**Text

Description automatically generated with medium confidence**

**Faculty of Contemporary Sciences and Technologies**

**Tetovo**

**Data Mining**

**Project:**

***“Detecting Phishing Websites using Data Mining”***

**Student(s)**: **Mentor**:

Baftjar Jusufi #30228 Prof. dr. Adrian Besimi

May 2022, Tetovo

**Detecting Phishing Websites using Data Mining Techniques**

**Abstract**

Phishing is one of the serious cyber threats now, where the victims’ credentials are acquired by an illegitimate website. As the number of web users increases, phishing attacks are gradually increasing.

Phishing sites which expects to take the victims confidential data by distracting them to surf a fake website page that resembles a honest to goodness one is another type of criminal develop through the internet and its one of the especially concerns toward numerous areas including emanaging an account and retailing.

In general, Phishing is a type of extensive fraud that appears when a malicious website serves like a real one. Phishing site detection is truly an uncertain and element issue along with numerous components and criteria that are not stable.

In this type of attack, unauthentic entities disguise themselves as genuine and trustworthy entities. The users are thus, tricked by the look and feel of the fake website which is almost identical to the legitimate one. Generally, attackers handle banking and payment sites, social media sites and E-Commerce sites to fake potential victims.

A phishing website is a common social engineering method that mimics trustful uniform resource locators (URLs) and webpages.

**Introduction**

As we stated above the objective of a phishing website is gaining personal information without permission, either by blackmail or through visiting an imitation webpage that resembles the real one, which requests that the user enters personal information. This results in information security breaches through compromises in confidential data whereby the victim might suffer a financial or asset loss. The attacker may additionally commit identity theft using the personal details of victims. Also, a phishing attack can harm the reputation of the financial institution which has been spoofed, as customers lose confidence that their account is secure.

**Types of phishing Attacks**

* **Clone phishing** -attack is an attack that attracts people by creating a homepage similar to a legitimate homepage that actually exists. A type of attack that involves phishing by replicating websites that users visit frequently. these sites usually ask users for login information. The replicate website stores the user's information on the attacker's server for use in future attacks.
* **Spear phishing** -is an attack that targets employees of a specific institution or company and induces access through e-mail or other methods. It is a type of Advanced Persistent Threat (APT) attack. In order to induce a user's click, it is often disguised as a similar organization sending mail. When an email attachment is executed, an attack that leads a user to a malicious code distribution site is executed, or a malicious code is directly executed to infect the user's PC. According to TREND MICRO, 91% of targeted attacks start with spear phishing emails, and 94% of spear phishing emails are attached files.
* **URL Attack** -This is an attack that can lead to a malicious site when a user clicks on a link disguised as a normal site. Attacks involving similar domain names or attacks using technically disguised links may be involved.
* **Search Engine Phishing Attack-** It is an attack that leads a user to a phishing site by manipulating it to be ranked high when a user searches through a search engine vulnerability. An attacker creates a phishing site and allows search engines to rank phishing websites at the top. If an attacker masquerades as a normal site and provides a product that is of interest to customers, it can be registered in a search engine. Therefore, the search engine displays both the normal site and the phishing site when displaying the search results of the user. Users trust the search results of search engines, so they connect to phishing sites without a doubt. When a site is visited, a malicious program is installed, or the personal information is provided to the phishing site through the membership registration process.
* **Drive-by-download Attack** -When using the Internet through a web browser, it is an attack in which malicious code is automatically downloaded and executed without user consent by simply accessing the website.

**Objective:**

So, the goal is to develop an ideal model for determining the possibility of the website being a phishing website or a legitimate one using data mining techniques.

We will start with the code we developed and explain step by step.

## ****Loading Data:****

The features are extracted and store in the csv file.

The reulted csv file is uploaded to this notebook and stored in the dataframe.

Text

Description automatically generated

## ****Familiarizing with Data****

In this step, few dataframe methods are used to look into the data and its features.

Text

Description automatically generated

So basically. shape shows us the shape of the dataset, such as 1354 columns and, 32 rows.

.info () tell us the information about our dataset if any entries are with NULL values, it counts them and the data type which is int64.

## ****Visualizing the data****

Few plots and graphs are displayed to find how the data is distributed and the how features are related to each other.

Graphical user interface, text

Description automatically generated

Scatter chart

Description automatically generatedAnd the result of these plots are: Diagram, engineering drawing

Description automatically generated

Text

Description automatically generatedHere we have the K’means model where we check how many clusters we have.

We first give the category of clustering, the maximum cycles of clustering and we find the data standardization.

We import from a library the function Kmeans , where it can make the k means model alone. And then we print the result where it shows the first and last columns with the result of clusters.

## ****Data Preprocessing****

Here, we clean the data by applying data preprocesssing techniques and transform the data to use it in the models.

Text

Description automatically generated

.describe() shows us the description of the dataset , as shown in the result it shows the count values of all columns, the mean ,std, min ,25% quartile,50,75 and maximum value of the columns.

Later we check for any empty entries with the isnull() function.

And as we can see most of the data is made of 0’s and 1’s so we can remove the column index because it is irrelevant for the models we are going to use later.

We do this with the .drop ([“which column we want”]) .

And we do the shuffling the rows , the reason for this is that when we split the data later this shuffle method helps that the dataset will be equally distributed to the train and test set.

