International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 5, ISSN No. 2455-2143, Pages 512-525
Published Online September 2019 in IJEAST (http://www.ijeast.com)

ANALYSIS, DESIGN AND DEVELOPMENT OF A PREDICTIVE MODEL TO PREDICT DIABETES OF HUMAN BODY USING DATA MINING ALGORITHM

Afzal Hossain
Department of Computer Science & Engineering
World University of Bangladesh, Dhaka, Bangladesh

Abstract - **Data mining is one of the most attractive interdisciplinary subfield of computer science. Data mining is one of the technique in which it can extract the useful information from the large data set. Then it transforms the data as human understandable format. Nowadays there are a large number of people affected by diabetic. So that analyzing the patient report using manual system is too much complicated to the doctor. The aim of the study is to make a model which can predict the diabetic and measure the level of diabetes of human body. To accomplish the research work the author have used J48 and Naive Bayes algorithm. The aim of research work is to work with maximum number of attributes to find out the appropriate outcomes which can predict the diabetes and then measure the level of diabetes of a human body. The authors have used several steps to reach the goal of the research. They proposed a proposal model which is the combination form of two systems which can predict the diabetes and after completing the prediction phase it can also measure the level of diabetes. Authors create a prediction model using J48 algorithm and a classification model using Naive Bayes algorithm. The prediction model which can predict the result of diabetes is yes or not by given the sample data if the result is yes then authors provide sample data to the classification model for classifying the level of diabetes into four types. After implementing the system the authors are tested their research work by using several sample data. According to the sample data they have found that their working result for prediction the diabetes is 96% and measuring the level of diabetes is 89%.**

Keywords – Data mining; Diabetes prediction; Classification; **J48 algorithm; Naive Bayes**

I. INTRODUCTION

Data mining is one of the most striking interdisciplinary subfield of computer science. Mukesh kumari, Dr. et al. (2014) said that Data mining must also be considered a process that requires goals and objectives to be specified. Data mining is delineated as the process of devising correlations, patterns and aptitude to search through the outsized amount of data stored in repositories, databases, and data warehouses. The overall goal of the data mining technique is to essence the useful information from the hefty data set and to renovate it into a comprehensible format. So that it can be accustomed for the future use. Two models are accustomed in data Mining Technologies. These are defined as Predictive model and classification model.

Diabetes is a group of metabolic disease. It happens when a body is not able to react or outgrowth to insulin, which is needed to maintain the rate of glucose. Diabetes conducts to other disease such as blindness, nerve damage, kidney disease, blood pressure, heart disease. It befall for hereditary and genetics factors, Infections caused by viruses, stress, obesity, increased cholesterol level, high carbohydrate diet, nutritional deficiency, excess intake of oil and sugar, no physical exercise, overeating, tension and worries, high blood pressure, insulin deficiency, insulin resistance. Vrushali Balpande et al. (2017) emphasize that this can be controlled by doing diet, exercise and make use of appropriate medications. Diabetes Mellitus is one of the major public health problems. There are two types of diabetes. Those are

1. Type 1 Diabetes
2. Type 2 Diabetes

Type 1 Diabetes
It usually starts in childhood or young adulthood. The body's immune system destroys the cells that release insulin, eventually eliminating insulin production from the body. Without insulin, cells cannot absorb sugar (glucose), which they need to produce energy.

Type 2 Diabetes
This type of Diabetes called adult onset. The human body does not produce sufficient insulin for proper function in the body. This type of Diabetes is developed at the age of 40.This can be prevented by earlier identification and regular exercise.

In this research paper, the authors mainly focused on prediction of diabetes and measuring the level of diabetes. To achieve the goal they have used two algorithms. Those are J48 and Naive Bayes algorithm. J48 algorithm was used to make a predictive model which can ensure that diabetes present on human body or not. Naive Bayes algorithm was used to make a classification model which can ensure that the level of diabetes.

II. LITERATURE REVIEW

In This Paper the Bayesian Network is used to predict diabetes. Classifier was applied to the modified dataset to construct the Bayesian Network model. WEKA Tool was used for Data Analysis. Mukesh kumari, Dr. Rajan Vohra et al. (2014) focused on Ccassification with Bayesian network shows the best accuracy is 99.51 percent and error in the classification is .48 percent when the results were compared to clinical diagnosis. The mean absolute error (MEA) =.0053 and root mean squared error (MRES =.0596).

Krishnaveni G et al. (2017) emphasize their paper presents the forecast of diabetes using classification technique. Here PIMA Indian diabetes dataset are used in this technique. This paper has proposed Naive Bayes algorithm which is more productive than other classifiers. The result showed that the Naive Bayes has the highest accuracy in the classification accuracy is 76.16 percent. The K Nearest Neighbor algorithm has lowest accuracy in the classifier is 71.13 percent. Matlab are used for data implementation.

Priyanka S.R et al. (2016) they focused on their paper the ID3 algorithm is used to generate decision tree from the data set. This algorithm creates a model that predicts the value of a target variables based on several input variables. Based on the category the proposed system can send a message through mail whether they are diabetic or not. The dataset includes 150 records and 14 attributes with one class attribute. The correctly classified instance is 94 percent and incorrectly classified is 6 percent for ID3 algorithm. The results obtained as sensitivity with 55 percent, specificity with 22 percent, and accuracy with 94 percent and error rate with 6 percent.

Saravananathan K et al. (2016) presents the diabetes data set which is used for this research work. This data set has 10 attributes. For the classification J48, CART, SVM, and KNN technique is accepted in this work. Totally, 545 patients' data is collected. In which, there are 366 male and 179 female patients whose age between 40 and 60 years. This research work mainly discusses about the accuracy of classification algorithms compared with the execution time and error rate using WEKA software. The accuracy of J48 method found to be 67.16 percent, CART is 62.29 percent, and Support Vector Machines is 65.05 percent and KNN 53.39 percent. The results show that the performance of J48 technique is significantly superior to the other three techniques for the classification of diabetes data.

Rajesh K et al. (2012) proposed that data mining relationship for efficient classification and applied data mining techniques to classify diabetes clinical data and predict the patient being affected with diabetes or not. They presented a system which gave training data on that data feature relevance analysis is done then comparison of classification algorithm, selecting classifier then improved classification algorithm is applied and then found out the evaluation that compared with training data. In their work Tthey applied C4.5 Algorithm gave classification rate of 91%.

Sathees Kumar B et al. (2014) present three types of decision tree algorithm are used such as Simple Cart, J48, and Naïve Bayes which are applied on type II diabetes. WEKA has been used to build the prediction model for decision tree. Type 2 diabetes is analysis in this paper. The data set consists of 10 attributes that are used to predict the type II diabetes. J48 is applied on type 2 diabetes data set in weka. Time taken to build the decision tree for J48 is .02 seconds where simple cart has .03 seconds and Naive Bayes has .06 seconds. The pre-processing was used to improve the quality of data. From the result the J48 has 93.5 percent accuracy rate.

Gaganjot Kaur et al. (2014) discussed prediction of diabetes in data mining. In every age group disease is common. A diabetes person has risk of having the other diseases as foot damage, nerve damage, heart diseases, kidney diseases etc. Attributes are involved the diabetes. Diabetes is generally 2 types: 1(insulin-dependent diabetes) 2(non-insulin-dependent diabetes).The international diabetes federation has claimed that presently 246 million people are suffering from diabetes worldwide and this number is expected to increase up to 380 million by 2025.

