SUPERVISED LEARNING CLASSIFICATION

TOTAL MARKS: 30 **DURATION: 2 HOURS**

INSTRUCTIONS: -

1. *Candidates should answer all the questions in the same order provided in the question paper.*
2. *Any activity that compromises the integrity of the examination will not be permitted.*
3. *Students should complete the examination within the provided timeline.*
4. *Candidates are expected to check and ensure that the correct answer file (in. ipynb format) is uploaded in LMS.*

DATASET INFORMATION: (Diabetes.CSV): The datasets consist of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age and etc.

Column Name	Discription
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body mass index (weight in kg/(height in m)^2)
DiabetesPedigreeFunction	Diabetes pedigree function
Age	Age (years)
Outcome	Class variable (0 or 1)

SECTION A: 5 MARKS

1. **Data Understanding (5 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.)? (1 MARK)*
   b. *Calculate five-point summary for numerical variables (1 MARK)*
   c. *Summarize observations for categorical variables – no. of categories, % observations in each category. (1 mark)*
   d. *Check for defects in the data such as missing values, null, outliers, etc and also check for class imbalance. (2 marks)*

SECTION B: 10 MARKS

2. **Data Preparation (10 marks)**
   a. *Fix the defects found above and do appropriate treatment if any. (3 marks)*
   b. *Visualize the data using relevant plots. Find out the variables which are highly correlated with Target? (3 marks)*
   c. *Do you want to exclude some variables from the model based on this analysis? What other actions will you take? (2 marks)*

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? (2 marks)*

SECTION C: 15 MARKS

3. **Model Building (15 marks)**

a. *Fit a base model and explain the reason of selecting that model. Please write your key observations. (3 marks)*

b. *What is the overall Accuracy? Please comment on whether it is good or not. (2 mark)*

c. *Evaluate the model built using Precision, Recall and F1 Score and what will be the optimization objective keeping in mind the problem statement. (3 marks)*

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. (5 marks)*

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. (2 marks)*