Identification of Models-Decision Tree and Random Forest Classifier using Rattle on Diabetes Disease

Aparna Chauhan, Ankur Garg

Abstract— ***Diabetes is the disease which is growing now a days in human body and there are a number of patient who are suffering by this diabetes in the world. The data related to medical area is very huge which is related to the many disease. So the first thing is that we have to choose a mining tool which give best result for the given databases. Because, this medical data is statistical and most of the researchers using this type of data. Data mining tool is used for the extracting better result in accuracy for the diabetes data base. By the data mining techniques the medical expert and researchers analyze the result and provide the best treatment for this disease. In this paper we are using diabetes data and apply it on the Rattle, an open source tool of data mining and perform two classification methods decision tree and random forest tree for classify the data and show that which classification algorithm is best for diabetes dataset.***

Index Terms**:** ***Data mining, Diabetes, Rattle tool, Decision Tree, Random Forest Tree***

I. INTRODUCTION

Data mining is used for the extraction of the meaningful information from huge amount of the data. It plays a very important role in the derive patterns by applying different type of techniques. Data mining have taken advantages to conclude the analysis in various areas like medical science, financial area, in organic compounds and also in whether forecasting. For the deep knowledge and medical data analysis, data mining is very efficient in the health care centers. The application of data mining is very helpful for the diminishing the error in the health care centers, and in making a decision on the policies related to health, it is very helpful in the detection of diseases in early stages, it also helpful in diminishing the rate of death by the analysis of result related to any disease. In the medical science area there are many diseases and the data is collected by the researchers like diabetes, cancer, heart related data and many more. If we are talking about the last 10 years the rate of death is very high due to the diabetes disease. because of increasing the number of diabetes patient and this data is very big in nature manually analysis is not possible on such type of data so researcher are using data mining for extracting the valuable information and result for the medical health care. There are a number of techniques which is used in diagnosis like neural network, Naïve Bayes, SVM(Support vector machine), decision tree & many more. Diabetes have basic three types which are Type 1, Type 2 and gestational diabetes. Due to lack of knowledge about the diabetes the treatment starts after a big amount of time basically patient don’t know that he/she is suffering from it or not. After some time when the body have some problem like vision, hearing, kidney and heart related problem is arises in body then the one go for the treatment and diabetes will check after the number of body tests. So for the purpose of analysis the researchers will store such type of patient data and analyze it with help of data mining techniques and get valuable information for the future.

II. LITERATURE SURVEY

The author explains about the disease like breast cancer, heart problem, diabetes and use classifier for the best result. For the heart disease they used KNN, Decision tree, Naive Bayes, the classification and clustering is used for the diabetes disease and extract pattern from the given data and prediction [1]. If the diabetes is present in the human body then it is very difficult to produce the insulin by the body. By the some test related to the blood sugar can help to detect the diabetes. They used back propagation techniques which is very helpful to reduce the cost for the test in medical area [2].

As the number of diseases are generating in the medical line so the data which is coming from these diseases are very big and it has to store in the electronic devices. Different type of machine learning & data mining techniques must be applied on it. Discussing is done on Prediction & Diagnosis, Diabetic Complications, Genetic Background, Environment, Health Care & Management. They perform machine learning on the diabetes data set which tells SVM techniques gives the most valuable result [3].

We worked on the non-diabetic and diabetic patient data by the help data mining tools. For the statistical analysis author used the weka tool. Subset evaluation, classification and associative rule mining is done in this paper. J48 and random tree is used for the classification purpose and by the result it is observed that j48 gives better result [4].Multi-layer perception, K-nearest neighbor and BLR (Binary Logistic Regression) is used for the diabetes data base. K-nearest neighbor perform good over the rest three techniques [5].

Revised Manuscript Received on July 18, 2019.

Aparna Chauhan, Computer Science and Engineering, Meerut Institute of Engineering and Technology, Meerut, Uttarpradesh, India.

Ankur Garg, Computer Science and Engineering, Meerut Institute of Engineering and Technology, Meerut, Uttarpradesh, India.

