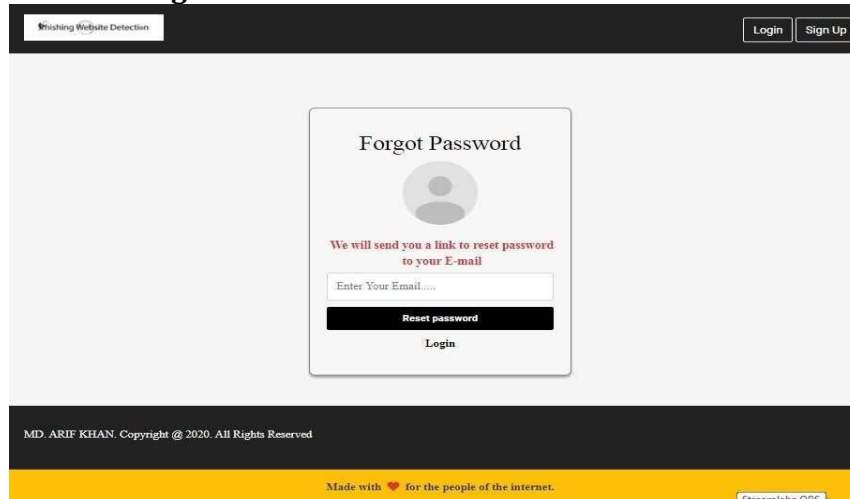
### 7.1.3 Signup Page



The screenshot shows a web application interface for a signup page. At the top, there is a dark header bar with a 'Flashing Website Detection' button on the left and 'Login' and 'Sign Up' buttons on the right. The main content area is light gray and features a central 'Register' form. The form has a title 'Register' and a subtitle 'Create your account. It's free and only takes a minute.' Below this, there are input fields for 'Username', 'First Name', 'Last Name', 'Email', 'Password', and 'Confirm Password'. A checkbox labeled 'I accept the Terms of Use & Privacy Policy' is located below the password fields. A dark 'Register Now' button is positioned below the checkbox. At the bottom of the form, there is a link 'Already have an account? Sign in'. The footer consists of a dark bar with the text 'MD. ARIF KHAN. Copyright @ 2020. All Rights Reserved' and a yellow bar below it with the text 'Made with ❤️ for the people of the internet.'

*Fig. 31. Signup Page*

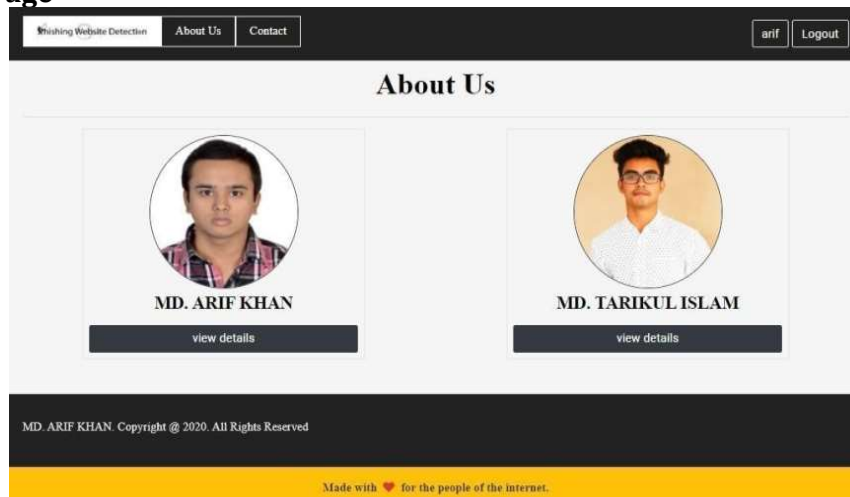
### 7.1.4 Forgot Password Page



The screenshot shows a web application interface for a forgot password page. At the top, there is a dark header bar with a 'Flashing Website Detection' button on the left and 'Login' and 'Sign Up' buttons on the right. The main content area is light gray and features a central 'Forgot Password' form. The form has a title 'Forgot Password' and a subtitle 'We will send you a link to reset password to your E-mail'. Below this, there is an input field labeled 'Enter Your Email.....'. A dark 'Reset password' button is positioned below the input field. At the bottom of the form, there is a link 'Login'. The footer consists of a dark bar with the text 'MD. ARIF KHAN. Copyright @ 2020. All Rights Reserved' and a yellow bar below it with the text 'Made with ❤️ for the people of the internet.' and a 'Streamlabs OBS' logo on the right.

*Fig. 32. Forgot Password Page*

### 7.1.5 About Page



The screenshot shows a web application interface for an about page. At the top, there is a dark header bar with a 'Flashing Website Detection' button on the left, 'About Us' and 'Contact' buttons in the center, and 'arif' and 'Logout' buttons on the right. The main content area is light gray and features a central 'About Us' section. This section contains two circular profile pictures. The first profile picture is of MD. ARIF KHAN, and the second is of MD. TARIKUL ISLAM. Below each profile picture is a dark 'view details' button. The footer consists of a dark bar with the text 'MD. ARIF KHAN. Copyright @ 2020. All Rights Reserved' and a yellow bar below it with the text 'Made with ❤️ for the people of the internet.'

