Prediction on Cardiovascular disease using Decision tree and Naïve Bayes classifiers

Prediction on Cardiovascular disease using Decision tree and Naïve Bayes classifiers

V Sai Krishna Reddy[1], P Meghana[1], N V Subba Reddy[2], B Ashwath Rao[3]

[1]MTech Computer science and Engineering, MIT Manipal 576104, India
[2]Professor, Department of Computer Science, MIT Manipal 576104, India
[3]Assistant Professor (Selection Grade), Department of Computer Science, MIT Manipal 576104, India

Abstract. Machine Learning is an application of Artificial Intelligence where the method begins with observations on data. In the medical field, it is very important to make a correct decision within less time while treating a patient. Here ML techniques play a major role in predicting the disease by considering the vast amount of data that is produced by the healthcare field. In India, heart disease is the major cause of death. According to WHO, it can predict and prevent stroke by timely actions. In this paper, the study is useful to predict cardiovascular disease with better accuracy by applying ML techniques like Decision Tree and Naïve Bayes and also with the help of risk factors. The dataset that we considered is the Heart Failure Dataset which consists of 13 attributes. In the process of analyzing the performance of techniques, the collected data should be pre-processed. Later, it should follow by feature selection and reduction.

1. Introduction

An important concern in medical establishments like hospitals is providing good services to the patients at reasonable prices such that the right diagnoses of patients and proper choices can be made to prevent serious consequences that are extremely intolerable.

Cardiovascular diseases (CVDs) are one the kind of hazardous diseases which involves the chance of death if it is not dealt with carefully. It's troublesome to identify cardiovascular disease but thanks to many tributary risk factors like polygenic disorder, high signs, high cholesterol, abnormal pulse, and plenty of different factors. By using the previous medical history of the patient, the data about the risk factors can be easily collected and can be used to classify the patient record whether the patient is prone to heart disease or not. By using a doctor's expertise, heart disease can be predicted through some of the traditional measures in the medical field such as ECG tests or ECHO tests. But the prediction of disease using these tests will depend upon the doctor's expertise. Thus it is advisable to use a machine learning model to predict the diseases where the only main concern is choosing the right model that gives the best accuracy.

Machine learning involves laptops to urge trained using given information set and use this coaching to predict the properties of given new information. for instance, we are going to train a laptop by feeding it one thousand pictures of cats and one thousand additional pictures that are not of a cat, and tell whenever to laptop whether or not an image is a cat or not. If we tend to show the computer a replacement image, then from the higher than coaching, the laptop ought to be able to tell whether or not this new image is a cat or not. Training and prediction involve the use of specialized algorithms. We tend to feed the coaching information to AN algorithmic rule, and thus the algorithmic rule uses this coaching information to supply predictions on a replacement check information. Numerous techniques in machine learning are used to hunt out the proneness of cardiovascular disease among humans. In this paper, the severity of the disease is observed using given risk factors supported by numerous Machine learning techniques like Decision Trees (DT), Naive Bayes (NB).

This paper is organized as follows: Section II discusses the related work on predicting heart disease. Section III discusses the techniques that we used to classify the Heart Failure dataset and gives an overview of preprocessing done before classification. Section IV illustrates the results of both techniques. We conclude the work done in Section V and lastly, Section VI suggests the future work on the considered dataset.