III. EXPERIMENTAL METHODOLOGY

A. *Algorithms* -J48

Let D_t be the set of training records that are associated with node t and $y = \{y_1, y_2, \cdots, y_c\}$ be the class labels. The following is a recursive definition of j48 algorithm.

- Step 1: If all the records in D_t belong to the same class y_t, then t is a leaf node labeled as y_t.
- Step 2: If D_t contains records that belong to more than one class, an attribute test condition is used to partition the records into smaller subsets. A child node is then created for each outcome of the test condition. The records in D_t are distributed to the children based upon their outcomes. This procedure is repeated for each child node.

B. Naive Bayes

- Each Tuple is an 'n' dimensional attribute vector

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 5, ISSN No. 2455-2143, Pages 512-525
Published Online September 2019 in IJEAST (http://www.ijeast.com)

- X : (x1,x2,x3,…. xn) Let there be 'm' Classes : C1,C2,C3…Cm Naive Bayes classifier predicts X belongs to Class Ci iff
- P (Ci/X) > P(Cj/X) for 1<= j <= m , j <>i Maximum Posteriori Hypothesis
- P(Ci/X) = P(X/Ci) P(Ci) / P(X) Maximize P(X/Ci) P(Ci) as P(X) is constant With many attributes, it is computationally expensive to evaluate P(X/Ci). Naïve Assumption of "class conditional independence" Here, P(X/Ci).P(Ci)=∏ P(Xk/Ci).P(Ci) & ∏ P(Xk/Ci) = P(x1/Ci) * P(x2/Ci) *…* P(xn/ Ci) .

B. *Working Flow Diagram*

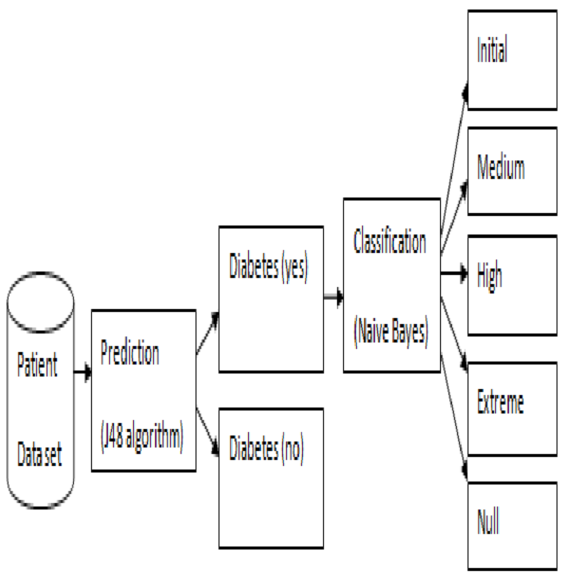

Fig.1. Flow chart of the system

In this design phase the authors have designed their research work in several phases. At first they were starting then they were affording the training data and then process these data then generate a predictive model. After generating a predictive model they were affording the test data. After these test data the authors ensure that diabetes yes or no and when they unearth the yes then they assess the classification model then they were taking a patient test report. According to this patient test report they have found that how many reports are successfully enumerated according to this post prandial blood sugar level. If the post prandial blood sugar is greater than 7.8 mmol/l then it is categorized as initial level and 12 – 15 mmol/l is medium level and 16 – 33 mmol/l is high level and greater than 33 mmol/l is extreme level and less than 7.8mmol/l is null.

C. System Design

The following system design is working in the system

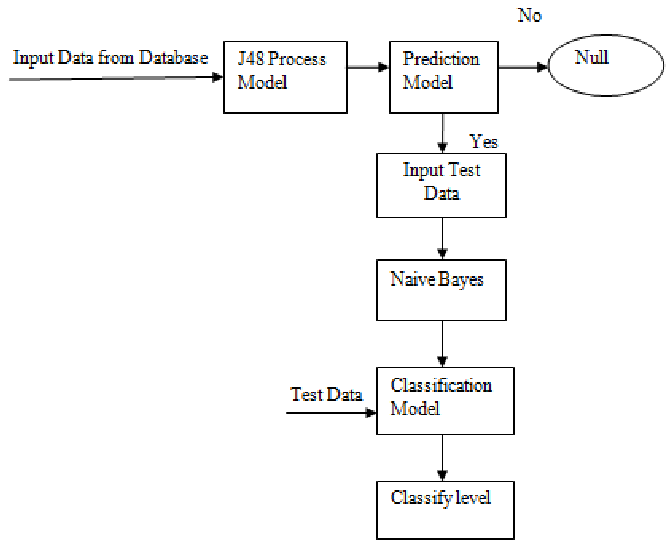

Fig.2. System Design for Diabetes Prediction and Classification Level

Input Data
The authors provide the sample data of patient data as input from database in the J48 process model. They have selected the following attribute for supplying the input data in the system.

J48 Process Model
In this stage this algorithm can analyze the data and create a decision tree for finding the leaf node. Each leaf node is assigned class level. This data are training set and we can train up the machine. Decontamination data and remove noise and isolated observation are happened in preprocessing.

Predictive Model
A predictive model is dressed by J48 for prediction a result. In this stage we must provide the test data. By this model we can assure that patient has diabetes or not. If a person has no diabetes then it goes to End state. If a person has diabetes then it goes to yes state.

Naive Bayes
Deliver the input data in Naive Bayes model which are patient test data for classification and this data is provided as training data. Naive Bayes can apply a technique on attribute set and class variable for creating a probabilistic relationship model.

Classification model
In this stage Naïve Bayes can create a classification model for measuring the level of diabetes. We provide the test data in

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 5, ISSN No. 2455-2143, Pages 512-525
Published Online September 2019 in IJEAST (http://www.ijeast.com)

this model and also provide a single patient data for testing the system.

Classify level

It can assure us about the level of diabetes. Which are initial level, medium level, high level, extreme level and null.

D. *Proposed Model*

Proposed model is described below

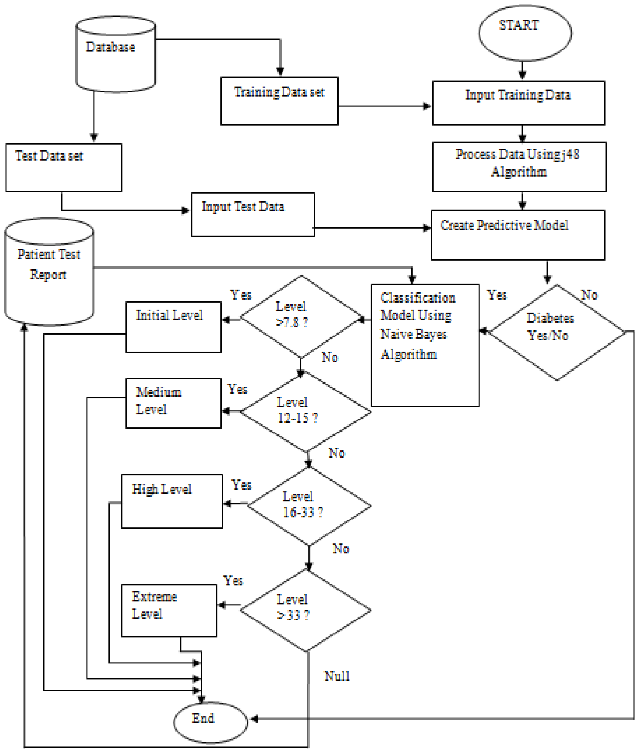

Fig.3. Proposed Model of Diabetes Prediction and Classification System

Authors are gone through several research papers and they proposed a proposal model which is the combination form of two systems which can predict the diabetes and after completing the prediction phase it can also measure the level of diabetes. They divided data of database in two parts which one is training dataset and another one is test data set. Authors input the training data into the system for processing data using J48 algorithm and create a prediction model. After making the predictive model they input test data into the system and the system confirms that the prediction of diabetes is yes or no. If the result is no then the system stop its processing but if the result is yes then its goes to the classification model which is made by Naive Bayes algorithm and classifying the level of diabetes. The authors give patients test reports as input data to identify the level of diabetic. If the level is greater than 7.8 mmol/l than its Initial level, if the level is greater than 12mmol/l and less than 15 mmol/l its Medium level, if the level is greater than 16mmol/l and less than 33 mmol/l its High level, greater than 33 mmol/l is Extreme level.

