# Prediction of Heart Disease Using Machine Learning Algorithms- Naïve Bayes, Introduction of Probabilistic Analysis and Classification (PAC) Algorithm, Comparison of Algorithms and Building a Centralized System deployed on Cloud Platform- HDPS

Prerana T H M
Software Engineer,
Samsung R & D Institute,
Bangalore, India
preranathm@gmail.com
Ph: 91-9880143938

N C Shivaprakash
Principal Research Scientist,
Indian Institute of Science,
Bangalore, India
shivaprakashiisc@gmail.com
Ph: 91-9449086370

N Swetha
Computer Science Engineering,
Vivekananda Institute of Technology,
Bangalore, India
swethaiisc@gmail.com
Ph: 91-9538023428

*Abstract*— **Diagnosis of the heart disease is a difficult task in real world. Prediction of heart disease is a major challenge faced by hospitals and medical centers, especially when it comes to accuracy. Classification techniques of Data Mining and Machine Learning Algorithms play a significant role in prediction and data exploration. The healthcare industry collects huge amounts of healthcare data which is not feasible to handle manually. With the tremendously growing population, the doctors and experts available are not in proportion with the population. Also, symptoms of heart disease may not be significant and thus are often neglected. The main objective of this research project is predicting the heart disease risk level of a patient using machine learning algorithms, Introduction of PAC, Comparison of PAC with existing ML Algorithms and creating a Centralized System for both doctors and patients to login and view the data on Cloud. Big Data like hospital records of patients is handled using Hadoop Map Reduce programming. Comparative study of the machine learning algorithms is done and graphical representation of the results is provided for easier understanding. This application i.e HDPS is made globally accessible by deploying it on Cloud Platform and can be accessed through any browser in any part of the world. This project can be extended for prediction risk level of other diseases like Cancer, Brain tumor etc.**

***Keywords-NaïveBayes, PAC-Probabilistic Analysis and Classificatio, Hadoop, BigData, MapReduce Programming, HDPS-Heart Disease Prediction System, PaI.***

## I. INTRODUCTION

In present world there are many scientific technologies which help doctors in taking clinical decisions but they might not be accurate. Heart disease prediction system can assist medical professionals in predicting state of heart, based on the clinical data of patients fed into the system. Doctors may sometimes fail to take accurate decisions while diagnosing the heart disease of a patient, therefore heart disease prediction systems which use machine learning algorithms assist in such cases to get accurate results. There are many tools available which use prediction algorithms but they have some flaws. Most of the tools cannot handle big data and most are not centralized, not deployed on cloud and hence not accessible on the internet. There are many hospitals and healthcare industries which collect huge amounts of patient data which becomes difficult to handle with currently existing systems. In this paper we are predicting the risk levels of patients for a huge data set and deploying the application on cloud platform where doctors and patients can login with a unique ID created by them. Doctors can upload the patient reports and patients on the other hand can view the reports on their laptops or personal computers. A collection of patient reports will be maintained online in the cloud.

Data mining techniques are used for preprocessing, machine learning algorithms are used for implementation and cloud computing is used for deployment. Popular machine learning algorithms have been implemented to determine the heart disease risk level and to help the doctors correctly predict the same. Finally a comparison between the algorithms is done which helps the user to determine which algorithm shows the highest accuracy. The interface is user friendly and the application is globally accessible on cloud.

This paper is divided into five phases first phase gives the theoretical background for reducing attributes from a data set, second phase gives the implementation of machine learning algorithms like Naïve Bayes and Introduction of PAC Algorithm for predicting heart disease risk level of a patient, third phase deals with processing Big Data using Hadoop Map Reduce programming, fourth phase talks about the centralized system deployed on cloud platform and fifth phase gives conclusion and future scope of the project.

## II. REDUCTION OF ATTRIBUTES USING DATA MINING TECHNIQUES

Data Mining is a process of extracting useful and important knowledge from huge data set. Data Preprocessing is a important process in Data Mining and Machine learning. Dimensionality reduction is an effective method for downsizing data.The important techniques for Dimensionality Reduction are Feature Selection and Feature Extraction.

