Advances and Applications in Mathematical Sciences
Volume 21, Issue 1, November 2021, Pages 269-282

HEART DISEASE PREDICTION SYSTEM USING ENSEMBLE PREDICTIVE MODELLING WITH PRINCIPAL COMPONENT ANALYSIS

M. MEHARUNNISA[1] and M. SORNAM[2]

[1]Assistant Professor
Department of BCA
Ethiraj College for women, India
E-mail: m.meharunnisa1987@gmail.com

[2]Professor
Department of Computer Science
University of Madras, Guindy Campus
Chennai, India
E-mail: madasamy.sornam@gmail.com

Abstract

In the recent years, data mining has been employed in the medical field for extracting and manipulating information, and aids within the higher process. Predicting the results of a process with a high level of accuracy is a difficult task. In this work, the advantage of the data mining models have been taken to predict the heart disease. The benchmark dataset, "Cleveland Heart Disease" dataset from UCI machine learning repository has been used. The main objective of this work is to propose an extensive data pre-processing task such as imputation of missing values and a feature engineering technique namely 'Principal Component Analysis' to be used to transform the dataset in a compressed form. Ensemble / classifier combination method called boosting method such as Gradient boosting machine and Random Forest are used. The results show that, the ensemble learners, with PCA attained ROC - AUC value of 0.90 and above with 100% accuracy of prediction. Moreover, it ensures that no missing information must be removed and might be imputed to confirm the data quality is enough. As a result, the model has shown to be helpful for the real time prediction of heart disease.

1. Introduction

2020 Mathematics Subject Classification: 34Dxx, 93Dxx.

Keywords: Ensembles, Gradient boosting machine, Prediction, machine learning, Random Forest, Principal Component Analysis.

Received June 2, 2020; Accepted January 18, 2021

An investigation study demonstrated [1] that heart diseases resulted in more than 2.1 million fatalities at all ages in India in 2015. The common cardiovascular disease comprises of ischaemic heart disease and stroke. According to World Health Organisation, the death rate globally is 17.7 million. In the course of recent years, unequal appropriation of the nation's fast monetary development and urbanization has presumably added to regional contrasts in the key hazard factors for cardiovascular mortality. Cardiovascular diseases [2] contributed to 28.1% of total deaths and 14.1% of total disability-adjusted life years in India during 2016.

Mokeddem and Atmani proposed [3] a model on Genetic Algorithm wrapped Naïve Bayes, in which the coronary heart disease dataset was used. In the first step, Genetic algorithm was used for feature reduction and the reduced features were used in the second step to evaluate Naïve Bayes algorithm. The classification accuracy (85.50%) thus obtained were compared with Support Vector Machine (83.5%), MultiLayer Perceptron (83.16%) and C4.5 decision tree (80.85%) algorithm.

Liu et al proposed [4] a hybrid classification system for heart disease diagnosis on stat log heart disease dataset from UCI machine learning repository. The proposed system used a mathematical approach called ReliefF and Rough Set (RFRS) method for feature reduction. The features were extracted using the ReliefF algorithm and reduced using heuristic RS reduction algorithm that was proposed. The C4.5 ensemble classifier was used for classification and the accuracy obtained was 92.5% according to jack-knife cross validation scheme.

Nasution et al. proposed [5] Principal Component Analysis (PCA) for feature reduction and used the reduced subset to overcome misclassification and over fitting on decision tree C4.5. They used cervical cancer dataset from UCI machine learning repository and evaluated the performance of their framework based on accuracy, specificity and precision. The accuracy rate was enhanced to 90.70% from 86.05% without PCA.

Divya Jain and Vijendra Singh reviewed [6] many feature selection techniques and classification systems to predict chronic diseases such as diabetes, heart disease, cancer at an early stage. The three traditional feature selection methods namely filter, wrapper and embedded methods were

discussed. Also, a hybrid approach have been used to remove insignificant and noisy features. It was observed that filter methods were more efficient than wrapper and embedded methods in terms of computation and efficiency.

