# Analysis of Supervised Machine Learning Algorithms for Heart Disease Prediction with Reduced Number of Attributes using Principal Component Analysis

Ayon Dey
Student, Dept. of CSE,
Gautam Buddha University,
Greater Noida, India

Jyoti Singh
Student,
Agra University,
Agra, India

Neeta Singh
Assistant Professor, Dept. of CSE,
Gautam Buddha University,
Greater Noida, India

## ABSTRACT
Globally, research on causes of death due to heart disease has shown that it is the number one cause of death. If current trends are allowed to continue, 23.6 million people will die from heart disease in coming 2030. The healthcare industry gathers enormous amounts of heart disease data which unfortunately, are not "mined" to discover hidden information for effective decision making. In this paper, study of PCA has been done which finds the minimum number of attributes required to increase the accuracy of various supervised machine learning algorithms. The objective of this research is to analyze supervised machine learning algorithms to predict heart disease.

## General Terms
Classification, Prediction, Heart disease, Reduced attributes, Algorithms, Supervised learning, Classification, Prediction.

## Keywords
Support Vector Machine, Naive Bayes, Decision Tree, Principal Component Analysis.

## 1. INTRODUCTION
According to a recent study by the Registrar General of India (RGI) and the Indian Council of Medical Research (ICMR), in the age group of 25-69 years about 25 percent of deaths occur because of heart diseases [1]. It is the single largest cause of death in the world. Table 1 shows some of the heart diseases and their causes.

Several researchers are using statistical and data mining tools in the diagnosis of heart disease. There are various complex data mining techniques and algorithms which are used in various areas [2] for prediction. Some of the application areas of data mining are given in Figure 2.

Data mining is an essential step of knowledge discovery. It combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from databases. Data mining uses two strategies: 1) supervised learning and 2) unsupervised learning.

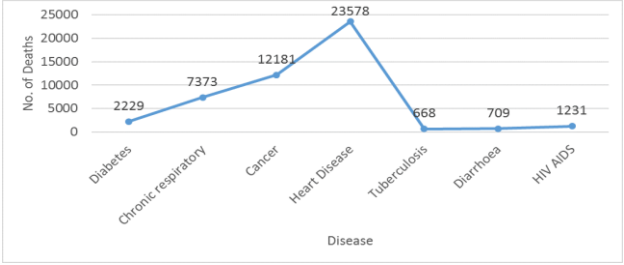


**Figure 1: Projected number of deaths worldwide by 2030 [1]**

**Table 1: Heart diseases and their major causes [1]**

| Diseases | Causes |
|---|---|
| • Coronary Heart Disease (CHD)<br>• Cerebrovascular Disease (stroke)<br>• Hypertensive Heart Disease<br>• Congenital Heart Disease<br>• Peripheral Artery Disease<br>• Rheumatic Heart Disease<br>• Inflammatory Heart Disease | • Tobacco use<br>• Physical inactivity<br>• Unhealthy diet<br>• Harmful use of alcohol |

In supervised learning, a training set is used to learn model parameters whereas in unsupervised learning no training set is used. The workflow [4] of supervised and unsupervised learning algorithms are given in
**Figure 3**.

