Proactive Health Management - Predicting Heart disease

Contents

Abstract

Artificial intelligence (AI) aims to mimic human cognitive functions. It is bringing a paradigm shift to healthcare, powered by increasing availability of healthcare data and rapid progress of analytics techniques. We survey the current status of AI applications in healthcare and discuss its future. AI can be applied to various types of healthcare data (structured and unstructured). Popular AI techniques include machine learning methods for structured data, such as the classical support vector machine and neural network, and the modern deep learning, as well as natural language processing for unstructured data. Major disease areas that use AI tools include cancer, neurology and cardiology. This project aims to see how AI applications can help in early detection of Coronary Heart disease.

Introduction to the project

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 cardio vascular diseases. The early diagnosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications. This research intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using logistic regression.

In our day to day life, people are undergoing a routine and busy schedule which leads to stress and anxiety. In addition to this, the percentage of people who are obese and addicted to cigarette goes up drastically. This leads to diseases like heart disease, cancer, etc. The challenge behind these diseases is its prediction. Each person has different values of pulse rate and blood pressure. But medically proven, the pulse rate must be 60 to 100 beats per minute and the blood pressure must be in the range of 120/80 to 140/90. Heart disease is one of the major cause of death in the world. The number of people affected by heart disease increases irrespective of age in both men and women. But other factors like gender, diabetes, BMI also contribute to this disease. In this paper, we have tried prediction and analysis of heart disease by considering the parameters like age, gender, blood pressure, heart rate, diabetes and so on. Since numerous factors are involved in heart disease, the prediction of this disease is challenging. Some of major symptoms of heart attack are:

- Chest tightness.
- Shortness of breath.
- Nausea, Indigestion, Heartburn, or stomach pain.
- Sweating and Fatigue.
- Pressure in the upper back Pain that spreads to the arm.

The following are the type of heart disease: Heart means “cardio”. Hence all heart diseases concern to category of cardiovascular diseases. The different kinds of heart disease are:

- Coronary heart diseases.
- Angina pectoris
- Congestive heart failure.
- Cardiomyopathy
- Congenital heart diseases

Coronary heart disease or coronary artery disease is the narrowing of the coronary arteries. It is one of the popular type of heart disease. . If a person has diabetes for a longer time, there are high chances for that person to have heart disease in future. With diabetes, there are other reasons which contribute to heart disease. They are smoking which raises the risk of developing heart disease, high blood pressure makes the heart work harder to pump blood and it can strain heart and damage blood vessels, abnormal cholesterol levels also contribute to heart disease and obesity. Also, family history of heart disease can be a cause of having heart disease. But this history is not considered in this paper for prediction of heart disease. The other risk factors include age, gender, stress and unhealthy diet. Chance of having a heart disease increases when a person is getting older. Men have a greater risk of heart disease. However, women also have the same risk after menopause. Leading a stressed life can also damage the arteries and increase the chance of coronary heart disease.

So, the aim of this project is that based on the above factors, we try to predict the risk of heart disease. A large amount of work has been done related to heart prediction system by using various techniques and algorithms by many people. These techniques may be based on deep-learning, machine-learning, data mining and so on. The aim of all those projects is to achieve better accuracy and to make the system more efficient so that it can predict the chances of heart attack.

Project problem

Today heart disease is one of the most important causes of death in the world. Even healthy people who stick on a regular exercise and adhere to healthy food habits are prone to stroke and heart attack. Few known cases like the SAP MD who died at a very young age of43, yesterday a 48 year old person participating in half marathon died of heart attack. So its early prediction and diagnosis is important in medical field, which could help in on timely treatment, decreasing health costs and decreasing death caused by it. In fact the main goal of using data mining algorithms in medicine by using patients' data is better utilizing the database and discovering tacit knowledge to help doctors in better decision making. Therefore using data mining and discovering knowledge in cardiovascular centres could create a valuable knowledge, which improves the quality of service provided by managers, and could be used by doctors to predict the future behaviour of heart diseases using past records. Also some of the most important applications of data mining and knowledge discovery in heart patients system includes: diagnosing heart attack from various signs and properties, evaluating the risk factors which increases the heart attack. In this article the effort focused on evaluating the heart diseases using some existing data. We would be trying out various supervised techniques with the available data and analyze the results. Finally we will provide the best algorithms to detect heart diseases using a comparison and will show the results in a table. It is obvious in the diagrams that the suggested method has the best performance and best quality in prediction.

