International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-11, September 2019

Heart Disease Prediction and Performance Assessment through Attribute Element Diminution using Machine Learning

M. Shyamala Devi, Mothe Sunil Goud, G. Sai Teja, MallyPally Sai Bharath

Abstract: ***In today's modern world, the human beings are affected with heart disease irrespective of the age. With the advancement of technological growth, predicting the availability of Heart diseases still remains a challenging issue. The difficulty of predicting the heart disease prevails due to the lack of availability of the symptoms. According to World Health Organization, 33% of population died due to heart diseases. For this, the diagnosis of heart diseases is made by complex combination of clinical data. With this overview, we have used Heart Disease Prediction dataset extracted from UCI Machine Learning Repository for predicting the level of heart disease. The prediction of heart disease classes are achieved in four ways. Firstly, the data set is preprocessed with Feature Scaling and Missing Values. Secondly, the raw data set is fitted to classifiers like logistic regression, KNN classifier, Support Vector Machine, Kernel Support Vector Machine, Naive Bayes, Random Forest and Decision Tree classifiers. Third, the raw data set is subjected to dimensionality reduction using Principal Component Analysis to project the dataset with important components. The dimensionality PCA reduced data set is fitted to the above-mentioned classifiers. Fourth, the performance comparison of raw data set and PCA reduced data set is done by analyzing the performance metrics like Precision, Recall, Accuracy and F-score. The implementation is done using python language under Spyder platform with Anaconda Navigator. Experimental results shows that Random forest is found to be effective with the accuracy of 89% without applying PCA, 85% with five component PCA and 86% with seven component PCA.***

Index Terms*: *Machine Learning, Classification, accuracy, precision, F-Score and recall.

I. INTRODUCTION

The Prediction of disease is a challenging task in the medical field so as to treat the patients in advance. By the medical survey, the population of elderly people is increasing rapidly and they need continuous care for their treatments. The prediction of the disease helps the chronically ill patients to progress their medical treatment and to save their life in advance. With the modern life, the people are lack of nutritious food which leads to poor health. The hospital management system also turns into using computerized devices to perform surgical operations. The people expect a world class treatment from the doctors. The disease makes the patients to be highly depressed because of the family background. They need to know the growth of the disease and the extent of spread of the disease. The prediction of disease in early stage remains to be a challenging issue.

The paper is organized in such a way that Section 2 deals with the related works. Section 3 discuss about dimensionality Reduction. Proposed work is discussed in Section 4 followed by the implementation and Performance Analysis in Section 5. The paper is concluded with Section 6.

II. RELATED WORK

A. Literature Review

The health care industry is moving towards technology by using machine learning in the prediction of diseases. The technology is used in the detection of Loco motor disorders, Heart diseases. This helps the doctors in diagnosing the disease earlier [1]. Heart is the important organ equal to brain in the existence of human body. Heart bumps the blood and transfer to all the organs of the body to function. The prediction of heart disease is a challenging task in the medical field. The patient's medical data is stored in the database and they are subjected to data mining techniques like artificial neural Network, Decision tree, Fuzzy Logic, K-Nearest Neighbor, Naive Bayes and Support Vector Machine [2].

A hybrid intelligent base system framework is designed for predicting the heart disease using machine learning algorithms. One of the most critical human diseases in the world is heart disease. When the person suffers from heart disease, the heart will not be able to push the needed blood to other parts of the body. The non invasive based machine learning algorithms are used to diagnose the heart disease [3]. The different parameters of the heart and the body can be used to predict the heart disease [4]. The analysis of the heart disease prediction is done for the male patients by using data mining techniques. They have three data mining techniques like Naive Bayes, Artificial neural network, and the J48 decision tree algorithms [5].

Manuscript published on 30 September 2019.
***Correspondence Author(s)**
M. Shyamala Devi, Associate Professor, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.
Mothe Sunil Goud, Final Year B.Tech, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.
G. Sai Teja, Final Year B.Tech, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India
MallyPally Sai Bharath, Final Year B.Tech, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India

The health care management assembles and collects the large amount of data with hidden and secured information. The effective decision and predictions are taken; advanced data mining techniques are used. An effective heart disease prediction system is designed using neural network which detects and categorizes the risk level of the heart disease [6]. A critical review on various feature selection, feature extraction methods, classification methods and the performances parameters are examined for predicting the wine quality [7]-[13].

