An Open Access Journal Available Online

Using Four Learning Algorithms for Evaluating Questionable Uniform Resource Locators (URLs)

Nureni Ayofe Azeez[1] & Opeyemi Imoru[2]

[1,2]Department of Computer Sciences,
University of Lagos, Nigeria.
[1]nazeez@Unilag.edu.ng;
[2]opeimoru@gmail.com

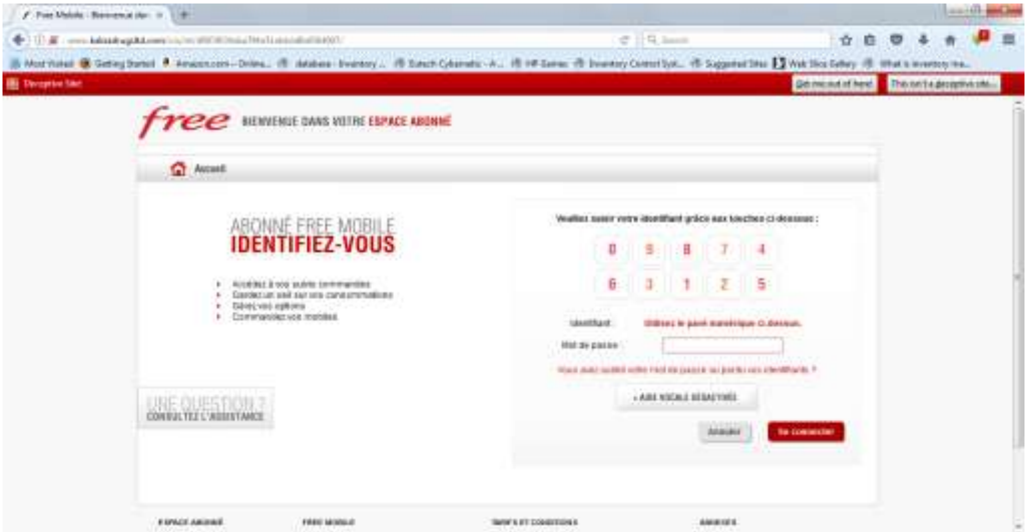
Abstract: Malicious Uniform Resource Locator (URL) is a common and serious threat to cyber security. Malicious URLs host unsolicited contents (spam, phishing, drive-by exploits, etc.) and lure unsuspecting internet users to become victims of scams such as monetary loss, theft, loss of information privacy and unexpected malware installation. This phenomenon has resulted in the increase of cybercrime on social media via transfer of malicious URLs. This situation prompted an efficient and reliable classification of a web-page based on the information contained in the URL to have a clear understanding of the nature and status of the site to be accessed. It is imperative to detect and act on URLs shared on social media platform in a timely manner. Though researchers have carried out similar researches in the past, there are however conflicting results regarding the conclusions drawn at the end of their experimentations. Against this backdrop, four machine learning algorithms:Naïve Bayes Algorithm, K-means Algorithm, Decision Tree Algorithm and Logistic Regression Algorithm were selected for classification of fake and vulnerable URLs. The implementation of algorithms was implemented with Java programming language. Through statistical analysis and comparison made on the four algorithms, Naïve Bayes algorithm is the most efficient and effective based on the metrics used.

Keywords: Malicious URLs; Pharming; Phishing, Attacks; Naïve Bayesian classifier; Decision tree, logistic Regression, k-means

1. Introduction

In the world today, online social networks have become powerful information diffusion platforms as they have attracted hundreds of millions of users. Online Social Networks (Guille et. al., 2013) (OSN) have changed the way people pursue social life and made it easy to connect with family members, classmates, friends and colleagues. In modern times, with increase in population the OSNs have become an easy and a much efficient platform in maintaining social relationships. Online Social Network sites like Facebook, YouTube, Badoo, Twitter, Linkedln, MySpace or Google+ have become popular sites on the Internet. They have attracted all ages from technicians to novice users. In the wide area sphere like research, working office, news media, organizations, entrepreneurship, industries, businesses, OSN have become a daily practice in use (Rao & Saleem, 2015). Most OSN are mainly used for information sharing and to express common interest views like political view, football discussion as well as fashion views etc. (Azeez et. al., 2014).

Its popular usage has been a major concern for the information technology society and experts and has alerted stakeholders to strengthen their defense against unauthorized entities such as malicious programs, Trojan horses, hackers, viruses etc. As online social networks sites have raised in popularity, cyber-criminals started to exploit these sites to spread malware and to carry out frauds (Rao & Saleem, 2015). Recent studies find that around 25% of all status messages in these systems contain URLs, amounting to millions of URLs shared per day. With this opportunity come challenges however from malicious users who seek to promote phishing, malware and other low quality content (Cao & Caverlee, 2015). The theft attacks such as phishing, pharming and spamming that are encountered by malicious e-mail URLs result in several loss to user and may lead to low usage of online services or e-commerce services. As a result of this negative occurrence and unfavorable experience, the authors propose a research work titled “investigating the performance of four learning algorithms for detecting fake and compromised urls”. Classification of URLs was based on their lexical features and host-based features and the Naïve Bayes Algorithm, Decision tree model algorithm (ID3) (Azeez & Iliyas, 2016), K means and Logical Regression model Algorithm were used as a probabilistic model to detect if a URL is malicious or legitimate. Figure 1 is a sample phishing website.



Deceptive Site!

Get me out of here! This isn't a deceptive site...

free BIENVENUE DANS VOTRE ESPACE ABONNÉ

Accueil

ABONNÉ FREE MOBILE
IDENTIFIEZ-VOUS

- Accédez à vos suivis commandes
- Gardez un oeil sur vos consommations
- Gérez vos options
- Commandez vos mobiles

UNE QUESTION ?
CONSULTEZ L'ASSISTANCE

Veuillez saisir votre identifiant grâce aux touches ci-dessous :

0 9 8 7 4
6 3 1 2 5

Identifiant Utilisez le pavé numérique ci-dessus.

Mot de passe

Vous avez oublié votre mot de passe ou perdu vos identifiants ?

AIDE VOCALE DÉSACTIVÉE

Annuler Se connecter

ESPACE ABONNÉ FREE MOBILE TARIFS ET CONDITIONS ANNEXES

Figure 1: Sample Phishing Website

2. Background/Related Work

Online learning algorithms like Perceptron, Logistic Regression with Stochastic Gradient Descent, Passive Aggressive (PA) Algorithm and Confidence Weighted (CW) Algorithms can be used to detect malicious URLs. Online algorithms are not only used to process large numbers of URLs more efficiently than batch algorithms they can also adapt more quickly to new features in the continuously evolving distribution of malicious URLs as compared to batch learning algorithms. These features include lexical URL features, IP address properties, WHOIS properties, domain name properties, blacklist membership, geographic properties and connection speed. (Ma et. al., 2011) developed a real time system for gathering URL features and compared it with a real time feed of labeled URLs from a large Web mail provider. Using these features and labels, they were able to train an online classifier that detected malicious Websites with 99% accuracy over a balanced dataset. (Ma et. al., 2011) Presented a novel two stage classification model to detect malicious Web pages (Azeez & Venter 2013). They divided the detection process into two stages. In the first stage they have estimated the maliciousness of Web pages using static features.

In the second stage, they used the potential malicious webpages found in the first stage for final identification of malicious web pages by extracting run time features of these webpages (Azeez, 2013). They extracted the static features from contents or properties of webpages without rendering fully or executing the webpages. Potential run time features like foreign contents, script contents and exploit contents were extracted by rendering webpages fully and executing them on specific systems. They used scoring algorithm for the classification.

(Qi & Davison, 2009) evaluated their scoring algorithm on the dataset of 20000 benign webpages for training and 13,646 instances of benign and malicious Web pages for testing. Web based classification approach was conducted which was a survey on the features and algorithms deployed for webpage classification.

The most common types of features used are the content features (text and HTML tags on the page), and Features of Neighbors (classification based on the class label of similar webpages). After the feature construction, standard classification techniques were applied, often with focus on multi-class classification and hierarchical classification (Azeez et. al., 2013). Like Spam detection, webpage classification also benefits significantly from text classification techniques.

