PROJECT REPORT

on

HEART DISEASE PREDICTION & MODEL COMPARISON

(CSE V Semester Mini project)

2023-2024

Submitted to:

Mr. Nitin Thapliyal

(CC-CSE-B-V-Sem)

Submitted by:

Mr. Abhinav Bansal

Roll. No:2118072

CSE-B-V-Sem

Session: 2023-2024

DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

GRAPHIC ERA HILL UNVERSITY, DEHRADUN

CERTIFICATE

Certified that Mr. Abhinav Bansal (Roll No.- 2118072) has developed mini project on “Heart Disease Prediction and Model Comparison” for the CS V Semester Mini Project Lab in Graphic Era Hill University, Dehradun. The project carried out by Students is their own work as best of my knowledge.

Mr. Nitin Thapliyal

Class Co-Ordinator

CSE-B-V-Sem

(CSE Department)

GEHU Dehradun

ACKNOWLEDGMENT

I would like to express our gratitude to my parents for their continuing support and encouragement. We also wish to thank them for providing us with the opportunity to reach this far in our studies.

We would like to thank particularly our project Co-ordinator Mr Nitin Thapliyal for his patience, support and encouragement throughout the completion of this project and having faith in us.

At last but not the least We greatly indebted to all other persons who directly or indirectly helped us during this work.

Mr. Abhinav Bansal

Roll No.- 2118072

CSE-B-V-Sem

Session: 2023-2024

GEHU, Dehradun

ABSTRACT

Nowadays, health disease are increasing day by day due to lifestyle, hereditary. Especially, heart disease has become more common these days, i.e. life of people is at risk. Each individual has different values for Blood pressure, cholesterol and pulse rate. But according to medically proven results the normal values of Blood pressure is 120/90, Cholesterol is 100-129 mg/dL ,Pulse rate is 72, Fasting Blood Sugar level is 100 mg/dL ,Heart rate is 60-100 bpm, ECG is normal, Width of major vessels is 25 mm (1 inch) in the aorta to only 8 µm in the capillaries.

This paper gives the survey about different classification techniques used for predicting the risk level of each person based on age, gender, Blood pressure, cholesterol, pulse rate.

"Disease Prediction" system based on predictive modelling predicts the disease of the user on the basis of the symptoms that user provides as an input to the system. The system analyses the symptoms provided by the user as input and gives the probability of the disease as an output. Disease Prediction is done by implementing 5 techniques such as Naïve Bayes, KNN, Decision Tree, Linear Regression and Random Forest Algorithms. These techniques calculate the probability of the disease. Therefore, average prediction accuracy probability 83% is obtained.

TABLE OF CONTENTS

CHAPTER: 1 INTRODUCTION

1.1 Introduction

In day to day life many factors that affect a human heart. Many problems are occurring at a rapid pace and new heart diseases are rapidly being identified. In today's world of stress Heart, being an essential organ in a human body which pumps blood through the body for the blood circulation is essential and its health is to be conserved for a healthy living. The health of a human heart is based on the experiences in a person's life and is completely dependent on professional and personal behaviors of a person. There may also be several genetic factors through which a type of heart disease is passed down from generations. According to the World Health Organization, every year more than 12 million deaths are occurring worldwide due to the various types of heart diseases which is also known by the term cardiovascular disease. The term Heart disease includes many diseases that are diverse and specifically affect the heart and the arteries of a human being. Even young aged people around their 20-30 years of lifespan are getting affected by heart diseases. The increase in the possibility of heart disease among young may be due to the bad eating habits, lack of sleep, restless nature, depression and numerous other factors such as obesity, poor diet, family history, high blood pressure, high blood cholesterol, idle behavior, family history, smoking and hypertension.

The diagnosis of the heart diseases is a very important and is itself the most complicated task in the medical field. All the mentioned factors are taken into consideration when analyzing and understanding the patients by the doctor through manual check-ups at regular intervals of time. The symptoms of heart disease greatly depend upon which of the discomfort felt by an individual. Some symptoms are not usually identified by the common people. However, common symptoms include chest pain, breathlessness, and heart palpitations. The chest pain common to many types of heart disease is known as angina, or angina pectoris, and occurs when a part of the heart does not receive enough oxygen. Angina may be triggered by stressful events or physical exertion and normally lasts under 10 minutes. Heart attacks can also occur as a result of different types of heart disease. The signs of a heart attack are like angina except that they can occur during rest and tend to be more severe. The symptoms of a heart attack can sometimes resemble indigestion. Heartburn and a stomach ache can occur, as well as a heavy feeling in the chest. Other symptoms of a heart attack include pain that travels through the body, for example from the chest to the arms, neck, back, abdomen, or jaw, lightheadedness and dizzy sensations, profuse sweating, nausea and vomiting.

