Lab Assignment 3 – Logistic Regression

Problem Statement:

World Health Organization has estimated 12 million deaths occur worldwide, every year due to Heart diseases. Half the deaths in the United States and other developed countries are due to cardio vascular diseases. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications. This research intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using logistic regression.

Goal: To predict whether the patient has 10-year risk of future Coronary Heart Disease (CHD).

1. Read heart disease dataset from https://github.com/Ravjot03/Heart-Disease-Prediction/blob/master/framingham.csv
   Dataset Description: It is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The dataset provides the patients' information. It includes over 4,000 records and 15 attributes.

   Framingham Heart study dataset includes several demographic risk factors:-
   i. sex: male or female
   ii. age: age of the patient
   iii. education: levels coded 1 for some high school, 2 for a high school diploma or GED, 3 for some college or vocational school, and 4 for a college degree.
   iv. currentSmoker: whether or not the patient is a current smoker
   v. cigsPerDay: the number of cigarettes that the person smoked on average in one day.
   vi. BPMeds: whether or not the patient was on blood pressure medication
   vii. prevalentStroke: whether or not the patient had previously had a stroke
   viii. prevalentHyp: whether or not the patient was hypertensive
   ix. diabetes: whether or not the patient had diabetes
   x. totChol: total cholesterol level
   xi. sysBP: systolic blood pressure
   xii. diaBP: diastolic blood pressure
   xiii. BMI: Body Mass Index
   xiv. heartRate: heart rate
   xv. glucose: glucose level
   xvi. **TenYearCHD:** 10 year risk of coronary heart disease CHD (**TARGET VARIABLE**)

2. Do the exploratory analysis of the dataset :
    - Perform univariate analysis by plotting various charts like: bar charts, distribution plots, boxplots.
    - Perform multivariate analysis
3. Impute the missing values if any.
4. Remove any undesirable feature from the dataset.
5. Check for the outliers in the columns and treat the outliers if present.
6. Split the dataset into train and test.
7. Construct logistic regression model to predict the heart disease and check the accuracy scores for train and test subsets.
8. Use cross validation and stratified cross-validation to construct another logistic regression models and compare the accuracy values of all constructed models.
9. Look for real world applications where you can apply logistic regression.