Exploration of Neighbor Kernels and Feature Estimators for Heart Disease Prediction using Machine Learning

Rincy Merlin Mathew, M. Shyamala Devi, Shakila Basheer

Abstract:** In the growing era of technological world, the people are suffered with various diseases. The common disease faced by the population irrespective of the age is the heart disease. Though the world is blooming in technological aspects, the prediction and the identification of the heart disease still remains a challenging issue. Due to the deficiency of the availability of patient symptoms, the prediction of heart disease is a disputed charge. With this overview, we have used Heart Disease Prediction dataset extorted from UCI Machine Learning Repository for the analysis and comparison of various parameters in the classification algorithms. The parameter analysis of various classification algorithms of heart disease classes are done in five ways. Firstly, the analysis of dataset is done by exploiting the correlation matrix, feature importance analysis, Target distribution of the dataset and Disease probability based on the density distribution of age and sex. Secondly, the dataset is fitted to K-Nearest Neighbor classifier to analyze the performance for the various combinations of neighbors with and without PCA. Thirdly, the dataset is fitted to Support Vector classifier to analyze the performance for the various combinations of kernels with and without PCA. Fourth, the dataset is fitted to Decision Tree classifier to analyze the performance for the various combinations of features with and without PCA. Fifth, the dataset is fitted to Random Forest classifier to analyze the performance for the various levels of estimators with and without PCA. The implementation is done using python language under Spyder platform with Anaconda Navigator. Experimental results shows that for KNN classifier, the performance for 12 neighbours is found to be effective with 0.52 before applying PCA and 0.53 after applying PCA. For Support Vector classifier, the rbf kernel is found to be effective with the score of 0.519 with and without PCA. For Decision Tree classifier, before applying PCA, the score is 0.47 for 7 features and after applying PCA, the score is 0.49 for 4 features. For, Random Forest Classifier, before applying PCA, the score is 0.53 for 500 estimators and after applying PCA, the score is 0.52 for 500 estimators.

***Index Terms:** Machine Learning, Classification, Kernel, Feature, Neighbor, Estimator.*

Revised Manuscript Received on October 30, 2019.
* Correspondence Author

Rincy Merlin Mathew*, Lecturer, Department of Computer Science, College of Science and Arts, Khamis Mushayt, King Khalid university, Abha, Asir, Saudi Arabia.

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.

Shakila Basheer, Assistant Professor, Department of Information Systems, College of Computer and Information Sciences, Princess Nourah Bint Abdul Rahman university, Riyadh-11671, Saudi Arabia

I. INTRODUCTION

For enabling the patients to treat in advance, the prediction of particular disease is necessary. Generally the heart disease is affected for the elderly people as per the survey. The heart disease prediction may help the immobile chronically ill patients to undergo the medical treatment in advance so as to save their life. As the symptoms needed to identify the heart disease is minimum, we are in need of predicting software to identify the existence of the heart disease. The doctors are unable to predict the disease by the lack of future symptoms of the patients. This requires the use of machine learning techniques for forecasting the future symptoms of the patients.

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

II. RELATED WORK

A. Literature Review

The health care industry is now focusing on using the appropriate technology in the prediction of health diseases. The diseases like Loco motor disorders and Heart diseases are using the technology for the detection of existence of the disease. This helps the doctors in diagnosing to forecast the disease in advance [1]. Heart is the significant limb which is equal to brain in the survival of human body. Heart smacks the blood and move to all the organs of the body for functioning. The prediction of heart disease still remains as a difficult 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].

When the person is suffering from heart disease, the heart will not be able to drive the desirable blood to other parts of the body. The non invasive based machine learning algorithms are used to diagnose the heart disease [3]. The different constraints of the heart and the body can be used to forecast the heart disease [4].

The analysis was done for the male patients for the heart disease prediction using data mining techniques.

They have used the three data mining techniques like Naive Bayes, Artificial neural network, and the J48 decision tree [5].

A overall technological review on various feature selection, feature extraction methods, classification methods and the performances parameters are examined for predicting the wine quality [6]-[22].

III. 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 five ways.

(i) Firstly, the analysis of dataset is done by exploiting the correlation matrix, feature importance analysis, Target distribution of the dataset and Disease probability based on the density distribution of age and sex

(ii) Secondly, the dataset is fitted to K-Nearest Neighbor classifier to analyze the performance for the various combinations of neighbors with and without PCA.

(iii) Thirdly, the dataset is fitted to Support Vector classifier to analyze the performance for the various combinations of kernels with and without PCA.

(iv) Fourth, the dataset is fitted to Decision Tree classifier to analyze the performance for the various combinations of features with and without PCA.

(v) Fifth, the dataset is fitted to Random Forest classifier to analyze the performance for the various levels of estimators with and without PCA.

