

URL Based Fake Website Detection Using Machine Learning Algorithm

BY

MD Mursalin Dowla

ID: 181-15-10582

AND

K. M. Nayem Ahmed

ID: 181-15-10870

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science & Engineering

Supervised By

Most. Hasna Hena

Assistant Professor

Department of CSE

Daffodil International University

Co-Supervised By

Ms. Afsara Tasneem Misha

Lecturer

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

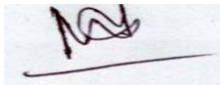
DHAKA, BANGLADESH

JANUARY 2022

APPROVAL

This Project titled “**URL Based Fake Website Detection Using Machine Learning**”, submitted by **MD Mursalin Dowla** and **K. M. Nayem Ahmed**, ID No: **181-15-10582** and **181-15-10870** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **02-01-2022**.

BOARD OF EXAMINERS

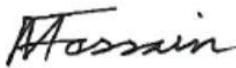


Chairman

Dr. Md. Ismail Jabiullah

Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

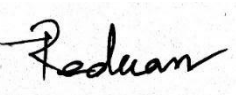


Internal Examiner

Dr. Md. Fokhray Hossain

Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Internal Examiner

Md. Reduanul Haque

Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



External Examiner

Dr. Mohammad Shorif Uddin

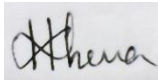
Professor

Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Most. Hasna Hena, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



Most. Hasna Hena

Assistant Professor
Department of Computer science and Engineering
Daffodil International University

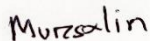
Co-Supervised by:



Ms. Afsara Tasneem Misha

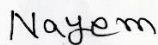
Lecturer
Department of Computer science and Engineering
Daffodil International University

Submitted by:



MD Mursalin Dowla

ID: 181-15-10582
Department of Computer science and Engineering
Daffodil International University



K. M. Nayem Ahmed

ID: 181-15-10870
Department of Computer science and Engineering
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes it possible to complete the final year project/internship successfully.

We are really grateful and wish profound our indebtedness to, **Most. Hasna Hena**, Assistant Professor, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Machine Learning*” to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to Dr. Touhid Bhuiyan, Professor and Head, Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Online platforms, mostly social media platforms have become a very important part of our life. We are very used to online content and sites. But many URLs lead us to fake sites. These are intentionally created to mislead users to gain certain information. This generally leads us to account hacking or information thieves. To identify these sites and stop users from using these URLs we will discuss machine learning algorithms and we will also have a dataset and apply all those algorithms on our dataset. Dataset was collected from various online open source platforms. A total of 20,000 data was collected and used, half of which was fake URLs and the other half was real URLs. First, we extracted many features from our initial dataset which was later used to train our model. We used an anaconda environment to implement our project. A Jupyter notebook was used to do the necessary codes. We were successful in extracting necessary features and applying machine learning algorithms. The dataset was divided into 80:20 ratio for training and testing purposes. The best supervised machine learning algorithms were chosen to train our model. Random Forest Classifier got the highest success from our model by gaining maximum accuracy. We got 97.50% accuracy from the Random Forest Classifier. Finally, the model was saved for later improvements. By this we believe we will have the best machine learning approach to detect fake content or sites that are online. Hopefully this will help detect online fake URLs and save users from its attacks.

TABLE OF CONTENTS

CONTENTS	PAGE
Approval Page	i
Declaration	ii
Acknowledgement	iii
Abstract	iv
List of Tables	vii
List of Figures	viii-ix

CHAPTER	PAGE
----------------	-------------

CHAPTER 1: INTRODUCTION	1-4
--------------------------------	------------

1.1 Introduction	1
1.2 Motivation	1-2
1.3 Rationale of the Study	2
1.4 Research Questions	3
1.5 Expected Output	3
1.6 Project Management and Finance	3-4
1.7 Report Layout	4

CHAPTER 2: BACKGROUND	5-9
------------------------------	------------

2.1 Terminologies	5
2.2 Related Works	5-7
2.3 Comparative Analysis and Summary	7-8
2.4 Scope of the Problem	9
2.5 Challenges	9

CHAPTER 3: RESEARCH METHODOLOGY	10-26
--	--------------

3.1 Research Subject and Instrumentation	10
3.2 Data Collection Procedure	10
3.3 Statistical Analysis	11

3.4 Proposed Methodology	11-26
3.5 Implementation Requirements.	26
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	27-30
4.1 Experimental Setup	27
4.2 Experimental Results & Analysis	27-29
4.3 Discussion	29-30
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	31-32
5.1 Impact on Society	31
5.2 Impact on Environment	31
5.3 Ethical Aspects	32
5.4 Sustainability Plan	32
CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH	33-34
6.1 Summary of the Study	33
6.2 Conclusions	33-34
6.3 Implication for Further Study	34
REFERENCES	35-37

LIST OF TABLES

TABLES	PAGE NO
Table 1: Accuracy of logistic regression	23
Table 2: Accuracy of decision tree classifier	24
Table 3: Accuracy of XGB classifier	25
Table 4: Accuracy of random forest classifier	25
Table 5: Accuracy of K-Nearest neighbors' classifier	25
Table 6: Accuracy of SVC classifier	26
Table 7: Accuracy of all algorithms	28
Table 8: Confusion matrix statistics	29

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.3.1: Initial datasets	11
Figure 3.4.1: Methodology	12
Figure 3.4.2: URL statistics	13
Figure 3.4.3: After Domain extraction	13
Figure 3.4.4: IP address-based feature extraction	14
Figure 3.4.5: Unsafe character-based feature extraction	14
Figure 3.4.6: URL length-based feature extraction	15
Figure 3.4.7: Storing domain length	16
Figure 3.4.8: Comparing domain length	16
Figure 3.4.9: Calculating URL depth	17
Figure 3.4.10: Slash based feature extraction	18
Figure 3.4.11: Checking for redirection in the URL	18
Figure 3.4.12: “Http” & “https” based feature extraction	19
Figure 3.4.13: Checking for URL shortening method	20
Figure 3.4.14: “-” & “@” based feature extraction	20

FIGURES	PAGE NO
Figure 3.4.15: Checking for dot in the URL	21
Figure 3.4.16: Checking sensitive words in the URL	22
Figure 3.4.17: Heatmap of the final dataset	22
Figure 3.4.18: Head of final dataset	23
Figure 3.4.19: Decision tree classifier	24
Figure 4.2.1: Accuracy of all algorithm	28
Figure 4.2.2: Confusion matrix of random forest classifier	28
Figure 4.2.3: Classification report	29

CHAPTER 1: INTRODUCTION

1.1 Introduction:

In this era of technology cyber bullying, Online harassment, Ransomware, Online prostitution, Account Hacking, Internet Fraud, Credit card Fraud. All these words are becoming our day to day problems who use online platforms very often [1]. The amount of Cybercrime is increasing day by day [2]. This problem is mostly initialized by fake content or sites as the content seems to be so much interesting to the user, he follows it and falls into the trap of these cyber criminals. The main targets of these criminals are mainly E-banking platforms, accounts that include personal information etc. Many models have been developed by different developers to prevent these attacks. Some models have been successfully built to block these sites. But when a user intentionally uses a fake URL link or agrees to its terms and conditions it's impossible for those blockers to block those sites. Now machine learning has always been a very effective approach to prevent these kinds of attacks. Machine learning is really helpful in automated platforms [3]. In our work we will create a model which will detect phishing websites automatically and warn the users. Our approach will be machine learning and different models will be applied and compared and only the best will be chosen for our work. Hopefully our work will be effective enough to detect fake content/sites and give users a crime free online service. Most of our data was collected from online open source websites such as phishtank, data for seo, alexa and fakebanklist. Rest of the data was collected from the data repository of the University of Brunswick. We used 20,000 data for our work. The highest accuracy that we gained was 97.50% from the Random Forest Classifier.

1.2 Motivation:

In the era of internet and technology everything is evolving day by day. People are getting the world in their hands. But some things are bad outcomes of technologies. Online crimes are increasing day by day. After watching the situation, we can say that

taking steps against these crimes are very much needed. We got our interest in this work after watching the efficiency of machine learning in this field. Automated cybercrime prevention can easily be done by machine learning. If we see the situation, we have to take action as soon as possible. Ransomware is increasing day by day. In the year of 2016 there were 4000 ransomwares alone in the U.S. [4]. The problem is not only among general online users. Healthcare industries use 6% of their budget for cybersecurity purposes in 2020. Because 50% of healthcare data was breached in that year. In the educational side we can also see the effect of cybercrimes. 66% of universities don't have the basic email security system. For that ransomware attacks on universities in 2019-2020 increased by 100%. As we can see the situation is getting out of hand. Specially for the covid-19 pandemic people are depending on online based platforms for various purposes. So, these things actually captured our eye and we got interested in this work. Our main motive is to provide in the process of a safer technological era.

1.3 Rational of the study:

Nowadays people are very comfortable with using online platforms. Entertainment, social platforms and online banking has become a part of our day to life. While Surfing the internet, users sometimes use fake websites which turn into some serious cybercrimes. So, there should be a process of verifying these sites.

Machine learning approach can be a very effective process of verifying these fake websites that users got into. Machine learning classifiers are mainly two types. supervised and unsupervised. In this matter supervised machine learning approach is more acceptable. Because supervised machine learning data already gets a label of correct answers. For that, verifying any wrong ones becomes easy. Data must get some important features and our model should get and verify those features and understand them. For that in our work we used supervised machine learning algorithms.

1.4 Research questions:

For doing the research perfectly it is very important to understand every concept very clearly. In the way of work, it is normal that we face many questions. So, to do our work we had to find answers to those questions. Below our research questions are presented,

1. What features should our data have for our research?
2. From where and how we should collect our data?
3. How can we process the data and get features from our data?
4. Which working environment should we use for our work?
5. Which algorithms should we choose to get the maximum number of accuracies of our model?

1.5 Expected output:

Our work was based on machine learning algorithms. We had to get the maximum number of accuracies for our models. We needed to get every feature perfectly for our work. Our model had to be trained with those features. Finally, our expected output had to verify the fake websites URLs for the users. So, the user can be alert before using those sites. Thus, many people can be safe from cybercrime while surfing the internet. We hope to gain the highest accuracy possible.

