ASSESSING PHISHING DETECTION APPROACHES USING VARIED DATA SETS AND CLASSIFICATION SCHEMES

Madiha Sadaf[1], Md. Abbas Qureshi[2], Md. Ateeq Ur Rahman[3]

[1]PG Scholar, Department of CSE, Shadan College of Engineering and Technology, Hyderabad, Telangana, India-500086, Madiha.sadaf1199@gmail.com

[2] Associate Professor, Department of CSE, Shadan College of Engineering and Technology, Hyderabad, Telangana, India-500086, mohammedabbas09@yahoo.com

[3]Professor, Department of CSE, Shadan College of Engineering and Technology, Hyderabad, Telangana, India-500086, mail_to_ateeq@yahoo.com

ABSTRACT:

Phishing attacks have evolved into a major cybersecurity concern, prompting extensive research to identify the most effective methods for classifying and detecting these deceptive tactics, which aim to deceive individuals and organizations into revealing sensitive information. This project addresses a notable gap in prior research by systematically evaluating various classification techniques under changing data conditions, ensuring that they are not limited to specific datasets or methods, thus offering a broader perspective on their effectiveness in combating phishing attacks. The study conducted assessments on thirteen contemporary classification techniques that are commonly utilized in preliminary research related to phishing. It subjected them to ten diverse performance measures, aiming to provide a comprehensive understanding of their capabilities. The findings of this research contribute valuable insights into the realm of phishing classification techniques, extending the knowledge base beyond what was previously explored in related studies, and ultimately assisting in the development of more effective countermeasures against phishing threats. The project incorporates the Stacking Classifier, a robust ensemble method, combining RF, MLP, and LightGBM models to achieve 100% accuracy in phishing attack classification. A user-friendly Flask-based front end enables easy user testing and performance evaluation. Implemented user authentication ensures secure access, contributing to a comprehensive evaluation of phishing classification techniques across diverse data sources and schemes.

***Index terms** - Benchmark testing, classification algorithms, performance evaluation, phishin*

.

1. INTRODUCTION

Phishing is a perilous threat to cybersecurity and according to The National Institute of Standards and Technology, it is attempts to get sensitive data, such as bank account numbers, or access to larger computerized systems by sending fraudulent requests through emails or websites. On average, the chances of being exposed to this attack in various sectors is 11% [1]. Phishing is also a socially engineered attack that tends to inflict physical or psychological harm on individuals and organizations [2]. The corporate sectors include technology, energy or utilities, retail, and financial services. These organizations are highly vulnerable to phishing. Therefore, cyber securitybased measures are needed to prevent these attacks [3]. Several studies have been carried out on phishing prevention, one based on its identification and classification.

Various techniques are used for the classification process, such as Random forest [4], [5], [6], [7], [8], [9], [10], support vector machine (SVM) [11], [12], [13], [14], Logistic regression [15], [16], [17], Multilayer perceptron (MLP) [18], C4.5 [19] and [20], and Naïve Bayes [21]. Each exhibits maximum performance according to the case it was applied. The results of the classification technique need not be generalized in all cases. Therefore, a comparative research must be carried out to resolve this gap.

However, only few studies have compared phishing classification techniques, such as [8], [18], [22], [23], and [24]. This comparative research is generally divided into four main parts, including phishing, the type of dataset, performance evaluation, and the techniques used. The data sources used by [8], [18], [22], [23], and [24] were obtained from a phishing website and URL, while [24] used raw emails sourced from Apache SpamAssassin and Nazario. The dominant performance evaluations are accuracy, precision, and F-measure. Random forest, SVM, and Naïve Bayes are the most widely used techniques. This comparative research has a gap, which is how the existing techniques affect various public datasets, including the balanced and unbalanced ones.

Interestingly, this research is based on the performance evaluation of the classification technique when using a specific unbalanced dataset for certain phishing types. This is similar to the processes adopted by studies that did not compare these classification techniques. Vaitkevicius and Marcinkevicius [18] used two balanced and one unbalanced datasets. It was reported that they obtained better results than previous comparisons. Gana and Abdulhamid [23] only used unbalanced public datasets, and it was proven that the classification performance changes in accordance with its subset scheme. This research is engineered by several studies that failed to prove how performance evaluation influences the techniques used to classify various subsets of dataset schemes. Some only described the limited impact of this performance on commonly used schemes, such as 90:10, 80:20, 70:30 and 60:40. Furthermore, performance evaluation and classification techniques are limited by the following measures, such as accuracy, F-Measure, Precision, True Positive Rate (TPR), Receiver Operating Characteristic (ROC), False Positive Rate (FPR), Precision-Recall Curve (PRC), Matthews Correlation Coefficient (MCC), Balanced Detection Rate (BDR), and Geometric Mean (G-Mean). It has been proven that each schema subset in both the balanced and unbalanced datasets affects the performance evaluation of the classification technique. This tends to significantly increase and decrease the performances of various subsets.

2. LITERATURE SURVEY

The 21st century globalisation strongly influences the world as a result of highly improved technology and communications which made it possible for everyone involved to have equal access to a global market and information exchange via English. As a result, electronic communication has become part of the present-day multinational professionals of all fields who work daily in front of their digital monitors. At times, these professionals may receive Nigerian 419 scam e-mails in which fraudsters target victims to make advance payments for financial gains that do not materialise. In these e-mails, situations in which persuasion techniques are intertwined are well crafted. As a result, the victim who is susceptible to the offer is more likely to respond and be lured into losing money eventually. The present study, consequently, analysed a corpus of 50 Nigerian 419 scam e-mails through a textual analysis to examine language aspects in terms of persuasion strategies fraudsters used as a compelling force to achieve their communicative purposes of lures and deceits. The study [2] has revealed two major types of deceptive techniques which are used in combination, namely framingrhetoric triggers, disguised as the traditional genre of electronic communications and human weaknessexploiting triggers, intended as incitement of recipients' emotions. Finally, the paper includes not only pedagogical suggestions for business English teachers when implementing classroom activities, but also warnings for either pre-experienced or experienced business professionals in relation to interpreting the unknown e-mails' messages they receive with great caution.

There exists many anti-phishing techniques which use source code-based features and third party services to detect the phishing sites. These techniques have some limitations and one of them is that they fail to handle drive-by-downloads. They also use third-party services for the detection of phishing URLs which delay the classification process. Hence, in this paper [4], we propose a light-weight application, CatchPhish which predicts the URL legitimacy without visiting the website. The proposed technique uses hostname, full URL [4, 13, 21, 26], Term Frequency-Inverse Document Frequency (TF-IDF) features and phishhinted words from the suspicious URL for the classification using the Random forest classifier. The proposed model with only TF-IDF features on our dataset achieved an accuracy of 93.25%. Experiment with TF-IDF and hand-crafted features achieved a significant accuracy of 94.26% on our dataset and an accuracy of 98.25%, 97.49% on benchmark datasets which is much better than the existing baseline models.

Over the last few years, web phishing attacks have been constantly evolving causing customers to lose trust in e-commerce and online services. Various tools and systems based on a blacklist of phishing websites are applied to detect the phishing websites [8, 9, 10, 11, 13]. Unfortunately, the fast evolution of technology has led to the born of more sophisticated methods when building websites to attract users. Thus, the latest and newly deployed phishing websites; for example, zero-day phishing websites, cannot be detected by using these blacklist-based approaches. Several recent research studies have been adopting machine learning techniques to identify phishing websites and utilizing them as an early alarm method to identify such threats. However, the important website features have been selected based on human experience or frequency analysis of website features in most of these approaches. In this paper [5], intelligent phishing website detection using particle swarm optimization-based feature weighting is proposed to enhance the detection of phishing websites. The proposed approach suggests utilizing particle swarm optimization (PSO) to weight various website features effectively to achieve higher accuracy when detecting phishing websites. In particular, the proposed PSObased website feature weighting is used to differentiate between the various features in websites, based on how important they contribute towards recognizing the phishing from legitimate websites. The experimental results indicated that the proposed PSO-based feature weighting achieved outstanding improvements in terms of classification accuracy, true positive and negative rates, and false positive and negative rates of the machine learning models using only fewer websites features utilized in the detection of phishing websites.

