Machine Learning for Cyber Security

7st SEMESTER

ACADEMIC YEAR 2023 (Aug-Nov)



SCHOOL OF COMPUTING SCIENCE & ENGINEERING

VIT BHOPAL UNIVERSITY

**Group Activity**

1. Akshay Chouke – 20MEI10044
2. Dipesh Dwivedi - 20MEI10046
3. Ariyan Sahoo - 20MEI10054
4. **Domain -** Phishing Detection
5. **Dataset Selection:**
   1. **Dataset 1** **–** This dataset is downloaded from the Kaggle platform. It has 2488 instances and 14 features. It contains information about the URL such as URL length, counts of @,.,-,& etc symbols.
   2. **Dataset 2 –** The provided dataset is downloaded from the Kaggle platform. It includes 11430 URLs (Instances) with 89 extracted features. The dataset is designed to be used as a benchmark for machine learning-based phishing detection systems. Features are from three different classes: 56 extracted from the structure and syntax of URLs, 24 extracted from the content of their correspondent pages, and 9 are extracted by querying external services.
   3. **Dataset 3 –** The provided dataset is collected from the research paper named as **Datasets for phishing websites detection** for the purpose of building and evaluating various classification methods for the task of detecting phishing websites based on the uniform resource locator (URL) properties, URL resolving metrics, and external services. It has 88647 instances and 111 features.
6. **Dataset Features Description:**
   1. Dataset 1 –

|  |  |  |  |
| --- | --- | --- | --- |
| **Seq. No.** | **Attribute Name** | **Description** | **Data Type** |
| 1 | url\_length | Length of the URL | Ratio |
| 2 | valid\_url | Whether the URL is valid or not | Nominal |
| 3 | at\_symbol | Presence of the "@" symbol in the URL | Nominal |
| 4 | sensitive\_words\_count | Count of sensitive words in the URL | Ratio |
| 5 | path\_length | Length of the path in the URL | Ratio |
| 6 | isHttps | Indicates if the URL uses HTTPS or not | Nominal |
| 7 | nb\_dots | Number of dots (periods) in the URL | Ratio |
| 8 | nb\_hyphens | Number of hyphens (dashes) in the URL | Ratio |
| 9 | nb\_and | Number of occurrences of "and" in the URL | Ratio |
| 10 | nb\_or | Number of occurrences of "or" in the URL | Ratio |
| 11 | nb\_www | Number of occurrences of "www" in the URL | Ratio |
| 12 | nb\_com | Number of occurrences of "com" in the URL | Ratio |
| 13 | nb\_underscore | Number of underscores in the URL | Ratio |
| 14 | target | Target variable (classification label) | Nominal |

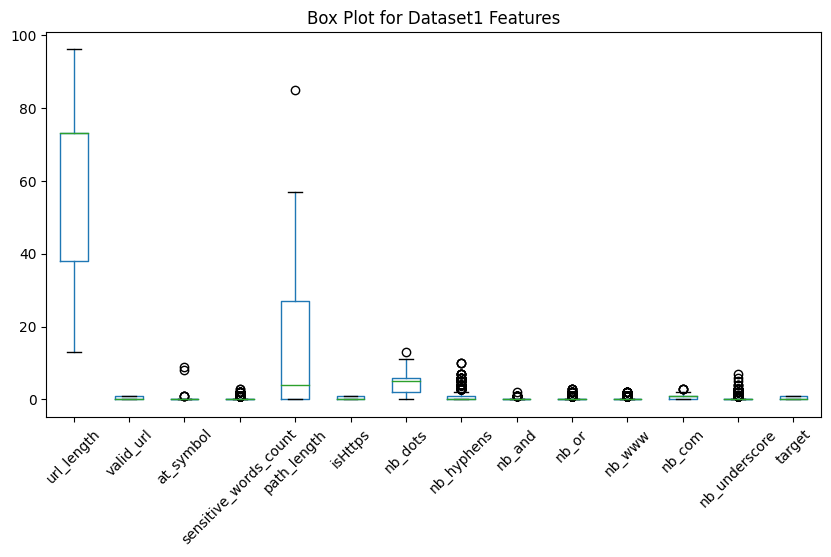
* 1. Dataset 2 –

|  |  |  |  |
| --- | --- | --- | --- |
| **Seq. No.** | **Attribute Name** | **Description** | **Data Type** |
| 1 | url | URL | Nominal |
| 2 | length\_url | Length of the URL | Ratio |
| 3 | length\_hostname | Length of the hostname in the URL | Ratio |
| 4 | ip | IP Address in the URL | Nominal |
| 5 | nb\_dots | Number of dots in the URL | Ratio |
| 6 | nb\_hyphens | Number of hyphens in the URL | Ratio |
| 7 | nb\_at | Number of "@" symbols in the URL | Ratio |
| 8 | nb\_qm | Number of question marks in the URL | Ratio |
| 9 | nb\_and | Number of occurrences of "and" in the URL | Ratio |
| 10 | nb\_or | Number of occurrences of "or" in the URL | Ratio |
| 11 | nb\_eq | Number of equal signs in the URL | Ratio |
| 12 | nb\_underscore | Number of underscores in the URL | Ratio |
| 13 | nb\_tilde | Number of tildes in the URL | Ratio |
| 14 | nb\_percent | Number of percent signs in the URL | Ratio |
| 15 | nb\_slash | Number of slashes in the URL | Ratio |
| 16 | nb\_star | Number of asterisks in the URL | Ratio |
| 17 | nb\_colon | Number of colons in the URL | Ratio |
| 18 | nb\_comma | Number of commas in the URL | Ratio |
| 19 | nb\_semicolon | Number of semicolons in the URL | Ratio |
| 20 | nb\_dollar | Number of dollar signs in the URL | Ratio |
| 21 | nb\_space | Number of spaces in the URL | Ratio |
| 22 | nb\_www | Number of occurrences of "www" in the URL | Ratio |
| 23 | nb\_com | Number of occurrences of "com" in the URL | Ratio |
| 24 | nb\_dslash | Number of double slashes in the URL | Ratio |
| 25 | http\_in\_path | Presence of "http" in the URL path | Nominal |
| 26 | https\_token | Presence of "https" token in the URL | Nominal |
| 27 | ratio\_digits\_url | Ratio of digits in the URL | Ratio |
| 28 | ratio\_digits\_host | Ratio of digits in the hostname | Ratio |
| 29 | punycode | Presence of Punycode in the URL | Nominal |
| 30 | port | Port number in the URL | Nominal |
| 31 | tld\_in\_path | Presence of TLD in the URL path | Nominal |
| 32 | tld\_in\_subdomain | Presence of TLD in the subdomain | Nominal |
| 33 | abnormal\_subdomain | Presence of abnormal subdomain | Nominal |
| 34 | nb\_subdomains | Number of subdomains in the URL | Ratio |
| 35 | prefix\_suffix | Presence of prefix or suffix in the URL | Nominal |
| 36 | random\_domain | Presence of random domain in the URL | Nominal |
| 37 | shortening\_service | Presence of URL shortening service | Nominal |
| 38 | path\_extension | Presence of file path extension | Nominal |
| 39 | nb\_redirection | Number of URL redirections | Ratio |
| 40 | nb\_external\_redirection | Number of external URL redirections | Ratio |
| 41 | length\_words\_raw | Length of words in the URL | Ratio |
| 42 | char\_repeat | Presence of repeated characters in the URL | Nominal |
| 43 | shortest\_words\_raw | Length of the shortest word in the URL | Ratio |
| 44 | shortest\_word\_host | Length of the shortest word in the hostname | Ratio |
| 45 | shortest\_word\_path | Length of the shortest word in the path | Ratio |
| 46 | longest\_words\_raw | Length of the longest word in the URL | Ratio |
| 47 | longest\_word\_host | Length of the longest word in the hostname | Ratio |
| 48 | longest\_word\_path | Length of the longest word in the path | Ratio |
| 49 | avg\_words\_raw | Average length of words in the URL | Ratio |
| 50 | avg\_word\_host | Average length of words in the hostname | Ratio |
| 51 | avg\_word\_path | Average length of words in the path | Ratio |
| 52 | phish\_hints | Presence of phishing hints in the URL | Nominal |
| 53 | domain\_in\_brand | Presence of domain in a brand | Nominal |
| 54 | brand\_in\_subdomain | Presence of brand in the subdomain | Nominal |
| 55 | brand\_in\_path | Presence of brand in the path | Nominal |
| 56 | suspecious\_tld | Presence of suspicious TLD in the URL | Nominal |
| 57 | statistical\_report | Presence of a statistical report in the URL | Nominal |
| 58 | nb\_hyperlinks | Number of hyperlinks in the URL | Ratio |
| 59 | ratio\_intHyperlinks | Ratio of internal hyperlinks | Ratio |
| 60 | ratio\_extHyperlinks | Ratio of external hyperlinks | Ratio |
| 61 | ratio\_nullHyperlinks | Ratio of null hyperlinks | Ratio |
| 62 | nb\_extCSS | Number of external CSS references | Ratio |
| 63 | ratio\_intRedirection | Ratio of internal URL redirections | Ratio |
| 64 | ratio\_extRedirection | Ratio of external URL redirections | Ratio |
| 65 | ratio\_intErrors | Ratio of internal URL errors | Ratio |
| 66 | ratio\_extErrors | Ratio of external URL errors | Ratio |
| 67 | login\_form | Presence of a login form in the URL | Nominal |
| 68 | external\_favicon | Presence of an external favicon | Nominal |
| 69 | links\_in\_tags | Presence of links in HTML tags | Nominal |
| 70 | submit\_email | Presence of email submission form | Nominal |
| 71 | ratio\_intMedia | Ratio of internal media references | Ratio |
| 72 | ratio\_extMedia | Ratio of external media references | Ratio |
| 73 | sfh | Server Form Handler (SFH) | Nominal |
| 74 | iframe | Presence of iframes in the URL | Nominal |
| 75 | popup\_window | Presence of popup windows in the URL | Nominal |
| 76 | safe\_anchor | Presence of safe anchors in the URL | Nominal |
| 77 | onmouseover | Presence of onmouseover attributes in the URL | Nominal |
| 78 | right\_clic | Presence of right-click protection in the URL | Nominal |
| 79 | empty\_title | Presence of empty title in the URL | Nominal |
| 80 | domain\_in\_title | Presence of the domain in the title | Nominal |
| 81 | domain\_with\_copyright | Presence of domain with copyright symbol | Nominal |
| 82 | whois\_registered\_domain | Domain registered in WHOIS | Nominal |
| 83 | domain\_registration\_length | Domain registration length in days | Interval |
| 84 | domain\_age | Age of the domain in days | Interval |
| 85 | web\_traffic | Web traffic rank | Ordinal |
| 86 | dns\_record | Presence of DNS records | Nominal |
| 87 | google\_index | Google index status of the URL | Nominal |
| 88 | page\_rank | PageRank score of the URL | Interval |
| 89 | target | Target variable (classification label) | Nominal |

