Regd.No. 114/2017

International Journal of Nature Science (IJNS)

Estd:2024

Predictive Analytics to Earlier Diagnosis of Heart Disease Using Univariate Feature Selection – A Data Driven Approach

U.Ramya
Assistant Professor, Department of Information Technology, Sri Krishna Arts and Science College, Coimbatore-641105,Tamilnadu, India,
ramyau@skasc.ac.in

S. Saraswathi
Associate Professor and Dean, Department of Academic Affairs, Nehru Arts and Science College, Coimbatore- 641105,Tamilnadu, India.
nascsaraswathi@nehrucolleges.com

Corresponding Author:
ramyau@skasc.ac.in

Article History

Volume:1, Issue:4, 2024

Received: 27th September, 2024

Accepted: 8th October 2024

Published: 20th December, 2024.

Abstract: In the current scenario, Heart-Disease or Cardiovascular-Disease (CVD) is the threating devil to the globe and major threat to human's life after COVID. It is a type of disease refers to a group of conditions that affect the heart and blood vessels. It is a very large team since it encompasses various disorders, the most common and serious one is Coronary-Artery-Disease (CAD. The Machine learning algorithms and strategies plays a predominant role in analysing the various crucial heart disease that provides effective prediction results. This study highlights the insights in the performance of Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM) and XGBoost algorithm using feature selection algorithm called Univariate Feature Selection (UFS). The Cleveland Heart Disease (CHD) dataset is accessed from Kaggle and it is a multivariate dataset contains 303 instances with 14 features. UFS uses chi-square feature test to assess the relationship between each feature and target variables. The performance dominant features (7) were identified and used to build the model using the above-mentioned popular computing algorithmic approaches. By developing the model through UFS, the attained result proves that the prediction accuracy is high in LR and SVM with 90% among the four er models. This article includes the effort taken in Feature Selection and prediction of dominant inputs to build the model.

Keywords: *Coronary-Artery-Disease (CAD), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM) and XGBoost*

Authors citation: U.Ramya and S.Saraswathi., Predictive Analytics to Earlier Diagnosis of Heart Disease Using Univariate Feature Selection – A Data Driven Approach.Int.J.Nat.Sci.Vol.1(4). 2024.Pp:13-19.

1. Introduction:

The cardiac arrest is a serious medical emergency with high fatality rate. It occurs mainly when the heart suddenly stops pumping, that leads to loss of blood flow to the vital organs. If the heart valves are damaged, then it does not function effectively and the blood is allowed to leak which leads to acute congestive heart failure. The main defect is plaque builds up inside the coronary arteries, these coronary arteries are the blood vessels capable of supplying oxygen and needed nutrients to the heart muscle. This leads to narrowing or blocking of the arteries that reduces blood flow of the heart [1]. In these cases, if the blood supply to the heart is insufficient will lead to chest pain and in most of the cases patients experience heart attack or myocardial infarction. It causes death worldwide and the major risk factors of this type of disease may include blood pressure, high cholesterol, smoking, diabetes, obesity or family history of heart disease. Early detection and findings will prevent the complications and that gives best outcome for the people to overcome the disease. High quality medical care is the barrier for healthcare groups. In order to provide the best and utmost care to the patients firstly the patient's problem should be identified or examined correctly. In order to do this there are many technological advancements are emerged today which supports the primary prediction at the earliest and supports Healthcare Domain to apply corrective measures.

