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 2, February 2024

DOI: 10.17148/IJARCCE.2024.13227

Machine Learning and Web Solution for Heart Disease Prediction

Suwarna Nimkarde[1], Omkar Chavan[2], Shlok Damudre[3] Bhagyashree Nikam[4]

Lecturer, Department of Computer Technology, Bharati Vidyapeeth Institute of Technology, Navi Mumbai, India[1]

Student, Department of Computer Technology, Bharati Vidyapeeth Institute of Technology, Navi Mumbai, India [2]

Student, Department of Computer Technology, Bharati Vidyapeeth Institute of Technology, Navi Mumbai, India [3]

Student, Department of Computer Technology, Bharati Vidyapeeth Institute of Technology, Navi Mumbai, India [4]

Abstract: According to WHO data, cardiac disorders cause around one crore twenty lakh deaths annually. Heart illness and cardiovascular disease have historically had a significant impact on the medical field, indicating their extreme hazards and widespread impact. Though it is not feasible to predict heart illnesses or CD in advance, nor is it feasible to monitor patients around the clock due to the high time and expertise requirements, treatment and diagnosis for heart disease can be extremely difficult, especially in developing or impoverished nations. Additionally, a person may pass away as a result of inadequate medical care or a delayed diagnosis. Researchers often use the wealth of data from the medical business to produce new science and technologies aimed at reducing the number of heart disease-related deaths. Numerous algorithms and data mining approaches are available to extract information from databases and utilize that information to make highly accurate predictions about cardiac ailments. We used machine learning in this heart disease model. The whole process was implemented on a dataset from Kaggle that contained 14 attributes and 303 rows in total. The model employs the following algorithms: Random Forest, SVM, NB, K-NN, Decision Tree, and Logistic Regression.

Keywords: Machine Learning, Heart Disease, Kaggle, Cardiovascular Disease, WHO, Classification, Dataset, ML Algorithm.

I. INTRODUCTION

In addition to being one of the most vital components of the cardiovascular system, the heart is one of the most vital organs of our body and aids in pumping. It consists of a network of different blood vessels, such as arteries, capillaries, veins, and lymphatic vessels. Blood travels through our system with the assistance of blood vessels. Unusual blood flow from the heart is the cause of certain cardiac conditions, such as heart attacks, strokes, and coronary heart disease (CVD). Heart defects of any form can lead to a variety of conditions known as cardiovascular illnesses, including congenital heart disease, arrhythmias, heart failure, and others. Numerous sectors focused on biological research have noted throughout time that cardiovascular illnesses are now the leading cause of death worldwide. Because cardiovascular disorders can be problematic and severe, prompt medical attention is required. Congenital heart disease, arrhythmia, coronary artery disease, heart failure, heart muscle disease, and heart valve disease are among the many different forms of cardiovascular disorders. In addition to individual behaviors and employment, a variety of additional factors, including excessive alcohol use, tobacco use, caffeine intake, physical inactivity, and prolonged sitting, can contribute to cardiovascular disease. In addition to these, there are additional psychological variables, such as stress, anxiety, depression, and obesity. As a result, we need to take precautions and make proactive efforts to stop these behaviors, as well as record the patient's symptoms and daily routines that put him at risk for cardiovascular illnesses. Patients must undergo a number of tests before a diagnosis of CVD can be made. These procedures include blood and blood sugar testing, electrocardiography (ECG), blood pressure measurement, cholesterol testing, and auscultation. These examinations are typically lengthy and complex if the patient's condition is critical and immediate medicine is needed. Making an accurate, timely, and effective medical diagnosis of heart disease is essential to taking preventative action to avert its repercussions. The main challenges facing the medical sciences today are providing high-quality services and producing precise and effective predictions. The final issue can be resolved with automation, which also presents a chance for healthcare-related research, particularly in the area of early disease identification, to increase survival rates in conditions like cancer and cardiovascular disease. Through ongoing experimentation and observation of the most recent advancements in programming and computing power, we have discovered that these enormous problems do, in fact, have an answer thanks to artificial intelligence and machine learning algorithms. Machine learning is being used for everything from creating systems to track vehicle safety to determining illness risk factors. With machine learning technologies, which offer popular predictive modeling methods to overcome present constraints, prediction algorithms based on huge amounts of data can

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 2, February 2024
DOI: 10.17148/IJARCCE.2024.13227

be developed. Classification is a prominent machine learning technique for result prediction. Classification models are useful in the identification of diseases when they are trained with sufficient data. The many methods for predicting and analyzing cardiac problems will be covered in this article.

