SUPERVISED LEARNING REGRESSION

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: (Fish.csv) This dataset is a record of 7 common different fish species in fish market sales.

1. **Species:** Species name of fish
2. **Weight:** Weight of fish in gram
3. **Length1:** Vertical length in cm
4. **Length2:** Diagonal length in cm
5. **Length3:** Cross length in cm
6. **Height:** Height in cm
7. **Width:** Diagonal width in cm (dependent variable)

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. (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 variable? (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 B: 10 MARKS

3. **Model Building (15 marks)**
    a. *Fit a base model and observe* the overall R- Squared, RMSE and MAPE values of the model. Please comment on whether it is good or not. (3 marks)

b. Check for multi-collinearity and treat the same. (2 marks)

c. How would you improve the model? Write clearly the changes that you will make before re-fitting the model. Fit the final model. (6 marks)

d. 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)

e. What changes from the base model had the most effect on model performance? (2 marks)