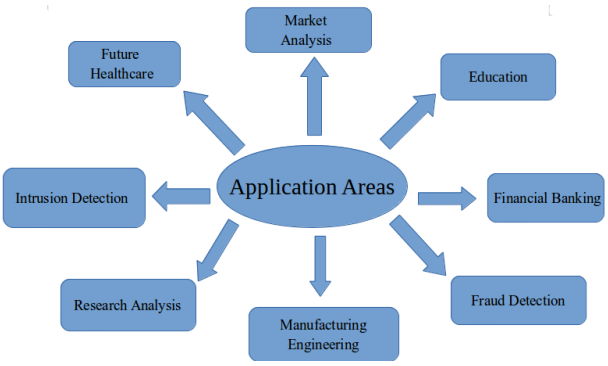


**Figure 2: Applications of Data Mining [2]**

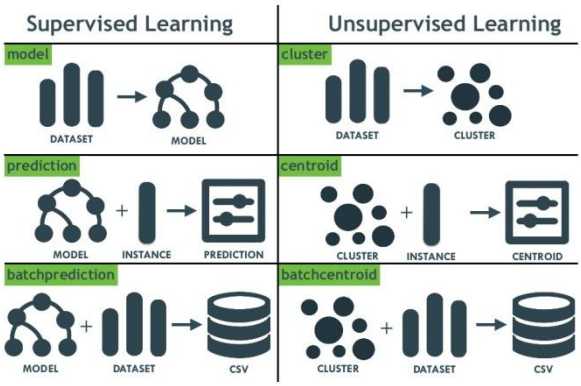


**Figure 3: Workflow of Supervised and Unsupervised Learning Algorithms [4]**

Each data mining technique serves a different purpose depending on the application objective. The two most common models to achieve the objectives are:

- **Classification** models classifies the data according to the discrete class labels.
- **Prediction** models learn from the classification models and predicts values. The values may be discrete or continuous

Some applications of Classification [5] and Predictive Modeling [6] are listed below in Table 2.

**Table 2: Applications of Classification and Predictive Modeling [5] [6]**

| Applications of Classification | Applications of Predictive Modelling |
|---|---|
| • Customer target marketing<br>• Medical disease diagnosis | • Identify trends<br>• Understand customers |
| • Event detection<br>• Multimedia data analysis | • Improve business performance<br>• Strategic decision making |
| • Biological data analysis<br>• Document categorization and filtering | • Predict behavior |
| • Social network analysis | |

With the help of pre-collected data, the future estimation can be examined [3]. Some of them are:

- Predict future trends
- Decision making
- Improve company annual revenue
- Market analysis
- Fraud/spam detection
- Study customer purchase habits

## 2. LITERATURE SURVEY

Several researchers are engaged in data mining study to recognize the pattern which the data shows. In recent years it has attracted great deal of interest in medical industry, market analysis, fraud detection, research analysis, manufacturing industry, etc. [2] 0. Various approaches for the prediction of heart attack risk levels from the heart disease database is done in medical industries. Firstly, the heart disease database is clustered using the K-means clustering algorithm. ID3 algorithm has been used as the training algorithm to show level of heart attack with the decision tree [8]. Data mining techniques like Naive Bayes, J48 decision tree and Bagging approaches are also used for prediction [1] [12]. Three classifiers Naive Bayes, Classification by clustering and Decision Tree were used for prediction along with Genetic algorithm to reduce the size of the data [9]. In another experiment, researchers [10] [11] used Decision Tree, Bayesian classification, KNN, Neural Networks and Classification based on clustering on the same dataset. Later, Genetic Algorithm has also been employed to reduce the size of the dataset.

A classification approach which uses Artificial Neural Network (ANN) and PCA for feature subset selection [13] is used to analyze the dataset. Similarly, several techniques had been used [18] [19] to improve the accuracy of the classifiers for better prediction. Some of them are tabulated in Table 3.

**Table 3: Data Mining techniques and Algorithms [18] [19]**

| No. | Algorithm | Super-vised | Unsuper-vised | Technique |
|---|---|---|---|---|
| 1. | Naive Bayes | ✓ | | Classification |
| 2. | Linear Regression | ✓ | | Regression |
| 3. | Artificial Neural Network | | ✓ | Neural Network |
| 4. | SVM | ✓ | | Classification, Regression |
| 5. | K-Means | | ✓ | Clustering |
| 6. | Quadratic Discriminant Analysis | ✓ | ✓ | Dimensionality Reduction |
| 7. | Decision tree | ✓ | | Tree |
| 8. | K Nearest Neighbor | | ✓ | Clustering |

In this paper analysis of the dataset by supervised machine learning technique has been done. The dataset contains both type of patients i.e. those who does and does not have heart disease. Here three supervised learning algorithms namely Naive Bayes, Decision Tree and Support Vector Machine are considered. Section 28 discusses the methods used. In Section 29 the implementation of supervised algorithms and analysis of the results are discussed. Lastly, this paper is concluded in Section 31.

