Enrollment No.:2020MSTA016

M.Sc. Statistics

Major Project Report

Heart Disease Prediction

-Using some statistical tools and Machine learning algorithms-

Submitted By:-

Shiwam Kumar Sahu

Supervisor:-

Dr. Arvind Pandey

(Associate Professor)

Department of Statistics

School of Mathematics, Statistics and Computational Sciences

Central University of Rajasthan, Ajmer - 305817

2020-2022

Contents

Abstract

Background: Heart disease is one of the leading causes of death in worldwide. In recent time, machine learning play a vital role in the medical field. The aim of our work is the proposal of a dimensionality reduction technique with the help of Principal Component Analysis(PCA) and then performing classification algorithm techniques K- Nearest Neighbor (K-NN), Naive Bayes, Decision tree, SVM kernel(RBF), Adaboost(AB), Gradient Boosting(GB), Decision Tree(DT),Random Forest(RF). In this work, we had used two Heart disease datasets which are available in Kaggle(an open-source where we find and publish datasets). One is from 1998, which contains 14 features, here Random Forest had the highest accuracy and the other one is from 2020, where we applied same model, SVM kernel had the highest accuracy. We hope the proposed system will be helpful and useful for the physician to diagnose heart disease accurately and effectively.

Keywords: K-nearest neighbor, Support vector machine (SVM)(RBF), Decision tree, Dimensonality Reduction(Using PCA), Random forest, Enemble methods.

1 Introduction

The epidemic of cardiovascular diseases(CVD) [8] in India is advancing rapidly. The state of affairs has been challenged in the recent by a series of studies, which shows an alarming rise in case of Cardio Vascular Disease mainly in women. This is not only the case of India but it has been shown that heart disease kills one American every 39 seconds. Heart disease kills more women than all forms of cancer combined. Nearly 40 % of all deaths in New York were due to Cardiovascular Disease in 2008. The Global Burden of Disease(GBD)[9] study reported elimated mortality from coronary heart disease(CHD)(most common type of heart disease) in India at 1.6 million in the year 2000. More women than men now die of heart disease. More than half of preventable deaths that are caused by heart disease and stroke happen to people under the age of 65. CVD claims life of more women every year than all forms of cancer combined. WHO revealed that CVD is expected to affect almost 23.6 million people by the year 2030. European Society of Cardiology (ESC) has published a report in which 26.5 million adults were identified having heart disease and 3.8 million were identifed each year. About 50–55% of heart disease patients die within the initial 1–3 years, and the cost of heart disease treatment is about 4% of the overall healthcare annual budget.[4]
The growth in medical data collection presents a new opportunity for physicians to improve patient

diagnosis. In recent years, practitioners have increased their usage of computer technologies to improve decision-making support. In the health care industry, machine learning is becoming an important solution to aid the diagnosis of patients. Recently, to solve difficult issues, a range of data mining techniques and machine learning techniques are built.

The contribution of the current work is to introduce an intelligent medical decision system for the diagnosis of heart disease based on contemporary machine learning algorithms. In this work, we are utilizing supervised learning and also some Ensemble techniques. In the disease prediction machine learning(ML) plays a signficant role. In this work, we predicts whether the patient has a particular disease type or not based on an efficient learning technique.

Several supervised learning algorithms like k-nearest neighbor (KNN), support vector machine (SVM),mdecision tree (DT), Naive Bayes (NB), and random forest (RF) and some Ensemble learning like Gradient Boosting, AdaBoost, XGBoost are used to classify whether the people tested belong to the class of heart disease or healthy people. Although, Principal component analysis(PCA) are used to select essential features and also reduces the dimension from the dataset which help to give more accurate result in less number of time. The rest of this work is structured as follows:-

Section2 - Describes the proposed architecture and methodology

Section3 - Experimental results:-

- Dataset discription
- Experiment setting
- Outcomes

Section4 - Conclusion of our work

2 The Proposed Methodology of predicting Heart Disease

Figure 1 describes the architecture of the proposed system. It is structured into six stages, including data collection, data preprocessing, Dimensonality reduction using Principle component analysis(PCA), data splitting, training models, and evaluating models.

The steps of the proposed approach are explained in detail as follows:-

2.1 Dataset Discription:

In this work, we use two heart disease datasets which are available in Kaggle:-

1) First heart disease dataset is from 1988 and consists of four databases: Cleveland, Hungary,

Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The ”target” field refers to the presence of heart disease in the patient. The target column includes two classes: 1 indicates heart diseases, and 0 indicates non heart disease. The 14 features together with their descriptions and data types are shown in Table(see table 1)