E. *Dataset*

TABLE1. Attributes set for prediction the Diabetes

S. No .	Attribute Name	Attribute Description	Attributes Values
1.	Patient Id	Patient Id Number	Numeric
2.	Patient Name	Name of patient	Ex: Mr. X
3.	Age	Age in Year	Newborn-1,2-5 ,6-10,10-15,16-20,21-40,41-60,61-80,>81
4.	Weight	Weight in kg's	4-10 kg,12–18 kg, 20 –32 kg,32 – 50 kg,51-60kg,61-70kg,71 80kg,>81kg
5.	Gender	Patient Male or female	M/F
6.	Physical Activity	Physical activity in a days	4-5 h, 6–7 h, 7–8 h, >8 h
7.	Urination	Number of times urination in a day	3-4 times, 5-6 times, 7-8 times, >8 times
8.	Water Consumption	Water consumption in a day	3-4 liters, 5-6 liters, 7-8 liters, >8 liters
9.	Hyper Tension	Person with hypertension	(Yes / No)
10.	Systolic blood pressure	Enter value of blood pressure upper limit in mmHg	Ideal-(80 mmHg to 120 mmHg) Pre-hypertension-(120.0 mmHg to 139.9 mmHg) Hypertension1-(140.0 mmHg to

			159.9 mmHg) Hypertension2-(160.0 mmHg and above)
11.	Diastolic Blood pressure	Enter value of blood pressure lower limit in mmHg	Ideal-(60 mmHg to 80 mmHg) Pre-hypertension-(80 mmHg to 89 mmHg) Hypertension1- (90 mmHg to 99 mmHg) Hypertension2-100 or above
12.	GTT (Glucose tolerance test)	The test is performed after consuming a concentrated amount of glucose dissolved in water.	Normal Range of GTT is 7.8 mmol/l
13.	High density Lipoprotein (HDL)	HDL (good) cholesterol– HDL helps remove cholesterol from your arteries	Normal Range of HDL is between (40-60 mg/dl)
14.	Low density Lipoprotein (LDL)	LDL (bad) cholesterol– the main source of cholesterol buildup and blockage in the arteries	Normal Range of LDL is less than 130 mg/dl
15.	Blurred Vision	Have blurred vision	(Yes / No)
16.	Wound healing	Number of the wound starts to dry	(Yes / No)
17.	Sleepy / drowsy	Always feel sleepy/drowsy	(Yes / No)
18.	Sudden Weight loss	Observe sudden weight loss	(Yes / No)
19.	Heredity	Elders found with diabetes	(Yes / No)
20.	Tiredness	Feel tiredness	(Yes / No)
21.	Eye Damage	Eye blindness problem	(Yes / No)
22.	Nerve Damage (neuropathy)	Loss of feeling in the hands, arms, feet, and legs	(Yes / No)
23.	Foot Damage	Numbness or reduced ability to feel pain or temperature change	(Yes / No)
24.	Pregnancy Complication	Blood sugar control	(Yes / No)
25.	Extreme Hunger	If the people are always feel hungry	(Yes / No)
26.	FBS (Fasting blood sugar)	The blood test has to be performed after twelve hours of fasting. That means the person must refrain from eating approximately 12 hours prior to this blood test.	Normal Range of FBS is (3.6-5.6 mmol/l)
27.	PPBS (Post Prandial Blood Sugar)	Prior to this test, the person fasts overnight and then consumes a carbohydrate	Normal Range of PPBS is 7.8 mmol/l

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 5, ISSN No. 2455-2143, Pages 512-525
Published Online September 2019 in IJEAST (http://www.ijeast.com)

		meal. Approximately between $1^1/_2$to 2 hours after eating, a blood sample from vein is drawn for testing.	
28	Diabetes Prediction	Person has Diabetes or not	(Yes / No)

The authors provide the sample data of patient data as input from database in the J48 process model. They have selected these attribute for supplying the input data in the system.

TABLE 2. Attribute set for Classification the level of Diabetes

S. No.	Testing Attributes Name	Attributes description	Attributes Value
1.	Gender	Patient Male or female	M/F
2.	Age	Age in Year	Newborn-1,2-5 ,6-10,10-15, 16-20,21-40, 41-60, 61-80,>81
3.	FBS (Fasting blood sugar)	The blood test has to be performed after twelve hours of fasting. That means the person must refrain from eating approximately 12 hours prior to this blood test.	Normal Range of FBS is (3.6-5.6 mmol/l)
4.	PPBS (Post Prandial Blood Sugar)	Prior to this test, the person fasts overnight and then consumes a carbohydrate meal. Approximately between $1^1/_2$to 2 hours after eating, a blood sample from vein is drawn for testing.	Normal Range of PPBS is 7.8 mmol/l
5.	GTT (Glucose tolerance test)	The test is performed after consuming a concentrated amount of glucose dissolved in water.	Normal Range of GTT is 7.8 mmol/l
6.	Systolic blood pressure	Enter value of blood pressure upper limit in mmHg	Ideal-(80 mmHg to 120 mmHg) Pre-hypertension-(120.0 mmHg to 139.9 mmHg) Hypertension1-(140.0 mmHg to 159.9 mmHg) Hypertension2-(160.0 mmHg and above)
7.	Diastolic Blood pressure	Enter value of blood pressure lower limit in mmHg	Ideal-(60 mmHg to 80 mmHg) Pre-hypertension-(80 mmHg to 89 mmHg) Hypertension1-(90 mmHg to 99 mmHg) Hypertension2-100 or above
8.	High density Lipoprotein (HDL)	HDL (good) cholesterol–HDL helps remove cholesterol from your arteries	Normal Range of HDL is between (40-60 mg/dl)
9.	Low density Lipoprotein (LDL)	LDL (bad) cholesterol–the main source of cholesterol buildup and blockage in the arteries	Normal Range of LDL is less than 130 mg/dl
10.	Diabetes class	Which type of diabetes he/she has	Initial level, Medium level, High level, Extreme level, null

The authors provide the sample data of patient data as input from database in the Classification model.

IV. RESULTS ANALYSIS

The dataset that is taken for this research work contains 500 records and 28 attributes for the purpose of predicting whether

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 5, ISSN No. 2455-2143, Pages 512-525
Published Online September 2019 in IJEAST (http://www.ijeast.com)

a person is diabetic or non-diabetic based on the symptoms. Data set are store in Microsoft Excel 2010 format.

