Contents

[1.0 Introduction. 2](#_Toc43655417)

[2.0 Data description 2](#_Toc43655418)

[3.0 Data Analysis 3](#_Toc43655419)

[3.1 HTTP methods in Acunetix 3](#_Toc43655420)

[3.2 HTTP methods in Netsparker 3](#_Toc43655421)

[3.3 HTTP methods in W3AF 4](#_Toc43655422)

[3.4 HTTP status codes in Acunetix. 4](#_Toc43655423)

[3.5 HTTP status codes in Netsparker 4](#_Toc43655424)

[3.6 HTTP status codes in W3AF. 5](#_Toc43655425)

[4.0 RESULTS. 6](#_Toc43655426)

[5.0 CONCLUSION 6](#_Toc43655427)

# 1.0 Introduction.

This report provides an analysis of web server access logs and the methods of analysis include the usage of programming modules such as pandas, pyspark, ua-parser, and arg-parser.

# 2.0 Data description

The data contains 3 datasets collected from web vulnerability scanners, for example, Acunetix, Netsparker, Web application attack and audit framework.

An APACHE ACCESS LOG PATTERN was used to obtain fields from the logs for example userid, host, timestamp, response code, protocols, HTTP methods etc.

The log files that we use for this assignment are in the [Apache Common Log Format (CLF)](http://httpd.apache.org/docs/1.3/logs.html#common). The log file entries produced in CLF will look something like this: 127.0.0.1 - - [01/Aug/1995:00:00:01 -0400] "GET /images/launch-logo.gif HTTP/1.0" 200 1839

Each part of log entry:

* 127.0.0.1 This is the IP address (or hostname, if available) of the client (remote host) which made the request to the server.
* - The "hyphen" in the output indicates that the requested piece of information (user identity from a remote machine) is not available.
* - The "hyphen" in the output indicates that the requested piece of information (user identity from local logon) is not available.
* [01/Aug/1995:00:00:01 -0400] The time that the server finished processing the request. The format is: [day/month/year:hour:minute:second time zone]
* "GET /images/launch-logo.gif HTTP/1.0" This is the first line of the request string from the client. It consists of three components: the request method (e.g., GET, POST, etc.), the endpoint (a [Uniform Resource Identifier](http://en.wikipedia.org/wiki/Uniform_resource_identifier)), and the client protocol version.
* 200 This is the status code that the server sends back to the client. This information is very valuable, because it reveals whether the request resulted in a successful response (codes beginning in 2), a redirection (codes beginning in 3), an error caused by the client (codes beginning in 4), or an error in the server (codes beginning in 5). The full list of possible status codes can be found in the HTTP specification ([RFC 2616](https://www.ietf.org/rfc/rfc2616.txt) section 10).
* 1839 The last entry indicates the size of the object returned to the client, not including the response headers. If no content was returned to the client, this value will be "-" (or sometimes 0)

|  |  |
| --- | --- |
| **Web Vulnerability Scanner** | **Data size** |
| Acunetix | 6539 |
| Netsparker | 7314 |
| Web application attack and audit framework | 3996, |

# 3.0 Data Analysis

## 3.1 HTTP methods in Acunetix

|  |  |
| --- | --- |
| **HTTP Method** | **Number** |
| POST | 5499 |
| GET | 1039 |
| **Total** | **6,538** |

## 3.2 HTTP methods in Netsparker

|  |  |
| --- | --- |
| **HTTP Method** | **Number** |
| GET | 4004 |
| HEAD | 2038 |
| POST | 1203 |
| OPTIONS | 34 |
| PROPFIND | 34 |
| NETSPARKER | 1 |
| **Total** | **7,314** |

## 3.3 HTTP methods in W3AF

|  |  |
| --- | --- |
| **HTTP Method** | **Number** |
| POST | 2656 |
| GET | 1340 |
| **Total** | **3,996** |

## 3.4 HTTP status codes in Acunetix.

|  |  |
| --- | --- |
| **HTTP status code** | **Number** |
| 303 | 3407 |
| 200 | 2036 |
| 500 | 706 |
| 404 | 258 |
| 302 | 99 |
| 301 | 18 |
| 400 | 12 |
| 403 | 3 |
| **Total** | **1,096** |

## 3.5 HTTP status codes in Netsparker

|  |  |
| --- | --- |
| **HTTP status code** | **Number** |
| 200 | 3784 |
| 500 | 6 |
| 404 | 2602 |
| 302 | 910 |
| 405 | 12 |
| **Total** | **7,314** |

## 3.6 HTTP status codes in W3AF.

|  |  |
| --- | --- |
| **HTTP status code** | **Number** |
| 200 | 2,124 |
| 500 | 513 |
| 404 | 490 |
| 302 | 108 |
| 301 | 77 |
| 403 | 2 |
| 304 | 1 |
| 303 | 681 |
| **Total** | **3,506** |

# 4.0 RESULTS.

Using the datasets provided training data was obtained and a model was created for each dataset testing for the three types of users.

Models tested to deduce results for all types of users, confusion matrices, precision-recall graphs and roc curves were generated to evaluate the performance of the models.

Within the Notebooks, all information was displayed

machine-learning/netsparker.ipynb,

/machine learning/w3af.ipynb,

/machine learning/acune\_dataset.ipynb

# 5.0 CONCLUSION

Datasets provided where small and could not facilitate proper machine learning results.