May - June 2020
ISSN: 0193-4120 Page No. 4955-4962

Heart Disease Classification: A Feature Engineering Approach

[1]S Mirudula, [2]Akram Pasha, [3]Sathya Rupa M, [4]Ritu Shenoy

[2]Professor, [1,2,3,4]C & IT, Reva University, Bangalore, India

Article Info
Volume 83
Page Number: 4955-4962
Publication Issue:
May - June 2020

Abstract

The healthcare applications are in the demand for rigorous medical data analytics algorithms. Machine Learning (*ML*) has taken a leading role in various data analytics field including healthcare applications. The ML algorithms are influenced by the various features of the medical data sets and eventually contribute to enhance the accuracy of classification of the diseases. In this paper, an attempt is made to experiment the level of influence of the various features of the Heart Disease Data set (*HDD*) through both feature selection and feature extraction techniques to enhance the classification accuracy of the various ML algorithms. Six ML classification algorithms have been deployed such as k-Nearest Neighbor (*kNN*), Decision Tree (*DT*), Gaussian Naive Bayes (*GNB*), Logistic Regression (*LR*), Support Vector Machines (SVM) and Random Forest (*RF*) in this study. The HDD consists of 303 records with 14 attributes of165 patients being tested on heart disease. The HDD was normalized and partitioned as Training and Testing sets in the ratio of 0.8 and 0.2 before training the ML classifiers. After scaling, it was observed that there was a hike in the accuracy of the SVM Classifier from 65% to 87% which is the highest compared to all other models. Weightage of all the attributes has been computed using RF-based feature importance. The Principal Component Analysis (*PCA*) based SVM was found to give the highest accuracy of 90.16% among all the classification models employed in the study.

Keywords: *Classification; Machine Learning; Feature Selection; Feature extraction; PCA*

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *16 May 2020*

1. Introduction

Across the world, the number one death cause is cardiovascular disease that is estimated to take about 17.9M lives every year. This is acknowledged by the World Health Organization (WHO). Out of these death counts, 85% are due to stroke and heart attack. The hazardous factors of heart disease are an imbalanced diet, lack of physical activities, usage of tobacco, alcohol, hoisting in blood pressure, glucose, blood lipids and overweight.

The enormous data generated in the field of biomedicine can be used to find the solution to prevent cardiac diseases, one such way is early detection of the disease to decrease the mortality rate of human [5]. Common heart attack signs and symptoms include tightness, fatigue and aching sensation in your chest or arms that may spread to your neck, jaw or back [16]. The most probable types of cardiac diseases are Atrial Fibrillation, Cardiomyopathy, Congenital heart defects, cardiovascular disease, Irregular Heart Rhythm, Arrhythmias and Coronary artery disease [5]. The long-established approaches to detect heart diseases are cardiac magnetic resonance imaging, blood test, cardiac computerized Tomography scan, electrocardiogram but these methods are invasive and time-consuming [21]. The information thus gathered is further processed by the health care department to find the important attributes, which contribute majorly to the diagnosis of the disease. In the health area, data mining techniques play a vital role to discover the pattern in the detection of the diseases.

May - June 2020
ISSN: 0193-4120 Page No. 4955-4962

Classification categorizes feature vectors based on some criteria and uses it into the data that depends on the training set and evaluation of distinct values [10]. ML is the widely available tool in various domains, as a different algorithm is not required for a different dataset but the efficiency of the algorithms might differ. It stimulates the use of different possible methods like decision trees, statistics, linear programming, etc. Feature engineering is a practice of using the realm knowledge to reduce the complexity of the data [17]. It is a crucial but work-intensive aspect of machine learning applications in terms of time and expertise. An endorsed set of attributes from the data set are recognized and extracted for model evaluation [8]. So, choosing the features that have more importance is a prerequisite that must be taken care of.

The major contribution is to develop a classification model employing features, aims and objectives of the study that summarizes the classification technique that the analysts have prevailed upon in the domain of healthcare. This also helps medical practitioners to diagnose new patients in their preliminary stages with higher speed and non-specialists to diagnose patients [14].

