HEART DISESASE PREDICTION USING LOGISTIC REGRESSION

[1]MR. N. THIRUMALA RAO, [2]CHILUVERU TARUN, [3]CHRISTINA RITHIKA, [4]MADIREDDY SRAVANI, [5]R. TEJKAMAL VARMA

[1](Assistant Professor) ,**ECE.** J.B. Institute Of Engineering & Technology

[2345]B,tech scholar ,**ECE.** J.B. Institute Of Engineering & Technology

ABSTRACT

Heart disease remains a significant global health concern, prompting the exploration of predictive models to aid in early detection and prevention. Machine learning, particularly logistic regression, offers a promising avenue for predicting the likelihood of heart disease based on various risk factors and patient characteristics. This abstract presents a summary of a study utilizing logistic regression to predict heart disease The study involves the collection of a comprehensive dataset comprising demographic information, medical history, lifestyle factors, and clinical indicators from a diverse population. Feature engineering techniques are employed to preprocess the data, including handling missing values, normalization, and encoding categorical variables Logistic regression, a widely used classification algorithm, is implemented to build a predictive model. The dataset is split into training and testing sets to evaluate the model's performance. Model evaluation metrics such as accuracy, precision, recall, and the area under the receiver operating characteristic curve (AUC-ROC) are utilized to assess the model's effectiveness in predicting heart disease. The results demonstrate the logistic regression model's capability to effectively predict the likelihood of heart disease based on the selected features. Moreover, feature importance analysis is conducted to identify the key factors contributing to the prediction, aiding in understanding the underlying relationships between risk factors and the presence of heart disease. The implications of this research are substantial, as accurate prediction models can assist healthcare professionals in early identification and intervention, potentially reducing the burden of heart disease by enabling targeted preventive measures. However, further validation on larger and more diverse datasets is essential to enhance the model's robustness and generalizability. In conclusion, the application of logistic regression in predicting heart disease presents a valuable approach in leveraging machine learning for proactive healthcare management. Continued refinement and validation of predictive models hold promise for improving early detection and personalized interventions to combat heart disease effectively.

1. INTRODUCTION

In the modern world, cardiovascular disease, often known as heart disease, is a severe sickness. spreading globally. For the past several years, the prevalence of heart disease has been rising quickly in our daily lives. Smoking, having high blood pressure, and having high cholesterol are three major risk factors for heart disease. “Women” typically exhibit distinct heart disease symptoms than men do, particularly when it comes to coronary artery disease (CAD) and other cardiovascular issues. A set of illnesses that impact the circulatory system are referred to as heart disease. There are a unique set of causes for each form of heart disease. The misbehaviour risk factors for myocardial infarction and brain attacks include poor eating habits, inactivity, tobacco use, and alcohol dependence. People may develop obesity, high blood lipids, high blood sugar, and high blood pressure as a result of behavioural risk factors. According to the primary care clinics, measurements of these "intermediate risk variables “include a higher probability of experiencing a myocardial infarction and “brain attacks”, congestive “heart failure”, or other problems. We must improve the performance of previous work architecture in this project using a preprocessing method that includes a normalization phase that fills in missing data with the mean value of each feature. A critical step in the machine learning process, data preparation improves the quality of the input data and extracts useful information.

The evaluation of various ML Algorithms for the identification of heart disease and the prediction of CVD are the main objectives of that work. Machine learning is a technique that enables a machine to learn without having to train it to do so explicitly. Artificial intelligence is a subfield that uses clever software to allow devices to perform tasks expertly. To overcome difficulties with classification, we employ logistic regression.

Instead of using linear regression, which denotes continuous progress, it does this by forecasting categorical outcomes. An example of a binomial, which has two possible results in the most straightforward instance, is the prediction of heart disease, which has been steadily rising in prevalence worldwide. When compared to the current method, using logistic regression increases Accuracy. Many academics are working to create various IoT-based medical gadgets as a result. Below are some of the researcher's findings.