Table 1: Dataset 1 discription

No.	Features	Descriptions
1	Age	Age of patient (years)
2	Sex	1: male, 0: female
3	Chest pain(CP)	CP types:- 1:typical angina, 2 :atypical angina, 3 : nonangina pain, 4 : asymptomatic
4	RestBP	Resting blood pressure
5	Chol	Serum cholesterol in mg/dl
6	FBS	Fasting blood sugar larger 120 mg/dl (1 true)
7	RestECG	Resting electrocardiographic result
8	Talach	Maximum heart rate accomplished
9	Exang	Exercise-induce angina (1 yes)
10	Oldpeak	ST depression induce: exercise relative to rest
11	CA	Number of major vessels (0–3)
12	Slope	Slope of peak exercise ST
13	Thal	No explanation provided, but probably thalassemia
14	Num	Diagnosis of cardiac disease: 1: yes 0: no

2) Second heart disease dataset is from 2020 annual CDC(Central For Disease Control and Prevention) survey data of 400k adults related to their health status. Originally, the dataset come from the CDC and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to gather data on the health status of U.S. residents. As the CDC describes: ”Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000

adult interviews each year, making it the largest continuously conducted health survey system in the world". The most recent dataset (as of February 15, 2022) includes data from 2020. It consists of 401,958 rows and 279 columns. The vast majority of columns are questions asked to respondents about their health status, such as "Do you have serious difficulty walking or climbing stairs?" or "Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]". The original dataset of nearly 300 variables was reduced to just about 20 variables. The dataset contains 18 variables (9 booleans, 5 strings and 4 decimals)and total number of rows is 319796. The 18 features together with their descriptions and data types are shown in Table(see table 2).

Table 2: Dataset 2 description

No.	Features	Descriptions
1	HeartDisease	Yes or No
2	BMI	12.02 - 94.85
3	Smoking	Yes or No
4	AlcoholDrinking	Yes or No
5	Stroke	Yes or No
6	PhysicalHealth	0-30
7	MentalHealth	0-30
8	DiffWalking	Yes or No
9	Sex	Male or Female
10	AgeCategory	18-24 to above 80
11	Race	American indian,Asian, Black, Hispanic,White,other
12	Diabetic	Yes or No
13	PhysicalActivity	Yes or No
14	GenHealth	Good,Fair,Very Good, Excellent
15	SleepTime	01-24
16	Asthma	Yes or No
17	KidneyDisease	Yes or No
18	SkinCancer	Yes or No

2.2 Data Preprocessing

The features are scaled to be in the interval $[-3, 3]$ with the help of standardization(or Z-score normalization) process. For standardized value(a z-score), using the formula:-

$$Z=\frac{X-\mu}{\sigma}$$

where the symbols are:-

X:observations

μ: Mean

σ:Standard deviation

2.3 Dimensionality reduction using Principal Component Analysis(PCA)

The extraction of the best features is a crucial phase because irrelevant features often affect the classification efficiency of the machine learning classifier. Principal Component Analysis(PCA) are used for dimensional reduction without loosing the much more information of the given features and also used to select essential features from the dataset.

2.4 Data Splitting

In this step, both the heart disease dataset is divided into a 80% training set and a 20% as the testing set. The training set is utilized for training the models, and the testing set is utilized to evaluate the models.

2.5 Training Models

In the training dataset we are applying various contemporary classifcation algorithms: Logistic Regression(LR), Support Vector Machine(SVM), Random Forest(RF), Decision Tree(DT), K-nearest neighbous(K-NN), Gradient Boosting(GB), Ada Boost(AB), XGBoost. **1)Logistic Regression** is equivalent of linear regression for categorical outcome variable. This can be used where predictors can be categorical or continuous. This is a statistical method for evaluating a dataset in which a result is calculated by one or more independent variables. It is a supervised learning technique similar to linear regression. It is used typically in cases when structured model is preferred over data driven models for classification tasks. In this model, categorical outcome variable cannot be

