Cardiovascular Disease Risk Prediction

A PROJECT REPORT

Submitted by

Khyaat Punetha	21BCE10086
Atharva Jaurkar	21BCE10300
Ujesh Mishra	21BCE10183
Swastik Yadav	21BCE10088
Sumit Jha	21BCE11656

in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

VIT BHOPAL UNIVERSITY

KOTHRIKALAN, SEHORE

MADHYA PRADESH - 466114

Sept 2022

VIT BHOPAL UNIVERSITY,KOTHRIKALAN, SEHORE MADHYA PRADESH – 466114

BONAFIDE CERTIFICATE

Certified that this project report titled **"Cardiovascular Disease Risk Prediction"** is the bonafide work of "**Khyaat Punetha (21BCE10086), Atharva Jaurkar (21BCE10300), Ujesh Mishra (21BCE10183), Swastik Yada (21BCE10088), Sumit Jha (21BCE11656)"** who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported here does not form part of any other project / research work on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

PROGRAM CHAIR	**PROJECT GUIDE**
Dr. Preetam Suman, Assistant Professor (Senior)	Dr. Gopal Singh Tandel, Assistant Professor
School of Computer Science and Engineering	School of Computer Science and Engineering
VIT BHOPAL UNIVERSITY	VIT BHOPAL UNIVERSITY

The Project Exhibition I Examination is held on 30 September 2022

Acknowledgement

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose Constant guidance and encouragement crown all efforts with success. I am very grateful to my project supervisor Dr. Gopal Singh Tandel, for the guidance, inspiration and constructive suggestions that helpful me in the preparation of this project. I won't forget to also mention my course mates; for their wonderful and skillful guidance in assisting me with the necessary support to ensure that my project is a success. I also thank my parents and family at large for their moral support for the project to ensure successful completion of the project.

ABSTRACT

Cardiovascular disease is one of the most heinous diseases, especially the silent heart attack, which attacks a person so abruptly that there's no time to get it treated and such disease is very difficult to be diagnosed. Various medical data mining and machine learning techniques are being implemented to extract the valuable information regarding the heart disease prediction. Yet, the accuracy of the desired results are not satisfactory. This Model proposes a heart disease prediction system using Machine learning techniques. Health care field has a vast amount of data, for processing those data certain techniques are used. Data Mining Is one of the techniques often used. Heart diseases are the Leading cause of death worldwide. This System predicts the arising possibilities of Heart Disease. The datasets used are classified in terms of medical parameters. This system evaluates those parameters using data mining classification technique. The datasets are processed in python programming using five main Machine Learning Algorithms namely Decision tree Algorithm and Naive Bayes Algorithm, Linear Regression, Knn Algorithm, Artificial Neural Networking which shows the best algorithm among these two in terms of accuracy level of heart disease.

LIST OF SYMBOLS

TP:Number of people with heart diseases.

TN:Number of people with heart diseases and no heart diseases.

FP:Number of people with no heart diseases.

FN:Number of people with no heart diseases and with heart diseases.

Table Of Contents

CHAPTER 1

INTRODUCTION

According to the World Health Organisation, every year 12 million deaths occur worldwide due to 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 data analysis. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in an attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduces the complications.

Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. This project aims to predict future Heart Disease by analysing data of patients which classifies whether they have heart disease or not using a machine-learning algorithm. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analysing to extract the desired data we can say that this technique can be very well adapted to do the prediction of heart disease

1.1 MOTIVATION FOR THE WORK

The main motivation of doing this research is to present a heart disease prediction model for the prediction of occurrence of heart disease. Further, this research work is aimed towards identifying the best classification algorithm for identifying the possibility of heart disease in a patient. This work is justified by performing a comparative study and analysis using three classification algorithms namely Naïve Bayes, Decision Tree, and Random Forest are used at different levels of evaluations. Although these are commonly used machine learning algorithms, the heart disease prediction is a vital task involving highest possible accuracy. Hence, the three algorithms are evaluated at numerous levels and types of evaluation strategies. This will provide researchers and medical practitioners to establish a better

1.2 PROBLEM STATEMENT

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either they are expensive or are not efficient to calculate the chance of heart disease in humans. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients everyday in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more patience, time and expertise. Since we have a good amount of data in today's world, we can use various machine learning algorithms to analyse the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

CHAPTER 2

LITERATURE SURVEY

2.1 Existing System

With growing development in the field of medical science alongside machine learning various experiments and research has been carried out in recent years releasing the relevant significant papers.

