IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.13315

Heart Disease Prediction System Using Machine Learning

Dr. Umesh Akare[1], Prof. Umme Ayeman Gani[2], Anushri Bhongade[3], Dhanashree Mure[4], Madhulika Chatterjee[5], Vanzuli Ramteke[6]

Assistant Professor, Department of Artificial Intelligence and Data Science Priyadrashini College of Engineering Hingna, Nagpur-19[1,2]

Student, Department of Artificial Intelligence and Data Science Priyadrashini College of Engineering, Hingna, Nagpur-19[3-6]

Abstract: Around the world, machine learning is utilized in a wide range of areas. The medical field is not an exception. Predicting whether there is a risk of heart diseases, issues with the loco motor system, and numerous other conditions can be significantly assisted by machine learning. Like that so information is predicted well in advance, it may offer physicians valuable insights that allow them to customize their diagnosis and treatment for each patient. We use machine learning techniques and methods for early prediction heart diseases in humans. In this project, we used machine learning techniques Logistic Regression & Decision Trees. We also suggest performing hybrid classification, as it can have numerous samples for both training and verifying the data.

Keywords: Machine learning, supervised learning, logistic regression, decision tree, python programming.

I. INTRODUCTION

The leading cause of disease is cardiovascular disease. So it is necessary that this type of disease is detected in an earlier stage. For this reason, deep research is required in field. Today, a large number of patients has been suffering from cardiovascular diseases and because of lack of instrumentation disease was identified too late, so we used machine learning to predict heart disease in an earlier stage so that we can prevent heart disease and reduce mortality rate. Machine learning is one of the emerging technologies that helps to predict disease in the earlier stage.

Machine learning is the subcategory of Artificial Intelligence (AI). AI is the branch of analyze and creating a model that behaves and think like a human. In machine learning we will train is the model called supervised learning. And supervised learning helps to predict the accurate percentage of disease that could occur in the future. We will use the data preprocessing process to train the data to predict the heart disease.

In our project, we are using medical specifications like blood pressure, cholesterol, sex, age, fasting sugar and etc. As the report of this parameter, we are predicting the accuracy of algorithms. we have used two algorithms logistic regression and decision tree algorithms.

In our research, we are computing two different machine learning techniques and determine which is the best based on the computation. In our study, we are predicting whether a person has heart disease or not using machine learning. We observe a number of patient characteristics, including cholesterol, fasting sugar, resting sugar, blood pressure, maxheartrate. We are using machine learning algorithms, that are Decision Tree, Logistic Regression. We are comparing the accuracy of the algorithms based on our selected attributes, and the algorithm that predicts heart disease providing result of highest accuracy.

II. LITERATURE REVIEW

[1] Chintan M. Bhatt, Parth Patel et al. (2023) proposed the system with the accuracies of all algorithms which were used in the paper were above 86% with the lowest accuracy of 86.37% given by decision trees and the highest accuracy given by multilayer perceptron. [1]

[2] Madhumita Pal, Smita Parija et al. (2020) developed and execute the model for predicting heart disease and obtained the 86.9 percent of accuracy and recognize rate of disease using random forest. [2]

DOI: 10.17148/IJARCCE.2024.13315

[3] Toukir Ahmed et al. (2019) developed the model by comparing two different algorithms that are Multilayer perceptron and Support vector machine have the best accuracy compared to algorithms to improve the accuracy of the model. [3]

[4] Ambrish. G, Bharathi Ganesh et al,(2023) developed the model using logistic regression and splitting the ratio by 90% training and 10% testing. The highest accuracy is 87.10% using logistic regression. [4]

[5] Nayab Akhtar et al. (2021) proposed the system for health care maintenance for detecting cardio disease. Heart disease is one of the leading cause of increase in mortality rate. They have used some algorithms like random forest, neural networks, decision tree and naive baye. By using this algorithms model provides 87% of accuracy. [5]

[6] In the year 2020, Keshav Srivastava et al. (2020) developed the model using some machine learning algorithm providing accuracy of 87%, support vector Machines gives an accuracy of 83%, and Random Forest gives an accuracy of 84%. [6]

[7] Archana Singh and Rakesh Kumar et al. (2020) prediction of the algorithms depends upon the dataset and dataset is divided into two parts training and testing. They analyze algorithms using confusion matrix. They had mentioned that KNN algorithm gives highest accuracy compared to other algorithms [7]

[8] Harshit Jindal et al. (2020) proposed the model using various algorithms such as SVM, decision tree for diagnosis of patients fatal heart disease, KNN, compared to random forest classifier and logistic regression yields best result. The maximum amount of accuracy is obtained by using KNN and logistic regression which provides with the maximum amount of accuracy that is 88.5%. [8]