(Gupta & McGrath, 2008) studied phishing infrastructure and the anatomy of phishing URLs. They pointed out the importance of features such as the length of the URL, age of linked/to domains, number of links present in the e/mails and the number of dots in the phishing URLs. (Sahoo et. al., 2017) Malicious URL Detection are broadly grouped into two major categories, (i) Blacklisting or Heuristics, and (ii) Machine Learning approaches.

- Blacklisting or Heuristic Approaches: Blacklisting approaches are a common and classical technique for detecting malicious URLs, which often maintains a list of URLs that are known to be malicious. Whenever a new URL is visited, a database lookup is performed. If the URL is present in the blacklist, it is considered to be malicious and then a warning will be generated; else it is assumed to be benign.
- Machine Learning: These approaches try to analyze the information of a URL and its corresponding websites or Webpages, by extracting good feature representations of URLs, and training a prediction model on training data of both malicious and benign URLs.

2.1 Url features

Phishing URLs can be examined based on two types of features: lexical features and host-based features of the URL. The lexical features analyse the format of the URL while the host based features identify the location, owner and how malicious sites are hosted and managed (Azeez & Ademolu 2016).

2.1.1 Lexical Features

According to (Azeez & Ademolu 2016), lexical features are the textual properties of the URL. It analyses the format of the URL not the content of the page it references. These properties include the length of the entire URL, presence of IP address in URL, the number of dots in the URL, presence of phishing keywords in URL, presence of suspicious characters such as @ symbol, hexadecimal characters and use of delimiters or special binary characters like “/”, “?”, “.”, “=”, “-”, “$”, “^” either in the host name or path (Dhanalakshmi & Chellappan, 2013).

a. Length of URL: Most phishing URLs use very large domain names to lure end-users so that the URL may appear legitimate. e.g. http://www.tsv1899benningen-ringen.de/chronik/update/alert/ibclogon.php.Thus, if the length of a URL is longer than 55 characters, the URL is flagged suspicious.
b. Use of IP address in URL : Some phishing websites contain an IP address in their URL instead of the domain name in order to hide the actual domain name which is malicious. When the URL in an email has its host name as an IP address. For example, in http://65.222.204.76/co/, we flag the URL suspicious.

c. Using the hexadecimal character codes : A malicious URL can also be represented using hexadecimal base values with a '%' symbol to hide the actual letters and numbers in the URL. Thus, a URL that has hexadecimal character codes will be flagged suspicious.

d. Use of @ symbol in URL : The '@' character is used by phishers to make host names difficult to understand. A @ symbol in a URL will enable the string to the left of the '@' symbol which is the actual legitimate URL to be discarded while the string to the right which leads to the phishing site is treated as the actual website. For example, in the URL http://www.worldbank.com@phishingsite.com, "www.worldbank.com" will be discarded

2.1.2 Host-Based Features

Host-based features describe the location of malicious sites, that is, where they are being hosted, who these sites are managed by and how they are managed. Some of these features are age of domain, page rank, number of domains (Azeez & Ademolu 2016).

a. Age of domain : The age of the domain identifies when a website is hosted such that a website that has less age or is relatively new is flagged suspicious. Many phishing sites have registered domain names that exist only for a short period of time to evade detection. They may be recently registered and some domains may not even be available at the time of checking. The WHOIS lookups on the WHOIS server is used to retrieve the domain registration date, and if the domain registration entry is not found on the WHOIS server, the URL is considered suspicious.

b. Presence of Form Tag : One of the methods phishers use to collect information from users is the use of form tag in URL. For example, <FORM action=http://www.paypalsite.com/profile.php method=post, the PayPal URL contains a form tag which has the action attribute actually sending the information to http://www.paypalsite.com/profile.php and not to http://www.paypal.com. Thus, a URL that has the form tag is flagged suspicious.

c. Number of Domains: A phishing URL may contain two or more domain names which are used to forward address from one domain to the other. For example, "http://www.google.com/url?sa=t&ct=res&cd=3&url=http%3A%2F%2Fwww.antiphishing.org%2F&ei=0qHRbWHK4z6oQLTmBM&usg=uIZX_3aJvESkMveh4uItI5DDUzM=&sig2=AVrQFpFvihFnLjpnGHVsxQ" has two domain names where "google.com" forwards the click to "antiphishing.org" domain name. The number of domain names in the URL extracted from an e-mail is counted and if more than one, we flag the URL suspicious.

3 Algorithms Considered

Four supervised machine learning classifiers (Naïve Bayes, Decision Tree, K-means and Logical Regression), were used for verification of fake URLS. They are briefly described below:

3.1. Naive-Bayes Classification Algorithm

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems (Mihaela, 2010).

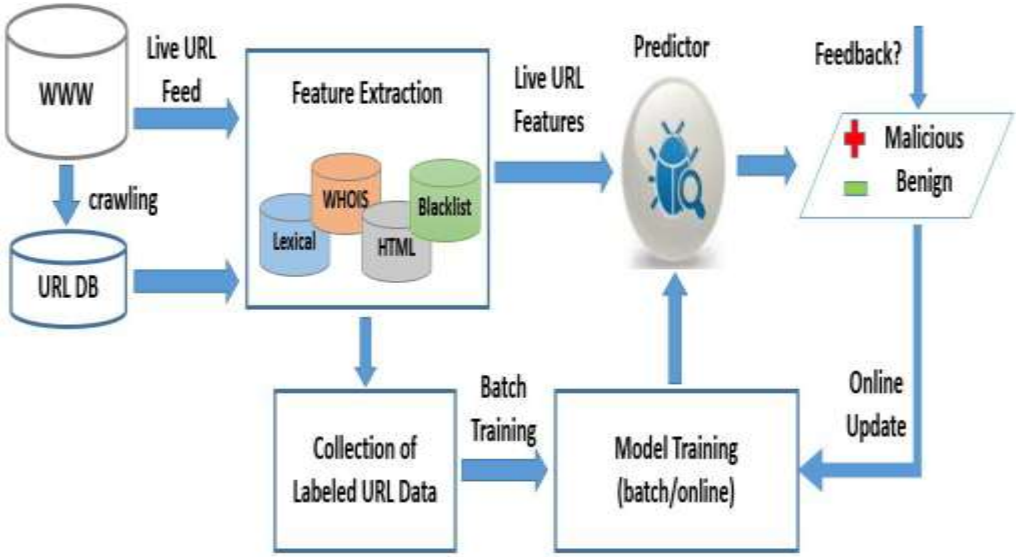


Figure 2: A framework for malicious url detection using machine learning (Sahoo et. al., 2017)

3.1.1 Uses of Naïve Bayes

1. Spam Filtering: It makes use of a naive Bayes classifier to identify spam e-mail. Bayesian spam filtering has become a popular mechanism to distinguish illegitimate spam email from legitimate email (sometimes called "ham" or "bacn").
2. Hybrid Recommender System Using Naive Bayes Classifier and Collaborative Filtering: Recommender Systems apply machine learning and data mining techniques for filtering unseen information and can predict whether a user would like a given resource. It is proposed a unique switching hybrid recommendation approach by combining a Naïve Bayes classification approach with the collaborative filtering.
3. Naive Bayes text classification: The Bayesian classification is used as a probabilistic learning method (Naive Bayes text classification). Naive Bayes classifiers are among the most successful known algorithms for learning to classify text documents (Mihaela, 2010).

