Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 40 Issue: 06-2021
DOI 10.17605/OSF.IO/D5FPZ

EPMD: EFFECTIVE PREDICTION MODEL FOR DISEASES BY REDUCING THE DIMENSIONS AND APPLYING CLUSTERING WITH DIFFERENT MACHINE LEARNING ALGORITHMS

A. P.Bhuvaneswari[1], Dr. R. Praveen Sam[2] & Dr. C. Shoba Bindu[3]

[1]Research Scholar, Dept. of Computer Science and Engineering, JNTUA University, Ananthapuramu, A.P, India.
[2]Professor, Dept. of Computer Science and Engineering, G.Pulla Reddy Engineering College, Kurnool, A.P, India.
[3]Professor, Dept. of Computer Science and Engineering, JNTUA College of Engineering, Ananthapuramu, A.P, India

Abstract

The field of machine learning need no explicit programming and tries to learn from the given data by identifying the patterns like how humans try to recognize. But at sometimes the humans may make mistakes but with machines the scope is less, and the basic requirement is only to come up with a quality data for training. As the data is generated from multiple sources the available data is in different formats, huge in volume and more unwanted is accumulated making it something as big. For quality results pre-processing must be done because accurate results come from the quality data. Unwanted features which are not required must be deleted to make it a quality one. Feature engineering on big data can result in quality data which when trained with machines will produce the accurate results. In this paper different dimensionality reduction algorithms are used to reduce the dimensions on different datasets and collected the quality results in the identification of diseases. Early identification of disease will help us in taking the necessary protective measures for increasing the life span. Disease datasets with various dependency variables are pre-processed and then features are reduced with different dimensionality reduction algorithms later identifying the similarity in the data points by applying the k_means clustering. The accuracy of the results are tested with different supervised machine learning algorithms for different diseases.

Keywords: Machine learning, dimensionality reduction, feature engineering, Accuracy, Prediction, Logistic regression, Clustering.

1. INTRODUCTION

The era of big data is dominated by five V’s like volume, velocity, variety, veracity, and value. The datasets are represented by huge volume with many rows and features. This huge data contains much of the useful information which needs to be identified. Different sources are used for gathering the data which comes from different varieties and with high speed which takes more processing time. The data needs to be analysed using different unsupervised machine learning algorithms in understanding the different patterns of useful information for different purposes. Large datasets with different features throw a lot of challenges with unavailable data,

Article to Cite:
Bhuvaneswari, Praveen Sam & Shoba Bindu, (2021). EPMD: Effective Prediction Model for Diseases by reducing the dimensions and applying clustering with different machine learning algorithms, Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition), 40(6).
DOI 10.17605/OSF.IO/D5FPZ

outlier data, and different varieties of data which is difficult for machines in understanding the given data. So, the data should be represented in easier format for the understanding of the machine which comes with accurate prediction. Medical datasets are huge with different dimensions which needs to be reduced for storing, processing and for understanding the dataset. Every disease will be having different features for identification purpose. Either we can select the important features by applying feature selection with different statical methods and reduce dimensions or transforming the given data into a new representation with few features for extracting the knowledge. Medical datasets are reduced for extracting useful information in prediction of the different diseases like diabetics, breast cancer, heart disease, much early in life for protective measures in increasing the life span. In medical data field each feature can be valuable in representing the disease.

1.1 Dimensionality Reduction

If the data gathered from different sectors is much big in volume with different features and many rows, then it takes much time to processing which is coined as Curse of Dimensionality. High dimensionality of the data needs to be reduced into lower dimensions for understanding the data in a better way and to extract useful knowledge from the data. Datasets are with many features with missing data and with a lot of duplication which needs to be removed so that machine can understand the given data in an easier and efficient manner. Even machine learning algorithms cannot predict the accurate results with high accuracy if quality data is not provided. Pre-processing of the dataset plays a vital role for the attaining the good results with machine learning algorithms. Dimensionality reduction can be done in two ways i.e either by selecting the specific features and removing the irrelevant and the other way by summarizing the exiting data by extracting the new components of representation of the available data by covering the maximum variance.