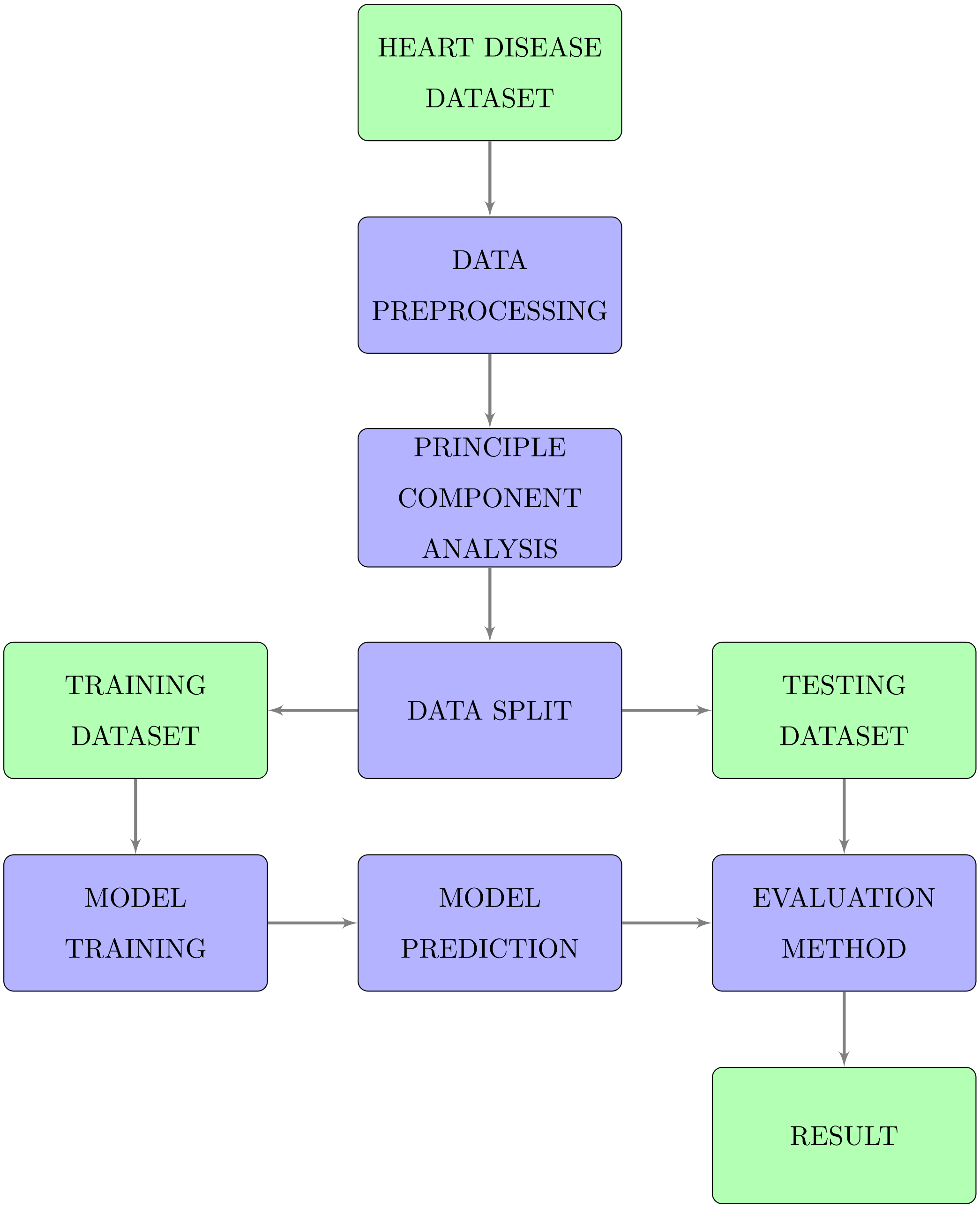


Figure 1: Flowchart of proposed methodology

directly modeled as an linear function of predictors. So, here instead of using outcome variable(Y) in the model, a function of Y, called **logit** is used.

logit: Think about modeling probability value as a linear function of predictors, specifically in a two class case. If P is the probability of class 1 membership,

$$P = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + ... + \beta_p * X_p$$

where p is the number of predictors.

Since, in logistic regression L.H.S range improves from $\{0, 1\}$ to $[0, 1]$ but R.H.S range is $\{-\infty, \infty\}$. So, typically a non linear function is used to approach to bring L.H.S range equal to R.H.S range that is,

$$P = \frac{1}{1 + \exp^{-(\beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + ... + \beta_p * X_p)}}$$

This function is called logistic response function. Now rearrange the previous two equations as below:

$$\frac{P}{1 - P} = \exp^{\beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + ... + \beta_p * X_p}$$

where $\frac{P}{1 - P} =$ **odds**.

Odds of belonging to a class is defined as ratio of probability of class 1 membership to probability of class 0 membership. Now the range of previous equation to be $\{0, \infty\}$. Taking log in both side,

$$\log(odds) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + ... + \beta_p * X_p$$

This is called **standard logistic model**. Now L.H.S and R.H.S both have range $\{-\infty, \infty\}$. $\log(odds)$ is called logit. It is used as the outcome variable in the model instead of categorical Y. Odds and logit can be written as a function of probability of class 1 membership.

2)Decision Tree use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, it can be said that the purity of the node increases with respect to the target variable. The decision tree splits the nodes on all available variables and then selects the split which results in the most homogeneous sub-nodes. Every internal node carries a test on features, and branches carry the test conclusion, and the class label is meant for each leaf node. It is utilized both for classifications and regression.

Given a database $D = \{t_1, t_2, ..., t_n\}$, where t_i denotes a tuple, which is defined by a set of attribute

$A = \{A_1, A_2, ..., A_m\}$. Also, given a set of classes $C = \{c_1, c_2, ..., c_k\}$.

