ASSIGNMENT TWO

Semester 1 - 2023

PAPER NAME: Data Mining and Machine Learning

PAPER CODE: COMP809

DUE DATE: Sunday 9th Jun 2024 at midnight

TOTAL MARKS: 100

Students' Names: ……………………………………………………………………………………………………

Students' IDs: ………………………………………………………………………………………………………

- **Due date: 09 Jun 2024 midnight NZ time.**
- Late penalty: maximum late submission time is 24 hours after the due date. In this case, a **5% late penalty** will be applied.
- Include your actual code (no screenshot) in an appendix with appropriate comments for each task.

Note: This assignment should be complemented by a group of two students.
Submission: a soft copy needs to be submitted through the canvas assessment link.

INSTRUCTIONS:

1. **The following actions may be deemed to constitute a breach of the General Academic Regulations Part 7: Academic Discipline,**
   - Communicating with or collaborating with another person regarding the Assignment
   - Copying from any other student work for your Assignment
   - Copying from any third-party websites unless it is an open book Assignment
   - Uses any other unfair means
2. **Please email DCT.EXAM@AUT.AC.NZ if you have any technical issues with your submission on Canvas immediately**
3. **Attach your code for all the datasets in the appendix section.**

Part A Assessment Tasks [20 marks]

The objective of this assignment is to conduct preliminary research on data mining methods used in variousapplication domains. The survey is intended to assist you in establishing a suitable framework (application area, tools, algorithms) on which your mining project will be based.

To achieve this objective, you need to follow the steps below:

1. Select a topic based on the **application domain** listed in Section 2 topics or you can come up with your own topic but need to confirm it with the teaching team.
2. Read and analyse **recent** peer-reviewed papers (minimum of 6 articles) on your specific topic.
3. From your research identify at least two themes and discuss these themes by comparing themto the various papers in your research. Some examples of themes are listed below:
    - Approaches/algorithms to solve the problem
    - Scientific results from experimentation
    - Perspectives on an issue
    - Advantages/disadvantages
4. Express your own opinions, e.g., new ideas, proposed approaches/models, how to extend the existing work, etc. Your opinions about **machine learning and data mining related issues** should be presented.
5. Write the report using LaTex / Word. **Minimum 4 pages (including references) and nomore than 6 pages** in 2 columnsIEEE proceedings format

2. Topics

You can pick one of the following topics or come up with a topic of your interest.

- Healthcare
- Banking and Finance
- Retail, Customer Relationship Management, Product Recommendation
- Computer Vision
- Fraud Analysis

3. Layout for Research Report

The research report must include:

- Title
- Abstract
- Introduction
- Background/motivation
- Comparison of related work (from peer-reviewed sources)
- Your opinion – new ideas, proposed approaches/models, how to extend the existing work
- Conclusion and future issues
- References

Part B: Predictions of Particulate Matter ($PM_{2.5}$ or PM_{10}) [80 marks]

Air pollution causes serious damage to public health and based on existing research; particulate matter (PM) smaller than $PM_{2.5}$ is currently considered to have the strongest correlation with the effects of cardiovascular disease. Therefore, making accurate predictions of $PM_{2.5}$ is a crucial task. In this part, you are required to build prediction models based on regression model, multi-layer perceptron (MLP) and long short-term memory (LSTM).

Dataset: The dataset for this experiment can be downloaded from the Environmental Auckland Data Portal. Your dataset includes $PM_{2.5}$ / PM_{10} (Output) and different predictors such as air pollution, Air Quality Index (AQI), and meteorological data collected on an hourly basis from **only one air quality monitoring stations station** listed below:

- Penrose Station (ID:7)
- Takapuna Station (ID:23)

Two PM_lag measurements, lag_1 and lag_2, should be included in your dataset. For example, lag_1 for $PM_{2.5}$ is the measurement for the previous hour (*h-1*) and lag_2 is $PM_{2.5}$ concentration for *h-2*.

Download relevant **PM** concentration, air pollution data (**SO_2, NO, NO_2**), and meteorological data **Solar Radiation** (W/m²), **Air Temperature** (°C), **Relative Humidity** (%), **Wind Direction** (°), and **Wind Speed** (m/s)). The dataset should be **hourly measurement** starting from January 2019 to December 2023 (5 years).

