IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

Heart Disease Prediction using Machine Learning Algorithms

Abhishek Umakant Pujare
Institute of Distance and Open Learning, Mumbai, Maharashtra, India

Abstract: *Heart disease cases are rising quickly every day, so it's crucial and worrisome to anticipate any potential illnesses in advance. This diagnosis is a challenging job that requires accuracy and efficiency. The primary focus of the study paper is on which patients, given different medical characteristics, are more likely to have heart disease. Using the patient's medical history, we developed a method to determine whether a heart disease diagnosis is probable or not for the patient. To forecast and categorize the patient with heart disease, we used a variety of machine learning algorithms, including KNN and logistic regression. The regulation of how the model can be used to increase the precision of heart attack prediction in any person was done in a very helpful way. When compared to the previously employed classifiers, such as naive bayes, etc., the suggested model's accuracy in predicting signs of having heart disease in a specific person was quite satisfactory. It did this by using KNN and Logistic Regression. Thus, using the provided model to determine the likelihood that the classifier will correctly and reliably recognize heart disease has relieved quite a bit of pressure. Given's system for predicting heart disease improves patient treatment while costing less. This research has provided us with a wealth of information that can be used to predict who will develop heart disease. It utilizes the .pynb file type.*

Keywords: Machine Learning, Heart Disease, Prediction, Detection, Naïve Bayes

I. INTRODUCTION

Machine learning is a technique for manipulating and extracting implicit, unknown or known information about data that may be helpful. Machine learning is a very broad and diverse subject, and its application and scope are expanding daily. In order to forecast and determine the accuracy of the given dataset, machine learning incorporates a variety of classifiers from supervised, unsupervised, and ensemble learning. Given that it will benefit many individuals, we can apply that knowledge to our HDPS project. These days, a wide variety of disorders that potentially harm your heart are referred to as cardiovascular diseases. According to the World Health Organization, there are 17.9 million CVD-related deaths worldwide.

It is the main factor in adult deaths. By using a person's medical history, our initiative can identify those who are most likely to be diagnosed with a cardiac condition. It can identify patients who are experiencing any heart diseases symptoms, such as chest pain or high blood pressure, and assist in making a diagnosis with fewer medical procedures and more efficient therapies so that the patient can be treated appropriately. Three data mining approaches, specifically: (1) Logistic regression and KNN, are the major topics of this study along with Random Forest Classifier. Our project's accuracy is 87.5%, which is higher than the accuracy of the prior system, which only used one data mining technique. Thus, increasing the use of data mining methods raised the HDPS precision and effectiveness. The field of supervised learning includes logistic regression. Logistic regression uses only discrete values. This project's goal is to determine, depending on the patient's medical characteristics—such as gender, age, chest discomfort, fasting blood sugar level, etc.— whether they are likely to be diagnosed with any cardiovascular heart illnesses. A dataset with the patient's medical background and characteristics is chosen from the UCI repository. We make a prediction about the patient's potential for heart disease using this dataset. We categorize a patient based on 14 medical characteristics to determine whether they are likely to develop a heart condition in order to anticipate this. Three algorithms— KNN, Random Forest Classifier, and Logistic Regression— were used to train these medical characteristics. KNN is the most effective algorithm here, providing an accuracy of 88.52%. Finally, we categorize individuals according to whether they are at risk of developing a cardiac condition or not. This procedure is also incredibly economical.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

II. LITERATURE REVIEW

Using the UCI Machine Learning dataset, extensive research has been done to predict cardiac disease. varied data mining approaches have been used to achieve varied accuracy levels, which are detailed below

Avinash Golande and colleagues investigate various ML algorithms that can be used to categorise cardiac disease. An investigation was conducted to examine the accuracy of the classification algorithms Decision Tree, KNN, and K-Means. The study found that Decision Trees had the highest accuracy, and it was concluded that by combining various methodologies and fine-tuning its parameters, it might be made more effective.

A system that combined the MapReduce algorithm with data mining techniques has been suggested by T. Nagamani et al. For the 45 instances in the testing set, the accuracy obtained according to this article was higher than the accuracy obtained using a traditional fuzzy artificial neural network. Here, the usage of dynamic schema and linear scaling increased the algorithm's accuracy.

AI ML model created by Fahd Saleh Alotaibi compares five alternative methods. When compared to Matlab and Weka, the Rapid Miner tool performed more accurately. This study compared the classification accuracy of Decision Tree, Logistic Regression, Random Forest, Naive Bayes, and SVM algorithms. The most accurate algorithm was the decision tree algorithm.