2. Literature survey:

The CVD is one of the major causes for death in the overall population of the world. In clinical data analysis, prediction of heart disease is considered as one of the most important aspects [2]. In data mining process, data preprocessing is one of the most important steps carried out to reduce the data redundancy and inconsistent data mainly to improve the classification accuracy [3]. In the medical field CVD is leading to fatal death. It is rapidly increasing due to obesity, cholesterol, blood pressure and usage of tobacco in many forms among the people. The CVD consists of CAD and CHD. There are many constraints in manual identification of heart disease, that leads to false predictions. In machine learning oriented research efforts have proved that more featured medical attributes can be used to predict so that it produces better model with greater accuracy and performance [4]. The author proposes several supervised learning techniques and compared the techniques such as Decision tree, Support vector machine, Random forest, Naive bayes, Logistic regression, Artificial neural networks, Logistic regression ,KNN and identifies the gaps in literature and structured the machine learning algorithm suffers from anyone of the limitations such as Dataset sample, Dataset distribution, Data-clinical correlation, feature selection, Hyper-parameter optimization, Meta heuristic techniques, Ensemble Techniques, Clinical aspects[5].There are many medical conditions that leads to CHD but diabetes is a common chronic disease worldwide. The authors statistically identified and analysed the features such as demographic information, laboratory indicators, medical examination and questionnaire and evaluated using XG boosting, Random Forest and Logistic Regression. By using the basic features identified feature selection is done and obtained the optimal feature subsets, then combined feature selection methods such as ANOVA and GINI and used machine learning algorithms and obtained the final feature set, thus leads to the generation of the final model [6]. The author conducted a metadata study that includes 451 papers from the year 2012 to 2021 and understood that researchers shown majority of their interest on SVM to predict CHD and the study aims to identify the ML based data driven approaches in identification of heart disease with the imbalanced data [7]. The authors used modelling techniques on the Cardiovascular dataset, which contains 70,000 records of patient information who got affected with coronary heart disease. Built a stacked model using 4 types of classifiers and identified the effective model with final accuracy of 75.1% [8]. Due to the increasing size of medical dataset, there is a major complexity in identification of the most important risk factors and understand the complex feature relationships to identify the disease. The study focuses on the process of identifying the risk factors that are most important for disease prediction. The authors performed correlation and interdependence of many different medical features for the heart disease prediction and applied a filter-based feature selection method and selected the most relevant features. If the number of features is reduced, the performance of the machine learning classification models improved with the reduced training time [9]. The predictions are made from the dataset obtained from UCI repository by using ML algorithm which contains 74 features and validated using six ML classifiers. The study proved the combination of chi-square with PCA (Principal Component analysis) produces greater accuracy in most of the classification algorithms [10]. The hybrid model builds the accuracy of 83% that is greater than the existing models [11]. The authors evaluated merged list of two feature sets and constructed a human-machine collaborative set and predicted the likelihood of 30 days readmission of patients who have Congestive heart failure(CHF) and achieved AUC above 0.8 and accuracy of 89% [12].The authors collected the heart disease prediction dataset from Cleveland repository that consists of 303 instances ,14 features and used for their investigation using Boruta feature selection method and acquired 88.2% accuracy for logistic regression model[13]. Thus, machine learning models that has revealed in this literature survey crops massive results that obliges as the unsurpassed decision support in the early finding of heart disease by compounding the prophetic tools into the EHR systems for risk accompanying with heart disease. This kind of prophetic techniques empowers the health care professionals to improve the primary findings and best action for the patients.

3. Proposed model:

The study focus on the implementation of UFS algorithm using Chi-Squared test to obtain the best features from the accessed dataset, a well-known for earlier predictions.

3.1 Feature selection

Feature selection is one of the predominant data preprocessing techniques to improve the efficiency and accuracy of machine learning prediction models. It is mainly applied to the high dimensionality dataset. The UFS algorithm is one of the suitable algorithms used to identify the best statistically important features. This method is mainly used to evaluate each feature individually and it will not consider the relationship between the features. Univariate method assesses each feature independently. This kind of approach is useful to identify the most relevant features by reducing dimensionality and complexity features. Some of the common techniques used in univariate feature selection method are ANOVA, t-test, Chi Square test, Information gain or Mutual information.

```
Begin
Step 1: Load the Cleveland dataset collected from Kaggle.
Step 2: Preprocess the dataset by handling missing values, encoding the categorical values and if needed scale the features.
Step 3: Define function for univariate feature selection method.
Step 4: Split the dataset into training and test dataset.
Step 5: Now apply the univariate feature selection on the training dataset.
Step 6: Extract the selected features from training and test dataset.
Step 7: Train a machine learning model on the selected features.
Step 8: Evaluate the model performance on the test dataset.
Step 9: Repeat the steps 5 to 8 with different parameters for comparison and optimization.
End
```

//Pseudocode 1: Univariate feature selection algorithm for heart disease prediction

The UFS method is a straightforward method to implement, efficiently handles high-dimensionality dataset by selecting relevant features based on the target variable. It has the advantage of reducing the risk of overfitting. It also selects a subset of features. This method is computationally less intensive when compared to the multivariate techniques. The UFS helps the healthcare professionals to focus on the clinically relevant factors to identify the important features that directly impact heart disease prediction.

