Abstract: This dataset collected mainly from: PhishTank archive, MillerSmiles archive, Googleâ€™s searching operators.

Associated Tasks:	Classification	**Number of Instances:**	2456	**Number of Attributes:**	30

Data Set Information:

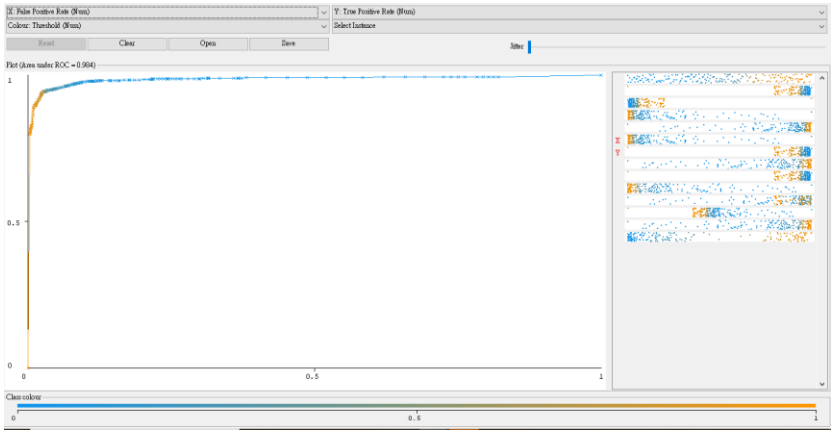
One of the challenges faced by our research was the unavailability of reliable training datasets. In fact this challenge faces any researcher in the field. However, although plenty of articles about predicting phishing websites have been disseminated these days, no reliable training dataset has been published publically, may be because there is no agreement in literature on the definitive features that characterize phishing webpages, hence it is difficult to shape a dataset that covers all possible features.
In this dataset, we shed light on the important features that have proved to be sound and effective in predicting phishing websites. In addition, we propose some new features.

1. Classification by using J48 algorithm

Test mode: 10-fold cross-validation.

```
Correctly Classified Instances       10599               95.8752 %
Incorrectly Classified Instances       456                4.1248 %
Kappa statistic                          0.9162
Mean absolute error                      0.0567
Root mean squared error                  0.1853
Relative absolute error                 11.4861 %
Root relative squared error             37.3035 %
Total Number of Instances            11055

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.942    0.028    0.964      0.942    0.953      0.916    0.984     0.983     -1
               0.972    0.058    0.955      0.972    0.963      0.916    0.984     0.978     1
Weighted Avg.  0.959    0.045    0.959      0.959    0.959      0.916    0.984     0.980

=== Confusion Matrix ===

    a    b   <-- classified as
 4615  283 |    a = -1
  173 5984 |    b = 1
```

ROC curve (J48):



N=11055(Positive=Phishing, Negative= None-Phishing)

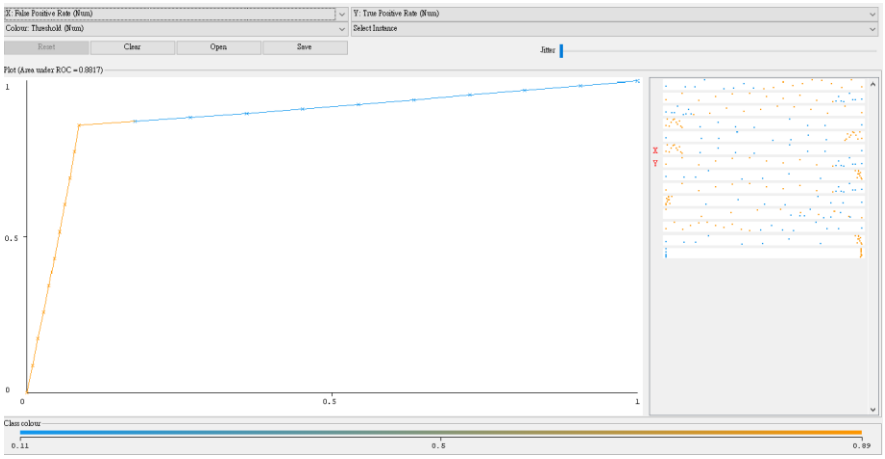
Actual \ Predict	Phishing	Not phishing
Phishing	4615(TP)	283(FN)
Not phishing	173(FP)	5984(TN)

