**ABSTRACT**

Data mining have used worldwide for mining information from different types of database and data mining techniques are quite popular in medical sectors. Nowadays coronary attorney is the most leading cause of death according to many surveys all over the world which also can be predicted by data mining. The main goal of this research is to predict and classify coronary attorney by designing predictive and classification model. Numbers of heart patients are increasing day by day so it is being quite difficult for doctors to analyze lot of patients data manually .So those predictive and classification models would be helpful for doctors to predict and classify coronary attorney.

Authors have studied lots of related paper and they have also gone through some websites for finding the framework and knowledge about this research .There was the description of different research methodology, information and guideline about different algorithms and how to find most accurate result.

Authors have developed a system which contains both predictive and classification model. Firstly they have planned for the whole research then they have selected 24 attributes on the basis of health condition and test reports of patients. Authors have collected 500 data of patients from Padma Diagnosis Center and Islamic general hospital according to the basis of preselected attributes and divided them into training and test dataset both for prediction and classification. WEKA 3.8.1 software have used for creating this system and authors have inserted training dataset and applied j48 algorithm for creating predictive model .Authors have also inserted training dataset for creating classification model by applying Naive Bayes algorithm on those dataset .Authors also have inserted test dataset in both predictive and classification model for testing and getting the outcome from the system.

Predictive model have designed for predicting the coronary attorney .Authors have used j48 algorithm for this model .The system will show the result of inserted test data. There would be two categories to express the result and they are -positive and negative. Positive would mean that the patient have the probability of having coronary attorney and negative would mean that they don’t have the probability of having coronary attorney. Negative would go to the null state of the classification model. Positive result would go for next stage which is classification. Naive Bayes have applied for classification and authors have classified the dataset on the basis of Age, Diabetes and LDL of patients report. Authors have fixed Null, Low, Medium and High categories for expressing the probable state of coronary attorney.

Authors have used a dataset of 300 data as training dataset and 200 data as test dataset in predictive model to predict coronary attorney. Here the prediction of coronary attorney has gained 86% accuracy by j48 algorithm. In classification model authors have used 300 data as training dataset and 200 data as test dataset. Here the classification of coronary attorney has gained 86.5% accuracy by Naïve Bayes algorithm.

This research has fulfilled all the objective of this research paper but there were some limitations too. Here authors have worked with lots of test reports, physical condition and habit of a patent but it’s not able to express the exact reason which is actually responsible for coronary attorney in a patient body and authors also didn’t create any sector in this system where patients could know about the treatment or guideline of each probable stage of coronary attorney. In future author would upgrade the those features which was the limitation of this paper

**CHAPTER 1 INTRODUCTION**

Nowadays Data Mining is one the most popular and effective technique which have used to find the real solution to each and every problem. This technique extract hidden predictive and expected data from a large database and all the processes works through some specific algorithms of Data Mining. Data mining is an important stage of KDD (Knowledge Discovery of Database) which have used for finding potentially important information even from the low level of data. The common purpose of the data mining in medical science is to diagnose and create the predictive model which can predict the specific outcome of any problems of medical science by analyzing the large data set.

Heart Disease and Coronary Attorney is already a thread to all over the world because it accounts to be the most leading cause of death worldwide. A Coronary Attorney is the death of a segment of heart muscle caused by a loss of blood supply. The blood is usually cut off when an artery supplying the heart muscle is blocked by a blood clot. Coronary Attorney called another name of fear and death everywhere and Heart Diseases are truly responsible for this. The people of low income and developing countries are mostly the patient of Heart Diseases. Here male and female both have the same chance percentage of having Heart disease who is the aged within 40-65. The prediction for Heart Disease is possible even though the prediction of Coronary Attorney is a really complex task for medical practitioners because it requires both experience and knowledge to predict. Feelings of pressure, tightness, pain, squeezing, or aching in the chest or arms that spread to the neck, jaw, or back can be a sign that a person have the probability of having coronary attorney. A perfect predictive model which contains accurate and effective attribute is the only way to predict both Heart Disease and Coronary Attorney.

A predictive model is the only possible way to predict Heart Disease and Coronary Attorney. In predictive model researchers have used training data set where data were decorated according to the selected attributes .Then authors have applied data mining algorithm on training data set for creating a predictive model.

In this research, the authors have designed and developed a predictive and classification model to predict Coronary Attorney using J48 and Naïve Bayes algorithms respectively which are both data mining algorithm. To conduct the research work the authors collected some real data of heart-related disease from several hospitals. After creating both predictive and classification model authors checked performance accuracy of the output and ensured that it’s able to perform as a perfect predictive model which would help medical practitioners and patients.

* 1. **Objectives**

1. To develop a predictive model to predict the coronary attorney of human body.
2. To design a classification model which would helps the user with appropriate outcomes.

**1.2 Justification of study**

The number of heart patients increasing day by day all over the world and it is very difficult to analyze lots of patient’s data manually for medical practitioners. So authors have designed a predictive model which would save both time and cost of diagnose. In this research, authors have analyzed all the symptom of the patients and worked with the maximum number of attributes in the expectation of getting an accurate outcome.

**1.3 Scope of study**

A Coronary Attorney is the death of a segment of heart muscle caused by a loss of blood supply. The authors have formed a predictive model which would help the doctors to unroll the accurate result through analyzing the test reports of patients and it would also minimize doctor’s workload. It will help the doctors to know that patients have the probability of having coronary attorney or not and if anyone have the probability the system would also be able to show the probable level of having coronary attorney

**CHAPTER 2 LITERATURE REVIEW**

The research undertook an experiment on performance of various data mining algorithms to predict the Heart attacks .J48, Naïve Bayes, REPTREE, CART and Bayes Net are applied in this research paper .Data was collected from medical practitioners in South Africa. Eleven attributes from database considered for the prediction. Authors described performance of each algorithm and they got J48, Naïve Bayes, REPTREE, CART performed similar and good in dataset while Bayes Net out-performed the Naïve Bayes algorithm. (Hlaudi Daniel Masethe, Mosima Anna Masethe 2014)

In this paper authors stated that Heart Disease affects most in low income countries and there it’s responsible for 82% of death that’s why 2360millions people died by CVD .Authors also stated that in Pakistan people (both male and female) are mostly facing Heart Disease when they are having 65(around) years old .Here authors designed a predictive model by using J48, Naïve Bayes ,Neural network. The research used 209 patient was collected form Faisalabad Institute of Cardiology .Work procedure was described by Data preprocessing, Knowledge discovery, Outer analysis, Clustering, Classification, prediction. WEKA 3.8.1 have used for this research and J48 predictive model showed 80% accuracy based on true positive rate ration and its performance was better than others algorithms.( Noreen Akhtar, Muhammad Ramzan Talib,Nosheen Kanwal 2018)

This paper presented a review of Heart Disease prediction for healthcare system using data mining techniques .Here authors stated that approximately 610,000 people die from heart disease in the United States every year .Though It’s the leading cause of death in both men and women there but the symptoms of heart disease vary between gender. Authors mentioned ‘irregular heartbeat ,congenital heart defects , weak heart muscles ,heart valve problems ,Heart infections , cardiovascular disease’ as the form of Heart disease .They also found ‘a fast or slow heartbeat ,dizziness, chest pains ,shortness of breath’ as symptoms of heart disease and ‘Cardiomyopathies ,Coronary Artery Disease , Diabetes, Diseases of the Heart Valves ,Heart Defects present at Birth ,High Blood Pressure ,Lung Disease such as Emphysema ,Past Heart Attacks’ as the cause of heart attack.(C.Sowmiya, Dr.P. Sumitra 2016 )

The aim of this research was to design a predictive model for heart disease detection using data mining techniques. Data collected from PGI, Chandigarh from the year 2008 to 2011 containing 7,339 instances and 15 attributes was selected. Authors have used Knowledge Discovery in Database (KDD) in their methodology. The authors used J48 Classifier, Naïve Bayes and Neural Network algorithms in WEKA 3.6.4 machine learning software to compare best performance among those algorithms. J48 algorithms provided 95.56%accuracy in predictive model. (Abhishek Taneja, 2015)