Different type of test are perform on the patient and prediction is done on the basis of factor age, hypertension, obesity, smoking, diabetes, physical inactivity, alcohol intake, high cholesterol. Density based clustering, hierarchal clustering, K-mean clustering is used in the weka tool for the analysis [6]

III. METHODOLOGY

A. Data Collection

For the analysis purpose we have collected data set from the UCI (machine learning repository). The data set is related to diabetes. The data set have 8 attributes and 1 is for class attribute. There are 768 instances, the attributes are shown in the Table 1 and all the values are numeric in the attributes. The original owner of the dataset is National Institute of Diabetes, Digestive & Kidney Diseases.

Table 1

S. No.	Attribute Name	Value
(i)	Glucose	Numeric
(ii)	Pregnancy	Numeric
(iii)	Blood Pressure	Numeric
(iv)	BMI	Numeric
(v)	Skin Thickness	Numeric
(vi)	Insulin	Numeric
(vii)	Age	Numeric
(viii)	DiabetesPedigreeFunction	Numeric
(ix)	Outcome	Categoric

B. Data Mining Tool Used

Rattle is used for statistical analysis of dataset .Rattle is stand for R Analytical Tool. To learn simply, Rattle is written in the R language and it is an analytical tool. It is a data mining application using graphical representation. Rattle has many data mining task standard, generating classification models, visualization, regression, correlation, clustering. In the Rattle we can use many type of data formats. In this paper, we used diabetes.csv data set for the analysis.

C. Data Mining

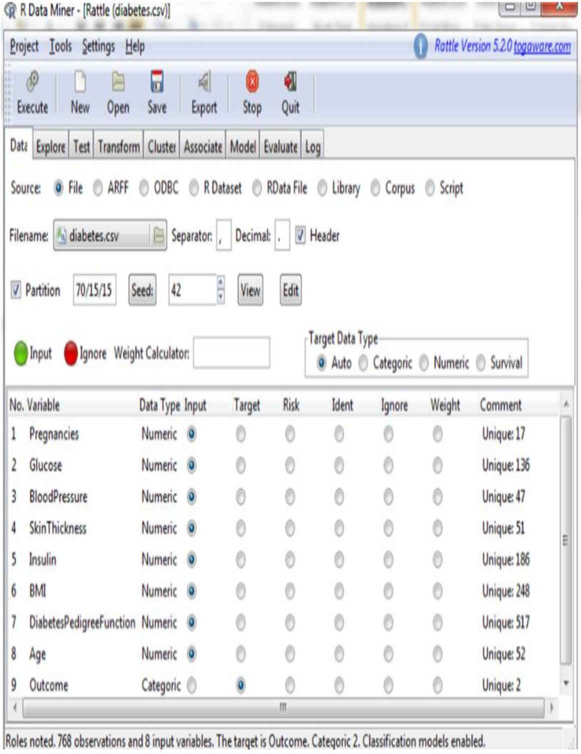
Data mining is the technique by which we can extract valuable pattern from the big data sets to conclude the meaningful information. As per the definition by Jiawei Han in his book Data Mining: Concepts and Techniques [6]. Data mining is the extraction of interesting, non-trivial, implicit, previously unknown, and potentially useful patterns or knowledge from a vast amount of data. Data mining has been used in various areas such as healthcare, business analytics, financial trading analysis, intrusion detection, etc. It is very helpful to check the behavior of the data and draw the useful pattern from it. Earlier the experts do all the analysis manually but it takes a lot of time but today we have many data mining automated tools which gives accurate result in a very short time, it has a good storage capacity.

Classification Method Used

In this work we have used two classification model namely Decision Tree and Random Forest tree. The Decision Tree model is very common and easy to understand Data mining model. We use minimum split 20, maximum depth 20, minimum bucket 7 and complexity 0.0100 for the decision tree model. When the data set is 100s or 1000s input variables we use Random Forest tree model so it is basically used when the data set is big. It describe the importance of the variables also. Random Tree is very useful in the machine learning process in present time, it draw 10s or 100s of decision tree. We use number of trees 500 and the variables are 2 in this case.

IV. RESULT AND ANALYSIS

A. For Decision tree model & Random Forest Tree creation first we have to load the data set in to the Rattle. Our data set name is diabetes.csv. Fig 1 shows after the loading the desired data set in the rattle.