III. DIMENSIONALITY REDUCTION

Dimensionality reduction process the conversion of very large data attributes into small set of dataset attributes. The dimensionality reduction can be performed under two categories. They are Feature selection and Feature Extraction methods. Feature selection method select the highly feature importance attributes based on variance between the attributes in the dataset. The Feature extraction method combines all the attributes as a matrix and selects the components based on the covariance relation between them.

IV. PROPOSED WORK

In our proposed work, machine learning algorithms are used to predict the disease of Heart disease data set. Our contribution in this paper is folded in two ways.

(i) Firstly, the data set is preprocessed with Feature Scaling and Missing Values.

(ii)Secondly, the raw data set is fitted to classifiers like logistic regression, KNN classifier, Support Vector Machine, Kernel Support Vector Machine, Naive Bayes, Random Forest and Decision Tree classifiers.

(iii) Third, the raw data set is subjected to dimensionality reduction using Principal Component Analysis to project the dataset with important components. The dimensionality PCA reduced data set is fitted to the above-mentioned classifiers.

(iv) Fourth, the performance comparison of raw data set and PCA reduced data set is done by analyzing the performance metrics like Precision, Recall, Accuracy and F-score.

A. Principal Component Analysis

Principal component analysis is the feature extraction method which takes out the new set of attributes from the data set attributes through the process of linear transformation. The selected new sets of attributes from the original dataset are called as the principal components and the steps of PCA are given below.

Step 1: Dataset Covariance matrix is constructed based on the mathematical approach.

Step 2: The Eigen vectors of the Covariance matrix is then calculated

Step 3: The high Eigen values from the eigen vectors are used to renovate the unique data set.

Step 4: The high variance attributes are selected as principal components.

B. System Architecture

The system architecture of our proposed work is shown in Fig. 1

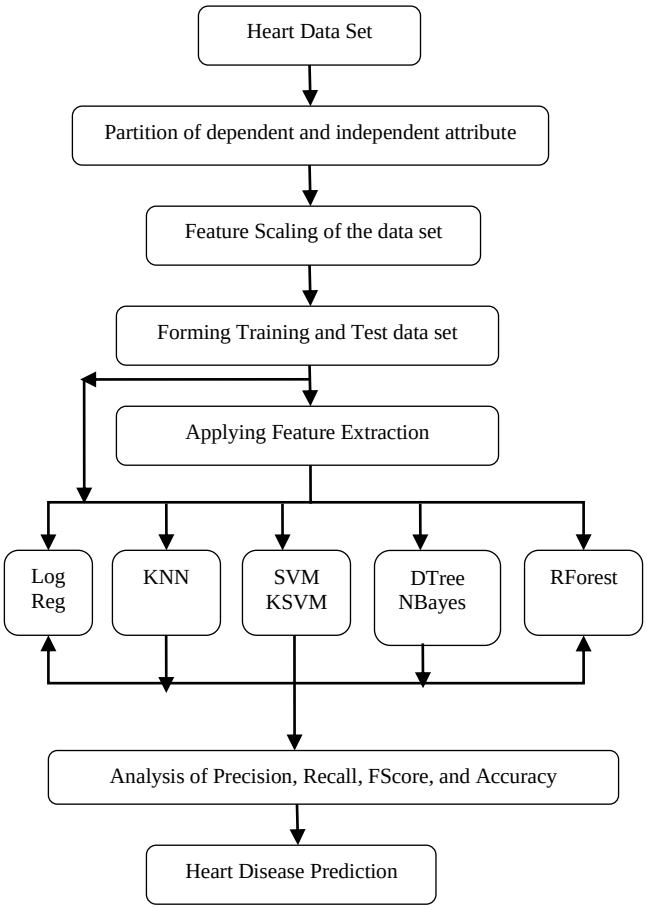

Fig. 1 System Architecture

V. IMPLEMENTATION AND PERFORMANCE ANALYSIS

A. Heart Disease Prediction for Feature Extraction

The Heart Disease dataset extracted from UCL ML Repository is used for implementation with 13 independent attribute and 1 diagnosis dependent attribute. The dataset consists of 779 individual's data. The attribute are shown below.

