FEATURE SELECTION BASED PHISHING URL DETECTION USING SUPERVISED MACHINE LEARNINGMETHODS

NANDHINI .N

(20PCA009)

Project Report Submitted

In partial fulfillment of the requirements for the Award of

Master of Computer Applications

DEPARTMENT OF COMPUTER SCIENCE

AVINASHILINGAM INSTITUTE FOR HOME SCIENCE AND

HIGHER EDUCATION FOR WOMEN

COIMBATORE – 641043

MAY – 2022

ABSTRACT

Due to the innovations in digital technologies the digital world is fast expanding and evolving towards cyber crimes. Cyber criminals are relied on the illegal use of digital assets, particularly personal credentials, financial data etc. Cyber criminals have expanded their data collection methods, but social engineering attacks remain their preferred way. Phishing is a sort ofsocial engineering crime in which an attacker attempts to steal someone's identity. Phishing is one of the major cyber attacks with many internet users falling victim to it. Phishing attack mostly target EMAILS, WEBSITE, URLs, SMS, VOICE and so on. Phishers develop cloned websitesand distribute the URL(s) to a large number of people by email, text, or social media. The aim of the project is to detect the Phishing URLs based on the various feature selection methods using supervised machine learning methods. Machine learning is the branch of artificial intelligence which helps to detect the phishing attack without any human intervention. The process of phishing URLs detection using supervised machine learning methods comprises of five phases.

The Phase 1 is the data collection in which Phishing URL dataset is used acquired from kaggle repository. Phase 2, deals with data preprocessing to remove the irrelevant data. In Phase 3, various feature selection techniques includes filter, wrapper and embedded feature selection methods are used to identify the significant features of the dataset which derive the appropriate result. Phase 4, deals with model building using supervised machine learning methods includes K-Nearest Neighbor (K-NN), Random forest and Logistic regression. In Phase 5, the comparative analysis is made between the supervised machine learning models to suggest the suitable model for Phishing URL detection. The Evaluation of the models are based on the performance metrics such as accuracy, precision, recall, f1 score and ROC curve in an effective way. Based on the comparative analysis embedded based feature selection attains 88% accuracy and Random forest Supervised Machine Learning model performs better with 97% accuracy in detecting Phishing URLs effectively with the proposed methodology.

METHODOLOGY

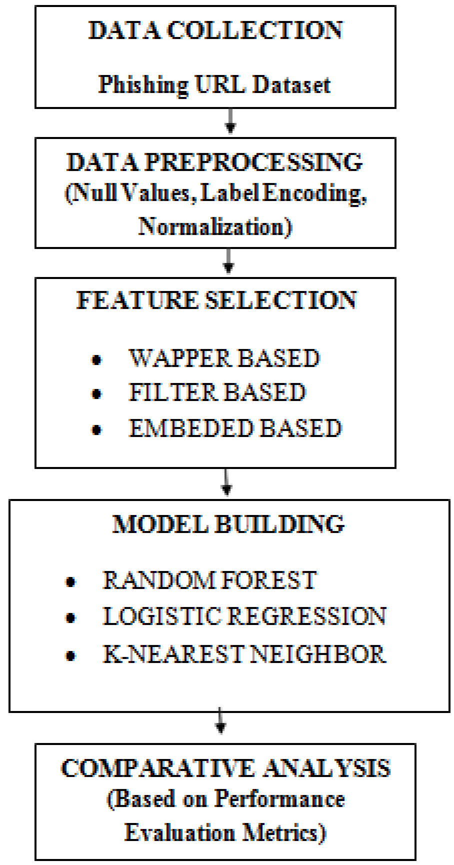


FIGURE 4.1 PROPOSED METHODOLOGY

MOTIVATION AND JUSTIFICATION

According to proof point survey nearly 90 percents of global organization were targeted with spear Phishing attacks in 2019, reflecting cyber criminals continued focus on compromising individual end users. Thus, the Phishing URL detection using Supervised Machine Learning techniques incorporated with various feature selection methods has its own space and necessity to be developed. This will further support automation of Phishing URL detection.

PROBLEM STATEMENT

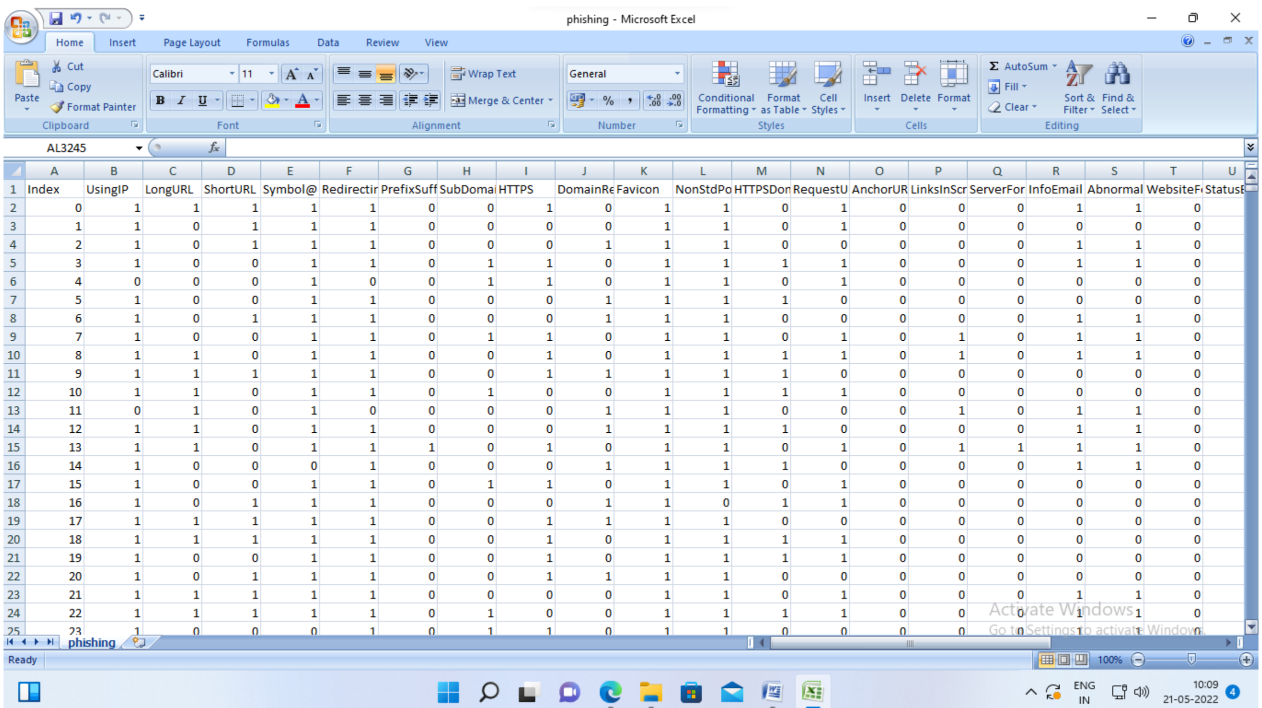
To identify the Phishing URLs in order to avoid the illegitimate user access controls that are attempting to acquire the user personal credentials and take over the device controls that are connected to a network.

