Northeastern University

Coronary Heart Disease Risk Prediction

INFO 7390 : Advanced Data Science and Architecture

Group Members

Anuja Naik

Eshanee Thakur

Ritesh Pendurkar

Rohit Gulati

Contents

Background

Heart disease is the major cause of morbidity and mortality globally, it accounts for more deaths annually than any other cause. According to the WHO, an estimated 17.9 million people died from heart disease in 2016, representing 31% of all global deaths. Over three quarters of these deaths took place in low- and middle-income countries. Of all heart diseases, coronary heart disease is by far the most common and the most fatal. In the United States, for example, it is estimated that someone has a heart attack every 40 seconds and about 805,000 Americans have a heart attack every year. The silver lining is that heart attacks are highly preventable and simple lifestyle modifications coupled with early treatment greatly improves its prognosis. It is, however, difficult to identify high risk patients because of the multi-factorial nature of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, et cetera. This is where machine learning and data mining come to the rescue. Doctors and scientists alike have turned to machine learning (ML) techniques to develop screening tools and this is because of their superiority in pattern recognition and classification as compared to other traditional statistical approaches.

Objective and Goals

- We will be implementing Machine Learning models to predict whether a patient will be diagnosed with coronary heart disease (CHD) in the next 10 years based on a few medical parameters associated with them
- Using exploratory data analysis, we will gain insights about the data by checking the distribution of different features, correlation of features with each other and the target variable
- We will be comparing classification metrics like accuracy, precision, F-1 score to identify the most efficient Machine Learning model

Methodology and Approach

There are different steps involved in predicting whether the patient has 10-year risk of coronary heart disease (CHD). Classification is a difficult activity as it requires pre-processing steps to convert the raw data into structured form. Classification process involves following main steps for predicting 10-year CHD. These steps are data collection, pre-processing, feature selection, classification techniques application, and evaluating performance measures.

Data Cleaning and Pre-processing

We checked and dealt with missing values and removed the irrelevant features from the data set as these can grossly affect the performance of different machine learning algorithms as many algorithms do not tolerate missing data.

Data pre-processing

```
# calculating the total percentage of missing data (na/null values)
missing_data = data.isnull().sum()
total_percentage = (missing_data.sum()/data.shape[0]) * 100

print(f'The total percentage of missing data is {round(total_percentage,2)}%')
```

```
The total percentage of missing data is 12.74%
```

```
# calculating the percentage of missing data for each column
total = data.isnull().sum().sort_values(ascending=False)
percent_total = round((data.isnull().sum()/data.isnull().count()).sort_values(ascending=False)*100,2)
missing = pd.concat([total, percent_total], axis=1, keys=["Total", "Percentage"])
missing_data = missing[missing['Total'] > 0]

missing_data
```

	Total	Percentage
glucose	388	9.15
BPMeds	53	1.25
totChol	50	1.18
cigsPerDay	29	0.68
BMI	19	0.45
heartRate	1	0.02

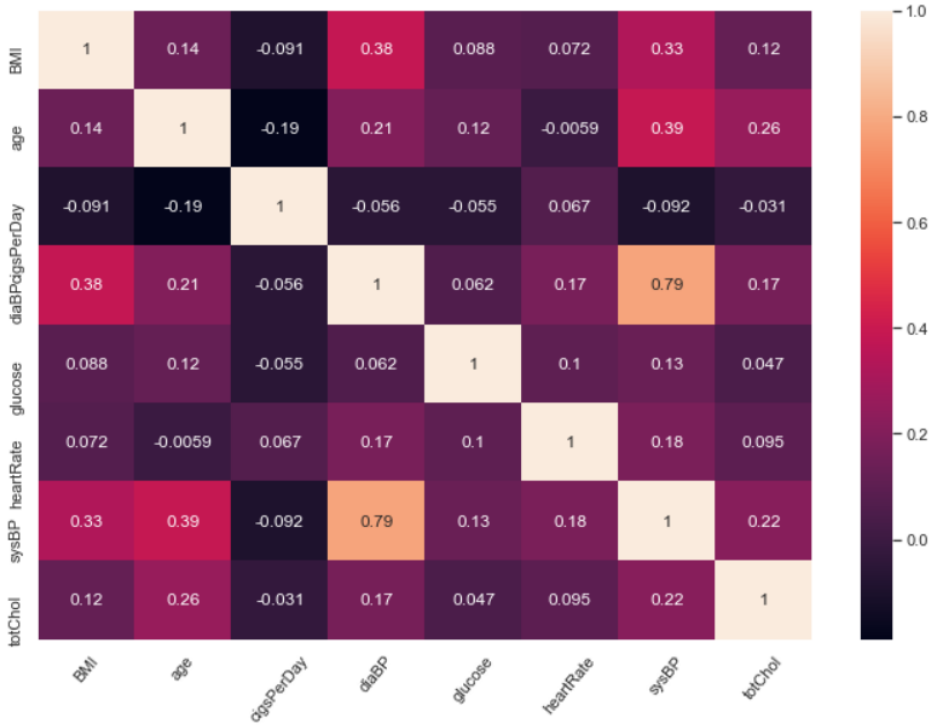
```
# Removing the missing data
data.dropna(axis=0, inplace=True)

data.shape
```

```
(3751, 15)
```

Exploratory Data Analysis

We wanted to gain important statistical insights from the data, so we checked for the distributions of different attributes, correlations of the attributes with each other and the target variable. We also calculated important odds and proportions for the categorical attributes.