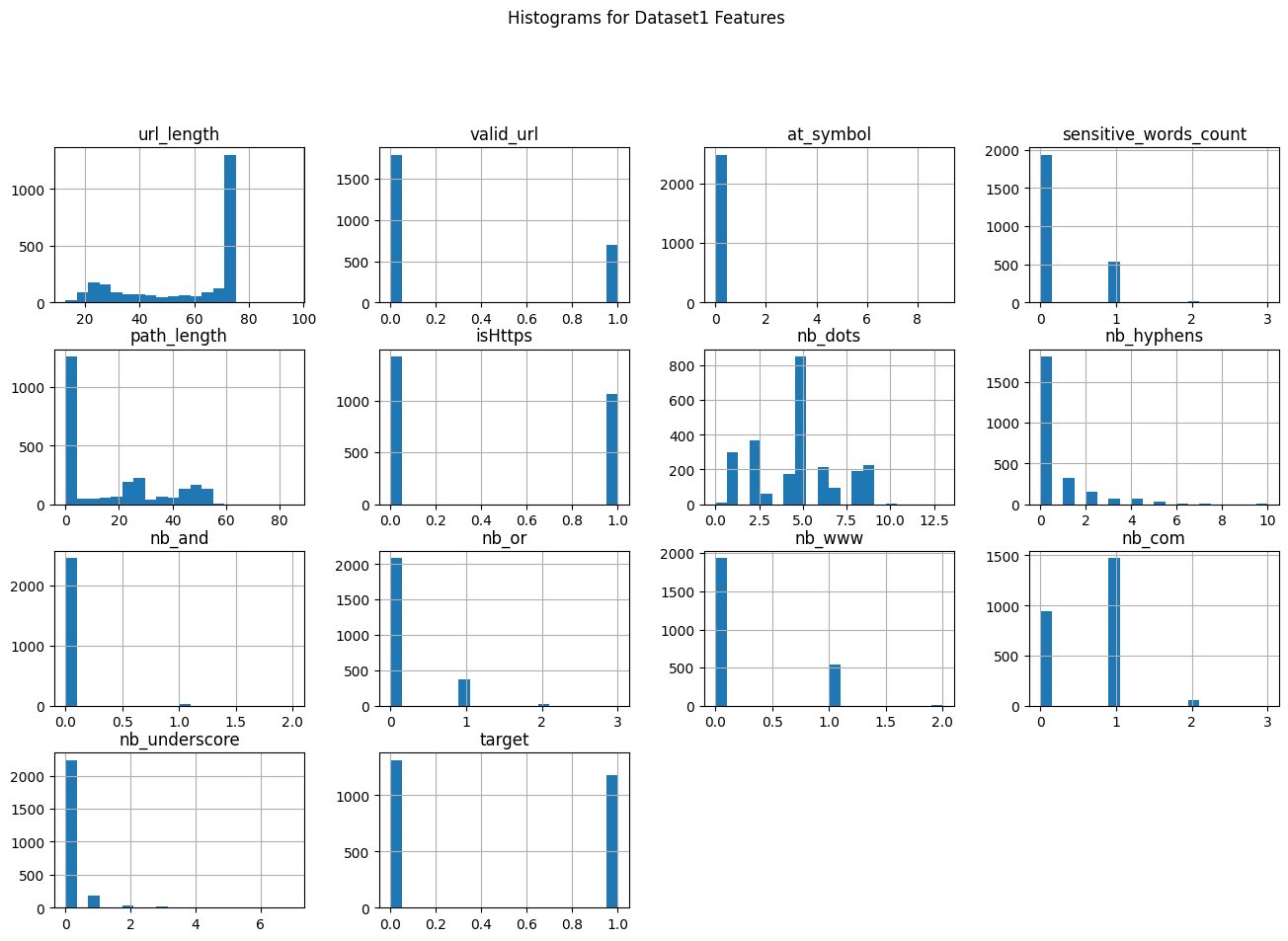
* 1. Dataset 3 –

|  |  |  |  |
| --- | --- | --- | --- |
| **Seq. No.** | **Attribute Name** | **Description** | **Data Type** |
| 1 | qty\_dot\_url | Quantity of dots in the URL | Ratio |
| 2 | qty\_hyphen\_url | Quantity of hyphens in the URL | Ratio |
| 3 | qty\_underline\_url | Quantity of underscores in the URL | Ratio |
| 4 | qty\_slash\_url | Quantity of slashes in the URL | Ratio |
| 5 | qty\_questionmark\_url | Quantity of question marks in the URL | Ratio |
| 6 | qty\_equal\_url | Quantity of equal signs in the URL | Ratio |
| 7 | qty\_at\_url | Quantity of "@" symbols in the URL | Ratio |
| 8 | qty\_and\_url | Quantity of "and" in the URL | Ratio |
| 9 | qty\_exclamation\_url | Quantity of exclamation marks in the URL | Ratio |
| 10 | qty\_space\_url | Quantity of spaces in the URL | Ratio |
| 11 | qty\_tilde\_url | Quantity of tildes in the URL | Ratio |
| 12 | qty\_comma\_url | Quantity of commas in the URL | Ratio |
| 13 | qty\_plus\_url | Quantity of plus signs in the URL | Ratio |
| 14 | qty\_asterisk\_url | Quantity of asterisks in the URL | Ratio |
| 15 | qty\_hashtag\_url | Quantity of hashtags in the URL | Ratio |
| 16 | qty\_dollar\_url | Quantity of dollar signs in the URL | Ratio |
| 17 | qty\_percent\_url | Quantity of percent signs in the URL | Ratio |
| 18 | length\_url | Length of the URL in characters | Ratio |
| 19 | qty\_dot\_domain | Quantity of dots in the domain name | Ratio |
| 20 | qty\_hyphen\_domain | Quantity of hyphens in the domain name | Ratio |
| 21 | qty\_underline\_domain | Quantity of underscores in the domain name | Ratio |
| 22 | qty\_slash\_domain | Quantity of slashes in the domain name | Ratio |
| 23 | qty\_questionmark\_domain | Quantity of question marks in the domain name | Ratio |
| 24 | qty\_equal\_domain | Quantity of equal signs in the domain name | Ratio |
| 25 | qty\_at\_domain | Quantity of "@" symbols in the domain name | Ratio |
| 26 | qty\_and\_domain | Quantity of "and" in the domain name | Ratio |
| 27 | qty\_exclamation\_domain | Quantity of exclamation marks in the domain name | Ratio |
| 28 | qty\_space\_domain | Quantity of spaces in the domain name | Ratio |
| 29 | qty\_tilde\_domain | Quantity of tildes in the domain name | Ratio |
| 30 | qty\_comma\_domain | Quantity of commas in the domain name | Ratio |
| 31 | qty\_plus\_domain | Quantity of plus signs in the domain name | Ratio |
| 32 | qty\_asterisk\_domain | Quantity of asterisks in the domain name | Ratio |
| 33 | qty\_hashtag\_domain | Quantity of hashtags in the domain name | Ratio |
| 34 | qty\_dollar\_domain | Quantity of dollar signs in the domain name | Ratio |
| 35 | qty\_percent\_domain | Quantity of percent signs in the domain name | Ratio |
| 36 | qty\_vowels\_domain | Quantity of vowels in the domain name | Ratio |
| 37 | domain\_length | Length of the domain name in characters | Ratio |
| 38 | domain\_in\_ip | Domain represented as an IP address | Nominal |
| 39 | server\_client\_domain | Server or client in the domain name | Nominal |
| 40 | qty\_dot\_directory | Quantity of dots in the directory path | Ratio |
| 41 | qty\_hyphen\_directory | Quantity of hyphens in the directory path | Ratio |
| 42 | qty\_underline\_directory | Quantity of underscores in the directory path | Ratio |
| 43 | qty\_slash\_directory | Quantity of slashes in the directory path | Ratio |
| 44 | qty\_questionmark\_directory | Quantity of question marks in the directory path | Ratio |
| 45 | qty\_equal\_directory | Quantity of equal signs in the directory path | Ratio |
| 46 | qty\_at\_directory | Quantity of "@" symbols in the directory path | Ratio |
| 47 | qty\_and\_directory | Quantity of "and" in the directory path | Ratio |
| 48 | qty\_exclamation\_directory | Quantity of exclamation marks in the directory path | Ratio |
| 49 | qty\_space\_directory | Quantity of spaces in the directory path | Ratio |
| 50 | qty\_tilde\_directory | Quantity of tildes in the directory path | Ratio |
| 51 | qty\_comma\_directory | Quantity of commas in the directory path | Ratio |
| 52 | qty\_plus\_directory | Quantity of plus signs in the directory path | Ratio |
| 53 | qty\_asterisk\_directory | Quantity of asterisks in the directory path | Ratio |
| 54 | qty\_hashtag\_directory | Quantity of hashtags in the directory path | Ratio |
| 55 | qty\_dollar\_directory | Quantity of dollar signs in the directory path | Ratio |
| 56 | qty\_percent\_directory | Quantity of percent signs in the directory path | Ratio |
| 57 | directory\_length | Length of the directory path in characters | Ratio |
| 58 | qty\_dot\_file | Quantity of dots in the file name | Ratio |
| 59 | qty\_hyphen\_file | Quantity of hyphens in the file name | Ratio |
| 60 | qty\_underline\_file | Quantity of underscores in the file name | Ratio |
| 61 | qty\_slash\_file | Quantity of slashes in the file name | Ratio |
| 62 | qty\_questionmark\_file | Quantity of question marks in the file name | Ratio |
| 63 | qty\_equal\_file | Quantity of equal signs in the file name | Ratio |
| 64 | qty\_at\_file | Quantity of "@" symbols in the file name | Ratio |
| 65 | qty\_and\_file | Quantity of "and" in the file name | Ratio |
| 66 | qty\_exclamation\_file | Quantity of exclamation marks in the file name | Ratio |
| 67 | qty\_space\_file | Quantity of spaces in the file name | Ratio |
| 68 | qty\_tilde\_file | Quantity of tildes in the file name | Ratio |
| 69 | qty\_comma\_file | Quantity of commas in the file name | Ratio |
| 70 | qty\_plus\_file | Quantity of plus signs in the file name | Ratio |
| 71 | qty\_asterisk\_file | Quantity of asterisks in the file name | Ratio |
| 72 | qty\_hashtag\_file | Quantity of hashtags in the file name | Ratio |
| 73 | qty\_dollar\_file | Quantity of dollar signs in the file name | Ratio |
| 74 | qty\_percent\_file | Quantity of percent signs in the file name | Ratio |
| 75 | file\_length | Length of the file name in characters | Ratio |
| 76 | qty\_dot\_params | Quantity of dots in URL parameters | Ratio |
| 77 | qty\_hyphen\_params | Quantity of hyphens in URL parameters | Ratio |
| 78 | qty\_underline\_params | Quantity of underscores in URL parameters | Ratio |
| 79 | qty\_slash\_params | Quantity of slashes in URL parameters | Ratio |
| 80 | qty\_questionmark\_params | Quantity of question marks in URL parameters | Ratio |
| 81 | qty\_equal\_params | Quantity of equal signs in URL parameters | Ratio |
| 82 | qty\_at\_params | Quantity of "@" symbols in URL parameters | Ratio |
| 83 | qty\_and\_params | Quantity of "and" in URL parameters | Ratio |
| 84 | qty\_exclamation\_params | Quantity of exclamation marks in URL parameters | Ratio |
| 85 | qty\_space\_params | Quantity of spaces in URL parameters | Ratio |
| 86 | qty\_tilde\_params | Quantity of tildes in URL parameters | Ratio |
| 87 | qty\_comma\_params | Quantity of commas in URL parameters | Ratio |
| 88 | qty\_plus\_params | Quantity of plus signs in URL parameters | Ratio |
| 89 | qty\_asterisk\_params | Quantity of asterisks in URL parameters | Ratio |
| 90 | qty\_hashtag\_params | Quantity of hashtags in URL parameters | Ratio |
| 91 | qty\_dollar\_params | Quantity of dollar signs in URL parameters | Ratio |
| 92 | qty\_percent\_params | Quantity of percent signs in URL parameters | Ratio |
| 93 | params\_length | Length of URL parameters in characters | Ratio |
| 94 | tld\_present\_params | Presence of TLD in URL parameters | Nominal |
| 95 | qty\_params | Quantity of parameters in the URL | Ratio |
| 96 | email\_in\_url | Presence of an email address in the URL | Nominal |
| 97 | time\_response | Time taken for the server response | Interval |
| 98 | domain\_spf | Presence of SPF in the domain | Nominal |
| 99 | asn\_ip | ASN (Autonomous System Number) of the IP address | Nominal |
| 100 | time\_domain\_activation | Time of domain activation | Interval |
| 101 | time\_domain\_expiration | Time of domain expiration | Interval |
| 102 | qty\_ip\_resolved | Quantity of IP addresses resolved from the domain | Ratio |
| 103 | qty\_nameservers | Quantity of nameservers for the domain | Ratio |
| 104 | qty\_mx\_servers | Quantity of MX (Mail Exchange) servers for the domain | Ratio |
| 105 | ttl\_hostname | Time to Live (TTL) for the hostname | Interval |
| 106 | tls\_ssl\_certificate | Presence of TLS/SSL certificate | Nominal |
| 107 | qty\_redirects | Quantity of URL redirects | Ratio |
| 108 | url\_google\_index | Google index status of the URL | Nominal |
| 109 | domain\_google\_index | Google index status of the domain | Nominal |
| 110 | url\_shortened | URL shortening status | Nominal |
| 111 | target | Target variable (classification label) | Nomina |