1.6 Project Management and Finance:

To do our project we needed some items which are a proper dataset, a functional computer, a proper coding environment, and a perfect language to implement machine learning. For the dataset we collected our dataset from various online open source platforms. And the data was arranged in google sheet and was saved in csv file. Two initial datasets were created this way. We had functional computers on our own. Our computer was sufficient to do all the necessary work for our project. Anaconda environment was used for our project implementation. We installed it on our computer.

The jupyter ide was used for necessary coding. Finally, we managed our project on our own.

1.7 Report Layout:

This report paper is divided into six chapters. And some other important information at the beginning and at the end.

1. The first chapter is for the introduction of our project. Past studies related to our work, our expectation, purpose of doing this work all are explained in this section.
2. This chapter is for our project background where past works and failings and improvements are discussed.
3. Here the main methodology and our main approach is explained.
4. Here the outcome has been discussed. Experimental results and discussion are important for the study.
5. The impact of our project on society and its necessity for our environmental improvements are discussed here.
6. The final conclusion and future plans of our work are discussed here.

CHAPTER 2: BACKGROUND

2.1 Terminologies:

For research purposes some paper was a must to review. Here in this section related works, comparative analysis and summary, scope of problem and challenges of our work have been described. Previous works have been reviewed. we tried to find out the approach of other works to find out the most efficient way. All the papers have been reviewed for gathering information from them. Comparative analysis and summary were also needed to compare some works that are done in this field. By this we could choose the best approach for our work. Scope of problem shows what problems we might get during doing this work. Finally, our challenges that we faced are described. Our main challenge was to extract the features which can be efficient for our model accuracy.

2.2 Related Works:

Some work has been done in this field. Kulkarni et al have created a system where websites are categorized based on their URLs [5]. They used a dataset which was collected from the University of California, Irvine machine learning repository. They included nine attributes and 1,353 samples. They used MATLAB scripts to create four machine learning models among them the decision tree algorithm delivered the highest number of accuracies. Some works are based on geometric deep learning methods.

Research has been done on the matter of phishing [6]. They collected a dataset from Kaggle which is an open source dataset collection site and applied different machine learning and deep learning techniques to gain the highest accuracy. While others cannot gain enough accuracy from random forest machine learning algorithms Rishikesh Mahajan and Irfan Siddavatam got 97.14% of accuracy from their work on phishing URLs [7]. One research has been done by data mining and machine learning techniques to verify bogus news. This project's outcome or result is in TF-IDF [8].

Tupsamudre et al did their work on phishing websites which were based on URL [9]. They tried to state that everything is in the name of the website. They extracted multiple features from their 110000 collected URLs and applied machine learning algorithms. Half of their URLs were phishing and half were legit URLs. They mainly focused on if IP of the URL, organization name-based statistics, domain statistics uncommon syntax, or sensitive words exist in the URL. Eshete et al did a very important research on their paper [10]. They talked about the effectiveness and efficiency issues of this kind of work. They talked about the importance of feature selection in this work. They focused on features which are commonly changing every day to get maximum success.

Here, they did a methodological overview on phishing website detection [11]. They did not complete their work they just explained their plan of working. They reviewed multiple papers and explained their approach will be machine learning. Their method also includes a notification-based feature which gives the users notification based on the danger. They planned to get 30 features of different types for their models.

In 2019 they mentioned that phishing attacks are the most major problems in information security [12]. They collected 16000 URLs data. Among them 12000 were phishing URLs which were collected from phishtank and other 4000 legitimate URLs were collected from 10 daily users. They mainly chose features based on URL, URL age, URL rank on the web. Their model was machine learning based. The maximum accuracy they got from 98% with 29 features.

Rashid et al used three machine learning algorithms FACA, SVM, Random forest [13]. They collected data by GNU webget python scripts. 70% of data was used to train and the rest is used to test. They got maximum accuracy from SVM 95.66%.

Sindhu et al used three machine learning algorithms SVM, Random forest, Neural Network [14]. They used a dataset from UCI machine learning repository. They had 11055 URL info among them 6157 are phishing and 4898 are legitimate. 30 features were extracted. They got maximum accuracy from SVM 97.45%.

They did research on phishing website detection using a heuristic way [15]. They collected datasets from phishtank and DMOZ. it includes 11660 phishing URLs and 5000 real URLs. They selected 4 features and calculated the value of 6 heuristics. Their approach got the highest accuracy of 97.16%.

Ahmed et al did multiple approaches in their work [16]. They did content-based approach, heuristic based approach, blacklist-based approach. They used 5 features. They tested their model with 100 URLs among them 59 were fake and 41 were real. The outcome shows 68 fake and 32 reals. Their accuracy was 96%.

They researched on fake website detection they got 98.77% accuracy on artificial neural networks [17]. They did their work on 16053 phishing and 7974 real URLs. They used 10 features for their model.

By reviewing all these researches, we were able to understand what could be our approach towards this work. We found out that there can be various ways of solving this problem. We understood that a machine learning approach could be the best way we can do this work.

2.3 Comparative analysis and summary:

Works in this field are not the same. Some use deep learning, some use machine learning, some use other methods. Some use a lot of data that includes URLs and some use a small amount of data. Feature extraction works also can be differentiated.

Machine learning works,

Alswailem et al have done a machine learning approach with a maximum accuracy of 98% [12]. They used data from open source platforms. They extracted features from URLs.

Rashid et al have also worked on phishing detection [13]. Their best model was SVM. They got 95.66% of accuracy. They didn't use many features but got a good accuracy result.

Sindhu et al got maximum accuracy from SVM 97.45% [14]. They used three machine learning algorithms SVM, Random forest, Neural Network.

They Researched on fake website detection and they got 98.77% accuracy on artificial neural networks [17]. They did their work on 16053 phishing and 7974 real URLs. They used 10 features for their model.

We can see here machine learning models gain a perfect number of accuracies in this work. The most common machine learning algorithms that are used in these works are

1. Decision Tree Classifier
2. Random Forest Classifier
3. Support Vector Machine

Almost every machine learning approach is based on feature extractions from datasets except some exceptions. Top common features are Domain age, URL rank, Ip address, URL length etc.

Moreover, the machine learning approach is pretty effective.

Other methods,

Ahmed, Abdulghani Ali; Abdullah, Nurul Amirah did multiple approaches in their work [16]. Their approaches,

1. content based approach
2. heuristic based approach
3. blacklist based approach.

They used 5 features. They tested their model with 100 URLs among them 59 were fake and 41 were real. The outcome shows 68 fake and 32 reals. Their accuracy was 96%.

L. A. T. Nguyen, B. L. To, H. K. Nguyen and M. H. Nguyen did research on phishing website detection using a heuristic way [15]. They collected datasets from phishtank and DMOZ. it includes 11660 phishing URLs and 5000 real URLs. They selected 4 features and calculated the value of 6 heuristics. Their approach got the highest accuracy of 97.16%.

Here we can see other approaches also have very good numbers. Most common features are Domain, Primary Domain, Subdomain and Path Domain. These are mainly heuristic based features

If we compare, the machine learning approaches are more convenient as accuracy scores are more than the other methods.

2.4 Scope of the Problem:

Doing this work we faced many problems. But eventually we got a solution for every problem. Some important problems that we faced are described here,

1. The first problem that we faced was starting the work with perfect strategy. Our work wasn't going well and certainly the reason was the strategies were not well organized. So, getting a perfectly organized strategy can be a problematic factor.
2. After that we faced a problem choosing the best approach for our model. As many models gained good results. We were researching each one very thoroughly to get our perfect approach.
3. Choosing a good environment and language was also a problem. But when we were sure that our approach would be machine learning our language and environmental decisions were easier.
4. Data preprocessing was also problematic as we had to use raw data for feature extraction. We couldn't use much preprocessing techniques in order not to lose any information.
5. As features are many, using the perfect features to gain the best accuracy was a problem. We saw that features help to improve accuracy but some features also impact negatively.

2.5 Challenges:

The biggest challenge that we faced was to collect the datasets. We know that the URL dataset exists on open source platforms. But we included a large portion of our dataset on our own. We cannot get URL information from surveying so we had to go to many platforms that could have URL information and collect data from there. We had to also collect two types of URLs. We know that real websites can be collected by surfing the web but phishing websites were the main challenge to collect. Another challenging task was to select the most effective features and machine learning algorithms. Features affect one model very much. It can lower the accuracy and also can higher the accuracy. There are a ton of machine learning algorithms. Choosing the best ones for this work was also challenging.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation:

The subject of our research is “URL based fake website detection using machine learning algorithm”. We reviewed a lot of papers and a lot of fields. We saw that machine learning fields are really needed right now. And cyber security is one of the most needed things in the era of technology. So, detecting fake websites which are the main reason for some cybercrimes got our attention. Nowadays people are really comfortable using online platforms. Especially in this pandemic situation we do most of our work online. Creating a safe online environment is really needed. We did our work in anaconda environment. A Jupyter notebook was used to do the necessary coding. Jupyter Notebook is really useful for Python development, Microsoft Windows platform, Multiple Editors for designing, various Python advance libraries such as Pandas, NumPy, Matplotlib, Seaborn, Cufflinks, NLTK, regular expression package, scikit-learn, word cloud etc to complete our work. The tools and environment were really user friendly.

3.2 Data Collection Procedure/Dataset Utilized:

Our initial dataset contains URLs of both fake and real websites. We kept the ratio of fake and real data 50:50. As we needed URLs so our best choice was online platforms. Data was collected from various platforms. phishtank [18] and fakebanklist [19] have a huge collection of fake URLs. We collected most of the fake URLs from there. Alexa [20] and dataforseo [21] got some real URL lists. And finally, some data was collected from a data repository of University of Brunswick [22]. A total of 20,000 data was collected. Two initial datasets were created fake and real. In the further processing features were extracted from both. And later after all preprocessing both datasets were combined into a final dataset. And the final dataset was used for our model.