This illuminates the vitality of ML in the healthcare discipline and how it can perform the accurate predictions and aids professionals from the medical field [19].

The subsequent sections of the paper are as follows. Section- II contains the Related Work of the different survey that is being conducted that has been a motivation of this paper. Section-III involves the various materials and methods that are being deployed. Section-IV is the discussion of the experimental results that are being carried out to compare and contrast the performance evaluation. Section-V contains the Conclusion and Future Outlook.

2. Related work

The literature review conducted so far concentrates on the appropriate work in the field of HDD availing the Feature Engineering Method. Referencing is done through the literature warehouse while combining the work done by various researchers. The studies determined that the various publications have challenges and analytical technique being deployed in their respective field. With different objectives and intentions, studies show that a lot of challenges have been overcome using ML algorithms and dimensionality reduction technique. The primary objective of this study is to determine the optimal subset of the features that contribute to maximal classification accuracy of an ML classifier, classifying the HDD.

For effective prediction, ML Algorithms are implemented for various dreadful disease outbreaks in different communities. Analysis of the accuracy is decreased when the data set has missing values [1]. The classifier implemented in the model using the preprocessing technique is a pivotal part to assemble the data set deployed by various ML classifiers to yield better outcomes [3]. Advantage has been noticed that once an algorithm learns the pattern in the data, the following tasks can be automatically carried out [7]. The other option is to create a ML pipeline that can recollect the same order and a complete set of preprocessing steps [17]. A Feature Vector is used in feature engineering model that could be expanded by adding contemporary features that are calculated based on the other based on other features[17]. In numerous ML tasks scaling the feature value is an important step [21]. A strategy that has been proven to be more effective especially in high dimensional data for different data mining or machine learning problems is feature selection. The aim is to build a comprehensive model [4]. Data mining is being used to reduce the number of tests that are being performed on the patients. The extraction of the hidden pattern is the aim to predict the presence of heart disease in the patient [12]. One such way of achieving the desired prediction with high accuracy is ML.

This makes use of various tools that utilize feature vectors and its various data types under the varied condition for prediction [14]. The analytical approach is determined to improve the accuracy of the poor classification algorithm and the implementation of the algorithm that helps in the detection of heart disease in the early stages [15]. In the work of [18], the problem of dimensionality reduction was addressed using the Bio-inspired approaches to find the optimal subset of features contributing to the highest classification accuracy for classifying Parkinson's disease. It is possible to use a Bio inspired algorithm for the highest classification of diseases in Feature Engineering in order to find the best set of features for maximal efficiency. Various survey papers have an overview of how the researchers have implemented the ML classification technique in the domain of detection of Heart Disease [3].

3. Methods and Materials

The methodology involved is shown in Fig -1. Here, an input i.e., an HDD is converted to a data frame, it is mandatory to split the data set for the evaluation of the model. The following step is the feature scaling where a MinMaxScaler is being deployed. We compare the output with and without dimensionality reduction and is evaluated. While building a classification model, it is necessary to train the classification model using various classifiers to obtain the best performing model. Finally, we evaluate the classification model using the test data set. The classifier that gives the best performance is used for further comparisons and computations of the results.

