Identifying Significant Features for Heart Disease Prediction using Data Classification Techniques

Kota Baby Sujana Priya

M.Tech Scholar, Department of CSE, R.V.R & J.C College of Engineering, AP, India.

Abstract— Forecast of heart disease is seen as one of the hugest subjects in the territory of clinical data assessment. The proportion of data in the social protection industry is tremendous. Heart disease is one of the best purpose behind somberness and mortality among the quantity of occupants of the planet. There are some flow examinations that applied data mining techniques in the heart disease estimate. Regardless, thinks about that have given thought towards the basic features that expect a basic activity in predicting heart disease are limited. It is basic to pick the correct blend of tremendous features that can improve the introduction of the gauge models. This investigation intends to recognize basic features and data mining systems that can improve the accuracy of anticipating heart disease. Predictive models were made using different blends of features, and strategies: Naive Bayes, Logistic Regression (LR), and Vote.

Keywords: *Data mining, Heart disease prediction, Recursive Feature Elimination, Support Vector Machine, Naive Bayes, Logistic regression ,Majority Voting .*

1. INTRODUCTION

The utilization of data mining [2] carries another measurement to cardiovascular ailment expectation. Different information digging methods are utilized for recognizing and removing valuable data from the clinical dataset with insignificant client sources of info and endeavors. Over the previous decade, scientists investigated different approaches to actualize data mining in medical services so as to accomplish a precise expectation of cardiovascular infections. The productivity of data mining to a great extent fluctuates on the procedures utilized and the features chosen. The clinical datasets in the human services industry are excess and conflicting. It is more earnestly to utilize data mining [3] procedures without earlier and proper arrangements. As indicated by Kavitha and Kannan, information excess and irregularity in a crude dataset influence the anticipated result of the calculations. Accordingly, to apply the AI calculations to its maximum capacity, a powerful planning is expected to propose the data sets. Besides, undesirable features can diminish the presentation of the data mining procedures also. In this way, alongside information planning, a legitimate component choice strategy is expected to accomplish high precision in the heart disease forecast utilizing critical features and data mining strategies [7]. In spite of the fact that it has been very evident that highlight choice is as significant as the determination of a reasonable strategy, analysts are as yet battling in joining fitting data mining procedure with an appropriate arrangement of features.

Heart disease or cardiovascular disease remains the main source of death all through the world for as far back as decades. Heart illnesses are the number one reason for death all inclusive: a greater number of individuals lost their life's yearly from heart maladies than from some other causes. On the off chance that we can anticipate the cardiovascular disease and give cautioning already, a bunch of passing's can be forestalled.

2. RELATED WORK

Xiao Liu et al [8], said that Heart disease is one of the most widely recognized disease on the planet. The target of this examination is to help the analysis of heart disease utilizing a mixture order framework depends on the ReliefF and Rough Set (RFRS) technique. This framework contains two subsystems: the RFRS include determination framework and a grouping framework with a troupe classifier. The outcomes exhibit that the presentation of the framework is better than the exhibitions of detailed arrangement methods.

Kindie Biredagn Nahato et al [10], said that the accessibility of clinical datasets and data mining strategies, urges the scientists to seek after exploration in removing information from clinical data sets. In this work harsh set confusion connection strategy with back propagation neural network (RS-BPNN) is utilized. This work has two phases. The principal stage is treatment of missing qualities to get a smooth informational collection and determination of fitting properties from the clinical dataset by disjointedness connection technique. The subsequent stage is grouping utilizing back propagation neural system on the chose reducts of the dataset.

Kavitha et al [11], said that a high dimensional data set is used in the preprocessing stage of data mining process.

This raw dataset consist of disposable and inconsistent data, thereby enlarge the search space and storage house of the data. To achieve the classification accuracy, we need to remove the unwanted and the irrelevant data present. The dimensionality reduction technique is used to compress the high dimensional data to lower dimensional data with some constraints. An architecture is integrated for the elementary prediction of the heart disease. The system is created by using the principal component analysis (PCA) to take out the features and mathematical model is computed to select the relevant.

