IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-2, Issue-7 E-ISSN: 2347-2693

# Predicting Heart Attack Using NBC, k-NN and ID3

S. A. Angadi[1] and Mouna M. Naravani[2*]

[1]*Professor, Center for P. G. Studies, VTU, Belgaum, India*
[2]*Center for P. G. Studies, VTU, Belgaum, India*
saangadi@vtu.ac.in, mouna139@gmail.com

**www.ijcaonline.org**

Received: 11 /06/ 2014 Revised: 24 /06/ 2014 Accepted: 12 /07/ 2014 Published: 30 /07/ 2014

***Abstract*—** We are living in a world full of data. Every day people encounter large amounts of data. Main problem here is dealing with this huge data. Data mining techniques can be used to handle such huge data. Health care environment collects vast amounts of data, but the unfortunate thing is that it is not efficient in extracting useful information from this wealthy data. Data mining techniques can be applied to extract valuable knowledge from the health care environment. In this paper, three supervised learning classification algorithms have been implemented to predict heart attack risk from heart disease database. The classification algorithms used are Naive Bayesian Classification (NBC), k-Nearest Neighbor (k-NN) Classification and ID3 Classification. As a pre-processing step Discretization of continuous variables is adopted. The heart disease data set is trained with these classifiers. A GUI is designed so that the user can input patient's record. The system is then able to predict whether or not the user has a risk of heart attack. The performance of these three algorithms is determined by computing accuracy. From the experiments, it is found that ID3 Classification outperforms other two classifiers with the accuracy of 91.72%.

***Keywords*—** Classification, ID3, Data mining, Supervised Learning, Naive Bayesian, k-Nearest Neighbor

## I. INTRODUCTION

In recent years, Data Mining has attracted the information technology (IT) industry. This is due to the wide availability of huge amounts of data. Therefore, there is a need for converting such huge data into useful information. Data Mining is one such area which can contribute in getting valuable or useful information from wide availability of data. Data Mining is the exploration of large datasets to extract hidden and previously unknown patterns, relationships and knowledge that are difficult to detect with traditional statistics [1]. In other words, Data Mining refers to extracting or mining knowledge from large amounts of data [2]. Thus, Data Mining is a powerful technique which has gained more attention with its great potential to help IT companies to concentrate on the information available in their data warehouses.

The health care system consists of huge amount of data. But there are no proper techniques for mining such huge data. That is, health care environment is "information rich", but "poor" in extracting useful information from the wealthy set of data. This is because there is lack of analysis tools that can efficiently mine the medical data [3]. The medical diagnosis is very complicated task and thus the results obtained after medical diagnosis must be accurate and efficient. Hospitals accumulate an enormous amount of patient data, which can be used for analysis of various diseases. Most of the medical decisions are made by doctors, which are dependent on doctor's experience and number of tests conducted for identifying a particular disease. Automating such medical systems using Data Mining techniques would definitely prove to be advantageous.

One such area in Medical Data Mining is to predict heart attack risk from the heart disease database. Heart disease or cardiovascular disease is the class of diseases that involve the heart or blood vessels. A sudden blockage of a coronary artery, generally due to a blood clot results in a heart attack [4]. Today most countries are facing increasing rates of heart disease and it has become a leading cause for death worldwide. According to [5], India is set to be the "heart disease capital of the world" in few years, with over 3 million deaths owing to cardiovascular diseases every year.

This paper is intended to predict the heart attack risk using three different supervised learning classification algorithms. The classification algorithms implemented are Naive Bayesian Classification, k-Nearest Neighbor Classification and ID3 Classification. Firstly, a heart disease data from UCI repository [6] is collected. Necessary measures are taken so that the data set does not contain any missing and duplicate values. Finally, Naive Bayesian Classification, k-Nearest Neighbor Classification and ID3 classification algorithms are used to train the classifiers. A GUI has been developed which accepts the patient's data as test cases and predicts the presence of heart attack risk of that patient.