## 3. DATASET AND ALGORITHMS

### 3.1. Heart Disease Data

The dataset used in this study is from Cleveland Clinic Foundation [7]. The dataset has 11 attributes and 303 rows. Each row corresponds to one particular patient and each attribute corresponds to the observations or type of tests for patients. The description of the attributes is shown in Table 4.

**Table 4: Attributes and their description [7]**

| No. | Attributes | Description |
|---|---|---|
| 1. | Age | Age in years |
| 2. | Gender | 1 Male<br>0 Female |
| 3. | Chest Pain Type | 1 Typical Angina<br>2 Atypical Angina<br>3 Non-Angina pain<br>4 Asymptomatic Pain |
| 4. | Blood Pressure | Blood pressure in mm Hg |
| 5. | Cholesterol | Cholesterol level in mg/dl |
| 6. | Blood Sugar | Is blood Sugar > 120 mg/dl<br>1 True<br>2 False |
| 7. | Resting ECG | 0 Normal<br>1 Having ST-T wave abnormality<br>2 Showing probable or define left ventricular hypertrophy |
| 8. | Max Heart Rate | Maximum Heart Rate Achieved |
| 9. | Exercise Angina | Exercise induced angina:<br>1 Yes<br>0 No |
| 10. | Slope | The slope of the peak exercise segment:<br>1 Up sloping<br>2 Flat<br>3 Down sloping |
| 11. | Diagnosis | 0 Healthy<br>1 Possible heart disease |

## 3.2. Algorithms

### *3.2.1. Naive Bayes*

The Naive Bayes algorithm represents a supervised machine learning method for classification. It uses a probabilistic model by determining probabilities of the outcomes. It is used in diagnostic and predictive problems. Naive Bayes is robust to noise in input dataset. An implementation of Naive Bayes has been illustrated in Figure 4.

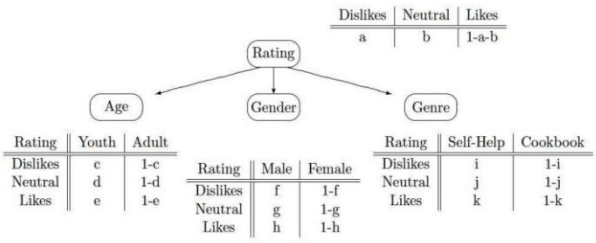


**Figure 4: Naive Bayes [14]**

### *3.2.2. Decision Tree*

Decision tree learning uses a decision tree as a predictive model which maps input about an item to output of the item. Tree models with finite classes of output are called classification trees. In these tree structures, leaves represent class labels and branches represent relation between attributes that results in those class labels. Decision trees with continuous output classes are called regression trees. In data mining, a decision tree can be an input for decision making. An example of decision tree is demonstrated in Figure **5**.

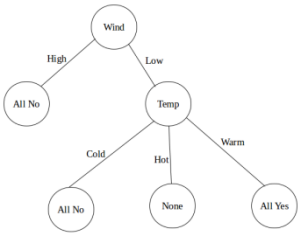


**Figure 5: Decision Tree [15]**

### *3.2.3. Support Vector Machine*

The SVM is a supervised machine learning algorithm for margin classification. It puts a hyperplane between the classes. SVM performs classification tasks by maximizing the margin which separates the classes while minimizing the classification errors. The working of Support Vector Machine is given in Figure **6**.

