A Novel Mechanism to predict Heart Disease using Machine Learning Techniques

Department of Computer Science and Engineering
Pabna University of Science and Technology, Pabna-6600

Course Title: Thesis
Course Code: CSE 4100 and CSE 4200

A thesis has been submitted to the Department of Computer Science and Engineering for the fulfillment of the requirement of Bachelor Degree in Computer Science and Engineering

Submitted By:
Md.Mahbub Alam Bablu
Roll Number:**150112**
Registration Number: 101637, Session: 2014-15

Supervised By:
S.M. Hasan Sazzad Iqbal
Assistant Professor, Department of Computer Science and Engineering
Pabna University of Science and Technology

February, 2020

DECLARATION

In accordance with rules and regulations of Pabna University of Science and Technology following declarations are made:

I hereby declare that this thesis has been done by me under the supervision of S.M. Hasan Sazzad Iqbal, Assistant Professor, Department of Computer Science and Engineering, Pabna University of Science and Technology, Pabna-6600.

I also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for awarding of any degree and any material reproduced in this thesis has been properly acknowledged.

Signature of the Examinee

CERTIFICATE

I am pleased to certify that Md. Mahbub Alam Bablu, Roll Number: 150112, Registration Number: 101637, Session: 2014-15 has performed a thesis work entitled ”A Novel Mechanism to predict Heart Disease using Machine Learning Techniques” under my supervision for the requirement of the completion of course entitled ”Thesis”. So far as I concern this is an original thesis that has been carried out for one year in the Department of Computer Science and Engineering, Pabna University of Science and Technology, Pabna-6600, Bangladesh.

To the best of my knowledge, this paper has not been duplicated from any other paper or submitted to elsewhere prior submission to the department.

S.M. Hasan Sazzad Iqbal
Assistant Professor,
Department of Computer Science and Engineering
Pabna University of Science and Technology, Pabna-6600.
Bangladesh.

Dedication

I dedicate this paper to my loving family specially my parents......!

Acknowledgements

First of all I would like to admit my gratefulness to the Almighty God for enabling me to perform this thesis successfully. I would like to express my deepest sense of gratitude to my honorable supervisor S.M. Hasan Sazzad Iqbal, Assistant Professor, Department of Computer Science and Engineering (CSE), Pabna University of Science Technology (PUST), for his scholastic supervision, valuable guidance, adequate encouragement and helpful discussion throughout the progress of this work. I am highly grateful to him for allowing me to pursuing this study under his supervision. I am deeply thankful to honorable chairman, Md. Mahmudul Hasan, and all the respectable teachers of the Department of Computer Science and Engineering, Pabna University of Science Technology, Pabna-6600, Bangladesh, for their encouragement to my research work. Finally, I am much grateful to my family members especially to my parents, all of my friends and well-wishers for their encouragement and supports.

February, 2020
Author

Abstract

Globally, heart disease is one of the most alarming disease nowadays. According to recent survey by WHO (World health organization) 17.9 million people die each year because of heart related diseases and it is increasing rapidly[1].So, this sector seek the attention of researchers to work with modern technology and tools for the early prediction of the disease in order to save millions of lives.There are number of research papers also have been done by using data mining and machine learning techniques[16].This paper analyzes various Machine Learning techniques namely Naive Bayes, Random Forest Classification, Decision tree and Support Vector Machine, K-Nearest Neighbor, Logistic Regression by using a qualified data set for Heart disease prediction.The main aim of this research paper includes finding the performances of these six different machine learning techniques and comparing their accuracies to find which perform better for prediction of the heart disease.The testing result showed that Support Vector Machine and Random Forest achieved model accuracies of 87.91%.

Keywords: Heart Disease; Machine Learning; Accuracy.

Contents

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

In this chapter we introduced our thesis overview, related work, motivation, objective and organization.In section 1.1 we discussed about related work; in section 1.2 we discussed about our thesis motivation; in section 1.3 we discussed about our thesis objective; in section 1.4 we discussed about the whole thesis paper organization; in section 1.5 we should give a shot discussion about this chapter.

Heart disease describes a range of conditions that affect our 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[7]. Other heart conditions, such as those that affect our heart's muscle, valves or rhythm, also are considered forms of heart disease. Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of clinical data analysis[15]. The amount of data in the healthcare industry is huge. Data mining turns the large collection of raw healthcare data into information that can help to make informed decisions and predictions. According to a news article, heart disease proves to be the leading cause of death for both women and men. The article states the following : About 610,000 people die of heart disease in the United States every year–that's 1 in every 4 deaths.1 Heart disease is the leading cause of death for both men and women. This makes heart disease a major concern to be dealt with. But it is difficult to identify heart disease because of several contributory risk factors such as high blood pressure, high cholesterol, abnormal pulse rate, and many other factors.Early stage detection of the disease and predicting the probability of a person to be at risk of heart disease can reduce the death rate[20]. Medical data mining techniques are used in medical data to extract meaningful patterns and knowledge. Medical information has redundancy, multi-attribution, incompleteness and a close relationship with time. The problem of using the massive volumes of data effectively becomes a major problem for the health sector. Data mining provides the methodology and technology to convert these data mounds into useful decision-making information[24]. This predication system for heart disease would facilitate Cardiologists in taking quicker decisions so that more patients can receive treatments within a shorter period of time, resulting in saving millions of life[14].

1.1 Related Works

1. A cloud based decision support system proposed by in order to helps the heart consultants during diagnosis process. This system used machine-learning methods for predicting the heart disease. The system was proposed to provide the assistance in affordable way, where the system have the capacity to integrate with existing system. In that research clustering method used for categorizing the dataset based on particular groups in unsupervised manner. The author in used an approach by implementing multiple clustering algorithms on heart disease dataset to understand the optimal solution, which can maximize the prediction accuracy ratio. ML approaches proved to be an effective in predicting the heart disease using historical data is further proved in a research conducted using Naïve Bayes, Decision Tree, support vector model and other models. The results indicated that the support vector machine provided the optimal results between other implemented approaches.

2. Bashir et al., (2019) attempted to improve the performance of heart disease prediction using feature selection approach. Different models such as Naïve Bayes, Random Forest and other used in the experiment implemented using Rapid Miner tool. The output indicated the high accuracy measured due to feature selection approach.

3. In another research, the Extreme Learning Machine techniques using feed forward neural network applied on Cleveland data based on 300 patients, suggested 80% accuracy in forecasting the heart disease in a patient.

4. In another research, the neural network applied using multi-layer perceptron, which also known as supervised learning. The system was proposed to determine the potential heart disease risk in a patient, using patient's historical data.

5. HF ratio using preserved ejection fraction is another work presented using multiple factors like strain rate, hypertensive situation, and velocity, where overall accuracy computed was more than 80%.

6. In the same way, another comparative study shown the performances of the multiple classifiers applied on two different tools; Matlab and Weka. Overall, the accuracy of the decision tree, Linear SVM and other models was recorded between 52% to 67.7%,

although the accuracy were considerably low.

7. Priti Chandra et al. proposed a early prediction of heart disease using Naive Bayes algorithm. And this research result of accuracy level was 86.29%.

8. Cemil et al.Proposed a application of knowledge discovering process on prediction of stroke patients they used Artificial Neural Network and Support Vector Machine and achieving accuracy in order to 85.09% for ANN and 84.26%.

1.2 Motivation

We discussed some thesis paper related to predicting heart disease using various machine learning techniques. All the previous work used traditional machine learning algorithms to predict the heart disease.And they gave a good accuracy.But we wanted to improve the previous accuracy by implementing of our selected machine learning algorithms. Here in this research, we try to tuning some of our applied machine learning algorithms to produce a better accuracy result.We get a better accuracy result by tuning support vector machine and random forest algorithm and they gave a satisfactory accuracy.

1.3 Thesis Objective

The main objective of this study is to predict weather a patient is affected with heart disease or not using different machine learning algorithms on a qualified dataset. Obtaining clear idea of our proposed machine learning techniques and analyze the result and comparing between the results of different machine learning techniques. We will analyze our techniques if there is any possibility to bring improvement for our results.

1.4 Organization of the Thesis

In these section we discussed about the organization of the thesis.

This chapter (CHAPTER 1: Introduction) presents an overview of the background of our work such as related work, motivation and our objective.

CHAPTER 2: Literature Review presents an overview of thesis literature, a clear concept about Machine Learning.Also discussed Importance of machine learning and their types. we also briefly discussed about our six machine learning algorithms that we want to apply.

CHAPTER 3: Simulation and Methodology represents of our dataset source, the structure of our dataset, data pre-possesing and how we applied our methods to the dataset.And also analyze our methods by some parameters inorder to check the performance difference.