R Data Miner - [Rattle (diabetes.csv)]
Project Tools Settings Help
Rattle Version 5.2.0 togaware.com
Execute
New
Open
Save
Export
Stop
Quit
Data
Explore
Test
Transform
Cluster
Associate
Model
Evaluate
Log

Source: ● File ○ ARFF ○ ODBC ○ R Dataset ○ RData File ○ Library ○ Corpus ○ Script

Filename: diabetes.csv Separator: , Decimal: . ☑ Header

☑ Partition 70/15/15 Seed: 42 View Edit

Input Ignore Weight Calculator:

Target Data Type: ● Auto ○ Categoric ○ Numeric ○ Survival

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	Pregnancies	Numeric	●	○	○	○	○	○	Unique: 17
2	Glucose	Numeric	●	○	○	○	○	○	Unique: 136
3	BloodPressure	Numeric	●	○	○	○	○	○	Unique: 47
4	SkinThickness	Numeric	●	○	○	○	○	○	Unique: 51
5	Insulin	Numeric	●	○	○	○	○	○	Unique: 186
6	BMI	Numeric	●	○	○	○	○	○	Unique: 248
7	DiabetesPedigreeFunction	Numeric	●	○	○	○	○	○	Unique: 517
8	Age	Numeric	●	○	○	○	○	○	Unique: 52
9	Outcome	Categoric	○	●	○	○	○	○	Unique: 2

Roles noted. 768 observations and 8 input variables. The target is Outcome. Categoric 2. Classification models enabled.

Fig 1: Loading data set image

Summary of the data set is as below in Table 2

Table 2

Attribute	Levels	Storage
Glucose		integer
Pregnancy		integer
Blood Pressure		integer
BMI		double
Skin Thickness		integer
Insulin		integer
Age		integer
DiabetesPedigreeFunction		double
Outcome	2	integer

Variable level outcome- yes, no

For the simple distribution, tables below-

The 1st and 3rd Qu. refers to the first and third quartiles, indicating that the 25% of observations have values of that variable which are less than or greater than the value listed.

Table 3

Pregnancies	Glucose
Min. : 0.00 1st Qu.: 1.00 Median : 3.00 Mean : 3.84 3rd Qu.: 6.00 Max. :17.00	Min. : 0 1st Qu.: 99 Median :116 Mean :121 3rd Qu.:142 Max. :199

Table 4

BloodPressure	SkinThickness
Min : 0.00 1st Qu.:62.00 Median:70.00 Mean : 68.23 3rd Qu.:78.00 Max. :114.00	Min:0.00 1st Qu.: 0.00 Median:22.00 Mean :20.47 3rdQu.:32.00 Max. :99.00

Table 5

Insulin	BMI
Min. : 0.00 1st Qu.: 0.00 Median : 36.00 Mean : 81.77 3rd Qu.:125.00 Max. :846.00	Min. : 0.00 1stQu.:26.60 Median :32.00 Mean :31.48 3rQu.:35.90 Max. :59.40

Table 6

DiabetesPedigreeFunction	Age
Min. :0.0780 1st Qu.:0.2440 Median :0.3700 Mean :0.4607 3rd Qu.:0.6010 Max. :2.4200	Min. :21.00 1st Qu.:24.00 Median:29.00 Mean :33.11 3rdQu.:41.00 Max. :72.00

Outcome
In case of No : 354
In case of Yes : 183

Now we have to check what is the distribution of the variables for this purpose click on the distribution tab and the output is as in fig 2.