III. METHODOLOGY

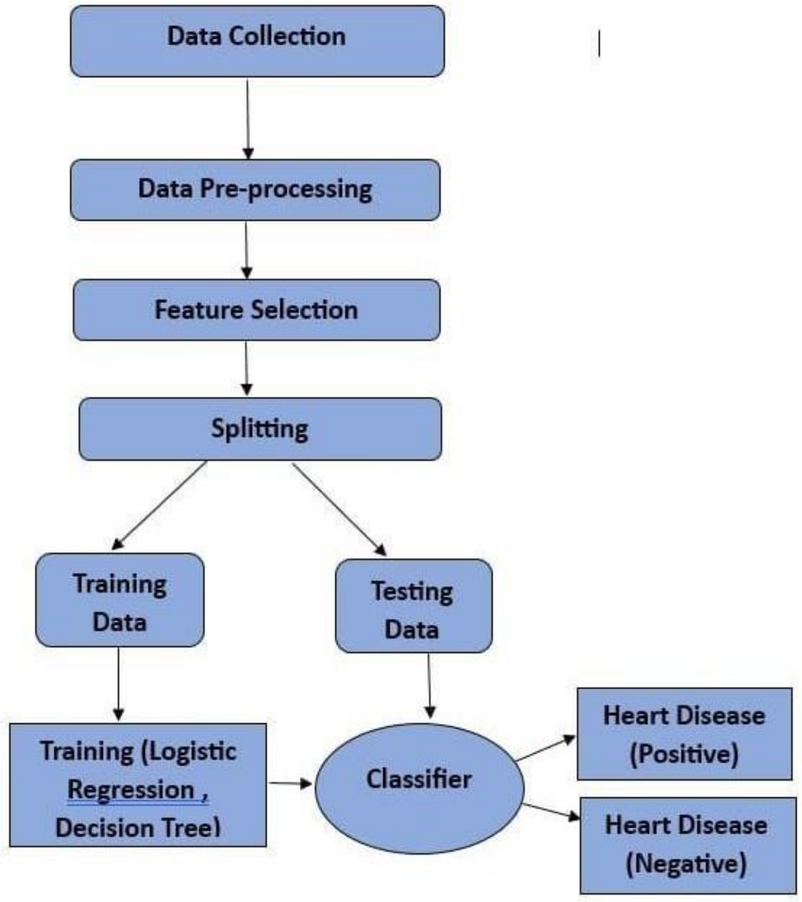

Fig.1

DOI: 10.17148/IJARCCE.2024.13315

This system utilizes Logistic Regression algorithm to analyze a variety of patient-related data, including medical history, lifestyle factors, and clinical measurements. The reason to prefer Logistic Regression over other algorithms is preferred because it has a capability to implicitly perform feature selection by assigning coefficient to the features, it is a good start for binary classification problems and it is robust in general.

Logistic Regression also is relatively efficient to train and make predictions and excels in classifying data into two distinct categories. By learning and understanding from patients factual data and patterns, machine learning models are capable of identifying correlations and risk factors that cannot be identified immediately by healthcare professional. As a result, the system can assist medical practitioners in making data driven decision, guiding them towards more targeted interventions and preventive measures.

Data collection: To begin with, we gather data for our heart prediction model. The source of this data was from a web-based repository named Kaggle [9] which has a .csv format files storing data. Over 900 patients. The process of data collection begins by gathering dataset which is then further divided into test and train sets. Test data helps to appraise the prediction model, and on other side train data helps to understand and learn about the prediction model.

Collection different type of data and representative dataset containing patient information, including measurements. Ensure the dataset is accurate, free from biases, and includes a sufficient number of positive and negative heart disease case.

Data Preprocessing: Handling missing values, outliers, and inconsistencies in the dataset are dealt with in this process. The pandas library, a standard data preprocessing tool in Python programming language, was employed for handling these issues pandas library was utilized. Divide the dataset into training and testing sets for model evaluation. Data preprocessing is an important step that is needed to be performed because raw data contains noise, missing values, inconsistencies, other irregularities that interrupts modeling tasks.

Feature Selection: Feature selection refers to the process of choosing a subset of the most relevant features from the original set to use in model building. Identify and select relevant features that contributes in predicting heart disease. We have used techniques like correlation analysis, regression, dimensionality reduction for feature selection.The features which had finalized were cholesterol, fasting sugar, blood pressure and max heart rate.

Data Splitting: Data Splitting is basically dividing the datasets into training and testing sets to evaluate model performance. In this project, 10% of the data is utilized for testing and 90% of the data is used for training. In our project we have divided dataset in various ratios in order to compare accuracy we can see this in detail in result and analysis.

Model Training: Use the training dataset to train the selected model. Optimize hyper parameters through techniques like grid search or random search to improve model performance. Training Set – 90% and Testing Set-10%.

