COMP615 – Foundations of Data Science

ASSIGNMENT ONE

Data Exploration and Classification

Semester 1, 2024

Due Date: Midnight Sunday 14th April 2024.

Late submissions will incur a 10-mark penalty per day.

Weighting: 25% of the final course mark

Submission: When submitting the assessment, the name and student ID must be indicated on the front page of the report.

Note: This assignment must be completed individually; all work submitted must be entirely your own.

Data Exploration and Classification

This assignment aims to evaluate your data exploration and classification skills using Python. You will work with a dataset, analyse its characteristics through visualisations and statistical measures, and construct a classification model to assess its performance.

Dataset Selection: Choose one of the datasets listed below. These datasets have been selected from the UCI Machine Learning repository. To understand the dataset attributes, carefully review the UCI Machine Learning repository website information. Explore the data thoroughly and apply classification techniques to gain insights and evaluate model performance.

1. Dry Bean Dataset
2. In-vehicle coupon recommendation
3. Maternal Health Risk
4. Estimation of Obesity Levels
5. Online Shoppers Purchasing Intention

Task 1: Introduction (200-300 words) [10 marks]

Provide a statement of the problem, outlining the problem your chosen dataset addresses. The statement of the problem should briefly address the question: What is the problem that you will investigate in this assignment?

Your introduction must describe:

- The aim of your work, what are you trying to achieve, and research questions you attempted to answer.
- All assumptions that your data must meet.

Task 2: Data Exploration (500-600 words) [20 marks]

This section of your report must discuss the dataset and any features you consider relevant to the analysis and modelling task.

- How many features (attributes) and instances exist, and what data types are these?
- Provide summary statistics of the continuous numerical features.
- Perform an initial exploration of the provided dataset to assess its cleanliness. Describe the steps taken to address both data cleanliness evaluation and data cleaning strategies.
- Illustrate the features of your dataset using **meaningful** boxplots, histograms and grouped scatter plots (remember, these plots allow you to analyse the individual distribution of features and the relationship between them).
- Explain what you can learn from your data exploration and visualisations provided.

Task 3: Classification Models (500-600 words) [40 marks]

You need to create a model using the **Decision Tree Classifier** and answer the following questions based on the model built. In building the model, use the 10-fold cross-validation option for testing. Your answers need to be supported by suitable evidence, wherever appropriate. Some examples of suitable evidence are Confusion Matrices, Model Visualizations, and Model Summary Reports.

a) You are required to report your preprocessing steps. The steps should include identifying any missing/duplicate data or outliers. Provide explanations of how you dealt with them. **[5 marks]**

b) Create a model using the Decision Tree algorithm. Adjust ***two*** suitable parameters (*one at a time*) to reduce the tree's size and improve your model's accuracy. Report the accuracy score for each parameter using the plots. Provide the final optimised classification tree and describe its structure. **[12 marks]**

c) Describe the role of the two parameters in the model building you used in part b) above. Do you expect that using the same values obtained for this dataset will improve the accuracy of other datasets? Justify your answer. **[8 marks]**

d) Find the feature importance based on the final classification model and explain your findings. **[5 marks]**

e) Generate and carefully examine the Confusion Matrix and explain your findings. Provide the model summary report and discuss the metrics (accuracy, precision, recall, and F1-score). **[10 marks]**

Task 4: Results and Discussions (500-600 words) [20 marks]

Describe and analyse your classification results. Compare the performance of the models and explain which performed better and why. Evaluate the performance using confusion matrices, recall, precision, and accuracy metrics.

Report presentation [10 marks]

Your report must include the following elements:

- **Title, Full Name, and Student ID:** Clearly state your title, full name, and student ID at the beginning of the report.
- **Table of Contents:** Include a table of contents to provide an overview of the report's structure.
- **List of Figures/Tables:** Provide a list of figures and tables used in your report for easy navigation.
- **Answers to Questions:** Present your answers to the questions asked. Explain your findings, insights, and observations clearly and concisely.
- **Figures (Plots) and Tables:** Include all relevant figures and tables that support your answers. However, DO NOT include the code used to generate these visualisations and tables.

- **Informative Labels and Captions:** Ensure that all visualisations and tables have informative labels and captions with suitable resolution to help the reader understand their significance.
- **Others:** Ensure that your report includes page numbers. All figures and tables should be clear and accompanied by descriptive captions. Thoroughly proofread your report to correct any typographical errors prior to submission. Follow a consistent formatting throughout the report (font size, colour etc.,).
- **Code:** Ensure that your code is clean, well-organised, and properly commented. The code must be well commented and ready to execute without errors. Do not provide the screenshot of your codes.

Submission Instructions

Please submit the following two files **separately** as part of your assignment:

1. Python Notebook or Code File (.ipynb, .py):
2. Report File (PDF Format)

Note: the report should focus on presenting your findings and insights rather than including the code itself. Please refrain from including the code file in your report, as including code in the report will result in a penalty.