1.2 Machine Learning Algorithms

Machine learning is a branch of data science basically used for the analysing the huge data for accurate results. The useful data will be extracted from the given dataset and is used in future. Machine learning algorithms also fails if the dataset is huge with more dimensions. The high dimensional data is reduced into lower dimensions with different machine learning algorithms. Later the extracted data is divided into training and test datasets in the ratio of 70% and 30% and tested for the results. The dataset should be pre-processed for getting the quality results. The different features of the dataset need to be represented in a similar format by applying standard scalers for the better understanding of the machines. Later different supervised algorithms can be used for calculating the accuracy of the results and for prediction of diseases accurately.

The medical dataset has different features for classifying the presence of disease. The data must be pre-processed by removing the noisy data, outliers, missing data, duplicated data, and later it should be scaled. Now the obtained pre-processed data is to be reduced into lower dimensions for extracting the knowledge. Different supervised machine learning algorithms are applied for classifying and predicting the presence of the disease with good accuracy.

The rest of the paper is structured as follows. Section 2 deals with the recent related works. Section 3 gives an account of the proposed methodology. Section 4 provides the experimental results and Section 5 gives the conclusion.

2.RELATED WORKS

There are several algorithms that have been developed to predict the presence of diseases like heart diseases which can cause major damage to the life if not detected early in the lifetime. Some of the recent related works are reviewed in this section.

Harshit Jindal et al. [1] predicts the presence of the heart disease by taking UCI repository heart dataset with 14 attributes and with classification algorithms likes RF,KNN and LR. Among them KNN outperforms with highest accuracy of 88.52% when compared with the previous models. After data collection the dataset is represented with the significant values and pre-processing is done for the better-quality data and finally splitting of data into training and test and calculating the accuracy by classification algorithms is done and a fine comparison of the results is done with the specific variables.

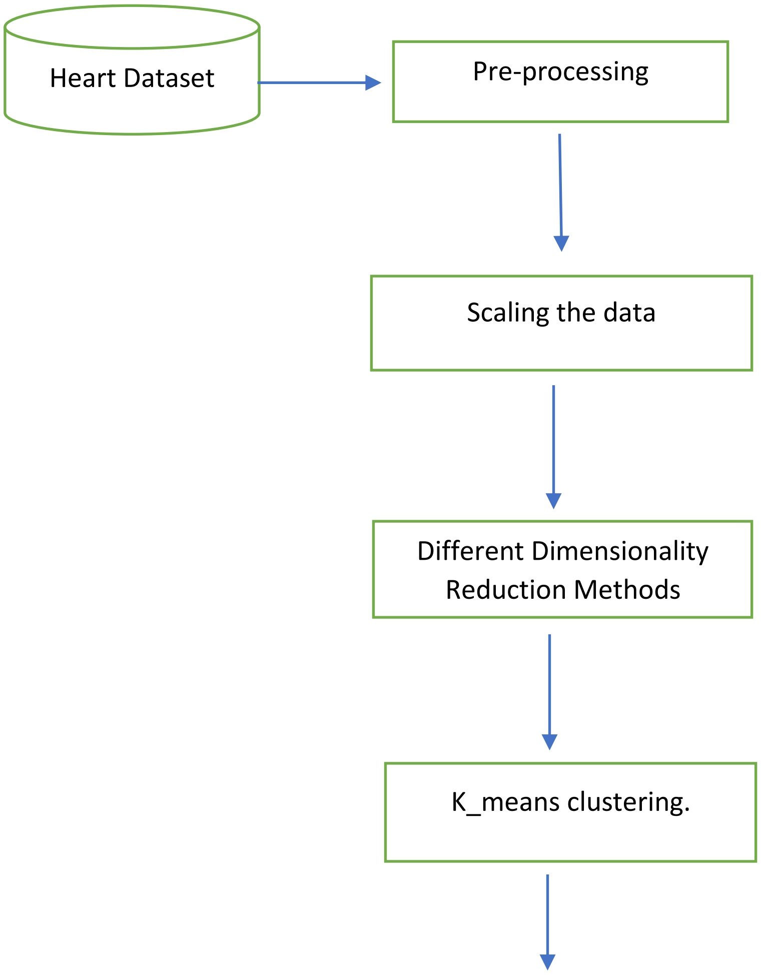
Aradhana et al. [2] identifies the cardiovascular disease presence using various machine learning algorithms like KNN, LR, RF, DT, NB. The dataset description is provided with the possible values for the attributes. The dependency factor is calculated by the correlations like spearman, Pearson and Kendal and found the strong relation between the variable by spearman. Based on the correlation results later the data is trained, and accuracy is tested with five cross-validation methods and finally based on ROC graphs gaussian naïve bayes have shown good results.

