International Journal of Engineering, Science and Mathematics
Vol. 6 Issue 8, December 2017 (Special Issue)
ISSN: 2320-0294 Impact Factor: 6.765
Journal Homepage: http://www.ijesm.co.in, Email: ijesmj@gmail.com Double-Blind
Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Data mining Techniques applied on Medical Data to Predict Various Diseases in Health care

PydipalaLaxmikanth*
Dr.RaviBhramaramba**

Abstract

The aim of writing this research paper is to present the importance of the data mining techniques, algorithm on medical data for finding interesting patterns and discover knowledge that are used in diagnosis and decision making to predict various diseases. The health care domain produces huge amount of data and stored in repositories. The Data mining is very useful in analyzing the medical data.it is very difficult to for medical practitioners to analyze the huge quantity of data and required enough knowledge to mine it. Classification, clustering, Association Data mining techniques widely used in healthcare domain and it improve the quality of prediction and diagnosis of various diseases. The goal of this paper is to discuss the importance of datamining techniques on a medical dataset to discovery the disease hidden patterns along the results of different techniques of datamining and its algorithms. The results shown in this paper are about classification accuracy, precision, recall.

Keywords:

Classification,Clustering ,Association,medicaldata, algorithm,datamining,knowledge discovery.

Author correspondence:

- PydipalaLaxmikanth,Department of Computer Science & Engineering, Chaitanya Engineering College, JNTUK, India.
- Dr.RaviBhramaramba,Department of Information Technology, GITAM Institute of Technology, GITAM University, India.

1.Introduction

Data mining is a multi-disciplinary field which combines statistics, machine learning, artificial intelligence and database technology. Data mining is the process of finding the interesting patterns in large data sets of any raw data. Data mining can also be knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. The popular term knowledge discovery from data or KDD is the synonym for data mining. It is the process of seven iterative sequence of the steps like Data cleaning, Data integration, Data Selection, Data transformation, Data mining, Pattern evaluation and knowledge presentation. The large amounts of data created by healthcare records are complex and large to be processed and analyzed by traditional ways. Data mining can be used for technical solutions to deal with the analysis of medical data and to find the predictions. At present several Data mining tools constructing to perform analysis for speed up the diagnosis of various diseases. In this paper consider Diabetic data and apply classification techniques and clustering techniques. The paper is organized as follows: Section–1 gives the introduction; The Classification and clustering techniques and its algorithms considered are presented and elaborated on

* Doctorate Program, Linguistics Program Studies, Udayana University Denpasar, Bali-Indonesia (9 pt)
** STIMIK STIKOM-Bali, Renon, Depasar, Bali-Indonesia

Section–2; in Section -3 disease description in Section–4 the literature survey is discussed; Section–5 deals with the considered datasets and the employed Data Mining tool WEKA; Section–6 Methodology followed in this paper; Section–7 presents the results and a discussion about them and finally the Section–8 presents the conclusions.

2. Classification and Clustering techniques on medical data:

2.1.Classification : Classification is the problem of categorizing a new observation in a set of categories, on the basis of a training set of data. Also known as supervised classification, it is used widely as the predictive data mining technique which uses the given class labels to order the objects in the data collection[1].Classification is a supervised machine learning technique that assigns labels or classes to different objects or groups[2].