CHAPTER 4: This chapter contains the accuracy performances and roc_auc score performances of different models and their graphical representations.Lastly there is a discussion to declare which models scores high by comparing the results.

CHAPTER 5: Conclusion is the last chapter in this paper. These chapter represents a clear discussion about all the workflows with results analysis. Then a short description about the future work availability in these research fields.

1.5 Discussion

This is the introduction chapter and this chapter just introduces about our thesis, previous related work done by different authors and our goals. Also its shows the blueprint of our work.

CHAPTER 2

LITERATURE REVIEW

In this chapter we introduced our thesis literature. In section 2.1 we discussed about Machine Learning and how it works; in section 2.2 we discussed importance of machine learning; in section 2.3 we discussed about types of machine learning technique; in section 2.4 we discussed about our machine learning algorithms that we want to use in this research;in section 2.5 we should summarize the chapter.

2.1 Machine Learning

Arthur Samuel, a pioneer in the field of artificial intelligence and computer gaming, coined the term "**Machine Learning**". He defined machine learning as – "**Field of study that gives computers the capability to learn without being explicitly programmed**". In a very layman manner, Machine Learning(ML) can be explained as automating and improving the learning process of computers based on their experiences without being actually programmed i.e. without any human assistance. The process starts with feeding good quality data and then training our machines(computers) by building machine learning models using the data and different algorithms. The choice of algorithms depends on what type of data do we have and what kind of task we are trying to automate[17].

How ML works?

- Gathering past data in any form suitable for processing.The better the quality of data, the more suitable it will be for modeling.

- Data Processing – Sometimes, the data collected is in the raw form and it needs to be pre-processed. Example: Some tuples may have missing values for certain attributes, an, in this case, it has to be filled with suitable values in order to perform machine learning or any form of data mining. Missing values for numerical attributes such as the price of the house may be replaced with the mean value of the attribute whereas missing values for categorical attributes may be replaced with the attribute with the highest mode. This invariably depends on the types of filters we use. If data is in the form of text or images then converting it to numerical form will be required, be it a list or array or matrix. Simply, Data is to be made relevant and consistent. It is to be converted into a format understandable by the machine.

- Divide the input data into training,cross-validation and test sets. The ratio between the respective sets must be 6:2:2

- Building models with suitable algorithms and techniques on the training set.

- Testing our conceptualized model with data which was not fed to the model at the time of training and evaluating its performance using metrics such as F1 score, precision and recall.

2.2 Importance of Machine Learning

Machine learning has several very practical applications that drive the kind of real business results – such as time and money savings – that have the potential to dramatically impact the future of your organization. At Interactions in particular, we see tremendous impact occurring within the customer care industry, whereby machine learning is allowing people to get things done more quickly and efficiently. Through Virtual Assistant solutions, machine learning automates tasks that would otherwise need to be performed by a live agent – such as changing a password or checking an account balance. This frees up valuable agent time that can be used to focus on the kind of customer care that humans perform best: high touch, complicated decision-making that is not as easily handled by a machine. At Interactions, we further improve the process by eliminating the decision of whether a request should be sent to a human or a machine: unique Adaptive Understanding technology, the machine learns to be aware of its limitations, and bail out to humans when it has a low confidence in providing the correct solution.

Machine learning has made dramatic improvements in the past few years, but we are still very far from reaching human performance[10]. Many times, the machine needs the assistance of human to complete its task. At Interactions, we have deployed Virtual Assistant solutions that seamlessly blend artificial with true human intelligence to deliver the highest level of accuracy and understanding.Data is the lifeblood of all business. Data-driven decisions increasingly make the difference between keeping up with competition or falling further behind.

2.3 Types of Machine Learning

Machine learning is sub-categorized to three types:

1.Supervised Learning

2.Unsupervised Learning

3.Reinforcement Learning

2.3.1 Supervised Learning

Supervised learning is the most popular paradigm for machine learning. It is the easiest to understand and the simplest to implement. It is very similar to teaching a child with the use of flash cards.

Given data in the form of examples with labels, we can feed a learning algorithm these example-label pairs one by one, allowing the algorithm to predict the label for each example, and giving it feedback as to whether it predicted the right answer or not[2]. Over time, the algorithm will learn to approximate the exact nature of the relationship between examples and their labels. When fully-trained, the supervised learning algorithm will be able to observe a new, never-before-seen example and predict a good label for it.

Supervised learning is often described as task-oriented because of this. It is highly focused on a singular task, feeding more and more examples to the algorithm until it can accurately perform on that task. This is the learning type that you will most likely encounter, as it is exhibited in many common applications.

2.3.2 Unsupervised Learning

Unsupervised learning is very much the opposite of supervised learning. It features no labels. Instead, our algorithm would be fed a lot of data and given the tools to understand the properties of the data. From there, it can learn to group, cluster, and/or organize the data in a way such that a human (or other intelligent algorithm) can come in and make sense of the newly organized data.

What makes unsupervised learning such an interesting area is that an overwhelming majority of data in this world is unlabeled. Having intelligent algorithms that can take our terabytes and terabytes of unlabeled data and make sense of it is a huge source of potential profit for many industries. That alone could help boost productivity in a number of fields[25]. For example, what if we had a large database of every research paper ever published and we had an unsupervised learning algorithms that knew how to group these in such a way so that we were always aware of the current progression within a particular domain of research. Now, you begin to start a research project yourself, hooking your work into this network that the algorithm can see. As you write your work up and take notes, the algorithm makes suggestions to you about related works, works you may wish to cite, and works that may even help you push that domain of research forward. With such a tool, your productivity can be extremely boosted. Because unsupervised learning is based upon the data and its properties, we can say that unsupervised learning is data-driven[3]. The outcomes from an unsupervised learning task are controlled by the data and the way its formatted.

2.3.3 Reinforcement Learning

Reinforcement learning is fairly different when compared to supervised and unsupervised learning. Where we can easily see the relationship between supervised and unsupervised (the presence or absence of labels), the relationship to reinforcement learning is a bit murkier. Some people try to tie reinforcement learning closer to the two by describing it as a type of learning that relies on a time-dependent sequence of labels, however, my opinion is that that simply makes things more confusing. We prefer to look at reinforcement learning as learning from mistakes. Place a reinforcement learning algorithm into any environment and it will make a lot of mistakes in the beginning. So long as we provide some sort of signal to the algorithm that associates good behaviors with a positive signal and bad behaviors with a negative one, we can reinforce our algorithm to prefer good behaviors over bad ones[11]. Over time, our learning algorithm learns to make less mistakes than it used to.

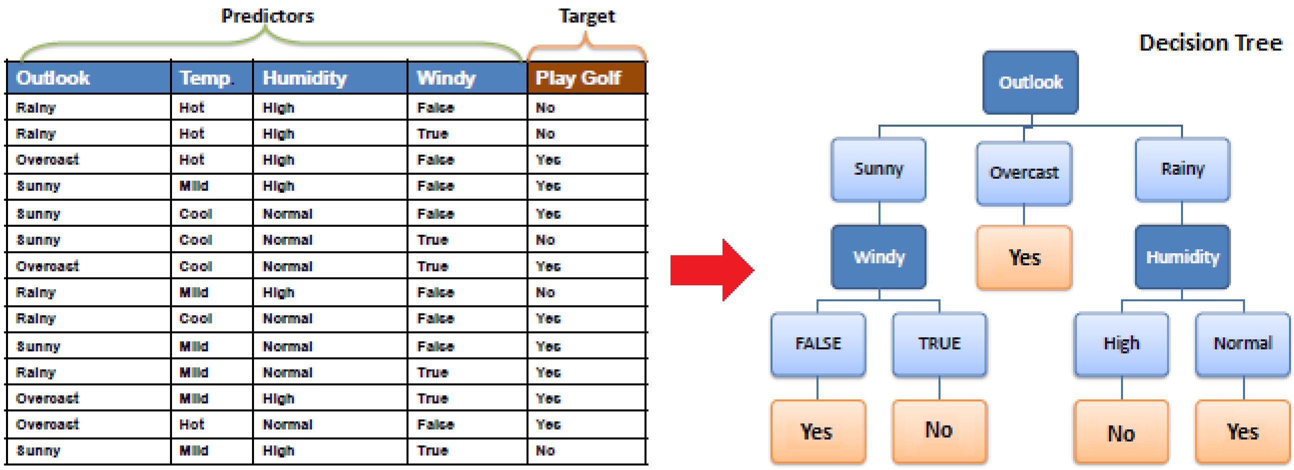
2.4 Algorithm Discussion

In machine learning we can use different algorithms otherwise known as classifiers to help us predict for our project. Here in our project we are looking forward to predict the number of patient that have heart disease and the number of patient that do not have heart disease running six algorithms to our data set. The reason we are going to use six is that it will allow us to get better and more reliable prediction. Because if we are using one algorithm or classifier and do not have anything else to compare it with then we cannot say that it a reliable prediction because it might be giving us a very good accuracy but this algorithm might not be the best or more appropriate one to use for our scenario. Whereas if we use more than one algorithm or classifier in our case six of them, we can compare them with one another and if we find one classifier is giving us accuracy that is not even in the ball park of the other algorithm provided accuracy we can understand that something is going wrong. It can be that the algorithm itself is not suitable for the job or we made a mistake in our coding. So using more than one algorithm is essential for any prediction based system. Now the algorithms that we have chosen to use in our project are: 1. Decision tree, 2. Naïve Bayes, 3.SVM (support vector machine) 4. Random Forest. 5.K-Nearest Neighbor(KNN) 6.Logistic Regression. We will be discussing each of those algorithms below.

