COMP615 – Foundations of Data Science

ASSIGNMENT TWO
Group assignment
Semester 1, 2024

Due Date: Midnight Sunday 2nd Jun 2024.

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**.

Naming of Submitted File: DatasetName_Student1Surname_ID_ Student2Surname_ID
Appendix: Submit the code separately.

AIMS

This assignment provides an opportunity to solve two real-world data mining problems using the machine learning workbench. You are required to conduct experiments for both parts and report them according to the specified requirements. Your answers below need to be supported by **suitable evidence**, **wherever appropriate**. Some examples of suitable evidence are the Confusion Matrices, Model Visualizations and Summary Statistics.

Note: This assignment should be completed in pairs (maximum of two students).

Part A (Predicting Bank Marketing Campaign Outcomes)

This part of the assignment is concerned with predicting the outcome of direct bank marketing campaigns (phone calls) of Portuguese banking. The dataset (Bank.zip) contains 17 attributes for which outcomes of subscribing to a term deposit (yes/no) are known. You are required to build models using the K-Nearest Neighbors (KNN) and Naïve Bayes (NB) algorithms.

a) **Explain the KNN and Naïve Bayes Algorithms** **[10 marks]**
In your own words, explain how each of the KNN and Naïve Bayes algorithms work.

b) **Perform Exploratory Data Analysis (EDA)** **[10 marks]**
Perform EDA and describe your dataset. Explain any pre-processing and data manipulation tasks you performed to prepare your dataset for building your models. Note: No grade will be given for presenting plots/tables without explanation.

c) **Feature Selection and Analysis** **[10 marks]**
Identify the most influential features in classifying this dataset using an appropriate method. Explain the process of the chosen feature selection method and use the **top five features** for building your models. Use a breakdown analysis for selected features by class and describe their distribution using appropriate plots.

d) **Independence Assumption in Naïve Bayes** **[5 marks]**
Discuss the independence assumption between the features in the Naïve Bayes algorithm and support your answer with respect to the selected features.

e) **Naïve Bayes Model Building and Evaluation** **[10 marks]**
Run the Naïve Bayes algorithm with the GaussianNB implementation for the selected features. Provide evaluation metrics, including the confusion matrix, showing the performance of the NB model. Discuss the results.

f) **KNN Model Building and Evaluation** **[10 marks]**
Fit a KNN model for a range of K values. Provide and examine the confusion matrix. Generate and provide a classification report showing precision, recall, F1 score, and overall accuracy to evaluate your model performance. Discuss the results.

g) **Model Comparison** **[5 marks]**
Compare the performance of your KNN and NB models. Discuss your findings.

Part B: Exploring Artificial Neural Networks

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

Tasks

a) Activation Function and Learning Rate in MLP **[5 marks]**

Explain the role of an activation function and learning rate in building a Multilayer Perceptron (MLP).

b) Baseline Model with MLPClassifier **[5 marks]**

Use the sklearn.MLPClassifier with default parameter values and a single hidden layer with k neurons ($k \leq 25$). Determine and report the best number of iterations that gives the highest accuracy. Use this classification accuracy as a baseline for comparison in later parts of this question.

c) Tracking Loss Value **[5 marks]**

Enable the loss value to be shown on the training segment and track the loss as a function of the iteration count. Explain any observed discrepancies between loss value and error value over consecutive iterations.

d) **Experimenting with Two Hidden Layers** **[10 marks]**

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, you had all k neurons in a single layer, in this part you 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 $k\text{-}1$ 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.

e) Explaining Accuracy Variation **[5 marks]**

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

f) Comparing MLP Classifier Performance **[5 marks]**

Compare the performance of the MLP Classifier with other classifiers on your dataset in part A. Choose the best-performing model and explain why you chose it. Discuss your findings from the experiments and provide your opinion on these classifiers.

Report Resentation **[5 marks]**

Submission Instructions

Ony one submission per group is required. Please submit the following two files **separately** as part of your assignment:

1. Python Notebook (.ipynb)
2. Report File (PDF Format)

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