# Efficient and Reliable Prediction of Heart Disease using integration of Genetic Algorithm and Naive Bayes algorithm in Python.

R. Suresh1 Research Scholar
Department of Computer Science & Engineering, Osmania university,Hyderabad
Dr. Nagaratna P. Hegde2 Professor
Department of Computer Science and Engineering Vasavi College of Engineering

**Abstract:**

Data mining process is discovering knowledge and patterns from extensive collection of data. Hospitals generate a boundless data daily. But, most of Data not used effectively. Use of organized tools for extracting knowledge from clinical databases is not used widely. As per World Health Organization 120 lakh people deaths per year because of heart diseases. Heart disease is a condition, where the heart is incompetent to drive necessary amount of blood to various parts in the body. Precise and timely diagnosis of heart disease is significant for heart failure avoidance and treatment. Traditional Diagnosis of heart disease through clinical findings is not reliable. Machine learning application is reliable and efficient to find healthy people and people with heart disease. The theme of the paper is to use machine learning algorithms to predict heart disease by summarizing the present researches. In this paper the Genetic algorithm and Naive Bayes algorithm is used in the health care dataset to classify the patients if possessing heart diseases or not based on the dataset attributes.

Keywords: Machine Learning, Naive Bayes Classification.

## 1. Introduction

Heart failure occurs because of heart unable to pump required blood to other parts of the body to perform normal functions. India heart disease rate is very high, symptoms of heart related diseases include high blood pressure, high fasting blood sugar, decrease of breath, weakness of body, puffed feet, and fatigue leads to cardiac abnormalities. In previous days heart disease prediction is complicated noticed by inspecting various research papers. In developing countries like India due to limited availability of diagnostic devices and shortfall of general physicians effected accurate prediction and treatment of heart disease patient. Heart associated risk issues can be improved by accurate diagnosis of heart disease patient .Machine learning uses computers for associating with process statistics to create accurate predictions.

Machine learning is part of data science, learning of computers is done with training from different datasets on the basis of generated patterns from the data. Algorithms developed for machine learning enables the computer to learn on its own based on input datasets.

Machine learning is extensively used on clinical data findings for effective prediction of various diseases. To overcome complexities in previous diagnosis of heart disease, a clinical decision

Decision making models developed based on machine learning models like Naïve Bayes, fuzzy logic ,logistic regression,k-means,neural networks, decision tree, and many more used by assorted researchers for diagnosis of heart disease due to early detection and associated care taken had decreased the ratio of heart disease deaths .

## 2. Literature Survey

Heart disease risk factors identified by statistical analysis on Medical dataset are age, BP, chol, fbs, thalach attributes .Doctors or medical staff with expertise of factors linked with heart disease aids to identify a individual with likelihood of experiencing a heart disease. Predictions obtained by different data mining algorithms cannot be compared because of different attributes, different datasets and different number of records.

Major research is carried out using a bench mark dataset in heart disease prediction called Cleveland heart disease dataset. Various researches investigated different data mining techniques used in heart disease diagnosis and determining various treatment procedures. Major issues in Heart disease diagnosis and prediction are obtaining accurate results. Doctors and health care professionals are now using data mining techniques to handle complexities and error in treatment of heart disease.

Mai Shouman, Tim Turner, and Rob Stocker
Given a model combining various early centroid selection for k-means algorithm used for clustering with Naive bayes classification for predicting heart disease patients showing accuracy of 84.5%.Advanatages of this model are, Various initial centroid selection are investigated in k-means clustering and Best accuracy with random row method **84.5** %.limitation are two clusters show better than others, No of instances are less, Target attributes have only two attributes health and sick.

M.Akhiljabbar B.LDeekshatulua and Priti Chandra Proposed an model to combine KNN and genetic algorithm for diagnosis of Heart disease patients. Advantages are It replies to nexus questions for determining existence of heart disease and assists Doctors to make wise resolution .Limitation are 15 medical attributes are incorporated more attributes needed to be considered and Only categorical data is used.