3.1.2 Naïve Bayes classifier

$p(c_j|d) = \mathit{Probability\ of\ class\ c_j,}$
$\mathit{given\ that\ we\ have\ observed\ d}$

$$p(c_j|d) = \frac{p(d|c_j)\,p(c_j)}{p(d)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots 1$$

$p(c_j|d)$ = Probability of instance d being in class c_j,

$p(d|c_j)$ = Probability of generating instance d given class c_j,

$p(c_j)$ = Probability of occurrences of class c_j,

$p(d)$ = Probaility of instance d occurring

Bayes classification for more features

To simplify the task naïve Bayesian classifiers assumes attributes have independent distribution and there by estimate

$p(d|c_j)=p(d_1|c_j) \times p(d_2|c_j) \times p(d_3|c_j) \times \ldots \times p(d_n|c_j)$ ……...2

$p(d|c_j) =$ Probability of class c_j generating instance d

$p(d_1|c_j) =$ The probability of class c_j generating the observed value for feature 1,

$p(d_2|c_j) =$ The probability of class c_j generating the observed value for feature 2,

$p(d_3|c_j) =$ The probability of class c_j generating the observed value for feature 3

3.2 Decision Tree Model Algorithm

The core algorithm for building decision trees called ID3 by J. R. Quinlan which employs a top-down, greedy search through the space of possible branches with no backtracking. ID3 uses Entropy and Information Gain to construct a decision tree.

Entropy is a measure of uncertainty associated with a random variable

For a discrete random variable Y taking m distinct values $\{y_1, \ldots, y_m\}$

$H(Y) = -\sum_{i=1}^{m} p_i \log(p_i)\ , where\ p_i = P(Y = y_i)$ .3

Conditional Entropy

$H(Y|X) = \sum_x p(x)\ H(Y|X = x)$………....4

Select the attribute with the highest information gain

Let pi be the probability that an arbitrary tuple in D belongs to class Ci, estimated by |Ci, D|/|D|

Expected information (entropy) needed to classify a tuple in D:

$info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$ ……….5

Information needed (after using A to split D into v partitions) to classify D:

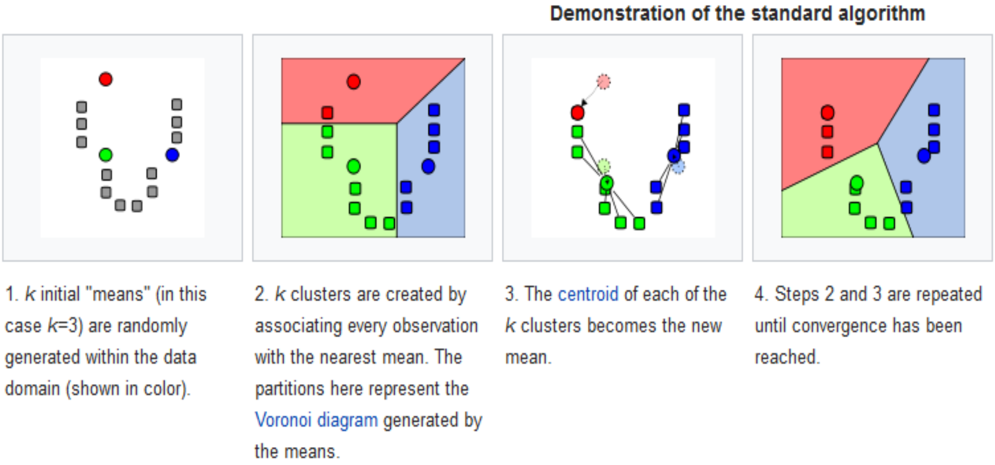
$info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times info(D_j)$…….6

Information gained by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$

3.3 K-Means

This is the most commonly used algorithm for an iterative refinement technique. Due to its ubiquity, it is often called the k-means algorithm; it is also referred to as Lloyd's algorithm, particularly in the computer science community. Lloyd's algorithm is based on the simple observation that the optimal placement of a center is at the centroid of the associated cluster (Faber, 1994). The main advantages of this algorithm are its simplicity and speed which allows it to run on large datasets. Its disadvantage is that it does not yield the same result with each run, since the resulting clusters depend on the initial random assignments (the k-means++ algorithm addresses this problem by seeking to choose better starting clusters).



Demonstration of the standard algorithm

1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3. The centroid of each of the *k* clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

Figure 3: Demonstration of the Standard Algorithm (Guido, 2014).

3.3.1 K-means Algorithm

Given a set of observations (x1, x2, …, xn), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k sets (k ≤ n) S = {S1, S2, …, Sk} so as to minimize the within-cluster sum of squares (WCSS):

$J = \sum_{j=1}^{K}\sum_{n\epsilon S_j} |x_n - \mu_j|^2$ ……………..7

Where $|x_n - \mu_j|^2$ is a chosen distance

measure between a data point and the cluster centre is an indicator of the distances of the n data points from their respective cluster centers.

Steps In k means algorithm

3.3.1.1 Assignment step: Assign each observation to the cluster with the closest mean

$S_i^{(t)} = \{x_j{:}||x_j - m_i^{(t)}|| \le || x_j - m_{i*}^{(t)}||\ for\ all\ i^* = 1, \ldots\ldots, k\}$

3.3.1.2 Update step: Calculate the new means to be the centroid of the observations in the cluster.

$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|}\sum_{x_j \in S_i^{(t)}} X_j$ ……………...…..8

Complexity of k means algorithm is given by: Complexity is O (n * K * I * d) n = number of points, K = number of clusters, I = number of iterations, d = number of attributes.

3.4 Logistic Regression

Logistic regression is used to obtain odds ratio in the presence of more than one explanatory variable. The procedure is quite similar to multiple linear regression, with the exception that the response variable is binomial. The result is the impact of each variable on the odds ratio of the observed event of interest. The main advantage is to avoid confounding effects by analyzing the association of all variables together (Sperandei, 2014).

The goal of logistic regression is to find the best fitting (yet biologically reasonable) model to describe the relationship between the dichotomous characteristic of interest.

Logistic regression models the probability of an event occurring depending on the values of the independent variables which can be categorical or numerical.

$P = {}^{outcome\ of\ interest}/_{all\ possible\ outcome}$

Odds of an event are the ratio of the probability that an event will occur to the probability that it will not occur. If the probability of an event occurring is p, the probability of the event not occurring is (1-p).

$$odds = \frac{p\,(occuring)}{P(not\,occuring)} = \frac{p}{1-p} \ldots\ldots 9$$

3.4.1 Odd ratio in logistic regression

Odd ratio is the ratio of two odd, the odds ratio (OR) is a comparative measure of two odds relative to different events

$$odds\,ratio = \frac{odds_1}{odds_2} = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_2}}$$

The dependent variable of logistic regression follows the Bernoulli distribution having an unknown probability. Bernoulli distribution is a special case of the binomial distribution where n =1 legitimate is "1" Malicious is "0".

$P(legitimate) = p\,and\,P(Malicious) = 1-$

p

In logistic regression we are estimating an unknown p for a given linear combination of the independent variable. We link together our independent variables to the Bernoulli distribution, the link is called Logit. The goal of logistic regression is to estimate p for a linear combination of the independent variables and, estimate of p is to tie together our linear combination of variables that could result unto the Bernoulli probability distribution with a domain from 0 to 1.

$$\text{logit}\,(p) = \ln(odds) = \ln\left(\frac{p}{1-p}\right) = a + \beta x \ldots\ldots 10$$

Where p is the probability of interested outcome and x is the explanatory variable. The parameters of the logistic regression are α and β. This is the simple logistic model. Taking the antilog of equation (1) on both sides, one can derive an equation for the prediction of the probability of the occurrence of interested outcome as

$$p = P(Y = interested\,outcome\,X = x, a\,specific\,value)$$

$$= \frac{e^{a+\beta x}}{1+\,e^{a+\beta x}} = \frac{1}{1+\,e^{a+\beta x}} \ldots\ldots\ldots 11$$

Extending the logic of the simple logistic regression to multiple predictors,

one may construct a complex logistic regression as

$$logit\,(y) = \ln\left(\frac{p}{1-p}\right) = a + \beta_1 X_1 + \cdots + \beta_k X_k \ldots 12$$

Therefore

$$p = P(Y = interested\,outcome\,X = x_1 \ldots\ldots \ldots X = x_k)$$

$$= \frac{e^{a+\beta_1 X_1+\cdots+\beta_k X_k}}{1+e^{a+\beta_1 X_1+\cdots+\beta_k X_k}} = \frac{1}{1+e^{a+\beta_1 X_1+\cdots+\beta_k X_k}} \ldots\ldots 13$$