[1] Dr. Purushottam proposed a paper “Efficient Heart Disease Prediction System” using hill climbing and decision tree algorithms .They used Cleveland dataset and preprocessing of data is performed before using classification algorithms. The Knowledge Extraction is done based on Evolutionary Learning (KEEL), an open source data mining tool that fills the missing values in the data set.A decision tree follows top-down order. For each actual node selected by hill-climbing algorithm a node is selected by a test at each level. The parameters and their values used are confidence. Its minimum confidence value is 0.25. The accuracy of the system is about 86.7%.

[2] Santhana Krishnan. J ,et al proposed a paper “Prediction of Heart Disease Using Machine Learning Algorithms” using den tree and Naive Bayes algorithm for prediction of heart disease. In the decision tree algorithm the tree is built using certain conditions which give True or False decisions. The algorithms like SVM, KNN are results based on vertical or horizontal split conditions depending on dependent variables. But a decision tree for a tree-like structure having root node, leaves and branches based on the decision made in each of tree Decision tree also helps in the understanding of the importance of the attributes in the dataset. They have also used the Cleveland data set. Dataset splits in 70% training and 30% testing by using some methods. This algorithm gives 91% accuracy. The second algorithm is Naive Bayes, which is used for classification. It can handle complicated, nonlinear, dependent data so it is found suitable for heart disease dataset as this dataset is also complicated, dependent and nonlinear in nature. This algorithm gives an 87% accuracy.

[3] Sonam Nikhar et al proposed paper " Prediction of Heart Disease Using Machine Learning Algorithms' ' their research gives point to point explanation of Naïve Bayes and decision tree classifiers that are used especially in the prediction of Heart Disease. 3 Some analysis has been led to think about the execution of prescient data mining strategy on the same dataset, and the result decided that Decision Tree has higher accuracy than Bayesian classifier.

[4] Aditi Gavhane et al proposed a paper "Prediction of Heart Disease Using Machine Learning", in which training and testing of dataset is performed by using neural network algorithm multi-layer perceptron. In this algorithm there will be one input layer and one output layer and one or more layers are hidden between these two input and output layers. Through hidden layers each input node is connected to the output layer. This connection is assigned with some random weights. The other input is called bias which is assigned with weight based on requirement the connection between the nodes can be feedforward or feedback.

[5] Avinash Golande et al, proposed "Heart Disease Prediction Using Effective Machine Learning Techniques" in which few data mining techniques are used that support the doctors to differentiate the heart disease. Usually utilised methodologies are k-nearest neighbour, Decision tree and Naïve Bayes. Other unique characterization-based strategies utilised are packing calculation, Part thickness, consecutive negligible streamlining and neural systems, straight Kernel selfarranging guide and SVM (Bolster Vector Machine).

CHAPTER 3

METHODOLOGY

3.1 EXISTING SYSTEM

Heart disease is even being highlighted as a silent killer which leads to the death of a person without obvious symptoms. The nature of the disease is the cause of growing anxiety about the disease & its consequences. Hence continued efforts are being done to predict the possibility of this deadly disease in advance. So that various tools & techniques are regularly being experimented with to suit the present-day health needs. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analysing to extract the desired data we can conclude. This technique can be very well adapted to the prediction of heart disease.

3.2 PROPOSED SYSTEM

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is preprocessed into the required format. The data is then divided into two parts: training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data. This system is implemented using the following modules.

1.) Collection of Dataset

2.) Selection of attributes

3.) Data Preprocessing

4.) Balancing of Data

5.) Disease Prediction

3.2.1 Collection of dataset

Initially, we collected a dataset for our heart disease prediction system. After the collection of the dataset, we split the dataset into training data and testing data. The training dataset is used for prediction model learning and testing data is used for evaluating the prediction model. For this project, 70% of training data is used and 30% of data is used for testing. The dataset used for this project is Heart Disease UCI. The dataset consists of 76 attributes; out of which, 14 attributes are used for the system.