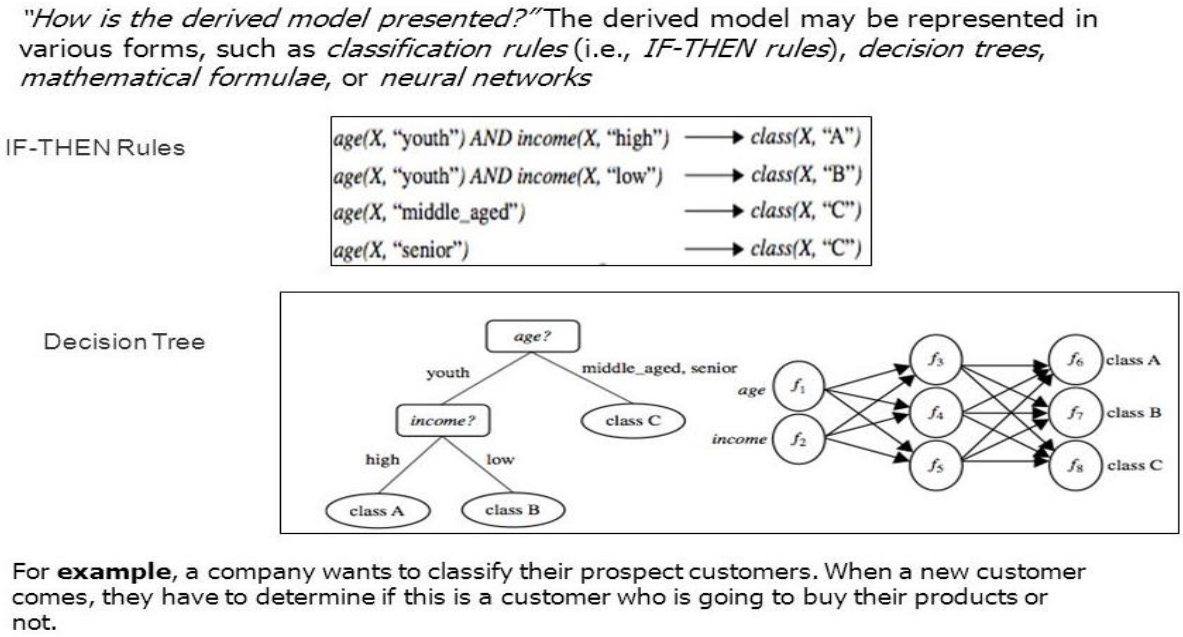


Fig 2.1 Data mining classification technique. (source from Han and Kamber book)

2.3 Clustering: It is a group that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster. it is a process of partitioning a set of data(or objects) into a set of meaningful sub-classes, called clusters.Clustering is a main task in data mining and a general technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bio-informatics[1].

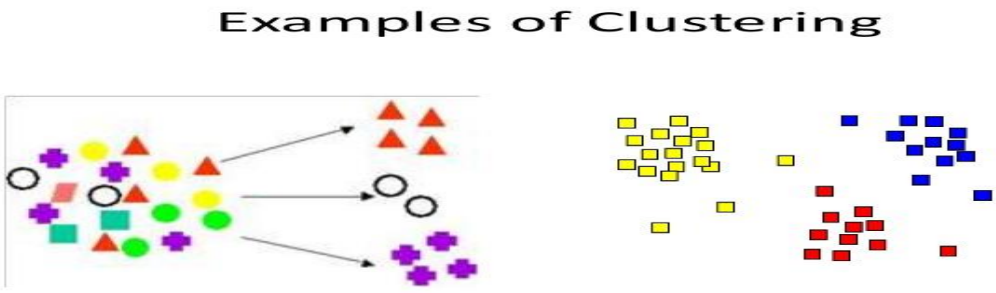


Fig 2.2 Data mining Clustering Anaysis. (source from web)

3. **Diabetic:**

3.1 **Diabetes:** Diabetes, often referred to by doctors as diabetes mellitus, describes a group of metabolic diseases in which the person has high blood glucose (blood sugar), either because insulin production is inadequate, or because the body's cells do not respond properly to insulin, or both. Patients with high blood sugar will typically experience polyuria (frequent urination), they will become increasingly thirsty (polydipsia) and hungry (polyphagia).[3]

3.1.1. Type 1 diabetes: In type 1 diabetes, the pancreas is unable to produce any insulin, the hormone that controls blood sugar levels.Insulin production becomes inadequate for the control of blood glucose levels due to the gradual destruction of beta cells in the pancreas. This destruction progresses without notice over time until the mass of these cells decreases to the extent that the amount of insulin produced is insufficient.Type 1 diabetes typically appears in childhood or adolescence, but its onset is also possible in adulthood.When it develops later in life, type 1 diabetes can be mistaken initially for type 2 diabetes. Correctly diagnosed, it is known as latent autoimmune diabetes of adulthood.[3]

The body does not produce insulin. Some people may refer to this type as insulin-dependent diabetes, juvenile diabetes, or early-onset diabetes. People usually develop type 1 diabetes before their 40th year, often in early adulthood or teenage years. Type 1 diabetes is nowhere near as common as type 2 diabetes. Approximately 10% of all diabetes cases are type 1. Patients with type 1 diabetes will need to take insulin injections for the rest of their life. They must also ensure proper blood-glucose levels by carrying out regular blood tests and following a special diet.[3]

