Heart Disease Prediction System Using Classification

Aditya Rai[1], Anvay Pakhale[2], Shashank Salian[3], Prachi Janrao[4]

[1](Computer Engineering, University of Mumbai, India)
[2](Computer Engineering, University of Mumbai, India)
[3](Computer Engineering, University of Mumbai, India)
[4](Computer Engineering, University of Mumbai, India)

Abstract: *Heart Disease Prediction System (HDPS) is an application that predicts occurrence of Heart Disease from medical data like pain location, smoking, diabetes, whether pain relieved after rest, years, family history using data mining technique.*

Early identification and prediction of Cardiovascular disease (CVD) is important for treatment which in turn can decrease mortality. Data mining plays an essential role in the field of heart disease prediction. Hence, there emerges a need to develop a computerized system to check the condition of a heart in a patient.

The main aim of this project is to reduce the individual and government healthcare expenditure. The traditional medical tests like MRI scan, ECG, CT scan are not only expensive but also requires heavy machinery and skilled operators to operate the machinery. The Heart Disease Prediction System (HDPS) will be able to reach a large amount of population residing in the remote areas through internet and will also eliminate the heavy machinery and operators.

The prediction of the HDPS will be based on machine learning algorithms. The algorithm which has highest accuracy while training on a dataset will be chosen for deployment of the system.

Keywords: *Cardiovascular Disease, Cost function , Decision Trees, Logistic regression, Random Forest Classifier, Sigmoid function.*

I. Introduction

Heart disease is the form of cardio-vascular disease which is prevailing in the whole world like a wave and becoming a bigger cause of deaths. Latest statistics suggest that in India, there are roughly 30 million heart patients. According to Heart Care Foundation of India this number will continue to increase due to things like stress, unhealthy eating habits, lack of physical exercise, lack of sleep and dependence on alcohol and cigarettes.

Cardiovascular diseases (CVD) are a class of diseases that involve heart and blood vessels. Cardiovascular diseases include coronary artery diseases (CAD) such as angina and myocardial infarction (commonly known as a heart attack)

Traditional methods of predicting Heart disease includes doctor's examination or number of medical tests such as Electrocardiography (ECG), Stress Test, and Heart MRI scan etc. Nowadays, Health care industries contain large amount of heath care data, which contains hidden information. The conventional diagnosis and tests like ECG, MRI scan are used for checking whether heart is healthy or not. But these tests are expensive as heavy imported machinery, skilled labour etc. are involved. The main aim of the healthcare policy of any country is to provide cheap but effective healthcare to every bonafide citizen. As India is a developing country, the government only spends 1.4% of the GDP (In FY18) on healthcare. So, there arises a demand for innovation, for developing products that can significantly reduce the costs.

Nowadays, Data mining techniques like classification are playing a vital role in the biomedical field as they are used to explore, analyze and extract medical data using complex algorithms to discover unknown patterns. Any data mining technique has two primary goals i.e. prediction and description. Prediction involves using parameters or fields present in the data set to predict unknown or future values of other variables of interest. While description focuses on finding patterns that describes the data suitable for human interpretation.Researchers are using data mining techniques for the diagnosis of many diseases such as heart disease, diabetes, stroke and cancer and many data mining techniques together with machine learning algorithms with good accuracy.Machine learning algorithms are useful techniques to extract flexible and comprehensible knowledge from huge datasets. These are less complicated to implement, and their results are more easily understood to users. Machinelearning techniques deal with a mix of quantitative, qualitative, missing or noisy data so common on engineering.

II. Literature Survey

The conventional clinical diagnosis is dependent on doctor's practice and expertise. Nowadays, health disease is increasing day by day due to life style, hereditary etc. Now a days, heart disease has become more common. By using data mining techniques, it takes less time for the prediction of the disease with more accuracy and reduced cost. Computer Aided Decision Support System plays a major task in medical field. Different types of studies have been done to focus on prediction of heart disease. Various data mining techniques are used for diagnosis and achieved different accuracy level for different methods.