1. Age
2. Sex
3. Chest-pain type
4. Resting Blood Pressure
5. Serum Cholestrol
6. Fasting Blood Sugar
7. Resting ECG
8. Max heart rate
9. Angina
10. ST depression
11. Peak exercise ST segment
12. Number of major vessels
13. Thal
14. Diagnosis of heart disease - Dependent Attribute

The Description of the attributes are shown in Table. 1.

Table. 1 Description of the dataset

S.No	Attribute	Description
1.	Age	Age of the individual.
2.	Sex	Gender of the individual with the following format: 1 = male 0 = female.
3.	Chest-pain type	Type of chest-pain as in Table. 2.
4.	Resting Blood Pressure	Resting blood pressure value in mmHg (unit)
5.	Serum Cholestrol	serum cholestrol in mg/dl (unit)
6.	Fasting Blood Sugar	Fasting blood sugar value of an individual with 120mg/dl. If fasting blood sugar > 120mg/dl then : 1 (true) else : 0 (false)
7.	Resting ECG	0 = normal 1 = having ST-T wave abnormality 2 = left ventricular hypertrophy
8.	Max heart rate	Max Heart Rate
9.	Angina	1 = yes 0 = no
10.	ST depression	Value (integer or float).
11.	Peak exercise ST segment	1 = Upsloping 2 = Flat 3 = Downsloping
12.	Number of major vessels	Values (0-3) colored by flourosopy
13.	Thal	displays the Thalassemia : 3 = Normal 6 = Fixed Defect 7 = Reversable Defect
14.	Diagnosis of heart disease	0 = Absence 1,2,3,4 = Present.

Table. 2 Description of Chest Pain Type

Chest Pain Format	Desription
1	Typical Angina
2	Atypical Angina
3	Non - Anginal Pain
4	Asymptotic

B. Performance Analysis

The relationship between the attributes in the dataset is shown as a histogram notation in the fig 2. The dataset is applied with principal component analysis with 5 components and the relation between the component and variance is shown in fig 3. The dataset is applied with principal component analysis with 5 components and the relation between the component and variance is shown in fig 4.