May - June 2020
ISSN: 0193-4120 Page No. 4955-4962

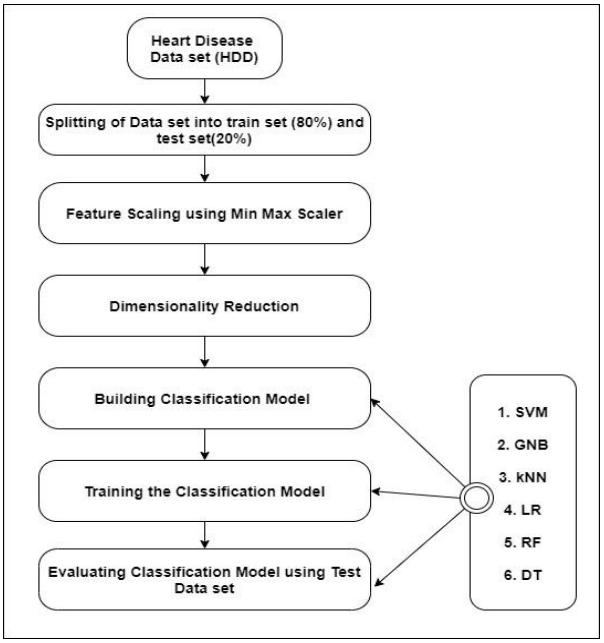

Figure 1: Proposed Framew

Data set Description

The root cause of the rise in the number of deaths worldwide is heart disease as referenced by WHO. Many experimental theses have already been conducted to predict the disease before it costs a person's life. The HDD chosen for the experimentation has been taken from Kaggle repository. The description of the data set is given in Table -1. The data set contains 14 attributes and 303 patient records, out of which 207 patients are found to be women and 96, are men. The selected attributes are the useful insight of the dataset. It is essential to keep in mind to avoid lateral features during this process. Each feature should only improvise the information contained in the dataset.

From the dataset, it was noted that 165 patients were more likely to have heart disease and 138 patients were healthy. Out of those 165 patients, 72 were found to be women and 93 were men. The whole data collection process was conducted as suggested by the experienced medical supervisor. Out of those 14 attributes, 4 attributes have binary values namely‘sex’, ‘fbs’, ‘exang’ and ‘target’. One attribute, i.e. old peak was found to have floating values. The decisive attribute in this data set having binary value is represented by the target variable. This variable can be used by a ML classifier to extract the required information of a patient suffering from heart disease.

Table 1: HDD Description

Seri al No.	Attribute	Description
1	age	age in years(29-77)
2	sex	Gender(1 = male; 0 = female)
3	cp	chest pain types: typical angina, atypical angina, non-angina pain, asymptomatic
4	trestbps	Resting blood pressure (in mmHg on admission to the hospital)
5	chol	serum cholesterol in mg/dl
6	fbs	Fastingbloodsugar120 mg/dL (1 = true; 0 =false)
7	restecg	Resting electrocardiographic results (values 0, 1, 2)
8	thalach	Maximum heart rate achieved (71-202)
9	exang	Exercise induced angina (1 = yes; 0 = no)
10	old peak	ST depression induced by exercise relative to rest
11	slope	The slope of the peak exercise ST segment (upsloping, flat, downsloping)
12	ca	Number of major vessels (0–3) colored by fluoroscopy
13	thal	Represent Thalassemia which is an inherited haemoglobin disorder.3 = normal; 6 = fixed defect; 7 = reversible defect
14	target	Diagnosis of heart disease (angiographic disease status: 0= absence; 1 = presence)

TEST Engineering & Management

May - June 2020
ISSN: 0193-4120 Page No. 4955-4962

Data Set Split

When working with a data set to obtain an efficient classification model, training the model is required. Thus, a divergent training set is necessary. In the evolution period, it is implausible to have an enormous amount of raw facts. Here, a better decision is to partition the dataset which we have into two sets, one for training and the other for testing. A tool called Model Selection from the Scikit Library aids the task. Train-test-split is a class in this library. This is usually used to split the dataset in various proportions. Consideration of different parameters can be noted such as test_size, train_size, random_state, etc. Splitting is done in such a way that it is partitioned into mutually exclusive datasets. x is an independent variable and y is dependant variable. x is split into two sets xTrain and xTest. Similarly, splitting for y takes place. Accordingly 80.00% is used as a training set and 20.00% as a testing set.