This paper is organized in the following manner. Section 2 describes the proposed model. Section 3 explains the exploratory data analysis such as missing value imputation, class imbalance and feature engineering. Section 4 includes the details of the predictive model based on ensemble machine learning methods and the discussion of the result. Section 5 presents a conclusion with Section 6 providing the future enhancement.

2. Proposed Model

In this paper, an accurate heart disease classification method is proposed. To illustrate the effectiveness and robustness of the proposed method, Cleveland Heart disease dataset [7] is utilized from UCI machine learning repository. A novel feature reduction and classification strategy is utilized. In the first step, data preprocessing is performed to remove outliers and to handle the missing data. In the second step, Variance Inflation factors for all the features are calculated in order to identify multicollinearity. Principal Component Analysis (PCA), a dimensionality reduction technique is utilized to reduce the multicollinearity. In the third step, ensemble machine learning algorithms are used on the resulting Principal Components (PCs) for classifying and predicting the heart disease. The performance of the classifiers are evaluated by measuring its accuracy, recall and precision. The proposed model can predict the heart disease accurately and effectively.

3. Exploratory Data Analysis

3.1. Description of the Dataset. The Heart disease dataset was obtained [7] from UCI machine learning repository. It consists of four databases from Cleveland, Hungary, Switzerland and the VA Long Beach. The dataset taken for this research is from Cleveland database. It consists of 303 records with 14 features. Each instance consists of 13 predictor variable and 1 class variable, which is independent.

Table 1. Description of the Cleveland Heart Disease dataset.

S.No	Attribute	Description
1	Age	Age in Years
2	Sex	Sex (1=male; 0=female)
3	Cp	Chest pain type Value 1: typical angina Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptotic
4	Trestbps	resting blood pressure (in mm Hg on admission to the hospital)
5	Chol	serum cholestoral in mg/dl
6	Fbs	fasting blood sugar >120 mg/dl) (1 = true; 0= false)
7	Restecg	Resting electrocardiographic results Value 0: normal Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05 mV) Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8	Thalach	maximum heart rate achieved
9	Exang	exercise induced angina (1 = yes; 0 = no)
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	the slope of the peak exercise ST segment Value 1: upsloping Value 2: flat Value 3: downsloping
12	Ca	number of major vessels (0-3) coloured by fluoroscopy
13	Thal	3 = normal; 6 = fixed defect; 7 = reversable defect
14	Num	diagnosis of heart disease (angiographic disease status) Value 0: <50% diameter narrowing

		Value 1: >50% diameter narrowing

The description of the dataset is presented in Table 1. There were 164 instances which were tested negative for healthy and 139 tested positive for possible heart disease.

4. Feature Engineering

4.1. Correlation Analysis. The correlation coefficient is used to measure the linear relationship between the variables. The value of 0, +1 and -1 represents no linear relationship, perfect positive linear relationship and a perfect negative linear relationship respectively. The values between 0 and -0.3, -0.3 and -0.7, -0.7 and -1.0 indicate [8] weak positive, moderate positive and strong positive relationship .The correlation among the independent variables in Figure 1 shows that the relationship between the slope and OP (OldPeak) variable is 0.57, which is slightly greater than 0.5. While developing the predictive model, high correlations among independent variables lead to "multicollinearity" problem. Hence, the estimate for the model become unstable.

4.2. Identification of Multicollinearity. Variance Inflation factor (VIF), a method to detect multicollinearity is utiliized in this work to identify the multicollinearity among the exploratory variables. VIF is calculated using equation (1)

$$VIF_j = \frac{1}{1 - R_j^2},$$

where R_j^2 is the R^2- value obtained by regressing the j^{th} predictor on the remaining predictors.

The Variance inflation of all the variables of the Cleveland heart disease dataset is tabulated in Table 2. If the variation inflation factor for a predictor variable is near or above 5, then there exists multicollinearity problem in the dataset. If there are one or additional factors with a high VIF [9], one amongst the factor ought to be removed from the model.

Table 2. Variance Inflation Factor of the exploratory variables.