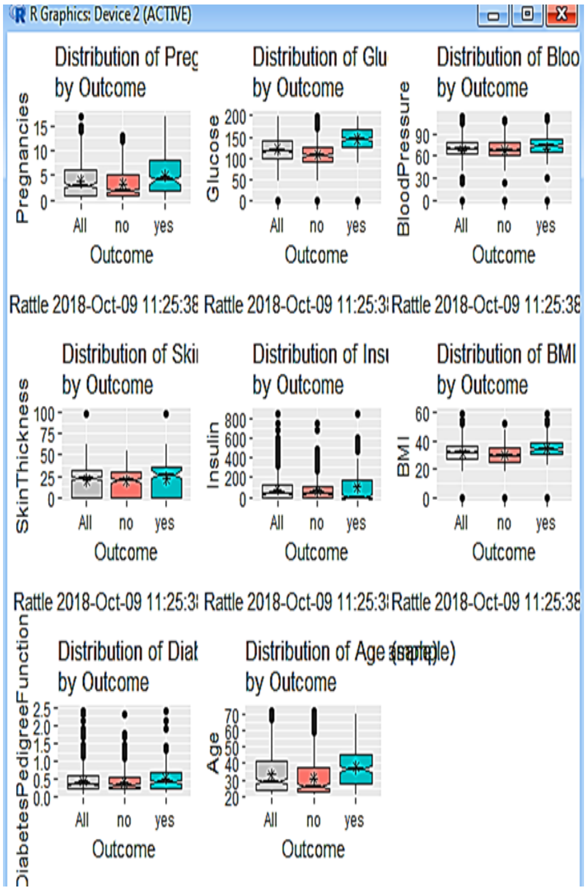

R Graphics: Device 2 (ACTIVE)
Distribution of Preg by Outcome
Distribution of Glu by Outcome
Distribution of Bloo by Outcome
Pregnancies
Glucose
BloodPressure
Outcome
All
no
yes
Rattle 2018-Oct-09 11:25:38 Rattle 2018-Oct-09 11:25:3 Rattle 2018-Oct-09 11:25:38
Distribution of Ski by Outcome
Distribution of Insu by Outcome
Distribution of BMI by Outcome
SkinThickness
Insulin
BMI
Rattle 2018-Oct-09 11:25:3 Rattle 2018-Oct-09 11:25:3 Rattle 2018-Oct-09 11:25:38
Distribution of Diat by Outcome
Distribution of Age (sample) by Outcome
DiabetesPedigreeFunction
Age

Fig 2: Distribution

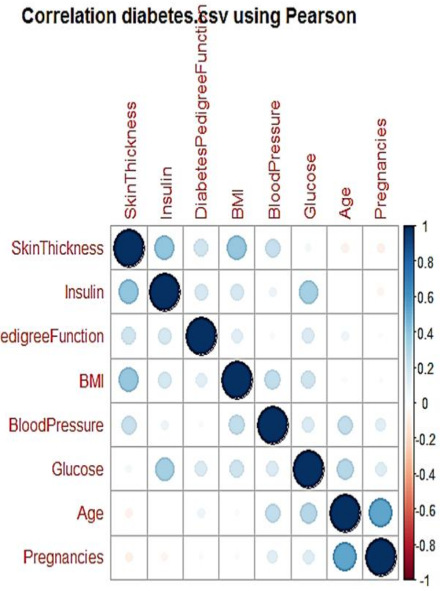


Fig 3: Correlation between the variables

B. Decision Tree

Tree now we are going to classify the data with the help of Decision tree model. For this we have to click on the model tab and select the tree radio button and selct the values as min. Split=20, max.depth=3, min.buckets=7 and, complexities=0.0100. After clicking on the tree model tab (decision tree) we have to click on the execute button on the top in the tool. Or we can code in the R console then it

will classify our data and the decision tree shown in fig 4.
Variables actually used in tree construction: [1]
Age , BloodPressure, BMI, Glucose, Insulin