2. Classification by Decision Stamp

Test mode: 10-fold cross-validation.

```
Correctly Classified Instances        9827               88.8919 %
Incorrectly Classified Instances      1228               11.1081 %
Kappa statistic                          0.7741
Mean absolute error                      0.1975
Root mean squared error                  0.3143
Relative absolute error                 40.0168 %
Root relative squared error             63.2617 %
Total Number of Instances            11055

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.857    0.086    0.888      0.857    0.872      0.774    0.882     0.838     -1
                 0.914    0.143    0.889      0.914    0.902      0.774    0.882     0.866     1
Weighted Avg.    0.889    0.118    0.889      0.889    0.889      0.774    0.882     0.854

=== Confusion Matrix ===

    a    b   <-- classified as
 4197  701 |    a = -1
  527 5630 |    b = 1
```

ROC curve(Decision Stump)



N=11055(Positive=Phishing, Negative= None-Phishing)

Actual \ Predict	Phishing	Not phishing
Phishing	4197(TP)	701(FN)
Not phishing	527(FP)	5630(TN)

3. "Confusion matrix" of 2 decision tree classifiers

	# of samples	# of classes	# of Attributes (features)	TP	TN	FP	FN	ROC area	PRC area
J48	11055	2	30	4615	5984	173	283	0.984	0.983
Decision Stump	11055	2	30	4197	5630	527	701	0.882	0.854

4. Analysis of the results by comparing the result for 2 decision trees.

By using the two decision tree, we want to analyze whether the website is the phishing website. Here, the true positive (TP) means the website is phishing website, true negative (TN) means that the website is not the phishing website.

Accuracy: It means that in all of the data, the percentage of the data that are correctly distributed. In J48 algorithm, the accuracy is 0.9587, and in decision stump algorithm, the accuracy is 0.8889.It can be interpreted that in all of the websites, almost 95% of the email in J48 algorithm is correctly distributed in the right categories, while there is only about 89% in decision stump algorithm.

Error Rate: It means that in all of the data, the percentage of the data that are not correctly distributed, which equals to 1-accuracy. In J48 algorithm, the error rate is 0.041248, and in decision stump algorithm, the error rate is 0.111081. It can be interpreted that in all of the emails, about 4% of the websites is not correctly distributed in the right categories in J48 algorithm; however, there is about 11% in decision stump algorithm.

Sensitivity: It means that the true Phishing websites are correctly classified into the Phishing website category, and the percentage in J48 algorithm is 94.2%, while in decision stump algorithm is 85.7%.

FP Rate: It is very important to find out false positive outcome (Real Non-Phishing website but in the Phishing websites category), and we hope it can be as low as possible, so we count the FP rate, which is calculated by FP divide by the Actual negative category, and in J48 algorithm, the percentage is 2.8%, while in decision stump algorithm is up to 8.6%.

Precision: It means that in the predicted Phishing categories, the percentage of the real Phishing websites; therefore, the higher the precision rate, the lower the FP rate. The precision rate is counted by TP divided by the Actual Predicted positive category, and the percentage in J48 algorithm is 96.4%, while in decision stump algorithm is 88.8%.

Recall: It counts the percentage of all of the Phishing websites which should be

detected (TP) in all of the detected Phishing website (TP+FN), and the result in J48 algorithm is 94.2%, while in decision stump algorithm is 85.7%.

F-Score: In general, we will combine the precision rate and the recall rate into the F-Score, so it's a comprehensive indicator, the higher the F-score, the higher the prediction rate and the recall rate in the model. The F-score in in J48 algorithm is 0.953, while in decision stump algorithm is 0.872.

MCC: It considers the TP, FP, TN, and FN, and can be seen as a balance measure. MCC is actually the correlation coefficient of the binary classification of the prediction data and the real data, and it usually in the range from -1 to +1. If the coefficient equals to 1, it means the model is perfectly predicted, while if the MCC equals to 0, the prediction can be seem as similar as the random prediction, and if the coefficient equals to -1, it tells the prediction is totally inconsistent with the real data. In this phishing website model, by using the J48 algorithm, MCC is equal to 0.916, while in decision stump algorithm is 0.774.

ROC area: For ROC curve, the x-axis is the recall rate, and the y-axis is precision rate. For PR curve, the x-axis is the FP rate, and the y-axis is TP rate. Area under the ROC Curve shows the effectiveness of the classification, and in this model, by using the J48 algorithm, ROC area is equal to 0.984, while in decision stump algorithm is 0.882. As the area under the ROC curve in J48 algorithm is larger than that in decision stump algorithm, so J48 algorithm gives us a better outcome.

PRC area: For PR curve, the x-axis is the recall rate, and the y-axis is precision rate. When the Positive=Phishing, by using the J48 algorithm, area under the PRC area equals to 0.938, while in decision stump algorithm is 0.838.

In conclusion, by examine all of the indicators above, we found that in all of the indicator, the J48 algorithm got the better performances than those in the decision stamp algorithm; therefore, the J48 algorithm gives a better result.