2. Literature Survey

Some of the previous classification techniques used to diagnose heart disease are discussed in this section. Purushottam et.al. divided the heart disease dataset (HDD) into two partitions and evaluated the performance in each partition by implementing the decision rules for the modified dataset [1]. It can be feature selection and feature extraction [1]. Rishabh Saxena et.al implemented KNN and Decision tree algorithms by using HDD (Heart disease dataset) in which irrelevant attributes are reduced and thereby exploring issues in terms of time complexity and accuracy [2]. Feng-Jang-Jan et.al. tells that how Naïve Bayes classifier is used for different problem domains for classification, by explaining the probabilistic computation involved in the Naïve Bayes classifier [3]. Mariam Benllarch et.al. proposed two decision tree models like Very fast Decision tree which is related to Hoeffding bound that determines no of samples needed for best splitting at a node, Extremely Fast Decision tree is an improvement to VFDT where it validates the splitting at a node. These are implemented on Cleveland HDD Dataset and results show that EFDT has more accuracy than VFDT [4]. S.Manikandan et.al developed a web-based interface with 81.25% to predict early heart disease by using Naive Bayes Classifier [5]. Shadab Adam Pattekari et al. used the Naïve Bayes to diagnose heart risk. This system provides effective results for the prediction of heart disease [6]. Ali Haghpanah Jahromi et.al used Gaussian Naïve Bayes for the classification of 12 UCI datasets and compared them with some of the strong classifiers. Results show that performance is improved when compared to related work of other classifiers [7].

Chaitrali et.al. implemented and compared Neural Networks, Decision tree, Naïve Bayes for predicting heart failure by considering the Heart failure dataset and adding two more attributes to it like obesity and smoking and thus evaluated the performance of the three algorithms. Results show that the Neural network gives the best accuracy for the dataset with newly added attributes [8]. Kanak Saxena et al. did work on data discovery and implemented an efficient prediction system that generates decision rules to classify the record from the Cleveland Heart disease dataset. Results show that this system is more accurate than other machine learning algorithms for the considered dataset [9]. Data processing techniques are accustomed to obtaining meaningful data from the information. Classification Techniques employed in Experiment for result analysis and accuracy [10]. Ajit Solanki et.al carried a survey on some data mining algorithms and analyzed the potential of algorithms on predicting heart disease. Results show that of all the data mining techniques, classification algorithms are best suited for forecasting heart disease and Multilayer Perceptron achieved the best accuracy [11]. Dr.T.Karthikeyan et.al. analyzed some of the classification algorithms and tabulated the results for certain criteria like Accuracy, speed, Robustness, Scalability, Interpretability. It is said that each algorithm has its pros and cons, selecting the right algorithm depends on the dataset, time, duration [12].

Most of the previous work on the UCI dataset gives valuable results and still many improvements on classification techniques are being made to increase the accuracy of prediction. But during our survey, we observed that a dataset called ‘Heart Failure Dataset’ on which there is no previous work done. So, our motivation for this paper is to implement an effective and accurate heart disease diagnosis by using machine learning algorithms on Heart Failure Dataset.

3. Methodology

3.1 Dataset Description:

The dataset that we considered is taken from Kaggle [13], which is called Heart Failure dataset. It contains a total of 299 samples and 13 attributes. Out of 13 attributes, 6 attributes are binary attributes and the remaining 7 attributes are non-binary. The outcome attribute is DEATH_EVENT which determines the possibility of heart failure for a person. The detailed description of attributes in the dataset are shown in Table .1

Table 1. Heart Failure Dataset Attributes & Detailed Description.

Attributes	Description
age	Patients age
anaemia	If the patient has anaemia or not
creatinine_phosphokinase	To know the CPK enzyme level within the blood (mcg/L)
diabetes	whether the patient has diabetes or not (boolean)
ejection_fraction(EF)	At every contraction how much percentage of blood leaving the heart (percentage)
high_blood_pressure	Whether the patient has high B.P or not (boolean)
platelets	Amount of Platelets (kiloplatelets/mL) within the blood
serum_creatinine(SC)	To know the SC level (mg/dL) within the blood
serum_sodium(SS)	To know the SS level (mEq/L) within the blood
sex	Woman or man (binary)
smoking	Whether the patient has the smoking habit or not (boolean)
time	Follow-up period (days)
DEATH_EVENT	In the follow-up period, the patient is demised or not (boolean)

3.2 Data Pre-Processing

It is the most important process when we are considering the machine learning model. The data (299 values) which we considered is not always clean and formatted data so while performing any operation we need to clean it and put it in a formatted way. For considered datasets, data pre-processing is applied and then the next steps are followed as shown in Fig-1. The details of the dataset that was considered are shown in Table 1.