OBJECTIVE

To develop a feature selection based Supervised Machine learning Models to detect and classify the Phishing URLs in order to handle the Phishing URLs. It further tries to intrude into the system through malicious link and gain personal information by user access control as a legitimate one. Based on the performance evaluation metrics, to suggest a suitable feature selection method and supervised machine learning classifier that detects Phishing URL appropriately.

RESULT AND DISCUSION

PHASE 1: DATA COLLECTION



Index	UsingIP	LongURL	ShortURL	Symbol@	Redirectir	PrefixSuff	SubDomai	HTTPS	DomainRe	Favicon	NonStdPo	HTTPSDon	RequestU	AnchorUR	LinksInScr	ServerFor	InfoEmail	Abnormal	WebsiteF	StatusE
0	1	1	1	1	1	0	0	1	0	1	1	0	1	0	0	0	1	1	0	
1	1	0	1	1	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	
2	1	0	1	1	1	0	0	0	1	1	1	0	0	0	0	0	1	1	0	
3	1	0	0	1	1	0	1	1	0	1	1	1	1	0	0	0	1	1	0	
4	0	0	0	1	0	0	1	1	0	1	1	0	1	0	0	0	0	0	0	
5	1	0	0	1	1	0	0	0	1	1	1	1	0	0	0	0	0	0	0	
6	1	0	1	1	1	0	0	0	1	1	1	0	0	0	0	0	1	1	0	
7	1	0	0	1	1	0	1	1	0	1	1	0	1	0	1	0	1	1	0	
8	1	1	0	1	1	0	0	1	0	1	1	1	1	0	1	0	1	1	0	
9	1	1	1	1	1	0	0	1	1	1	1	1	0	0	0	0	0	0	0	
10	1	1	0	1	1	0	1	0	0	1	1	1	1	0	0	0	0	0	0	
11	0	1	0	1	0	0	0	0	1	1	1	0	0	0	1	0	1	1	0	
12	1	1	0	1	1	0	0	0	1	1	1	1	0	0	0	0	1	1	0	
13	1	1	0	1	1	1	0	1	0	1	1	0	1	0	1	1	1	1	0	
14	1	0	0	0	1	0	0	0	1	1	1	1	0	0	0	0	1	1	0	
15	1	0	0	1	1	0	1	1	0	1	1	0	1	0	0	0	0	0	0	
16	1	0	1	1	1	0	0	0	1	1	0	1	1	0	0	0	0	0	0	
17	1	1	1	1	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	
18	1	1	1	1	1	0	0	1	0	1	1	1	1	0	0	0	0	0	0	
19	1	0	0	1	1	0	0	1	0	1	1	1	1	0	0	0	0	0	0	
20	1	0	1	1	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	
21	1	1	1	1	1	0	0	0	0	1	1	0	1	0	0	0	1	1	0	
22	1	1	1	1	1	0	1	0	0	1	1	1	1	0	0	0	1	1	0	
23	1	0	0	0	1	0	1	1	0	1	1	0	0	0	0	0	1	1	0	

Figure 5.1 Phishing URL Dataset

Description

The dataset is gathered from Kaggle repository, Dataset Source: https://www.kaggle.com/eswarchandt/phishing-website-detector. A collection of website URLs for 11054 websites. Each sample has 30 website parameters and a class label indicating a phishing website or not (1 or 0). The overview of this dataset, it has 11054 samples with 32 features.

PHASE 2: DATA PREPROCESSING

Null values

```
df.isnull()
```

	Index	UsingIP	LongURL	ShortURL	Symbol@	Redirecting//	PrefixSuffix-	SubDomains	HTTPS	DomainRegLen	...	UsingPopupWindow	IframeRedir
0	False	False	False	False	False	False	False	False	False	False	...	False	
1	False	False	False	False	False	False	False	False	False	False	...	False	
2	False	False	False	False	False	False	False	False	False	False	...	False	
3	False	False	False	False	False	False	False	False	False	False	...	False	
4	False	False	False	False	False	False	False	False	False	False	...	False	
...	...	...	...	...	...	...	...	...	...	...	...	...	
11049	False	False	False	False	False	False	False	False	False	False	...	False	
11050	False	False	False	False	False	False	False	False	False	False	...	False	
11051	False	False	False	False	False	False	False	False	False	False	...	False	
11052	False	False	False	False	False	False	False	False	False	False	...	False	
11053	False	False	False	False	False	False	False	False	False	False	...	False	

```
df.notnull()
```

	Index	UsingIP	LongURL	ShortURL	Symbol@	Redirecting//	PrefixSuffix-	SubDomains	HTTPS	DomainRegLen	...	UsingPopupWindow	IframeRedirect
0	True	True	True	True	True	True	True	True	True	True	...	True	T
1	True	True	True	True	True	True	True	True	True	True	...	True	T
2	True	True	True	True	True	True	True	True	True	True	...	True	T
3	True	True	True	True	True	True	True	True	True	True	...	True	T
4	True	True	True	True	True	True	True	True	True	True	...	True	T
...	...	...	...	...	...	...	...	...	...	...	...	...	
11049	True	True	True	True	True	True	True	True	True	True	...	True	T
11050	True	True	True	True	True	True	True	True	True	True	...	True	T
11051	True	True	True	True	True	True	True	True	True	True	...	True	T
11052	True	True	True	True	True	True	True	True	True	True	...	True	T
11053	True	True	True	True	True	True	True	True	True	True	...	True	T