Key contributions

The understanding about heart disease was mainly due to research known as the Framingham Heart Study (FHS), the most influential investigation in the history of modern medicine. It is a long-term, ongoing cardiovascular study on residents of the town of Framingham, Massachusetts, USA. The study began in 1948 with 5209 adult subjects from Framingham and is now on its third generation of participants. Much of our appreciation of the pathophysiology of heart disease came from the results of studies from the FHS. It established the traditional risk factors, such as high blood pressure, diabetes, and cigarette smoking for coronary heart disease. Over the years, careful monitoring of the Framingham study population has led to the identification of the major CVD risk factors – high blood pressure, high blood cholesterol, smoking, obesity, diabetes, and physical inactivity – as well as a great deal of valuable information on the effects of related factors, such as blood triglyceride and high-density lipoprotein cholesterol levels, age, gender, and psychosocial issues

Hence, basis all this knowledge and facts and the datasets provided by FHS, we tried to apply various supervised techniques like Random Forest classifier, Decision Tree, SVM, Logistic Regression to detect the class variable.

Machine learning

Following operational stages were followed to preparing a model, which could be used as software application in predicting heart disease:

- Selecting Data: We used the existing dataset available in kaggle.
- Analysis and understanding of the data
- Pre-processing the data, cleaning of the data
- Feature analysis
- Splitting the data into train and test
- Applying different ML techniques to get the best model
- Analysis of the results

➢ Data set recourse

The dataset is publically available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is 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. Each attribute is a potential risk factor and can be categorised as demographic, behavioural and medical risk factors.

➢ Attribute Description

There were 16 attributes were as follows:

Attribute	**Description and value**	**Type**
Male	Gender. 0-Female, 1 Male	Demographic
Age	Age of patient	Demographic
Education	Education of the patient (1-5)	Demographic
Current smoker	1 if current smoker and 0 otherwise	Behavioural
Cigarette per day	If current smoker then number of cigarette per day	Behavioural
BP Meds	Blood Pressure (1 if BP is there 0 otherwise)	Medical
Prevalent BP	Prevalent blood pressure	Medical
Prevalent Hyp	Prevalent hyper tension	Medical
Diabetes	1 if diabetes 0 otherwise	Medical
Total cholesterol	Cholesterol level	Medical
Sys BP	Systolic blood pressure	Medical
Dia BP	Diastolic blood pressure	Medical
BMI	Body mass index	Medical
Heart rate	Heart rate or pulse of the patient	Medical
Glucose	Glucose level	Medical
Ten Year CHD- **Predict Variable**	10 year risk of coronary heart disease CHD "1"-"Yes", "0" -"No"	Medical