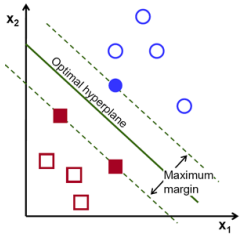


**Figure 6: Support Vector Machine [16]**

### *3.2.4. Principal Component Analysis*

PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The first principal component has the largest possible variance. The variance decreases after each subsequent principal components. The resulting vectors are an uncorrelated basis set 31. The working of PCA is given in **Figure 7**.

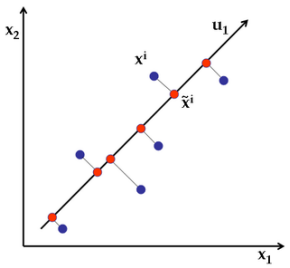


**Figure 7: Principal Component Analysis [17]**

## 4. IMPLEMENTATION AND FINDINGS

First the dataset is divided into training dataset and test dataset. The training dataset is being fed into the algorithms. The algorithms learn from this dataset. Later, in the test dataset, all the columns except the last one are fed in the algorithms. The last column is the actual outcome. The algorithm with the input

data, forms a column of its own. It can do so because it has learned the pattern from the training dataset. The predicted column given by the algorithm is then compared to the actual column in the dataset. This comparison gives the required accuracy. The work-flow of this research work has been depicted as a flowchart in Figure 8. This work has been implemented in Python.

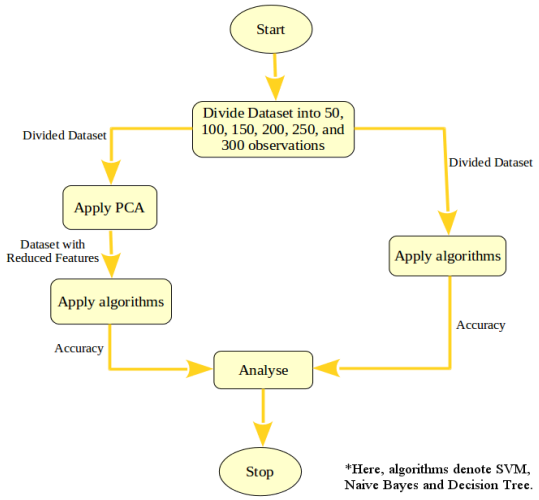


**Figure 8: Flowchart of the approach**

In this work, the dataset is fed in the intervals of 50. The sequence in which the observations are given are 50, 100, 150, 200, 250 and 303 observations respectively. These observations are then divided into train and test dataset. Both datasets are then combined with three classifiers to predict values. Based on their performance, their accuracy is generated.

When the accuracies of algorithms are plotted against the number of observations, a graph is generated. The graph is shown in Figure 9

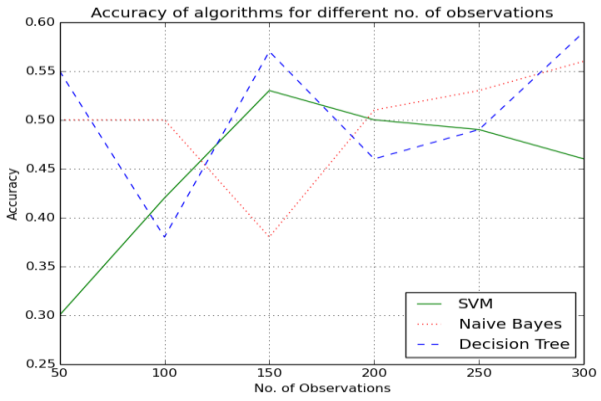


**Figure 9: Number of observations vs Accuracy**

After applying PCA on these three algorithms, their accuracies are being evaluated which is shown in Figure 10. It is also observed from the graph that different algorithms show different accuracy when combined with PCA and when different number of attributes are selected. Based on the accuracies of the classifiers, a score is calculated. The equation for the score is:

$$Score = 3 \times \frac{svm \times nb \times dt}{svm + nb + dt}$$

where, svm = accuracy of SVM,

nb = accuracy of Naive Bayes

and, dt = accuracy of Decision Tree for no. of selected feature(s)

It can be observed from

Figure 10, that the score is maximum with 6 attributes. Thus the dataset of 10 input attributes is reduced to 6 attributes.

