Classification and Prediction of Heart Disease: A Machine Learning Approach

Minal Chaudhari

Terna Engineering College Navi Mumbai, India

Rohini Patil

Terna Engineering College Navi Mumbai, India

Corresponding author: Minal Chaudhari, Email: minalchaudhari.comp@gmail.com

Heart condition refers to any condition affecting heart. There are many sorts, a number of which are preventable. Unlike disorder, which incorporates problems with the whole cardiovascular system, heart condition affects only the heart. During this research, the dataset is collected from UCI repository which brings together data from 4 other databases Cleveland, Hungary, Switzerland and long beach. Dataset contains 1025 patient with 14 attributes is employed with a set value. The results of proposed model are comparing with the previous model where the marginally changes in accuracy with feature selection attribute by using cfs evaluator using genetic search method. There the multi layer perceptron algorithm increases their accuracy 91.21% to 91.41%. In data preprocessing some outliers and extreme value within the dataset would be removed. The whole 869 instances were used for classification. An information gain evaluator method was performed on heart disease dataset that shows increasing the performance of classification accuracy. The naïve bayes, SVM, MLP, KNN, J48 gives the accuracy 81.93, 82.62, 91.59, 99.65, 98.84 respectively.

Keywords: Heart Disease Prediction, Machine Learning, Classification Algorithm, Feature Selection.

2021. In Rahul Srivastava & Aditya Kumar Singh Pundir (eds.), *New Frontiers in Communication and Intelligent Systems*, 151–156. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-95502-00-4-18

1 Introduction

Heart disease is additionally referred to as cardiovascular diseases are the one reason for globally death, 17.9 million lives every year in which 32% of deaths globally. CVD are a bunch of disorders of the heart and blood vessels including coronary heart condition, cerebrovascular disease, rheumatic heart disease, and other conditions. Out of four hearts related deaths are because of heart attacks and strokes. The term disorder might be wont to refer to heart conditions that specifically affect the blood vessels. Some people with heart condition have symptoms like this is often when there are changes or pain within the body to point out a disease.

The presented dataset contains 76 attributes including the class attribute, for 1025 patients collected, but in this paper, only a subset of 14 attributes is age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca thal, target.

2 Literature Survey

Mustafa's [1] prediction of heart condition and classification sensitivity analysis comparative analysis of various classifier was finished positive and negative diagnosed participants. Some extension and type of limitation to figure out here, differing types of classifiers are often in included within the analysis and sensitivity analysis are often performed on these classifications

Effective heart disease prediction [2] using hybrid machine learning techniques enhances the result accuracy 88.7% with hybrid random forest with linear model. The extension of this paper is extremely desirable.

Heart disease prediction using Machine Learning Techniques [3] system supported machine learning algorithm and techniques are very accurate in predicting the guts related disease. Plenty of scope of research work to be done a way to handle high dimensional data and over fitting. Analysis of supervised Machine Learning algorithms for heart condition prediction using Principal Component Analysis (PCA) was done in [4]. Moreover study using statistical and data processing tools within the diagnosis of heart condition was done in [5]. Enhanced cardiovascular disease prediction using Ensemble Learning method was given in [6].

The main objective of this paper is to analyze the heart patient by using various stages. They used the ensemble learning methods like boosting, bagging and voting.

3 Data Preprocessing

3.1 Outliers & Extreme values

An outlier is an observation that is unlike the other observations. Causes of outliers on a dataset:

1. Data Entry Errors: - Errors caused during data collection.
2. Measurement Error: - It is the most common source of outliers.
3. Experimental Errors: - Data extraction or executing errors.
4. Intentional: - Dummy outliers made to test detection method.
5. Data processing Errors: - Data manipulation / Data set unintended mutation.
6. Sampling errors.
7. Natural Outliers: - It is a natural outlier not artificial.

3.2 Feature Selection

By comparing with existing paper [1] on same dataset genetic search method gives best accuracy than best first search method in CFS feature selection technique. On the preprocessed dataset by removing outliers and extreme value the information gain feature selection technique gives the best attribute selection for J48 algorithm as shown in Table 2.