A system that employs NB (Naive Bayesian) approaches for dataset categorization and the AES (Advanced Encryption Standard) algorithm for safe data transport was proposed by Anjan Nikhil Repaka, et al.

A survey was conducted by Theresa Princy, R., et al., using various classification algorithms for heart disease prediction. The classifiers' accuracy was examined for a variety of variables using Naive Bayes, KNN (K-Nearest Neighbor), Decision Trees, and Neural Networks as the classification methodologies.

Heart disease was predicted by Nagaraj M. Lutimath et al. using Naive Bayes classification and SVM. (Support Vector Machine). Mean Absolute Error, Sum of Squared Error, and Root Mean Squared Error are the performance measurements utilised in analysis. It has been determined that SVM outperformed Naive Bayes in terms of accuracy.

After reading the aforementioned publications, the fundamental idea behind the suggested system was to build a heart disease prediction system based on the inputs presented in Table 1. By comparing the accuracy, precision, recall, and f-measure scores of the classification algorithms Decision Tree, Random Forest, Logistic Regression, and Naive Bayes, we were able to determine which classification algorithm would be most effective at predicting heart disease.

Shah et al.'s study from 2020 [18] sought to create a model for predicting cardiovascular illness using machine learning methods. The 303 cases and 17 attributes of the Cleveland heart disease dataset, which was sourced from the UCI machine learning repository, were utilised to generate the data for this project. The authors used a range of supervised classification techniques, including k-nearest neighbor, naive Bayes, decision trees, and random forests. (KKN). The study's findings showed that, at 90.8% accuracy, the KKN model had the best level of precision.

The study emphasises the potential value of machine learning methods in anticipating cardiovascular illness and the the significance of choosing the right models and methods to get the best results.

In a study by Drod et al. (2022), the goal was to identify the most important risk factors for cardiovascular disease (CVD) in patients with metabolic-associated fatty liver disease using machine learning (ML) approaches. (MAFLD). 191 MAFLD patients had their blood biochemically analyzed, and subclinical atherosclerosis was evaluated. Using ML techniques, such as multiple logistic regression classifier, univariate feature ranking, and principal component analysis, a model to identify those with the highest risk of CVD was created. (PCA). The most important clinical traits, according to the study, were hypercholesterolemia, plaque scores, and length of diabetes. With an AUC of 0.87, the ML method worked well, correctly classifying 114/144 (79.17%) low-risk patients and 40/47 (85.11%) high-risk patients. The results of the study show that using straightforward patient criteria, an ML technique is beneficial for identifying MAFLD patients with extensive CVD.

The author of a study by Alotalibi (2019) [19] set out to look into the effectiveness of machine learning (ML) approaches for diagnosing heart failure condition. The study made use of

Using a dataset from the Cleveland Clinic Foundation, we developed prediction models using a variety of ML algorithms, including decision trees, logistic regression, random forests, naive bayes, and support vector machines (SVM). During the model development process, a 10-fold cross-validation strategy was used. The findings showed that the decision tree algorithm, which had a rate of 93.19%, and the SVM method, which had a rate of 92.30%, had the

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

highest accuracy in predicting heart disease. This work highlights the decision tree algorithm as a potential useful tool for forecasting heart failure disease and sheds light on the possibilities of ML approaches as such.

III. METHODOLOGY

This study seeks to estimate the likelihood of developing heart disease using computerised heart disease prediction, which may be useful for patients and medical professionals.

We used a dataset and many machine learning methods to accomplish this goal, and the findings are presented in this study report. We intend to sanitise the data, get rid of extraneous details, and add new characteristics like MAP and BMI to improve the technique. The dataset will then be divided depending on gender, and k-modes clustering will be used. Finally, we will use the cleaned data to train the model. As shown in Figure, the revised process will result in more accurate results and greater model performance..

3.1 Data collection and pre-processing.

This study seeks to estimate the likelihood of developing heart disease using computerised heart disease prediction, which may be useful for patients and medical professionals.

We used a dataset and many machine learning methods to accomplish this goal, and the findings are presented in this study report. We intend to sanitise the data, get rid of extraneous details, and add new characteristics like MAP and BMI to improve the technique. The dataset will then be divided depending on gender, and k-modes clustering will be used.

Finally, we will use the cleaned data to train the model. As shown in Figure, the revised process will result in more accurate results and greater model performance.