A simple logistic function is defined by the formula

$$y = \frac{e^x}{1+e^x} = \frac{1}{1-e^{-x}} \ldots\ldots\ldots\ldots\ldots\ldots 14$$

To provide flexibility, the logistic function can be extended to the form

$$y = \frac{e^{a+\beta x}}{1+\,e^{a+\beta x}} = \frac{1}{1+\,e^{a+\beta x}} \ldots\ \ldots\ldots 15$$

Where α and β determine the logistic intercept and slope. Logistic regression fits α and β, the regression coefficients.. The logistic or logit function is used to transform an 'S'-shaped curve into an approximately straight line and to change the range of the proportion from

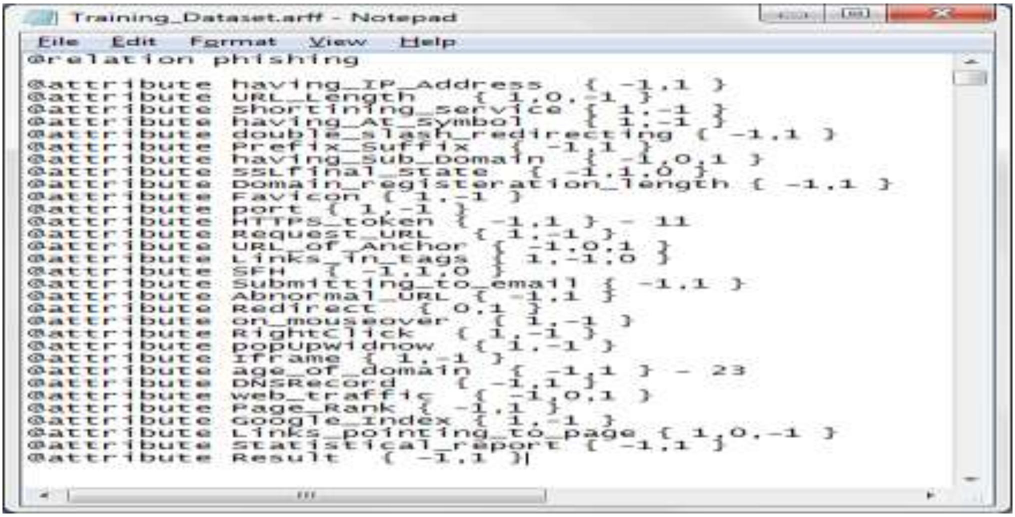
0 – 1 to -∞ - +∞ as

$$\text{logit}\,(p) = \ln(odds) = \ln\left(\frac{p}{1-p}\right) = a + \beta x \ldots 16$$

Where p is the probability of interested outcome, α is the intercept parameter,β is a regression coefficient, and χ is a predictor.

4 Implementation, Findings and Results

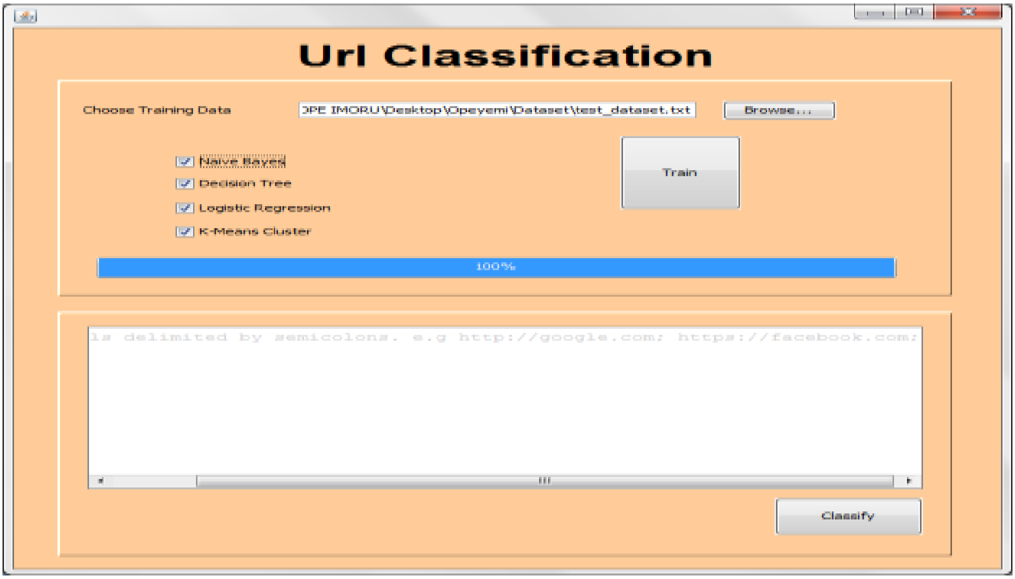
The system is a web based application, it classifies a URL as malicious or legitimate based on lexical features and host based features. Four machine learning algorithms which are all supervised learning algorithms (Naïve Bayes algorithm, decision tree algorithm, k means algorithm and logistic regression algorithm) were used to classify the URL. Based on the trained features, the system classifies the URL as malicious else it is classified as legitimate. The collected features include both URL-based features and host-based features. The verification of fake urls using supervised learning algorithm based on repetitive and redundancy values have been implemented with java programming language in the Netbeans integrated Development Environment (IDE) and are tested against 200 URLs. This has been done to determine the algorithm that has the highest maximal level of effectiveness, accuracy and efficiency. Some of the collected features hold categorical values termed as ‘‘Legitimate ’and Malicious’’, these values have been replaced with numerical values 1, 0 and -1 instead of ‘‘Legitimate’’, ‘‘Malicious’’ and ‘‘Suspicious’’ respectively.



```
@relation phishing

@attribute having_IP_Address  { -1,1 }
@attribute URL_Length   { 1,0,-1 }
@attribute Shortining_Service { 1,-1 }
@attribute having_At_Symbol   { 1,-1 }
@attribute double_slash_redirecting { -1,1 }
@attribute Prefix_Suffix  { -1,1 }
@attribute having_Sub_Domain  { -1,0,1 }
@attribute SSLfinal_State  { -1,1,0 }
@attribute Domain_registeration_length { -1,1 }
@attribute Favicon { 1,-1 }
@attribute port { 1,-1 }
@attribute HTTPS_token { -1,1 } - 11
@attribute Request_URL  { 1,-1 }
@attribute URL_of_Anchor { -1,0,1 }
@attribute Links_in_tags { 1,-1,0 }
@attribute SFH  { -1,1,0 }
@attribute Submitting_to_email { -1,1 }
@attribute Abnormal_URL { -1,1 }
@attribute Redirect  { 0,1 }
@attribute on_mouseover  { 1,-1 }
@attribute RightClick  { 1,-1 }
@attribute popUpWidnow  { 1,-1 }
@attribute Iframe { 1,-1 }
@attribute age_of_domain  { -1,1 } - 23
@attribute DNSRecord   { -1,1 }
@attribute web_traffic  { -1,0,1 }
@attribute Page_Rank { -1,1 }
@attribute Google_Index { 1,-1 }
@attribute Links_pointing_to_page { 1,0,-1 }
@attribute Statistical_report { -1,1 }
@attribute Result  { -1,1 }
```