Apurb Rajdhan et al. [3] highlights model by having a comparative study among the different machine learning algorithms by using the Cleveland dataset. Four data mining algorithms are used in predicting the heart disease. Among the four algorithms random forest is giving the high accuracy in prediction of heart disease with an accuracy of 90.16%. Identified the features with its distinct values which are considered for identifying the presence of problem.

Indra Kumari *et al.* [4] Proposes an effective heart disease prediction model by performing the exploratory data analysis in identifying the risk factors and the correlation between the features and applying k_means with heart data and the results are represented by tableau. The Cleveland dataset is reduced to 209 with 8 attributes which helps in calculating the accuracy for prediction of occurring heart attack.

Li yang *et al.* [5] suggest the possibility of cardiovascular disease in eastern china for around 101056 patients. Random Forest algorithm prediction gives high accuracy in predicting the presence or absence of cardiovascular disease compared with different other models with AUC of 78.71%.

3. PROPOSED METHODOLOGY



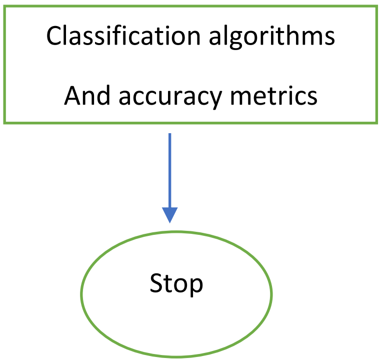


Figure 1: Effective Prediction Model for Diseases.

The proposed heart prediction model uses the Cleveland heart dataset which is available in UCI repository with 303 patient’s data with 14 featured columns. The dataset is Pre-processed for checking the missing values and after that the data will be scaled with the standard scaler for obtaining the uniqueness in the data. Dimensionality reduction is about reducing the available features in identifying the patterns available in the data. The heart dataset is tested with the different dimensionality reduction algorithms for extracting the valuable data from the existing data and classified with different algorithms for achieving the highest accuracy in predicting the presence of the heart disease.

3.1 Different Dimensionality Reduction Algorithms

Dimensionality reduction is reducing the dimensions of the given dataset for the effective visualization and the identification of the patterns present in the data. Feature engineering can be done either as feature selection or feature extraction. Feature selection is selecting the important features based on some statistical measures and then classifying the data and testing the accuracy with different algorithms. Feature extraction is extracting the data without loosing maximum of information and reducing the data by different dimensionality reduction algorithms like PCA [6], LDA, ICA, NMF, LLE, SVD, TNSE etc. and applying the different classification algorithms for checking the highest accuracy in predicting the heart disease prediction.

4.EXPERIMENTAL RESULTS

The results show the different accuracies with different supervised classification algorithms after reducing their dimensions by different dimensionality reduction algorithms. The common myth regarding heart attack is the age factor.