*Fig 33. About Page*

### 7.1.6 User Profile

**Fig. 34.** User Profile

### 7.1.7 Contact Page

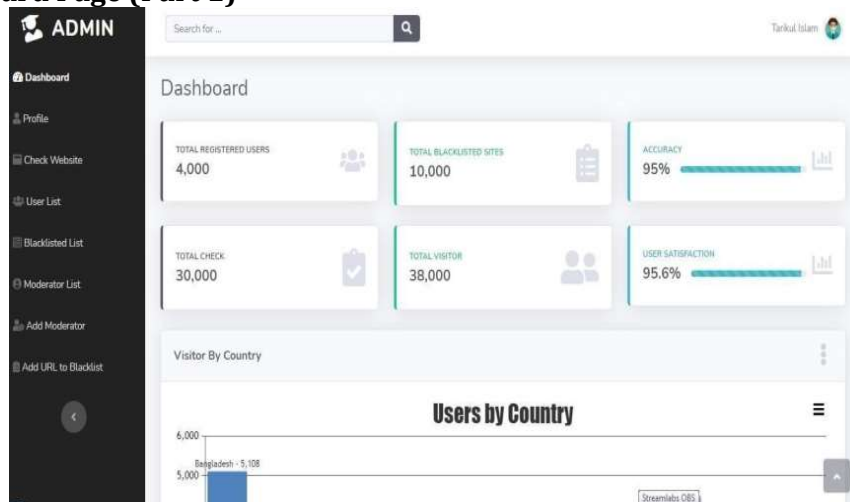
**Fig. 35.** Contact Page

## 7.2 Admin side

### 7.2.1 Login Page

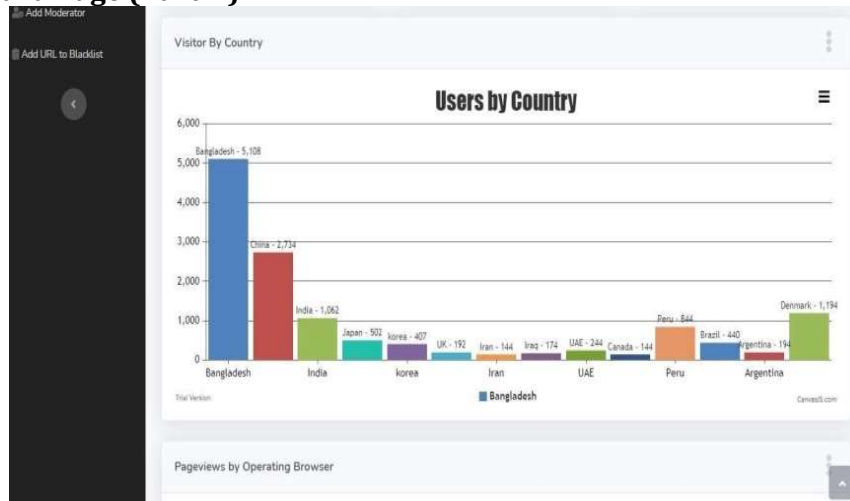
**Fig. 36.** Login Page

### 7.2.2 Dashboard Page (Part 1)



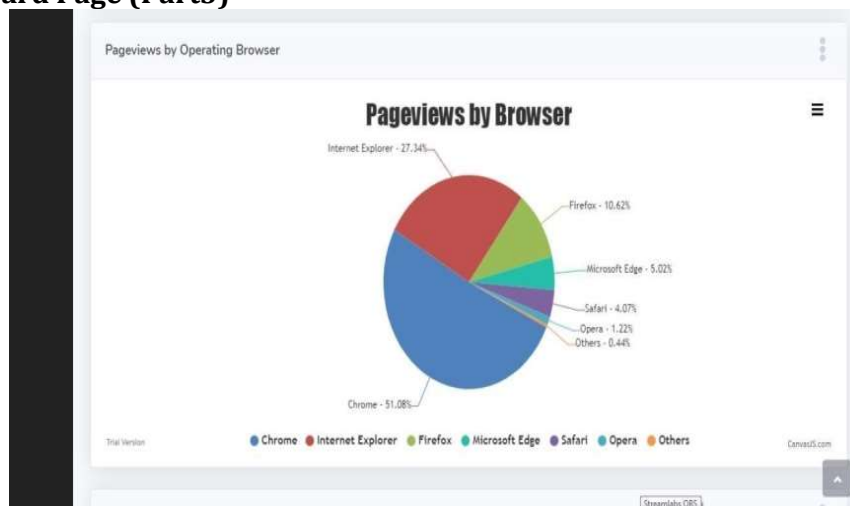
*Fig. 37. Dashboard Page (part-1)*

### 7.2.3 Dashboard Page (Part-2)



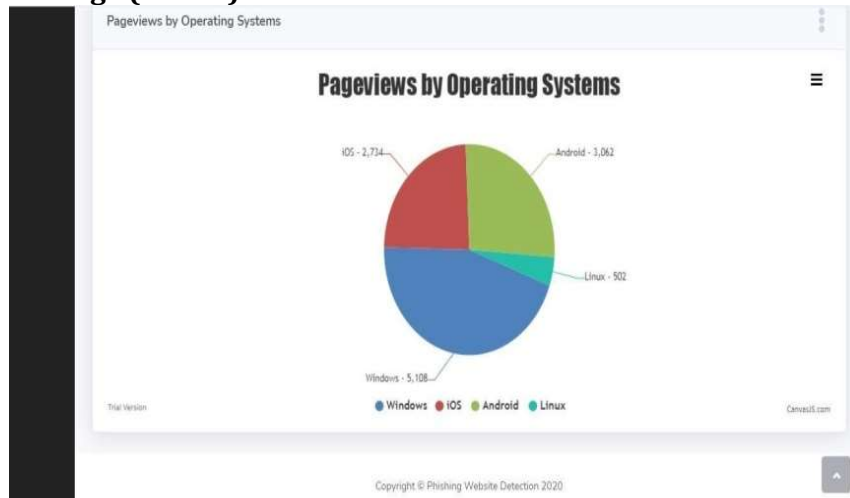
*Fig. 38. Dashboard Page (part-2)*

### 7.2.4 Dashboard Page (Part3)



*Fig. 39. Dashboard Page (part-3)*

### 7.2.5 Dashboard Page (Part-4)



**Fig. 40. Dashboard Page (Part-4)**

### 7.2.6 Admin Profile Page

The figure shows the Admin Profile Page. It features a sidebar on the left with the same menu items as the dashboard page. The main content area is titled "Profile" and contains a user profile card. The profile card includes a profile picture placeholder with a "Choose File" button and a "Save