3.2 Data Collection

The process of gathering, measuring, and analysing precise insights for study is known as data collection. A researcher can assess their hypothesis using the data that they have gathered. Regardless of the field of study, data gathering is typically the first and most crucial phase in the research process. A structured data set of Algerians who have completed analyses at the Mohand Amokrane EHS Hospital ex CNMS in Algiers, Algeria, is used in this study. It has 1200 rows and 20 columns, and the variables age, sex, cp, trestbps, chol, Ex-Ang, Col-Ves, fbs, restecg, thalach, exang, oldpeak, slope, RBP, ca, thal, smoking, alcohol use.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
1	29	1	1	130	204	0	0	202	0	0.0	2	0	2
2	29	1	1	130	204	0	0	202	0	0.0	2	0	2
3	29	1	1	130	204	0	0	202	0	0.0	2	0	2
4	29	1	1	130	204	0	0	202	0	0.0	2	0	2
5	34	1	3	118	182	0	0	174	0	0.0	2	0	2
6	34	0	1	118	210	0	1	192	0	0.7	2	0	2
7	34	1	3	118	182	0	0	174	0	0.0	2	0	2
8	34	0	1	118	210	0	1	192	0	0.7	2	0	2
9	34	1	3	118	182	0	0	174	0	0.0	2	0	2
10	34	0	1	118	210	0	1	192	0	0.7	2	0	2
11	35	0	0	138	183	0	1	182	0	1.4	2	0	2
12	35	1	1	122	192	0	1	174	0	0.0	2	0	2
13	35	1	0	120	198	0	1	130	1	1.6	1	0	3

Fig : Datasets for the study

3.3 Manual investigation.

Data exploration, also known as manual exploration, is the first stage of data analysis, during which users examine a sizable data collection informally to find the first patterns, traits, and areas of interest. This approach is intended to help establish a broad picture of significant trends and key areas to investigate in more detail rather than to show every piece of information that a data set contains. To begin the pre- processing for our study, we add a column to our data collection called Target that has the values 0 or 1 (0 = not sick, 1 = sick). Figure explains this.

IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 1, January 2024

fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	1	192	0	0.7	2	0	2	1
0	0	174	0	0.0	2	0	2	1
0	1	192	0	0.7	2	0	2	1
0	1	182	0	1.4	2	0	2	1
0	1	174	0	0.0	2	0	2	1
0	1	130	1	1.6	1	0	3	0
0	0	156	1	0.0	2	0	3	0
0	1	182	0	1.4	2	0	2	1
0	1	174	0	0.0	2	0	2	1
0	1	130	1	1.6	1	0	3	0
0	0	156	1	0.0	2	0	3	0
0	1	182	0	1.4	2	0	2	1
0	1	174	0	0.0	2	0	2	1
0	1	130	1	1.6	1	0	3	0
0	0	156	1	0.0	2	0	3	0

Fig: Manual Investigation

3.4 Data Preprocessing

We prepare the data to be implemented before beginning the application of machine learning algorithms. This phase is accomplished in two steps:

features choice The correlation matrix is the basis for this stage. Initially, we possessed 20 of the previously mentioned qualities. We identified 13 attributes (age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal) that are connected to and reliant on one another after using the Pearson correlation matrix.Table explains the specifics of the chosen characteristics.

Attribute	Description	Values
Age	Age	29 to 62 years
Sex	Sex	1 - male 2 - female
CP	Chest pain type	1- typical angina pectoris 2- atypical angina 3- non-anginal pain 4- asymptomatic
trestbps	Resting blood pressure in mm/Hg	Numeric value : example: 140mm/Hg
Chol	Serum cholesterol in mg/dl	Numeric value : example: 289mg/hg
Fbs	Fasting blood pressure>120mg/dl	Numeric value : example: 129mm/Hg
Restecg	Resting electrocardiographic results	0- normal, 1- have the ST-T 2- hypertrophy
thalach	Maximum heart rate achieved	Numeric value : Example: 140,173
Exang	Exercise induced angina	1 - Yes 2 - No
Oldpeak	ST depression induced by exercise relative to rest	Numeric Value
Slope	The slope of the peak exercise ST segment	1 - on the rise 2 - flat 3- the downhill slope
Ca	Number of major vessels colored by flourosopy	0 to 3 vessels
Thal	Thalassemia	3- normal, 6- defect repaired, 7- reversible defect

Table: The details of selected Features

IV. SYSTEM IMPLEMENTATION

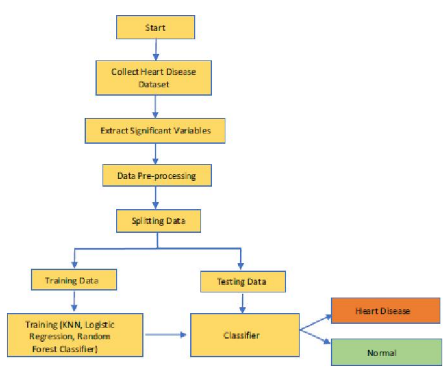


Fig: System Flow

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

4.1 Split Data Set:

Dividing a data set our data set is divided into two sections: The test portion is 20% larger than the training data set in the first half. Fig. shows the division of our data set..