3. FRAMEWORK

This research focuses on finding the data mining techniques with significant features that will perform well in predicting heart disease. However, it is challenging to identify the proper technique and select the significant features. Existing investigations have demonstrated that data mining procedures utilized in the cardiovascular infection forecast are lacking, and a legitimate assessment is required to distinguish the critical features and data mining methods that will improve the exhibition. The heart disease datasets were gathered from the information source, UCI Machine Learning Repository. Cleveland dataset was chosen since it is a generally utilized database of AI specialists with records that are generally completed. An efficient feature selection technique has been developed using Support Vector Machine (SVM) Recursive Feature Elimination (RFE) for identifying the significant features. The methods Vote, Naïve Bayes and Logistic Regression were applied to make forecast models for this analysis utilizing the readied dataset.

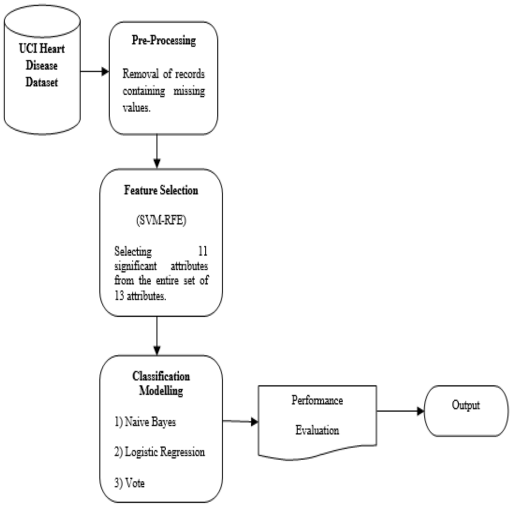
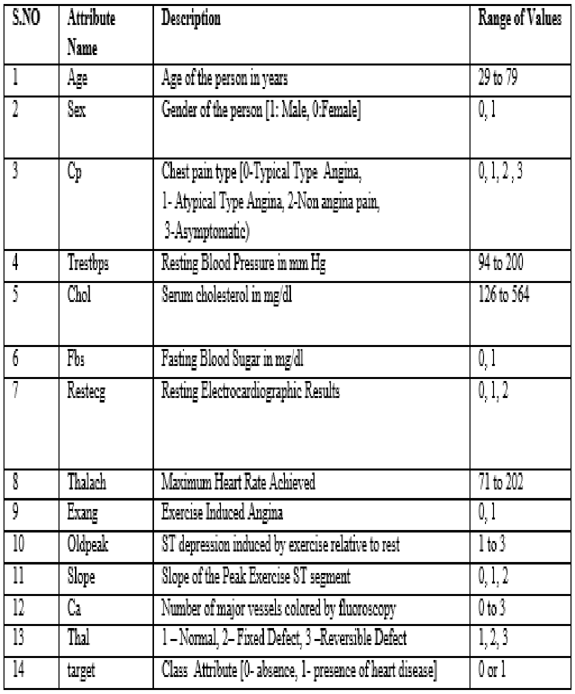

UCI Heart Disease Dataset
Pre-Processing
Removal of records containing missing values.
Feature Selection
(SVM-RFE)
Selecting 11 significant attributes from the entire set of 13 attributes.
Classification Modelling
1) Naive Bayes
2) Logistic Regression
3) Vote
Performance Evaluation
Output

Fig.1: System Architecture

DATASET DESCRIPTION:

A Cleveland heart disease dataset from the UCI machine learning repository has been used for the experiments. The dataset consists of 14 attributes and 1000 instances. There are 8 categorical attributes and 6 numeric attributes. The description of the dataset is shown in Table 1. Patients from age 29 to 79 have been selected in this dataset.

Table 1. Description of attributes from Cleveland Heart disease dataset.



S.NO	Attribute Name	Description	Range of Values
1	Age	Age of the person in years	29 to 79
2	Sex	Gender of the person [1: Male, 0:Female]	0, 1
3	Cp	Chest pain type [0-Typical Type Angina, 1- Atypical Type Angina, 2-Non angina pain, 3-Asymptomatic)	0, 1, 2 , 3
4	Trestbps	Resting Blood Pressure in mm Hg	94 to 200
5	Chol	Serum cholesterol in mg/dl	126 to 564
6	Fbs	Fasting Blood Sugar in mg/dl	0, 1
7	Restecg	Resting Electrocardiographic Results	0, 1, 2
8	Thalach	Maximum Heart Rate Achieved	71 to 202
9	Exang	Exercise Induced Angina	0, 1
10	Oldpeak	ST depression induced by exercise relative to rest	1 to 3
11	Slope	Slope of the Peak Exercise ST segment	0, 1, 2
12	Ca	Number of major vessels colored by fluoroscopy	0 to 3
13	Thal	1 – Normal, 2– Fixed Defect, 3 –Reversible Defect	1, 2, 3
14	target	Class Attribute [0- absence, 1- presence of heart disease]	0 or 1

