SCS4204/IS4103/CS4104 - Data Analytics

Assignment 2

The given data set in the UGVLE was collected from North East of Andhra Pradesh, India. This dataset has 894 samples and the Training Dataset consists of 583 samples and the Testing Dataset consists of 311 samples.

The Training Dataset set contains 416 liver patient records and 167 non-liver patient records. The given dataset has only two (2) classes, **Yes** for the liver patient and **No** for the non-liver patient. There are eleven (11) attributes available in the dataset and 'Gender' and 'Class' attributes are nominal attributes while all the others are numerical attributes. The last attribute is a class field used to divide the dataset into two groups as a liver patient or not. This dataset contains several missing (unavailable) attribute values, denoted by "?".

Class distribution of the Training and the Test datasets are as follows.

	Training Dataset	Test Dataset
Yes	416	221
No	167	90
	583	311

Attribute Information:

1. Age - Age of the patient
2. Gender - Gender of the patient
3. TB - Total Bilirubin
4. DB - Direct Bilirubin
5. ALK - Alkaline Phosphatase
6. SGPT - Alamine Aminotransferase
7. SGOT - Aspartate Aminotransferase
8. TP - Total Proteins
9. ALB - Albumin
10. AG_Ratio - Albumin and Globulin Ratio
11. Class - Used to split the data into two sets (labeled by the experts)

Download the file "**SCS4204_IS4103_CS4104 _dataset.xlsx**" in the UGVLE for the Training and Testing datasets.

You are required to **train a supervised machine learning model** for the given data set using **any software tool** available.

Prepare a PDF file including,

a. Screenshots with an explanation of the tools you used for the above training process.
b. Brief explanation of the pre-processing steps you followed.
c. Generated *Confusion matrix* for the Test dataset.
d. List of below measures calculated for the Test dataset.
   I. Accuracy
   II. Precision
   III. Sensitivity
   IV. Specificity
   V. Error Rate
e. Append your full code lines at the end of the PDF file.

Submit a Zip file including the above-prepared **PDF file** and **all code files** you implemented.

Don't just copy-paste someone else's code for the sake of completion. If you are found to commit plagiarism or an act of cheating, you will be penalized depending on the circumstances.

Deadline: Submit on or before 11.55 PM, 14th August 2022 to the UGVLE.