Photo" button. To the right of the profile card are two sections: "User Settings" and "Contact Settings". The "User Settings" section contains fields for Username (user.name), Email Address (user@example.com), First Name (John), and Last Name (Doe), along with a "Save Settings" button. The "Contact Settings" section contains an Address field with the value "Sunset Blvd, 38". The footer of the page includes the text "Copyright © Phishing Website Detection 2020".

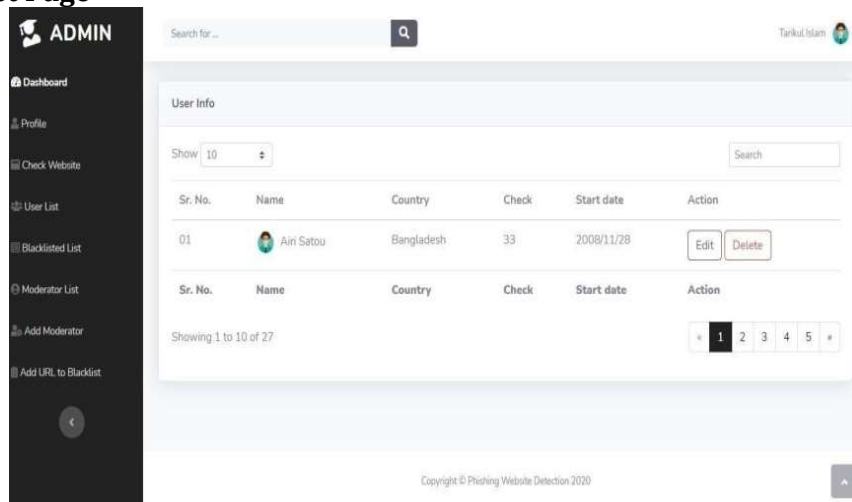
**Fig. 41. Admin Profile Page**

### 7.2.7 Admin Url Check Page

The figure shows the Admin URL Check Page. It features a sidebar on the left with the same menu items as the dashboard page. The main content area is titled "Check Website By URL :". It contains a form with a "URL :" label, a text input field with the example value "https://www.google.com", and a "Check" button. The footer of the page includes the text "Copyright © Phishing Website Detection 2020".