3.2.2 Selection of attributes

Attribute or Feature selection includes the selection of appropriate attributes for the prediction system. This is used to increase the efficiency of the system. Various attributes of the patient like gender, chest pain type, fasting blood pressure, serum cholesterol, exang, etc are selected for the prediction. The Correlation matrix is used for attribute selection for this model

3.2.3 Pre-processing of Data

Data preprocessing is an important step for the creation of a machine learning model. Initially, data may not be clean or in the required format for the model which can cause misleading outcomes. In pre-processing of data, we transform data into our required format. It is used to deal with noises, duplicates, and missing values of the dataset. Data pre-processing has the activities like importing datasets, splitting datasets, attribute scaling, etc. Preprocessing of data is required for improving the accuracy of the model.

3.3 Comparisons from Literature Survey

Serial Number	**Model Used**	**Existing Methods**	**Our Models**
		Accuracy(%)	
1	Linear Regression Model	89.13	76.92
2	Multinomial naive bayes model	81.48	73.77
3	Decision Tree	79	78.02
4	K- Nearest Neighbour	73.77	82.41
5	Artificial Neural Network	84	83.51

CHAPTER 4

WORKING OF SYSTEM

4.1 MACHINE LEARNING

In machine learning, classification refers to a predictive modelling problem where a class label is predicted for a given example of input data.

- Supervised Learning

  Supervised learning is the type of machine learning in which machines are trained using well "labelled" training data, and on the basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output. In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher. Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y)

- Unsupervised learning

  Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data. The goal of unsupervised learning is to find the underlying structure of the dataset, group that data according to similarities, and represent that dataset in a compressed format. Unsupervised learning is helpful for finding useful insights from the data. Unsupervised learning is much similar to how a human learns to think by their own experiences, which makes it closer to the real AI. Unsupervised learning works on unlabeled and uncategorized data which make unsupervised learning more important. In real-world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.

- Reinforcement learning

Reinforcement learning is an area of Machine Learning. It is about taking suitable action to maximise reward in a particular situation. It is employed by various software and machines to find the best possible behaviour or path it should take in a specific situation. Reinforcement learning differs from supervised learning in a way that in supervised learning the training data has the answer key with it so the model is 11 trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of a training dataset, it is bound to learn from its experience.

4.2 Algorithms

We have used 5 main models for predicting our result. All are discussed below

4.2.1 Logistic Regression

Logistic regression is a process of modelling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on. Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best into a category. As aspects of cyber security are classification problems, such as attack detection, spam detection, fraud detection, emergency detection, logistic regression is a useful analytic technique.

4.2.2 Multinomial Naive Bayes

The Multinomial Naive Bayes algorithm is a Bayesian learning approach popular in Natural Language Processing (NLP). The program guesses the tag of a text, such as an email or a newspaper story, using the Bayes theorem. It calculates each tag's likelihood for a given sample and outputs the tag with the greatest chance. The Naive Bayes classifier is made up of a number of algorithms that all have one thing in common: each feature being classed is unrelated to any other feature. A feature's existence or absence has no bearing on the inclusion or exclusion of another feature.

4.2.3 Decision Tree

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute

Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.

4.2.4 K-Nearest Neighbour

The KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

4.2.5 Artificial Neural Network

A brain's neural network is made up of around 86 billion neurons. The neurons are connected by what are called synapses. There are about 100 trillion synapses in the human brain. The neurons send signals to each other through the synapses.

To create a neural network we combine artificial neurons together so that the outputs of some neurons are inputs of other neurons. We will be working with feed forward neural networks which means that the neurons only send signals in one direction. In particular, we will be working with what is called a Multi-Layer Perceptron (MLP).

CHAPTER 5

Performance Analysis

5.1 Performance Measures

In this project, various machine learning algorithms like Logistic Regression, Naive Bayes, Decision Tree,K-Nearest Neighbour, Artificial Neural Network are used to predict heart disease. Heart Disease UCI dataset, has a total of 76 attributes, out of those only 14 attributes are considered for the prediction of heart disease. Various attributes of the patient like gender, chest pain type, fasting blood pressure, serum cholesterol, exang, etc are considered for this project. The accuracy for individual algorithms has to be measured and whichever algorithm is giving the best accuracy,that is considered for the heart disease prediction. For evaluating the experiment, various evaluation metrics like accuracy, confusion matrix, precision, recall, and f1-score are considered.