Razali and Ali (2009) investigated using a
Decision tree generated treatment process for upper respiratory infection disease patient. Obtained accuracy of 94.73 % by
Recommending drugs through treatment. Best performances are obtained by Association rules and decision trees. In future Comparisons with data mining techniques like genetic algorithms means and Naive bayes needs to be investigated.

AH Chen et al. presented heart disease prediction model which helps doctors in predicting a heart disease found on the clinical data of patients. Thirteen important clinical data features such as age, sex, cholesterol were selected. Machine learning, datasets from clinical Repositories and ANN are used for classifying heart disease and given accurate prediction model with 80%.

Carlos Ordonez had used association rule mining with the training and testing idea on a dataset for heart related disease prediction. The main disadvantage is it produced large number of rules which are complex and irrelevant to medical field. Validation on independent sample is not done. To reduce the number of rules devised an algorithm which uses search constraints. The rules obtained had valuable medical features and obtained accurate prediction model.

Serdar AYDIN et al. had used various DM techniques for diagnosing heart disease patients, studied and compared accuracy of various methods .Various DM techniques used are Random forest, Bagging, AdaBoostM1, RBF, Naive Bayes, NN etc. The dataset used from Long Beach VA contained 200 records each had 14 attributes. Data mining techniques are analyzed using WEKA tool. The comparative Results showed that RBF Network has the highest accuracy of 88.20%,in the determining of heart disease

## 3. Proposed method

The proposed approach combined genetic algorithm and Naïve Bayes to enhance the accuracy of determining heart disease based on data set obtained from various Medical records. Genetic algorithm is used as a fitness count to crop inessential and duplicate attributes, and to give feature importance values to the attributes towards effective prediction. Less value of feature importance attributes are pruned, and Naive Bayes classification algorithm is built based on selected attributes. The Naive Bayes algorithm is trained and tested to classify a person with a heart disease or not.

**Genetic algorithm**

Data mining used ideas from biological theory for problem Optimization. John Holland in 1975 invented Genetic algorithm useful for search and optimization problems. Each problem solution denoted by chromosomes. Chromosomes are collection of, independent attributes which constitute the problem. Population is defined as group of all chromosomes.

Advantages with genetic algorithm are discussed below

1) Wider result space
2) Global optimum is simple to uncover
3) Function evaluations are only used
4) Incompetent data are controlled well.

Limitations of genetic algorithms are

1) Complexity involved in identifying suitable fitness function
2 Greater no of fitness assessments are required
3) Genetic algorithm is not appropriate to identify local optima.

**Naive Bayes**

Naive Bayes is uncomplicated technique for probabilistic classifier based on Bayes theorem. Given class attribute Naive Bayes assumes the value of any attribute is independent of the value of any other attribute. Bayes theorem is: P (A|B) = P (B|A) * P (A)/P(B)

Proposed algorithm is divided into two parts.

1) First Module deals with statistical Analysis and removing irrelevant attributes by genetic Algorithm.
2) Second module deals with training and testing Naive Bayes algorithm and obtains accuracy in prediction.

**Proposed algorithm**

1. Selected dataset is loaded
2. Apply genetic algorithm on attributes of the data set.
3. Rank all the Attributes based on their feature importance value.
4. Select the subdivision of high ranked attributes.
5. Apply Naïve Bayes algorithm on the subset of selected attributes to maximize accuracy.
6. Calculate classifier accuracy.

$$\text{Accuracy} = \frac{\text{no. of records correctly classified}}{\text{Total no.of records}}$$

**4. Results and discussion**

Four combined datasets are taken for research. Three data sets are taken from UCI Repository and and one from heart disease set taken from various diagnosis centres of Telangana state . Attributes are selected from feature selection by genetic algorithm .

Used various PYTHON libraries like PANDAS to takeout the data frame to perform data processing. Sklearn library for implementation of Genetic algorithm and Naive Bayes . MATPLOTLIB to Visualize the outputs obtained. Used both Genetic and Naive Bayes algorithm on heart disease dataset obtained accuracy of prediction and various other parameters defined in confusion matrix .In proposed work 16 Attributes used are, Age, cp, sex, chol, trestbps , restecg, fbs ,exang,thalach,slope thalach, slope, oldpeak,ca,tropt,bnp,thal and target. Figure 1 to show Mat plot for important attributes based on feature importance. Figure 2 shows Mat plot for Heart disease frequency with respect to Attribute Age. Figure 3 shows Scatter plot between Attributes Maximum Heart Rate and Age. Figure 4: Scatter plot Between Attributes Resting Blood Pressure and Thalach.

