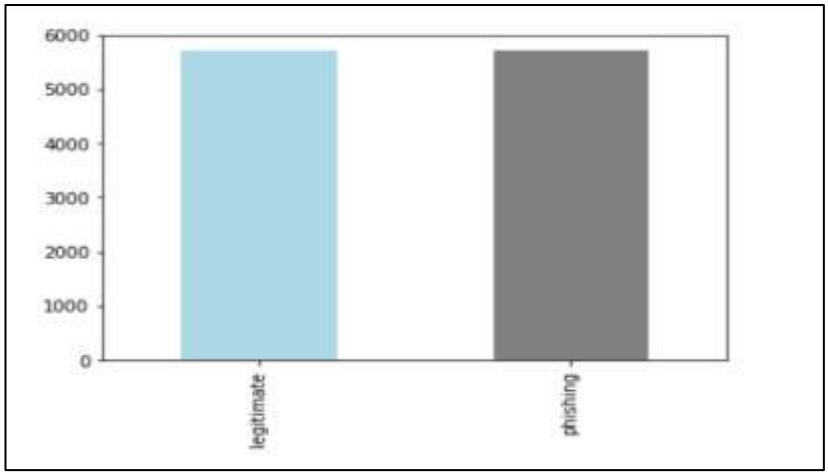
DETECTING PHISHING WEBSITES

Group members:
Shreya Kakade (14)
Apeksha Kamath (15)
Atharva Kulkarni (26)

Google Colab link: https://colab.research.google.com/drive/1bK0QblEbgQYkTjSwH21DMms_-zOn7x6W?usp=sharing#scrollTo=vuKZrlRgs8gk

Inferences:

1. There are a total of 11430 rows and 89 columns in our dataset.
2. There were no null values in the dataset.
3. The dataset is perfectly balanced.



4. There were two columns in our dataset in which were of object datatype. The first column was ‘url’ and the second column was our target variable ‘status’. Since ‘url’ column is not significant for the analysis, we dropped that column.
5. We performed categorical encoding for the target variable which had categorical variables viz. ‘legitimate’ and ‘phishing’. So, legitimate was encoded as ‘1’ and phishing was encoded as ‘0’.
6. On analysis we found that our dataset set had too many columns. Since too many features do not always ensure better results and sometimes too much data can be confusing, difficult to sort out and analyse, and can lead businesses in the wrong direction, we performed feature selection to reduce the number of columns and only focus on the columns significant for analysis.
7. We performed Filter based feature selection methods which included Removing Constant features, Quasi constant features, duplicate features, correlated features, univariate features and mutual information, etc. and embedded feature selection methods which included lasso regularization and compared the accuracies obtained by the two methods.

8. We also compared the accuracies by splitting the dataset in different ratios. The results obtained are as follows:
   1. For train - test split ratio of 70:30

	Type	Steps	KNN	DT	RF	NB	No. of Features
0	Raw data	None	83.87	93.64	96.00	74.94	87
1	Filter based(basic)	Remove Constant, Quasi Constant & Duplicated F...	83.87	93.58	95.71	74.94	64
2	Filter based (correlated)	Remove Correlated features (Spearman)	83.84	93.52	95.77	74.65	58
3	Statistical	Univariate	83.84	93.67	95.36	74.62	56
4	Filter-Based	Mutual Information	86.67	89.03	93.72	74.39	50
5	Embedded	Lasso	84.28	92.38	94.60	63.38	7

   2. For train – test split ratio of 80:20

	Type	Steps	KNN	DT	RF	NB	No. of Features
0	Raw data	None	84.03	93.96	96.54	74.80	87
1	Filter based(basic)	Remove Constant, Quasi Constant & Duplicated F...	84.03	93.48	95.66	74.80	65
2	Filter based (correlated)	Remove Correlated features (Spearman)	83.98	94.22	96.15	74.71	59
3	Statistical	Univariate	83.98	94.09	95.75	74.71	57
4	Filter-Based	Mutual Information	87.40	89.58	94.22	75.06	51
5	Embedded	Lasso	83.90	93.08	94.75	69.64	7

9. After performing feature selection, we were able to reduce the number of columns from 89 to 51 without compromising on accuracy. The number was further reduced to 7 with a decrease in accuracy.
10. On analysis, we found that the best accuracies were obtained by using Random Forest model.