Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.777577

CARDIOVASCULAR DISEASE DETECTION USING SUPERVISED DECISION TREE CLASSIFIER

Mr. Hardik J. Prajapati
Ph.d. Scholar, Faculty of Engineering and Technology, Computer Engineering Department Sankalchand Patel College of Engineering, Sankalchand Patel University, Visnagar, Gujarat, India, Hardikjp2707@gmail.com

Dr. Dushyantsinh B. Rathod
Professor& HOD, Computer Engineering Department, Ahmedabad Institute of Technology GotaOgnaj Road, Gota, Ahmedabad, Gujarat, India, Dushyantsinh.rathod@gmail.com

Abstract
The heart plays an important character in living things. Diagnosis & prognosis of heart disease needs greater completeness and accuracy because a small mistake can lead to extreme problems or loss of the person, there are many heart-related deaths and the number is expanding rapidly everyday . To solve this problem, a disease awareness prediction system is a key requirement. Machine learning is a type of AI (artificial intelligence). It provides outstanding support for the prediction of all types of events caused by natural disasters. In this article, we calculate the correctness of machine learning algorithms for heart-disease prediction, as these algorithms are SVM, LOR, GNB (Gaussian Naive Bayes)and Decision Treein using UCI benchmark data sets for training and testing. The best tool to implement Python programming is the Anaconda (Jupyter) notebook, which contains many kinds of libraries and header files that make the task crisp and efficient.
Key Words: supervised; reinforced; confusion matrix;linear regression; unsupervised; python

1 .Introduction

Because the heart is one of the biggest and most important systems of the human body, it requires special attention. Because most diseases are linked to the heart, it is necessary to know the most providing knowledge for disease prediction. A comparative research in this topic is essential for this reason.Today, most patients die because their diseases are discovered close to the deadline owing to a lack of accuracy in the instrument; as a result, it is critical to understand the most helpful information for disease prediction..One of the most successful testing methods is machine learning, which is based on training and testing. [4] Machine learning is a subset of artificial intelligence (AI), which is among the many learning areas in which robots mimic human abilities. Machine learning methods, on the other hand, are taught how to perceive and use data, leading to the term "artificial intelligence" becoming assigned to the combination of the two technologies.Machine learning, by definition, learns from natural phenomena and things, so in this project we use physiological parameters as test data, such as fats, heart rate, biological sex, age, and so on, and try comparing the algorithms' accuracy based on these, because we used three algorithms in this project: SVM, LOR, GNB (Gaussian Naive Bayes).The first section of this article provides an overview of machine learning and heart

problems. The Algorithm of Data Mining is discussed in second Section. The third section is a review of literature. The planned architecture is discussed in forth Section. The dataset and properties of this project are briefly described in Section fifth. The last section of this document's summary finishes with a brief glimpse into the document's future scope.

2. Data Mining Algorithm

Gaussian Naive Bayes (GNB)

Gaussian Naive Bayes (GNB) is a classification technique used in Machine Learning (ML) based on the probabilistic approach and Gaussian distribution. Gaussian Naive Bayes assumes that each parameter (also called features or predictors) has an independent capacity of predicting the output variable. For classification issues, Naive Bayes is a machine learning technique. It is based on Bayes' theory of probability. It is mostly used for text categorization with large training datasets. Emotion recognition, spam filtering, and news item classification are some examples.

It is well-known for its efficiency. The Naive Bayes method allows for rapid prediction and model creation. This is the first method that has been considered to tackle the problem of text categorization.

Support Machine Vector (SVM)

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane.

Decision Tree

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.The decisions or the test are performed on the basis of features of the given dataset.

Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.Logistic regression predicts the output of a

categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

3. Literature survey:

No	Paper Title	Authors	Publication	Related Work
1	**Heart Disease Prediction Using Data Mining Algorithm[1]**	Varun Kumar, UmeshDevagade, VinayKaranth, K. Rohitaksha, VirenViraj Shankar,	Springer (2020)	The Convolutional Neural Network method uses structured data to determine early heart disease risk.Their model can achieve an accuracy of up to 85 percent.The CNN method may also be applied to unstructured data and pictures.
2	**Cardiac Arrest Disease Prediction System Using Classification Algorithms[2]**	P. K. Gupta, SarthakVinayaka	Springer (2020)	They ran the dataset through multiple machine learning algorithms and tested accuracy by predicting cardiac disease. With the suggested modified random forest technique, they were able to attain a maximum accuracy of 86.84 percent. The suggested approach performs as well in real time, and the accuracy of the system may be improved by collecting additional data and using other deep learning &CNN techniques.
3	**A Hybrid Approach for Cardiac Disease Prediction Using Machine Learning Techniques[3]**	MenaouerBrahami, Nada Matta, FatmaZahraAbdeldjouad	Springer (2020)	Efficient categorization of healthcare datasets was and continues to be a key machine learning topic.This study looked at a variety of classification techniques, including Logistic Regression, Adaptive Boosting, and Multi-Objective Evolutionary Fuzzy Classifier (MOEFC). When used without ensemble, Majority Voting had the accurate results of 80.20, LR had the lowest accuracy, and AdaBoostM1 had the best accuracy.
4	**Cardiac Disease Diagnosis Using Machine Learning Algorithms[4]**	Rakesh Kumar, Archana Singh	IEEE 2020	The accuracy of four distinct machine learning algorithms was measured in this study, and KNN came out on top with an accuracy of 87 percent.
5	**Robust Cardiac Disease Diagnosis &Prediction:**	Shamsheela Habib , Muhammad AffanAlim	IEEE 2020	In this study, we suggest using a novel strategy for early cardiac disease prediction that includes machine learning algorithms. The paper's main goal is to uncover correlation-based characteristics

	A Novel Approach based on Significant Feature and Ensemble learning Model[5]			that can aid in producing reliable prediction outcomes. The UCI vascular heart disease dataset is utilised for this purpose, and our findings are compared to a previously published publication. The accuracy of our proposed model was 85.43 percent.
6	**Cardiac Disease Diagnosis &Prediction Model Based on Model Ensemble[6]**	XuWenxin	IEEE 2020	The study established a novel model orchestral composition heart disease prediction strategy that included three independent models (SVM, decision tree, and ANN) to obtain an accuracy of 87 percent.
7	**Prediction of Heart Disease Patients using Data Mining Technique[7]**	Shaicy P Shaji, Mamatha Alex P	IEEE 2019	The goal of this initiative is to detect various cardiac illnesses and take all necessary actions to avoid them at a reasonable charge as early as feasible.For the prediction of cardiac ailments, they use the 'Data mining' approach, in which characteristics are input into Random forest, SVM and KNN classification algorithms.SVM has an accuracy of 85 percent, Random forest has an accuracy of 85 percent, and KNN has an accuracy of 83 percent.
8	**Classification Technique for Cardiac Arrest Disease Prediction in Data Mining[8]**	SaumyaYadav, Rajiv Rajan, MohiniChakarverti,	IEEE 2019	Future potentials can be projected using the prediction analysis technique using the current data collection. For prediction analysis, an earlier SVM classifier is employed in this work. Because the KNN classifier uses the same number of hyper planes as the number of classes, it has a greater accuracy of 83 percent than the SVM classifier.

4.Proposed Architecture

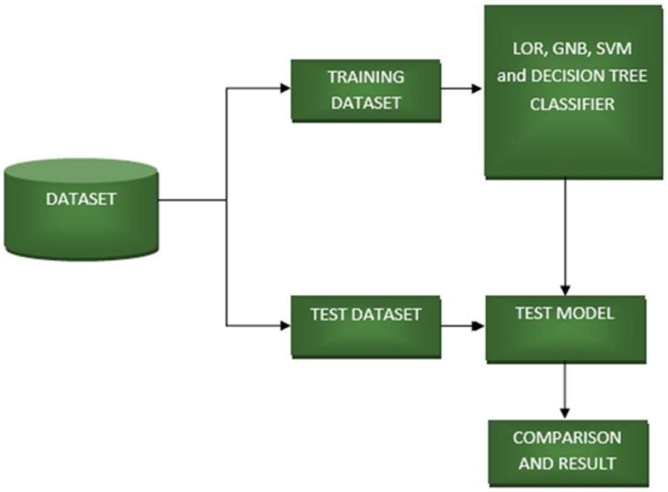

Fig 4.1 :Proposed heart disease prediction model[1]

How does the model work?

