Question Set 3 | Target Course Outcome: CO2

Department of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham – Coimbatore

III Year B.Tech. CSE V Sem B Section
19CSE305 – Machine Learning
Lab Evaluation – I

Instructions:

- There are 5 questions in the set and the total mark is 30.
- The questions should be neatly worked out in Google Colab/Anaconda's Jupyter IDE and it needs to be made sure that the python notebook is named as Roll_Number_Eval1.ipynb (e.g. CB.EN.U8CSE96108_Eval1.ipynb).
- The pdf (the file name should be same as the name of ipynb) exported version of ipynb should be uploaded to the outlook form whose link is: **https://forms.office.com/r/A1gPEXnygA**.
- The original i-python notebook should be uploaded to the assignment created in AUMS by name **19CSE305_ML_Lab_Eval_1.**

---

1. **Loading the dataset (3 x 1 = 3 marks)**
    a. Load the Advertising dataset given along with the question set in a file named "advertising.csv" and display the shape for dependent and independent variables.
    b. Calculate and print the central tendency measures, such as the means of the columns "Radio" & "Newspaper", and the median of "TV".
    c. Count and print the number of observations that have values for the column "Sales" between 10.5 and 15.5 (both inclusive).

2. **Basic operations with data frame (use the df created for question 1) (4 x 1 = 4 marks)**
    a. List out the first 8 instances, last 6 instances and 100 random instances of df.
    b. Display the column data types and memory usage information of the df.
    c. Generate and display descriptive statistics for all columns (both discrete if any & continuous) in the df.
    d. Print the number of null values present in each column of df.

3. **Data visualization (use the df created for question 1) (3 x 2 = 6 marks)**
    a. Draw a pairplot to show the correlation between all possible pairs of variables in df and write your observation in a text cell.
    b. Draw lmplot from Seaborn for each of all three independent variables versus the output variable (the three regression plots should be rendered in a 1 X 3

grid as subplots) and delineate which of the three independent variables you chose has the highest correlation with the dependent variable.

c. Draw two distplots for the column “Sales” in df in such a way that plot 1 uses 50 as number of bins and plot 2 uses 70 as number of bins.

4. **Simple LR Model Training (use the df created for question 1) (8 marks)**
    a. Separate out the input (only one input variable that was found to have the highest correlation with the output variable) and output variables from df. (1 mark)
    b. Perform train-test-split with 20% of instances to be in the test set and rest should belong to the train set. (1 mark)
    c. Create a simple linear regression model and fit it to the training data and get the predictions made for the test set as soon as the training is done. (2 marks)
    d. Print the slope and intercept of the trained model. (1 mark)
    e. Investigate the model’s performance in terms of MAE, MSE, & RMSE. (3 marks)

5. **Multiple linear regression and Gradient Descent parameter estimation (use the df created for question 1) (9 marks)**
    a. Train a linear regression model in manner that the model takes all input features in order to accurately make predictions for the output variable (separating of inputs and the output from df, train-test split, model training, prediction, printing the coefficients, and reporting the model skill through performance metrics). (8 marks)
    b. Write your observation about SLR Vs. MLR in a text cell of your ipynb. (1 marks)