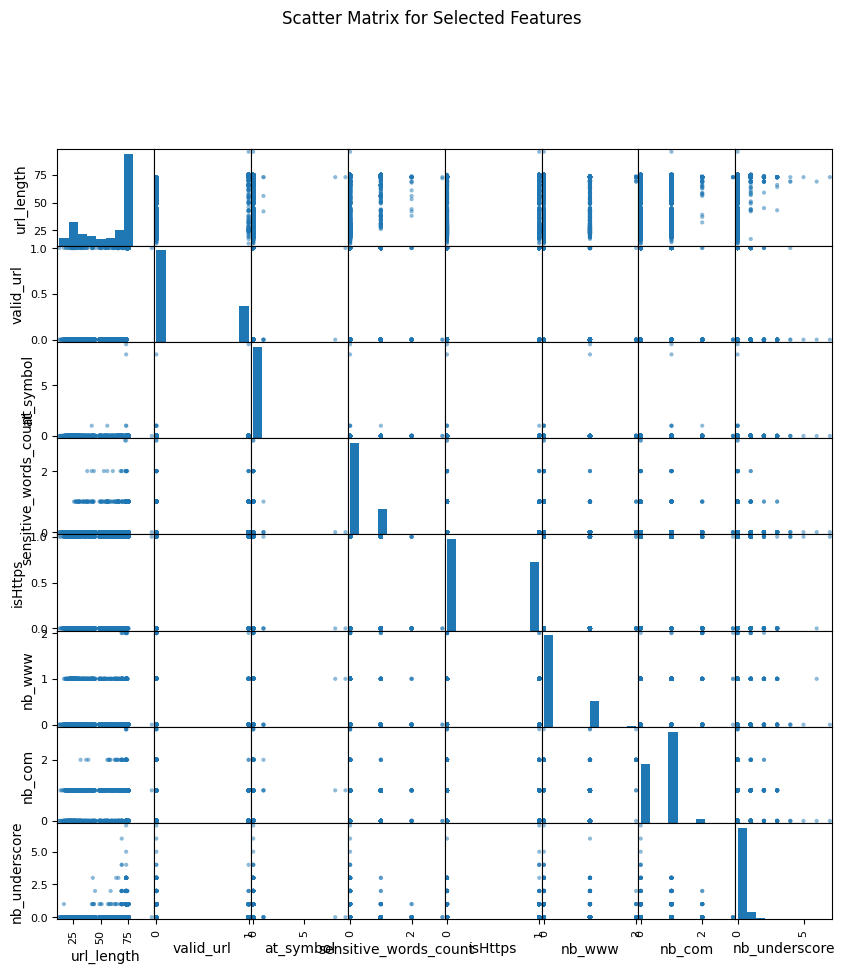
1. **Data Visualization:** 
   1. Dataset 1 –
      1. Box plots