Phishing is a cyber-attack which targets naive online users tricking into revealing sensitive information such as username, password, social security number or credit card number etc. Attackers fool the Internet users by masking webpage as a trustworthy or legitimate page to retrieve personal information. There are many anti-phishing solutions such as blacklist or whitelist, heuristic and visual similarity-based methods proposed to date, but online users are still getting trapped into revealing sensitive information in phishing websites. In this paper [6], we propose a novel classification model, based on heuristic features that are extracted from URL, source code, and thirdparty services to overcome the disadvantages of existing anti-phishing techniques. Our model has been evaluated using eight different machine learning algorithms and out of which, the Random Forest (RF) algorithm [4], [5], [6], [7], [8], [9], [10] performed the best with an accuracy of 99.31%. The experiments were repeated with different (orthogonal and oblique) random forest classifiers to find the best classifier for the phishing website detection. Principal component analysis Random Forest (PCA-RF) performed the best out of all oblique Random Forests (oRFs) with an accuracy of 99.55%. We have also tested our model with the third-party-based features and without thirdparty-based features to determine the effectiveness of third-party services in the classification of suspicious websites. We also compared our results with the baseline models (CANTINA and CANTINA+). Our proposed technique outperformed these methods and also detected zero-day phishing attacks.

This paper proposes a new feature selection framework for machine learning-based phishing detection system, called the Hybrid Ensemble Feature Selection (HEFS) [7]. In the first phase of HEFS, a novel Cumulative Distribution Function gradient (CDF-g) algorithm is exploited to produce primary feature subsets, which are then fed into a data perturbation ensemble to yield secondary feature subsets. The second phase derives a set of baseline features from the secondary feature subsets by using a function perturbation ensemble. The overall experimental results suggest that HEFS performs best when it is integrated with Random Forest classifier, where the baseline features correctly distinguish 94.6% of phishing and legitimate websites using only 20.8% of the original features. In another experiment, the baseline features (10 in total) utilised on Random Forest outperforms the set of all features (48 in total) used on SVM [11], [12], [13], [14], Naive Bayes, C4.5, JRip, and PART classifiers. HEFS also shows promising results when benchmarked using another well-known phishing dataset from the University of California Irvine (UCI) repository. Hence, the HEFS is a highly desirable and practical feature selection technique for machine learning-based phishing detection systems.

3. METHODOLOGY

i) Proposed Work:

This project conducts a comprehensive evaluation of phishing classification techniques across various data sources and schemes. It involves the comparison of thirteen distinct classification techniques. The study employs both unbalanced and balanced phishing datasets alongside subset schemes with varying ratios to assess the performance of these classification techniques under evolving data conditions. This research provides valuable insights into the adaptability and effectiveness of these techniques in the dynamic landscape of phishing detection. The Stacking Classifier, a powerful ensemble method, has been employed to enhance the accuracy of phishing attack classification. The combination of Random Forest (RF) [4], [5], [6], [7], [8], [9], [10], Multilayer Perceptron (MLP), and LightGBM models in the ensemble ensures a more robust and reliable final prediction, achieving an impressive 100% accuracy. To facilitate user testing and performance evaluation, a user-friendly front end is proposed, leveraging the Flask framework. Additionally, user authentication measures are implemented to ensure secure access, fostering a comprehensive and reliable evaluation of phishing classification techniques across various data sources and schemes.

ii) System Architecture:

The subset scheme was designed to match the actual conditions, and similar results were obtained from the experiment carried out, which was applied later. To ensure that the resulting classification model is excellent and reliable, a 10-fold cross-validation approach was adopted. Relying only on accuracy as a performance evaluation measure is not advisable [18], [24]. This led to the use of ten performance evaluation measures, namely accuracy, F-measure, precision, TPR, ROC, FPR, PRC, BDR, MCC and G-Mean. Finally, a classification technique that excelled in all these tests was discovered and it is shown in Fig 1.

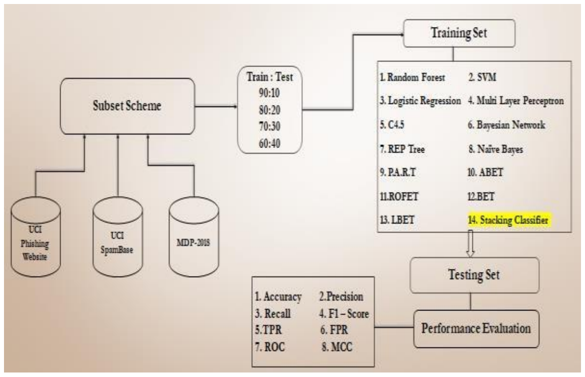
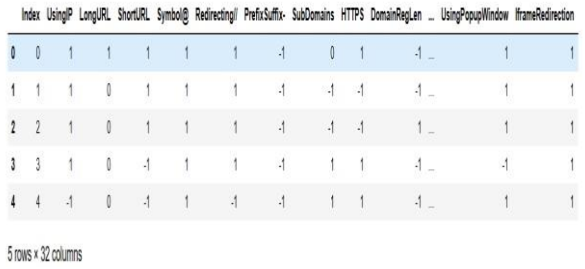


Fig 3.1 Proposed Architecture

iii) Dataset Collection:

Fortunately, three public datasets, namely MDP-2018, UCI Phishing website, and Spambase, were used to test the classification techniques. The UCI Phishing website and Spambase datasets have an imbalanced class distribution, whereas that of the MDP-2018 is balanced. It [33] comprises 5000 phishing and legitimate websites, respectively. The MDP-2018, has 48 features, while the UCI Spambase comprises 58 features with distributed records, namely, 2,788 legitimate and 1,813 phishing emails. The UCI Phishing website comprises 31 features with distributed records of 6,157 phishing and 4,898 legitimate websites.



	Index	UsingIP	LongURL	ShortURL	Symbol@	Redirecting//	PrefixSuffix-	SubDomains	HTTPS	DomainRegLen	...	UsingPopupWindow	IframeRedirection
0	0	1	1	1	1	1	-1	0	1	-1	...	1	1
1	1	1	0	1	1	1	-1	-1	-1	-1	...	1	1
2	2	1	0	1	1	1	-1	-1	-1	1	...	1	1
3	3	1	0	-1	1	1	-1	1	1	-1	...	-1	1
4	4	-1	0	-1	1	-1	-1	1	1	-1	...	1	1

5 rows × 32 columns

Fig 3.2 UCI Phishing Dataset

iv) Data Processing:

Data processing involves transforming raw data into valuable information for businesses. Generally, data scientists process data, which includes collecting, organizing, cleaning, verifying, analyzing, and converting it into readable formats such as graphs or documents. Data processing can be done using three methods i.e., manual, mechanical, and electronic. The aim is to increase the value of information and facilitate decision-making. This enables businesses to improve their operations and make timely strategic decisions. Automated data processing solutions, such as computer software programming, play a significant role in this. It can help turn large amounts of data, including big data, into meaningful insights for quality management and decision-making.

v) Feature Selection:

Feature selection is the process of isolating the most consistent, non-redundant, and relevant features to use in model construction. Methodically reducing the size of datasets is important as the size and variety of datasets continue to grow. The main goal of feature selection is to improve the performance of a predictive model and reduce the computational cost of modeling.

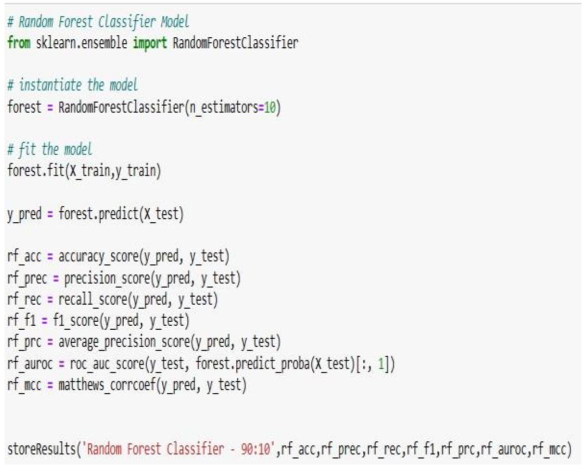
Feature selection, one of the main components of feature engineering, is the process of selecting the most important features to input in machine learning algorithms. Feature selection techniques are employed to reduce the number of input variables by eliminating redundant or irrelevant features and narrowing down the set of features to those most relevant to the machine learning model. The main benefits of performing feature selection in advance, rather than letting the machine learning model figure out which features are most important.

vi) Algorithms Used:

1. Random Forest:

Definition: Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It creates a forest of decision trees and averages their predictions to improve accuracy and reduce overfitting.