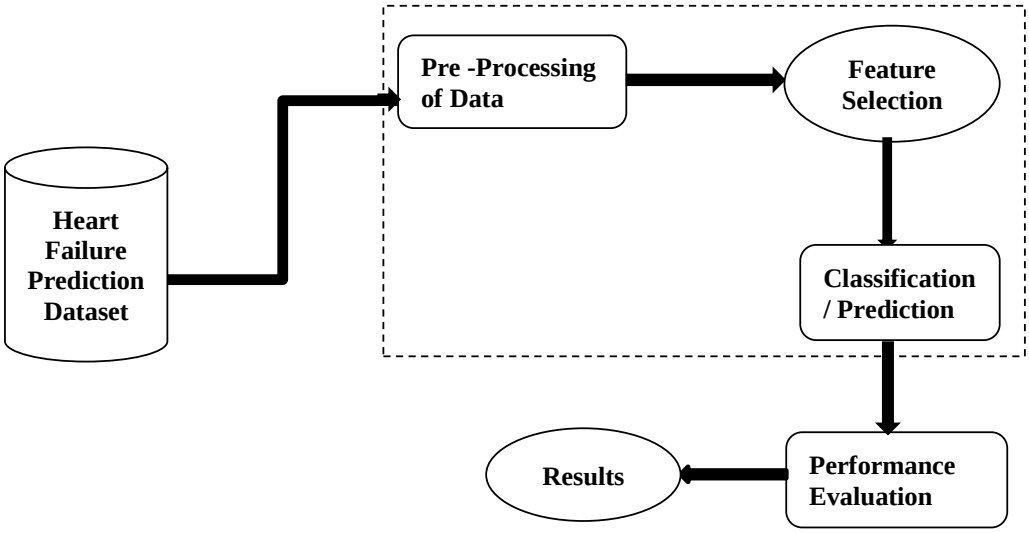

Fig - 1 Experiment workflow with the dataset

3.3 Feature Selection
Among the 13 attributes, Death-Event is the output binary attribute that we want to predict. So, from the remaining 12 attributes that we tend to the thought of from 2 attributes like age and sex area unit might not be a good factor to predict the health of the patient. The remaining attributes are vital as they help to observe the current medical records of patients and also to know the severity of the disease. So, we removed two attributes like age, sex and trained the model using the remaining 10 attributes.

3.4 Classification
Classification could be a technique that wants to assign data points to the gathering of target categories. The main goal of classification techniques is how accurate the algorithm can classify each and every input data into target class labels. In our case classification techniques are employed to predict if a person is prone to heart failure or not given some risk factors. Before applying these algorithms for classifications, out of 299 samples, we considered 200 as training data and 99 as test data.

3.4.1 Decision tree: When represented pictorially, the decision tree is a tree-like model which follows a top-down approach to assign the data to its actual target classes by the use of decision rules. It is a kind of supervised learning rule that's largely used for classification issues. Astonishingly, it works for each categorical and continuous dependent variable. During this rule, we tend to split the population into 2 or additional uniform sets. This is often done supported most important attributes/ freelance variables to create as distinct teams as attainable.

Entropy: It defines the amount of uncertainty present in Dataset D.
The formula for binary classification is given by

$$\text{Entropy} = -((p(0)) * \log(P(0)) + (p(1)) * \log(P(1))) \quad (1)$$

Information Gain: On which attribute should the next splitting takes place is determined by calculating Information Gain.

$$\text{Gain} = \text{Entropy}(D) - I(\text{Attribute}) \quad (2)$$

Step to implement decision tree algorithm is given by

1. First Find the Entropy for the whole dataset, say DATASET-ENTROPY(S)
2. For each attribute/feature:
    1. Find entropy for all other attributes ENTROPY (A) to get an output.
    2. Calculate the consider attribute Average information Entropy.
    3. Then find gain for the attribute which is considered.
3. Select the best gain attribute.
4. Repeat the above steps to get a suitable output.