In this research author analyzed various data mining tools and techniques for prediction of heart disease and to find most the efficient system for diagnosis. Authors explained the difference between data mining (DM) and knowledge Discovery in Database (KDD). Extract expected data from large database by algorithms is Data Mining and KDD is a full process of extraction of implicit, unique, and potentially useful information from data. They stated that data mining is just an important stage KDD and DM is not a full process and main objective of using KDD is to abstract high-level information from low-level data. Authors divided the risk factors of heart disease in two parts .They found age ,sex and family history as the uncontrollable risk factor and smoking, weight, cholesterol, diabetes, blood pressure as the controllable risk factor of heart disease. Authors also described the implementation of Naïve Bayes, ID3, Neural Network, Support Vector Machine and Hybrid.(S. Kiruthika Devi, S. Krishnapriya and Dristipona Kalita,2016)

The researchers compared the performance of the data mining algorithms (Naïve Bayes, Back Propagation, SVM, K-Nearest Neighbor and C4.5 classifiers) for predicting heart disease, and diabetes on respective patient datasets. Authors described the work procedure briefly by Collect the data, Preprocess the data to reduce the noise , Select the important attributes ,Input the dataset with reduced set of attributes to the classifiers ,Split the dataset into training and testing datasets, Apply data mining techniques to the training dataset ,Develop the predictive model, Evaluate the performance of model using test dataset , Repeat this process with other all classification techniques , Compare the performance among these data mining techniques. Average accuracy of used classification techniques was 75%.Authors selected 13 attributes for their dataset and database contained 270 data after removing the instances with missing attribute values. Authors compared of Performances of Classification Algorithms on CHDD .The result of this prediction model showed 120 positive and 150 negative cases.

(Sambasiva Rao Voletiz, Kiran Kumar Reddi, 2016)

In this research authors stated that “In Europe and the United States, the Framingham Risk Score (FRS) is utilized as a guideline for predicting the risk of

Cardiovascular disease using various factors that influence the disease” .Authors developed a prediction model using data mining, in order to predict a risk of cardiovascular disease at early stage. The model used the data set of Korean National Health and Nutrition Examinations Survey 5 .Data of total 6,742 people and 8 attributes were selected for this research. Discriminate analysis method used to classify the data and the Result of proposed model showed a higher accuracy among Korean patients than the FRS guideline, and showed much lower errors than the FRS. (Yang et al., 2013)

The authors found that 12 million deaths occurs worldwide and more 50% occurred in United States and other developed countries. The research focused on using different algorithms and combinations of several target attributes for intelligent and effective heart attack prediction using data mining. For predicting heart attack, significantly 15 attributes were selected. KNN, Neural Networks, Bayesian classification, Classification based on clustering, Decision Tree. The predictive data mining technique on the same dataset showed that Decision Tree outperforms and sometime Bayesian classification is having similar accuracy as of decision tree but other predictive methods like KNN, Neural Networks, Classification based on clustering are not performing well. The accuracy of the Decision Tree and Bayesian Classification improved after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction.(Joyti Soni et al. 2011)

The aim of this research is to predict the risk of coronary heart disease. A total of 140 males and 207 females between 30 and 74 years of age were selected purposively from three Government colonies at Dhaka city and data were collected by interview, blood pressure measurements and laboratory analyses. Three fifth of the respondents were females and more than two fifth (44%) were between age 30 and 39 years. More than two fifth (45.7%) male participants were current smoker (p<0.001), around one fourth of the total participants were diabetic, two-fifth proportion had hypertension and most of them (87%) had dyslipidemia where females were higher in proportion (p<0.01). Around three fourth of the participants had low risk (<10% CHD risk in 10 years), followed by more than one fifth (21.6%) in the intermediate risk (10-20 %) and around three percent had high risk (>20%).Males were in higher risk than females (p<0.001). (Monower MM et al., 2018)

**CHAPTER 3 METHODOLOGY**

**3.1 Methodology**

In this research authors used six phases to explain the system. The phases are planning, requirement analysis, data collection and preprocess proposed model, design & development, system testing &outcomes.

**Planning**

**Requirement Analysis**

**Data collection and preprocess**

**Proposed Model**

**Design & Development**

**System testing & outcomes**

**Fig 3.1: Block diagram of Coronary attorney prediction and classification system**

**3.1.1 Planning**

A perfect planning is the precondition of any successful project. So authors choose it as the first phase of this research. Authors have studied lots of paper which were related to their research and they have also gone through some websites then they have got the framework of their research how they could complete their research successfully so they have classified the full process of work in multiple phases. Planning is the most important part of the methodology because there is a guideline exists for each and every phase of planning.

**3.1.2 Requirements analysis**

Requirements analysis is the tasks that an analyst performs to structure and organize requirements, specify and model requirements and designs, validate and verify information, identify solution options that meet business needs, and estimate the potential value that could be realized for a solution option. There are two types of requirement –

1. System analysis
2. User analysis

**3.1.2.1 System Requirement**

Systems requirement can be called as a set of all those elements which is essential to run the system in proper way .System requirement can divided into two part and both of them are responsible equally for an accurate output. They are:

* **Software Requirements:**

WEKA, Web browser and Drivers.

* **Hardware Requirements:**

Laptop, Display Adapter, Memory, Secondary Storage, Peripheral, Processing Power.

**3.1.2.2 User Requirement**

A product which fulfills user requirement always gets fame and success .Researchers was careful about user requirement of the system. Sometimes it takes lots of time to predict coronary attorney of a patient and lots of people find it as a problem so authors were hopeful to develop a system which would predict and classify coronary attorney.

**3.1.3 Data collection and preprocess**

Data collection and preprocess was the most complicated work for authors. Authors have studied about all the health condition and habits of a patient which can be the responsible for the coronary attorney through the related paper and websites. They have selected 24 attributes which were really effective and essential to predict coronary attorney. The attributes were Patient Id, Patient Name, Gender, Age, weight, weight loss, Drug dependence, Drug addiction, Hyper tension, Systolic blood pressure, Diastolic blood pressure, Cholesterol HDL, Cholesterol LDL, Fasting blood sugar, Post Mensal Blood sugar ,Heredity, Physical activity, Blurred vision, Tiredness, Chest pain, Shoulder \_Pain, EPIGASTRIC, ECG change, Heart Prediction. Authors have needed actual data of the patient from the hospital .So authors have applied at Padma Diagnosis Center, Islamia General Hospital, Popular Medical College & Hospital, LABAID Specialized Hospital, Ibrahim Cardiac Hospital, and National Heart Foundation of Bangladesh for patient’s data. At last authors have got 500 patients data from Padma Diagnosis Center and Islamia general hospital. Then authors have decorated the data set according to the structure of each and every attribute. There were some missing values of data there authors have used null/0 then authors have inserted the data set into Microsoft Office Excel.

**3.1.4 Proposed model**

In this research authors have used WEKA software. There were lots of data mining algorithms exist for several tasks in WEKA software. But the performance accuracy and execution time is not same for all algorithms. Authors tested multiples algorithms for prediction and classification model. They have found J48 and Naïve Bayes as the most effective and time-consuming algorithm. According to many other researchers these algorithms are best. So authors have used J48 for prediction of Coronary attorney and Naïve Bayes to classify the probability of coronary attorney.

Insert Patient Dataset

Null

Predicted Result

Prediction

Apply Algorithm (J48)

Negative

Positive

Classification

Apply Algorithm (Naïve Bayes)

Classified Result

High

Medium

Null

Low

**Fig 3.2 Flow chart of the system**

Fig 3.2 is expressing the flow chart of the system. Firstly authors have inserted patient dataset in the system. Then authors have created prediction model by applying j48 algorithm on the dataset. Authors have selected two categories to express the predicted result and it was positive and negative. According to the system negative results go to the null state and positive results go the next stage which is classification. In classification stage authors have applied Naïve Bayes algorithm on positive predicted results and created classification model which is able to show the classified result in four categories .The categories are – Null ,Low ,Medium and High .