3.3 Classification Algorithm

1. **Naive Bayes:** Naïve Bayes algorithm is supervised learning algorithm which is used classification Bayes theorem.

2. **Support Vector Machine (SVM) [8]:** The goal of the SVM algorithm is to make the simplest line or decision boundary which will divide n-dimensional space into classes in order that we will easily put the info point within the correct category. This best decision boundary is named hyperplane[6].

3. **Multilayer Perceptron [10]:** The multilayer perceptron are networks of perceptron networks of linear classifiers. They will implement arbitrary decision boundaries using "Hidden layer". MLP forms the idea for the entire neural network and have greatly improved the facility of the computers when applied to classification and regression problems.

4. **K-Nearest Neighbors [7]:** The KNN Algorithm may be a simple and easy to implement supervised machine learning algorithm which will be solved both classification & regression problems. It extracts the knowledge supported the samples Euclidean distance function *d(xi, xj)* and therefore the majority of k-nearest neighbors.

5. **J48 [6]: -** The J48 Algorithm is simple classification algorithms to see the information categorically & continuously. J48 decision tree can cater lost or missing attributes estimations of the information and ranging attributes costs. Here, accuracy is often expanded by pruning.

4 Performance Measure

The performance measure gives results for classification of the heart disease data using various classifier algorithms. 10 folds cross validation method are used. A subset evaluator was used for feature selection of heart disease datasets feature to supply the proposed prediction models for various classifiers.

The Table2 shows the outcome generated by different classifier on the dataset using tool i.e., WEKA 3.9. Table1 indicated that without feature selection algorithm and data preprocessing using different classification algorithm [9] like Naïve Bayes (NB), SVM, KNN, J48, Multilayer Perceptron, with accuracy respectively.

Table 1. Result of various classification algorithms

Algorithm	**Accuracy (%) Without Feature Selection**
Naïve Bayes	83.12
SVM	84.19
KNN	99.70
J48	98.04
Multilayer Perceptron	95.51

Table 2. Comparison of Best First Search and Genetic Search method

Algorithm	**Accuracy (%) With Feature Selection CFS Evaluator (Best First Search)**	**Accuracy (%) With Feature Selection CFS Evaluator (Genetic Search)**
Naïve Bayes	82.82	83.02
SVM	83.31	83.21
KNN	100	100
J48	90.43	98.43
Multilayer Perceptron	91.21	91.41

The Table 2 shows the comparison of existing model (Best First Search method) with proposed model (Genetic Search Method).The data set used for this model is same i.e.,1025 instances. The accuracy of J48 and Multilayer Perceptron algorithm using genetic search method is 98.43 and 91.41.

Table 3. Classification algorithm using different feature selection methods

	Chi-square Attribute Evaluator Method	**Correlation Attribute Evaluator Method**	**Info Gain Attribute Evaluator Method**	**Cfs Evaluator (Genetic Search Method)**
Naïve Bayes	81.93	82.27	81.93	83.19
SVM	83.08	81.70	82.62	83.19
MLP	96.08	86.42	91.59	92.86
KNN	100	99.65	**99.65**	100
J48	98.73	97.92	**98.84**	98.96

The dataset by removing outliers and extreme value 869 instances is used for feature selection technique. By using feature selection method, the Information Gain filter method is best for the selection attributes which is less as compare to other. The selected attributes are 13, 12, 10, 8, 3, 9, 11, 5. The KNN algorithm gives more accuracy with less time i.e., 99.65 in 0 sec. as shown in Table 4.

Table 4. Statistical Analysis of various classifiers

Classifier	**Accuracy (%)**	**Kappa**	**RAE**	**ROC**	**MAE**	**Time (sec)**
Naïve Bayes	81.93	0.63	43.74	0.878	0.218	0.02
SVM	82.62	0.65	34.80	0.824	0.172	0.12
MLP	**91.59**	**0.83**	**22.08**	**0.944**	**0.11**	0.99
KNN	**99.65**	**0.99**	**0.81**	**0.993**	**0.0041**	**0**
J48	**98.84**	**0.97**	**2.69**	**0.999**	**0.01**	**0.1**