II. PROBLEM STATEMENT

The heart is an essential component of the circulatory system, as we have already seen. Impaired cardiac function can result in life-threatening illnesses, including heart failure. We have been monitoring some recent advancements in machine learning (ML) approaches, which are employed for data extraction, pattern identification, prediction, and decision-making in a variety of computer science domains. Because deep learning uses layers of machine learning algorithms, it is also appropriate for edge computing. In the past, a number of studies have looked into the prediction and classification of cardiac disease using machine learning approaches. However, these kinds of studies focus more on the effects of certain machine learning techniques than on the application of optimized techniques to improve these techniques. This work describes the development of an intelligent heart disease prediction system utilizing machine learning techniques. The system predicts heart failure in patients at an early stage, allowing for an earlier start to diagnosis and therapy.

III. LITERATURE REVIEW

- Vishal Dineshkumar Soni [1] used the UCI dataset. Specifically, Decision Table, Naïve Bayes, SMO, and Lazy Kstar machine learning algorithms are employed. According to the experimental data, the Naive Bayes and Decision Table exhibit superior performance. The experimental results have an accuracy of 85.58% for Naïve Bayes and 85.15% for Decision Table.
- Dr. M. Kavitha, G. Gnaneswar, R. Dinesh, Y. Rohith Sai, and R. Sai Suraj [2] used the Cleveland heart disease dataset. Three algorithms were utilized: hybrid (Decision Tree + Random Forest), decision tree, and random forest. Upon applying this method to the dataset, the outcome demonstrated that the hybrid model outperformed the other models, achieving 88% accuracy.
- Rohit Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh [3] Using the UCI Heart Disease Dataset, the author of this work applied machine learning and deep learning techniques. This work makes use of the following techniques: decision tree, random forests, K-neighbors, SVM, logistic regression, and deep learning. Following the application of every technique to the dataset, the outcome demonstrates that, with an accuracy of 94.2%, the DL technique yields superior results.
- Pabitra K. Bhunia, Arijit Debnath, Poulami Mondal, Monalisa D. E., Kankana Ganguly, and Pranati Rakshit [4] Heart.csv is a dataset that was used. The random forest classifier, decision tree, K nearset neighbor, support vector machine, and logistic regression methods were applied to the dataset. According to the implementation results, the RFC and SVM algorithms perform better, with 90.32% accuracy for the RFC and 90.32% accuracy for the SVM.
- Karna V. V. Reddy, Irraivan Elamvazuthi, Azrina Abd Aziz, Sivajothi Paramasivam, Hui Na Chua, and S. Pranavanand [5] The research made use of the Cleveland Heart Dataset. NB, LR, SMO, IBk/KNN, AdaBoostM1 + DS, AdaBoostM1 + LR, Bagging + REPTree, Bagging + LR, JRip, and RF are the techniques that were employed. Upon executing every technique on the dataset, the findings indicate that the SMO and AdaBoostM1 + LR exhibit superior accuracy, scoring 85.148% and 84.818%, respectively.

IV. DATA ANALYSIS

Our goal in data analysis is to grasp and comprehend the provided data better. Data was obtained from Kaggle UCI. There are 303 records in the data set, each with 14 distinct properties. Our dependent variable, the target column in the data set, exists. Binary values 0 and 1 are used to encode the dependent variable. We created a few graphs and used bar plots, count plots, heat maps, and other visualization tools to analyze the data.
Based on the data analysis, we came to some significant and insightful conclusions. In this study, we shall examine each of those individually.

A. Understanding Dataset

DataSet Information: The attributes of the data collection are explained in this table, along with their meaning.

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 2, February 2024

DOI: 10.17148/IJARCCE.2024.13227

TABLE I DATA INFORMATION

Column Name	Description
age	Age
sex	1: male, 0: female
cp	Chest pain type, 1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic
trestbps	Resting blood pressure
chol	Serum cholestoral in mg/dl
fbs	Fasting blood sugar > 120 mg/dl
restecg	Resting electrocardiographic results (values 0,1,2)
thalach	Maximum heart rate achieved
exang	Exercise induced angina
oldpeak	oldpeak = ST depression induced by exercise relative to rest
slope	The slope of the peak exercise ST segment
ca	Number of major vessels (0-3) colored by flourosopy
thal	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
target	0: Healthy, 1: Problem with Heart

B. Analysis attribute and Correlation

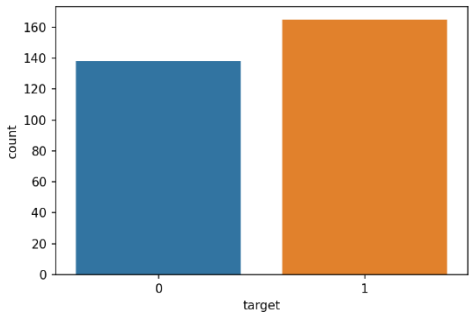

Fig. 1 Target

Figure 1 indicates that a greater number of records had heart-related issues. 54% of the data set's records have goal value 1, meaning they have a cardiac condition. 45% of the records in the data collection have a target value of 0. However, there is hardly any difference between the two. Thus, there's no need to preprocess the data in advance.

