School of Computing, Engineering and Built Environment

Software Development for Data Science

Module Code: MMI226822

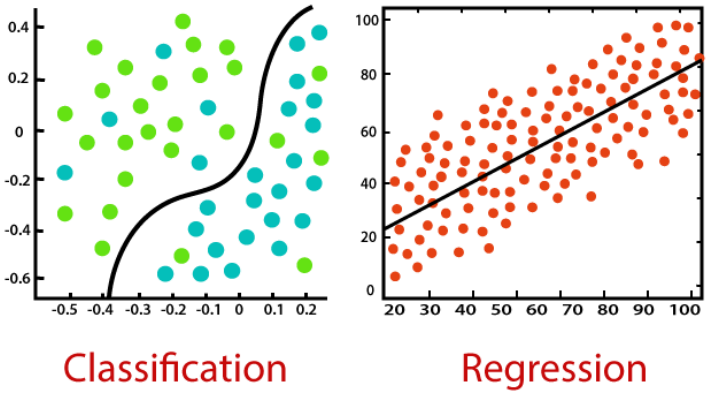
Coursework 2

Issue date: 30 November 2023

This coursework comprises 50% of the overall mark for the module.

Attention is drawn to the university regulations on plagiarism. Whilst discussion of the coursework between individual students is encouraged, the actual work has to be undertaken individually. Collusion may result in a zero mark being recorded for the coursework for all concerned and may result in further action being taken.

Machine Learning: Regression or Classification



1. Introduction

The goal of this coursework is for you to demonstrate proficiency in the Exploratory Data Analysis and Machine Learning techniques we have covered in this class (and beyond if you like) using Python and apply them to a novel dataset in a meaningful way.

In this coursework you will use Google Colab to generate a Jupyter notebook for exploring the application of Regression or Classification to analyse a dataset. Specifically, the goal of this assignment is to explore the process of selecting, implementing and evaluating either a Classification or a Regression machine learning algorithm for a specific dataset. This means that you will be required to **choose a dataset (see below), define a Classification or Regression task, select an appropriate algorithm, implement it, and thoroughly evaluate its performance**.

This will involve researching about the algorithm, *justifying* the choice of your algorithm and the evaluation metric(s), and critically evaluate and discuss the results.

2. Jupyter Notebook

This notebook should provide a discussion and demonstration of the steps you have undertaken to select the algorithm and the dataset, the training and testing of the algorithms and evaluation of the performance of the algorithms on the chosen dataset.

The notebook should include the following:

- Introduction to the topic and the dataset
- Formulation of a machine learning task (a type of classification or regression)
- The algorithm with appropriate reasoning why this was chosen
- The training and testing of the algorithm
- Performance evaluation using appropriate metrics
- Critical Analysis / Discussion
- Conclusion

It is not expected that you develop the software for the algorithm implementation, it is totally acceptable to use open-source software. You are required to detail each of the steps that you have undertaken to use the software, demonstrate the complete workflow, documentation and any appropriate visualisations as required to demonstrate the algorithm performance.

2. Datasets

You need to choose one of the following datasets for the Coursework:

1. **Customer Churn Prediction**. Customer churn refers to the phenomenon where customers or subscribers cease their relationship with a company, typically by discontinuing the use of its products or services. Churn is a crucial metric for some businesses.

The *Telco customer churn* data (IBM) contains information about a fictional telco company that provided home phone and Internet services to 7043 customers in California. It indicates which customers have left, stayed, or signed up for their service.

The dataset can be found here:
https://www.kaggle.com/datasets/blastchar/telco-customer-churn
For further details, see also:
https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113.

2. **Predicting student's dropout and academic success**. This dataset is created from a higher education institution related to students enrolled in different undergraduate degrees. The dataset can be used to identify students at risk at an early stage of their academic path, so that strategies to support them can be put into place. The dataset includes information known at the time of student's enrollment – academic path, demographics and social-economic factors, and whether they dropped out at the end of the course.

More information on this dataset can be found here:
https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

3. **515K Hotel Reviews Data in Europe.** The data was scraped from Booking.com. This dataset contains 515,000 customer reviews and scoring of 1493 luxury hotels across Europe.

More information on this dataset can be found here:
https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe

The datasets are included as csv files in the Coursework 2 Data folder on GCULearn.

3. Instructions

Here's what you will need to do:

1. **Download** the dataset of your choosing from GCU Learn/Coursework 2 folder and upload these to your Google Drive.

2. Using a Google Colab Notebook, employ the **Python** libraries that we have covered in class to **ingest, understand, clean and transform (if needed), analyse, evaluate and visualise the data**. The notebook must provide a clear demonstration of the steps you have undertaken to analyse the dataset, detailing both the methods required and the **justification** for undertaking each step.