It has been demonstrated that less sophisticated systems, like logistic regression and support vector machines with linear kernels, produce results that are more accurate than those of more complex systems. ROC curves and F1 ratings have been utilized as evaluation methods. This study's work analyses heart illness by applying machine learning methodologies to gauge the levels of disease severity. On the UCI heart disease dataset, experiments are conducted. The intention of this work is to help people better understand their conditions and to encourage them to seek professional care early when necessary. The study's foundation is publicly accessible medical data on heart disease. There are 208 entries in this dataset, and each contains eight details on the patient, including their age, type of chest pain, blood sugar level, blood pressure, heart rate, ECG, and more . This paper proposes a system to uses a logistic regression classification algorithm to classify the risk level. This paper will cover some of the most recent studies employing data mining approaches to forecast cardiac diseases to ascertain whether data mining methodologies are relevant and practical. Additionally, it will evaluate the various mining algorithm combinations used. The standard data set for heart disease include the “UCI Machine Learning Repository”.

There are 270 records total, some of which pertain to patients without cardiac disease and others to those who do. There are a total of 13 features in full, including “age”, “gender”, “chest pain”, “resting blood pressure level”, “cholesterol”, “fasting blood sugar”, resting “ECG” results, maximal “heart rate”, exercise-induced angina, old peak ST depression brought on by exercise relative to sleep, the slope of the peak exercise ST segment, and the number of significant vessels coloured by fluoroscopy. cardiovascular data, machine learning looks at how computers may learn (or enhance their performance) . This section displays the risk level drawn from the heart disease database. The cardiovascular disease database's patient clinical treatment data has undergone pretreatment to improve the mining process. Such proactive measures can avoid both the onset of sickness and the progression of the disease into a severe stage. As a result of the collection of various risk factors as a set of data, the data were then grouped into a number of risk factors that people are known to experience in their daily lives.

1.1 MOTIVATION OF 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 it’s detection. There are instruments available which can predict heart disease but either it are expensive or not efficient to calculate chance of heart disease in human. Early detection of cardiac disease 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 require more sapience, time and expertise. Since we have a good amount of data in today’s world, we can use various machine learning algorithms to analysis the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

2.LITERATURE SURVEY

In this paper, they calculate the accuracy of four different machine learning approaches and on the basis of calculation we conclude that which one is best among them. In this paper they have mentioned the introduction about the machine learning and heart diseases and described the machine learning classification. They illustrated the related work of researchers. This is about the methodology used for this prediction system. Authors have briefly described the dataset and their analysis with the result of this project. In this paper, the authors have described the demonstration algorithms like Support Vector Machine, Hoeffding Decision Tree, Logistic Model Tree (LMT), Naïve Bayes, Random Forest. And they concluded that Random Forest algorithm performs good predictions that can be understood easily. Five datasets are combined to develop a larger and more reliable dataset in this paper. Two selection techniques, Relief and LASSO, are utilized to extract the most relevant features based on rank values in medical references. This also helps to deal with over fitting and under fitting problems of machine learning. In this paper, heart disease prediction system is developed using various algorithms of Machine learning techniques. The approach followed is, the Nan values are replaced by the mean of the column. Due to which accuracy for the prediction gets improved. In this paper, they have used decision trees in predicting the accuracy of events related to heart disease.

They have also introduced Computer Aided Decision Support System (CADSS) in the field of medicine and research.. In this paper, they have used MLP perceptron; it provides the users with a prediction result that gives the state of a user leading to CAD. Due to the recent advancements in technology, the machine learning algorithms are evolved a lot and hence here they have used Multi Layered Perceptron (MLP) in the system because of its efficiency and accuracy. In this paper, two supervised data mining algorithm was applied on the dataset to predict the possibilities of having heart disease of a patient, were analysed with classification model. The classification technique is used for classifying the entire dataset into two categories namely yes and No [7]. In this system, the output consists of horizontal and vertical line splits based on the condition depends on the dependent variables. The accuracy level of this algorithm is quite higher than the other algorithms. The reason for the higher accuracy of this algorithm is these model analyses the dataset in the tree shape format.