Selection of Attributes: This is the most important step which includes choosing the right attributes for the prediction system is part of the attribute or feature selection approach. The prediction is based on a number of patient characteristics, gender, age, sex, chest pain, cholesterol, Fasting Sugar, Resting Sugar, blood Pressure, Max Heart Rate. This model’s attribute selection process makes use of the correlation matrix

Table-2. Attribute Selection

Sr. No	Attributes	Description	Data Type
1	Age	Patients Age	Numerical
2	Sex	Gender of Patient (Male:1,Female:0)	Nominal
3	Chest Pain	1:Typical Angina, 2:Atypical Angina 3:Non-Angina Pain 4:Asymptomatic	Nominal

4	Cholesterol	NormalRange: 200mg/dl(5.17mmol/L) BorderlineHigh: 200-239(5.17 to 6.18mmol/L) Highest: 240mg/dl(6.21mmol/l)	Numerical
5	Fasting_sugar	70-100mg/dL	Numerical
6	Resting_sugar	AfterMeals: 80-140mg/dL 1 Hour after eating:90-130mg/dL 2 Hour after eating :80-120mg/dL	Numerical
7	Blood Pressure	HighBloodPressure: 140/90 LowBlood Pressure: 90/60	Numerical
8	Max Heart Rate	30 – 39: 185bpm 40 – 49: 175bpm 50 – 59: 165bpm 60 – 69: 155bpm	Numerical

Technologies used: In this paper, we used Python Programming Language for data analysis and machine learning tasks. This model uses sklearn and pandas' libraries for data cleaning and interpretation. We have selected logistics regression as the algorithm. For front end development, we use React JS, to ensure consistent UX design and create reusable components.

Machine Learning: A systematic of several algorithms, machine learning is a strong technology that gives the system the ability to mimic human learning and thinking processes.

Logistic Regression: Logistic regression can be identified as supervised learning which is helpful for solving classification and regression problems. As we know variables are divided or classified discrete or binary values, which is helpful while dealing with classification problems,0 or 1. Logistic Regression uses sigmoid function that helps one to determine values of categorical variables, this answers can be 0 or 1, yes or no, true or false. Logistic regression can be used for predictive analysis that are based on different mathematical functions and operations. This type of function is called logistic function, often known as sigmoid function that is used in logistic regression. The logistic regression gives accuracy of 70% to 90%. In our model logistic regression gives accuracy of 90%.

Decision Tree: Another supervised learning method is the decision tree algorithm, which works best for tackling classification problems but can also be applied to regression issues. A decision tree is a hierarchical classifier in which the features of a dataset are represented by internal nodes, the decision is represented by each leaf node, which is devoid of any additional branches. It is a graphical tool that shows all of the analysis or decision-making. A decision tree only poses a query, and then divides tree into subsequent sub-queries based on the response. If same application has a wider range of data, chances are that accuracy and precision matters may change. Further research and review is needed as this technology is showing promising future.

IV. RESULT AND ANALYSIS

In our model we have used logistic regression and decision tree. In this model, we have divided the datasets into training and testing sets for evaluation of model performance. Our project 10% of the data is utilized for testing and 90% of the data is used for training. Differing from previous work we have used Logistics Regression and increased its accuracy with increasing training from 50% to 90% and 90% training set and 10% testing set provides highest accuracy of 90.75%. We can observe this in given below Table-1.

Our model in comparison to other models is different and more effective because in order to receive accuracy we have used only two algorithms which includes logistic regression and decision tree.

Table-1 displays the datasets with five distinct ratios and their corresponding accuracy values, which are used to test the logistic regression and decision tree models.

For a split ratio of 90:10 between training and testing, logistic regression and decision trees yielded an accuracy of 90.75%.

Figure 2 illustrates how the model's accuracy increases with further training, and Table 3 displays the correctness of the results.

Table-1. Split percentage of training and test set

Sr No.	Training Set	Testing Set	Accuracy
1	50	50	82.71%
2	60	40	83.58%
3	70	30	85.35%
4	80	20	87.25%
5	90	10	90.71%