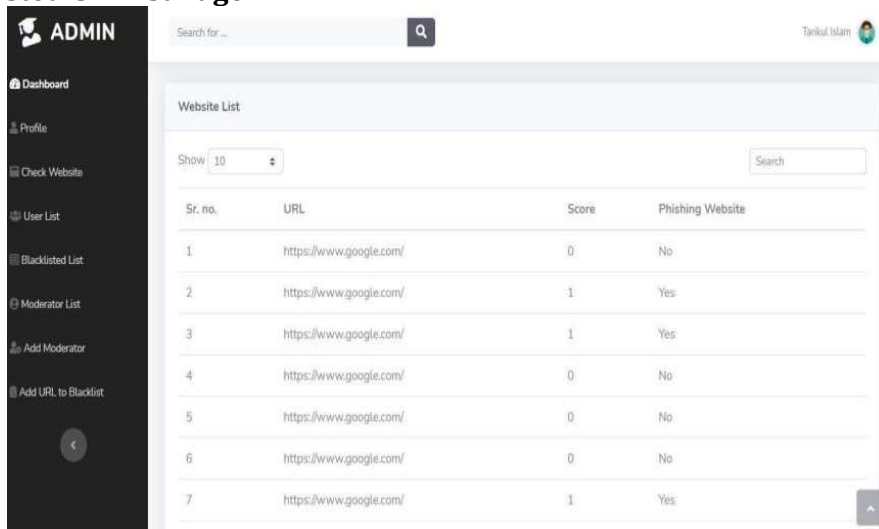
**Fig 42. Admin URL Check Page**

### 7.2.8 User List Page



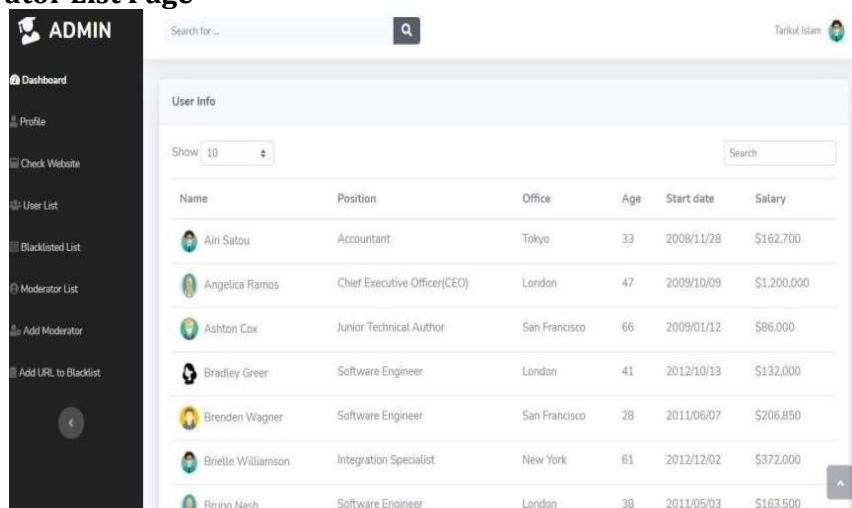
**Fig. 43.** User List Page

### 7.2.9 Blacklisted Url List Page



**Fig. 44.** Blacklisted URL List Page

### 7.2.10 Moderator List Page



**Fig. 45.** Moderator List Page

### 7.2.11 Add Moderator Page

**Fig 46.** Add Moderator Page

### 7.2.12 Add URL to Blacklist Page

**Fig. 47.** Add URL to Blacklist Page

## 7.3 Software Testing Report

After the framework is created, the method of framework testing must be carried on in arrange to test in case the framework is free of bugs. If amid the framework testing, there are bugs or mistakes detected, the engineer may haveto adjust and settle the bugs quickly.

### 7.3.1 Test forms on all pages

Shapes are an indispensable portion of any site. Shapes are utilized for accepting data from clients and to associate with them. So what ought to be checked on these shapes?

- To begin with, check all the validations on each field.
- Check for default values of the areas.
- Off-base inputs within the forms to the areas within the shapes.
- Alternatives to make shapes on the off chance that any, shape erase, see, or adjust shapes.

Let's take a case of the look motor venture right now I am working on, To this extent we have sponsor and partner signup steps. Each signup step is distinctive but it's subordinate to the other steps. So sign-up stream ought to be executed accurately. There are distinctive field validations like e-mail IDs, Client budgetary data validations, etc. All these validations ought to be checked in manual or robotized web testing.

### 7.3.2 Cookies Testing

Treats are little records put away on the client's machine. These are essentially utilized to preserve the session- the login sessions. Test the application by empowering or crippling the treats in your browser

choices. Test on the off chance that the treats are scrambled sometime recently composing to the user machine. If you're testing the session treats (i.e. treats that lapse after the session closes) check for login sessions and client stats after the session closes. Check the impact on application security by erasing the treats. (I will before long compose an isolated article on cookie testing as well).

### 7.3.3 Validate your HTML/CSS

On the off chance that you're optimizing your location for Look motors at that point, HTML/CSS validation is the foremost imperative one. Approve the location for HTML sentence structure blunders.

### 7.3.4 Database testing

Information consistency is additionally exceptionally critical in web applications. Check for information astuteness and mistakes when you alter, erase, adjust the shapes, or do any DB-related usefulness. Check-in case all the database inquiries are executing correctly, data is recovered additionally updated accurately. More on database testing might be stacked on DB, we will address this in web stack or execution testing underneath.

### 7.3.5 Test for Navigation

Route implies how a client surfs the internet pages, distinctive controls like buttons, and boxes, or how the client employments the joins on the pages to surf diverse pages. Ease-of-use testing incorporates the taking after:

- Websites ought to be simple to utilize.
- Informational given ought to be exceptionally clear.
- Check in case the information given is idealized to fulfill its reason.
- Primary menu ought to be given on each page.
- It ought to be steady and sufficient.

### 7.3.6 Content checking

Content should be logical and easy to understand. Check for spelling errors. Usage of dark colors annoys the users and should not be used in the site theme. You can follow some standard colors that are used for web pages and content building. These are the commonly accepted standards like what I mentioned above about annoying colors, fonts, frames, etc.

### 7.3.7 Interface Testing

The most interfacing are:

Web server and application server interface, Application server and Database server interface. Check if all the intelligence between these servers is executed and mistakes are dealt with legitimately. If the database or web server returns any blunder message for any inquiry by the application server at that point application server ought to capture and show these mistake messages fittingly to the clients. Check what happens on the off chance that the client hinders any exchange in between. Check what happens if the association to the net server is reset in between.

