EEN1083/EEN1085 Data Analysis and Machine Learning I

Release Date = 14-10-2024
Due Date: 29-11-2024 (5pm)
Total Marks: 25

This assignment is designed for students to demonstrate their ability to perform data analysis and machine learning with an aim to gain hands-on experience.

There are three datasets provided for this assignment.

Dataset 1: Energy (Gas and Electricity) Consumption Dataset

The Energy Consumption Dataset contains information related to the number of units of (electricity or gas) consumed at different time intervals throughout the day. The dataset consists of a full 12 months period in 2023 and 2024. The dataset can be accessed at: Shared Google Drive Folder

Dataset 2: Dublin Air Quality Dataset

This data was collected by Google and Dublin City Council as part of Project Air View Dublin. Google's first electric Street View car equipped with Aclima's mobile air sensing platform drove through the roads of Dublin City measuring street by street air quality. Driving predominantly took place Monday–Friday between 9:00 am and 5:00 pm from May 2021 through August 2022, so the dataset primarily represents typical daytime, weekday air quality. The car measured pollution on each street and highway at 1-second intervals, driving with the flow of traffic at normal speeds. The pollutants determined are: Carbon Monoxide(CO), Carbon Dioxide(CO2), Nitrogen Dioxide (NO2), NO (nitric oxide), Ozone (O3), and Particulate Matter PM2.5 (including size resolved particle counts from 0.3 - 2.5 µm). Further details of the dataset can be found at: https://data.gov.ie/dataset/google-airview-data-dublin-city

Dataset 3: Mars surface image (Curiosity Rover) Dataset

This is an image dataset containing 6691 images of the Mars surface collected by the Mars Science Laboratory (MSL, Curiosity). The dataset has labels and spans 24 classes. The provided dataset has low resolution images of roughly 256 X 256 pixels each. There is a high resolution version of the dataset available, however for this assignment you should only use this low resolution image dataset. You can get further details of the dataset at their website: https://zenodo.org/records/1049137

https://data.nasa.gov/Space-Science/Mars-surface-image-Curiosity-rover-labeled-dat a-se/cjex-ucks

Assignment Tasks:

Students enrolled in the 5 credit version of this module are required to attempt any 2 of the following 3 questions, students enrolled in the 7.5 credit module are required to attempt all 3 questions.

Q1: Data Analysis **[50 Marks]**

This question relates to Dataset 1. Using Dataset 1, you are required to;

1.1. Briefly describe the dataset. (Report) **[5 Marks]**

1.2. Formulate a real-world problem that could be solved by applying data analysis over the given dataset. (Report) **[5 Marks]**

1.3. Apply different data pre-processing techniques e.g. data cleansing, data reshaping, and data preparation. (Report + Code) **[10 Marks]**

1.4. Perform exploratory data analysis by generating relevant summary statistics. (Report + Code) **[10 Marks]**

1.5. Use a set of visualization techniques to convey the outcomes of the data analysis and convey a message for the general audience in order to address your formulated problem (Task 1.2).

(Report + Code)**[20 Marks]**

Q2: Machine Learning on Sensor Dataset **[50 Marks]**

This question relates to Dataset 2. Using Dataset 2, you are required to;

2.1. Briefly describe the dataset. (Report) **[5 Marks]**

2.2. Formulate two different real-world problems (Regression and Classification) that could be solved by applying data analysis and machine learning over the given dataset. (Report) **[5 Marks]**

2.3. Apply different data pre-processing techniques e.g. data cleansing, reshaping, resize and preparation. (Report + Code) **[5 Marks]**

2.4. Divide the dataset into training, testing and validation datasets and explain your approach. (Report + Code) **[5 Marks]**

2.5. Perform exploratory data analysis by generating relevant summary statistics and visualization. (Report + Code) **[5 Marks]**

2.6. Train a machine learning model over the given dataset to solve a regression based problem formulated in Task 2.2.

(Report + Code) **[10 Marks]**

2.7. Train a machine learning model over the given dataset to solve a classification based problem formulated in Task 2.2.

(Report + Code) **[10 Marks]**

2.8. Evaluate performance of your trained models using different error detection or accuracy measurement techniques.

(Report + Code) **[5 Marks]**

Q3: Machine Learning on Images Dataset **[50 Marks]**

This question relates to Dataset 3. Using Dataset 3, you are required to;

3.1. Briefly describe the dataset. (Report) **[5 Marks]**

3.2. Formulate two different real-world problems (Clustering and Classification) that could be solved by applying data analysis and machine learning over the given dataset. (Report) **[5 Marks]**

3.3. Apply different data pre-processing techniques e.g. data cleansing, reshaping, resize and preparation. (Report + Code) **[5 Marks]**

3.4. Divide the dataset into training, testing and validation datasets and explain your approach. (Report + Code) **[5 Marks]**

3.5. Perform exploratory data analysis by generating relevant summary statistics and visualization. (Report + Code) **[5 Marks]**

3.6. Train a machine learning model over the given dataset to solve a clustering problem formulated in Task 3.2. (Report + Code) **[10 Marks]**

3.7. Train a machine learning model over the given dataset to solve a classification based problem formulated in Task 3.2.

(Report + Code) **[10 Marks]**

3.8. Evaluate performance of your trained models using different error detection or accuracy measurement techniques.

(Report + Code) **[5 Marks]**

Submission Guidelines

The final submission should be submitted on the loop before the due deadline. A final submission should contain;

1. A report (PDF or Word document) documenting all assumptions, design decisions, and findings. Include visualizations, plots, and tables. You should strive to make your work completely reproducible using only the report document: include details on everything you tested and all results. Document and justify all design decisions.

2. Two or more separate notebooks (at least one for each question) containing all of your relevant code. In case you have tested multiple versions of the datasets (e.g. after cleansing) or trained different models for the same task, do not overwrite the previous code. You can either submit multiple versions of the notebook or separate code sections for each iteration/repetition of any task.

3. Any additional resources (e.g. datasets or installation, preparation or execution guidelines) bundled as archive file(s) e.g. zip. If your zip folder size is greater than 250MB, you have to break it down into multiple zip files ensuring each file is less than 250MB so it can be uploaded on the loop.

Important: Do **NOT** include code as images (e.g. screenshots of code) in your report. Include code snippets as text.

Plagiarism

Please read and strictly adhere to the DCU Academic Integrity and Plagiarism Policy. Note that reports are automatically checked against each other and against external web sources for plagiarism. Any suspected plagiarism will be treated seriously and may result in penalties and a zero grade (see Sec 6.2 of the DCU Academic Integrity and Plagiarism Policy). You are not allowed to copy any code (in full or parts) from jupyter notebooks provided by the dataset owner or third party for their own analysis. In case of suspected code plagiarism, you will be asked for an interview to demonstrate your programming skills and ability to code independently.

Grading

The assignment is worth 25% of the overall mark for the module. Marks will be awarded based on the quality of the resulting report. In particular, I will be checking to see if you are handling data correctly, carrying out exploratory analysis to gain insights, correctly performing model selection, and critically, documenting everything in a clear and concise way. The submitted code will also be checked to ensure that the work is your own.

Late Submission:

Please note that any last minute requests to extend assignment submission deadlines will not be entertained. Late submission will incur a **penalty of 30%** on the awarded grades. Loop will remain open for late submission until **02-12-2024 (5pm)**. No submissions will be accepted after the late submission deadline.