The accuracy of SVM, Naive Bayes and Decision Tree after applying PCA are shown in Figure 11, Figure 12 and Figure 13 respectively. It can be observed that the accuracy of SVM has been increased dramatically with PCA as compared to Decision Tree and Naive Bayes.

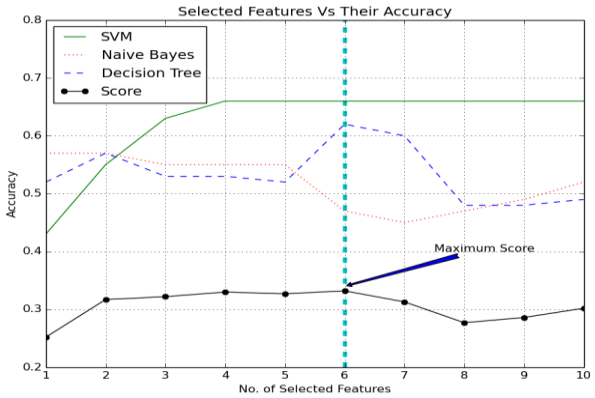


**Figure 10: Selection of attributes using PCA**

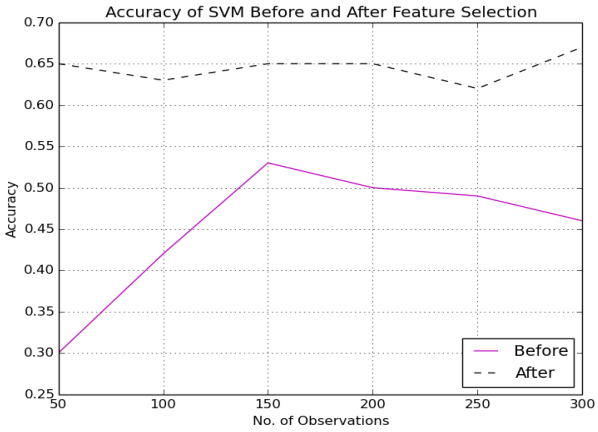


**Figure 11: Accuracy of SVM Before and After Feature Selection**

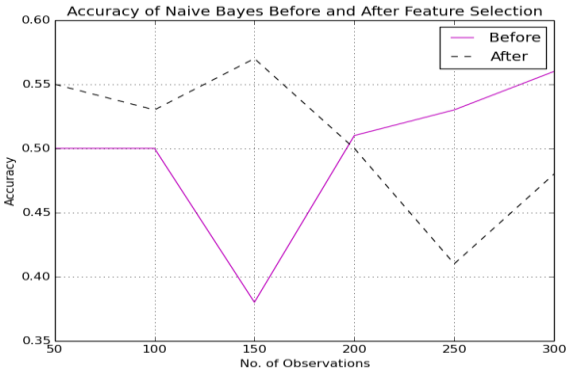


**Figure 12: Accuracy of Naive Bayes Before and After Feature Selection**

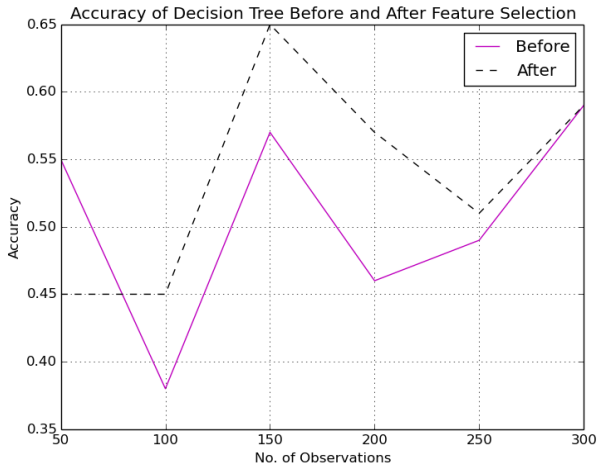


**Figure 13: Accuracy of Decision Tree Before and After Feature Selection**

## 5. CONCLUSION

In this paper, SVM, Naive Bayes and Decision tree has been applied with and without using PCA on the dataset. We used PCA to reduce the number of attributes. After reducing the size of the dataset, SVM outperforms Naive Bayes and Decision tree. SVM can further be used to predict heart disease. A GUI desktop application can be built using SVM and this dataset to predict the possibility of cardiovascular disease in a patient.

## 6. REFERENCES

[1] Vikas Chaurasia, Saurabh Pal, - "Early Prediction of Heart Diseases Using Data Mining Techniques", Carib.J.SciTech, 2013, Vol. 1, 208-2017

[2] http://bigdata-madesimple.com/14-useful-applications-of-data-mining/

[3] http://bus237datamining.blogspot.in/2012/11/advantages-disadvantages.html

[4] http://www.slideshare.net/bigml/big-ml-spring-2014-webinar-clustering

[5] Charu C. Aggarwal - "Data Classification: Algorithms and Applications", Chapman and Hall, CRC Press

[6] http://www.sas.com/en_us/insights/analytics/predictive-analytics.html

[7] http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[8] Hnin Wint Khaing, "Data Mining based Fragmentation and Prediction of Medical Data", IEEE, 2011

[9] M. Anbarasi, E. Anupriya, N. Ch. S. N. Iyengar, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", International Journal of Engineering Science and Technology, Vol. 2(10), 2010, 5370-5376