2.4.1 DECISION TREE

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed[12]. The final result is a tree with decision nodes and leaf nodes[22]. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy). Leaf node (e.g., Play) represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data[15].

To build decision tree we have to consider two things. First one is entropy and Second one is information gain.



Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Figure 2.1: Decision Tree

Entropy E(X) = -$\sum p(X) \log p(X)$

Information Gain $I(X, Y) = E(X) - E(X|Y)$

Steps to evaluate the Decision tree:

Step 1: Calculate entropy of the target.

Step 2: The dataset is then split on the different attributes. The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain, or decrease in entropy.

Step 3: Choose attribute with the largest information gain as the decision node, divide the dataset by its branches and repeat the same process on every branch.

Step 4a: A branch with entropy of 0 is a leaf node.

Step 4b: A branch with entropy more than 0 needs further splitting.

Step 5: The ID3 algorithm is run recursively on the non-leaf branches, until all data is classified.

2.4.2 Naive Bayes

Naive Bayes is a kind of classifier which uses the Bayes Theorem. It predicts membership probabilities for each class such as the probability that given record or data point belongs to a particular class[23]. The class with the highest probability is considered as the most likely class[9]. This is also known as Maximum A Posteriori (MAP).

The MAP for a hypothesis is: MAP(H) = max(P(H| E)) = $max((P(E|H) * P(H))/P(E)) = max(P(E|H) * P(H))$

P(E) is evidence probability, and it is used to normalize the result. It remains same so, removing it won't affect.

Naive Bayes classifier assumes that all the features are unrelated to each other. Presence or absence of a feature does not influence the presence or absence of any other feature[18]. We can use Wikipedia example for explaining the logic i.e.,

A fruit may be considered to be an apple if it is red, round, and about 4 in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

In real datasets, we test a hypothesis given multiple evidence(feature). So, calculations become complicated. To simplify the work, the feature independence approach is used to 'uncouple' multiple evidence and treat each as an independent one.

P(H| $MultipleEvidences) = P((E1|H)) * P(E2|H)\ldots\ldots * P(En|H) * P(H)/P.$

2.4.3 Support Vector Machine

Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. But generally, they are used in classification problems. In 1960s, SVMs were first introduced but later they got refined in 1990. SVMs have their unique way of implementation as compared to other machine learning algorithms. Lately, they are extremely popular because of their ability to handle multiple continuous and categorical variables.

Working of SVM

An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH).

The followings are important concepts in SVM

Support Vectors Datapoints that are closest to the hyperplane is called support

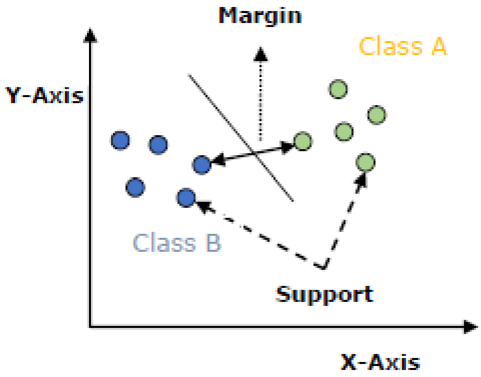


Figure 2.2: Support Vector Machine Classifier

vectors. Separating line will be defined with the help of these data points.

Hyperplane As we can see in the above diagram, it is a decision plane or space which is divided between a set of objects having different classes.

Margin It may be defined as the gap between two lines on the closet data points of different classes. It can be calculated as the perpendicular distance from the line to the support vectors. Large margin is considered as a good margin and small margin is considered as a bad margin.

The main goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH) and it can be done in the following two steps First, SVM will generate hyperplanes iteratively that segregates the classes in best way.

Then, it will choose the hyperplane that separates the classes correctly.

2.4.4 Random Forest

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

Working of Random Forest Algorithm:

We can understand the working of Random Forest algorithm with the help of following steps-

Step 1 First, start with the selection of random samples from a given dataset.

Step 2 Next, this algorithm will construct a decision tree for every sample.

Then it will get the prediction result from every decision tree.

Step 3 In this step, voting will be performed for every predicted result.

Step 4 At last, select the most voted prediction result as the final prediction result.

2.4.5 Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

Before diving into the implementation of logistic regression, we must be aware of the following assumptions about the same :-

- In case of binary logistic regression, the target variables must be binary always and the desired outcome is represented by the factor level 1.

- There should not be any multi-collinearity in the model, which means the independent variables must be independent of each other.

- We must include meaningful variables in our model.

- We should choose a large sample size for logistic regression.

2.4.6 K-Nearest Neighbor

K-nearest neighbors (KNN) algorithm uses 'feature similarity' to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set. We can understand its working with the help of following steps

Step 1

For implementing any algorithm, we need dataset. So during the first step of KNN, we must load the training as well as test data.

Step 2 Next, we need to choose the value of K i.e. the nearest data points. K can be any integer.

Step 3 For each point in the test data do the following

3.1 Calculate the distance between test data and each row of training data with the help of any of the method namely: Euclidean, Manhattan or Hamming distance. The most commonly used method to calculate distance is Euclidean.

3.2 Now, based on the distance value, sort them in ascending order.

3.3 Next, it will choose the top K rows from the sorted array.

3.4 Now, it will assign a class to the test point based on most frequent class of these rows.

Step 4 End.

2.5 Discussion

Heart Disease is a global health concern.How machine learning techniques play a vital role to predict heart diseases are the main topics discussed in this chapter. This chapter provides the basic concepts about the theory of the thesis.

CHAPTER 3

SIMULATION AND METHODOLOGY

In this chapter we discussed about simulation and methodology of our research. Section 3.1 we discussed about data collection; in section 3.2 we discussed about dataset structure and description; in section 3.3 we discussed about data preprocessing methods; in section 3.4 we discussed about simulation tools; in section 3.5 we discussed about methods analysis; in section 3.6 we discussed about cross validation; in section 3.7 we discussed about performance matrices; in section 3.8 we should give a short discussion about these chapter.

3.1 Dataset Collection

The dataset used for this study was taken from UCI machine learning repository known as cleveland dataset.(http://archive.ics.uci.edu/ml/datasets/Heart+Disease).

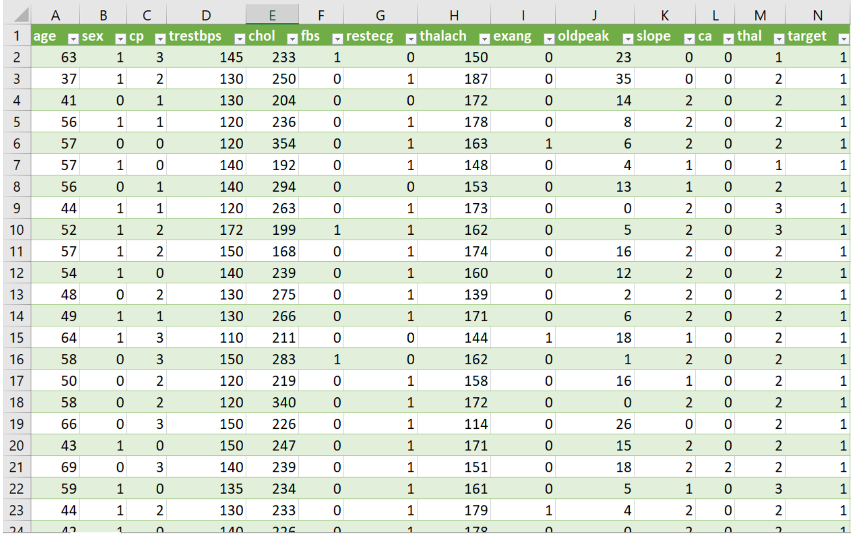
3.2 Dataset structure & description

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[13]. Experiments with the Cleveland database have concentrated on endeavours to distinguish disease presence (values 1, 2, 3, 4) from absence (value 0). There are several missing attribute values, distinguished with symbol '?'. The missing values are handled in the data processing section to prepare for the modeling of the algorithms.