Feature Scaling

The HDD chosen in this study is unfettered from missing values but it consists of distributed data which might result in the erroneous analysis of the data. Feature scaling is a technique to normalize independent attributes present in the data set. It is performed as a preprocessing step to deal with fluctuating values. While analyzing, Data is generalized within a range without modifying the quality of the data and it promotes computation at a faster rate. It is essential for ML classifiers which majorly depend on distance. There are few ML algorithms such as kNN and SVMfor which there is a need for scaling before being fed as input as they are sensitive to attribute transformations. In this study, MinMaxScaler has been employed to scale the features, this technique preserves the shape of the data set and also re-scales the attribute value with distribution value in the range of 0 to 1.The scaled features are then trained to ML classifiers such as KNN and SVM for a better performance rate using the Eq-1.Where $x=(x_1,\ldots x_n)$ and $X_{new}$is the rescaled value, $X_i$is the original value, min(x) is the minimum value in future, and max(x) is the maximum value in future.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

Eq-1: Formula used in MinMaxScaler

Feature Importance and Feature Selection

It is essential to build a model that can be interpreted conveniently many real-time cases. Feature importance helps in a better understanding about the logic of the model which helps to improve the efficiency of the model. Here, we can assess how much each attribute is contributing to reduce the weighted impurity. A model is designed to predict if the patients are prone to have heart disease. Identifying which attributes are important can help us in early detection and improves the quality of service. The phenomenon of selecting attributes that help in predicting response conveniently is feature selection. This is an availed technique to recognize the set of attributes from the dataset which are consequential when building models. The several reasons to perform feature selection are simplification of the model, to avoid the Curse of Dimensionality and enhanced generalization by reducing overfitting. Any straightforward algorithm helps to decrease the error rate in the model. The importance of feature selection is that it allows the ML classifier to train at a faster rate. Analyzing the model is made easier with feature selection which is an underlying cause. Filter is one such known method to perform feature selection among various other methods. This category of feature selection is known to be rapid to compute, while it is still capturing the productiveness of the featureset.

Feature Extraction

The key logic behind implementing feature extraction is to reduce the number of features in the data set by innovating new features from the existential features. This approach produces improved accuracy, enhanced speed in training, improved data visualization, etc. The size complexity of the dataset can be minimized using the technique called PCA. Here, we take our original set of attributes and find the possible combinations of the input attributes which can best summarize the original data distribution to scale down its original dimension. For better behavior, Normalization of the data is carried out. All attributes will have equivalent standard deviation and it escalates the variance. For standardization of the features, standard scalar function has been adopted. Explained-variance-ratio explores how much of the original data variance was preserved.

4. Experimentation and Discussion of Results

Experimental Setup

The complete work was developed on a computing platform with the specifications of 2GB RAM, 1TB ROM and i3 processor. The software packages used in this work are Pandas, NumPy, Scikit Learn, Matplotlib, Seaborn and XGBoost. All these packages were employed on a web-based IDE for Python 3.7 version.

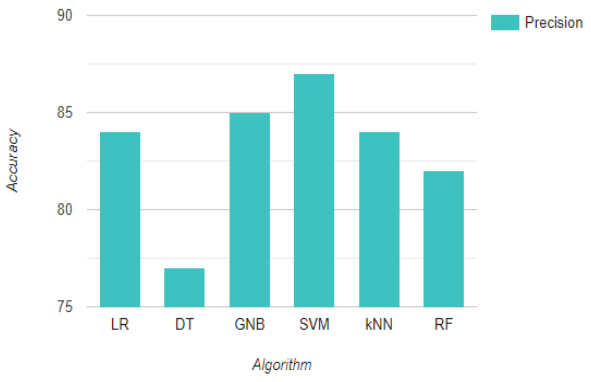

Figure 2: Performance Evaluation of Different Classifiers

May - June 2020
ISSN: 0193-4120 Page No. 4955-4962

In the Fig-2, the performance of different classifiers is being compared. Following are the results that have been obtained. LR has an accuracy of 84%, DT has about 77%, GNB is 85%, SVM is 87%, KNN is 84% and RF is 82%.These results are obtained after performing feature selection. SVM classifier has outperformed other classifiers with an accuracy percentage of 87.

Confusion Matrix

The confusion matrix is a table that is used to describe the performance of the classification model on a data set. It is a table with four different combinations of predicted and actual values. The four different sections are True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN).

