Clizia Giorgia Manganaro
ID. 2017897

Project Proposal

Phishing URL Classification with PySpark

Task:

Phishing is a type of social engineering attack often used to steal personal and financial information.
This attack usually takes place by sending emails that contain links within them that allow you to access the malicious site with the hope that the unfortunate user enters a username, password and/or other information.
Therefore, **the main task of the project is to identify and classify phishing URLs in order to counter this phenomenon which is growing rapidly.**

Dataset:

The dataset chosen to use for the project can be found on Kaggle at the following link: Kaggle and it is composed by ~11500 Urls with 89 extracted features.
Features are from three different classes:

1. 56 extracted from the structure and syntax of URLs;
2. 24 extracted from the content of their correspondent pages;
3. 7 are extracted by querying external services.

In addition, the dataset is balanced and it contains exactly 50% phishing and 50% legitimate URLs.

ML methods:

Different binary classification models will be analyzed to determine if a URL is legitimate or used in a phishing attempt and compare the performance achieved by each.
In detail:

- SVM
- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier

Evaluation:

After analyzing the various models, a comparison and evaluation will be made between them.
It will be carried out graphically using the ROC curve and the accuracy of the models will also be calculated to verify which model will obtain better performance.