Following is the code snippet that we used to implement the Decision tree, where entropy is used and max- depth is kept 5

1. It will find the entropy values.
    a. clf_entropy=DecisionTreeClassifier(criterion="entropy",random_state=1000,max_depth=5,min_samples_leaf=3)
    b. clf_entropy.fit(X_train,y_train)

2. With the entropy values it will give theaccuracy.
    c. y_pred_en=clf_entropy.predict(X_test)
    d. y_pred_en
    e. print(accuracy_score(y_test,y_pred_en))

3.4.2 Naïve Bayes. The main base of this classifier is Bayes Theorem which is computed using conditional probability. This classifier computes the probability of occurrence of the output feature given a set of input attribute values. Internally, the probability of every value with respect to the feature will be calculated. After calculating the probability of all the values of the output attribute, we consider the value which has maximum probability. The naïve Bayes model is extremely simple and fast for large datasets ad works well for binary classification. Bayes algorithm is given by

$$P(c|f)=P(f|c)*P(c)*P(f)^{-1} \quad (3)$$

p(c|f): The posterior probability of category (c, target) considering (f, attributes) already happened.
p(c): The prior probability of target c without any given condition
p(f|c): The likelihood of attributes when target c is given.
p(c): The marginal probability of target c.

After modifying the dataset, the Naive Bayes algorithm follows the process as:

1. Calculate probabilities so that it will lead to finding calculate class probability values.
2. After finding the probability values classifier will start predictions with input samples.
3. At last, it will get predictions of test data so that it will give the accuracy value.
4. This method is also helpful to find new instance outcomes with the help of the class with the highest posterior probability as the outcome of predictions.

Gaussian Naïve Bayes is one of the variants of the Naïve Bayes algorithm, which can be used when features of the given dataset are not in binary form. It is useful to model the probabilities of those features using Gaussian distribution. The dataset which is considered consists of some continuous data, so it is suitable for using the gaussian model.

If the data is continuous then introduce the Gaussian model to make predictions. It follows the process as:

1. In the gaussian model to learn data first find the mean and standard deviation.
2. Then using the below equation find the probability density function values.

$$P(x_i|y)= \frac{1}{\sqrt{2\pi\sigma_y^2}} * exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \quad (4)$$

3. After completing the above steps, it can plug the values with the suitable probability equation to find the output.

4. Performance Evaluation & Results

Obtain a confusion matrix (CM) to analyze the performance of a classification model. In machine learning, a confusion matrix is like a matrix that consists of 4 different values for binary classification which are used to compute performance measures like accuracy, recall, precision. In this study, the outcome attribute is a binary attribute in which 0 represents the person who has no chance of heart failure whereas 1 represents the person who is prone to heart failure.

Table 2. Confusion matrix for DT

	Predicted 0	Predicted 1
Actual 0	64	10
Actual 1	8	18

Table 3. Confusion matrix for GNB

	Predicted 0	Predicted 1
Actual 0	67	7
Actual 1	7	19

In Table 2, 64 and 18 are TN and TP values respectively which are correctly classified, and 8, 10 are FN and FP values respectively which are incorrectly classified. In Table 3, for Gaussian Naïve Bayes (GNB), 67 and 19 records are correctly classified into target classes but a total of 14 records are incorrectly classified. Accuracy, precision, recall for Decision tree and Naïve Bayes is computed from TP, TN, FP, FN values from the confusion matrix.

$$\text{Accuracy} = \frac{Number\ of\ correct\ predicions}{Total\ number\ of\ predictions} \quad (5)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

The accuracy values found are listed in the table below

Table 4. Accuracy Values

Classifier	Accuracy
Decision Tree	82%
Gaussian Naïve Bayes	86%

Table 5. Major values for performance evaluation

Classifier	Precision	Recall
Decision Tree	64.28	69.23
Gaussian Naïve Bayes	73.07	73.07

During the experiment, we also observed that if we increase the random states from 100 to 1000 the accuracy will increase. If we considered basic vital attributes from the dataset like B.P, Ejection- fraction, and smoking then the accuracy will increase.