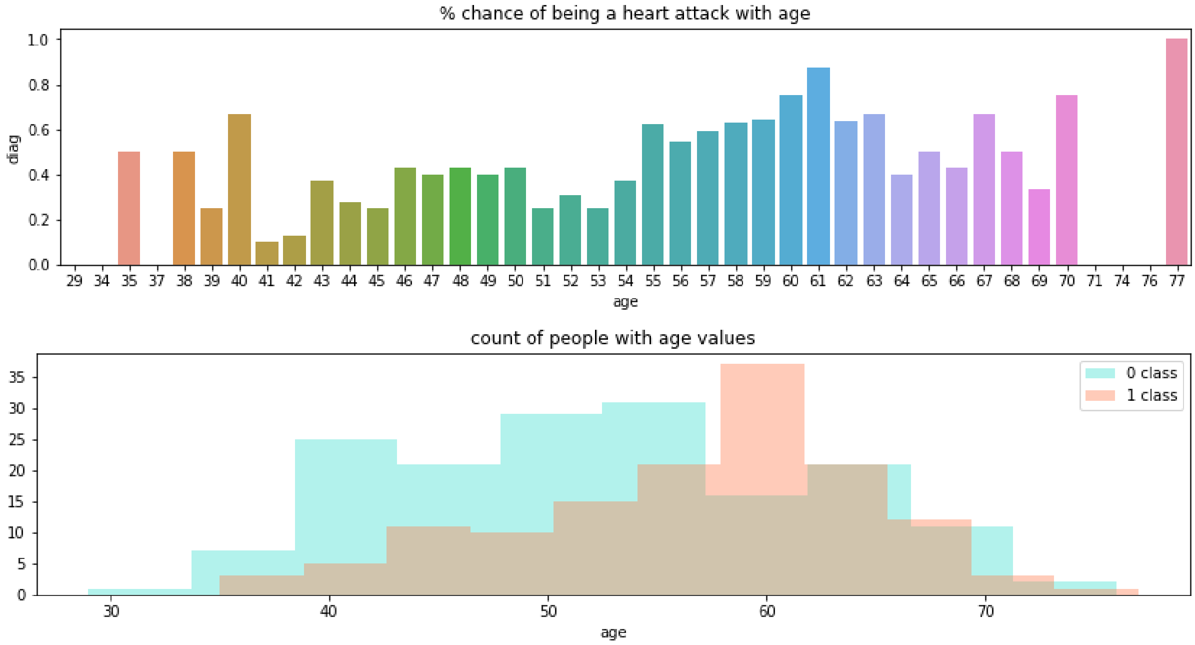


Figure 2: Visualization of heart data based on age factor.

Algorithm	PCA	LDA	SVD	LLE	TNSE	NMF	KPCA	FastICA
KNN	65	82	59	56	56	57	65	68
SVC	54	85	50	51	51	56	54	51
LSVC	52	86	50	50	50	59	65	64
LR	67	86	58	51	51	59	67	60
DTC	62	74	54	53	54	51	63	61
GNB	64	85	58	50	50	59	64	64
RFC	67	79	57	58	56	61	59	62
GB	63	80	53	57	57	58	63	64

Table 1: Different dimensionality reduction algorithms with accuracies.

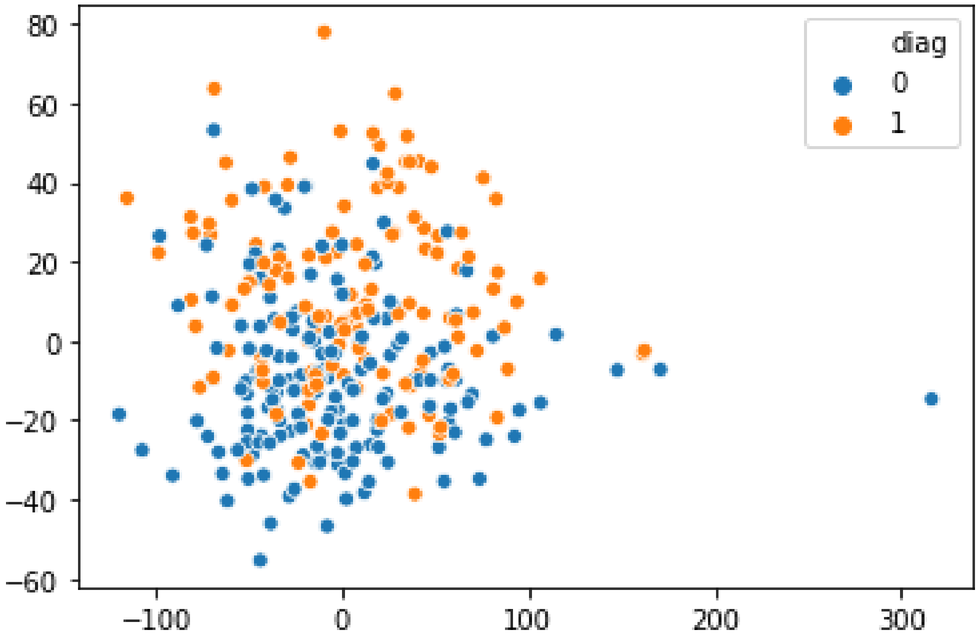


Figure 3: Visualization after reducing dimensions.