Feature selection is the process of selecting a subset of relevant features. Feature selection techniques are a subset of the more general field of feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature Subset selection is one of the method for Feature selection. In subset selection we find the best subset of the set of features. The best subset contains the least number of dimensions that most contribute to accuracy. We discard the remaining, unimportant dimensions.

There are two approaches in subset selection
a. Forward Selection
b. Backward Selection

Forward Selection starts with no variables and we add them one by one, at each step adding the one that decreases the error the most, until any further addition does not decrease the error.
In backward selection, we start with all variables and remove them one by one, at each step removing the one that decreases the error the most, until any further removal increases the error significantly. Let us denote by F, a feature set of input dimensions, xi, i = 1,...,d. E(F) denotes the error incurred on the validation sample when only the inputs in F are used. In sequential backward selection, we start with F containing all features and we remove one attribute at a time from F, and we remove the one that causes the least error. $j = \text{argmin}_i E(F - x_i)$ and we remove $x_j$ from F if $E(F - x_j) < E(F)$. We stop if further removal does not decrease the error.

On theoretical basis Backward Selection is more favourable in our case, since we are considering only 13 out of 76 heart parameters.

### DATA SET USED

The dataset under consideration has been taken from University of California Irvin (UCI).13 attributes are involved in prediction of heart disease. These thirteen attributes have been shown in Table 1. This data set is fed into the classification model i.e. Naïve Bayes Classification and Probabilistic Analysis and Classification. The Big data file containing patient's record is given as the input and the result is processed.

**Predictable attributes :**

Value=0 ( safe level)

Value=1 (low risk level)

Value=2(medium risk level)

Value=3 (high risk level)

Value=4(Very high risk level)

**Input Attributes :**

1.Age in Year
2. Sex (value 1: Male; value 0: Female)
3. Chest Pain Type (value 1:typical type 1 angina, value 2: typical type angina, value
3:non-angina pain; value 4: asymptomatic)
4. Fasting Blood Sugar (value 1: >120 mg/dl; value 0: <120 mg/dl)
5. Restecg – resting electrographic results (value 0:normal; value 1: having ST-T
wave abnormality; value 2: showing probable or definite left ventricular
6. Exang - exercise induced angina (value 1: yes; value 0: no)
7. Slope – the slope of the peak exercise ST segment (value 1:unsloping; value 2:
flat; value 3: down sloping)
8. CA – number of major vessels colored by fluoroscopy (value 0-3)
9. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
10. Trest Blood Pressure (mm Hg on admission to the hospital)
11. Serum Cholesterol (mg/dl)
12. Thalach – maximum heart rate achieved
13. Oldpeak – ST depression induced by exercise

TABLE 1- INPUT ATTRIBUTES.

## III. MACHINE LEARNING ALGORITHMS:

### A. Naive Bayes Classifier

Naïve Bayes classifier is used as the first method for prediction of heart disease. Here we use the preprocessed 13 attributes as input. With Naïve Bayes' assumption all attributes are independent of each other, this significantly reduces the calculations shown later. The Naïve Bayes formula is given by

$$\underset{\text{Posterior Probability}}{P(c \mid x)} = \frac{\overset{\text{Likelihood}}{P(x \mid c)}\,\overset{\text{Class Prior Probability}}{P(c)}}{\underset{\text{Predictor Prior Probability}}{P(x)}}$$

- fp(c|x) is the posterior probability of class (target) given predictor (attribute).
- P(c) is the prior probability of class, also called prior. It is the probability of observing a class in general.
- P (x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor also called evidence.

The term evidence is constant for all the class values hence the posterior probability of a class is proportional to product of likelihood and prior value for that class only. Here x is a tuple of 13 reduced attributes and can be expressed as (x1, x2, x3…, x13).Using Naïve Bayes' assumption likelihood can be split into continuous product of class conditional probabilities of 13 attributes as shown below.