3. METHODOLOGY

3.1 EXISTING SYSTEM

The before all existing system works on sets of both Deep learning and data mining. The existing system modules generates comprehensive report by implementing the strong prediction algorithm The main aims of the existing system to compare and check the before patient whose having disease outputs and new patient disease and determine future possibilities of the heart disease to a particular patient By Implementing the above mentioned model we will get the goal of developing a system with increased rate of accuracy of estimating the new patient getting heart attack percentage. The model which is proposed for Heart Attack Prediction System is invented for using Deep learning algorithms and approach. But by using all the existing systems the accuracy is very less.

3.1.1 DRAWBACKS OF EXISTING SYSTEM

Deep Learning, although touted as a revolutionary subset of Machine learning, its distinctive features can also result in its downside. Along with its advantages, deep learning is also known to cause inherent issues in implementation owing to technical issues. Let us also delve into the potential disadvantages of Deep learning in detail. Potential Disadvantages of Deep Learning

- Requires a Large Amount of Data Deep Learning's advantage of using massive data as its training dataset can cause a big advantage. A significant amount of High-quality data is required for the proper functioning of a deep learning model. This massive requirement demands a significant amount of time as well as resources for obtaining data.

- Extensive computing Needs This is one of the major disadvantages of Deep learning. For training a specific model with huge datasets necessitates more computing resources than other machine learning models.

Some of the examples are -Powerful central processors and graphics processing units, large amounts of storage and random-access memories, etc.

- Overfitting Tendencies One of the biggest disadvantages of Deep learning is its problem of Overfitting.

In the case of Overfitting, the model performs well on training data but comparably poor on unseen data. This may result in the model rendering irrelevant or incorrect answers. This further results in undermining automatic and transfer learning.

- Issues with Interpretation Another significant limitation of deep learning is that its models can be complicated to interpret or explain, unlike the case with traditional machine learning algorithms and models.

Some people may struggle to comprehend the operating mechanism of the model or its decision-making processes.

3.2 PROPOSED SYSTEM

This proposed system have a data which classified if patients have heart disease or not according to features in it. This proposed system can try to use this data to create a model which tries predict (reading data and data Exploration) if a patient has this disease or not. In this proposed system,we use logistic regression (classification) algorithm. By using sckit learn library to calculate score. Implements Navie Bayes algorithm to getting accuracy result. Finally analysing the results by the help of Comparing Models and Confusion Matrix. From the data we are having, it should be classified into different structured data based on the features of the patient heart. From the availability of the data, we must create a model which predicts the patient disease using logistic regression algorithm. First, we have to import the datasets. Read the datasets, the data should contain different variables like age, gender, sex, cp(chest pain),slope, target. The data should be explored so that the information is verified. Create a temporary variable and also build a model for logistic regression. Here, we use sigmoid function which helps in the graphical representation of the classified data. By using logistic regression, naïve bayes the accuracy rate increases.

COLLECTION OF DATASET

Initially, we collect 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

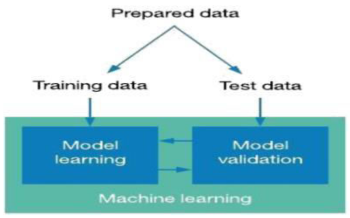


Figure 1 : Collection of Data

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.

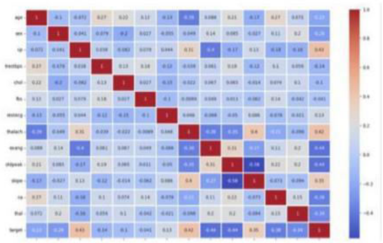


Figure 2 : Correlation matrix

4. WORKING OF SYSTEM

4.1 SYSTEM ARCHITECTURE

The system architecture gives an overview of the working of the system. The working of this system is described as follows: Dataset collection is collecting data which contains patient details. Attributes selection process selects the useful attributes for the prediction of heart disease. After identifying the available data resources, they are further selected, cleaned, made into the desired form. Different classification techniques as stated will be applied on pre-processed data to predict the accuracy of heart disease. Accuracy measure compares the accuracy of different classifiers.

