Mini Project Report on

HEART DISEASE PREDICTION SYSTEM

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

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE & ENGINEERING

Submitted by:

NAINA **2118815**

Department of Computer Science and Engineering
Graphic Era Hill University

Dehradun, Uttarakhand
January 2023

Table of Contents

CHAPTER 1
Introduction

"Data is the new science. Big Data holds the answers." –Pat Gelsinger

Rising healthcare costs have been a major issue for developed nations. (Dadgostar, 2019). According to CDC, an estimated 859,000 people in the US die from cardiovascular disease or 1 in every 3 deaths. Cardiovascular diseases cost $216 billion in the healthcare system and $147 lost in productivity (Mayo, 2022). This cost has been a major concern in the US, and therefore early detection is important. In light of the rapid advancement of biotechnology, and an era of big data generated for healthcare by mainly EHR(electronic health records) in various structures, it is increasingly more important to intelligently use this information to make sense of hidden patterns, detect abnormalities, and predict heart diseases.

Artificial intelligence has certainly made computers smarter. Machine learning which is a subset of artificial intelligence plays an important role in mining large datasets and extracting valuable knowledge from them. Training a machine appropriately with proper train data set, the machine's algorithm can learn patterns and therefore detect any abnormalities in the initial stage of a disease which can help patients save overall cost and time. This project will examine the opportunities of machine learning and data mining in the healthcare industry especially in heart diseases, how early diagnosis can minimize healthcare costs; and how data generated by EHR can provide insights for medical professionals in terms of detecting abnormalities for potential chronic diseases. We begin by providing a brief research background, followed by the problem statement, research questions, objectives, and the organization of this culminating experience project.

Heart Disease

Cardiovascular diseases are the dominant cause of cost and disease burden in the world. (Roth et al., 2020). Cardiovascular diseases refer to any disorder in the heart and blood vessels. Major blood vessels that supply to the heart muscles are affected by a heart condition. These blood vessels build up on cholesterol deposits called plaque reducing blood flow to major parts of the body and the heart (Heart disease, 2022). Over time if left untreated, this can lead to stroke, heart attack, or heart failure. Heart diseases are considered silent killers and at times not diagnosed until life-threatening symptoms start to emerge. Diagnosis of these diseases can include various blood tests, MRIs and CT scans, ECGs, or Holter monitoring. All this medical big data is collected and stored in various databases, which do not provide value on their own, but if integrated and analyzed using Artificial Intelligence, machine learning and data mining techniques it is possible to generate diagnostic information that can lives while minimizing costs.

Machine Learning: In the modern era, humans are experiencing exponential growth of data like never before. With the availability of online data and inexpensive computational computer power, machine learning algorithms can learn and develop models without human intervention (Jordan & Mitchell, 2015). Machine learning, a subset of artificial intelligence, can collect meaningful

knowledge from its training data and automatically improve through exposure without having to be programmed. The machine's algorithm can be classified into four main types, which are Supervised, Unsupervised, Semi-supervised, and Reinforcement Learning (Sarker, 2021). Supervised Learning can be split into two categories: Classification and Regression. Unsupervised learning can be classified into Clustering and Association (Delua, 2021). Both learning approaches are mainly distinguished by using labeled or unlabeled datasets to anticipate the outcome. Each of these has a distinctive set of guidelines when applied to medical data and effectively using it will help extract vital knowledge
(Gupta et al., 2021).

Medical Big Data: Big data can be categorized as any data set that is too extensive, complex, and diverse to be handled by typical desktop software (Camm et al.,2021). As medical information technology advances, so do various forms of medical data. Medical big data is widely used to improve healthcare quality. Such data include audio, lab tests, previous diagnostic reports, clinical records, research, and images (Sun et al., 2019). There are various sources of these data which are stored in different datasets. However, extracting value from a single dataset can be undesirable, but can attain excellent insights by potentially linking various datasets (Lee & Yoon, 2017). A medical data warehouse serves as the centralized repository for the medical data recovered from various data sources such as lab databases, electronic health records (EHR), electronic medical records (EMR), and it has the potential to provide better insights than the analysis of data in a single database.