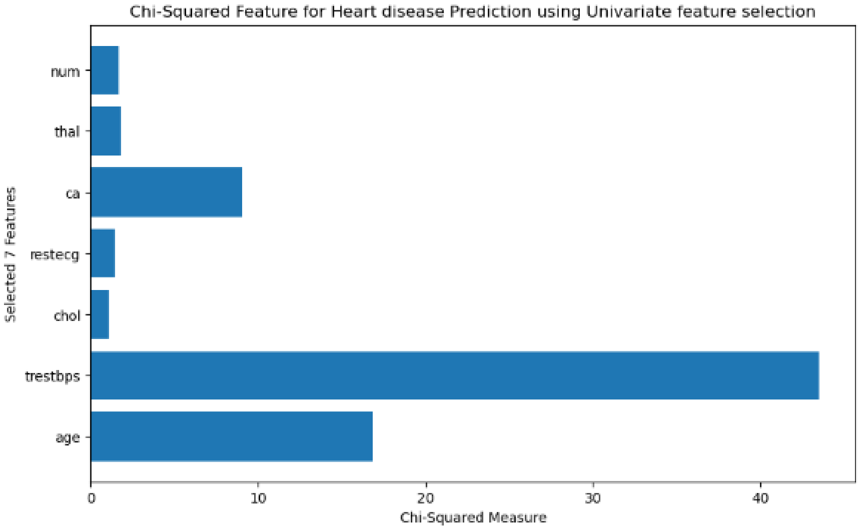

Fig 1: Feature Selection using UFS (Chi-Squared Test)

4. Health Care Data Analysis on CHD Dataset:

4.1. Dataset Representation:

The CHD dataset is collected from Kaggle that contains 303 instances with 14 attributes. In this study, feature selection is done considering all the 14 attributes and the best 5 attributes is selected using Chi-Squared test. Age and sex attributes are the major risk factor for heart disease. Cp(Chest pain type) denotes the various levels of heart disease .Resting blood pressure(trestbps) and Cholestrol(chol) indicates the high blood pressure level and cholesterol level both are also considered as the major risk factors of heart disease. Fasting blood pressure(fbs) indicates the blood sugar level(diabetes) a significant risk factor in the heart disease prediction. Resting ECG(restecg) will produce the abnormal conditions found in the electrocardiogram).Maximum heart rate(thalach) it mainly indicates the poor functioning of the heart. Exercise Induced angina(exang) denotes the restricted blood flow to the heart. STDepression(oldpeak) is used to measure the changes in the ST segment of the ECG. Slope of ST segment(slope) denotes the different types of heart stress or damage. Number of major vessels(ca) denotes the extent of blood flow restriction in the heart. Thalassemia(thal) indicates the different types of thalassemia. num is the attribute to identify the presence and severity of heart disease.