TP: prediction is positive and it is true. TN: prediction is negative and it is true. FP: prediction is positive and it is false. FN: prediction is negative and it is false.

We usually describe the predictive values as positive and negative and the actual values as true and false.

Recall = TP / $(TP + FN)$
Precision = TP/ $(TP + FP)$
Accuracy = $(TP + TN)$/ $(TN + FP + FN + TP)$
Error rate = $(FP + FN)$/ $(TN + FP + FN + TP)$

The general outline of the confusion matrix is represented in the Fig-3. The maximum value that recall and precision can hold is one.

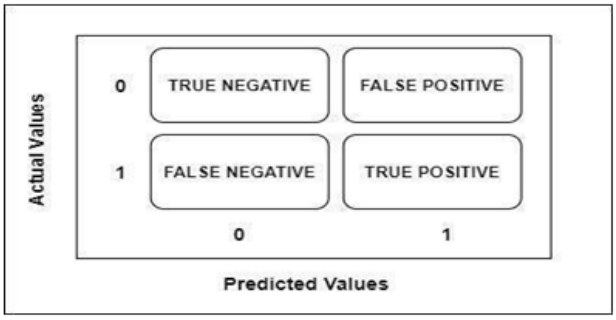

Figure 3: Confusion Matrix

Feature Importance

The result obtained after the experimentation of feature importance is shown in Fig-4. It was observed that the attribute 'thalach' which stands for maximum achieved heart rate is playing a major role with the importance value of 0.1618 and then 'thal' with a value of 0.1213followed by ‘ca’ with the value of 0.1169 and 'trestbps' with 0.0943as their importance value. 'slope', 'restecg' and 'fbs' have the least importance with the value of 0.0409, 0.0214 and 0.0165 respectively. This weightage is very useful for medical practitioners to deal with the different kinds of cases which they might encounter in the near future.

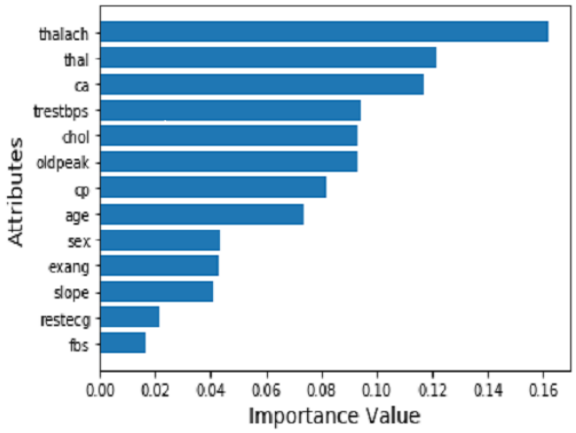

Figure 4: Random Forest Feature Importance

Feature Selection

With reference to Xticks graph in Fig-5, It was observed that the attribute ‘thalach’ has the selective importance of 0.157904 with the feature ranking of ‘1’. The attribute ‘thal’ with selective importance of0.124128 has the feature ranking of ‘2’ followed by the attribute ‘ca’ with a selective importance of 0.118652 and feature ranking of ‘ 3’. The succeeding attributes are‘cp’ with selective importance of0.110772, ’chol’ with selective importance of 0.0902,’age’ with selective importance of 0.07824,’old peak’ with selective importance of 0.073005, ’trestbps’ with selective importance of 0.069764.’sex’, ’slope’, ’exang’, ’restecg’ , ‘fbs’ have selective importance of 0.055914, 0.04379, 0.03465, 0.02503, 0.017889 respectively. The performance of the 14 attributes yields a percentage of 87% which is the same as that of the first five attributes. Henceforth, the dimensionality reduction technique is proven to be more effective.