3. Regarding the Notebook in Google Colab, please take note of the following:

   - **Sections and Headings.** Use markdown headings (**#**, **##**, **###**, etc.) to structure the report by creating different sections. Add subheadings as needed to organise content.
   - **Dataset Selection**: Choose a dataset for either a classification or regression task. The number of variables included should be complex enough to demonstrate the capabilities and challenges of the selected algorithm.
   - **Problem Formulation**: Clearly formulate the problem you aim to address, defining your objective as a well-defined machine learning task.
   - **Algorithm selection**: Select a machine learning algorithm (a type of classificatio or regression) that you believe is most suitable for your task. Justify your choice.
   - **Data Cleaning and Wrangling:** Depending on the dataset, you may need to perform data cleaning and wrangling. Identify and handle missing values, duplicates, and outliers if present in the data.
   - **Model Training**: Split your dataset into training and testing sets. Train your model on the training set and tune hyperparameters where required. Clearly document your steps.
   - **Model Evaluation an Analysis with relevant Metric(s):** Choose appropriate evaluation metric(s) for your task with justification, conduct a comprehensive evaluation considering your metric(s), and provide insightful and critical analysis of model performance.
   - **Summary, Discussion and Conclusion**: Summarise your findings, highlighting the strengths and limitations of the chosen algorithm for the given task. Propose possible improvements or future research directions.
   - Please include **references** (aim for 3-5 references).
   - **Exporting the Report:** Go to File -> Print - save as PDF, or open the Notebook in Anaconda/Jupyter, or use an online conversion tool e.g. https://www.vertopal.com/en/convert/ipynb-to-pdf.

   By following the above steps, you can create a well-structured report from a Jupyter Notebook, combining code and explanations effectively.

4. Final deliverables

The Coursework submission should be in the format of a **Jupyter notebook** and a generated **pdf report**.

Coursework reports should be submitted to GCULearn via Turnitin no later than **Friday 12th of January 2024 23.59**.

5. Marking criteria

Coursework 2 is worth 50% of the Software Development for Data Science module assessment. A rubric for Coursework 2 is provided in Figure 1.

Criteria	Excellent	Good	Acceptable	Substandard	Insufficient
Introduction, Problem understanding and formulation [/15%]	Clear introduction to the data and the problem. Demonstrates a deep understanding, defining objectives as a well-defined machine learning task.	Introduces data and the problem, showing a good understanding. Defines objectives as a machine learning task with some clarity improvements needed.	Some introduction provided. Demonstrates a basic understanding of the problem, with a somewhat vague definition of the objective and task.	Limited introduction provided. Does not show a clear understanding of the problem or formulates the task inaccurately.	Not shown or not sufficient.
Data pre-processing [/10%]	Clear demonstration of data processing steps, justifies choices, and effective use of statistics and visualisations for exploration.	Makes mostly appropriate choices but lacks some details and justification. Considers missing values, duplicates, and outliers, using generally appropriate stats and visualisations.	Data processing steps generally appropriate but lack sufficient justification and clarity. Limited consideration of missing values, duplicates, and outliers, requires further clarification.	Evidence of minimal data processing is shown, but the choices are mostly not correct and/or not sufficiently explained.	Not shown or not sufficient.
Model selection and implementation [/30%]	Selects an appropriate model, justifies the choice, and implements it effectively with appropriate techniques or optimizations. A clear explanation of the algorithm is provided.	Chooses a suitable model, provides some justification, and implements it competently. Some explanation of the algorithm is provided.	Selects an appropriate model without clear justification and demonstrates basic implementation skills. Limited or no explanation of algorithm.	Algorithm is not appropriate to the data and problem definition, and choice of algorithm is not justified.	Not shown or not sufficient.
Evaluation and Analysis with relevant metrics and visualisations [/25%]	Conducts comprehensive evaluation, considering multiple metrics, and provides insightful and critical analysis of model performance.	Conducts a reasonably thorough evaluation, considering relevant metrics, and provides a good critical analysis of model performance.	Conducts a basic evaluation with limited consideration of metrics and provides a basic analysis of model performance.	Does not conduct a meaningful evaluation or misinterprets the results.	Not shown or not sufficient (e.g. no evaluation of model performance provided).
Conclusions [5%]	Concludes with a well-detailed summary of key steps and in-depth discussion on implications or recommendations for further analysis.	Conclusion is provided with sufficient detail, but requiring some further clarification.	Basic conclusion provided, with some omissions and/or lacking detail.	Very limited onclusions shown, or the conclusions do not follow from the report.	Not shown or not sufficient (eg. no conclusion section is provided).
Code Quality and Documentation (/5%]	Code is well-organized, well-documented, and follows best practices for readability and reproducability.	Code is organized, adequately documented, and follows general best practices. A few errors or further clarification needed.	Code is presented for main sections in the report and works for the most part with some errors present. Code organization and documentation are somewhat lacking.	Code is disorganised, poorly documented, and does not follow best practices.	No evidence of Python code, or code not working.
Overall report structure and presentation [/10%]	Exceptional report with high effort, creativity, and clarity in presenting research. Professional formatting, structured with headings, clear paragraphs, and references.	Meets expectations with good effort and clarity, though some improvements are needed. Structure is generall fine and includes headings and references.	The report meets minimum requirements but lacks depth. Basic information presented for most of the material, with some omissions or errors.	Report falls short of expectations. Formatting attempts with errors, omissions, and disorganized paragraphs make the narrative hard to follow.	Not sufficient. No evidence of formatting. Narrative is entirely unclear, paragraphs are poorly organized; or no evidence that report was created using Jupyter Notebook.

Figure 1. Coursework 2 Marking Rubric.