Figure 4: Lexical and Host-Based Features for url Classification



Url Classification

Choose Training Data: OPE IMORU\Desktop\Opeyemi\Dataset\test_dataset.txt

☑ Naive Bayes
☑ Decision Tree
☑ Logistic Regression
☑ K-Means Cluster

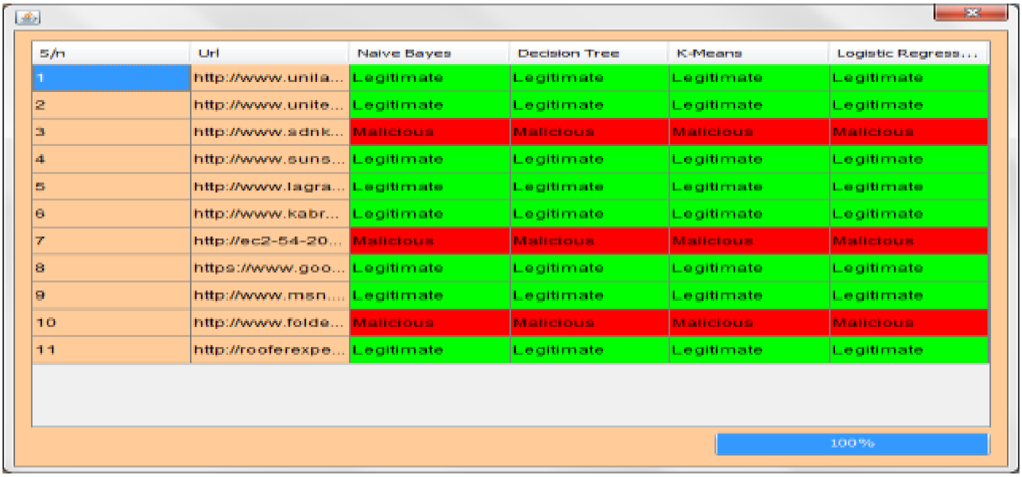
100%

ls delimited by semicolons. e.g http://google.com; https://facebook.com;

Figure 5: Initialization of the Program/ Training Dataset File

Sample of the url classification (2 url examples) for all features is shown in diagrams below:

http://www.Unilag.edu.ng;
http://www.unitedhealthgroup.com
http://www.sdnkasepuhan02btg.sch.id/cana/a2a7938099b2075bd8b9b69804524753/;
http://www.sunsofttec.com/eeeettt/2185266aadae98f002016e352372bba8/;
http://www.lagranherramienta.com/easy/ayo/ayo1/;
http://www.kabradrugsltd.com/css/nt/df6f3f034aba794e31abbdd8a0564007/;
http://ec2-54-200-151-255.us-west-2.compute.amazonaws.com/-/accord2/;
https://www.google.com/;
http://www.msn.com/en-us?cobrand=hp-notebook.msn.com&OCID=HPDHP&pc=HPNTDF;
http://www.folder365.world/yawa/aptgd/;
http://rooferexpert.com/css/8933617-dosar-nr-1817842015/394c-4735-82399c8f64a5248/botosani_firme/ec77154aef4d9311a65613d9a59cf370/;



S/n	Url	Naive Bayes	Decision Tree	K-Means	Logistic Regress...
1	http://www.unila...	Legitimate	Legitimate	Legitimate	Legitimate
2	http://www.unite...	Legitimate	Legitimate	Legitimate	Legitimate
3	http://www.sdnk...	Malicious	Malicious	Malicious	Malicious
4	http://www.suns...	Legitimate	Legitimate	Legitimate	Legitimate
5	http://www.lagra...	Legitimate	Legitimate	Legitimate	Legitimate
6	http://www.kabr...	Legitimate	Legitimate	Legitimate	Legitimate
7	http://ec2-54-20...	Malicious	Malicious	Malicious	Malicious
8	https://www.goo...	Legitimate	Legitimate	Legitimate	Legitimate
9	http://www.msn....	Legitimate	Legitimate	Legitimate	Legitimate
10	http://www.folde...	Malicious	Malicious	Malicious	Malicious
11	http://rooferexpe...	Legitimate	Legitimate	Legitimate	Legitimate

Figure 6: Classification of 10 Samples url

Table 1: Breakdown of Naïve Bayes Classifier for www.Unilag.Edu.Ng

URL FEATURES	LEGITIMATE	MALICIOUS
NOIPADDRESS	0.855932	0.872549
LEGITIMATEURL	0.235294	0.145631
NORMALURL	0.79661	0.823529
NOATSYMBOL	0.974576	0.941176
NODOUBLESPLASH	0.813559	0.77451
NOPREFIXSUFIX	0.288136	0.009804
LEGITIMATEDOMAIN	0.352941	0.495146
MALICIOUSSSL	0.016807	0.242718
MALICIOUSREGISTRATIONLENGTH	0.220339	0.421569
NOHTTPSTOKENDMAIN	0.635593	0.627451
DOMAINAGEOLDERTHAN6MONTHS	0.677966	0.411765
HASDNSRECORD	0.542373	0.264706

Table1 shows the Naïve Bayes mathematical breakdown of the url features for www.Unilag.edu.ng. From Table 1, it was deduced that the Unilag url is a legitimate url based on the features.

4.1 Breakdown and Graphical Classification of Legitimate url

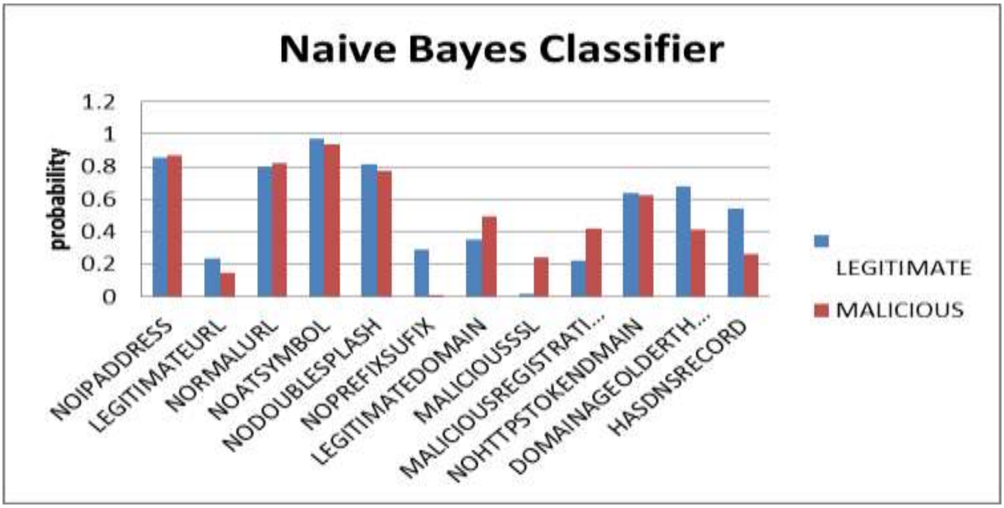


Figure 7: Graphical Figure of Naïve Bayes Classifier for www.unilag.edu.ng

Figure 7 is a graphical representation that shows the url features of legitimate and malicious breakdown as depicted in Table 1.

Table 2: Breakdown of Decision Tree www.unilag.edu.ng

URL FEATURES	LEGITIMATE	MALICIOUS
NOIPADDRESS	0.14	0.13
LEGITIMATEURL	0.24	0.15
NORMALURL	0.8	0.82
NOATSYMBOL	0.97	0.94
NODOUBLESPLASH	0.81	0.77
NOPREFIXSUFIX	0.29	0.01
LEGITIMATEDOMAIN	0.35	0.5
MALICIOUSSSL	0.02	0.24
MALICIOUSREGISTRATIONLENGTH	0.22	0.42
NOHTTPSTOKENDMAIN	0.64	0.63
DOMAINAGEOLDERTHAN6MONTHS	0.32	0.59
HASDNSRECORD	0.54	0.26

Table 2 shows the mathematical breakdown of Decision Tree showing url features of Unilag website. From the table it was deduced that the Unilag url is a legitimate url based on the features as shown in Table 2.