3.1.2. Type 2 diabetes: Type 2 diabetes mellitus most commonly develops in adulthood and is more likely to occur in people who are overweight and physically inactive.Unlike type 1 diabetes which currently cannot be prevented, many of the risk factors for type 2 diabetes can be modified. For many people, therefore, it is possible to prevent the condition.The International Diabetes Foundation highlight four symptoms that signal the need for diabetes testing:Frequent urination .Weight loss ,Lack of energy,Excessive thirst.[3]

4. Related works: Data mining techniques used on medical data for predicting of the disease. In this paper we diagnosis the diabetes and breast cancer by using clustering and classification techniques and find the statistics of the disease in order to finding its level.

The author [1] creation of expert clinical system for the diagnosis of the diabetic mellitus using clustering and classification techniques of data mining. However with suitable modification the same can be extended to evolve similar systems in other application areas in health care.

The author of [2] used the diabetic data set and using classification finding the Accuracy, Sensitivity, Specificity and RMS when compared to the other two datasets.

The author of [6] is used the CART Method of decision Tree classification to monitoring and diagnosing the diabetes.

The author of [7] used the Support Vector Machines is used to in order to find out the diagnosis of type-2 diabetes.

The author of [8] is c4.5 classification algorithm is used to diagnosis the diabetes and finding the Accuracy and Recall and Precision.

5. Data sets and Tools used: Record set with attributes was used from the Pima Indians Diabetes Data Set [9]. The records Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Waikato Environment for Knowledge Analysis (WEKA) came about through the perceived need for a unified work-bench that would allow researchers easy access to state-of- the-art techniques in machine learning [10]. WEKA aims to provide a complete collection of machine learning algorithms and data pre-processing tools to researchers and practitioners equally. It allows users to promptly try out and compare different machine learning methods on new data sets. The work bench having algorithms for regression, classification, clustering, association rule mining and attribute selection.

6. Methodology:

6.1 Algorithms Applied

The model created for this paper uses Simple K-Means clustering algorithm for predicting the person having diabetes or not among the patients. After finding the person diabetes or not the classification algorithms like Navie Bayes,J48,Decision Table, Random are used to find the accuracy of classification.

6.2 Discussion :

The proposed model having three stages.

1. Data pre-processing.

2. Application of Simple K-Means algorithm to the dataset for clustering the data into two clusters as cluster-0 (tested_negative), cluster-1 (tested_negative).Application of Classification algorithms to classify the patient's threat of diabetes levels.

6.3 Data preprocessing : As the collected data contain some inconsistencies data preprocessing was done to remove the inconsistencies.

Attribute	Description	Type
Preg	Number of times pregnant	Numeric
plas	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Numeric
pres	Diastolic blood pressure (mm Hg)	Numeric
skin	Triceps skin fold thickness (mm)	Numeric
insu	2-Hour serum insulin (mu U/ml)	Numeric
mass	Body mass index (weight in kg/(height in m)^2)	Numeric
pedi	Diabetes pedigree function	Numeric
age	Age (years)	Numeric
class	Class variable (0 or 1)	Numeric

Table 6.1 The attributes used in the experimentation.

6.4 Accuracy Measures:

Accuracy- Proportion of correct classifications (true positives and negatives) from overall number of cases.
Accuracy = (TP + TN)/(TP + TN + FP + FN)

Precision-Proportion of correct positive classification (true positive) from cases that are predicted as positive.

Precision = TP/(TP + FP).

Recal- Proportion of correct positive classification (true positive) from cases that are actually potitive.
Recall = TP/(TP + FN).

TP(True Positive) - Positive tuples.(positive and predicted postitive)
TN(True Negative) - Negative tuples.(Negative and predicted negative)
FP(False Positive) - Incorrectly classified positive tuples.(case was negative but predicted positive)
FN (False Negative)- Incorrectly classified negative tuples.(case was positive but predicted negative)

Actual class	Predicted class	
	Negative	Positive
Negative	TN(True Negative)	FP(False Positive)
Positive	FN (False Negative)-	TP(True Positive)

Table 6.4Accuracy Measure Table

7. Results and Discussions:

The model was executed in three stages:
Stage 1 - The entire dataset was preprocessed.
Stage 2 - After preprocessing the clean dataset was clustered to find the patient is diabetic or not .
Stage 3 - The clustered dataset was classified into three classes as mild, moderate and severe. This was done in order to predict the threat levels of diabetes for each patient.