## ****Splitting the Data****

Text

Description automatically generated

So, we assign 2 variables such as X and Y to our target column that is “Result” , we drop the Result area so we can manipulate.

We do the split and 80/20 split and the result is:



## ****Machine Learning Models & Training****

From the dataset above, it is clear that this is a supervised machine learning task. There are two major types of supervised machine learning problems, called classification and regression.

This data set comes under classification problem, as the input URL is classified as phishing (1) or legitimate (0) so if the result is closer to 0 the website it’s not phishing as closer as to 1 the website is a phishing website. The machine learning models (classification) considered to train the dataset in this notebook are:

* Decision Tree
* Random Forest
* Multilayer Perceptons
* Bagging Classifier

And then we have a comparison model that will compare all these methods result to see if we have a phishing case and if the methods are close to the same results , so if the models have similar result, we can clearly say that the website is a phishing or not a phishing one.

Text

Description automatically generated

**Importing the library needed , creating holders for storing the models result.**

**Creating a function for calling to store the results.**

## ****Decision Tree****

Decision trees are widely used models for classification and regression tasks. Essentially, they learn a hierarchy of if/else questions, leading to a decision. Learning a decision tree means learning the sequence of if/else questions that gets us to the true answer most quickly.

In the machine learning setting, these questions are called tests (not to be confused with the test set, which is the data we use to test to see how our model is). To build a tree, the algorithm searches over all possible tests and finds the one that is most informative about the target variable. So, what is the code , we instantiate the model from sklearn called decisiontreeclassifier we fit the decision tree model with our values from the dataset and we use predict to predict our entries. We check the accuracy with the accuracy score, and we print the result , and we show the plot with our results.

Text

Description automatically generated

Text

Description automatically generated**Plotting the decision tree^**

Chart

Description automatically generated

## Random Forest Classifier

**Random forests** or **random decision forests** is an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned.

Random forests for regression and classification are currently among the most widely used machine learning methods.A random forest is essentially a collection of decision trees, where each tree is slightly different from the others. The idea behind random forests is that each tree might do a relatively good job of predicting, but will likely overfit on part of the data**.**

To build a random forest model, you need to decide on the number of trees to build (the n\_estimators parameter of RandomForestRegressor or RandomForestClassifier).

We import the random forest model , we instantiate the model and fit the random forest classifier model to our x and y values, we predict the test and training set data , print and store the results.

Text

Description automatically generatedChart

Description automatically generated

### **Multilayer Perceptrons (MLPs): Deep Learning**

The field of artificial neural networks is often just called neural networks or multi-layer perceptrons after perhaps the most useful type of neural network. A perceptron is a single neuron model that was a precursor to larger neural networks.

Multilayer perceptrons (MLPs) are also known as (vanilla) feed-forward neural networks, or sometimes just neural networks. Multilayer perceptrons can be applied for both classification and regression problems.

As the previous models we instantiate the model from sklearn.neural\_network and we fit the model with the x train and y train values, we predict using the multi-layer perceptron classifier.

We compute the accuracy of the model performance we print the results, and we store the values so we can use them later in the comparison model.

Text

Description automatically generated

## Bagging Classifier

A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction.

Such a meta-estimator can typically be used as a way to reduce the variance of a black-box estimator (e.g., a decision tree), by introducing randomization into its construction procedure and then making an ensemble out of it.

Text

Description automatically generatedSo, we import the classifier , and we instantiate the model, we fit the bagging classifier and check the test and training data and show the results. We store the results also.

Text

Description automatically generated

So, as we can see from the console , the results of all the models are similar close to number 1 so the website is considered as a phishing one based on these models that we used.

The results are also pretty close so we can say for sure this is the correct result.

Also, we have the Comparison model that we compare all these methods between them.

Text

Description automatically generatedA dataframe is created to store the results of the models.

And we print the stored results we did before.

And the printed version of this code is:

Text

Description automatically generated

So, this is the result of all these models we used.

**References:**

https://www.tutorialspoint.com/data\_mining/dm\_applications\_trends.htm#:~:text=Trends%20in%20Data%20Mining&text=Application%20Exploration.,of%20data%20mining%20query%20language.

https://programs.wiki/wiki/python-data-analysis-data-mining-modeling-cluster-analysis.html

https://vitalflux.com/bagging-classifier-python-code-example/#:~:text=Bagging%20classifier%20is%20an%20ensemble,random%20features%20are%20drawn)%2C%20random

https://www.section.io/engineering-education/implementing-bagging-algorithms-in-python/

https://en.wikipedia.org/wiki/Autoencoder

https://www.geeksforgeeks.org/data-mining-cluster-analysis/

https://github.com/VitaminDhruv/Detecting-Phishing-websites/blob/master/phising.py

https://www.researchgate.net/publication/282408351\_Phishing\_Websites\_Detection\_Using\_Data\_Mining\_Classification\_Model

https://www.youtube.com/watch?v=qrHhO9E4lIs

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html

https://github.com/shreyagopal/t81\_558\_deep\_learning/blob/master/t81\_558\_class\_14\_03\_anomaly.ipynb

https://machinelearningknowledge.ai/decision-tree-classifier-in-python-sklearn-with-example/#:~:text=Decision%20tree%20is%20a%20type,internal%20nodes%2C%20and%20leaf%20nodes.