A. System Architecture

The overall design of this paper is shown in Fig. 1

IV. 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

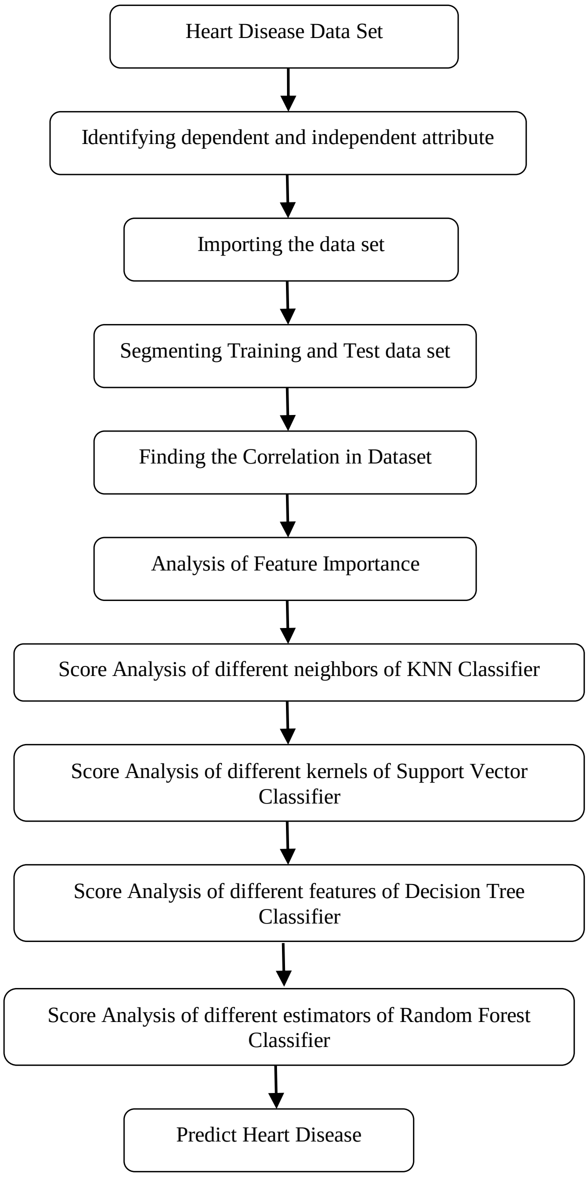
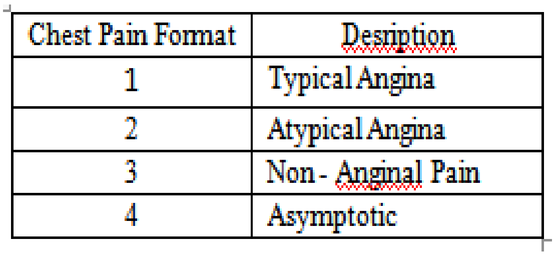

Heart Disease Data Set
Identifying dependent and independent attribute
Importing the data set
Segmenting Training and Test data set
Finding the Correlation in Dataset
Analysis of Feature Importance
Score Analysis of different neighbors of KNN Classifier
Score Analysis of different kernels of Support Vector Classifier
Score Analysis of different features of Decision Tree Classifier
Score Analysis of different estimators of Random Forest Classifier
Predict Heart Disease

Fig. 1 System Architecture

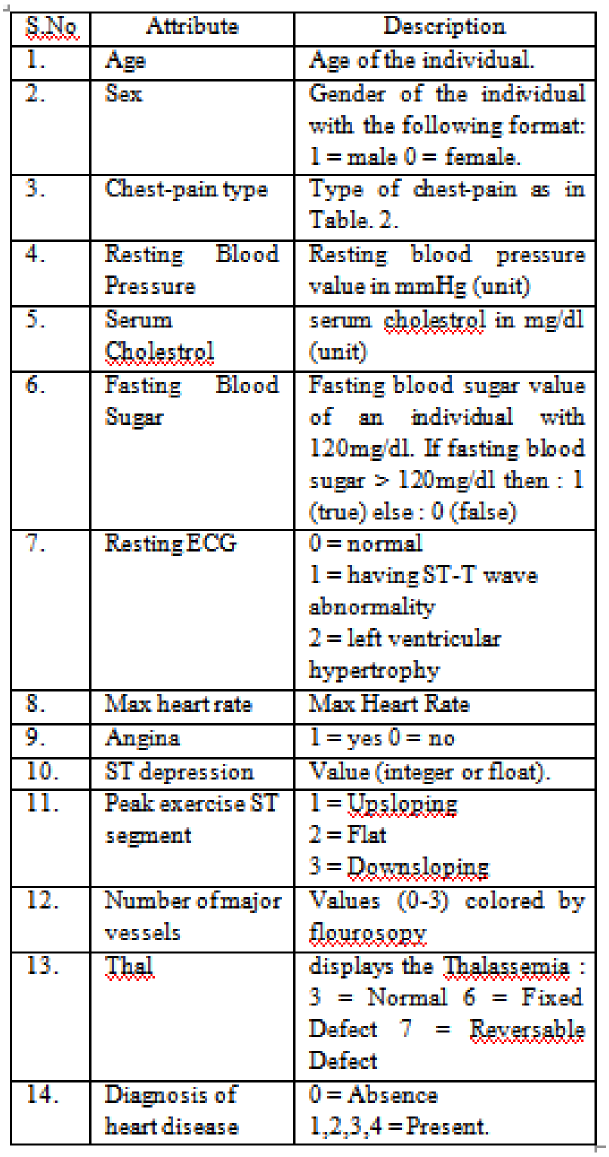
The portrayal of the dependent attribute is shown in Fig. 2.



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

Fig. 2. Chest Pain Type Distribution

The expansion of the dataset attributes shown in Fig. 3.



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.

Fig. 3. Heart Disease Dataset Schema

B. Performance Analysis

The important features distribution is shown in fig 2. The relationship between the components of the heart disease dataset is shown in fig 3. The allotment of the target dependent attribute of heart disease dataset is shown in fig 4.