The remaining sections of the paper are as follows: Section 2 explains the work carried out on predicting heart disease in data mining. Data mining classification techniques is presented in Section 3. Data set collection process is depicted in Section 4 as Collecting the Data Sets. System

 

Architecture is presented in Section 5. Section 6 consists of Implementation Details. Results are presented in Section 7.

## II. LITERATURE SURVEY

Intelligent and effective heart attack prediction methods were developed in [7] using data mining techniques. It discussed briefly about classification based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network to massive volumes of healthcare data. One Dependency Augmented Naïve Bayes classifier (ODANB) and naive credal classifier 2 (NCC2) were used for data preprocessing and effective decision making. Supervised Discretization algorithm of Fayyad and Irani (1993) was used to discretize all numerical features. ODANB was compared with other improved Naïve Bayes methods and Naïve Bayes itself. Results showed that ODANB performed better than other methods for the disease prediction not related to heart attack. But for prediction of heart disease it was observed that Naïve Bayes had better results. The objective of [8] was to predict the diagnosis of heart disease with reduced number of attributes. Classification algorithms were implemented like Naive Bayes, Classification by Clustering and Decision Tree. Out of fourteen attributes only six attributes were used for diagnoses which were reduced using Genetic algorithm. The experiments were conducted using weka tool. All attributes were made categorical and inconsistencies were resolved. Results showed that Decision tree technique outperformed other two techniques with high model construction time. Naïve Bayes performed consistently and Classification via clustering performed poor compared to other two methods. Distance based classification techniques were implemented in [9] for the diagnoses of heart disease, one is K-Nearest Neighbor (KNN) classification and another is integrating voting with KNN. The distance measure used for KNN was Euclidean distance. Voting technique used here was an aggregation technique which combines decisions of multiple classifiers. The training data was divided into small subsets. These subsets were then used to build the classifier. In this case each classifier gives the predicted results. The final prediction depends on the class which gets majority votes. The final decision was selected by summing up all votes and by choosing the class with the highest aggregate. The number of voting divisions used in this paper ranged between three and eleven subsets. The accuracy achieved for KNN without voting was between 94% and 97.4 % for different values of K. The highest accuracy achieved was 97.4% and specificity was 99% for K=7. The results showed that KNN with voting did not enhance the accuracy of the system in diagnosis of heart disease. In [10] authors have developed data mining algorithms for predicting survival of Coronary Heart Disease (CHD). Authors have carried out a clinical observation and a 6-month follow up to include 1000 CHD cases. 502 cases were employed to develop the prediction models using three popular data mining algorithms namely, Support Vector Machine (SVM), Artificial Neural Networks and Decision Trees. The analysis showed that the best predictor was SVM with 92.1% accuracy, second was the artificial neural network with 91% accuracy and the decision tree came out to be the worst with accuracy of 89.6%. The results were achieved using average value of 10 fold cross-validation for each algorithm.

## III. CLASSIFICATION TECHNIQUES

Classification is a data mining technique, which classifies data into predefined groups or classes. It is a supervised learning technique where a class label of a training data is known in advance. It contrasts with unsupervised learning technique (also known as Clustering), where a class labels of a training data is not known in advance [2]. Algorithms implemented in this paper are Naive Bayesian Classification, K Nearest Neighbor (k-NN) Classification and ID3 Classification. These are supervised learning methods.

**Naive Bayesian Classification** is a supervised learning technique and a statistical method for classification. Bayesian Classifiers are able to predict class membership probabilities such as the probability that a given tuple belongs to a particular class. Naive Bayes Classifier assumes that given a class label of the tuple, its attributes are not dependent, i.e, attribute values are conditionally independent of one another. This assumption is called as "*Class Conditional Independence*". Naive Bayesian Classification algorithm is based on Bayes Theorem (1). Bayes Theorem is named after Thomas Bayes, a nonconformist English clergyman who did early work in probability and decision theory during the 18th century [2]. Let X be a data tuple and H be some hypothesis. Bayes Theorem is:

$$P(H|X) = P(X|H)\cdot P(H) / P(X) \quad (1)$$

Bayes Classifier combines prior knowledge with observed data. Given a training data, it assigns a posterior probability to a class based on its prior probability and its likelihood. Naive Bayes Classifier computes "*maximum posterior hypothesis*". It assumes class conditional independence among attributes and assigns the class to new instance that has highest maximum posterior hypothesis [2].