A decision tree T is a tree associated with D that has the following properties:

- Each internal node is labeled with an attribute A_i
- Each edges is labeled with predicate that can be applied to the attribute associated with the parent node of it
- Each leaf node is labeled with class c_j

In principle, there are exponentially many decision tree that can be constructed from a given database(also called training data). Some of the tree may not be optimum and some of them may give inaccurate result. There are mainly two approaches are known in DT:

- Greedy strategy
- Modification of greedy strategy
  - ID3
  - C4.5
  - CART

3)Random Forest is considered as a highly accurate and robust method because of the number of decision trees participating in the process. It tries to build k different decision trees by picking a random subset S of training samples. It generates fully Iterative Dichotomiser 3 (ID3) trees with no pruning. It makes a final prediction based on the mean of each prediction, and it tends to be robust to overfitting, mainly because it takes the average of all the predictions, which cancels out biases. As the name suggest as forest the random forest classifier is an ensemble of decision trees where a random vector sample produce each classifier from input vector and each tree cast a unit vote for the most popular class to classify an input vector, most of the time trained with a bagging method.

A random forest is a classifier consisting of a collection of tree structured classifiers $\{h(x, \theta_k), k = 1, 2, 3, ...\}$ where the θ_k are independently, identically distributed random trees and each tree casts a unit vote for the final classification of input x. Like CART(classification and regression trees), RF uses the **gini index** for determining the final class in each tree. The final class of each tree is aggregated and voted by weighted values to construct the final classifier.

Gini Index: The gini index is mainly the impurity measures of the node. The gini index of node impurity is the measure most commonly chosen for classification-type problems. If a dataset T contains examples from n classes

$$\text{Gini}(T) \text{ is defined as: } Gini(T)=1-\sum_{j=1}^{n}(p_j)^2$$

If a datset T is split into two subsets $T1$ and $T2$ with sizes $N1$and $N2$ respsectively, the gini index of the split data contains examples from n classes, the gini index(T) is defined as:

$$Gini_{split}(T)=\frac{N_1}{N}\text{gini}(T_1) + \frac{N_2}{N}\text{gini}(T_2)$$

Working algorithm of Random Forest(RF):

1. A random seed is chosen which pulls out at random a collection of samples from the training dataset while maintaining the class distribution.

2. With this selected dataset, a random set of attributed from the original dataset is chosen based on user defined values. All the input variables are not considered because of enourmous computation and high chances of overfitting.

3. In a dataset where M is the total number of input attributes in the dataset, onLy R attributes are chosen at random for each tree where R<M.

4. The attributes from this set creates the best possible split using the gini index to develop a decision tree model. The process repeats for each of the branches until the termination condition starting that leaves are the nodes that are too small to split.

4)Naive Bayes model is very effective for large datasets because of its simplicity. It works on the probability basis $p(c|x)$, where $p(c|x)$ is the posterior probability of the class (c) and predictor (x). This is a probabilistic statistical base classifier based upon Bayes' theorem which is strong supervised machine learning classification technique. It assumes that all the features are conditionally independent which means the effect of an attribute value has no effect on other attribute value. Naïve Bayes is a very light weight classifier can be used to classify big dataset easily. It is very robust to ignore noise and irrelevant attributes. It is very easy to construct and no need of complicated iterative parameter estimation schemes.

In Naive bayes, all records are used instead of relying on just the matching records.

Naive bayes modification:

- For class i of outcome variable, compute the probabilities $(P_1, P_2, ..., P_p)$ of belonging to class i for each predictor's value $(X_1, X_2, ..., X_p)$ taken by the new observation to be classified.

- compute $P_1 * P_2 * ... * P_p * P(c_i)$, here P_{c_i} is the proportion of record that are belonging to class i.
- Execute previous two steps for all the classes
- To compute the probability of the new observation belonging to class i, divide the value computed in step 2 by the summation of values computed in step 2 for all the classes
- Execute previous step for all the classes
- Classify the new observation to the class with the highest probability value.

Naive bayes formula:

$$P(C_i/X_1, X_2, ..., X_p) = \frac{[P(X_1/C_i)*P(X_2/C_i)*...*P(X_p/C_i)]*P(C_i)}{[P(X_1/C_1)*P(X_2/C_1)*...*P(X_p/C_1)]*P(C_1)+...+[P(X_1/C_m)*P(X_2/C_m)*...*P(X_p/C_m)]*P(C_m)}$$

Naive bayes formula is directly derived fromt the exact bayes formula after making following assumptions:

- Predictor's values $(X_1, X_2, ..., X_p)$ occur independent of each other for a givene class.
  $P(X_1, X_2, ..., X_p/C_i) = P(X_1/C_i) * P(X_2/C_i) * ... * P(X_p/C_i)$
- For classification, naive bayes formula works quite well
- Since, we don't require probability values to be accurate in absolute term, rather just a reasonably accurate rank ordering of these values.
- For the same reason, we should use the numerator only and drop the denominator which is common for all the classes.