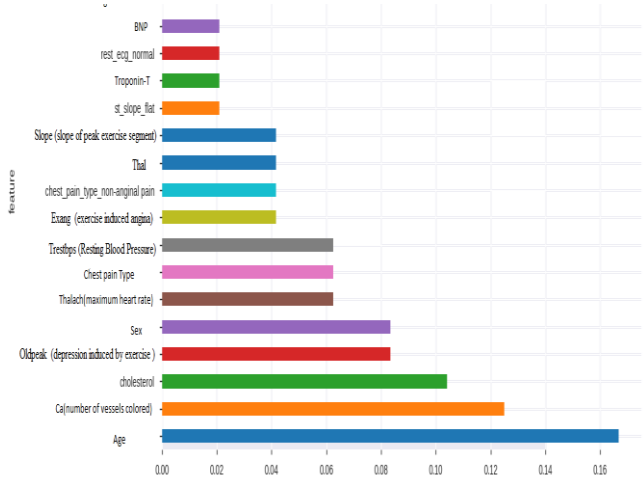


Figure1: Mat plot attributes with Feature importance

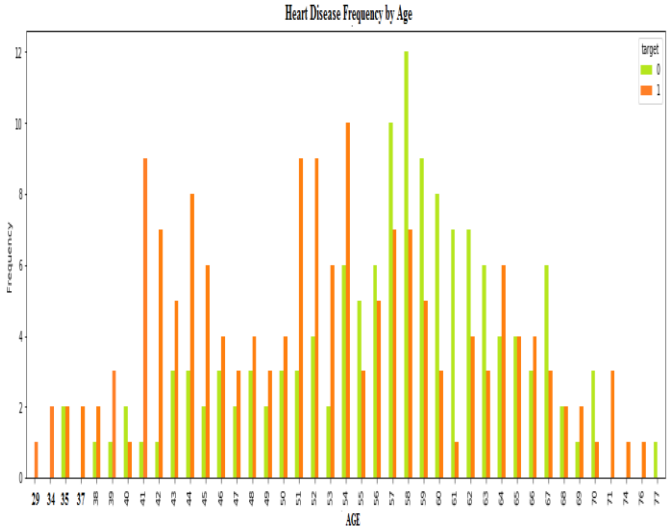


Figure2: Heart Disease Frequency with respect to attribute Age.

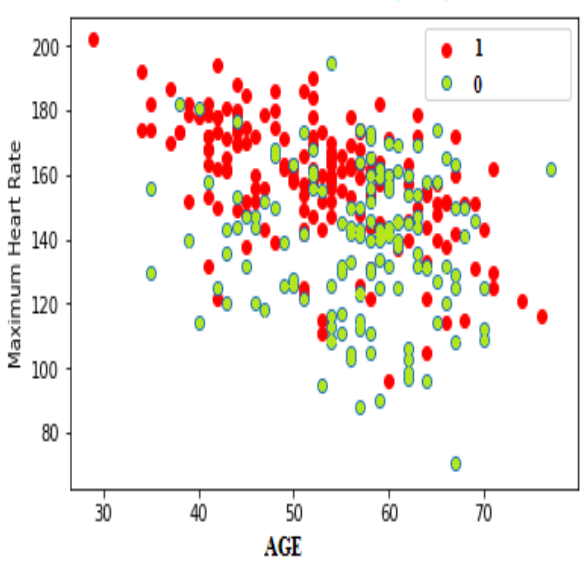


Figure 3: Scatter Plot between Attribute Maximum Heart Rate and Age.