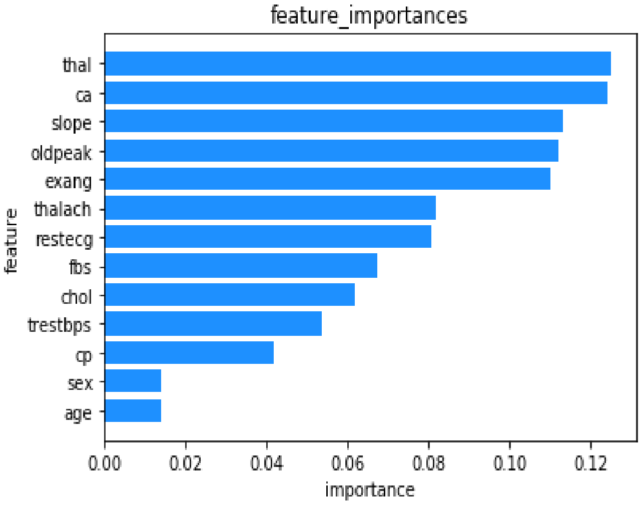


Fig. 2. Important Features of Heart Disease Data Set

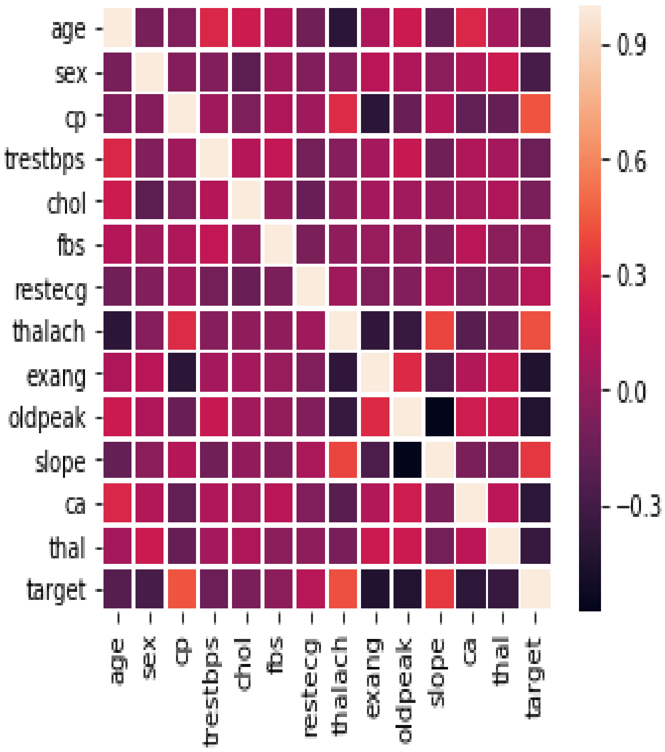


Fig. 3. Correlation Matrix of Heart Disease Data Set

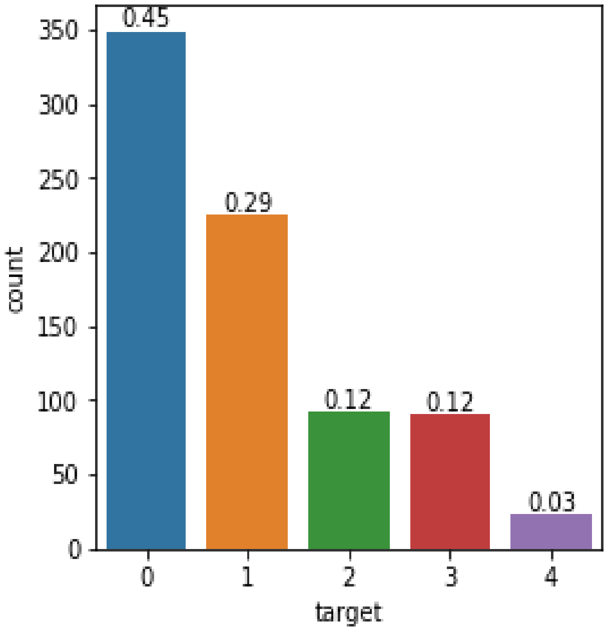


Fig. 4. Target Distribution of Heart Disease Data Set

The Disease Probability and Density Distribution based on Age for Heart Disease Data Set is shown in fig 5. The Distribution of Important features with Target for Heart Disease Data Set is shown in fig 6. The Target Distribution for Heart Disease Prediction for Heart Disease Data Set is shown in fig 7.