3.3 Statistical Analysis:

Supervised machine learning needs some predefined features which helps the model to learn certain things. Our initial dataset contains just URL information. Below steps will describe the statistics of our dataset.

Initial dataset and its statistics:

Initially we had two dataset one was a fake dataset and the other one was a real dataset. Both had 10,000 URLs with no other columns than URLs. So, a total of 20,000 data was collected for our work. Samples are given below.

URLs		URLs	
0	https://privatbank.ua/ua/perevod-zarplaty-s-ka...	0	http://digilander.libero.it/senza.filtr...
1	http://squarespace.com/press/2014/7/9/juniper-...	1	http://r8ezc.info/NCVxgy6dZO/
2	https://quizlet.com/login?redir=https%3A%2F%2F...	2	http://r5eum.info/KxKC6qrU0
3	https://quizlet.com/sign-up?redir=https%3A%2F%...	3	http://alsope.covidharsh.com/
4	http://ssa.gov/disabilityfacts/materials.html?...	4	http://528b05t.r7evz.info/fnsT
5	http://stackoverflow.com/questions/21494979/fi...	5	https://www.registroslbknet.com/cliente/
6	http://steamcommunity.com/sharedfiles/filedeta...	6	http://registroslbknet.com/cliente
7	http://steamcommunity.com/sharedfiles/filedeta...	7	http://sbcglobalflsff.yolasite.com/
8	http://superuser.com/questions/630156/how-do-i-...	8	https://sbcglobalflsff.yolasite.com/
9	http://superuser.com/questions/874217/run-a-st...	9	http://anvpwy.covidharsh.com/

Figure 3.3.1: Initial datasets

We did check for Null values in our dataset and cleansing of these two datasets. After that no further action was needed. We did feature extraction procedure next.

3.4 Proposed Methodology:

As our approach is machine learning we know that many machine learning algorithms exist. From them we chose some very good supervised machine learning algorithms for our model. As people are very used to online platforms it will be easy for them to do some necessary steps before using any site. All our model needs is a website URL and it will predict if the URL is real or fake.

As URLs are in string format our models cannot use this directly. So, some necessary features were extracted from them and then it was used to train and test six machine learning algorithms. Based on performance the best algorithm was used for our model. Down below a design methodology is shown.

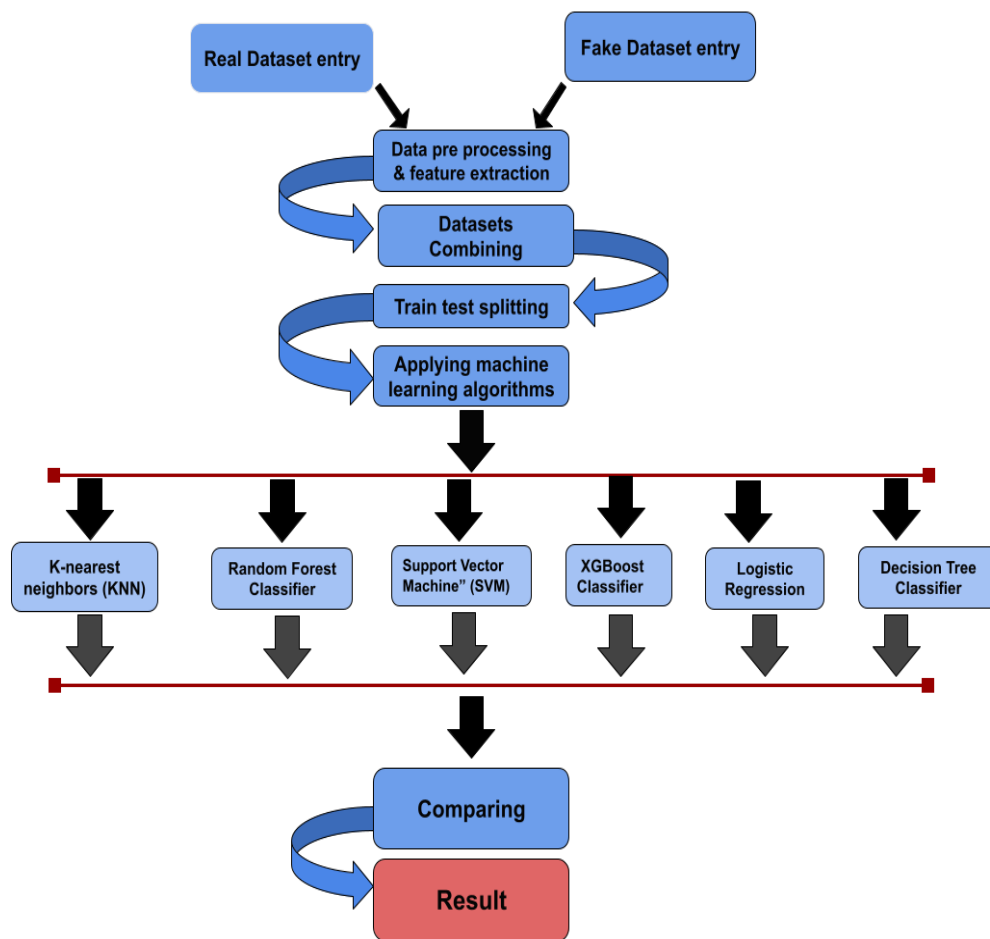


Figure 3.4.1: Methodology

After entering the datasets, we did check for Null values in our dataset and cleansing of these two datasets. After that no further action was needed. We did feature extraction procedure next.

Feature extraction:

As we have done our work based on URLs. Fake and real website URLs have some statistical differences. For example, real website URLs do not consist of IP addresses. So if a URL has an IP address it is fake. Like that a total of 16 features were extracted from our datasets. All the features are described down below,

1. Domain name extraction:

Usually a domain is always presented in the URL. A URL will take users to any of the website's pages if a domain name is available. Every URL contains a domain name with other information to locate the desired page of the users.

We are extracting the domains from the URLs. This feature isn't very effective for machine learning algorithms. Later it will be deducted from the final dataset.

A URL Generally includes below structure:

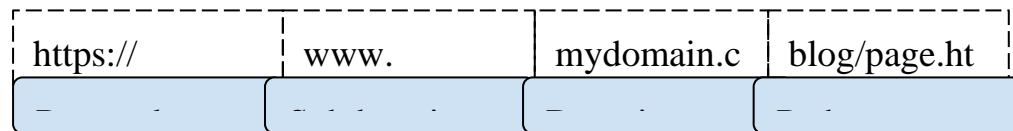


Figure 3.4.2: URL statistics

After extracting this feature, we can see the domain names of our final dataset like example given below,

```

0          privatbank.ua
1      squarespace.com
2          quizlet.com
3          quizlet.com
4              ssa.gov
5      stackoverflow.com
6      steamcommunity.com
7      steamcommunity.com
8          superuser.com
9          superuser.com
Name: Domain, dtype: object

```

Figure 3.4.3: After Domain extraction

2. IP address instead of domain name:

Most URLs do not include an IP address [23] instead it uses domain name as the nickname of the IP address. So, if the URL consists of an IP address instead of domain name that is set to 1 means it is fake or else it is set to 0 means it is real. Attackers might use IP addresses to steal sensitive information. We extracted this feature from our datasets. The extracted feature looks like this,

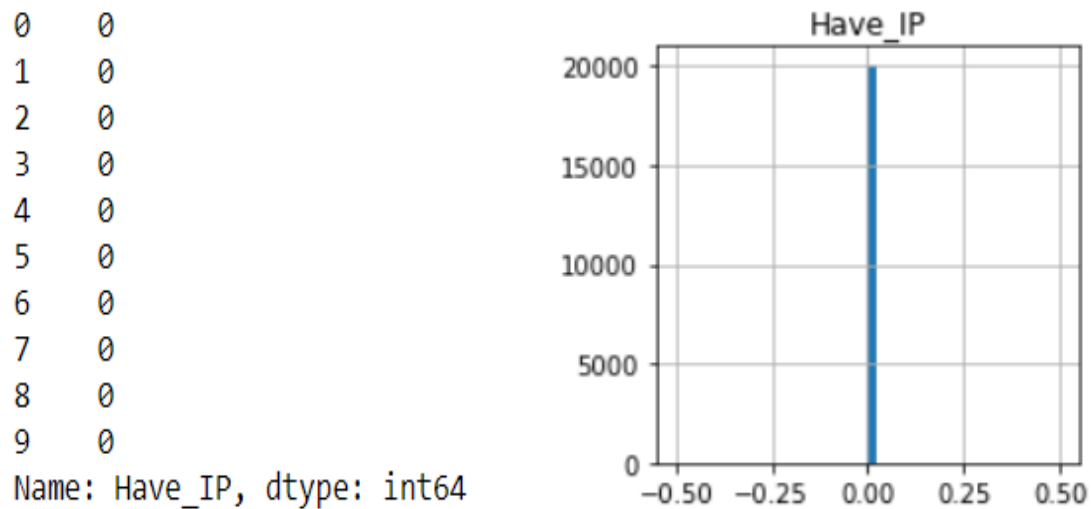


Figure 3.4.4: IP address-based feature extraction

We can see that the column includes 0 or 1 means the value is stored in binary form.

3. Unsafe character in the URL:

Some special characters are allowed in the URLs. Such as (\$ _ . + ! * ' () ,)

If any special character besides these ones are used in any URLs that might be harmful [24]. So, if a URL consisting of any unsafe characters such as ([] { } | \ " % ~ # < >) is considered fake. Otherwise it is real.