5)K-Nearest Neighbours(K-NN)

K-NN is non-parametric method, as it does not consider the dimensionality of dataset for diagnosis because it relies upon nearest training data points. The "GridSearchCv" was used to figure out the total number of neighbors for the KNN training needed to achieve superior performance. In K-NN useful information for modeling is extracted using the similarities between the records based on predictors values, typically disctance based similarity measures are used. Most popular metric is euclidean distance for measuring the distance between two records. Consider two recors having values of the predictors denoted by $(X_1, X_2, X_3, ..., X_p)$ and $(W_1, W_2, W_3, ..., W_p)$ then:

$$D_{eu}=\sqrt{(X_1 - W_1)^2 + (X_2 - W_2)^2 + (X_3 - W_3)^2 + ... + (X_p - W_p)^2}$$

Euclidean distance is preferred in K-NN due to many distance computations. The main idea of K-NN is to find K record in the training partition which are neighboring the new observation to be classified. These K neighbors are used to classify the new observation into a predominant class among the neighbors.

6)Boosting means producing a model sequence that aims to correct the errors that have arisen in the models. In boosting based on the previous model's elements that are not properly classified, new samples are produced. Then, by combining the weak models, the ensemble method increases its efficiency.

7)SVM(support vector machine) Given a set of data with N attributes, Support Vector Machine (SVM) classifier is to find a suitable hyper plane in N-Dimensional space that clearly classify the dataset with a maximum margin between data points, where it segregates the two main classes hyper-plane and line to separate the available sets of points, and it is considered a supervised machine learning algorithm which can be used for classification. Support vector machine (SVM) is considered as a supervised machine learning classification technique that is built, based on the concept of decision planes that define decision boundaries. This algorithm function by making "hyperplane" and categories the data based on class values, SVM algorithm performs margin maximization which means it tries to make maximum difference between classes . SVM creates complex non-linear boundaries that are robust to over fitting and the major advantage is high classification accuracy. There are two two types of SVM one is linear and other one is non linear SVM. On this two heart disease datasets we are applying non linear svm. In non linear SVM kernel trick is introduced.Kernel trick helps to map from a low dimension space to the high dimension space. There is a simple operation on two vectors in the low-D space that can be used to compute the scalar product of their two images in the high-D space.

$$K(x^a, x^b)=\phi(x^a).\phi(x^b)$$

Some commonly used kernels:-

1. Polynomial:$K(x,y)=(x.y+1)^p$
2. Gaussian radial basis function (R.B.F)=$K(x,y)$=exp $\frac{-||x-y||^2}{2\sigma^2}$
3. Neural net: $K(x,y)=\tanh(k.x.y-\delta)$ In this work we have used kernel(RBF) SVM, which gives best resut in the second HDD.

2.6 Evaluating Models

For evalauation of the proposed model we are focusing on some criteria, Accuracy, Recall, Precision, F-score. The confusion matrix (see table 3) helps practitioners to form a clear idea of whether the results have a high performance.

Table 3: Confusion Matrix

Class	Predicted Class=0	Predicted Class=1
Actual Class=0	True Negative(TN)	False Positive(FP)
Actual Class=1	False Negative(FN)	True Positive(TP)

The confusion matrix elements were:

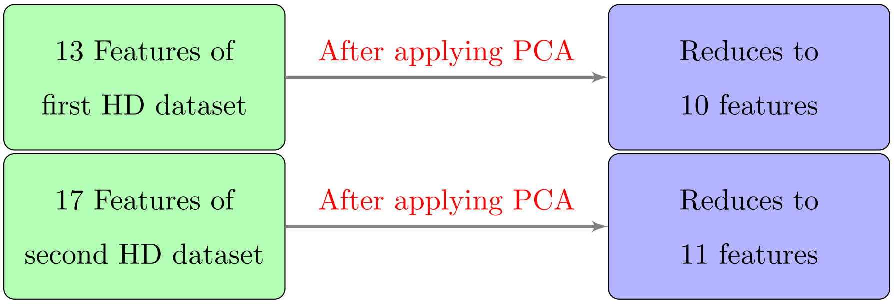
1. True positive (TP), which were patients who had heart disease and were correctly diagnosed;
2. True negative (TN), which were patients who did not have heart disease and were correctly diagnosed;
3. False negative (FN), which were patients who had heart disease and were misdiagnosed; and
4. False positive (FP), which were patients who did not have heart disease and were misdiagnosed. In the medical field, false negatives are the most dangerous predictions

The different performance metrics were calculated using a confusion matrix. Accuracy (Acc) measured the properly classified instances.