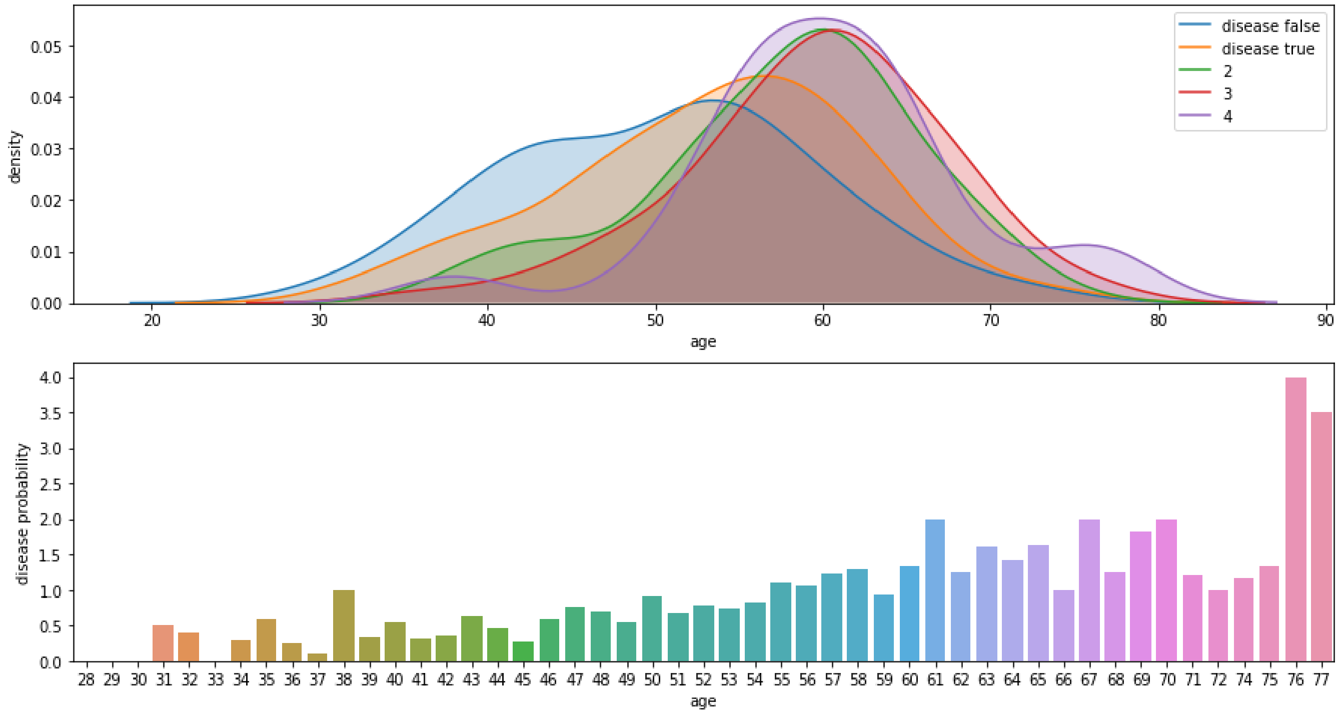


Fig. 5. Disease Probability and Density Distribution based on Age for Heart Disease Data Set

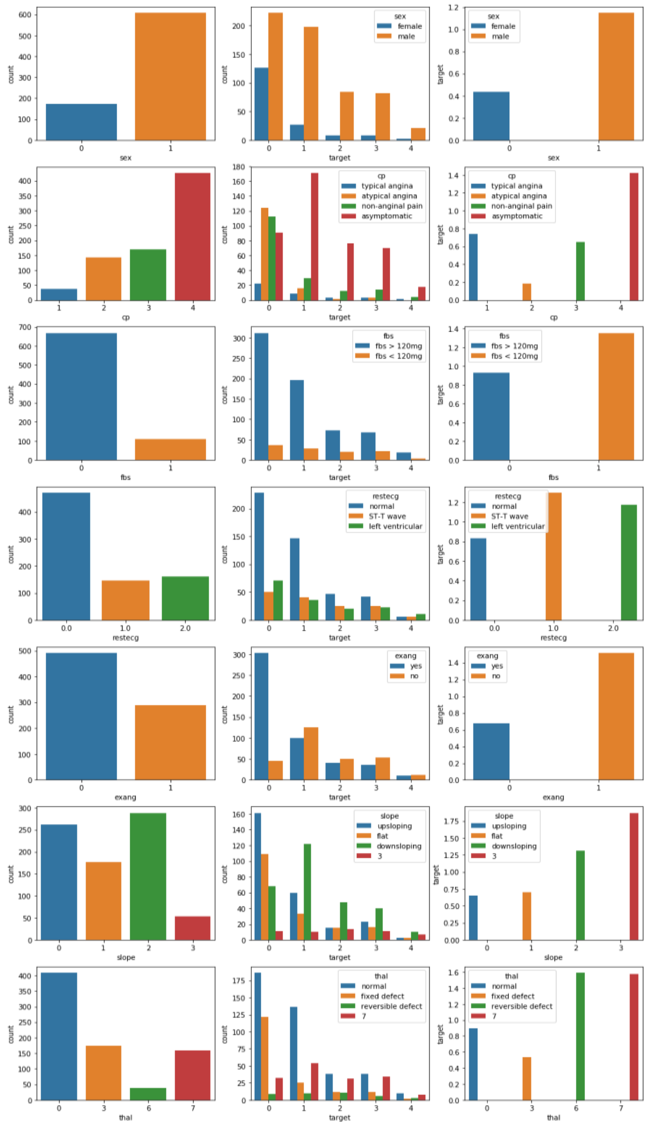


Fig. 6. Distribution of Important features with Target for Heart Disease Data Set

