IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, July 2022
Impact Factor 6.252

Analyzing Heart Disease Dataset using a Classification Algorithm

Mr. Yogesh Patil[1] and Dr. Priya Chandran[2]
Bharati Vidyapeeth's Institute of Management and Information Technology, Navi Mumbai, India

Abstract: *One of the most prevalent ailments today is heart disease, and for many healthcare professionals, early detection of this condition is essential to both protecting their patients from it and saving lives. For the classification of the Heart Disease dataset in this research, a comparison examination of several classifiers was carried out in order to accurately identify and forecast instances with few variables. This research shows the comparative study of classification algorithms on the Heart disease dataset.*

Keywords: KNN, Naïve Bayes algorithm, Classification, Heart disease prediction

I. INTRODUCTION

Heart disease is the main cause of demise worldwide. The target of the information extracting method is to extract the data from the dataset and makeover it into a clear construction for additional use. This is a diagnostic method planned to examine the information in seek of reliable patterns or organized associations connecting variables, and then to confirm the finding by applying the detected patterns.

When the heart cannot pump enough blood to meet the needs of the body's other organs, heart failure results. It is responsible for the annual death of approximately 17 million people worldwide [18]. With encouraging outcomes, a number of studies have been conducted to predict survival from heart failure. Therefore, the goal of this study is to deal with a class imbalance in the classification dataset and pick relevant predictive features in order of their ranking in order to improve the accuracy of earlier works on predicting heart failure survival.

II. LITERATURE REVIEW

Heart disease is the leading cause of demises in the world over the past 10 years. Extensive research is going on this area. Naïve bayes and weighted associative classifier is [12] to predict the chance of patient's heart attack. Ensemble based feature extraction methods were used to predict the probability of heart disease [13]. The authors used LDA and PCA in their research study. Heart disease is a term that assigns to a vast number of medical conditions related to heart. These medical conditions describe the unhealthy condition that directly affects the heart and all its parts. The healthcare industry gathers large amount of heart disease data which are not "mined" to discover hidden information for effective decision making.

Data mining is one of the crucial areas of research that is more popular in health organization. Data mining plays an effective role for finding new trends in healthcare organization which is helpful for all the parties associated with this field. The authors [4] used pattern recognition and data mining methods in predicting models in the domain of cardiovascular diagnoses. In [3], authors proposed a new approach for association rule mining based on sequence number and clustering transactional data set for heart disease predictions.

Takci [11] used twelve classification algorithms from various categories and four feature selection methods for heart attack prediction. The models were assessed based on the ROC analysis results, processing speed, and model correctness. Spencer et al. [14] conducted experiments on four frequently used heart disease datasets using four different feature selection techniques: principal component analysis, Chi-squared testing, Relief, and symmetrical uncertainty. The authors point out that the advantages of feature selection vary based on the machine learning strategy used for the cardiac datasets. Medical data mining has high potential for exploring the hidden patterns in the data sets of the medical domain [17]. These patterns can be utilized for clinical diagnosis for widely distributed in raw medical data which is heterogeneous in nature. This paper explores the utility of various Classification algorithms in classify and predict the disease.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, July 2022
Impact Factor 6.252

III. RESEARCH METHODOLOGY

3.1 Objective

The objective of the research is to predict possibility of heart attacks from the patient dataset using data mining techniques and determine which model gives the highest percentage of correct predictions for the diagnoses. This study uses the machine KNN and Naïve Bayes classification algorithm for the prediction. This study also compares both the machine learning algorithms based on the parameters used.

3.2 Methodology

The proposed study focuses on increasing classification accuracy by decreasing the number of features in a dataset and selecting a suitable classification algorithm for this type of dataset. The task of data mining in this research is to build models for prediction of the class based on selected attributes.

The flow of the process has been shown in the figure data pre-processing is done on the dataset then that data is processed using KNN and Naïve Bayes classification algorithms. The result of the both classification algorithm is compared and the algorithm is selected which is suitable for this type of dataset and gives best result.

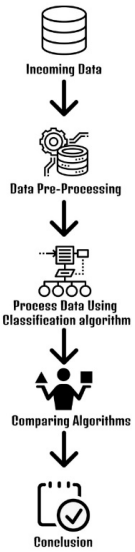

Figure 1: Flowchart of the study

A. KNN

K Nearest Neighbors algorithm stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique [15]. The three distance functions used in KNN are 'Euclidean', 'Manhattan' and 'Minkowski', where k – is the number of nearest neighbours, X – denotes predictor and Y is target to be predicted.