### 7.3.8 Browser compatibility

In my web-testing career, I have experienced this as the foremost affecting portion of web location testing. A few applications are exceptionally subordinate to browsers. Distinctive browsers have different arrangements and settings that your web page ought to be consistent along. With your site, coding ought to be a cross-browser stage consistent. On the off chance that you're utilizing Java scripts or AJAX calls for UI usefulness, performing security checks or validations at that point donates more stretch on browser compatibility testing of your web application. Test web applications on diverse browsers like Web Pilgrim, Firefox, Netscape Pilot, AOL, Safari, and Musical Drama browsers with distinctive forms.

### 7.3.9 OS compatibility

A few usefulness in your web application is that it may not be congruous with all working frameworks. All unused advances utilized in web advancement like realistic plans, and interface calls like diverse API may not be accessible in all Working Frameworks.

Subsequently test your web application on diverse working frameworks like Windows, UNIX, MAC, Linux, and Solaris with diverse OS flavors.

### 7.3.10 Versatile browsing

We are in a modern innovation time. So in the future portable browsing will shake. Test your web pages on mobile browsers. Compatibility issues may be there on portable gadgets as well.

### 7.3.11 Printing alternatives

On the off chance that you're giving page-printing choices at that point make beyond any doubt textual styles, page arrangement, page design, etc., getting printed legitimately. Pages ought to fit the paper estimate or as per the size mentioned in the printing choice.

### 7.3.12 Performance testing

Web applications ought to maintain an overwhelming stack. Web execution testing ought to incorporate:

- Web Stack Testing
- Web Push Testing

Test application execution on distinctive web association speeds.

#### Web stack testing

You wish to test on the off chance that numerous clients are getting to or asking on the same page. Can a framework maintain top stack times? The location ought to handle numerous synchronous client demands, expansive input information from clients, synchronous association to DB, overwhelming stack on particular pages, etc.

#### Web Stretch testing

For the most part, push implies extending the framework past its indicated limits. Web push testing is performed to break the location by giving push and it's checked as how the framework responds to stretch and how it recoups from crashes. Stretch is for the most part given on input areas, login, and sign-up zones.

In web execution testing site usefulness on distinctive working frameworks and diverse equipment stages is checked for program and equipment memory spillage blunders.

### 7.3.13 Security Testing

Taking after are a few of the test cases for web security testing:

- Test by gluing inside the URL straightforwardly onto the browser address bar without login. Inside pages ought to not open.
- On the off chance that you're logged in utilizing username and watchword and browsing inner pages at that point attempt changing URL alternatives specifically.
- Web catalogs or records ought to not be open specifically unless they are given a download choice.
- All exchanges, mistake messages, and security breach endeavors ought to get logged in log files somewhere on the internet server.

### 7.3.14 Testing Using Software

#### Test View

- Load Testing
- Error Testing

Load testing using JMeter. Three times testing and there is no error.

#### 7.3.14.1 View Result in Table

Sample #	Start Time	Thread Name	Label	Sample Time(ms)	Status	Bytes	Sent Bytes	Latency	Connect Time(...)
1	01:38:30.712	Thread Group 1-4 HTTP Request		1349	✓	330593	270	390	49
2	01:38:30.812	Thread Group 1-5 HTTP Request		1644	✓	330601	270	390	49
3	01:38:31.011	Thread Group 1-7 HTTP Request		2716	✓	330600	270	400	60
4	01:38:31.212	Thread Group 1-9 HTTP Request		2980	✓	330601	270	398	49
5	01:38:30.513	Thread Group 1-2 HTTP Request		4078	✓	330601	270	404	50
6	01:38:30.612	Thread Group 1-3 HTTP Request		4174	✓	330601	270	407	50
7	01:38:32.062	Thread Group 1-4 HTTP Request		2844	✓	330601	270	882	0
8	01:38:30.411	Thread Group 1-1 HTTP Request		4615	✓	330601	270	1505	53
9	01:38:31.111	Thread Group 1-8 HTTP Request		4042	✓	330601	270	396	46
10	01:38:31.311	Thread Group 1-6 HTTP Request		4230	✓	330601	270	1969	50
11	01:38:32.457	Thread Group 1-5 HTTP Request		3253	✓	330601	270	587	0
12	01:38:33.727	Thread Group 1-7 HTTP Request		2779	✓	330593	270	484	98
13	01:38:30.911	Thread Group 1-6 HTTP Request		5794	✓	330601	270	402	49
14	01:38:34.591	Thread Group 1-2 HTTP Request		2549	✓	330601	270	495	100
15	01:38:35.027	Thread Group 1-1 HTTP Request		2468	✓	330601	270	1540	75
16	01:38:35.153	Thread Group 1-8 HTTP Request		2675	✓	330601	270	441	83
17	01:38:35.541	Thread Group 1-3 HTTP Request		2608	✓	330601	270	1444	49
18	01:38:34.786	Thread Group 1-3 HTTP Request		3632	✓	330601	270	1517	83
19	01:38:34.192	Thread Group 1-9 HTTP Request		5056	✓	330601	270	3859	2368
20	01:38:36.706	Thread Group 1-6 HTTP Request		3565	✓	330790	270	1504	50