**3.1.5 Design and development**

It is the most important phase of the methodology. The authors have selected 24 and 10 attributes for prediction and classification respectively. Here authors have divided data set into two parts, training and test data set both for prediction and classification. Training dataset of prediction was structured according to the basis of preselected 24 attributes then authors have selected j48 algorithm and generated a predictive model. After generating predictive model they have inserted test data set. Authors have decided to show the result of predictive model into two categories, they are- positive and negative. Positive means that the patient has the probability of coronary attorney and negative means that the patient don’t have the probability of having coronary attorney. In classification, authors have used 10 attributes for both training and test dataset. Age, and test reports were presented on this 10 attributes. Authors have decided to classify the result into four categories and they were High, Medium, Low and Null. This would explain the probability level of coronary attorney of a patient. In coronary attorney Age, LDL, Systolic blood pressure, and Diabetes are the most important health condition.

|  |  |  |  |
| --- | --- | --- | --- |
| Age | Diabetes | Cholesterol (LDL) | Categories |
| <=20 | ------ | ----- | Null |
| 20> | Yes | <100 | Null |
| 100-129 | Low |
| 130-159 | Medium |
| 160-190 | High |
| 20> | No | 100-129 | Null |
| 130-159 | Low |
| 160-189 | Medium |
| 190=> | High |

**Table 3.1: Classification table of Coronary attorney**

**3.1.6 System testing & Outcomes**

Authors have used training and test data for both prediction and classification of Coronary Attorney of the human body. In prediction, authors have known about the positive and negative probability of having Coronary Attorney. Authors have put 300 data in training and 200 data in test data set which have used for prediction and dataset had 24 attributes .J48 have showed 86% accuracy of the predictive model. The correctly classified instances were 86% and the incorrectly classified instances were 14%. In classification, authors have set the outcomes within (Medium, High, Low, Null) four categories. Authors have put 300 data in training and 200 data in test dataset which have used for classification and datasets have contained 10 attributes. Naive Bayes showed 86.5% accuracy in the classification model. The correctly classified instances were 86.5% and the incorrectly classified instances were 13.5%.

**CHAPTER 4 REQUIREMENT ANALYSES, DESIGN & DEVELOPMENT**

**4.1 Requirements analysis**

In any kind of project requirement analysis is actually analyze all those tasks which have used to determine the necessity, expectation and condition both for the users and developers. Requirement analysis is truly responsible for success or failure of a system. The requirement needs to be documented and related to the needs and expectation of users. System analyst divides the activity of requirement analysis into two parts, first one is requirement gathering and analysis and second one is requirement specification. Most of the time analyst gather and analyze requirement by observing existing project, studying existing procedures or by discussing with users and requirement specification is a scale to determine that analyzed requirements are actually expected by the users and developer or not . The main aim of the specification is to organize the requirements in systematical and documental way. In this research authors found two of requirements

1. User requirements
2. System requirements

**4.1.1 User Requirement**

A product which fulfills user requirement always gets fame and success .So it’s necessary to be careful about user requirement. Sometimes it takes lots of time, complexity and expense to predict coronary attorney by doctors by analyzing patient’s reports and the researchers have found that this system which contains both predictive and classification model which can be useful for doctor to know the patient’s probability of having coronary attorney and its level too in less time, complexity and expense.

**4.1.2 System Requirements**

Systems requirement can be called as a set of all those elements which are essential to run and develop the system in a proper way. According to analyst to run any system there are basically two types of system requirements and both of them are equally responsible for an accurate output of the system. One is software requirement and another is hardware requirement. They are:

* **Software Requirements:**

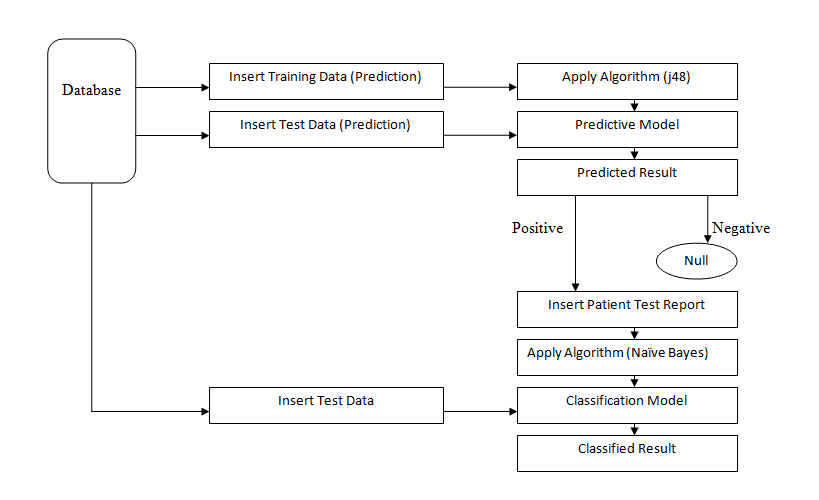
WEKA, Web browser and Drivers.

* **Hardware Requirements:**

Laptop, Display Adapter, Memory, Secondary Storage, Peripheral, Processing Power.

**4.2 System Design**

System design expresses the working process of any system. The system design of this system have drew and described below.

****

**Fig 4.2: System design for prediction and classification of Coronary Attorney**

There is a predictive and a classification model exists in the system. In this system database has stored all the datasets for prediction and classification. System will insert training data set and will apply j48 algorithm for creating predictive model .After that the system would be ready to predict coronary attorney and it can be tested by inserting test data to know the positive or negative probability of having Coronary Attorney of patients. If results are negative it means patients have no probability of having Coronary Attorney and the system will save this on Null state. Otherwise if the results are positive then those data will go to the next stage which is insert patent test report .In this stage authors have reduced the attributes of those positive predicted data and they have only used age and test reports of patients on the data . Authors have changed the structure of data because test reports are the only essential things which can be used for the classification of patient data. Then author have applied Naïve Bayes algorithm on those data and created a classification model .After that the system was ready to classify the probability level of coronary attorney by inserting test data into the classification model.

**4.2.1 Predictive Data**

Authors designed a predictive model to predict coronary attorney of human by inserting patient’s real data as training data in j48 algorithm. Authors have selected 24 attributes and decorated data according to attributes.