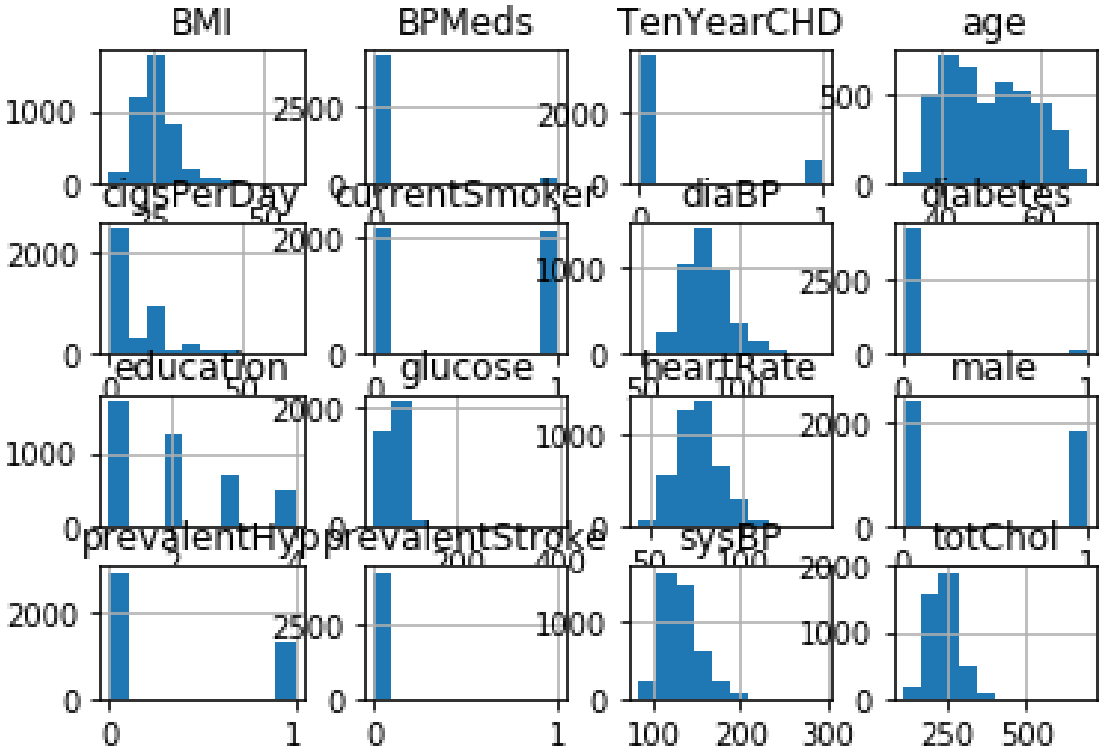
➢ Sample output

	male	age	education	Smoker	Cigs_Day	BPMeds	Pre_Stroke	Pre_Hyp	diabetes	totChol	sysBP	diaBP	BMI	Heart_Rt	glucose
0	1	39	4	0	0	0	0	0	0	195	106	70	26.97	80	77
1	0	46	2	0	0	0	0	0	0	250	121	81	28.73	95	76
2	1	48	1	1	20	0	0	0	0	245	127.5	80	25.34	75	70
3	0	61	3	1	30	0	0	1	0	225	150	95	28.58	65	103
4	0	46	3	1	23	0	0	0	0	285	130	84	23.1	85	85

- Data Spread

Basic spread of the fields is as given below:

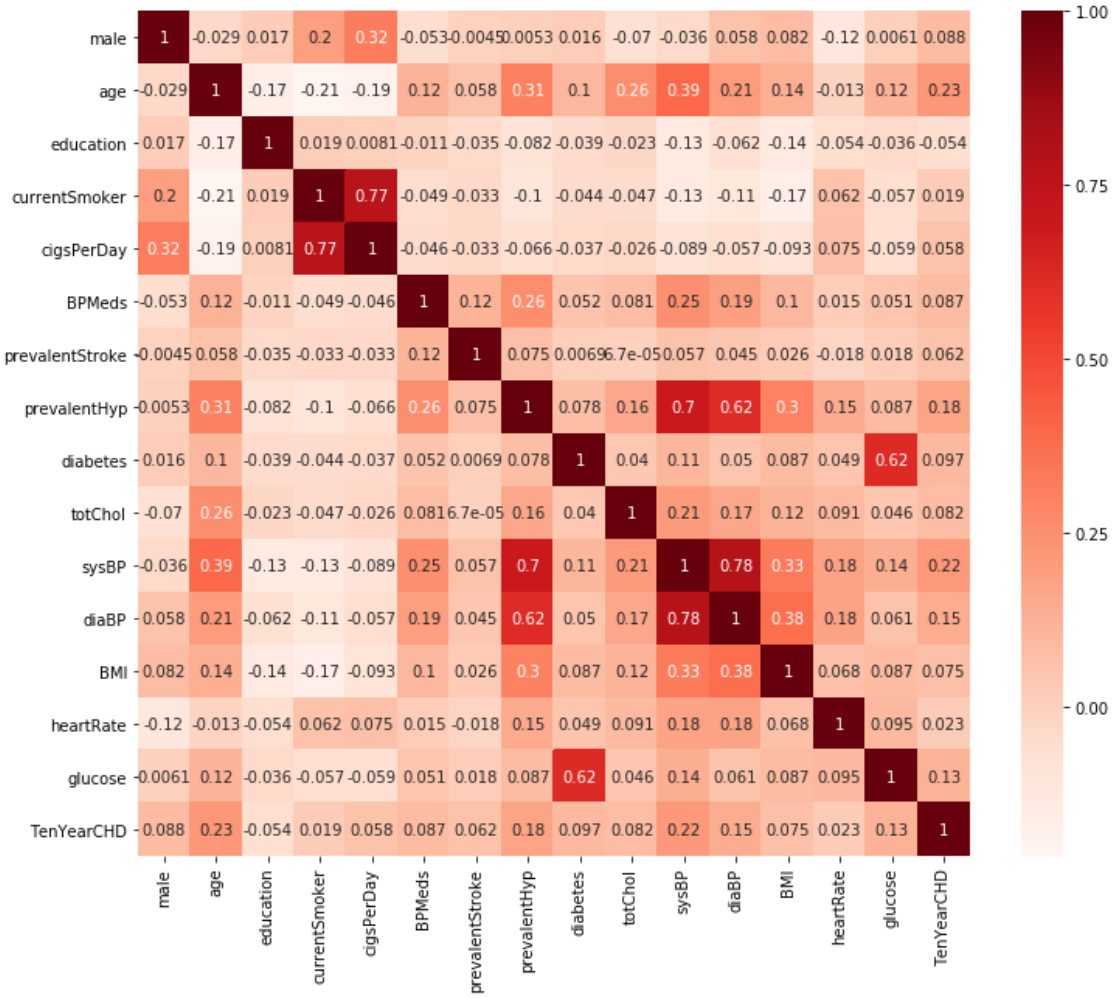
	male	age	Edu	Smoke	Cigs_day	BP	Pre_Stroke	Pre_Hyp	Dia	totChol	sysBP	diaBP	BMI	Heart_Rt	Glucose
mean	0.43	49.58	1.98	0.49	9.00	0.03	0.01	0.31	0.03	236.72	132.35	82.89	25.80	75.88	81.97
std	0.50	8.57	1.02	0.50	11.92	0.17	0.08	0.46	0.16	44.59	22.04	11.91	4.08	12.03	23.96
min	0	32	1	0	0	0	0	0	0	107	83.5	48	15.54	44	40
max	1	70	4	1	70	1	1	1	1	696	295	143	56.8	143	394



