Assumptions: -1: Legitimate, 0: Suspicious, 1: Phishing

Use python to check how many sites are actually legitimate and how many are phishing. Check this using result column and counting all 1’s and all -1’s. This should match up with 2456.

**Our Research Goal**

Our project aims to identify the most relevant subset of features that can accurately identify phishing URLs, using Decision Tree and Naïve Bayes algorithms together with five different subsets of features.

**Data**

We collected our training and test data from the UCI phishing dataset that is publicly available. This dataset contained 2456 unique URL instances and a total of 11,055 URLs of which 6,157 are phishing and 4,898 are legitimate sites. Each row represented a URL and each URL was previously parsed and represented according to 30 features which could determine whether or not the URL is used for phishing or just identify a specific feature as suspicious for a particular URL. The features considered includes whether or not an IP Address is used instead of a URL, the length of the URL, the presence of link tags to the same domain as the webpage and whether or not the webpage uses IFrames. At the end of each row, there is a result which identifies the true nature of the URL, 1 if it is a phishing site and -1 if it a legitimate site.

**Experiment**

We chose to use the classification algorithm, Random Forest, because our dataset is made up of discrete data. This means each feature can only be represented by -1, 1 and sometimes 0. Random Forest also proves to be highly efficient when working with large datasets. We used 10,000 URLs for training and 1,055 for testing. Our belief is that using more URLs to train the algorithm should result in better predictions, and hence higher accuracy, recall and F1 scores, for the test data.

From the 30 features we identified five subsets. These were:

Set A: Features we think are most important.

Set B: Features without 3 possible outcomes. This makes the data more binary and eliminates the possibility of uncertain URLs.

?Set C: Web-presence related features: Domain age, Website traffic, Page Rank, Google Index

?Set D:

Set E: All 30 features

Is dataset discrete or continuous? Determine this.