Amrita VishwaVidyapeetham

Amrita School of Computing, Coimbatore

Lab Evaluation -3,

Fifth Semester, Computer Science and Engineering

19CSE304 Foundations of Data Science

Duration: One hour Maximum: 20 Marks

Note:

- Your submitted solution for both parts (Part1 and Part2) should be in a single .ipynb file.
- Each question carries 2 marks.
- State the question numbers against your answers clearly.

Part1(10marks)

Consider the dataset "data1.csv". Sales are proposed to be enhanced with advertisement in TV, Radio and Newspaper.

1. Draw (a) a scatter plot of money spent on *TV advertisements* versus *Sales* (b) Pair plots and Heatmap.
2. Develop a Linear Regression model based on money spent on TV advertisements versus Sales.
3. With the regression line so developed, predict the sales that can be anticipated based on the money spent on TV advertisements.
4. Draw the Regression Line superimposing on the data.
5. Employ statsmodels.api and run an OLS regressor on the data. Plot the line of regression and residuals employing libraries of statsmodel. Comment on the heteroscedasticity.

Part2 (10marks).

Consider the dataset "data2.csv". Columns A through H are various health parameters. Label (Column I) indicates the presence of diabetes(1) or its absence(0).Objective is to prepare a KNN classifier for the dataset.

1. Split the data for training and testing in the ratio of 80:20.
2. Rescale the distribution of values so that the mean of observed values is 0 and the standard deviation is 1.
3. Develop a KNN classifier model and predict for the test data.
4. Draw up the confusion matrix.
5. Identify an optimum k value based on minimum mean errors (consider a range of 20). Draw a corresponding graph between Mean error and k-value.

==========