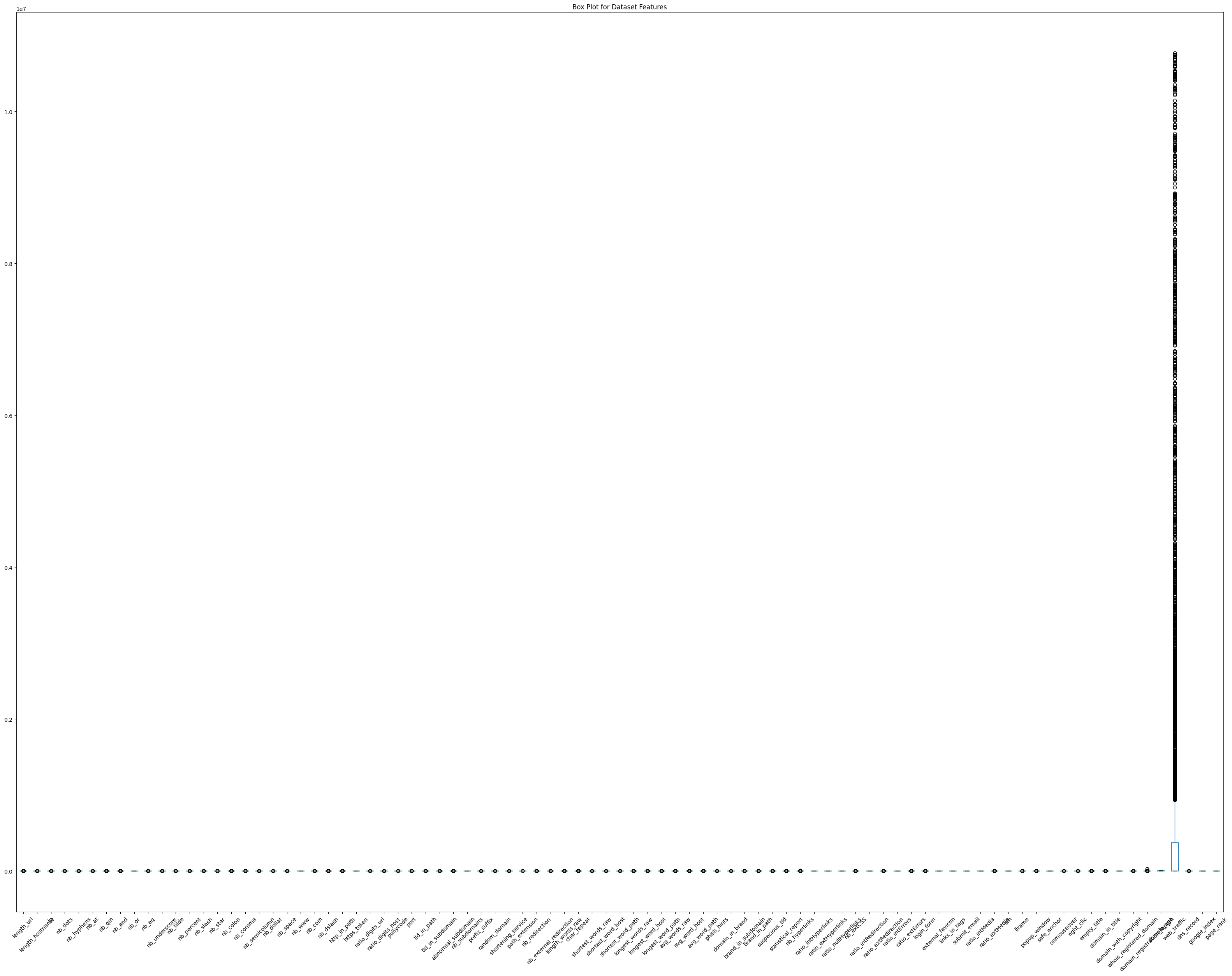
* + 1. Histograms

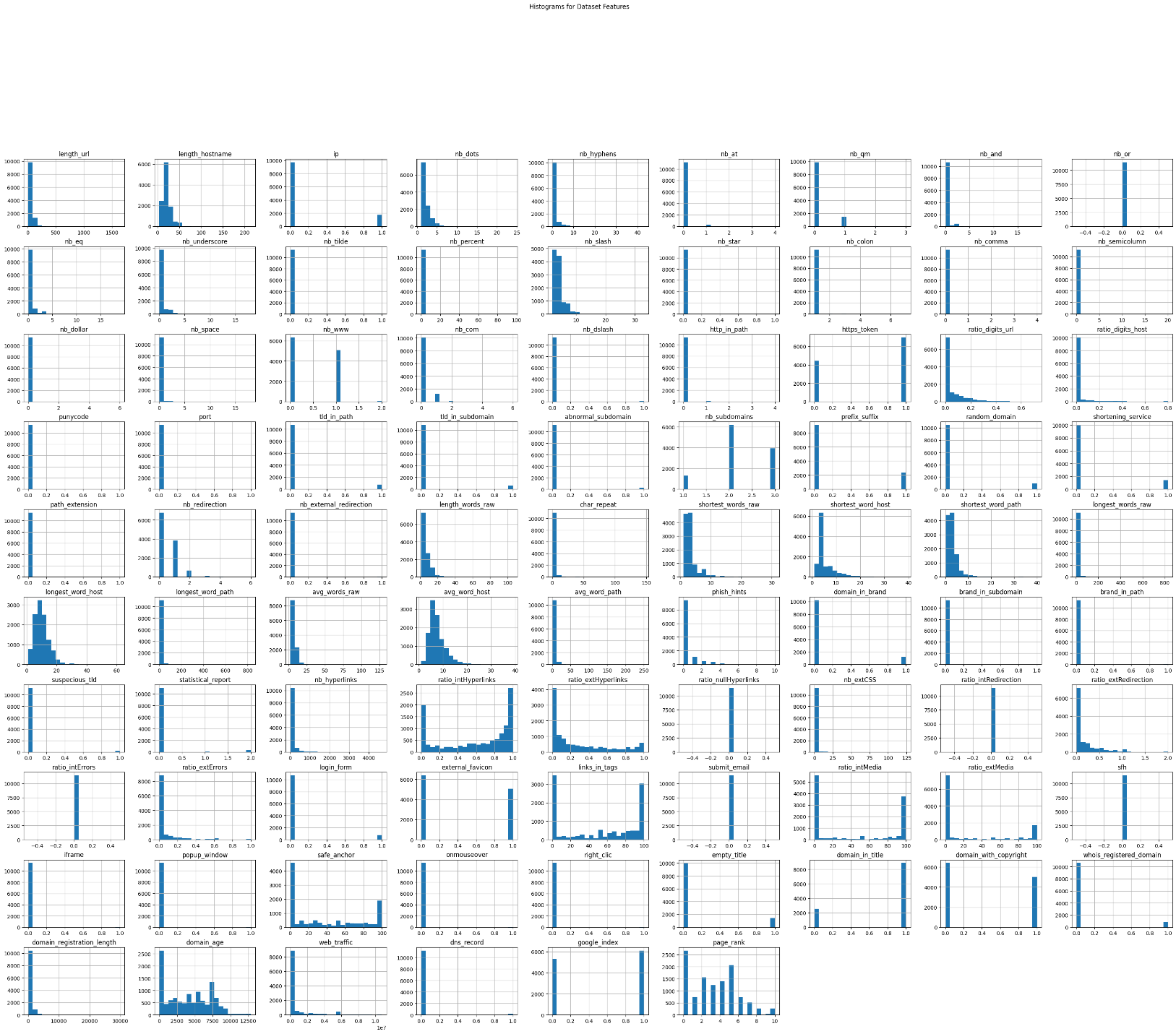


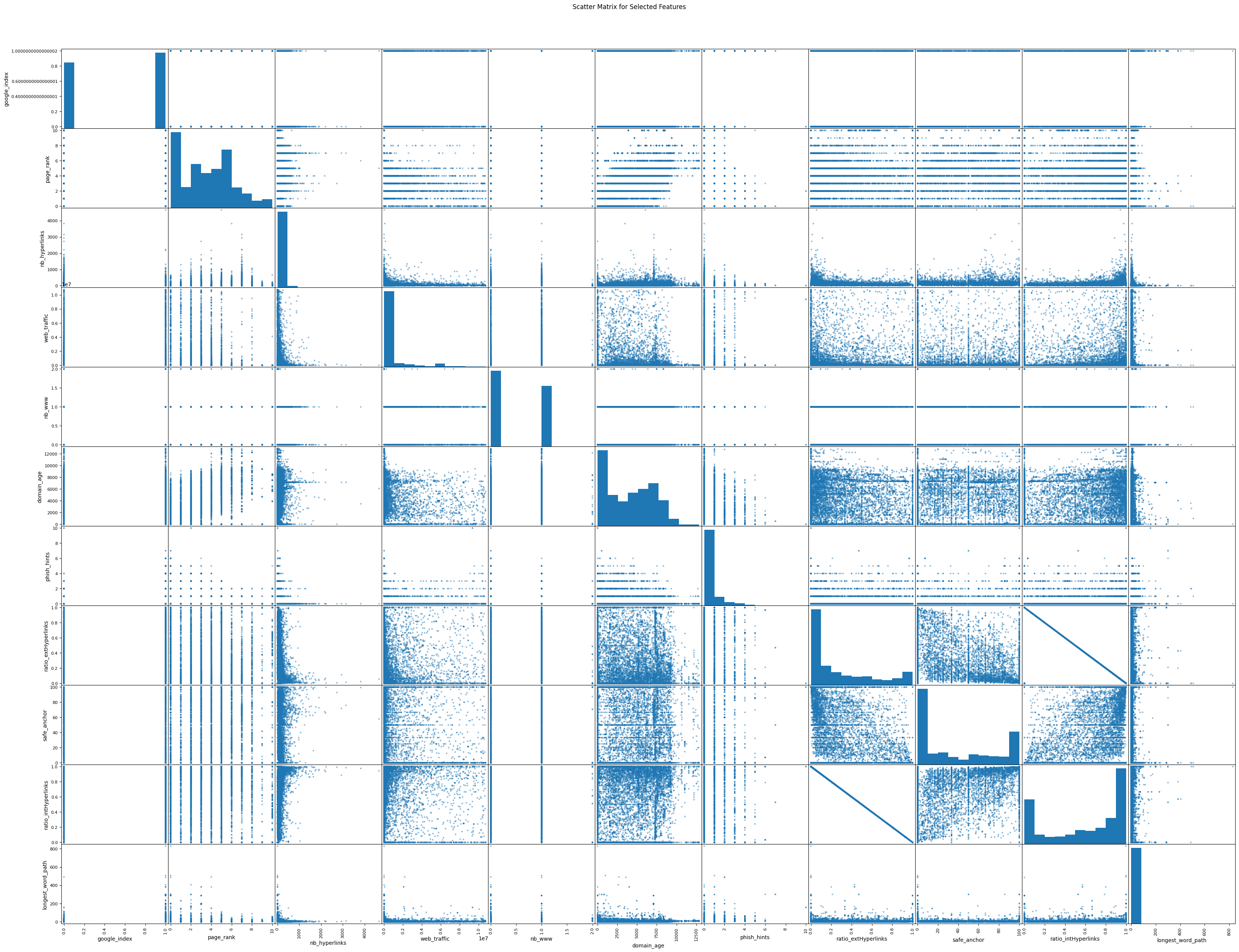
III. Scatter Matrix



* 1. Dataset 2 –
     1. Box plots

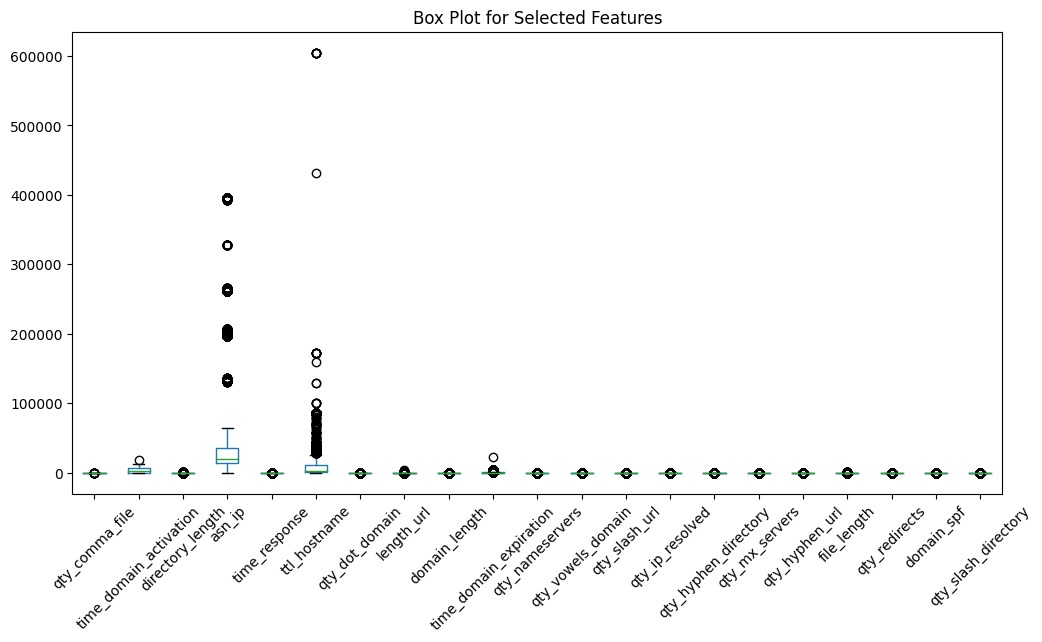


* + 1. Histograms
    2. Scatter Matrix

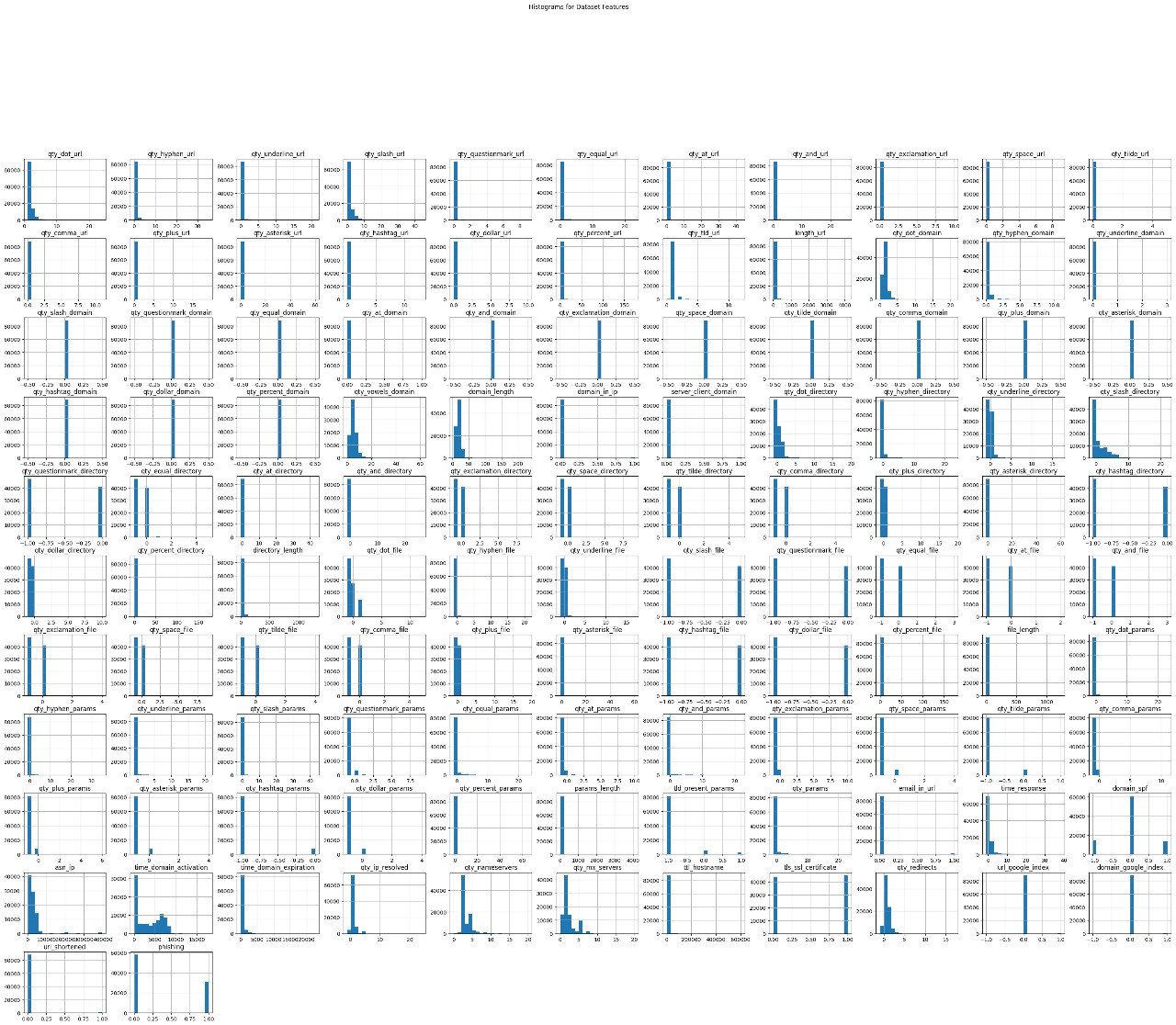


* 1. Dataset 3 –

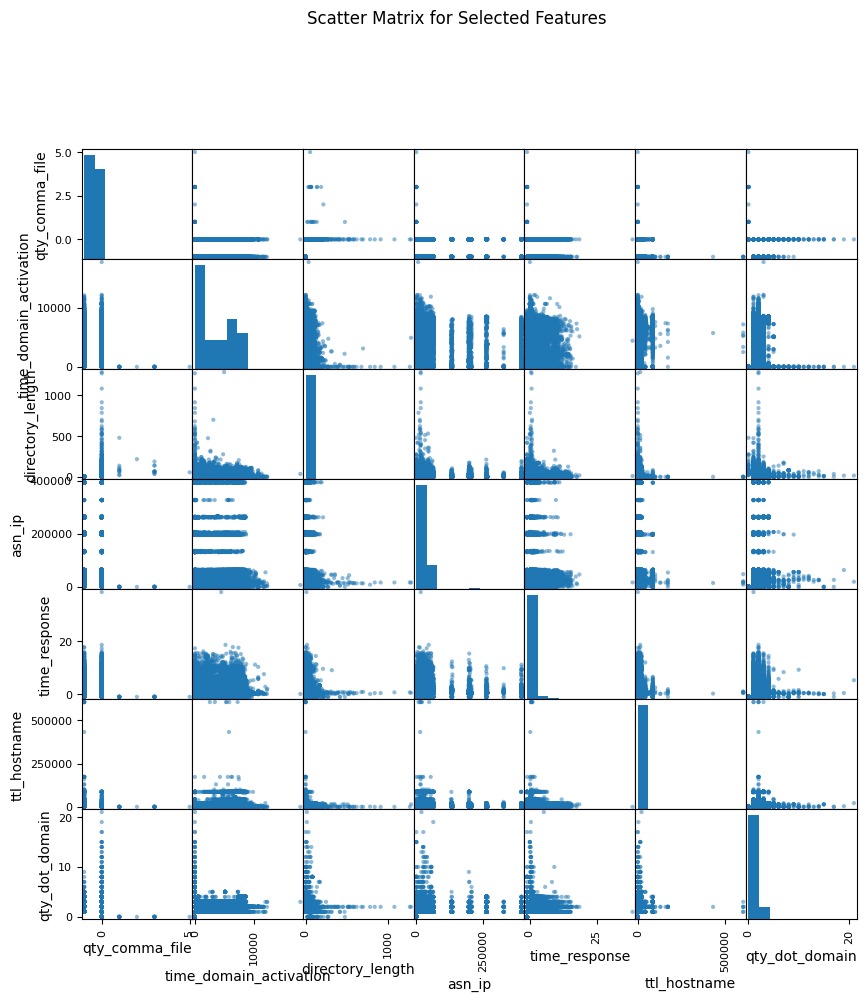
1. Box Plots-



1. Histograms



1. Scatter Matrix



1. **Feature Ranking:** 
   1. Dataset 1 –

|  |  |
| --- | --- |
| Rank | Attribute |
| 1 | valid\_url |
| 2 | url\_length |
| 3 | path\_length |
| 4 | nb\_dots |
| 5 | nb\_hyphens |
| 6 | nb\_www |
| 7 | sensitive\_words\_count |
| 8 | isHttps |
| 9 | nb\_or |
| 10 | nb\_underscore |
| 11 | nb\_com |
| 12 | nb\_and |
| 13 | at\_symbol |

* 1. Dataset 2 –

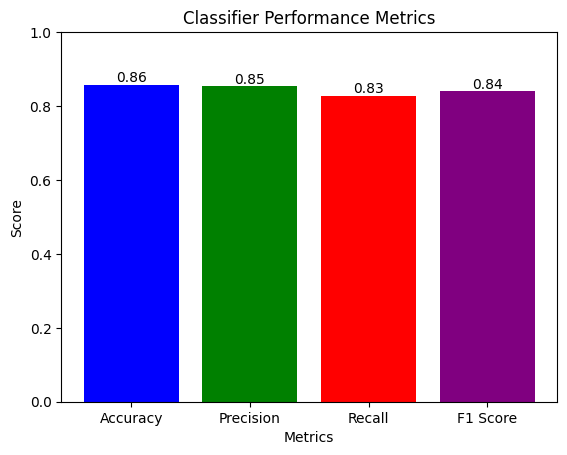
|  |  |
| --- | --- |
| Rank | Feature |
| 1 | google\_index |
| 2 | page\_rank |
| 3 | nb\_hyperlinks |
| 4 | web\_traffic |
| 5 | nb\_www |
| 6 | domain\_age |
| 7 | phish\_hints |
| 8 | ratio\_extHyperlinks |
| 9 | safe\_anchor |
| 10 | ratio\_intHyperlinks |
| 11 | longest\_word\_path |
| 12 | ratio\_extRedirection |
| 13 | ratio\_digits\_url |
| 14 | length\_url |
| 15 | length\_hostname |
| 16 | longest\_words\_raw |
| 17 | avg\_word\_path |
| 18 | domain\_in\_title |
| 19 | char\_repeat |
| 20 | length\_words\_raw |
| 21 | links\_in\_tags |
| 22 | shortest\_word\_host |
| 23 | domain\_registration\_length |
| 24 | shortest\_word\_path |
| 25 | nb\_slash |
| 26 | avg\_words\_raw |
| 27 | ratio\_digits\_host |
| 28 | nb\_dots |
| 29 | nb\_hyphens |
| 30 | avg\_word\_host |
| 31 | nb\_eq |
| 32 | longest\_word\_host |
| 33 | ip |
| 34 | shortest\_words\_raw |
| 35 | domain\_with\_copyright |
| 36 | ratio\_extMedia |
| 37 | ratio\_extErrors |
| 38 | ratio\_intMedia |
| 39 | nb\_subdomains |
| 40 | domain\_in\_brand |
| 41 | nb\_extCSS |
| 42 | nb\_qm |
| 43 | nb\_redirection |
| 44 | nb\_underscore |
| 45 | https\_token |
| 46 | tld\_in\_subdomain |
| 47 | shortening\_service |
| 48 | prefix\_suffix |
| 49 | empty\_title |
| 50 | external\_favicon |
| 51 | nb\_percent |
| 52 | nb\_and |
| 53 | dns\_record |
| 54 | nb\_space |
| 55 | whois\_registered\_domain |
| 56 | suspecious\_tld |
| 57 | random\_domain |
| 58 | login\_form |
| 59 | http\_in\_path |
| 60 | nb\_com |
| 61 | statistical\_report |
| 62 | tld\_in\_path |
| 63 | nb\_at |
| 64 | nb\_colon |
| 65 | nb\_semicolumn |
| 66 | abnormal\_subdomain |
| 67 | nb\_dslash |
| 68 | port |
| 69 | brand\_in\_path |
| 70 | popup\_window |
| 71 | iframe |
| 72 | nb\_tilde |
| 73 | nb\_comma |
| 74 | right\_clic |
| 75 | brand\_in\_subdomain |
| 76 | onmouseover |
| 77 | path\_extension |
| 78 | nb\_external\_redirection |
| 79 | nb\_dollar |
| 80 | nb\_or |
| 81 | nb\_star |
| 82 | punycode |
| 83 | ratio\_nullHyperlinks |
| 84 | ratio\_intRedirection |
| 85 | ratio\_intErrors |
| 86 | submit\_email |
| 87 | sfh |

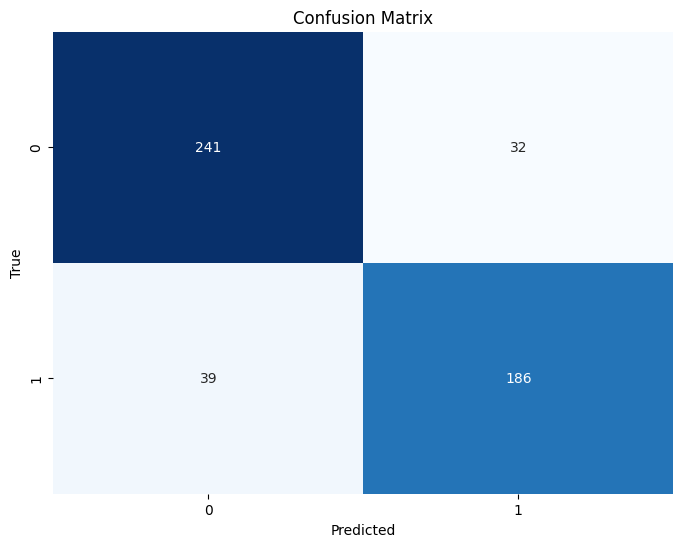
* 1. Dataset 3 –

|  |  |
| --- | --- |
| Rank | Feature |
| 1 | qty\_comma\_file |