A **decision tree algorithm**, ID3 (Iterative Dichotomiser) was developed by J. Ross Quinlan during the late 1970s and early 1980s. Decision trees are constructed in a top-down recursive divide-and-conquer manner. The training set is recursively partitioned into smaller subsets as the tree is being built. Given a tuple, X, for which the associated class label is unknown, the attribute values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node, which holds the class prediction for that tuple [2]. The ID3 algorithm uses the information gain measure to select among the candidate attributes at each step while growing the tree [12]. Information gain (or simply gain) is the expected reduction in entropy. In ID3, an attribute with highest information gain is chosen as a splitting attribute. A node is created and is labeled with this attribute. Branches are created for each value of attribute and the samples are

partitioned accordingly [2]. Information Gain (2), Gain (S, A) of an attribute A, relative to a collection of example S, is computed as [12]:

$$\text{Gain(S,A)} = \text{Entropy(S)} - \sum_{v \in \text{Values(A)}} \frac{|Sv|}{|S|} \text{Entropy}(Sv) \quad (2)$$

Entropy or expected information (3) is a measure of the impurity in a collection of training sets. The smaller the entropy value is, the greater the purity of subset partition. It is computed as follows [12]:

$$\text{Entropy(S)} = \sum_{i=1}^{c} Pi \ \log_2 Pi \quad (3)$$

Where $P_i$ is the proportion of S belonging to class i.

**K-Nearest Neighbor (k-NN)** Classification is a lazy learner. It simply stores a training tuple and waits until it is given a test tuple. Once the test tuple is given, it classifies the tuple based on its similarity to the stored training tuples. The k-Nearest Neighbor was first described in early 1950s [2]. When given an unknown tuple, a k-nearest-neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. These k training tuples are the k "nearest neighbors" of the unknown tuple. "Closeness" is defined in terms of a distance metric [2], such as Euclidean distance (4). The Euclidean distance between two points or tuples, say, X1 = ($x_{11}$, $x_{12}$, ... , $x_{1n}$) and X2 = ($x_{21}$, $x_{22}$, ... , $x_{2n}$) is :

$$dist(x_1, x_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \quad (4)$$

## IV. COLLECTING THE DATASETS

In this paper, the heart disease data is collected from UCI machine learning Repository [6]. The totals of 296 instances were collected. Each instance consists of 14 attributes. The records were split into two data sets. A rule of thumb is to use 2/3rd data as training set and 1/3rd data as testing set [11]. Hence, training set consists of 200 records and testing set consists of 96 records. In order to avoid bias, records for both training and testing set was selected randomly. Data set consists of both numerical and categorical attribute values. The data sets are present in the form of text files which makes processing easier and faster. Each instance is separated by a comma in a text file. The detailed description of attributes is shown in Table I.

The attribute "Diagnosis" is identified as predictable attribute with values "Yes" for the patients having heart disease risk and value "No" for the patients having no heart disease risk. If the data set contains missing value, that record was discarded and care was taken so that the data set does not contain any duplicate values.

