Comparison of Data Mining Algorithms for Heart Disease Prediction

Neal Karpe
IIIT Hyderabad
Gachibowli, Hyderabad
India. 500032
20161159
neal.karpe@students.iiit.ac.in

Gunjan Karamchandani
IIIT Hyderabad
Gachibowli, Hyderabad
India. 500032
20161086
gunjan.karamchandani@students.iiit.ac.in

Dr. Kamalakar Karlapalem
IIIT Hyderabad
Gachibowli, Hyderabad
India. 500032
kamal@iiit.ac.in

ABSTRACT
Data Mining has enormous power which hasn't been harnessed completely especially in the field of health sciences. Even the ratio of the amount of data available and the amount of knowledge being generated from it is extremely sad. This research paper aims to analyze and compare different classifiers to predict whether a person has heart disease or not. Heart diseases is the leading cause of death in today's world due to improper lifestyle and eating habits. An algorithm which predicts the heart disease with high accuracy will be a blessing for the health industry and will save a lot of lives. Comparison between two classifiers - Naive Bayes and the Decision Tree has been presented after careful implementation and decision of parameters involved in implementation.

CCS Concepts
• Applied computing→Life and Medical Sciences

Keywords
Data Mining; Heart Disease; Prediction; Data Science; Naive Bayes Classifier

1. INTRODUCTION
There is a lot of medical data available but not a lot of information has been generated from it. This is probably because people often fail to believe that there can be patterns in health sciences and believe that all the "prediction" techniques work well only in the stocks industry. However this is a myth. According to the World Health Organisation (WHO), heart diseases have been the leading cause of deaths across the world in the last 10 years. It is not that enough data isn't available, it's just the lack of interest data scientists have shown in the field of health care. Understanding what causes a particular disease and how much they contribute, we can predict whether a person has chances of having a heart disease or not. This will ensure early detection of disease and better treatment of patient and hence reduce the deaths caused by the disease. Hence, we present comparison of accuracies and implementation details of 2 prediction algorithms - Naive Bayes and Decision Tree for prediction of Heart disease.

The data for training and testing our models was obtained from the Heart Disease Data Set of UCI Machine Learning Repository. The data set has 304 instances, each having 14 attributes.

2. FEATURE COMPREHENSION
The dataset has 14 attributes each affecting the chances of having a heart disease. The values each of them take and how they affect the chances of having heart disease has been explained in the following subsections.

2.1 Age
Age attribute takes value in years and the age of the specimens lie in the range [29, 77]. Biological sciences have proven that as a person ages, so does his/her blood vessels, thus making it harder for the blood to flow through them. Also, over the years the deposits in the arteries also slow the blood flow through them.

2.2 Sex
Sex attribute takes binary value - 1 if male, 0 if female. Symptoms, causes, cures, after effects of heart diseases are different for males and females. There are several differences between the 2 genders:

- Women are more likely to suffer from heart disease in their later years than men due to the presence of estrogen in them whose levels significantly reduces after Women reach their menopause state. Hence the probability of a male having a heart attack at 40 years of age is more than a female having the same.
- Heart diseases leave the arteries and ventricles weaker and more damaged in case of women than men, thus leaving women with a greater risk of having heart diseases again.

2.3 Chest Pain Type

Chest pain can be of 4 types - typical angina, atypical angina, non angina, asymptomatic. Chest pain type play a very important role in predicting heart attacks. Anginal pains usually occur due to less blood or oxygen reaching the heart. Typical angina more common in men and atypical angina more common in women both indicate a high risk of having heart disease. While other types of chest pains may arise due to indigestion or a similar non-serious cause. Hence, people who face Anginal pains are more likely to have a heart disease.

2.4 Resting Blood Pressure

Measured in mm of Hg, blood pressure in one of the key attributes in predicting whether the person has a heart disease or not. High blood pressure causes coronary arteries serving the heart to slowly become narrowed from a buildup of fat, cholesterol and other substances that together are called plaque which interrupts the blood flow to the heart resulting in heart diseases.

2.5 Serum Cholesterol level

Measurement of high- and low-density lipoprotein cholesterol in the blood is the serum cholesterol level. Low density cholesterol can build up in a the arteries, clogging them and reducing blood flow and hence leading to heart diseases. Hence, high serum cholesterol level often indicate a high risk of heart disease.