*Fig. 48. View Result in Table*

### 7.3.14.2 View Result in Tree

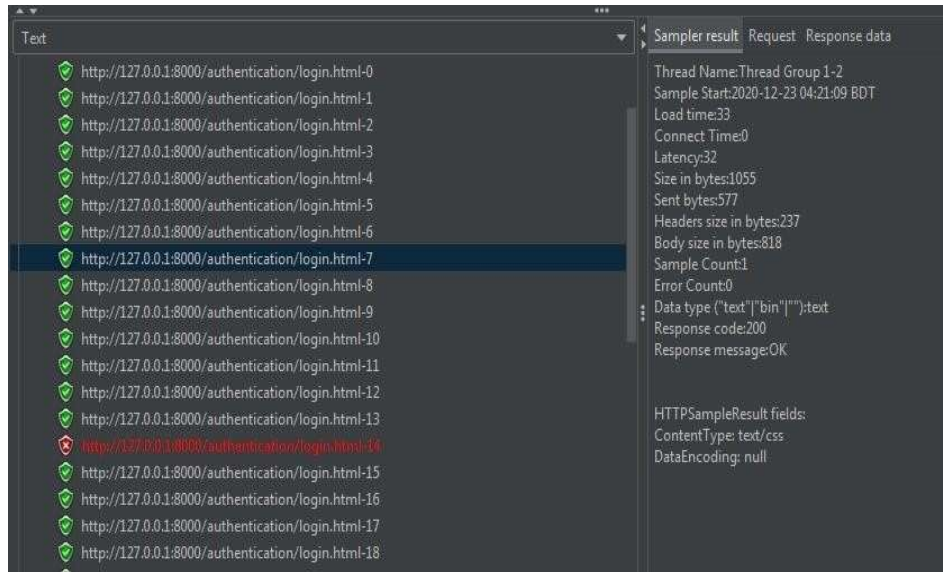


Fig. 49. View Result in Tree

### 7.3.14.3 Summary Report

Label	# Samples	Average	Min	Max	Std. Dev.	Error %	Throughput	Received KB/s	Sent KB/sec	Avg. Bytes
HTTP Request	5	2108	1562	2567	426.48	0.00%	1.7/sec	550.37	0.45	330506.2
TOTAL	5	2108	1562	2567	426.48	0.00%	1.7/sec	550.37	0.45	330506.2

Fig. 50. Summary Report

## 7.3.15 Manual Testing Report

### 7.3.15.1 Using phishing website

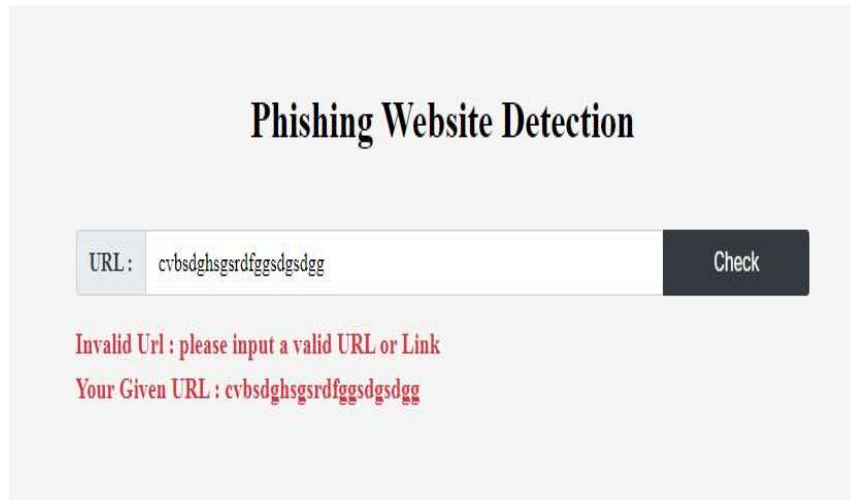
Phishing Website	Result
http://norcaltc-my.sharepoint.com	Phishing
http://sieck-kuehlsysteme.de	Phishing
http://wvk12-my.sharepoint.com	Phishing
http://currentlyattreply.weebly.com	Phishing
http://portalstone.joomla.com	Phishing
http://keypointtraining-my.sharepoint.com	Phishing
http://tiseopavingco-my.sharepoint.com	Phishing
http://gillas-first-project.webflow.io	Phishing
http://appriver3651012784-my.sharepoint.com	Phishing
http://mundovirtualhabbo.blogspot.com	Phishing
http://web-kiwi.org/	Phishing
https://www.ysense.com	Phishing

### 7.3.15.2 Using the legitimate website

Legitimate Website	Result
https://facebook.com	Legitimate
https://google.com	Legitimate
http://graphicriver.net	Legitimate
http://ecnavi.jp	Legitimate
http://icicibank.com	Legitimate
http://nypost.com	Legitimate
http://tune.pk	Legitimate
https://yahoo.com	Legitimate
http://kienthuc.net.vn	Legitimate
https://youtube.com	Legitimate

### 7.3.15.3 Invalid URL Test

If the URL is invalid or without a registered top-level domain, then the page is back with an invalid URL.



*Fig. 51. System catch simple text*



*Fig. 52. System catch URL without TLD*

## 8. Advantages and Disadvantages

### 8.1 Limitations

- All e-banking websites related to data will be stored in one place.
- It fails to detect when attackers use a compromised domain for hosting their site.

### 8.2 Benefits

- Many e-commerce websites can use this technique to maintain positive customer relationships. Payments can be made safely online by users.
- Comparing this system's data mining algorithm to other conventional classification algorithms, it performs better.
- The user can confidently make online product purchases with the aid of this technology.