5. Conclusion

In this paper, we picked two algorithms Decision Tree (DT), Naïve Bayes (NB), and implemented them on the Heart Failure dataset and, recorded the results. By comparing Table 4 & Table 5 in this experimental work, we have analyzed 2 classifiers in-depth namely Gaussian Naïve Bayes, Decision tree and we recorded the best accuracies for the considered dataset which are 86.0% and 82.0% respectively. Recognizing the problem with less time helps to save lives and precautions to be taken for avoiding heart disease with help of ML Techniques. The main aim of this paper provides an understanding of prediction models for the considered Heart Failure Dataset, which will help in the medical field using DT and GNB. We conclude that the Gaussian Naïve Bayes algorithm performs well on the Heart Failure Dataset with an accuracy of 86.0%.

6. Future Work

Most of the previous work is done on the UCI Heart disease dataset. Our work is done on the Heart failure dataset, which is a different dataset from the UCI Heart disease dataset. As the DT algorithm follows a rule-based approach, it helps predict cardiovascular disease. For prediction, it considers some set of rules and performs the final decision. When we need to predict the disease in the medical field by considering the assumption of independence then Naïve Bayes is a better algorithm. But this assumption of independence may not apply for some attributes and moreover, Naïve Bayes gives better results for relatively larger datasets. So for this Heart Failure Dataset, in the future, the study can be improved by merging a diverse mixture of algorithms to urge better accuracy. It can be developed for other dangerous diseases to detect problems early to make decisions.

References

[1] Purushottam, Kanak Saxena, and Richa Sharma 2015 Efficient Heart Disease Prediction System using decision tree *International Conference on Computing Communication and Automation* (ICCCA) pp. 72-77

[2] Rishabh Saxena, Aakriti Johri, Vikas Deep, and Purushottam Sharma 2019 Heart Diseases Prediction System Using CHC-TSS Evolutionary, KNN, and Decision Tree Classification Algorithm, © *Springer, Advances in Intelligent Systems and Computing*, p.p.809-819

[3] Yang F 2018 An Implementation of Naive Bayes Classifier *International Conference on Computational Science and Computational Intelligence* (CSCI) pp. 301-306.

[4] Benllarch M, El Hadaj S and Benhaddi M, 2019 Improve Extremely Fast Decision Tree Performance through Training Dataset Size for Early Prediction of Heart Diseases *2019 International Conference on Systems of Collaboration Big Data, Internet of Things & Security (SysCoBIoTS)* pp. 1-5

[5] Manikandan S 2017 Heart attack prediction system, *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)* pp.817-820

[6] Shadab Adam Pattekari and Asma Parveen 2012 Diagnose Heart Disease using Naïve Bayes Classifier *International Conference on Computing, Communication and Automation (ICCCA2015*)

[7] Jahromi and Ali Haghpanah 2017 A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features *Artificial Intelligence and Signal Processing Conference* (AISP)

[8] Chaitrali S. Dangare, Sulabha S. Apte, 2012 Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques *International Journal of Computer Applications* (0975 – 888) Volume **47**

[9] Kanak Saxena, Richa Sharma 2016 Efficient Heart Disease Prediction *Procedia Computer Science* Volume **85**, pp. 962-969

[10] Monika Gandhi, Shailendra Narayan Singh 2015 Predictions in Heart Disease Using Techniques of Data Mining *2015 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management (ABLAZE-2015)*

[11] Ajit Solanki, Mehul P. Barot 2019 Study of Heart Disease Diagnosis by Comparing Various Classification Algorithms *International Journal of Engineering and Advanced Technology (IJEAT)* ISSN: 2249 – 8958, Volume-**8**, Issue-2S2

[12] Dr.T.Karthikeyan, Dr.B.Ragavan and V.A.Kanimozhi 2016 A Study on Data mining Classification Algorithms in Heart Disease Prediction, *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* Vol. **5**, Issue 4, April 2016, ISSN: 2278 – 1323

[13] https://www.kaggle.com/andrewmvd/heart-failure-clinical-data last accessed on July 1st, 2021