International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Detecting Heart Disease Using a Supervised Decision Tree Classifier

Hardik J. Prajapati[1], Dr. Dushyantsinh B. Rathod[2]

Submitted:25/03/2024 **Revised**: 06/05/2024 **Accepted**: 17/05/2024

Abstract: The heart is vital to all living organisms and accurate diagnosis and prognosis of heart disease are crucial, as minor errors can result in severe consequences or loss of life. The incidence of heart-related deaths is increasing rapidly each day. To address this issue, an effective disease prediction system is essential. Machine learning, a branch of artificial intelligence (AI), offers significant support for predicting various events, including those triggered by natural disasters. In this study, we assess the ability of machine learning algorithms to predict heart disease. The algorithms evaluated include SVM (Support Vector Machine), Logistic Regression (LOR), Gaussian Naive Bayes (GNB) and Decision Tree, using UCI benchmark datasets for testing and training. Python, implemented through the Anaconda (Jupyter) notebook, is the preferred tool, offering an assortment of libraries and headers that enhance efficiency and precision.

Keywords: *supervised, confusion matrix, linear regression, unsupervised, python, reinforced*

1. Introduction

Given the heart's critical role as Among the majority vital systems in the human anatomy, it requires focused attention. Since many diseases are linked to heart health, it is essential to have accurate information for predicting such conditions. A comparative study in this area is crucial for this purpose. Numerous individuals today experience diseases that are diagnosed too late, frequently due to a lack of precision in diagnostic tools. Therefore, identifying the most valuable data for disease prediction is vital.

Machine learning, a highly effective testing method, is particularly relevant here. Artificial intelligence functions through testing and training processes. One of its subfields, known as machine learning, involves training machines to replicate human abilities. Such methods are taught to recognize and utilize data, which is why the term "artificial intelligence" is often associated with the integration of these technologies. Machine learning inherently learns from natural occurrences, and in this project, we utilize physiological parameters such as cholesterol levels, heart rate, biological sex, age, and more as test data to contrast the precision of various algorithms. Specifically, this project employs three algorithms: Gaussian Naive Bayes, Support Vector Machine and Logistic Regression.

The first section of this article offers an overview of artificial intelligence and heart-related issues. The second section delves into the Data Mining Algorithm. The third section reviews existing literature. The proposed architecture is discussed in fourth section. The fifth section briefly outlines the project's dataset and attributes. The final section concludes with a summary and a brief look into the future scope of the study.

2. Data Mining Algorithm

There are so many Data mining algorithms, But following algorithms are referred for research study.

2.1. Gaussian Naive Bayes (GNB)

GNB is a probabilistic machine learning classification technique. According to this approach, every characteristic or predictor makes an independent contribution to the outcome variable's prediction. Based on Bayes probability theorem, Naive Bayes is a kind of artificial intelligence function used for problems with classification. It is particularly effective in text classification, especially with large training datasets. Applications of Naive Bayes include emotion detection, spam filtering, and categorization of news articles. The method is renowned for its efficiency, allowing for quick predictions and model building. It was among the earliest methods employed to address text categorization problems.

2.2. Support Vector Machine (SVM)

The well-liked supervised learning technique SVM is typically employed for challenges involving machine learning classification, although it also performs well in regression and classification problems. Identifying the ideal line or decision boundary to divide an n-dimensional space into separate classes is the primary intent of the SVM technique. This separation will make it easier to classify fresh data points in the future. A hyper plane is the best decision boundary that SVM was capable to identify. Support vectors are selected using SVM in arrangement to build the hyper plane. The algorithm is named Support Vector Machine knowing that it uses these support vectors

[1] *PhD Scholar , Faculty of Engineering and Technology, Sankalchand Patel University, Visnagar, Gujarat*
[2] *Professor & Head, Ahmedabad Institute of Technology, Gota, Ahmedabad, Gujarat*
** Corresponding Author Email: hardikjp2707@gmail.com*

to figure out the hyper plane.

2.3. Decision Tree

A type of supervised learning method called decision trees can be applied to tasks combining regression as well as classification. They are frequently used in classification-related issues. In this tree-structured model, each internal node stands for a dataset property, every branch for a decision rule, and every leaf node for an outcome.

In a decision tree, the two main kinds of nodes are nodes that make decisions and leaf nodes. Leaf nodes show the outcomes of decisions made and don't extend further, whereas decision nodes are in charge of making decisions and may have several branches. The characteristics of the information set given serve as the basis for the choices or tests made inside the tree.