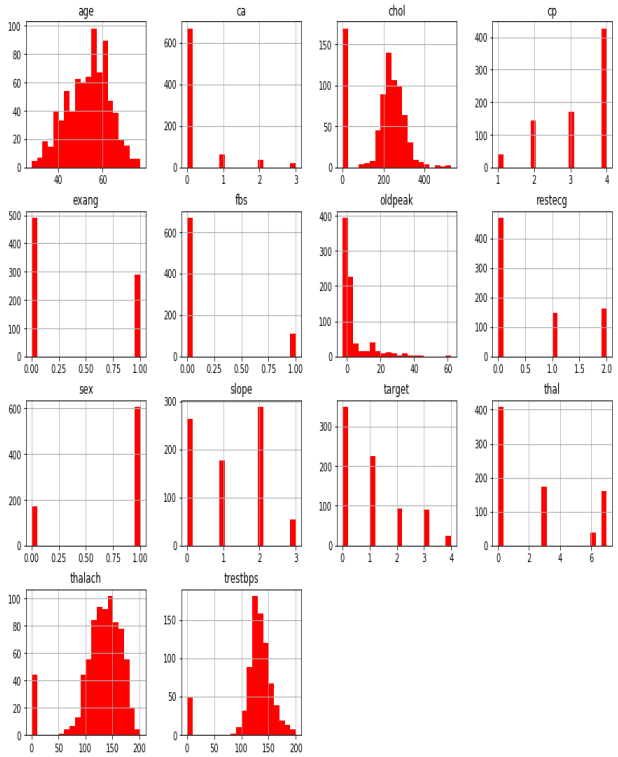

Fig. 2. Histogram Relation of the heart disease data set

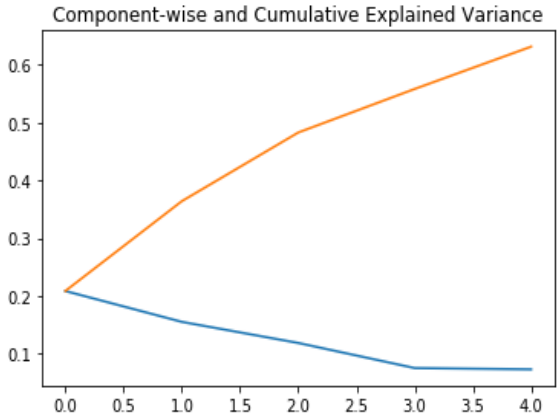

Fig. 3. Component VS Variance for PCA with 5 components

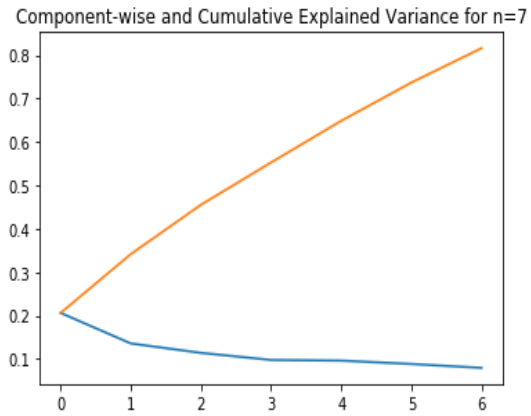

Fig. 4. Component VS Variance for PCA with 7 components

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The obtained confusion matrix for each of the classifiers like logistic regression, KNN classifier, Support Vector Machine, Kernel Support Vector Machine, Naive Bayes, Random Forest and Decision Tree classifiers without applying principal component analysis is shown in fig 5.

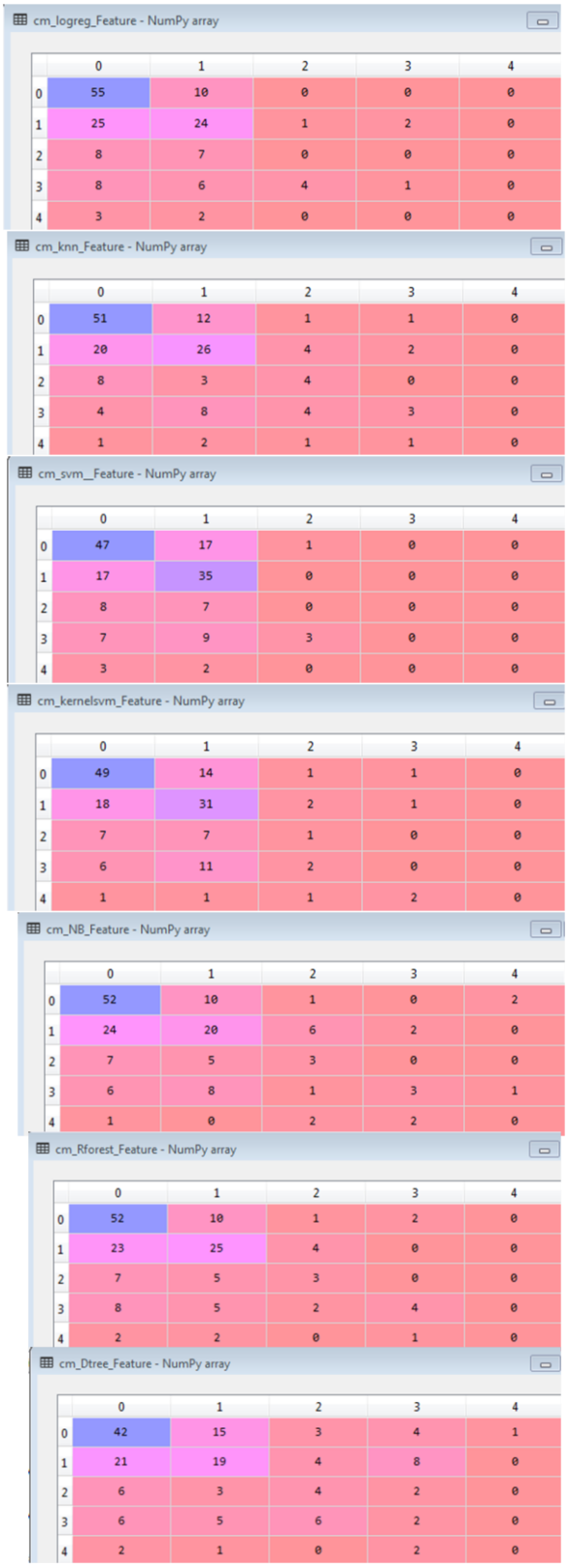

Fig. 5 Confusion Matrix for Logistic, KNN, Kernel SVM, Naive Bayes, Random Forest and Decision Tree Classifier without PCA