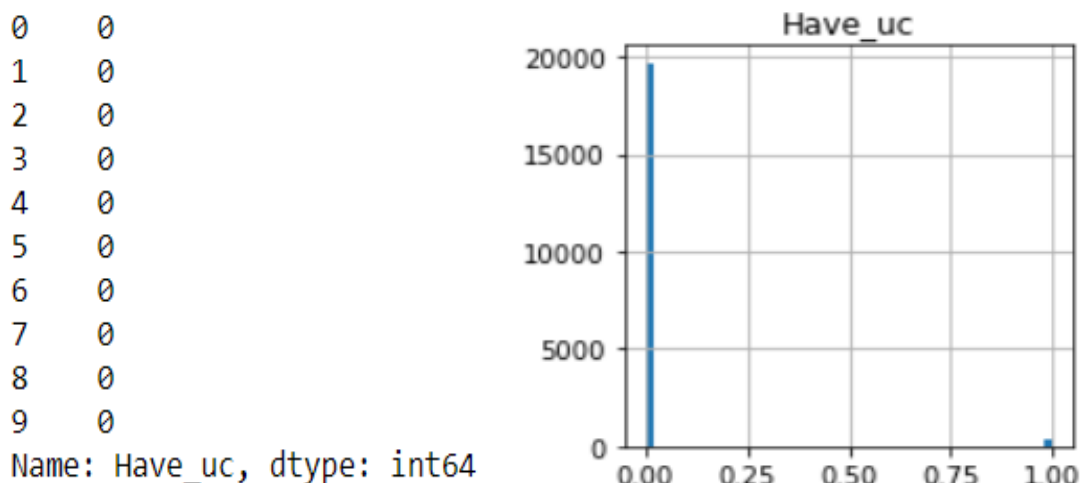


Figure 3.4.5: Unsafe character-based feature extraction

Here, the fake URL is set to the value of 1 and the real URL is set to the value of 0. After extraction the values are returned in binary form.

4. Length of URL:

We calculated the length of the URL and created a feature based on it. An average real website's URL length is 62[25].

Nowadays for doing seo of any website URLs are mostly in the length of 50-60. But the fake websites do not intend to care about this feature. So, most of the fake URLs are more than 62.

So, if a website's length is more than 62 it is considered fake, otherwise it is considered real.

Here, the fake URL is set to the value of 1 and the real URL is set to the value of 0. After extraction the data sample looks like this,

```
0    1
1    1
2    1
3    1
4    1
5    1
6    1
7    1
8    1
9    1
Name: URL_Length, dtype: int64
```

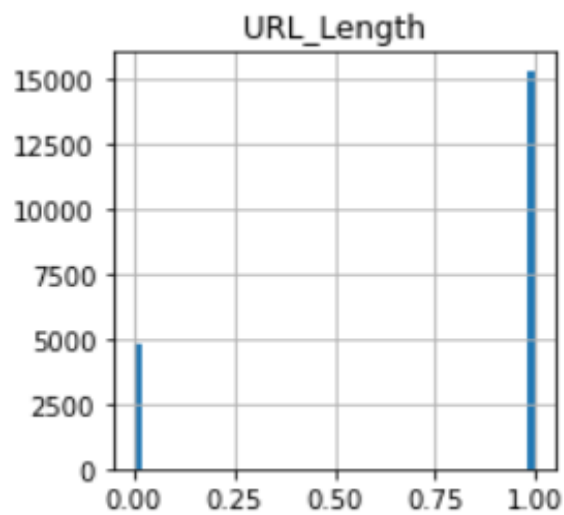


Figure 3.4.6: URL length-based feature extraction

This feature is also stored in binary value.

5. Storing domain length:

A maximum domain name can be 253 characters [26]. We calculate the length of our URLs domain and store it as a numerical value. The outcomes are as below example.

As we can see the lengths of the domain can also be calculated and it is stored as numerical values.

```

0    13
1    15
2    11
3    11
4     7
5    17
6    18
7    18
8    13
9    13
Name: Domain_Length_value, dtype: int64

```

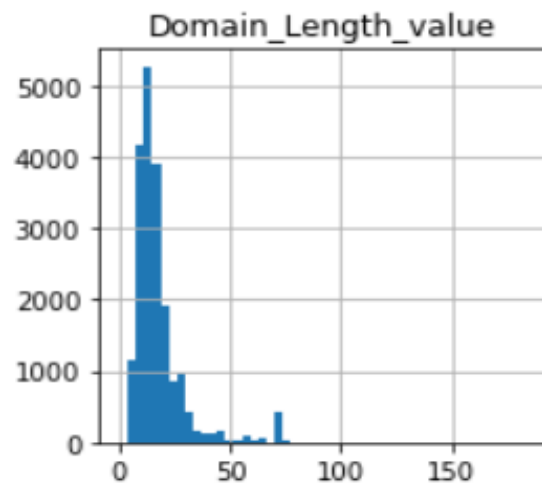


Figure 3.4.7: Storing domain length

6. Comparing domain length:

As we know, the domain can be 253 characters. But real websites tend to use shorter domain length. According to gaebler on domain length top 10,000 websites are not longer than 8 characters [27]. And more than 1 million real websites are barely more than 10 characters long. It is also said that real websites are always below 20 characters. So, for our feature extraction we set that if a URLs domain length is less than 20 characters it is real thus the value is set to 0. Otherwise the value is said to be 1. Below we can see the extracted feature.

```

0    0
1    0
2    0
3    0
4    0
5    0
6    0
7    0
8    0
9    0
Name: domain_Length, dtype: int64

```

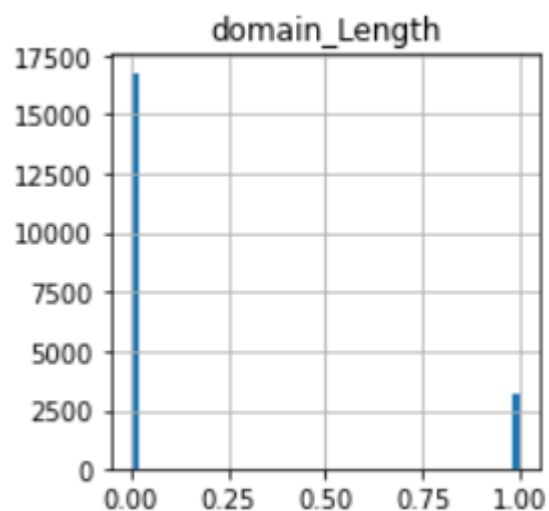


Figure 3.4.8: Comparing domain length

So, we can see domain length-based features are stored as binary values.

7. Depth of URL:

Here we calculate the depth of the URLs. Depth refers to how many subpages exist in any URL or how many clicks it actually takes to reach that certain page. This value is stored numerically. Here we can see the stored depth value of our URLs.

```
0    3
1    5
2    1
3    1
4    2
5    3
6    7
7    7
8    3
9    3
Name: URL_Depth, dtype: int64
```

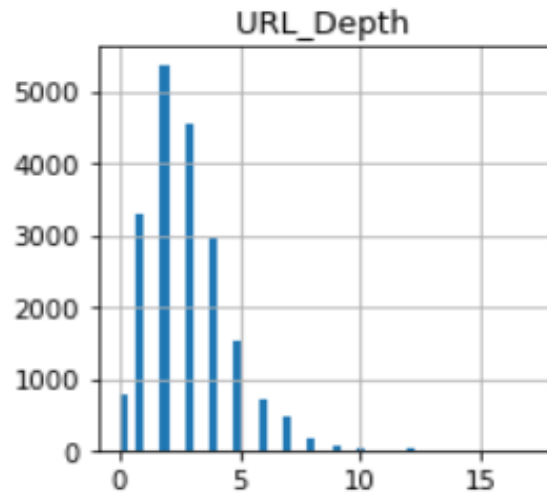


Figure 3.4.9: Calculating URL depth

8. How many slashes exist in the URL:

This feature is also based on depth as depth is calculated based on the existing slashes. The most best performing websites have a depth of 5 or less. Fake websites tend to use more than 5 subpages to confuse the users [28]. So, websites containing more than 5 depth is considered to be fake.

Down below feature extraction is shown.

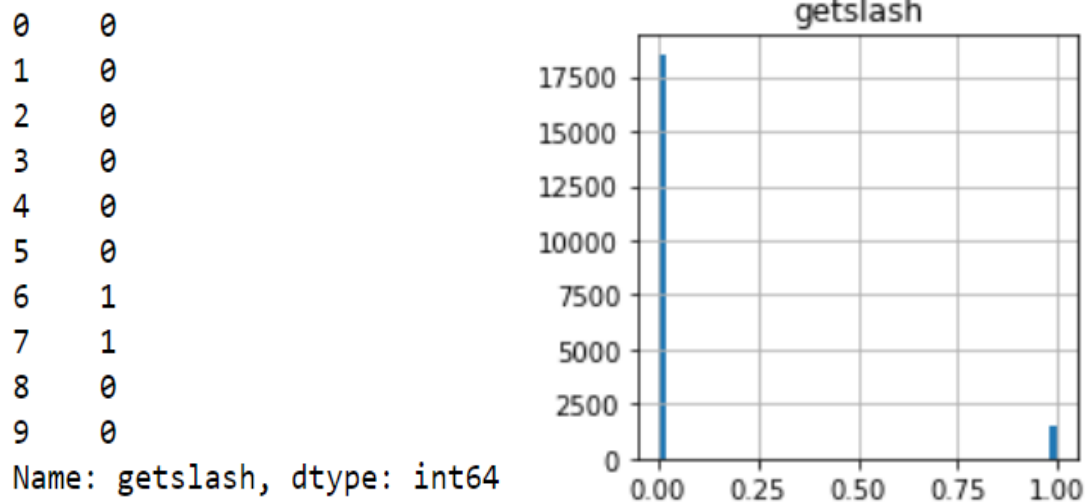


Figure 3.4.10: Slash based feature extraction

Here the values are stored in binary form.

9. Redirection of URL:

When a URL has “//” in it. It means it will redirect to another page. Generally, a redirected real URL has “//” in the 6th or 7th position based on “http” or “https” [29]. So, if the position of “//” exceeds 7th position it is redirecting the user to an unknown path. So, if any URL has “//” in more than 7th position it is fake, otherwise it is real. Here we can see the outcome of this feature extraction.