| 2 | time\_domain\_activation |
| 3 | directory\_length |
| 4 | asn\_ip |
| 5 | time\_response |
| 6 | ttl\_hostname |
| 7 | qty\_dot\_domain |
| 8 | length\_url |
| 9 | domain\_length |
| 10 | time\_domain\_expiration |
| 11 | qty\_nameservers |
| 12 | qty\_vowels\_domain |
| 13 | qty\_slash\_url |
| 14 | qty\_ip\_resolved |
| 15 | qty\_hyphen\_directory |
| 16 | qty\_mx\_servers |
| 17 | qty\_hyphen\_url |
| 18 | file\_length |
| 19 | qty\_redirects |
| 20 | domain\_spf |
| 21 | qty\_params |
| 22 | tls\_ssl\_certificate |
| 23 | qty\_slash\_directory |
| 24 | qty\_dot\_url |
| 25 | params\_length |
| 26 | qty\_equal\_url |
| 27 | qty\_underline\_url |
| 28 | qty\_hyphen\_domain |
| 29 | qty\_dot\_directory |
| 30 | qty\_hyphen\_file |
| 31 | qty\_tld\_url |
| 32 | qty\_underline\_directory |
| 33 | qty\_underline\_file |
| 34 | qty\_dot\_file |
| 35 | qty\_hyphen\_params |
| 36 | qty\_plus\_url |
| 37 | url\_shortened |
| 38 | qty\_dot\_params |
| 39 | qty\_percent\_file |
| 40 | domain\_in\_ip |
| 41 | qty\_and\_url |
| 42 | qty\_underline\_params |
| 43 | qty\_and\_params |
| 44 | tld\_present\_params |
| 45 | qty\_at\_params |
| 46 | qty\_slash\_params |
| 47 | qty\_equal\_params |
| 48 | qty\_percent\_params |
| 49 | qty\_space\_params |
| 50 | domain\_google\_index |
| 51 | qty\_comma\_params |
| 52 | qty\_and\_directory |
| 53 | qty\_hashtag\_params |
| 54 | qty\_tilde\_params |
| 55 | qty\_comma\_url |
| 56 | qty\_questionmark\_params |
| 57 | qty\_tilde\_directory |
| 58 | qty\_questionmark\_url |
| 59 | qty\_at\_url |
| 60 | qty\_exclamation\_url |
| 61 | qty\_space\_url |
| 62 | qty\_tilde\_url |
| 63 | qty\_asterisk\_url |
| 64 | qty\_hashtag\_url |
| 65 | qty\_dollar\_url |
| 66 | qty\_percent\_url |
| 67 | qty\_underline\_domain |
| 68 | qty\_slash\_domain |
| 69 | qty\_questionmark\_domain |
| 70 | qty\_equal\_domain |
| 71 | qty\_at\_domain |
| 72 | qty\_and\_domain |
| 73 | qty\_exclamation\_domain |
| 74 | qty\_space\_domain |
| 75 | qty\_tilde\_domain |
| 76 | qty\_comma\_domain |
| 77 | qty\_plus\_domain |
| 78 | qty\_asterisk\_domain |
| 79 | qty\_hashtag\_domain |
| 80 | qty\_dollar\_domain |
| 81 | qty\_percent\_domain |
| 82 | server\_client\_domain |
| 83 | qty\_questionmark\_directory |
| 84 | qty\_equal\_directory |
| 85 | qty\_at\_directory |
| 86 | qty\_exclamation\_directory |
| 87 | qty\_space\_directory |
| 88 | qty\_comma\_directory |
| 89 | qty\_plus\_directory |
| 90 | qty\_asterisk\_directory |
| 91 | qty\_hashtag\_directory |
| 92 | qty\_dollar\_directory |
| 93 | qty\_percent\_directory |
| 94 | qty\_slash\_file |
| 95 | qty\_questionmark\_file |
| 96 | qty\_equal\_file |
| 97 | qty\_at\_file |
| 98 | qty\_and\_file |
| 99 | qty\_exclamation\_file |
| 100 | qty\_space\_file |
| 101 | qty\_tilde\_file |
| 102 | qty\_plus\_file |
| 103 | qty\_asterisk\_file |
| 104 | qty\_hashtag\_file |
| 105 | qty\_dollar\_file |
| 106 | qty\_exclamation\_params |
| 107 | qty\_plus\_params |
| 108 | qty\_asterisk\_params |
| 109 | qty\_dollar\_params |
| 110 | email\_in\_url |
| 111 | url\_google\_index |

