Heart Failure Prediction using Data Mining

Pranav Sirodaria
School of Technology
Pandit Deendayal Energy University
Gandhinagar, Gujarat
pranav.sce20@sot.pdpu.ac.in

Kshitij Raj Burnwal
School of Technology
Pandit Deendayal Energy University
Gandhinagar, Gujarat
Kshitij.bce20@sot.pdpu.ac.in

Suyamoon Pathak
School of Technology
Pandit Deendayal Energy University
Gandhinagar, Gujarat
suyamoon.pce20@sot.pdpu.ac.in

Kkshitij Kapadia
School of Technology
Pandit Deendayal Energy University
Gandhinagar, Gujarat
Email: kkshitij.kce@sot.pdpu.ac.in

Abstract: Congestive heart failure has been one of the primary causes of death in the world today, estimated to take 17.9 million lives yearly,. The health care sectors gather enormous amounts of data that contain some hidden information and are useful for making effictive decisions. The use of data mining techniques like Logistic Regression, SVM, Naïve Bayes, and Decision trees to predict heart disease is only partially explored in a number of studies. Using machine learning techniques, we propose a novel approach in this paper that aims to identify significant features, thereby finding the accuracy of cardiovascular disease prediction. At the end, we compared our outputs with the main reference paper titled - Prediction of heart disease and classifiers' sensitivity analysis authored by Almustafa, K. M.

Keywords: heart failure, data mining techniques, logistic regression, naïve bayes, decision trees

I. Introduction

Because of numerous contributing risk factors, including diabetes, high blood pressure, high cholesterol, abnormal pulse rate, and many other factors, it is challenging to diagnose heart disease.The severity of heart disease in humans has been determined using various data mining and neural network techniques.

Data mining is the practice of getting useful insights out of numerous databases.Due to the nontrivial information found in large amounts of data, data mining is most beneficial in exploratory analyses.The process of extracting data in order to discover hidden patterns that can be transformed into significant is known as data mining.[1] Knowledge of data mining enables a user-oriented approach to new and hidden patterns in the data.

The suggested method may extract patterns and correlations related with heart disease from a historical heart disease database. The cleaveland dataset is the most used dataset by numerous researchers in their studies.

It can also respond to difficult questions about heart disease diagnosis, which can help medical professionals make wise clinical judgements. [2]

II. Methodology

In this Exploratory Data Analysis a variety of Model Classifications including Logistic Regression (LR), Support Vector Machine (SVM), Desicion Tree (DT), and Naive Bayes have been used .Various datasets have been tried upon and we have chosen the dataset displaying highest accuarcy across all 4 algorithms.

This study will examine the dataset on heart failure prediction from the processod dataset 'cleveland.csv'.A total of 302 records make up the dataset, which was split into two sets: a training set and a testing set.

After gathering various records, preprocessing of heart disease data occurs.There are 302 patient records in the dataset overall, 7 of which have some missing data. For the attributes of the given dataset, multiclass variables and binary classification are introduced.The presence or absence of heart disease is determined using the multi-class variable. If the patient has heart disease, the value is set to 1, otherwise it is set to 0 to indicate that the patient is heart disease-free. [3] Pre-processing of data is done by converting diagnosis values from medical records.The results of 297 patient records' pre-processing show that 137 of the records reflect a value of 1, indicating the presence of heart disease, while the other 160 records reflected a value of 0, indicating the absence of heart disease.

From among the 76 attributes of the data set, total of 14 attributes are used, mainly, age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, old peak, the slope of the peak exercise ST segment, number of major vessels flourosopy and defect along with the class attribute. [4][5]

III. DATA MINING CLASSIFIERS

i. LOGISTIC REGRESSION

When the output or dependent variable is categorical or dichotomous, logistic regression is used for classification problems.When using logistic regressions, there are a few presumptions to be aware of, including the numerous forms of logistic regression, the various kinds of independent variables, and the readily available training data. [6]

Probability is always between 0 and 1.In logistic regression, probability is computed using the sigmoid function.The logistic function is a straightforward S-shaped curve that transforms data into a value between 0 and 1.