Figure 4.1 depicts the many processes involved in predicting heart disease.

1. It begins with data collection; in this paragraph, several forms of data, mostly structured, semi-structured, or unstructured, may be acquired from a variety of sources, including hospitals.
2. Once the data has been gathered, it is cleaned to eliminate missing values and report at a lower degree of granularity, and the clean data is then categorised into training and test data sets.
3. Following data separation, the data is fed into SMOT class Imbalance and a variety of machine learning techniques, including LOR, GNB, SVMand Decision Tree Technology. This stage primarily involves teaching the computer to improve its predicted accuracy by utilising training data.
4. Once our model has learnt enough from the data, it will be ready to be tested.
5. The learnt model is validated by putting it to the test with the test data.
6. The model is disseminated after the predicted accuracy reaches the specified level.

```
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 4.2 :Report of Decision TreeClassification (Recall,Precision,F1-score)

Figure 4.2 illustratesDecision Tree classification report. Decision Treealgorithm provided the accuracy 0.89. Its recall value is 0.89 and f1-score is 0.89.

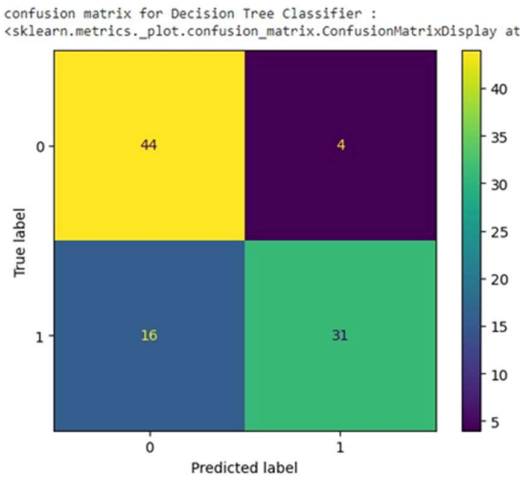

Fig 4.3 :Decision Tree Confusion Matrix

Figure 4.3 illustratesDecision Tree confusion matrix. The confusion matrix and group names are returned by the function. The heatmap function may be used tovisualise the confusion matrix.

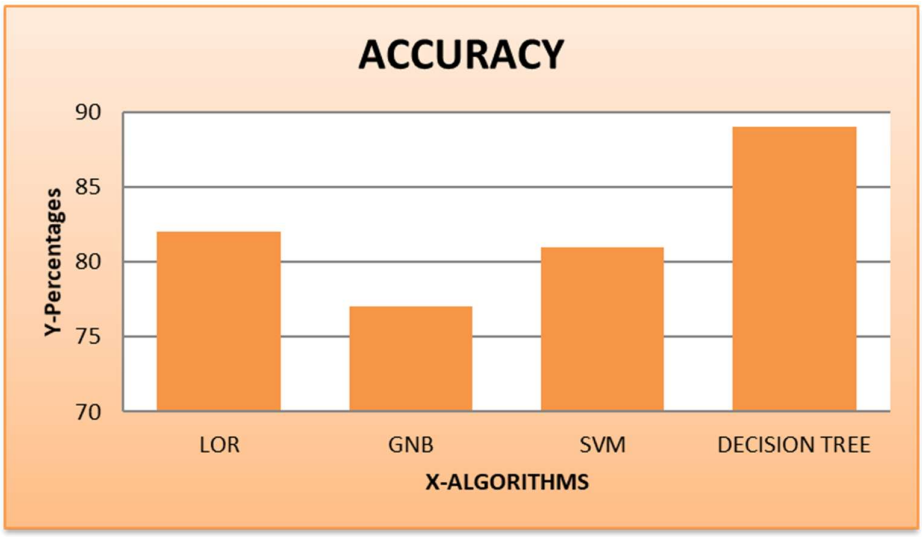

Fig 4.4 :Accuracy Chart of ML algoirthms

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

Table 4.1 :Accuracy Table of ML algorithms

Figure 4.4 illustrates Accuracy chart of ML algorithms. LORprovides 82% accuracy, GNB provides 77% accuracy, SVM provides 81% accuracy and Decision Tree provides 89% accuracy.

5. Dataset & Model

Hospital data

The medical datasets we looked at in this study are hospital records that are stored in our database [1]. There are 14 characteristics that are being processed in total. Laboratory data, as well as basic patient information such as age, sex, and cholesterol level, are among the structured data analysed. All of these information is required and essential in order to detect heart disease in patients. Unstructured data, some of which is included in table 5.1, might be viewed as a future scope.

Prediction of disease risk We have primarily concentrated on heart disease predicting in our model. The created model's goal is to forecast whether or not the individual has present or future heart disease. The model asks the user to enter values that are connected with different patient-relevant attribute values = (x1; x2; ;xn). This will include general, laboratory, and medical data, which will be transmitted to the algorithm, which will provide findings that are more accurate than the other algorithms analysed.