11054 rows × 32 columns

```
df.isnull()
```

	Index	UsingIP	LongURL	ShortURL	Symbol@	Redirecting//	PrefixSuffix-	SubDomains	HTTPS	DomainRegLen	...	UsingPopupWindow	IframeRedir
0	False	False	False	False	False	False	False	False	False	False	...	False	
1	False	False	False	False	False	False	False	False	False	False	...	False	
2	False	False	False	False	False	False	False	False	False	False	...	False	
3	False	False	False	False	False	False	False	False	False	False	...	False	
4	False	False	False	False	False	False	False	False	False	False	...	False	
...	...	...	...	...	...	...	...	...	...	...	...	...	
11049	False	False	False	False	False	False	False	False	False	False	...	False	
11050	False	False	False	False	False	False	False	False	False	False	...	False	
11051	False	False	False	False	False	False	False	False	False	False	...	False	
11052	False	False	False	False	False	False	False	False	False	False	...	False	
11053	False	False	False	False	False	False	False	False	False	False	...	False	

```
df.notnull()
```

	Index	UsingIP	LongURL	ShortURL	Symbol@	Redirecting//	PrefixSuffix-	SubDomains	HTTPS	DomainRegLen	...	UsingPopupWindow	IframeRedirect
0	True	True	True	True	True	True	True	True	True	True	...	True	T
1	True	True	True	True	True	True	True	True	True	True	...	True	T
2	True	True	True	True	True	True	True	True	True	True	...	True	T
3	True	True	True	True	True	True	True	True	True	True	...	True	T
4	True	True	True	True	True	True	True	True	True	True	...	True	T
...	...	...	...	...	...	...	...	...	...	...	...	...	
11049	True	True	True	True	True	True	True	True	True	True	...	True	T
11050	True	True	True	True	True	True	True	True	True	True	...	True	T
11051	True	True	True	True	True	True	True	True	True	True	...	True	T
11052	True	True	True	True	True	True	True	True	True	True	...	True	T
11053	True	True	True	True	True	True	True	True	True	True	...	True	T

11054 rows × 32 columns

Figure 5.2.1 Finding missing values using Null Values

```python
# Import label encoder
from sklearn import preprocessing

# label_encoder object knows how to understand word labels.
label_encoder = preprocessing.LabelEncoder()

# Encode labels in column 'species'.
df['class']= label_encoder.fit_transform(df['class'])

df['class'].unique()
```

```
array([0, 1], dtype=int64)
```

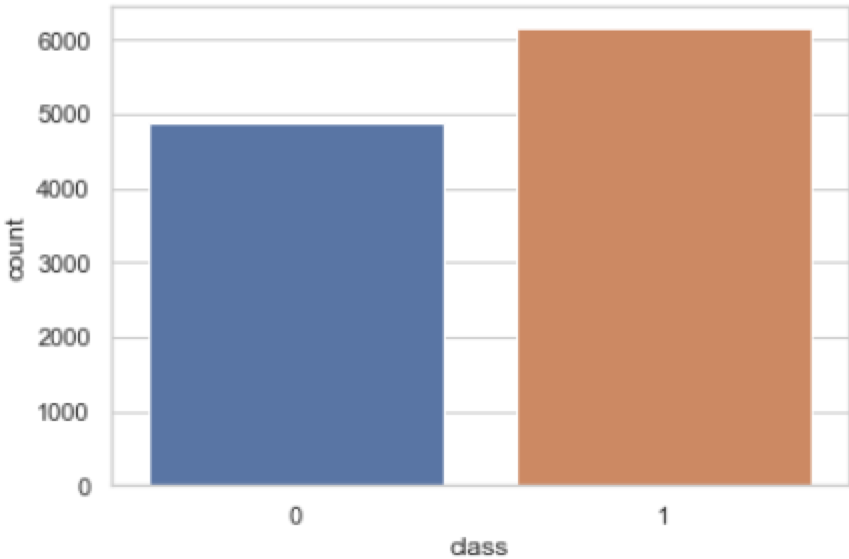


Figure 5.2.2 Exploratory Data Analysis and Label Encoding

Description

Figure 5.2.2 represents the Phishing URL attacks consisting of a large number of related variables contains 6157 Phishing URL and real URL data consists of 4897 records. It shows the count of Phishing URL and real URL data.

