UMBC

AN HONORS UNIVERSITY IN MARYLAND

Department of Computer Science and Electrical Engineering

CMSC 462 – Introduction to Data Science

Assignment 3

Due: 10/20/2024

Total Points - 45

This assignment consists of two parts. For each part, please answer all the questions in a single document. Also submit both the R files. You will find both the datasets in the CourseDataSet folder in MS Teams.

Using the CovidMortality dataset do the following:

1. As you can understand that there is no effective way to use the name of the state. There are many other ways to use that in the multiple linear regression model. As an example, you can divide the states into regions. You can also divide the states into say blue states and red states or you can divide the states into high income and low income. Please use one of such way and explain why you did that. [10]
2. Build a multiple linear regression model, use deaths as dependent variables. Discuss, the model and about different independent variables. Discuss that by creating your own categorical variable did it make the model better? I want a detailed discussion. [15]

Using LendingClub dataset, please do the following

1. Do the descriptive statistics. Your goal is to tell a story about this dataset. I don't have a specific answer in mind. So be creative. [5]

2. Build Logistic regression and Naïve Bayes model with the data. Paste your outputs here. [5]

3. Compare the models and discuss about each model's accuracy. Make sure you also include ROC in your discussion. Explain the outputs that you pasted above in details. [10]