EDS 6340 INTRODUCTION TO DATA SCIENCE:
GROUP 2 – Website Phishing

Vijaya Durga Pantham-2146476

Manas Anand Nissankara – 2158728

Sahithya Reddy Kadapakonda-2147286

Rohith Bejjam - 2138034

1. Introduction

The dataset taken from UCI repository

https://archive.ics.uci.edu/ml/datasets/Website+Phishing

i) Instances :1353

ii) Attributes :10

Website phishing is one of the most serious security concerns for the online community. Our goal is to examine the issue of online phishing and to uncover characteristics that separate phishing websites from legal ones. Every day, individuals and private people are targeted by website phishing. Over $100 million was scammed from tech companies, and attacks are being undertaken on small and medium-sized firms.

Our dataset is well-structured and well-defined. In our dataset, three categories (Legitimate, Suspicious, and Phishing) are represented by numbers.

1 – Legitimate websites

0 – Suspicious

-1 – Phishing

Attributes

SFH, popUpWindow, SSLfinal State, Request URL, URL of Anchor, web traffic, URL length, domain age, having IP address, Result attributes. We only have nominal values and not a Multimodel dataset. Multimodel data refers to data of several types and contexts (image, text..etc)

These files contain a collection of legal and phishing website examples. Each website is represented by a collection of attributes that indicate whether the website is real. Data may be used to feed into the machine learning process.

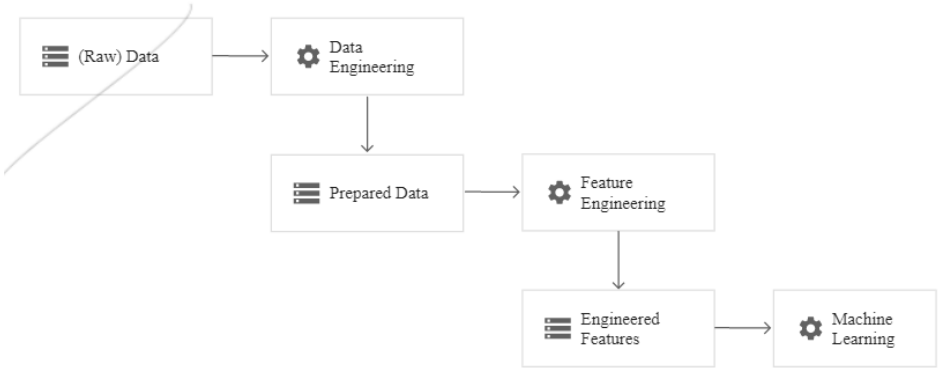
• These datasets can help machine learning and data mining researchers, as well as security researchers and practitioners. Computer security aficionados might find these datasets intriguing for creating firewalls, clever ad blockers, and virus detection systems.

• Because the presented datasets contain easily extracted attributes, this dataset can assist researchers and practitioners in easily building classification models in systems preventing phishing attacks.

Finally, the datasets given might be utilized as a performance benchmark for creating cutting-edge machine learning approaches for phishing website categorization.

2. Pre-Processing of the data

Data preprocessing includes several operations. Each operation is designed to help ML build better predictive models.



Handling missing data

```
In [8]: sum_1=0
        for column in data.columns[:]:
            sum_1 += data[column].isnull().sum()
            if sum_1!= 0:
                print("data is not clean")
        if sum_1 == 0:
            print("data is clean")

        data is clean

In [7]: data.isnull().sum()

Out[7]: SFH                0
        popUpWidnow        0
        SSLfinal_State     0
        Request_URL        0
        URL_of_Anchor      0
        web_traffic        0
        URL_Length         0
        age_of_domain      0
        having_IP_Address  0
        Result             0
        dtype: int64
```

Cleaned Data

	SFH	popUpWidnow	SSLfinal_State	Request_URL	URL_of_Anchor	web_traffic	URL_Length	age_of_domain	having_IP_Address	Result
0	1	-1	1	-1	-1	1	1	1	0	0
1	-1	-1	-1	-1	-1	0	1	1	1	1
2	1	-1	0	0	-1	0	-1	1	0	1
3	1	0	1	-1	-1	0	1	1	0	0
4	-1	-1	1	-1	0	0	-1	1	0	1

3. Results for all single models

We have implemented the following models to our classification dataset and checked which model is performing better by checking the accuracy of each model. In our dataset total three classes are there 1, 0, -1. We have oversampled the data and fitted the model and performed hyperparameter tuning to each model using GridSearchCV or Randomized Search.