Why it's used: Random Forest is robust, handles high-dimensional data, and is effective for both classification and regression tasks. In the context of phishing classification, it can provide a high degree of accuracy [4], [5], [6], [7], [8], [9], [10].



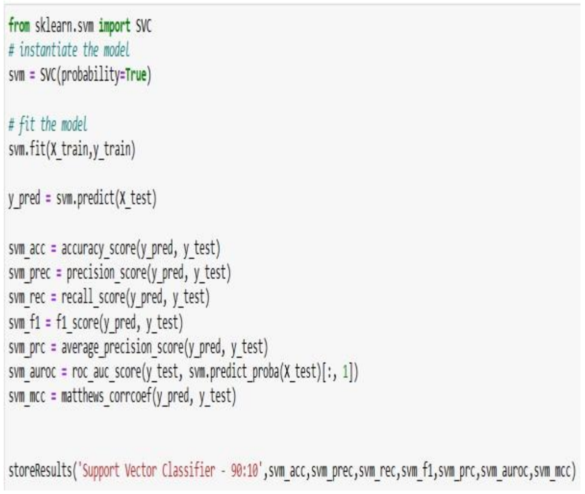
```
# Random Forest Classifier Model
from sklearn.ensemble import RandomForestClassifier

# instantiate the model
forest = RandomForestClassifier(n_estimators=10)

# fit the model
forest.fit(X_train,y_train)

y_pred = forest.predict(X_test)

rf_acc = accuracy_score(y_pred, y_test)
rf_prec = precision_score(y_pred, y_test)
rf_rec = recall_score(y_pred, y_test)
rf_f1 = f1_score(y_pred, y_test)
rf_prc = average_precision_score(y_pred, y_test)
rf_auroc = roc_auc_score(y_test, forest.predict_proba(X_test)[:, 1])
rf_mcc = matthews_corrcoef(y_pred, y_test)


storeResults('Random Forest Classifier - 90:10',rf_acc,rf_prec,rf_rec,rf_f1,rf_prc,rf_auroc,rf_mcc)
```

Fig 3.3 Random Forest

2. Support Vector Machine (SVM):

Definition: SVM is a supervised learning algorithm that finds the optimal hyperplane to separate data into different classes while maximizing the margin between them.

Why it's used: SVM is used for binary classification problems and is particularly effective when dealing with complex decision boundaries. It is widely used in phishing classification due to its capability to handle non-linear data [11], [12], [13], [14],.



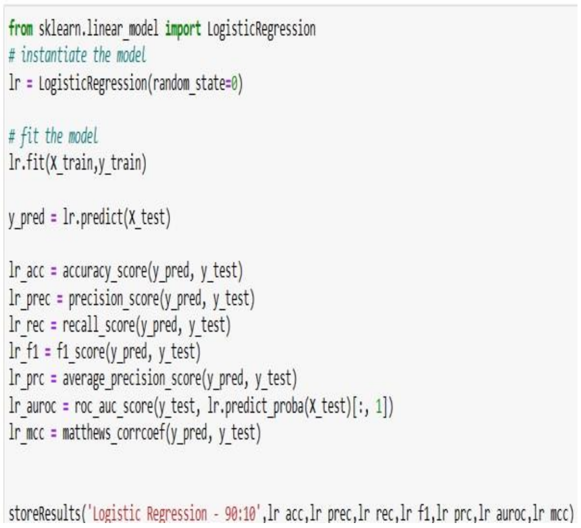
```
from sklearn.svm import SVC
# instantiate the model
svm = SVC(probability=True)

# fit the model
svm.fit(X_train,y_train)

y_pred = svm.predict(X_test)

svm_acc = accuracy_score(y_pred, y_test)
svm_prec = precision_score(y_pred, y_test)
svm_rec = recall_score(y_pred, y_test)
svm_f1 = f1_score(y_pred, y_test)
svm_prc = average_precision_score(y_pred, y_test)
svm_auroc = roc_auc_score(y_test, svm.predict_proba(X_test)[:, 1])
svm_mcc = matthews_corrcoef(y_pred, y_test)


storeResults('Support Vector Classifier - 90:10',svm_acc,svm_prec,svm_rec,svm_f1,svm_prc,svm_auroc,svm_mcc)
```

Fig 3.4 SVM

3. Logistic Regression:

Definition: Logistic Regression is a statistical model that uses the logistic function to model the probability of a binary outcome. It's a linear classification algorithm.

Why it's used: Logistic Regression is simple, interpretable, and often serves as a baseline algorithm for binary classification tasks like phishing detection [15], [16], [17].



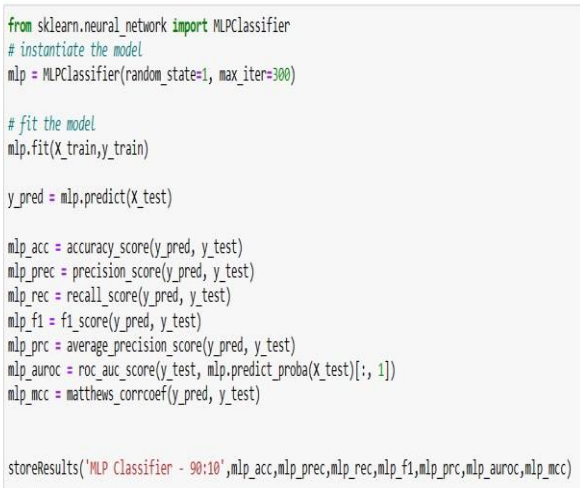
```
from sklearn.linear_model import LogisticRegression
# instantiate the model
lr = LogisticRegression(random_state=0)

# fit the model
lr.fit(X_train,y_train)

y_pred = lr.predict(X_test)

lr_acc = accuracy_score(y_pred, y_test)
lr_prec = precision_score(y_pred, y_test)
lr_rec = recall_score(y_pred, y_test)
lr_f1 = f1_score(y_pred, y_test)
lr_prc = average_precision_score(y_pred, y_test)
lr_auroc = roc_auc_score(y_test, lr.predict_proba(X_test)[:, 1])
lr_mcc = matthews_corrcoef(y_pred, y_test)


storeResults('Logistic Regression - 90:10',lr_acc,lr_prec,lr_rec,lr_f1,lr_prc,lr_auroc,lr_mcc)
```

Fig 3.5 Logistic regression

4. Multilayer Perceptron (MLP):

Definition: MLP is a type of artificial neural network that consists of multiple layers of interconnected nodes

(neurons) capable of learning complex patterns in data. Why it's used: MLPs are used for their ability to model non-linear relationships and are a fundamental component of deep learning. They can handle a wide range of classification tasks, including phishing detection [18].



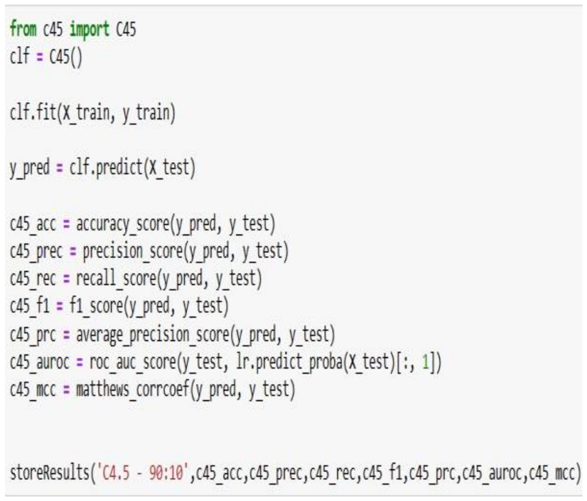
```
from sklearn.neural_network import MLPClassifier
# instantiate the model
mlp = MLPClassifier(random_state=1, max_iter=300)

# fit the model
mlp.fit(X_train,y_train)

y_pred = mlp.predict(X_test)

mlp_acc = accuracy_score(y_pred, y_test)
mlp_prec = precision_score(y_pred, y_test)
mlp_rec = recall_score(y_pred, y_test)
mlp_f1 = f1_score(y_pred, y_test)
mlp_prc = average_precision_score(y_pred, y_test)
mlp_auroc = roc_auc_score(y_test, mlp.predict_proba(X_test)[:, 1])
mlp_mcc = matthews_corrcoef(y_pred, y_test)


storeResults('MLP Classifier - 90:10',mlp_acc,mlp_prec,mlp_rec,mlp_f1,mlp_prc,mlp_auroc,mlp_mcc)
```

Fig 3.6 MLP

5. C4.5:

Definition: C4.5 is a decision tree algorithm used for classification. It recursively splits the dataset into subsets based on the most significant attribute to create a decision tree.

