COSC2753 - Machine Learning
Assignment 1 - Individual

Assessment Type	Individual assignment. Submit online via Canvas → Assignments →Assignment 1. Marks awarded for meeting requirements as closely aspossible. Clarifications/updates may be made via announcements/relevant discussion forums.
Due Date	Week 6, Friday 11th April 2025, 23:59 pm Late submission: 20%/day, until 15th April 2024, 23:59 pm
Marks	30%

1 Overview

This assignment is designed to help you as a student to become more confident in applying machine learning. In this assignment, you will explore a real data-set to practice the typical machine learning process which includes:

- Exploratory Data Analysis
- Selecting the appropriate ML techniques and applying them to solve a real world ML problem.
- Analyzing the output of the above algorithm(s).
- Research how to extend the modelling techniques that are taught in class.
- Providing an ultimate judgment of the final trained model that you would use in a real-world setting.

To complete this assignment, you will require skills and knowledge from lecture and lab material for Weeks 1 to 5 (inclusive). You may find that you will be unable to complete some of the activities until you have completed the relevant lab works. However, you willbe able to commence work on some sections. Thus, you can initially do the work, and

continue to build in new features as you learn the relevant skills. *A machine learning model cannot be developed within a day or two. Therefore, please start early.*

This assignment has four (04) deliverables:

Please note that you must submit all the 04 deliverables as required. However, your final grade will be based solely on your PDF report, your prediction file and presentation. That means any part of your code that is not covered or explained in your report and presentation will not be considered for grading.

1. **A PDF report** *preferably converted form of notebook, following bellow criteria:*
   - ***Bullet point format:*** Bullet point is where you raise each point in one bullet point, using clear topic, then explain the important detail as summary under the title (this description is a clear example). **A report must be no more than 05 pages**, (plus up to 2 pages for possible references and graphs).
   - ***Graphs:*** Your report should include the graphs produced by your analysis.
   - ***Markdown:*** If you are using notebook, it needs to be in the format of the provided tutorials. That means the report should include markdown text explaining the rational, critical analysis of your approach and ultimate judgement.
   - ***Specification:*** The report needs to be self-explanatory, well structured, and fulfill all the assignment specifications.
2. **A video presentation*, following bellow criteria:***
   - ***Presentation format:*** You need to use your PDF report as the basis of your presentation. In the presentation you will go through each bullet point and explain it in detail. That should include your judgment.
   - ***Presentation length:*** Your presentation should be 10 minutes (minimum of 09, and maximum of 11 minutes). You should not exceed 11 minutes, as you will be marked only based on the first 11 minutes of your presentation. You will lose mark if it is less than 09 minutes.
   - ***Must cover:*** Fulfill all the assignment specifications, based on your PDF report. You should share your window containing your PDF report while presenting, and have your camera on so your face also captured in the video.
3. **A set of predictions from your ultimate judgment**. Your prediction must be based on your final method and your ultimate judgement. The sample solution is included, the ID need to include the ID of the selected data from the test set .
4. **Your Python scripts or Jupyter notebooks** used to perform your modelling & analysis with instructions on how to run them, which need to have embedded explanatory comments. Remember that code is only used for reference, and unless you

also include your comments in the report and presentation, you will not receive any mark for them.

More detail is provided in Section.3, Assignment detail, bellow.

2 Learning Outcomes

This assessment relates to the following course learning outcomes (CLOs):

- **CLO 1:** Understand the fundamental concepts and algorithms of machine learning and applications.
- **CLO 3:** Set up a machine learning configuration, including processing data and performing feature engineering, for a range of applications.
- **CLO 4:** Apply machine learning software and tool-kits for diverse applications.

3 Assessment details

3.1 Task

In this assignment, you will develop a machine learning model to ***predict human life expectancy*** based on various attributes associated with the region of birth. Your goal is to analyze the given dataset, train different machine learning models, evaluate their performance, and determine the most suitable model for prediction. You'll be working with a dataset provided on Canvas to develop and refine your models.

- **Preprocess the Data:** Analyze and clean the dataset as needed to ensure effective model training. You need to include Exploratory Data Analysis which helps you to come up with the reasoning behind each approach's performance.
- **Approach:** You are required to use a suitable approach to find a predictive model. Your approach ***must follow the restrictions in 3.2***. And each element of the system is *justified* using data analysis, performance analysis, your analytical argument and/or knowledge from relevant literature.
- **Train Machine Learning Model:** Implement and train ML models to predict life expectancy. As one of the aims of the assignment is to become familiar with the machine learning paradigm, you should evaluate multiple different models to determine which one is most appropriate for this task, remember to **report at least three (03) models** (that follows the restrictions in 3.2)
- **Evaluate Model Performance**: setup an evaluation framework, including selecting appropriate performance measures, and determining how to split the data into training and validation data.