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 2, February 2024

DOI: 10.17148/IJARCCE.2024.13227

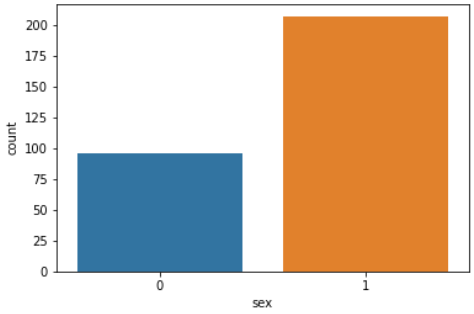

Fig. 2 Sex

We discovered that women are more likely than men to develop heart disease after examining the graph above. Since 0 values are more prevalent on target values than male values, based on our data, The graph above helps to visualize this.

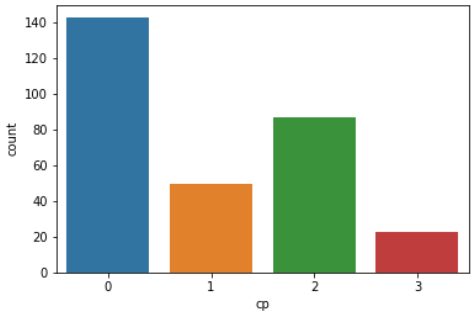

Fig. 3 CP

According to the graph above, there is a lower likelihood of heart disease in people experiencing normal angina-like chest pain. Others nearby are also at an equal risk of developing heart disease.

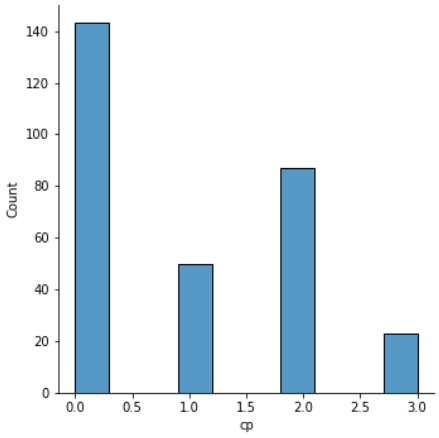

Fig. 3.1 CP

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 2, February 2024

DOI: 10.17148/IJARCCE.2024.13227

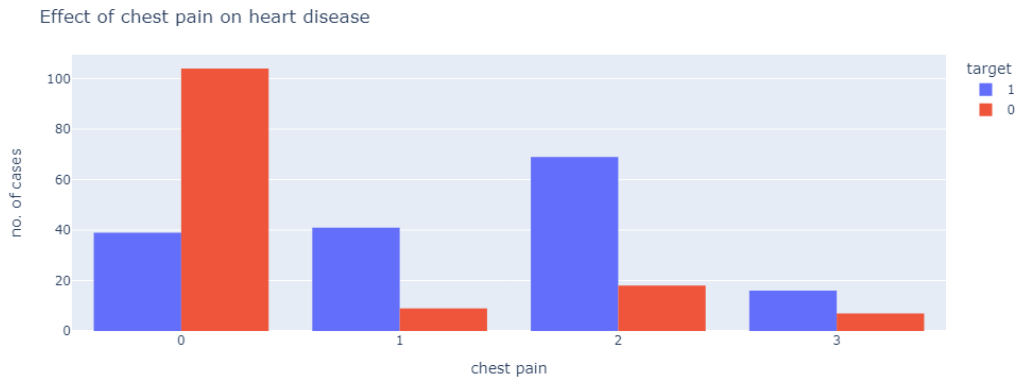

Fig. 3.2 CP

According to the two figures above, there are a lot of records with a value of 0, while values of 4 are extremely uncommon. Yet, value 4 has the biggest effect on the goal.

C. Correlation of Heatmap

A heatmap is a useful tool for visualizing and comprehending the relationship between several column values. The x and y axes are both covered by columns. Additionally, the color darkness is used to convey the desity of the link. The relationship is more dense in dark colors. Since the heat map's diagonal just depicts the relationship between the columns, we may ignore it. We deduced from the heat map that there isn't a meaningful correlation between the various attributes. We can therefore presume that every characteristic in the data set has the same significance for predicting the dependent variable target. We conducted this analysis in order to comprehend the data and determine the necessary pre-processing processes.

V. OUR APPROACH

Machine learning algorithms perform well when the dataset is small; however, deep learning algorithms outperform ML methods when the dataset is vast. We used machine learning techniques to achieve high disease prediction accuracy.