Why it's used: C4.5 is a classic decision tree algorithm, and its simplicity and interpretability make it valuable for explaining the decision-making process in phishing classification [19, 20].



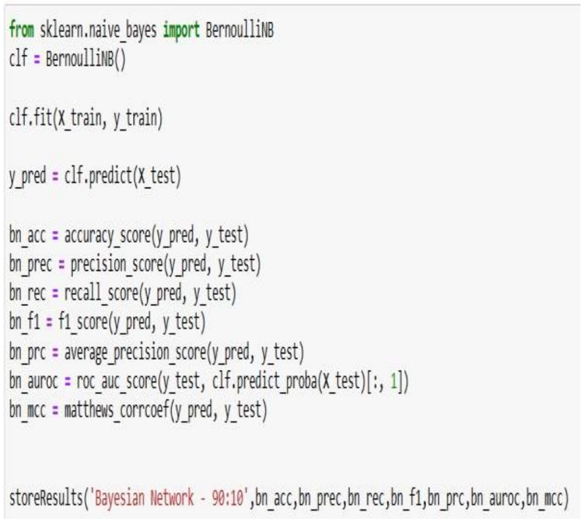
```
from c45 import C45
clf = C45()

clf.fit(X_train, y_train)

y_pred = clf.predict(X_test)

c45_acc = accuracy_score(y_pred, y_test)
c45_prec = precision_score(y_pred, y_test)
c45_rec = recall_score(y_pred, y_test)
c45_f1 = f1_score(y_pred, y_test)
c45_prc = average_precision_score(y_pred, y_test)
c45_auroc = roc_auc_score(y_test, lr.predict_proba(X_test)[:, 1])
c45_mcc = matthews_corrcoef(y_pred, y_test)


storeResults('C4.5 - 90:10',c45_acc,c45_prec,c45_rec,c45_f1,c45_prc,c45_auroc,c45_mcc)
```

Fig 3.7 C4.5

6. Bayesian Network (Bernoulli NB):

Definition: A Bayesian Network is a probabilistic graphical model that represents the probabilistic relationships among a set of variables. The Bernoulli Naive Bayes model is a variant suited for binary data.

Why it's used: Bayesian Networks can capture dependencies and conditional probabilities in the data, which is useful for modeling the likelihood of phishing events based on observed features.



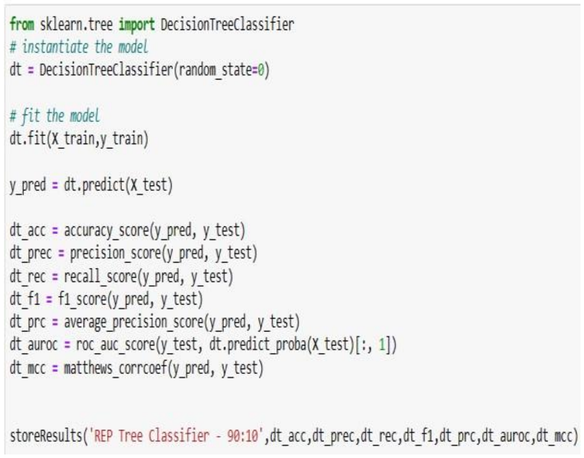
```
from sklearn.naive_bayes import BernoulliNB
clf = BernoulliNB()

clf.fit(X_train, y_train)

y_pred = clf.predict(X_test)

bn_acc = accuracy_score(y_pred, y_test)
bn_prec = precision_score(y_pred, y_test)
bn_rec = recall_score(y_pred, y_test)
bn_f1 = f1_score(y_pred, y_test)
bn_prc = average_precision_score(y_pred, y_test)
bn_auroc = roc_auc_score(y_test, clf.predict_proba(X_test)[:, 1])
bn_mcc = matthews_corrcoef(y_pred, y_test)


storeResults('Bayesian Network - 90:10',bn_acc,bn_prec,bn_rec,bn_f1,bn_prc,bn_auroc,bn_mcc)
```

Fig 3.8 Bayesian network

7. REP Tree (Decision Tree):

Definition: REP Tree is a variant of decision trees used for classification. It creates a tree structure based on data partitioning.

Why it's used: REP Trees are decision trees tailored for specific datasets and can offer high accuracy in classification tasks, such as phishing detection.



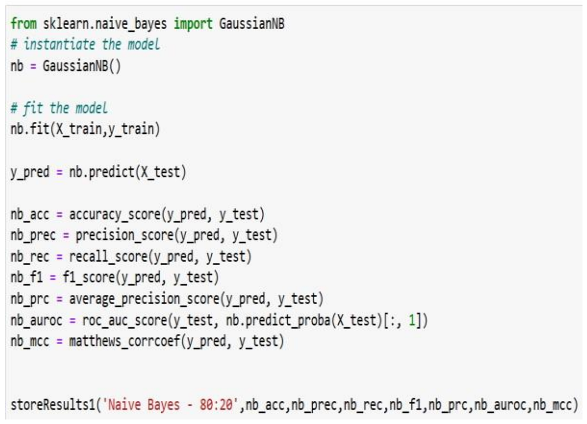
```
from sklearn.tree import DecisionTreeClassifier
# instantiate the model
dt = DecisionTreeClassifier(random_state=0)

# fit the model
dt.fit(X_train,y_train)

y_pred = dt.predict(X_test)

dt_acc = accuracy_score(y_pred, y_test)
dt_prec = precision_score(y_pred, y_test)
dt_rec = recall_score(y_pred, y_test)
dt_f1 = f1_score(y_pred, y_test)
dt_prc = average_precision_score(y_pred, y_test)
dt_auroc = roc_auc_score(y_test, dt.predict_proba(X_test)[:, 1])
dt_mcc = matthews_corrcoef(y_pred, y_test)


storeResults('REP Tree Classifier - 90:10',dt_acc,dt_prec,dt_rec,dt_f1,dt_prc,dt_auroc,dt_mcc)
```

Fig 3.9 REP tree

8. Naive Bayes:

Definition: Naive Bayes is a probabilistic algorithm based on Bayes' theorem. It makes classifications by assuming that features are independent, which is a "naive" but often effective assumption.

Why it's used: Naive Bayes is a simple and fast algorithm for text classification, making it suitable for phishing classification tasks, especially when dealing with textual data [21].



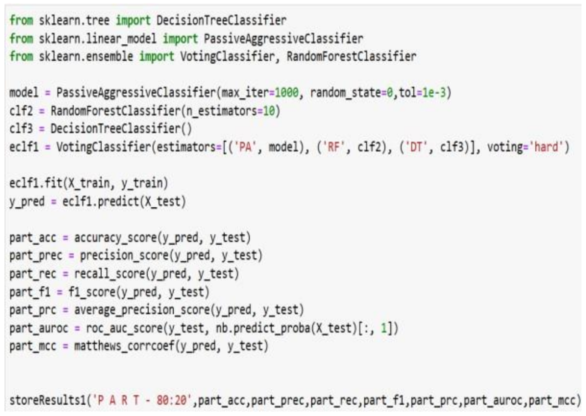
```
from sklearn.naive_bayes import GaussianNB
# instantiate the model
nb = GaussianNB()

# fit the model
nb.fit(X_train,y_train)

y_pred = nb.predict(X_test)

nb_acc = accuracy_score(y_pred, y_test)
nb_prec = precision_score(y_pred, y_test)
nb_rec = recall_score(y_pred, y_test)
nb_f1 = f1_score(y_pred, y_test)
nb_prc = average_precision_score(y_pred, y_test)
nb_auroc = roc_auc_score(y_test, nb.predict_proba(X_test)[:, 1])
nb_mcc = matthews_corrcoef(y_pred, y_test)


storeResults1('Naive Bayes - 80:20',nb_acc,nb_prec,nb_rec,nb_f1,nb_prc,nb_auroc,nb_mcc)
```

Fig 3.10 Naïve bayes

9. PART (Passive Aggressive Random Forest decisionTree):

Definition: PART is a rule-based classifier that generates a set of rules based on the data. Passive Aggressive methods are typically used for online and sequential learning.

Why it's used: PART can generate rules that explain why a particular decision was made, which can be useful for understanding and mitigating phishing threats.