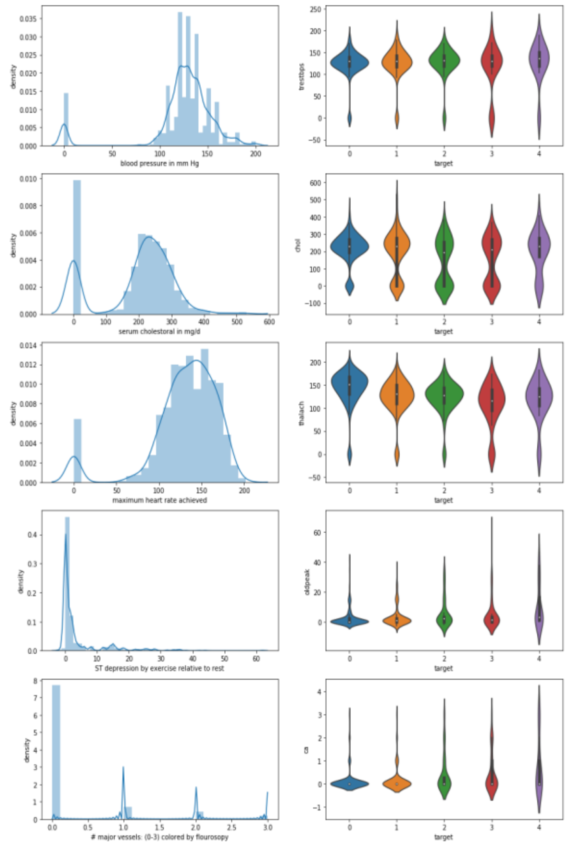
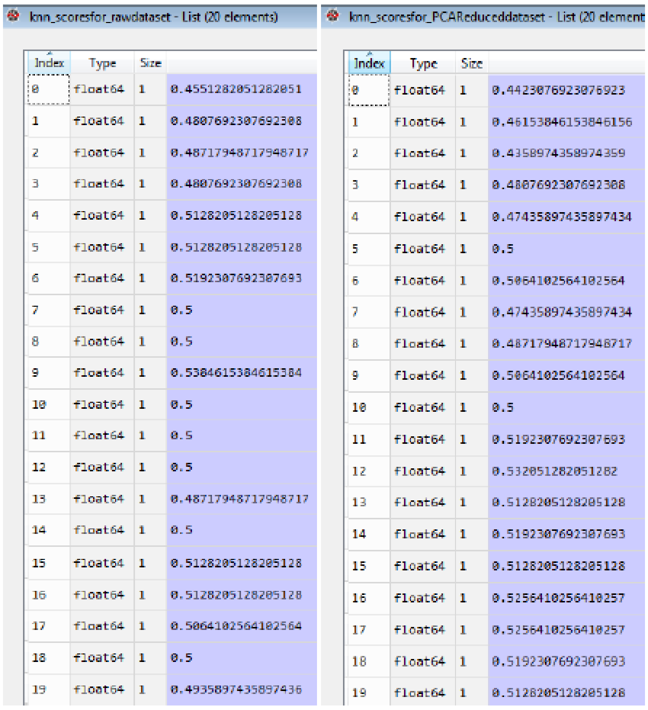


Fig. 7. Target Distribution for Heart Disease Prediction

The dataset is fitted to K-Nearest Neighbor classifier to analyze the performance for the various combinations of neighbors with and without PCA. The Scores for the raw and PCA reduced dataset for KNN Classifier for different number of neighbors is shown in fig 8 – Fig 10.



knn_scoresfor_rawdataset - List (20 elements)

Index	Type	Size	
0	float64	1	0.4551282051282051
1	float64	1	0.4807692307692308
2	float64	1	0.48717948717948717
3	float64	1	0.4807692307692308
4	float64	1	0.5128205128205128
5	float64	1	0.5128205128205128
6	float64	1	0.5192307692307693
7	float64	1	0.5
8	float64	1	0.5
9	float64	1	0.5384615384615384
10	float64	1	0.5
11	float64	1	0.5
12	float64	1	0.5
13	float64	1	0.48717948717948717
14	float64	1	0.5
15	float64	1	0.5128205128205128
16	float64	1	0.5128205128205128
17	float64	1	0.5064102564102564
18	float64	1	0.5
19	float64	1	0.4935897435897436

knn_scoresfor_PCAReduceddataset - List (20 element

Index	Type	Size	
0	float64	1	0.4423076923076923
1	float64	1	0.46153846153846156
2	float64	1	0.4358974358974359
3	float64	1	0.4807692307692308
4	float64	1	0.47435897435897434
5	float64	1	0.5
6	float64	1	0.5064102564102564
7	float64	1	0.47435897435897434
8	float64	1	0.48717948717948717
9	float64	1	0.5064102564102564
10	float64	1	0.5
11	float64	1	0.5192307692307693
12	float64	1	0.532051282051282
13	float64	1	0.5128205128205128
14	float64	1	0.5192307692307693
15	float64	1	0.5128205128205128
16	float64	1	0.5256410256410257
17	float64	1	0.5256410256410257
18	float64	1	0.5192307692307693
19	float64	1	0.5128205128205128

Fig. 8. Scores for raw and PCA reduced dataset for KNN Classifier.

