Aim: To classify websites as Phishy or Legitimate based on features based on URLs, Content, and 3rd party sources.

- **Phase-1: Dataset Collection**
  In this Phase, we collected the URLs and content (HTML, Javascript, CSS, Screenshot, Favicons) of confirmed Phishy as well as confirmed Legitimate websites from **PhishTank.**
  Confirmed Legitimate websites
  Confirmed Phishy websites

  Initially, we tried extracting the website content using pywebcopy but were unable to download the resources properly (i.e., the content used to get downloaded in an unorganized manner) Also, the code was hard to debug. Also, the majority of the content had *'Not Found' and 'Forbidden'* as the content. So, we wrote the code to extract the HTML, Javascript (Inline+External), Favicons, and CSS ourselves. The screenshots of the websites are downloaded from the link: https://cdn.phishtank.com/{phishID}.jpg. Where the PhishID is the unique ID given to each website on PhishTank.

  The code for extracting the resources will collect the resources in folders named after their PhishID (so that each website URL can be mapped to the resources using PhishID). However, the HTML files, JS files, CSS files, etc., will be saved in their respective subfolders under that parent folder having PhishID as a name. The URL along with PhishID will also be stored in an Excel file which will have columns: **["PhishID", "URL", "HTML", "JS", "CSS", "Images", "Not Found", "Forbidden", "Favicon", "ScreenShot", "Status Code"']**
  Where the values of ['HTML', 'JS', 'CSS', 'Imgaes', 'Favicon', 'Screenshot''] will be **one** if the code is able to download that particular resource, 0 otherwise, similarly the 'Not Found' will be **one** if the website is not found (i.e., error: 404), 0 otherwise, the Status code is the code returned to the request made to the URL (In case of response as **OK**, the status code will be **200**).

- **Phase-2 Feature Extraction and Classification**
  In this phase, we used the URLs and content to extract URL-based and content-based features for each website. We also used some 3rd party features, which are extracted using whois.
  The URL-based features are as follows:
  1. ***Length of URL***: Value will be 1 if the length of the URL is less than 54, 0 if the length is between 54 and 75 (inclusive), and -1 otherwise.
  2. ***Has IP address:*** If the domain has an IP address, then the value will be -1, else 1.
  3. ***Shortening Service:*** If the URL has used any URL shortening services, then the value will be -1 else, the value will be 1.
  4. ***Having @ Symbol:*** If the URL has '**@**', then the value will be -1, 1 otherwise.