B. Functions in KNN

Euclidean	$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$
Manhattan	$\sum_{i=1}^{k}\lvert x_i - y_i\rvert$
Minkowski	$\left(\sum_{i=1}^{k}(\lvert x_i - y_i\rvert)^q\right)^{1/q}$

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, July 2022
Impact Factor 6.252

3.3 Dataset Table

Clinical feature	Description
Age	Age in year
Sex	Value 1:Male,value 0:Female
CP	Patient's CP Level
Trtbps	Patient's Trest BPS Level
Chol	Patient's Cholesterol Level
Fbs	Patient's FBS Level
Restecg	Patient's Resting ECG Levels
Thalachh	Patient's Thalach Levels
Exng	Patient's Exang Levels
Oldpeak	Patient's Old Peak History Recorded
Slp	Patient's Slope Levels
Caa	Patient's CA Levels
Thall	Patient's Thal Levels
Output	0 - Healthy Individual 1 - Heart-Disease Patient

3.4 Naïve Bayes Classification Algorithm

The Naive Bayes algorithm is a simple probability classifier that calculates a set of probability by counting the frequency and combinations of values in a given data set. The algorithm uses Bayes's theorem and assumes that all variables are independent considering the value of the class variable. This conditional independence assumption is rarely valid in real-world applications, so it is characterized as Naive, but the algorithm tends to learn quickly in a variety of controlled classification problems. [16].

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

where,
P(A|B) = Conditional probability of A given B.
P(B|A) = Conditional probability of B given A.
P(A) = Probability of event A.
P(B) = Probability of event B.

V. IMPLEMENTATION

The libraries “Class” and “caret” are used for KNN Algorithm and “e1071” is used for Naïve Bayes algorithm. The dataset is imported in RStudio and then using these libraries the corresponding algorithms are implemented on the dataset.

5.1 Dataset

In this paper, we use the heart disease data from Kaggle.com. The data contains fields related to different measuring units used in calculating the probability of a person for prediction of heart disease.
This dataset contains 304 records.

5.2 Preprocessing

The data has many irrelevant and missing records. To handle this data cleaning is done. It involves handling of missing data, noisy data. Below is the graphical representation of some Columns after completing the pre-processing. The graph represents Mean, Standard Deviation, Quantiles and validity of the data. Below figure 2.1, 2.2 and 2.3 shows the result after pre-processing of data of some columns of dataset. These figures contains Mean, Standard Deviation, Quantiles and Validity of data.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, July 2022
Impact Factor 6.252

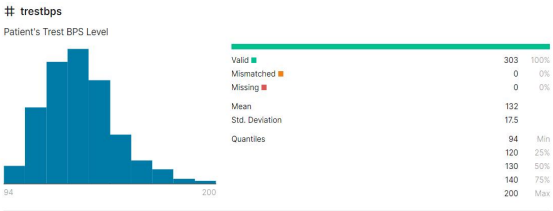

Figure 1: Trest BPS Level

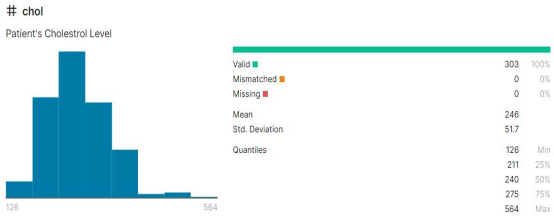

Figure 2: Patient's Cholesterol Level

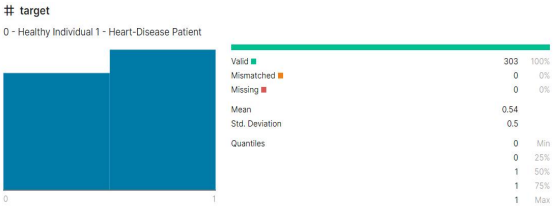

Figure 3: Final output

VI. RESULTS AND DISCUSSIONS

The results of different classification algorithms after training the model are in this section. In this analysis KNN and Naïve Bayes algorithm is used for doing the comparative study for Heart disease dataset. Below table shows the Classification algorithm and their accuracy on Heart disease dataset.