The obtained confusion matrix for each of the classifiers like logistic regression, KNN classifier, Support Vector Machine, Kernel Support Vector Machine, Naive Bayes, Random Forest and Decision Tree classifiers with five components of principal component analysis is shown in fig 6.

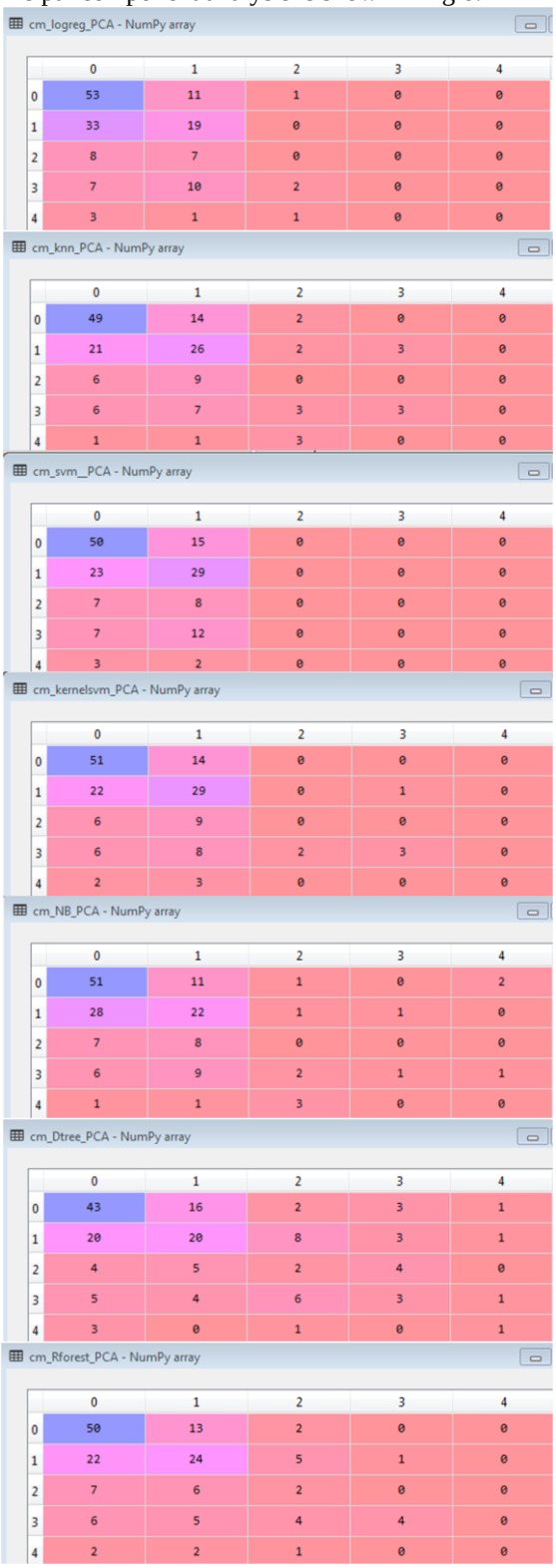

Fig. 6 Confusion Matrix for Logistic, KNN, Kernel SVM, Naive Bayes, Random Forest and Decision Tree Classifier with 5 Component PCA

Retrieval Number: K15970981119/19©BEIESP
DOI: 10.35940/ijitee.K1597.0981119
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

The obtained confusion matrix for each of the classifiers like logistic regression, KNN classifier, Support Vector Machine, Kernel Support Vector Machine, Naive Bayes, Random Forest and Decision Tree classifiers with seven components of principal component analysis is shown in fig 7.

