SEGi UNIVERSITY

FACULTY OF ENGINEERING, BUILT-ENVIRONMENT, AND INFORMATION TECHNOLOGY (FOEBEIT)

BACHELOR OF INFORMATION TECHNOLOGY (HONS)

JANUARY-MAY 2024 INTAKE

TCS3393 DATA MINING

GROUP ASSIGNMENT [2-3 members per group]

This assignment is worth 25% of the overall marks available for this module. This assignment aims to help the student explore and analyse a set of data and reconstruct it into meaningful representations for decision-making.

Introduction

The online landscape is ever-evolving, with websites serving as crucial assets for businesses, organizations, and individuals. As the internet continues to grow, the need for accurate and efficient website classification becomes paramount. Understanding the nature of websites, their content, and the user experience they provide is vital for various purposes, including online security, marketing strategies, and content filtering.

Embarking on a data science project, you collaborate with a cybersecurity firm dedicated to enhancing web security measures. The firm provides you with a rich dataset encompassing various attributes of websites, including their URLs, user comments, and assigned categories. Your objective is to develop a classification model capable of accurately categorizing websites based on these variables.

The dataset includes information on the URLs of different websites, user comments associated with those websites, and pre-existing categories assigned to them. The challenge lies in creating a model that not only accurately classifies websites but also adapts to the dynamic nature of the online environment, where new types of websites constantly emerge.

Your goal is to implement advanced data analysis techniques to train a model that enhances the efficiency of web classification.

Techniques

The techniques used to explore the dataset using various data exploration, manipulation, transformation, and visualization techniques are covered in the course. As an additional feature, you must explore further concepts which can improve the retrieval effects. The datasetprovided for this assignment is related to the `website classification`.

Dataset

This dataset contains information on 1407 websites URL. It includes 3 variables that describe various categories of websites. The dataset will be analyzed using subsets of these variables for descriptive and quantitative analyses, depending on the specific models used.

Assignment Task: Websites Classification

Objective:

Develop a classification model to categorize websites using advanced data science techniques.The model should robustly classify the website based on comments stated in the dataset.

Tasks:

1. Data Exploration:
   - Conduct an initial exploration of the dataset to understand its structure, size, and variables.
   - Examine the distribution of website categories to identify any imbalances in the dataset.
   - Explore the distribution of URLs and user comments length to gain insights into the data.

2. Descriptive Analysis:

A. Basic Exploration:

- Describe the structure of the dataset. How many observations and variables does it contain?
- What are the data types of the variables in the dataset?

B. Statistical Summary:

- Provide a statistical summary of the 'Category' variable. What are the most common website categories?
- Calculate basic descriptive statistics (mean, median, standard deviation) for relevant numeric variables.

C. URL Analysis:

- Analyze the distribution of website URLs. Are there any patterns or commonalities?
- Are there any outlier URLs that need special attention?

3. Data Preprocessing:

A. Cleaning Text Data:

- Explore the 'cleaned_website_text' variable. What preprocessing steps would you take to clean text data for analysis?
- Implement text cleaning techniques and explain their importance in preparing data for text-based analysis.

B. Handling Missing Values:

- Identify if there are any missing values in the dataset. Propose strategies for handling missing values, specifically in the 'cleaned_website_text' column.

4. Visualization:

A. Category Distribution Visualization:

- Create a bar chart or pie chart to visually represent the distribution of website categories.
- How does the visualization help in understanding the balance or imbalance of the dataset?

B. Text Data Visualization:

- Generate word clouds or frequency plots for the 'cleaned_website_text' variable. What insights can be gained from these visualizations?

5. Model Development

   A. Data Mining Analysis:

   - Split the dataset into training and testing sets for model evaluation.
   - Implement various machine learning algorithms for classification, such as logistic regression, decision trees, or random forests.

   B. Training and Evaluation

   - Evaluate the performance of each model using metrics like accuracy, precision, recall, and F1-score.
   - Discuss the challenges and considerations specific to evaluating a model for website classification.

6. Advanced Techniques:

   i. Feature Engineering:

      - Propose additional features that could enhance the model's performance. How might these features capture more nuanced information about websites?

   ii. Dynamic Nature of Websites:

      - Given the dynamic nature of the online environment, how could the model adapt to newly emerging website types? Discuss strategies for model adaptation.

7. Create Dashboard, Report and Conclusions:

   - Summarize the findings, including insights gained from exploratory data analysis and the performance of the classification model.
   - How interpretable is the chosen model? Can you explain the decision-making process of the model in the context of website classification?
   - Provide recommendations for further improvements or considerations in the dynamic landscape of web classification.
   - Reflect on the challenges encountered during the analysis. What potential improvements or future work would you recommend to enhance the model's performance?