2.4. Logistic Regression (LOR)

Within the category of supervised learning, the LOR method is a well-known machine learning technique. It is used with a number of independent factors for forecasting a categorical dependent variable. Discrete numbers, such as true or false, yes or no, 0 or 1, etc., can be utilized in response to a logistic regression since the goal is to forecast a categorical dependent variable. The probabilistic values derived by logistic regression analysis fall between 1 and 0, instead of to exact binary numbers like 0 and 1.

3. Literature survey

Table 1 describe literature survey.

4. Proposed Architecture

Table 1. Literature Survey

No	*Paper Title*	*Authors*	*Publication*	*Related Work*
1	Heart Disease Prediction Using CNN Algorithm [1]	Varun Kumar, Umesh Devagade, Vinay Karanth, K. Rohitaksha, VirenViraj Shankar	Springer (2020)	With an accuracy of up to 85%, the convolutional neural network method analyses the peril of early heart disease using structured data. In addition, photos and unstructured data can be handled using the CNN technique.
2	Heart Disease Prediction System Using Classification Algorithms [2]	P. K. Gupta, Sarthak Vinayaka	Springer (2020)	They investigated the accuracy of numerous machine learning algorithms by applying them to the dataset to predict cardiac ailments. With an ultimate accuracy of 86.84%, they used a modified random forest approach. This method operates well in real-time, and by acquiring more data and including other CNN and deep learning algorithms, the system's accuracy could be substantially increased.
3	Heart Disease Prediction System Using Classification Algorithms [3]	Menaouer Brahami, Nada Matta , FatmaZahra Abdeldjouad	Springer (2020)	Effective classification of healthcare datasets has always been an important focus in machine learning. This study explored various classification functions, such as Logistic Regression, Adaptive Boosting, and Multi-Objective Evolutionary Fuzzy Classifier (MOEFC). When evaluated individually, Majority Voting achieved an accuracy of 80.20%, Logistic Regression had the lowest accuracy, and AdaBoostM1 yielded the highest accuracy.
4	Heart Disease Prediction Using Machine Learning Algorithms [4]	Rakesh Kumar, Archana Singh	IEEE 2020	In this study, the accuracy for the four different machine learning algorithms was examined; KNN delivered the best outcome, with an accuracy of 87%.
5	Robust Heart Disease Prediction: A Novel Approach based on Significant Feature and Ensemble learning Model [5]	Shamsheela Habib , Muhammad Affan Alim	IEEE 2020	In current study, we offer a novel machine learning technique as the foundation for our advanced cardiac disease prediction approach. Finding correlation-based features that improve prediction accuracy is its main objective. We utilise the UCI Vascular Cardiovascular Disease Dataset in our work, and we juxtapose our results with those of an earlier investigation. The accuracy of the model we recommended was 85.43%.
6	Heart Disease Prediction Model Based on Model Ensemble [6]	Xu Wenxin	IEEE 2020	The study established an original approach to predict heart disease using three different models: SVM, decision tree, and ANN, with an accuracy of 87%.
7	Prediction and Diagnosis of Heart Disease Patients using Data Mining Technique [7]	Shaicy P Shaji, Mamatha Alex P	IEEE 2019	The scheme aims to diagnose various cardiac diseases and to swiftly set preventive measures into place at an affordable cost. The method employs data mining addresses to forecast cardiac problems by feeding information into the Random Forest, SVM, and KNN classification methods. While KNN obtained an accuracy of 83 percent, SVM & random forest model both reached an accuracy of 85 percent.

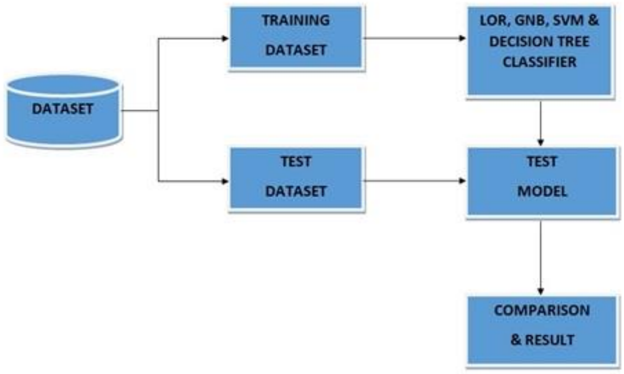


Fig 1. Proposed Architecture

How does the model work?

Figure 1 depicts the many processes involved in predicting heart disease.

1. The data gathering is the first step in the procedure, where various types of data (structured, semi-structured, or unstructured) are gathered from resources like hospitals.