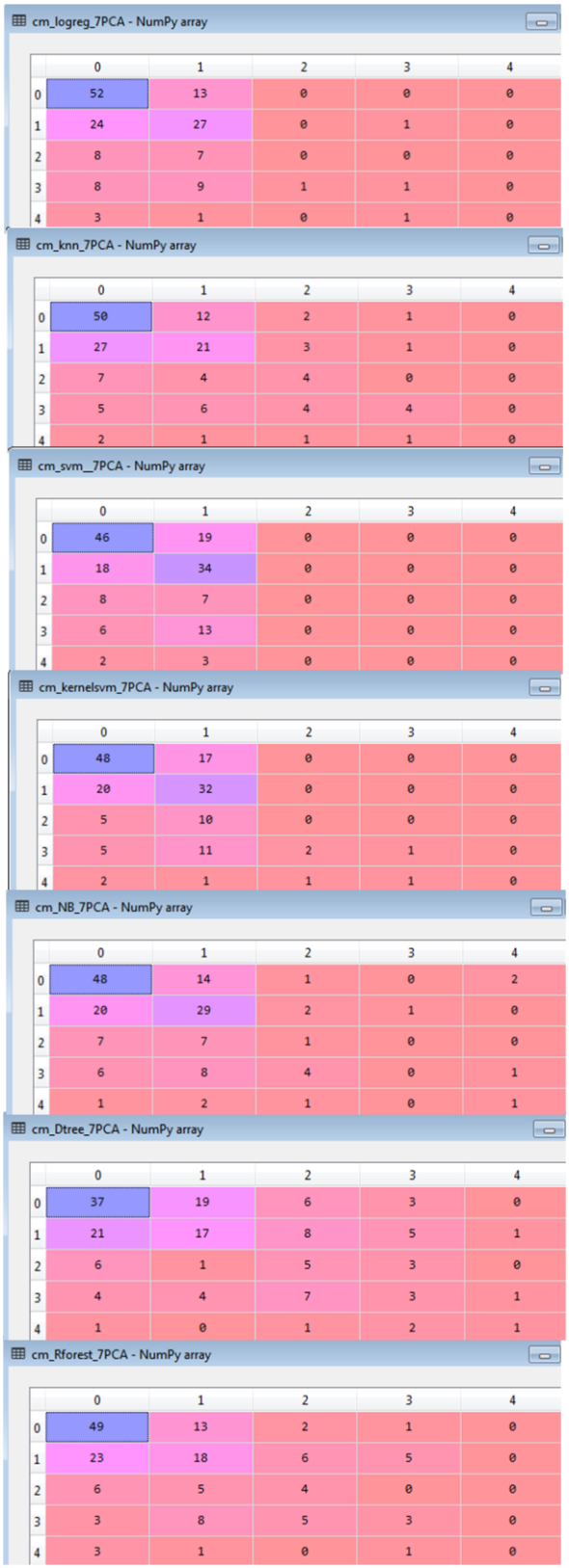

Fig.7 Confusion Matrix for Logistic, KNN, Kernel SVM, Naive Bayes, Random Forest and Decision Tree Classifier with 7 Component PCA

The performance analysis is done based on precision, Recall and FScore for all the classifiers like logistic regression, KNN classifier, Support Vector Machine, Kernel Support Vector Machine, Naive Bayes, Random Forest and Decision Tree classifiers and is shown in the Table 3 -Table 5.

Table. 3 Performance Comparison of Precision, Recall and FScore for all the classifiers before applying PCA

Classifier	Performance Metrics without PCA		
	Precision	Recall	FScore
Logistic Reg	0.44	0.51	0.45
KNN	0.5	0.54	0.51
Kernel SVM	0.43	0.52	0.47
Naive Bayes	0.47	0.51	0.47
Random Forest	0.44	0.47	0.45
Decision Tree	0.41	0.43	0.42
SVM	0.41	0.53	0.46

Table. 4 Performance Comparison of Precision, Recall and FScore for all the classifiers with 5 Component PCA

Classifier	Performance Metrics without PCA		
	Precision	Recall	FScore
Logistic Reg	0.34	0.46	0.46
KNN	0.46	0.5	0.46
Kernel SVM	0.49	0.53	0.48
Naive Bayes	0.43	0.47	0.42
Random Forest	0.51	0.51	0.48
Decision Tree	0.43	0.44	0.43
SVM	0.38	0.51	0.43

Table. 5 Performance Comparison of Precision, Recall and FScore for all the classifiers with 7 Component PCA

Classifier	Performance Metrics without PCA		
	Precision	Recall	FScore
Logistic Reg	0.43	0.51	0.45
KNN	0.49	0.51	0.48
Kernel SVM	0.46	0.52	0.46
Naive Bayes	0.42	0.51	0.46
Random Forest	0.47	0.55	0.49
Decision Tree	0.41	0.41	0.42
SVM	0.42	0.51	0.46

The performance analysis is done based on accuracy for all the classifiers and is shown in the Table 6.