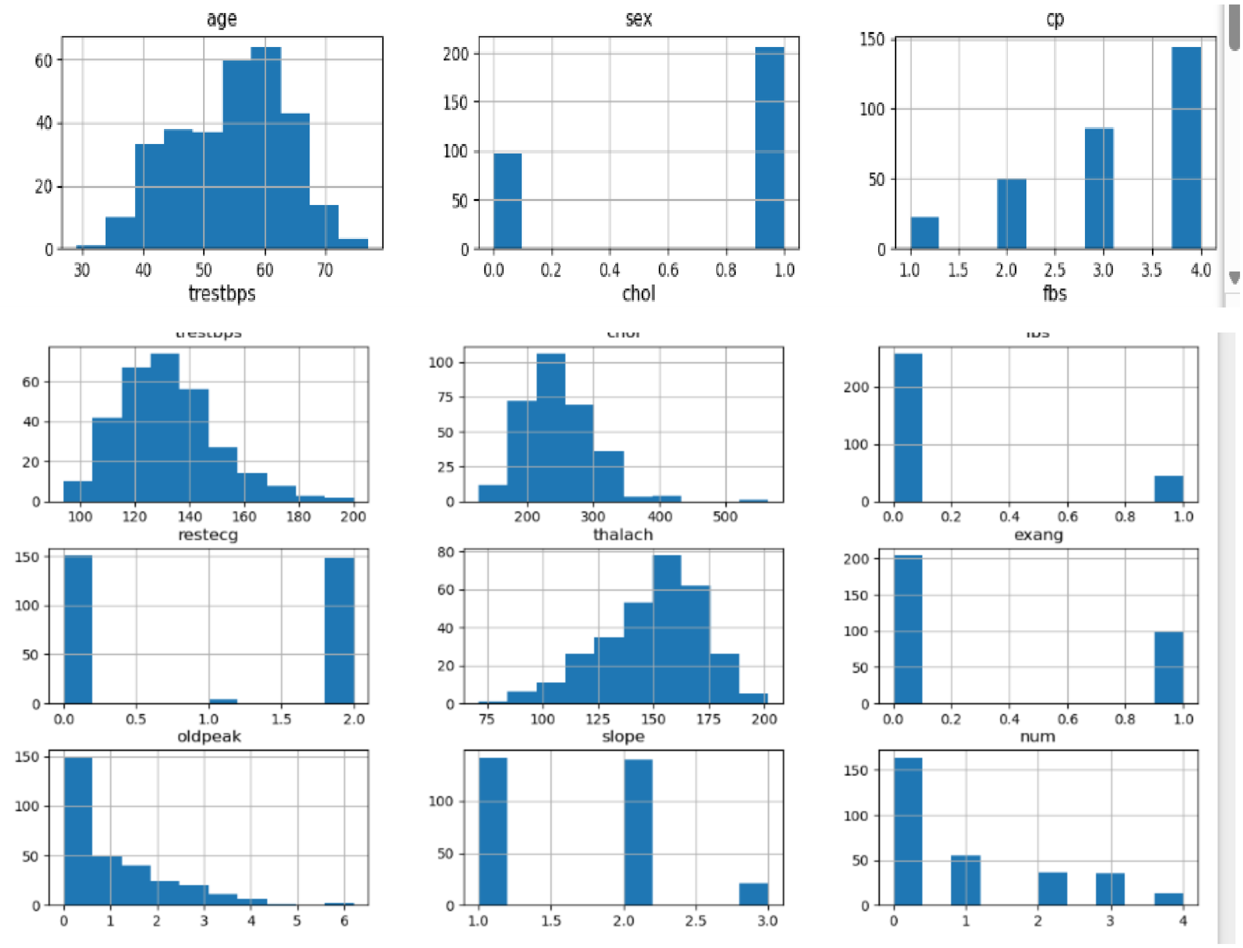

Fig 2: Dataset Representation-CHD dataset

5.Results and Discussion:

The upshots highlight the implication of Machine learning algorithms for heart disease prediction using the feature selection method. The enactment of the univariate feature selection algorithm using Chi-Squared test evidenced that is the best approach in improving the actual performance of the models. Best features are selected from the dataset and helps to improve the evaluation metrics of the models. After the feature selection using univariate feature selection methods, selected features are applied to the machine learning models such as Logistic regression, Random Forest, SVM and XG Boost final accuracy is calculated. Regular supervision and re-evaluation of the rules generated from the learning set are necessary for the responsible usage of these models to guarantee their continued effectiveness and reliability in a range of clinical contexts.

6.Conclusion:

This study aimed to propose an effective model that uses feature selection and classifier models for predicting heart disease. The data that has been used for prediction is obtained from the CHD Dataset. Before applying the UFS for preprocessing of the data is done to ensure there is no irrelevant or missing data in the dataset. This ensures better performance of the model. This research effort used four classifier methods and taken various measures such as accuracy, precision, recall and f1 score. Among all the four classifier models Logistic Regression and SVM has the highest accuracy 90%, precision 0.9, f1 score 0.95 and recall 1. CHD dataset has only limited number of instances. In the future study many real-world data can be taken which contain diverse dataset with huge number of instances and attributes more than 14 considering the major risk factors. Additionally, the future work can extend predicting heart disease along with other medical conditions which is considered to be the severe cause for death.