A. Preparing data for prediction model

This is the sample of dataset used for prediction Diabetes in human body. The dataset contains 500 instances. Prediction Model has 28 attributes and provides this data as input. Here our first work is about predicting diabetes whether a person has diabetes or non-diabetes in a dataset by applying J48 algorithm. The dataset variables which are used for prediction of diabetes are Systolic blood, Diastolic Blood pressure, GTT (Glucose tolerance test),Nerve Damage (neuropathy),FBS (Fasting blood sugar),PPBS (Post Prandial Blood Sugar) etc. If the value of fasting plasma glucose is less than 5.6 mmol/l and value of Post Prandial Blood Sugar test is less than 7.8 mmol/lthan it will be given score 0, means a person is non-diabetes If the value of fasting plasma glucose lies in the range of 3.6-5.6 mmol/l and value of casual glucose tolerance test lies in the range of 7.8 mmol/l than it will be given score 1, means a person has diabetes.

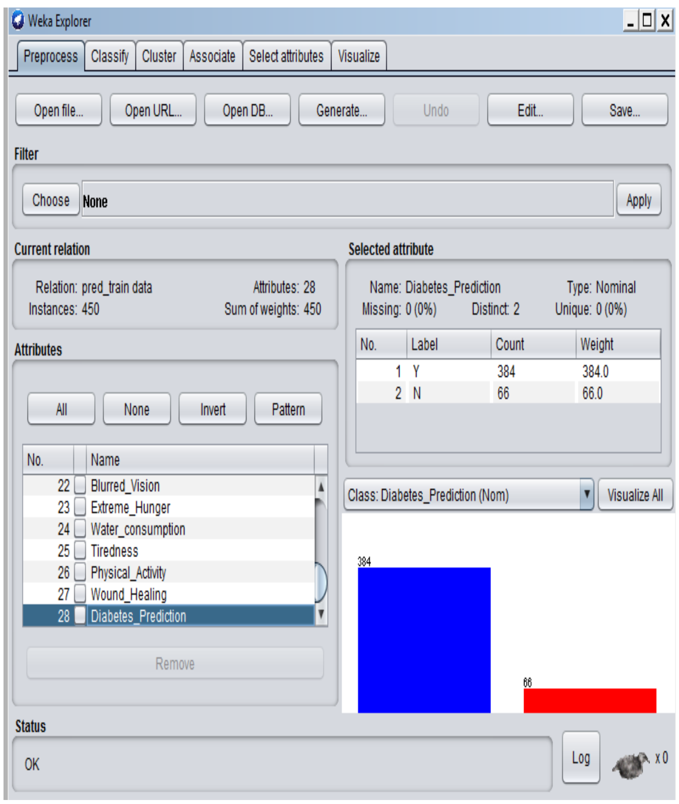

Fig.4. Representing data load into WEKA.

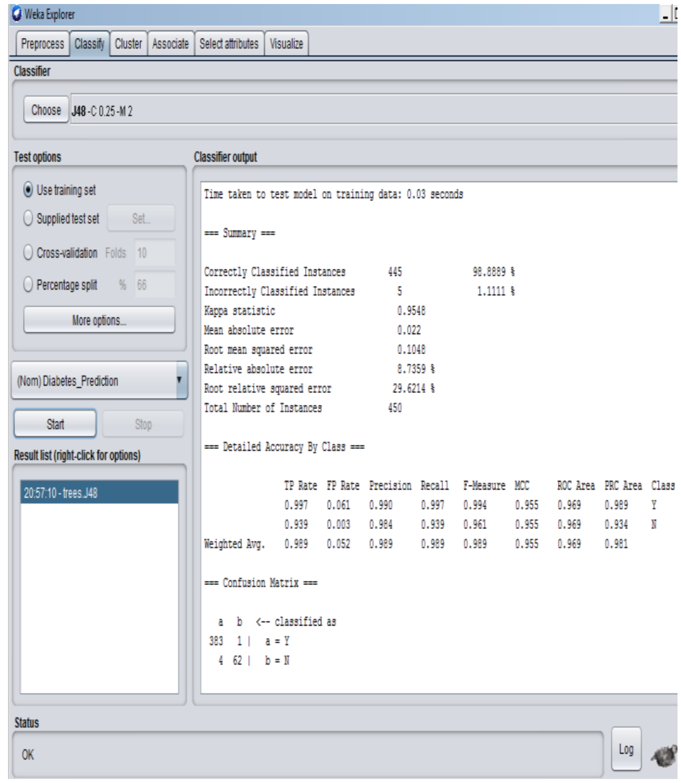

Fig.5. Training data results obtained using J48.

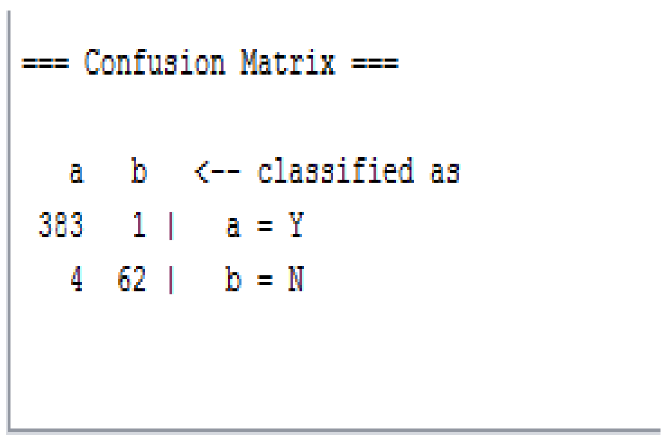

Fig.6. Confusion Matrix of Training data results obtained using J48.

TABLE 3.Training data results obtained using J48.

Prediction result	Total Data : 450
Diabetes : Yes	383
Diabetes : No	62

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 5, ISSN No. 2455-2143, Pages 512-525
Published Online September 2019 in IJEAST (http://www.ijeast.com)

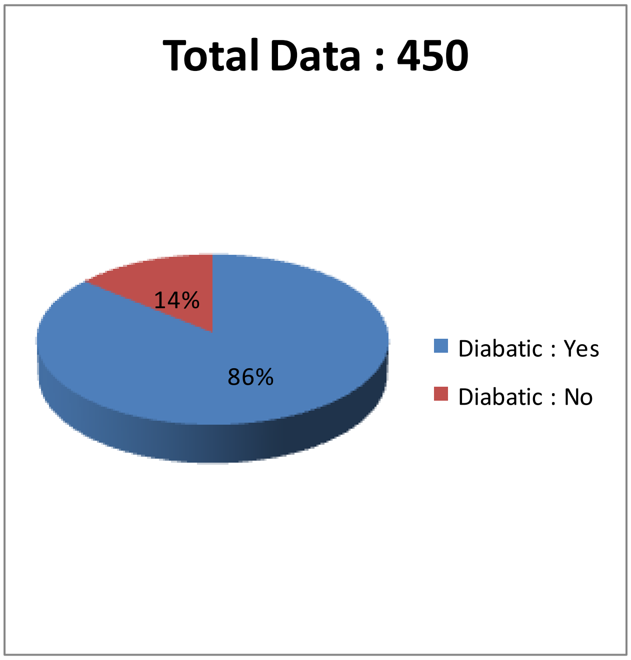

Fig.7. Graphical view of Training data results obtained using J48

Correctly Classified Instances of training data set 98.8%.and Incorrectly Classified Instances are 1.1%. Dataset are trained properly for prediction model.