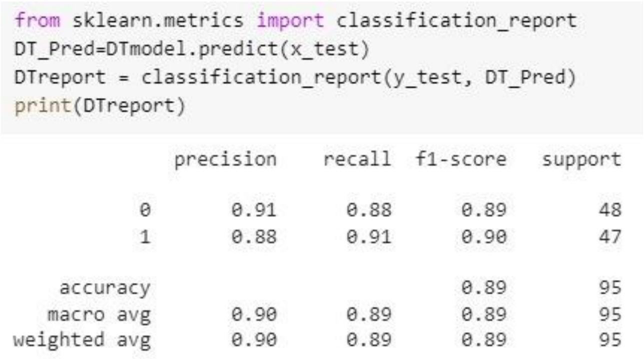
2. Following the acquisition of data, the data is cleaned to remove missing values and to consolidate it into a lower level of granularity. Next, the cleaned data is separated into test and training datasets.

3. When the data is divided, it is processed using SMOTE to address class imbalance and integrated into different machine learning algorithms, such as Logistic Regression (LOR), Gaussian Naive Bayes (GNB), SVM, and Decision Tree. This phase focuses on training the model to enhance its predictive accuracy using the training data.

4. After the model has received sufficient training, it is ready for testing.

5. The trained model is validated by assessing its performance on test in order information.

6. After achieving the desired level of predicted accuracy, The example is deployed.



```
from sklearn.metrics import classification_report
DT_Pred=DTmodel.predict(x_test)
DTreport = classification_report(y_test, DT_Pred)
print(DTreport)
```

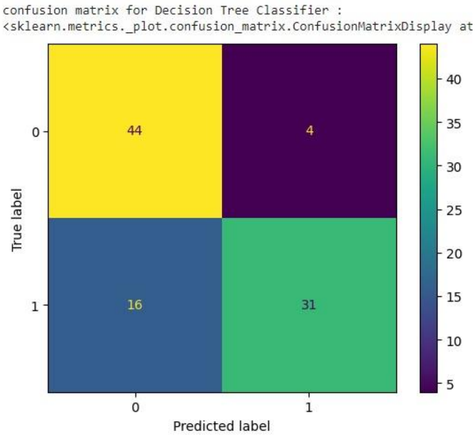
```
              precision    recall  f1-score   support

           0       0.91      0.88      0.89        48
           1       0.88      0.91      0.90        47

    accuracy                           0.89        95
   macro avg       0.90      0.89      0.89        95
weighted avg       0.90      0.89      0.89        95
```

Fig 2. Report of Decision Tree Classification (Recall, Precision, F1-score)

Figure 2 illustrates Decision Tree classification report. Decision Tree algorithm provided the accuracy 0.89. Its recall value is 0.89 and f1-score is 0.89.



```
confusion matrix for Decision Tree Classifier :
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at
```

Fig 3. Decision Tree Confusion Matrix

Figure 3 illustrates Decision Tree confusion matrix. The confusion matrix and group names are returned by the function. The heat map function may be used to visualize the confusion matrix.

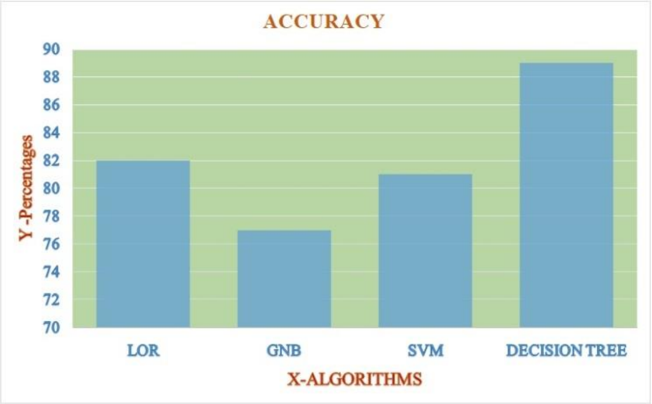


Fig 4. Accuracy Chart of ML algorithms

Table 2. Accuracy Table of ML algorithms

Algorithms	***Accuracy %***
LOR	82
GNB	77
SVM	81
DECISION TREE	89

Figure 4 illustrates Accuracy chart of ML algorithms. LOR provides 82% accuracy, GNB provides 77% accuracy, SVM provides 81% accuracy and Decision Tree provides 89% accuracy.

5. Dataset & Model

5.1. Hospital Data

In this paper, we examined medical datasets from hospital records stored in our database [1], encompassing a total of

14 features. They include fundamental patient data like sex, age and cholesterol levels, as well as structured data like laboratory findings, which are all important for diagnosing heart failure. Unstructured data, some of which is outlined in Table 3, is also considered for future exploration. Our focus has been on predicting heart disease risk with our model. The goal is to establish if a person has heart disease now or is at risk of getting it in the future. The model requires users to input values related to various patient attributes = (x1; x2 ;; xn). This data, which includes general, laboratory, and medical information, It is handled by the algorithm, leading to predictions which are more accurate compared to each of the other algorithms which have been studied.