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_cholestoral - serum cholestoral 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

.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
2	63	1	3	145	233	1	0	150	0	23	0	0	1	1
3	37	1	2	130	250	0	1	187	0	35	0	0	2	1
4	41	0	1	130	204	0	0	172	0	14	2	0	2	1
5	56	1	1	120	236	0	1	178	0	8	2	0	2	1
6	57	0	0	120	354	0	1	163	1	6	2	0	2	1
7	57	1	0	140	192	0	1	148	0	4	1	0	1	1
8	56	0	1	140	294	0	0	153	0	13	1	0	2	1
9	44	1	1	120	263	0	1	173	0	0	2	0	3	1
10	52	1	2	172	199	1	1	162	0	5	2	0	3	1
11	57	1	2	150	168	0	1	174	0	16	2	0	2	1
12	54	1	0	140	239	0	1	160	0	12	2	0	2	1
13	48	0	2	130	275	0	1	139	0	2	2	0	2	1
14	49	1	1	130	266	0	1	171	0	6	2	0	2	1
15	64	1	3	110	211	0	0	144	1	18	1	0	2	1
16	58	0	3	150	283	1	0	162	0	1	2	0	2	1
17	50	0	2	120	219	0	1	158	0	16	1	0	2	1
18	58	0	2	120	340	0	1	172	0	0	2	0	2	1
19	66	0	3	150	226	0	1	114	0	26	0	0	2	1
20	43	1	0	150	247	0	1	171	0	15	2	0	2	1
21	69	0	3	140	239	0	1	151	0	18	2	2	2	1
22	59	1	0	135	234	0	1	161	0	5	1	0	3	1
23	44	1	2	130	233	0	1	179	1	4	2	0	2	1
24	42	1	0	140	226	0	1	178	0	0	2	0	2	1

Figure 3.1: Heart Disease dataset

- slope - the slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping)

- no_of vessels - number of major vessels (0-3) colored by flourosopy

- thal(thalassemia) - 3 = normal; 6 = fixed defect; 7 = reversable defect

- diagnosis - the predicted attribute - diagnosis of heart disease (angiographic disease status)

Types of features:

Categorical features (Has two or more categories and each value in that feature can be categorised 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, thal(thalassemia) and 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_cholestoral, max_heart_rate, ST_depression.

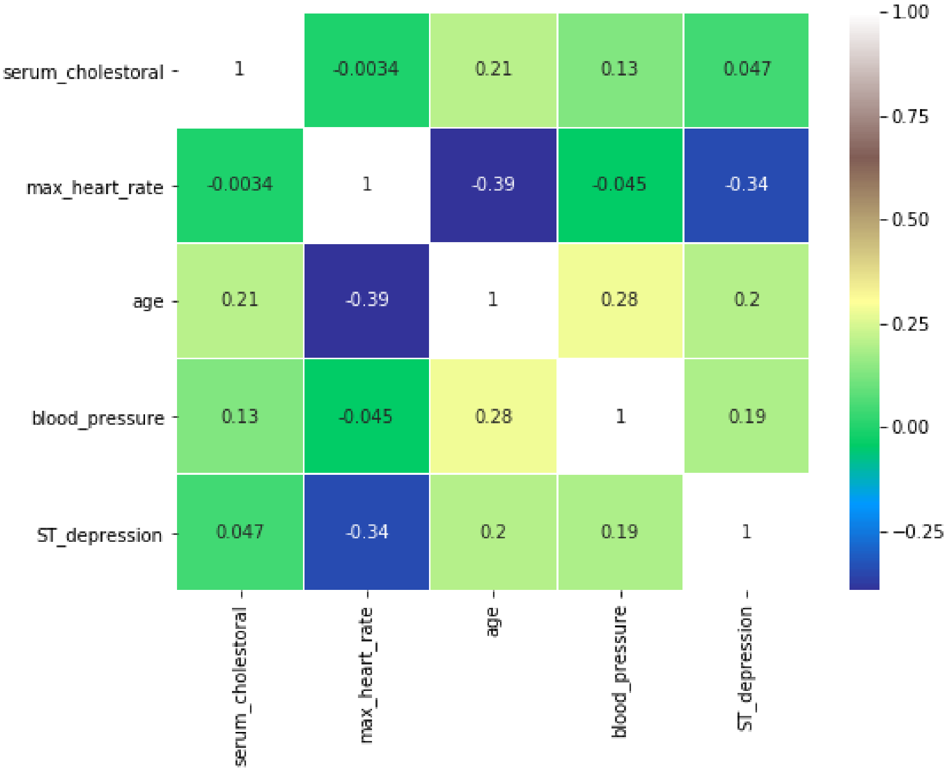


Figure 3.2: Data Correlation Matrix

3.3 Data Preprocessing

In any Machine Learning process, Data Pre-processing is that step in which the data gets transformed, or Encoded, to bring it to such a state that now the machine can easily parse it[8]. In other words, the features of the data can now be easily interpreted by the algorithm.

In this Heart Disease Prediction we need to first preprocess the data in order to compatible with our machine learning algorithms contained in Sci-kit Learn library.We preprocess our data by detecting missing values, detecting outliers and standardized our data to prepare our dataset more clean in order to modeling our machine learning algorithms.

3.3.1 Handling Missing Values

In our dataset, when we see the description of the dataset then we can see that the features **"no_of_vessels"** and **"thal"(thalassemia)** have missing values. We detect these missing values by "pandas" library in python. **"no_of_vessels"** has 4 missing values and **"thal"(thalassemia)** has 2 missing values.

There are many options we could consider when replacing a missing value, for example:

- A constant value that has meaning within the domain, such as 0, distinct from all other values.
- A value from another randomly selected record.
- A mean, median or mode value for the column.
- A value estimated by another predictive model.

We know that the two columns that have missing values, both are categorical feature.So in this case **mode** (most frequently occurring value in a given vector) is usually used for filling 'nans'.

3.3.2 Data Standardization

Standardization is an important technique that is mostly performed as a pre-processing step before many Machine Learning models, to standardize the range of features of input data set.

Standardization comes into picture when features of input data set have large differences between their ranges.

Z-score is one of the most popular methods to standardize data, and can be done by subtracting the mean and dividing by the standard deviation for each value of each feature.[24]

$$z = \frac{value - mean}{standard\ deviation}$$

Standardization scales the data and gives information on how many standard deviations the data is placed from its mean value[24]. Effectively, the mean of the data (µ) is 0 and the standard deviation $(\sigma) is 1$.

In our project we scale the data by the **StandardScaler** library from **sklearn.preprocessing** package using python language.

3.3.3 Data Train-Test Splitting

Now we will split the data for Training phase and Testing phase.

We will split the data into $(7:3) ratio, that means$

70% data are for training the model

and 30% data are for the testing to check the accuracy score.

In order to train our model we need the test data but there is a caution, before training our models we should implement cross validation for our data in order to check the effectiveness of our model.it is also of use in determining the hyper parameters of our model, in the sense that which parameters will result in lowest test error. cross validation reduces the problem of overfitting of data.

3.4 Simulation Tools

To easily perform various Machine Learning methods, we use **Anaconda Navigator**, **Jupyter Notebook** and **Python** version 3.0.

3.4.1 Anaconda Navigator

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows users to launch applications and manage conda packages, environments and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository, install them in an environment, run the packages and update them.

3.4.2 Jupyter Notebook

JupyterLab is a web-based interactive development environment for Jupyter notebooks, code, and data. JupyterLab is flexible: configure and arrange the user interface to support a wide range of workflows in data science, scientific computing, and machine learning. JupyterLab is extensible and modular: write plugins that add new components and integrate with existing ones.

3.4.3 Python

Python is a multi-paradigm programming language: a sort of Swiss Army knife for the coding world and its free, open source software. It supports object-oriented programming, structured programming, and functional programming patterns, among others. There's a joke in the Python community that "Python is generally the second-best language for everything." The creator of python is Guido Van Possum. Data scientists coming from engineering and scientific backgrounds like python because of its inherent readability, simplicity, many numbers of dedicated analytical libraries, extensibility and general purpose nature.

3.5 Methods Analysis

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.

First of all, let's equip ourselves with a handy tool that benefits from the cohesion of SciKit Learn library and formulate a general function for training our models.