7.1 Performance of the Simple K-Means Algorithm :

The Simple k-means algorithm clusters the entire dataset into 3 clusters as cluster-0 – tested negative
cluster-1 for tested positive. The time taken to build the model was 0.11 seconds. Among the 768 instances of the data, 500 were in cluster-0 (– tested negative), 268 were in cluster-1 (i.e. tested positive) .This clustered dataset was given as input to the model which classified each patient's threat levels of diabetes as mild, moderate and severe.

kMeans
======
Number of iterations: 4
Within cluster sum of squared errors: 149.5177664581119
Initial starting points (random):
Cluster 0: 1,126,56,29,152,28.7,0.801,21,tested_negative
Cluster 1: 8,95,72,0,0,36.8,0.485,57,tested_negative
Missing values globally replaced with mean/mode.

Attribute	Full Data (768.0)	Cluster 0(500.0)	Cluster1(268)
preg	3.8451	3.298	4.8657
plas	120.8945	109.98	141.2575
pres	69.1055	68.184	70.8246
skin	20.5365	19.664	22.1642

insu	79.7995	68.792	100.3358
mass	31.9926	30.3042	35.1425
pedi	0.4719	0.4297	0.5505
age	33.2409	31.19	37.0672
class	tested_negative	tested_negative	tested_positive

Table 7.1 Final cluster centroids

Time taken to build model (full training data) : 0.11 seconds

Clustered Instances	Number of instances	%percentage
0	500(tested_negative)	65
1	268 (tested_positive)	35
Total instances	700	100

7.2 Classifier Performances: considered four classification algorithms namely Naïve Bayes, J48, Decision Table, Random forest algorithms. To find the accuracy of classification of dataset considered the accuracy, precision, Recall measure. After watching the these values found the Random forest classification algorithm classifying the dataset instances accurately. Below can find the confusion matrix and the calculations of accuracy, precision, Recall in the below table.

Classifier	**Accuracy**	**precision**	**Recall**
Naive Bayes	0.763	0.676	0.616
J48	0.841	0.664	0.745
Decision Table	0.776	0.735	0.560
Random forest	1.000	1.000	1.000

Table 7.2 Accuracy, precision and recall Measures

7.2.1Naive Bayes:

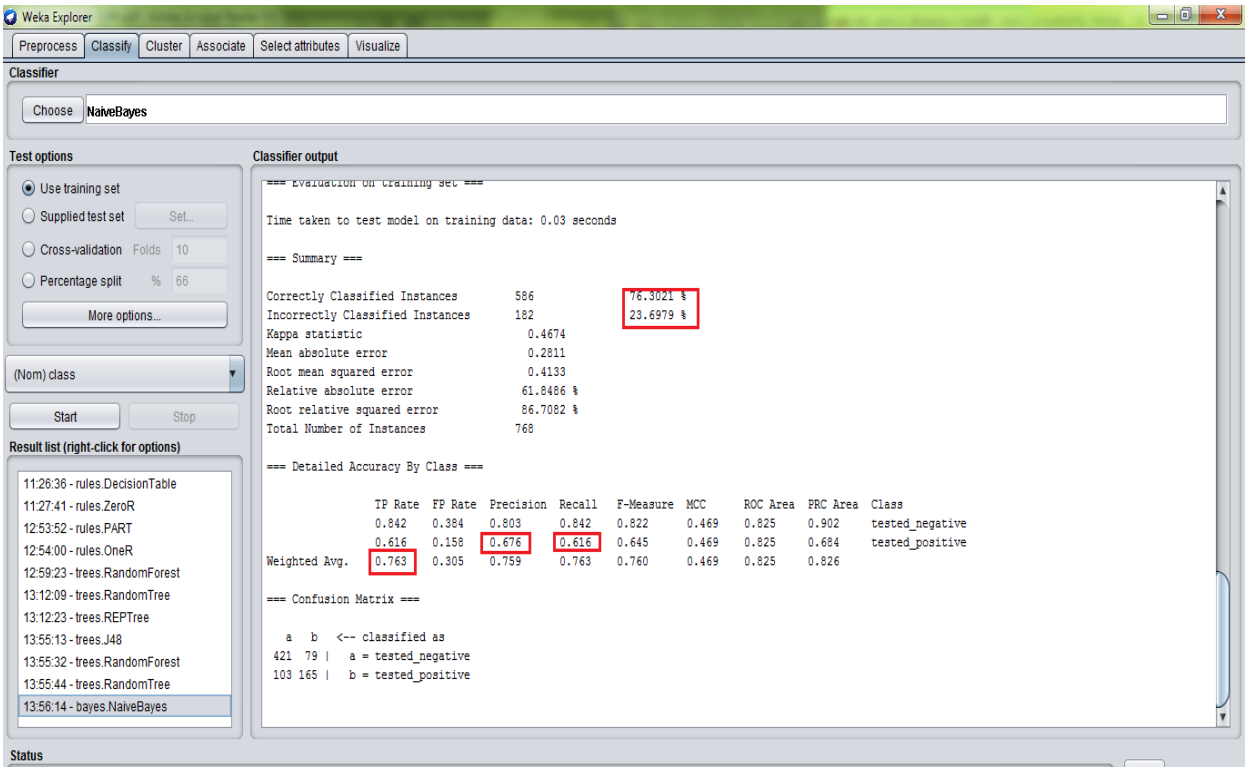
Confusion matrix:

Actual class	Predicted class	
	Negative	Positive
Negative	TN(True Negative) -421	FP(False Positive)-79
Positive	FN (False Negative)-103	TP(True Positive)-165

Accuracy = (TP + TN)/(TP + TN + FP + FN) =421+165/768 =0.7630

Precision = TP/(TP + FP)=165/165+79=0.676

Recall = TP/(TP + FN)=165/165+103=0.616



7.2.2 J48:

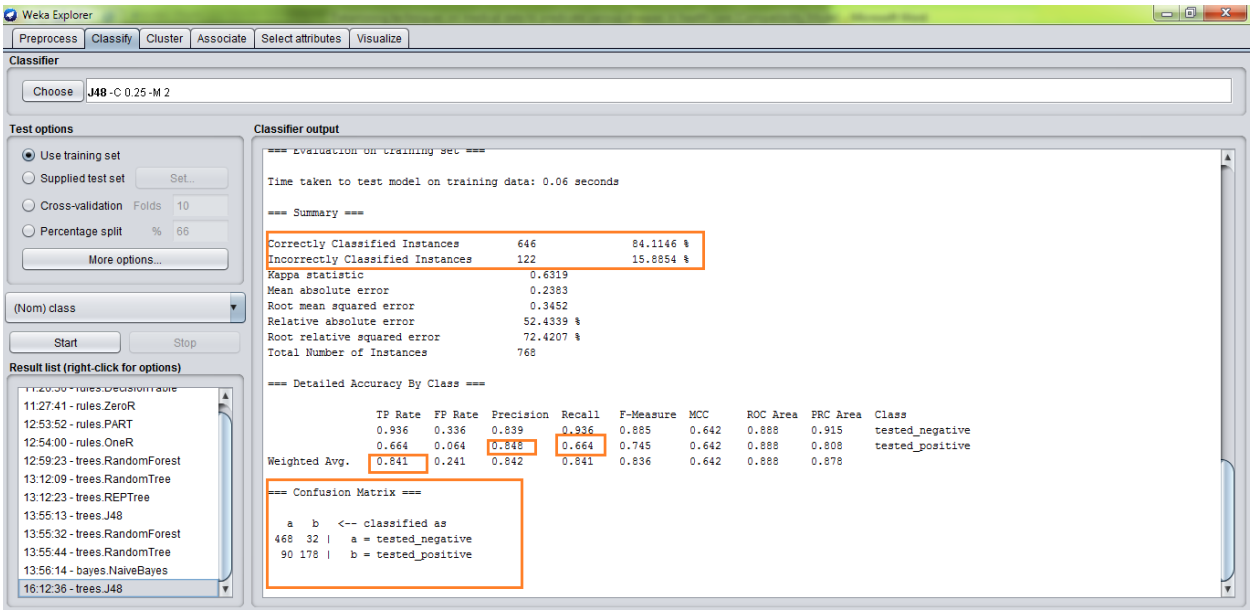
Confusion matrix:

Actual class	Predicted class	
	Negative	Positive
Negative	TN(True Negative) -468	FP(False Positive)-32
Positive	FN (False Negative)-90	TP(True Positive)-178