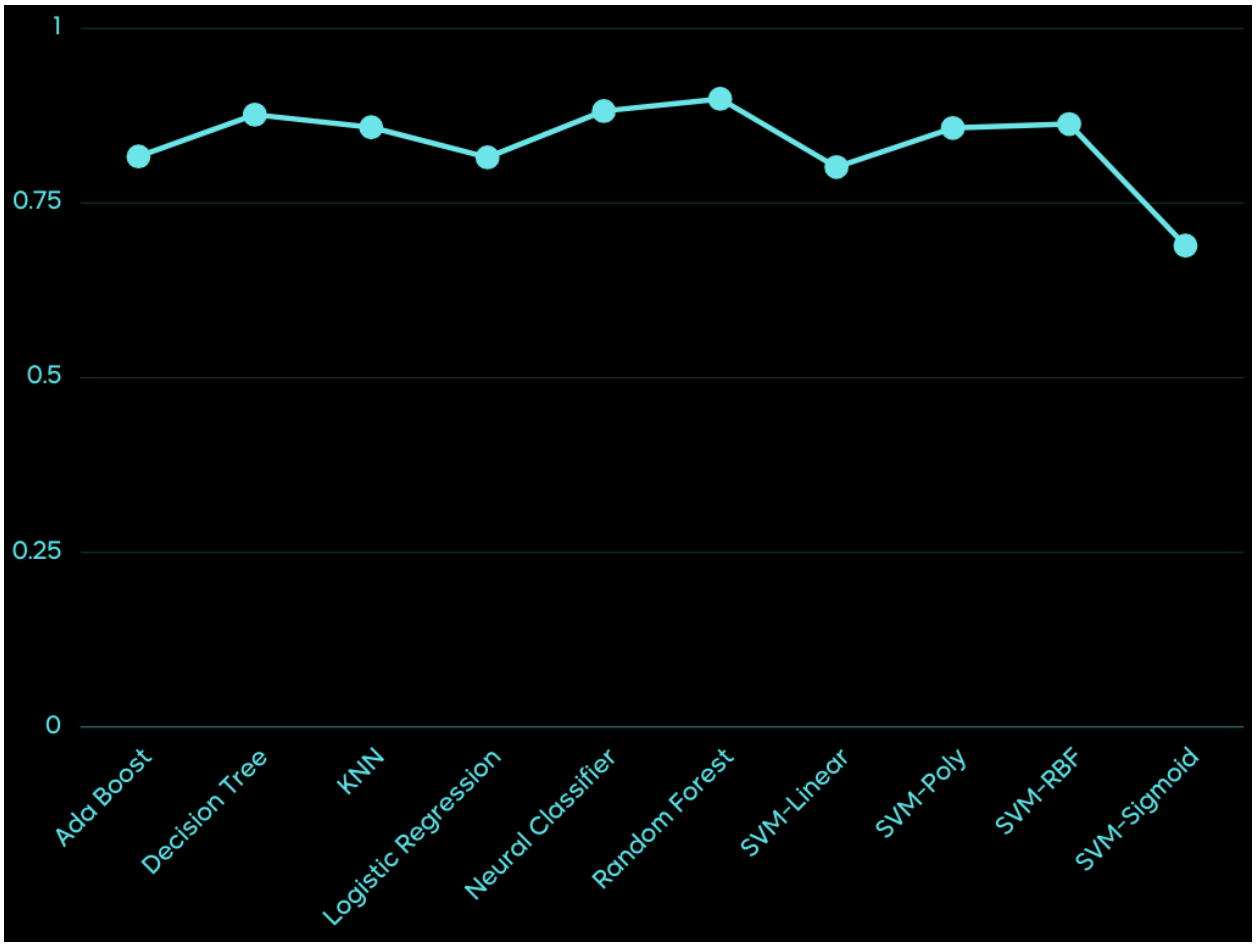
5. ***Double Slash Redirecting:*** If there is a double slash at the 7th or 8th position of the URL, then the value will be 1, If the position is greater than 7, then the value will be -1.
6. ***Prefix-Suffix***: Checks if '**-**' is present or not in the URL; if present, then return -1, 1 otherwise.
7. ***CTLD:***
8. ***HTTPS in Domain:*** If 'HTTPS'/'https' or 'HTTP'/'http' is present in the URL, then return -1, 1 otherwise.
9. ***Sensitive Words:*** If sensitive words like *'secure', 'account', 'webscr', 'login', 'ebayisapi', 'signin', 'banking', 'confirm', 'credit', 'verify', 'reset', 'verification', 'authenticate'* are present in the URL then return -1, 1 otherwise
10. ***Has Tilde:*** Check if '**~**' is present in the URL; if present, then return -1, else 1
11. ***Has Port:*** Check if the port is present in the URL; if present, then return 1, and -1 otherwise.
12. ***Current Age of Domain:*** This feature calculates the current age of the domain (i.e., current date - domain registration date). If the current age is greater than 365 days (i.e., 1 year), then the value will be -1, and 1 otherwise.
13. ***Length of Domain:*** This feature calculates the difference between the domain registration date and expiration date, if the difference is less than or equal to 365 days, then the value will be -1, 1 otherwise.
14. ***Match Domain Name:*** The value of this feature will be '1' if the domain name in the URL matches with the domain name registered on the Whois server and -1 otherwise.
15. ***Frequency of <a> tags***: compares the maximum frequent link domain of the hrefs present in the <a> tags with the base urls domain.
16. ***Frequency of all_links(links, images, scripts)***: compares the maximum frequent link domain of the hrefs present in the <a> tags with the base urls domain.
17. ***IFrame Redirection:*** IFrame is an HTML tag used to display an additional webpage into one that is currently shown. Phishers can make use of the "iframe" tag and make it invisible i.e., without frame borders. In this regard, phishers make use of the "*frameBorder*" attribute, which causes the browser to render a visual delineation.
18. ***Checking popups:*** It is unusual to find a legitimate website asking users to submit their personal information through a pop-up window. On the other hand, this feature has been used in some legitimate websites, and its main goal is to warn users about fraudulent activities or broadcast a welcome announcement, though no personal information was asked to be filled in through these pop-up windows.
19. ***Right-click disabling:*** Phishers use JavaScript to disable the right-click function so that users cannot view and save the webpage source code. This feature is treated exactly as "Using onMouseOver to hide the Link." Nonetheless, for this feature, we will search for the event **"event.button==2"** in the webpage source code and check if the right click is disabled.