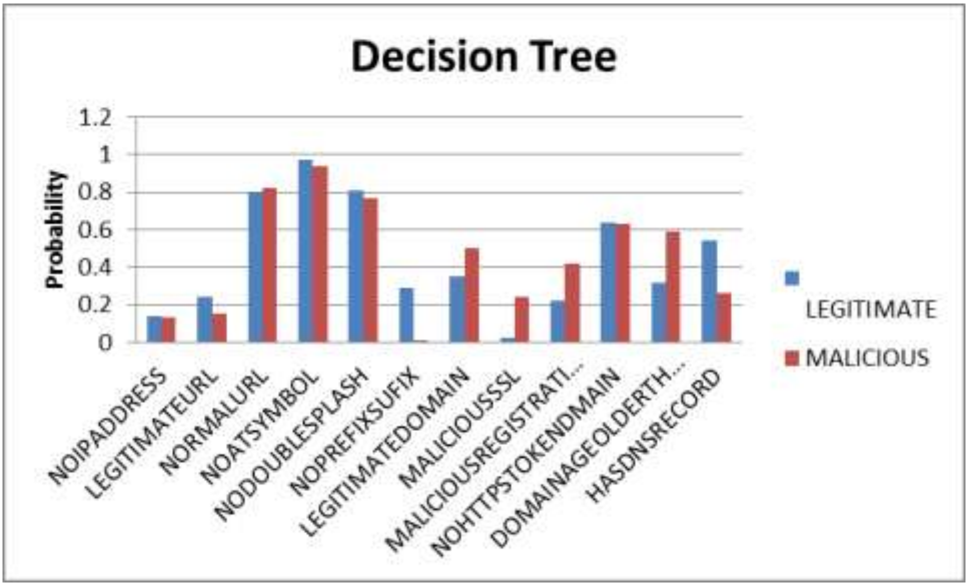


Figure 8: Graphical Figure of Decision Tree Classifier for www.unilag.edu.ng

Figure 8 is a graphical representation that shows the url feature of legitimate and malicious breakdown from Table 2.

Table 3: Breakdown of K-Means www.unilag.edu.ng

URL FEATURES	LEGITIMATE	MALICIOUS
NOIPADDRESS	0.862069	0.88
LEGITIMATEURL	0.232759	0.14
NORMALURL	0.801724	0.83
NOATSYMBOL	0.982759	0.95
NODOUBLESPLASH	0.181034	0.22
NOPREFIXSUFIX	0.284483	0
LEGITIMATEDOMAIN	0.353448	0.5
MALICIOUSSSL	0.008621	0.24
MALICIOUS REGISTRATION LENGTH	0.215517	0.42
NOHTTPSTOKENDMAIN	0.637931	0.63
DOMAINAGEOLDERTHAN6MONTHS	0.681034	0.41
HASDNSRECORD	0.543103	0.26

Table 3 shows the numerical values of k-means url features of Unilag website. From the table, it was deduced that the Unilag url is a legitimate url based on the features depicted in Table 3.

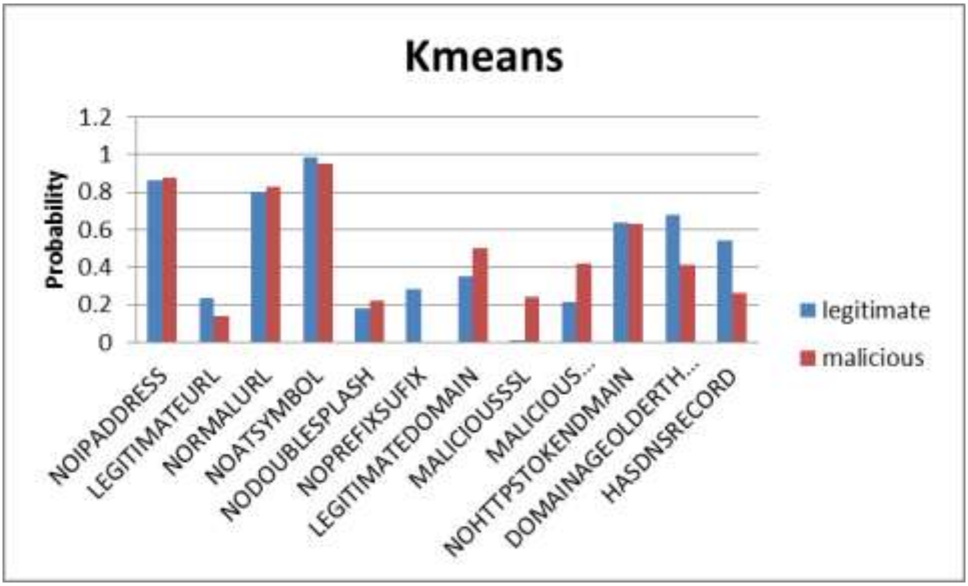


Figure 9: Graphical Figure of K Means Classifier for www.unilag.edu.ng

Figure 9 is a graphical representation that shows the k-means interpretation of url feature of legitimate and malicious breakdown as depicted in Table 3.

Table 4: Breakdown of Logistic Regression www.unilag.edu.ng

URL FEATURES	WEIGHT
NO IPADDRESS	-0.246116026
LEGITIMATE URL LENGTH	0.406965719
NORMAL URL	-0.023747636
NO ATSYMBOL	0.222137496
NO DOUBLESPLASH	0.159693131
NO PREFIXSUFIX	1.045097936
LEGITIMATE DOMAIN	-0.391132699
MALICIOUS SSL	-1.463416988
MALICIOUS REGISTRATION LENGTH	-0.084222514
NOHTTPSTOKENDMAIN	0.031575833
DOMAIN AGE OLDER THAN6MONTHS	0.526511068
HASDNSRECORD	0.580109969

Table 4 shows the numerical values obtained for Logistic Regression of url features of Unilag website. From the table it was deduced that the Unilag url is a legitimate url based on the features used.

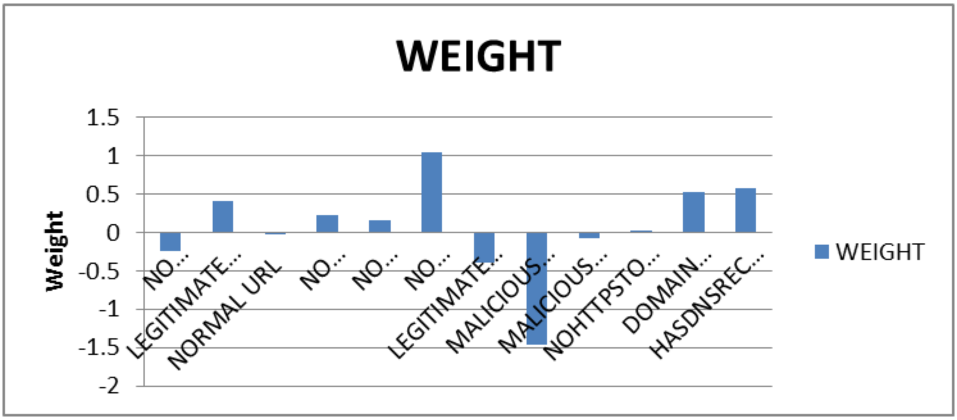


Figure 10: Graphical Figure of Logistic Regression Classifier for www.unilag.edu.ng

The above graphical representation shows the Logistic Regression interpretation of url feature of legitimate and malicious breakdown as depicted in Table 4.

4.2 Breakdown and Graphical Classification Of Malicious url

Table 5; Naïve Bayes Breakdown Details of http://www.sdnkasepuhan02btg.sch.id/cana/A2a7938099b2075bd8b9b69804524753/

URL FEATURES	Legitimate	Malicious
NOIPADDRESS	0.855932	0.872549
LEGITIMATEURL	0.058824	0.087379
NORMALURL	0.79661	0.823529
NOATSYMBOL	0.974576	0.941176
NODOUBLESPLASH	0.813559	0.77451
HASPREFIXSUFIX	0.711864	0.990196
MALICIOUSDOMAIN	0.369748	0.174757
MALICIOUSSSL	0.016807	0.242718
MALICIOUSREGISTRATIONLENGTH	0.220339	0.421569
NOHTTPSTOKENDMAIN	0.635593	0.627451
DOMAINAGEOLDERTHAN6MONTHS	0.322034	0.588235
HASDNSRECORD	0.457627	0.735294
	6.233513	7.279363

Table 5 shows the values obtained for Naïve Bayes of url features of Unilag website. From the table it was deduced that the url is malicious based on the url features in the table.

