University for the Common Good

SCHOOL OF COMPUTING, ENGINEERING AND BUILT ENVIRONMENT
Department of Computing

MSc Big Data Technologies
MSc Financial Technology

Artificial Intelligence and Machine Learning

Module code:

MMI226824-22-A

Coursework Specification and Guidance

First Diet Coursework 1

Session 2022/2023, Trimester A
Module Leader: Jacob Koenig

This coursework is to be submitted electronically via GCULearn, no later than:
Sunday, 04.12.2022 23:59.

Please note that **you can make only one submission attempt**.
Ensure that your files to be submitted are correct and you are happy with them.
All submission attempts are FINAL.

Coursework 1

This module has two coursework components, this is Coursework 1. The pass mark for this coursework element is 45%.

In this coursework (CW) you will implement the first six steps of the Machine Learning Pipeline from as shown in the book AI with Python by Artasanchez and Prateek [1] as shown in Figure 1.

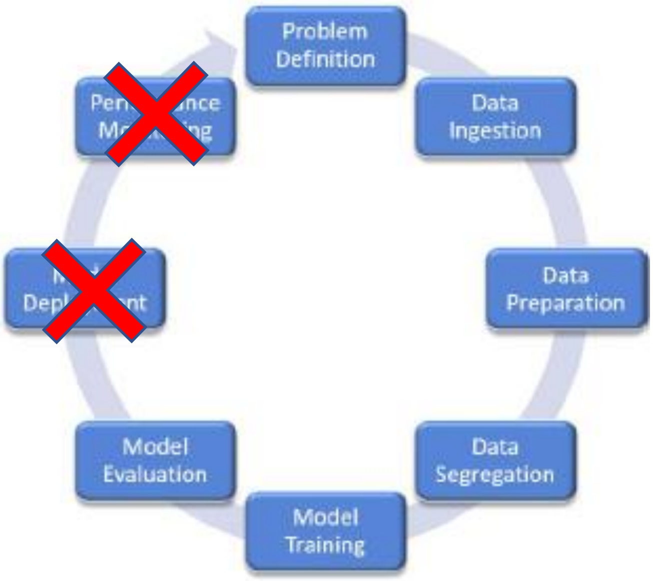


Figure 1 First six steps in a Machine Learning Pipeline [1]

You will use a modified version of the Bike Sharing dataset [2] which is posted on GCU Learn (*bike-dataset.zip*). This dataset describes the hourly and daily count of rental bikes between years 2011 and 2012 in a bikeshare system with the corresponding weather and seasonal information.

This CW will require you to outline a detailed problem definition, ingest, prepare, and segregate the data, for subsequent training and evaluation of machine learning models to predict the bike rental count.

You are required to submit a Jupyter Notebook based on the template available on GCU Learn (*template.ipynb*).

Jupyter Notebook Contents

The various sections in the notebook should include code, code comments and appropriate Markup cells describing your approach chosen.

In detail, sections should include the following:

Introduction and Problem Definition

- Textual description providing an overview over the data
- A discussion on why this problem is a regression problem
- A detailed problem statement question as discussed in Lecture 3.

Data Ingestion

- Code to load the data into a suitable format to be used in the notebook
- A description of the statistical data types for each field in the file "*bike-dataset hour.csv*"

Data Preparation

You should assume that exploratory data analysis has taken place and the following was concluded:

- Missing values are in the 'temp' and 'atemp' columns.
- The peak usage hours are: 7-9AM and 4-7PM on working days, and 10am-4pm on non-working days.
- At night (10pm-4am) the bike rentals are low
- If the humidity or wind-speed is high, the number of rentals decreases.

Your data preparation steps should therefore include the following:

- Fill the missing values in the temperature columns automatically with values that would most closely mirror the actual temperature.
- Create a new field that indicates whether it is a peak time or not
- Create a new field that indicates whether it is night time or not
- Remove all fields containing information about specific dates ('yr', 'mnth', 'dteday'), 'casual' and 'registered' and any other variables that you deem irrelevant.
- A justification (and potential application) of whether you should use data binning or not
- Suitable encoding of the data

Data Segregation

- Code and justification for the selection and application of a suitable data split

Model Training

- Selection of *two* different Regression models and justification why they are suitable. **Only one** of those models should be Tree-Based (e.g. Random Forest or Decision Tree)
- Application of those models as a baseline on the data

- Utilisation of manual **or** automatic hyperparameter optimization and justification of your choices to create "optimized" versions of each regression model

Model Evaluation

- Selection of appropriate regression metrics and a written outline why they are suitable for this data
- A comparison of the baseline models to the "optimized" versions and an evaluation of the results

Conclusion

- A conclusion and interpretation of the results and suggestion of potential improvements

You should provide sufficient written documentation in the notebooks Markdown cells to show that you understood and have justified the steps that have been implemented. Marks may not be awarded if the code has insufficient explanations and is all contained only within very few code cells.
The **maximum word limit** for all Markdown cells (excluding inline code comments or the "Sources" Markdown cell) is ***1500* Words.**

You can use machine learning libraries such as scikit-learn to assist your implementation and training of algorithms. However, ***copying code steps*** from external sources which have used the same, or similar dataset is ***not permitted*** and may lead to zero marks being awarded for this particular section. You should write the code yourself and demonstrate your understanding using the textual explanations.