KNN

With our phishing data we have implemented KNN model we got an accuracy of 85%.

	precision	recall	f1-score	support
-1	0.91	0.82	0.86	134
0	0.64	0.86	0.73	21
1	0.83	0.88	0.85	99
accuracy			0.85	254
macro avg	0.79	0.85	0.82	254
weighted avg	0.86	0.85	0.85	254

Logistic Regression

For Logistic Regression we got an accuracy of 77%

	precision	recall	f1-score	support
-1	0.91	0.77	0.83	134
0	0.29	0.62	0.39	21
1	0.83	0.81	0.82	99
accuracy			0.77	254
macro avg	0.68	0.73	0.68	254
weighted avg	0.83	0.77	0.79	254

Random Forest

For Random Forest we got an accuracy of 91%

	precision	recall	f1-score	support
-1	0.95	0.88	0.91	134
0	0.83	0.90	0.86	21
1	0.87	0.94	0.90	99
accuracy			0.91	254
macro avg	0.88	0.91	0.89	254
weighted avg	0.91	0.91	0.91	254

Linear Regression

For Linear Regression we got an accuracy of 82%

	precision	recall	f1-score	support
-1	0.86	0.90	0.88	134
0	0.00	0.00	0.00	21
1	0.77	0.90	0.83	99
accuracy			0.82	254
macro avg	0.55	0.60	0.57	254
weighted avg	0.76	0.82	0.79	254

Non-Linear SVM

For Non Linear SVM we got an accuracy of 94%

	precision	recall	f1-score	support
-1	0.97	0.93	0.95	134
0	0.81	1.00	0.89	21
1	0.94	0.94	0.94	99
accuracy			0.94	254
macro avg	0.91	0.96	0.93	254
weighted avg	0.94	0.94	0.94	254

Linear SVM

For Linear SVM we got an accuracy of 83%

	precision	recall	f1-score	support
-1	0.92	0.82	0.87	134
0	0.45	0.90	0.60	21
1	0.88	0.83	0.85	99
accuracy			0.83	254
macro avg	0.75	0.85	0.78	254
weighted avg	0.87	0.83	0.84	254

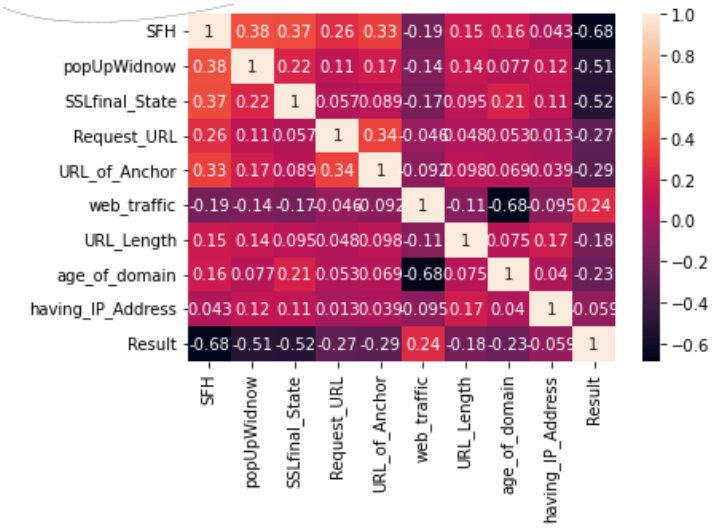
4. First Variable Selection

- Variable selection is performed using the correlation technique.
- Correlation statistics is performed on our dataset to find the first variable selection. Features with high correlation are more linearly dependent.