TABLE I

INPUT ATTRIBUTES USED FOR EXPERIMENT

| Attributes | Description |
|---|---|
| age | Age in years |
| gender | Male, Female |
| cp (Chest pain type) | typical angina , atypical angina, non-anginal pain, asymptomatic |
| trestbps | resting blood pressure |
| chol | serum cholestoral in mg/dl |
| fbs (fasting blood sugar > 120 mg/dl) | Yes, No |
| restecg - resting electrocardiographic results | Normal,<br>Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV),<br>Showing probable or definite left ventricular hypertrophy by Estes' criteria |
| thalach | maximum heart rate achieved |
| exang (exercise induced angina) | Yes, No |
| oldpeak | ST depression induced by exercise relative to rest |
| slope - the slope of the peak exercise ST segment | Upsloping, flat,<br>Downsloping |
| ca (number of major vessels colored by flourosopy) | 0, 1, 2, 3 |
| thal | Fixed defect, Normal, Reversable defect |
| Diagnosis (the predicted attribute) | Yes: has heart disease<br>No: no heart diseases |

## V. SYSTEM ARCHITECTURE

The architectural design gives a high-level view of the system. It also shows how the components or sub-systems interact with each other. Fig. 1 depicts the flow of the system. The heart disease data set is first pre-processed to eliminate missing values. Discretization of continuous data is also performed in pre-processing phase. Once the data set is pre-processed it is applied to classification phase. Different classification algorithms, namely Naive Bayes Classifier (NBC), ID3 and K-Nearest Neighbor (k-NN) are trained in this phase. Test data set is used to evaluate the trained algorithm. Further, the user can also input new data to test the

system. In this case, the system predicts whether or not the new input provided by the user has a heart attack risk. Also the system finds the accuracy of three classification algorithm.

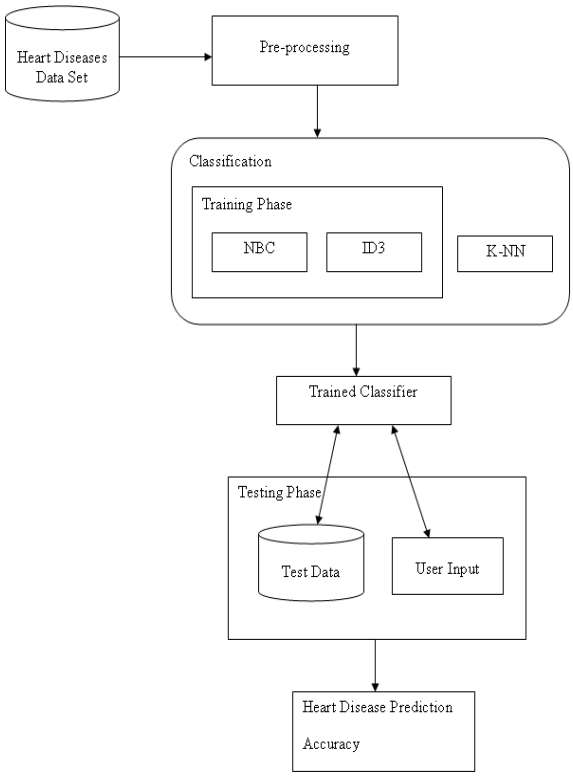


Fig. 1 System Architecture

## VI. IMPLEMENTATION DETAILS

The goal of this paper is to use data mining techniques to predict heart attack risk. Classification techniques are adopted to develop a predictive model. The model is built using Java Programming language. As a pre-processing step missing and redundant values identified in the data set were corrected manually. For Naive Bayesian Classifier data set remains unchanged. It was taken as it is in [6]. For K-Nearest Neighbor Classifier data set was changed to numeric form. For ID3 Classifier requires data set to be in the form of categorical type. Entropy based Discretization method was employed to form categories for continuous-valued attributes [13]. This is a supervised learning method which is used to convert continuous-data into categorical form. It is characterized by finding the split with the maximal information gain [14]. The instances are then categorised according to splitting interval. The Discretization technique is computed on training data set. It is then applied unchanged to the test data set. The outcome of pre-processing step is the data set appropriate for next process. Once the data became ready, classifiers were trained using this data. When the system is loaded, initially the user interface (UI) will be displayed with the 13 attributes. The user must input the clinical data. The "Predict" button is used to get the predicted result.