Modules:

1. Data preprocessing:

Data preprocessing is to hold the missing data in the datasets. The Cleveland heart disease dataset contains six records that have missing values. All the records with missing qualities were expelled from the dataset, therefore diminishing the quantity of records. Next, the estimations of anticipated quality for the nearness of heart disease in the data set was changed from multi class values (0 for absence and 1, 2, 3, 4 for presence) to the double qualities (0 for absence; 1 for presence of heart disease).

2. Feature selection:

Feature selection suggests not just cardinality decrease, which means forcing a self-assertive or predefined cutoff on the quantity of properties that can be viewed as when constructing a model, yet additionally the selection of qualities, implying that either the expert or the demonstrating instrument effectively chooses or disposes of characteristics dependent on their handiness for examination.

3. Performance measure

The performance of the characterization models was estimated utilizing three execution measures: exactness, f-measure and accuracy. Precision is the level of accurately anticipated occurrences among all occasions. F-measure is the weighted mean of the accuracy and review. Accuracy is the level of right forecasts for the positive class.

ALGORITHMS:

SVM - RFE

The fundamental reason for SVM (Support Vector Machine)-RFE (Recursive Feature Elimination) [1] is to process the positioning loads for all features and sort the features as per weight vectors as the arrangement premise. SVM-RFE is an emphasis procedure of the retrogressive evacuation of features. It's a means for highlight set determination are appearing as follows. SVM-RFE's determination of capabilities can be for the most part separated into three stages, specifically, (1) the input of the datasets to be classified, (2) estimation of weight of each element, and (3) the cancellation of the element of least weight to acquire the positioning of features.

Naive Bayes:

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem [5]. It is not a single algorithm, but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

Logistic Regression:

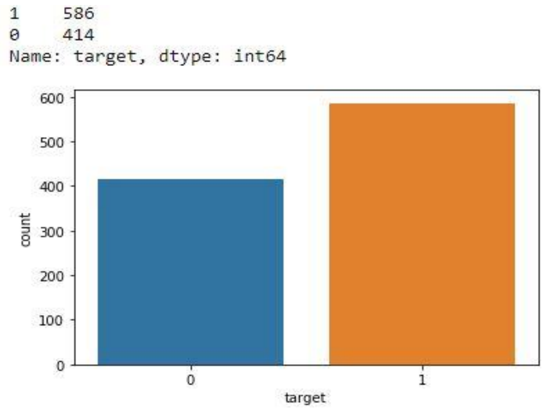
Logistic regression [4] is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of the target or dependent variable is dichotomous, which means there would be only two possible classes.

Vote:

The Vote technique [6] used in the proposed model is a hybrid technique that combines Naïve Bayes and Logistic Regression.

4. EXPERIMENTAL RESULTS

A heart disease dataset from the UCI machine learning repository, Total- 14 attributes (13 features and one target attribute). Records - 1000 The distribution of 1000 records for ‘target’ attributes resulted in 414 records for ‘0’(absence of disease) and 586 records for ‘1’(presence of disease). The classification models were developed using the three data mining techniques (i.e. Vote, Naive Bayes and Logistic Regression).

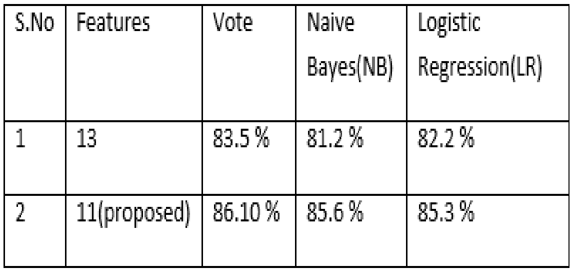


```
1    586
0    414
Name: target, dtype: int64
```

