To utilize this method we can detect the problem of heart algorithms (naïve bayes, IBK, random forest) on the basis of disease, it may help to reduce death rate. we can detect whether the person attack heart disease or not. is known in earlier. Using data mining classification techniques on heart disease data set with help of weka tool, Classification analysis is one of the main techniques used in Data Mining. The classification method analyses the data from different perspectives and evaluates their accuracies and detecting a problem in various aspects. In this work, we focus on Heart Disease problem; specifically we took University of California, Irvine (UCI) heart disease dataset and WEKA (Waikato Environment for Knowledge Analysis) data mining tool. Various researches have investigated this dataset for better performance measures. In our paper does a comparative study of commonly used machine learning algorithms to detecting heart diseases. The aim of this research to judge the accuracy of different data mining algorithms such as NaiveBayes, IBK, RandomForest on heart disease dataset and determine the optimum algorithm for detection of heart disease.

The classification is the process of finding a model. A model is called as constructing a class label to a set of unclassified classes. In supervised classification, the set of possible classes is known in earlier. Using data mining classification techniques on heart disease data set with help of weka tool, we can detect whether the person attack heart disease or not. To utilize this method we can detect the problem of heart disease, it may help to reduce death rate.

In this paper we described the analysis of three different algorithms (naïve bayes, IBK, random forest) on the basis of various performance parameters. Simple definitions about these three algorithms are,
- The first approach is NaiveBayes classifier. It is one of the techniques of bayes classification. It uses probabilistic feature based on applying bayes theorem with strong assumptions between the features. It is extremely scalable classifier. It mainly used in medical field.
- The second approach is IBK classifier. IBK may refer to: IBK algorithm is implements from the k-nearest neighbor algorithm, In weka it is called as IBK (Instance-Based learning with parameter k) and it is in lazy class folder. In machine learning it sometimes called as memory-based learning. k- Nearest neighbor algorithm, kernel machines and RBF networks are example of IBK classifier.
- The third approach is RandomForest classifier. RandomForest algorithm is comes under tree classification technique. It is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees. In machine learning way of saying the RandomForest classifier. RandomForest algorithm can use both for classification and the regression kind of problems.
algorithm, K-means clustering algorithm for their analysis. They collected patient’s databases from experience specialists. Diagnosis of heart disease is important task in medicinal data mining. Maximal Frequent Item set Algorithm used association rule mining procedure. Item set algorithm is used to classify the data based on cross validations and partitions techniques. C4.5 algorithm is a training set algorithm; it shows about heart attack rank with a tree structure. K-means clustering algorithm is used to cluster database, which will remove positive feedbacks from database. Due to these methods they predict heart disease symptoms successfully.

They used data mining classification techniques [4]. Those are support vector machine (SVM), logistic regression and neural networks, to predict the prevalence of heart disease. They used Cleveland dataset for their analysis and evaluate the accuracy. From their research Logistic Regression and SVM gives lesser complex models and give more accurate result. They used F1 score and ROC curves to evaluate measures. Through this process they told about early prediction of heart disease.

They proposed a new method of homogenous data mining technique and hybrid data mining algorithms. It is used to predict the early prognosis of cardiovascular diseases [5]. In this paper they gave detailed account of support vector machine, decision trees, and logistic regression algorithms and rule based approach. Rule based model compare the accuracy of applying rules to individual algorithms on Cleveland heart disease database in order to provide an accurate model of predicting heart disease.

The relative study on another research, these authors analyzed different data mining tools using different data mining techniques [6]. This paper compares performance of different data mining tools like WEKA, XLMiner and KNIME with the methods of classification, clustering and association rule mining. They used Statlog heart disease dataset for analyzing performance of tools and techniques.

They are developed prediction models for heart disease survivability [7]. They used three famous algorithms extracted from a decision tree or rule-based classifier to develop the prediction model using a large set of database. They used 10-fold cross validation methods and rule based classifier on CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and Decision Table (DT) algorithms.

### III. Methodology

A classifier is a supervised function (machine learning tool) where the learned (target) attribute is categorical (nominal) [7]. Methodology incorporate a brief explanation about data set, data mining tool, classification of algorithms etc., The main aim of this paper is to find best classification algorithm for heart disease detection between Bayesian (naïve bayes), Lazy (IBK) and Tree (random forest) classifiers.

#### A. Dataset Description

We have tested the Statlog Heart Disease Dataset (SHDD) on different data mining classification methods such as bayes, lazy, trees. This dataset is taken from the University of California, Irvine (UCI) Machine Learning Database. It contains totally 270 instances and 14 attributes of healthy persons and patients with heart problem [8-10].

1) age
2) sex
3) chest
4) resting_blood_pressure
5) serum_cholesteral
6) fasting_blood_sugar
7) resting_electrocardiographic_results
8) maximum_heart_rate_achieved
9) exercise_induced_angina
10) oldpeak
11) slope
12) number_of_major_vessels
13) thal
14) class

The type of class attribute is nominal and it have two distinct values, it included the label as absent and present regarding the absence and presence of heart disease respectively.