- **Select the Best Model:** you need to analyze the model and the results from your model using appropriate techniques and establish how adequate your model is to perform the task in real world and discuss limitation if there are any (**ultimate judgment**). Justify your choice of the most suitable regression model based on its predictive performance and interpretability. Note that *the best model is not just the one with the lowest error, but the one you can justify based on the data and evaluation.*
- **Prediction:** Finally, using the best model to predict the result for the test set.
- **Report Your Findings:** Summarize your approaches, results, and reasoning behind selecting the final model.

3.2 Restrictions

- Your models should not include the "ID" features as input, which is not an attribute.
- You may analyze feature importance using data analysis techniques. However, **removing features without proper justification** may lead to an incomplete assignment and result in **a loss of marks**.
- You must report results for at least **three different models** in this assignment.
    - **At least two models** should be selected from the techniques covered in class up to **week 5**.
    - You are **encouraged** to explore and report **one (or more) additional model(s)** beyond the techniques taught in class up to **Week 5**.

3.3 Dataset

The data set for this assignment is available on Canvas. There are the following files:

- "readme.txt": this file contains some brief description of each of the fields (attribute names)
- "train.csv": Contain the train set, attributes and target for each people. This data is to be used in developing the models. Use this for your own exploration and evaluation of which approach you think is "best" for this prediction task.
- "test.csv": Contain the test set, attributes for each people. You need to make predictions for this data and **submit the prediction via Canvas**. The teaching team will use this data to evaluate the performance of the model you have developed.
- "s1234567_prediction.csv": Shows the expected format for your predictions on the unseen test data. ***You should organize your predictions in this format. Any deviation from this format will result on zero marks for the results part.*** Changethe number "1234567" in filename to your student ID.

License agreement: The provided data is a modified version of a publicly available data source, and is subject to copyright. The dataset can only be used for the purpose of this assignment. Sharing or distributing this data or using this data for any other commercial or non-commercial purposes is prohibited.

4 Submission

You have to submit all the relevant material as listed below via Canvas.

1. **A PDF report** *preferably converted form of notebook* used for the model development including critical analysis of your approach and ultimate judgment. The report should be inPDF format. Search for instructions on converting the notebook to PDF.

2. **A video presentation** around 10 minutes (minimum of **09**, and maximum of 11 minutes). You should not exceed 11 minutes, as you will be marked only based on the first 11 minutes of your presentation. You will lose mark if it is less than **0**9 minutes.
3. **A set of predictions** from your ultimate judgment and it should be in the CSV format. Note that the file "s1234567_predictions.csv" will only show the expected format for your predictions on the unseen test data, and **please do NOT change format or order** of this file.
4. **Your Python scripts or Jupyter notebooks** used to perform your modelling & analysis with instructions on how to run them, which need to have embedded explanatory comments. Should be a ZIP file containing all the support files. It will be used for plagiarism checking.

Please name your report, video and source code by following this convention:

COSC2753_A1_YourStudentID

And your prediction file should be:

COSC2753_A1_Predictions_YourStudentID.csv

where YourStudentID is your student ID, such as s3726118

If your submission does not follow the name convention, the mark deduction will be applied.

The submission portal on canvas consists of ***four sub-pages***.

- First page for the PDF submission – ***only PDF file***
- The second page for the video presentation submisson – only video file.

- The third page for code submission. Should be a ZIP file containing source code and all the support files. We strongly recommend youto attach a README file with instructions on how to run your application. Make sure that *your assignment can run only with the code included in your zip file*!
- The fourth page for submitting predictions on test set (CSV file "s1234567_predictions.csv": shows the expected format for your predictions on the unseen test data. **Please do NOT change format or order** of this file.)

After the due date, you will have 5 days to submit your assignment as a late submission. Late submissions will incur a penalty of 20% per day. After these five days, Canvas will be closed and you will lose ALL the assignment marks.

Assessment declaration:

When you submit work electronically, you agree to the assessment declaration https://www.rmit.edu.au/students/student-essentials/assessment-and-exams/assessment/ assessment-declaration

5 Teams

Not relevant. This is an individual assignment.