4.2 Testing Algorithm:

We verify the accuracy of each algorithm on the various data sets after running the three algorithms on the four data sets (600, 800, 1000, and 1200 lines). Table 3 shows how we determined the accuracy for the three methods based on the confusion matrix (we used the latest data set, which had 1200 lines)

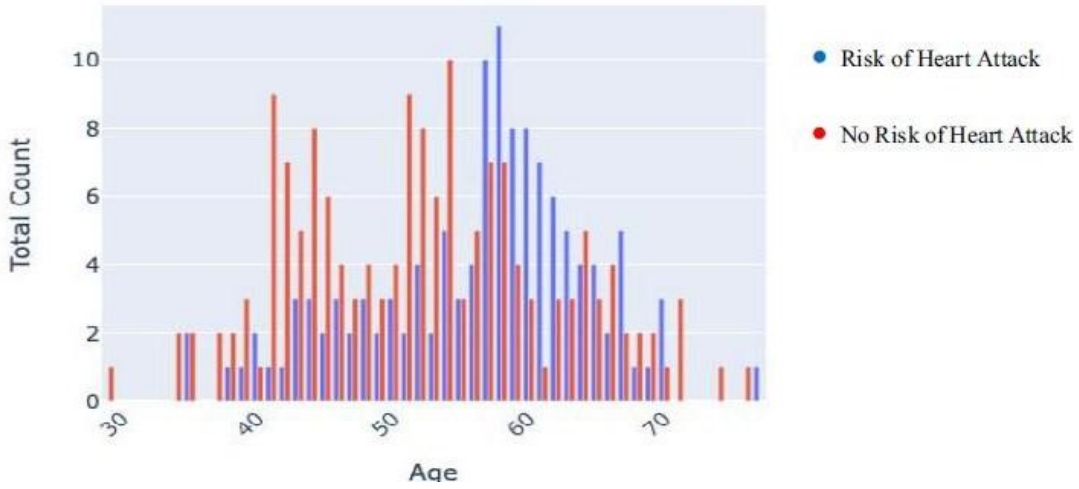
Now that we have the 4 data sets, we can show the accuracy of each algorithm, In terms of accuracy and stability against changes in the data sets, Fig. compares the three algorithms

	Neural Network Dataset (1200 lines)		SVM Dataset (1200 lines)		KNN Dataset (1200 lines)	
	Sick	*Not sick*	*Sick*	*Not sick*	*Sick*	*Not sick*
Sick	94	8	90	11	84	18
Not sick	6	92	9	90	11	87
Accuracy	**93%**		**90%**		**85,5%**	

V. RESULTS

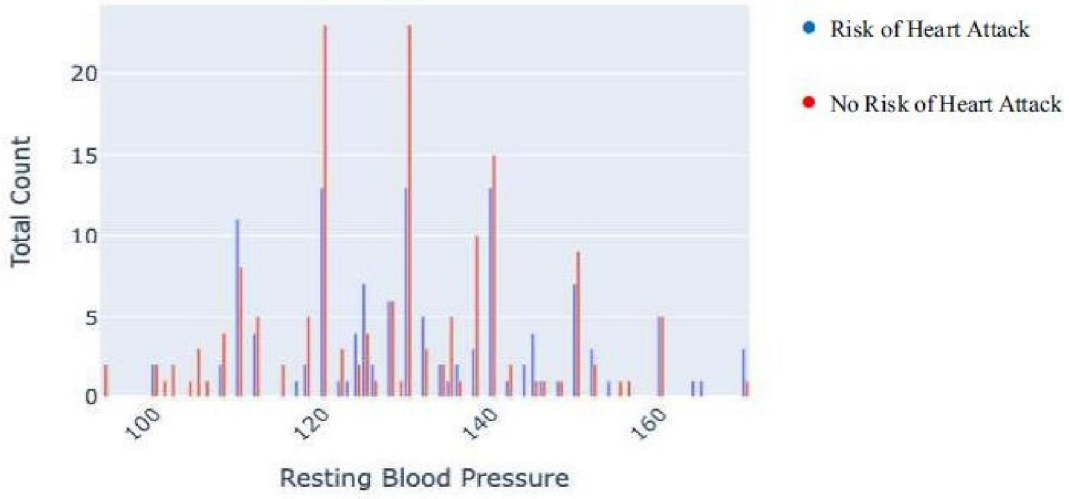
Shows the Risk of Heart Attack on the basis of their age.