Table. 6 Accuracy for all the classifiers with and without PCA

Classifier	Accuracy		
	Without PCA	5 Component PCA	7 Component PCA
Logistic Reg	72	70	71
KNN	77	75	73
Kernel SVM	79	76	74
Naive Bayes	82	80	79
Random Forest	89	85	86
Decision Tree	74	70	71
SVM	75	72	73

The performance analysis is done based on Accuracy, precision, Recall and FScore for all the classifiers like logistic regression, KNN classifier, Support Vector Machine, Kernel Support Vector Machine, Naive Bayes, Random Forest and Decision Tree classifiers and the analysis is depicted as a pictorial representation and is shown in the fig. 8 – fig. 11.

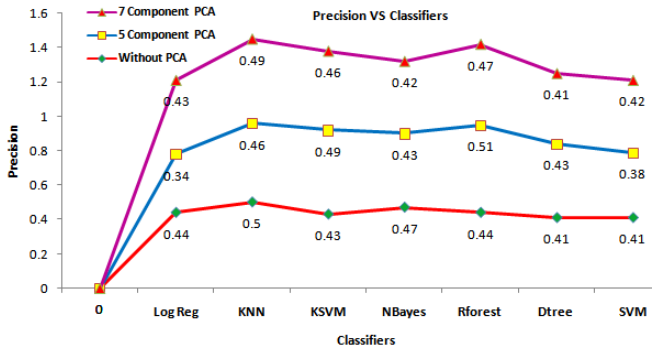

Fig 8. Precision VS Classifiers

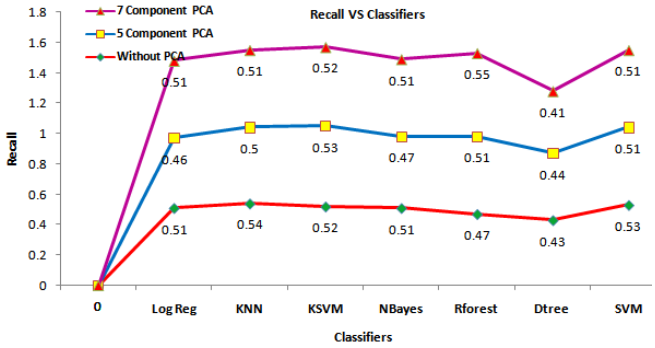

Fig 9. Recall VS Classifiers

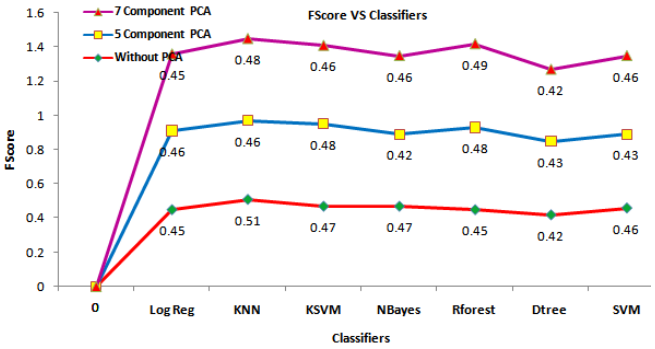

Fig 10. FScore VS Classifiers

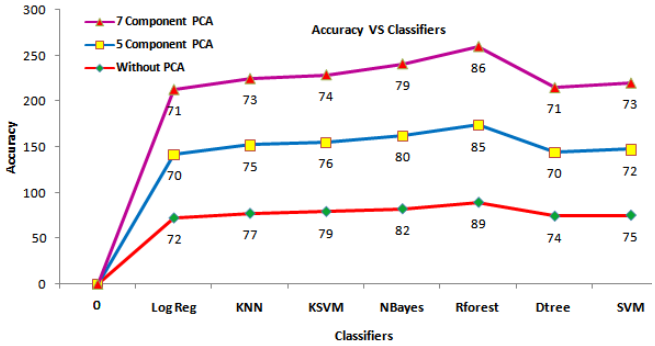

Fig 11. Accuracy VS Classifiers

VI. CONCLUSION

This paper analyzes the method to predict the heart disease of the heart dataset through the dimensionality reduction thereby improving the accuracy prediction. An attempt is made to implement the dimensionality reduction through principal component analysis for the heart disease dataset. The performance of all the classifiers of the heart disease dataset is compared before and after applying PCA. Experimental results shows that after applying PCA, the Random forest is found to be effective with the accuracy of 86% compared to other classifiers with the seven component PCA.

REFERENCES

1. Samir B Patel," Heart disease prediction using Machine Learning and Data Mining", International journal of computer sciences and Engineering, vol 7,2015,pp.129-137
2. V.V.Ramalingam, "Prediction of Heart Diseases Using Machine Learning", International Journal of Engineering and Technology, vol. 7,no. 2, 2018.
3. Amin Ul Haq ,"A Hybrid Intelligence System Frame Work for Prediction of Heart Disease Using ML", Mobile Information Systems ,article id 3860146,21 pages.
4. Shadman Nashif, "Heart Disease Detection by Using Machine Learning Algorithm and a Real time Cardiovascular Health Monitoring System ", World journal of Engineering and Technology,vol.06 no.04,2018,article id 88650.
5. Poornima Singh ," Effective heart disease prediction system using data mining techniques", International journal of Nano medicine, Issued on 2018 , pp 121-124.
