**Section 1: Case Study Report**

**Introduction to the problem**

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 cardiovascular diseases. The most behavioral risk factors for cardiovascular disease and stroke are unhealthy food, lack of physical activity, smoking, and alcohol drinking. A heart attack occurs when the heart’s blood circulation is obstructed by arteries plaque build-up. A thrombus in an artery causes a stroke by impeding blood flow to the brain. The symptoms are common to other illnesses and might be confused with indicators of ageing, making diagnosis difficult for practitioners.

**Justify why it is important/useful to solve this problem**

Precision prediction and timely identification of cardiac disease are essential for improving patient survival rate. Because of the increased collection of medical data, practitioners now have a great opportunity to promote healthcare diagnosis. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications.

**How would you pitch this problem to a group of stakeholders to gain buy-in to proceed?**

The current methods for predicting and diagnosing cardiac disease are mostly dependent on practitioners’ evaluation of a patient’s medical history, signs, and physical assessment reports. Nowadays, information about patients with clinical reports is widely accessible in databases in the healthcare field, and it is rising rapidly day by day. This research intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using logistic regression.

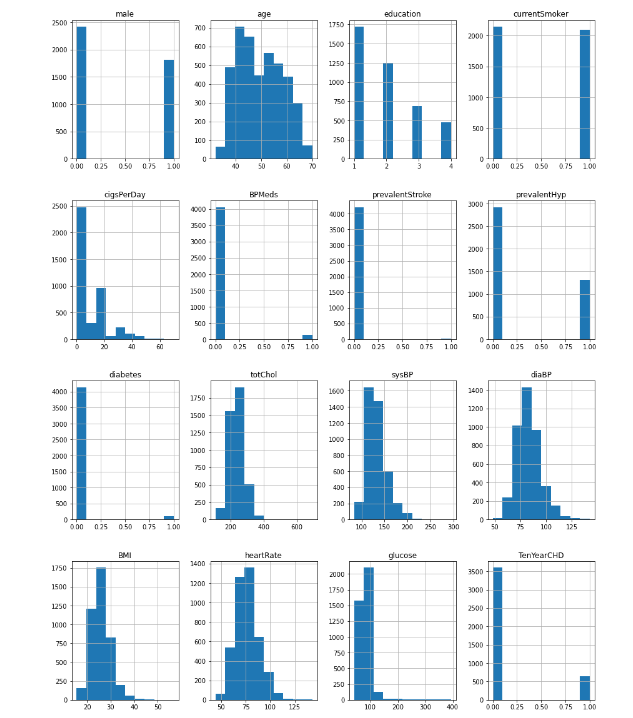
**Explain where you obtained your data**

The dataset can be downloaded from the link <https://raw.githubusercontent.com/matackett/sta210/master/data/framingham.csv> and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. Our classification goal is primarily to predict whether the patient has 10-year risk of future coronary heart disease (CHD).The dataset provides the patients’ information. It includes over 4,000 records and 15 attributes. Predict variable (desired target) in our data set is 10 year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”). As part of the Analysis on the missing values, we'll identify the no. of rows that have null values for at least one column and drop them if the

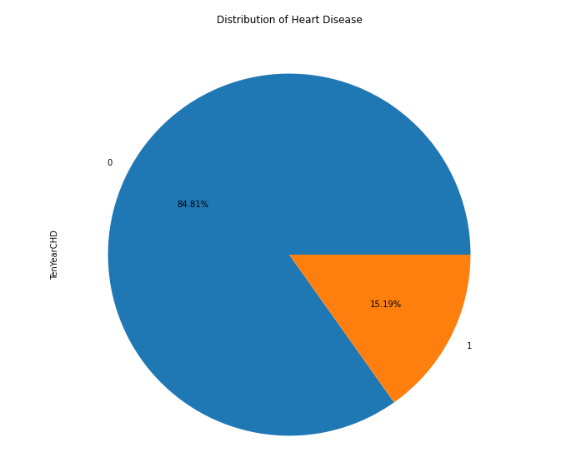
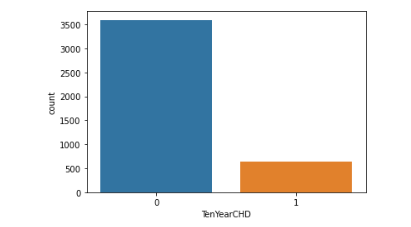
**Section 2:**