5.2. Data Pre Processing

As expected, missing data which can come from numerous places, including human error can impair the accuracy of predicting. To maintain accuracy, it is essential to address this data loss. The model receives the data after any redundant features are eliminated and any missing attributes are filled in. This process is managed during the pre-processing phase, where the dataset is randomly divided into test & training sets. This division allows for the calculation of accuracy, It gets utilized in assessment the model's performance.

Table 3. Attributes of Dataset

Sr	Attribute	Description
1	Age	Age of the patient (25 to 75)
2	Sex	Patient's gender (female-1 male-0)
3	cp	type of chest pain (4 values)
4	trestbps	blood pressure at rest
5	ca	fluoroscopy coloration of a no. of significant vessels (0-3)
6	thalach	Angina caused by exercise
7	restecg	Attained maximal heart rate
8	oldpeak	Exercise-induced ST depression compared to rest
9	fbs	blood sugar levels after a fast > 120 mg/dl
10	target	0 = no disease, 1 = disease
11	Chol	cholesterol levels in the blood in mg/dl
12	slope	ST portion slope of the peak workout
13	thal	1 indicates normal, 2 indicates a permanent abnormality and 3 indicates a reversible defect.
14	exang	resting electrocardiographic results

6. Conclusion & Future Work

The study employed SVM, LOR, GNB, and Decision Tree classification models to assess the effectiveness of three supervised data mining methods in predicting the likelihood of a patient developing heart disease. All Algorithms have been investigated. on the same dataset to identify which method provided the most accurate results.- conclusion might elaborate on the importance of the work or suggest applications and extensions.

As a result, Logistic Regression achieved 82% accuracy, GNB attained 77% accuracy, SVM reached 81% accuracy, while Decision Tree classifier achieved the highest accuracy at 89% in predicting heart disease patients.

In the future, the developed system and An algorithm for machine learning classification could be applied to predict or diagnose various other diseases. Additionally, the automation of research on cardiac conditions could be refined or increased by using more machine learning techniques.

7. References

Author contributions

Hardik Prajapati: Role of Primary Author (a) Feasibility Study (b) Requirement Analysis from Health Stakeholder (c) Requirement Gathering from Hospital (d) Planning (e) Designing (f) Implementation / Coding [Python / Google Colab] (g) Testing/Validation

Dr. Dushyantsinh Rathod: Guidance about research problem, Documentations, Deadline Maintenance

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Varun Kumar, Umesh Devagade, Vinay Karanth, K. Rohitaksha, VirenViraj Shankar. Heart Disease Prediction Using CNN Algorithm Springer Nature Singapore Pte Ltd 2020.

[2] P. K. Gupta, Sarthak Vinayaka. Heart Disease Prediction System Using Classification Algorithms. Springer Nature Singapore Pte Ltd 2020.

[3] Menaouer Brahami, Nada Matta, Fatma Zahra Abdeldjouad, Heart Disease Prediction System Using Classification Algorithms. 18th International Conference, ICOST 2020.

[4] Rakesh Kumar, Archana Singh. Heart Disease Prediction Using Machine Learning Algorithms. 2020 IEEE.

[5] Shamsheela Habib, Muhammad Affan Alim. Robust Heart Disease Prediction: A Novel Approach based on Significant Feature and Ensemble learning Model

2020 IEEE.

[6] Xu Wenxin. Heart Disease Prediction Model Based on Model Ensemble. 2020 IEEE.

[7] Shaicy P Shaji, Mamatha Alex P . Prediction and Diagnosis of Heart Disease Patients using Data Mining Technique. 2019 IEEE.

[8] Saumya Yadav, Rajiv Rajan, Mohini Chakarverti. Classification Technique for Heart Disease Prediction in Data Mining. 2019 IEEE.

[9] Abhishek Kumar1, Pardeep Kumar, Ashutosh Srivastava, V. D. Ambeth Kumar,K. Vengatesan, Achintya Singhal. Comparative Analysis of Data Mining Techniques to Predict Heart Disease for Diabetic Patients. Springer Nature Singapore Pte Ltd 2020.

[10] Anbarasi, M., Anupriya, E., Iyengar, N.: Enhanced prediction of heart disease with feature subset selection using genetic algorithm. Int. J. Eng. Sci. Technol. 2(10),(2010)