A challenging task in k-Nearest Neighbor Classifier was on deciding the correct value for k. By estimating the error rate of classifier appropriate k value can be determined. Starting with k =1 error rate was computed each time by incrementing the value of k. For even number of classes, k value should be odd. Hence, k value was chosen as k = 1, 3, 5 ... upto 15. Table II shows the error rate and accuracy obtained for various values of k.

TABLE II

ANALYSIS OF K-NN CLASSIFIER

| K value | Accuracy (in %) | Error Rate |
|---|---|---|
| 1 | 87.5 | 0.125 |
| 3 | 76.01 | 0.2398 |
| 5 | 76.01 | 0.2398 |
| 7 | 72.29 | 0.277 |
| 9 | 68.24 | 0.3175 |
| 11 | 68.91 | 0.3108 |
| 13 | 67.22 | 0.3277 |
| 15 | 65.87 | 0.3412 |

The analysis showed that error rate is low and accuracy is high for k = 1 compared to other values of k. Therefore, for this paper k = 1 was determined. Thus, the classifier is the Nearest Neighbor Classifier.

The steps followed in building the system are as follows:

1. Read training file.
2. Train the classifiers NBC, K-NN and ID3 using training data set.
3. Take user input.
4. Apply user input to the trained classifiers and get predicted results.
5. Compute accuracy of classifiers.
6. Display the predicted class and accuracy of classifiers.

Fig. 2 shows the prediction as NO for user input and also accuracy values for the input data. Once user inputs all values, Predict button has to be clicked to obtain whether or not the input data has a risk of heart attack. On clicking Accuracy button, accuracy of prediction is displayed. On

clicking of Accuracy of NB, Accuracy of k-NN and Accuracy of ID3 buttons, accuracy of these classifiers is displayed.

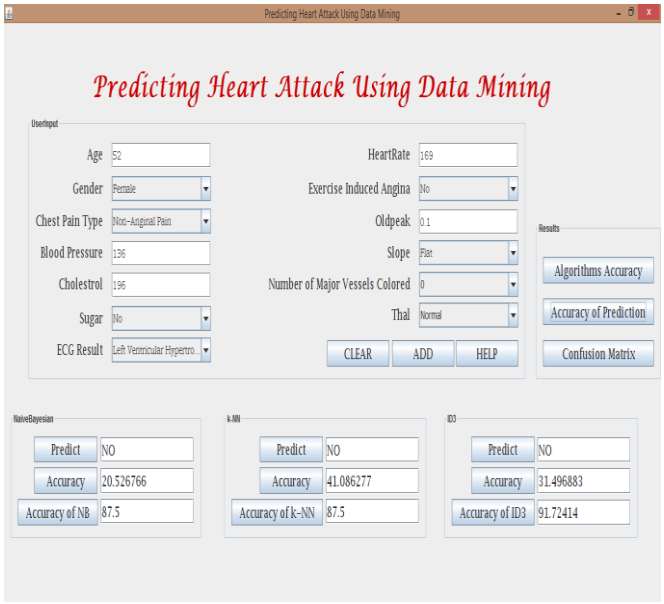


Fig. 2 User interface to predict heart attack risk

## VII. EXPERIMENTAL RESULTS

This section consists of experimental results and analysis of the system. When a user clicks "Accuracy" button, the performance of NBC, K-NN and ID3 is displayed in terms of accuracy.

The performance of classification algorithm in predicting the correct class for a given instance is measured in terms of Accuracy. A test data set is used to find the accuracy of classification algorithm. Accuracy is defined as the percentage of test data correctly classified by the classifier. Accuracy is calculated by generating a confusion matrix (5). A confusion matrix displays the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data. A sample confusion matrix is shown in Table III.