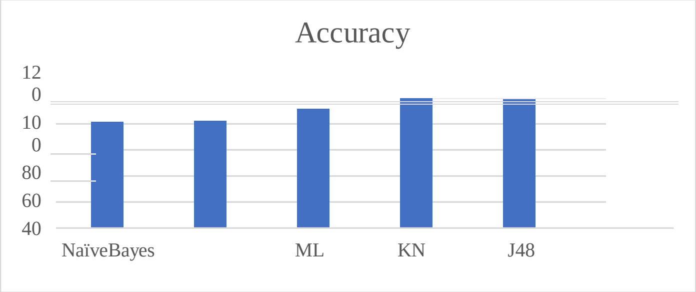

Fig 1. Visual representation of accuracy in Table 3

Table3 shows the kappa statistics, ROC (Receiver Operating Characteristic) curve and MAE output of the classifier and we can see the exceed classification of the KNN, Decision Tree J48, and Multilayer Perceptron MLP classifier. For KNN k is 1 kappa=0.99, ROC=0.993 and MAE=0.0041, for Decision Tree J48 kappa=0.97, ROC=0.99 and MAE=0.01, and kappa=0.83, ROC=0.944 and MAE=0.11 for Multilayer Perceptron MLP.

5 Conclusion

The paper consist results comparative analysis of various classifiers. The algorithm was used Naïve Bayes, SVM, Multilayer Perceptron, KNN, Decision Tree J48 and Random Forest. In Data preprocessing some outliers and extreme value within the dataset would be removed. The whole 869 instances were used for classification, a feature extraction method on the dataset by using information gain eval to gauge the classification performance. The Naïve Bayes, SVM, MLP, KNN, J48 gives the accuracy 81.93, 82.62, 91.59, 99.65, 98.84 respectively. The statistical measure kappa=0.99 for KNN classifier where k=1. Some decision trees have performed extremely well but decision trees have performed very poorly in other cases which might be because of over fitting. The benefit of this analysis by comparing different classification algorithm and methods on the dataset. Therefore, the advantage of having a valid and suitable feature extraction method for heart related disease prediction with minimum data rather than all available ones.

References

[1] K. M. Almustafa, “Prediction of heart disease and classifiers’ sensitivity analysis”, *BMC Bioinformatics*, vol. 21, Article number: 278, 2020.

[2] S. Mohan, C. Thirumalai and G. Srivastava, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques,” in *IEEE Access*, vol. 7, pp. 81542-81554, 2019.

[3] V. V. Ramalingm, A. Dandapath and M. K. Raja, “Heart disease prediction using machine learning techniques: A survey”, *International Journal of Engineering & Technology*, vol. 7, no.2, 684, 2020.

[4] S. Ardabili, A. Mosavi and A. R. Várkonyi-Kóczy, “Advances in Machine Learning Modeling Reviewing Hybrid and Ensemble Methods”, *Preprints*, 2019080203, 2019.

[5] M. Llamedo and J. PabloMart´ınez, “Heartbeat Classification Using Feature Selection Driven by Database Generalization Criteria”, *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 616-625, 2011.

[6] M. Fatima and M. Pasha, “Survey of machine learning algorithms for disease diagnostic”, *Journal of Intelligent Learning Systems and Applications*, vol. 9, no. 1, pp. 1-16, 2017.

[7] A. Malav, K. Kadam and P. Kamat, “Prediction of heart disease using k-means and artificial neural network as a hybrid approach to improve accuracy”, *International Journal of Engineering and Technology*, vol. 9, no. 4, pp. 3081-3085, 2017.

[8] S. Radhimeenakshi, “Classification and prediction of heart disease risk using data mining techniques of Support Vector Machine and Artificial Neural Network”, in *3rd International Conference on*

Computing for Sustainable Global Development, 2016, pp. 3107–3111.

[9] C. C. Aggarwal, *Data Classification: Algorithms and Applications*, CRC Press, 2014.

[10] J. S. Sonawane and D. R. Patil, "Prediction of heart disease using multilayer perceptron neural network", in *International Conference on Information Communication and Embedded Systems*, 2014, pp. 1-6.