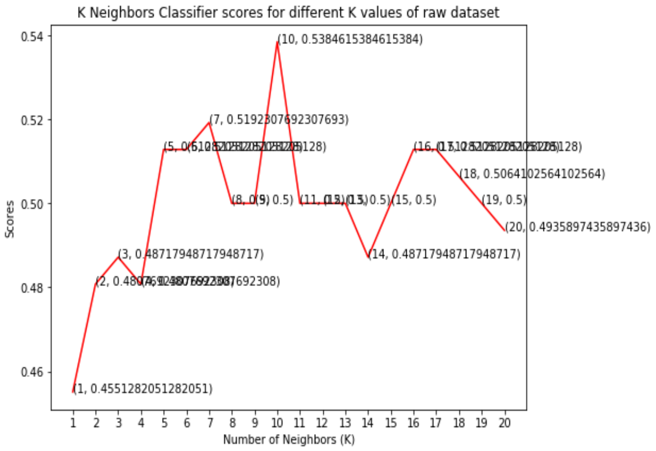


Fig. 9. RAW Dataset Score Analysis for different K values of KNN

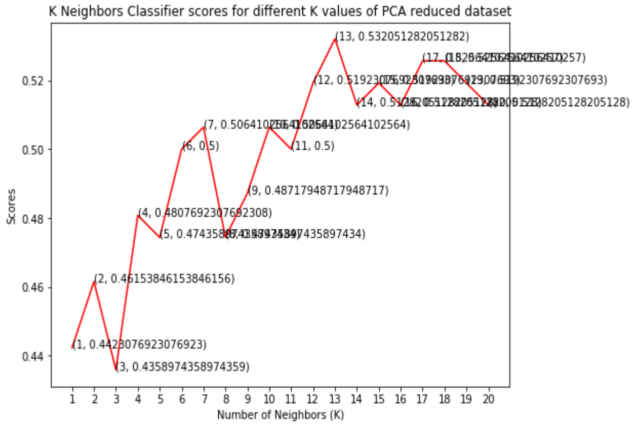
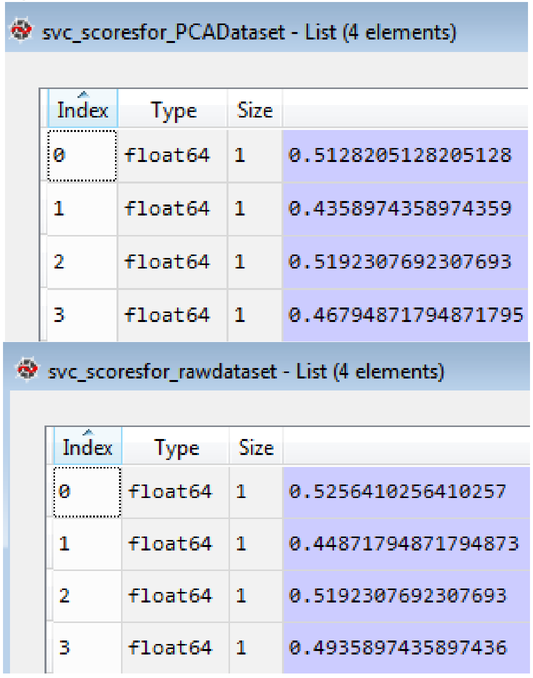


Fig. 10. PCA Reduced Dataset Score Analysis for different K values of KNN Classifier.

The dataset is fitted to Support Vector classifier to analyze the performance for the various combinations of kernels with and without PCA. The Raw Dataset and PCA Reduced Dataset Scores for Support Vector Classifier analysis is shown in fig 11 – Fig 13.



svc_scoresfor_PCADataset - List (4 elements)

Index	Type	Size	
0	float64	1	0.5128205128205128
1	float64	1	0.4358974358974359
2	float64	1	0.5192307692307693
3	float64	1	0.46794871794871795

svc_scoresfor_rawdataset - List (4 elements)

Index	Type	Size	
0	float64	1	0.5256410256410257
1	float64	1	0.44871794871794873
2	float64	1	0.5192307692307693
3	float64	1	0.4935897435897436

Fig. 11. Raw Dataset and PCA Reduced Dataset Scores for Support Vector Classifier.

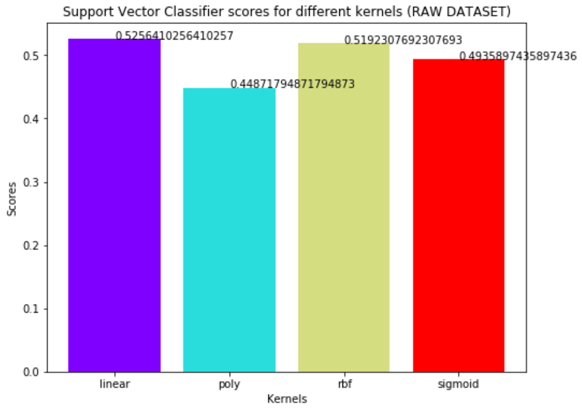


Fig. 12. RAW Dataset Score Analysis for different Kernels of Support Vector Classifier.

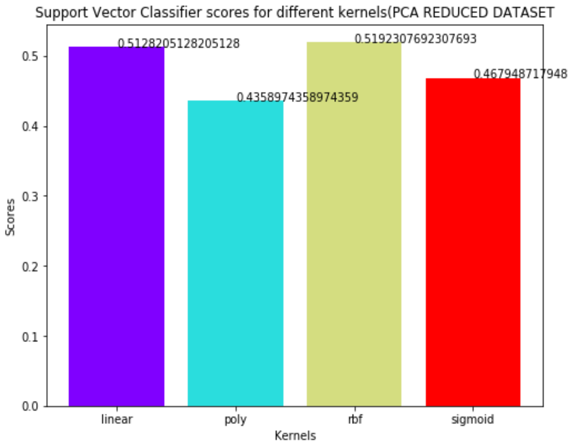
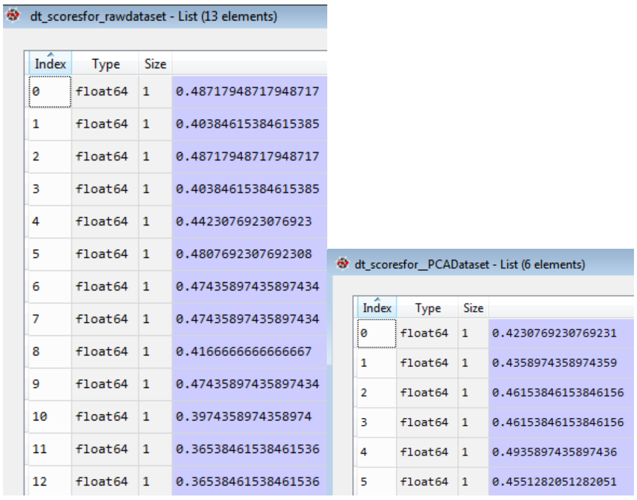


Fig. 13. PCA Reduced Dataset Score Analysis for different Kernels of Support Vector Classifier.