S.No	Attribute	VIF
1	Age	1.48
2	Sex	1.30
3	Cp	1.32
4	Trestbps	1.19
5	Chol	1.14
6	Fbs	1.07
7	Restecg	1.09
8	Thalach	1.62
9	Exang	1.38
10	Oldpeak	1.75
11	Slope	1.68
12	Ca	1.35
13	Thal	1.52

From Table 2, the variance inflation factor of all the variables is less than 5, we can safely conclude that there is no multicollinearity problem as far as the given dataset is concerned. This low value of VIF also indicates that the estimates developed using these variables will be stable.

4.3. Principal Component Analysis. The intention of PCA is to discover a new set of attributes that better captures the variability of the facts. Principal components are linear aggregate of normalized variables. Since principal components are orthogonal to every other; the correlation among them will become zero. This solves the trouble of multicollinearity while there may be greater quantity of correlated unbiased variable. This method is based on some mathematical ideas on standard deviation, variance, covariance, Eigen vectors and Eigen values.

The goal of PCA is to find a transformation of the data that satisfies the following properties: (a) each pair of new attributes has 0variance. (b)The characteristics are instructed as to how much information variance is captured by each attribute. The attributes are ordered with respect to how much information variance is captured by each attribute. (c) The first attribute captures as much of the variance of the data as possible. (d) Subject to the orthogonality requirement, each successive attribute captures as much of the remaining meaning as possible.

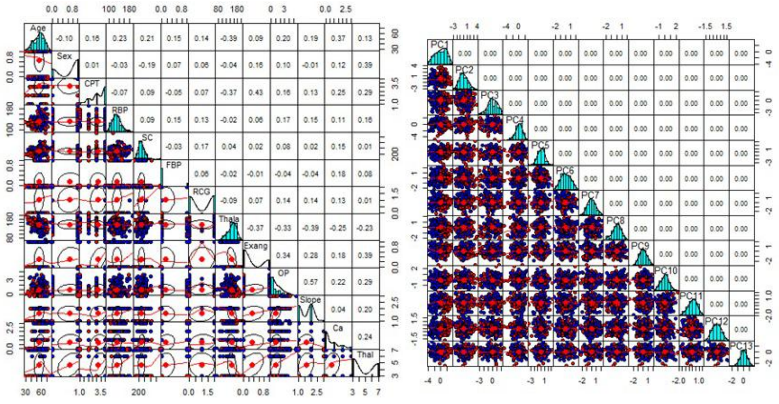
4.4. Significance of PCA on Heart disease dataset. Importances of principal components obtained are tabulated in Table 3.

Table 3. PCA Analysis report.

Principal Components	Standard Deviation	Proportion of Variance	Cumulative Proportion
PC1	1.7462	0.2346	0.2346
PC2	1.2504	0.1203	0.3548
PC3	1.1632	0.1041	0.4589
PC4	1.10629	0.09414	0.55306
PC5	1.01864	0.07982	0.63288
PC6	0.94687	0.06897	0.70184
PC7	0.89684	0.06187	0.76371
PC8	0.85191	0.05583	0.81954
PC9	0.78600	0.04752	0.86706
PC10	0.71132	0.03892	0.90598
PC11	0.66467	0.03398	0.93997
PC12	0.64323	0.03183	0.97179
PC13	0.60554	0.02821	1.00000

It can be observed from Table 3, PCA analysis reports as many PCs as the number of independent variables in the dataset. The standard deviation of each component is shown in the second column of the table. The last column

shows the cumulative proportion of variance. It appears that first PC accounts for 23.46% of the total variance, second PC for 35.48% of the total variance and so on. From Table 3, first ten principal components explains 90% variance in the dataset, first eleven principal components explains 93% variance and first twelve principal components explains 97% variance in the dataset.



(a) Before PCA (b) After PCA

Figure 1. Scatter plot and Correlation among the variables.

Figure 1 shows the correlations among the principal components are reduced to zero. Figure 2 shows the scree plot for determining how many principal components need to be kept to capture most of the variability of the data.