20. ***Number of redirects(Website forwarding):*** The fine line that distinguishes phishing websites from legitimate ones is how many times a website has been redirected. Our dataset finds that legitimate websites have been redirected one-time max. On the other hand, phishing websites containing this feature have been redirected at least four times.
21. ***Server Form Handler(SFH):*** SFHs that contain an empty string or "**about: blank**" is considered doubtful because an action should be taken upon the submitted information. In addition, if the domain name in SFHs is different from the domain name of the webpage, this reveals that the webpage is suspicious because the submitted information is rarely handled by external domains.
22. ***Request_url***:- Request URL examines whether the external objects contained within a webpage, such as images, videos, and sounds, are loaded from another domain. In legitimate webpages, the webpage address and most of the objects embedded within the webpage share the same domain.
23. ***Url of <a> tags***:- An anchor is an element defined by the <a> tag. This feature is treated precisely as a "Request URL." However, for this feature, we examine:
    a. If the <a> tags and the website have different domain names. This is similar to the request URL feature.
    b. If the anchor does not link to any webpage, e.g.:
        i. *<a href="#">*
        ii. *<a href="#content">*
        iii. *<a href="#skip">*
        iv. *<a href="JavaScript ::void(0)">*
24. ***Links in <Meta>, <Script>, and <Link> tags:*** Given that our investigation covers all angles likely to be used in the webpage source code, we find that it is common for legitimate websites to use *<Meta>* tags to offer metadata about the HTML document; *<Script>* tags to create a client-side script; and <Link> tags to retrieve other web resources. It is expected that these tags are linked to the same domain of the webpage.