In this portion we use classification techniques of Machine Learning, to classify the dataset. From the training and testing data, the classification algorithm is applied[12]. For the classifier algorithm, here we input some training dataset. For finding a result these algorithms perform with some independent test dataset. By input the dataset we get some evaluation results, and, separately, we can deploy the classifier in some real situation to make predictions on fresh data coming from the environment. It's really important in classification that when we're looking at our evaluation results, we only get reliable evaluation results if the test data is different from the training data[12]. For the reason of having only one dataset, we divided into two datasets. We use some of it for training and some of it for testing perhaps two-thirds of it for training and one-third of it for testing. It's really important that the training data is different from the test data. For the better performance sometimes we add more attribute with the dataset. Sometimes it would need to minimize.

We applied our training dataset to these 6 algorithms KNN,Decision Tree, Random Forest, Logistic Regression, Naive Bayes, Support Vector Machine and while doing it we also applied by various parameters to check the Accuracy, Area Under the Curve Score(AUC).

3.5.1 K-Nearest-Neighbor(KNN):

KNN is an instance-based learning algorithm that store all available data points and classifies the new data points based on similarity measure such as distance.

Despite applying the KNN algorithm, the result is very promising. We also see if KNN can perform even better by trying different **'n_neighbours'** inputs. Here in our work, **n_neighbors** range from 1 to 20.And by doing this we find a improved accuracy for 8 no. neighbor out of 20 neighbors that we test.

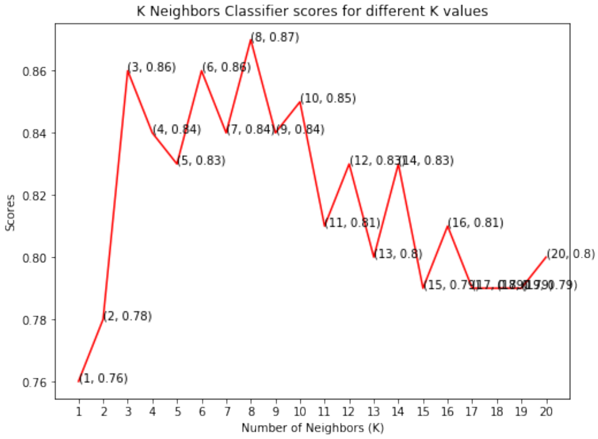


Figure 3.3: K Nearest Neighbor Classifier for different k values

3.5.2 Decision Tree:

Decision Tree decides the target class of a new sample based on selected features from available data using the concept of information entropy. The nodes of the tree are the attributes, each branch of the tree represents a possible decision and the end nodes or leaves are the classes.

In Desicion tree classifier, our data will be overfitting if we do not mention the **maximum depth** of the tree. And mentioning the maximum depth of the tree we can reduce the overfitting problem and find a very good accuracy. Otherwise classifier will randomly build it's depth and causing much more time and error and accuracy will be decresed[7].

Higher value of max_depth is cuasing overfitting of the training data and lower value of max_depth causing underfitting. So define maximum depth of the tree is very important for better accuracy of the Decision tree.So in order to train our model we iterate over 1 to 8 max_depth and find 79.12% accuracy on max_depth = 6.

3.5.3 Support Vector Machine:

SVM discriminates a set of high-dimension features using a or sets of hyperplanes that gives the largest minimum distance to separates all data points among classes.From the Sklearn package we get SVM classifier. By this classifier the accuracy of

SVM model is very good.we used C parameter in order to test our data on

'linear kernel'of SVM.

C is essentially a regularisation parameter, which controls the trade-off between achieving a low error on the training data.Tuning C correctly is a vital step in best practice in the use of SVM.

kernel parameters selects the type of hyperplane used to separate the data. Using 'linear' will use a linear hyperplane (a line in the case of 2D data).

So we trained our Support Vector Machine classifier with our training data and then we test our data.We determined c = 0.05 and kernel = 'linear' and this gave us a better accuracy.

3.5.4 Random Forest:

Random Forest works by constructing multiple decision trees on various sub-samples of the datasets and output the class that appear most often or mean predictions of the decision trees.

In this Random Forest classifier from Sklearn package. The accuracy is quite good, because Random Forest is an ensemble learner, it consist of several decision trees and by voting technique maximum of accuracy of the decision trees we can find the Random Forest classifier accuracy. After we trained the Random Forest model with our training data then at the time of testing we should set a decision tree estimator so we can find more robust accuracy. So for our work we estimate the decision tree is 110. And this give us a better accuracy.

3.5.5 Logistic Regression:

Logistic regression uses an equation as the representation, very much like linear regression.Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modeled is a binary values (0 or 1) rather than a numeric value.Below is an example logistic regression equation:

$$y = e^{(b0+b1*x)}/(1 + e^{(b0+b1*x)}) \tag{3.1}$$

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in our input data has an associated b coefficient (a constant real value) that must be learned from your training data.From our Sklearn.linear.model package we can get Logistic Regression classifier, and for prediction by Logistic Regression we trained this classifier with our training data and then we test our data.

3.5.6 Naive Bayes:

Bayes Rule is: P (Z $| W) = P(W | Z) * P(Z)/P(W))$

In above the Bayes rule determines the probability of Z over given W. Now when it comes to the independent feature we will go for the Naive Bayes algorithm. The algorithm is called naive because we consider W's are independent to one another.

In the case of multiple Z variables, we will assume that Z's are independent.

The Bayes rule will be: P(W=k $| Z) = P(Z | W = k)*P(W = k)/P(Z)KisaclassofW.$

The Naive Bayes will be: P(W=k $| Z1.....Zn) = P(Z1 | W = k) * P(Z2 | W = k)... * P(Zn | W = k) * P(W = k)/P(Z1) * P(Z2).... * P(Zn)$

Which we can say: P (Outcome $| Evidence) = Likelihoodprobabilityofevidence * prior/Evidenceprobability$

The left-hand side of the equation is posterior while the right-hand side there is two numerators:

The first one represents the likelihood of the evidence which is accounted as the conditional probability of each Z given W of a particular class 't'.

From our Sklearn.linear.model package we can get Naive Bayes classifier, and for prediction by Naive Bayes classifier we trained this classifier with our training data and then we test our data.After testing our data we observe a better accuracy.

3.6 Cross Validation

Cross validation is an essential step in model training. It tells us whether our model is at high risk of overfitting.Before feeding the dataset to our models, we should perform cross validation in order to evaluate our models performances.cross validation often called k-fold cross validation. Because in this process the dataset are divided into k subsets.And in these k subsets, k-1 subsets are used for training the model.We continue this process by changing the testing part in each iteration and training the model over the other parts[5].

So, in this research we used 10-fold cross validation.Each time our dataset are divided into 10 subsets and (10-1) = 9 subsets are used for training our model and 1 subset for testing.In this way we can use our each part of the dataset into training and testing phase.Each iteration we found a validation score and finally we take the mean value of the array of validation score to generalized our model and make performance efficient for our training and testing operation. Reason behind we take k = 10 because according to our size of dataset, and as cross validation works iteratively so taking higher k value can slow our process.According to many research, 5 or 10 folds are standard for our data processing and generalized our model.Cross Validation reduced bias. Every data points get to be tested exactly once and is used in training k-1 times. The variance of the resulting estimate is reduced as k increases.

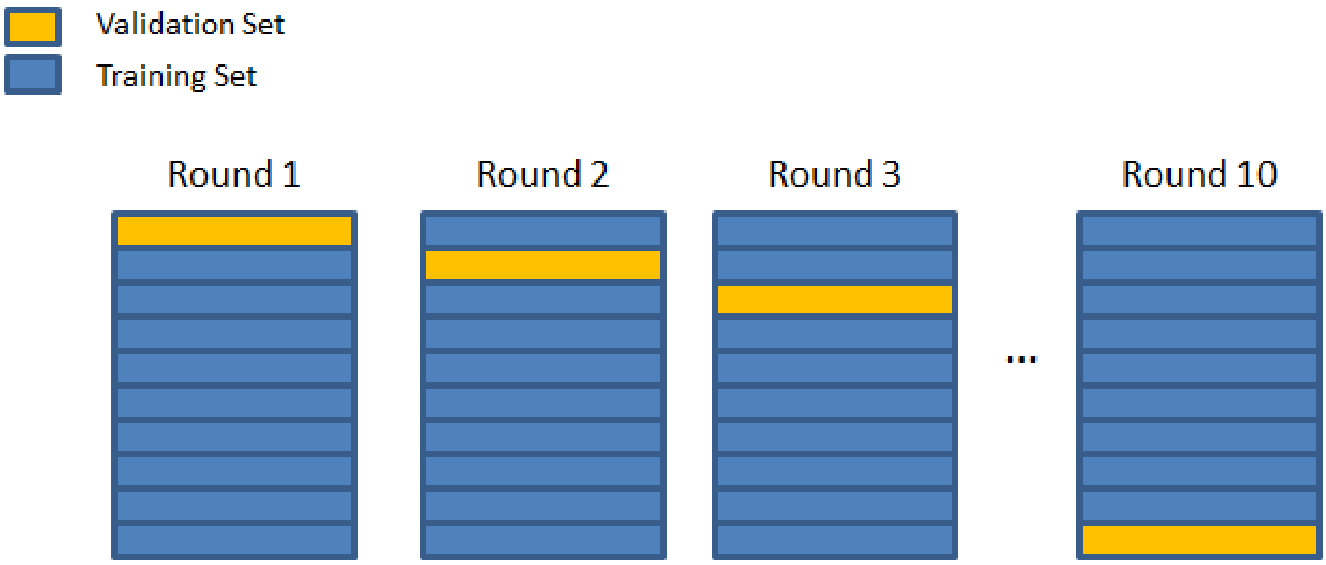


Figure 3.4: 10-fold Cross Validation

3.7 Performance Matrices

When evaluating a machine learning algorithm on a classification problem, we are given a vast amount of performance information to digest. This is because classification may be the most studied type of predictive modelling problem and there are so many different ways to think about the performance of classification algorithms[4].

