COMP809: Data Mining & Machine Learning

Assignment 1 – Part B (weight: 20%)

2021 Semester 2

Data Mining Applications

This is an **individual assignment**.
Submission: A soft copy needs to be submitted through Turnitin (a link for this purpose will be set up in Blackboard) Include your actual code **(no screenshot)** in **Appendix with appropriate comments** for each task.

Due date**: Friday 03 Sep at 12 midnight.**
Late penalty: maximum late submissions time is 24 hours after the due date. In this case, 5% **late penalty** will be applied.

AIMS

This assignment gives you an opportunity to solve two real-word data mining problems using the machine learning workbench. In the two questions given below justification of your answers carries a high proportion of the marks awarded. You are required to conduct experiments for both case studies and report them according to the specified requirements.

1. Study Area I (Dataset is bank.csv use the Bank.zip)

This application is concerned with predicting the outcome of direct bank marketing campaigns (phone calls) of a Portuguese banking. The dataset contains 17 attributes for which outcomes of subscribe a term deposit (yes/no) on a term deposit are known.

You are required to build a model using the **Decision Tree Classifier** and answer the following questions based on the model built. Use the data segment on the subscriptions whose outcomes are known. In building the model, use the **10-fold cross validation** option for testing.

Your answers below need to be supported by suitable evidence, **wherever appropriate**. Some examples of suitable evidence are the Confusion Matrices, Model Visualizations (from Python, Weka, MATLAB, or any other tools) and Summary Statistics.

a) Describe the pre-processing you have performed to prepare your data. [**4 marks]**

b) Using an appropriate method identify the **top five** most influential features in classifying this dataset. Explain the process of the chosen feature selection method. **[5 marks]**

c) Now build a model using the Decision Tree algorithm. By adjusting ***two*** suitable parameters (*one at a time*) reduce the size of the tree to not more than 10 to 15 nodes in order to improve interpretability of the model generated. Which of the two parameters yielded better accuracy while producing smaller trees? [**5 marks]**

d) Describe the role of the **two parameters** in model building that you used in b) above. Do you expect that manipulating the parameter in the same way will improve accuracy for other types of datasets? Justify your answer. [**8 marks]**

e) Provide and carefully examine the Confusion Matrix. You will notice that the client subscribed a term deposit (yes) outcome is significantly smaller than the (no) outcome. Why do you think this happens? Will a suitable visualization help to explain this phenomenon?

[8 marks]

2. Study Area II (Dataset is Autism-Child-Data.arff)

This application is from the medical domain and is concerned with diagnosis of childhood Autistic Spectrum Disorder Screening (ASDS) for a collection of individuals from whom relevant medical data has been obtained. The dataset contains 10 behavioural features (AQ-10-Child), 10 individuals characteristics, and the outcome (effectiveness of detection). The objective is to predict whether the given individual characteristics are effective in detecting the ASD cases. The effectiveness of ASDS detection is labelled as 'Yes' or 'No' in this dataset.

For this dataset you will also use both the **Decision Tree classifier** and **Naïve Bayes (NB)** algorithms to build a predictive mode for the ASDS. For both methods use the 10-fold cross validation option for testing.

a) Describe what is the autism spectrum disorder (ASD) and discuss the significance of early diagnosis of ASD. Briefly describe the Autism-Spectrum Quotient (AQ) and include two recent references to support your answer (no more than one page). **[6 Marks]**

b) Use an appropriate method of feature selection to **identify the top five significant features**. State the method used and list the features produced and explain why this feature reduction method was used. Discuss the independence assumption between the features in Naïve Bayes algorithm and support your answer with reference to the selected features.

[6 marks]

c) Run the **Naïve Bayes** algorithm with the *GaussianNB* implementation for the selected features. Provide the metrics to evaluate the performance of NB model and discuss the results. **[8 marks]**

d) Run the **Decision Tree Classifier** algorithm and compare the list produced in part (b) with the top five features produced by the Decision Tree model. Identify similarities and differences. Discuss any differences. **[10 marks]**