## 9. Conclusion

In this work, we executed six classifiers utilizing Google Colab. The classifiers were utilized to distinguish phishing URLs. In identifying phishing URLs, there are two steps. The primary step is to extricate highlights from the URLs, and the moment step is to classify URLs utilizing the demonstration

that has been created with the assistance of the preparing set information. In this work, we utilized the information set that gave the extricated highlights. The information set, from the Kaggle contains 30 input highlights.

Phishing could be a way to get a user's private data through mail or a site. As the utilization of the web is exceptionally tremendous, nearly all things are accessible online presently it is either shopping for dresses, electronic contraptions, and ceramics or paying for portable, TV & power bills. Instead of standing out in line for hours, individuals are becoming aware of utilizing online strategies. Due to this phishers contain a wide scope to actualize phishing tricks. As there's a part of inquiries about work tired this range, there's not any single method, which is sufficient to distinguish all sorts of phishing assaults. As innovation increments, phishing assaults utilize unused strategies day by day. This empowers us to discover compelling classifiers to identify phishing. We can say tree-based classifiers within the ML approach are way more reasonable than others.

## 9.1 Future Work

- All e-banking websites related to information will be put away in one put.
- It comes up short of distinguishing when assailants utilize a compromised space to facilitate their location
- Web-browser plug in or add-on
- Web API

## Compliance with Ethical Standards

**Conflicts of interest:** Authors declared that they have no conflict of interest.

**Human participants:** The conducted research follows the ethical standards and the authors ensured that they have not conducted any studies with human participants or animals.

## References

- [1] N. Abdelhamid, A. Ayesh, F. Thabtah, "Phishing Detection based Associative Classification", Data Mining. Expert Systems with Applications (ESWA), vol. 41, pp 5948-5959, 2014.
- [2] Arun Kulkarni, Leonard L. Brown, III, "Phishing Websites Detection using Machine Learning", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 7, 2019.
- [3] Nur Sholihah Zaini<sup>1</sup>, Deris Stiawan<sup>2</sup>, Mohd Faizal Ab Razak<sup>3</sup>, Ahmad Firdaus<sup>4</sup>, Wan IsniSofiah Wan Din<sup>5</sup>, Shahreen Kasim<sup>6</sup>, Tole Sutikno<sup>7</sup>, "Phishing detection system using machine learning classifiers", Indonesian Journal of Electrical Engineering and Computer Science Vol. 17, No. 3, pp. 1165~1171, March 2020.
- [4] Bhagyashree E. Sananse, Tanuja K. Sarode, "Phishing URL Detection: A Machine Learning and Web Mining-based Approach ", International Journal of Computer Applications (0975 – 8887) Volume 123 – No.13, August 2015.
- [5] Babagoli, Mehdi &Aghababa, Mohammad & Vahid Solouk, " Heuristic nonlinear regression strategy for detecting phishing websites", Soft Computing. 23.10.1007/s00500-018-3084-2 , June 2019.
- [6] Muhammad Taseer Suleman and Shahid Mahmood Awan, "Optimization of URL-Based Phishing Websites Detection through Genetic Algorithms" Autom. Control Comput. Sci., vol. 53, no. 4, pp. 333–341, 2019.
- [7] Mohammad, Rami, McCluskey, T.L. and Thabtah, FadiAbdeljaber, "Intelligent Rule based Phishing Websites Classification", IET Information Security, Vol. 8 (3). pp. 153-160, ISSN 1751-8709, 2014.
- [8] M. Hazim, N. B. Anuar, M. F. Ab Razak, and N. A. Abdullah, "Detecting opinion spams through supervised boosting approach", PLoS One, vol. 13, no. 6, pp. 1–23, 2018.
- [9] PhishMe, "Analysis of Susceptibility, Resiliency, and Defense Against Simulate and Real Phishing Attacks", 2017.
- [10] W. S. Cybersecurity, "Nearly 1.5 Million New Phishing Sites Created Each Month," Webroot Smarter Cybersecurity, 2017.
- [11] APWG, "APWG Phishing Attack Trends Reports", APWG Unifying Global Response to Cybercrime, 2018.
- [12] M. F. A. Razak, N. B. Anuar, F. Othman, A. Firdaus, F. Afifi, and R. Salleh, "Bio-inspired for Features Optimization and Malware Detection", Arab. J. Sci. Eng., 2018.
- [13] C. Whittaker, B. Ryner, M. Nazif, "Large-Scale Automatic Classification Of Phishing Pages", In Proc 17th Annual Network and Distributed System Security Symposium, NDSS 10, San Diego, CA, USA, 2010.
- [14] S. Garera, N. Provos, M. Chew, A.D. Rubin, "A Framework For Detection And Measurement Of Phishing Attacks". In Proc. 5th ACM Workshop on Recurring Malcode, WORM'07, ACM, New York, NY, USA, 2007.
- [15] Y. Zhang, J. Hong, L. Cranor, "CANTINA: A ContentBased Approach To Detecting Phishing Web Sites", In Proc. 15th Int. Conf. World Wide Web, WWW'07, Banff, Alberta, Canada, 2007.