- Correlation Matrix

Correlation between the features and the class variable and also amongst the features. We observed that few of the fields are highly correlated like

- Current Smoker and Cigsperday
- BP Meds & Sys & Dia BP, SyBP and DiaBP
- Diabetes and Glucose

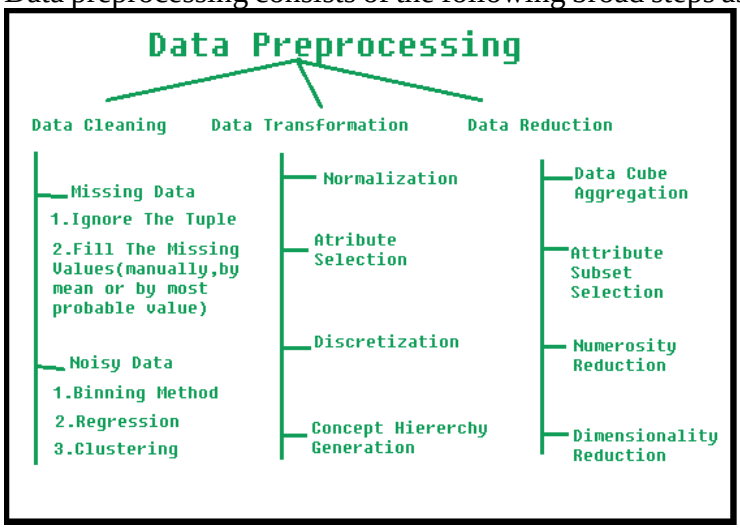


- Top 5 highly correlated features are:

Column	Corr%
age	0.23
sysBP	0.22
prevalentHyp	0.18
diaBP	0.15
glucose	0.13

- Data set pre-processing

Data preprocessing consists of the following broad steps as described in the figure given below:



- Data Cleaning

While analysing the data, we found that following fields have null data:

Column Name	**Logic used to handle the null data**
education	We took the median and substituted. This field contains value from 1-4 and not much information was provided
cigsPerDay	This field was replaced with the 0 or 9, basis the field currentSmoker. Only if the person smokes would the cigsperday have value.
BPMeds	This field was substituted with 1
totChol	Mean value was used for filling up the null
BMI	Mean value was used for filling up the null
heartRate	Mean value was used for filling up the null
glucose	This field was filled basis another field diabetes, if diabetes is 0 then 100 else 220

Basis the histogram and general data spread noisy data treatment has not been done.

- Data Transformation

Most of the fields were categorical in nature i.e. they were having 0 or 1 value only, only the following fields needed normalization:
age, cigsPerDay, totChol , sysBP , diaBP , BMI , heartRate & glucose.

