# Prediction of Coronary Heart Disease in the next ten years

Muthukumar Kadhirvel

2022-11-11

**Introduction**

## Problem Statement

Heart disease is a generic term used to represent all conditions that affect the structure and functioning of the heart. Coronary heart disease (CHD) is a specific type of heart disease where the heart’s arteries cannot pump enough blood to the heart that has enough oxygen. In USA alone there are around 18.2 million people with CHD and that makes it one of the deadliest heart diseases in USA according to CDC. (NHLBI 2022).

Coronary Heart Disease is caused by two main factors - Cholesterol – Waxy substance whose build-up can block flow of blood in the heart’s arteries or Blood vessels of heart do not work properly. But the symptoms of heart disease can vary by person and there is no indication of a person having CHD until they have chest pain or heart attack (Heart’s blood flow blocked) or cardiac arrest (Heart suddenly stops working). By then it is too late to alleviate the condition and stop the person from a sudden death or otherwise we can have the person change their lifestyle, medicine, surgery etc. to prevent the same. (NHLBI 2022).

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

CHD is a severe problem and detecting the presence of it in the earlier stages would help the person get cured or do not let it become serious by changing their lifestyle, medicine, surgery etc. If we identify the prospective patients by a robust model, then hospitals and clinics can use that model to identify undiagnosed patients who are at risk for the disease. This can improve the spending for the patient and improve the quality of life as well.

**Data Processing - EDA**

## Dataset

<https://www.kaggle.com/datasets/jiantay33/coronary-prediction>

The above dataset from Kaggle various information of patients and we have 4238 rows and 16 columns. The continuous variables are Age/Number of cigarettes smoked in a day/Cholesterol Level/sysBP (Systolic Blood Pressure)/diaBP (Diastolic Blood Pressure)/BMI (Body Mass Index)/Heart Rate Reading/Glucose Level etc.

The categorical variables are already in the form of ordinal values like Highest education received or binary values like Male or Female/If a person smokes or not/Patient on Blood Pressure Medication or not/Patient had stroke already or not/Patient had hypertension already or not/Patient had diabetes already or not etc. The target variable is already in the binary format of whether the patient got diagnosed with CHD in the next 10 years.

## Data Preparation

Since the categorical variables are binary/ordinal and continuous variables are numeric, we need not create any dummy variables. The dataset had missing values from at seven variables, and I used four different techniques to replace the missing values.

Education – Replaced the missing values with mode as the overall dataset had more of Level 1 and it is an ordinal variable.

cigsPerDay – Though this is a typical continuous variable and can be replaced with mean, we had a column called currentSmoker which identifies if the patient is a smoker or not. So, we made use of that to identify only smokers and replaced missing values for them with their mean.

BPMeds – During analysis noticed that this has a good correlation with prevalentHyp field and any time when BPMeds had a value of 1, prevalentHyp also had a value of 1 and so replaced missing values with the corresponding value of prevalentHyp since it was always present.

totChol/BMI/heartrate/glucose – Replaced missing values with mean as these are continuous variables.

## Visualization

We start off with building Histograms of the age and education with the two levels of the target variable to see if there are any changes to distribution. We can see that for persons less than 35 there is very little risk of CHD and even for people aged less than 40 it is the same. As far as education is concerned, we can see that most patients have education level of 1 but the percentage of people having CHD does not seem to depend on education.

Chart, histogram

Description automatically generatedChart, bar chart

Description automatically generated

We run scatter plots of sysBP/diaBP and totChol/diaBP as these are typically tied together in medical analysis and we can see there is a high positive correlation between sysBP and diaBP while there is no correlation between totChol and diaBP features.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

We run descriptive analysis on the dataset along with creating correlation matrix/heatmap as the correlation matrix/heatmap can give us a clear idea of the importance of each variable in predicting the target variable. We are using point-biserial correlation as our dataset has a mixture of continuous and binary variables and we can see that there are no clear variables that have a good correlation with TenYearCHD variable and so we move on with modelling.

**Predictive Analytics - Modelling**

## Models and Metrics

## Ours is a binary classification problem because our end goal is to predict a person can have CHD in the next ten years or not and so we used Logistic Regression/ Random Forest Classifier/ Decision Tree Classifier. We will first split the data into training dataset and test data in the ratio of 80% to 20% where the Target is the Ten-Year CHD, and the Features series has the rest of the columns.

As ours is a binary classification problem we will use accuracy as the evaluation metric and Industry standard for good accuracy for a predictive model is around 70% to 90%. The other important metric to consider for classification problem is the confusion matrix as even models with good accuracy might not predict the correct positive predictions and that is where confusion matrix comes into the picture to make sure the number of false positive and false negatives are less.

## Model Evaluation

Logistic Regression yielded good accuracy but was bad at predicting TenYearCHD value of 1 which was the same with Random Forest Classifier. As far as Decision Tree Classifier was concerned it was far better than the above two models in terms of predicting TenYearCHD value of 1.

Since Decision Tree Classifier was yielding better results, we went ahead and performed MinMaxScaler on the dataset and then performed the modelling again and we can see that both the accuracy and Confusion Matrix yielded better results. But still the model was not performing well as expected and so went ahead with SMOTE (Synthetic Minority Oversampling Technique) which is used in case of datasets with predicting minority class.

Text

Description automatically generated

As we can see above this created a larger dataset with equal TenYearCHD value of 0 and 1 and applying Decision Tree Classifier on the resample dataset we were able to yield better results. Chart

Description automatically generatedChart

Description automatically generatedA picture containing chart

Description automatically generatedChart

Description automatically generatedChart

Description automatically generated

**Conclusion**

## Conclusion/Recommendations

## The final Decision Tree Classifier model with resample dataset on the scaled data of the original dataset has yielded the best results and will be used to go ahead with implementation. This model is ready to be deployed and we can continue to input more sample data that becomes available to retrain our model and make it a recurring mechanism to keep the model in line with the current data that may change because of changing lifestyles.

## Potential Challenges

## The one challenge with this project is that the dataset has specific factors but there may be other factors that can contribute to the Ten-Year CHD diagnosis. Factors like Race/ Ethnicity/ Income/ Region etc. could also be crucial factors but our dataset does not have that information.

## References

NHLBI (2022). U.S. Department of Health and Human Services. (n.d.). *What is coronary heart disease?* National Heart Lung and Blood Institute. Retrieved September 11, 2022, from https://www.nhlbi.nih.gov/health/coronary-heart-disease