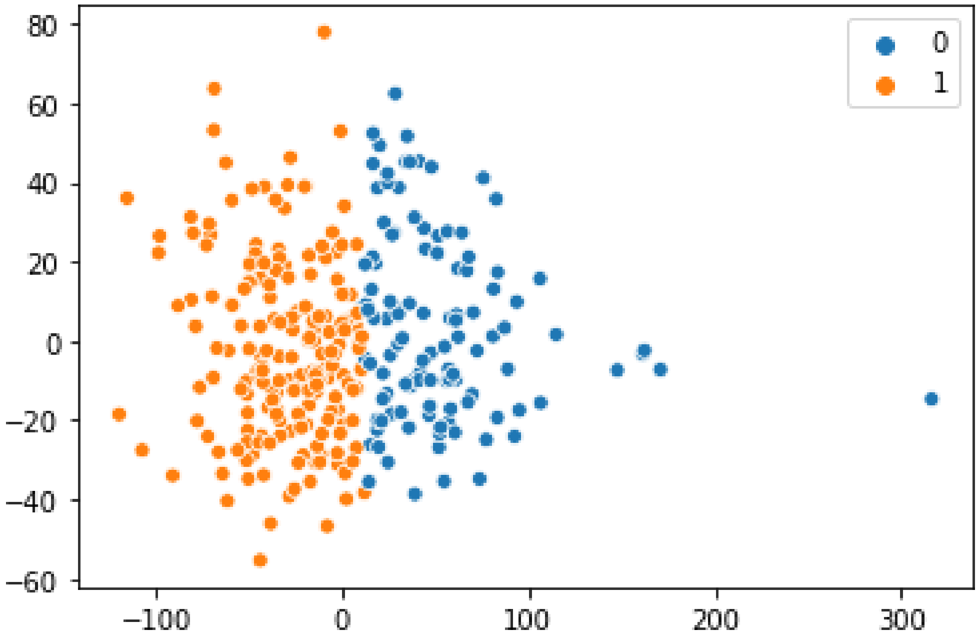


Figure 4: Visualization after dimensionality reduction and clustering.

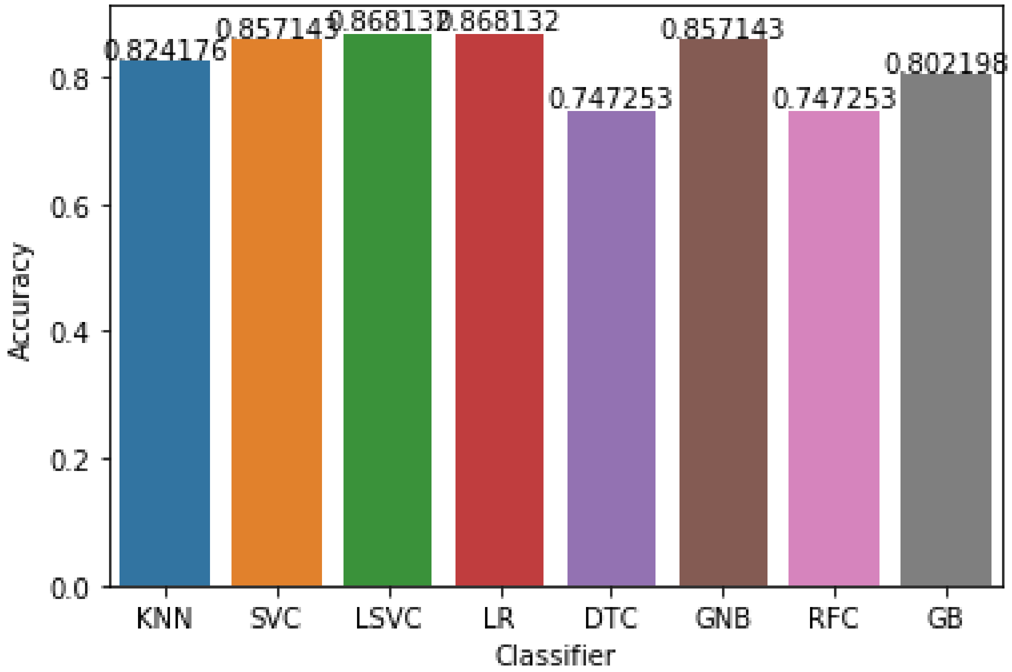


Figure 5: Classifiers accuracy for heart data with different dimensionality reduction algorithms where LDA + LR has good accuracy.