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl. No.** | **Attribute Name** | **Attribute Description** | **Attribute Values** |
| 1. | Patient Id | Patient Id Number | Numerical |
| 2. | Patient Name | Name of Patient | Examples: Mr. ,Mrs. |
| 3. | Gender | Patient Male or Female | M/F |
| 4. | Age | Age in Year | 16-20,21-40,41-60,61-80,>80 |
| 5. | Weight | Weight in kg’s | 5-10 kg , 11-18 kg , 19-32 kg , 33-50 kg , 51-60 kg , 61-70 kg , 71-80 kg , >81 kg |
| 6. | Weight Loss | Observe suddenly weight loss | Yes /No |
| 7. | Drug Dependence | Dependent on medicines | Yes / No |
| 8. | Drug Addiction | Dependent on drugs | Yes /No |
| 9. | Hyper Tension | Person with hypertension | Yes /No |
| 10. | Systolic Blood Pressure | Enter value of blood pressure upper limit in mmHg | Normal: less than 120 mmHg  Pre-hypertension: 120–139 mmHg  Hypertension: 140 mmHg or higher |
| 11. | Diastolic Blood Pressure | Enter value of blood pressure lower limit in mmHg | Normal: less than 80mmHg  Pre-hypertension :80–89 mmHg  Hypertension:90 mmHg or higher |
| 12. | Cholesterol HDL(High-density lipoprotein cholesterol) | HDL (Good)cholesterol removes harmful and bad cholesterol | To low : Below 40 mg/dL  Acceptable : 40-49 mg/dL  Very Good : 60mg/dL above |
| 13. | Cholesterol LDL(low-density lipoprotein cholesterol) | LDL It is called the "bad" cholesterol because a high LDL level leads to a buildup of cholesterol in your arteries. | Very High : Above 190 mg/dL  High : 160-189 mg/dL  Border line : 130-159 mg/dL  Acceptable : Below 100 mg/dL |
| 14. | Fasting Blood Sugar | Fasting blood glucose: A test to determine how much glucose (sugar) is in a blood sample after an overnight fast. The fasting blood glucose test is commonly used to detect diabetes mellitus. | Normal: 3.9 to 5.4 mmols/l  Pre-diabetes: 5.5 to 6.9 mmol/l  Diabetes: Above 7.0 mmol/l |
| 15. | Post Mensal Blood Sugar | Pre and post meal testing allows you to see how your meal and, where relevant, your medication for that meal affects your blood glucose levels. | Normal :5-7 mmols/l  Pre-diabetes:7-10 mmols/l  Diabetes: Above 10 mmol/l |
| 16. | Heredity | Hereditary heart problem | Yes /No |
| 17. | Physical Activity | Physical activity in a days | 5-4 h ,6-7 h , 7-8 h ,> 8h |
| 18. | Blurred Vision | Have Blurred Vision | Yes /No |
| 19. | Tiredness | Feel tiredness | Yes /No |
| 20. | Chest pain | There are chest pain | Yes /No |
| 21. | shoulder pain | There are shoulder pain | Yes /No |
| 22. | Epigastria | There are epigastria | Yes /No |
| 23. | ECG change | Patient ECG report | Yes/No |
| 24. | Heart attack prediction | Predict heart attack possibility or not | Positive or negative |

**Table 4.1 Attributes details for Coronary attorney prediction**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl.**  **No.** | **Testing**  **Attribute Name** | **Attribute Description** | **Attribute Values** |
| 1. | Patient Id | Patient Id Number | Numerical |
| 2. | Gender | Patient Male or Female | M/F |
| 3. | Age | Age in Year | Newborn-1,2-5,6-10,11-15,16-20,21-40,41-60,61-80,>80 |
| 4. | Weight | Weight in kg’s | 5-10 kg , 11-18 kg , 19-32 kg , 33-50 kg , 51-60 kg , 61-70 kg , 71-80 kg , >81 kg |
| 5. | Drug Dependence | Dependent on medicines | Yes / No |
| 6. | Drug Addiction | Dependent on drugs | Yes /No |
| 7. | Systolic Blood Pressure | Enter value of blood pressure upper limit in mmHg | Normal: less than 120 mmHg  Pre-hypertension: 120–139 mmHg  Hypertension: 140 mmHg or higher |
| 8. | Diastolic Blood Pressure | Enter value of blood pressure lower limit in mmHg | Normal: less than 80mmHg  Pre-hypertension :80–89 mmHg  Hypertension:90 mmHg or higher |
| 9. | Cholesterol HDL(High-density lipoprotein cholesterol) | HDL (Good)cholesterol removes harmful and bad cholesterol | To low : Below 40 mg/dL  Acceptable : 40-49 mg/dL  Very Good : 60mg/dL above |
| 10. | Cholesterol LDL(low-density lipoprotein cholesterol) | LDL It is called the "bad" cholesterol because a high LDL level leads to a buildup of cholesterol in your arteries. | Very High : Above 190 mg/dL  High : 160-189 mg/dL  Border line : 130-159 mg/dL  Acceptable : Below 100 mg/dL |
| 11. | Fasting Blood Sugar | Fasting blood glucose: A test to determine how much glucose (sugar) is in a blood sample after an overnight fast. The fasting blood glucose test is commonly used to detect diabetes mellitus. | Normal: 3.9 to 5.4 mmols/l  Pre-diabetes: 5.5 to 6.9 mmol/l  Diabetes: Above 7.0 mmol/l |
| 12. | Post Mensal Blood Sugar | Pre and post meal testing allows you to see how your meal and, where relevant, your medication for that meal affects your blood glucose levels. | Normal :5-7 mmols/l  Pre-diabetes:7-10 mmols/l  Diabetes: Above 10 mmol/l |
| 13. | ECG change | Patient ECG report | Yes/No |
| 14 | Coronary attorney class | Which type of heart attack possibility he/she has | High ,Medium ,Low ,Null |

**Table 4.1 Attributes details for the classification of Coronary Attorney**

**4.2.2 J48**

In j48 algorithm, a decision tree is grown in a recursive fashion by partitioning the training records into successively purer subsets. Let DT Be the set of training records that are associated with node t and y = {y1, y2. . . yc} be the class labels. The following is a recursive definition of j48 algorithm.

Step 1: If all the records in Data belong to the same class yt, then t is a leaf node labeled as yt.

Step 2: If Data contains records that belong to more than one class, an attribute test condition is selected to partition the records into smaller subsets. A child node is created for each outcome of the test condition and the records in Dt are distributed to the children based on the outcomes. The algorithm is then recursively applied to each child node.(Pang-Ning , Tan,Michael Steinbach & Vipin Kumar,2006)

**4.2.3 Predictive Model**

A predictive model is a model which used to predict any specific outcomes .Here Designers have inserted training dataset for creating a prediction model to predict Coronary Attorney by using j48 algorithm. If anyone wants to test this system or wants to predict result then they have to insert patient’s test data into the system .This system shows two specific types of outcome , positive and negative .Positive means the patient have the probability of having Coronary Attorney and negative means patient don’t have the probability.

**4.2.4 Naïve Bayes**

A training set of tuples and their associated class labels**.** Each tuple is represented by n-dimensional vector X(x1,…,xn), nmeasurements of n attributes A1,…,An

**Classes:** suppose there are mclasses C1… Cm

**Principle:**

* Given a tuple X, the classifier will predict that X belongs to the class having the highest posterior probabilityconditioned on X
* Predict that tuple X belongs to the class Ci if and only if

P(Ci\X)>P(Cj\X) for 1<= j<=m , j<>i

* Maximize P(Ci\X): find the maximum posteriori hypothesis

P(Ci\X)=P(X\Ci)P(Ci)\P(X)

* P(X) is constantfor all classes, thus, maximize P(X\Ci)P(Ci)
* To maximize P(X\Ci)P(Ci), we need to know class prior probabilities

If the probabilities are not known, assume that P(C1)=P(C2)=…=P

(Cm) ⇒ maximize P(X\Ci)

Class prior probabilities can be estimated by P(Ci)=|Ci,D|/|D|

* Assume Class Conditional Independenceto reduce computational cost of P(X\Ci)

Given X(x1,…,xn), P(X\Ci) is:

P (X\Ci)=P(x1\ci)\*P(x2\ci)\*….\*P(xn\ci)

* The probabilities P(x1\Ci), …P(xn\Ci) can be estimated from the training tuples. ([Johann Gamper](http://www.inf.unibz.it/~gamper/) and [Mouna Kacimi](http://www.inf.unibz.it/~mkacimi/) ,2012)

**4.2.5 Classification Model**

Classification model is a model which is able to classify the level of any instance which is already inserted into the model as test data. In this research authors have also created a classification model which can classify the probable level of Coronary Attorney of patient and this model is capable to run with single and multiple data(test data).

