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. Fourthly, 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 neighbors 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.

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

Fig. 1 System Architecture

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

<table>
<thead>
<tr>
<th>Chest Pain Format</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Typical Angina</td>
</tr>
<tr>
<td>2</td>
<td>Atypical Angina</td>
</tr>
<tr>
<td>3</td>
<td>Non - Anginal Pain</td>
</tr>
<tr>
<td>4</td>
<td>Asymptotic</td>
</tr>
</tbody>
</table>

Fig. 2. Chest Pain Type Distribution

The expansion of the dataset attributes shown in Fig. 3.
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. 3. Heart Disease Dataset Schema

Fig. 2. Important Features of Heart Disease Data Set
Exploration of Neighbor Kernels and Feature Estimators for Heart Disease Prediction using Machine Learning

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.
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.

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.

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.

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.
V. CONCLUSION

This paper analyzes the parameters in the classification algorithms. Experimental results show 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