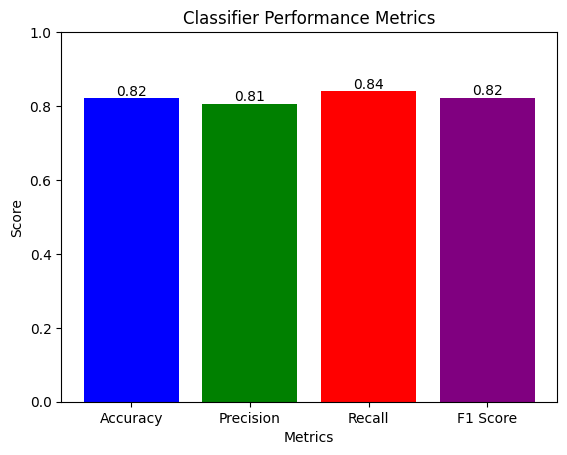
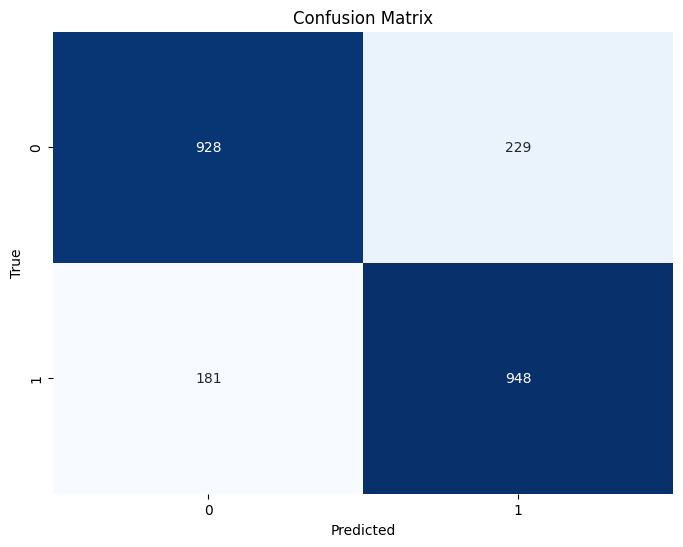
1. **Classification Algorithms and Performance Metrices:** 
   1. **KNN –**