**4.2.6 Classification Level**

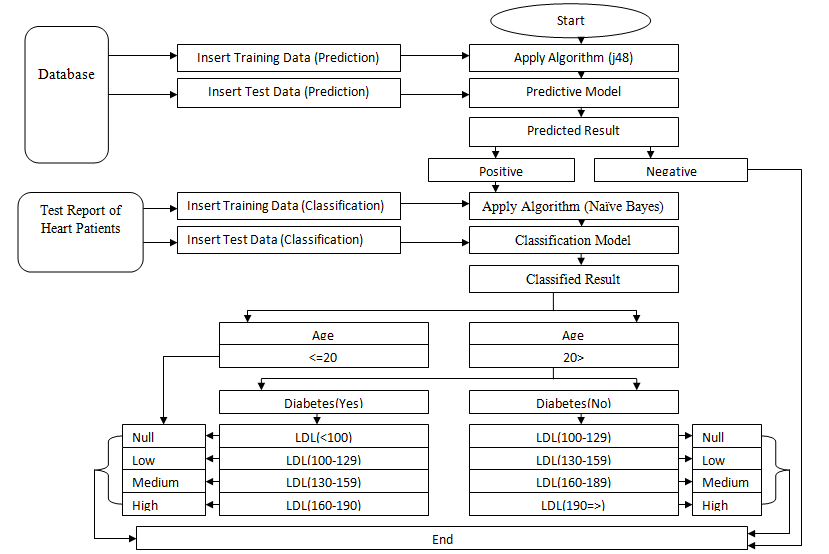
Authors have fixed 4 types or levels of outcomes as result of this classification model. Levels are High, Medium, Low, Null and each of this level express the condition of patients on the basis of Coronary Attorney.

**4.2.7 WEKA**

WEKA is software where lots algorithm and techniques are available for data mining .Authors have used WEKA software to create and run the system which have expressed in this research. Preprocess, Classify, Cluster, Associate, Select attribute and visualization features are exist in WEKA. Authors Design both predictive and classification models by using preprocess, classify and visualization.

**4.3 Proposed model**

The propose model of this system have drew and described below.

****

**Fig 4.2: Proposed for prediction and classification system of Coronary Attorney**

This propose model is expressing the full process of system .For designing this model authors have studied and analyzed lots of research paper they have also gone through many websites. Authors have firstly divided the preprocessed data into training and test dataset both for prediction and classification of the system. They have inserted training data after applying j48 algorithm for creating a predictive model. After the creation of predictive model authors have inserted test dataset or data to know the positive or negative probability of having Coronary Attorney of patients. If result is negative it means patients have no probability of having Coronary Attorney and the system will save this on the end state of the system .Otherwise if the results are positive then those data will go on classification model. Authors can use positive data (mainly patient test reports) of prediction model or also can insert data (mainly patient test reports) after applying Naïve Bayes algorithm. Naïve Bayes algorithm have used for creating classification model. Authors classified outcomes of classification model on the basis of Age, Diabetes and LDL in four fixed categories (High, Medium, Low and Null).If patients age is less than or equal to 20 then the result would be Null there is no necessity of observing anything. If age is over 20 then it will check the state of patient’s Diabetes. Here authors fixed two different ways to classify level. First way for those patients who are over 20 and have Diabetes. If there Cholesterol (LDL) range within160-190 mg/dl then the classification result would be High, if Cholesterol (LDL) range within130-159 mg/dl then the classification result would be Medium, if Cholesterol (LDL) range within100-129 mg/dl then the classification result would be Low and if Cholesterol (LDL) <100 mg/dl then the classification result would be Null.

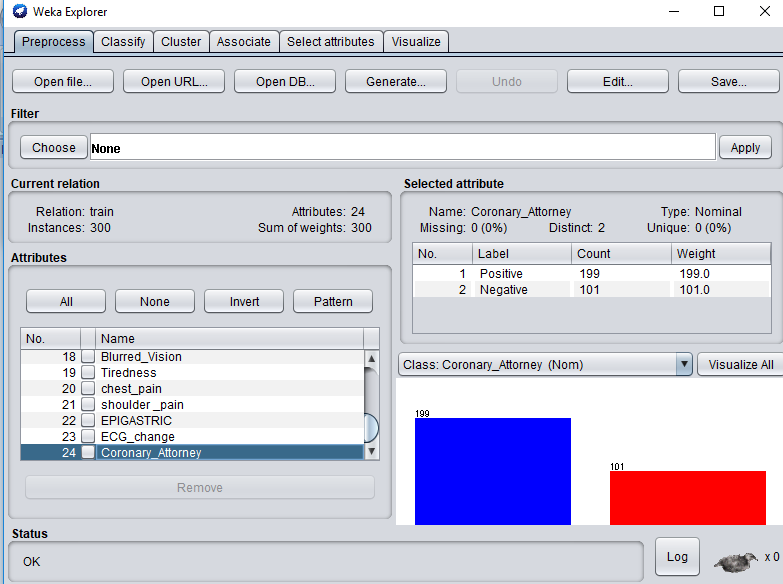
Second way for those patients who are over 20 but don’t have Diabetes. If there Cholesterol (LDL) range within100-129 mg/dl then the classification result would be Null , if Cholesterol (LDL) range within130-159 mg/dl then the classification result would be Low , if Cholesterol (LDL) range within160-189 mg/dl then the classification result would be Medium and if Cholesterol (LDL) 190=> mg/dl then the classification result would be High.

**CHAPTER 5 RESULT ANALYSIS**

Authors collected a dataset of 500 patient’s data and decorated them according to the structure of preselected 24 attributes .Each data is able express each and every health condition and habit of a patient which can be responsible for Coronary Attorney.

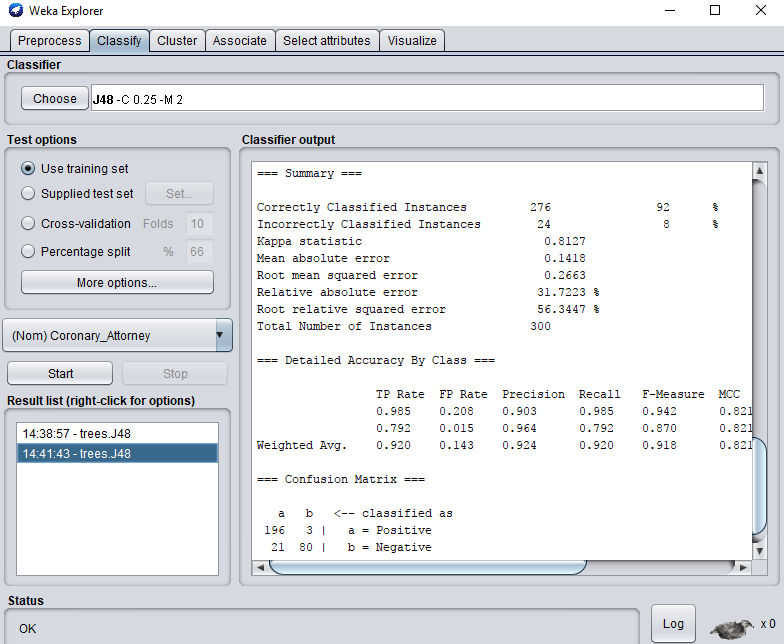
**5.1 prediction model**

In prediction, both training (300data) and test (200data) datasets were classified by 24 attributes then authors have selected j48 algorithm and generated a predictive model. After generating predictive model they have inserted test dataset. After delivering the test dataset authors have got the result as positive and negative. Positive means patient have the probability of coronary attorney and negative means the patient doesn’t have the probability of coronary attorney.

****

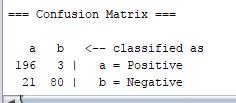
**Fig 5.1: load data into WEKA.**

Fig 5.1 is showing that patient’s dataset have inserted into the WEKA software.

****

**Fig 5.2: Training data results gained by j48.**

Authors have inserted training dataset and applied j48 on those data .Fig 5.2 is expressing the result of inserted data after applying j48 algorithm.