KNN is a non-parametric strategy which is utilized for order and relapse. Contrasted with other machine learning calculation KNN is the most straightforward calculation. This calculation comprise K-wardrobe preparing models in the element space. In this calculation K is a client characterized consistent. The test information are grouped by allocating a consistent esteem which is most interminable among the K-preparing tests closest to the point. Writing demonstrates the KNN has the solid consistency result.

Decision tree fabricates grouping models as a tree structure. It breaks the dataset into littler subset while at the same time a related choice tree is steadily created. The choice tree utilizes a best down methodology strategy. The foundation of the choice tree is the informational collection and the leaf is the subset of the informational index. Mixture Algorithm is the mix of KNN calculation and ID3. These calculations are utilized for coronary illness expectation. The KNN calculation is utilized to preprocess the information; it is called as preprocessed calculation. The preprocessed information are considered as preparing set and after that the information has been grouped into a tree structure. The ID3 calculation is connected for the classifier to foresee the heart ailment. The mistaken qualities are ordered through KNNCalculation.

Heart Disease Prediction using Machine Learning Techniques[1]:This paper byV.V. Ramalingam*, Ayantan Dandapath, M Karthik Raja presents a survey of various models based on such algorithms and techniques and analyse their performance. Models based on supervised learning algorithms such as Support Vector Machines (SVM), K-Nearest Neighbour (KNN), Naïve Bayes, Decision Trees (DT), Random Forest (RF) and ensemble models have been studied.

Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review[2]:This attempt by Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee and Asmita Mukherjee has presented various data mining algorithms like Decision Trees, K-means algorithm, Naïve Bayes ,etc. and data mining tools such as WEKA, RapidMiner, TANAGRA.

REVIEW OF HEART DISEASE PREDICTION SYSTEM USING DATA MINING AND HYBRID INTELLIGENT TECHNIQUES[6]:This paper by R. Chitra and V. Seenivasagam. In this paper Heart attack prediction system methodology is categorized in three types. At first type data mining technique (mainly classification technique) are analysed. The second type intelligent techniques used for heart disease prediction are analysed. The final type the role of feature subset in the heartdisease prediction is discussed.

III. Proposed Work

Algorithm:

A. Logistic Regression:

Logistic regression falls under the category of supervised learning. Sigmoid function is used in logistic regression. Despite the name logistic regression, this is not used for regression problem where the task is to predict the real-valued output. It is a classification problem which is used to predict a discrete output (1/0, -1/1, True/False).

y=x+x1*w1+x2w2+x3w3+…..+xnwn ...(i)

$$y = logistic(c + x1 * w1 + x2 * w2 + x3 * w3 + \cdots . + xnwn)$$

…(ii)

$$y = 1/1 + e[-(c + x1 * w1 + x2 * w2 + x3 * w3 + \cdots + xnwn)]$$

Sigmoid Function (Logistic Function):

The predicted value of the logistic regression is between negative infinity to positive infinity. The output of the algorithm is 0-no, 1-yes.

$$z = \theta 0 + \theta 1 \cdot x1 + \theta 2 \cdot x2 + \cdots \cdot \qquad \text{…(iii)}$$

h=g(x)= $\frac{1}{1+e^{-x}}$ …(iv)

Cost Function:

The cost function for logistic regression is given by,

$$Cost(h_\theta, y) = -\log\big(h_\theta(x)\big) \; if \; y = 1$$
$$Cost(h_\theta, y) = -\log\big(1 - h_\theta(x)\big) if \; y = 0$$

…(v)

In the below Fig 1, the green line represents the decision boundary and red and blue points in the plot indicate two discrete outputs.