Classification Accuracy

This the ratio of the number of correct predictions out of all predictions made, often presented as a percentage where 100% is the best an algorithm can achieve[19]. The percentage of correctly classified instances is often called accuracy or sample accuracy. It is the ratio of the number of correct predictions out of all predictions made.

Accuracy = TP+TN/TP+TN+FP+FN

Where, TP (True Positives): TP is the number of examples predicted positive that are actually positive

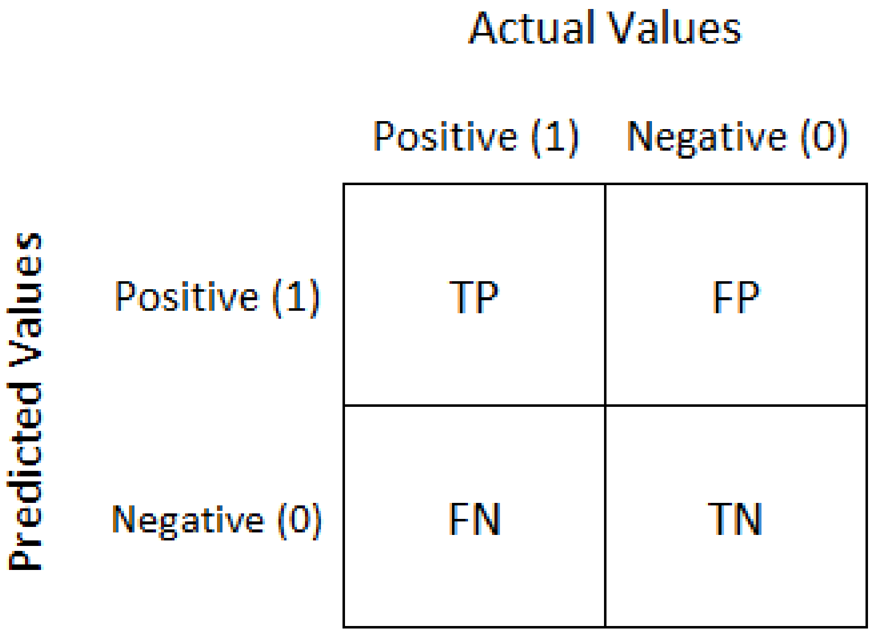
FP (False Positives): FP is the number of examples predicted positive that are actually negative

TN (True Negatives): TN is the number of examples predicted negative that are actually negative

FN (False Negatives): FN is the number of examples predicted negative that are actually positive

Confusion Matrix

A table showing the number of predictions for each class compared to the number of instances that actually belong to each class. This is very useful to get an overview of the types of mistakes the algorithm made. A confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix). The name stems from the fact that it makes it easy to see if the system is confusing two classes.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 3.5: Confusion Matrix

3.8 Discussion

This chapter showed a clear description about the methodology we used for the analysis of our thesis.We have discussed how we prepare the dataset for our models for training and testing phase.We analyzed our methods by the testing data and discussed how their performance can be improved.

CHAPTER 4

RESULT AND DISCUSSION

The most important chapter, mainly focused about all analysis results and discussion. Section 4.1 we discussed about the accuracy of our models; in section 4.2 we discussed about the AUC score of the ROC curve showing ROC curves;in section 4.3 we should give a short discussion about these chapter.

4.1 Models Accuracy Performances

The results have been obtained by applying different classification algorithms. In our first experiment we used the whole dataset with all features and applied Support Vector Machine, Decision Tree, Random Forest, Gaussian Naive Bayes. Table 4.1 contains accuracy of the different algorithms that we applied on our dataset.

Sr. No.	Classifier	Accuracy(%)
1	SVM	87.91%
2	Decision Tree	79.12%
3	Random Forest	87.91%
4	Gaussian Naive Bayes	86.81%
5	Logistic Regression	85.71%
6	KNN	86.81%

Table 4.1: Accuracy of our Algorithms

Now we will plot the Bargraph of the accuracy table 4.1. And from that we can graphically visualize the accuracies of the models. From the table 4.1 we can see that Random Forest, SVM, KNN, Naive Bayes are coparatively best classifiers.

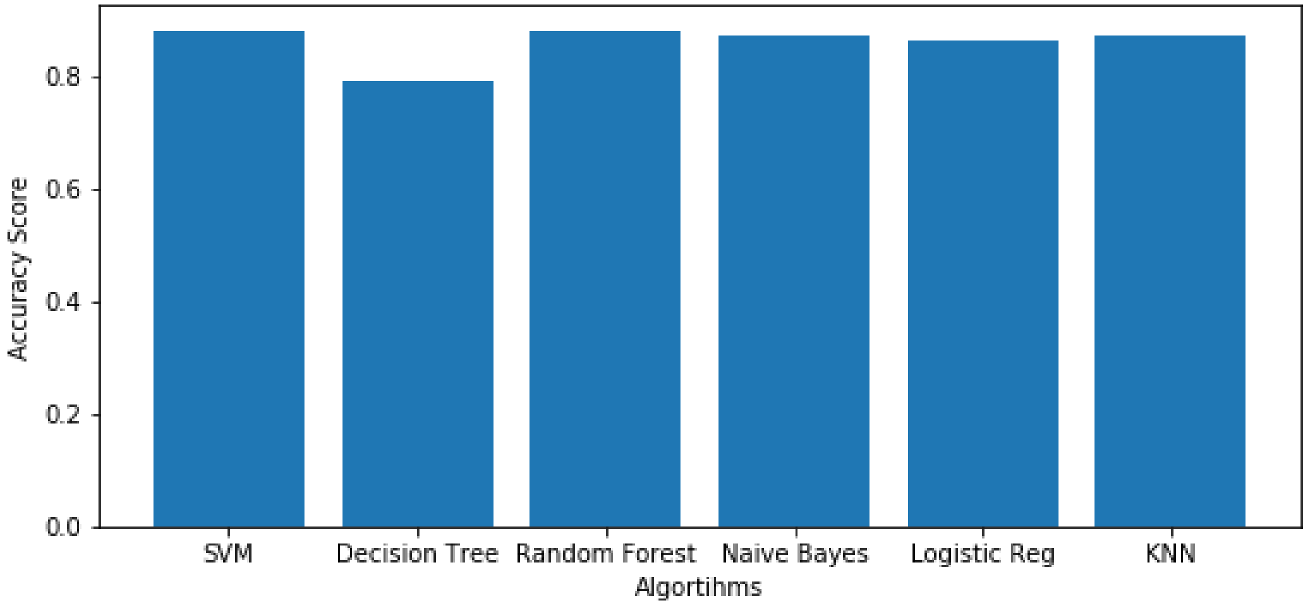


Figure 4.1: Bargraph of Accuracy of the models

From the bargraph of the Accuracy's of the algorithms clearly shows the accuracy level of the algorithms. All the algorithms perform good performance accordingly to our dataset.

4.2 ROC_AUC Score Performance

In this section we will see the AUC(Area Under the Curve) score of the models. AUC score measurement defines that how well a classifier can distinguish the positive and negative class of a predictive problem. Below we will see the tabular representation of AUC scores each of the models.

Sr. No.	Classifier	AUC Score
1	SVM	0.87
2	Decision Tree	0.77
3	Random Forest	0.87
4	Gaussian Naive Bayes	0.86
5	Logistic Regression	0.85
6	KNN	0.86

Table 4.2: AUC Scores of our Algorithms

From table 4.2 we can see that more complex algorithm like Random Forest, SVM comparatively scores very good number.Other algorithm such as Decision tree scores good but comparatively lower than others.Now we will see the graphical representation of AUC Scores.