Normalization

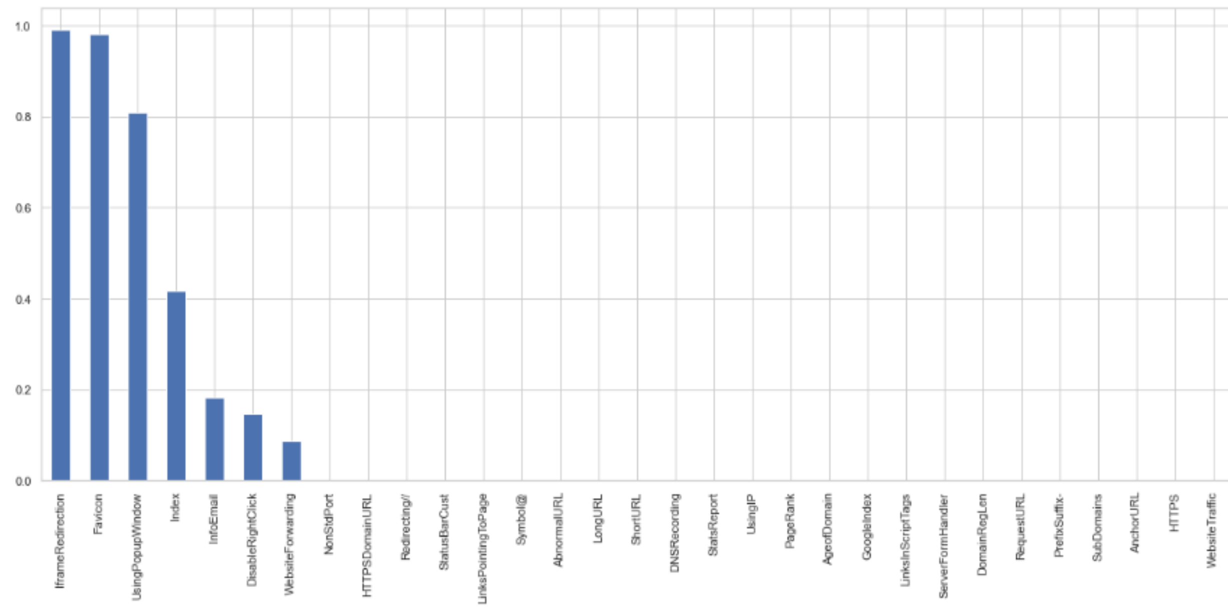
```
          Index  UsingIP  LongURL  ShortURL  Symbol@  Redirecting//  \
0      0.000000      1.0      1.0       1.0      1.0            1.0
1      0.000090      1.0      0.0       1.0      1.0            1.0
2      0.000181      1.0      0.0       1.0      1.0            1.0
3      0.000271      1.0      0.0       0.0      1.0            1.0
4      0.000362      0.0      0.0       0.0      1.0            0.0
...         ...      ...      ...       ...      ...            ...
11049  0.999638      1.0      0.0       1.0      0.0            1.0
11050  0.999729      0.0      1.0       1.0      0.0            0.0
11051  0.999819      1.0      0.0       1.0      1.0            1.0
11052  0.999910      0.0      0.0       1.0      1.0            1.0
11053  1.000000      0.0      0.0       1.0      1.0            1.0

       PrefixSuffix-  SubDomains  HTTPS  DomainRegLen  ...  UsingPopupWindow  \
0                0.0         0.0    1.0           0.0  ...               1.0
1                0.0         0.0    0.0           0.0  ...               1.0
2                0.0         0.0    0.0           1.0  ...               1.0
3                0.0         1.0    1.0           0.0  ...               0.0
4                0.0         1.0    1.0           0.0  ...               1.0
...              ...         ...    ...           ...  ...               ...
11049            1.0         1.0    1.0           0.0  ...               0.0
11050            0.0         1.0    0.0           0.0  ...               0.0
11051            0.0         1.0    0.0           0.0  ...               1.0
11052            0.0         0.0    0.0           1.0  ...               0.0
11053            0.0         0.0    0.0           1.0  ...               1.0

       IframeRedirection  AgeofDomain  DNSRecording  WebsiteTraffic  PageRank  \
0                    1.0          0.0           0.0             0.0       0.0
1                    1.0          1.0           0.0             1.0       0.0
2                    1.0          0.0           0.0             1.0       0.0
3                    1.0          0.0           0.0             0.0       0.0
4                    1.0          1.0           1.0             1.0       0.0
```

Figure 5.2.3 Normalization using Feature Scaling

PHASE 3: FEATURE SELECTION

Univariate Method

```
IframeRedirection        9.902037e-01
Favicon                  9.812869e-01
UsingPopupWindow         8.094827e-01
Index                    4.171068e-01
InfoEmail                1.827377e-01
DisableRightClick        1.472192e-01
WebsiteForwarding        8.688614e-02
NonStdPort               1.582756e-03
HTTPSDomainURL           1.526618e-04
Redirecting//            1.331781e-04
StatusBarCust            4.568094e-05
LinksPointingToPage      1.352631e-06
Symbol@                  1.430011e-07
AbnormalURL              7.911640e-08
LongURL                  3.524346e-08
ShortURL                 2.323022e-11
DNSRecording             1.901198e-11
StatsReport              5.153490e-15
UsingIP                  1.094420e-20
PageRank                 3.656009e-22
AgeofDomain              2.952538e-30
GoogleIndex              6.719496e-35
LinksInScriptTags        3.636019e-61
ServerFormHandler        2.406392e-85
DomainRegLen            9.003742e-106
RequestURL              1.665025e-134
PrefixSuffix-           2.503531e-253
SubDomains              6.510602e-290
AnchorURL                0.000000e+00
HTTPS                    0.000000e+00
WebsiteTraffic           0.000000e+00
dtype: float64
```



Forward Feature Selection

```
#Get the selected feature index.
model.k_feature_idx_
```

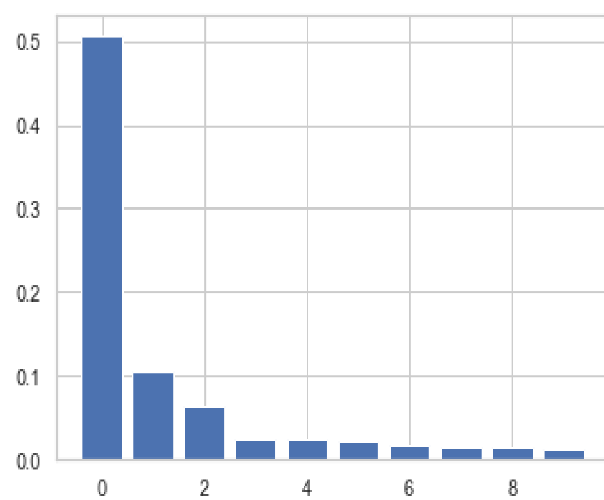
```
(0, 1, 2, 3, 4, 5, 6, 7, 8, 9)
```

```
#Get the column name for the selected feature.
model.k_feature_names_
```

```
('IframeRedirection',
 'Favicon',
 'UsingPopupWindow',
 'Index',
 'InfoEmail',
 'DisableRightClick',
 'WebsiteForwarding',
 'NonStdPort',
 'HTTPSDomainURL',
 'Redirecting//')
```