Note 1: Not all mentioned independent variables are collected at these monitoring stations.

Note 2: The unit of measurement for PM and air pollution data should be (μg/m³).

Introduction and Data Pre-processing [10 marks]

Make sure your dataset all has the same temporal resolution (i.e. hourly measurement). Perform data exploration and identify missing data and outliers (data that are out of the expected range). For example, unusual measurements of air temperature of 40(°C) for Auckland, Relative Humidity measurements above 100, and negative or unexplained high concentrations are outliers.

- Introduce the problem being addresses in this assignment.
- Provide attribute-specific information about outliers and missing data. How can these affect dataset quality?
- Based on this analysis, decide, and justify your approach for data cleaning. Once your dataset is cleaned move to the next step for feature selection.

Data Exploration and Feature Selection [10 marks]

Choose **five attributes** of your dataset that has the highest correlation with $PM_{2.5}$ or PM_{10} concentration using Pearson Correlation or any other feature selection method of your choice with justification.

- Provide the correlation plot (or results of any other feature selection method of your choice) and elaborate on the rationale for your selection.
- Describe your chosen attributes and their influence on PM concentration.
- Provide graphical visualisation of variation of PM variation.
- Provide summary statistics of the PM concentration.
- Provide summary statistic of predictors of your choice that has the highest correlation in tabular format.

Experimental Methods

Use 70% of the data for training and the rest for testing the MLP and LSTM models. Use a Workflow diagram to illustrate the process of predicting PM concentrations using the MLP and LSTM models. **[5 marks]**

For both models, provide root mean square error (RMSE), Mean Absolute Error (MAE), and correlation coefficient (R^2) to quantify the prediction performance of each model.

Multilayer Perceptron (MLP)

1) In your own words, describe multilayer perceptron (MLP). You may use one diagram in your explanation (one page). **[5 marks]**

2) Use the *sklearn.neural_network.MLPRegressor* with default values for parameters and **a single hidden** layer with k= 25 neurons. Use default values for all parameters and experimentally determine the best learning rate that gives the highest performance on the testing dataset. Use this as a baseline for comparison in later parts of this question. **[5 marks]**

3) 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 accuracy. In part 2, 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. **[5 marks]**

4) From the results in part 3 of this question, you will observe a variation in the obtained performance metrics with the split of neurons across the two layers. Give explanations for some possible reasons for this variation and which architecture gives the best performance. **[5 marks]**

Long Short-Term Memory (LSTM)

1) Describe LSTM architecture including the gates and state functions. How does LSTM differ from MLP? Discuss how does the number of neurons and batch size affect the performance of the network? **[5 marks]**

2) To create the LSTM Model and determine the optimal architecture, apply Adaptive Moment Estimation (ADAM) to train the networks. Identify an appropriate cost function to measure model performance based on training samples and the related prediction outputs. To find the best epoch, based on your cost function results, complete up to 30 runs keeping the learning rate and the number of batch sizes constant (e.g. at 0.01 and 4 respectively). Provide a line plot of the test and train cost function scores for each epoch. Report the summary statistics (Mean, Standard Deviation, Minimum and Maximum) of the cost function as well as the run time for each epoch. Choose the best epoch with justification. **[5 marks]**

3) Investigate the impact of differing the number of the batch size, complete 30 runs keeping the learning rate constant at 0.01 and use the best number of epochs obtained in previous step 2. Report the summary statistics (Mean, Standard Deviation, Minimum and Maximum) of the cost.

function as well as the run time for each batch size. Choose the best batch size with justification. **[5 marks]**

4) Investigate the impact of differing the number of neurons in the hidden layer while keeping the epoch (step 2) and Batch size (step 3) constant for 30 runs. Report the summary statistics (Mean, Standard Deviation, Minimum and Maximum) of the cost function as well as the run time. Discuss how does the number of neurons affect performance and what is the optimal number of neurons in your experiment? **[5 marks]**

Model Comparison

1) Plot model-specific actual and predicted PM to visually compare the model performance. What is your observation? **[2.5 marks]**

2) Compare the performance of both MLP and LSTM using RMSE. Which model performed better? Justify your finding. **[2.5 marks]**

Report Presentation **[10 marks]**