BIN2023R01 – INTRODUCTION TO DATAMINING & MACHINE LEARNING FOR BIOINFORMATICS

Lab Exercise 5- Multiple linear regression

Aim: To perform multiple linear regression of the given dataset

Procedure:

1. Import necessary libraries.
2. Load the given dataset.
3. Drop the unnecessary columns for building the model and understand the data distribution with Seaborn
4. Evaluate the quality of the datasets by checking the presence of any missing values, and eliminate outliers.
5. Perform data pre-processing such as normalization and standardization.
6. Separate independent and dependent variables.
7. Split dataset into training and testing datasets.
8. Create and fit the linear regression model.
9. Predict the Test set results.
10. Evaluate the model using appropriate performance metrics.

Questions:

1. Distinguish simple linear regression and multiple linear regression.
2. How many independent and dependent variables are considered in multiple linear regression?
3. How does the complexity of the relationship between variables differ between simple and multiple linear regression?
4. What are the limitations of simple linear regression compared to multiple linear regression?
5. Provide examples of scenarios where simple linear regression and multiple linear regression would be appropriate.
5. How does the code define the independent and dependent variables? Explain the process of fitting a multiple linear regression model in the code.
6. How is the performance of the multiple linear regression model evaluated?
7. What are the preprocessing steps performed on the data before fitting the multiple linear regression model?
8. What improvements or modifications could be made to enhance the performance of the model?

Soft copy deadline: 26th February 11:59PM
Hard copy deadline: 27th February 3:15PM