Problem Statement

Modern information technology tools and techniques such as AI, machine learning and data mining could help support healthcare professionals by providing them with the information they need to make decisions that will minimize deaths caused by heart disease at minimal cost. For example, machine learning algorithms can mine large databases to identify frequent patterns that eventually lead to heart disease and death.

Objectives

The main objective of this project is to explore how Machine Learning algorithms can be used in the diagnosis of heart disease by building an optimized model that can be used to predict heart diseases.

Chapter 2

Literature Survey

Bardhwaj et al., (2017), Shailaja et al., (2018), Sun et al., (2019), and Lee & Yoon, (2017) studied a broad overview of machine learning techniques used in healthcare for various diseases. They provided insights into the potential value of medical big data that can be used for clinical decision support, diagnostics, treatment decisions, fraud detection, and prevention. They briefly summarized the nine-step data mining process along with focusing on why efficient decision support was required by the healthcare system. The results from their experiment showed that machine learning models can be used for the early diagnosis of diseases. Their research is applicable

to this project to an extent; however, their research is less focused on the diagnosis of heart diseases. Therefore, we move forward to review the literature that aligns with our project objective which is how machine learning algorithms can be used in the diagnosis of heart disease. A comprehensive review by Tripoliti et al., (2017) focused on machine learning methodologies evaluating heart failure. They researched severity estimation of heart failure and the prediction of re-hospitalization, mortality, and destabilizations. They performed an extensive study on related works of heart
failure.

A study by J. & S., (2019) used two supervised classifiers called Naïve Bayes Classifier and Decision Tree Classifiers to predict heart diseases on a dataset. Their Decision Tree model predicted the heart disease patients with an accuracy of 91 percent and the Naïve Bayes Classifier had an accuracy of 87 percent. A study by Kamal kant et al.(2014) proposed a model using the Naïve Bayes algorithm to predict heart diseases. The naïve Bayes algorithm is used to assign no dependency between the features. Their study concluded that the Naïve
Bayes algorithm is the most effective for heart disease prediction after that Neural Networks and Decision Trees.

Nidhi Bhatla et al., (2012) used different data mining techniques to predict heart diseases. Their study revealed that the Neural Networks algorithm has performed with higher accuracy than Decision Trees. Their research project included two additional features such as obesity and smoking other than the common attributes.

A review by Rishi Dubey et al., (2015) studied different machine learning algorithms for the prediction of heart disease. Their study concluded that Neural Network is an efficient technique for heart disease prediction. Further adding that this method can also be used to select appropriate treatment.

Ashish Chhabbi et al., (2016) used a dataset collected from UCI repository to perform different data mining techniques to predict heart disease. They applied K-means algorithm and Naïve Bayes and their results revealed that tuning the number of clusters of the k-means algorithm gave better results than the default
K-means.

Boshra Baharami et al., (2015) evaluated various classification methods such as, Decision Tree, K-Nearest Neighbors(k-NN), SMO (used to train Support Vector Machines). On their dataset, they used feature selection techniques to only select the important attributes and achieved the highest accuracy of
83.732% with Decision Trees.

Mrudula Gudadhe et al., (2010) studied heart disease classification using a decision support system. The methods they used were Support Vector Machine (SVM) They incorporated a multilayer perceptron neural network (MLPNN) with three layers in their decision support system and revealing that MLPNN can be used for successfully diagnosing heart disease.

The literature review reveals emerging and advanced machine learning and data mining algorithms involved in predicting heart diseases. It is evident from the above literature review that data mining algorithms have effectively predicted heart diseases. The trustworthiness of the model for predicting heart diseases with

different risk factors is a high concern, however, SVM, Decision Trees, Bagging and Boosting, and RandomForest have achieved reliable results in the diagnosis of heart disease (Jan et al., 2018).