2.6 Fasting Blood Sugar

High blood sugar can damage the blood vessels and the arteries thus leading to slower blood flow to the heart and hence increasing the chance of heart diseases. The threshold value of the blood sugar is considered to be 120 mg/dl above which a person is highly likely to have a heart disease, hence this attribute takes binary values, 1 if the level is greater than 120 mg/dl and 0 if not.

2.7 Resting electrocardiographic results

This attribute may take 3 values:

- 0: Normal
- 1: Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
- 2: Showing probable or definite left ventricular hypertrophy by Estes' criteria

2.8 Maximum heart rate achieved

Maximum heart rate plays a major role in setting one's aerobic capacity—the amount of oxygen one is able to consume. According to different studies, high aerobic capacity is associated with a lower risk of heart diseases. Hence, a lower maximum heart rate means lower risk of heart disease.

2.9 Exercise induced angina

Binary attribute which is 1 for the person for whom exercise induced angina and 0 for whom it didn't. As discussed before, Angina is lesser blood reaching the heart, hence causing pain and discomfort and is responsible for heart diseases. Hence, exercise induced angina also is a very key contributor to heart diseases.

2.10 ST depression induced by exercise relative to rest

ST depression refers to a finding on an electrocardiogram,wherein the trace in the ST segment is abnormally low below the baseline. Often a lower depression implies a higher chance of having certain types of heart diseases.

2.11 Slope of the peak exercise ST segment

This attribute takes 3 possible values: 0(upslope), 1(flat), 2(downslope) indicating the nature of slope.

2.12 Number of major vessels colored by fluoroscopy

This parameter takes a value between 0-3 and indicates the number of vessels that were colored by fluoroscopy - an imaging technique that uses X-rays to obtain real-time moving images of internal organs. This feature has some limitation though. It is believed that number of vessels have a clear impact on chances of having a heart disease however it is also well known that fluoroscopy is not very sensitive and the

number of records we have is not enough to compensate for the sensitivity bias.

2.13 Thallium stress test

A thallium stress test is an imaging test that shows how well blood flows into the heart while one is exercising or is at rest. This attribute may take 3 values: Normal (blood flow is normal through the coronary arteries), Fixed defect, Reversible defect. According to the studies, having fixed defects (indicated by attribute value 2), increases the chances of a person having cardiac arrest or other cardiac events than people who have *normal* Thallium Stress Test or have *reversible* defects.

3. HEART DISEASE PREDICTION

3.1 Using Data Mining Methodologies

In each row of the dataset, the *target* attribute tells us whether or not the patient was diagnosed with heart disease. The *target* attribute takes two values: 1 meaning presence of heart disease and 0 meaning absence of heart disease. Hence for this dataset, heart disease prediction boils down to a two-class classification problem. Based on the training samples (rows of the dataset), the aim is to build a classifier so that any future patient can predict heart disease by extracting the 13 feature values using simple medical methods. In this paper, we are implementing two different classifiers(Naive Bayes Classifier and Decision Tree) to solve this problem of heart disease prediction. The implementation was optimized wherever possible to get the best running times so that a lot of iterations of training and predicting could happen. We also used the random forest method to find out the most important attributes. The results were interesting and there was a huge difference in the minimum and maximum contribution from the attributes.

3.2 Naive Bayes

Naive Bayes is a probabilistic modelling of each class. This classifier model assumes that the features are independent of each other. On studying the biological significance of different features, we found out that they could safely assumed to be independent of each other, hence we selected the naive bayes as one of the classifiers. The aim is to compute the likelihood probability for a sample to belong to a particular class, for all given classes. For any new test sample, we calculate the posterior probability for that sample in every class, and we predict that sample belongs to the class with the highest posterior probability. Mathematically predicted class is given by the formula,

$$max\,(P(Y=0)P(X|Y=0), P(Y=1)P(X|Y=1)$$

where

$$P(X|Y=y) = \prod_{i=1}^{13} P(x_i|Y=y)$$

and the likelihood is computed as:

$$P(x_i|y) = e^{(-(X_i - m_{(y,i)})^2/2v_{(y,i)})}/\sqrt{(2\pi v_{(y,i)}{}^2)}$$

where $m_{(y,\,i)}$ and $v_{(y,\,i)}$ denote the mean and standard variation of the i^{th} feature of the training samples belonging to class y.

The advantage of Naive Bayes is that it is a 'generative model', meaning we can 'generate' and example sample from each class. Since there are no parameters involved in the naive bayes algorithm, the tough problem of selecting the parameters which give the best accuracy is eliminated. Apart from that a lot of preprocessing steps are eliminated since there's no scaling involved. In our implementation we modelled the likelihood as a Gaussian distribution for each class.