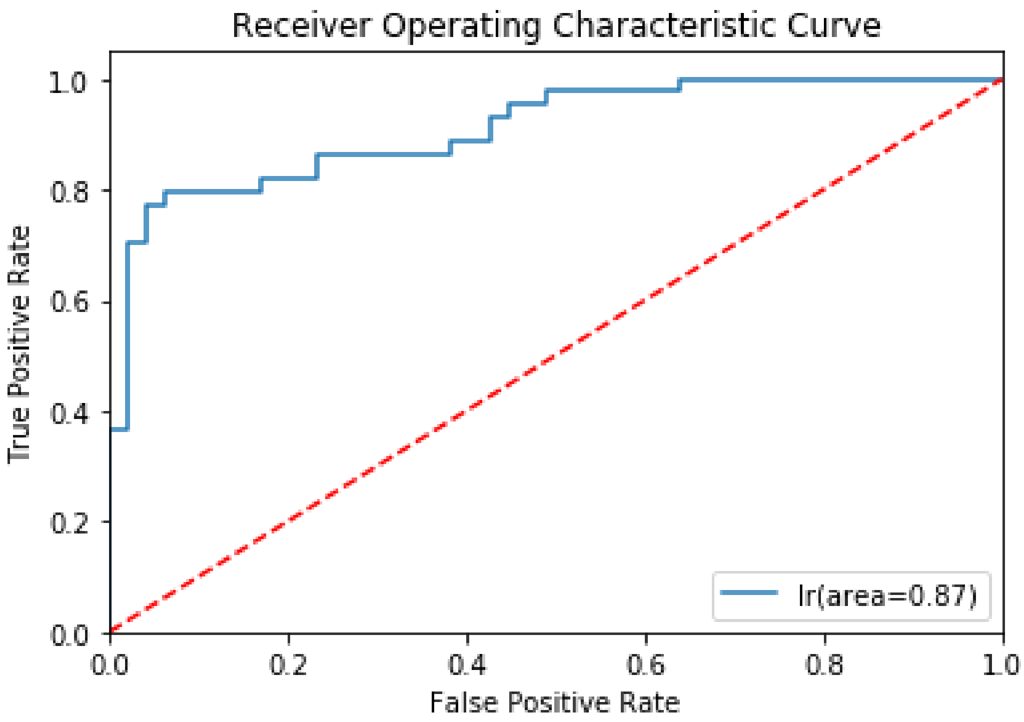


Figure 6: ROC curve for heart dataset without clustering.