- The formula for calculating accuracy was given by:
  Accuracy=$\frac{TP+TN}{TP+FN+FP+TN}$
- Recall identified the proportion of patients with heart disease given by:
  Recall=$\frac{TP}{TP+FN}$
- Precision was the positive predictive value defined by:
  Precision=$\frac{TP}{TP+FP}$
- The F1 score considered a harmonic average between precision in and recall defined by:
  F1 score=$2\left(\frac{Precision * Recall}{Precision + Recall}\right)$

3 Experimental Results

In this section of the work discusses the experimental results of various classification algorithms. At first, we performs the feature scaling with the help of standardization(or Z- score) technique in which the values lie between $[-3, 3]$. Then we are applying Principle component analysis(PCA) which reduce the dimension of the dataset. Here the below flowchart show that total number of features present in both the dataset reduces to the 10 and 11 features.



The main purpose of applying PCA in the dataset is that, it not only reduces the dimension but it also reduces the time taken by the model to the dataset which are generated by us. Then the performance of all used classifcation models i.e. K-Nearest Neighbors (KNN), Decision Tree(DT), Random Forest (RF), Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), Adaboost (AB), Gradient Boosting (GB), XGBoost along with this reduced feature space is evaluated.

Table 4 shows that the Random Forest(RF) achieved the highest performance for the first dataset with accuracy, recall, and precision, which are 100%, 100% and 97% respectively. The worst performance achieved by Logistic Regression with accuracy, recall, and precision, which are 86.34% 78% and 94%. Table 4 has constructed as decresing order of accuracy.

Table 4: Results of different classification algorithms

Techniques	Accuracy(in %)	Recall(in %)	Precision(in %)
Random forest	100	100	97
XGBoost	98.53	100	97
Decision tree	98.53	100	97
AdaBosst	96.58	95	98
Gradient boost	96.09	95	97
SVM (RBF)	90.73	86	95
K-NN	90.24	98	85
Logistic regression	86.34	78	94

Now, after achieving this accuracy, recall and precisions, similar model has applied to the second heart diasease dataset which has large number of rows-319796 and columns-18 then we have evaluating which classification algorithm best perform in this dataset with how much percentage of accuracy?. Similarly,Table 5 shows that theSVM(Kernel =RBF) achieved the highest performance for the second dataset with accuracy, recall, and precision, which are 91.38% 100% and 91% respectively.The worst performance achieved by Decision Tree with accuracy, recall, and precision, which are 87.57% 93% and 93%. Table 5 has constructed as decresing order of accuracy.

Table 5: Performance of the same model which has applied to the 2nd dataset

Techniques	Accuracy(in %)	Recall(in %)	Precision(in %)
SVM (RBF)	91.38	100	91
Logistic regression	91.34	100	91
AdaBosst	91.31	99	92
XGBoost	91.29	99	92
Gradient boost	91.28	99	92
Random forest	90.43	98	92
K-NN	90.39	98	92
Decision tree	87.57	93	93

The obtained results are also illustrated in the figures given below:

Figure 2 shows the performance of the different algorithms. In this figure, we see that the Random forest predicts whether the person has heart disease or not with 100% accuracy, Decision Tree and XGBoost were also predicting a better but slightly less percentage of accuracy in comparison to the Random Forest. In the first heart disease dataset, Naive Bayes gives the worst result with an accuracy of 80%.

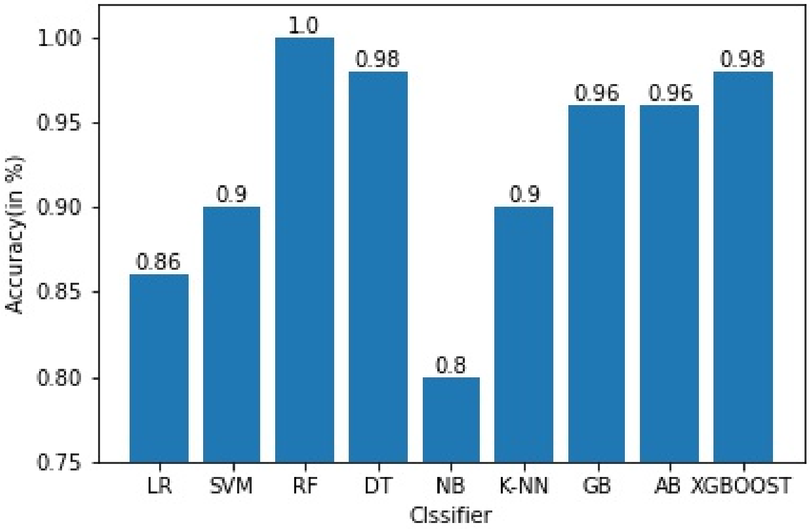


Figure 2: Accuracy comparision of different ML algorithms

Similarly, Figure 3 shows the same model performance to the second dataset. Here we see that Kernel(RBF) SVM predicts well in comparison to the other classification algorithms. Here, except for Decision Tree and Naive Bayes, approximately all the algorithms perform quite well, only the 0.000-0.003% minor variations in the accuracy had shown. Decision Tree gives the worst accuracy which is 86.43%, but Kernel SVM predicts 91.38%.

