NIT3202 Data Analytics for Cyber Security Assignment

Objectives

- To apply skills and knowledge acquired throughout the block in classification algorithms and machine learning process.
- To rationalize the use of machine learning algorithms to process of data effectively and efficiently in big size.
- To demonstrate ability to use R to perform email classification tasks that are common for corporate security analyst.
- To scientifically conduct and document machine learning experiments for analytics purposes.

This assignment consists of a report worth 30 marks. Delays caused by student's own computer downtime cannot be accepted as a valid reason for late submission without penalty. Students must plan their work to allow for both scheduled and unscheduled downtime.

Submission instructions: You must submit an electronic copy of all your assignment files via VU Collaborate Dropbox. You must include both your report, source codes and necessary data files. Assignments will not be accepted through any other manner of submission.
Students should note that email and paper-based submissions will ordinarily be rejected.

Late submissions: Submissions received after the due date are penalized at a rate of 30% (out of the full mark) per day, no exceptions. Late submission after 3 days would not be accepted, and mark would be recorded as zero (0).

Copying, Plagiarism Notice

This is an individual assignment. You are not permitted to work as a part of a group when writing this assignment. The University's policy on plagiarism can be viewed online at https://policy.vu.edu.au/view.current.php?id=27

Overview

Machine learning methods use effectively to detect malicious websites. In this assignment, you are required to classify malicious websites by using provided datasets. The features have been extracted and clearly structured in CSV format, as summarized in Table 1.

Table 1: features description of malicious and benign websites dataset

COLUMN NAME	description
URL	the anonymous identification of the URL analyzed in the study
URL_LENGTH	the number of characters in the URL
NUMBER_SPECIAL_CHARACTERS	the number of special characters identified in the URL, such as /, %, #, &, =

CHARSET	the character encoding standard (also known as the character set)
SERVER	the operating system of the server obtained from the packet response.
CONTENT_LENGTH	the content size of the HTTP header
WHOIS_COUNTRY	the country of the server
WHOIS_STATEPRO	the state of the country of the server (if known)
WHOIS_REGDATE	the server date and time
WHOIS_UPDATED_DATE	the last update of the server
TCP_CONVERSATION_EXCHANGE	the number of TCP packets exchanged between the server and our honeypot client
DIST_REMOTE_TCP_PORT	the number of the ports detected and different to TCP
REMOTE_IPS	the total number of IPs connected to the honeypot
APP_BYTES	the number of bytes transferred
SOURCE_APP_PACKETS	packets sent from the honeypot to the server
REMOTE_APP_PACKETS	packets received from the server
APP_PACKETS	the total number of IP packets generated during the communication between the honeypot and the server
DNS_QUERY_TIMES	the number of DNS packets generated during the communication between the honeypot and the server
TYPE	1 is for malicious websites and 0 is for benign websites

Problem Statement

This is an individual assessment task. Each student is required to submit a report of approximately 1000 words along with exhibits to support findings with respect to the provided malicious and benign websites. This report should consist of:

- Literature review in malicious websites detection
- Construction of datasets, data pre-processing and features
- Workflow of malicious website detection that describes the process of conducting malicious website detection
- Technical findings of classification results
- Justified discussion of the performance evaluation outcomes for different classifiers

Requirements

To demonstrate your achievement of these goals, you must write a report of at least 1,000 words (1,500 words maximum and excluding references). Your report should consist of the following chapters:

1. A proper title that matches the contents of your report. It should be no more than 20 words. (1 points)

2. Your name and student number in the author line. (1 points)
3. An executive summary that summarizes the context, aim, method, findings, recommendations, and limitations of the paper. The summary should be at most **150 words**. (You may find hints on writing good executive summaries from https://students.unimelb.edu.au/academic-skills/explore-our-resources/report-writing/executive-summaries.) (2 points)
4. An introduction chapter that provides the background of the paper, the classification algorithms of your choice (at least three algorithms), the data used for classification, the performance evaluation metrics (at least three evaluation metrics), the summary of your findings, and the organization of the rest of your report. It should be at most **250 words**. (2 points)
5. A literature review chapter that surveys the latest techniques from academic research papers regarding malicious website detection. Besides machine learning techniques, you need to include at least another two approaches to detect malicious websites. You are advised to identify and cite at least one paper published by ACM and IEEE journals or conference proceedings. In addition, the aim of this part of the report is to demonstrate a deep and thorough understanding of the existing body of knowledge encompassing multiple classification techniques for security data analytics. Specifically, your argument should explain why machine learning algorithms should be used rather than human readers. It should be at most **250 words**. (Please read through the hints on this web page before writing this chapter http://www.uq.edu.au/student-services/learning/literature-review.) (4 points)
6. A chapter that describes the dataset you used for malicious website detection, including the construction of datasets, data pre-processing, and the features identified for classification. It should be less than **200 words**. (4 points)
7. A methodology chapter that depicts the workflow of malicious website detection which describes the process of conducting malicious website detection. Figure 1 gives an example of Twitter Spam Detection workflow. You must have a paragraph of description of the workflow within **100 words**. (4 points)

