What are the important risk factors that can predict heart disease? and Can we effectively cluster heart disease patients to identify similar patients which can provide insights for treatment plan?

Nisi Mohan Kuniyil 300321388

21/11/2020

Research Question

What are the important risk factors that can predict heart disease? and Can we effectively cluster heart disease patients to identify similar patients which can provide insights for treatment plan?

According to WHO over 17 million people die from Cardiovascular Diseases (CVD) every year, representing 31% of all global deaths. Most cardiovascular diseases can be prevented by addressing behavioral risk factors such as tobacco use, unhealthy diet, obesity, physical inactivity and harmful use of alcohol using population-wide strategies. People with cardiovascular disease or who are at high risk due to the presence of one or more risk factors such as hypertension, diabetes, and hyperlipidaemia need early detection and management using counseling and medicines, as appropriate. This research intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using logistic regression and decision tree. Also, explore clustering techniques to group similar patients together and analyse the clusters for common features.

Source

The dataset is publicly available on the edX platform, and it is from a cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether a patient has a risk of developing cardiovascular disease in the next 10-years. The dataset provides the patients' information from the longitudinal study containing repeated observations of over 17 variables. It includes over 4,000 records.

Project Plan

1. Data is collected from https://courses.edx.org/courses/HarvardX/PH207x/2012_Fall/datasets/

   There are 74 columns and 4434 observations in this dataset. Out of these 74 columns, approximately 17 variables are measured repeatedly as part of the longitudinal study. Key columns are shown below.

```
##     sex1 totchol1 age1 sysbp1 diabp1 cursmoke1 cigpday1  bmi1 diabetes1 bpmeds1
## 1   Male      195   39  106.0     70        No        0 26.97        No      No
## 2 Female      250   46  121.0     81        No        0 28.73        No      No
## 3   Male      245   48  127.5     80       Yes       20 25.34        No      No
## 4 Female      225   61  150.0     95       Yes       30 28.58        No      No
## 5 Female      285   46  130.0     84       Yes       23 23.10        No      No
## 6 Female      228   43  180.0    110        No        0 30.30        No      No
##   heartrte1 glucose1 prevchd1 prevap1 prevmi1 prevstrk1 prevhyp1 hdlc1
## 1        80       77       No      No      No        No       No    NA
```

```
## 2        95       76       No      No      No        No       No    NA
## 3        75       70       No      No      No        No       No    NA
## 4        65      103       No      No      No        No      Yes    NA
## 5        85       85       No      No      No        No       No    NA
## 6        77       99       No      No      No        No      Yes    NA
```

```
## $var.labels
##  [1] "Random ID"
##  [2] "Death indicator"
##  [3] "Incident Angina Pectoris"
##  [4] "Incident Hospitalized MI"
##  [5] "Incident Hosp MI-Fatal CHD"
##  [6] "Incident Hosp MI, AP, CI, Fatal CHD"
##  [7] "Incident Stroke Fatal/non-fatal"
##  [8] "Incident Hosp MI or Stroke, Fatal or Non"
##  [9] "Incident Hypertension"
## [10] "Time (years) to Angina"
## [11] "Time (years) to Hosp MI"
## [12] "Time (years) to MI-Fatal CHD"
## [13] "Time (years) to CHD"
## [14] "Time (years) to Stroke"
## [15] "Time (years) to CVD"
## [16] "Time (years) to Death"
## [17] "Time (years) to Hypertension"
## [18] "Sex, exam 1"
## [19] "Total cholesterol (mg/dL), exam 1"
## [20] "Age (years), exam 1"
## [21] "Systolic blood pressure (mmHg), exam 1"
## [22] "Diastolic blood pressure (mmHG), exam 1"
## [23] "Current smoker, exam 1"
## [24] "Number of cigarettes per day, exam 1"
## [25] "Body mass index, exam 1"
## [26] "Diabetic, exam 1"
## [27] "Use anti-hypertension medication medication, exam 1"
## [28] "Heart rate (beats per minute), exam 1"
## [29] "Glucose level (mg/dL), exam 1"
## [30] "Prevalent coronary heart disease, exam 1"
## [31] "Prevalent angina pectoris, exam 1"
## [32] "Prevalent myocardial infarction, exam 1"
## [33] "Prevalent stroke, exam 1"
## [34] "Prevalent hypertension, exam 1"
## [35] "HDL cholesterol (mg/dL), exam 1"
## [36] "LDL cholesterol (mg/dL), exam 1"
## [37] "Sex, exam 2"
## [38] "Total cholesterol (mg/dL), exam 2"
## [39] "Age (years), exam 2"
## [40] "Systolic blood pressure (mmHg), exam 2"
## [41] "Diastolic blood pressure (mmHG), exam 2"
## [42] "Current smoker, exam 2"
## [43] "Number of cigarettes per day, exam 2"
## [44] "Body mass index, exam 2"
## [45] "Diabetic, exam 2"
## [46] "Use anti-hypertension medication medication, exam 2"
## [47] "Heart rate (beats per minute), exam 2"
```

```
## [48] "Glucose level (mg/dL), exam 2"
## [49] "Prevalent coronary heart disease, exam 2"
## [50] "Prevalent angina pectoris, exam 2"
## [51] "Prevalent myocardial infarction, exam 2"
## [52] "Prevalent stroke, exam 2"
## [53] "Prevalent hypertension, exam 2"
## [54] "HDL cholesterol (mg/dL), exam 2"
## [55] "LDL cholesterol (mg/dL), exam 2"
## [56] "Sex, exam 3"
## [57] "Total cholesterol (mg/dL), exam 3"
## [58] "Age (years), exam 3"
## [59] "Systolic blood pressure (mmHg), exam 3"
## [60] "Diastolic blood pressure (mmHG), exam 3"
## [61] "Current smoker, exam 3"
## [62] "Number of cigarettes per day, exam 3"
## [63] "Body mass index, exam 3"
## [64] "Diabetic, exam 3"
## [65] "Use anti-hypertension medication medication, exam 3"
## [66] "Heart rate (beats per minute), exam 3"
## [67] "Glucose level (mg/dL), exam 3"
## [68] "Prevalent coronary heart disease, exam 3"
## [69] "Prevalent angina pectoris, exam 3"
## [70] "Prevalent myocardial infarction, exam 3"
## [71] "Prevalent stroke, exam 3"
## [72] "Prevalent hypertension, exam 3"
## [73] "HDL cholesterol (mg/dL), exam 3"
## [74] "LDL cholesterol (mg/dL), exam 3"
```

2. Data Tidying.
    - Handling NAs.
    - Converting categorical variables into factors.
    - Aggregating Longitudinal Variables.
    - Creating binary labels for predictiion.
3. Visualization for exploratory data analysis.
    - Box plots can be used to visualise continuous variables distribution like age against the target variable.
    - Bar plots can be used to visualise distributions for variables such as gender, smoker/non-smoker etc.
    - Correlation plot to show feature correlations with each other.
    - Data Preparation and Partitioning.
4. Building and Training a Logistic Regression Model
    - Evaluating the Logistic Regression Model.
    - How accurate is the model for predicting heart disease?
    - What are the most important features that predict heart disease?
5. Building and Training a Decision Tree Model.
    - Evaluating the Decision Tree Model
    - Compare the performance of decision trees with Logistic Regression.
    - Analyse feature importance.
6. Clustering

- Using K-means or hierarchical clustering, check whether the patient data can be grouped into similar clusters.
- Analyse the clusters to get insights into similarities between patients in the dataset.