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Hence given as input a patient record of 13 attributes we can calculate posterior probability for all risk levels. Patient has the risk level for which the posterior probability is maximum. Training data set is used for calculation of class conditional probabilites. Given an attribute xi we can calculate P(xi|Cj) for class Cj. For this we can use basic definition of probability that is.

$$\mathrm{P}(\mathrm{x}_i \mid \mathrm{Cj}) = \frac{\text{Number of times } \mathrm{x}_i \text{ occurs in rows of training data set } \mathrm{X}^t \text{ for class Cj}}{\text{Number of times Cj occurs in training data set } \mathrm{X}^t}$$

xi € X and j=0,1,2,3,4. Hence for the calculation of likelihood entire training dataset is used. However this method of calculation holds good if and only if variables are discrete in nature like sex, chest pain type etc. for patient record. In this dataset exactly 5 attributes i.e., age, cholesterol, resting blood pressure, thalach and oldpeak are continuous. Hence the initial approach is calculation of class conditional densities using probability density function assuming Normal distribution for all the continuous variables as shown

$$p(x = v \mid c) = \frac{1}{\sqrt{2\pi\sigma_c^2}}\, e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

- Here $\sigma_c^2$ is the variance for variable x given class C.
- µc is the mean for variable x given class C.

Using normal distribution for age, cholesterol and thalach is an approximately good assumption. However resting blood pressure and old peak do not fit into this distribution and this results in over fitting of model.As a result,we get partially accurate results leading to low accuracy.To avoid dealing with distribution of variables, we can use another approach, i.e.assuming these variables to be discrete.In this case,calculation of class conditional probabilities for these variables is done in the same way as done for other discrete variables. This assumption holds good in this case since the dataset is large and also leads to improved results and high accuracy.

### B. Probabilistic Analysis and Classification (PAC)

Probabilistic Analysis and Classification is a supervised machine learning algorithm derived from Naïve Bayes Algorithm. It uses the concept of weighted average probability calculation over the entire training data set {Xt} It is built over Naïve Bayes model to overcome the shortcomings of Naïve Bayes algorithm. One advantage is complete reduction of continuous variables to discrete variables using discretization technique, hence the hard work to find suitable distribution for continuous variables is not required which else was resulting in "over fitting the model". Another advantage is due to complete conversion of continuous variable to discrete variable, Laplacian smoothing used in Naïve Bayes Classification is not required, which in turn reduces unnecessary comparisons and instructions.

The main idea of algorithm is to use weighted average calculation for all heart disease attributes until unless we find an exact same tuple in the training data set, in this case the risk level of tuple is assigned the risk level of the input patient record. However this doesn't happen often and so we have to use weighted average calculation for the entire training data set and calculate the contribution of each and every attribute for that particular risk level and what the different contributions for entire training data set are. For considering the entire data set we have used number of supporting tuples for various risk levels in the training data set. This is similar to concept of "Prior" in Naïve Bayes algorithm but in Naïve Bayes algorithm where prior probabilities give more weight to risk levels on the basis of their own values. In PAC it substantially reduces this weight, resulting in error, due to difference in percentage increase in numerator and denominator in the term $\alpha i$. So to overcome this drawback, we multiply by normalizing factor to reduce this error and give appropriate results. Finally the maximum term $\mu i$ among all risk levels is returned as the risk level for the patient.

Other variances from Naïve Bayes implementation are feeding the training data and Big Data files, which have to be parsed in pre-initial step to convert continuous variables to discrete variables. Another step is conversion of continuous variables supplied by user to discrete form so that algorithm can read and process them.

**PAC algorithm:**

```
PAC (input.csv){
 fp = input.csv
fw = training.csv
fq = make_discrete(fp)
while fq!=EOF
for each line in fq
for each line in fw
$\alpha i = \sum 1$ (for each matching attribute)
 13 Where i= different risk levels
End For
$\beta i = \alpha i/SPi$
Where SPi is the number of supporting cases.
For each risk level
 End Forf
 $\mu i$= Normalizing Factor j x $\beta i$
 r=maximum ($\mu i$)
 Output r as the risk level
End While
}
maximize ($\mu i$){
Return index for which $\mu i$ is maximum for all i=0, 1,2,3,4
}
made_discrete(fq){
Assign continuous variables discrete values Vi by splitting into equal intervals with varying ranges.
Return the dataset which has maximum accuracy
 }
```

## IV. HADOOP MAP REDUCE PROGRAMMING FOR PROCESSING BIG DATA

In this paper we have successfully designed an algorithm for accurate prediction of heart disease risk level. PAC algorithm is built using existing machine learning algorithms. It covers up the disadvantages of the existing algorithms and in turn increases accuracy of prediction of disease risk level. Many hospitals and health care industries have huge amounts of patient data. With the tremendously growing population, the doctors and experts available are not in proportion with the population. Doctors may sometime fail to correctly diagnose the severity of the disease. Hadoop single node cluster is used to process Big Data. Map Reduce code is implemented for the designed algorithms. Figure 1.1 shows the flow diagram for implementation of Map Reduce Programming for the uploaded data set.