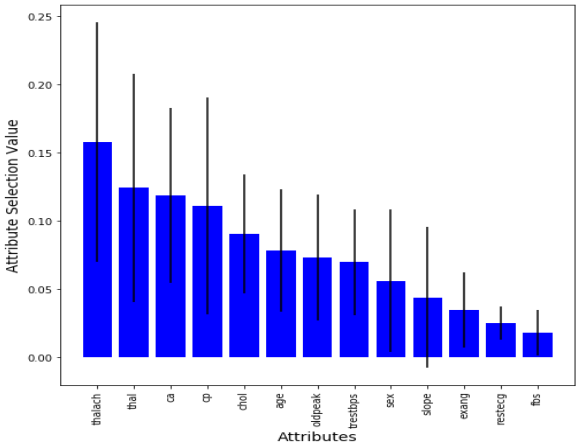

Figure 5: Tree-Based Feature Selection

Feature Extraction

The experimentation of feature extraction using PCA resulted in two elements namely Principal Component 1 and Principal Component2. The explained variations per

principal component is 0.7475, 0.15037. There was a hike in the accuracy percentage of the SVM model from 87% to 90.16% after performing feature extraction. The Scatter Matrix demonstrates the impact of Principal Component 1 and Principal Component 2 of the target variable as shown in the Fig-6.

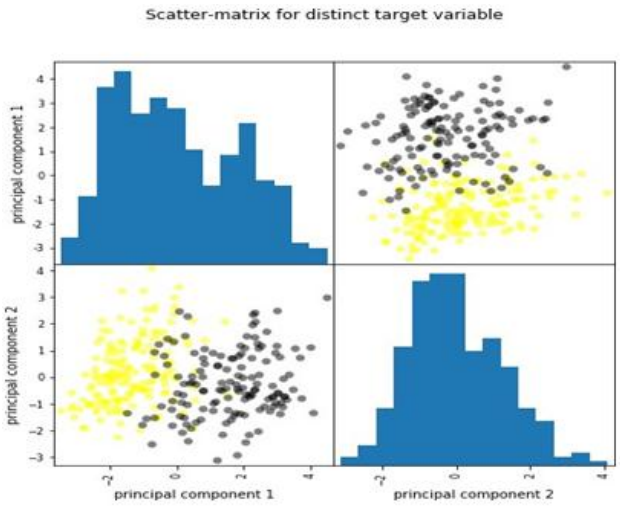

Figure 6: Feature Extraction for distinct target Variable

The performance of the SVM classifier has been compared with respect to the accuracy rate as shown in the Fig-7. It was observed that the outcome was insignificant compared to other classifiers before scaling. The performance increased to 87% after feature selection. There was a significant increase in the accuracy rate after feature extraction with the accuracy percentage of 90.16%. This demonstrates that feature extraction yields a better result when compared to feature selection.

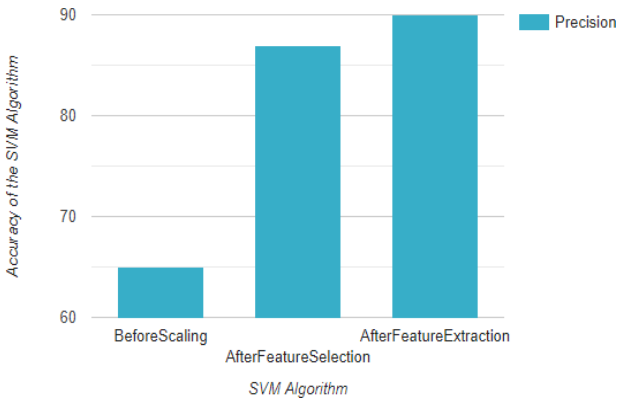

Figure 7: Performance Evaluation of SVM Classifiers

Performance Evaluation using Confusion Matrix

A Confusion Matrix has been implemented for the three different scenarios as shown in Fig-8, Fig-9 and Fig-10 to compare their efficiencies. Precision and Recall has been computed as a comparative measure for the matrices. Precision is an important measure of quality and Recall is an important measure of quantity. Noticeable result is gathered when compared and contrasted.