$$p(x) = \sigma(t) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Table 1. Logistic Regression With Different Training Set

Method used	Train Accuracy	Test accuracy
20% training	56.67%	56.37%
40% training	65.28%	58.79%
60% training	63.53%	55.73%
80% training	60.74%	55.73%

ii. SUPPORT VECTOR MACHINE

Let the training samples having dataset Data = {yi, xi}; i = 1, 2, . . . , n where xi ∈ R n represent the i th vector and yi ∈ R n represent the target item. The linear SVM finds the optimal hyperplane of the form f (x) = w T x + b where w is a dimensional coefficient vector and b is an offset. This is done by solving the subsequent optimization problem:

$$\text{minimize } \frac{1}{n}\sum_{i=1}^{n}\zeta_i + \lambda\|\mathbf{w}\|^2$$

$$\text{subject to } y_i\left(\mathbf{w}^\mathsf{T}\mathbf{x}_i - b\right) \geq 1 - \zeta_i \text{ and } \zeta_i \geq 0, \text{ for all } i.$$

Table 2. SVM With Different Training Set

Method used	Train Accuracy	Test accuracy
20% training	65%	60.08%
40% training	71.07%	62.08%
60% training	71.27%	59.01%
80% training	66.94%	57.37%

iii. Naive Bayes

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods. Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

Table 3. Naïve Bayes With Different Training Set

Method used	Train Accuracy	Test accuracy
20% training	70%	59.67%
40% training	58.67%	64.28%
60% training	56.90%	63.93%
80% training	56.19%	63.93%

iv. DECISION TREES

For training samples of data D, the trees are constructed based on high entropy inputs. These trees are simple and fast constructed in a top-down recursive divide and conquer (DAC) approach. Tree pruning is performed to remove the irrelevant samples on D.

$$\mathrm{H}(X) := -\sum_{x \in \mathcal{X}} p(x) \log p(x) = \mathbb{E}[-\log p(X)],$$

Table 4. Decision Tree With Different Training Set

Method used	Train Accuracy	Test accuracy
20% training	100%	72.83%
40% training	100%	70.87%
60% training	100%	74.59%
80% training	100%	77.04%