Numerous models using different algorithms have been proposed in the past, producing unique ways to talk about reliability and accuracy for heart disease. In the above literature review, many different data mining prediction models have been introduced such as SVM, Naïve Bayes, Decision Trees, Bagging and Boosting, and RandomForest for heart disease. The models using these algorithms to predict heart disease produced very high accuracy. Therefore, based on these data mining algorithms, we move forward with our research objective in this project to explore these machine learning algorithms and build an optimized model.

Chapter 3
Methodology

We introduce the following Machine Learning algorithms used in predicting heart disease; SVM, KNN, Decision Trees ,and Random Forest, we also list some of the advantages and disadvantages of using these algorithms. Finally, we move forward using RandomForest Classifier algorithm for building our optimized model.

Machine Learning Algorithms for Heart Disease Prediction

Machine learning has been widely employed in a variety of medical prediction datasets, and heart disease prediction is the primary among them. In particular, the medical problem of identifying high-risk patients early on is notably important, as cardiovascular incidents are often fatal; clearly, a timely diagnosis, or even better, preventative care, is a worthy goal.

Decision Trees

One of the most often used forms of supervised learning, a decision tree is a powerful prediction-making/ categorization algorithm that uses previous data to progress from root nodes to decision nodes to leaf nodes. In its most basic form, information is split along branches and ultimately into leaf nodes.

Decision trees form the basis of a more powerful algorithm: random forest

Random Forest:

The fundamental principle that governs random forest prediction is that there is wisdom in crowds: a large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models (Yiu, 2019).

Support Vector Machines:

A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be employed for both classification and regression purposes. Support Vector Machine algorithms have been used effectively to predict cardiovascular disease in several studies. Alty uses a simple digital volume pulse to effectively predict (over 85% accuracy) CVD risk (Alty et al., 2003).

SVM in linear non-separable cases

In the linearly separable case, SVM is trying to find the hyperplane that maximizes the margin, with the condition that both classes are classified correctly. But in reality, datasets are probably never linearly separable, so the condition of 100% correctly classified by a hyperplane will never be met
(Bambrick, 2022).

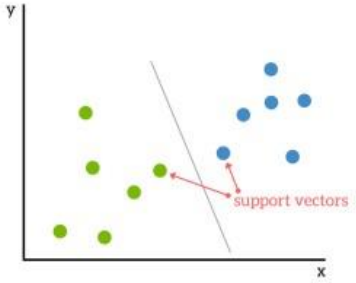

Figure 7. Support Vectors Data Points (Bambrick, 2022)

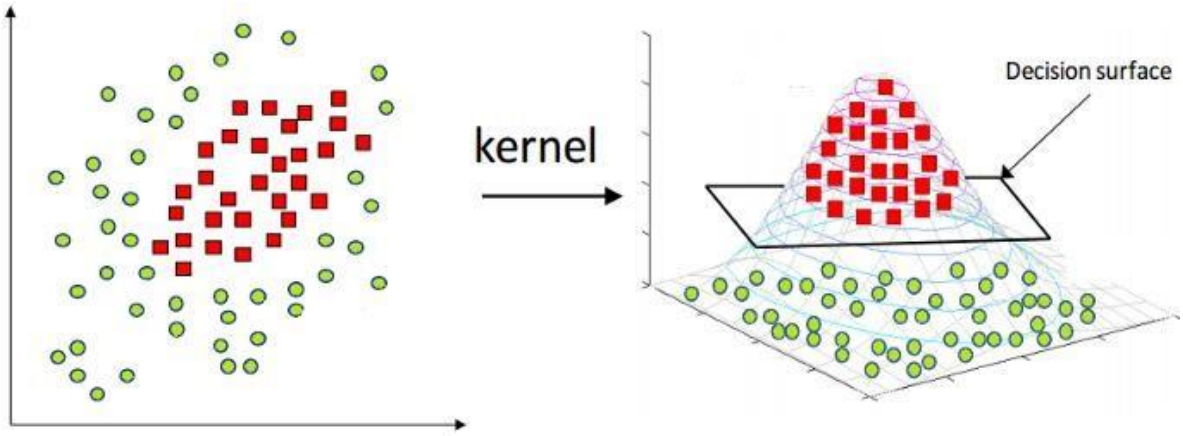

Figure 8. Kernel Trick (Zhang, 2018)

