**Data Mining Task**

The task is to tell whether a given website URL is a phishing site or not. The task will classify based on the characteristics of the URL.

**Datasets**

https://www.kaggle.com/datasets/akashkr/phishing-website-dataset?resource=download

The dataset has 11055 entries. With 9 different variables like if the URL has an IP address in it, if the URL has a long length, if a shortening service is used, if the @ symbol is used in the url, if // appears in the url, and other variables

data.phishtank.com/data/online-valid.csv

This data comes from phishtank.com which is a site where people can report suspecting phishing websites. This includes variables such as their ID in the phishtank.com servers, their url, when it was submitted, if it was a verified address, if it’s online, and what site it targets. It contains 93920 entries.

**Methodology**

For this task I will either be using weka or scikit-learn to build a model that can predict if a given site is a phishing site or not. Using scikit-learn I will probably be using a logistic regression classifier and a decision tree classifier and comparing the efficiency and accuracy of the two. Which could develop to more types of models or programs as we progress through this course.

**Final Product**

The final product for this project is to supply a series of URLs from a testing dataset (probably from an 80/20 training/testing split) and the model should accurately (above 75%) discern if a URL is a phishing URL based on its characteristics. I don’t think I’ll learn anything new about the topic itself, but I will learn about training and testing ML models and other data structures (like decision trees) that help influence decisions.