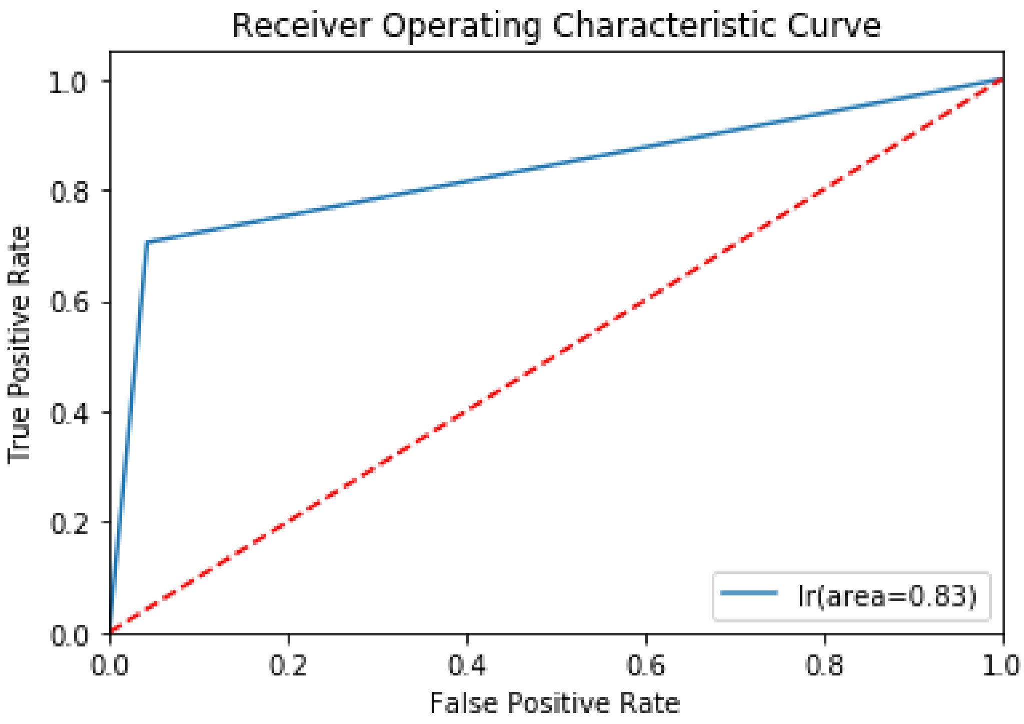


Figure 7: ROC curve for heart dataset with k_means clustering.

5. CONCLUSION

Heart attack is the common disease now a days throughout the world and as soon as it attacks the chance of life well is very less usually. Most of the times the people are not aware that will undergo heart attack at some time and by the time they release it will be too late. So detecting the disease much early in the life is mandatory. Based on some of the features the prediction of the chances of heart disease can be identified. The machines if trained with the quality can predict accurately the presence of disease. For that dimensions more in number are reduced in Cleveland heart dataset with different algorithms and finally tested the accuracy. Among the different dimensionality reduction methods LDA combined with logistic regression is predicting with high accuracy before and after clustering when compared with other algorithms. LDA algorithm have attained a good accuracy when compared with other dimensionality reduction algorithms. This can also be applied to different major diseases like breast cancer and diabetes in future for knowing which dimensionality reduction will give best accuracy.

6. REFERENCES

[1] Harshit Jindal, Sarthak Agarwal, Rishabh Khera, Rachna Jain and Preeti Nagrath. "Heart Disease Prediction Using Machine Learning Algorithms." Materials Science and Engineering,2021.

[2] S Aradhana, P Jankisharan, SK Virendra and M Ashish. "Cardiovascular Diseases Prediction using Various Machine Learning Techniques." Materials Science and Engineering,2021.

[3] Apurb Rajdhan, Milan Sai, Dr.Poonam Ghuli . "Heart Disease Prediction Using Machine Learning." International Journal of Engineering Research and Technology." April 2020.

[4] R. IndraKumari , T . Poongodi, Soumya Ranjan Jena. "Heart Disease Prediction Using Exploratory Data Analysis." Procedia Computer Science,2020.

[5] LiYang1,4,5, HaibinWu2,5, Xiaoqing Jin3, PinpinZheng4, Shiyun Hu1, XiaolingXu1, WeiYu1 & JingYan. "Study of cardiovascular disease prediction model based on random forest in eastern China." Nature research scientific reports,2020.

[6] Devansh Shah, Samir Patel, Santosh Kumar Bharti. "Heart Disease Prediction Using Machine Learning Techniques." Springer Nature Singapore Pte Ltd 2020.

[7] Mohammad Reza Mahmoudi, Shahab S. Band. "Principal component analysis to study the relations between the spread rates of COVID-19 in high risks countries." Alexandria Engineering Journal , Vol.60 , pp.457-464, 2020.

[8] Hager Ahmed,Eman M.G. Younis,Abdeltawab Hendawi,Abdelmgeid A. Ali. "Heart Disease Identification from Patients' Social Posts, Machine Learning Solution on Spark." Elsevier,2019.

[9] Md. Abu Bakr Siddique1, Shadman Sakib2, Md. Abdur Rahman3. "Performance Analysis of Deep Autoencoder and NCA Dimensionality Reduction Techniques with KNN, ENN and SVM Classifiers." International Conference on Innovation in Engineering and Technology,2019.

[10] Albert Nguessan Ngo, David Joseph Turbow. "Principal Component Analysis of Morbidity and Mortality among the United States Homeless Population: A Systematic Review and Meta-Analysis." Public Health and Community Medicine, 2019.

[11] Jianqing Fan, Qiang Sun, Wen-Xin Zhou, Ziwei Zhu . "Principal component analysis for big data." Researchgate,2018.

[12] Santhanam T, Padmavathi MS. Application of K_means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. Procedia Comput Sci 2015;47:76-83.