Classifiers	Accuracy(%)
KNN Algorithm	67.03
Naive Bayes Algorithm	78.94

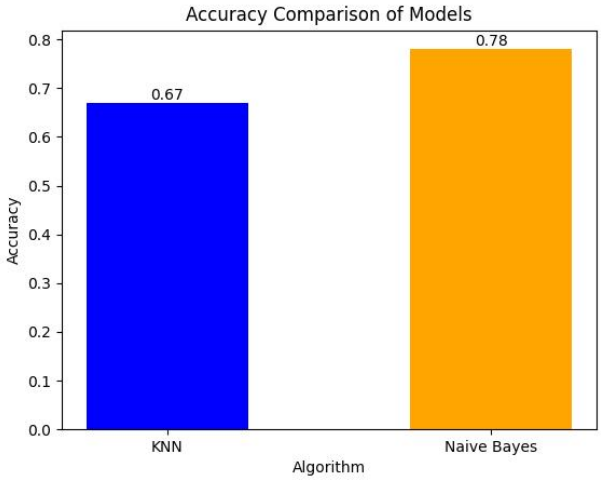

Figure 3 : Accuracy graph

Figure 3 shows the accuracy graph of final result of KNN and Naive Bayes algorithm.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, July 2022
Impact Factor 6.252

6.1 KNN ALGORITHM

Below figure shows the results after performing KNN algorithm on the dataset.

```
> table(knn.15,test.heart_labels)
     test.heart_labels
knn.15  0  1
     0 20 14
     1 16 41
> knn.15
 [1] 1 1 1 1 1 1 1 0 1 1 0 0 0 1 0 1 0 0 0 0 1 1 1 1 1 0 1 1 1 1 1 1 1 0 1 0 1 1 1 1
[39] 1 1 1 1 1 1 1 1 0 1 1 0 0 1 1 1 1 1 1 0 1 1 0 1 0 0 1 0 0 0 0 1 1 0 1 0 0 1 1 1
[77] 0 0 0 0 1 1 1 1 0 1 0 1 0 0 0
Levels: 0 1
```

```
> ACC.15
[1] 67.03297
```

From above figure we can conclude that the accuracy of the KNN algorithm is 67.03% and the misclassification rate for KNN algorithm is 32.97%.

6.2 Naïve Bayes Algorithm

For graphical representation we used “ggplot2” library. From Below figure we can conclude that the age group of 40 to 60 has the highest probability of getting heart diseases compared to age above 60.

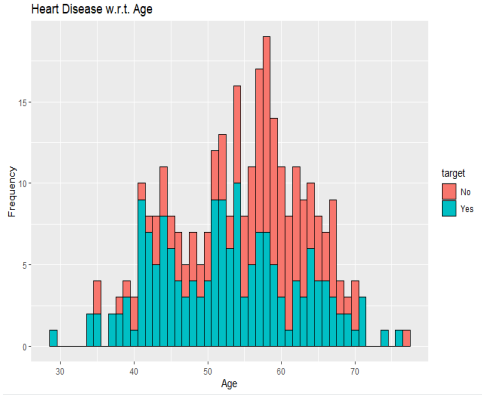

Figure 4 : Heart Disease w.r.t Age

```
> confusionMatrix(test$pred,test$target)
Confusion Matrix and Statistics

          Reference
Prediction No Yes
       No  34   0
       Yes  0  42

               Accuracy : 1
                 95% CI : (0.9526, 1)
    No Information Rate : 0.5526
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.4474
         Detection Rate : 0.4474
   Detection Prevalence : 0.4474
      Balanced Accuracy : 1.0000

       'Positive' Class : No

> acc_tt_nb<-(23+37)/76;acc_tt_nb
[1] 0.7894737
```

From above figure we can conclude that the model generated with help of Naive Bayes algorithm is 78.95% accurate or we can also say that the misclassification rate for Naive Bayes algorithm is 21.05%.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, July 2022
Impact Factor 6.252

VII. CONCLUSION

The study conducted an experiment to examine the best strategy for predicting heart attacks using several data mining methods. When employing several categorization methods in data mining, the research results show no significant differences in prediction. Data Mining's value in the medical domain is recognized, and measures are being attempted to use relevant approaches in disease prediction. Different persons conducted diverse research projects using certain successful strategies, which were studied. We compared two classification algorithms, KNN and Naive Bayes, in this study, and found that the Naive Bayes method produces the best results for the dataset and is more accurate for the heart disease prediction dataset.