**Dataset 1**

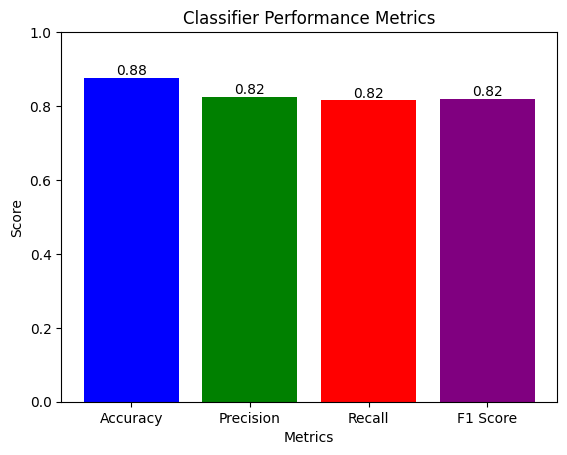




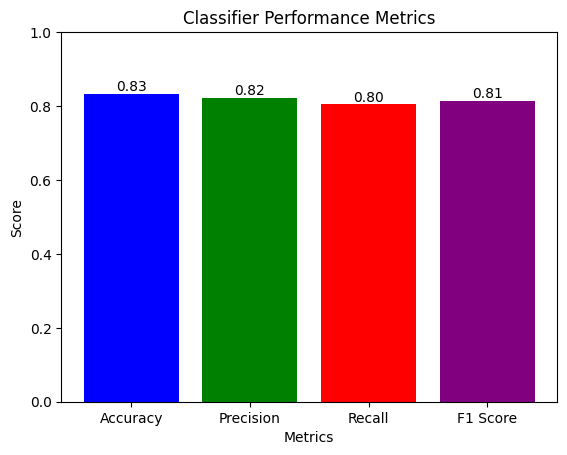
**Dataset 2**

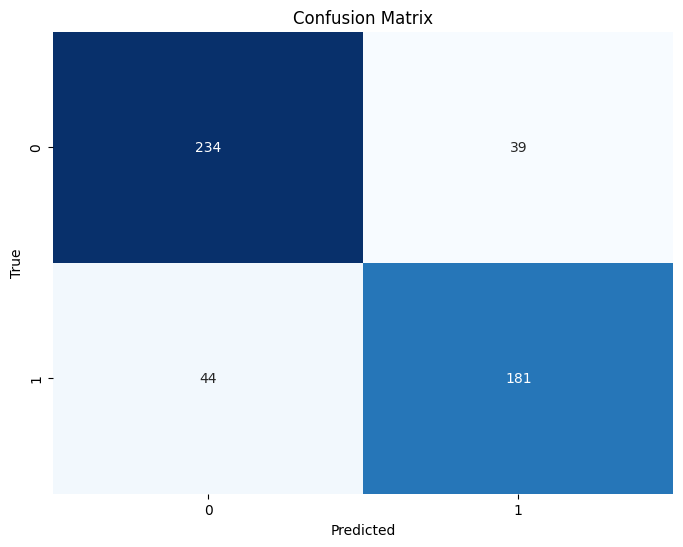


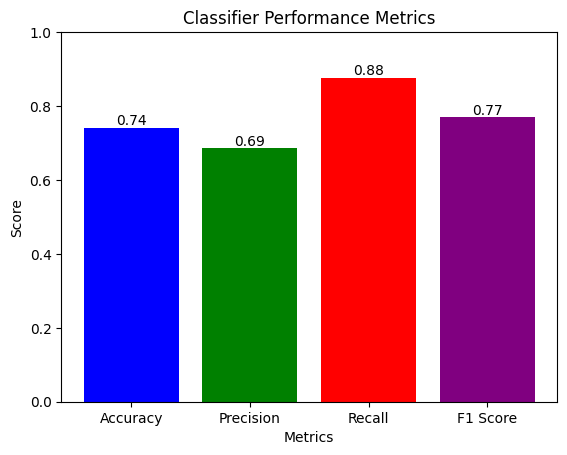
**Dataset 3**

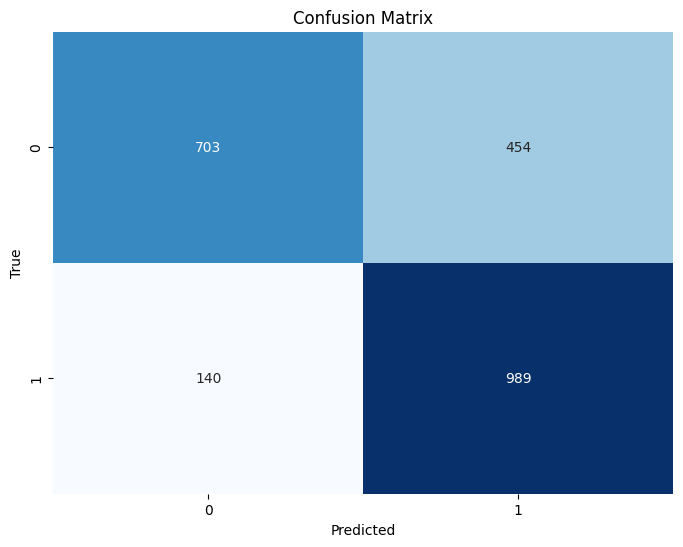
****

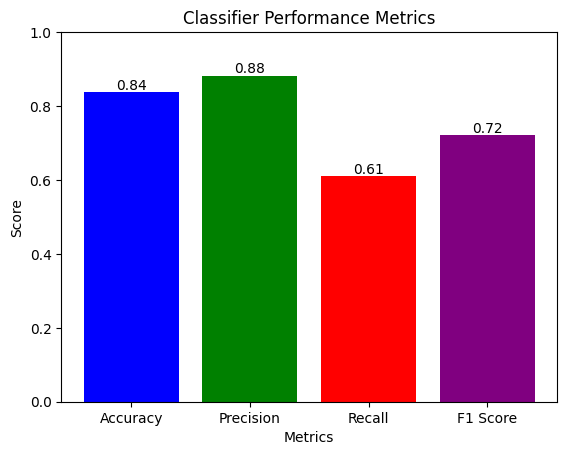
* 1. **Naïve Bayes –**

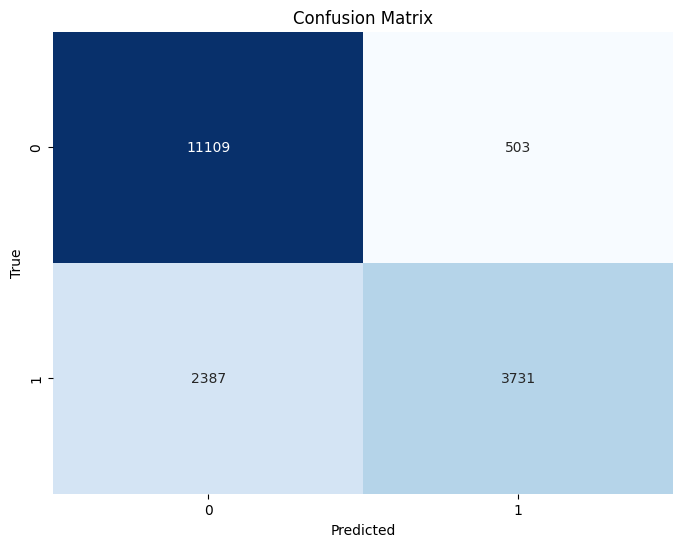
**Dataset 1**



**D****ataset 2**

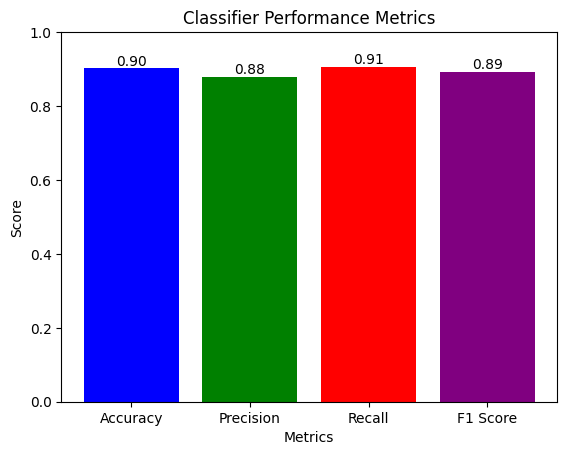
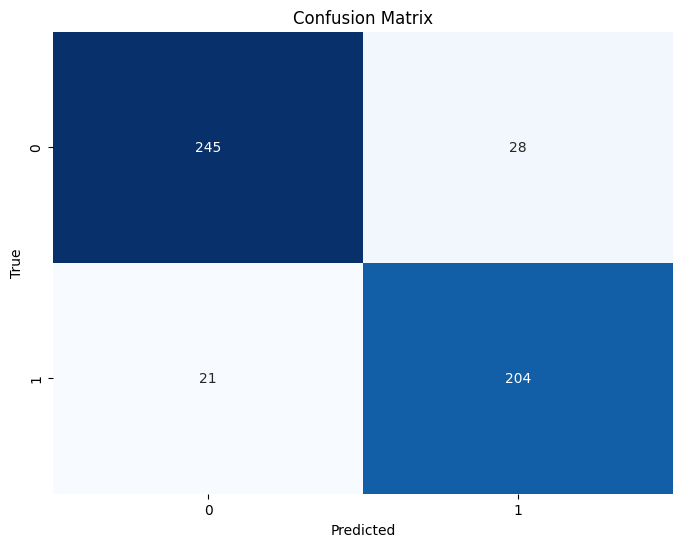