Accuracy- Accuracy is the ratio of the number of correct predictions to the total

number of inputs in the dataset. It is expressed as:

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

Confusion Matrix- It gives us a matrix as output and gives the total performance of the system.

Where

TP: True positive

FP: False Positive

FN: False Negative

TN: True Negative

5.2 Performance Analysis

Accuracy of Logistic Regression: 76.92%

Accuracy of Naive Bayes: 73.77%

Accuracy of Decision Tree: 78.02%

Accuracy of K-Nearest Neighbour: 82.41%

Accuracy of Artificial Neural Network: 83.51

The highest accuracy is given by the **Artificial Neural Network.**

5.3 Summary

After performing the machine learning approach for training and testing we find that accuracy of the Artificial Neural Network is better compared to other algorithms. Accuracy is calculated with the support of the confusion matrix of each algorithm, here the number count of TP, TN, FP, FN is given and using the equation of accuracy, value has been calculated and it is concluded that extreme gradient boosting is best with 83.51% accuracy.

Chapter 6

FUTURE ENHANCEMENT AND CONCLUSION

6.1 Future Enhancements

Further for its enhancement, we can train on models and predict the types of cardiovascular diseases providing recommendations to the users, and also use more enhanced models.

6.2 Conclusion

Heart diseases are a major killer in India and throughout the world. Application of promising technology like machine learning to the initial prediction of heart diseases will have a profound impact on society. The early prognosis of heart disease can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. The number of people facing heart diseases is on a raise each year. This prompts for its early diagnosis and treatment. The utilisation of suitable technology support in this regard can prove to be highly beneficial to the medical fraternity and patients. In this paper, the five different machine learning algorithms used to measure the performance are Logistic Regression Naïve Bayes, Decision Tree, K-Nearest Neighbour and Artificial Neural Network applied on the dataset

All the five machine learning methods are compared based on which one prediction model is generated. Hence, the aim is to use various evaluation metrics like confusion matrix, accuracy, precision, recall, and f1-score which predicts the disease efficiently. Comparing all seven the extreme gradient boosting classifier gives the highest accuracy of 83.51%.

Appendix

Python

Python is an interpreted, high-level, general purpose programming language created by Guido Van Rossum and first released in 1991, Python's design philosophy emphasises code Readability with its notable use of significant White space. Its language constructs and object oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically typed and garbage collected. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming.

Sklearn

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modelling including classification, regression, clustering and dimensionality reduction via a consistent interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

Numpy

NumPy is a library for the python programming language, adding support for large, multi- dimensional arrays and matrices, along with a large collection of high level mathematical functions to operate on these arrays. The ancestor of NumPy, Numeric, was originally created by Jim with contributions from several other developers. In 2005, Travis created NumPy by incorporating features of the competing Numarray into Numeric, with extensive modifications. NumPy is open source software and has many contributors.

Matplotlib

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged.

REFERENCES

[1] Pal M, Parija S, Panda G, Dhama K, Mohapatra RK. Risk prediction of cardiovascular disease using machine learning classifiers. Open Med (Wars). 2022 Jun 17;17(1):1100-1113. doi: 10.1515/med-2022-0508. PMID: 35799599; PMCID: PMC9206502.

[2] A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of Heart Disease Using Machine Learning, " 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1275-1278, doi: 10.1109/ICECA.2018.8474922.

[3] Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked, 16, 100203. https://doi.org/10.1016/j.imu.2019.100203

[4] Qian, X., Li, Y., Zhang, X., Guo, H., He, J., Wang, X., Yan, Y., Ma, J., Ma, R., & Guo, S. (2022). A Cardiovascular Disease Prediction Model Based on Routine Physical Examination Indicators Using Machine Learning Methods: A Cohort Study. Frontiers in Cardiovascular Medicine. https://doi.org/10.3389/fcvm.2022.854287

[5] Dimopoulos, A.C., Nikolaidou, M., Caballero, F.F. et al. Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. BMC Med Res Methodol 18, 179 (2018). https://doi.org/10.1186/s12874-018-0644-1

[6] M. Dhilsath Fathima, S. Justin Samuel, R. Natchadalingam, V. Vijeya Kaveri. (2022) Majority voting ensemble feature selection and customised deep neural network for the enhanced clinical decision support system. International Journal of Computers and Applications 0:0, pages 1-11.