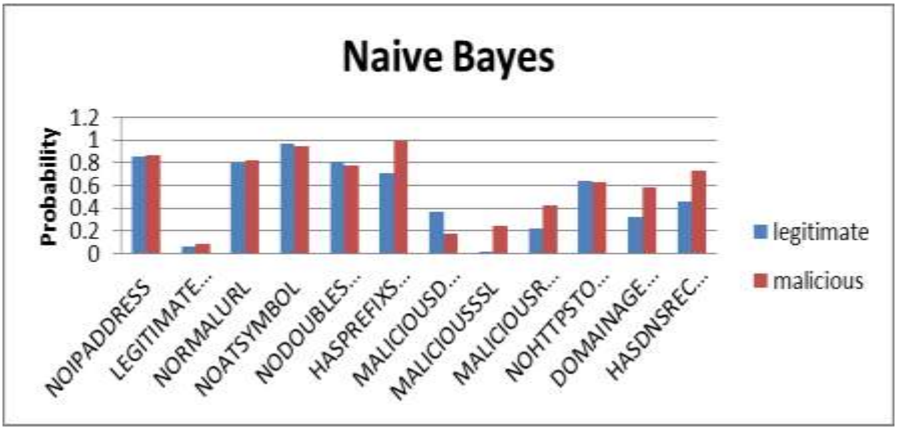


Figure 11: Graphical Figure of Naïve Bayes Classifier for http://www.sdnkasepuhan02btg.sch.id/cana/a2a7938099b2075bd8b9b69804524753/

Figure 11 shows a graphical representation of the values obtained in Table 5 for the Naïve Bayes.

Table 6: Decision Tree Breakdown Details of http://www.sdnkasepuhan02btg.sch.id/cana/a2a7938099b2075bd8b9b69804524753/

URL CLASSIFICATION	LEGITIMATE	MALICIOUS
NOIPADDRESS	0.86	0.87
LEGITIMATEURL	0.24	0.15
NORMALURL	0.8	0.82
NOATSYMBOL	0.97	0.94
NODOUBLESPLASH	0.81	0.77
HASPREFIXSUFIX	0.71	0.99
MALICIOUSDOMAIN	0.37	0.17
MALICIOUSSSL	0.02	0.24
MALICIOUSREGISTRATIONLENGTH	0.22	0.42
NOHTTPSTOKENDMAIN	0.64	0.63
DOMAINAGEOLDERTHAN6MONTHS	0.68	0.41
N0DNSRECORD	0.46	0.76

The features depicted in Table 6 shows the url features used for Decision Tree classification. It also shows the values obtained for the malicious url.

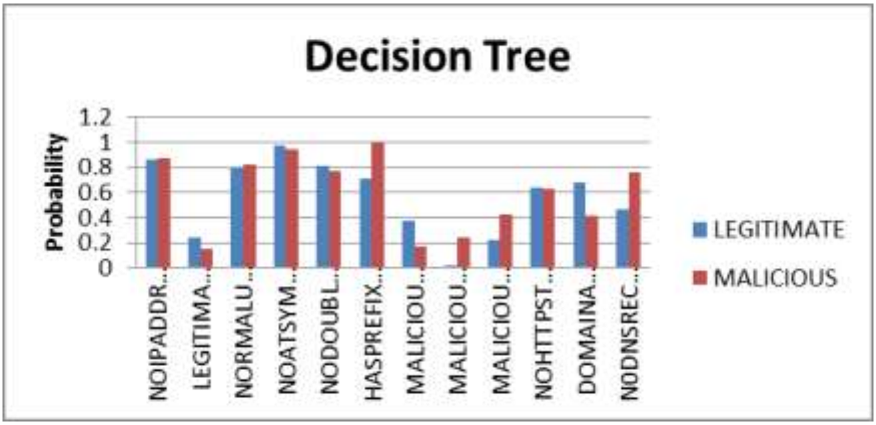


Figure 12: Graphical Figure of Decision Tree Classifier for http://www.sdnkasepuhan02btg.sch.id/cana/a2a7938099b2075bd8b9b69804524753/

Figure 12 is a graphical representation that shows the interpretation of Decision Tree for url features of both legitimate and malicious as depicted in Table 6 above.

Table 7: K-Means Showing the Breakdown of http://www.sdnkasepuhan02btg.sch.id/cana/a2a7938099b2075bd8b9b69804524753/

URL FEATURES	LEGITIMATE	MALICIOUS
NOIPADDRESS	0.862069	0.88
SUSPICIOUS URL LENGH	0.051724	0.08
NORMALURL	0.801724	0.83
NOATSYMBOL	0.982759	0.95
NODOUBLESPLASH REDIRECTING	0.181034	0.22
HASPREFIXSUFIX	0.715517	1
MALICIOUSDOMAIN	0.37069	0.17
MALICIOUSSSL	0.008621	0.24
MALICIOUSREGISTRATIONLENGTH	0.215517	0.42
NOHTTPSTOKENDMAIN	0.637931	0.63
DOMAINAGEOLDERTHAN6MONTHS	0.681034	0.41
N0DNSRECORD	0.456897	0.74

Table 7 shows the url features used for k-means classification. It shows the values obtained for both malicious and legitimate urls.

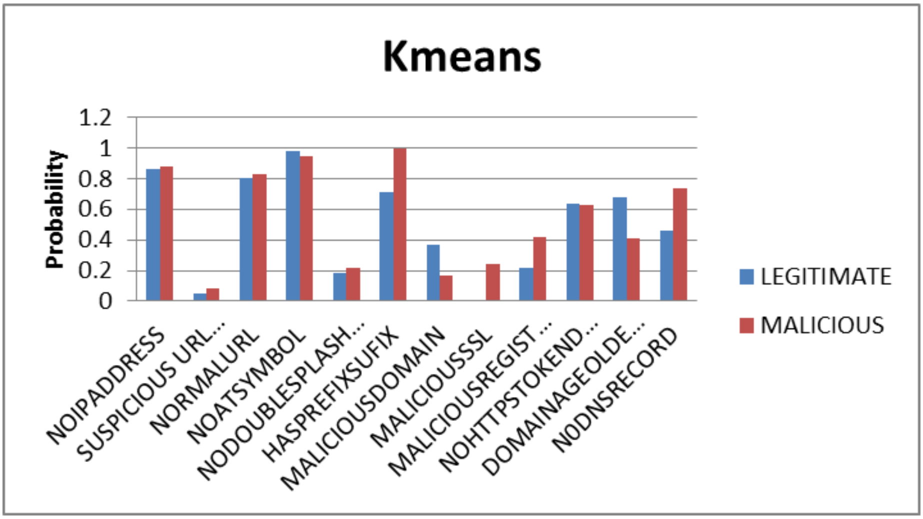


Figure 13: Graphical Figure of K-Means Classifier for http://www.sdnkasepuhan02btg.sch.id/cana/a2a7938099b2075bd8b9b69804524753/

Figure 13 is a graphical representation that shows the k-means interpretation of url features for legitimate and malicious as depicted in Table 7.

Table 8: Logical Regression Breakdown Details of HTTP://WWW.SDNKASEPUHAN02BTG.SCH.ID/CANA/A2A7938099B2075BD8B9B69804524753/

FEATURES	WEIGHT
NOIPADDRESS	-0.246116027
SUSPICIOUS URL LENGH	-0.063673963
NORMALURL	-0.023749636
NOATSYMBOL	0.222137497
NODOUBLESPLASH REDIRECTING	0.159693132
HASPREFIXSUFIX	-1.018083616
MALICIOUSDOMAIN	0.006899523
MALICIOUSSSL	-1.463416988
MALICIOUSREGISTRATIONLENGTH	-0.084222514
NOHTTPSTOKENDMAIN	-0.004561513
DOMAINAGEOLDERTHAN6MONTHS	0.526510685
N0DNSRECORD	-0.553095649

Table 8 shows the url features used for Logistic Regression classification. It shows the corresponding values obtained.