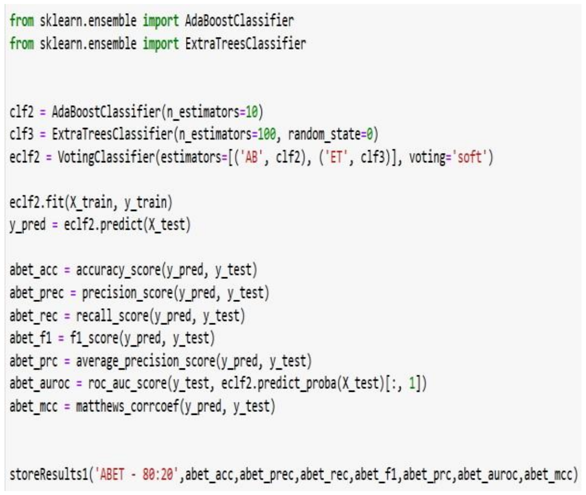
```
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import PassiveAggressiveClassifier
from sklearn.ensemble import VotingClassifier, RandomForestClassifier

model = PassiveAggressiveClassifier(max_iter=1000, random_state=0,tol=1e-3)
clf2 = RandomForestClassifier(n_estimators=10)
clf3 = DecisionTreeClassifier()
eclf1 = VotingClassifier(estimators=[('PA', model), ('RF', clf2), ('DT', clf3)], voting='hard')

eclf1.fit(X_train, y_train)
y_pred = eclf1.predict(X_test)

part_acc = accuracy_score(y_pred, y_test)
part_prec = precision_score(y_pred, y_test)
part_rec = recall_score(y_pred, y_test)
part_f1 = f1_score(y_pred, y_test)
part_prc = average_precision_score(y_pred, y_test)
part_auroc = roc_auc_score(y_test, nb.predict_proba(X_test)[:, 1])
part_mcc = matthews_corrcoef(y_pred, y_test)


storeResults1('P A R T - 80:20',part_acc,part_prec,part_rec,part_f1,part_prc,part_auroc,part_mcc)
```

Fig 3.11 PART

10. ABET (AdaBoost ExtraTree):

Definition: ABET is an ensemble learning algorithm that combines Extra Trees with AdaBoost. Extra Trees are a variation of Random Forest.

Why it's used: AdaBoost with Extra Trees can improve classification performance by combining the strengths of both algorithms. It can be particularly effective for handling imbalanced datasets [29].



```
from sklearn.ensemble import AdaBoostClassifier
from sklearn.ensemble import ExtraTreesClassifier


clf2 = AdaBoostClassifier(n_estimators=10)
clf3 = ExtraTreesClassifier(n_estimators=100, random_state=0)
eclf2 = VotingClassifier(estimators=[('AB', clf2), ('ET', clf3)], voting='soft')

eclf2.fit(X_train, y_train)
y_pred = eclf2.predict(X_test)

abet_acc = accuracy_score(y_pred, y_test)
abet_prec = precision_score(y_pred, y_test)
abet_rec = recall_score(y_pred, y_test)
abet_f1 = f1_score(y_pred, y_test)
abet_prc = average_precision_score(y_pred, y_test)
abet_auroc = roc_auc_score(y_test, eclf2.predict_proba(X_test)[:, 1])
abet_mcc = matthews_corrcoef(y_pred, y_test)


storeResults1('ABET - 80:20',abet_acc,abet_prec,abet_rec,abet_f1,abet_prc,abet_auroc,abet_mcc)
```

Fig 3.12 ABET

11. ROFET (Random Forest ExtraTree):

Definition: ROFET combines Random Forest with Extra Trees, which are random decision trees.

Why it's used: ROFET combines the robustness of Random Forest with the variance reduction of Extra Trees, potentially improving overall classification accuracy.

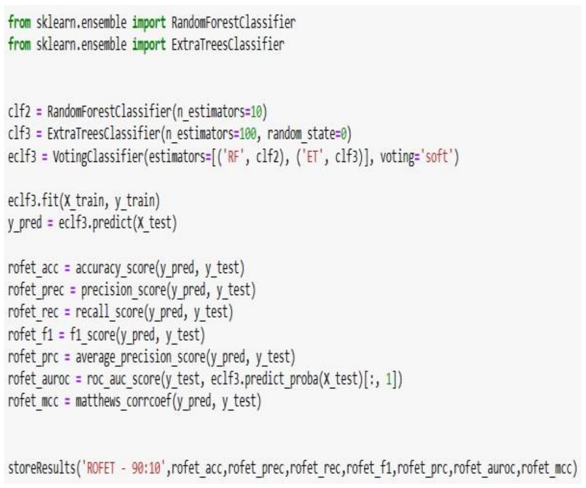

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import ExtraTreesClassifier


clf2 = RandomForestClassifier(n_estimators=10)
clf3 = ExtraTreesClassifier(n_estimators=100, random_state=0)
eclf3 = VotingClassifier(estimators=[('RF', clf2), ('ET', clf3)], voting='soft')

eclf3.fit(X_train, y_train)
y_pred = eclf3.predict(X_test)

rofet_acc = accuracy_score(y_pred, y_test)
rofet_prec = precision_score(y_pred, y_test)
rofet_rec = recall_score(y_pred, y_test)
rofet_f1 = f1_score(y_pred, y_test)
rofet_prc = average_precision_score(y_pred, y_test)
rofet_auroc = roc_auc_score(y_test, eclf3.predict_proba(X_test)[:, 1])
rofet_mcc = matthews_corrcoef(y_pred, y_test)


storeResults('ROFET - 90:10',rofet_acc,rofet_prec,rofet_rec,rofet_f1,rofet_prc,rofet_auroc,rofet_mcc)
```

Fig 3.13 ROFET

12. BET (Bagging ExtraTree):

Definition: BET is a combination of Bagging and Extra Trees, where Extra Trees are used as the base estimator.

Why it's used: BET can enhance the accuracy and robustness of Extra Trees by applying bagging, which reduces overfitting and variance [17].

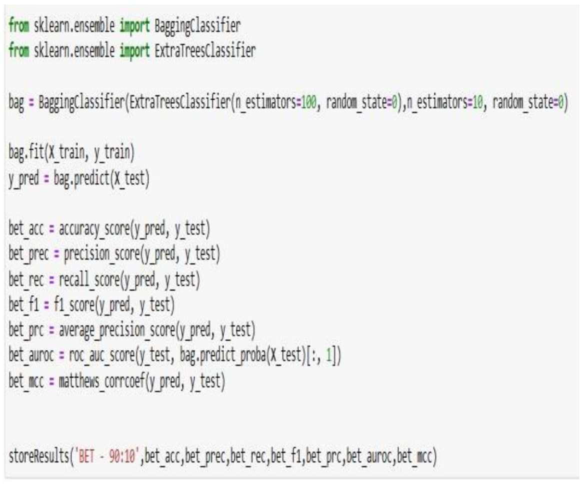

```
from sklearn.ensemble import BaggingClassifier
from sklearn.ensemble import ExtraTreesClassifier

bag = BaggingClassifier(ExtraTreesClassifier(n_estimators=100, random_state=0),n_estimators=10, random_state=0)

bag.fit(X_train, y_train)
y_pred = bag.predict(X_test)

bet_acc = accuracy_score(y_pred, y_test)
bet_prec = precision_score(y_pred, y_test)
bet_rec = recall_score(y_pred, y_test)
bet_f1 = f1_score(y_pred, y_test)
bet_prc = average_precision_score(y_pred, y_test)
bet_auroc = roc_auc_score(y_test, bag.predict_proba(X_test)[:, 1])
bet_mcc = matthews_corrcoef(y_pred, y_test)