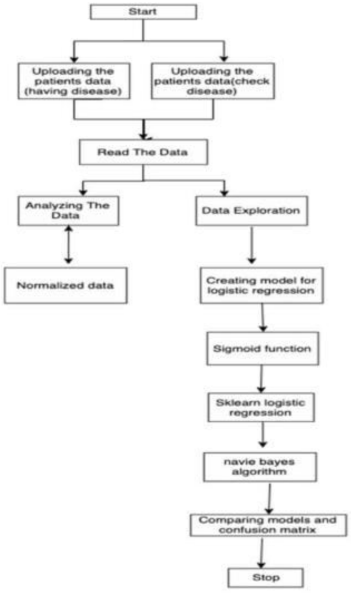
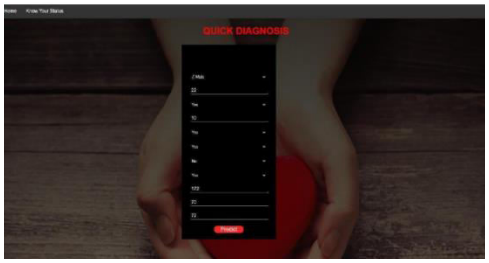
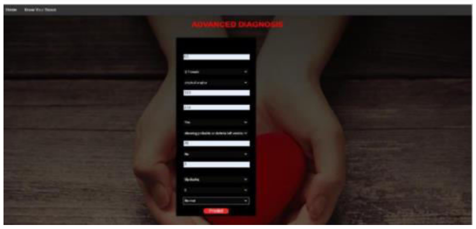

Start
Uploading the patients data (having disease)
Uploading the patients data(check disease)
Read The Data
Analyzing The Data
Data Exploration
Normalized data
Creating model for logistic regression
Sigmoid function
Sklearn logistic regression
navie bayes algorithm
Comparing models and confusion matrix
Stop

