CN5021 – Programming for Data Science

Final Assessment (Total : 50 marks)

Deadline for Submission: Thursday 19 December 2024 at 11 am

Submission Method: Moodle Platform

Number of files to submit: 2

- **Question 1 (word document or pdf)**
- **Question 2 (.ipynb file including code and comments)**

Question 1 (25 marks)

Transportation is a necessity and globally, authorities have been embarking on new techniques to deliver **precise traffic related data,** for informed decision making, within the transportation sector. For instance, authorities require precise data about the different categories of vehicles on the roads. This data can help in addressing concerns related to congestion in different parts of road networks (intersections, flyovers, roundabouts amongst others).

You are part of a Research and Development team called Apex Research, which specialises in Intelligent Traffic Management Systems. As such, you have been requested to provide a report of how you will **apply the different phases of the Data Science Lifecycle to this project**. In your report, you will need to also include the following:

- The different sources of data
- The techniques that you can potentially use for detecting and counting the vehicles
- The types of reports that you will generate as part of this project

The description will need to be maximum of 5 pages, including relevant and appropriate diagrams and tables as required. You will need to submit your report as a word document or pdf.

Question 2 (25 marks)

As part of a team of data scientists, you have been provided with a dataset and requested to develop a KNN model to predict whether someone has a potential risk of heart disease or not.

The dataset is in the form of a csv file called **data-heart.csv**. This dataset is available on the Moodle site under Week 11 Assessment. This dataset is complete, as it is, and therefore, there is no need to check for missing values or zero values.

Your task is to create a KNN model and predict the potential risk of heart disease (You can refer to the KNN that you developed in lab week 7 for inspiration). The break down of marks is as follows:

- Importing the appropriate libraries (2 marks)
- Loading the dataset (1 mark)
- Printing the length of the dataset (1 mark)

- Printing the first 5 rows of the dataset (1 mark)
- Splitting the dataset for training and testing, 80% for training and 20% for testing (3 marks)
- Feature Scaling (2 marks)
- Identifying the appropriate number of neighbours for the KNN algorithm (2 marks)
- Defining and Fitting the model (5 marks)
- Predicting results from the test set (3 marks)
- Evaluating the model using Confusion matrix and other measures (5 marks)

Submit your code as a .ipynb file. Use the following file name StudentID-StudentFirstName-Question2.ipynb. Ensure that you include appropriate comments in your code.