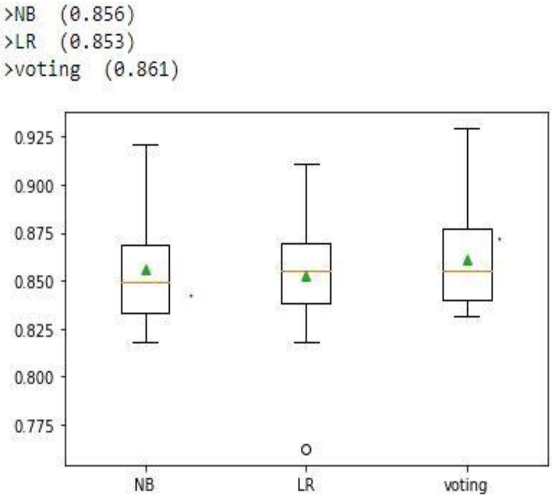
Fig.2: Disease prediction

An experiment was first conducted without using any feature selection techniques. Next, the SVM-RFE based feature selection are used to obtain the significant features. In light of the investigation results, eleven (sex, cp,trestbps,fbs,restecg,thalach,exang,oldpeak, slope, ca and thal) noteworthy features are obtained. A resulting table shows the accuracy obtained by each data mining technique.

Table 2: Resulting table



S.No	Features	Vote	Naive Bayes(NB)	Logistic Regression(LR)
1	13	83.5 %	81.2 %	82.2 %
2	11(proposed)	86.10 %	85.6 %	85.3 %



```
>NB  (0.856)
>LR  (0.853)
>voting  (0.861)
```

Fig.4: Performance chart

5. CONCLUSION

An investigation was directed utilizing the UCI Cleveland data set to distinguish the noteworthy features and the data mining techniques. For predicting heart disease , the significant features are obtained using feature selection. The data mining techniques that produce high exactness in prediction are recognized in this exploration as Naïve Bayes, Logistic Regression and Vote. Among the said three strategies, Vote has beat the other two techniques. In anticipation of heart disease, further enhancement can be done with a dynamic dataset with real time inputs.

ACKNOWLEDGEMENTS

The authors are greatly acknowledged to DST-FIST (Government of India) for funding to setting up the research computing facilities at RVR&JC College of Engineering.

REFERENCES

[1] SVM-RFE Based Feature Selection and Taguchi Parameters Optimization for Multiclass SVM Classifier Mei-Ling Huang,1 Yung-Hsiang Hung,1 W. M. Lee,2 R. K. Li,2 and Bo-Ru Jiang1.

[2] Bhatla, N., Jyoti, K., 2012. An analysis of heart disease prediction using different data mining techniques. Int. J. Eng. 1 (8), 1–4.

[3] Chaurasia, V., Pal, S., 2013. Early prediction of heart diseases using data mining techniques. Carib. J. SciTech. 1, 208–217.

[4] Predicting Heart Diseases In Logistic Regression Of Machine Learning Algorithms By Python Jupyterlab A. S. ThanujaNishadi.

[5] Heart disease prediction using Naïve Bayes Garima Singh1, Kiran Bagwe2, Shivani Shanbhag3, Shraddha Singh4, Sulochana Devi5.

[6] Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques C. Beulah Christalin Latha, S. CarolinJeeva.

[7] Khemphila, A., Boonjing, V., 2011. Heart disease classification using neural network and feature selection. In: 21st International Conference on Systems Engineerin(ICSEng). IEEE, Las Vegas, pp. 406–409.

[8] Liu, X., Wang, X., Su, Q., Zhang, M., Zhu, Y., Wang, Q., Wang, Q., 2017. A hybrid classification system for heart disease diagnosis based on the RFRS method. Comput. Math. Methods Med.

[9] Nahar, J., Imam, T., Tickle, K.S., Chen, Y.P.P., 2013. Computational intelligence for heart disease diagnosis: a medical knowledge driven approach. Expert Syst. Appl. 40 (1), 96–104.

[10] Nahato, K.B., Harichandran, K.N., Arputharaj, K., 2015. Knowledge mining from clinical datasets using rough sets and backpropagation neural network. Comput. Math. Methods Med. 2015, 1–13.

[11] An Efficient Framework for Heart Disease Classification using Feature Extraction and Feature Selection Technique in Data Mining R.Kavitha, E.Kannan.