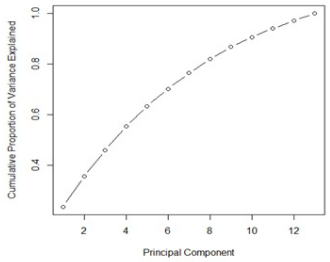


Figure 2. Scree Plot representing the variances of all the Principal Components.

The advantage of principal components is to find the straight line that

best spreads the data out when it is projected along it. PCA is a kind of linear transformation on a given dataset that has values for a certain number of coordinates for a certain amount of areas. This linear transformation suits the given dataset to a new coordinate space that the most widespread variance is discovered on the primary coordinate, and every next coordinate is orthogonal to the ultimate and has a lesser variance. In this manner, it is possible to remodel a set of x correlated variables over y samples to a set of p uncorrelated principal components over the same samples. With few principle components, it is possible to identify the initial variables that are strongly correlated with each other.

5. Predictive Ensembles Modeling with Principal Components

The basic idea of ensemble modeling is to construct multiple base classifiers C_i from each training set D_i and then aggregating their predictions when classifying unknown records. The ensemble of classifiers can be constructed by manipulating the training set, input features, the class labels and the learning algorithm. An unknown sample x is classified by combining the predictions made by the base classifiers $C_i(x)$.

Algorithm 1: Ensemble method

Input: D-original training data, k-no of base classifiers, T-Test data

1 for $i = 1$ to k **do**

2 Create training set D_i from D

3 Build base classifier C_i from D_i.

4 end

5 for each unknown sample $x \in T$ **do**

6 $C^*(x) = Vote(C_1(x), C_2(x), \cdots, C_k(x))$

7 end

The output class can be obtained by taking a majority vote on the individual predictions of the classifiers or by weighting the predictions of

each classifier with the base classifier. Ensemble methods work better with unstable classifiers, i.e., base classifiers that are sensitive to minor noise in the training set. Examples of unstable classifiers [10] include decision trees, rule-based classifiers and artificial neural network. The different methods [11] involved in ensembling are boosting, bagging and stacking. The ensemble methods used in our research work are gradient boosting machine and random forest.

6. Results and Discussion

There were two phases of experiment for this study: (1) training phase – 75% of dataset i.e., 226 records (2) Testing phase – 25% of dataset i.e., 77 records One common measure discussed in the literature is accuracy, which is defined as correctly classified instances divided by total number of instances. There are other performance metrics such as, Recall, F-Measure, and Receiver operating characteristic (ROC) and Area under the curve (AUC) which are used for evaluation of algorithm.

Receiver operating characteristic (ROC) curve could be a graphical approach for classification result between true positive rate and false positive rate of a classifier. In the curve, true positive rate i.e., sensitivity is plotted on the y axis and also the false positive rate i.e., 1-specificity is plotted along the x axis.

Area under the Curve (AUC) represents degree of disconnectedness. It tells what proportion model is capable of differentiating between categories. Higher the AUC, higher the model is at predicting true as true and false as false.

Table 4. Results of the Ensemble model without using PCA.

Metrics	Ensemble Learning Algorithm	
	GBM	RF
Precision	0.80	0.81
Recall	0.80	0.81
F-Measure	0.80	0.81
AUC	0.90	0.90

Accuracy	85.71%	84.42%

The result in Table 5 shows that, without using principal component analysis, the ensemble learning algorithm, GBMand RF predicted the disease and its metrics. Maximum of AUC=1 means that the diagnostic test is perfect in distinguishing the healthy and diseased subjects [12]. From Table 5, the AUC of both the model is 0.90.

Table 5. Performance evaluation of the proposed system for heart disease prediction.