TABLE III

SAMPLE CONFUSION MATRIX

| | | Predicted Class | |
|---|---|---|---|
| | | $C_1$ | $C_2$ |
| Actual Class | $C_1$ | True Positive (TP) | False Negative (FN) |
| | $C_2$ | False Positive(FP) | True Negative (TN) |

Accuracy is computed as follows [2]:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Where,
**TP (True Positive)** = If the instance is positive and it is classified as positive, it is counted as a true positive.
**TN (True Negative)** = If the instance is negative and it is classified as negative, it is counted as a true negative.
**FP (False Positive)** = If the instance is negative and it is classified as positive, it is counted as a false positive.
**FN (False Negative)** = If the instance is positive and it is classified as negative, it is counted as a false negative.

The error rate or misclassification rate (6) is calculated as follows [2]:

$$\text{Error Rate} = \frac{FP+FN}{P+N} \quad (6)$$

Where, P = TP + FN and N = FP + TN. Error rate is used to determine value of k in k-Nearest Neighbor Classifier.

In order to compute accuracy, each instance from a test data set is read. This instance is then passed to all the three classifiers and predicted results are obtained. Now, the predicted class is compared with actual class values of the test data set. The number of correct classification is obtained. A confusion matrix is generated which is then used to calculate the accuracy of the classifier.

Fig. 3 shows the confusion matrix for Naive Bayesian Classification (NBC), k-Nearest Neighbor (k-NN) Classification and ID3 Classification. Based on the values in confusion matrix accuracy of algorithms is calculated. This confusion matrix is displayed upon clicking of Confusion Matrix button.

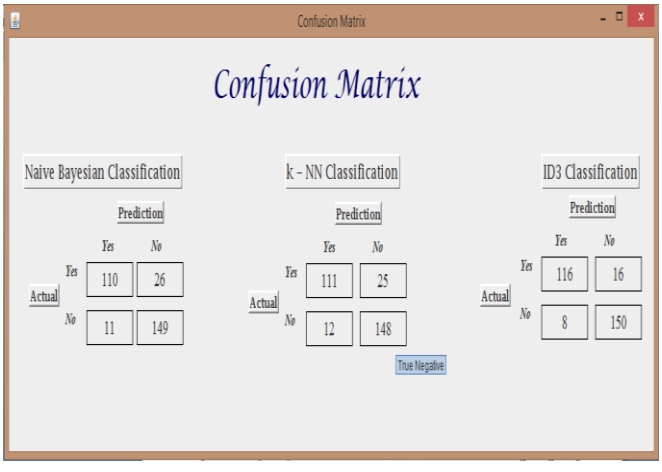


Fig. 3 Confusion Matrix

Fig. 4 depicts the bar chart showing the results of confusion matrix. Confusion matrix consists of True Positive, False

Negative, False Positive and True Negative values. Bar chart for all these values for the three algorithms is generated. This confusion matrix bar chart is displayed upon clicking of Confusion Matrix button.

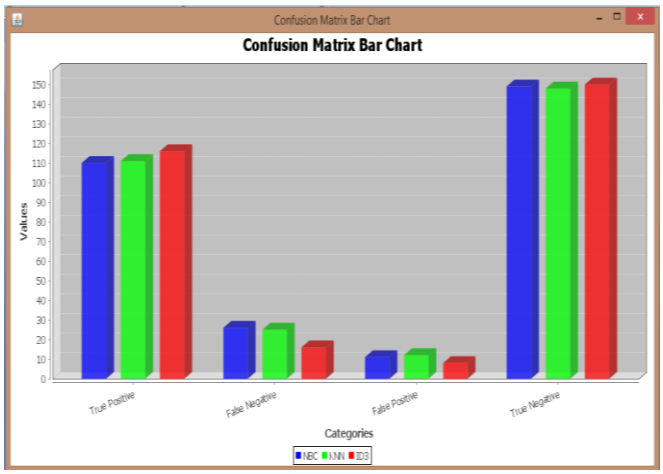


Fig. 4 Confusion Matrix Bar Chart

Fig. 5 shows the bar chart comparing the accuracy of classification algorithms namely Naive Bayesian Classification (NBC), k-Nearest Neighbor (k-NN) Classification and ID3 Classification. This bar chart is generated upon clicking of Algorithms Accuracy button.