- [16] J. Ma, L.K. Saul, S. Savage, G.M. Voelker, "Beyond Blacklists: Learning To Detect Malicious Web Sites From Suspicious URLs", In Proc. 15th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, Paris, France, 2009.
- [17] Sandeep Kumar Satapathy, Shruti Mishra, Pradeep Kumar Mallick, Lavanya Badiginchala, Ravali Reddy Gudur, Siri Chandana Guttha, "Classification of Features for detecting Phishing Web Sites based on Machine Learning Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-8 Issue-8S2, June 2019.
- [18] Ms. SophiyaShikalgarg, Dr. S. D. Sawarkar, Mrs. Swati Narwane, 2019, "Detection of URL-based Phishing Attacks using Machine Learning", INTERNATIONAL JOURNAL OFENGINEERING RESEARCH & TECHNOLOGY (IJERT), Volume 08, Issue 11 (November 2019).
- [19] Kiruthiga, R. and Akila, D.,. Phishing websites detection using machine learning. International Journal of Recent Technology and Engineering, Vol. 8(2), pp.111-114, 2019.
- [20] Doshi, J., Parmar, K., Sanghavi, R. and Shekokar, N., "A comprehensive dual-layer architecture for phishing and spam email detection", Computers & Security, Vol. 133, pp.103378, 2023.

#### Others :

**Project Idea :** <https://nevonprojects.com/detecting-phishing-websites-using-machine-learning>

**Google-colab :** <https://colab.research.google.com>

#### For Dataset :

<https://www.unb.ca/cic/datasets/url-2016.html>

#### For Webpage Design :

<https://getbootstrap.com/docs/4.5/getting-started/introduction>

#### Impact Learning :

**Document:** <https://pypi.org/project/ImpactLearning>

#### Dataset Load Google colab :

<https://www.youtube.com/watch?v=p5gnRGaw2aA&list=PLKdU0fuY4OfFWY36nDJDII26jXwInSm8f&index=16>

#### Feature Importance :

##### Theoretical :

[https://www.youtube.com/watch?v=Y1O\\_3TuVs\\_8&list=PLKdU0fuY4OfFWY36nDJDII26jXwInSm8f&index=32](https://www.youtube.com/watch?v=Y1O_3TuVs_8&list=PLKdU0fuY4OfFWY36nDJDII26jXwInSm8f&index=32)

##### Practical :

<https://www.youtube.com/watch?v=VXv0-Hv9cT4&list=PLKdU0fuY4OfFWY36nDJDII26jXwInSm8f&index=33>

#### Confusion Matrix :

1. <https://www.youtube.com/watch?v=7B5wz-s4pBE&list=PLKdU0fuY4OfFWY36nDJDII26jXwInSm8f&index=23>

2. <https://www.youtube.com/watch?v=Ihq-w9MW6Nc&list=PLKdU0fuY4OfFWY36nDJDII26jXwInSm8f&index=24>

#### Decision tree :

##### Theoretical

<https://www.youtube.com/watch?v=r7P3zmIcyw0&list=PLKdU0fuY4OfFWY36nDJDII26jXwInSm8f&index=18>

<https://www.youtube.com/watch?v=ivA04FnJrwY&list=PLKdU0fuY4OfFWY36nDJDII26jXwInSm8f&index=19>

##### Practical :

[https://www.youtube.com/watch?v=2cyHLn\\_WLqM&list=PLKdU0fuY4OfFWY36nDJDII26jXwInSm8f&index=20](https://www.youtube.com/watch?v=2cyHLn_WLqM&list=PLKdU0fuY4OfFWY36nDJDII26jXwInSm8f&index=20)

<https://www.youtube.com/watch?v=xV1SCO3fCrw&list=PLKdU0fuY4OfFWY36nDJDII26jXwInSm8f&index=21>

#### SVM :

##### Practical :

<https://www.youtube.com/watch?v=11hs9ilCO24&list=PLKdU0fuY4OfFWY36nDJDII26jXwInSm8f&index=27>

#### Regression Tree :

##### Theoretical :

<https://www.youtube.com/watch?v=ZeZhhZWVe88&list=PLKdU0fuY4OfFWY36nDJDII26jXwInSm8f&index=29>

##### Practical :

<https://www.youtube.com/watch?v=msQSmT3Sf1I&list=PLKdU0fuY4OfFWY36nDJDII26jXwInSm8f&index=30>

**KNN :****Theoretical :**

[https://www.youtube.com/watch?v=L5RUA2\\_eCIg&list=PLKdU0fuY4OFfWY36nDJDII26jXwInSm8f&index=34](https://www.youtube.com/watch?v=L5RUA2_eCIg&list=PLKdU0fuY4OFfWY36nDJDII26jXwInSm8f&index=34)

**Practical :**

<https://www.youtube.com/watch?v=NXt1fazDGu0&list=PLKdU0fuY4OFfWY36nDJDII26jXwInSm8f&index=35>

**Naive Bayes :****Theoretical :**

<https://www.youtube.com/watch?v=nQ2YHQagK3g&list=PLKdU0fuY4OFfWY36nDJDII26jXwInSm8f&index=36>

**Practical :**

<https://www.youtube.com/watch?v=kTf6a4DiAEY&list=PLKdU0fuY4OFfWY36nDJDII26jXwInSm8f&index=37>