Heart failure is also an outcome of heart disease, and breathlessness can occur when the heart becomes too weak to circulate blood. Some heart conditions occur with no symptoms at all, especially in older adults and individuals with diabetes. The term 'congenital heart disease' covers a range of conditions, but the general symptoms include sweating, high levels of fatigue, fast heartbeat and breathing, breathlessness, chest pain. However, these symptoms might not develop until a person is older than 13 years. In these types of cases, the diagnosis becomes an intricate task requiring great experience and high skill. A risk of a heart attack or the possibility of the heart disease if identified early, can help the patients take precautions and take regulatory measures. Recently, the healthcare industry has been generating huge amounts of data about patients and their disease diagnosis reports are being especially taken for the prediction of heart attacks worldwide. When the data about heart disease is huge, the machine learning techniques can be implemented for the analysis.

Data Mining is a task of extracting the vital decision-making information from a collective of past records for future analysis or prediction. The information may be hidden and is not identifiable without the use of data mining. The classification is one data mining technique through which the future outcome or predictions can be made based on the historical data that is available. The medical data mining made a possible solution to integrate the classification techniques and provide computerized training on the dataset that further leads to exploring the hidden patterns in the medical data sets which is used for the prediction of the patient's future state. Thus, by using medical data mining it is possible to provide insights on a patient's history and is able to provide clinical support through the analysis. For clinical analysis of the patients, these patterns are very much essential. In simple English, the medical data mining uses classification algorithms that is a vital part for identifying the possibility of heart attack before the occurrence. The classification algorithms can be trained and tested to make the predictions that determine the person's nature of being affected by heart disease.

In this research work, the supervised machine learning concept is utilized for making the predictions. A comparative analysis of the three data mining classification algorithms namely Random Forest, Decision Tree and Naïve Bayes are used to make predictions. The analysis is done at several levels of cross validation and several percentage of percentage split evaluation methods respectively. The Stat Log dataset from UCI machine learning repository is utilized for making heart disease predictions in this research work. The predictions are made using the classification model that is built from the classification algorithms when the heart disease dataset is used for training. This final model can be used for prediction of any types of heart diseases.

Health-care is a field of the most needed service and an economically 2ndlargest industry in 21stcentury. While we talk about the affordability and quality assurance in health-care industry, several statistical analyses are carried on making health solutions more precise and flawless in this current era of increasing health problems and chronic diseases. Advancements on data driven intelligent technologies is disease diagnosis and detection, treatment and research are remarkable. Medical image analysis, symptom-based disease prediction is the part where the most sought-after brains are working. In this paper we aim to present our proposed model on the prediction on diagnosis of cardio vascular disease with ECG analysis and symptom-based detection. The model aims to be researched and advance in further to become robust and end to end reliable research tool. We will discuss about the classical methods and algorithms implemented on CVD prediction, gradual advancements, draw comparison of performance among the existing systems and propose an enhanced multi-module system performing better in terms of accuracy and feasibility. Implementation, training and testing of the modules have been done on datasets obtained from UCI and Physio net data repositories. Data format have been modified in case of the ECG report data for betterment of action by the convolutional neural network used in our research and in the risk prediction module we have chosen attributes for training and implementing the multi-layered neural network developed by us. The further research and advancement possibilities are also mentioned in the paper.

1.2 Literature Review

According to Ordonez the heart disease can be predicted with some basic attributes taken from the patient and in their work have introduced a system that includes the characteristics of an individual human being based on totally 13 basic attributes like sex, blood pressure, cholesterol and others to predict the likelihood of a patient getting affected by heart disease. They have added two more attributes i.e. fat and smoking behavior and extended the research dataset. The data mining classification algorithms such as Decision Tree, Naive Bayes, and Neural Network are utilized to make predictions and the results are analyzed on Heart disease database.