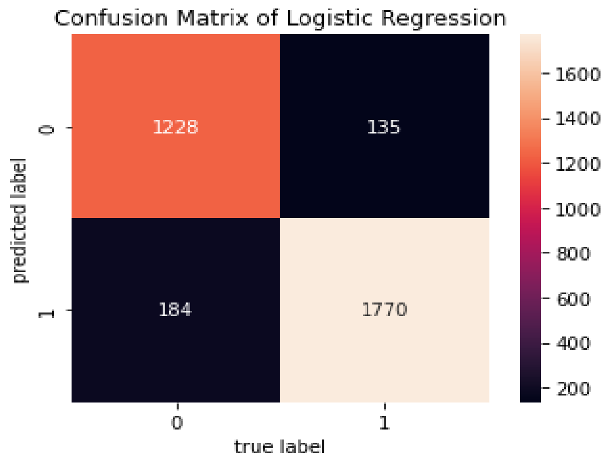
Tree Based Regression Feature Selection

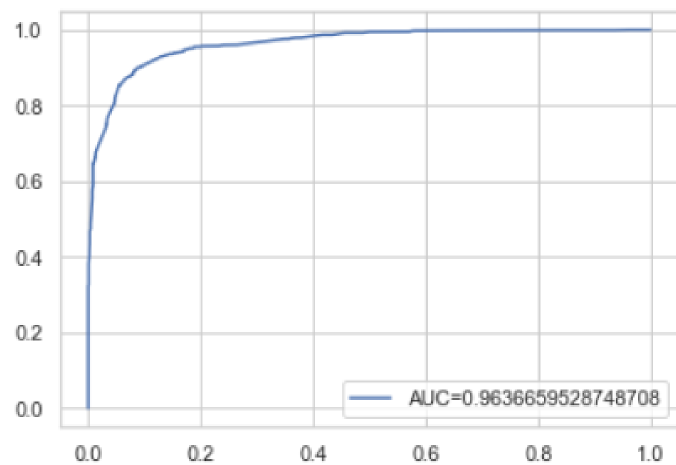
	Specs	Score
8	HTTPS	0.505778
6	PrefixSuffix-	0.106829
14	AnchorURL	0.064345
26	WebsiteTraffic	0.024049
16	ServerFormHandler	0.023941
7	SubDomains	0.023870
25	DNSRecording	0.017083
28	GoogleIndex	0.014940
17	InfoEmail	0.014504
2	LongURL	0.013841