- **Mapper**: Inside Mapper function each line from input file is taken as input to map phase and is fed to different map-tasks in parallel, considering multi-node cluster each node follows the same procedure in parallel. If there are N lines in input file and we have default M map tasks then number of lines processed by each map task is N/M.

  The mapper function executes our algorithm on each and every map task of node and hence on each and every node in a multi-node cluster. Every time it takes single line from Big Data as input and processes machine learning algorithm to calculate risk level. Here the line number is taken as key and entire line is taken as value. The risk level is supplied as key to reducer and value is assigned to whatever attribute we wish to evaluate with. The context file is the intermediate output given by mapper function as input to reducer function.

- **Reducer:** The reducer shuffles the risk level provided by context file and sorts them according to key values provided to reducer function in ascending order and stores the sorted output in a file. The map-reduce jobs are used to process Big Data in both the algorithms. Different Map-Reduce functions are implemented to calculate the graphs for different attributes versus number of people with and without disease. This can be employed for various population surveys.

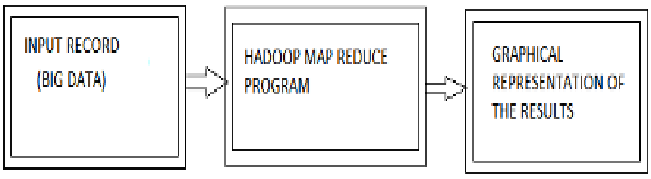

INPUT RECORD
(BIG DATA)
HADOOP MAP REDUCE PROGRAM
GRAPHICAL REPRESENTATION OF THE RESULTS

Figure 1.1 Flow diagram of Big Data processing

### Graphical Analysis Results

The output of the project can be either a report of a single patient for form based input or graphical output if Big Data file is provided as input. A comparative study of machine 8learning algorithms explained above is made and an accuracy graph is plotted to determine the best algorithm for disease prediction. This includes multiple aspects of the study such as the total number of patients who have and do not have heart disease, number of patients of a particular age who have and do not have disease etc. all these aspects are shown in graphical format so that it is easier for the user to understand. Figure 1.6 shows the comparative study of the Machine Learning algorithms as explained in the paper Naïve Bayes for continuous variables (red), Naïve Bayes for discrete variables (blue) and PAC Algorithm (green).