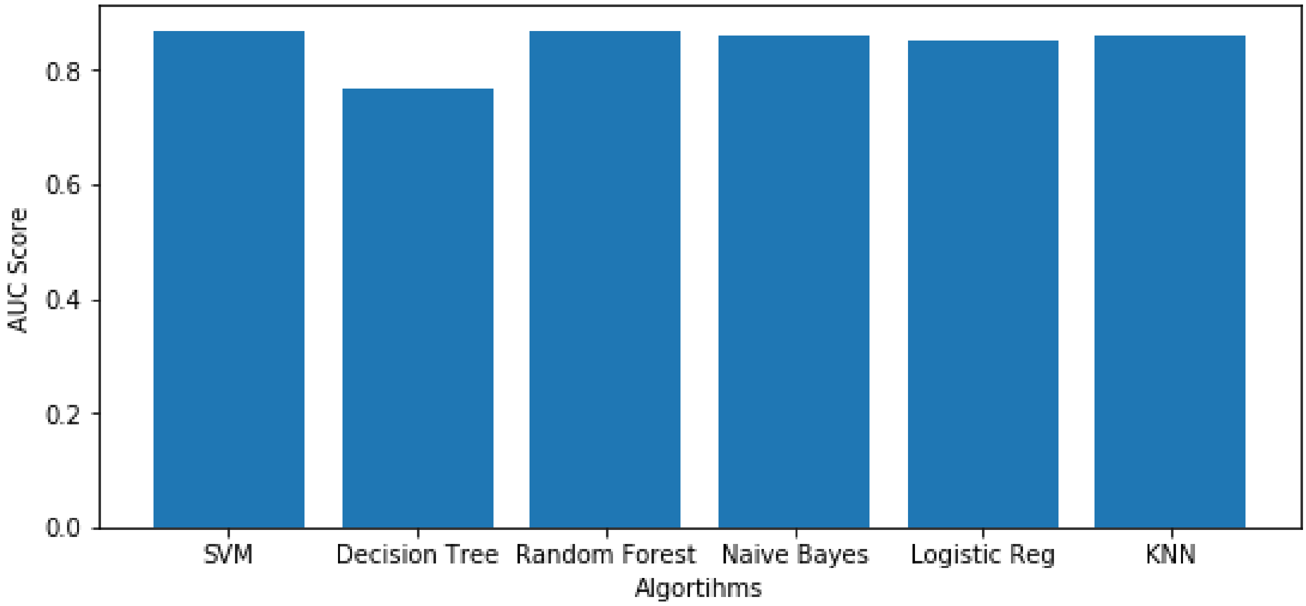


Figure 4.2: Bargraph of AUC Scores of models

4.2.1 Receiver operating characteristic(ROC)Curves

A **Receiver Operating Characteristic** curve, or **ROC** curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

TPR = TP/TP+FN.

FPR = FP/FP+TN.

Here TP = True Positive, FP = False Positive, TN = True Negative and FN = False Negative from the confusion matrix Fig 3.1.

ROC analysis provides tools to select possibly optimal models and to discard sub-optimal ones independently from (and prior to specifying) the cost context or the class distribution[6]. ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making.

Now we will see the graphical representation of ROC curve of our algorithms.

Logistic Regression and Naive Bayes ROC Curve:

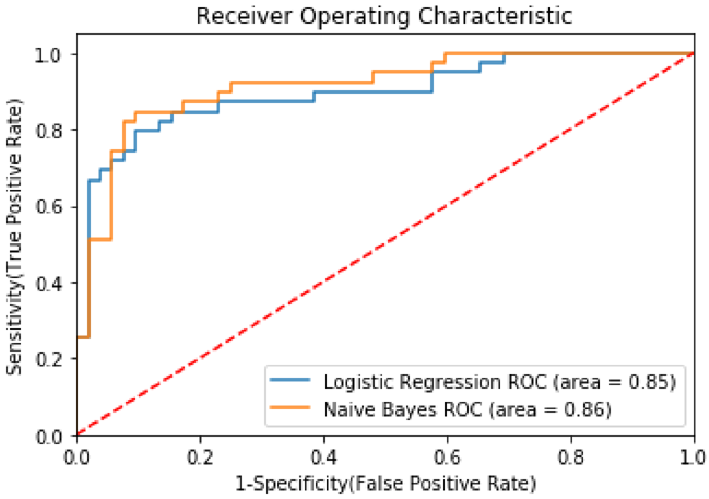


Figure 4.3: Logistic Regression and Naive Bayes ROC Curve

From the figure 4.3 we can say that Naive Bayes performs better than the Logistic Regression to measure the positive class. AUC Score of the Naive Bayes is 0.86 and AUC Score of Logistic Regression is 0.85 that means it measures the proportion of actual positives that are correctly identified as the patient who have the heart disease.

Random Forest and Decision Tree ROC Curve:

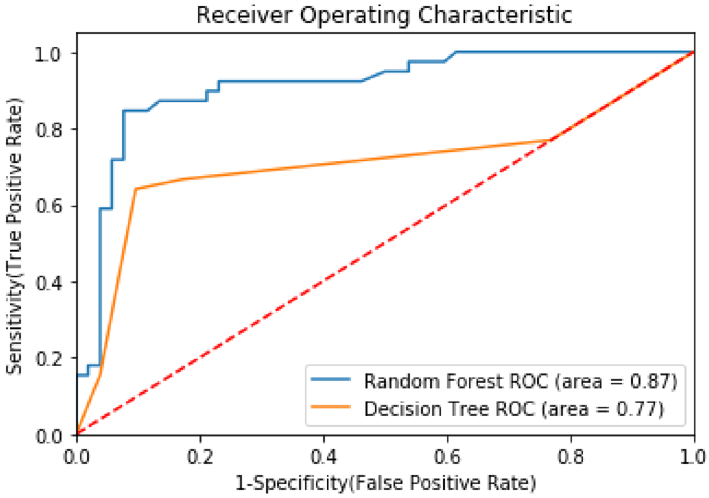


Figure 4.4: Random Forest and Decision Tree ROC Curve

Random Forest classifier is an ensemble learner, that means random forest actually built on several number of decision trees and random forest classifier make prediction by the voting sytem.So every decision tree in the random forest classifier predict a result by subset of our main dataset. this way majority voting result as counted as final prediction[21].So the AUC score is higher than decision tree. Decision tree causes overfitting and it causes bad effect on the accuracy. From figure 4.4 we can say that Random Forest performs better than Decision tree in terms of classify the true positve class that who have the heart disease.

SVM and KNN ROC Curve:

From Fig 4.5 we can say that SVM and KNN are quite similar in terms of predicting the positive class of having disease of a patient.

AUC, or Area Under Curve, is a metric for binary classification. It's probably the second most popular one, after accuracy.Accuracy deals with ones and zeros, meaning we either got the class label right or we didn't. But many classifiers are able to quantify their uncertainty about the answer by outputting a probability value. To compute accuracy from probabilities we need a threshold to decide when zero turns into one. The most

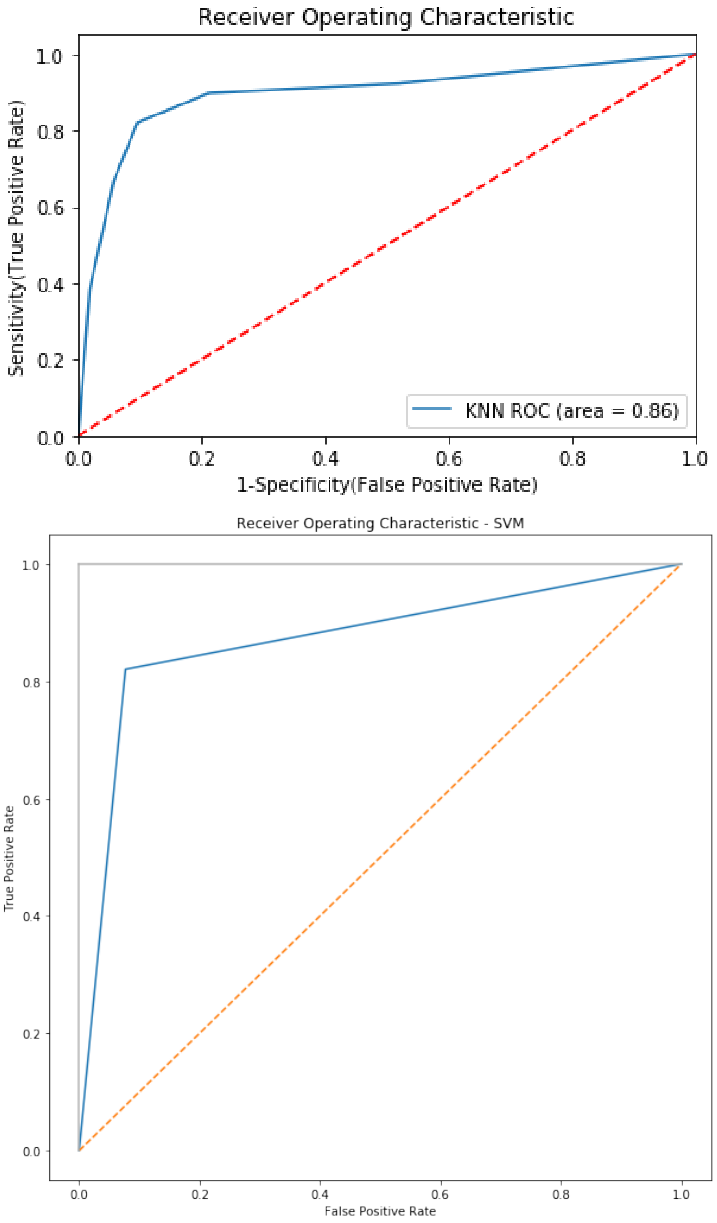


Figure 4.5: SVM and KNN ROC Curve

natural threshold is of course 0.5.