Accuracy = (TP + TN)/(TP + TN + FP + FN) =468+178/768 =0.841

Precision = TP/(TP + FP)=178/178+32=0.848

Recall = TP/(TP + FN)=178/178+90=0.664



7.2.3 Decision Table:

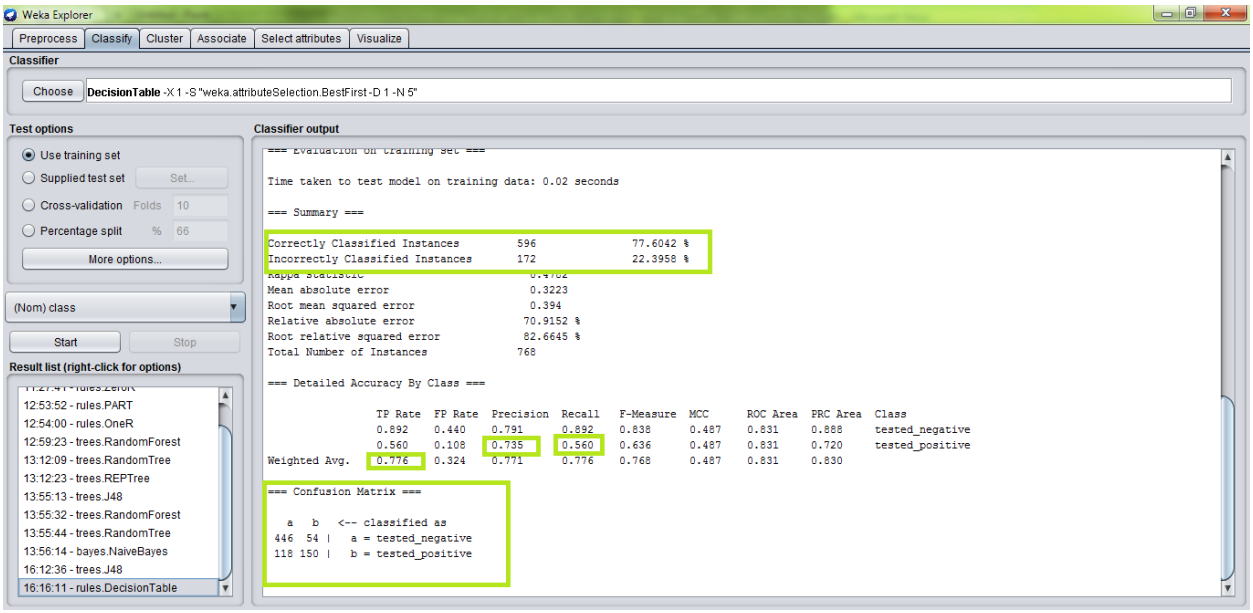
Confusion matrix:

Actual class	Predicted class	
	Negative	Positive
Negative	TN(True Negative) -446	FP(False Positive)-54
Positive	FN (False Negative)-118	TP(True Positive)-150

Accuracy = (TP + TN)/(TP + TN + FP + FN) =446+150/768 =0.776

Precision = TP/(TP + FP)=150/150+54=0.735

Recall = TP/(TP + FN)=150/150+118=0.560



7.2.4Random Forest:

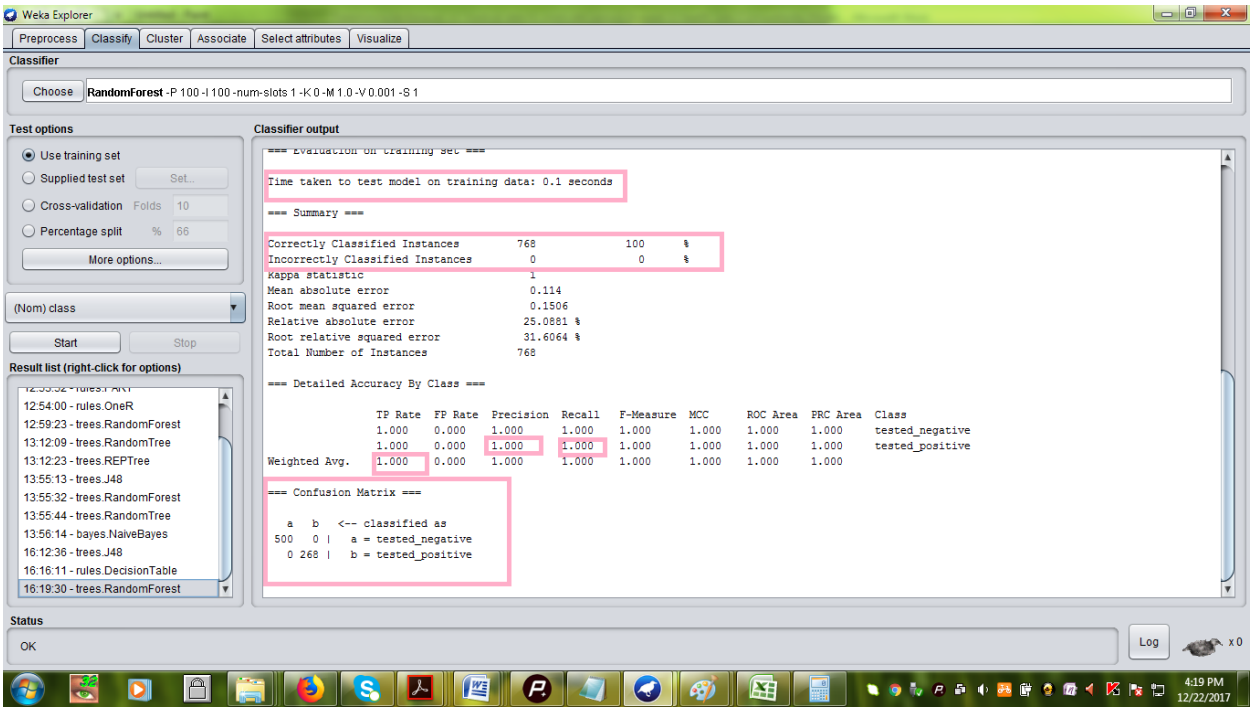
Confusion matrix:

Actual class	Predicted class	
	Negative	Positive
Negative	TN(True Negative) -500	FP(False Positive)-0
Positive	FN (False Negative)-0	TP(True Positive)-268

Accuracy = (TP + TN)/(TP + TN + FP + FN) =500+268/768 =1.000

Precision = TP/(TP + FP)=268/268+0=1.000

Recall = TP/(TP + FN)=268/268+0=1.000



8.Conclusion: From the Results depicted in the previous section the classification activity is very sensitive. whilechosing the classification algorithm more conscious .the best algorithm chosen then more accurately classification happens.

9. REFERENCES:

1.SrideivanaiNagarajan, and R. M. Chandrasekaran, Design and Implementation of Expert Clinical System for Diagnosing Diabetes using Data Mining Techniques. **Indian Journal of Science and Technology,** April 2015;Vol 8(8):771–776.

2.Ch. Rakesh, D.N.D.Harini, M. Bhanu Sridhar.An Empirical Analysis of Classification Algorithms for Medical Data. International Conference on Artificial Intelligence, Robotics and Embedded Systems,2015 OCT16, At Andhra University, Visakhapatnam, India.

3. By The MNT Editorial Team, 2016 January 5: Diabetes: Symptoms, Causes and Treatments: Available from: https://www.medicalnewstoday.com/info/diabetes : (Electronic Newspaper Article on the Internet).

4. By The MNT Editorial Team, 2017 November 27: What you need to know about breast cancer: Available fromhttps://www.medicalnewstoday.com/articles/37136.php : (Electronic Newspaper Article on the Internet).

5.D.Lavanya, Dr. K.Usha Rani, Performance evaluation of decision tree classifiers on medical datasets.2011july. International Journal of Computer Applications (0975 – 8887),Volume 26– No.4, P:2-4.

6.KavithaK ,Sarojamma R M , Monitoring of Diabetes with Data Mining via CART Method.2011NOV. International Journal of Emerging Technology and Advanced Engineering. ISSN 2250-2459, Volume 2, Issue 11,P-158-162.

7.BayuAdhi Tama, Rodiyatul F. S., Hermansyah ,An Early Detection Method of Type-2 Diabetes Mellitus in Public Hospital.2011 August,TELKOMNIKA, Vol.9, No.2, pp. 287~294

8. K. Rajesh, V. Sangeetha, Application of Data Mining Methods and Techniques for Diabetes Diagnosis,2011 Sep. International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3,P-225-229.

9. UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes.

10. Hall, Mark, et al. "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter 11.1 (2009): 10-18.