Assessment Task 3: Problem solving task.

This document supplies detailed information on Assessment Task 3 for this unit.

Key information

• Due: **Sunday 24 September 2022 by 8.00 pm (AEST),**

Learning Outcomes

This assessment assesses the following Unit Learning Outcomes (ULO) and related Graduate Learning Outcomes (GLO):

Unit Learning Outcome (ULO)	Graduate Learning Outcome (GLO)
ULO3 - Perform linear regression, classification using logistic regression and linear Support Vector Machines. **ULO4 -** Perform non-linear classification using KNN and SVM with different kernels. **ULO5 -** Perform non-linear classification using Decision trees and Random forests. **ULO6 -** Perform model selection and compute relevant evaluation measure for a given problem. **ULO7 -** Use concepts of machine learning algorithms to design solution and compare multiple solutions.	**GLO1 -** through the assessment of student ability to apply advanced data processing techniques through programming for prediction. **GLO5 -** through assessment of student ability to deal with defined data set and solve problems.

Purpose

Students will be given a specific data set for analysis and will be required to develop and compare various classification techniques. Each student must demonstrate skills acquired in data representation, classification, and evaluation.

Assessment 3 **Total marks = 60**

Submission Instructions

a) Submit your solution codes into a **notebook file with ".ipynb"** extension. Write discussions and explanations including outputs and figures into a separate file and **submit as a PDF file**.

b) Submission other than the above-mentioned file formats will not be assessed and given **zero** for the entire submission.

c) Insert your Python code responses into the cell of your submitted ".ipynb" file **followed by the question** i.e., copy the question by adding a cell before the solution cell. If you need multiple cells for better presentation of the code, add question only before the first solution cell.

d) Your submitted code should be executable. If your **code does not generate** the submitted solution, then you will **get zero** for that part of the marks.

e) Answers must be **relevant and precise**.

f) No **hard coding** is allowed. Avoid using specific value that can be calculated from the dataprovided.

g) Use **topics covered till week 10** for answering this assignment.

h) Submit your assignment **after running each cell individually**.

i) The submitted notebook **file name** should be of this form _A3_studentID.ipynb". For example, if your student ID is 1234, then the submitted file name should be "_A3_1234.ipynb".

Assessment Task 3: Problem solving task.

Questions

1. What are the differences between hyperparameter and parameter of a machine learning (ML) model. Explain your answer using at least two machine learning models that you have learned in this unit. **(6 marks)**

2. Prove that Elastic net can be used as either LASSO or Ridge regulariser. **(4 marks)**

Background

The recently started human and other genome projects are likely to change the situation of molecular biology. Comprehensive analyses of whole genomic sequences will enable us to understand the general mechanisms of how protein and nucleic acid functions are encoded in the sequence data.

Dataset filename: yeast2vs4.csv

Dataset description: There are 8 features and one target in the dataset. All the features are in a numerical format, and the target is in text format. For further information about the attributes, please read "Data Set Information.pdf".

Questions

3. Analyse the importance of the features for predicting presence or absence of protein using two different approaches. Explain the similarity/difference between outcomes. **(10 marks)**

4. Create three supervised machine learning (ML) models except any ensemble approach for predicting presence or absence of protein. **(20 Marks)**

   a. Report performance score using a suitable metric. Is it possible that the presented result is an overfitted one? Justify.
   b. Justify different design decisions for each ML model used to answer this question.
   c. Have you optimised any hyper-parameters for each ML model? What are they? Why have you done that? Explain.
   d. Finally, make a recommendation based on the reported results and justify it.

N. B. Question 5 is a HD (High Distinction) level question. Those students who target HD grade should answer this question (including answering all the above questions). For others, this question is an option. This question aims to demonstrate your expertise in the subject area and the ability to do your own research in the related area.

5. Build three ensemble models for predicting presence or absence of protein. **(20 Marks)**
   a. When do you want to use ensemble models over other ML models?
   b. What are the similarities or differences between these models?
   c. Is there any preferable scenario for using any specific model among the set of ensemble models?
   d. Write a report comparing performances of models built in question 5 and 6. Report the best method based on model complexity and performance.
   e. Is it possible to build ensemble model using ML classifiers other than decision tree? If yes, then explain with an example.

Assessment Task 3: Problem solving task.