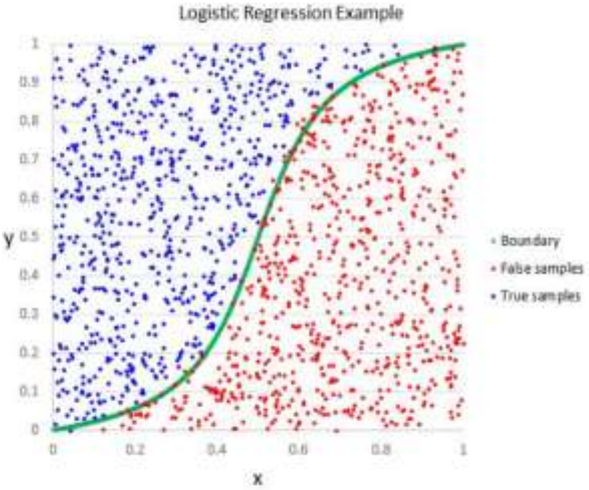


Fig 1: Fig Courtesy: Data Science Central

D. Random Forest:

Precondition: A training set S(et) := (x1, y1), . . . ,(xn, yn), features F, and number of trees in forest B.

```
i) function RandomForestClassifier(Set , F)
ii) h ← Ø
iii) for i ∈ 1, . . . , B do
iv) S (i) ← A sample from Set
v) hi ← RandomizedTreeLearn(S (i) , F)
vi) (h)← (h)H ∪ {hi}
vii) end for
viii) return h
ix) end function
x) function RandomizedTreeLearn(Set , F)
xi) At each node:
xii) f ← very small subset of F
xiii) Split on best feature in f
xiv) return the learned tree
xv) end function
```

Fig 2 given below indicates based on dataset X and features it is further divided into trees and based on majority voting the final output is obtained.

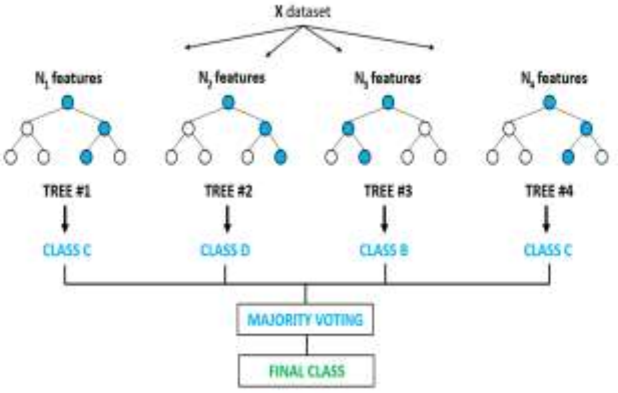


Fig2: Fig Courtesy: Global Software Support

C. Decision Trees:

```
decisionTreeLearning(examples,attributes,parent_examples)

 if len(examples) == 0:
return pluralityValue(parent_examples)
   #return most probable answer as no training data left

eliflen(attributes)==0:return pluralityValue(examples)
elif (all examples classify the same): return their classification

A = max(attributes, key(a)=importance(a, examples)
# choose the most promising attribute to condition
```

In fig 4 given below, based on Boolean logic decisions are made at every node of the decision tree.

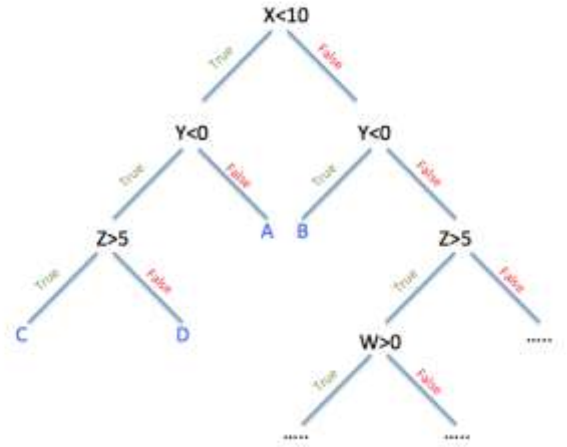


Fig 3: Figure Courtesy: Medium

E. K Nearest Neighbors (KNN):

Algorithm
1. Compute a distance value between the item to be classified and every item in the training data-set
2. Pick the k closest data points (the items with the k lowest distances)
3. Conduct a “majority vote” among those data points — the dominating classification in that pool is decided as the final classification**.**

Pseudocode:

```
1. Load the training and test data
2. Choose the value of K
3. For each point in test data:
       - find the Euclidean distance to all training data points
       - store the Euclidean distances in a list and sort it
       - choose the first k points
       - assign a class to the test point based on the majority of the classes present in the chosen points
4. End
```

General formula for Euclidean distance:

$$E(x, y) = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2}$$

General formula for Euclidean distance

Based on proximity of training data points and majority voting the output is obtained as shown in Fig 4 below.