Yilmaz, [17]have proposed a method that uses least squares support vector machine (LS-SVM) utilizing a binary decision tree for classification of cardiotocogram to find out the patient condition.

Duff, et al. have done a research work involving five hundred and thirty-three patients who had suffered from cardiac arrest and they were integrated in the analysis of heart disease probabilities. They performed classical statistical analysis and data mining analysis using mostly Bayesian networks.
Frawley, et al. have performed a work on prediction of survival of Coronary heart disease (CHD) which is a challenging research problem for medical society. They also used 10-fold cross-validation methods

to determine the impartial estimate of the three prediction models for performance comparison purposes.

Lee et al. proposed a novel methodology to expand and study the multi-parametric feature along with linear and nonlinear features of Heart Rate Variability diagnosing cardiovascular disease. They have carried out various experiments on linear and non-linear features to estimate several classifiers, e.g., Bayesian classifiers, CMAR, C4.5 and SVM. Based on their experiments, SVM outperformed the other classifiers.

Noh et al. suggested a classification method which is an associative classifier that is constructed based on the efficient FP-growth method. Because the volume of patterns can be diverse and huge, they offered a rule to measure the cohesion and in turn allow a tough choice of pruning patterns in the pattern-generating process.

Parthiban, et al. have proposed a new work in which the heart disease is identified and predicted using the proposed Coactive Neuro-Fuzzy Inference System (CANFIS). Their model works based on the collective nature of neural network adaptive capabilities and based on the genetic algorithm along with fuzzy logic in order to diagnose the occurrence of the disease. The performance of the proposed CANFIS model was evaluated in terms of training performances and classification accuracies. Finally, their results show that the proposed CANFIS model has great prospective in predicting the heart disease.

Singh, et al. have done a work using, one partition clustering algorithm (K-Means) and one hierarchical clustering algorithm (agglomerative). K-means algorithm has higher effectiveness and scalability and converges fast when production with large data sets. Hierarchical clustering constructs a hierarchy of clusters by either frequently merging two smaller clusters into a larger one or splitting a larger cluster into smaller ones. Using WEKA data mining tool, they have calculated the performance of k-means and hierarchical clustering algorithm on the basis of accuracy and running time.

1.3 What is heart disease?

Heart disease describes a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease; heart rhythm problems (arrhythmias); and heart defects you're born with (congenital heart defects), among others. The term "heart disease" is often used interchangeably with the term "cardiovascular disease." Cardiovascular disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke. Other heart conditions, such as those that affect your

heart's muscle, valves or rhythm, also are considered forms of heart disease. Heart failure is a serious condition with high prevalence (about2% in the adult population in developed countries, and or than8%inpatients olderthan75years). About 3 – 5%ofhospitaladmissions are linked with heart failure incidents. Heart failure is the first cause of admission by healthcare professionals in their clinical practice. The costs are very high, reaching up to 2% of the total health costs in the developed countries. Building an effective disease management strategy requires analysis of large amount of data, early detection of the disease, assessment of the severity and early prediction of adverse events. This will inhibit the progression of the disease, will improve the quality of life of the patients and will reduce the associated medical costs. Toward this direction machine learning techniques have been employed. The aim of this paper is to present the state-of-the-art of the machine learning methodologies applied for the assessment of heart failure. More specifically, models predicting the presence, estimating the subtype, assessing the severity of heart failure and predicting the presence of adverse events, such as destabilizations, re-hospitalizations, and mortality are presented. According to the authors' knowledge, it is the first time that such a comprehensive review, focusing on all aspects of the management of heart failure, is presented.

1.4 Risk factors

Risk factors for developing heart disease include

- Age. Aging increases your risk of damaged and narrowed arteries and weakened or thickened heart muscle.
- Sex. Men are generally at greater risk of heart disease. However, women's risk increases after menopause.
- Family history. A family history of heart disease increases your risk of coronary artery disease, especially if a parent developed it at an early age (before age 55 for a male relative, such as your brother or father, and 65 for a female relative, such as your mother or sister).
- Smoking. Nicotine constricts your blood vessels, and carbon monoxide can damage their inner lining, making them more susceptible to atherosclerosis. Heart attacks are more common in smokers than in nonsmokers.
- Certain chemotherapy drugs and radiation therapy for cancer. Some chemotherapy drugs and radiation therapies may increase the risk of cardiovascular disease.
- Poor diet. A diet that's high in fat, salt, sugar and cholesterol can contribute to the development of heart disease.