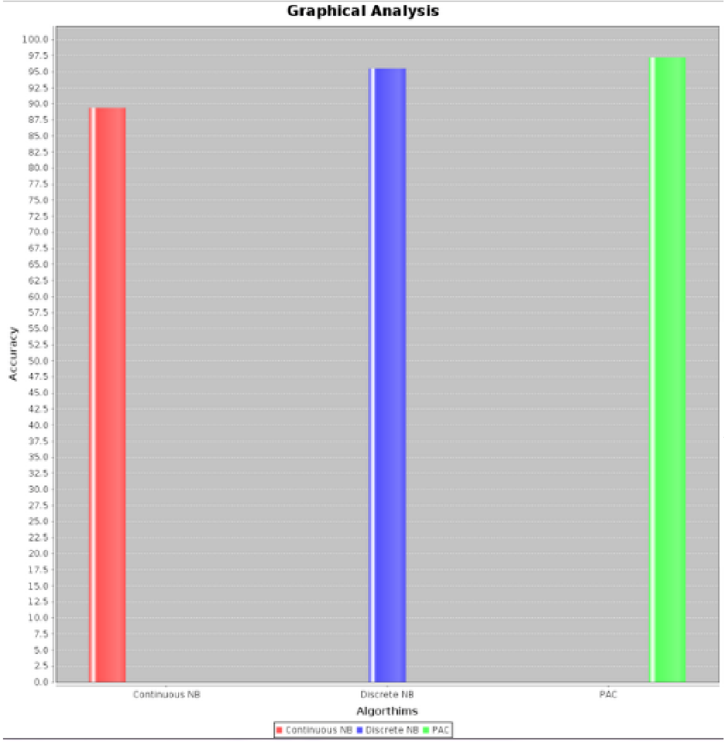


Figure 1.1 Comparison of Machine learning Algorithms

| **Machine Learning Algorithms** | **Accuracy** |
|---|---|
| NAÏVE BAYES CONTINUOUS VARIABLE | 89.80% |
| NAÏVE BAYES DISCRETE VARIABLE | 95.21% |
| PROBABILISTIC ANALYSIS & CLASSIFICATION | 97.48% |

TABLE 2: COMPARISON TABLE

## V. BUILDING A CENTRALIZED SYSTEM AND DEPLOYMENT ON CLOUD PLATFORM- HDPS

Cloud computing involves distributed computing over a network, where a program or application may run on many connected computers at the same time. It specifically refers to a computing hardware machine or group of computing hardware machines commonly referred as a server connected through a communication network such as the Internet, a local area network (LAN) or wide area network (WAN). Any individual user who has permission to access the server can use the server's processing power to run an application, store data, or perform any other computing task. This project is deployed on Cloud Platform called Jelastic.

Jelastic is a Platform-as-Infrastructure(PAI) cloud computing service that provides networks, servers, and storage solutions to software development clients, enterprise businesses, OEMs and web hosting providers. The company has developed technologies for moving Java and PHP based on applications onto the cloud.

It has international hosting partners and data centers. The company can add memory, CPU and disk space to meet customer needs. The main competitors of Jelastic are Google App Engine, Amazon Elastic Beanstalk, Heroku, and Cloud Foundry. Jelastic is unique in that it does not have limitations or code change requirements, and it offers automated vertical scaling, application lifecycle management and availability from multiple hosting providers around the world.

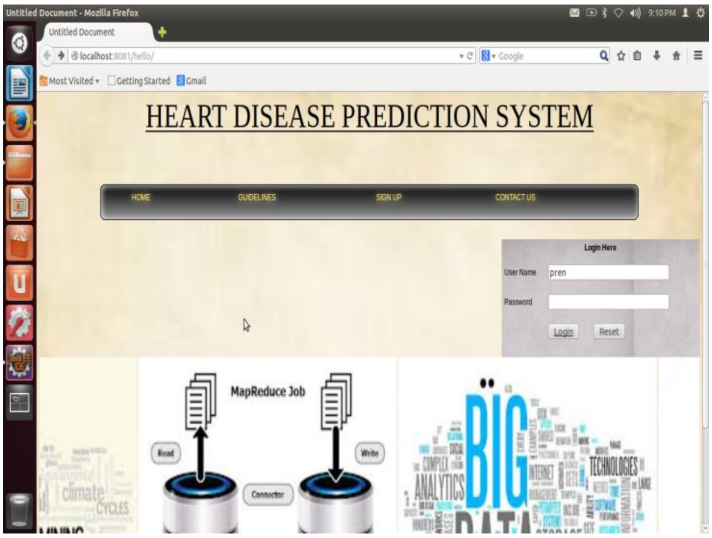
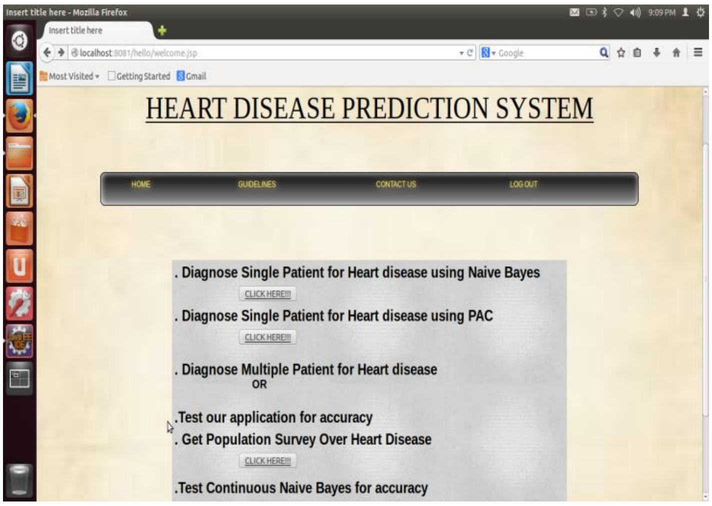