Pseudo code

```
for every test sample X
        probabilityOfClass0 = P(Y=0)*P(X|Y=0)
        probabilityOfClass1 = P(Y=1)*P(X|Y=1)
        if probabilityOfClass0>probabilityOfClass1
                X belongs to class 0
        else
                X belongs to class 1
```

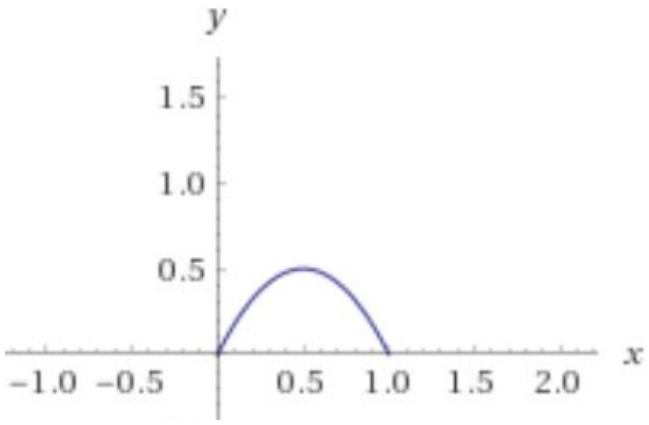
3.3 Random Forest

A random forest is an ensemble classification technique that fits a number of weak decision tree classifiers on various subsamples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The subsample size is always the same as the original input sample size but the samples are drawn with replacement. For every subsample we choose a subset of the features in order to ensure that the decision tree is weak. This also helps in finding

out the most important features because if the subset containing the most important features is used to make the decision tree, the accuracy will be the most on the test set. We used gini impurity criteria to decide on the split attribute. Gini impurity being faster to compute than entropy (involves logarithm computation) allows for more training and prediction iterations. Studies have concluded that entropy is a better criteria for deciding the split attribute for this specific problem, however since our goal is to create weak classifiers and Gini impurity being faster to calculate best suits our purpose. Gini impurity for M classes with p denoting the fraction of items labeled with class 0 in the set is given by:

$$Gini\_Impurity = 1 - p^2 - (1-p)^2 = -2p^2 + 2p$$

Figure 1. Plot showing variation of gini impurity with different values of p



We used no early stopping technique because even if the tree overfits a particular subset, the overfitting is neutralized by other sub-sampling.

Pseudo code

```
outputs = []
for i in [1, 100]
      temporaryTrainingData = subsample(trainingData)
      ithDecisionTreePredictions = DecisionTree(temporaryTrainingData)
      outputs = outputs +  ithDecisionTreePredictions

for every column in outputs
      predLabel[i] = 0 or 1 according to majority voting in that column

def DecisionTree(trainingData)
      take a small subset of features
      split on the features in decreasing order of impurity reduction until
            no more features left to split on
                  or
            no more samples remaining
                  or
            100% purity achieved
      return trainedModel
```

4. Accuracy measures

4.1 Holdout method

This idea involves randomly selecting 70% of the rows as training data and the remaining 30% as test data. In our dataset, the challenge is to divide 304 rows into 212 training samples and 92 testing samples. Since there are $\binom{304}{212}$ possible ways of doing this (which is of the order 10^{79}), we chose 10000 random shufflings of the rows and split them in the ratio 70-30.

4.2 Cross validation

This idea is to pick out 1 sample from the dataset as a test sample, and train the model on the remaining *(n-1)* samples. This experiment is repeated *n* times, as there are *n* ways to achieve this split. The number of times that the left out test sample is correctly classified upon the total number of such tests gives the average accuracy of the model.

5. RESULTS

Following table shows the minimum accuracy, maximum accuracy and the average accuracy of the algorithms measured using both the above accuracy measures.