<table>
<thead>
<tr>
<th>No</th>
<th>Age</th>
<th>sex</th>
<th>chest</th>
<th>resting</th>
<th>serum</th>
<th>fasting</th>
<th>Electro</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>70.0</td>
<td>1.0</td>
<td>4.0</td>
<td>130.0</td>
<td>322.0</td>
<td>0.0</td>
<td>2.0</td>
</tr>
<tr>
<td>2</td>
<td>67.0</td>
<td>0.0</td>
<td>3.0</td>
<td>115.0</td>
<td>564.0</td>
<td>0.0</td>
<td>2.0</td>
</tr>
<tr>
<td>3</td>
<td>57.0</td>
<td>1.0</td>
<td>2.0</td>
<td>124.0</td>
<td>261.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>4</td>
<td>64.0</td>
<td>1.0</td>
<td>4.0</td>
<td>128.0</td>
<td>263.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>5</td>
<td>74.0</td>
<td>0.0</td>
<td>2.0</td>
<td>120.0</td>
<td>269.0</td>
<td>0.0</td>
<td>2.0</td>
</tr>
<tr>
<td>6</td>
<td>65.0</td>
<td>1.0</td>
<td>4.0</td>
<td>120.0</td>
<td>177.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>7</td>
<td>56.0</td>
<td>1.0</td>
<td>3.0</td>
<td>130.0</td>
<td>256.0</td>
<td>1.0</td>
<td>2.0</td>
</tr>
<tr>
<td>8</td>
<td>59.0</td>
<td>1.0</td>
<td>4.0</td>
<td>110.0</td>
<td>239.0</td>
<td>0.0</td>
<td>2.0</td>
</tr>
<tr>
<td>9</td>
<td>60.0</td>
<td>1.0</td>
<td>4.0</td>
<td>140.0</td>
<td>293.0</td>
<td>0.0</td>
<td>2.0</td>
</tr>
<tr>
<td>10</td>
<td>63.0</td>
<td>0.0</td>
<td>4.0</td>
<td>150.0</td>
<td>407.0</td>
<td>0.0</td>
<td>2.0</td>
</tr>
<tr>
<td>11</td>
<td>59.0</td>
<td>1.0</td>
<td>4.0</td>
<td>135.0</td>
<td>234.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>12</td>
<td>53.0</td>
<td>1.0</td>
<td>4.0</td>
<td>142.0</td>
<td>226.0</td>
<td>0.0</td>
<td>2.0</td>
</tr>
<tr>
<td>13</td>
<td>44.0</td>
<td>1.0</td>
<td>3.0</td>
<td>140.0</td>
<td>235.0</td>
<td>0.0</td>
<td>2.0</td>
</tr>
<tr>
<td>14</td>
<td>61.0</td>
<td>1.0</td>
<td>1.0</td>
<td>134.0</td>
<td>234.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Figure 1: Data set sample

### B. Tool Description

WEKA (Waikato Environment for Knowledge Analysis) is open source software issued under the GNU General public License [11]. Named after a flightless New Zealand bird, Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. Weka GUI Chooser is appears on the first screen on the weka tool. The GUI Chooser window encompasses Simple CLI, Explorer, Experimenter, Knowledge Flow methods.

Machine Learning is nothing but a type of artificial intelligence which enables computers to learn the data without help of any explicit programs. Machine learning systems crawl through the data to find the patterns and when these are found, adjust the programs actions accordingly.

Weka data formats: Weka uses the Attributes Relation File Format for data analysis by default. But listed below are some formats that weka supports from where data can be imported:

- CSV
- ARFF
- Database using ODBC

### C. Classification

Classification is finding a set of model to applying further process of an analysis. The goal of classification is to
accurately predict the target class for each case in the data [6]. Different classification algorithms use different techniques for finding relationships [12]. These relationships are summarized in a certain format, to evaluate the result in a different manner to find out accuracy. In this research, we have analyzed three classifiers namely Bayesian, lazy and trees. In Bayesian classifier, we have analyzed the classification algorithm namely naïve bayes, in lazy classifier we have analyzed the classification algorithm IBK, in tree classifier we have analyzed the classification algorithm namely random forest.

D. Bayesian Classification

Bayesian classifiers are statistical classifiers. It’s based on bayes theorem. A class is to predict the features value of the class. The Bayes rule is used for a given a set of features for the known class [13]. A bayes classifier includes NaïveBayes, NaiveBayes Multinominal, Naivebayes Simple, NaiveBayes Updateable algorithms and so on. From these NaiveBayes is simple and better to apply large set of databases.

Naïve Bayes Classifier

Naïve Bayesian classifiers assume attribute value on a given class is individual value of the other attributes. This assumption is called class conditional independence. It made to simplify the computations involved and in this sense, it is called as naïve.

Bayes theorem formula: bayes theorem formula is an important method for calculating conditional probabilities. It is used to calculate posterior probabilities. Bayes theorem describes the probability of an event, based on conditions that sometimes might be related to the event [14].