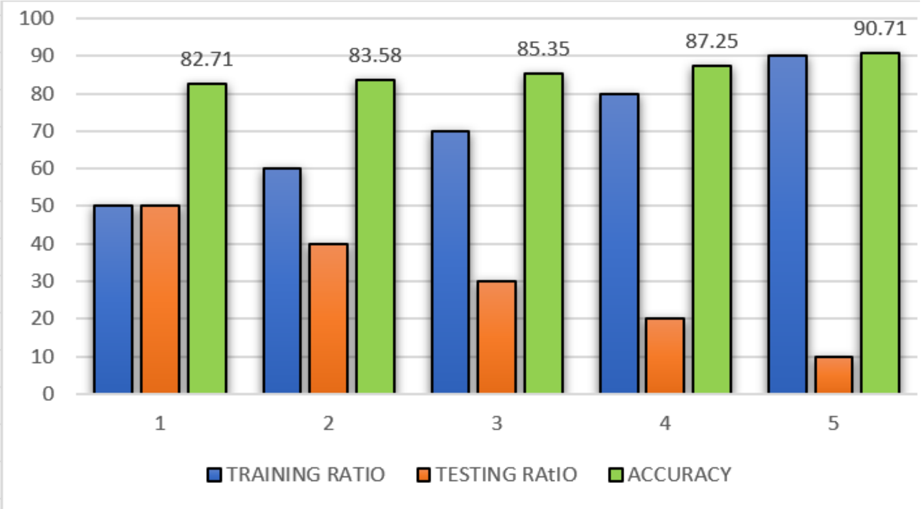

Fig. 2. Logistic regression and Decision tree accuracy for training and testing ratio

The accuracy of 90.71% is obtained by using logistic regression for divided ratio of training and testing 90:10.
The Logistics Regression increases its accuracy with increasing training by 50% to 90% and by taking 90% training and 10% testing provides highest accuracy of 90.71%.

V. CONCLUSION

In conclusion, the application that was developed has an accuracy of 90% after many trial and error attempts and consideration of a few key attributes. Yes, it is true that this accuracy has been achieved for a relatively small amount of data, which has been used in a very controlled environment with a restricted number of attributes, but this clearly indicates that the algorithm functions for this scenario. Using machine learning to predict heart disease has proven to be a successful strategy. Integrating a range of data, including patient demographics, medical history, and lifestyle traits, can help develop robust prediction models. When it comes to accurately detecting those who are in danger, many algorithms vary in their usefulness, including logistic regression with decision trees. Using the feature importance analysis, which helps identify the key components influencing predictions, clinicians can more effectively target their interventions.

DOI: 10.17148/IJARCCE.2024.13315

However, there are still problems to be tackled, including the necessity for large and diverse datasets, the interpretability of complex models, and ethical concerns regarding patient privacy. Validating the model and continuously improving it are essential. There's a risk that the accuracy and precision issues could alter if the application has a larger variety of data. Given the prospective future of this technology, more investigation and analysis are required.

REFERENCES

[1]. Researchers Chintan M. Bhatt and Parth Patel et al. "Effective Heart Disease Prediction Using Machine Learning techniques”. Multidisciplinary Digital Publishing Institute (MDPI), 2023.

[2]. Scientist Madhumita Pal et al. proposed "Prediction of Heart Diseases using Random Forest". Institute of Physics (IOP), 2020.

[3]. Toukir Ahmed et al, proposed “Prediction of heart disease Multi-Layer Perceptron Neural Network and Support vector Machine” International Conference on Electrical Information and Communication Technology (EICT),2019

[4]. Ambrish G,Bharathi Ganesh et al, proposed "Logistic regression technique for prediction of cardiovascular disease" in KeAi Chinese root global impact,2023.

[5]. Nayab Akhtar et.al,Heart disease Prediction,2021

[6]. Keshav Srivastava and Dilip Kumar Choubey published a paper on “Heart Disease Prediction using Machine Learning and Data Mining” in International Journal of Recent Technology and Engineering (IJRTE), 2020.

[7]. Archana Singh and Rakesh Kumar et al, proposed “Heart Disease Prediction Using Machine Learning Algorithms” International Conference on Electrical and Electronics Engineering, 2020.

[8]. Harshit Jindal, Sarthak Agarwal, Rishabh Khera Jain and Preeti Nagrath proposed “Heart disease prediction using machine learning algorithms” 2020.

[9]. Researchers Prathyusha Ammisetty, Dr. Chiranjeevi Paritala et al, proposed "Prediction of Heart Disease Using Machine Learning Algorithm". International Journal for Research Trends and Innovation (IJRTI) 2022.

[10]. Tulika Lodh, Anirban Dey, Naorem Rinita, Sunil Kumar, Subodh Kumar et al, proposed a paper "Analysis of Heart Disease Prediction using Machine Learning Techniques”. International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET) 2021

[11]. A.Vindhya Sree, CH. Neha, K. Hima Bindu ,Sumera published a paper on "heart disease prediction". International Journal for Innovative Engineering and Management Research (IJIEMR) 2022.

[12]. Hossam Magdy Balaha1 Ahmed Osama Shaban published a paper on "A multi-variate heart disease optimization and recognition framework". Neural Computing and Applications (2022).

[13]. Researchers Hafsa Binte Kibria and Abdul Matin published a paper on “The Severity Prediction of the Binary and Multi-Class Cardiovascular Disease a Machine Learning Based Fusion Approach”,2022.