Data Collection and Preprocessing Phase

Date	06-06-2024
Team ID	740031
Project Title	DETECTION OF PHISHING WEBSITE FROM URLS
Maximum Marks	6 Marks

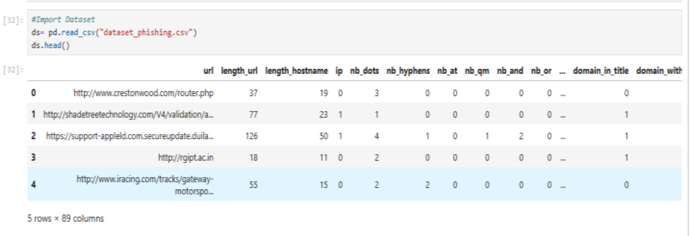
Data Exploration and Preprocessing Report

Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

Section	**Description**
Data Overview	Dimension: 614 rows × 13 columns Descriptive statistics: -
Univariate Analysis	

Bivariate Analysis	-
Multivariate Analysis	-

<table>
<tr><td>Outliers and Anomalies</td><td>-</td></tr>
<tr><td colspan="2"><b>Data Preprocessing Code Screenshots</b></td></tr>
<tr><td>Loading Data</td><td>



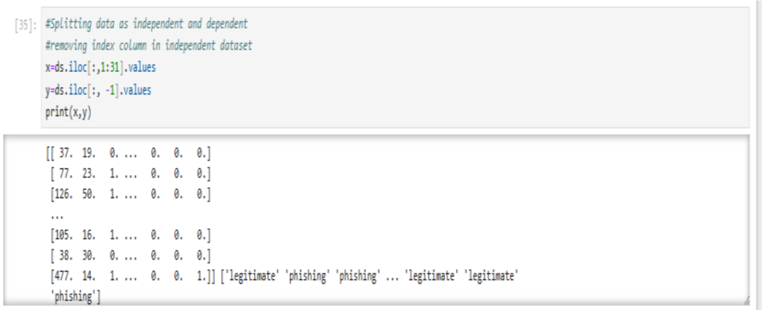
```
[32]: #Import Dataset
      ds= pd.read_csv("dataset_phishing.csv")
      ds.head()
```

[32]:

<table>
<tr><th></th><th>url</th><th>length_url</th><th>length_hostname</th><th>ip</th><th>nb_dots</th><th>nb_hyphens</th><th>nb_at</th><th>nb_qm</th><th>nb_and</th><th>nb_or</th><th>...</th><th>domain_in_title</th><th>domain_with</th></tr>
<tr><td>0</td><td>http://www.crestonwood.com/router.php</td><td>37</td><td>19</td><td>0</td><td>3</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>...</td><td>0</td><td></td></tr>
<tr><td>1</td><td>http://shadetreetechnology.com/V4/validation/a...</td><td>77</td><td>23</td><td>1</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>...</td><td>1</td><td></td></tr>
<tr><td>2</td><td>https://support-appleld.com.secureupdate.duila...</td><td>126</td><td>50</td><td>1</td><td>4</td><td>1</td><td>0</td><td>1</td><td>2</td><td>0</td><td>...</td><td>1</td><td></td></tr>
<tr><td>3</td><td>http://rgipt.ac.in</td><td>18</td><td>11</td><td>0</td><td>2</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>...</td><td>1</td><td></td></tr>
<tr><td>4</td><td>http://www.iracing.com/tracks/gateway-motorspo...</td><td>55</td><td>15</td><td>0</td><td>2</td><td>2</td><td>0</td><td>0</td><td>0</td><td>0</td><td>...</td><td>0</td><td></td></tr>
</table>

5 rows × 89 columns

</td></tr>
<tr><td>Handling Missing Data</td><td>



```
[35]: #Splitting data as independent and dependent
      #removing index column in independent dataset
      x=ds.iloc[:,1:31].values
      y=ds.iloc[:, -1].values
      print(x,y)
```

```
[[ 37.  19.   0. ...   0.   0.   0.]
 [ 77.  23.   1. ...   0.   0.   0.]
 [126.  50.   1. ...   0.   0.   0.]
 ...
 [105.  16.   1. ...   0.   0.   0.]
 [ 38.  30.   0. ...   0.   0.   0.]
 [477.  14.   1. ...   0.   0.   1.]] ['legitimate' 'phishing' 'phishing' ... 'legitimate' 'legitimate'
 'phishing']
```

-

</td></tr>
<tr><td>Data Transformation</td><td>-</td></tr>
<tr><td>Feature Engineering</td><td>Attached the codes in final submission.</td></tr>
<tr><td>Save Processed Data</td><td>-</td></tr>
</table>