References

[1] Seckeler MD, Hoke TR. The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. Clin Epidemiol 2011:67–84.

[2] Mohammad Shafenoor Amin, Yin Kia Chiam, Kasturi Dewi Varathan,Identification of significant features and data mining techniques in predicting heart disease,Telematics and Informatics,Volume 36,2019,Pages 82-93,ISSN 0736-5853,

[3] Kavitha, R., Kannan, E., 2016. An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining.International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS), pp. 1–5.

[4] M. Swathy, K. Saruladha,A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques,ICT Express,Volume 8, Issue 1,2022,Pages 109-116,ISSN 2405-9595,https://doi.org/10.1016/j.icte.2021.08.021.

[5] Javed Azmi, Muhammad Arif, Md Tabrez Nafis, M. Afshar Alam, Safdar Tanweer, Guojun Wang,A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data,Medical Engineering & Physics,Volume 105,2022,103825,ISSN 1350-4533,https://doi.org/10.1016/j.medengphy.2022.103825.

[6] Cai-Yi Ma, Ya-Mei Luo, Tian-Yu Zhang, Yu-Duo Hao, Xue-Qin Xie, Xiao-Wei Liu, Xiao-Lei Ren, Xiao-Lin He, Yu-Mei Han, Ke-Jun Deng, Dan Yan, Hui Yang, Hua Tang, Hao Lin, Predicting coronary heart disease in Chinese diabetics using machine learning, Computers in Biology and Medicine, Volume 169,2024,107952,ISSN 0010-4825,https://doi.org/10.1016/j.compbiomed.2024.107952.

[7] Md Manjurul Ahsan, Zahed Siddique,Machine learning-based heart disease diagnosis: A systematic literature review,Artificial Intelligence in Medicine,Volume 128,2022,102289,ISSN 0933-3657,https://doi.org/10.1016/j.artmed.2022.102289.

[8] Vardhan Shorewala,Early detection of coronary heart disease using ensemble techniques, Informatics in Medicine Unlocked,Volume 26,2021,100655,ISSN 2352-9148,https://doi.org/10.1016/j.imu.2021.100655.

[9] Muhammad Salman Pathan, Avishek Nag, Muhammad Mohisn Pathan, Soumyabrata Dev, Analyzing the impact of feature selection on the accuracy of heart disease prediction,HealthcareAnalytics,Volume2,2022,100060,ISSN2772-4425,https://doi.org/10.1016/j.health.2022.100060.

[10] Anna Karen Gárate-Escamila, Amir Hajjam El Hassani, Emmanuel Andrès,Classification models for heart disease prediction using feature selection and PCA,Informatics in Medicine Unlocked,Volume19,2020,100330,ISSN23529148,https://doi.org/10.1016/j.imu.202 0.100330

[11] T. Vivekanandan, N. Ch Sriman Narayana Iyengar,Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease,Computers in Biology and Medicine,Volume 90,2017,Pages 125-136,ISSN 0010-4825,https://doi.org/10.1016/j.compbiomed.2017.09.011.

[12] Ofir Ben-Assuli, Tsipi Heart, Robert Klempfner, Rema Padman,Human-machine collaboration for feature selection and integration to improve congestive Heart failure risk prediction,Decision Support Systems,Volume 172,2023,113982,ISSN 0167-9236,https://doi.org/10.1016/j.dss.2023.113982.

[13] G. Manikandan, B. Pragadeesh, V. Manojkumar, A.L. Karthikeyan, R. Manikandan, Amir H. Gandomi,Classification models combined with Boruta feature selection for heart disease prediction,Informatics in Medicine Unlocked,Volume 44,2024,101442,ISSN 2352-9148,https://doi.org/10.1016/j.imu.2023.101442.

[14] Anna Karen Gárate-Escamila, Amir Hajjam El Hassani, Emmanuel Andrès,Classification models for heart disease prediction using feature selection and PCA,Informatics in Medicine Unlocked,Volume 19,2020,100330,ISSN 2352-9148,https://doi.org/10.1016/j.imu.2020.100330.