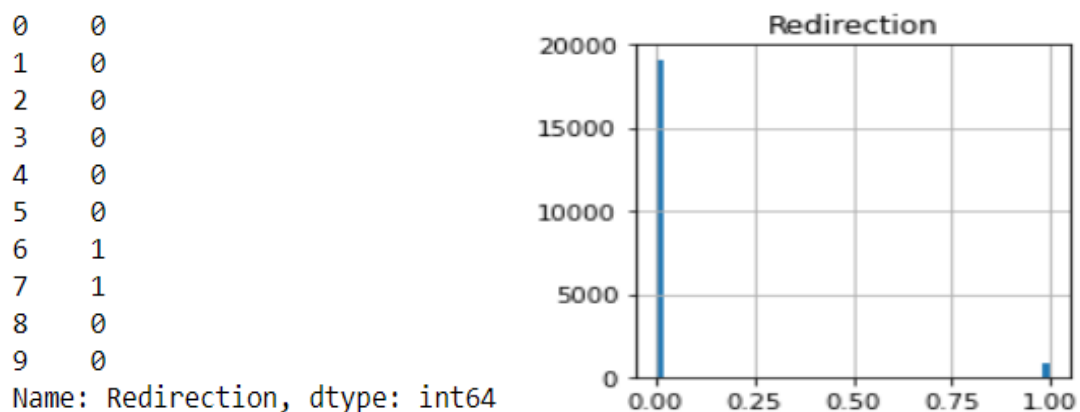


Figure 3.4.11: Checking for redirection in the URL

10. “Http” or “https” in domain name:

Https is basically http with encryption. HTTPS uses TLS (SSL) to encrypt normal HTTP requests and responses. For that https is far more secure than the http. A http URL starts with http:// where a https URL starts with https:// [30]. So, we decided to select all the websites with http a threat because fake websites might use less secure website URLs. All the URLs with http are marked as 1 and the rest are 0. The extracted data looks like the figure below.

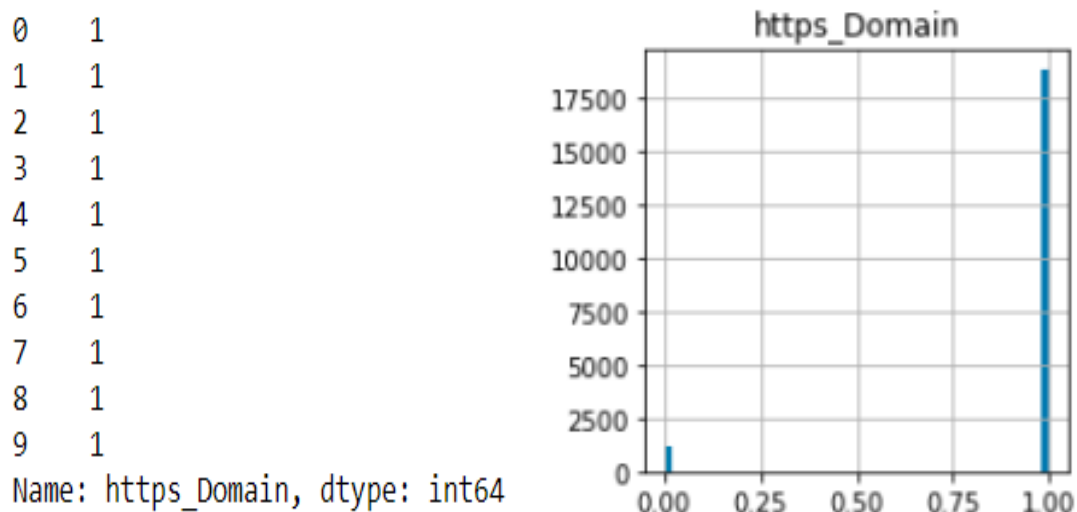


Figure 3.4.12: “Http” & “https” based feature extraction

11. URL shortening method:

URL shortening is a method in which the real URL is made considerably smaller in length which works the same as the real URL. We know this is a great feature for online URL sharing but fake websites makers love this fact because this feature makes the links opaque. Fake URLs use this service to mislead users [31]. So, if any URL contains a shortening method it is considered fake. The values are assigned in binary format.

```

0    0
1    0
2    1
3    1
4    0
5    0
6    0
7    0
8    0
9    0
Name: TinyURL, dtype: int64

```

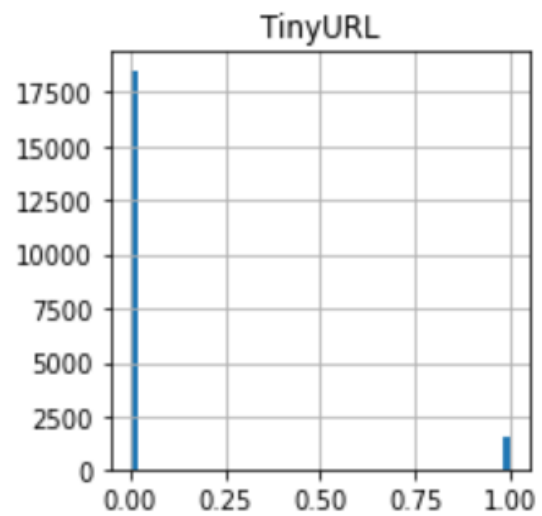


Figure 3.4.13: Checking for URL shortening method

12. “-” & “@” in domain:

Fake websites tend to use special characters like “-” and “@” in the domain to confuse the users where real websites rarely use these [28]. So, if the domain contains “-” or “@” the features returns 1. Otherwise it is 0.

The feature looks like this.

```

0    0
1    0
2    0
3    0
4    0
5    0
6    0
7    0
8    0
9    0
Name: Characters, dtype: int64

```

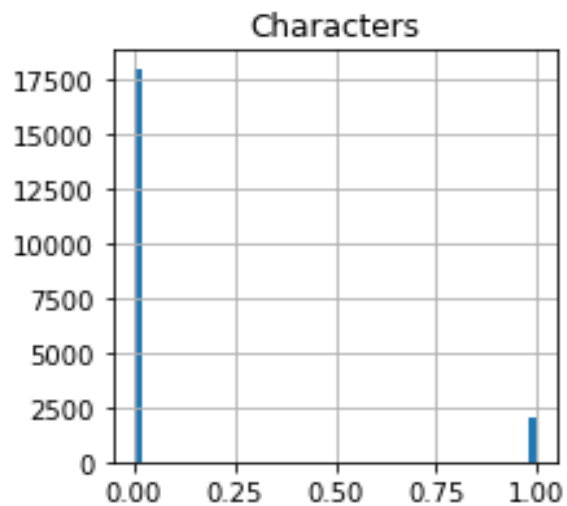


Figure 3.4.14: “-” & “@” based feature extraction

13. “.” in the URL:

Most of the real websites do not contain more than 5 “.” in their URL [28]. So, if any URL contains more than 5 “.” It is considered fake. So, the return value will be set to 1. Otherwise the return value will be 0.

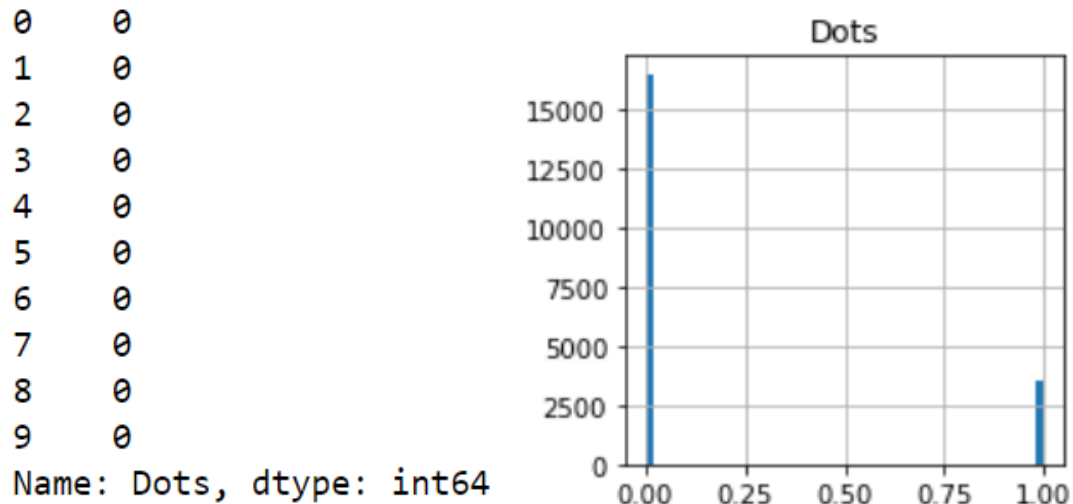


Figure 3.4.15: Checking for dot in the URL

So, this feature will also store values in binary format.

14. Sensitive words in URL:

Sensitive words like 'confirm' 'account' 'banking' 'password' 'secure' 'backup' 'toolbar' 'webscr' 'username' 'install' 'ebyisapi' might be included in a fake URL website [28]. So, if the URL contains these words the value is set to be 1, otherwise the value is 0.

Here we can see some examples of stored features.

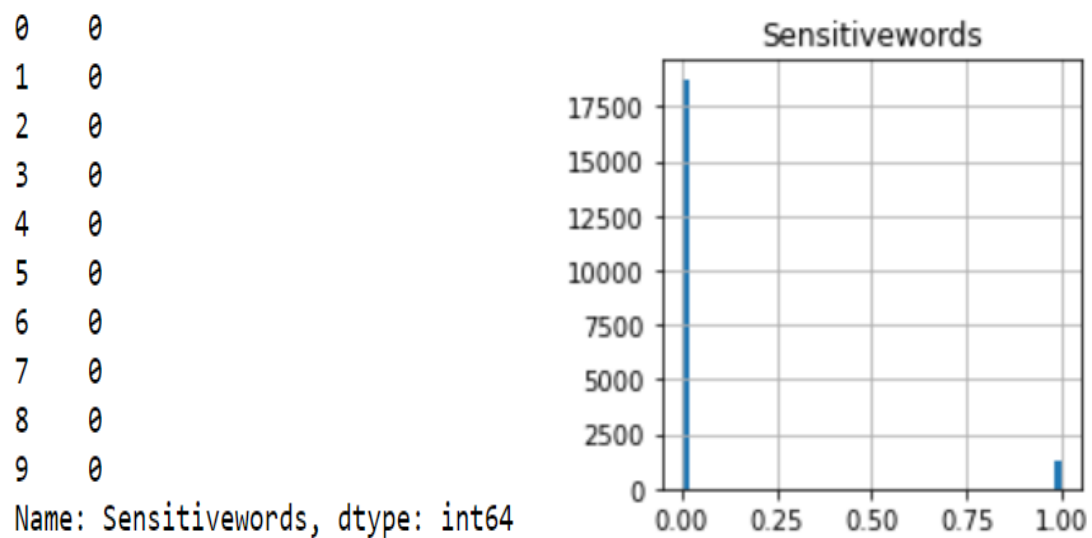


Figure 3.4.16: Checking sensitive words in the URL

Finalizing Dataset:

So, all these features are extracted from the initial URL datasets.