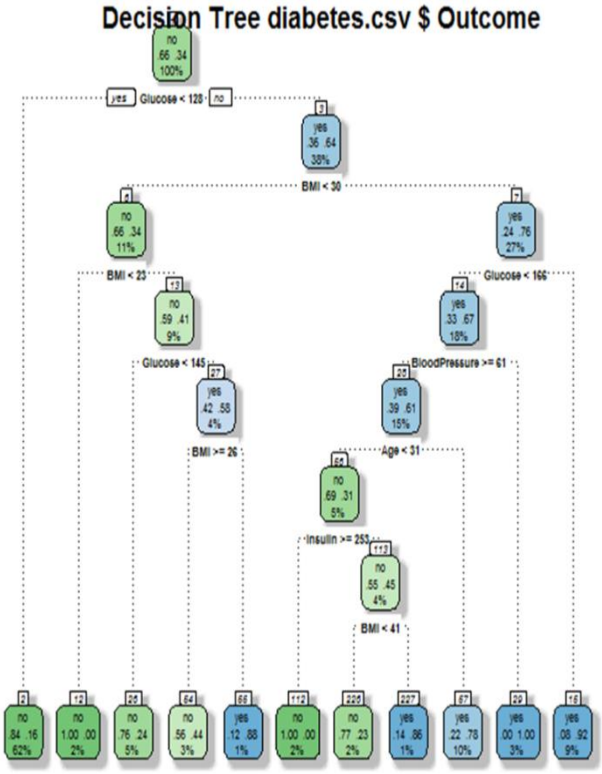


Fig 4: Decision Tree

Error matrix of decision tree classifier for our data is-

Table 7

Actual	Predicted		
	No	yes	Error
No	65	6	8.5
Yes	28	16	63.6

Out comes from the error matrix is in table 8

Table 8: Result from decision model

True positive rate	0.636
False positive rate	0.085
Precision	0.0289
Recall	0.636

C. For the Random Forest Tree model result, we have to classify our data with this model. For this we have to apply this model in the rattle with the number of trees 500 and 2 variables and then code in R console for Random forest model.

Error matrix of Random Forest tree classifier for our data is-

Table 9

Actual	Predicted		
	no	yes	Error
No	59	12	16.9
Yes	21	23	47.7

Out comes from the error matrix is in table 10

Table 10: Result from Random Forest Model

True positive rate	0.169
False positive rate	0.477
Precision	0.3428
Recall	0.169
F-measure	0.00933
ROC Area	0.79

V. ACCURACY BY THE RATTLE TOOL

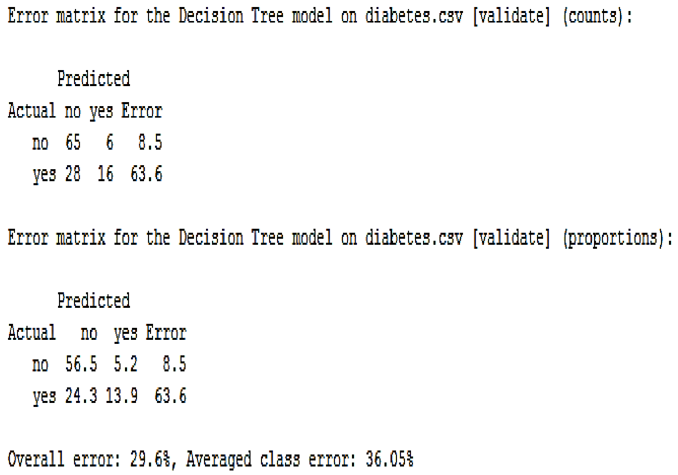
We have applied our classifier algorithm on our data and the result displayed. Percentage of accuracy is coming from the classifier which is displayed in Table 11 which is coming from the output of our classifiers.

This table is constructed on the basis of output by the RATTLE tool on applying Decision Tree, random Tree classification algorithm.

Table 11: Accuracy by Rattle Tool

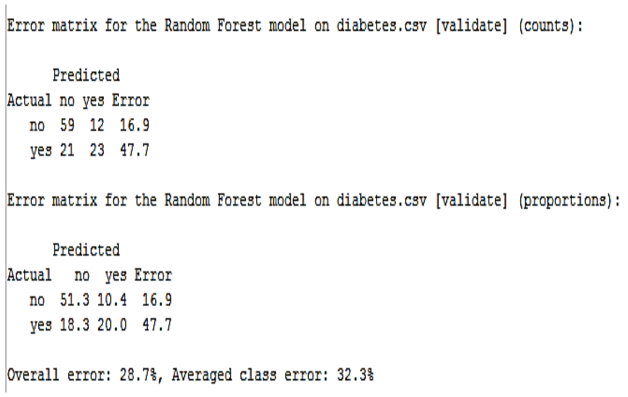
RATTLE	Decision tree	Random Forest Tree
	70.43 %	71.30 %