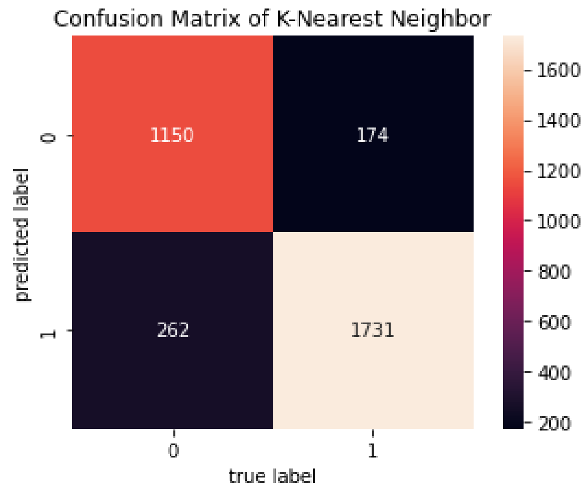
PHASE 4: MODEL BULDING

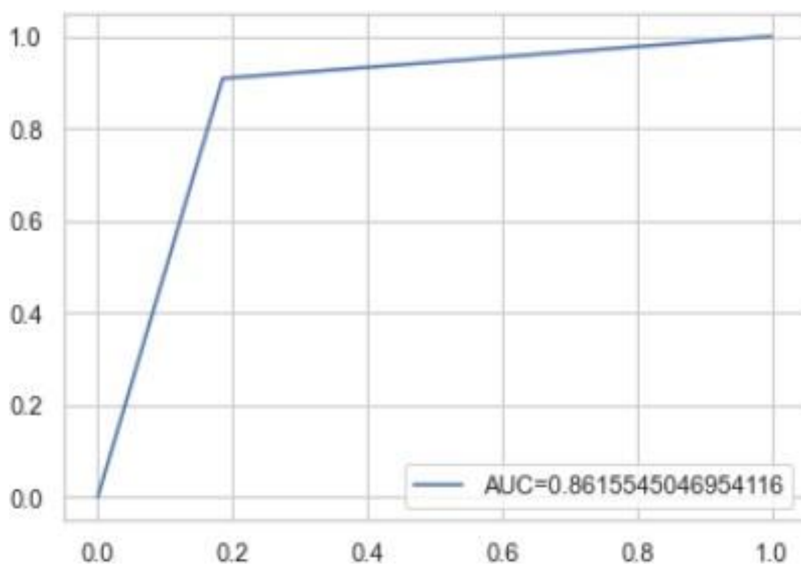
Logistic Regression



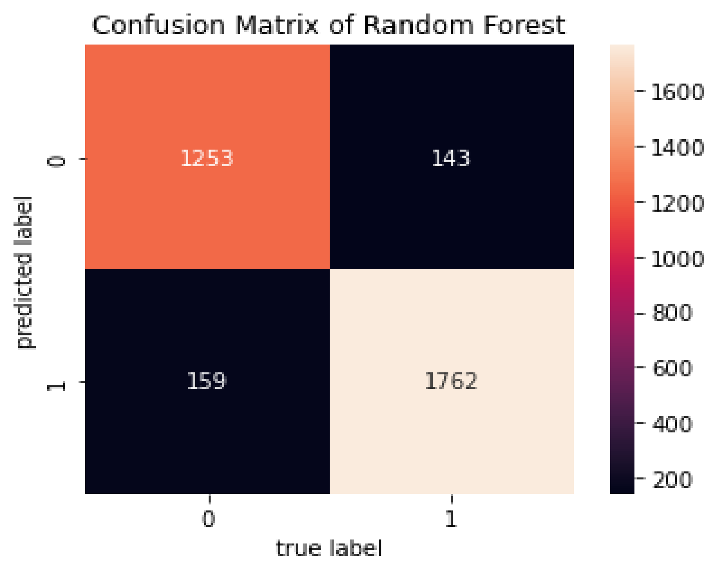


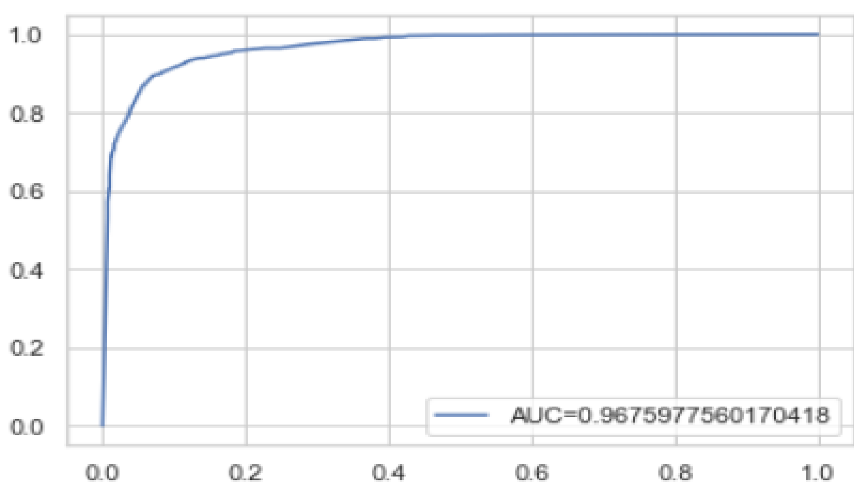
K-Nearest Neighbor



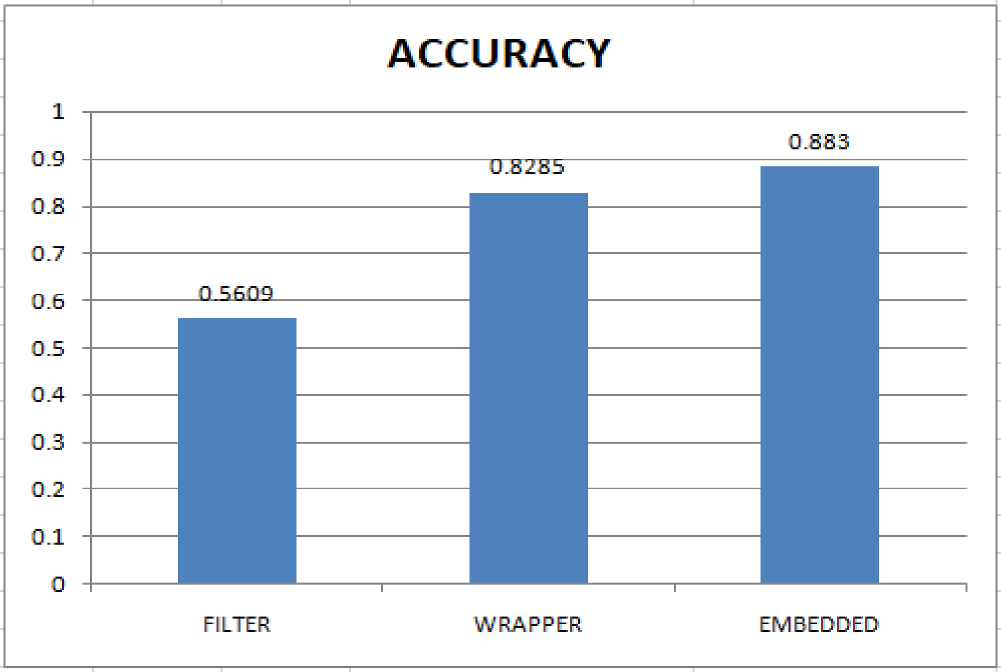


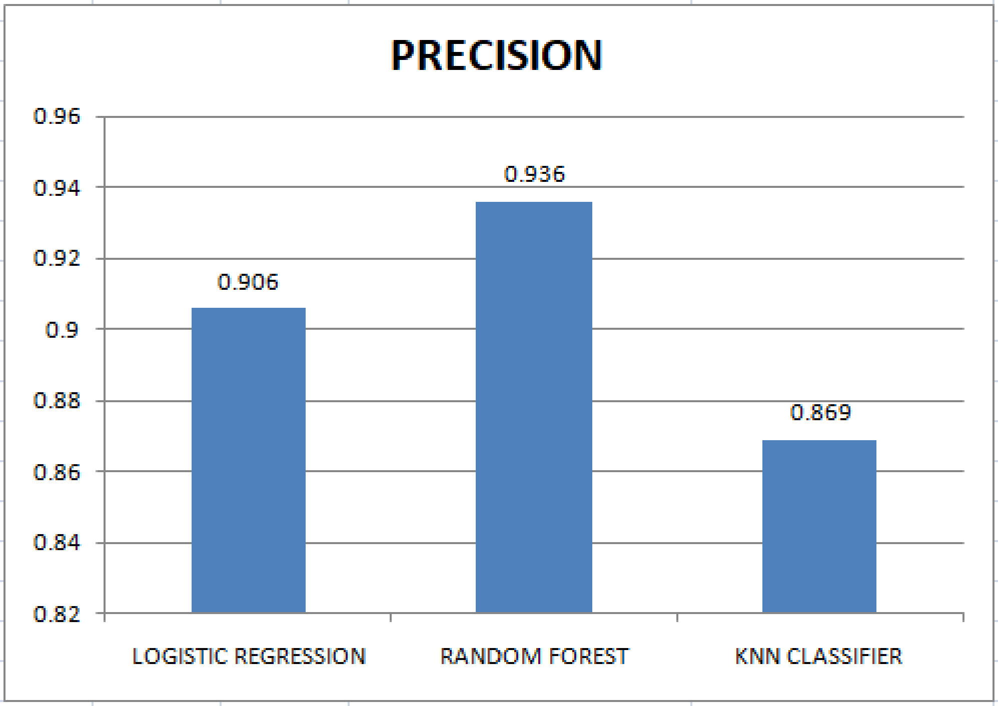
Random Forest

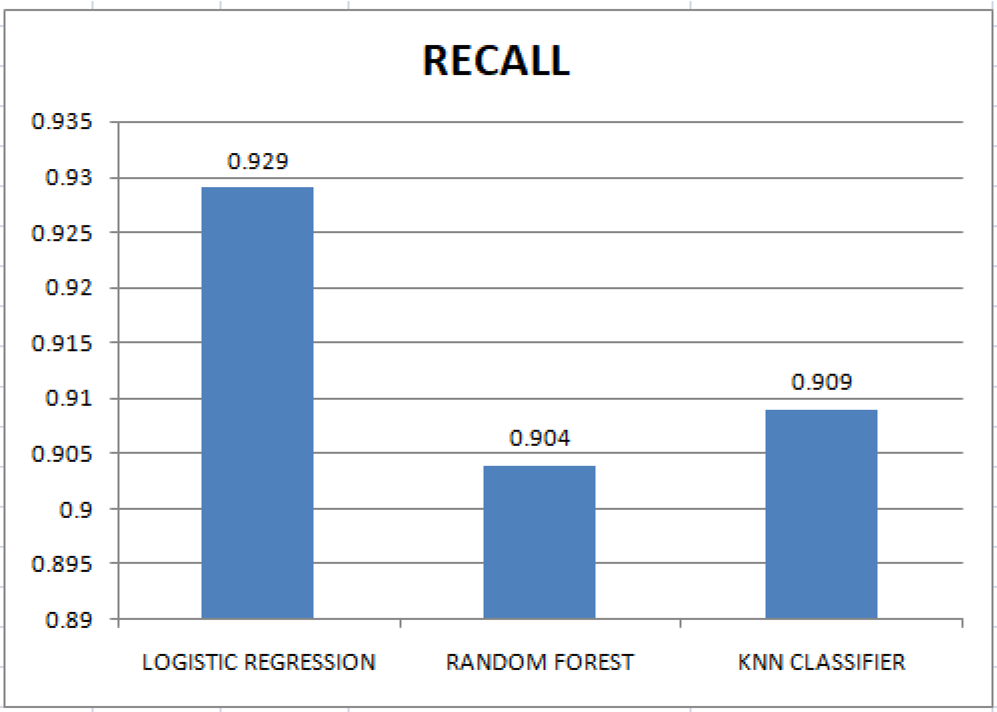


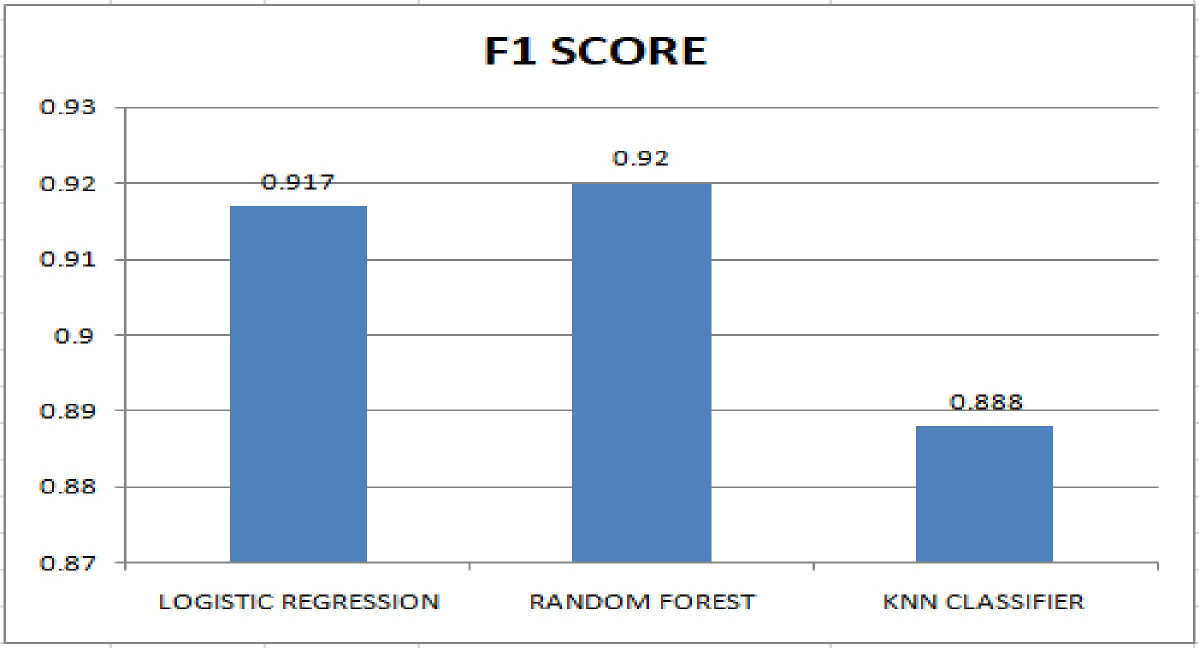


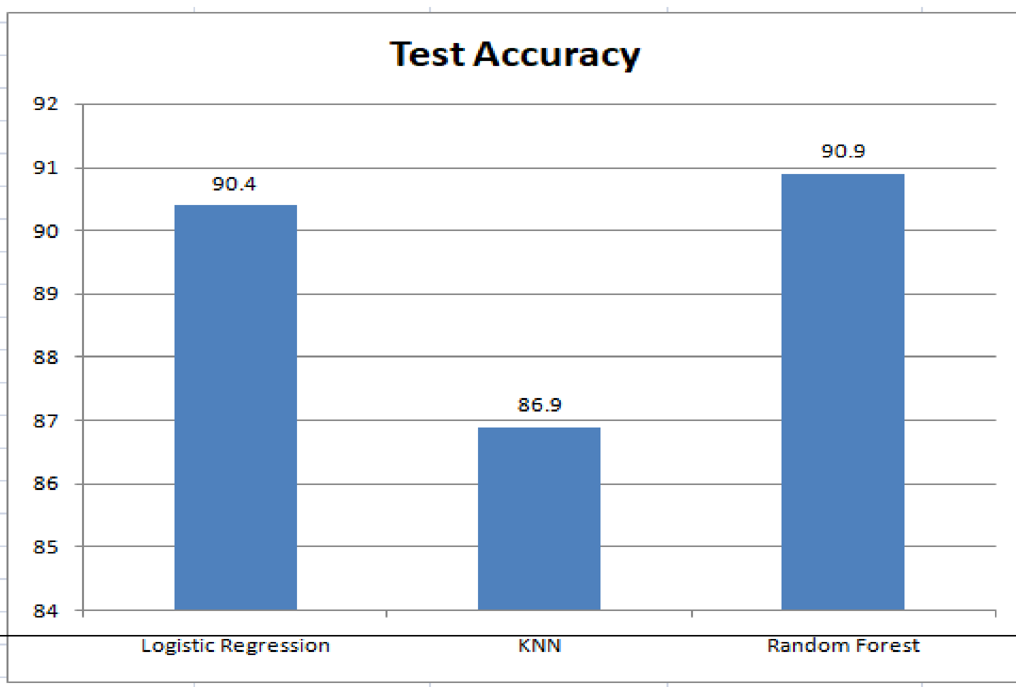
PHASE 5: COMARATIVE ANALYSIS

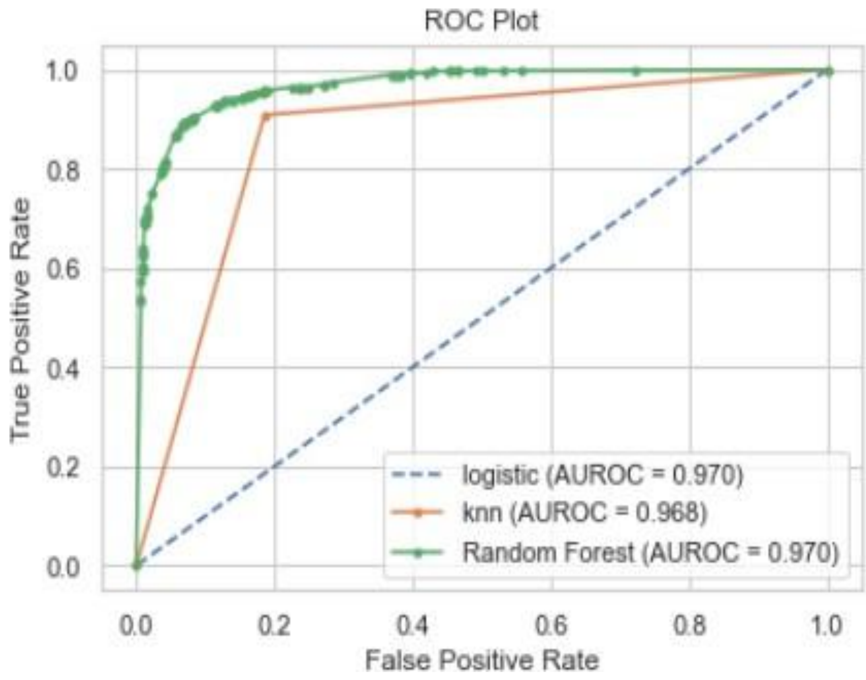



PRECISION
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.82
0.906
0.936
0.869
LOGISTIC REGRESSION
RANDOM FOREST
KNN CLASSIFIER


RECALL
0.935
0.93
0.925
0.92
0.915
0.91
0.905
0.9
0.895
0.89
0.929
0.904
0.909
LOGISTIC REGRESSION
RANDOM FOREST
KNN CLASSIFIER


F1 SCORE
0.93
0.92
0.91
0.9
0.89
0.88
0.87
0.917
0.92
0.888
LOGISTIC REGRESSION
RANDOM FOREST
KNN CLASSIFIER


Test Accuracy
92
91
90
89
88
87
86
85
84
90.4
86.9
90.9
Logistic Regression
KNN
Random Forest



SCOPE FOR FUTURE DEVELOPMENT

In this project, it deals with Supervised Machine Learning models to detection Phishing URLs based on various feature Selection techniques. It is found that Phishing attacks are very crucial and it is important to get a robust mechanism to detect it. In future, this project will be enhanced using deep learning algorithms to design an effective framework to detect the Phishing URLs accurately with more improved results.

CONCLUSION

