Distinction TASK

About this task

Step-1

This task is designed to assess the Distinction level expectations. It is a classification task, please get familiar with the task and data set, and then answer the questions with your code and report.

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: Monday 13 May (No submission before this date means no feedback!)
Submission deadline: Before creating and submitting portfolio.

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 312 patients in the data set (224 for train and 88 for test), 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)

Specifically, the problem you are going to solve is: Can you

- Accurately predict the survival status given the labelled data?
- Well explain your prediction and the associated findings? For example, identify the key factors which are strongly associated with the response variable, i.e., survival status.

Data set

The training data contains 224 rows and the test data contains 88 rows, each of which have 19 columns (excluding the ID column): the N_Days attribute is the number of days between registration and the earlier of death, transplantation, or study analysis time in July 1986, the status attribute is the target variable that we will predict, and the rest 17 columns can be used as the input features. The details of the original data set can be found and downloaded in the original UCI repository. The values of the "status" column in the test set is leaved with empty to simulate real world predictions.

Evidence of Learning – SIT307/SIT720

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.

1. Load and explore the training and test dataset, do necessary pre-processing.
    a. Show both training and test dataset size.
    b. Based on the training and test data, show the feature types, and indicate which feature has missing values.
    c. Use an appropriate method to deal with the missing values for both the training and test set.
    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. What can you do with the label imbalance issue?
    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. Save your prediction (the prediction should contain two columns only: testID and Status), and submit it to the specific Kaggle in-class platform, do a screenshot of your model performance and report it.

4. (This question is for SIT720 students only) Analyse the importance of the features for predicting "Status" using two different approaches. Give statistical reasons of your findings.