- Before Feature Selection:

The Precision value is 0.86 and the Recall is 0.95

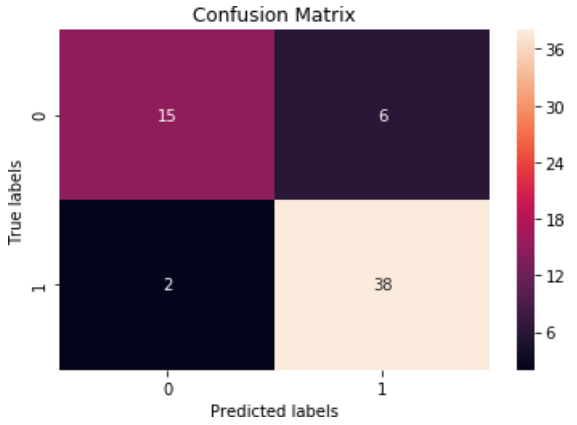

Figure 8: Before Feature Selection

- After Feature Selection:

The Precision value is 0.93 and the Recall is 0.93

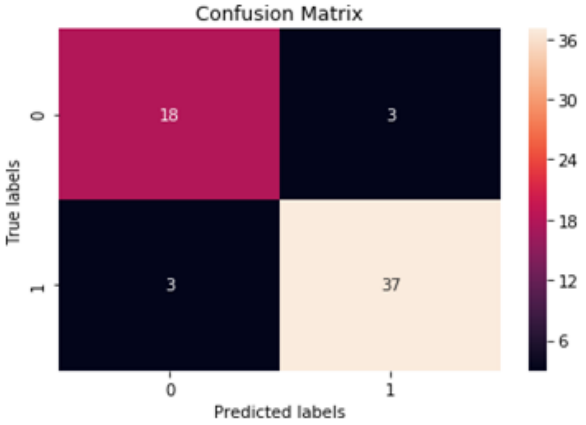

Figure 9: After Feature Selection

- After Feature Extraction

The Precision value is 0.88 and Recall is 0.93

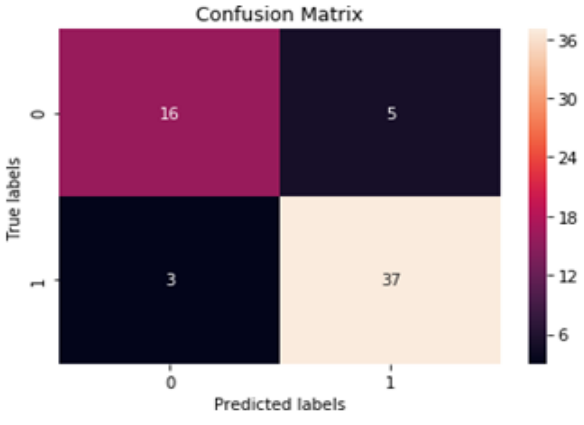

Figure 10: After Feature Extraction

May - June 2020
ISSN: 0193-4120 Page No. 4955-4962

5. Conclusion and Future Outlook

As the health care sector calls for modern solutions, there is a demand for implementing an efficient ML model for the classification of the data. In this experiment, MinMaxScaler is deployed first to scale down the attributes and then the HDD is trained using six ML classifiers. After the process of feature selection, it was observed that the top five attributes gave an accuracy of 87%. This result is equivalent to the accuracy percentage of the model, after scaling with 13 attributes. Thus, we conclude that feature selection justifies being more efficient in terms of time and space. After performing feature extraction, the resultant accuracy rate attained by using PCA was 90.16% for SVM classification model. Therefore, feature extraction produces an enhanced accuracy rate when compared to feature selection.

The future scope of this work can be further extended on cloud-based storage to store huge amount of data. Distributed computing platform enhances the computational rate and parallel processing of the data. Further, this work can be implemented on various disease data sets that can be developed using computational intelligence techniques which will be beneficial for the healthcare application and automation.

References

[1] Chen M, Hao Y, Hwang K, Wang L, Wang L. Disease prediction by machine learning over big data from healthcare communities. Ieee Access. 2017 Apr26;5:8869-79.

[2] Peker M. A decision support system to improve medical diagnosis using a combination of k-medoids clustering based attribute weighting and SVM. Journal of medical systems. 2016 May1; 40(5):116.