****

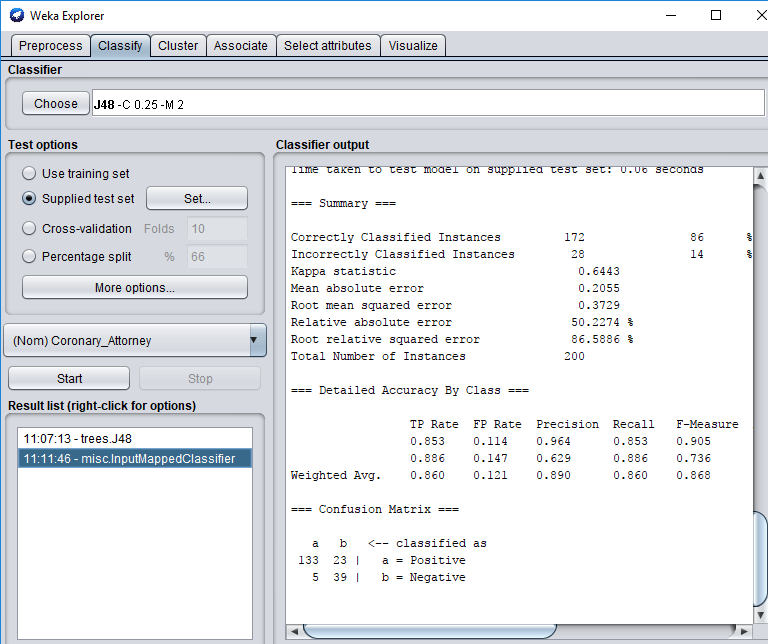
**Fig 5.2.1: Confusion matrix of Training data results gained by j48.**

Fig 5.2.1 is expressing that true positive instances is 196, true negative instance is 80, false positive instances is 3 and false negative instances is 21 of that confusion matrix.

|  |  |
| --- | --- |
| Prediction result | Total Data:300 |
| Coronary Attorney Prediction(Yes) | 196 |
| Coronary Attorney Prediction(No) | 80 |

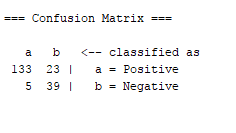
**Fig 5.3: Graphical view of training data results gained by j48.**

Correctly classified instance of training data set is 92% and incorrectly classified instance is 8% Here fig 5.3 is expressing that 71% of data have the probability of Coronary Attorney and 29% don’t have the probability and this classification actually based on correctly classified instance .



**Fig 5.4: Test data results gained by j48.**

Authors have inserted test dataset and appliedj48 algorithm on those data .Fig 5.4 is expressing the result of inserted data after applying j48 algorithm. Here authors have used 200 data and got 86% accuracy and there correctly classified instances are 172 and incorrectly classified instances are 28.This fig also expressing the classified result by confusion matrix.



**Fig 5.4.1: Confusion matrix of Test data results gained by j48.**

Fig 5.4.1 is expressing that true positive instances are 133, true negative instances are 39, false positive instances are 23 and false negative instances are 5 of that confusion matrix.

|  |  |
| --- | --- |
| Prediction result | Total Data:200 |
| Coronary Attorney Prediction(Yes) | 133 |
| Coronary Attorney Prediction(No) | 39 |

**Fig 5.5: Graphical view of Test data results gained by j48**

Correctly classified instances of training data set are 86% and incorrectly classified instances are 14% Here fig 5.5 is expressing that 77% instances have the probability of Coronary Attorney and 23% don’t have the probability .

|  |  |
| --- | --- |
| Result | Accuracy rate |
| Correct accuracy | 86% |
| Incorrect accuracy | 14% |

**Fig 5.6: Accuracy of Coronary Attorney gained by j48.**

The Graphical view of Prediction model gained by j48 algorithm. Fig 5.6 is expressing 86% accuracy of predictive model of coronary attorney.

**5.2 Confusion matrix**

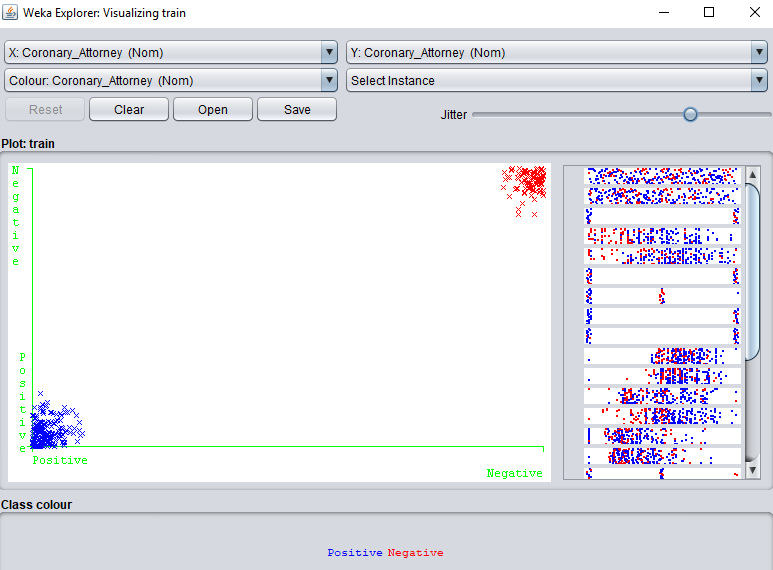
Confusion matrix is actually way to express the result of different kind of models or systems. A classification system combines and compares the output of actual and predicted classifications and final output express by four categories of confusion matrix. They are true positive (TP), false positive (FP), false negative (FN), true negative (TP).Higher percentage of TP and TN in confusion matrix proves that performance of the system is accurate otherwise not accurate.

|  |  |  |
| --- | --- | --- |
| Actual Class | Predicted Class | |
| TP | FP |
| FN | TN |

P N

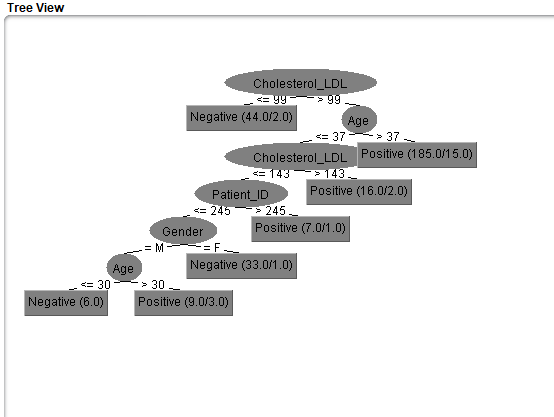
**Table 5.1: Confusion Matrix.**

* **True Positive**: The result which is actually positive and also predicted as positive by this system is called True Positive.
* **False Positive**: The result which is actually not positive (negative) but predicted as positive by this system is called False Positive.
* **False Negative**: The result which is actually not negative (positive) but predicted as negative by this system is called False Negative.
* **True Negative**: The result which is actually negative and also predicted negative by this system is called True Negative



**Fig 5.7: Prediction of Coronary Attorney after applying technique.**

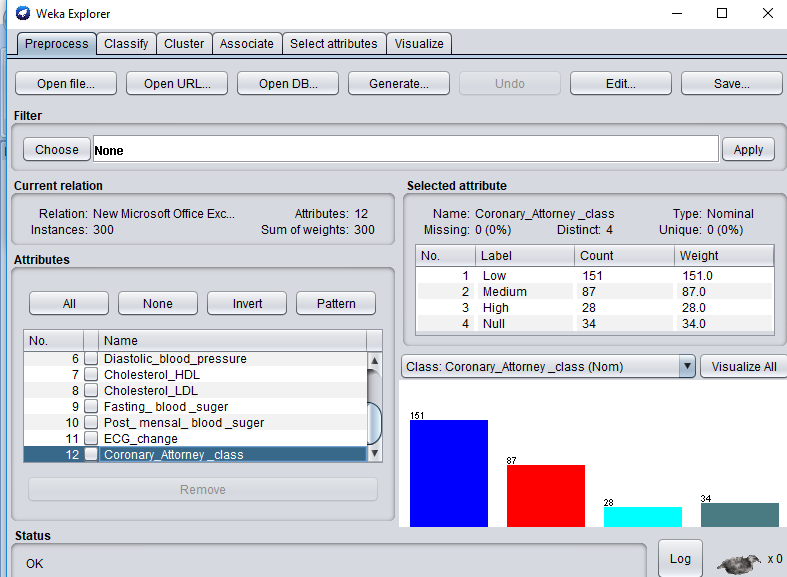
The Graphical view of Predicted result of Coronary Attorney after applying j48 algorithm.