Plagiarism

You should also pay attention to the university's codes and practices[1] as well as their plagiarism regulations[2].

Any kind of content (images, text, ***code***) that was copied from any source and used in your coursework **without** acknowledging the source is bad academic practice and could fall under plagiarism.
The discussion of coursework between students is encouraged but the work must be undertaken individually. ***Collusion*** (copying work between students) may result in a zero mark being recorded for everyone involved and further action being taken.

[1] https://www.gcu.ac.uk/academicquality/regulationsandpolicies/universityassessmentregulationsandpolicies/

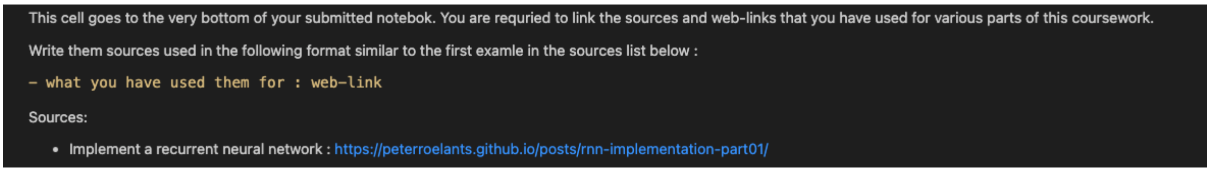
[2] https://www.gcu.ac.uk/library/smile/plagiarismandreferencing/

Sources

To avoid plagiarism or poor academic practice you need to ensure that you specify where you obtained any material you use and how you have modified it.

This **MUST** include specific web addresses (not just google.co.uk).

A template for this is included on at the very bottom of the supplied template notebook and shown in Figure 2.



This cell goes to the very bottom of your submitted notebok. You are requried to link the sources and web-links that you have used for various parts of this coursework.

Write them sources used in the following format similar to the first examle in the sources list below :

```
- what you have used them for : web-link
```

Sources:

- Implement a recurrent neural network : https://peterroelants.github.io/posts/rnn-implementation-part01/

Figure 2 Sources Cell to be included at the very end of the submission

Coursework Submission

You should submit **two files** for your coursework:

- A Jupyter Notebook with your code and markup
- A PDF version of your Jupyter Notebook.
-

Both files should follow the following naming scheme:

- CW1_LastName_FirstName_StudentID.ipynb
- CW1_LastName_FirstName_StudentID.pdf

-

If your name is *Nicola Sturgeon* with the StudentID *S12345*, name the files:

- *CW1_sturgeon_nicola_S12345.ipynb*
- *CW1_sturgeon_nicola_S12345.pdf*

Coursework files are submitted using GCU Learn.

You should double check your files before submission to ensure that you did not miss any files.

Omitting either the Jupyter Notebook or the PDF file will lead to less marks being awarded.

Marking Scheme

The marking scheme which will be used to assess the coursework is appended below.

References

[1] Artasanchez, Alberto, and Prateek Joshi. Artificial Intelligence with Python: Your complete guide to building intelligent apps using Python 3. x. Packt, (2020).

[2] Fanaee-T, Hadi, and Joao Gama. "UCI machine learning repository bike sharing data set." (2012).

Marking Rubric **Total: 100%**	**Coursework 1 – 50% of the overall module mark.**			
Topic	Fail 0%-49%	Low Pass 50%-59%	Moderate Attempt 60%-80%	Very Good Attempt 80%-100%
Introduction, Problem Definition and Conclusion **15%**	Not included or insufficient attempt up to a very limited introduction, definition of the problem and conclusion of the work.	Some introduction has been given and a very basic problem definition as well as a conclusion have been stated with some omissions.	The introduction is provided and can be used to understand the necessity of this coursework. The conclusion contains some detail.	A clear introduction and problem definition is provided supplemented with a clear justification as to why this is a regression problem and detailed concluding statements.
Data Ingestion, Preparation and Segregation **40%**	Not included or insufficient attempt up to a very poor attempt at ingesting and preparing and segregating the data.	Only some data preparation steps have been included and the textual explanations are lacking. The ingestion may have minor errors and the segregation has not been explained sufficiently	All data ingestion and preparation steps have been included but some more detail in the explanations would be necessary. The segregation includes a sensible split and justifications.	All data ingestion and preparation and segregation steps have been included and clearly outlined in detail.
Model Training and Evaluation **45%**	Not included or insufficient attempt up to a very poor attempt at selecting, training, and evaluating the models.	Model training is included but not completed according to the specification and model comparison is lacking.	Two models have been selected, trained, optimized, and compared using relevant metrics, with justified choices and some discussion	There is a clear, justified selection of models and appropriate selection of hyperparameters. The metric choice is explained in detail, the results are presented well by comparing baselines to the "tuned" versions.