Predicting Heart Disease using Machine Learning

A Project Work Synopsis

Submitted in the partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

IN

Computer Science

Submitted by:

NAME OF THE STUDENT & University Roll Number

Nikhil	**17BCS4273**
Pawan Kumar	**17BCS4302**
Sachin	**17BCS4309**
Viraj	**17BCS4301**

Under the Supervision of:

Miss Aaisha Makkar

CHANDIGARH UNIVERSITY, GHARUAN, MOHALI - 140413, PUNJAB

August-December 2020

Table of Contents

Abstract:

The health care industry produces a huge amount of data. This data is not always made use to the full extent and is often underutilized. Using this huge amount of data, a disease can be detected, predicted or even cured. A huge threat to human kind is caused by diseases like heart disease, cancer, tumour and Alzheimer's disease. In this paper, we try to concentrate on heart disease prediction. Using machine learning techniques, the heart disease can be predicted. The medical data such as Blood pressure, hypertension, diabetes, cigarette smoked per day and so on is taken as input and then these features are modelled for prediction. This model can then be used to predict future medical data.

Thus we propose to develop an application which can predict the vulnerability of a heart disease given basic symptoms like age, sex, pulse rate etc. The machine learning algorithm neural networks has proven to be the most accurate and reliable algorithm and hence used in the proposed system.

INTRODUCTION

1.1 Heart is an important organ of the human body. It pumps blood to every part of our anatomy. If it fails to function correctly, then the brain and various other organs will stop working, and within few minutes, the person will die. Change in lifestyle, work related stress and bad food habits contribute to the increase in rate of several heart related diseases

.

Heart diseases have emerged as one of the most prominent cause of death all around the world. According to World Health Organisation, heart related diseases are responsible for the taking 17.7 million lives every year, 31% of all global deaths. In India too, heart related diseases have become the leading cause of mortality [1]. Heart diseases have killed 1.7 million Indians in 2016, according to the 2016 Global Burden of Disease Report, released on September 15,2017. Heart related diseases increase the spending on health care and also reduce the productivity of an individual. Estimates made by the World Health Organisation (WHO), suggest that India have lost up to $237 billion, from 2005-2015, due to heart related or Cardiovascular diseases [2]. Thus, feasible and accurate prediction of heart related diseases is very important.

Medical organisations, all around the world, collect data on various health related issues. These data can be exploited using various machine learning techniques to gain useful insights. But the data collected is very massive and, many a times, this data can be very noisy. These datasets, which are too overwhelming for human minds to comprehend, can be easily explored using various machine learning techniques. Thus, these algorithms have become very useful, in recent times, to predict the presence or absence of

heart related diseases accurately.

1 . Types of Cardiovascular Diseases

Heart diseases or cardiovascular diseases (CVD) are a class of diseases that involve the heart and blood vessels. Cardiovascular disease includes coronary artery diseases (CAD) like angina and myocardial infarction (commonly known as **a heart attack**).There is another heart disease, called coronary heart disease(CHD), in which a waxy substance called plaque develops inside the coronary arteries. These are the arteries which supply oxygen-rich blood to heart muscle. When plaque begins to build up in these arteries, the condition is called atherosclerosis. The development of plaque occurs over many years. With the passage of time, this plaque can harden or rupture (break open). Hardened plaque eventually narrows the coronary arteries which in turn reduces the flow of oxygen-rich blood to the heart. If this plaque ruptures, a blood clot can form on its surface. A large blood clot can most of the time completely block blood flow through a coronary artery. Over time, the ruptured plaque also hardens and narrows the coronary arteries. If the stopped blood flow isn't restored quickly, the section of heart muscle begins to die. Without quick treatment, a heart attack can lead to serious health problems and even death. Heart attack is a common cause of death worldwide. Some of the common symptoms of heart attack [2] are as follows.

1.1. Chest pain

It is the most common symptom of heart attack. If someone has a blocked artery or is having a heart attack, he may feel pain, tightness or pressure in the chest.