For instance, a patient can observed some symptoms about heart disease, then we can apply patient database to bayes formula to easily find out that person affected by heart disease or not. Because, bayes theorem telling about probabilistic nature for a patient need treatment or not. In simple words, suppose a doctor is interested in whether a person has heart disease, and knows the persons age. If heart disease is related to age, then using bayes theorem, the person’s age can be used to access more accurate probability that the patient have heart disease [15].

Thomas Bayes introduced Bayes theorem, who first provided an equation that allows new evidence to update beliefs.

\[ P(Q|P) = \frac{P(P|Q) \cdot P(Q)}{P(P)} \]
\[ P(P|Q) = \frac{P(Q|P) \cdot P(P)}{P(Q)} \]

- \( P(P|Q) \) is the probability of \( P \) and \( Q \) has occurred and is known as possibility.
- \( P(Q) \) is notable as prior probability and \( P(Q|P) \) is notable as posterior probability.

Figure 2: Implementation of Naïve Bayes Classifier

E. Lazy Classification

Lazy is a classification technique, it includes IBK, KStar, LWL methods to classify the database. Lazy method doesn’t do anything until last minute. They always use any other model for classification purpose. The lazy learners use the same dataset as both training set and testing set. Before doing construction model in order to classify a given test set [12]. That is when a given training set, a lazy learner first stores training set and waits until a test set. When it sees the test set it will do generalization in order to classify the training set based on its similar property, at last the classifier store the training set. Unlike eager learning methods, lazy learners do less work. Based on this unique nature lazy learners store the training set before classification. Lazy learning solve multiple problems consecutively and deal the problem area in successful.

IBK Classifier

In weka it’s called IBK (instance based learning with parameter k) and it’s the lazy class folder. The term learning is indicating store all training instances. K- Nearest Neighbor (KNN) is an example of IBK Classifier [9]. KNN algorithm is used to specify the number of nearest neighbors to use when classifying a test instances and the outcome is determined by majority vote. Opening of IBK classifier have following steps. The first step to choose weka Explorer initially, then choose dataset, and choose classify tap to get options from IBK implementation. It has the cross-validation option that can help by choosing the best value automatically. weka uses cross-validation to select the best value for KNN. Example of instance-based learning algorithm is k-nearest neighbor algorithm, weighted regression, case-based reasoning, kernel machines and RBF networks.

K-Nearest neighbor

Features

All instances correspond to data point in an n-dimensional Euclidean space. Classification is postponed until a new instance arrives. Classification is done by compare feature vector of different data points. The end function may be real valued or discrete function.

The arbitrary instance is represented by, 
\((x1(a),x2(a),x3(a),...,xn(a))\)

- Let \(xi(a)\) denotes instances

Euclidean distance between two instances
NaïveBayes is a part of bayes approach. It gave 83.70 percentages accuracy, and 226 correctly classified instances also take less time to build a model. On the other hand, other two algorithms are IBK and RandomForest, these are comes under lazy and tree approach, these algorithms are took incorrectly classified instances is high and process time also high. The accuracy of IBK classifier is 75.18 percentages and the accuracy of 81.48 percentages for random forest. From these process IBK and RandomForest algorithms are not suitable for better heart disease detection. NaïveBayes is a best technique in heart disease database to get an accurate result.

In future we have desired to develop web mining and text mining methods to get a better prediction of heart disease symptoms.

System architecture of this research work,

**IV. PERFORMANCE OF THE CLASSIFIERS**

The classifiers performance is evaluating under the following factors,
(1) The time to building a model
(2) Classifying correctly and incorrectly instances depends on their attributes
(3) The accuracy of the classifiers

<table>
<thead>
<tr>
<th>Evaluation Criteria</th>
<th>Classifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NaïveBayes</td>
</tr>
<tr>
<td>Timing to build model (in Sec)</td>
<td>0.09</td>
</tr>
<tr>
<td>Correctly classified instances</td>
<td>226</td>
</tr>
<tr>
<td>Incorrectly classified instances</td>
<td>44</td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>83.70%</td>
</tr>
</tbody>
</table>

Correctly classified instances are taken as accuracy in our research. NaïveBayes is a better algorithm for heart disease detection compare to other classification algorithms.
VI. CONCLUSION AND FUTUREWORK

This paper initially gives brief introduction about Data Mining, Weka tool and heart disease dataset. Weka is a powerful tool used in data mining techniques. The dataset contains 270 instances. The classification algorithms are classifying the data into correctly; incorrectly instances and also it classify the database in different manner. Through this effort we can find out Naive bayes algorithm give better result compare to IBK and random forest algorithm. These algorithms are used in many areas like banking, medical and stock market analysis to predict good result.

Futurework

Future research should concentrate on collecting data from a more recent time period and real symptoms record of the patients. Our dataset contains only 270 instances. This is very small work of a research. This work is extended up to large set of database and identifies the result in the help of two or three machine learning algorithms and then takes better approach from that performance. In future we have compare tools using classification techniques which tool execute accurately.

VII. ACKNOWLEDGMENT

The authors thank to Selvammm Arts and Science College for giving an opportunity to publish the paper. The authors also thank University of Waikato for WEKA tool availability as an open source.

VIII. REFERENCES

[2] Jiawei Han, Micheline Kamber, “Data Mining: Concepts and Techniques, Third Edition”