	feature names	coef
0	SFH	1.562888
1	popUpWidnow	1.674898
2	SSLfinal_State	1.305973
3	Request_URL	0.926274
4	URL_of_Anchor	0.295214
5	web_traffic	-0.087095
6	URL_Length	-0.105277
7	age_of_domain	0.184485
8	having_IP_Address	-0.597664

- We consider only those features that have a coefficient different from 0. So, from the features above we can discard web_traffic as it is nearer to 0 i.e it has 0 importance**.**

- Feature Selection helps to improve the model accuracy, low computational cost and easier to understand and explain



5. Bi-directional Elimination, Wrapper Method

In backward elimination, we start with the full model including all the independent variables and then remove the insignificant feature with the highest p-value. This process repeats again and again until we have the final set of significant features. Bi -directional is mixture of two methods:

- Forward
- Backward

With this method we removed the features that are not required. And build the model, the best models are Random forest with accuracy 94% and Nonlinear SVM with accuracy 92%

Random Forest

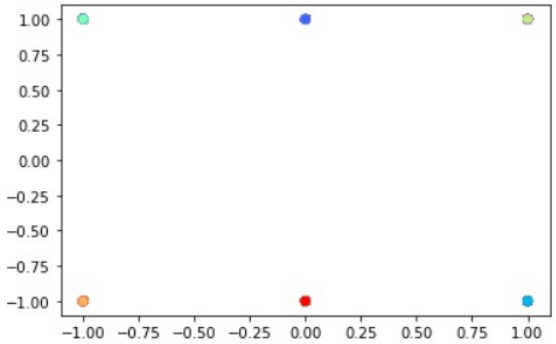
	precision	recall	f1-score	support
-1	0.98	0.92	0.95	134
0	0.84	1.00	0.91	21
1	0.92	0.96	0.94	99
accuracy			0.94	254
macro avg	0.91	0.96	0.93	254
weighted avg	0.94	0.94	0.94	254

Non-Linear SVM

	precision	recall	f1-score	support
-1	0.97	0.92	0.94	134
0	0.76	0.90	0.83	21
1	0.90	0.93	0.92	99
accuracy			0.92	254
macro avg	0.88	0.92	0.89	254
weighted avg	0.93	0.92	0.92	254

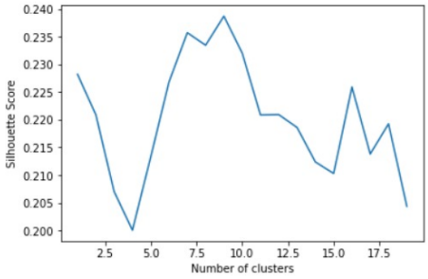
6. Clustering

Clustering is performed totally 9 clusters are identified and plotted the clusters. By taking K value as 9, data has been fit to the KMeans model. Below is the representation of cluster



Silhouette Coefficient

- Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.1
- Means clusters are well apart from each other and clearly distinguished.0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.-1: Means clusters are assigned in the wrong way



With these set of clusters, we calculated the inertia, and k-means score. The performance of clustering is low on our dataset, so this is not the best model.

The clustering useful to build a predictive model**:**

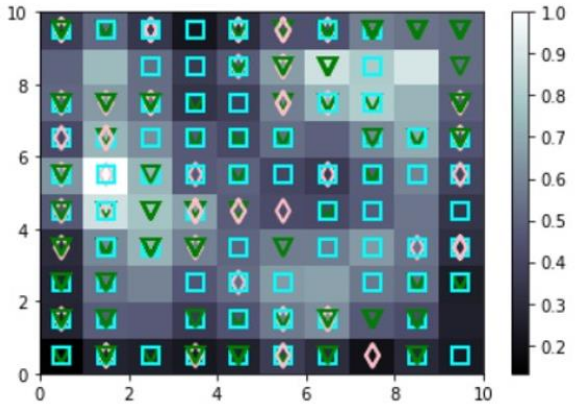
Clustering can also serve as a useful data-preprocessing step to identify homogeneous groups on which to build predictive models.