Metrics	PC1 - PC10		PC1 - PC11		PC1 - PC12	
	GBM	RF	GBM	RF	GBM	RF
Precision	0.83	0.80	0.81	0.81	0.86	0.82
Recall	0.89	0.89	0.86	0.89	0.84	0.80
F-Measure	0.86	0.84	0.84	0.85	0.85	0.90
AUC	0.90	0.89	0.90	0.90	0.92	0.90
Accuracy	84.42%	83.12%	100%	100%	100%	100%

The results in Table 6 shows that, our proposed model i.e., ensemble learners along with principal component can give better model with reduced dimension. Using the first ten principal components, the AUC of GBM and RF becomes 0.90 and 0.89 respectively. The AUC of GBM and RF becomes 0.90, while using the first 11 principal components. The AUC of GBM and RF becomes 0.92 and 0.90 respectively while using the first 12 principal components rather than using all the dimensions. The Accuracy of 100% is achieved when the model is used with 11 and 12 principal components.

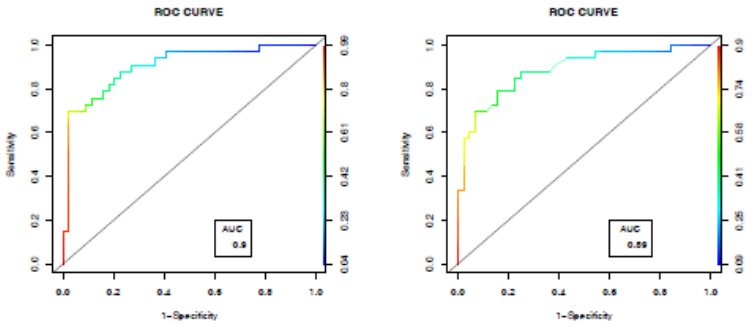


Figure 3. ROC OF GBM and RF using PC1-PC10.

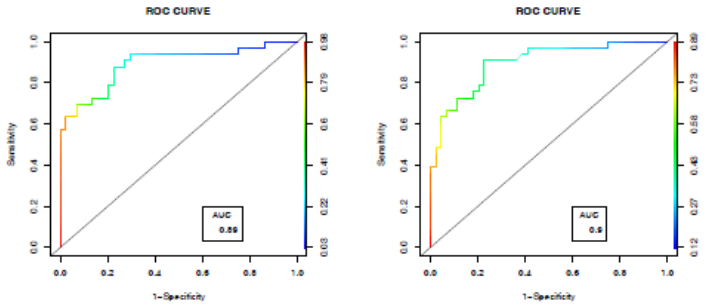


Figure 4. ROC OF GBM and RF using PC1-PC11.

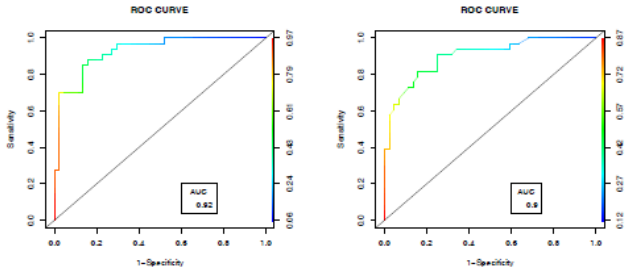


Figure 5. ROC OF GBM and RF using PC1-PC12.

It can be seen from the above figure, a good classification model ought to be located as close to the higher left corner of the diagram, whereas a model that produces random guesses ought to reside on the most diagonal, connecting the points $(TPR = 0, FPR = 0)$ and $(TPR = 1, FPR = 1)$, in which TPR is commonly referred as sensitivity and FPR as 1-specificity.

Finally in this research work, ggbiplot is used to visualize the components

of PCA. It is used to plot the principal components obtained and infers the charactistics of chest pain type (CP).The four types of chest pain type are grouped as 1, 2, 3 and 4 namely typical angina, atypical angina, non-angina pain and asymptotic pain.

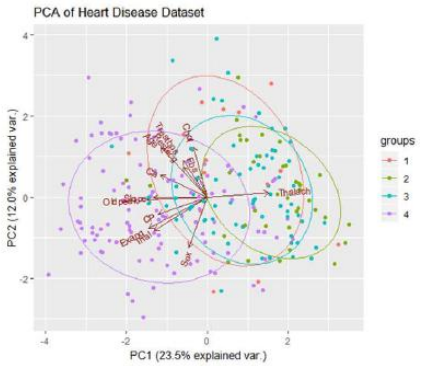


Figure 6. GGBIPLOT representation of PCA.