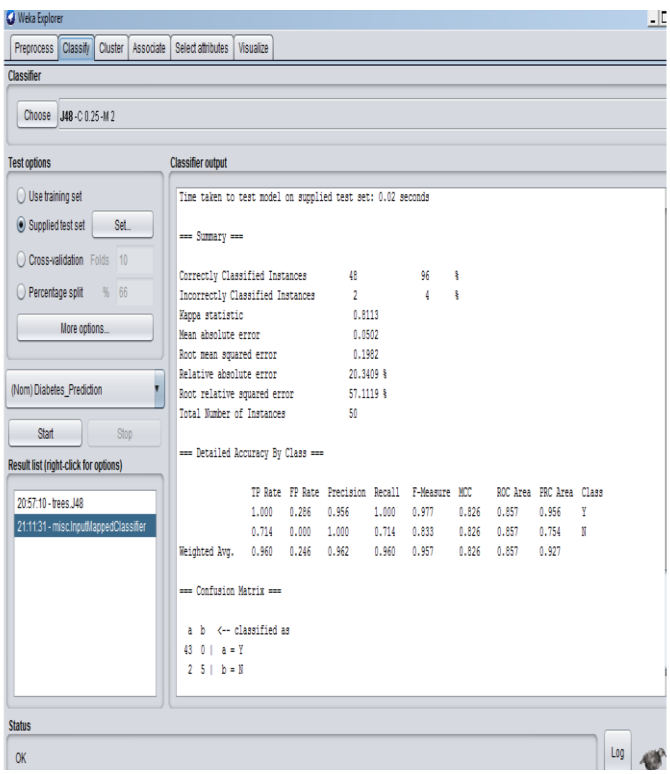

Fig.8. Test data results obtained using J48.

```
=== Confusion Matrix ===

  a  b   <-- classified as
 43  0 |  a = Y
  2  5 |  b = N
```

Fig.9. Confusion Matrix of Test data results obtained using J48.

TABLE 4.Test data results obtained using J48.

Prediction result	Total Data : 50
Diabetes : Yes	43
Diabetes : No	5

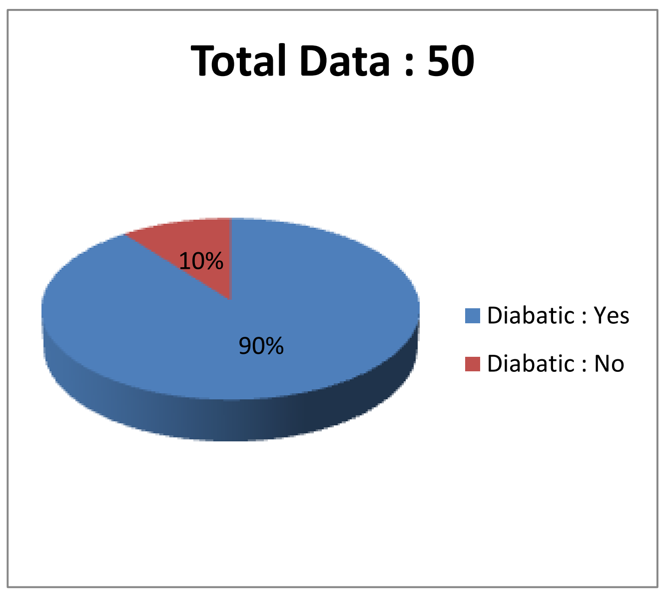

Fig.10. Graphical view of Test data results obtained using J48

Now Correctly Classified Instances of test data set 96%.and Incorrectly Classified Instances are 4%.

TABLE 5. Accuracy of Diabetes prediction using J48.

Result	Accuracy rate
Correct Accuracy	96%
Incorrect Accuracy	4%

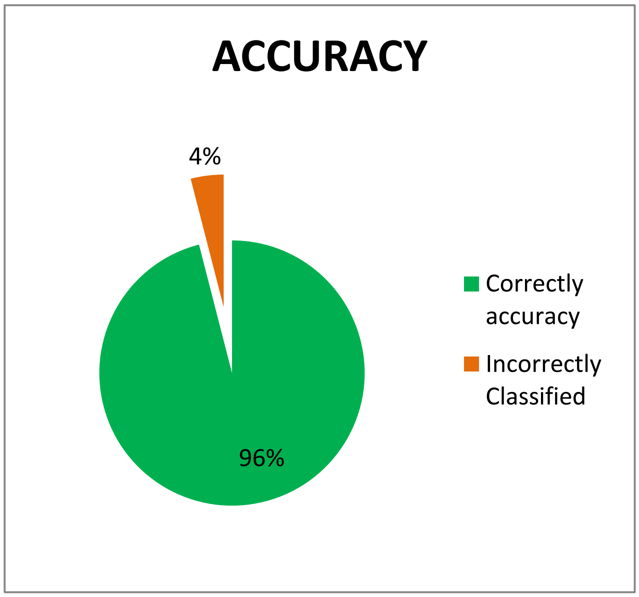

Fig.11. Accuracy of Diabetes prediction using J48.

The Graphical view of prediction model after using J48 algorithm

B. Confusion matrix

A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier [8].

TABLE 6. Confusion Matrix

Actual classes \ Predicted classes	P	N
P	TP	FP
N	FN	TN

• True positive (TP)- These are the positive tuples that were correctly labeled by the classifier. If the outcome from a prediction is p and the actual value is also p, then it is called a true positive (TP).

• True Negative (TN)-These are the negative tuples that were correctly labeled by the classifier.

• False Positive (FP)-These are the negative tuples that were incorrectly labeled as positive. However if the actual value is n then it is said to be a false positive (FP).

• False Negative (FN)-These are the positive tuples that were mislabeled as negative. Accuracy is calculated as (TP+TN)/(P+N) where, P=TP+FN and N=FP+TN. Or TP+TN/(TOTAL) [8]. According to experimental results, correctly classified instances for J48 is 48. Accuracy of J48 is 96%.

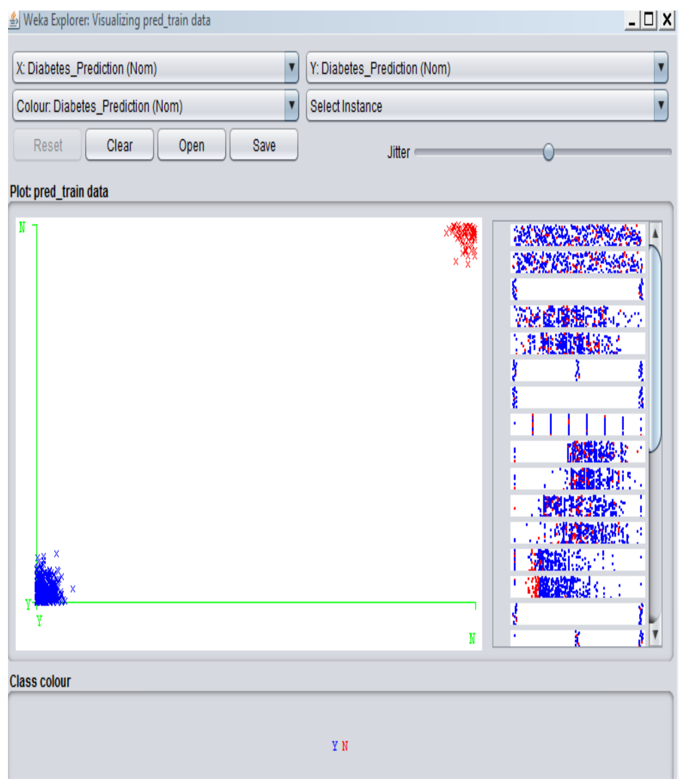

Fig.12. Prediction of diabetes shown after applying technique.

