ML2 Internal Graded Assessment

Answer the following questions for the following datasets:

1. **Campus Recruitment (Placement Dataset)**
2. **Patient Treatment (data-ori)**
3. **Heart disease data**
4. **Bank Client Data**
5. **Employee data**

SECTION A

1. Data Understanding (10 marks)

a. Read the dataset (tab, csv, xls, txt, inbuilt dataset). What are the number of rows and no. of cols & 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.

d. Check for defects in the data such as missing values, null, outliers, etc and also check for class imbalance.

SECTION B

2. Data Preparation (15 marks)

a. Fix the defects found above and do appropriate treatment if any.

b. Visualize the data using relevant plots. Find out the variables which are highly correlated with Target?

c. Do you want to exclude some variables from the model based on this analysis? What other actions will you take before moving ahead with model creation?

d. Split dataset into train and test (70:30). Are both train and test representative of the overall data? How would you ascertain this statistically?

SECTION C

3. Model Building (15 marks)

a. Fit a base model and explain the reason of selecting that model. Please write your key observations.

b. What is the overall Accuracy? Please comment on whether it is good or not.

c. Evaluate the model built using Precision, Recall and F1 Score and what will be the optimization objective keeping in mind the problem statement.

d. 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.

e. 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.