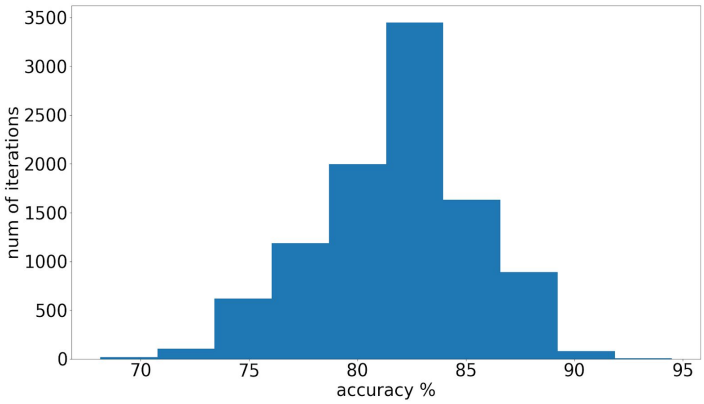
Table 1. Accuracy comparison

Algorithm	Accuracy Measure	Minimum Accuracy	Maximum Accuracy	Average Accuracy
Naive Bayes	70-30 split	68.13%	95.6%	**81.73%**
Naive Bayes	Leave out 1	0%	100%	**81.19%**
Random Forest	70-30 split	62.64%	92.31%	**79.34%**
Random Forest	Leave out 1	0%	100%	**81.52%**

From the table we can see that Naive Bayes algorithm gives an accuracy of 81.73% measured using the holdout technique. Random Forest (with 10 decision trees) on the other hand gives an accuracy of 79.32% using the same technique. The high accuracy of Naive Bayes proves that our assumption that the features are independent of each other is indeed the case. Slightly poor accuracy of Random Forest might be because we are considering only 10 decision trees, and many times a majority of the decision trees can overfit the data, leading to poor accuracy on the test samples. Another interesting result is that the average accuracy (on 70-30 train-validation) for Random Forest with 100 decision trees, goes up to **82%**. This is because even if some trees overfit, majority voting neutralizes

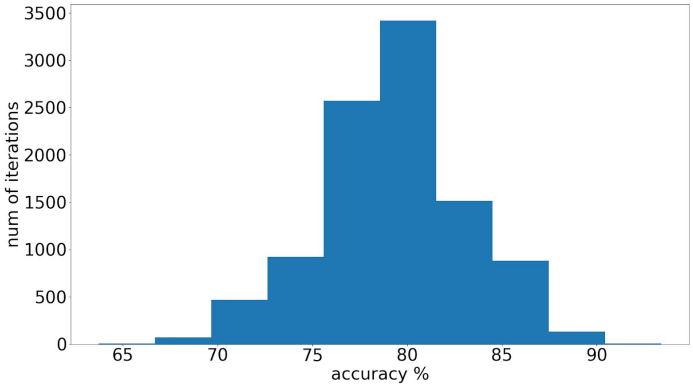
the results. This result is not very surprising as ensemble methods mostly strengthen the classification model, rather than weakening it.

Figure 2. Histogram of accuracy % of Naive Bayes in 10000 iterations of random 70-30 split testing



From the above histogram, we see that most of the iterations of Naive Bayes classification give an accuracy of 79%-85%.

Figure 3. Histogram of accuracy % of Random Forest (with 10 decision trees) in 10000 iterations of random 70-30 split testing



From the above histogram, we see that most of the iterations of Random Forest classification give an accuracy of 76%-82%.

Table 2. Running time comparison

Algorithm	Accuracy Measure	Running time
Naive Bayes	70-30 split	2 min, 53 sec
Naive Bayes	Leave out 1	0.5 sec
Random Forest	70-30 split	2 min, 1 sec
Random Forest	Leave out 1	3.5 sec

6. RANKING THE FEATURES

We know that all features cannot contribute equally to the prediction model. We use the random forest technique to find the features most relevant to the prediction model. For every feature, an *importance* measure is calculated. As previously mentioned, each node represents a condition on an attribute, and contributes to a reduction in impurity. Hence for each feature, we find the average reduction caused by it, across all nodes in all the 10000 trees. These values are scaled to proportions on a percentage scale. We ran the Random Forest Algorithm on the entire dataset and got the following results.

Figure 4. % importance of each feature

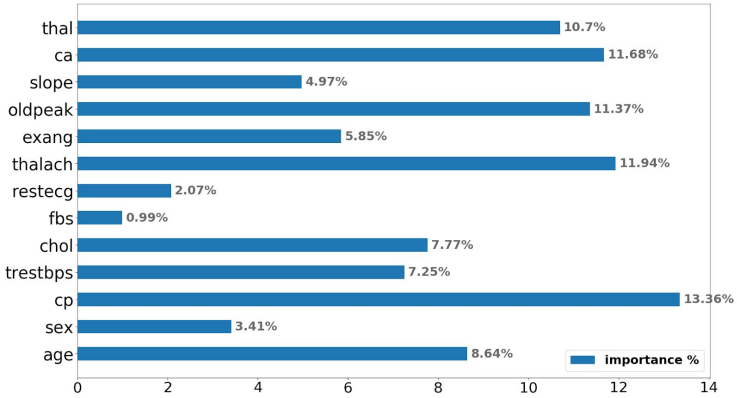


Table 3. Feature importance ranks

Rank	Feature	Importance
1.	**cp**	**13.36%**
2.	thalach	11.94%
3.	ca	11.68%
4.	oldpeak	11.37%
5.	thal	10.7%
6.	age	8.64%
7.	chol	7.77%
8.	trestbps	7.25%
9.	exang	5.85%
10.	slope	4.97%
11.	sex	3.41%