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 5, ISSN No. 2455-2143, Pages 512-525
Published Online September 2019 in IJEAST (http://www.ijeast.com)

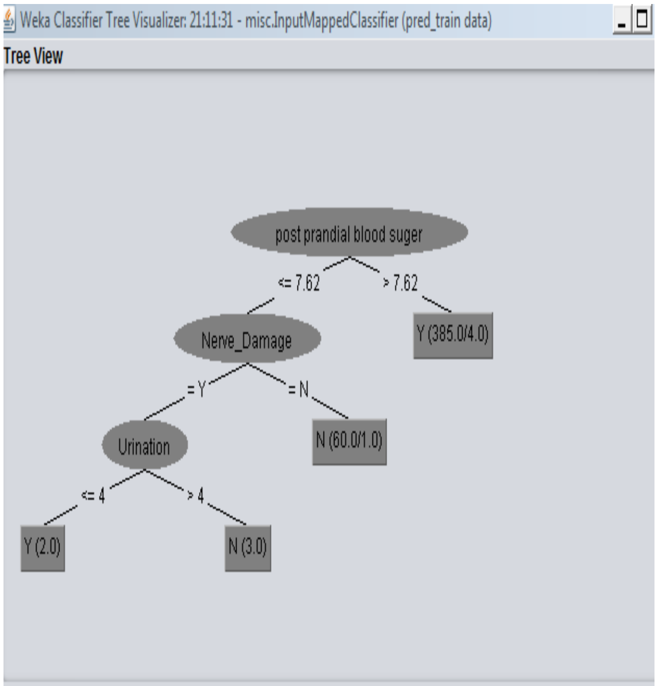

Fig.13. Tree Visualization on Prediction of diabetes shown after applying technique.

C. Preparing data for Classification model

This is the sample of dataset used for Classification the level of Diabetes in human body. The dataset used contains 500 instances. All instances have 10 input attribute's five output attribute.

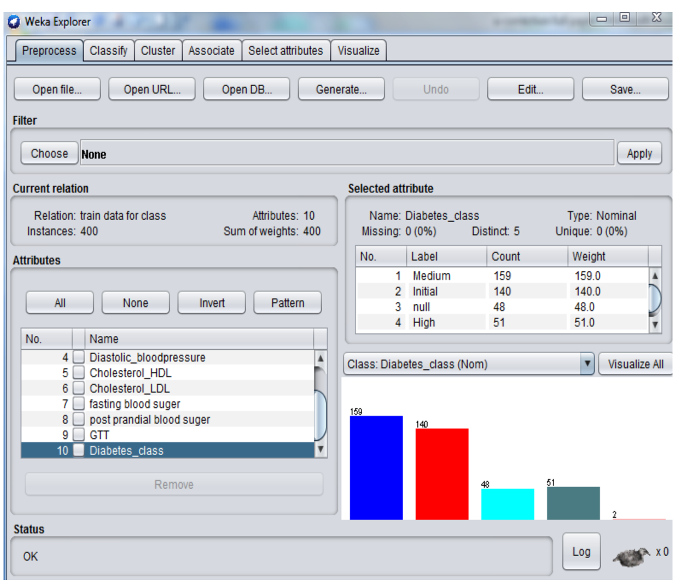

Fig.14. Representing data load into WEKA.

After predicting the diabetic of human body now we are given dataset in WEKA to find out the level of diabetes using Nave Bayes algorithm. This is the sample of dataset used for prediction Diabetes in human body. The dataset used contains 500 instances. All instances have 10 input attribute's five output attribute. The dataset variables which are used for classifying the level of diabetes are Systolic blood, Diastolic Blood pressure, GTT (Glucose tolerance test),FBS (Fasting blood sugar),PPBS (Post Prandial Blood Sugar), High-density Lipoprotein (HDL), Low-density Lipoprotein (LDL) etc. If the value of fasting plasma glucose is less than 5.6 mmol/l and value of Post Prandial Blood Sugar test is less than 7.8 mmol/l than it will be given score 0 which is Null. If the Post Prandial Blood Sugar test is gather than 7.8 mmol/l than it will be given score 1 which is Initial level. If the Post Prandial Blood Sugar test is gather than 12 mmol/l and less than 19 mmol/l than it will be given score 2 which is Medium level. If the Post Prandial Blood Sugar test is gather than 20 mmol/l and less than 33 mmol/l than it will be given score 3 which is High level.If the Post Prandial Blood Sugar test is gather than 33 mmol/l than it will be given score 4 which is Extreme level.

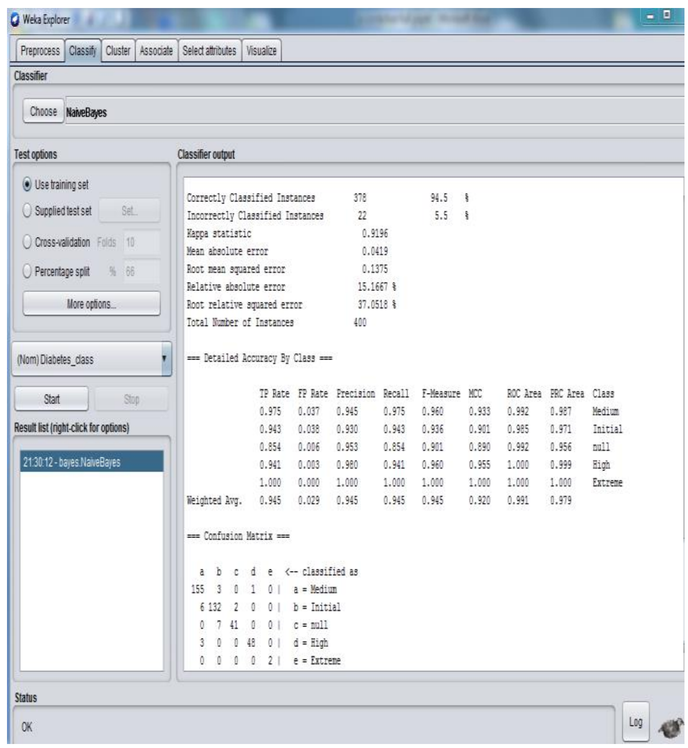

Fig.15. Training data results obtained using Nave Bayes.

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 5, ISSN No. 2455-2143, Pages 512-525
Published Online September 2019 in IJEAST (http://www.ijeast.com)

```
=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
 155   3   0   1   0 |   a = Medium
   6 132   2   0   0 |   b = Initial
   0   7  41   0   0 |   c = null
   3   0   0  48   0 |   d = High
   0   0   0   0   2 |   e = Extreme
```

Fig.16. Confusion Matrix of Training data results obtained using Naive Bayes.

TABLE7. Training data results obtained using Naive Bayes.