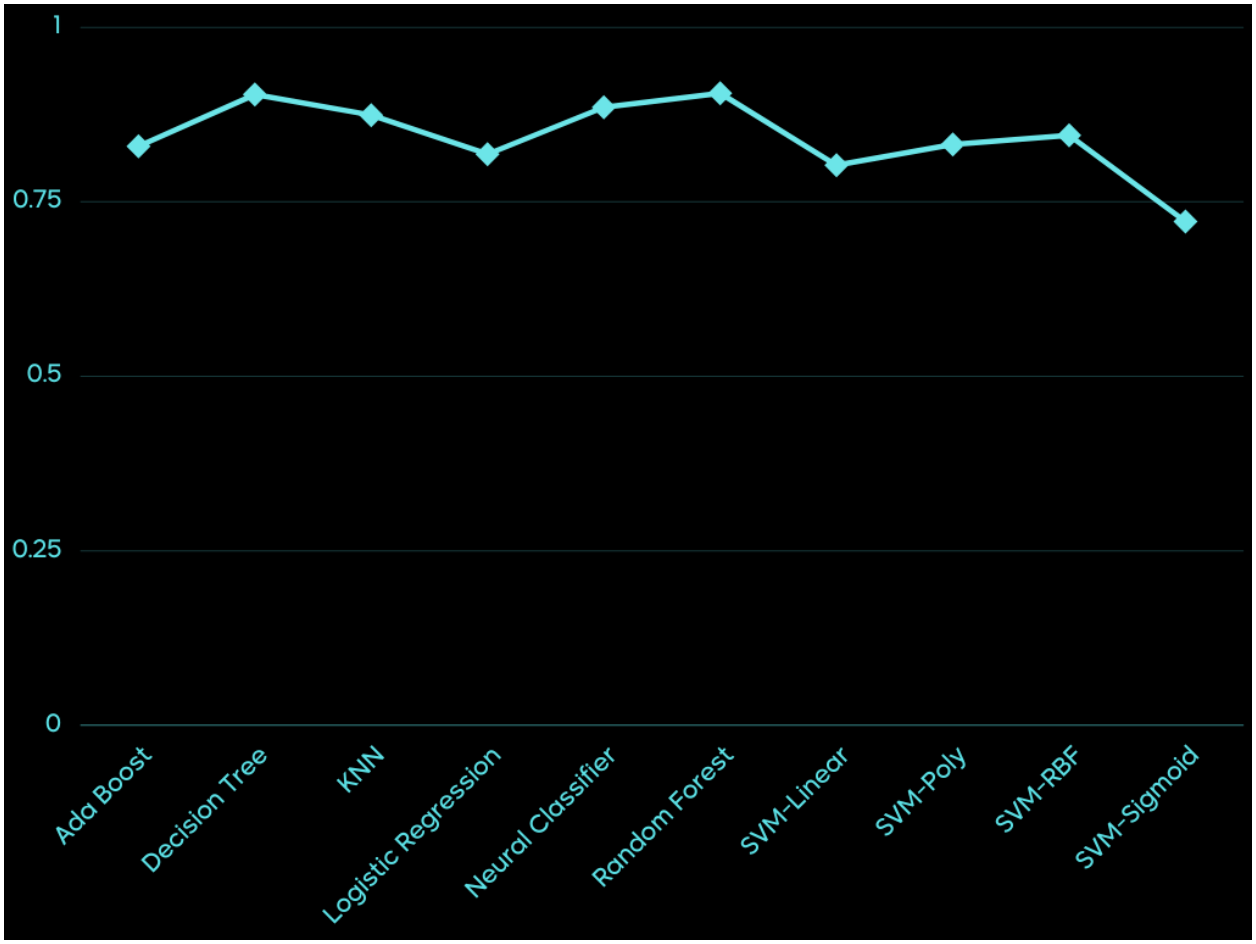
Results of Classification:

Classifier	**Accuracy** *(out of 1)*	**Recall** *(out of 1)*	**Precision** *(out of 1)*	**F1-Score** *(out of 1)*
Ada Boost	0.817	0.862	0.830	0.846
Decision Tree	0.877	0.882	0.903	0.893
KNN	0.858	0.884	0.875	0.879
Logistic	0.815	0.877	0.819	0.847