REFERENCES

[1]. V.Manikandan and S.Latha, " Predicting the Analysis of Heart Disease Symptoms Using Medical Data Mining Methods" ,International Journal of Advanced Computer Theory and Engineering, Vol. 2, Issue. 2, 2013.

[2]. V. Chaurasia, "Early Prediction of Heart Diseases Using Data Mining," Caribbean Journal of Science and Technology, vol. 1, pp. 208–217, 2013..

[3]. M. Jabbar, P. Chandra, and B. Deekshatulu, "CLUSTER BASED ASSOCIATION RULE MINING FOR," Journal of Theoretical & Applied Information Technology, vol. 32, no. 2, pp. 196–201, 2011.

[4]. T. J. Peter and K. Somasundaram, "AN EMPIRICAL STUDY ON PREDICTION OF HEART DISEASE USING CLASSIFICATION DATA MINING TECHNIQUES," 2012.

[5]. S. B. Patil and Y. S. Kumaraswamy, "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction," International Journal of Computer Science and Network Security (IJCSNS), vol. 9, no. 2, pp. 228–235, 2009.

[6]. P. Chandra, M. . Jabbar, and B. . Deekshatulu, "Prediction of Risk Score for Heart Disease using Associative Classification and Hybrid Feature Subset Selection," in 12th International Conference on Intelligent Systems Design and Applications (ISDA), 2012, pp. 628– 634.

[7]. K. Srinivas, K. Raghavendra Kao, and A. Govardham, Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques," in The 5th International Conference on Computer Science & Education, 2010, pp. 1344– 1349.

[8]. C. Ordonez, "Association rule discovery with the train and test approach for heart disease prediction.," IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society, vol. 10, no. 2, pp. 334–43, Apr. 2006

[9]. D. S. Chaitrali and A. S. Sulabha, "A Data Mining Approach for Prediction of Heart Disease Using Neural Networks," International Journal of Computer Engineering & Technology (IJCET), vol. 3, no. 3, pp. 30–40, 2012.

[10]. R. Chitra and V. Seenivasagam, "REVIEW OF HEART DISEASE PREDICTION SYSTEM USING DATA MINING AND HYBRID INTELLIGENT TECHNIQUES," Journal on Soft Computing (ICTACT), vol. 3, no. 4, pp. 605–609, 2013.

[11]. H. Takci, "Improvement of heart attack prediction by the feature selection methods," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 26, pp. 1–10, 2018

[12]. N. A. Sundar, P. P. Latha, and M. R. Chandra, "PERFORMANCE ANALYSIS OF CLASSIFICATION DATA MINING TECHNIQUES OVER HEART DISEASE DATA BASE," International Journal of Engineering Science & Advanced Technology, vol. 2, no. 3, pp. 470– 478, 2012.

[13]. X.-Y. Gao, A. A. Ali, H. S. Hassan, and E. M. Anwar, "Improving the accuracy for analyzing heart diseases prediction based on the ensemble method," *Complexity*, vol. 2021, Article ID 6663455, 10 pages, 2021.

[14]. R. Spencer, F. Thabtah, N. Abdelhamid, and M. Thompson, "Exploring feature selection and classification methods for predicting heart disease," *Digital Health*, vol. 6, Article ID 2055207620914777, 2020.

[15]. K Nearest Neighbors - Classification – Data Mining Map https://www.saedsayad.com/k_nearest_neighbors.html.

[16]. Dimitoglou, G., Adams, J. A., & Jim, C. M. (2012). Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability. arXiv preprint arXiv:1206.1121.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, July 2022
Impact Factor 6.252

[17]. Predicting Diabetes by cosequencing the various Data Mining Classification Techniques P. Radha , Dr. B. Srinivasan, IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 6, August 2014.

[18]. Boutayeb, A., Boutayeb, S. The burden of non communicable diseases in developing countries. *Int J Equity Health* **4,** 2 (2005). https://doi.org/10.1186/1475-9276-4-2.