Classifying the level of diabetes	Total Data : 400
Null	41
Initial	132
Medium	155
High	48
Extreme	2

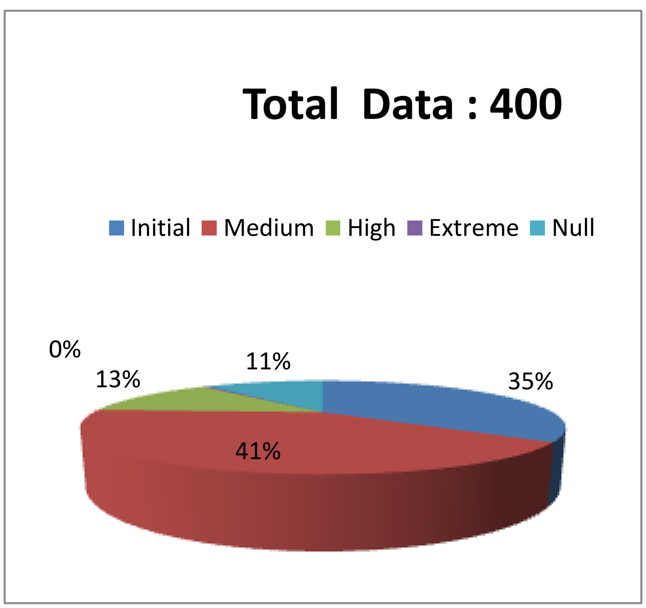

Fig.17. Graphical view of Training data results obtained using Naive Bayes.

Now Correctly Classified Instances of training data set 94%.and Incorrectly Classified Instances are 5% from 400 dataset.

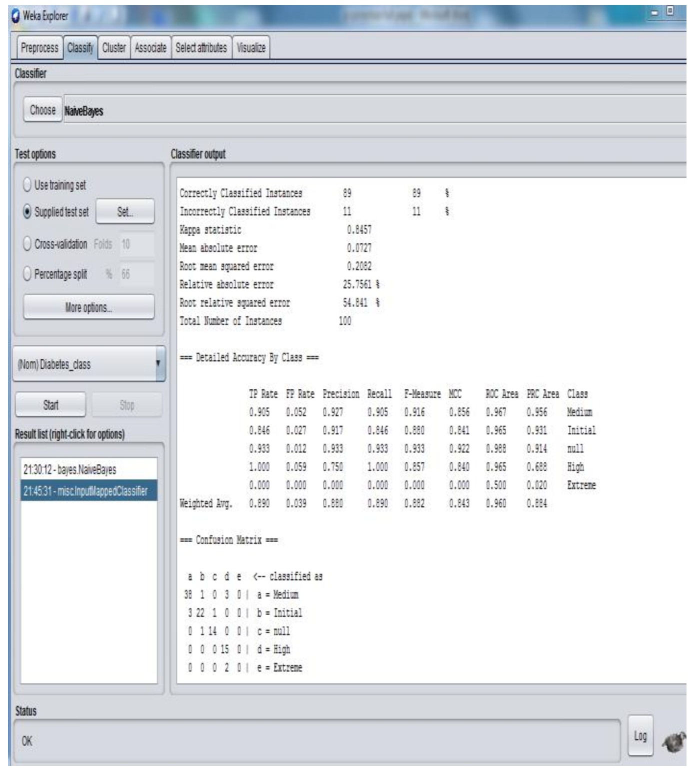

Fig.18. Test data results obtained using Naive Bayes.

```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 38  1  0  3  0 |  a = Medium
  3 22  1  0  0 |  b = Initial
  0  1 14  0  0 |  c = null
  0  0  0 15  0 |  d = High
  0  0  0  2  0 |  e = Extreme
```

Fig.19. Confusion Matrix of Test data results obtained using Naïve Bayes.

TABLE8. Test data results obtained using Naïve Bayes.

Classifying the level of diabetes	Total Data : 100
Null	14
Initial	22
Medium	38
High	15
Extreme	0

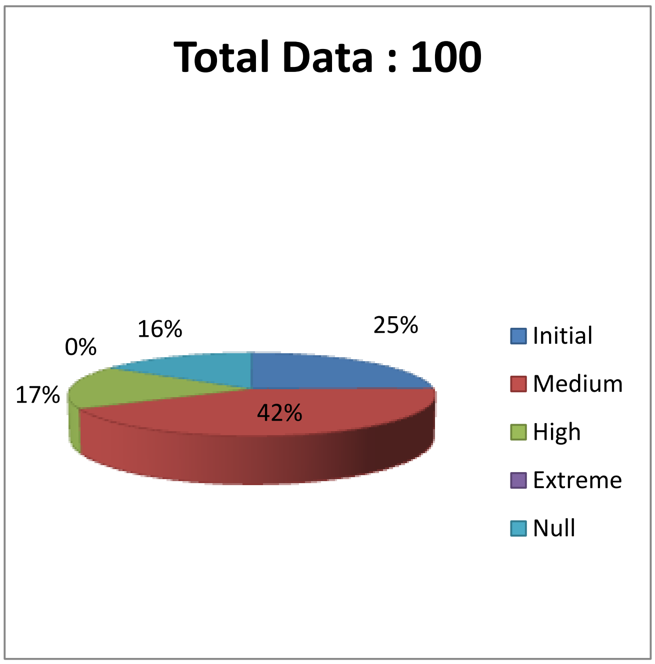

Fig.20. Graphical view of Test data results obtained using Naive Bayes.

Now Correctly Classified Instances of test data set 89%.and Incorrectly Classified Instances are 11% from 100 dataset.

TABLE 9. Accuracy of classifying the level of diabetes using Naive Bayes.

Result	Accuracy rate
Correct Accuracy	89%
Incorrect Accuracy	11%

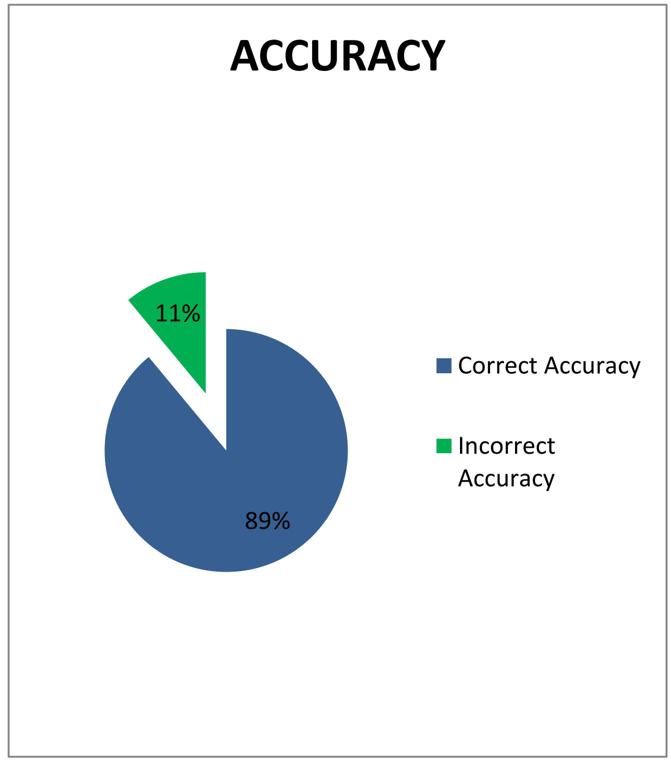

Fig.21. Accuracy of classifying the level of diabetes using Naive Bayes.