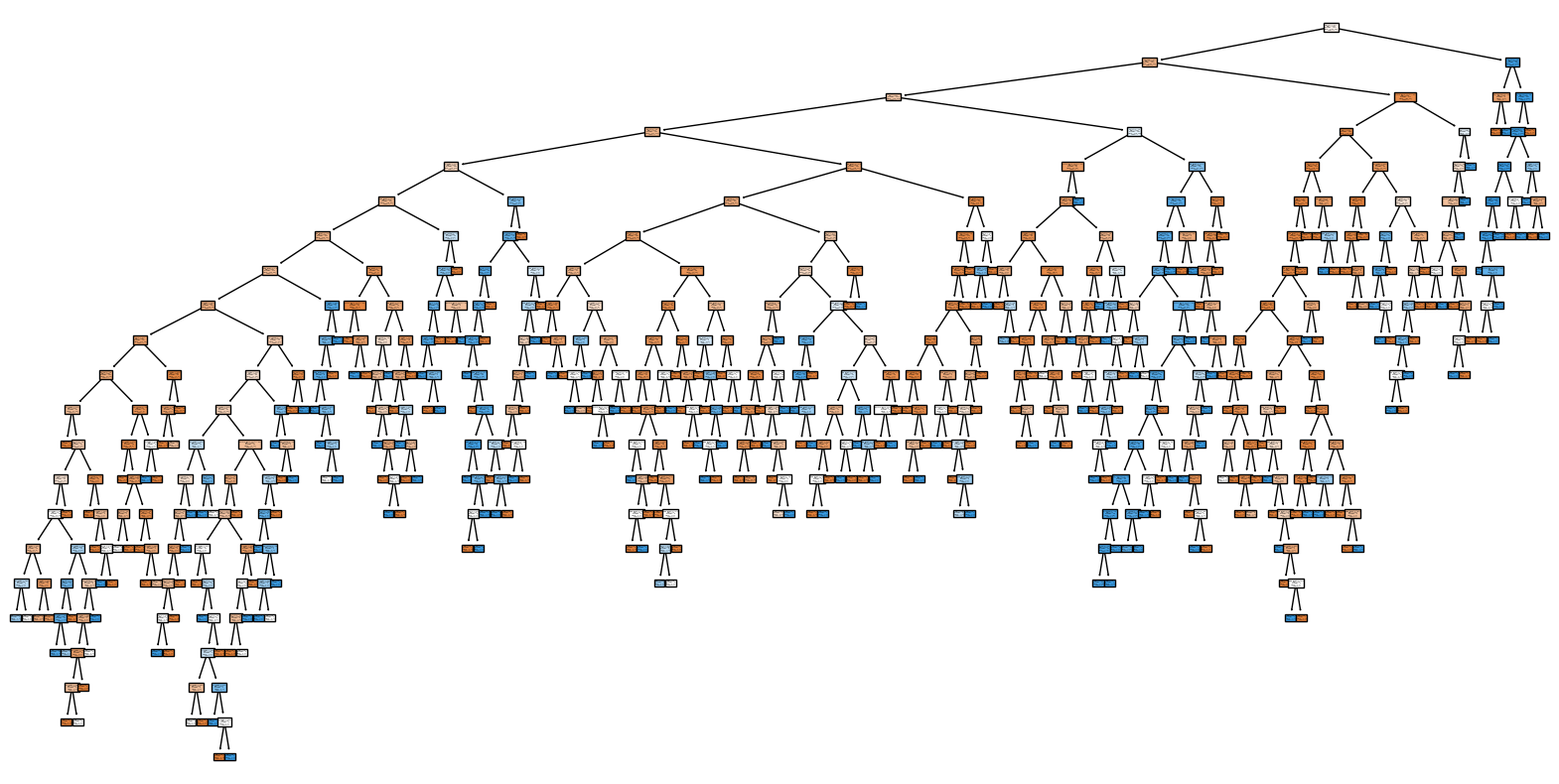
**Dataset 3**



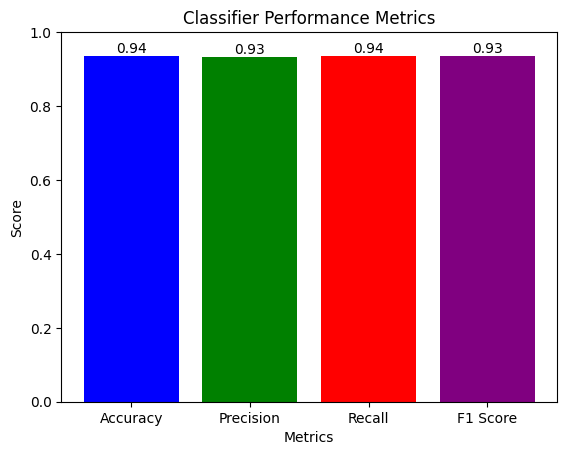
* 1. **Decision Tree–**

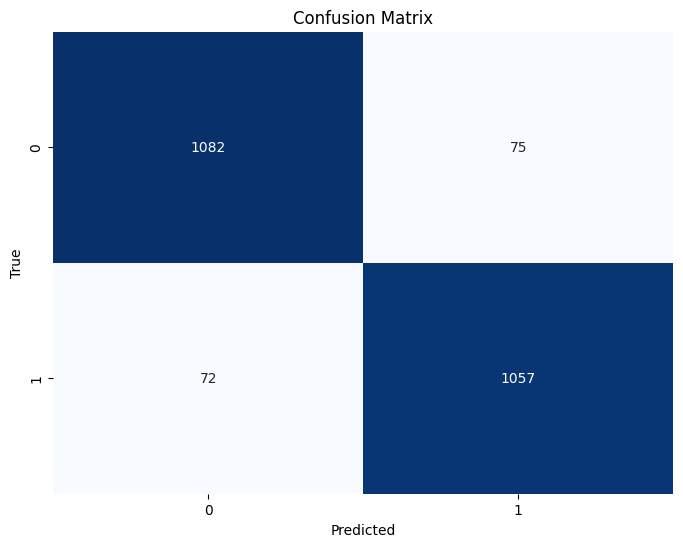
**Dataset 1**

****

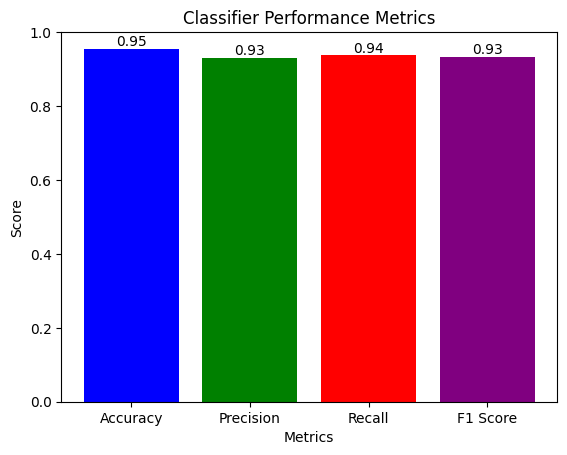
****

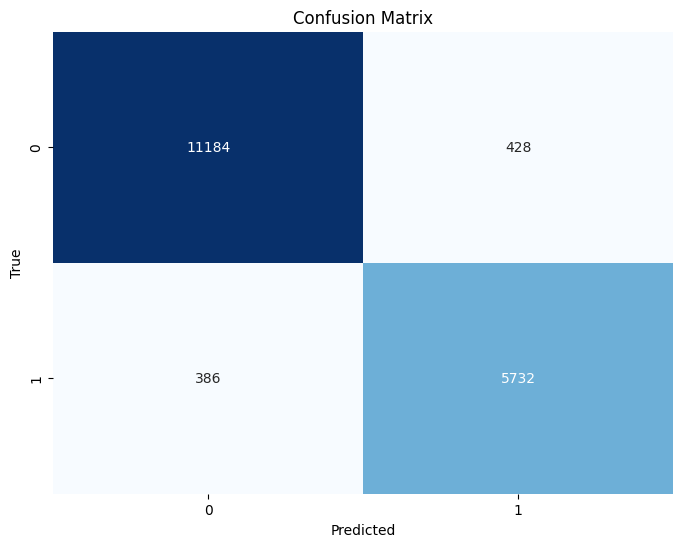
**Dataset 2**

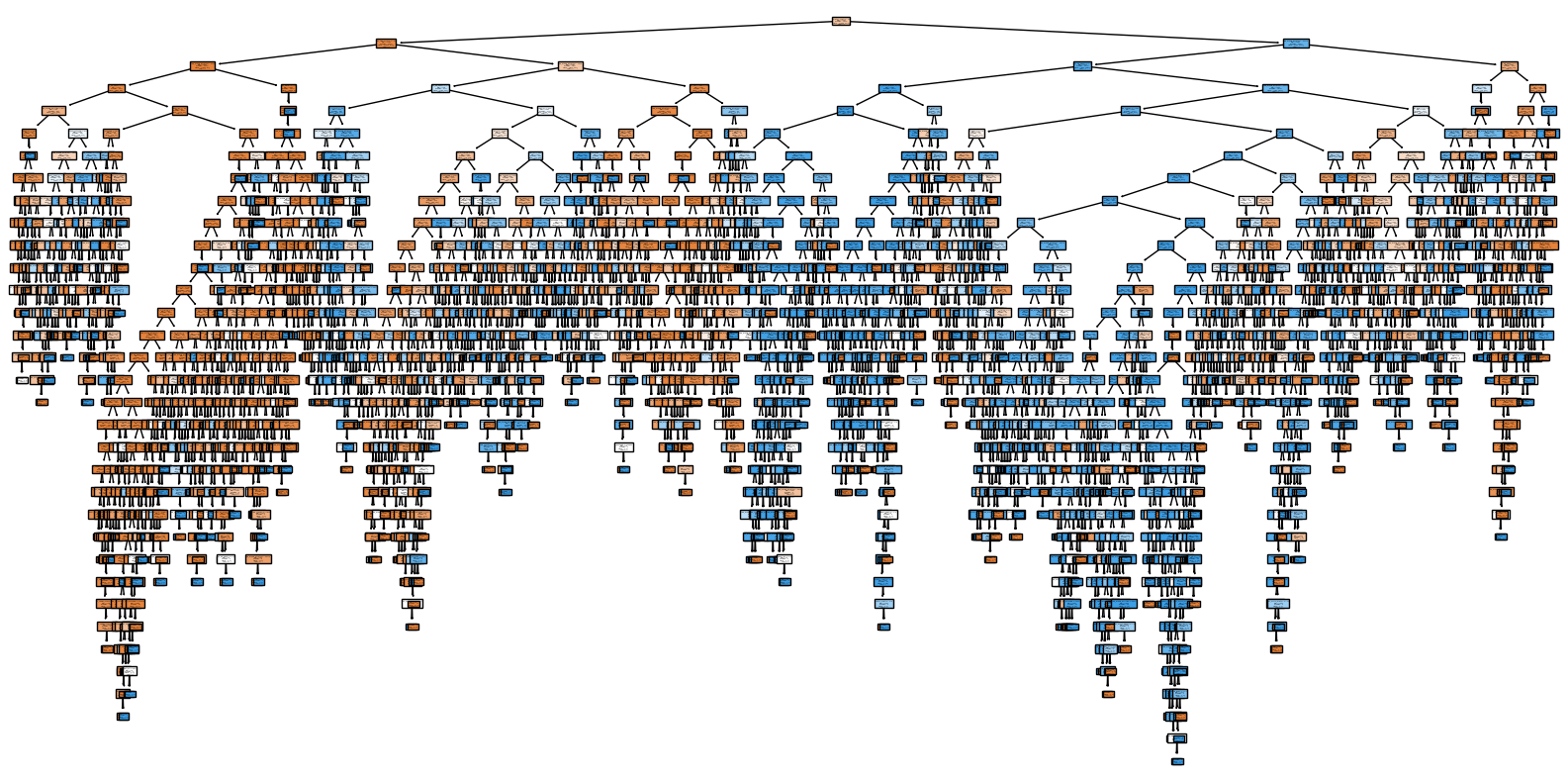
****



****

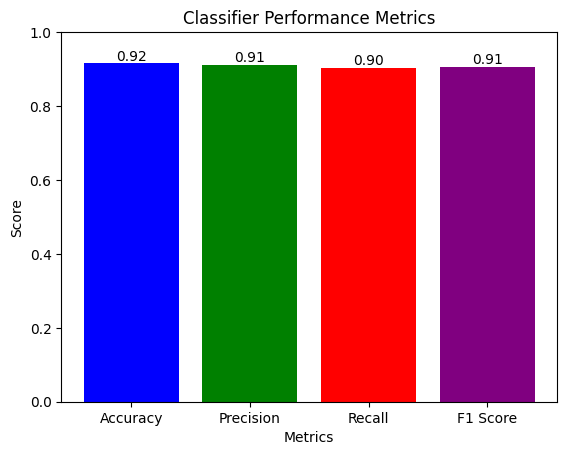
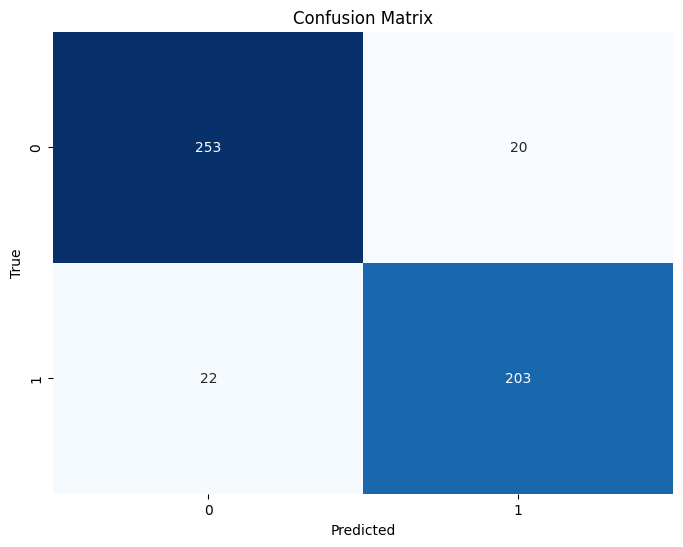
**Dataset 3**



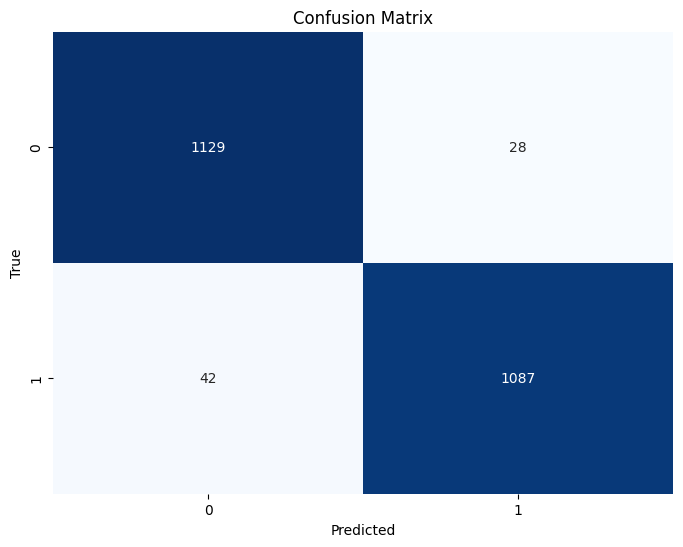
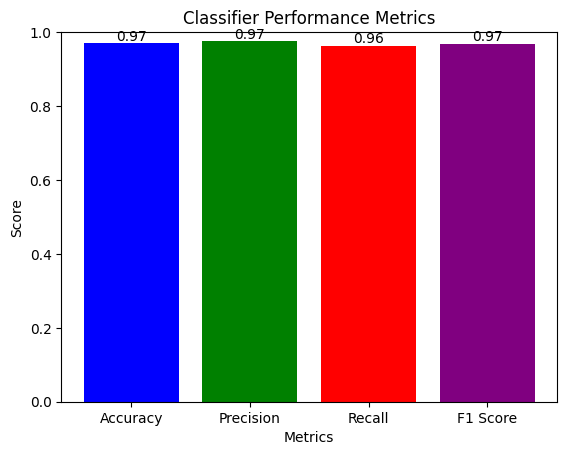
****

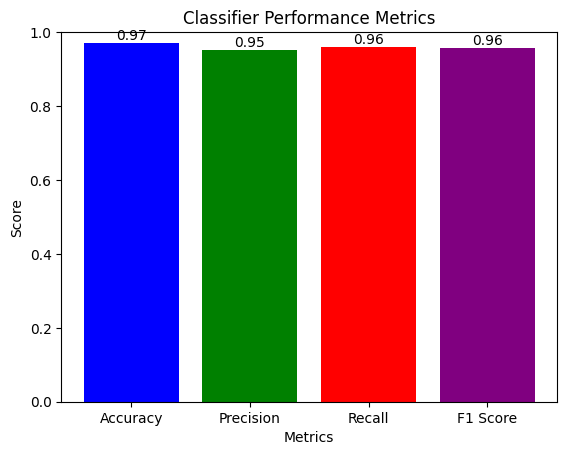
* 1. **Random Forest–**

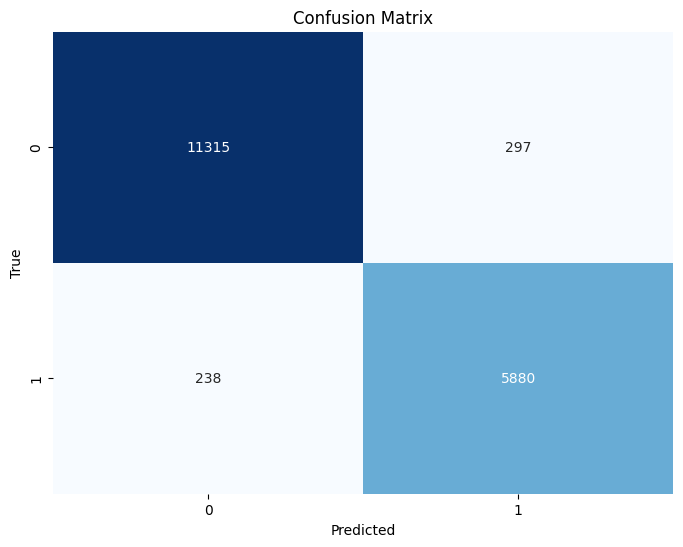
**Dataset 1**

****

**Dataset 2**

****

**Dataset 3**



1. **Analysis and Comparison:** 
   1. **Dataset 1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| KNN | 0.86% | 0.85% | 0.83% | 0.84% |
| Naïve Bayes | 0.83% | 0.82% | 0.80% | 0.81% |
| Decision Tree | 0.90% | 0.88% | 0.91% | 0.89% |
| Random Forest | 0.92% | 0.91% | 0.90% | 0.91% |

* 1. **Dataset 2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| KNN | 0.82% | 0.81% | 0.84% | 0.82% |
| Naïve Bayes | 0.74% | 0.69% | 0.88% | 0.77% |
| Decision Tree | 0.94% | 0.93% | 0.94% | 0.93% |
| Random Forest | 0.97% | 0.97% | 0.96% | 0.97% |

* 1. **Dataset 3**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| KNN | 0.88% | 0.82% | 0.82% | 0.82% |
| Naïve Bayes | 0.84% | 0.88% | 0.61% | 0.72% |
| Decision Tree | 0.95% | 0.93% | 0.94% | 0.93% |
| Random Forest | 0.97% | 0.95% | 0.96% | 0.96% |