Figure 1.2 Login screen of HDPS



. Diagnose Single Patient for Heart disease using Naive Bayes

. Diagnose Single Patient for Heart disease using PAC

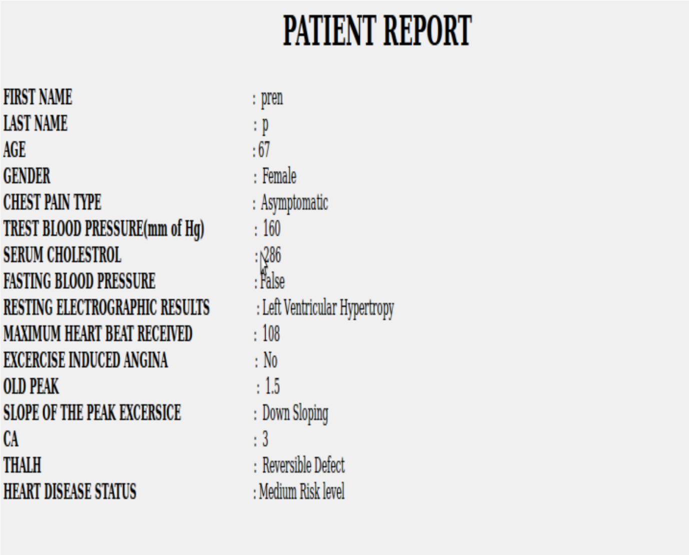
. Diagnose Multiple Patient for Heart disease
OR

.Test our application for accuracy

. Get Population Survey Over Heart Disease

.Test Continuous Naive Bayes for accuracy

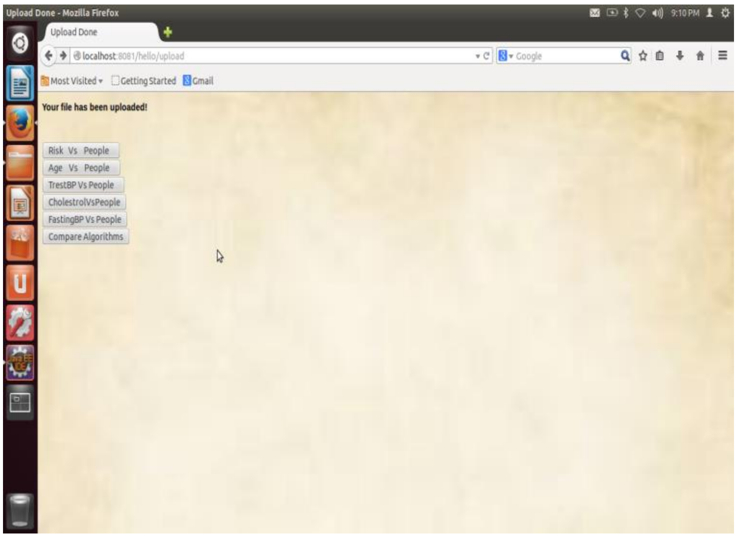
Figure 1.3 various options to choose for Heart Disease Risk Level Prediction



PATIENT REPORT

| | |
|---|---|
| FIRST NAME | : pren |
| LAST NAME | : p |
| AGE | : 67 |
| GENDER | : Female |
| CHEST PAIN TYPE | : Asymptomatic |
| TREST BLOOD PRESSURE(mm of Hg) | : 160 |
| SERUM CHOLESTROL | : 286 |
| FASTING BLOOD PRESSURE | : False |
| RESTING ELECTROGRAPHIC RESULTS | : Left Ventricular Hypertropy |
| MAXIMUM HEART BEAT RECEIVED | : 108 |
| EXCERCISE INDUCED ANGINA | : No |
| OLD PEAK | : 1.5 |
| SLOPE OF THE PEAK EXCERSICE | : Down Sloping |
| CA | : 3 |
| THALH | : Reversible Defect |
| HEART DISEASE STATUS | : Medium Risk level |

Figure 1.4 Single Patient Report of Heart Disease Risk Level after uploading Patient Data



Your file has been uploaded!