Error matrix is divided into two parts one is predicted and other is actual. In our output of the confusion matrix there are two attribute one is no and second is yes. So if we have m attribute then the confusion matrix will be m x m. Here we have m=2, our confusion matrix will be 2x2.



```
Error matrix for the Decision Tree model on diabetes.csv [validate] (counts):

      Predicted
Actual no yes Error
   no  65   6   8.5
   yes 28  16  63.6

Error matrix for the Decision Tree model on diabetes.csv [validate] (proportions):

      Predicted
Actual   no  yes Error
   no  56.5  5.2   8.5
   yes 24.3 13.9  63.6

Overall error: 29.6%, Averaged class error: 36.05%
```

Fig 5: Error Matrix for Decision Tree Model



```
Error matrix for the Random Forest model on diabetes.csv [validate] (counts):

      Predicted
Actual no yes Error
   no  59  12  16.9
   yes 21  23  47.7

Error matrix for the Random Forest model on diabetes.csv [validate] (proportions):

      Predicted
Actual   no  yes Error
   no  51.3 10.4  16.9
   yes 18.3 20.0  47.7

Overall error: 28.7%, Averaged class error: 32.3%
```

Fig 6: Error Matrix for Random Forest Model

If we combine all the error matrices then the output will show as below-

Table 12: Combined Error Matrix

Rattle	**Decision Tree**		**Forest Tree**	
	NO	**YES**	**NO**	**Yes**
NO	65	6	59	12
YES	28	16	21	23

VI. CONCLUSION AND FUTURE WORK

From the both model is clear that the accuracy of decision tree is 70.43% and overall error is 29.6 % and average class error is 36.05% and by the Random forest tree model is 71.30% and overall error is 28.7% and average class error is 32.3%so by this result we can say that for our data the random forest tree gives the best result than decision tree.

We have try to conclude the result on the basis of two classifier decision tree and random forest tree but if we try with the SVM , neural network and regression or other classifier it may be we get some other good result also. We did only classification in this work but we can also apply clustering technique for the clustering the data and get the desired result correspond to it.

VII. ACKNOWLEDGEMENT

I am highly grateful and thankful to Dr. Ankur Garg Sir, Professor, Department of Computer Science & Engg., MIET, Meerut, for their strict control, ongoing support, inspiration & guidance, which ensure the value of my research work. Work under his guidance was enriched by experience. In last, I express gratitude to my parents and my siblings for their lot of encouragement, love and inspiration.

REFERENCES

1. S.Vijiyarani S.Sudha,“ Disease Prediction in Data Mining Technique”– A Survey,International Journal of Computer Applications & Information Technology Vol. II, Issue I, January 2013 (ISSN: 2278-7720)
2. Prof.Sumathy, Prof.Mythili, Dr.Praveen Kumar, Jishnujit T M, K Ranjith Kumar, “Diagnosis of Diabetes Mellitus based on Risk Factors”, International Journal of Computer Applications, Vol.10, Issue No.4, November.2010
3. Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvarda,” Machine Learning and Data Mining Methods in Diabetes Research”, Computational and Structural Biotechnology Journal 15 (8 January 2017) 104–116
4. Miss. N. Vijayalakshmi, Miss. T. Jenifer, “An Analysis of Risk Factors For Diabetes Using Data Mining Approach”, IJCSMC, Vol. 6, Issue. 7, July 2017, pg.166 – 172
5. S.Selvakumar, K.Senthamarai Kannan, S.GothaiNachiyar,” Prediction of Diabetes Diagnosis Using Classification Based Data Mining Techniques”, International Journal of Statistics and Systems ISSN 0973-2675 Volume 12, Number 2 (2017), pp. 183-188.
6. P.Yasodha, M. Kannan, “Analysis of a Population of Diabetic Patients Databases in WEKA Tool”. International Journal of Scientific & Engineering Research, Volume 2, Issue 5, May-2011 ISSN 2229-5518 Analysis of a Population of Diabetic Patients Databases in WEKA Tool.
7. Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2nd edition, 2006.
8. [8] Assal, J. P., and L. Groop. "Definition, diagnosis and classification of diabetes mellitus and its complications." World Health Organization (1999): 1-65.
9. [9].Koh, Hian Chye, and Gerald Tan. "Data mining applications in healthcare."Journal of healthcare information management 19.2 (2011): 65.