Name: Carol Muriithi

Purpose: Understanding Epidemiology of the Heart Disease

CS : Project in Python

**INTRODUCTION: Data mining on Framingham Heart study dataset**

The Framingham Heart Study is one of the longest ongoing cardiovascular cohort study in the world. It started in 1948 and the initial number of adult subjects was 5,209. The study is now its third generation and was conducted on the inhabitants of the city of Framingham Massachusetts. The results from the study have been used to better understand the epidemiology hypertensive or arteriosclerotic cardiovascular disease.

The idea of this project is to perform comprehensive data analysis on the Framingham Heart Study Dataset where we verify the previously made inferences while also making new inferences from this data by using different Statistical Methods such as PCA and Logistic Regression.

**METHODS**

**Understanding the dataset**

**Datasets:-**

Number of Datasets: 2

NIH: Framingham Heart Dataset (1)

Kaggle: Framingham Heart Dataset (Subset of NIH dataset: 2)

The following demographic risk factors were included in the Framingham Heart Dataset:

1.Sex: 1 for male,2 for female

2.Age: age of the patient

3.Education: Different education Levels were 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.

4.The data set also includes behavioral risk factors associated with smoking

CurrentSmoker: If the patient is a current smoker or not

CigsPerDay: The number of cigarettes that the person smoked on an average in one day.

5.Medical history risk factors:

BPMeds: Whether or not the patient was on blood pressure medication

PrevalentStroke: Stroke incidence in the past

PrevalentHyp: Hypertensive (1) or not (0)

Diabetes: Diabetic (1) or not (0)

6.Risk factors from the first physical examination of the patient:

TotChol: Total cholesterol levels (HDL + LDL)

SysBP: Systolic blood pressure

DiaBP: Diastolic blood pressure

BMI: Body Mass Index

HeartRate: Resting Heart beat (per minute)

Glucose: Fasting glucose levels (Taken after fasting for at least 7 hours)

TenYearCHD: 10 year risk of coronary heart disease CHD(This extra feature or Variable was only included in the Kaggle Dataset and not the NIH Dataset). This was calculated using framingham risk score.

**Tasks for the data analysis**

1. Perform data visualizations to identify the associations between different variables to get a better understand our dataset.
2. Use PCA to determine the most significant variable(s) in our dataset.
3. Perform Data visualizations based on the BMI classes and compare how each of them compare. Check if the assumption that- Super Obese and Obese classes are the most unhealthy in terms of Heart health is true or not.
4. Use Logistic Models for predicting Ten Year CHD risk.

**Exploratory Data Analysis (EDA)**

1. I did a 'Framingham Heart Study Data Distribution'

This showed how different variables were distributed across the population

1. I did a visualization of how many people had a risk of CHD in 10 years
2. Correlation Plots. I also compared this to results in R.

Reason for Analysis: The purpose of the correlation plot was to show how much one variable is affected by another.

Results: From the results of our analysis, it was clear that SYSBP(Systolic blood pressure) and DIASBP (Diastolic blood pressure) are the most highly positively correlated variables in the entire dataset.That means if one increased then the other would increase and vice-versa. The correlation value was 0.72. However, SYSBP is a better variable for predicting cardiovascular health than DIABP.

The next slightly correlated variables were Age and SYSBP which had a low correlation value of 0.38. The DIABP and BMI had an almost similar value of 0.34.

This confirms the assumption that as you grow older, you run a slightly higher risk of increasing your SYSBP.

The next slightly correlated variables were Age and SYSBP which had a low correlation value of 0.38. The DIABP and BMI had an almost similar value of 0.34. This confirms the assumption that as you grow older, you run a slightly higher risk of increasing your SYSBP. It makes sense because older population have higher cases of High Blood pressure as opposed to younger populations.

**Principal Component Analysis(PCA): Exploratory data analysis**

Description of the data exploration/analysis using PCA:

A multivariate Analysis begins with having several substantial correlated variables. In such an event, one can use a statistical analysis called Principal Component Analysis (PCA). PCA functions as a dimension-reduction tool that is often used to reduce a large set of variables to a small set that still contains most of the information in the large set. The smaller number of variables in the small set is called principal components. The PCAs are numbered 1,2,3,4….

The first PCA value generally accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.

We did a PCA on both sets of the Framingham Heart Data that were both from the NIH and Kaggles. Our reason for doing PCA was to find which variables were the most important in the entire dataset: Which variable contributed the most to disease risk all factors considered?

From our analysis of The Framingham Heart Dataset(subset):

The histogram above plots the highest eigenvectors which corresponds to the highest PCA . The plot shows that PCA1 explains over 60% of the variance in the dataset, PCA2 explains over 19% and PC3 explains about 18%  of the variance. The remaining variance is explained by the rest of the dimensions.

Our reason for doing PCA was to find which variables were the most significant in the entire dataset ,i.e, Which variable contributed the most to disease risk when all factors are considered? After our PCA analysis, we identified that Total Cholesterol level was the most important variable for this group of people. This can be explained by the diet and lifestyle choices. Foods  which are high in low density lipoproteins (LDL) cholesterol puts one at a higher risk for cardiovascular disorders. The amount of dietary cholesterol, saturated fats and trans fats in the diet have the major influences on the LDL cholesterol levels.

**Modelling and prediction: Logistic Regression**

In this part, we finalized our logistic model that is predicting the risk of developing CHD in Ten years.

The variable TENYEARCHD was provided in the kaggle dataset and I used it to create a following logistic model-

TenYearCHD ~ sysBP + diaBP + BMI + heartRate + glucose + age + cigsPerDay + education + totChol

**Accuracy of our model**

The model was able to predict a 10 year risk of getting CHD with 85%.

Specificity and Sensitivity should be as close to one as possible. However, the

Sensitivity for my model was low while the specificity is very high The model is specific and not sensitive

*Sensitivity- Also called as True Positive rate, it measures the proportion of actual positives that were correctly identified using the logistic model.*

*Specificity- Also called as True Negatives rate, measures the proportion of actual negatives that are correctly identified using the logistic model.*

**Conclusion**

From our preliminary analysis, we were able to identify several relationships between variables in the Framingham Heart Dataset. After doing a dimensionality reduction using PCA, we were able to identify that variables such as Total Cholesterol, Glucose, Systolic Blood Pressure and Diastolic Blood Pressure were the most significant variables that captured most of the variance in the dataset. This summarizes the essence that Cardiovascular Diseases, Blood Pressure and Diabetes are a direct result of lifestyle factors pertaining to diet, exercise, common medications, such as aspirin and smoking cigarettes.

Overall, Heart diseases are complex and over the years they have been linked to various risk factors but it is still challenging to make an accurate prediction of an individual’s risk of developing any heart disease over time. The Framingham Heart study is a historically significant heart study that helped us better understand many of these risk factors. Over time, as more data is made available related to these risk factors, we will be able to make predictions in the future and understand the relationship between these risk and even protective factors (eg- Physical Activity) like: What’s the percentage reduction in the risk of developing any Heart disease based on the amount of physical activity each day ?

**Citations/References**

Mahmood, S. S., Levy, D., Vasan, R. S., & Wang, T. J. (2013). The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet (London, England)*, *383*(9921), 999-1008.