So we can conclude that a ROC curve plots the performance of a binary classifier under various threshold settings. this is measured by true positive rate and false positive rate. If our classifier predicts “true” more often, it will have more true positives (good) but also more false positives (bad). If our classifier is more conservative, predicting “true” less often, it will have fewer false positives but fewer true positives as well. The ROC curve is a graphical representation of this tradeoff.

A perfect classifier has a 100% true positive rate and 0% false positive rate, so its

ROC curve passes through the upper left corner of the square. A completely random classifier (ie: predicting “true” with probability p and “false” with probability 1-p for all inputs) will by random chance correctly classify proportion p of the actual true values and incorrectly classify proportion p of the false values, so its true and false positive rates are both p. Therefore, a completely random classifier’s ROC curve is a straight line through the diagonal of the plot.

4.3 Discussion

In this chapter we mainly showed our various machine learning techniques performance result and analyze their result to decide which machine learning technique is better to predict correctly heart disease patients.We also showed the Area Under the Curve score of each algorithms to show the performances of the models more clearly.After seeing the comparison of the performances of the models we can say that SVM achieved 87.91%,Random Forest achieved 87.91%,KNN achieved 86.81% and Gaussian Naive Bayes achieved 86.81% accuracies which are quite satisfactory results.

CHAPTER 5

CONCLUSIONS AND FUTURE WORKS

We should summarize our work in this chapter. Section 5.1 We should conclude the whole in a summary;in section 5.2 we discussed about the future work opportunities

5.1 Conclusion

In this reasearch we have tried to compare different machine learning algorithms and predict if a certain person, given various personal characteristics and symptoms, will get heart disease or not. The main motive of our research was to comparing the accuracy and analyzing the reasons behind the variation of different algorithms.We have used UCI Machine Learning heart disease dataset, we have considered 13 attributes which is very much related for heart disease and implemented six different algorithms to analyze the accuracy.After the implementation part, we have found Random Forest and Support Vector Machine(SVM) are giving the maximum accuracy level in our dataset which is 87.91% and Decision Tree is performing the lowest level of accuracy which is 77.12%. Moreover, if we increase the attributes, maybe we can found more accurate result but it will take more time to process and the system will be slower than now. As it will be little more complex and will be handling more data's.

5.2 Future Works

We found a satisfactory accuracy from Support Vector Machine and Random Forest classifier from our dataset.In future,by using more efficient methods with more informative features related to heart disease, it is possible to improve the overall accuracy.

Bibliography

[1] https://https://en.wikipedia.org/wiki/cardiovascular$_{d}isease$.

[2] M Elizabeth Brickner, L David Hillis, and Richard A Lange. Congenital heart disease in adults. *New England Journal of Medicine*, 342(5):334–342, 2000.

[3] Austin H Chen, Shu-Yi Huang, Pei-Shan Hong, Chieh-Hao Cheng, and En-Ju Lin. Hdps: Heart disease prediction system. In *2011 Computing in Cardiology*, pages 557–560. IEEE, 2011.

[4] R Chitra and V Seenivasagam. Review of heart disease prediction system using data mining and hybrid intelligent techniques. *ICTACT journal on soft computing*, 3(04):605–09, 2013.

[5] Chaitrali S Dangare and Sulabha S Apte. Improved study of heart disease prediction system using data mining classification techniques. *International Journal of Computer Applications*, 47(10):44–48, 2012.

[6] Colleen Hadigan, James B Meigs, Peter WF Wilson, Ralph B D'agostino, Benjamin Davis, Nesli Basgoz, Paul E Sax, and Steven Grinspoon. Prediction of coronary heart disease risk in hiv-infected patients with fat redistribution. *Clinical Infectious Diseases*, 36(7):909–916, 2003.

[7] Balraj S Heran, Jenny MH Chen, Shah Ebrahim, Tiffany Moxham, Neil Oldridge, Karen Rees, David R Thompson, and Rod S Taylor. Exercise-based cardiac rehabilitation for coronary heart disease. *Cochrane database of systematic reviews*, (7), 2011.

[8] Julien IE Hoffman and Samuel Kaplan. The incidence of congenital heart disease. *Journal of the American college of cardiology*, 39(12):1890–1900, 2002.

[9] MA Jabbar and Shirina Samreen. Heart disease prediction system based on hidden naïve bayes classifier. In *2016 International Conference on Circuits, Controls, Communications and Computing (I4C)*, pages 1–5. IEEE, 2016.

[10] Sun Ha Jee, Yangsoo Jang, Dong Joo Oh, Byung-Hee Oh, Sang Hoon Lee, Seong-Wook Park, Ki-Bae Seung, Yejin Mok, Keum Ji Jung, Heejin Kimm, et al. A coronary heart disease prediction model: the korean heart study. *BMJ open*, 4(5):e005025, 2014.

[11] Bernard Lown and Marshall Wolf. Approaches to sudden death from coronary heart disease. *Circulation*, 44(1):130–142, 1971.

[12] Hlaudi Daniel Masethe and Mosima Anna Masethe. Prediction of heart disease using classification algorithms. In *Proceedings of the world Congress on Engineering and computer Science*, volume 2, pages 22–24, 2014.

[13] Irvine H Page, JN Berrettoni, Antanas Butkus, and F Mason Sones Jr. Prediction of coronary heart disease based on clinical suspicion, age, total cholesterol, and triglyceride. *Circulation*, 42(4):625–645, 1970.

[14] Sellappan Palaniappan and Rafiah Awang. Intelligent heart disease prediction system using data mining techniques. In *2008 IEEE/ACS international conference on computer systems and applications*, pages 108–115. IEEE, 2008.

[15] Atul Kumar Pandey, Prabhat Pandey, KL Jaiswal, and Ashish Kumar Sen. A heart disease prediction model using decision tree. *IOSR Journal of Computer Engineering (IOSR-JCE)*, 12(6):83–86, 2013.

[16] Jaymin Patel, Dr TejalUpadhyay, and Samir Patel. Heart disease prediction using machine learning and data mining technique. *Heart Disease*, 7(1):129–137, 2015.

[17] Shantakumar B Patil and YS Kumaraswamy. Extraction of significant patterns from heart disease warehouses for heart attack prediction. *IJCSNS*, 9(2):228–235, 2009.

[18] Shadab Adam Pattekari and Asma Parveen. Prediction system for heart disease using naïve bayes. *International Journal of Advanced Computer and Mathematical Sciences*, 3(3):290–294, 2012.

[19] G Purusothaman and P Krishnakumari. A survey of data mining techniques on risk prediction: Heart disease. *Indian Journal of Science and Technology*, 8(12):1, 2015.

[20] Bryan L Roth et al. Drugs and valvular heart disease. *N Engl J Med*, 356(1):6–9, 2007.

[21] Kanak Saxena, Richa Sharma, et al. Efficient heart disease prediction system. *Procedia Computer Science*, 85:962–969, 2016.

[22] Jyoti Soni, Ujma Ansari, Dipesh Sharma, and Sunita Soni. Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, 17(8):43–48, 2011.

[23] G Subbalakshmi, K Ramesh, and M Chinna Rao. Decision support in heart disease prediction system using naive bayes. *Indian Journal of Computer Science and Engineering (IJCSE)*, 2(2):170–176, 2011.

[24] K Sudhakar and Dr M Manimekalai. Study of heart disease prediction using data mining. *International journal of advanced research in computer science and software engineering*, 4(1), 2014.

[25] Peter WF Wilson, Ralph B D'Agostino, Daniel Levy, Albert M Belanger, Halit Silbershatz, and William B Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847, 1998.