**Fig 5.8: Tree visualization on prediction of coronary attorney after applying algorithm.**

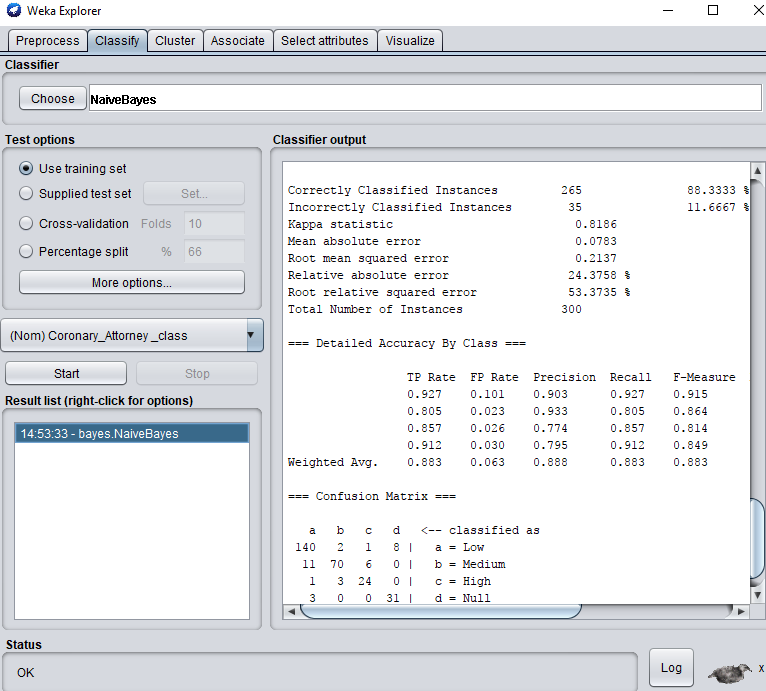
**5.3 Data preparing for classification model**

Authors have used a dataset of 300 data as training and 200 data as test dataset to classify the probability level of Coronary Attorney. In classification, authors have used 10 attributes for both training and test data set. Age and test reports would be present on this 10 attributes. Authors have decided to classify the result into four categories and there were High, Medium, Low and Null. This would explain the probability level of coronary attorney of a patient. In coronary attorney Age, LDL, Systolic blood pressure, and Diabetes are the most important health condition. If patients age is less than or equal to 20 then the result would be Null there is no necessity to observe anything. If age is over 20 then it will check the state of patient’s diabetes. Here authors have fixed two different ways to classify level. First way for those patients who are over 20 and have diabetes too .If the Cholesterol (LDL) range within160-190 mg/dl then the classification result would be High, if Cholesterol (LDL) range within130-159 mg/dl then the classification result would be Medium, if Cholesterol (LDL) range within100-129 mg/dl then the classification result would be Low and if Cholesterol (LDL) <100 mg/dl then the classification result would be Null. Second way for those patients who are over 20 but don’t have diabetes. If Cholesterol (LDL) range within100-129 mg/dl then the classification result would be Null, if Cholesterol (LDL) range within130-159 mg/dl then the classification result would be Low , if Cholesterol (LDL) range within160-189 mg/dl then the classification result would be Medium and if Cholesterol (LDL) 190=> mg/dl then the classification result would be High.



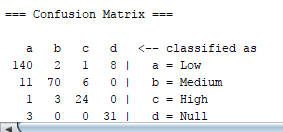
**Fig 5.9: Load data into WEKA for classification.**

Fig 5.9 is showing that the dataset of patients test report’s have inserted into the WEKA software for classification.



**Fig 5.10: Training data results gained by Naive Bayes**

Authors have inserted training dataset and applied Naive Bayes on those data .Fig 5.10 is expressing the result of inserted data after applyingNaive Bayes algorithm.



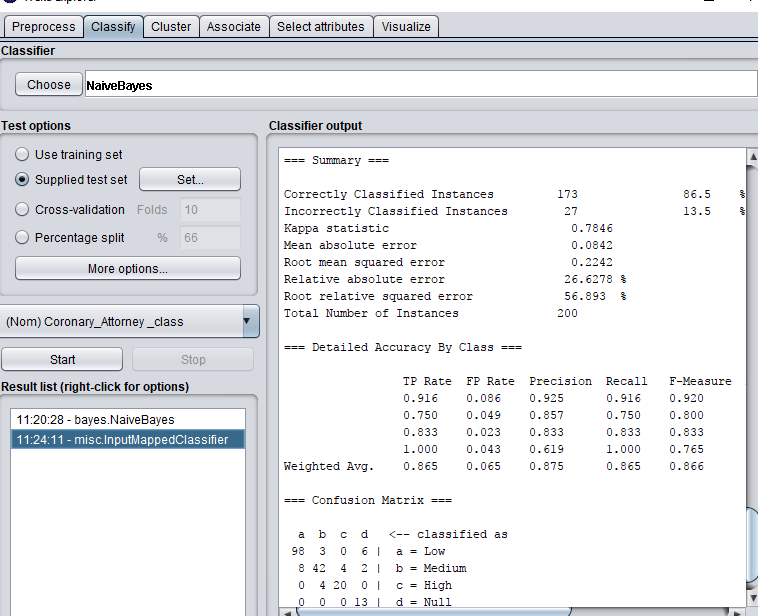
**Fig 5.10.1: Confusion Matrix of Training dataset gained by Naive Bayes**

Fig 5.10.1 is expressing that correctly classified Low instances are 140, correctly classified Medium instances are 70, correctly classified High instances are 24 and correctly classified Null instances are 31 of that confusion matrix.

|  |  |
| --- | --- |
| Classifying the level of coronary attorney | Total data:300 |
| Null | 31 |
| Low | 140 |
| Medium | 70 |
| High | 24 |

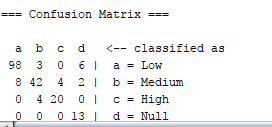
**Fig 5.11: Graphical view of Training data results gained by Naïve Bayes**

Correctly classified instances of training dataset are 88% and incorrectly classified instances are 11%. Here fig 5.11 is expressing the probability level of Coronary Attorney and 12% instances have the Null probability, 53% instances have the Low probability, 26% instances have the Medium probability and 9% instances have the High probability of having coronary attorney.



**Fig 5.12: Test data results gained by Naive Bayes**

Authors have inserted test dataset and appliedNaive Bayes algorithm on those data .Fig 5.12 is expressing the result of inserted data after applying Naive Bayes algorithm. Here authors have used 200 data and got 86.5% accuracy and there correctly classified instances are 173 and incorrectly classified instances are 27.This fig also expressing the classified result by confusion matrix.



**Fig 5.12.1: Confusion Matrix of Test data results gained by Naïve Bayes.**

Fig 5.12.1 is expressing that correctly classified Low instances are 98, correctly classified Medium instances are 42, correctly classified High instances are 20 and correctly classified Null instances are 13 of that confusion matrix.

|  |  |
| --- | --- |
| Classifying the level of coronary attorney | Total data:200 |
| Null | 13 |
| Low | 98 |
| Medium | 42 |
| High | 20 |

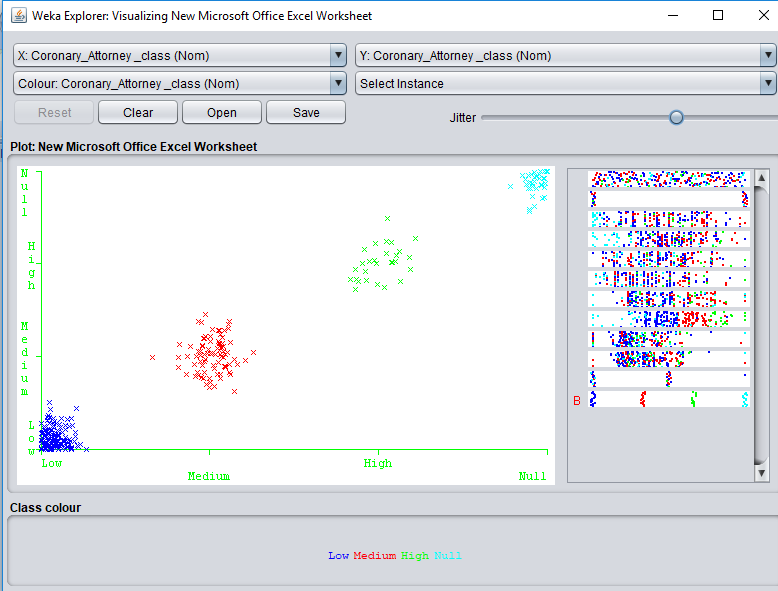
**Fig 5.13: Graphical view of Test data result gained by Naïve Bayes**