6. M. Gunay, and T. Ensarı, "Predictive churn analysis with machine learning methods," 26th Signal Processing and Communications Applications Conference (SIU), Izmir, 2018, pp. 1- 4.
7. M. Shyamala Devi, Rincy Merlin Mathew, and R. Suguna,"Attribute Heaving Extraction and Performance Analysis for the Prophesy of Roof Fall Rate using Principal Component Analysis", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.8, June 2019, pp. 2319-2323.
8. R. Suguna, M. Shyamala Devi, and Rincy Merlin Mathew, " Customer Churn Predictive Analysis by Component Minimization using Machine Learning", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.8, June 2019, pp. 2329-2333.
9. Shyamala Devi Munisamy, and Suguna Ramadass Aparna Joshi, "Cultivar Prediction of Target Consumer Class using Feature Selection with Machine Learning Classification", Springer's book series "Learning and Analytics in Intelligent Systems, Springer, LAIS vol. 3, pp. 604-612, 2019.
10. Suguna Ramadass, and Shyamala Devi Munisamy, Praveen Kumar P, Naresh P, "Prediction of Customer Attrition using Feature Extraction Techniques and its Performance Assessment through dissimilar Classifiers", Springer's book series "Learning and Analytics in Intelligent Systems, Springer, , LAIS vol. 3, pp. 613-620, 2019.
11. R.Suguna, M. Shyamala Devi, Rupali Amit Bagate, and Aparna Shashikant Joshi, "Assessment of Feature Selection for Student Academic Performance through Machine Learning Classification", Journal of Statistics and Management Systems, Taylor Francis, vol. 22, no. 4, 25 June 2019, pp. 729-739.
12. M. Shyamala Devi, Rincy Merlin Mathew, R. Suguna, "Feature Snatching and Performance Analysis for Connoting the Admittance Likelihood of student using Principal Component Analysis", International Journal of Recent Technology and Engineering, Volume-8 Issue-2, 30 July 2019.pp. 4800-4807.
13. R. Suguna, M. Shyamala Devi, Rincy Merlin Mathew, "Customer Segment Prognostic System by Machine Learning using Principal Component and Linear Discriminant Analysis", International Journal of Recent Technology and Engineering, Volume-8 Issue-2, 30 July 2019. pp. 6198-6203.

Retrieval Number: K15970981119/19©BEIESP
DOI: 10.35940/ijitee.K1597.0981119
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)