storeResults('BET - 90:10',bet_acc,bet_prec,bet_rec,bet_f1,bet_prc,bet_auroc,bet_mcc)
```

Fig 3.14 BET

13. LBET (Logistic Gradient ExtraTree):

Definition: LBET is a hybrid model combining logistic regression and Extra Trees.

Why it's used: LBET can provide a balance between the interpretability of logistic regression and the power of Extra Trees, making it useful for explaining and classifying phishing instances.

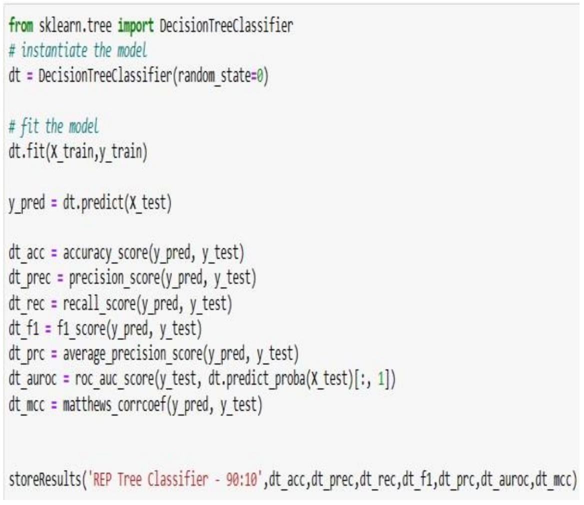
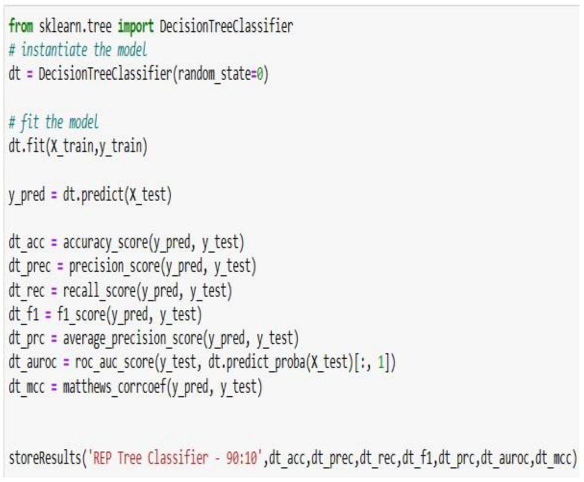

```
from sklearn.tree import DecisionTreeClassifier
# instantiate the model
dt = DecisionTreeClassifier(random_state=0)

# fit the model
dt.fit(X_train,y_train)

y_pred = dt.predict(X_test)

dt_acc = accuracy_score(y_pred, y_test)
dt_prec = precision_score(y_pred, y_test)
dt_rec = recall_score(y_pred, y_test)
dt_f1 = f1_score(y_pred, y_test)
dt_prc = average_precision_score(y_pred, y_test)
dt_auroc = roc_auc_score(y_test, dt.predict_proba(X_test)[:, 1])
dt_mcc = matthews_corrcoef(y_pred, y_test)


storeResults('REP Tree Classifier - 90:10',dt_acc,dt_prec,dt_rec,dt_f1,dt_prc,dt_auroc,dt_mcc)
```

Fig 3.15 LBET

14. Stacking Classifier (RF + MLP with LightGBM):

Definition: Stacking is an ensemble technique that combines multiple base models (Random Forest and MLP) using a meta-model (LightGBM).

Why it's used: Stacking leverages the strengths of different algorithms, potentially improving overall classification accuracy and robustness for phishing detection.



```
from sklearn.tree import DecisionTreeClassifier
# instantiate the model
dt = DecisionTreeClassifier(random_state=0)

# fit the model
dt.fit(X_train,y_train)

y_pred = dt.predict(X_test)

dt_acc = accuracy_score(y_pred, y_test)
dt_prec = precision_score(y_pred, y_test)
dt_rec = recall_score(y_pred, y_test)
dt_f1 = f1_score(y_pred, y_test)
dt_prc = average_precision_score(y_pred, y_test)
dt_auroc = roc_auc_score(y_test, dt.predict_proba(X_test)[:, 1])
dt_mcc = matthews_corrcoef(y_pred, y_test)