[3] Al-Janabi MI, Qutqut MH, Hijjawi M. Machine learning classification techniques for heart disease prediction: a review. International Journal of Engineering & Technology. 2018; 7(4):5373-9.

[4] Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H. Feature selection: A data perspective. ACM Computing Surveys (CSUR). 2017 Dec6;50(6):1-45.

[5] Singh J, Kamra A, Singh H. Prediction of heart diseases using associative classification. In2016 5th International Conference on Wireless Networks and Embedded Systems (WECON) 2016 Oct 14 (pp. 1-7).IEEE.

[6] Philip K. Optimization of feature selection in machine learning using genetic algorithms. http://www.phili pkali nda.com/ds8.html,2017.

[7] Dey A. Machine learning algorithms: a review. International Journal of Computer Science and Information Technologies.2016;7(3):1174-9.

[8] Chen Y, Wang Y, Cao L, Jin Q. CCFS: A Confidence-based Cost- effective feature selection scheme for healthcare data classification. IEEE/ACM transactions on computational biology and bioinformatics. 2019 Mar7.

[9] Khurana U, Turaga D, Samulowitz H, Parthasrathy S. Cognito: Automated feature engineering for supervised learning. In2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW) 2016 Dec 12 (pp. 1304-1307).IEEE.

[10] Nargesian F, Samulowitz H, Khurana U, Khalil EB, Turaga DS. Learning Feature Engineering for Classification. InIJCAI 2017 Aug 19 (pp.2529-2535).

[11] Reddy NS, Nee SS, Min LZ, Ying CX. Classification and Feature Selection Approaches by Machine Learning Techniques: Heart Disease Prediction. International Journal of Innovative Computing. 2019 May 31;9(1).

[12] Patel J, TejalUpadhyay D, Patel S. Heart disease prediction using machine learning and data mining technique. Heart Disease. 2015 Sep;7(1): 129-37.

[13] Fatima M, Pasha M. Survey of machine learning algorithms for disease diagnostic. Journal of Intelligent Learning Systems and Applications. 2017;9(01):1.

[14] SharmaH, RizviMA. Prediction of heart disease using machine learning algorithms: A survey. International Journal on Recent and Innovation Trends in Computing and Communication. 2017 Aug;5(8):99-104.

[15] Latha CB, Jeeva SC. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked. 2019 Jan1;16:100203.

[16] Kumar M, Shambhu S, Sharma A. Classification of heart diseases patients using data mining techniques.

[17] Heaton J. An empirical analysis of feature engineering for predictive modeling. In Southeast Con 2016 2016 Mar 30 (pp. 1-6). IEEE.

[18] Pasha A, Latha PH. Bio-inspired dimensionality reduction for Parkinson's disease (PD) classification. Health Information Science and Systems. 2020 Dec;8(1):1-22.

[19] Osisanwo FY, Akinsola JE, Awodele O, Hinmikaiye JO, Olakanmi O, Akinjobi J. Supervised machine learning algorithms: classification and comparison. International Journal of Computer Trends and Technology (IJCTT). 2017Jun; 48(3):128-38.

[20] Shu T, Zhang B, Tang YY. Effective heart disease detection based on quantitative computerized traditional chinese medicine using representation based classifiers. Evidence-Based Complementary and Alternative

May - June 2020
ISSN: 0193-4120 Page No. 4955-4962

Medicine.2017;2017.
[21] Bollegala D. Dynamic feature scaling for online learning of binary classifiers. Knowledge-Based Systems. 2017 Aug1;129:97-105.
[22] API design for machine learning software: experiences from the scikit- learn project, Buitinck et al.,2013.
[23] 1.McKinney W, others. Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference. 2010. p.51–6.
[24] Oliphant, T.E., 2006. A guide to NumPy, Trelgol Publishing USA.
[25] Hunter, J.D., 2007. Matplotlib: A 2D graphics environment. Computing in science & engineering, 9(3), pp.90–95.