RMIT UNIVERSITY

COSC2753 - Machine Learning
Assignment 1

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 14th April 2023, 23:59 pm Late submission: 20%/day, until 18th April 2023, 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 ina 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 completesome of the activities until you have completed the relevant lab work.

However, you willbe able to commence work on some sections. Thus, do the work you can initially, and continue to build in new features as you learn the relevant skills. *A machine learningmodel cannot be developed within a day or two. Therefore, start early.*

This assignment has three deliverables:

1. A PDF version of the python notebook. The notebook should include markdowntext explaining the rational, critical analysis of your approach and ultimate judgment. Note: remember to start with Data Analysis, and bullet point yourfinding.
2. A set of predictions from your ultimate judgment.
3. Your Python scripts or Jupyter notebooks used to perform your modelling &analysis with instructions on how to run 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

Intensive Care Units (ICUs) are constantly challenged to monitor their patients for the risk of sepsis development (an infection that can accrue while staying in ICU). While thischallenge has been around for many years, the recent COVID-19 pandemic has increased its prominence. For an ICU, the ability to predict if a patient in ICU will develop a sepsis is very beneficial. That would assist with reducing the risk of health complications, and managing the ICU resources (such as bed availability, etc.).
In this assignment, you will develop a ML model to predict if a patient will

develop sepsis in the period of their stay in the ICU, based on provided attributes (features) related to: patient characteristics, diagnoses, treatments, services, hospital charges and patients socio-economic background.

The machine learning **task** we are interested in is: **"*Predict if a given in ICU would not develop a sepsis (Sepsis Negative / class 0) or will develop sepsis (Sepsis Positive / class 1) during their ICU stay*".**

The data set to develop your models is given to you on canvas. Note that you needto transform the target column ("Sepsis") to match the two classes mentioned in the above task. The attributes describe patients' condition, test results and age.

- In your Jupyter notebook, you need to include Exploratory Data Analysis (This stepwill help you to come up with the reasoning behind each approach's performance).
- You need to come up with an ***approach*** (that follows the restrictions in 3.2), where each element of the system is *justified* using data analysis, performance analysis and/or knowledge from relevant literature.
- As one of the aims of the assignment is to become familiar with the machine learning paradigm, you should evaluate multiple different models (only use techniques taught in class up to week 5 - inclusive) to determine which one is most appropriate for this task, remember to report at least 3.
- Setup an evaluation framework, including selecting appropriate performance measures, and determining how to split the data.
- Finally, 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**).
- Finally, predict the result for the test set.

3.2 Restrictions

- Your models should **NOT** have features (attributes) of "ID" and "insurance" fields, which are not related to patients' condition and therefore are not attributes.
- You may analyze the importance of the features based on data analysis, but please note iffeature removal is not justified then you will not be able to complete the assignment correctly and will lose mark.
- You are only allowed to **use techniques taught in class up to week 5 (inclusive)**

for this assignment. That is, you are **NOT** allowed to use ML techniques such as: Neural networks or SVM for this task.

3.3 Dataset

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

- "README.md": Description of dataset.
- "train_data.csv": Contain the train set, attributes and target for each patient. 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_data.csv": Contain the test set, attributes for each patient. 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_predictions.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 in filename to your student ID.

License agreement: 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. **The PDF version of the python notebook** used for the model development including critical analysis of your approach and ultimate judgment. Should be inPDF format. Search for instructions on converting the notebook to PDF.
2. A **set of predictions** from your ultimate judgment. Should be in CSV format. If your model predicts the patient will not develop Sepsis during their stay time at ICU, the associated "Sepsis" value in CSV should be Negative (Positive otherwise). Note that "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.

3. Your **code** (Jupyter notebooks) used to perform your analysis. Should be a ZIP file containing all the support files. will be used for plagiarism checking - notebookshould match PDF

Please name your report 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 ***three sub-pages***.

- First page for PDF- Notebook submission – ***only PDF file***
- The second 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 third 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.

Approach: You are required to use ML technique taught in class during week 2-5, including: linear, non- linear and regularization techniques. Each element of the approach need to be *justified* using exploratory data analysis (EDA), performance analysis 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 version of the notebook (inserted as Markdown text). 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 statementssuch 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 data.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.