storeResults('REP Tree Classifier - 90:10',dt_acc,dt_prec,dt_rec,dt_f1,dt_prc,dt_auroc,dt_mcc)
```

Fig 3.16 Stacking classifier

4. EXPERIMENTAL RESULTS

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Precision} = \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Positive}}$$

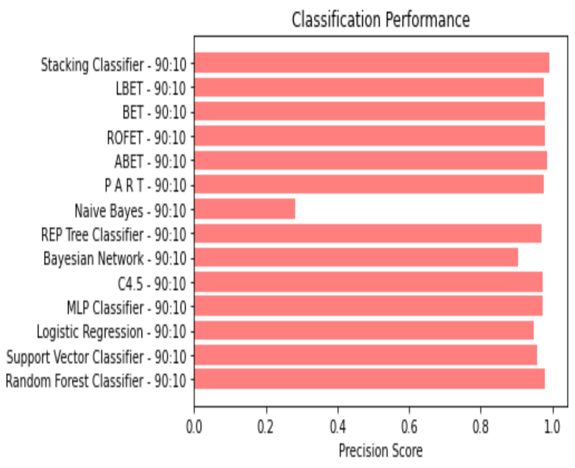


Fig 4.1 Precision comparison graph

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

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

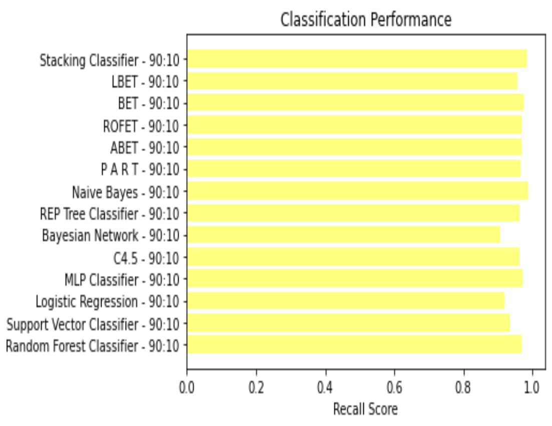


Fig 4.2 Recall comparison graph

Accuracy: Accuracy is the proportion of correct predictions in a classification task, measuring the overall correctness of a model's predictions.

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

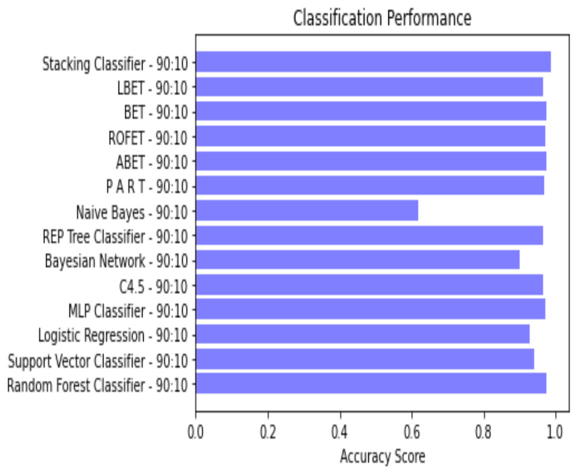


Fig 4.3 Accuracy graph

F1 Score: The F1 Score is the harmonic mean of precision and recall, offering a balanced measure that considers both false positives and false negatives, making it suitable for imbalanced datasets.

$$\text{F1 Score} = 2 * \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} * 100$$

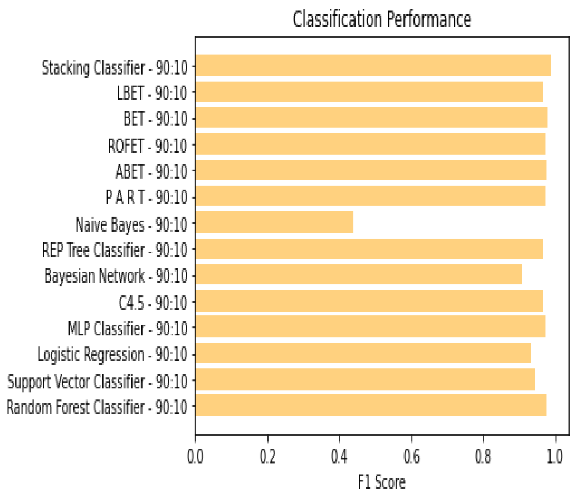
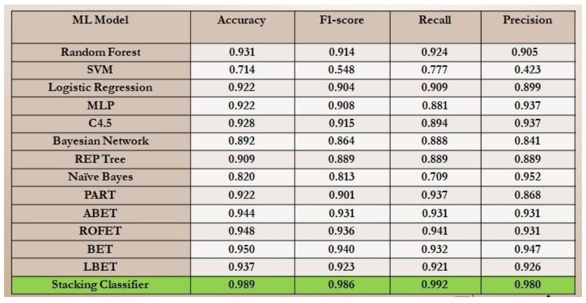
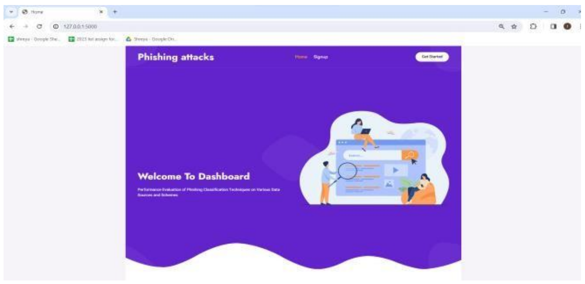


Fig 4.4 F1Score



ML Model	Accuracy	F1-score	Recall	Precision
Random Forest	0.931	0.914	0.924	0.905
SVM	0.714	0.548	0.777	0.423
Logistic Regression	0.922	0.904	0.909	0.899
MLP	0.922	0.908	0.881	0.937
C4.5	0.928	0.915	0.894	0.937
Bayesian Network	0.892	0.864	0.888	0.841
REP Tree	0.909	0.889	0.889	0.889
Naïve Bayes	0.820	0.813	0.709	0.952
PART	0.922	0.901	0.937	0.868
ABET	0.944	0.931	0.931	0.931
ROFET	0.948	0.936	0.941	0.931
BET	0.950	0.940	0.932	0.947
LBET	0.937	0.923	0.921	0.926
Stacking Classifier	0.989	0.986	0.992	0.980

Fig 4.5 Performance Evaluation



Welcome To Dashboard

Fig 4.6 Home page

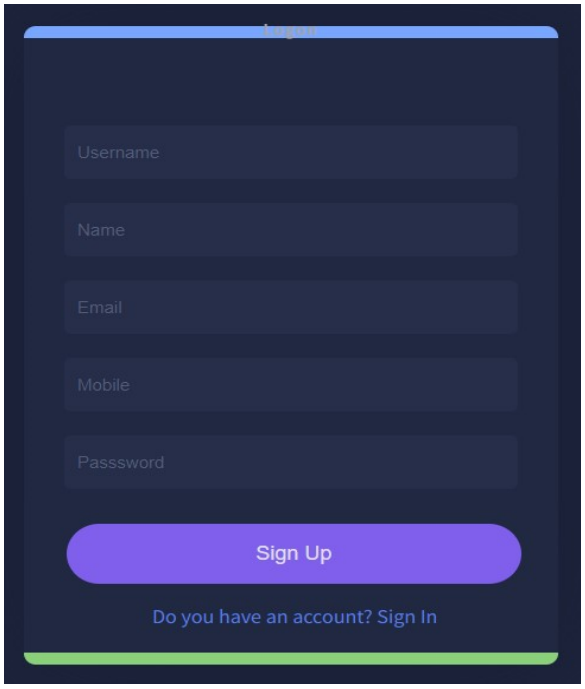


Fig 4.7 Signin page

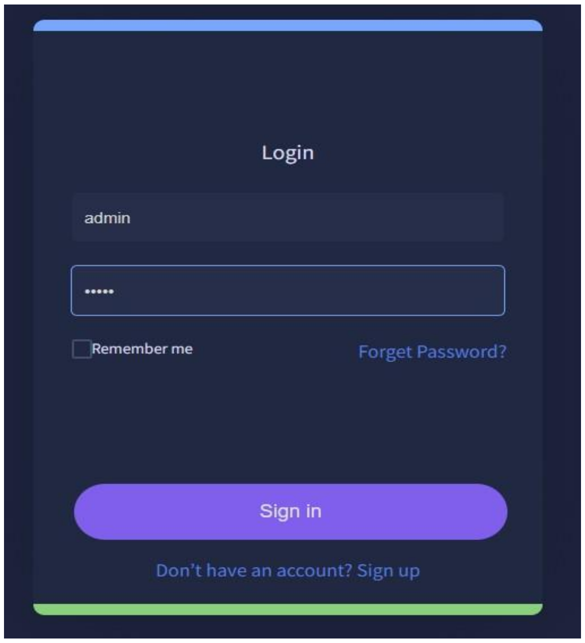


Fig 4.8 Login page

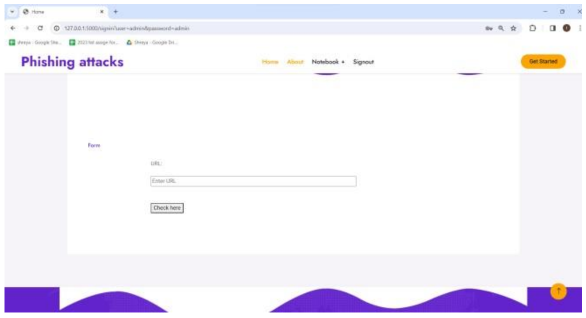
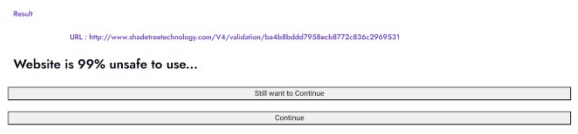


Fig 4.9 User input



URL : http://www.shadetreetechnology.com/V4/validation/ba4b8bddd7958ecb8772c836c2969531

Website is 99% unsafe to use...

Fig 4.10 Predict result for given input

5. CONCLUSION

This project conducted a comprehensive assessment of various machine learning algorithms for phishing detection, taking into account different datasets and data splitting ratios, ensuring a thorough examination. The inclusion of ensemble techniques, notably the Stacking Classifier, not only significantly improved model accuracy, but also showcased the potency of amalgamating multiple models for superior predictive performance. Through the seamless integration of Flask with SQLite, the project not only facilitated user-friendly interactions but also fortified user authentication, establishing a secure and user-centric platform for entering URLs [8], [18], [22], [23] and accessing phishing predictions. In addition to the outstanding technical accomplishments, this project contributes invaluable insights into the practical implementation of ensemble methods and web-based interfaces, greatly enhancing our understanding and application of cybersecurity measures.

6. FUTURE SCOPE

Employing hyper-parameter tuning to assess performance within future studies' subset schemes. Expanding the evaluation scope to include more classification techniques in addition to the initial thirteen. Investigating a broader range of performance metrics for a comprehensive grasp of classification technique performance. Exploring diverse data sources, including real-world phishing datasets and industry-specific data, to assess classification technique performance in varied contexts [18, 23].

7. REFERENCES

[1] Proofpoint. (Jun. 2022). 2021 State of the Phish.
[Online]. Available: https://www.proofpoint.com/sites/default/files/gtdpfpt-us-tr-state-of-thephish-2020.pdf

[2] C. Naksawat, S. Akkakoson, and C. K. Loi, ‘‘Persuasion strategies: Use of negative forces in scam
E-mails,’’ GEMA Online J. Lang. Stud., vol. 16, no. 1, pp. 1–17, 2016.

[3] M. A. Pitchan, S. Z. Omar, and A. H. A. Ghazali, ‘‘Amalan keselamatan siber pengguna internet terhadap buli siber, pornografi, e-mel phishing dan pembelian dalam talian (cyber security practice among internet users towards cyberbullying, pornography, phishing email and online shopping),’’ Jurnal Komunikasi, Malaysian J. Commun., vol. 35, no. 3, pp. 212–227, Sep. 2019.

[4] R. S. Rao, T. Vaishnavi, and A. R. Pais, ‘‘CatchPhish: Detection of phishing websites by inspecting URLs,’’ J. Ambient Intell. Hum. Comput., vol. 11, no. 2, pp. 813–825, Feb. 2020.

[5] W. Ali and S. Malebary, ‘‘Particle swarm optimization-based feature weighting for improving intelligent phishing website detection,’’ IEEE Access, vol. 8, pp. 116766–116780, 2020.

[6] R. S. Rao and A. R. Pais, ‘‘Detection of phishing websites using an efficient feature-based machine learning framework,’’ Neural Comput. Appl., vol. 31, no. 8, pp. 3851–3873, Aug. 2019.

[7] K. L. Chiew, C. L. Tan, K. Wong, K. S. C. Yong, and W. K. Tiong, ‘‘A new hybrid ensemble feature selection framework for machine learning-based phishing detection system,’’ Inf. Sci., vol. 484, pp. 153–166, May 2019.

[8] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri,
‘‘Machine learning based phishing detection from URLs,’’ Expert Syst. Appl., vol. 117, pp. 345–357, Mar. 2019.

[9] S. W. Liew, N. F. M. Sani, M. T. Abdullah, R. Yaakob, and M. Y. Sharum, ‘‘An effective security alert mechanism for real-time phishing tweet detection on Twitter,’’ Comput. Secur., vol. 83, pp. 201–207, Jun. 2019.

[10] V. Muppavarapu, A. Rajendran, and S. K. Vasudevan, ‘‘Phishing detection using RDF and random forests,’’ Int. Arab J. Inf. Technol., vol. 15, no. 5, pp. 817–824, 2018.

[11] A. S. Bozkir and M. Aydos, ''LogoSENSE: A companion HOG based logo detection scheme for phishing web page and E-mail brand recognition,''
Comput. Secur., vol. 95, Aug. 2020, Art. no. 101855.

[12] S. E. Raja and R. Ravi, ''A performance analysis of software defined network based prevention on phishing attack in cyberspace using a deep machine learning with CANTINA approach (DMLCA),'' Comput. Commun., vol. 153, pp. 375–381, Mar. 2020.