To analyse the correlation between continuous features we plotted a heatmap and deduced that some of the features were highly correlated i.e. sysBP and diaBP, age and sysBP, BMI and diaBP.

Feature Selection

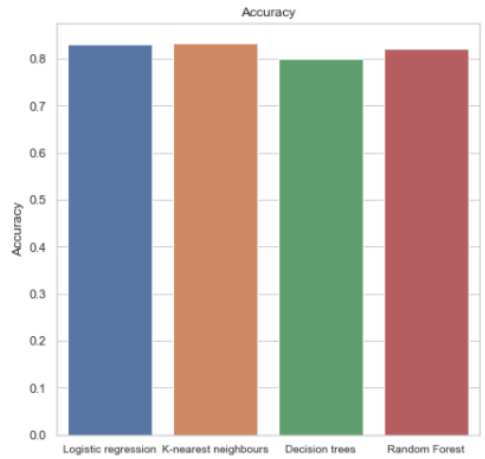
Since having irrelevant features in a data set can decrease the accuracy of the models applied, based on the exploratory data analysis, we selected important features and later used them to build different models. The features are as below:

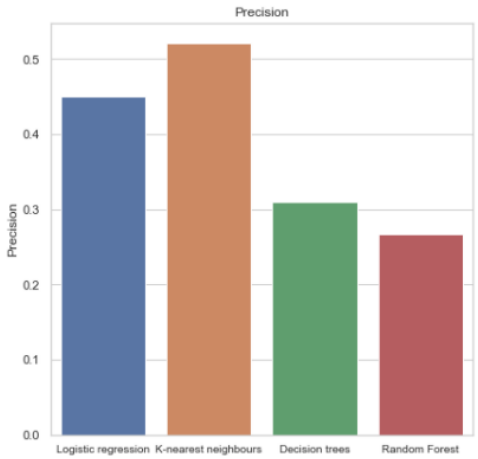
- Age
- Total Cholestrol (totChol)
- Systolic BP (sysBP)
- Diastolic BP (diaBP)
- BMI
- Heart Rate (heartRate)
- Glucose level (glucose)

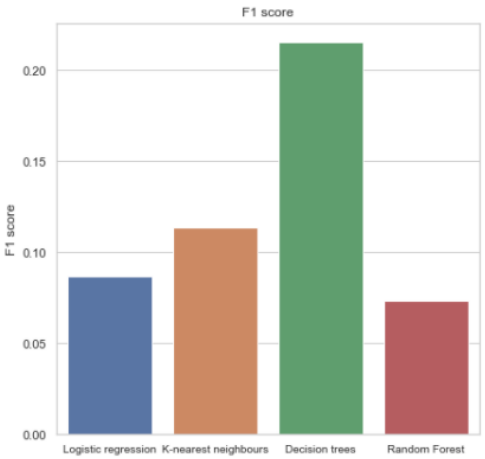
Results and Analysis

In our project we have implemented the four machine learning algorithms to predict Ten-year CHD based on the medical history and other relevant attributes. Based on our model performance and analysis, the K-Nearest Neighbor classifier model displayed the best accuracy amongst the other models. K-Nearest Neighbor classifier could predict the risk of CHD with 83.39% accuracy. Below table displays a consolidated view of accuracy and other parameters for all the models.

MODELS	ACCURACY	PRECISION	F1-SCORE
Random Forest Classifier	82.06%	0.26	0.07
Logistic Regression Analysis	83.12%	0.45	0.08
Decision Tree Classifier	79.92%	0.31	0.21
K-Nearest Neighbor Classifier	83.39%	0.52	0.11







Conclusion & Future Scope

An implementation of Coronary Heart Disk risk prediction is achieved in this project. The purpose of this project was to gain insights into whether Machine Learning models could predict if a patient will be diagnosed with coronary heart disease (CHD) in the next 10 years based on a few medical parameters associated with the patients. We performed exploratory data analysis on the dataset and created relevant visualizations for the same. We implemented four machine learning classification algorithms to determine which of the models were able to predict 10-year CHD risk with the highest accuracy rate. We performed data cleaning and data pre-processing to convert raw dataset into a meaningful and useable dataset. In continuation from the above step, we performed feature selection based on exploratory data analysis and trained the model accordingly. We concluded that K-Nearest Neighbor Classifier model displayed the highest rate of accuracy amongst all the classifier models.

In future these algorithms can be tested on larger dataset. Moreover, these algorithms can be improved so that efficiency of categorization could be enhanced. Additionally, for feature engineering, we could implement other approach like Smote technique which would help in improving the models' sensitivity by balancing the datasets.