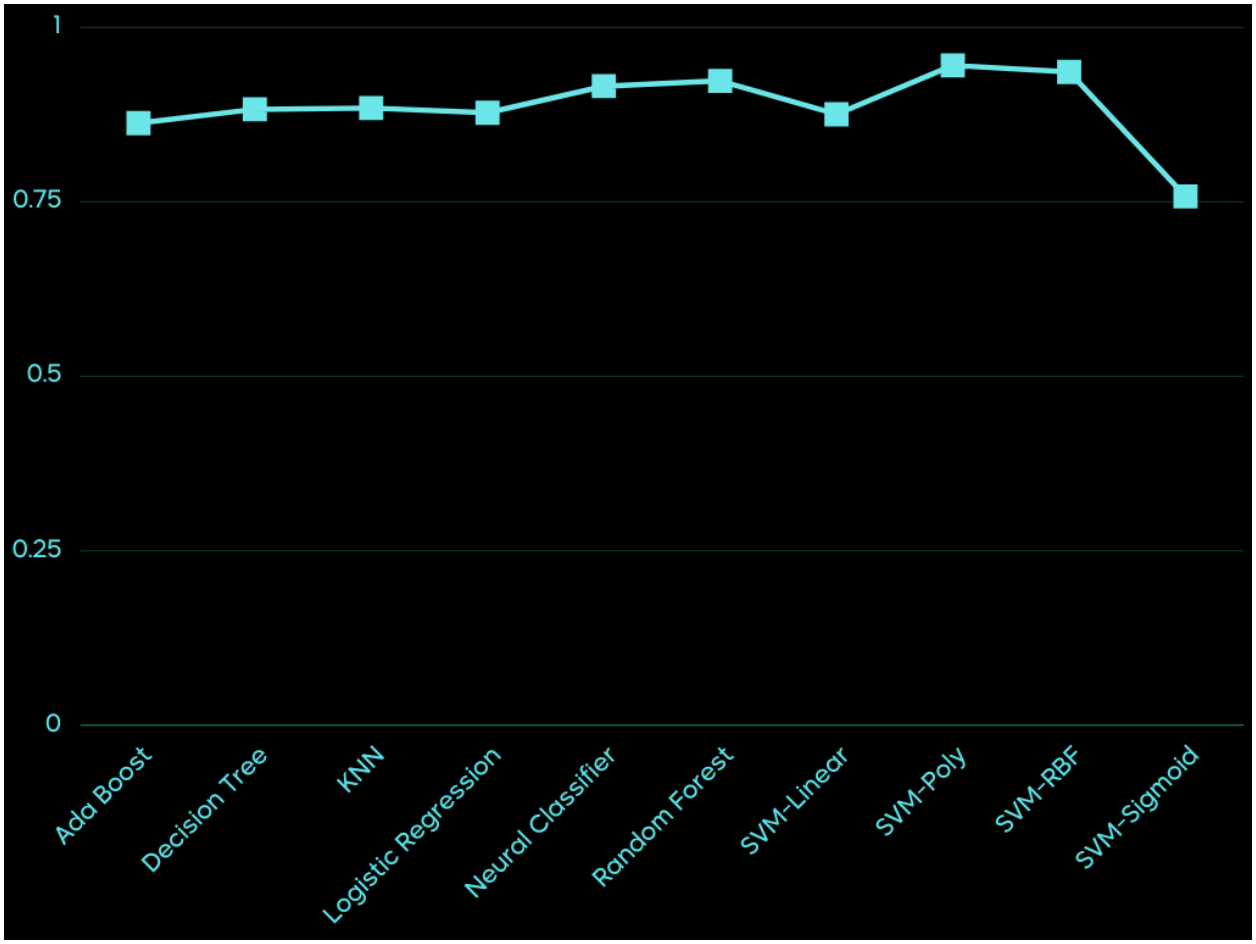
Regression				
Neural Classifier	0.882	0.915	0.886	0.900
Random Forest	0.899	0.923	0.906	0.914
SVM-Linear	0.801	0.875	0.803	0.837
SVM-Polynomial	0.857	0.945	0.833	0.885
SVM-RBF	0.863	0.936	0.846	0.889
SVM-Sigmoid	0.689	0.757	0.722	0.740