Here correctly classified instances of test dataset are 86.5% and incorrectly classified instances are 13.5%. Here fig 5.13 is expressing the probability level of Coronary Attorney and 7% instances have the Null probability, 57% instances have the Low probability, 24% instances have the Medium probability and 7% instances have the High probability of having coronary attorney.

|  |  |
| --- | --- |
| Result | Accuracy rate |
| Correct accuracy | 86.5% |
| Incorrect accuracy | 13.5% |

**Fig 5.14: Accuracy level of Coronary Attorney gained by Naïve Bayes.**

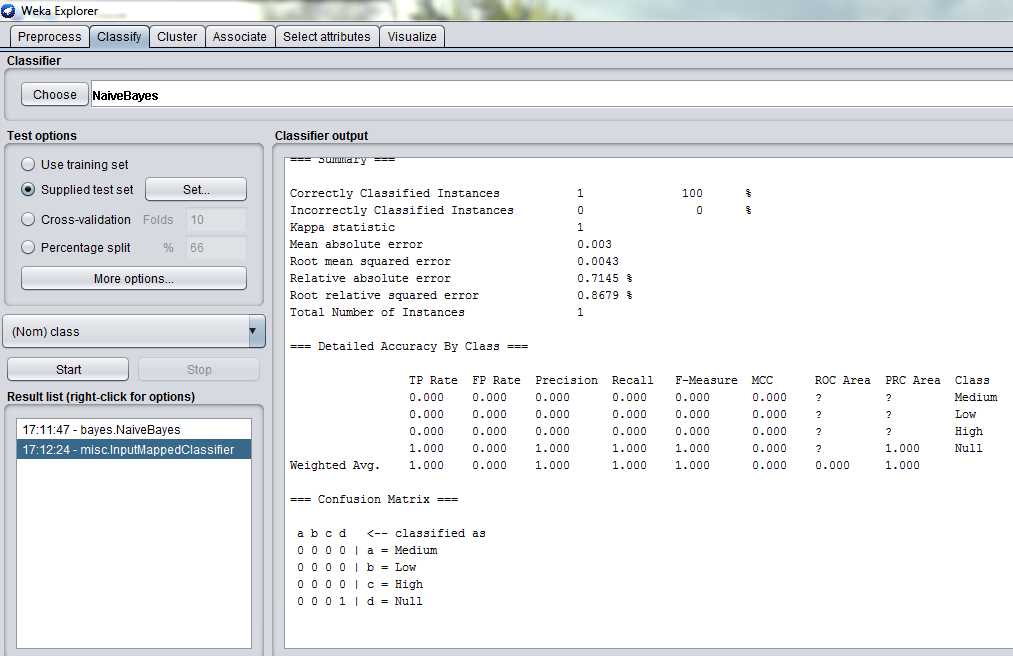
The Graphical view of Prediction model gained by Naïve Bayes. Fig 5.14 is expressing 87% accuracy of classification model of coronary attorney.



**Fig 5.15: Classification of Coronary Attorney after applying technique.**

The Graphical view of Classification level of Coronary Attorney after applying Naïve Bayes.

**5.4 Testing result of Classification Model**

****

**Fig 5.16: Single patient test data result for classifying the level of Coronary Attorney**

The correctly classified instance of single patient test report was 100% .This model can perform perfectly with single data.

**CHAPTER 6 CONCLUSIONS**

**6.1 Conclusions**

This research work has introduced a way to predict and classify coronary attorney of medical patients by creating predictive and classification model .Researchers have collected real data of heart patients from Medicals on the basis of preselected attributes and used those data for creating both predictive and classification model .This model is capable to predict the probability of Coronary Attorney of human body and which was the main purpose of this research work. Authors used two data mining techniques, j48 and Naïve Bayes in this research.J48 used for design a predictive model which shows the probability of coronary attorney of a patient and Naïve Bayes for classification model which show probability level of coronary attorney of a patient and both of the model help user with appropriate outcomes .

**6.2 Limitations**

1. In this research authors have worked with lots of test reports, physical condition and habit of a patent but here patients will not be able to know the exact reason which was responsible for coronary attorney.
2. Here authors also didn’t create any sector in this system where patients could know about the treatment or guideline of each probable stage of coronary attorney.

**6.1 Future Work**

In future authors will upgrade their system by adding multiple features. The system will help patient to know about the actual reason of coronary attorney in his body and the system will also suggest the treatment procedure, medicine and guideline on the basis of each probable stage of coronary attorney.

**References**

Abhishek Taneja, “Heart Disease prediction system using data mining techniques”, An International Open Free Access and Peer Reviewed Research Journal, Vol.3, No.4,December 2013.

C.Sowmiya , Dr.P. Sumitra ,“Comparative Study of Predicting Heart Disease By Means of Data Mining”,International Journal Of Engineering And Computer Science ISSN: 2319-7242 Volume 5 Issue 12 Dec. 2016.

Hlaudi Daniel Masethe & Mosima Anna Masethe , “Prediction of Heart Disease using Classification Algorithms”,World Congress on Engineering and Computer Science(WCECS)22-24October,2014,San Francisco,USA.

Jung Gi Yang, Hyun Bok Choi, Jung Tae Kim, Mi Hee Jang, Un Gu Kang and Young Ho Lee, “A Study of Cardiovascular Disease Prediction Model using Discriminant analysis”, 191, Hambakmoero, Yeonsu-gu, Incheon, South Korea,2013.

Joyti Soni,Ujma Ansari,Dipesh Sarma,Sunita Soni, “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction ”,International journal of computer allpication,Vol 7,No. 8,March 2011.

Monower MM, Flora MS, Akhter A, Akter K, Choudhury SR, “Framingham Risk Assessment of Coronary Heart Diseases in Selected Area of Dhaka City: A Cross-Sectional Study”,JNHFB, Jan 2018.

Noreen Akhtar, Muhammad Ramzan Talib, Nosheen Kanwa l, “Data Mining Techniques to Construct a Model: Cardiac Diseases”, International Journal of Advanced Computer Science and Applications, Vol. 9, No.1, 2018, Faisalabad, Pakistan.

S. Kiruthika Devi\*, S. Krishnapriya and Dristipona Kalita, “Prediction of heart disease using data mining techniques”, Indian journal of Science and technology,Vlo9(39),October 2013,TamilNadu,India.

Sambasiva Rao Voleti, Kiran Kumar Reddi, “Design of an Optimal Method for Disease Prediction using Data Mining Techniques”,International Journal of Advanced Research in Computer Science and Software Engineering,Vol.6,Issue 12,December 2016.

[Johann Gamper](http://www.inf.unibz.it/~gamper/) and [Mouna Kacimi](http://www.inf.unibz.it/~mkacimi/), 1st semester 2012, Data warehousing and Data Mining, (DWDM),Available from: http://www.inf.unibz.it/dis/teaching/DWDM/slides2012/lesson10-Classification2.pdf

Pang-Ning, Tan, Michael Steinbach & Vipin Kumar, 2006, “Introduction to data mining”, published by Dorling Kindersley (India) Pvt.Ltd. Available from: http://www.uokufa.edu.iq/staff/ehsanali/Tan.pdf

|  |  |
| --- | --- |
|  |  |