- High blood pressure. Uncontrolled high blood pressure can result in hardening and thickening of your arteries, narrowing the vessels through which blood flows.
- High blood cholesterol levels. High levels of cholesterol in your blood can increase the risk of formation of plaques and atherosclerosis.
- Diabetes. Diabetes increases your risk of heart disease. Both conditions share similar risk factors, such as obesity and high blood pressure.
- Obesity. Excess weight typically worsens other risk factors.
- Physical inactivity. Lack of exercise also is associated with many forms of heart disease and some of its other risk factors, as well.
- Stress. Unrelieved stress may damage your arteries and worsen other risk factors for heart disease.
- Poor hygiene. Not regularly washing your hands and not establishing other habits that can help prevent viral or bacterial infections can put you at risk of heart infections, especially if you already have an underlying heart condition. Poor dental health also may contribute to heart disease.

1.5 Complications

Complications of heart disease include:

- Heart failure. One of the most common complications of heart disease, heart failure occurs when your heart can't pump enough blood to meet your body's needs. Heart failure can result from many forms of heart disease, including heart defects, cardiovascular disease, valvular heart disease, heart infections or cardiomyopathy.
- Heart attack. A blood clot blocking the blood flow through a blood vessel that feeds the heart causes a heart attack, possibly damaging or destroying a part of the heart muscle. Atherosclerosis can cause a heart attack.
- Stroke. The risk factors that lead to cardiovascular disease also can lead to an ischemic stroke, which happens when the arteries to your brain are narrowed or blocked so that too little blood reaches your brain. A stroke is a medical emergency — brain tissue begins to die within just a few minutes of a stroke.
- Aneurysm. A serious complication that can occur anywhere in your body, an aneurysm is a bulge in the wall of your artery. If an aneurysm bursts, you may face life-threatening internal bleeding.
- Peripheral artery disease. Atherosclerosis also can lead to peripheral artery disease. When you develop peripheral artery disease, your extremities — usually your legs — don't receive enough blood flow. This causes symptoms, most notably leg pain when walking (claudication).

- Sudden cardiac arrest. Sudden cardiac arrest is the sudden, unexpected loss of heart function, breathing and consciousness, often caused by an arrhythmia. Sudden cardiac arrest is a medical emergency. If not treated immediately, it is fatal, resulting in sudden cardiac death.

1.6 Prevention

Certain types of heart disease, such as heart defects, can't be prevented. However, you can help prevent many other types of heart disease by making the same lifestyle changes that can improve your heart disease, such as:

- Quit smoking
- Control other health conditions, such as high blood pressure, high cholesterol and diabetes
- Exercise at least 30 minutes a day on most days of the week
- Eat a diet that's low in salt and saturated fat
- Maintain a healthy weight
- Reduce and manage stress
- Practice good hygiene

CHAPTER: 2 PRESENT WORK

2.1 About Heart Disease

Heart Diseases affect a large population in today's world, where the lifestyle is moved from active to comfort-oriented. We live in era of fast foods. Which build up cholesterol, diabetes and many more factors which in turn affects the heart in some way or the other. According to the World Health Organization Cardiovascular Diseases (CVD) or Heart Diseases cause more death than any other diseases globally. The amount of data in medical sectors is quite large and computerized as well. They are not utilized or put to any use. This data if studied and analyzed could be put to good use like prediction of diseases or even prevent them. Diseases such as cancer can be detected, and the stage can also be predicted by training dataset with pictures of cancer cells. Similarly, heart disease can be predicted based on aspects like cholesterol, diabetes, heart rate etc. The prediction of heart diseases is a challenge and very risky. We observed that in some cases solutions of problems does not rely on a single method. It varies from situation to situation. It is also a challenge as most of the data are sparse or missing as they were not stored in the motive of analyzing. We therefore set out goal to finding which method would be best for predicting the diseases using data of four different hospitals from four different places. This is a comparative study on the efficiency of different data mining techniques such as Logical Regression, Random forest, K-Nearest Neighbors, Decision Tree in predicting heart diseases. The Data Mining techniques are analyzed, and the accuracy of prediction is noted for each method used. The result showed that heart diseases can be predicted with accuracy of above 90%. Cardiovascular diseases are the leading cause of death globally, resulted in 17.9 million deaths (32.1%) in 2015, up from 12.3 million (25.8%) in 1990. It is estimated that 90% of CVD is preventable. There are many risk factors for heart diseases that we will take a closer look at. The main objective of this study is to build a model that can predict the heart disease occurrence, based on a combination of features (risk factors) describing the disease. Different machine learning techniques will be implemented and compared upon standard performance metric such as accuracy.

