COMP615 – Foundations of Data Science

ASSIGNMENT TWO

Classification - group assignment

Semester 1, 2022

Due Date: Midnight Sunday 5th Jun 2022.

Late submissions will incur a 10 marks penalty per day.

Weighting: 25% of the final course mark

Submission: When submitting the assessment **the names and student IDs must be indicated on the front page of the report**.

AIMS

This assignment allows you to solve a real-world problem using the machine learning workbench. The analysis and justifications of your answers carry a high proportion of the marks awarded. Please make sure you read through the entire assignment before you start.

1 Introduction [**5 marks]**

You are expected to provide information on the dataset you are assigned to use for your assignment. Provide a statement of the problem, outlining the problem that your dataset addresses.

2 Data Exploration [**10 marks]**

In this section of your report, you are required to discuss the dataset and any missing or outliers found in the dataset.

- How many features (attributes), instances and what data types are these?
- Cleary identify the inputs and the output(Class). Do you consider your class distribution balance or not, justify your answer.
- Provide summary statistics of the continuous numerical features.
- Illustrate the features of your dataset using meaningful visualisation (eg. boxplots, histograms, etc.).

3 Decision Tree Classifier

You are required to build 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 the Confusion Matrices, Model Visualizations and Summary Statistics.

a) Now build a model using the Decision Tree algorithm. Adjusting ***two*** suitable parameters (*one at a time*) to reduce the size of the tree to improve the accuracy of your model. Report the accuracy score for each parameter using the plots. Provide the final optimized classification tree and describe its structure. [**10 marks]**

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

c) Generate and examine the Confusion Matrix carefully and explain your findings. Provide the model summary report and discuss the metrics (accuracy, precision, recall, and F1-score).

[8 marks]

d) Find the feature importance based on the final classification model and explain your findings. **[4 marks]**

4 Artificial Neural Network (ANN)

In this part, you are required to explore various architectures for building an Artificial Neural Network (ANN). In building the model, use the 10-fold cross-validation option for testing.

a) Use an appropriate feature selection method to identify the top **five most significant features**. State the method used and list the features produced. Compare the list produced in the previous section by the Decision Tree model. Identify similarities and differences. Discuss any differences. **[10 marks]**

b) Use the sklearn.MLPClassifier with default values for parameters and **a single hidden layer** with k neurons (k <=25). Use default values for all parameters other than the number of iterations. Determine the best number for iteration that gives the highest accuracy. Use this classification accuracy as a baseline for comparison in later parts of this question. **[10 marks]**

c) Plot the loss for training and test segments as a function of the iteration count and discuss your observation. **[5 marks]**

d) Experiment with **two hidden layers** and experimentally determine the split of the number of neurons across each of the two layers that gives the highest classification accuracy. In part 1, we had all k neurons in a single layer, in this part we will transfer neurons from the first hidden layer to the second iteratively in step size of 1. Thus, for example in the first iteration, the first hidden layer will have k-1 neurons whilst the second layer will have 1, in the second iteration k-2 neurons will be in the first layer with 2 in the second and so on. Summarise your classification accuracy results in a 25 by 2 table with the first column specifying the combination of neurons used (e.g., 12, 13) and the second column specifying the classification accuracy. **[15 marks]**

e) From the table created in part d of this part, you will observe the accuracy variation with the split of neurons across the two layers. Give explanations for some possible reasons for this variation. **[5 marks]**

5 Performance Evaluation

Compare the performance of the Decision Tree and MLP Classifiers on your dataset. Choose the best-performing model for your dataset and explain why you have chosen it. Discuss the overall findings from your experiments. **[5 marks]**

6 Report Presentation **[5 marks]**