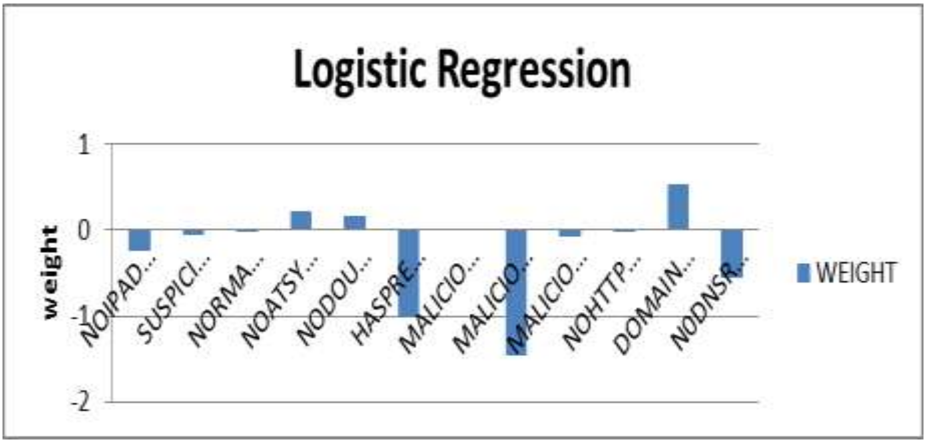


Figure 14: Graphical Figure of Logistic Regression Classifier for http://www.sdnkasepuhan02btg.sch.id/cana/a2a7938099b2075bd8b9b69804524753/

Figure 14 is a graphical representation that shows the logistic regression interpretation of url features for legitimate and malicious as contained in Table 8.

5. Conclusion

A study and evaluation of four Machine Learning Algorithms for evaluating legitimacy of urls has been successfully carried out. The algorithms were implemented and tested with different dataset. A comparison of all four algorithms was done to know their level of efficiency and effectiveness in detecting and evaluating both legitimate and malicious urls. It is of note that twelve different url features were considered and evaluated for each of the algorithms. With the available results, as observed in the numerical values and graphical representations for the experimentation, the Naïve Bayes Algorithm is considered to be the most effective and efficient of all the four machine learning algorithms evaluated. Naïve Bayes Algorithm yielded good results for detecting legitimate and malicious values when tested with the same url under the same features. Some future works, therefore for this admirable research work include the development of a new algorithm that can be more accurate than Naïve Bayes algorithm. This can be achieved by hybridizing two or more supervised learning algorithms in order to have a more accurate, efficient and reliable url legitimate evaluation.

References

Ayofe, A.N, Adebayo, S.B, Ajetola, A.R, Abdulwahab, A.F (2010) "A framework for computer aided investigation of ATM fraud in Nigeria" International Journal of Soft Computing, Vol. 5, Issue 3 pp. 78-82

Azeez, N.A., and Lasisi, A. A. (2016). Empirical and Statistical Evaluation of the Effectiveness of Four Lossless Data Compression Algorithms. Nigerian Journal of Technological Development, Vol. 13, NO. 2, December 2016, 64-73.

Azeez, N.A and Otudor, A.E. (2016) "Modelling and Simulating Access Control in Wireless Ad-Hoc Networks" Fountain Journal of Natural and Applied Sciences. Vol 5(2), pp 18-30

Azeez, N. A., & Ademolu, O. (2016). CyberProtector: Identifying Compromised URLs in Electronic Mails with Bayesian Classification. 2016 International Conference Computational Science and Computational Intelligence (CSCI) (pp. 959-965). Las Vegas, NV, USA: IEEE.

Azeez, N. A., & Iliyas, H. D. (2016). Implementation of a 4-tier cloud-based architecture for collaborative health care delivery. Nigerian Journal of Technological Development, 13(1), 17-25.

Azeez, N. A., & Venter, I. M. (2013). Towards ensuring scalability, interoperability and efficient access control in a multi-domain grid-based environment. SAIEE Africa Research Journal, 104(2), 54-68.

Azeez, N.A Abidoye, A.P Adesina, A.O Agbele, K.K Venter, I.M Oyewole, A.S (2013) "Statistical Interpretations of the Turnaround Time Values for a scalable 3-tier grid-based Computing architecture" Computer Science & Telecommunications, Vol 39 (3), pp 67-75.

Azeez, N. A., Iyamu, T., and Venter, I. M. (2011). Grid security loopholes with proposed countermeasures. In E. Gelenbe, R. Lent, and G. Sakellari (Ed.), 26th International Symposium on Computer and Information Sciences (pp. 411-418). London: Springer.

Azeez, N.A and Venter, I.M (2012). Towards achieving scalability and interoperability in a triple-domain grid-based environment (3DGBE)- Information Security for South Africa (ISSA), 2012, pp 1-10.

Azeez, N. A., and Babatope, A. B. (2016). AANtID: an alternative approach to network intrusion detection. The Journal of Computer Science and its Applications. An International Journal of the Nigeria Computer Society, 129-143.

Azeez, N. A. (2012). Towards Ensuring Scalability, Interoperability and Efficient Access Control In a Triple-Domain Grid-Based Environment. Cape Town: University of the Western Cape.

Cao, C. & Caverlee, J. 2015 Detecting Spam URLs in Social Media via Behavioral Analysis, Department of Computer Science and Engineering, Texas A&M University College Station, Texas, USA.

Dhanalakshmi, R., & Chellappan, C. (2013). Detecting Malicious URLs in E-mail - An Implementation. AASRI Procedia , 125-131.

Faber, V., 1994 Clustering and the Continuous k-means Algorithm, Los Alamos Science, vol. 22, pp. 138-144, 1994.

Gupta, M. & McGrath D. 2008, Behind phishing: an examination of phisher modi operandi, in proceedings of the USENIX Workshop on Large/scale Exploits and Emergent Threats (LEET), San Fransicso, CA, Apr 2008.

Guido, S., 2014 Kmeans Clustering With Scikit- Learn Https://Www.Slideshare.Net/Sarahguido/Kmeans-Clustering-With-Scikitlearn.

Guille, A., Hacid,H., Favre, C., Zighed, D.A. 2013: Information Diffusion in online Social Networks: A

Survey. ERIC Lab, Lyon 2 University, France, Bell Labs France, Alcatel-Lucent, France Institute of Human Science, Lyon 2 University, France. ACM. 2013 published in SIGMOD Record, Vol 42 ISS2, June 2013.

Ma, J., Saul, L., Savage, S. & Voelker,G. 2011. "Learning to Detect Malicious URLs",ACM Transactions on Intelligent Systems and Technology, vol. 2, no. 3, no. 30, (2011),pp. 30:1-30:24.

Mihaela M. 2010 : Naïve-Bayes Classification Algorithm. 7 May 2010 http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab4-NaiveBayes.pdf . .

Nureni, A. A., and Irwin, B. (2010). Cyber security: Challenges and the way forward. Computer Science & Telecommunications, 29, 56-69.

Qi, X. & Davison, B. 2009 Web Page Classification: Features and Algorithms ACM Comput. Surv. 41, 2, Article 12 (February 2009), 31 pages DOI = 10.1145/1459352.1459357 http://doi.acm.org/10.1145/1459352.1459357.

Rao, V. & Saleem, P.A., 2015 Twitter Adoption And Analysis Online Social Networks International Journal Of Global Innovations - Vol.2, Issue .I Paper Id: Sp-V2-I1-257 Issn, Dept Of Cse, Kkr And Ksr Institute Of Technology And Sciences (Kits) Guntur, A.P., India. 2015.

Sahoo, D. Liu,C & Steven C.H. Hoi 2017. Malicious URL Detection using Machine Learning: A Survey " arXiv:1701.07179v2 [cs.LG] 16 Mar 2017. https://arxiv.org/pdf/1701.07179.pdf

Sperandei, S. 2014 Understanding logistic regression analysis. Biochem Med (Zagreb). 2014 Feb; 24(1): 12–18.Published online 2014 Fe doi: 10.11613/BM.2014.003PMCID: PMC3936971

Srivastava, V.T. 2007 : Phishing and Pharming- The Deadly Duo, SANS Institute 2007 Accepted January 29, 2007. https://www.sans.org/reading-room/whitepapers/privacy/phishing-pharming-evil-twins-1731