Programming Assignment 04

Assignment:
You are the new junior member of a data mining consulting firm. A new client has come to the firm with a new project. They have health data and would like you to help them draw a connection between key biological metrics and Chronic Heart Disease (CHD).

The client has provided anonymized patient data that contains:
Age (Age)
Count of Cigarettes Per Day (cigPerDay)
Total Cholesterol (totChol)
Systolic Blood Pressure (sysBP)
Diastolic Blood Pressure (diaBP)
Body Mass Index BMI
Heart Rate (heartRate)
Blood Glucose level (glucose)

The client has classified each patient as either having CHD (CHD = 1) or does not have CHD (CHD = 0).

The details:
There are missing data values. Determine if there is a relation between the attribute that is missing the data and another attribute, then use Simple Linear Regression to impute values to fill in the missing data.

Partition the data into a training set and a test set using an 80% / 20% split

Create a model for the client using k Nearest Neighbor. Use three different values for k.

Evaluate the performance of the three different models using Confusion Matrices and Accuracy and Error rates. Then select which of the three models should be presented to the client.

Post a report on the model using the corporate report template into Canvas.