We decided to move forward using RandomForest algorithm to build our optimized model. The RandomForest Classification algorithm provides one of the highest accuracies among all classification methods. To build a model, we collected our data set from the UCI repository, which had 303 patients, with 14 features. We imported the data as a CSV file to PyCharm. With the help of NumPy, Pandas, and Scikit-Learn libraries in Python, we can clean, extract features, split into training and test datasets, and then train the model using the RandomForest algorithm. To optimize the model, we change the hyperparameters, the results of each hyperparameter change will be shown in the results section. Next chapter we will start by introducing our dataset and then a detailed analysis and results .

KNN classifier:

A machine learning algorithm used for classification and regression problems. It works by finding the K nearest points in the training dataset and uses their class to predict the class or value of a new data point.

Chapter 4

DATA COLLECTION AND ANALYSIS

The Cleveland heart disease dataset used to build the machine learning model in this project was collected from the UCI machine learning repository (Latha & Jeeva, 2019). The dataset has 303 instances and 14 attributes; the dataset's description is given in the table below.

S.No	Attribute Name	Description	Range of Values
1	Age	Age of the person in years	29 to 79
2	Sex	Gender of the person [1: Male, 0: Female]	0, 1
3	Cp	Chest pain type [1-Typical Type 1 Angina 2- Atypical Type Angina 3-Non-angina pain 4-Asymptomatic)	1, 2, 3, 4
4	Trestbps	Resting Blood Pressure in mm Hg	94 to 200
5	Chol	Serum cholesterol in mg/dl	126 to 564
6	Fbs	Fasting Blood Sugar in mg/dl	0, 1
7	Restecg	Resting Electrocardiographic Results	0, 1, 2
8	Thalach	Maximum Heart Rate Achieved	71 to 202
9	Exang	Exercise Induced Angina	0, 1
10	OldPeak	ST depression induced by exercise relative to rest	1 to 3
11	Slope	Slope of the Peak Exercise ST segment	1, 2, 3
12	Ca	Number of major vessels colored by fluoroscopy	0 to 3
13	Thal	3 – Normal, 6 – Fixed Defect, 7 – Reversible Defect	3, 6, 7
14	Num	Class Attribute	0 or 1

To summarize a few of the attributes in the Cleveland dataset we can conclude that the dataset included patients from the range 29 to 79 age, the male patients were given the value of 1 and the females were given 0. To indicate any sort of heart disease four types were denoted, Type 1 is Typical Angina which is when blood flow to the heart is reduced resulting in chest pain. (Mayo Clinic, 2022). Type 2 is Atypical Angina, Type 3 is indicated as Non-Angina pain, and
Type 4 was considered Asymptomatic. The fourth feature of the dataset was
Trestbps which is the resting blood pressure measure ranging from 94 to 200. The next attribute is Chol ranging from 126 to 564. Fasting blood sugar (Fbs) was denoted as 1 if the blood sugar is below 120mg/dl and 0 if it was above. Thalach is the maximum heart rate achieved ranging from 71 to 202. Exerciseinduced Angina (exang) was given the value of 0 if there is no pain and 1 if there is pain. The target or the num attribute is denoted as 1 if the patient is diagnosed with heart disease and 0 for normal patients.

Given the dataset, there are several competing algorithms that can be considered immediately. They are considered in no order and where appropriate, their general relative merits and disadvantages will be considered as pertaining to this dataset. We will not perform a full training regime on the data.

