**FRAUD URL DETECTION**

**ABSTRACT :-**

Phishing attack is a simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Aim of the paper is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm.

1. **INTRODUCTION :-**

Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. In United States businesses, there is a loss of US$2billion per year because their clients become victim to phishing. In 3rd Microsoft Computing Safer Index Report released in February 2014, it was estimated that the annual worldwide impact of phishing could be as high as $5 billion. Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques. The general method to detect phishing websites by updating blacklisted URLs, Internet Protocol (IP) to the antivirus database which is also known as “blacklist" method. To overcome the drawbacks of blacklist method, many security researchers now focused on machine learning techniques. Machine learning technology consists of a many algorithms which requires past data to make a decision or prediction on future data. Using this technique, algorithm will analyze various blacklisted and legitimate URLs and their features to accurately detect the phishing websites.

1. **DATASET:**

The URL\_Dataset of Phishing websites were collected from www.kaggle.com . The data set consists of total 11430 rows and 89 columns which include 5715 Legitimate URLs and 5715 Phishing URLs. Legitimate URLs are labelled as “0” and Phishing URLs are labelled as “1”.

There are 88 Independent Variables and 1 Dependent Variable representing “Legitimate” and “Phishing”.

1. **DATA PRE-PROCESSING PHASE:**

This Phase Inspects the several pre-processing cases that provides us the most appropriate and cleaned data for the further Use In Detection Technique using ML Algorithms .

* **Detection Of Missing Values** **:-** Through this Method, We try to check if the data contains any missing values or not . If Yes, we probably try to estimate the missing values using the Strategy either “Mean (where the data is mostly affected by the Outliers)”, “Median (where the data is least affected by Outliers with respect to mean)” and “Mode (which can be used in case of qualitative data)”. We can use several techniques to detect and fix the missing values like by checking the individual rows or Columns and so on.
* **Detection Of Outliers** **:-** Outliers are those characteristics that holds certain different anomalies behaviour from other attributes/features. We can detect the Outliers from a known method by using the BOXPLOT visualization. In BOXPLOT Representation , the data points lying above 100%-percentile or below 0%-percentile represents the Outliers. There are several other methods to detect Outliers like Scatter Plot ,Z-Score, etc.
* **Detection of Correlation among the features using Correlation Heatmap :-**

Using Correlation Heatmap , We try to find the relationship among the features. The relationship implies to be either Strong , Weak or No Relation between the features.

Before using the Correlation Heatmap, we are provided with the Correlation Matrix which consists of all the correlated values among the features that ranges according to the dataset.

Correlation ranges between -1 to 1. The more closer is the Correlated value to -1 or 1, more the strong and positive or negtive correlation to be. No Correlation holds when the Correlated value between any two feature becomes zero implies the two variables to be independent.

* **Detection of Skewness in the Dataset :-**

Skewness is a measure of the asymmetry of a distribution. A distribution can have right (or positive), left (or negative), or zero skewness. By the help of Skewness , one can check the dataset to be positively or negatively skewed as The larger the skew, the greater are the proportion of correlations that are negative . In this dataset , most of the features are positively skewed. If the dataset are badly skewed, bimodal, or otherwise violate the assumptions of the general linear model, then one can better use Spearman's rho (rank-order correlation) or Kendall's tau.

We can enhance our dataset through the data cleaning phase which provides with a dataset free from outliers, missing values, etc. and check the several insights using the visualization plots like Box Plot, Distribution Plot, Pair Plot, Joint Plot, etc.

* **Checking for the Categorical Data in the dataset :-**

The Dataset should be comprised of features of the Same Data Type. If this Condition is violated implies presence of features with different data types, One can use two important pre-processing module named as “ OneHotEncoder ” (For Independent variable) and

“ LabelEncoder ” (For Dependent variable) to convert all the categorical features into integer based features.

* **Feature Scaling :-**

Feature Scaling is a technique to standardize the independent and dependent features present in the data in a fixed range. It is performed during the data pre-processing. Feature scaling through standardization (or Z-score normalization) can be an important pre-processing step involves rescaling the features such that they have the properties of a standard normal distribution with a mean of zero and a standard deviation of one.

1. **MACHINE LEARNING ALGORITHM :-**

Six machine learning classification model Logistic, Decision Tree, Random Forest and Support Vector Machine, K-Nearest Neighbour, Naïve Bayes Classifier has been selected to detect the phishing websites.