[10] Jyoti Soni, Ujma Ansari, Dipesh Sharma and Sunita Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011

[11] M. Pechenizkiy, A. Tsymbal and S. Puuronen, "PCA-based feature transformation for classification: issues in medical diagnostics", IEEE, Computer-Based Medical Systems, 2004. CBMS 2004. Proceedings (1063-7125), Page No. 535 – 540, June 2004

[12] V. Chaurasia and S. Pal, "Data Mining Approach to Detect Heart Diseases", International Journal of Advanced Computer Science and Information Technology (IJACSIT), Vol. 2, No. 4, 2013, Page 56-66.

[13] Akhil Jabbar, B.L Deekshatulu and Priti Chandra, "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection", Global Journal of Computer Science and Technology Neural & Artificial Intelligence, Volume 13, Issue 3, Version 1.0, Year 2013, ISSN: 0975-4172

[14] https://webdocs.cs.ualberta.ca/~greiner/C-651/Homework2_Fall2008.html

[15] http://jeremykun.com/2012/10/08/decision-trees-and-political-party-classification/

[16] http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html

[17] https://alliance.seas.upenn.edu/~cis520/wiki/index.php%3Fn=Lectures.PCA

[18] Saba Bashir, Usman Qamar, Farhan Hassan Khan, "BagMOOV: A novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting", Australasian Physical & Engineering Sciences in Medicine, Volume 38, Issue 2, pp 305-323, March 2015

[19] Neeraj Shah, Valay Parikh, Nileshkumar Patel, Nilay Patel, Apurva Badheka, Abhishek Deshmukh, Ankit Rathod, James Lafferty, "Neutrophil lymphocyte ratio significantly improves the Framingham risk score in prediction of coronary heart disease mortality: Insights from the National Health and Nutrition Examination Survey-III", International Journal of Cardiology, Volume 171, Issue 3, Pages 390–397, February 2014

[20] Mai Shouman, Tim Turner, Rob Stocker,(2012),"Using Data Mining Techniques In Heart Disease Diagnosis And Treatment ",Proceedings in Japan-Egypt Conference on Electronics, Communications and Computers,IEEE,Vol.2 pp.174-177