1.2. Nausea ,Indigestion, Heartburn and Stomach Pain

These are some of the often overlooked symptoms of heart attack. Women tend to show these symptoms more than men.

1.3. Pain in the Arms

The pain often starts in the chest and then moves towards the arms, especially in the left side.

1.4. Feeling Dizzy and Light Headed

Things that lead to the loss of balance.

1.5. Fatigue

Simple chores which begin to set a feeling of tiredness should not be ignored.

1.6. Sweating

Some other cardiovascular diseases which are quite common are stroke, heart failure, hypertensive heart disease, rheumatic heart disease, Cardiomyopathy, Cardiacarrhythmia, Congenital heart disease, Valvular heart disease, Aortic aneurysms, Peripheral artery disease and Venous thrombosis. Heart diseases may develop due to certain abnormalities in the functioning of the circulatory system or may be aggravated by certain lifestyle choices like smoking, certain eating habits, sedentary life and others. If the heart diseases are detected earlier then it can be treated properly and kept under control. Here, early detection is the main key. Being well informed about the whys and wherefores of heart disease will help in prevention summarily.

Prevalence of Cardiovascular Diseases

An estimated 17.5 million deaths occur due to cardiovascular diseases worldwide. More than 75% deaths due to cardiovascular diseases occur in the middle-income and low-income countries. Also, 80% of the deaths that occur due to CVDs are because of stroke and heart attack [3]. India too has a growing number of CVD patients added every year. Currently, the number of heart disease patients in India is more than 30 million. Over two lakh open heart surgeries are performed in India each year. A matter of growing concern is that the number of patients requiring coronary interventions has been rising at 20% to 30% for the past few years.

1. Problem Definition

In a statement,
Given clinical parameters about a patient, can we predict whether or not they have heart disease?

2. Data

The original data came from the Cleavland data from the UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/heart+Disease
There is also a version of it available on Kaggle. https://www.kaggle.com/ronitf/heart-disease-uci

3. Evaluation

If we can reach 95% accuracy at predicting whether or not a patient has heart disease during the proof of concept, we'll pursue the project.

4. Features

This is where you'll get different information about each of the features in your data. You can do this via doing your own research (such as looking at the links above) or by talking to a subject matter expert (someone who knows about the dataset).
Create data dictionary

1. age - age in years
2. sex - (1 = male; 0 = female)
3. cp - chest pain type
    - 0: Typical angina: chest pain related decrease blood supply to the heart
    - 1: Atypical angina: chest pain not related to heart
    - 2: Non-anginal pain: typically esophageal spasms (non heart related)
    - 3: Asymptomatic: chest pain not showing signs of disease
4. trestbps - resting blood pressure (in mm Hg on admission to the hospital) anything above 130-140 is typically cause for concern
5. chol - serum cholestoral in mg/dl
    - serum = LDL + HDL + .2 * triglycerides
    - above 200 is cause for concern
6. fbs - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
    - '>126' mg/dL signals diabetes
7. restecg - resting electrocardiographic results
    - 0: Nothing to note

- 1: ST-T Wave abnormality
  - can range from mild symptoms to severe problems
  - signals non-normal heart beat
- 2: Possible or definite left ventricular hypertrophy
  - Enlarged heart's main pumping chamber

8. thalach - maximum heart rate achieved
9. exang - exercise induced angina (1 = yes; 0 = no)
10. oldpeak - ST depression induced by exercise relative to rest looks at stress of heart during excercise unhealthy heart will stress more
11. slope - the slope of the peak exercise ST segment
    - 0: Upsloping: better heart rate with excercise (uncommon)
    - 1: Flatsloping: minimal change (typical healthy heart)
    - 2: Downslopins: signs of unhealthy heart
12. ca - number of major vessels (0-3) colored by flourosopy
    - colored vessel means the doctor can see the blood passing through
    - the more blood movement the better (no clots)
13. thal - thalium stress result
    - 1,3: normal
    - 6: fixed defect: used to be defect but ok now
    - 7: reversable defect: no proper blood movement when excercising
14. target - have disease or not (1=yes, 0=no) (= the predicted attribute)

Hardware Specification

CPU: 2 x 64-bit, 2.8 GHz, 8.00 GT/s CPUs or better. Verify machine architecture. Memory: minimum RAM size of 32 GB, or 16 GB RAM with 1600 MHz DDR3 installed, for a typical installation with 50 regular users. Verify memory requirements.