A. Understanding Dataset

Support Vector machines normally use a classification strategy, but they can also be used for regression problems. SVMs are ideally suited for high-dimensional scenarios and are also memory-efficient because they employ a subnet of training points in the decision function.
This model achieves an accuracy of approximately 86.89%.

B. Logistic Regression

Logistic regression is commonly used to solve and predict binary classification problems. The objective outcome is generally binary, which means there are only two possible classes for the target variable: 0 or 1.
The accuracy of this model is 88.52%.

C. Naïve Bayes

The Naive Bayes technique is a supervised machine learning algorithm that is fundamentally based on the Bayes theorem. The Bayes theorem is applied under the 'naive' premise of conditional independence.
The accuracy of this model is 80.33%.

D. K-NN

The K Nearest Neighbour approach is a lazy technique that generates the model without using any training data points because all of the training data is used during the testing phase. Also, there is no prior assumption for the underlying data distribution, indicating that this is a non-parametric technique.
The model's accuracy is 78.69%.

E. Logistic Regression

Random forest is a supervised learning algorithm that may be used for both classification and regression. A forest is made up of many trees; consequently, random forests simply construct decision trees based on randomly selected data samples and present forecasts for each tree, resulting in the selection of the best tree through voting.
The accuracy of this model is 88.52%.

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 2, February 2024

DOI: 10.17148/IJARCCE.2024.13227

VI. CONCLUSION

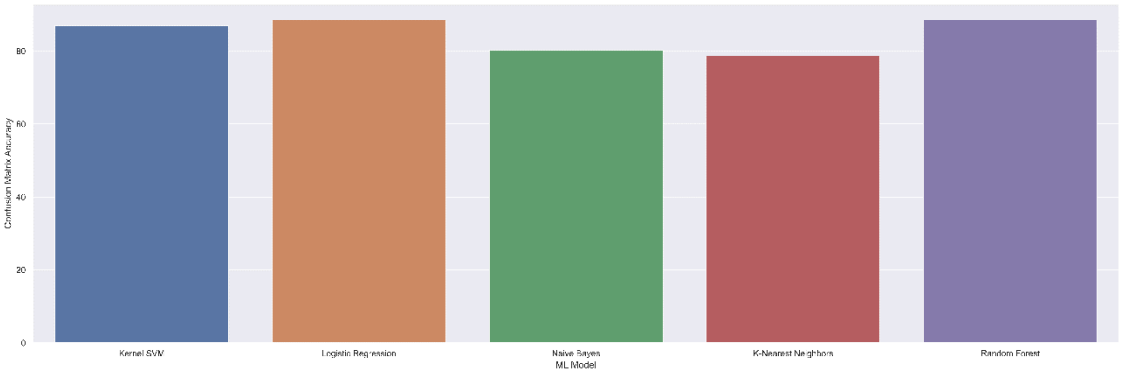

Fig. 4 ML Model Accuarcy

In this study, we investigated various algorithms and used them on a dataset. The solution employed the Kaggle heart dataset, which has 303 rows of data and 14 characteristics. We employed both machine learning algorithms for the implementation because we noticed that all methods performed exceptionally well. According to our analysis and the algorithms used, the Logistic Regression and Random Forest algorithms outperformed the other techniques in terms of accuracy. The accuracy achieved by Logistic Regression and Random Forest is 88.52%, respectively. This heart disease model will help us grasp and gain a better understanding of a person's health, which will be beneficial to both doctors and patients. This model allows us to quickly track the patient's health risk based on their age and any other data that was used. In the future, we will merge multiple datasets with a larger number of observations to conduct more experiments.

REFERENCES

[1]. Vishal Dineshkumar Soni, "Detection Of Heart Disease Using Machine Learning Techniques", *International Journal of Scientific & Technology Research*, 2020.

[2]. Dr. M. Kavitha, G. Gnaneswar, R. Dinesh , Y. Rohith Sai1, R. Sai Suraj,"Heart Disease Prediction using Hybrid machine Learning Model", *International Conference on Inventive Computation Technologies*, 2021.

[3]. Rohit Bharti , Aditya Khamparia , Mohammad Shabaz , Gaurav Dhiman , Sagar Pande , and Parneet Singh,"Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning", *Hindawi Computational Intelligence and Neuroscience*, 2021.

[4]. Pabitra Kumar Bhunia, Arijit Debnath, Poulami Mondal, Monalisa D E, Kankana Ganguly, Pranati Rakshit", Heart Disease Prediction using Machine Learning, 2021.

[5]. Karna Vishnu Vardhana Reddy, Irraivan Elamvazuthi, Azrina Abd Aziz, Sivajothi Paramasivam, Hui Na Chua and S. Pranavanand,"Heart Disease Risk Prediction Using machine Learning Classifiers with Attribute Evaluators", 2021.