Python and Libraries: Python is a high-level programming language that is currently widely used for scientific computing. Its interactive nature and powerful libraries such as Scikit learning, NumPy, Matplolib, and Pandas have positively impacted Data Science. Scikit- learn is a comprehensive and open-sourced machine-learning package that includes a collection of efficient machine-learning methods. (Hao & Ho, 2019). This collection of methods includes data transformation, supervised and unsupervised learning, selection, and model evaluation which are important topics related to machine learning. (Hao & Ho, 2019)

Supervised Learning: Supervised learning is mapping between feature variables and correlating target variables implemented by machine learning algorithms, (Hao & Ho, 2019). One of the main conditions of supervised learning is that both the feature and target variable's labels are known. The labeled datasets are then used to train the machine learning algorithm until it can find patterns between feature and target variables. Once the supervised learning algorithm is finished training on a given dataset to find a pattern and thus build a model, the trained model is then introduced to the testing dataset where labels are intentionally not revealed. The purpose of this is to measure the accuracy the model accomplishes on an unlabeled dataset. In addition, depending on the results the model can be fine-tuned to achieve higher accuracy.

CHAPTER FIVE
DISCUSSION, CONCLUSION, AND AREAS FOR FURTHER STUDY

Discussion
The research questions are:
What Machine learning algorithms are used in the diagnosis of heart disease?
How can Machine Learning techniques be used to minimize misdiagnosis (additional tests, and wrong treatment all resulting in greater monetary impact to the patient)?
How can Machine Learning be used to detect early abnormalities, thus benefiting both patients and the healthcare system?

What follows is the discussion of the findings and conclusion, followed by suggestions for areas for further study.

The findings and conclusion for each question are:

1)Machine learning algorithms used in predicting heart disease are Naïve Bayes, Decision Trees, Support Vector Machine, Bagging and Boosting, and RandomForest, concluding that these algorithms can achieve high accuracy in predicting heart disease.

2) Machine learning algorithms can analyze a large amount of data to assist medical professionals in making more informed decisions cost-effectively. 3) Machine Learning algorithms allowed us to analyze clinical data, draw relationships between diagnostic variables, design the predictive model, and tests it against the new case. The predictive model achieved an accuracy of 89.4 percent using RandomForest Classifier's default setting to predict heart diseases.

Machine learning and data mining techniques are a major turning point in medical diagnosis and this project has shown how important information from medical records can be utilized to diagnose heart disease patients. The project's objective to explore how machine learning algorithms can be used in the diagnosis of heart disease has been achieved by identifying 5 different algorithms covered in Chapter 3, additionally developing an optimized model with one of them. Finally, the model we built to predict heart disease can save enormous medical bills, improve diagnosis capability on large scale, and most importantly save lives.

Conclusion
Heart disease is a life-threatening disease affecting millions of people around the world every year (Asadi et al., 2021). Hence, early prediction of heart disease can benefit patients and healthcare professionals by providing the information they need to minimize death and reduce costs. Since medical big data has been increasing daily and data storage costs decreasing, machine learning algorithms can play an important part in processing these medical data and predicting diseases.

With the help of the RanfomForest Classifier algorithm, we were able to build a machine-learning model. Our model was trained and tested by a dataset from the UCI repository. The dataset consisted of labeled 303 patients, it included both diagnosed heart disease patients and normal patients. After the model was trained

and then tested, we achieved an accuracy of 89.4% with the default hyperparameter. While we tried to tune RandomForest Classifier's hyperparameter; N_estimator in the hope of higher accuracy, we noticed that the default resulted in the highest.

We can conclude that machine learning and data mining can play an important role in our healthcare system. Traditionally, diagnosis of the disease was performed by standard procedures and doctor's intuitions which had limitations and led to costly expenses, but with machine learning models, diagnosis can be done on large datasets cost-effectively.

Areas for Further Study

As we have developed a supervised machine-learning model using the Random Forest algorithm and tuning one of its hyperparameters called 'N_Estimator', in the future this model can be trained and tested using a larger set of data with additional attributes. Additionally, our model holds an opportunity for further research to be performed by modifying different hyperparameters.