This assignment allows students to apply knowledge of data exploration, preprocessing, data modelling, and model building to solve a real-world problem in the business domain. It also encourages them to explore additional concepts for improving model performance.

Deliverables

- The complete **Python program** (source code (ipynb)) and **report** must be submitted to **Blackboard.**
- Python Script (Program Code):
  - ○ Name the file under your name and SUKD number.
  - ○ Start the first two lines in your program by typing your name and SUKD number. For example:

    # Nor Anis Sulaiman

    #SUKD20231234

○ For each question, give an ID and explain what you want to discover. For example:

a. Explore the distribution of website categories in the dataset. Are there any specific categories that are more prevalent than others?

b. Visualize the distribution of URL lengths and user comments lengths. Are there patterns or outliers that could be informative for the classification model?

c. What steps would you take to clean and preprocess the URLs and user comments for effective analysis?

d. How might you handle any missing values in the dataset, and what impact could they have on the classification model?

e. Provide descriptive statistics for key variables such as URL lengths and user comments lengths. What insights can be derived from these statistics?

f. Explore potential additional features that could enhance the model's ability to classify websites accurately.

g. How might the inclusion of features derived from URLs or user comments contribute to the overall model performance?

h. Choose a classification algorithm suitable for website classification. Explain your choice.

i. Implement the chosen algorithm using Python and relevant libraries. What considerations should be taken into account during the model implementation phase?

j. Split the dataset into training and testing sets. How would you assess the performance of the model using metrics like accuracy, precision, recall, and F1-score?

k. Discuss potential challenges in evaluating the model's effectiveness and generalization to new websites.

l. Create visualizations to interpret the model's predictions and showcase its classification performance.

Documents: Coursework Report

As part of the assessment, you must submit the project report in printed and softcopy form, which should have the following format:

A) Cover Page:

All reports must be prepared with a *front cover*. A protective transparent plastic sheet can be placed in front of the report to protect the front cover. The front cover should be presented with the following details:

- Module
- Coursework Title
- Intake
- Student name and ID
- Date Assigned (the date the report was handed out).
- Date Completed (the date the report is due to be handed in).

B) Contents:

- Introduction and assumptions (if any)
- Data import / Cleaning / pre-processing / transformation
- Each question must start in a separate page and contains:
  - Analysis Techniques - data exploration / manipulation / visualization
  - Screenshot of source code with the explanation.
  - Screenshot of output/plot with the explanation.
  - Outline the findings based on the results obtained.
- The extra feature explanation must be on a separate page and contain:

- Screenshot of source code with the explanation.
- Screenshot of output/plot with the explanation.
- Explain how adding this extra feature can improve the results.

C) *Conclusion*

- Depth and breadth of analysis
- Quality and depth of feedback on the analysis process
- Reflection on learning and areas for improvement

D) *References*

- The font size used in the report must be 12pt, and the font is Times New Roman. Full source code is not allowed to be included in the report. The report must be typed and clearly printed.
- You may source algorithms and information from the Internet or books. Proper referencing of the resources should be evident in the document.
- All references must be made using the **APA (American Psychological Association)** referencing style as shown below:
  - *The theory was first propounded in 1970 (Larsen, A.E. 1971), but since then has been refuted; M.K. Larsen (1983) is among those most energetic in their opposition………*
  - */**Following source code obtained from (Danang, S.N. 2002)*/*
    int noshape=2;
    noshape=GetShape();
- A list of references at the end of your document or source code must be specified in the following format:

  Larsen, A.E. 1971, A Guide to the Aquatic Science Literature, McGraw-Hill, London.

  Larsen, M.K. 1983, British Medical Journal [Online], Available from http://libinfor.ume.maine.edu/acquatic.htm (Accessed 19 November 1995)

  *Danang, S.N., 2002, Finding Similar Images [Online], The Code Project, *Available from http://www.codeproject.com/bitmap/cbir.asp, [Accessed 14th *September 2006]*

  Further information on other types of citation is available in *Petrie, A., 2003, UWE Library Services Study Skills: How to reference [online], England, University of Western England, Available from http://www.uweac.uk/library/resources/general/info_study_skills/harvard2.htm, [Accessed 4th September 2003].*

Assignment Assessment Criteria

Documentation (30%)		
Criteria	**Criteria**	**Marks Allocated**
Data Exploration	*Structure of the report and references.* Content: • Description and justification of the python conceptsincorporate. • Program out screenshots,graphs • Project description, limitation, and conclusion	100%
Data Manipulation		
Data Transformation		
Data Visualization		
Conclusion/ Findings		

Development Tools

Your program should be written in Jupyter Notebook or any available data analysis tools.

Academic Integrity

- You are expected to maintain the utmost level of academic integrity during the duration of the course.
- Plagiarism is a serious offence and will be dealt with according to SEGi and UCLAN University regulations on plagiarism.