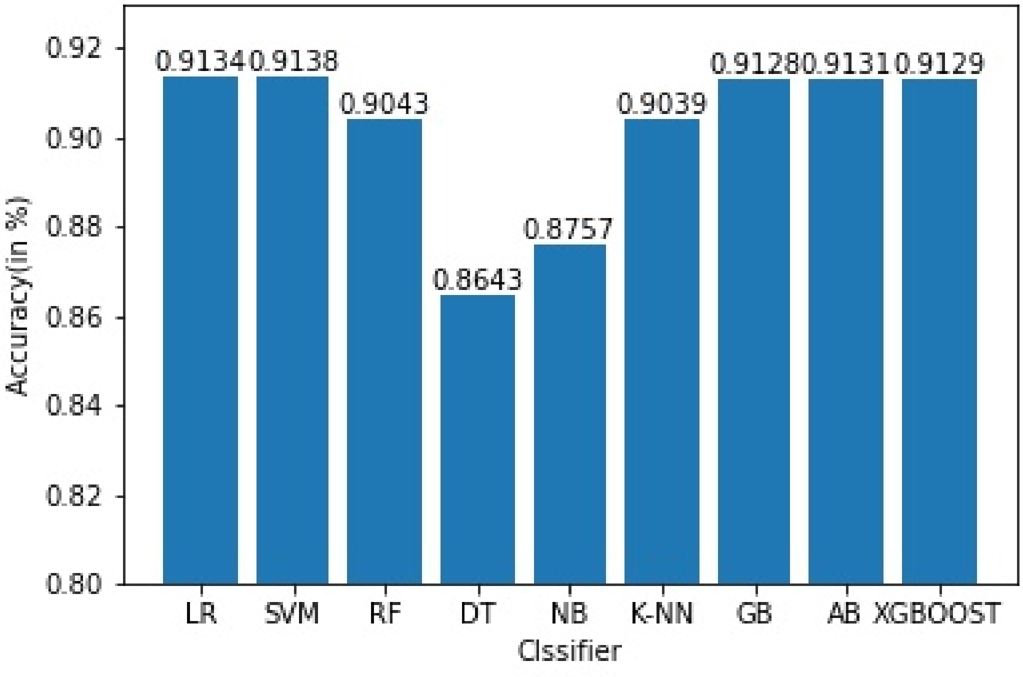


Figure 3: Accuracy comparision of same ML algorithms to the second dataset

Fifure 4 and Figure 5 Present the mixed bar plots of the Accuracy and F1 Score of the different classification algorithms. The greatest result of the first dataset was **Random forest** with 100% accuracy and 100% F1 score. and for the second dataset **kernel SVM** with 91.38% accuracy and 95% F1 score. In Figure 3(1st dataset) we see that F1 score and accuracy have not much variation. But In figure 5 (2nd dataset) F1 score and accuracy of all the algorithms were approximately similar except DT and Naive Bayes.