Data preprocessing

Forecasting accuracy will suffer as a result of missing data, as predicted. This data loss might be attributed to a variety of factors, including human mistake. As a result, we must compile it in order to preserve correctness. Missing attributes are filled and superfluous characteristics, if any, Before the data is fed into the model, it is deleted. This is handled during the preprocessing step, when the dataset is randomly split either training and test data to provide an accuracy value that may be used to assess the design.

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	flourosopy coloration of a number 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

Table 5.1. Attributes of Dataset

7. Conclusion& Future Work

- TheSVM, LOR, GNB (Gaussian Naive Bayes) and Decision Tree classification models were used to assess the outcomes of three supervised data mining technique used to a dataset in this study to predict the probability of a patient acquiring heart disease.To decide which method is the most accurate, all of these algorithms are performed on the same dataset.
- As a consequence, LORprovides 82% accuracy, GNB provides 77% accuracy, SVM provides 81% accuracy whileDecision Treeclassifier had an 89% accuracy level in predicting heart disease patients.
- In the future, the developed system and the machine learning classification algorithm might be utilised to predict or diagnose different illnesses. Other machine learning techniques might be utilised to expand or improve the work in heart disease research automation.

References

[1].Varun Kumar, UmeshDevagade, VinayKaranth, K. Rohitaksha, VirenViraj Shankar, “CardiacDisease Prediction Using Data Mining Algorithm” © Springer Nature Singapore Pte Ltd 2020.

[2]. P. K. Gupta, SarthakVinayaka, “Cardiac Arrest DiseasePrediction System Using Classification Algorithms” © Springer Nature Singapore Pte Ltd 2020.

[3].MenaouerBrahami, Nada Matta, Fatma Zahra Abdeldjouad, “A Hybrid Approach for CardiacDisease Prediction Using Machine Learning Techniques” 18th International Conference, ICOST 2020.

[4]. Rakesh Kumar, Archana Singh “CardiacDisease DiagnosisUsing Machine Learning Algorithms” ©2020 IEEE. .

[5]. Shamsheela Habib, Muhammad AffanAlim “Robust CardiacDisease Diagnosis &Prediction: A Novel Approach based on Significant Feature and Ensemble learning Model ” ©2020 IEEE. .

[6].XuWenxin “CardiacDisease Diagnosis &Prediction Model Based on Model Ensemble.” ©2020 IEEE.

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

[8].SaumyaYadav, Rajiv Rajan, MohiniChakarverti, "Classification Technique for Cardiac Arrest Disease Prediction in Data Mining" ©2019 IEEE.

[9].Abhishek Kumar1, Pardeep Kumar, Ashutosh Srivastava, V. D. AmbethKumar,K. Vengatesan, Achintya Singhal1 "Comparative Analysis of Data Mining Techniques to Predict Heart Disease for Diabetic Patients" © Springer Nature Singapore Pte Ltd 2020.
https://www.kaggle.com/ronitf/heart-disease-uci