7. Visualization using Dimensionality Reduction

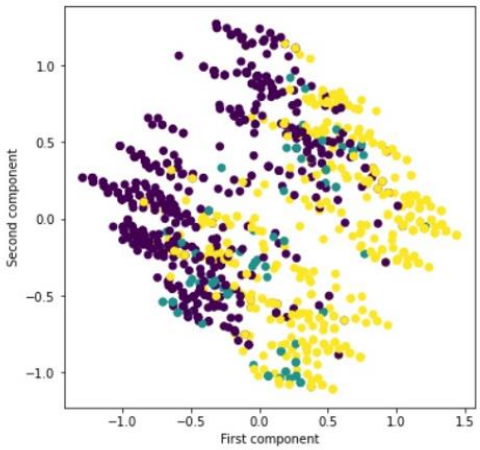
PCA

Principal Component Analysis goal is to minimize the dimensionality of datasets comprising of several highly or weakly correlated variables while keeping as much of the dataset's variety as possible. Similar work is done by the variables are changed into a new set of variables known as principal components.

SOM



PCA



8. Ensemble Modelling

Using ensemble learning, numerous model accuracies are pooled to provide results with greater accuracy than a single model could produce on its own. Kernel SVM and Random Forest were the models which performed remarkably well at validating data. The accuracy of the voting classifier, an ensemble model with 20 classifiers, was 0.92, not significantly different from the accuracy of the single models and even extremely near to the accuracy of the Random Forest model but the recall for the ensemble model has higher.

Single model result

	precision	recall	f1-score	support
-1	0.93	0.91	0.92	141
0	0.83	0.75	0.79	20
1	0.89	0.93	0.91	110
accuracy			0.91	271
macro avg	0.89	0.86	0.87	271
weighted avg	0.91	0.91	0.91	271

Ensemble result of all models

VC Classification Report on Test data using best model

```
[[131   2   8]
 [  1  18   1]
 [  8   1 101]]
```

	precision	recall	f1-score	support
-1	0.94	0.93	0.93	141
0	0.86	0.90	0.88	20
1	0.92	0.92	0.92	110
accuracy			0.92	271
macro avg	0.90	0.92	0.91	271
weighted avg	0.92	0.92	0.92	271

10. General Discussion

In the end, we were successful in achieving the project's core goals, however area where we could have improved was visualization and comparing the models, as the train and test split was different for different model. There is still a lot to be shown even though we have included numerous visuals whenever and wherever it was appropriate. Due to the size of our dataset, this problem has less solid information on the phishing data set. The attributes are not highly correlated with the output attribute. Execution of some programs, particularly the hyperparameter adjustment, took a very long time. However, we are proud of what we have accomplished because we learned how to analyze data, clean it, visualize it, deal with different data scales, outliers, choose a model's structure, validate the model's predictions, and put the right classification algorithms, feature selection methods, and dimensionality reduction techniques into practice.

11. Conclusion

We have been effective in utilizing several classification algorithms to comprehend the nature of website phishing and recommend the necessary course of action. We can maximize the value of the data. We discovered elements (factors) that are important for generating predictions. By analyzing this data, it is possible to detect the websites which are not trustable. We were able to gather and examine 9 different characteristics that set phishing websites away from trustworthy ones. Making predictions enabled us to locate the important feature. Compared the outcomes of various classification methods to choose the model that satisfying our demands. Recognized the importance of features in a classification model and visualized them.

12.References

1. https://www.kaggle.com/code/davidfumo/comparing-11-classification-models
2. https://towardsdatascience.com/feature-selection-for-machine-learning-in-python-wrapper-methods-2b5e27d2db31
3. https://www.kaggle.com/datasets/shashwatwork/web-page-phishing-detection-dataset