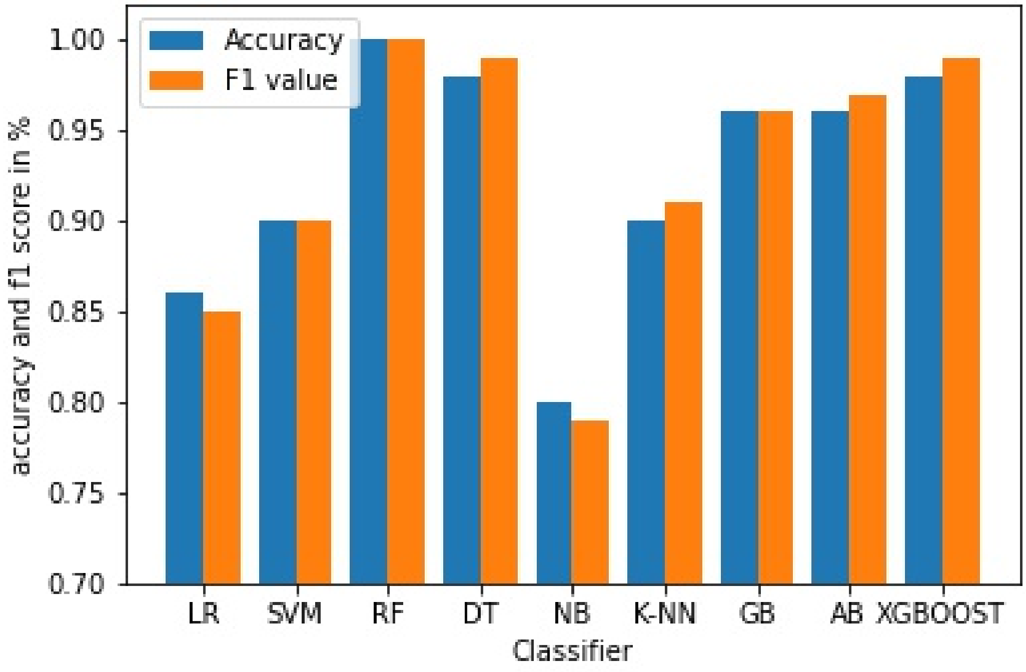


Figure 4: Accuracy vs F1 score for 1st HDD

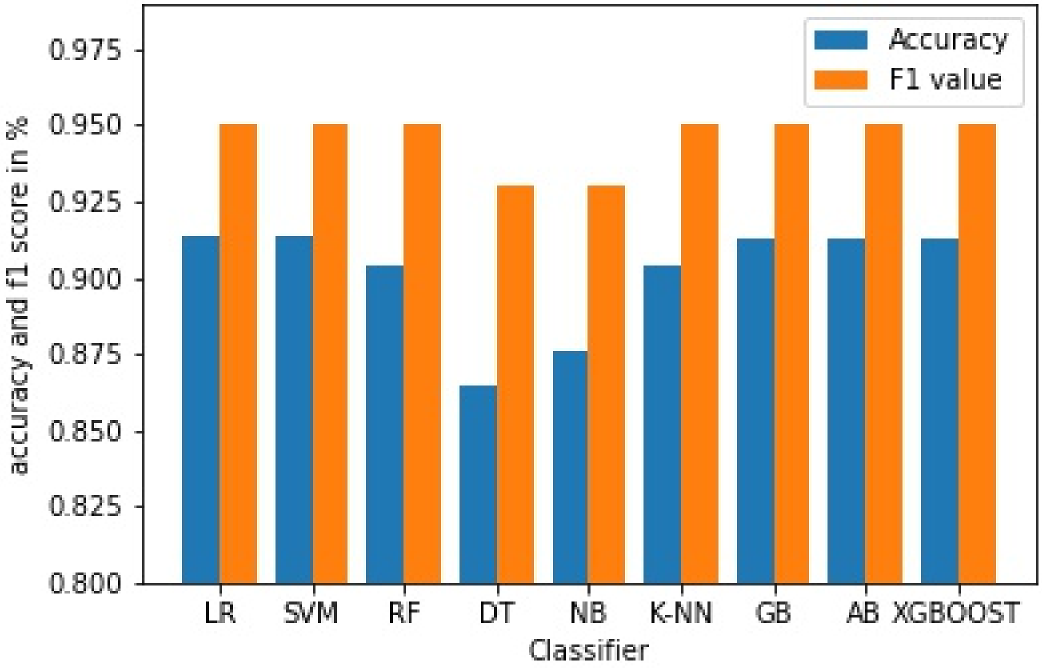


Figure 5: Accuracy vs F1 score for 2nd HDD

3.1 Conclusions

We apply different classification algorithms and some statistcal tools to predict heart disease in this work. Some Ensemble methos (Boosting) with dimension reduction techniques (PCA) were used which help to reduce the time laps covered by different algorithms and predict the heart disease with best accuracy as well. In this work we also compare the different accuracy and F1 score between ensemble methods(GB,AB,XGBoost) and six classifiers(LG,K-NN,SVM(RBF),DT,RF,NB) which were applied after the PCA.

The experimental results showed that the **Random Forest** in first Heart Disease Dataset(HDD) and **kernel(RBF)SVM** in second Heart Disease Dataset had achieved the best performance

3.2 Availability of the Data

1. The first heart disease dataset are available at `https://www.kaggle.com/johnsmith88/heart-disease-dataset`
2. The second heart disease dataset are available at `https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease`

References

[1] Arabasadi, Z., Alizadehsani, R., Roshanzamir, M., Moosaei, H. and Yarifard, A. A. Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. *Comput. Methods Programs Biomed. 141,* 19–26 (2017).

[2] Xiao-Yan Gao,Abdelmegeid Amin Ali, Hassan Shaban Hassan, and Eman M. Anwar. Academic Editor: Ahmed Mostafa Khalil. *Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method.*19 December 2020

[3] Anna Karen Garate-Escamila , Amir Hajjam El Hassani , Emmanuel Andres*Classification models for heart disease prediction using feature selection and PCA*8 January 2020

[4] Yar Muhammad, MuhammadTahir , Maqsood Hayat1 and KilTo Chong *Early and accurate detection and diagnosis of heart disease using intelligent computational model*

[5] Armin Yazdani, Kasturi Dewi Varathan2 , Yin Kia Chiam , Asad Waqar Malik and Wan Azman Wan Ahmad *A novel approach for heart disease prediction using strength scores with signifcant predictors*

[6] Khaled Mohamad Almustafa. *Prediction of heart disease and classifiers' sensitivity analysis*

[7] Machine learning A-Z : Hands-On python and R in Data Science *UDEMY*

[8] Cardiovascular diseases (CVDs). Retrieved from `http://www.who.int/cardiovascular_diseases/en/`

[9] Global Burden of Disease Study. Retrived from `https://en.wikipedia.org/wiki/Global_Burden_of_Disease_Study`