Table 1. Comparison of Confusion Matrix of the Proposed Model

| Method used | Class | Precision | Recall | F-Measure | Accuracy |
|---|---|---|---|---|---|
| Naive Bayes | Positive | 0.74 | 0.89 | 0.81 | 0.754 |
| | Negative | 0.85 | 0.76 | 0.95 | |
| Genetic Algorithm | Positive | 0.89 | 0.78 | 0.83 | 0.823 |
| | Negative | 0.79 | 0.86 | 0.84 | |
| **Proposed Work Naive Bayes +Genetic** | Positive | 0.9 | 0.79 | 0.86 | **0.894** |
| | Negative | 0.75 | 0.91 | 0.87 | |

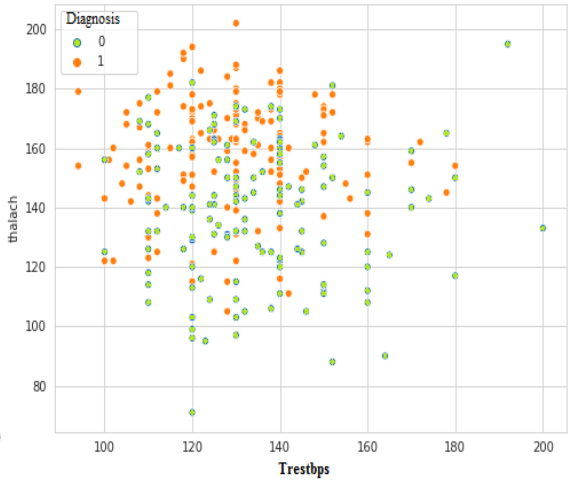


Figure 4: Scatter plot Between Attributes Resting Blood Pressure and Thalach

Used Genetic algorithm for feature selection and selected features are taken in a subset and Naive Bayes algorithm is applied on subset ,accuracy of classifier is obtained, other parameters defining accuracy is defined in the form of confusion matrix .

Results showed the mean of accuracy obtained from the present model is high compared to Naïve Bayes and Genetic algorithm taken individually .The percentage of increase in prediction accuracy by 6 to 14 %.

## 5. Conclusion

In Proposed work presented a way for effective classification to predict heart disease with greater accuracy considering various parameters defined in the form of confusion matrix. To validate, used k fold cross validation. The model proposed can be used by doctors for in time heart disease diagnosis to prevent premature deaths.

## References

[1] Mai Shouman, Tim Turner and Rob Stocker, Integrating Naive Bayes and K-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients, 2012

[2] I.H. Witten, E. Frank, Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann 2005.

[3] SellappanPalaniappan, RafiahAwang, Intelligent Heart Disease Prediction System using Data Mining Techniques, 2008 IEEE.

[4] V. SreeHariRao, M.NareshKumar, Novel Approaches for Predicting Risk Factors of Atherosclerosis, 2013 IEEE.

[5] Mai Shouman, Tim Turner and Rob Stocker, Using Data Mining Techniques in Heart Disease Diagnosis and Treatment, 2012 IEEE.

[6] Monika Gandhi, Dr. Shailendra Narayan Singh, Predictions in Heart Disease using Techniques of Data Mining, 2015 IEEE conference.

[7] Theresa Princy. R, J. Thomas, Human Heart Disease Prediction System using Data Mining Techniques, 2016 IEEE Conference.

[8] Mohammad HosseinTekieh, BijanRaahemi,Importance of Data Mining in Healthcare: A Survey, 2015 IEEE Conference.

[9] AnkitaDewan, Meghna Sharma, Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification, 2015 IEEE.

[10] ZoubidaAlaouiMdaghri, Mourad El Yadari, AbdelillahBenyoussef, Abdellah El Kenz, Study and analysis of Data Mining for Healthcare, 2016 IEEE Conference.

[11] J. Thomas, R Theresa Princy,Human heart disease prediction system using data mining techniques ,in: Circuit, Power and Computing Technologies (ICCPCT), 2016 .

[12] SeyedaminPouriyeh, Sara Vahid, Giovanna Sanninoy, Giuseppe De Pietroy, Hamid Arabnia, Juan Gutierrez ,A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease, 22nd IEEE Symposium on Computers and Communication (ISCC 2017).