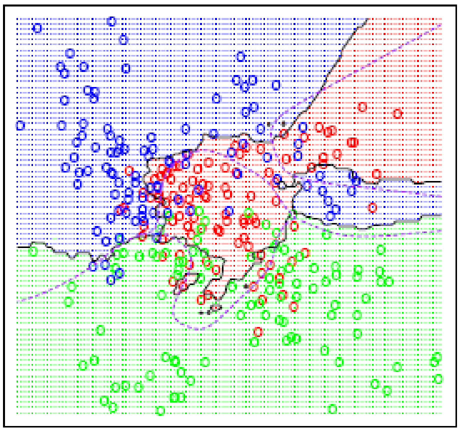


Fig 4: Figure Courtesy: Towards Data Science

F. Flowchart of HDPS

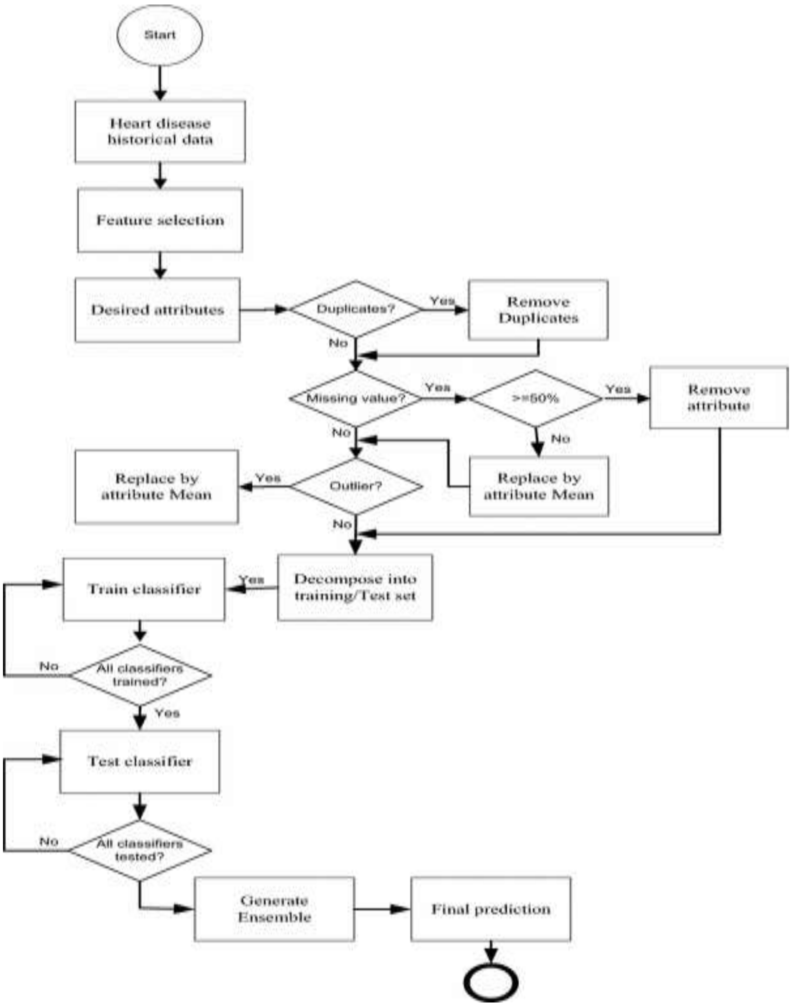


Fig 5

Fig 5 shows the flow of the heart disease prediction system from feature selection, cleaning the dataset by removing the duplicates or by filling up the missing values, training the model on training dataset and prediction by the system based on real time input of data by the user.