Finally, we combined the both data frame in one final dataset which has a total of 15 columns and 10000 rows.

Let's check the heatmap and sample of our final dataset,

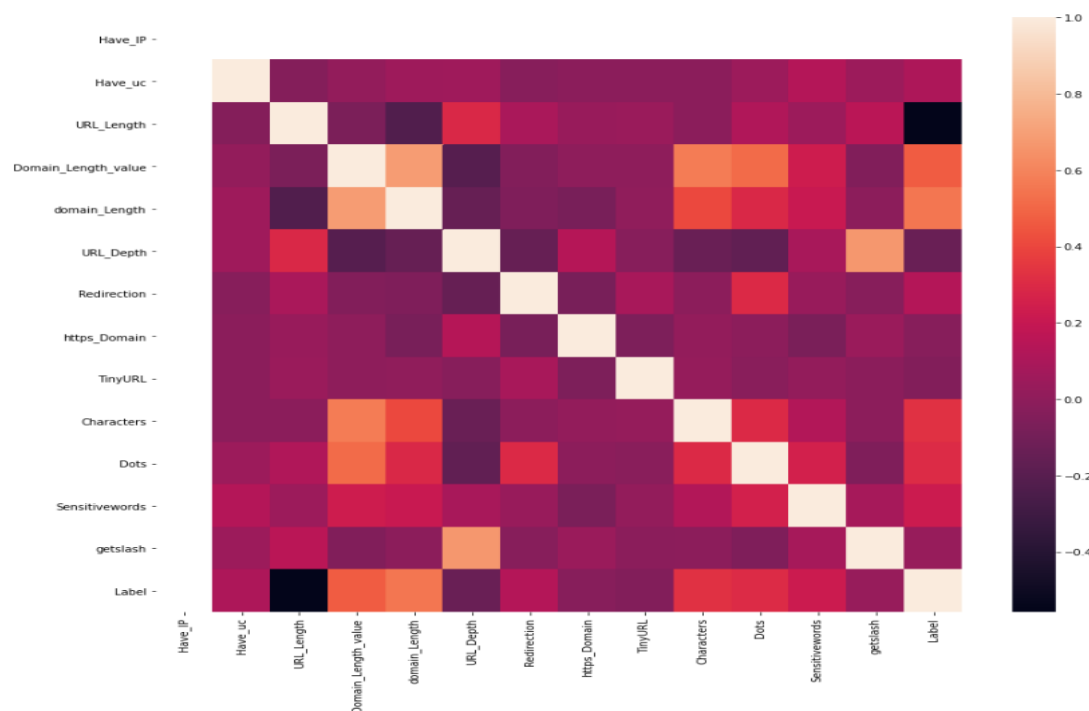


Figure 3.4.17: Heatmap of the final dataset

	Domain	Have_IP	Have_uc	URL_Length	Domain_Length_value	domain_Length	URL_Depth	Redirection	https_Domain	TinyURL	Characters	Dots
0	privatbank.ua	0	0	1	13	0	3	0	0	0	0	0
1	squarespace.com	0	0	1	15	0	5	0	1	0	0	0
2	quizlet.com	0	0	1	11	0	1	0	0	1	0	0
3	quizlet.com	0	0	1	11	0	1	0	0	1	0	0
4	ssa.gov	0	0	1	7	0	2	0	1	0	0	0

Figure 3.4.18: Head of final dataset

Now our dataset is again checked for data cleansing and finally the final dataset is ready for our model.

Applying Machine Learning Algorithms:

At first, we dropped the domain column as it had nothing to do with machine learning algorithms. Then we divided our processed final dataset into an 80:20 ratio in order to train and test. For that we used anaconda's default library functions.

After that we applied six different supervised machine learning algorithms to get the maximum performance of our model. All the algorithms are described below.

A. Logistic Regression:

Logistic Regression is a Machine Learning technique that makes predictions to identify the value of a dependent variable such as tumor status, email categorization, or university entrance etc by learning from independent variables [32].

Logistic regression is a supervised Machine Learning algorithm, which means that the data used for training is labeled, implying that the solutions are already in the training set.

The results from Logistic Regression:

Table 1: Accuracy of logistic regression

Name	Data Used	Training Accuracy in %	Testing Accuracy in %
Logistic Regression	80:20 Ratio	93.19%	92.87%

B. Decision Tree Classifier:

This classification has two steps. One step is where the model uses training data to train the model. And the second step is where the model uses test data for prediction steps [33].

It's basically a tree structured classifier. The internal nodes represent the features of the dataset and branches represent the decision rules and leaf represents the outcome.

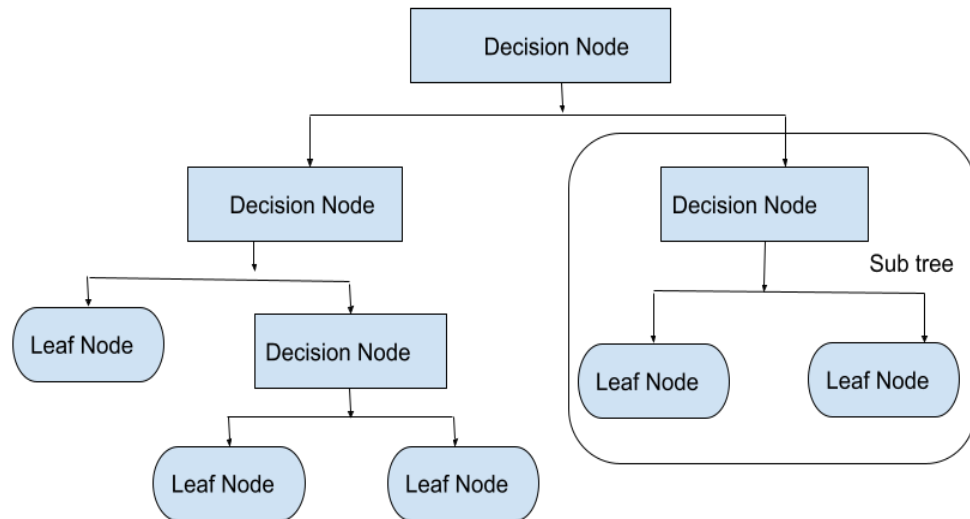


Figure 3.4.19: Decision tree classifier

The results from Decision tree Classifier:

Table 2: Accuracy of decision tree classifier

Name	Data Used	Training Accuracy in %	Testing Accuracy in %
Decision tree Classifier	80:20 Ratio	97.50%	96.80%

C. XGB Classifier:

XGBoost Classifier is a gradient boosted decision tree implementation which is aimed for competitive machine learning for speed and performance [34].

XGBoost is used because it offers a very efficient implementation of the stochastic gradient boosting technique as well as access to a set of model hyperparameters. So, it gives users control over the model training process. The most significant XGBoost's success is its scalability in all fields.

The results from XGB Classifier:

Table 3: Accuracy of XGB classifier

Name	Data Used	Training Accuracy in %	Testing Accuracy in %
XGB Classifier	80:20 Ratio	97.38%	96.87%

D. Random Forest Classifier:

Random forest is a very flexible machine learning algorithm. Even without hyper parameter tuning it produces good results.

It is one of the most used machine learning algorithms because of its simplicity and its diversity because it can be used both as classification and regression [34].

The results from Random Forest Classifier:

Table 4: Accuracy of random forest classifier

Name	Data Used	Training Accuracy in %	Testing Accuracy in %
Random Forest Classifier	80:20 Ratio	97.50%	96.90%

E. K-Nearest Neighbors Classifier:

It is a supervised machine learning algorithm. It predicts any targeted variable depending on one or many dependent variables [35].

It mainly checks feature similarity of any given data. And gives output based on which feature was the most similar. It means the data will be assigned a value based on how close it is to the training dataset.

The results from K-Nearest Neighbors Classifier:

Table 5: Accuracy of K-Nearest neighbors classifier

Name	Data Used	Training Accuracy in %	Testing Accuracy in %
KNN Classifier	80:20 Ratio	96.96%	96.10%

F. SVC:

SVC, known as Linear Support Vector Classifier method, which uses a linear kernel function to perform classification on datasets [36].

This classifier works well with a large number of samples like we have here.

It is a nonparametric clustering algorithm. It does not make any assumptions on the numbers or shapes in the dataset. If a dataset contains high complex parameters it needs preprocessing steps or this algorithm is safe to use.

The results from SVC Classifier:

Table 6: Accuracy of SVC classifier

Name	Data Used	Training Accuracy in %	Testing Accuracy in %
SVC Classifier	80:20 Ratio	92.35%	91.97%

All these algorithms were compatible for our work. And the best algorithms were chosen. Also, all the algorithms were supervised as we needed for our model.

All the algorithms have their unique positivity and negativity.

3.5 Experimental Requirements:

This project needs some essential tools to develop properly. It needed some important tools. In order to do our work perfectly we had to manage all these requirements. Down below all the essential requirements are mentioned,

Software and Hardware Requirements:

1. OS (Windows / Linux / Mac) that supports Python 3.0
2. Compatible computer to run anaconda environment including python 3.0
3. RAM (Minimum Requirements of 4 GB)
4. HDD (Minimum requirements of 128 GB)

Development Tools:

1. Jupyter Notebook / Google colab / PyCharm
2. Microsoft Word / Google Docs / Text Editor
3. Drawing Tools
4. Anaconda environment.
5. Microsoft Excel/Google sheets.

CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup:

We used python 3 for our model implementation. For that anaconda environment was used. At first, we installed an anaconda environment on our computer. Then the Jupyter notebook was used to do the essential codes of our model. This is basically a text editor very useful for coding. Different python libraries including NumPy, Pandas, Matplotlib, Seaborn, NLTK, Cufflinks, scikit-learn, OS, Warnings, Strings, Word Cloud etc were used for our work. We installed many packages for using all these libraries.

4.2 Experimental Results & Analysis:

We applied seven different machine learning algorithms and the machine learning algorithm was chosen for our model.

Seven machine learning algorithms were,

1. Logistic Regression
2. Decision Tree Classifier
3. XGB Classifier
4. Random Forest Classifier
5. K-Nearest Neighbors Classifier
6. SVC

Each algorithm gave different outcomes. At first, we didn't get our expected accuracy. Later we understood that our algorithms were fine but the features were less effective. So, we added more new features and at last we got an accuracy of highest 97.50%. All the algorithms performed well.

All the results are,

Table 7: Accuracy of all algorithms

Algorithm	Training Accuracy	Testing Accuracy
Logistic Regression	93.19%	92.87%
Decision Tree Classifier	97.50%	96.80%
XGB Classifier	97.38%	96.87%
Random Forest Classifier	97.50%	96.90%
K-Nearest Neighbors Classifier	96.96%	96.10%
SVC	92.35%	91.97%

Accuracy

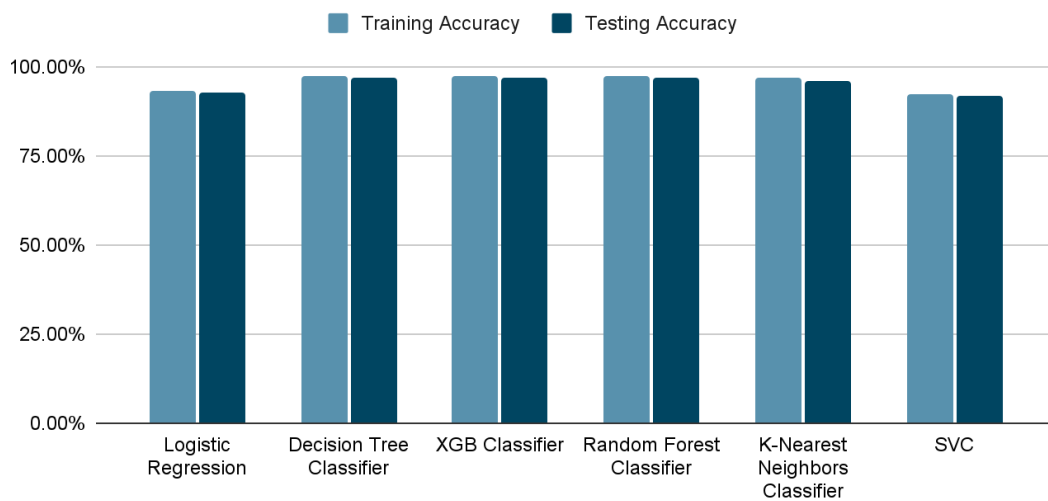


Figure 4.2.1: Accuracy of all algorithms

Confusion Matrix:

If we look at the confusion matrix of our model, we can say that our model performed pretty well.

```
array([[1964, 29],
       [ 95, 1912]], dtype=int64)
```

Figure 4.2.2: Confusion matrix of random forest classifier

Down below a proper classification report is given based on testing data.

	precision	recall	f1-score	support
0	0.95	0.99	0.97	1993
1	0.99	0.95	0.97	2007
accuracy			0.97	4000
macro avg	0.97	0.97	0.97	4000
weighted avg	0.97	0.97	0.97	4000

Figure 4.2.3: Classification report

4.3 Discussion:

Our model gave a satisfying number of accuracies. The highest accuracy we got on training data was 97.50%. And testing data got a highest of 96.90% of accuracy.

Model based accuracy comparison,

1. Logistic Regression: Training accuracy (93.19%) Testing Accuracy (92.87%)
2. Decision Tree Classifier: Training accuracy (97.50%) Testing Accuracy (96.80%)
3. XGB Classifier: Training accuracy (97.38%) Testing Accuracy (96.87%)
4. Random Forest Classifier: Training accuracy (97.50%) Testing Accuracy (96.90%)
5. K-Nearest Neighbors Classifier: Training accuracy (96.96%) Testing Accuracy (96.10%)
6. SVC: Training accuracy (92.35%) Testing Accuracy (91.97%)

We can see the maximum accuracy was obtained from the Random Forest classifier.

So finally, the model was selected for our project. And finally, it was saved by Pickle, a library function of Python for future works.

Confusion Matrix:

Table 8: Confusion matrix statistics

Confusion Matrix	Actual Positive	Actual Negative
Predict Positive	True Positive (TP)	False Positive (FP)

Predict Negative	False Negative (FN)	True Negative (TN)
------------------	---------------------	--------------------

From figure 4.2.2 We can see that true positive and True Negative outcomes are really good compared to false Negative and true False Negative.

So, we can say our model performed well.

Classification Report:

Precision: It is the number of positive class predictions.

Precision= $TP/(TP+FP)$.

Recall: Positive class predictions which are out of all positive data.

Recall= $TP/(TP+FN)$

F1 Score: Calculated by precision and recall.

F1 Score = $(Precision * Recall) / (Precision + Recall)$

CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society:

We people are social beings. Now in the era of internet and technology online platforms are evolving day by day. People are getting the world in their hands. Online crimes are also increasing day by day. Taking steps against these crimes are very much needed. We can see people are suffering from online hassle every day. It is because they are depending on it very much which we cannot change. But our project should make their online experience much safer. And by using online platforms people can save a lot of money, time and energy which can help our society very much. It doesn't seem very important for sociological improvement. Our project should help sociological improvements smoother by making online platforms easier and safer to use.

5.2 Impact on Environment:

Day by day our environment is getting worse. We are hurting our own environment very much. All the traffic, dust we throw and many more are giving our environment negative impacts. So, if we can do some of our work virtually it will be very efficient for our environment. Like if we do not hurry for a physical meeting causing traffic jams instead, we do that meeting online it will be great use for environmental improvement. Most of our bank related work can be done online nowadays. Also, many official works can be done online. All these can cause less use of papers and plastics or some elements which are hurting our environment. So, we can see doing online work can help grow our environment positively. But for that we need safer online platforms for that our project can help in various ways.

So, we can say our project also has a great positive impact on our environment.

5.3 Ethical Aspects:

Our main goal is to make online platforms much safer. Because we need a safer online user experience. Our project will make online uses safer which will be beneficial for online platforms. Such as banking, social media, surfing the web and different platforms will be highly beneficial. Our project was carefully done by keeping in mind the user's side. Because whatever happens keeping the users safe is our main priority. Our model will keep 100% privacy of the users. Finally, as we say, a safer web for everyone.

5.4 Sustainability Plan:

Our first sustainability plan will be “Easy Access”. We planned to do a web version of our project. For that our project should be easier for the users to access. No installation process will be needed to access our model. And users will be directly given all access to our project.

Secondly, we will be giving the users to perform very easy detection. So as the user will be detecting fake websites that might harm them so a very easy platform will be created for the users to do easy detection.

Our other sustainability plan will be to make sure if the users are getting confident to use our platform. Many suggestions will be given to the users so that users can easily get confidence while using our platforms.

A feedback option will be provided for the users. This will be another sustainability plan. As the users can give feedback and will be treated with a reply from the admin sides.

So, these are our sustainability plans.

CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study:

Our main priority was to make the online platforms user experience safer to use. We propose a model which will detect fake websites and by fake we mean websites that cause cybercrimes. Because it is a really big problem nowadays. Here our model depends just on the URL of the websites. So, we can say that the model is less complicated for the users. Means users don't need to give extra unnecessary information instead the user will just provide the website URL and our model will detect if the website is safe to use or not. We focused on the URL statistics of fake websites and extracted features from our dataset which was collected from different open source platforms. Finally, with many features a targeted column was included and the dataset was used to train our model. 80% of the data was used for training and the rest was used for testing our model. We used a machine learning approach for our model. Seven different supervised machine learning algorithms (Logistic Regression, Decision Tree Classifier, XGB Classifier, Random Forest Classifier, K-Nearest Neighbors Classifier, SVM) were used for selecting the best performing model. Random Forest Classifier performed the best for our approach with 97.50% of highest accuracy.

6.2 Conclusions:

As people are suffering from very different online attacks which cause them very much including money lost. In order to stop that we have to stop cybercrimes and make the online platforms safer. Our model can help people going to unknown websites without falling into any cyber criminals' traps. Our model just needs the URL of the websites and can detect if it's fake or not. So, the users will not be so confused and can detect fake websites before clicking on them. It will be a great step to stop online crimes as the user won't be using dangerous URL links anymore. The features were selected based

on the best statistics of the fake URLs. The outcome that we gained was very efficient so our model was a success. We think our model will impact very much on making the online platforms safer.

6.3 Implication for Further Study:

As we estimated the outcome of our work was very satisfactory. But as the web changes simultaneously we have to improve our model further in the future. We plan to make a web and android application version of our work. We intend to create a customer service provider point for our users. There will be an AI based reply system which will reply to the questions of users by itself. Also, there will always be an admin waiting for users to give them feedback in any emergency situation. We plan to take feedback from the users who are directly becoming the victim of cybercrimes. By these we can know the new ways that will be invented in the future. And according to the changes that the cyber criminals are making to their URLs we will take necessary steps. New features will be created based on the new changes so that the model does not fall behind over time.

