**TOTAL MARKS:70**

**ATTRIBUTE INFORMATION:** This data set contains 416 liver patient records and 167 non liver patient records. The data set was collected from north east of Andhra Pradesh, India. Selector is a class label used to divide into groups (liver patient or not). This data set contains 441 male patient records and 142 female patient records.

- Age of the patient: Any patient whose age exceeded 89 is listed as being of age "90
- Gender of the patient
- TB Total Bilirubin: A bilirubin test measures the amount of bilirubin in your blood. It’s used to help find the cause of health conditions like jaundice, anaemia, and liver disease. Bilirubin is an orangeyellow pigment that occurs normally when part of your red blood cells break down.
- DB Direct Bilirubin: Higher than normal levels of bilirubin may indicate different types of liver problems.
- Alkphos Alkaline Phosphotase: If you show signs of liver disease or a bone disorder, your doctor may order an alkaline phosphatase (ALP) test to measure the amount of the enzyme in your blood and help in diagnosing the problem.
- Sgpt Alamine Aminotransferase: The alanine aminotransferase (ALT) test is a blood test that checks for liver damage.
- Sgot Aspartate Aminotransferase: AST is an enzyme your liver makes. Other organs, like your heart, kidneys, brain, and muscles, also make smaller amounts. AST is also called SGOT.
- TP Total Protiens
- ALB Albumin: An albumin blood test is a type of liver function test. Liver function tests are blood tests that measure different enzymes and proteins in the liver, including albumin. An albumin test may also be part of a comprehensive metabolic panel, a test that measures several substances in your blood.
- A/G Ratio Albumin and Globulin Ratio
- Selector field used to split the data into two sets (labeled by the experts): Selector is a class label used to divide into groups(liver patient or not).

1. **Read the dataset (tab, csv, xls, txt, inbuilt dataset)**
2. **Summarize important observations from the data set (5 MARKS)**

   *Some pointers which would help you, but don’t be limited by these*

   a. *Find out number of rows; no. & types of variables (continuous, categorical etc.)*

   b. *Calculate five-point summary for numerical variables*

   c. *Summarize observations for categorical variables – no. of categories, % observations in each category*

3. **Check for defects in the data. Perform necessary actions to 'fix' these defects (5 MARKS)**

   ***Some pointers which would help you, but don't be limited by these***

   a. *Do variables have missing/null values?*
   b. *Do variables have outliers?*
   c. *Is the Target distributed evenly? Is it a defect? If Yes, what steps are being taken to rectify the problem.*

4. **Summarize relationships among variables (10 marks)**

   a. *Plot relevant categorical plots. Find out which are the variables most correlated or appear to be in causation with Target? Do you want to exclude some variables from the model based on this analysis? What other actions will you take?*
   b. *Plot all independent variables with the target & find out the relationship? Perform the Relevant Tests to find out if the Independent variables are associated with the Target Variable.*

      *Hint: based on your observations you may want to transform features or create additional features.*

5. **Split dataset into train and test (70:30) (5 MARKS)**

   *a. Are both train and test representative of the overall data? How would you ascertain this statistically?*

6. **Fit a base model and explain the reason of selecting that model. Please write your key observations. (15 MARKS)**

   a. *What is the overall Accuracy? Please comment on whether it is good or not.*
   b. *What is Precision, Recall and F1 Score and what will be the optimization objective keeping in mind the problem statement.*
   c. *Which variables are significant?*
   d. *What is Cohen's Kappa Value and what inference do you make from the model*
   e. *Which other key model output parameters do you want to look at?*

7. **How do you improve the accuracy of the model? Write clearly the changes that you will make before re-fitting the model. Fit the final model. (20 MARKS)**

   *Please feel free to have any number of iterations to get to the final answer. Marks are awarded based on the quality of final model you are able to achieve.*

8. **Summarize as follows (10 MARKS)**

   1. *Summarize the overall fit of the model and list down the measures to prove that it is a good model*
   2. *Write down a business interpretation/explanation of the model – which variables are affecting the target the most and explain the relationship. Feel free to use charts or graphs to explain.*
   3. *What changes from the base model had the most effect on model performance?*
   4. *What are the key risks to your results and interpretation?*