12.	restecg	2.07%
13.	**fbs**	**0.99%**

7. CHALLENGES

While the classifiers are able to predict the chances of a person having heart diseases to a decent accuracy, there are some challenges that still stop us from performing the best:

- Lack of training data: As specified above the dataset just contains the data for 303 specimens which is an extremely small percentage of our population. We know that more data implies better training of the model and hence better prediction for the test data. Hence, data is the biggest enemy in our case.
- Label problems: The labelling for instances is binary. The class label which is the diagnosis of heart disease (angiographic disease status) is recorded as follows:
    - Value 0: < 50% diameter narrowing
    - Value 1: > 50% diameter narrowing

  Discretization of the heart disease severity has several problems. Firstly, the patients that are around the 50% border have to always be classified into one of the extremes. Secondly, a person with 90% diameter narrowing and 51% diameter narrowing will be predicted into the same class, even though their conditions are clearly different.

- Feature problems: Similar to the issue with labelling above, even some features like fasting blood sugar are binary. As described above it takes value one when sugar level >120 mg/dl and 0 otherwise. This means there's no distinction between a person having 121 mg/dl blood sugar level and 200 mg/dl blood sugar level which is clearly misleading and false.

8. REFERENCES

[1] Uci Machine Learning Repository: Heart Disease Data Set: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[2] Gender Differences in Coronary Heart Disease: Ramzi Khamis-Tareq Ammari-Ghada Mikhail - https://heart.bmj.com/content/102/14/1142

[3] How Age and Gender Affect Your Heart: https://wa.kaiserpermanente.org/healthAndWellness/index.jhtml?item=%2Fcommon%2FhealthAndWellness%2Fconditions%2FheartDisease%2FageAndGender.html

[4] Gender Matters: Heart Disease Risk in Women - Harvard Health Harvard Health Publishing - https://www.health.harvard.edu/heart-health/gender-matters-heart-disease-risk-in-women

[5] Gender and Heart Disease - Go Red For Women American Heart Association - https://www.goredforwomen.org/know-your-risk/find-out-your-risk/gender-heart-disease/

[6] Typical and Atypical Angina: What To Look For: https://www.harringtonhospital.org/typical-and-atypical-angina-what-to-look-for/

[7] How High Blood Pressure Can Lead To a Heart Attack: http://www.heart.org/en/health-topics/high-blood-pressure/health-threats-from-high-blood-pressure/how-high-blood-pressure-can-lead-to-a-heart-attac

[8] Hamid Marateb-Sobhan Goudarzi - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4468223/

[9] DataScience.com, Oracle +. "Introduction to Random Forests." *Oracle + DataScience.com*, www.datascience.com/resources/notebooks/random-forest-intro.

[10] Shouman, Turner, Stocker (2011), Using Decision Tree for Diagnosing Heart Disease Patients, Proceedings of the 9-th Australasian Data Mining Conference (AusDM'11), Ballarat, Australia.

[11] K.Anderson, P.Odell, P.Wilson and W.Kannel, "Cardiovascular disease risk profiles,"American Heart Journal, vol. 121, no. 1, 1991

[12] A. Methaila, P. Kansal, H. Arya, and P. Kumar, "Early heart disease prediction using data mining techniques," Computer Science & Information Technology Journal, 2014

[13] V.Chaurasia and S.Pal, "Early Heart Disease Prediction Using Data Mining Techniques,"

Caribbean Journal of Science and Technology, Vol. 1, 208-217, 2013.

[14] 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)

[15] Franck Le Duff,Cristian Munteanb,Marc Cuggiaa,Philippe Mabob,"Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method", Studies in health technology and informatics, Vol. 107, No. Pt 2, pp. 1256-9, 2004.

[16] R.Bhuvaneshwari and K.Kalaiselvi,"Naïve Bayesian Classification approach In Healthcare Applications",International Journal of Computer Science and Telecommunication ",vol 3,no 1,pp.106-112,2012.