6 Academic integrity and plagiarism (standard warning)

Academic integrity is about honest presentation of your academic work. It means acknowledging the work of others while developing your own insights, knowledge and ideas. You should take extreme care that you have:

- Acknowledged words, data, diagrams, models, frameworks and/or ideas of others you have quoted (i.e. directly copied), summarized, paraphrased, discussed or mentioned in your assessment through the appropriate referencing methods
- Provided a reference list of the publication details so your reader can locate the source if necessary. This includes material taken from Internet sites. If you do not acknowledge the sources of your material, you may be accused of plagiarism because you have passed off the work and ideas of another person without appropriate referencing, as if they were your own.

RMIT University treats plagiarism as a very serious offence constituting misconduct. Plagiarism covers a variety of inappropriate behaviors, including:

- Failure to properly document a source
- Copyright material from the internet or databases
- Collusion between students

For further information on our policies and procedures, please refer to the following: https://www.rmit.edu.au/students/student-essentials/rights-and-responsibilities/ academic- integrity.

7 Marking guidelines

A detailed rubric is attached on canvas. In summary:

Approach: You must adopt a suitable approach to building predictive models. At least **two models** should be chosen from the techniques covered in class between **Weeks 2 to 5**, including **linear, non-linear, and regularization methods**. Additionally, you are **encouraged to explore at least one more technique** beyond what was taught in class during this period. Each element of the approach needs to be *justified* using exploratory data analysis (EDA), performance analysis, your analytical argument, and/or published work inliterature. *This assignment isn't just about your code or model, but* ***the thought process behind your work****, why you think one model worked better than another and how you make the connection to your data analysis step.* The elements of your approach may include:

- Exploratory data analysis (EDA)
- Setting up the evaluation framework
- Selecting models, loss function and optimization procedure.
- Hyper-parameter setting and tuning
- Identify problem specific issues/properties and solutions.
- Analyzing model and outputs.

All the elements of your approach should be justified and the justifications should be visible in the PDF file (inserted as Markdown text), and your video presentation. The justificationsyou provide may include:

- How you formulate the problem and the evaluation framework.
- Modelling techniques you select and why you selected them.
- Parameter settings and other approaches you have tried.
- Limitation and improvements that are required for real-world implantation.

This will allow us to understand your rationale. We encourage you to explore this problem and not just focus on maximizing a single performance metric. By the end ofyour report, we should be convinced that of your ultimate judgment and that you have considered all reasonable aspects in investigating this problem.

Remember that good analysis provides *factual statements, evidence and justifications for conclusions* that you draw. A statement such as:

"I did xyz because I felt that it was good"

is not analysis. This is an unjustified opinion. Instead, you should aim for statements such as:

"I did xyz because it is more efficient. It is more efficient because...<evidences>..".

Ultimate Judgment & Analysis: You must make an *ultimate judgment* of the "best"model that you would use and recommend in a real-world setting for this problem. It is up to you to determine the criteria by which you evaluate your model and determinewhat is means to be "the best model". You need to provide evidence to support your ultimate judgment and discuss limitation of your approach/ultimate model if there areany in the notebook as Markdown text.

Performance on test set (Unseen data): You must use the model chosen in your ultimate judgment to predict the target for unseen testing data (provided in test.csv).Your ultimate prediction will be evaluated, and the performance of all of the ultimate judgments will be published.

Implementation: Your implementation needs to be efficient and understandable by the instructor. You should follow good programming practices.

8 Additional Information

8.1 Getting Started

To help you get started, we suggest the following:

- *Load dataset into your Jupyter or your favourite Python IDE*
- *Do some preliminary data exploration, to understand it better (this will help you later on with trying to figure which regression approach is ideal and how to improve it)*
- *Setup your data into training and testing datasets*
- *Selectthebasiclinearregressionalgorithmandtrainitthenevaluateit*
- *Analyse the results and see what is going on (to help you determine what needs to be changed to improve the regression model)*
- *Now you can continue with your method development, discussion and ultimate judgment, etc.*

8.2 Source of Help

Most questions should be asked on Canvas, however, please do not post any code. There is a FAQ, and anything in the FAQ will override what is specified in this specifications, if there is ambiguity.

Your lecturer is happy to discuss questions and your results with you. Please feel free to come talk to us during consultation, or even a quick question, during lecture break.