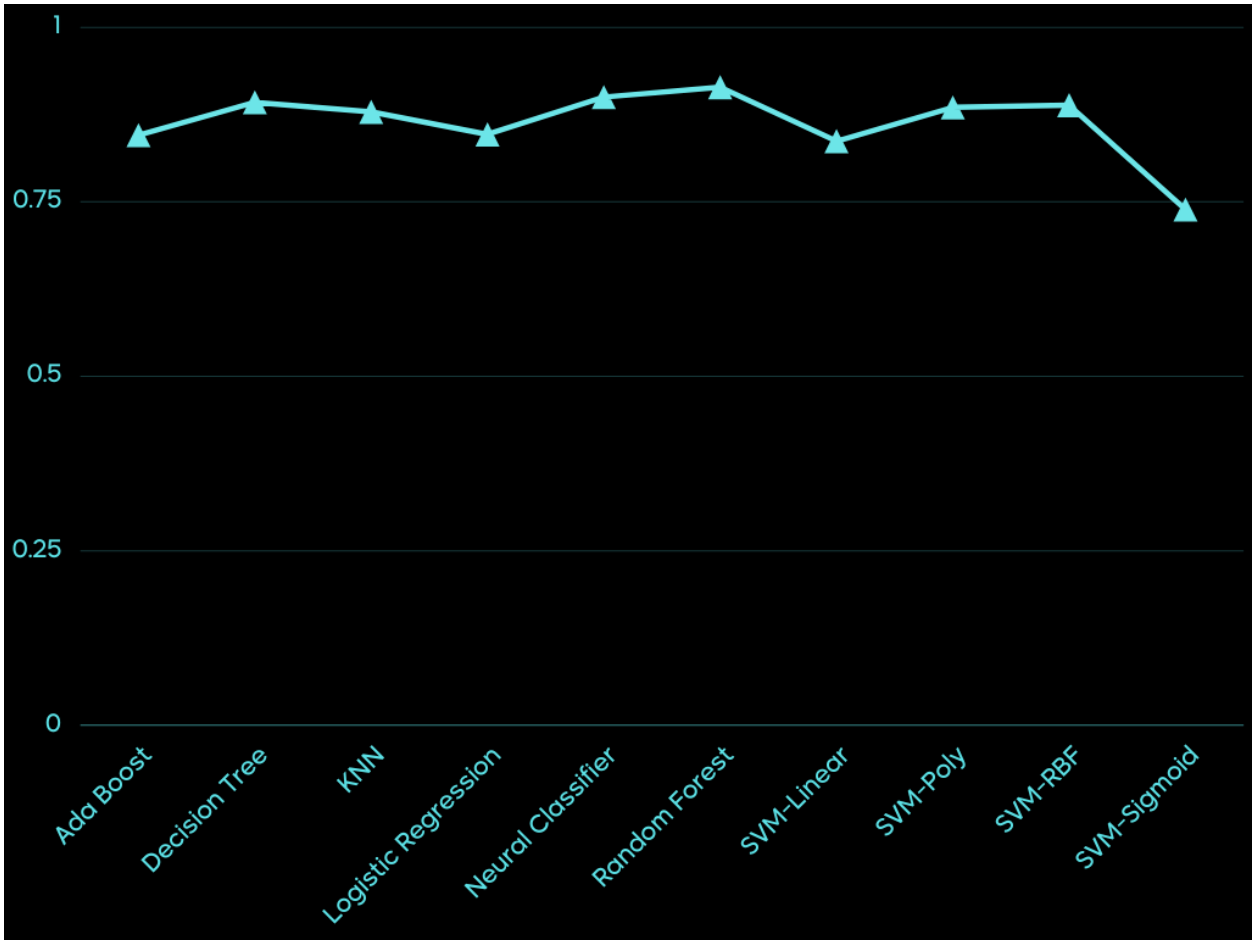
Accuracy


1
0.75
0.5
0.25
0
Ada Boost
Decision Tree
KNN
Logistic Regression
Neural Classifier
Random Forest
SVM-Linear
SVM-Poly
SVM-RBF
SVM-Sigmoid

Precision


1
0.75
0.5
0.25
0
Ada Boost
Decision Tree
KNN
Logistic Regression
Neural Classifier
Random Forest
SVM-Linear
SVM-Poly
SVM-RBF
SVM-Sigmoid

Recall


1
0.75
0.5
0.25
0
Ada Boost
Decision Tree
KNN
Logistic Regression
Neural Classifier
Random Forest
SVM-Linear
SVM-Poly
SVM-RBF
SVM-Sigmoid

F1-Score