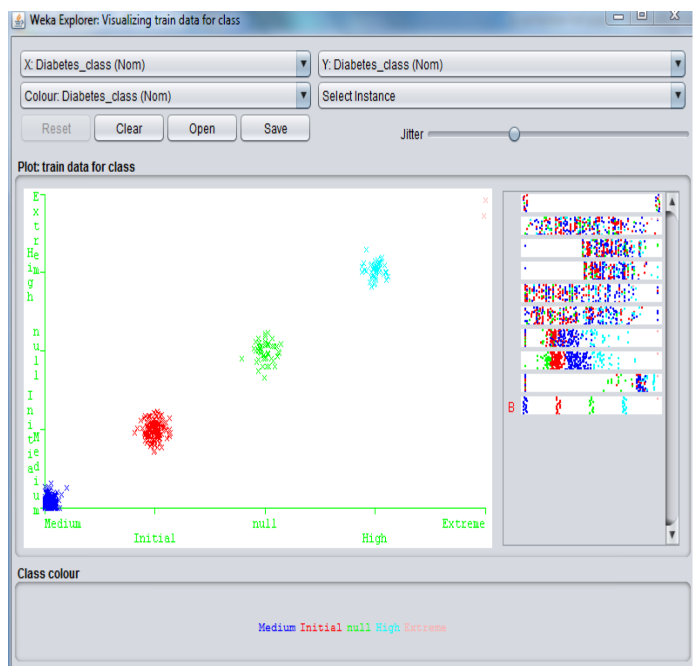

Fig.22. Classifying level of diabetes shown after applying technique.

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 5, ISSN No. 2455-2143, Pages 512-525
Published Online September 2019 in IJEAST (http://www.ijeast.com)

The Graphical view of Classification level after applying Naïve Bayes

D. Testing Result for classifying the level of diabetes

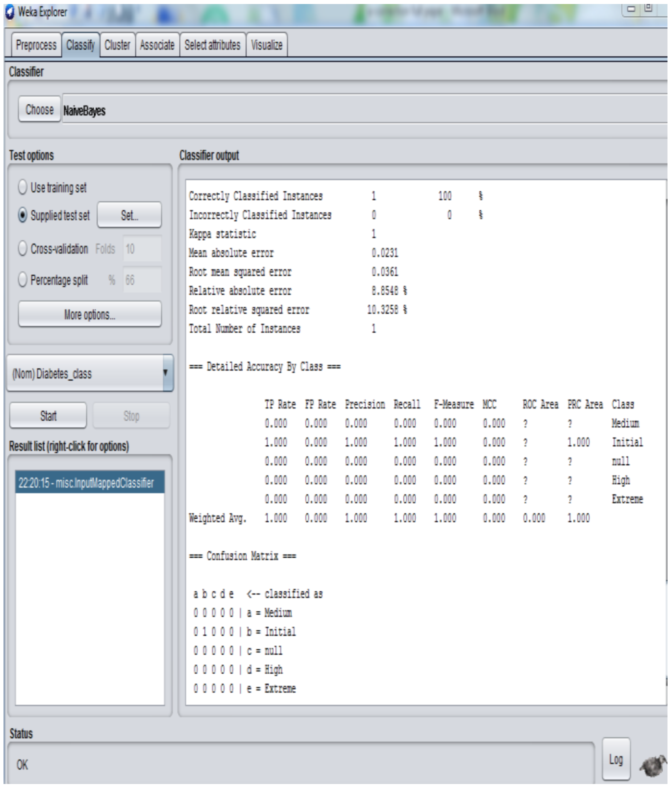

Fig.23. Single patient test data result for classifying the level of diabetes.

The single patient test report can perform correctly. Correctly Classified Instances of single patient test report is 100%

V. CONCLUSION

The purpose of this study was to predict the diabetes. The study covered two algorithms. Those are used in prediction & Classification of Data Mining Algorithm. The aim of data mining is to extract knowledge from information stored in dataset and generate clear and understandable description of patterns. One of the objectives was to provide smart way to predict the diabetes and measure the level of diabetes and also analyze the patient report in an effortless and efficient manner. The main objective was to create a diabetes prediction and classification model. To get the result of occurrence of the diabetic disease by providing the details to the application that is designed to help out the users with appropriate outcomes.

The Purpose of the study is to make a Predictive model by using data mining algorithms to predict the diabetes in human body and also measure the level of diabetic by analyzing all the symptoms of diabetic patients. Throughout the review the concept of their approaches are discussed in briefly with elementary examples. Though the research has been completed as it's goaled and find out the level of diabetes. There were some limitations of the research paper.

VI. BIBLIOGRAPHY

[1] Devi M., Maria J (2016), "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus" International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 1, pp 727-730

[2] Krishnaveni J, Sudha T (2017),"A NOVEL TECHNIQUE TO PREDICT DIABETIC DISEASE USING DATA MINING – CLASSIFICATION TECHNIQUES", International Journal of Advanced Scientific Technologies, Engineering and Management Sciences (IJASTEMS-ISSN: 2454-356X) Volume.3,Special Issue.1.

[3] Piatetsky G, Fayyed U and Smith P (1996): From data mining to Knowledge discovery: An overview. Advances in knowledge Discovery and Data Mining.pages 1-35, MIT Press.

[4] Gaganjot Kaur & Amit Chhabra (2014)," Improvement J48 Classification Algorithm for the Prediction of Diabetes, " International Journal of Computer Application (0975 – 8887), Vol.98 – No. 22.

[5] HianChyeKoh and Gerald Tan (2012): Data Mining Applications in Healthcare. Journal of Healthcare Information Management, Vol 19, No 2.

[6] Saravananathan ,T. Velmurugan, (2016) " Analyzing Diabetic Data using Classification Algorithms in Data Mining",Indian Journal of Science and Technology, Vol 9(43), DOI: 10.17485/ijst/2016/v9i43/93874.

[7] Rajesh K& Sangeetha V (2012),"Application of Data MiningMethods and Techniques for Diabetes Diagnosis," inproceedings of International journal of Engineering andInnovative Technology, vol.2, Issue 3.

[8] Mukesh kumari, Dr. RajanVohra,&Anshularora (2014),"Prediction of Diabetes Using Bayesian Network",International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 5174-5178

[9] Ms. Nilam chandgude,Prof. Suvarna pawar (2016), "A survey on diagnosis of diabetes using various classification algorithm",International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169, Volume: 3 Issue: 12 ,6706 – 6710

[10] Durairaj M, Kalaiselvi G (2015) " Prediction Of Diabetes Using Soft Computing Techniques- A Survey", INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 4, ISSUE 03, ISSN 2277-8616

[11] Giudici P (2003): Applied Data Mining Statistical Methods for Business and Industry.

[12] Priyanka S. R. Shetty, Sujata Joshi (2016)," A Tool for Diabetes Prediction and Monitoring Using Data Mining Technique", I.J. Information Technology and Computer Science, 2016, 11, 26-32 ,DOI: 10.5815/ijitcs.2016.11.04

[13] Sathees Kumar B, Gayathri P (2014), "Analysis of Adult-Onset Diabetes Using Data Mining Classification Algorithms",International Journal of Modern Computer Science (IJMCS) Volume No.-2, Issue No.-3, ISSN: 2320-7868 (Online)

[14] Vrushali Balpande , Rakhi Wajgi (2017), "Review on Prediction of Diabetes using Data Mining Technique", International Journal of Research and Scientific Innovation (IJRSI) , Volume IV, Issue IA, ISSN 2321–2705.

[15] Veena Vijayan V., Aswathy Ravikumar (2017), "Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus", International Journal of Computer Applications (0975 – 8887) Volume 95– No.17.

[16] Abhishek Taneja (2013), "Heart Disease prediction system using data mining techniques", An International Open Free Access and Peer Reviewed Research Journal, Vol.3, No.4.

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