Credit TASK (Task 6.2)

About this task

Step-1 This task is designed to assess the Credit level expectations.

Step-2 Your tutor will then review your submission and will give you feedback. If your submission is incomplete the tutor will ask you to include missing parts. Tutor can also ask follow-up questions, either to clarify something that you have submitted or to assess your understanding of certain topics.

Feedback and submission deadlines

Feedback deadline: Friday 5 Sep (No submission before this date means no feedback!)
Submission deadline: Before creating and submitting portfolio.

Required documents

Execute your code into a jupyter notebook (.ipynb file) and keep the output, write a report (.pdf file) to answer the following questions, and submit your code and report to OnTrack.

Background: Cirrhosis results from prolonged liver damage, leading to extensive scarring, often due to conditions like hepatitis or chronic alcohol consumption. The data provided is a subset sourced from a Mayo Clinic study on primary biliary cirrhosis (PBC) of the liver carried out from 1974 to 1984.

This is a dataset to develop and validate machine learning algorithms for predicting the survival status of the collected patients. There are 418 patients in the data set, and each patient has 17 collected features. The aim of this task is to utilize 17 clinical features for predicting survival state of patients with liver cirrhosis. The survival states include 0 = D (death), 1 = C (censored), 2 = CL (censored due to liver transplantation)

1. Load and explore the dataset.
    a. Use an appropriate method to deal with the missing values for the data set.
    b. Split the data set into training and test set with a ratio of (8:2).
    c. Based on the training and test data, show the feature types and indicate which features are continuous or categorical.
    d. Do necessary encoding for the categorical features.
    e. Show the label distribution based on the training data, is it a balanced training set?

2. Based on the **pre-processed training data from question 1**, create three supervised machine learning (ML) models for predicting "Status".

    a. Use an appropriate validation method, report performance score using a suitable metric. Is it possible that the presented result is an underfitted or 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. Use a method to deal with the label imbalance issue and indicate whether there is a model improvement after you balance the dataset.
    e. Finally, make a model recommendation based on the reported results and justify it.

3. Use the best model that you get from question 2, do prediction on the pre-processed test set and report the model performance.

4. Analyse the importance of the features for predicting “Status” using two different approaches. Give statistical reasons of your findings.