From the Figure 6, it is clear that typical angina is highly influenced by serum cholesterol and age. Atypical angina is influenced by maximum heart rate achieved. Non anginal pain is due to cholesterol, fasting blood sugar, resting blood pressure and number of blood vessels. Asymptotic pain is highly influenced by all the features except maximum heart rate achieved.

7. Conclusions and Future Work

In this research, it is explored that the proposed model contributed a lot to the prediction model without losing any information from the dataset due to outliers, noise and missing data. With the help of principal component analysis, the dimensionality reduction has been achieved. The proposed model also addressed the issue of multicolinearity and made use of ensemble learners to create a stable model. All the models are evaluated on Precision, Recall, F-Measure and AUC. Through intensive experiments, it is found that ensemble method along with principal components outperforms the machine learning methods in the prediction of heart disease. ROC and AUC is used to measure the robustness of the model. Finally, for the benchmarking of model correctness, the performance of GBM and random forest model with PC's is compared with the same model without PC's. Taken together, the model identifies the risk factors associated with each of the chest pain type. The proposed method can be very helpful to the physicians for their final decision

on their patients as by using such an efficient model to make accurate decision. The same method can also be applied to other disease prediction such as type 2 diabetes mellitus, liver disease and identification of breast cancer and so on.

This system can further be expanded, by using genetic algorithms which helps to deal with micro array datasets with more number of predictor variables. The data set used in this work is available in UCI machine learning repository.

References

[1] Geneva World Health Organization. Global health estimates, Deaths by cause, age, sex, by country and by region, 2000-2015. 2016.

[2] Valery L. Feigin, Gregory Roth, Mohsen Naghavi, Priya Parmar, Rita Krishnamurthi, Sumeet Chugh, George A Mensah, Bo Norrving, Ivy Shiue and Marie Ng, et al. Global burden of stroke and risk factors in 188 countries, during 1990-2013: a systematic analysis for the global burden of disease study 2013, The Lancet Neurology 15(9) (2016), 913-924.

[3] Sidahmed Mokeddem, Baghdad Atmani and Mostéfa Mokaddem, Supervised feature selection for diagnosis of coronary artery disease based on genetic algorithm, arXiv preprint arXiv: (2013), 1305-6046.

[4] Liu Xiao, Wang Xiaoli, Su Qiang, Mo Zhang, Zhu Yanhong, Wang Qiugen and Wang Qian, A hybrid classification system for heart disease diagnosis based on the rfrs method, Computational and mathematical methods in medicine, 2017.

[5] M. Z. F. Nasution, O. S. Sitompul and M. Ramli, Pca based feature reduction to improve the accuracy of decision tree c4. 5 classification, In Journal of Physics: Conference Series, volume 978, page 012058. IOP Publishing, 2018.

[6] Divya Jain and Vijendra Singh, Feature selection and classification systems for chronic disease prediction, A review, Egyptian Informatics Journal 19(3) (2018), 179-189.

[7] P. M. Murphy, Uci repository of machine learning databases, department of information and computer science, university of california. http://www. ics. uci. edu/AI/ML/MLDBRepository. html, 1992.

[8] Bruce Ratner, The correlation coefficient: Its values range between+ 1/- 1, or do they? Journal of targeting, measurement and analysis for marketing 17(2) (2009), 139-142.

[9] Michael Olusegun Akinwande, Hussaini Garba Dikko and Agboola Samson, Variance inflation factor: as a condition for the inclusion of suppressor variable (s) in regression analysis, Open Journal of Statistics 5(07) (2015), 754.

[10] Leo Breiman, Bagging predictors, Machine learning 24(2) (1996), 123-140.

[11] Giorgio Valentini and Francesco Masulli, Ensembles of learning machines, In Italian

workshop on neural nets, Springer (2002), 3-20.

[12] Rajeev Kumar and Abhaya Indrayan, Receiver operating characteristic (roc) curve for medical researchers, Indian pediatrics 48(4) (2011), 277-287.