The dataset is fitted to Decision Tree classifier to analyze the performance for the various features with and without PCA. The Scores for the raw and PCA reduced dataset for Decision Tree Classifier for different number of features is shown in fig 14 – Fig 16.



dt_scoresfor_rawdataset - List (13 elements)

Index	Type	Size	
0	float64	1	0.48717948717948717
1	float64	1	0.40384615384615385
2	float64	1	0.48717948717948717
3	float64	1	0.40384615384615385
4	float64	1	0.4423076923076923
5	float64	1	0.4807692307692308
6	float64	1	0.47435897435897434
7	float64	1	0.47435897435897434
8	float64	1	0.4166666666666667
9	float64	1	0.47435897435897434
10	float64	1	0.3974358974358974
11	float64	1	0.36538461538461536
12	float64	1	0.36538461538461536

dt_scoresfor__PCADataset - List (6 elements)

Index	Type	Size	
0	float64	1	0.4230769230769231
1	float64	1	0.4358974358974359
2	float64	1	0.46153846153846156
3	float64	1	0.46153846153846156
4	float64	1	0.4935897435897436
5	float64	1	0.4551282051282051

Fig. 14. Scores for the raw and PCA reduced dataset for Decision Tree Classifier.

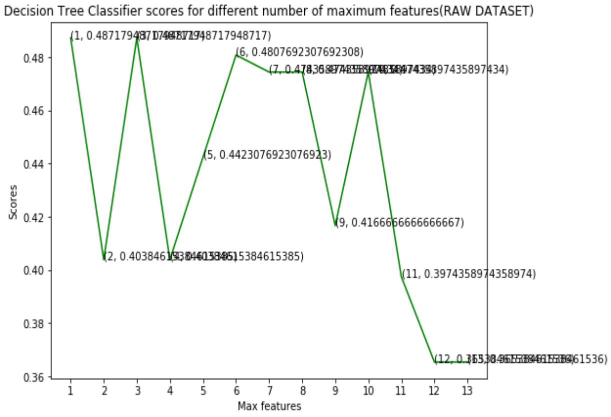


Fig. 15. Raw Dataset Score Analysis for different features of Decision Tree Classifier.

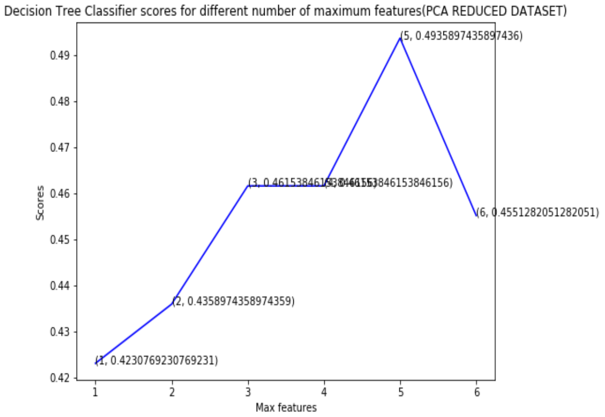
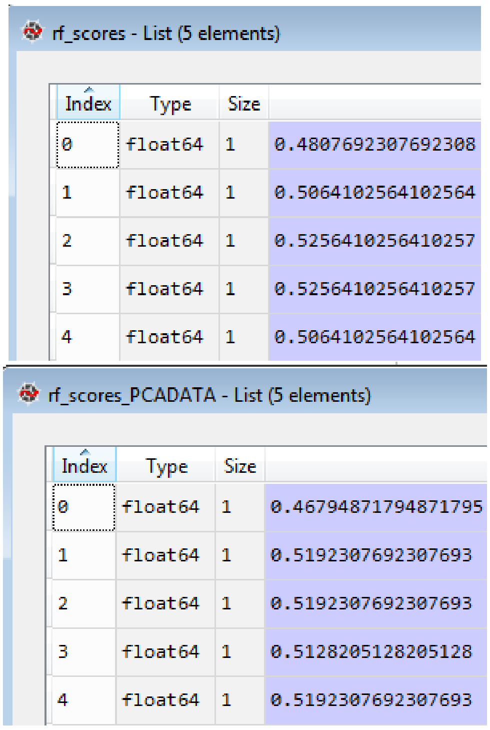


Fig. 16. PCA Reduced Dataset Score Analysis for different features of Decision Tree Classifier.

The dataset is fitted to Random Forest classifier to analyze the performance for the various estimators with and without PCA. The Scores for the raw and PCA reduced dataset for Random Forest Classifier for different number of estimators is shown in fig 17 – Fig 19.



rf_scores - List (5 elements)

Index	Type	Size	
0	float64	1	0.4807692307692308
1	float64	1	0.5064102564102564
2	float64	1	0.5256410256410257
3	float64	1	0.5256410256410257
4	float64	1	0.5064102564102564

rf_scores_PCADATA - List (5 elements)

Index	Type	Size	
0	float64	1	0.46794871794871795
1	float64	1	0.5192307692307693
2	float64	1	0.5192307692307693
3	float64	1	0.5128205128205128
4	float64	1	0.5192307692307693

Fig. 17. Scores for raw and PCA dataset for Random Forest.