**Milestone 1**

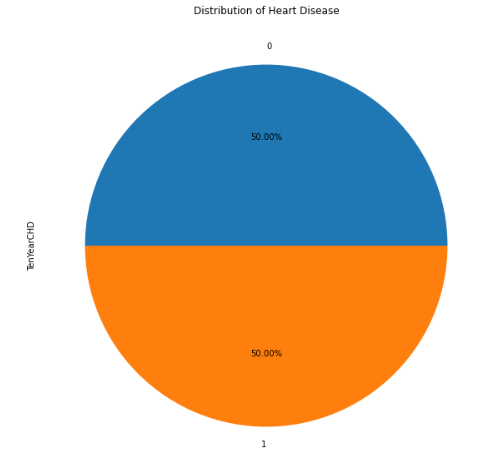
Checking Distribution using Histograms

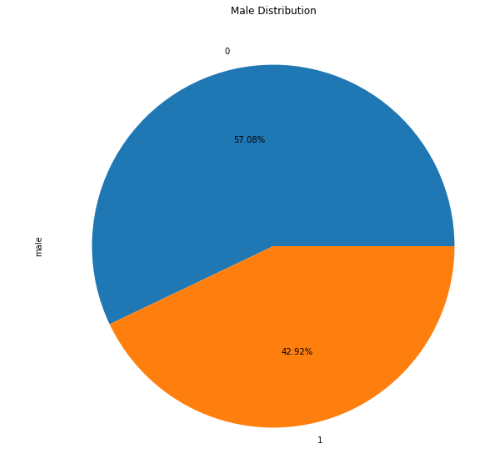


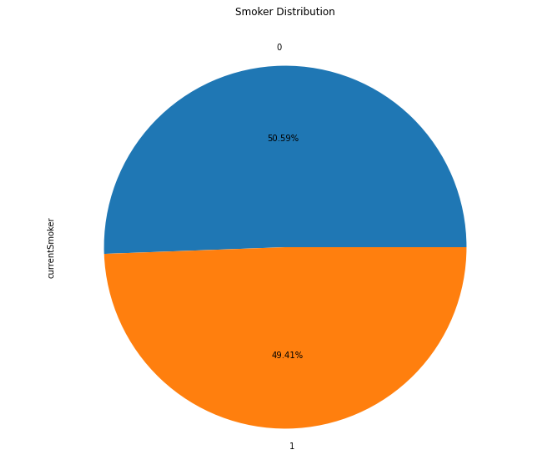
Checking the distribution of the target variable.



The above plot shows that a total of 84.81% are healthy and the remaining are suffering from heart disease. We have applied SMOTE to the training data to overcome this uneven distribution issue.



Checking Male distribution:

Checking smoker distribution:

**Data Preparation:**

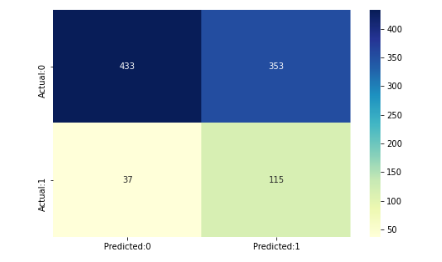
In the final model, I did not include the column Education as generally it’s not considered to be a reason for heart disease. So, we dropped it from our dataset. Also, we have renamed the column male to gender to sound neutral. We did not need any new features out of the dataset that we had and hence we did not engineer any new features.

We checked the no. of null values in each column and as the total no. of rows with at least one missing value is 12% of the total data set, we decided to drop those rows to make the data cleaner. Also, we do not have any categorical columns and hence we don’t need to create any Dummy variables.

**Milestone 3:** Model building and evaluation

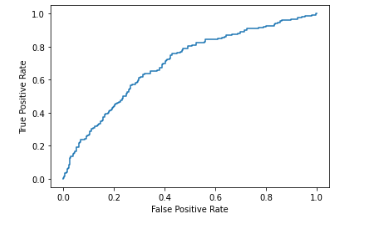
We have selected the logistic regression model for our problem as this is a classification problem. This is a supervised classification algorithm as we provide the training data initially and use the model to predict on the test data. Later we'll be using the confusion matrix to understand the true and false predictions. We divided our data set into training and test data sets in the ratio of 75:25. We wanted to make sure that our model does a good job while dealing with the unknown target as well. We used PCA transformation to make sure that we pick only the columns that retain 90% of the variance. This reduced the features to 10 from 14.

**Confusion Matrix:**

****

Here we used the Logistic regression and got an accuracy of 58% which means that it is a good model. Our model has predicted 548 correct predictions and a total of 390 incorrect predictions. We used PCA transformed data with variance of 90 %. The % of variance explained by the PCA representation reflect the % of information that this representation brings about the original structure. Higher is the % of variance, higher is the % of information and less is the information loss. The minimum % suggested is 60% but we took 90% which means that the attributes we get will be useful.

**ROC Curve:**



The area under the ROC curve quantifies model classification accuracy, the higher the area, the greater the disparity between true and false positives, and the stronger the model in classifying members of the training dataset. An area of 0.5 corresponds to a model that performs no better than random classification and a good classifier stays as far away from that as possible. An area of 1 is ideal. The closer the AUC to 1 the better. Our Model stands at 71% which means that it is a good model.

**Conclusion:**

What does the analysis/model building tell you?

Our data set analysis and the model indicate that there are some features that could cause heart disease such as if they are older, they have higher systolic blood pressure, they are hypertensive, they have higher diastolic blood pressure, they have higher glucose level glucose,  
they have diabetes, they are males.

The accuracy of the model is 58% which means that it is not ready for deployment yet. Training the model with much more vast data set may help us in improving the accuracy of the model. Before the SMOTE model was applied, the accuracy was more than 84% but it was affected by the imbalance in the target classes.

We can explore other models as well to see if the overall accuracy of the model increases. Once an appropriate model is found, it could help the public to be more cautious of the lifestyle choices.

**References:**

<https://www.kaggle.com/code/neisha/heart-disease-prediction-using-logistic-regression/data>

<https://towardsdatascience.com/applying-smote-for-class-imbalance-with-just-a-few-lines-of-code-python-cdf603e58688>