From these findings, it is clear that even if the majority of studies use other algorithms, such as SVC and Decision trees, to identify patients with heart disease, KNN, Random Forest Classifier, and Logistic Regression produce a superior outcome to them [23]. Our algorithms are faster and more precise than those employed by earlier studies. They also save a significant amount of money, making them very cost- effective. Furthermore, the combined maximum accuracy of KNN and Logistic Regression is 88.5%, which is higher than or nearly equal to the accuracy of earlier studies. So, to sum up that our accuracy is improved due to the increased medical attributes that we used from the dataset we took. Our project also tells us that Logistic Regression and KNN outperforms Random Forest Classifier in the prediction of the patient diagnosed with a heart Disease. This proves that KNN and Logistic Regression are better in diagnosis of a heart disease. The following ‘figures ’ shows a plot of the number of patients that are been segregated and predicted by the classifier depending upon the age group, Resting Blood Pressure, Sex, Chest Pain



Shows the Risk of Heart Attack on the basis of their age

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024



Shows the Risk of Heart Attack on the basis of their Resting Blood Pressure

VI. CONCLUSION

Heart disease has increased in prevalence throughout the world, including in our country. (Algeria). Consequently, diagnosing the illness before contracting it reduces the danger of dying. There has been extensive research done in this prediction field.

Our study is a component of the investigation into the identification and prognosis of cardiac disease.

It is based on the use of machine learning algorithms, of which we have selected the three most popular ones (Neural Network, SVM, and KNN), on a real data set of Algerian people, with excellent results; we reached 93% accuracy with Neural Network. The key finding of our study was that, after testing the algorithm's stability on a variety of data sets of varying sizes, it was clear that neural networks produced the greatest results. Additionally, we conducted research on feature selection and employed a correlation matrix to identify attribute dependencies. This strategy can be improved in a number of ways, including by using deep learning algorithms, other attribute selection techniques, and even bigger data sets..

REFERENCES

[1] Avinash Golande, Pavan Kumar T, "Heart Disease Prediction Using Effective Machine Learning Techniques", International Journal of Recent Technology and Engineering, Vol 8, pp.944-950,2019.

[2] T.Nagamani, S.Logeswari, B.Gomathy," Heart Disease Prediction using Data Mining with Mapreduce Algorithm", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019.

[3] Fahd Saleh Alotaibi," Implementation of Machine Learning Model to Predict Heart Failure Disease", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.

[4] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, "Design And Implementation Heart Disease Prediction Using Naives Bayesian", International Conference on Trends in Electronics and Information(ICOEI 2019).

[5] Theresa Princy R,J. Thomas,'Human heart Disease Prediction System using Data Mining Techniques', International Conference on Circuit Power and Computing Technologies,Bangalore,2016.

[6] Nagaraj M Lutimath,Chethan C,Basavaraj S Pol.,'Prediction Of Heart Disease using Machine Learning', International journal Of Recent Technology and Engineering,8,(2S10), pp 474-477, 2019.

[7] UCI, —Heart Disease Data Set.[Online]. Available (Accessed on May 12020): https://www.kaggle.com/ronitf/heart-disease-uci.

[8] Sayali Ambekar, Rashmi Phalnikar,"Disease Risk Prediction by Using Convolutional Neural Network",2018 Fourth International Conference on Computing Communication Control and Automation.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

[9] C. B. Rjeily, G. Badr, E. Hassani, A. H., and E. Andres, ―Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field, ‖ in Machine Learning Paradigms, 2019, pp. 71 – 99.
[10] Jafar Alzubi, Anand Nayyar, Akshi Kumar. "Machine Learning from Theory to Algorithms: An Overview", Journal of Physics: Conference Series, 2018
[11] Fajr Ibrahem Alarsan., and Mamoon Younes ‘Analysis and classification of heart diseases using heartbeat features and machine learning algorithms’,Journal Of Big Data,2019;6:81.
[12] Internet source [Online].Available (Accessed on May 1 2020): http://acadpubl.eu/ap
.