[13] M. Sameen, K. Han, and S. O. Hwang, ''PhishHaven—An efficient real-time AI phishing
URLs detection system,'' IEEE Access, vol. 8, pp. 83425–83443, 2020.

[14] R. S. Rao, T. Vaishnavi, and A. R. Pais,
''PhishDump: A multi-model ensemble based technique for the detection of phishing sites in mobile devices,'' Pervasive Mobile Comput., vol. 60, Nov. 2019, Art. no. 101084.

[15] Y. Ding, N. Luktarhan, K. Li, and W. Slamu, ''A keyword-based combination approach for detecting phishing webpages,'' Comput. Secur., vol. 84, pp. 256–275, Jul. 2019.

[16] A. K. Jain and B. B. Gupta, ''A machine learning based approach for phishing detection using hyperlinks information,'' J. Ambient Intell. Hum. Comput., vol. 10, no. 5, pp. 2015–2028, May 2019.

[17] A. E. Aassal, S. Baki, A. Das, and R. M. Verma, ''An in-depth benchmarking and evaluation of phishing detection research for security needs,'' IEEE Access, vol. 8, pp. 22170–22192, 2020.

[18] P. Vaitkevicius and V. Marcinkevicius, ''Comparison of classification algorithms for detection of phishing websites,'' Informatica, vol. 31, no. 1, pp. 143–160, Mar. 2020.

[19] Y.-H. Chen and J.-L. Chen, ''AI@ntiPhish—
Machine learning mechanisms for cyber-phishing attack,'' IEICE Trans. Inf. Syst., vol. E102.D, no. 5, pp. 878–887, May 2019.

[20] C. L. Tan, K. L. Chiew, K. S. C. Yong, S. N. Sze, J. Abdullah, and Y. Sebastian, ''A graph-theoretic approach for the detection of phishing webpages,''
Comput. Secur., vol. 95, Aug. 2020, Art. no. 101793.

[21] S. Mishra and D. Soni, ''Smishing detector: A security model to detect smishing through SMS content analysis and URL behavior analysis,'' Future
Gener. Comput. Syst., vol. 108, pp. 803–815, Jul. 2020.

[22] M. Karabatak and T. Mustafa, ''Performance comparison of classifiers on reduced phishing website dataset,'' in Proc. 6th Int. Symp. Digit. Forensic Secur. (ISDFS), Mar. 2018, pp. 1–5.

[23] N. N. Gana and S. M. Abdulhamid, ''Machine learning classification algorithms for phishing detection: A comparative appraisal and analysis,'' in Proc. 2nd Int. Conf. IEEE Nigeria Comput. Chapter (NigeriaComputConf), Oct. 2019, pp. 1–8.

[24] T. Gangavarapu, C. D. Jaidhar, and B. Chanduka, ‘‘Applicability of machine learning in spam and phishing email filtering: Review and approaches,’’ Artif. Intell. Rev., vol. 53, no. 7, pp. 5019–5081, Oct. 2020.

[25] S. Priya, S. Selvakumar, and R. L. Velusamy, ‘‘Evidential theoretic deep radial and probabilistic neural ensemble approach for detecting phishing attacks,’’ J. Ambient Intell. Hum. Comput., vol. 14, no. 3, pp. 1951–1975, Jul. 2021.

[26] P. L. Indrasiri, M. N. Halgamuge, and A. Mohammad, ‘‘Robust ensemble machine learning model for filtering phishing URLs: Expandable random gradient stacked voting classifier (ERGSVC),’’ IEEE Access, vol. 9, pp. 150142–150161, 2021.

[27] A. Ozcan, C. Catal, E. Donmez, and B. Senturk, ‘‘A hybrid DNN-LSTM model for detecting phishing URLs,’’ Neural Comput. Appl., vol. 35, no. 7, pp. 4957–4973, Aug. 2021.

[28] S.-J. Bu and H.-J. Kim, ‘‘Optimized URL feature selection based on genetic-algorithm-embedded deep learning for phishing website detection,’’ Electronics, vol. 11, no. 7, p. 1090, Mar. 2022.

[29] V. Zeng, S. Baki, A. E. Aassal, R. Verma, L. F. T. De Moraes, and A. Das, ‘‘Diverse datasets and a customizable benchmarking framework for phishing,’’ in Proc. 6th Int. Workshop Secur. Privacy Anal., Mar. 2020, pp. 35–41.

[30] A. Ihsan and E. Rainarli, ‘‘Optimization of knearest neighbour to categorize Indonesian’s news articles,’’ Asia–Pacific J. Inf. Technol. Multimedia, vol. 10, no. 1, pp. 43–51, Jun. 2021.

[31] Y. A. Alsariera, V. E. Adeyemo, A. O. Balogun, and A. K. Alazzawi, ‘‘AI meta-learners and extra-trees algorithm for the detection of phishing websites,’’ IEEE Access, vol. 8, pp. 142532–142542, 2020.

[32] E. Sukawai and N. Omar, ‘‘Corpus development for Malay sentiment analysis using semi supervised approach,’’ Asia–Pacific J. Inf. Technol. Multimedia, vol. 9, no. 1, pp. 94–109, Jun. 2020. [33] C. L. Tan, ‘‘Phishing dataset for machine learning: Feature evaluation,’’ Mendeley Data, V1, 2018, doi: 10.17632/h3cgnj8hft.1.

[34] X.-Y. Lu, M.-S. Chen, J.-L. Wu, P.-C. Chang, and M.-H. Chen, ‘‘A novel ensemble decision tree based on under-sampling and clonal selection for web spam detection,’’ Pattern Anal. Appl., vol. 21, no. 3, pp. 741–754, Aug. 2018.

[35] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, ‘‘A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches,’’ IEEE Trans. Syst., Man C, Appl. Rev., vol. 42, no. 4, pp. 463–484, Jul. 2012.

[36] E. S. Gualberto, R. T. De Sousa, T. P. D. B. Vieira, J. P. C. L. Da Costa, and C. G. Duque, ‘‘From feature engineering and topics models to enhanced prediction rates in phishing detection,’’ IEEE Access, vol. 8, pp. 76368–76385, 2020.

[37] H. Zhang, G. Liu, T. W. S. Chow, and W. Liu, ‘‘Textual and visual contentbased anti-phishing: A Bayesian approach,’’ IEEE Trans. Neural Netw., vol. 22, no. 10, pp. 1532–1546, Oct. 2011.

[38] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Amsterdam, The Netherlands: Elsevier, 2017.

[39] T. Saito and M. Rehmsmeier, ‘‘The precisionrecall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,’’ PLoS ONE, vol. 10, no. 3, pp. 1–21, Mar. 2015.

[40] T. Fawcett, ‘‘An introduction to ROC analysis,’’ Pattern Recognit. Lett., vol. 27, no. 8, pp. 861–874, Dec. 2006.

[41] A. E. Aassal, L. Moraes, S. Baki, A. Das, and R.

Press, 2011.

Verma, ‘‘Anti-phishing pilot at ACM IWSPA 2018: Evaluating performance with new metrics for unbalanced datasets,’’ in Proc. Anti-Phishing Shared Task Pilot 4th ACM IWSPA, 2018, pp. 2–10.

[42] H. A. Alshalabi, S. Tiun, and N. Omar, ‘‘A comparative study of the ensemble and base classifiers performance in Malay text categorization,’’ Asia–

Pacific J. Inf. Technol. Multimedia, vol. 6, no. 2, pp. 53–64, Dec. 2017.

[43] N. Japkowicz and M. Shah, Evaluating Learning Algorithms. Cambridge, U.K.: Cambridge Univ.

[44] R. Gowtham and I. Krishnamurthi,

‘‘PhishTackle—A web services architecture for antiphishing,’’ Cluster Comput., vol. 17, no. 3, pp. 1051– 1068, Sep. 2014.