The dataset used for this study was taken from UCI machine learning repository.

2.2 Data Structure & Description

2.2.1 Importing Libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

%matplotlib inline

import os
print(os.listdir())

import warnings
warnings.filterwarnings('ignore')
```

```
['.config', 'heart.csv', 'sample_data']
```

2.2.2 Load Data

```
data = pd.read_csv("heart.csv")
```

2.2.3 Check the type of the Dataset

```
type(data)
```

```
pandas.core.frame.DataFrame
```

2.2.4 Check the shape of the data

```
data.shape
```

```
(303, 14)
```

2.2.5 Check the top four columns of the dataset

```
1 data.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

2.2.6 Dataset Description

```
1 data.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

```
1 data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
age         303 non-null int64
sex         303 non-null int64
cp          303 non-null int64
trestbps    303 non-null int64
chol        303 non-null int64
fbs         303 non-null int64
restecg     303 non-null int64
thalach     303 non-null int64
exang       303 non-null int64
oldpeak     303 non-null float64
slope       303 non-null int64
ca          303 non-null int64
thal        303 non-null int64
target      303 non-null int64
dtypes: float64(1), int64(13)
memory usage: 33.2 KB
```

The dataset used in this project contains 14 variables. The independent variable that needs to be predicted, 'diagnosis', determines whether a person is healthy or suffer from heart disease.

Experiments with the Cleveland database have concentrated on endeavors to distinguish disease presence (values 1, 2, 3, 4) from absence (value 0). There are several missing attribute values, distinguished with symbol '?'. The header row is missing in this dataset, so the column names have to be inserted manually.

Features information:

- age - age in years
- sex - sex (1 = male; 0 = female)
- chest pain - chest pain type (1 = typical angina; 2 = atypical angina; 3 = non anginal pain; 4 = asymptomatic)
- blood pressure - resting blood pressure (in mm Hg on admission to the hospital)
- serum cholesterol - serum cholesterol in mg/dl
- fasting blood sugar - fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
- electrocardiographic - resting electrocardiographic results (0 = normal; 1 = having ST-T; 2 = hypertrophy)
- max heart rate - maximum heart rate achieved
- induced angina - exercise induced angina (1 = yes; 0 = no)
- ST depression - ST depression induced by exercise relative to rest
- slope - the slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = down sloping)
- no of vessels - number of major vessels (0-3) colored by fluoroscopy
- thalassemia - 3 = normal; 6 = fixed defect; 7 = reversable defect
- diagnosis - the predicted attribute - diagnosis of heart disease (angiographic disease status) (Value 0 = < 50% diameter narrowing; Value 1 = > 50% diameter narrowing)

2.2.7 Types of Features

Categorical features (Has two or more categories and each value in that feature can be categorized by them): sex, chest pain

Ordinal features (Variable having relative ordering or sorting between the values): fasting blood sugar, electrocardiographic, induced angina, slope, no of vessels, thalassemia, diagnosis

Continuous features (Variable taking values between any two points or between the minimum or maximum values in the feature column): age, blood pressure, serum cholesterol, max heart rate, ST depression

2.2.8 Some Random Data Columns

```
data.sample(5)
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
7	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
269	56	1	0	130	283	1	0	103	1	1.6	0	0	3	0
254	59	1	3	160	273	0	0	125	0	0.0	2	0	2	0
64	58	1	2	140	211	1	0	165	0	0.0	2	0	2	1
103	42	1	2	120	240	1	1	194	0	0.8	0	0	3	1

2.2.9 Checking the missing Data

```
[10]  data.isnull().sum()
```

```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

```
data.isnull().sum().sum()
```

```
0
```

No data is missing, which is good.

2.3 Uniqueness of Sex Column –

Two sex types: 1 is male and 0 is female.