For this we used the **StandardScaler** function of **sklearn.preprocessing**

- Data Reduction

The dataset used was not too huge to reduce it further. There were only 15 features, hence the attribute reduction also was not needed

- Finding the best machine learning model for the problem objective

A look at machine learning algorithms, there is no one solution or one approach that fits all. There are several factors that can affect the decision to choose a machine learning algorithm. Following are some of the salient features of our dataset:

- The data is labelled (Supervised problem)
- The output variable is binary i.e. it has only 2 outputs 0 or 1. (Binary Classification problem)
- All the features are numeric in nature and most of the feature are categorical in nature.
- Class imbalance exists in the dataset
- The model needs to make a decision

Basis all the above factors, we used the following ML techniques:

Model	**Brief description**
Logistic Regression	Logistic regression performs binary classification, so the label outputs are binary. It takes linear combination of features and applies non-linear function (sigmoid) to it, so it's a very small instance of neural network.
KNN	**k-nearest neighbors** algorithm (**k-NN**) is a non-parametric method used for classification and regression. In both cases, the input consists of the **k closest** training examples in the feature space.
Decision Tree	Decision trees easily handle feature interactions and they're non-parametric, so you don't have to worry about outliers or whether the data is linearly separable. One disadvantage is that they don't support online learning, so you have to rebuild your tree when new examples come on. Another disadvantage is that they easily overfit, but that's where ensemble methods like random forests (or boosted trees) come in
SVM	SVM is a supervised machine learning technique that is widely used in pattern recognition and classification problems—when your data has exactly two classes. High accuracy, nice theoretical guarantees regarding overfitting, and with an appropriate kernel they can work well even if you're data isn't linearly separable in the base feature space. Especially popular in text classification problems where very high-dimensional spaces are the norm.
Random Forest	Random Forest is an ensemble of decision trees. It can solve both regression and classification problems with large data sets. It also helps identify most significant variables from thousands of input variables. Random Forest is highly scalable to any number of dimensions and has generally quite acceptable performances
XG Boost	Both the ensemble methods were also used to see if the performance and accuracy improves.
Light GBM	

➢ Training and testing the model

We used the train_test_split model of Sklearn library. Used the 33% data for testing and 67% for training the dataset.

➢ Model Evaluation

Choosing the right evaluation technique is an integral part of any ML project. Since this was a classification problem, we used the following metric to evaluate the model:

- Precision – Recall
- ROC-AUC
- Accuracy

Analysis of results

After processing the data using application and applying different ML algorithms following were the scores:

	LR	XG Boost	Random Forest	SVM	Decision Tree
Precision	0.82	0.81	0.81	0.72	0.79
Recall	0.85	0.85	0.85	0.85	0.79
F1-score	0.8	0.79	0.8	0.78	0.79
Support	1399	1399	1399	1399	1399
Accuracy	85%	85%	85%	85%	79%
AUC	73%	72%	73%	73%	73%
True +ve	13	3	16	0	60
True -ve	1181	1187	1175	1189	1039
False +ve	8	2	14	0	150
False -ve	197	207	194	210	150

On analysis of the results by various models, though the AUC for every model was same but Logistic Regression is having better Precision & Recall. Due to class imbalance the AUC and Accuracy scores are lower.

Conclusion

Heart is a pump or pulsating pump which composed of four compound holes with two atriums and two ventricles, which delivers blood to all body organs. So the heart is a vital organ of body. Unfortunately most of death is caused by heart diseases. Todays the cardiovascular diseases are the most important challenges of healthcare in the worldwide. Prevention and management of cardiovascular diseases requires a pervasive and comprehensive system for recording data. Information of patient records are one of the most important data, which must be classified for easy and fast treatment process. Goal of current project was to detect cardiovascular diseases so that preventive treatment / action can be taken on time.

In this article we introduced some of the most useful algorithms and techniques of artificial intelligence which recently used, and briefly described their properties. In recent years several studies carried out in artificial intelligence subject on heart patients' data, and many of algorithms were successful. But the important point is that the level of success in these algorithms depends on various factors and it is not possible to choose a method as the best one. Factors like data type of database, selecting sub-set of properties and risk factors, number of properties, the larger size of database, low number of missed data and access to suitable and correct data increases the success chance in exploration and increase the quality of algorithms' results.

References

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4966216/
https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset
https://towardsdatascience.com/
https://www.analyticsvidhya.com/