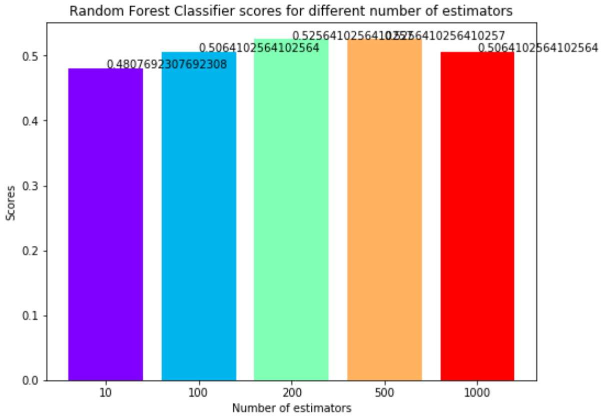


Fig. 18. PCA Reduced Dataset Score Analysis for different features of Random Forest Classifier.

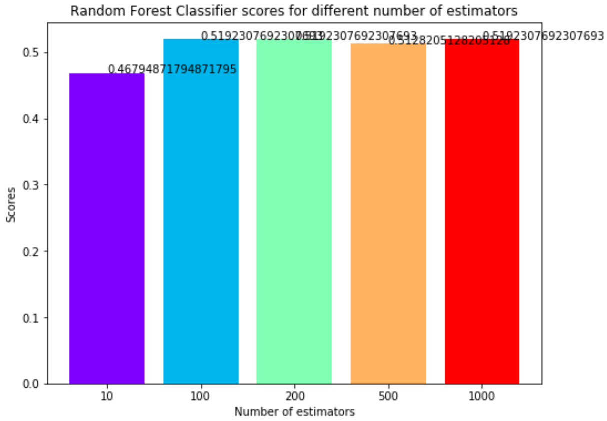


Fig. 19. PCA Reduced Dataset Score Analysis for different features of Random Forest Classifier.

V. CONCLUSION

This paper analyzes the parameters in the classification algorithms. Experimental results shows that for KNN classifier, the performance for 12 neighbours is found to be effective with 0.52 before applying PCA and 0.53 after applying PCA. For Support Vector classifier, the rbf kernel is found to be effective with the score of 0.519 with and without PCA. For Decision Tree classifier, before applying PCA, the score is 0.47 for 7 features and after applying PCA, the score is 0.49 for 4 features. For, Random Forest Classifier, before applying PCA, the score is 0.53 for 500 estimators and after applying PCA, the score is 0.52 for 500 estimators.

ACKNOWLEDGEMENT

Thanks to the Deanship of scientific research princess Nourah Bint Abdul Rahman University for supporting. This research was funded by the Deanship of Scientific Research at Princess Nourah Bint Abdul Rahman University through the fast track Research Funding Program.

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. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
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. M. Shyamala Devi, Shefali Dewangan, Satwat Kumar Ambashta, Anjali Jaiswal, Sairam Kondapalli, “Recognition of forest Fire Spruce Type Tagging using Machine Learning Classification”, International Journal of Recent Technology and Engineering, Volume-8 Issue-3, 30 September 2019.
14. M. Shyamala Devi, Usha Vudatha, Sukriti Mukherjee, Bhavya Reddy Donthiri, S B Adhiyan, Nallareddy Jishnu, “ Linear Attribute Projection and Performance Assessment for Signifying the Absenteeism at Work using Machine Learning”, International Journal of Recent Technology and Engineering, Volume-8 Issue-3, 30 September 2019.
15. M. Shyamala Devi, Mothe Sunil Goud, G. Sai Teja, MallyPally Sai Bharath, “Heart Disease Prediction and Performance Assessment through Attribute Element Diminution using Machine Learning”, International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.11, 30 September 2019
16. M. Shyamala Devi, Rincy Merlin Mathew, R. Suguna, ” Regressor Fitting of Feature Importance for Customer Segment Prediction with Ensembling Schemes using Machine Learning”, International Journal of Engineering and Advanced Technology, Volume-8 Issue-6, pp. 952 – 956, 30 August 2019
17. R. Suguna, M. Shyamala Devi, Rincy Merlin Mathew, “Integrating Ensembling Schemes with Classification for Customer Group Prediction using Machine Learning”, International Journal of Engineering and Advanced Technology, Volume-8 Issue-6, pp. 957 – 961, 30 August 2019.
18. Rincy Merlin Mathew, R. Suguna, M. Shyamala Devi, “Composite Model Fabrication of Classification with Transformed Target Regressor for Customer Segmentation using Machine Learning”, International Journal of Engineering and Advanced Technology, Volume-8 Issue-6, pp. 962 – 966, 30 August 2019.
19. M. Shyamala Devi, Shefali Dewangan, Satwat Kumar Ambashta, Anjali Jaiswal, Nariboyena Vijaya Sai Ram, “Backward Eliminated Formulation of Fire Area Coverage using Machine Learning Regression”, International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.12, 10 October 2019. (Accepted for Publication)
20. M. Shyamala Devi, Ankita Shil, Prakhar Katyayan, Tanmay Surana, “Constituent Depletion and Divination of Hypothyroid Prevalance using Machine Learning Classification”, International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.12, 10 October 2019. (Accepted for Publication)
21. Shakila Basheer, Rincy Merlin Mathew, M. Shyamala Devi, “Ensembling Coalesce of Logistic Regression Classifier for Heart Disease Prediction using Machine Learning”, International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.12, 10 October 2019, pp. (Accepted for Publication)
22. M. Shyamala Devi, Shakila Basheer, Rincy Merlin Mathew, “Exploration of Multiple Linear Regression with Ensembling Schemes for Roof Fall Assessment using Machine Learning”, International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.12, 10 October 2019. (Accepted for Publication)