```
data["sex"].unique()
```

```
array([1, 0])
```

2.4 Check the percentage

```
countFemale = len(data[data.sex == 0])
countMale = len(data[data.sex == 1])
print("Percentage of Female Patients:{:.2f}%".format((countFemale)/(len(data.sex))*100))
print("Percentage of Male Patients:{:.2f}%".format((countMale)/(len(data.sex))*100))
```

```
Percentage of Female Patients:31.68%
Percentage of Male Patients:68.32%
```

2.5 Heart Disease frequency for ages

```
pd.crosstab(data.age,data.target).plot(kind="bar",figsize=(20,6))
plt.title('Heart Disease Frequency for Ages')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.savefig('heartDiseaseAndAges.png')
plt.show()
```

2.6 Heart Disease frequency for Sex

```
pd.crosstab(data.sex,data.target).plot(kind="bar",figsize=(20,10),color=['blue','#AA1111' ])
plt.title('Heart Disease Frequency for Sex')
plt.xlabel('Sex (0 = Female, 1 = Male)')
plt.xticks(rotation=0)
plt.legend(["Don't have Disease", "Have Disease"])
plt.ylabel('Frequency')
plt.show()
```

2.7 Making the data column names easily recognizable

```
data.columns = ['age', 'sex', 'chest_pain_type', 'resting_blood_pressure', 'cholesterol', 'fasting_blood_sugar', 'rest_ecg', 'max_heart_rate_achieved',
       'exercise_induced_angina', 'st_depression', 'st_slope', 'num_major_vessels', 'thalassemia', 'target']
```

2.8 Heart Disease frequency for Fasting Blood Sugar

```
pd.crosstab(data.fasting_blood_sugar,data.target).plot(kind="bar",figsize=(20,10),color=['#4286f4','#f49242'])
plt.title("Heart disease according to FBS")
plt.xlabel('FBS- (Fasting Blood Sugar > 120 mg/dl) (1 = true; 0 = false)')
plt.xticks(rotation=90)
plt.legend(["Don't Have Disease", "Have Disease"])
plt.ylabel('Disease or not')
plt.show()
```

2.9 Analyzing the chest pain

```
data["chest_pain_type"].unique()
```

```
array([3, 2, 1, 0])
```

```
plt.figure(figsize=(26, 10))
sns.barplot(data["chest_pain_type"],y)
```

2.10 Analyzing the Resting Blood Pressure

```
data["resting_blood_pressure"].unique()
```

```
array([145, 130, 120, 140, 172, 150, 110, 135, 160, 105, 125, 142, 155,
       104, 138, 128, 108, 134, 122, 115, 118, 100, 124,  94, 112, 102,
       152, 101, 132, 148, 178, 129, 180, 136, 126, 106, 156, 170, 146,
       117, 200, 165, 174, 192, 144, 123, 154, 114, 164])
```

```
plt.figure(figsize=(26, 10))
sns.barplot(data["resting_blood_pressure"],y)
```

2.11 Data Preparation

Total Among 303 data's randomly 242 are chosen for Training and 61 are chosen for Testing.

```
[128]  1 X_train.shape
```

```
(242, 13)
```

```
[129]  1 X_test.shape
```

```
(61, 13)
```

```
[130]  1 Y_train.shape
```

```
(242,)
```

```
[131]  1 Y_test.shape
```

```
(61,)
```

CHAPTER: 3 MODELLING & PREDICTING WITH ML

The main goal of the entire project is to predict heart disease occurrence with the highest accuracy. In order to achieve this, we will test several classification algorithms. This section includes all results obtained from the study and introduces the best performer according to accuracy metric. We have chosen several algorithms typical for solving supervised learning problems throughout classification methods.

Modelling and predicting with Machine Learning

The main goal of the entire project is to predict heart disease occurrence with the highest accuracy. In order to achieve this, we will test several classification algorithms. This section includes all results obtained from the study and introduces the best performer according to accuracy metric. We have chosen several algorithms typical for solving supervised learning problems throughout classification methods.

3.1 Logistic Regression

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes.

Sigmoid function:

$$S(z)=1/1+e^{\wedge}-z$$

Cost Function:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \text{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$\text{Cost}(h_\theta(x), y) = -\log(h_\theta(x)) \quad \text{if y} = 1$$

$$\text{Cost}(h_\theta(x), y) = -\log(1 - h_\theta(x)) \quad \text{if y} = 0$$

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))]$$

Vectorized cost function:

$$h = g(X\theta)$$

$$J(\theta) = \frac{1}{m} \cdot \left(-y^T \log(h) - (1-y)^T \log(1-h)\right)$$

Code:

```
from sklearn.linear_model import LogisticRegression

lr = LogisticRegression()

lr.fit(X_train,Y_train)

Y_pred_lr = lr.predict(X_test)
```

```
Y_pred_lr.shape
```

```
Y_an = lr.predict([[44,1,1,130,219,0,0,188,0,0,2,0,2]])
print(Y_an)
Y_an = lr.predict([[65,0,0,150,225,0,0,114,0,1,1,3,3]])
print(Y_an)
```

#prediction

```
score_lr = round(accuracy_score(Y_pred_lr,Y_test)*100,2)

print("The accuracy score achieved using Logistic Regression is: "+str(score_lr)+" %")
```

3.2 Random Forest

Random Forest is a supervised learning algorithm. Random forest can be used for both classification and regression problems, by using random forest regressor we can use random forest on regression problems. But we have used random forest on classification in this project so we will only consider the classification part.

Code:

```
from sklearn.ensemble import RandomForestClassifier

max_accuracy = 0


for x in range(2000):
    rf = RandomForestClassifier(random_state=x)
    rf.fit(X_train,Y_train)
    Y_pred_rf = rf.predict(X_test)
    current_accuracy = round(accuracy_score(Y_pred_rf,Y_test)*100,2)
    if(current_accuracy>max_accuracy):
        max_accuracy = current_accuracy
        best_x = x

print(max_accuracy)
print(best_x)

rf = RandomForestClassifier(random_state=best_x)
rf.fit(X_train,Y_train)
Y_pred_rf = rf.predict(X_test)
```

```
Y_pred_rf.shape
```

```
Y_an = rf.predict([[44,1,1,130,219,0,0,188,0,0,2,0,2]])
print(Y_an)
Y_an = rf.predict([[65,0,0,150,225,0,0,114,0,1,1,3,3]])
print(Y_an)
```
#prediction

```
score_rf = round(accuracy_score(Y_pred_rf,Y_test)*100,2)

print("The accuracy score achieved using Decision Tree is: "+str(score_rf)+" %")
```

3.3 Naïve Bayes

Bayes' Theorem is stated as:

$$P(h|d) = (P(d|h) * P(h)) / P(d)$$

- P(h|d) is the probability of hypothesis h given the data d. This is called the posterior probability.
- P(d|h) is the probability of data d given that the hypothesis h was true.
- P(h) is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.
- P(d) is the probability of the data (regardless of the hypothesis). we are interested in calculating the posterior probability of P(h|d) from the prior probability p(h) with P(D) and P(d|h). After calculating the posterior probability for a number of different hypotheses, we will select the hypothesis with the highest probability. This is the maximum probable hypothesis and may formally be called the (MAP) hypothesis. This can be written as: MAP(h) = max(P(h|d))

Code:

```
from sklearn.naive_bayes import GaussianNB

nb = GaussianNB()

nb.fit(X_train,Y_train)

Y_pred_nb = nb.predict(X_test)
```

```
Y_pred_nb.shape
```

```
Y_an = nb.predict([[44,1,1,130,219,0,0,188,0,0,2,0,2]])
print(Y_an)
Y_an = nb.predict([[65,0,0,150,225,0,0,114,0,1,1,3,3]])
print(Y_an)
```
#prediction

```
score_nb = round(accuracy_score(Y_pred_nb,Y_test)*100,2)