Phishing attack is one of the common types of cyber-attacks where the attackers steal user's credential information in the form of URLs, E-mails, SMS, or through phone calls where the user loses their sensitive information which may leads to cyber-threat. In this project the Phishing URL dataset from Kaggle repository has been used, Data Preprocessing methods are applied to refine the data, various Feature selection techniques includes Filter, Wrapper and Embedded Feature Selection methods was implemented to acquire the appropriate features that thrive to detect the Phishing URL detection, Supervised Machine Learning models such as logistic regression, K-NN and Random forest classifiers are built and the performance of the classifiers are evaluated using the performance metrics such as accuracy, precision, recall, F1 Score and ROC Curve. Based on the comparative analysis between the performances of the classifiers, it shows that embedded feature selection method attains 88% accuracy towards selection of top ten features and Random Forest Classifier achieve better accuracy of 97% compared with other supervised machine learning models in terms of detecting the Phishing URLs more precisely.

REFERENCE

1. G. K., . S., . N., . S. P., . S., & . V. (2019). URL PHISHING DATA ANALYSIS AND DETECTING PHISHING ATTACKS USING MACHINE LEARNING IN NLP. *International Journal of Engineering Applied Sciences and Technology*, *3*(10), 26–31. https://doi.org/10.33564/ijeast.2019.v03i10.007

2. Abusaimeh, H. (2021). Detecting the Phishing Website with the Highest Accuracy. *TEM Journal*, 947–953. https://doi.org/10.18421/tem102-58

3. Assegie*, T. A. (2021). K-Nearest Neighbor Based URL Identification Model for Phishing Attack Detection. *Indian Journal of Artificial Intelligence and Neural Networking*, *1*(2), 18–21. https://doi.org/10.35940/ijainn.b1019.041221