* **Logistic Regression Classifier :-**

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

* **Decision Tree Classifier :-**

One of the most widely used algorithm in machine learning technology. Decision tree algorithm is easy to understand and also easy to implement. Decision tree begins its work by choosing best splitter from the available attributes for classification which is considered as a root of the tree. Algorithm continues to build tree until it finds the leaf node. Decision tree creates training model which is used to predict target value or class in tree representation each internal node of the tree belongs to attribute and each leaf node of the tree belongs to class label. In decision tree algorithm, Gini index and Information gain methods are used to calculate these nodes.

* **Random Forest Classifier :-**

Random Forest algorithm is one of the most powerful algorithms in machine learning technology and it is based on concept of decision tree algorithm. Random Forest algorithm creates the forest with number of decision trees. High number of tree gives high detection accuracy. Creation of trees are based on bootstrap method. In bootstrap method features and samples of dataset are randomly selected with replacement to construct single tree. Among randomly selected features, Random Forest algorithm will choose best splitter for the classification and like decision tree algorithm; Random Forest algorithm also uses Gini index and Information gain methods to find the best splitter. This process will get continue until Random Forest creates n number of trees. Each tree in forest predicts the target value and then algorithm will calculate the votes for each predicted target. Finally Random Forest algorithm considers high voted predicted target as a final prediction.

* **Support Vector Machine :-**

Support Vector Machine is another powerful algorithm in machine learning technology. In Support Vector Machine algorithm each data item is plotted as a point in n-dimensional space and Support Vector Machine algorithm constructs separating line for classification of two classes, this separating line is well known as hyperplane. Support Vector Machine seeks for the closest points called as support vectors and once it finds the closest point it draws a line connecting to them. Support Vector Machine then construct separating line which bisects and perpendicular to the connecting line. In order to classify data perfectly the margin should be maximum. Here the margin is a distance between hyperplane and support vectors. In real scenario it is not possible to separate complex and non linear data, to solve this problem Support Vector Machine uses kernel trick which transforms lower dimensional space to higher dimensional space.

* **K-Nearest Neighbour :-**

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. It assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. It stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. It can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data. It is simple to implement. It is robust to the noisy training data KNN Algorithm can be more effective if the training data is large.

* **Naïve Bayes Classifier :-**

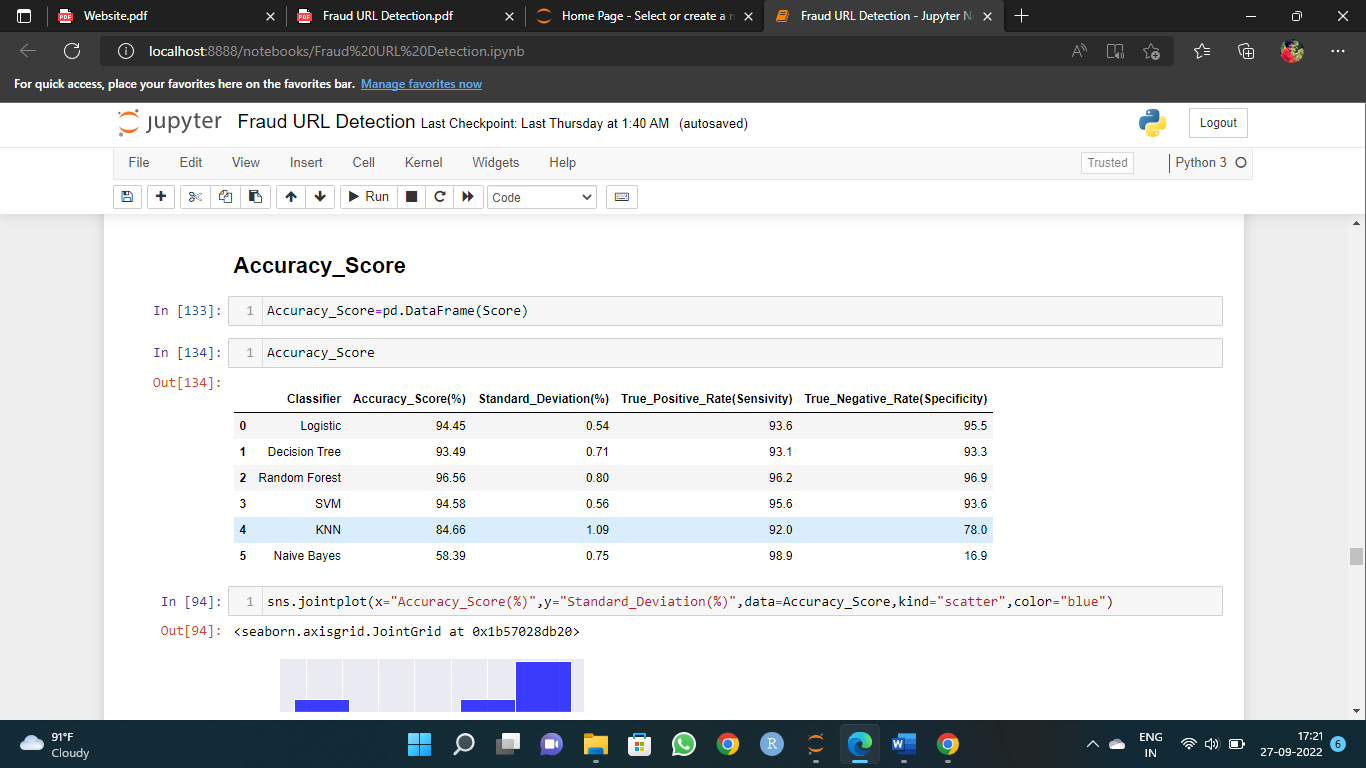
Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles. The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:

**Naïve**: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of colour, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.

**Bayes**: It is called Bayes because it depends on the principle of Bayes' Theorem.

1. **IMPLEMENTATION AND RESULT :-**

Scikit-learn tool has been used to import Machine learning algorithms. Dataset is divided into training set and testing set in 70:30 ratio respectively. Each classifier is trained using training set and testing set is used to evaluate performance of classifiers. Performance of classifiers has been evaluated by calculating classifier's accuracy score, standard deviation true positive rate(Sensitivity) and true negative rate(Specificity).



**Fig. 1:- REPRESENTATION OF ACCURACY SCORE**

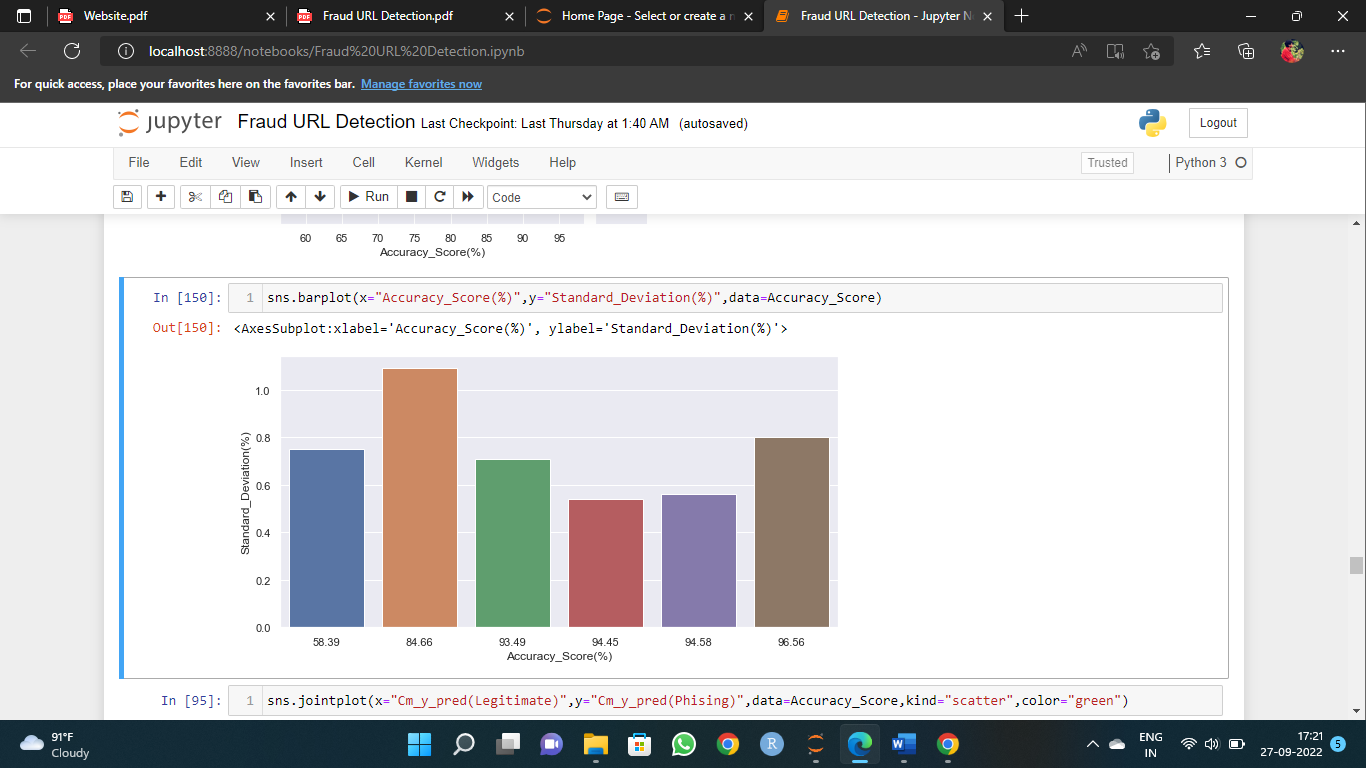
 **Fig. 2:- ACCURACY COMPARISON**

Fig. 1 Shows that the Random Forest Classifier has the Highest Accuracy Score 96.56% with second highest deviation 0.80% with respect to other techniques whereas Naïve Bayes has the Least Accuracy Score 58.39% with a good amount of deviation 0.75%.

Result also shows that detection accuracy of phishing websites increases as more dataset used as training dataset. All classifiers performed good when 70% of data used as training dataset except Naïve Bayes.

1. **Feature Extraction Using Principal Component Analysis :-**

Principal Component Analysis is an unsupervised learning algorithm that is used for the dimensionality reduction in machine learning. It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation. These new transformed features are called the **Principal Components**. It is one of the popular tools that is used for exploratory data analysis and predictive modelling. It is a technique to draw strong patterns from the given dataset by reducing the variances. PCA generally tries to find the lower-dimensional surface to project the high-dimensional data. PCA works by considering the variance of each attribute because the high attribute shows the good split between the classes, and hence it reduces the dimensionality. Some real-world applications of PCA are image processing, movie recommendation system, optimizing the power allocation in various communication channels. It is a feature extraction technique, so it contains the important variables and drops the least important variable.

1. **Cross- Validation Technique :-**

Cross-validation is a technique for validating the model efficiency by training it on the subset of input data and testing on previously unseen subset of the input data. We can also say that it is a technique to check how a statistical model generalizes to an independent dataset*.* In machine learning, there is always the need to test the stability of the model. It means based only on the training dataset; we can't fit our model on the training dataset. For this purpose, we reserve a particular sample of the dataset, which was not part of the training dataset. After that, we test our model on that sample before deployment, and this complete process comes under cross-validation. This is something different from the general train-test split.