print("The accuracy score achieved using Naive Bayes is: "+str(score_nb)+" %")
```

3.4 Result

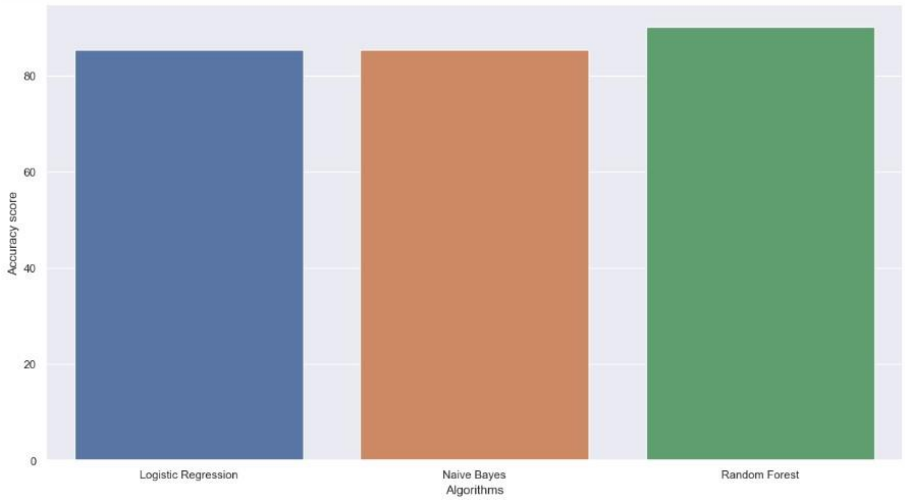
```
scores = [score_lr,score_nb,score_rf]
algorithms = ["Logistic Regression","Naive Bayes","Random Forest"]

for i in range(len(algorithms)):
    print("The accuracy score achieved using "+algorithms[i]+" is: "+str(scores[i])+" %")
```

```
The accuracy score achieved using Logistic Regression is: 85.25 %
The accuracy score achieved using Naive Bayes is: 85.25 %
The accuracy score achieved using Random Forest is: 90.16 %
```

```
sns.set(rc={'figure.figsize':(15,8)})
plt.xlabel("Algorithms")
plt.ylabel("Accuracy score")

sns.barplot(x=algorithms,y=scores)
```



CHAPTER: 4 CONCLUSION

4.1 Inference

The overall objective of our project is to predict accurately with less number of tests and attributes the presence of heart disease. In this project, fourteen attributes are considered which form the primary basis for tests and give accurate results more or less. Many more input attributes can be taken but our goal is to predict with less number of attributes and faster efficiency to predict the risk of having heart disease at a particular age span. Five data mining classification techniques were applied namely K-Nearest Neighbor, Naive Bayes, Decision Tree, Random Forest & Logistic Regression. It is shown that Random Forest has better accuracy than the other techniques.

This is the most effective model to predict patients with heart disease. This project could answer complex queries, each with its own strength with respect to ease of model interpretation, access to detailed information and accuracy.

This project can be further enhanced and expanded. For example, it can incorporate other medical attributes besides the 14 attributes we used. It can also incorporate other data mining techniques, e.g., Time Series, Clustering and Association Rules. Continuous data can also be used instead of just categorical data. Another area is to use Text Mining to mine the vast amount of unstructured data available.

This project is presented using data mining techniques. From logistic regression, KNN, Naive Bayes, Decision Tree, Random forest are used to develop the system. Random Forest proves the better results and assists the domain experts and even the person related to the medical field to plan for a better and early diagnosis for the patient. This system performs realistically well even without retraining.

4.2 Drawbacks

The Algorithms used in our project does not give a 100% accuracy, so the prediction is not 100% feasible. Clinical diagnosis and diagnosis using our project may differ slightly because the prediction is not 100% accurate. Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Clinical decisions are often made based on doctor’s intuition and experience rather than on the knowledge rich data collected from the dataset.

4.3 Future Scope

We are planning to introduce an efficient disease prediction system to predict the heart disease with better accuracy using Support Vector Machine (SVM). Our project aims to provide a web platform to predict the occurrences of disease based on various symptoms. The user can select various symptoms and can find the diseases with their probabilistic figures. Our project can be improved by implementing medicine suggestion to the patient along with the results. We can implement a feedback from the experienced doctors who can give their views and opinions about certain medicines /practices done by the doctor on the patient. We can implement a live chat option where the patient can chat with a doctor available regarding medication for the respective result for their symptoms. Our project could be used as a training tool for Nurses and Doctors who are freshly introduced in the field related to heart diseases. The patient can have a choice in choosing the medicines he/she should take in order to have a healthier life. Moreover, if implemented on a large scale it can be used in medical facilities like hospital, clinics where a patient wouldn't have to wait in long queues for treatment if he is feeling symptoms related to heart disease.

References

[1] M. K. Awang and F. Siraj, “Utilization of an artificial neural network in the prediction of heart disease,” Int. J. Bio-Science Bio-Technology, vol. 5, no. 4, pp. 159–165, 2013. [2] https://scholar.google.com/