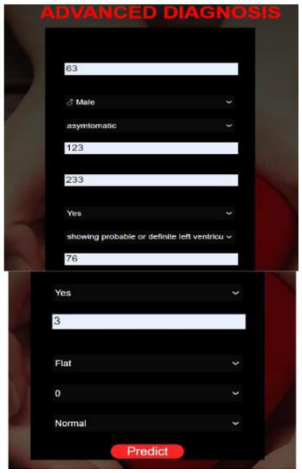
Figure 4.1 System architecture

5. INPUT AND OUTPUT

5.1 INPUT







ADVANCED DIAGNOSIS

63

♂ Male

asymtomatic

123

233

Yes

showing probable or definite left ventricu

76

Yes

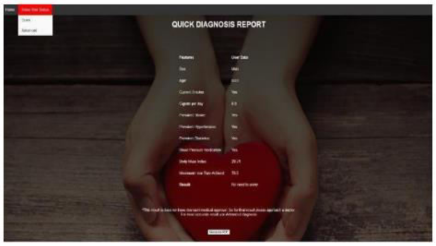
3

Flat

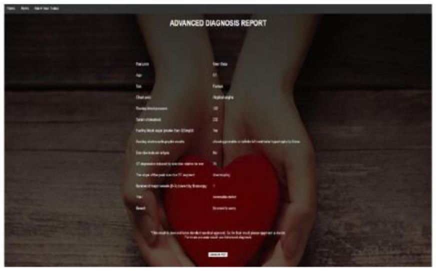
0

Normal

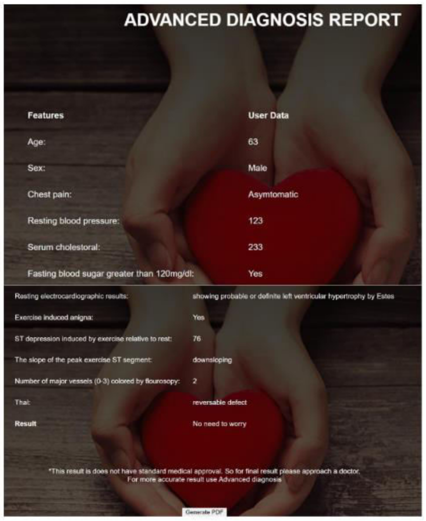
5.2 OUTPUT



QUICK DIAGNOSIS REPORT



ADVANCED DIAGNOSIS REPORT



ADVANCED DIAGNOSIS REPORT

Features	User Data
Age:	63
Sex:	Male
Chest pain:	Asymtomatic
Resting blood pressure:	123
Serum cholestoral:	233
Fasting blood sugar greater than 120mg/dl:	Yes
Resting electrocardiographic results:	showing probable or definite left ventricular hypertrophy by Estes
Exercise induced anigna:	Yes
ST depression induced by exercise relative to rest:	76
The slope of the peak exercise ST segment:	downsloping
Number of major vessels (0-3) colored by flourosopy:	2
Thal:	reversable defect
Result	No need to worry

*This result is does not have standard medical approval. So for final result please approach a doctor.
For more accurate result use Advanced diagnosis

6. RESULT

After performing the machine learning approach for training and testing we find that accuracy of the Logistic Regression 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 84% accuracy and the comparison is shown below.

TABLE : Accuracy comparison of algorithms Algorithm Accuracy

Algorithm	Accuracy
Logistic Regression	84.4%
Naïve Bayes'	77.9%
K-nearest Neighbour	75.3%
Decision Tree	OVERFITTING MODEL
Support vector machine	83.1%

7.CONCLUSION AND FUTURE WORK

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 utilization 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 SVM, Decision Tree, Naïve Bayes, Logistic Regression, K-nearest neighbour applied on the dataset. The expected attributes leading to heart disease in patients are available in the dataset which contains 76 features and 14 important features that are useful to evaluate the system are selected among them. If all the features taken into the consideration then the efficiency of the system the author gets is less. To increase efficiency, attribute selection is done. In this n features have to be selected for evaluating the model which gives more accuracy. The correlation of some features in the dataset is almost equal and so they are removed. If all the attributes present in the dataset are taken into account then the efficiency decreases considerably. All the five machine learning methods accuracies 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 five the logistic regression gives the highest accuracy of 84%.

8.REFERENCES

[1] Soni J, Ansari U, Sharma D & Soni S (2011). Predictive data mining for medical diagnosis: an overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-8

[2] Dangare C S & Apte S S (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-8.

[3] Ordonez C (2006). Association rule discovery with the train and test approach for heart disease prediction. IEEE Transactions on Information Technology in Biomedicine, 10(2), 334-43.

[4] Shinde R, Arjun S, Patil P & Waghmare J (2015). An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm. International

Journal of Computer Science and Information Technologies, 6(1), 637-9.

[5] Bashir S, Qamar U & Javed M Y (2014, November). An ensemble-based decision support framework for intelligent heart disease diagnosis. In International Conference on Information Society (i-Society 2014) (pp. 259-64). IEEE. ICCRDA 2020 IOP Conf. Series: Materials Science and Engineering 1022 (2021) 012072 IOP Publishing doi:10.1088/1757-899X/1022/1/012072 9

[6] Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). A coronary heart disease prediction model: the Korean Heart Study. BMJ open, 4(5), e005025.

[7] Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Ärnlöv J & Ingelsson E (2013). Multilocus genetic risk scores for coronary heart disease prediction. Arteriosclerosis, thrombosis, and vascular biology, 33(9), 2267-72.

[8] Jabbar M A, Deekshatulu B L & Chandra P (2013, March). Heart disease prediction using lazy associative classification. In 2013 International MutliConference on Automation, Computing,Communication, Control and Compressed Sensing (iMac4s) (pp. 40- 6). IEEE. 53

[9] Brown N, Young T, Gray D, Skene A M & Hampton J R (1997). Inpatient deaths from acute myocardial infarction, 1982-92: analysis of data in the Nottingham heart attack register. BMJ, 315(7101), 159-64.

10] Folsom A R, Prineas R J, Kaye S A & Soler J T (1989). Body fat distribution and selfreported prevalence of hypertension, heart attack, and other heart disease in older women. International journal of epidemiologyy, 18(2), 361-7.