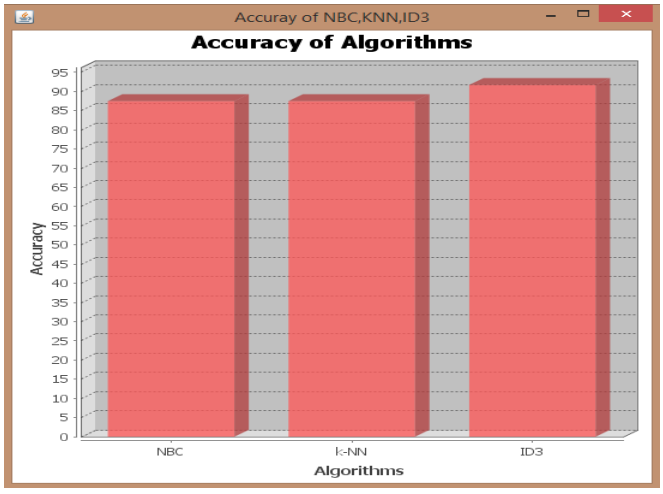


Fig. 5 Accuracy of algorithms

Table IV depicts the accuracy obtained for NBC, K-NN and ID3 Classifications.

TABLE IV

ACCURACY OF NBC, K-NN, ID3

| Classifiers | Accuracy (in %) |
|---|---|
| NBC | 87.5% |
| k-NN | 87.5% |
| ID3 | 91.72% |

Fig. 6 depicts the bar chart comparing the accuracy of prediction for classification algorithms namely Naive Bayesian Classification (NBC), k-Nearest Neighbor (k-NN) Classification and ID3 Classification. It clearly shows that k-NN Classification has the highest prediction accuracy compared to other two classification techniques. This bar chart is generated upon clicking of Accuracy of prediction button.

From the experiments, it is found that accuracy of Naive Bayesian Classification, k-Nearest Neighbor Classification and ID3 Classification are 87.5%, 87.5% and 91.72% respectively. Based on the analysis it is found that accuracy is better in ID3 Classification compared to other two classification techniques.

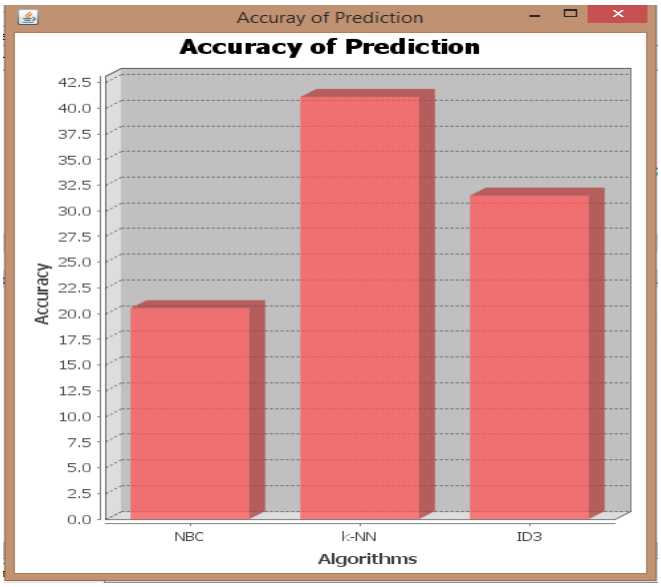


Fig. 6 Accuracy of prediction

## CONCLUSION

The aim of this paper was to develop a prediction model for predicting heart attack risk using data mining techniques. The data mining technique applied here was classification. Three different classification algorithms have been implemented namely, Naive Bayesian Classification, k-Nearest Neighbor Classification and ID3 Classification. For pre-processing of data set, Entropy based Discretization method was employed. A GUI has been designed so that the user can input patient's record. The dataset for heart disease was collected from UCI Repository. Total of 296 records were available. Each record consists of 14 attributes including class attribute. The records were split into training set and testing set randomly. The performance of each classifier was evaluated using a standard metric known as Accuracy. From the experiments, it is found that accuracy of Naive Bayesian Classification, k-Nearest Neighbor Classification and ID3 Classification are 87.5%, 87.5% and 91.72% respectively. Based on the analysis it is found that

accuracy is better in ID3 Classification compared to other two classification techniques. Hence, this study showed that data mining techniques can be efficiently applied in medical domain to model and predict heart attack risk.