1. **Feature Selection Technique :-**

Feature selection is a way of selecting the subset of the most relevant features from the original features set by removing the redundant, irrelevant, or noisy features. While developing the machine learning model, only a few variables in the dataset are useful for building the model, and the rest features are either redundant or irrelevant. If we input the dataset with all these redundant and irrelevant features, it may negatively impact and reduce the overall performance and accuracy of the model. Hence it is very important to identify and select the most appropriate features from the data and remove the irrelevant or less important features, which is done with the help of feature selection in machine learning. Below are some benefits of using feature selection in machine learning:

It helps in avoiding the curse of dimensionality. It helps in the simplification of the model so that it can be easily interpreted by the researchers. It reduces the training time. It reduces overfitting hence enhance the generalization.

Two methods of Feature Selection are used :-

* **Mutual Information Classification :-**

It checks the dependency between the two variables and if the two variables are independent then there correlation is cosiderered as zero. Higher the values of the features , more is the Dependency between the variables. Mutual Information always gives either Positive Value for Dependency or Zero Value for Independency.

* **Feature Importance :-**

This Technique gives the Scores of all features individually, The Higher the Score More Relevant it is. Feature Importance refers to techniques that calculate a score for all the input features for a given model — the scores simply represent the “importance” of each feature. A higher score means that the specific feature will have a larger effect on the model that is being used to predict a certain variable. Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature.

**CONCLUSION :-**

This paper aims to enhance detection method to detect phishing websites using machine learning technology. We achieved 96.58% detection accuracy using random forest algorithm with some amount of True Positive Rate. Also result shows that classifiers give better performance when we used more data as training data. In future hybrid technology will be implemented to detect phishing websites more accurately, for which random forest algorithm of machine learning technology and blacklist method will be used.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*END\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*