**ANALYSIS REPORT FOR THE LOG DATA ANALYSIS**

**INTRODUCTION**

The Web serves as a convenient portal between end-users and company resources such as user accounts. A Web application is a computer program that allows end-users to interact with a Web server by sending and receiving data, possibly accessing a database or other system resources, via a browser. Such applications that facilitate a rich online experience (personal financing, social networking, business transactions, et cetera) are burdened with securing the privacy of sensitive data while permitting data exchange. A Web application accepting input is vulnerable to the class of attacks known as injection attacks, which include SQL Injection and Cross-Site Scripting attacks.

**DATA DESCRIPTION.**

**Netsparker, w3af and Acunetix data file.**

**INTRODUCTION TO THE DATASETS**

The data is about detection of vulnerabilities, XSS and SQL injection attacks on the requests being made on the web browser.

Our data sets consist of 7313rows and 2colums for netsparker,3995 rows and 3columns for w3af and in text format. After data cleaning I happen to filter out 1o new columns which include the following

**IP** : this includes the user’s ip address used when accesing the website.

**Date**: this column contains the dates when the user access the website.

**Domain:** this the user’s domain name/host names.

**Method**: this contains the http methods which was used for example GET, POST when accessing or uploading data to the web.

**url:** this is basically the webs url link to the data being accessed.

**HttpProto**: this contains the web’s prototypes to access the information.

**TimeTaken:** this contains the time taken by the request to be replied from when it was submitted to when areplied was received.

**UrlRef:** this contains the web status error codes.

**UserAgent**: this contains the agent the user used to access the web.

**Heade**r: this contains the request’s headers from the web.

**DATA ANALYSIS AND VISUALISATION**

**Graph1**

I plotted the first graph which indicates the variations of error massage replies in the netsparker log data file.

Its visually seen that the 200 error message has the highest number of requests meaning that most of the requests submitted where successfully replied without any attack.

The 404 massage comes next with the second highest replies which indicates that the request was unsuccessful due client error side which may be caused by an attack between the client/user and the web servers and this indicates some kind of vulnerabilities in the web browsers like traffic, denial of service attacks etc.

The other remaining visuals for 302,405,500 they also stand for unsuccessful access to the information which may be caused by a number of attacks, server errors, client errors and others.

But from the graph its clearly seen that the sum of unsuccessful massages is higher than the successful ones hence indication of vulnerabilities and attacks in between the client and the web servers and this needs to be worked on.

**Graph 2**

This graph indicates the categories of code class from the web data files which includes success(2xx) with the highest bar visual indicating that most requests were successfully processed, client errors(4xx), redirect(3xx), server errors(5xx) which indicates unsuccessful processing of request from the client.

As a result, this indicates also presences of vulnerabilities of web browsers since there requests which were not successfully processed due to un known reasons.

**MODEL BUILDING AND THE RESULTS**

Here I started with dropping the un necessary columns from the data Frame for better plotting, and then then I converted the string data columns into numerical data types

Which can be easily clustered using unsupervised machine learning model kmeans

Elbow method

Since iwas going to use the kmeans model it’s a necessity to first determine the the distribution of data to be used which helps to find the value of optimal k .

From the graph plotted its visually indicated that the values of optimal k can be 2 or 3 but choose to take 3 since its more optimal can best determine the number of clusters in the data Frame which is three.

And for the w3af data file took k=4 as the optimal value of clusters

**KMEANS MODEL**

From the plot, its shows the clear distribution of clusters which shows the variations between the data frame values. the results show that the values are clearly clustered with high concertation at the lower values.

**ACCURRENCY OF THE MODEL**

The model was a100% accurate since it alternated all the rows in the data frame and every value was plotted and I did this by matching the means plot labels.

**RESULTS**

The results from the visualizations shows presence of attacks in the data files ,this is seen from the various error messages being sent back to the clients which includes the redirects, client errors, server errors and hence the websites should be revised for better services and reducing on the number of attack which causes some of the failures to process the requests.

**DATA ANALYTICS DASHBOARD**

For the dashboard I used Django environment for the development.

On the menu part I included the log file, data analysis, visualization, models, and traffic detection links. And the file is included in the zipped file.

**DOCUMENTING THE CODE**

For the documentation I tried to use comments in the notebook for better understanding of the code for every piece of code in the notebook.