Software Specification

Anaconda Navigator

Anaconda is a conditional free and open-source distribution of the Python and R programming languages for scientific computing, that aims to simplify package management and deployment. The distribution includes data-science packages suitable for Windows, Linux, and macOS.

Jupyter Notebook

Project Jupyter is a nonprofit organization created to "develop open-source software, open-standards, and services for interactive computing across dozens of programming languages". Spun off from IPython in 2014 by Fernando Pérez, Project Jupyter supports execution environments in several dozen languages.

1 LITERATURE REVIEW

Dimensionality Reduction involves selecting a mathematical representation such that one can relate the majority of, but not all, the variance within the given data, thereby including only most significant information. The data considered for a task or a problem, may consists of a lot of attributesor dimensions, but not all of these attributes may equally influence the output. A large number of attributes, or features, may affect the computational complexity and may even lead to overfitting which leads to poor results. Thus, Dimensionality Reduction is a very important step considered while building any model. Dimensionality Reduction is generally achieved by two methods -Feature Extraction and Feature Selection.

A. Feature Extraction

In this, a new set of features is derived from the original feature set. Feature extraction involves a transformation of the features. This transformation is often not reversible asfew, or maybe many, useful information is lost in the process .In and Principal Component Analysis (PCA)is used for feature extraction. Principal Component Analysis is a popularly used linear transformation algorithm. In the feature space, it finds the directions that maximize variance and finds directions that are mutually orthogonal. It is a global algorithm that gives the best reconstruction.

B. Feature Selection

In this, a subset of original feature set is selected. In [5],key features are selected by CFS(Correlation based Feature Selection) Subset Evaluation combined with Best First Search method to reduce dimensionality. In [6]chi-square statistics test is used to select the most significant features.

PROBLEM FORMULATION

Machine Learning is used across many spheres around the world. The healthcare industry is no exception. Machine Learning can play an essential role in predicting presence/absence of Locomotor disorders,

Heart diseases and more. Such information, if predicted well in advance, can provide important insights to doctors who can then adapt their diagnosis and treatment per patient basis.

2 RESEARCH OBJECTIVES

The proposed research is aimed to carry out work leading to the development of an approach for Predicting Heart Disease with Classification Machine Learning Algorithms

The proposed aim will be achieved by dividing the work into following objectives:

- **Predict** whether a patient should be diagnosed with Heart Disease. This is a **binary** outcome.
- Experiment with various **Classification Models** & see which yields greatest **accuracy**.
- Examine **trends** & **correlations** within our data
- Determine which **features** are **most important** to Positive/Negative Heart Disease diagnosis

ADVANTAGES

1. Increased accuracy for effective heart disease diagnosis.
2. Handles roughest(enormous) amount of data using random forest algorithm and feature selection
3. Reduce the time complexity of doctors.
4. Cost effective for patients.

DISADVANTAGES

1. Prediction of cardiovascular disease results may not accurate.
2. Data mining techniques does not help to provide effective decision making.
3. Cannot handle enormous datasets for patient record

METHODOLOGY

The following methodology will be followed to achieve the objectives defined for proposed research work:

1. Detailed study of Predicting Heart disease using machine learning will be done.
2. Installation and hand on experience on existing approaches of Anaconda, Jupyter Notebook will be done. Relative pros and cons will be identified.
3. Various parameters will be identified to evaluate the proposed system.
4. Comparison of new implemented approach with exiting approaches will be done.