IV. EXPLORATRY DATA ANALYSIS AND VISUALIZATION

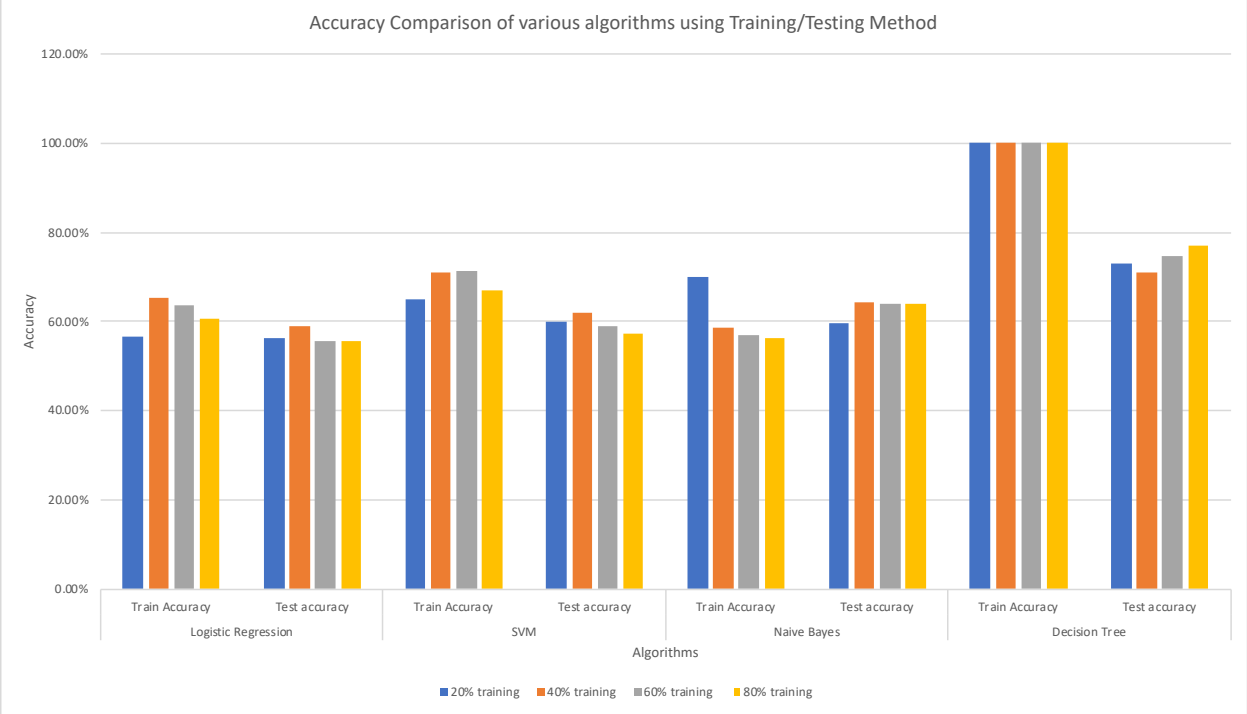

Fig.1. Accuracy Comparison of various algorithms for all the training testing split

In the above figure, the accuracy percentage have been compared in different algorithms that we have used in this study. All the models have been implemented for different variations of training and testing datasets.

All the graphs show a comparison of all 4 algorithms on the splits in training, testing of (80, 20) (60,40), (40,60) and (20,80) respectively.

The prediction models are developed using 14 features and the accuracy is calculated for modeling techniques. The best classification methods are given above in the graph. This table compares the accuracy of all the models for training and testing datasets. The highest accuracy is achieved by logistic regression method in comparison with existing methods.

V. Conclusion

Identifying the processing of raw healthcare data of heart information will help in the long term saving of human lives and early detection of abnormalities in heart conditions. In this paper, a comparative analysis of different classifiers was done for the classification of the Heart Disease dataset for positive and negative diagnosed participants. The algorithms were used are Logistic Regression, Decision Tree, Support Vector Machine and Naïve Bayes. The highest accuracy obtained is from Logistic Regression at 80% training and 20% testing which is equal to 83.95.

Table 5: Accuracy Comparison in reference paper[3]

Classifier used	Accuracy (Reference model)	Accuracy (Our Model)
SVM	84.1951	62.08
Naïve Bayes	83.122	64.28
Logistic Regression	-	58.79
Decision Tree	98.0488	77.04

VI. Acknowledgment:

Our group would like to thank our professor 'Rashmi Bhattad,' under whose guidance we were able to write this paper, our college Pandit Deendayal Energy University, which gave us this opportunity to participate and resources for this project. Now, a special thanks to every member this group who worked day and night in this project.

References:

[1] Bhajibhakare, M. M., Shaikh, N., & Patil, D. (2019). Heart disease prediction using machine learning. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, *7*(XII), 455-460.

[2] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, *7*, 81542-81554.

[3] Almustafa, K. M. (2020). Prediction of heart disease and classifiers' sensitivity analysis. *BMC bioinformatics*, *21*(1), 1-18.

[4] Wang, M., Yao, X., & Chen, Y. (2021). An imbalanced-data processing algorithm for the prediction of heart attack in stroke patients. *IEEE Access*, *9*, 25394-25404.

[5] Memon, B., & Ghulamani, S. (2022). A RELATIVE STUDY OF DIFFERENT MACHINE LEARNING CLASSIFICATION ALGORITHMS TO FORECAST THE HEART DISEASE. *Journal of Information Systems and Digital Technologies*, *4*(1), 11-27.

[6] Singh, P., Singh, S., & Pandi-Jain, G. S. (2018). Effective heart disease prediction system using data mining techniques. *International journal of nanomedicine*, *13*(T-NANO 2014 Abstracts), 121.

[7] Alkeshuosh, A. H., Moghadam, M. Z., Al Mansoori, I., & Abdar, M. (2017, September). Using PSO algorithm for producing best rules in diagnosis of heart disease. In *2017 international conference on computer and applications (ICCA)* (pp. 306-311). IEEE.