G. System Architecture:

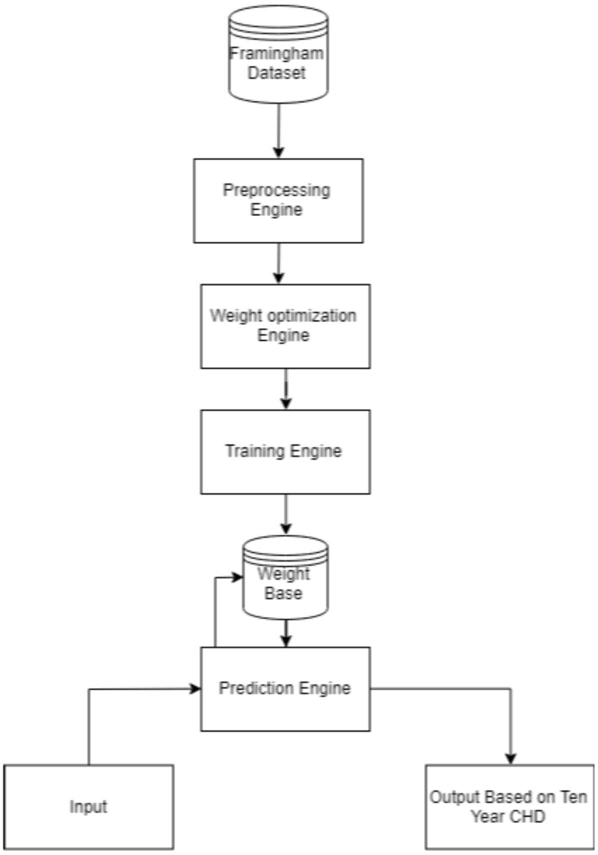
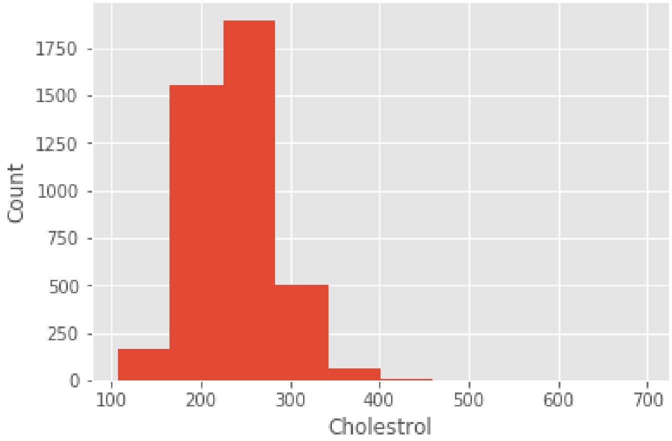


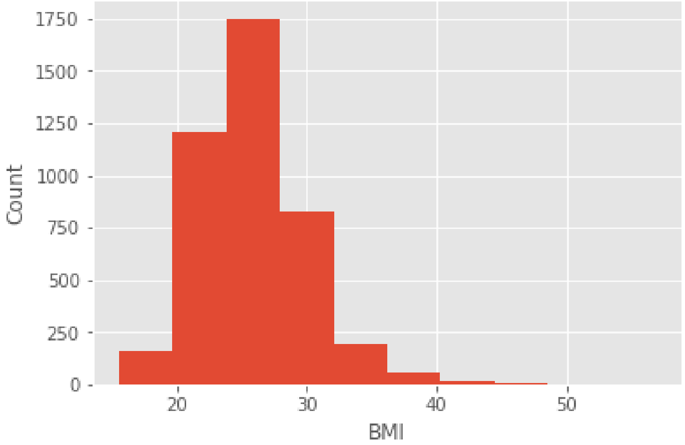
Fig 6

IV. Results and Discussions

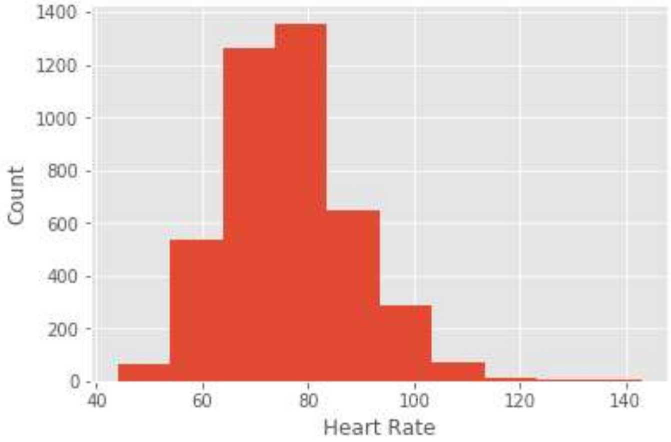
To do performance analysis, different metrics are considered to evaluate the performance of proposed technique. The overall objective of this study is to predict the presence of heart disease accurately in minimum time. The numerous heart attack prediction techniques presented in this paper. The data mining approach involves screening of medical data of heart patients to check the vulnerability. Based on the Framingham Heart Study (FHS)dataset accuracy of Logistic Regression was 85.48%, Quadratic Discriminant Analysis was 82.82% and K Nearest Neighbourswas 77.34%. Given below are the histogram plots of some of the important parameters like cholesterol, heart rate, BMI and glucose versus the count of the individuals.