3 TENTATIVE CHAPTER PLAN FOR THE PROPOSED WORK

CHAPTER 1: INTRODUCTION

This chapter will cover the overview of

CHAPTER 2: LITERATURE REVIEW

This chapter include the literature available for Predicting Heart Disease using Machine learning ,

The findings of the

researchers will be highlighted which will become basis of current implementation.

CHAPTER 2: BACKGROUND OF PROPOSED METHOD

This chapter will provide introduction to the concepts which are necessary to understand the proposed system.

CHAPTER 4: METHODOLOGY

This chapter will cover the technical details of the proposed approach.

CHAPTER 5: EXPERIMENTAL SETUP

This chapter will provide information about the subject system and tools used for evaluation of proposed method.

CHAPTER 6: RESULTS AND DISCUSSION

The result of proposed technique will be discussed in this chapter.

CHAPTER 7: CONCLUSION AND FUTURE SCOPE

The major finding of the work will be presented in this chapter. Also directions for extending the current study will be discussed.

CONCLUSION

Heart diseases when aggravated spiral way beyond control. Heart diseases are complicated and take away lots of lives every year .When the early symptoms of heart diseases are ignored, the patient might end up with drastic consequences in a short span of time. Sedentary lifestyle and excessive stress in today's world have worsened the situation. If the disease is detected early then it can be kept under control.

However, it is always advisable to exercise daily *and discard unhealthy habits at the earliest.* Tobacco consumption and unhealthy diets increase the chances of stroke and heart diseases. Eating at least 5 helpings of fruits and vegetables a day is a good practice. For heart disease patients, it is advisable to restrict the intake of salt to one teaspoon per day.

One of the major drawbacks of these works is that the main focus has been on the application of classification techniques for heart disease prediction, rather than studying various data cleaning and pruning techniques that prepare and make a dataset suitable for mining. It has been observed that a properly cleaned and pruned dataset provides much better accuracy than an unclean one with missing values. Selection of suitable techniques for data cleaning along with proper classification algorithms will lead to the development of prediction systems that give enhanced accuracy.

In future *an intelligent system may be developed that can lead to selection of proper treatment methods for a patient diagnosed with heart disease. A lot of work has been done already in making models that can predict whether a patient is likely to develop heart disease or not. There are several treatment methods for a patient once diagnosed with a particular form of heart disease. Data mining can be of very good help in deciding the line of treatment to be followed by extracting knowledge from such*
suitable databases.

8.REFERENCES

[1] Ramadoss and Shah B et al.“A. Responding to the threat of
chronic diseases in India”. Lancet. 2005; 366:1744–1749. doi: 10.1016/S0140-6736(05)67343-6.

[2] Global Atlas on Cardiovascular Disease Prevention and Control.
Geneva, Switzerland: World Health Organization, 2011

[3] Dhomse Kanchan B and Mahale Kishor M. et al. “Study of
Machine Learning Algorithms for Special Disease Prediction Conference on Global Trends in Signal Processing, Information Computing and Communication.

[4] R.Kavitha and E.Kannan et al. “An Efficient Framework for
Heart Disease Classification using Feature Extraction and Feature

[5] Shan Xu ,Tiangang Zhu, Zhen Zang, Daoxian Wang, Junfeng Hu
Based on CFS Subset Evaluation and Random Forest Conference on Big Data Analysis

[7] Kanika Pahwa and Ravinder Kumar et al. “Prediction of Heart
4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON).

[8] Seyedamin Pouriyeh, Sara Vahid, Giovanna Sannino, Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease”, 22nd IEEE Symposium on Computers and Communication (ISCC 2017): Workshops - ICTS4eHealth 2017

[9] Hanen Bouali and Jalel Akaichi et al. “Comparative study of
Different classification techniques, heart Diseases use Case.”, 2014 13th International Conference on Machine Learning and Applications