4. Banik, B., & Sarma, A. (2018). Phishing URL detection system based on URL features using SVM. *International Journal of Electronics and Applied Research*, *5*(2), 40–55. https://doi.org/10.33665/ijear.2018.v05i02.003

5. E., B., & K., T. (2015). Phishing URL Detection: A Machine Learning and Web Mining-based Approach. *International Journal of Computer Applications*, *123*(13), 46–50. https://doi.org/10.5120/ijca2015905665

6. Glăvan, D. (2020). Detection of Phishing attacks using the anti-Phishing framework. *Scientific Bulletin of Naval Academy*, *XXIII*(1), 208–212. https://doi.org/10.21279/1454-864x-20-i1-028

7. Joshi, A., & Pattanshetti, P. T. R. (2019). Phishing Attack Detection using Feature Selection Techniques. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3418542

8. Kim, T., Hong, J., Park, N., & Kim, S. W. (2020). Graph-Based Phishing URL Detection. *KIISE Transactions on Computing Practices*, *26*(3), 156–160. https://doi.org/10.5626/ktcp.2020.26.3.156

9. Patil, S. S., & Dinesha, H. A. (2022). URL Redirection Attack Mitigation in Social Communication Platform using Data Imbalance Aware Machine Learning Algorithm. *Indian Journal of Science and Technology*, *15*(11), 481–488. https://doi.org/10.17485/ijst/v15i11.1813

10. Xiu, W. (2021). Malicious URL Detection Algorithm Based on Multi Neural Network Series. *CONVERTER*, 579–590. https://doi.org/10.17762/converter.209

Book Reference

- Ganapathi, P. ., Shanmugapriya, D. ., & Roshni, A. . (2022). Dynamic Analysis Based Mobile Malware Classification Using Supervised Machine Learning Methods. *Dynamic Analysis Based Mobile Malware Classification Using Supervised Machine Learning Methods,* 1–168. https://doi.org/10.9734/bpi/mono/978-93-5547-441-4