## REFERENCES


[1] Sivagowry .S, Dr. Durairaj. M, Persia.A, "An Empirical Study on applying Data Mining Techniques for the Analysis and Prediction of Heart Disease", Int, Conference on Information Communication and Embedded System (ICICES), ISBN: **978-1-4673-5786-9**, Page No (**265-270**), Feb **21-22, 2013**

[2] Jiawei Han, Micheline Kamber, and Jian Pei, "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, Third (3rd) Edition, ISBN: **1-55860-901-6**, **2012**

[3] Jyoti Soni, Uzma Ansari, Dipesh Sharma, Sunita Soni, "Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers", Int. Journal on Computer Science and Engineering (IJCSE), Volume-**03**, Issue-**06**, Page No (**2385-2392**), **2011**

[4] Asha Rajkumar, Mrs. G.Sophia Reena, "Diagnosis Of Heart Disease Using Datamining Algorithm", Global Journal of Computer Science and Technology, Vol ume-**10**, Issue--**10**, Page No (**38-43**), **2010**

[5] Indian Express: http://archive.indianexpress.com/news/india-set-to-be--heart-disease-capital-of-world--say-doctors/1009607/

[6] UCI Machine Learning Repository [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[7] K.Srinivas, B.Kavihta Rani, Dr. A.Govrdhan, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", International Journal on Computer Science and Engineering (IJCSE), Volume-**02**, Issue-**02**, Page No (**250-255**), **2010**

[8] Shamsher Bahadur Patel, Pramod Kumar Yadav, Dr. D. P.Shukla, "Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques", IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS), Volume -**04** Issue-**02**, Page No (**61-64**), **2013**

[9] Mai Shouman, Tim Turner, Rob Stocker, "Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients", International Journal of Information and Education Technology, Volume-**02** Issue-**03**, Page No (**220-223**), **2012**

[10] Yanwei Xing, Jie Wang, Zhihong Zhao, Yonghong Gao, "Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease", International Conference on Convergence Information Technology, ISBN: 0-7695-3038-9, Page No (**868 – 872**), Nov **21-23, 2007**

[11] Mary Slocum, "Decision Making Using ID3 Algorithm", Rivier Academic Journal, Volume-**08**, Number-**02**, Page No (**1-12**), **2012**

[12] Hnin Wint Khaing, "Data Mining based Fragmentation and Prediction of Medical Data", Int, Conference on Computer Researh and Development(ICCRD), ISBN: **978-1-61284-839-6**, Page No (**480-485**), March **11-13, 2011**

[13] EntropyBasedBinning: http://www.saedsayad.com/supervised_binning.htm

[14] Pang-Ning Tan, Vipin Kumar, Michael Steinbach, "Introduction to Data Mining", Addison-Wesley, **2006**


## AUTHORS PROFILE

Prof. S. A. Angadi is currently working as Professor in the Department of Computer Science & Engineering, Center for P. G. Studies, Visvesvaraya Technological University, Belgaum. His research interests include Image Processing, Intelligent Systems and Internet of Things.

Name: Mouna M. Naravani
D/O: Mallikarjun B. Naravani
Has got B. E. in Information Science and Engineering from B. V. Bhoomaraddi College of Engineering and Technology, Hubli, Karnataka and is currently studying in Fourth semester M.Tech (CSE) at Center for P. G. Studies, Visvesvaraya Technological University, Belgaum, Karnataka, India.