REFERENCES

1. 9 most common computer and internet cyber crimes/The Law Office of Elliott Kanter, available at <<<https://www.enkanter.com/article/9-most-common-computer-and-internet-cyber-crimes>>>, last accessed on 05-11-2021 at 10.50 PM
2. National Crime Records Bureau, available at <<<https://ncrb.gov.in/en/cyber-crimes-statesuts>>>, last accessed on 05-11-2021 at 10.58 PM.
3. Automated machine learning - datarobot: Ai cloud/DataRobot, available at <<<https://www.datarobot.com/platform/automated-machine-learning/>>>, last accessed on 05-11-2021 at 11.02 PM.
4. Sponsored: 81 ransomware statistics, data, trends and facts for 2021/SIGNAL Magazine, available at <<<https://www.afcea.org/content/sponsored-81-ransomware-statistics-data-trends-and-facts-2021>>>, last accessed on 05-11-2021 at 10.28 PM.
5. Kulkarni, A., & L., L. (2019). "Phishing websites detection using machine learning", *International Journal of Advanced Computer Science and Applications*, vol. 10(7). <https://doi.org/10.14569/ijacsa.2019.0100702>.
6. Selvakumari, M., Sowjanya, M., Das, S., & Padmavathi, S. (2021, May). "Phishing website detection using machine learning and deep learning techniques", In *Journal of Physics: Conference Series* (Vol. 1916, No. 1, p. 012169). <https://doi.org/10.1088/1742-6596/1916/1/012169>.
7. Mahajan, R., & Siddavatam, I. (2018). "Phishing Website Detection using Machine Learning Algorithms", *International Journal of Computer Applications*, vol. 181(23), pp. 45–47, <https://doi.org/10.5120/ijca2018918026>.
8. al Asaad, B., & Erascu, M. (2018). "A Tool for Fake News Detection", *2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*. Published. <https://doi.org/10.1109/synasc.2018.00064>.
9. Tupsamudre, H., Singh, A. K., & Lodha, S. (2019). "Everything Is in the Name – A URL Based Approach for Phishing Detection", *Lecture Notes in Computer Science*, pp. 231–248, https://doi.org/10.1007/978-3-030-20951-3_21
10. B. Eshete, A. Villafiorita and K. Weldemariam, "Malicious Website Detection: Effectiveness and Efficiency Issues," 2011 First SysSec Workshop, 2011, pp. 123-126, doi: 10.1109/SysSec.2011.9.
11. M. D. Bhagwat, P. H. Patil and T. S. Vishawanath, "A Methodical Overview on Detection, Identification and Proactive Prevention of Phishing Websites," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 1505-1508, doi: 10.1109/ICICV50876.2021.9388441.

12. A. Alswailem, B. Alabdullah, N. Alrumayh and A. Alsedrani, "Detecting Phishing Websites Using Machine Learning," 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), 2019, pp. 1-6, doi: 10.1109/CAIS.2019.8769571.
13. J. Rashid, T. Mahmood, M. W. Nisar and T. Nazir, "Phishing Detection Using Machine Learning Technique," 2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH), 2020, pp. 43-46, doi: 10.1109/SMART-TECH49988.2020.00026.
14. S. Sindhu, S. P. Patil, A. Sreevalsan, F. Rahman and M. S. A. N., "Phishing Detection using Random Forest, SVM and Neural Network with Backpropagation," 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), 2020, pp. 391-394, doi: 10.1109/ICSTCEE49637.2020.9277256.
15. L. A. T. Nguyen, B. L. To, H. K. Nguyen and M. H. Nguyen, "A novel approach for phishing detection using URL-based heuristic," 2014 International Conference on Computing, Management and Telecommunications (ComManTel), 2014, pp. 298-303, doi: 10.1109/ComManTel.2014.6825621.
16. A. A. Ahmed and N. A. Abdullah, "Real time detection of phishing websites," 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2016, pp. 1-6, doi: 10.1109/IEMCON.2016.7746247.
17. K. Gajera, M. Jangid, P. Mehta and J. Mittal, "A Novel Approach to Detect Phishing Attack Using Artificial Neural Networks Combined with Pharming Detection," 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), 2019, pp. 196-200, doi: 10.1109/ICECA.2019.8822053.
18. PhishTank | Join the fight against phishing, available at <<<https://phishtank.org/>>>, last accessed on 24-11-2021 at 12.10 AM.
19. aa419 - Fake Sites Database, available at <<<https://db.aa419.org/fakebankslist.php>>>, last accessed on 24-11-2021 at 12.15 AM.
20. Alexa - Top sites, available at <<<https://www.alexa.com/topsites>>>, last accessed on 24-11-2021 at 12.18 AM.
21. Top 1000 Websites, available at <<<https://dataforseo.com/top-1000-websites>>>, last accessed on 24-11-2021 at 12.20 AM.
22. URL 2016 | Datasets | Research | Canadian Institute for Cybersecurity | UNB, available at <<<https://www.unb.ca/cic/datasets/url-2016.html>>>, last accessed on 24-11-2021 at 12.23 AM.
23. Error 429 (Too Many Requests) - Quora, available at <<<https://www.quora.com/Why-do-you-usually-find-domain-names-instead-of-IP-addresses-in-a-URL>>>, last accessed on 24-11-2021 at 12.27 AM.

24. Valid URL Characters: Safe & Unsafe Character List, available at <<<https://abramillar.com/2018/01/15/special-characters-short-words-urls/>>>, last accessed on 24-11-2021 at 12.30 AM.
25. How to Create SEO Friendly URLs, available at <<<https://neilpatel.com/blog/seo-urls/>>>, last accessed on 24-11-2021 at 12.34 AM.
26. Maximum domain name length, available at <<<https://webmasters.stackexchange.com/questions/16996/maximum-domain-name-length>>>, last accessed on 24-11-2021 at 12.36 AM.
27. 8 Characteristics of Top Domain Names, available at <<<https://www.networksolutions.com/blog/establish/domains/8-characteristics-of-top-domain-names>>>, last accessed on 24-11-2021 at 12.43 AM.
28. Wang, W., Zhang, F., Luo, X., & Zhang, S. (2019). "PDRCNN: Precise Phishing Detection with Recurrent Convolutional Neural Networks", Security and Communication Networks, 2019, pp. 1–15. <https://doi.org/10.1155/2019/2595794>
29. URL redirection, available at <https://en.wikipedia.org/wiki/URL_redirection>, last accessed on 24-11-2021 at 12.54 AM.
30. Why is HTTP not secured, available at <<<https://www.cloudflare.com/en-gb/learning/ssl/why-is-http-not-secure/>>>, last accessed on 24-11-2021 at 12.56 AM.
31. TechCrunch is part of the Yahoo family of brands, available at <<<https://techcrunch.com/2009/04/06/are-url-shorteners-a-necessary-evil-or-just-evil/>>>, last accessed on 24-11-2021 at 01.00 AM.
32. X. Zou, Y. Hu, Z. Tian and K. Shen, "Logistic Regression Model Optimization and Case Analysis," 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), 2019, pp. 135-139, doi: 10.1109/ICCSNT47585.2019.8962457.
33. Python Decision Tree Classification with Scikit-Learn DecisionTreeClassifier, available at <<<https://www.datacamp.com/community/tutorials/decision-tree-classification-python>>>, last accessed on 03-12-2021 at 12.41 AM.
34. Candice Bentéjac, Anna Csörgő, Gonzalo Martínez-Muñoz, "A Comparative Analysis of XGBoost", Research gate, pp. 1-20, November 2019
35. k-nearest neighbor algorithm in Python, available at <<<https://www.geeksforgeeks.org/k-nearest-neighbor-algorithm-in-python/>>>, last accessed on 03-12-2021 at 12.54 AM.
36. Sklearn SVM (Support Vector Machines) with Python, available at <<<https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python>>>, last accessed on 03-12-2021 at 12.56 AM.

Turnitin Originality Report

Processed on: 04-Dec-2021 21:36 +06
ID: 1720414140
Word Count: 7285
Submitted: 1

Doula By Most. Hena

Similarity Index	Similarity by Source
6%	Internet Sources: 4% Publications: 3% Student Papers: 3%

< 1% match (Internet from 07-Apr-2021) http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/5444/192-25-773%20%2823%25%29.pdf?isAllowed=y&sequence=1
< 1% match (Internet from 07-Apr-2021) http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/5149/162-15-7878%20%2818%29.pdf?isAllowed=y&sequence=1
< 1% match (Internet from 01-Oct-2021) http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/5263/161-15-7251%20%20%2820%29.pdf?isAllowed=y&sequence=1
< 1% match (Internet from 07-Apr-2021) http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/5158/152-15-6060%20%2817%29.pdf?isAllowed=y&sequence=1
< 1% match (Internet from 04-Aug-2020) http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/4092/P15438%20%286%29.pdf?isAllowed=y&sequence=1
< 1% match (Internet from 07-Apr-2021) http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/5207/162-15-7914%20%2825%29.pdf?isAllowed=y&sequence=1
< 1% match (student papers from 23-Oct-2018) Submitted to University of Sydney on 2018-10-23
< 1% match (student papers from 07-Nov-2021) Submitted to University of Sydney on 2021-11-07
< 1% match (Internet from 25-Nov-2021) https://lineaproblem-krat.com/community/tutorials/logistic-regression-R4q-4397h2atl3
< 1% match (Internet from 14-Jan-2021) https://www.freecodecamp.org/news/how-i-built-a-mobile-app-for-online-shopping-amid-covid-19-lock-down/
< 1% match (student papers from 25-Aug-2016) Submitted to King's College on 2016-08-25
< 1% match (student papers from 19-Apr-2021) Submitted to Colorado Technical University on 2021-04-19
< 1% match (publications) "Machine Learning for Healthcare Applications", Wiley, 2021
< 1% match (Internet from 24-Nov-2021) https://medium.com/analytics-vidhya/algorithmic-momentum-trading-strategy-747e726d04b4
< 1% match (student papers from 05-Mar-2021) Submitted to Asia Pacific Institute of Information Technology on 2021-03-05
< 1% match (publications) "Proceedings of Research and Applications in Artificial Intelligence", Springer Science and Business Media LLC, 2021
< 1% match (Internet from 21-Jan-2020) https://pdfs.semanticscholar.org/080a/efaae4269cfb5ca69f3497136caee6c54e64.pdf
< 1% match () Verbelen, Moira Inez, "Statistical Dissection of Pharmacogenetics with Machine Learning", 2017