Figure 1.5 Results Page after uploading a Big Data File

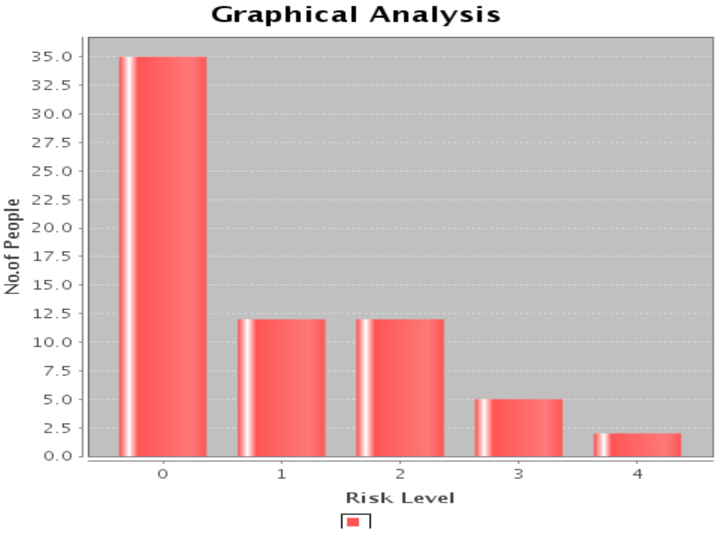


Figure 1.6 Results- Risk Level Vs Number of People

## VI. CONCLUSION & FUTURE SCOPE

Health care related data are huge in nature and they arrive from various birthplaces which are not suitable in structure or quality. These days, the utilization of knowledge and experience of copious specialists and medical screening data of patients collected in a database during the diagnosis process, has been widely accepted.

Using Hadoop node cluster for processing big data is one of the recent technologies.

Implementing accurate machine learning algorithms to determine the heart disease risk and comparison of algorithms is done to determine the accuracy using graphs. It is easier to understand the graphs and the user can also determine his own risk level and get the report for the same. The project is globally accessible using cloud service and Big Data can be easily processed.

The project can be used in hospitals and research centers to analyze the heart disease attributes and how they contribute to the disease and its prediction. Multiple efficient machine learning algorithms can be included in this to more accurately predict the heart disease and compare it with other algorithms.

Statistical study can be done precisely by focusing on important attribute for medical study of heart disease over big population set. The project for prediction of heart disease can be extended for prediction of similar disease by collecting the medical data from hospitals and medical centers and create a similar diagnosis web forum as done in heart disease. Depending on the increasing requirement multi nodes can be added to the cluster to decrease the execution time and process more data.

## VII REFERENCES

1. http://www.rohitmenon.com/index.php/introducing-mapreduce-part-i/
2. http://www.javacodegeeks.com/2013/08/writing-a-hadoop-mapreduce-task-in-java.html
3. http://developer.yahoo.com/hadoop/tutorial/module4.html
4. http://nxhoaf.wordpress.com/2013/01/04/hadoop-mapreduce-word-count-using-eclipse/
5. http://bigdatacircus.com/2012/09/09/hadoop-map-reduce-introduction-and-internal-data-flow/
6. https://www.harding.edu/fmccown/r/#barcharts
7. http://stackoverflow.com/questions/19510656/how-to-upload-files-on-server-folder-using-js
8. http://www.uniweimar.de/medien/webis/teaching/lecturenotes/machine-learning/unit-en-decision-trees-algorithms.pdf
9. Prediction System for heart disease using Naïve Bayes *Shadab Adam Pattekari and Asma Parveen Department of Computer Science and Engineering Khaja Banda Nawaz College of Engineering.
10. Data mining in Cloud Computing Ruxandra-Ştefania PETRE Bucharest Academy of Economic Studies ruxandra_stefania.petre@yahoo.com
11. Review of Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques R.Chitra1and V.Seenivasagam2 1Department of Computer Science and Engineering, Noorul Islam Centre for Higher Education, India.