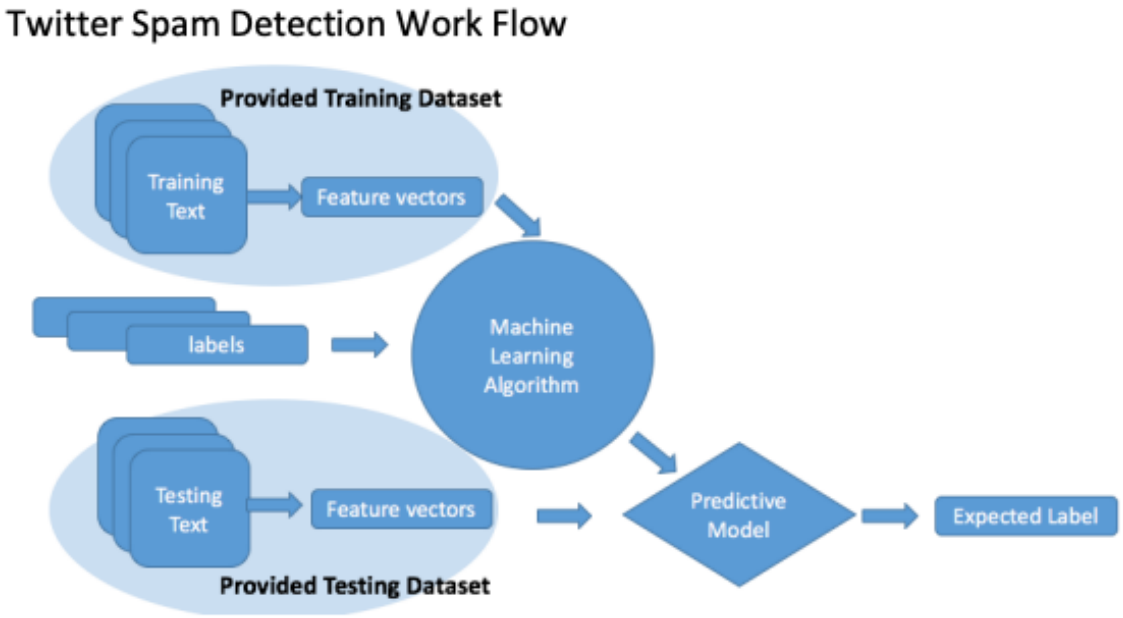

Figure 1: Twitter Spam Detection workflow

8. A technical demonstration chapter that consists of fully explained screenshots of your experiments conducted in R. You should explain each step of the classification procedure, including data processing, model training and testing, and performance calculation. You should use at least three machine learning algorithms. You can explain one algorithm in detail, and the other two can be described in short, such as "similarly, we run the other two algorithms as shown in Figure XX." This part should be within **150 words** and include screenshots of the codes. (4 points)
9. A performance evaluation chapter that evaluates the performance of classifiers. You should analyse each classifier's performance with respect to the performance metrics of your choice. In addition, you should compare the performance results in terms of evaluation metrics, e.g., accuracy, false positive, recall, F-measure, speed, and so on, for the selected classifiers and datasets. This part should be within **250 words**. (You should use at least three evaluation metrics to evaluate the performance of classifiers and compare the performance of different classifiers. You can demonstrate your experiment results in the form of table and plots) (4 points)
10. A conclusion chapter that summarizes the study, gives the study's major findings, and recommends the best-performing classification algorithm. It should be within **150 words**. (2 points)
11. A bibliography lists of all cited papers and other resources. You must use in-text citations in **Harvard style** and each citation must correspond to a bibliography entry. **There must be no bibliography entries that are not cited in the report**. (You should know the contents from this page https://www.vu.edu.au/library/get-help/referencing/referencing-guides.) (2 points)