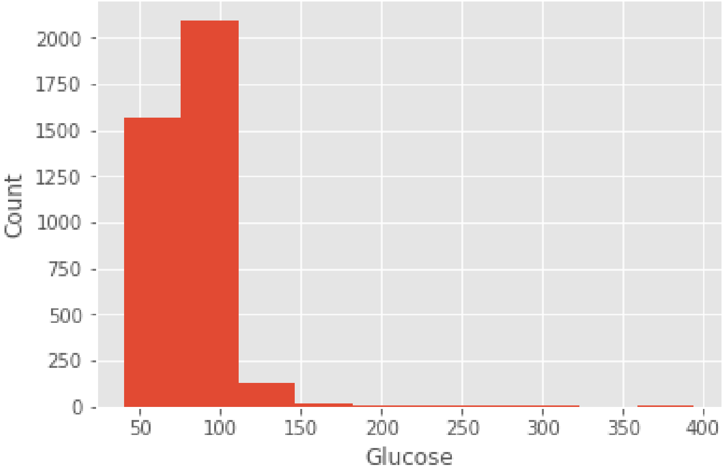
From the above histograms it is conclusive that cholesterol levels of the quite many individuals are beyond safe range (up to 100 mg/dL is the safe range).



Majority of the individuals have safe levels of BMI (up to 25 is the safe range).



The above histogram of heart rate versus count indicates that majority have heart rate within safe range i.e 60 to 100 beats per minute.



From the above histogram we can conclude that majority of the individuals surveyed have safe glucose levels i.e. 100 to 125 mg/dL.

V. Conclusion and Future Scope

Diagnosis of heart disease by using machine learning methods is one of the challenges in the health field. Various Data mining techniques and classifiers are discussed in many studies which are used for efficient heart disease prediction.

Different technologies give different precision depending on several attributes considered. These systems improve the quality of clinical evidence - based decisions and help reduce the financial and timing cost taken by patients. All the techniques discussed could prove valuable in helping to address some of the challenges associated with reducing the expected healthcare spending due to CVD by informing, engaging and empowering individuals to actively participate in modifying their most significant risk factors for CVD. Hence, understanding the usefulness of data mining for early diagnosis of various heart diseases becomes important.

Acknowledgements

We sincerely thank and laud the efforts of DrNitin Shah for providing his medical expertise and his approval of parameters required for the project.

References

[1]. V.V. Ramalingam, Ayantan Dandapath, M Karthik Raja ,Heart disease prediction using machine learning techniques: a survey, International Journal of Engineering & Technology, 7 (2.8),(2018) 684-687.

[2]. Animesh Hazra,Subrata Kumar Mandal,Amit Gupta,Arkomita Mukherjee and Asmita Mukherjee, Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review, Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 10, Number 7 (2017) pp. 2137-2159.

[3]. Asha Rajkumar and B. Sophia Reena, Diagnosis of Heart Disease Using Data mining Algorithm , Global Journal of Computer Science and Technology, Vol. 10, No. 10, pp.38 - 43, 2010

[4]. Latha Parthiban and R. Subramanian, Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm , International Journal of Biological and LifeScience, Vol. 15, pp. 157 - 160, 2007.5.1.

[5]. Chaitrali S. Dangare, Sulabha S. Apte, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques,International Journal of Computer Applications (0975 – 888)Volume 47– No.10, June 2012

[6]. R. Chitra and V. Seenivasagam, Review ofHeart disease predictionsystem using data mining and hybrid intelligent techniques, ICTACT Journal on Soft Computing, JULY 2013, Volume: 03, Issue: 04.