COSC2753 - Machine Learning
Assignment 1 - Individual

Assessment Type	Individual assignment. Submit online via Canvas → Assignments → Assignment 1. Marks awarded for meeting requirements as closely as possible. Clarifications/updates may be made via announcements/relevant discussion forums.
Due Date	Week 6, Friday 12th April 2024, 23:59 pm Late submission: 20%/day, until 16th 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 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 complete some of the activities until you have completed the relevant lab work. However, you will be 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 your report summarizing the methodology, results, and conclusions drawn from the analysis. It should include some parts of your python notebook and the explanation of the rational, critical analysis of your approach, ultimate judgment.
2. A set of predictions from your ultimate judgment.

3. Your Python scripts or Jupyter notebooks containing code for data preprocessing, feature engineering, model implementation, analysis evaluation, and 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

The objective of this assignment is to apply machine learning techniques to predict the onset of diabetes in individuals based on various features. Students will propose, implement, and evaluate machine learning models to accurately predict the status of diabetes development in patients.

In this assignment, students will work on a dataset containing various health metrics and demographic information of individuals, with the target variable being the presence or absence of diabetes. The main tasks include of the followings:

Exploratory Data Analysis (EDA):

- Perform thorough exploratory data analysis to understand the distribution of features, identify correlations, and visualize patterns in the dataset.
- Handle missing values, outliers, and perform necessary data preprocessing steps.

Model Proposal:

- **Propose at least three different machine learning models** suitable for predicting the status of diabetes development in patients (only use techniques taught in class up to week 5 - inclusive).
- Justify the selection of each model based on its strengths, weaknesses, and suitability for the task.

Model Implementation:

- Implement the proposed machine learning models using appropriate libraries (e.g., scikit-learn, TensorFlow, Keras, etc.).
- Fine-tune hyperparameters using cross-validation techniques to optimize model performance.

Model Evaluation:

- Submit the prediction for patients in a given test set whose dependent variable (i.e., the presence or absence of diabetes) is hidden

Discussion and Conclusion:

- Provide a detailed discussion on the effectiveness of different machine learning models for diabetes prediction.
- Discuss the implications of the findings and potential applications of the predictive models in healthcare settings.
- Highlight areas for future research or improvements in predictive modeling for diabetes.

3.2 Restrictions

- Your models should **NOT** have features (attributes) of "Id" fields, which is not related to patients' condition and therefore is not attributes.
- You may analyze the importance of the features based on data analysis, but please note if feature 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, SVM … for this task.

3.3 Dataset

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

Feature Description (code_book.txt):

This text file provides a detailed description of the features present in both the training and test datasets. It includes information about each feature's name, data type, and potential significance in predicting diabetes. Understanding these features is crucial for data preprocessing, feature engineering, and model building.

Training Data (data_train.csv):

This CSV file contains the training data for the diabetes prediction task. Each row represents an individual with various features related to health metrics and demographic information. The

last column of the dataset indicates the target variable, i.e., the presence or absence of diabetes.

Test Data (data_test.csv):

This CSV file contains the test data for the diabetes prediction task. Similar to the training data, each row represents an individual with the same set of features. However, the target variable (diabetes status) is hidden in this file, and labels are not provided. Students are required to use their trained models to predict the status of diabetes for each individual in this dataset. The true labels will be used for evaluation but are not provided to students in this file.

Prediction file (s1234567_predictions.csv)

The file "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.*** Change the 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 version of your report summarizing the methodology, results, and conclusions drawn from the analysis. It should include some parts of your python notebook and the explanation of the rational, critical analysis of your approach**,** ultimate judgment.
2. Your Python scripts or Jupyter notebooks containing code for data preprocessing, feature engineering, model implementation, analysis evaluation, and instructions on how to run them. **Should be a ZIP** file containing all the support files and will be used for plagiarism checking.
3. A set of predictions from your ultimate judgment and it should be in the provided CSV format. If your model predicts the patient will not develop Diabete, the associated "Status" value in CSV should be 0 (and 1 otherwise). 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.

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.

Submission Instruction

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

- First page for the report in PDF version - ***only PDF file*** submission.
- 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/my-course/assessment-results/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.