International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-1, November 2019

Experimental Evaluation of Open Source Data Mining Tools: R, Rapid Miner and KNIME

Hemlata, Preeti Gulia

Abstract: ***In the current scenario of Big Data, open source Data Mining tools are very popular in business data analytics. The paper presents a comprehensive study of three most popular open source data mining tools – R, RapidMiner and KNIME. The tools are compared by implementing them on two real datasets. Performance is evaluated by creating a decision tree of the datasets taken. Our objective is to find the best tool for classification. The study can help researchers, developers and users in selecting a tool for accuracy in their data analysis and prediction. Experiments depict that accuracy level of the tool changes with the quantity and quality of the dataset. The results show that RapidMiner is the best tool followed by KNIME and R.***

Keywords **: *Classification, Data Mining tools, Decision tree, KNIME , R, RapidMiner.***

I. INTRODUCTION

Data mining is the fundamental part of knowledge discovery. It is the activity of finding new and useful information from the available vast and unorganized data. For extracting the possible information in an efficient way the data should be prepared [1]. After preparation, different models are built and standard statistical procedures are followed for analysis. Now-a-days, a number of big data mining tools and software are available for mining knowledge from Big Data.

Open source data mining software tools are available for Big Data Analytics. Recently in 2017, KD Nugget Software poll, R and RapidMiner were among the top 5 used data mining tools list [2]. The poll shows that more than 45% users use R language consistently from 2015- 2017. More than 30% users use Rapid Miner as a Big Data Analytical tool. KNIME was used by approximately 20% users. Fig. 1 shows the poll results.

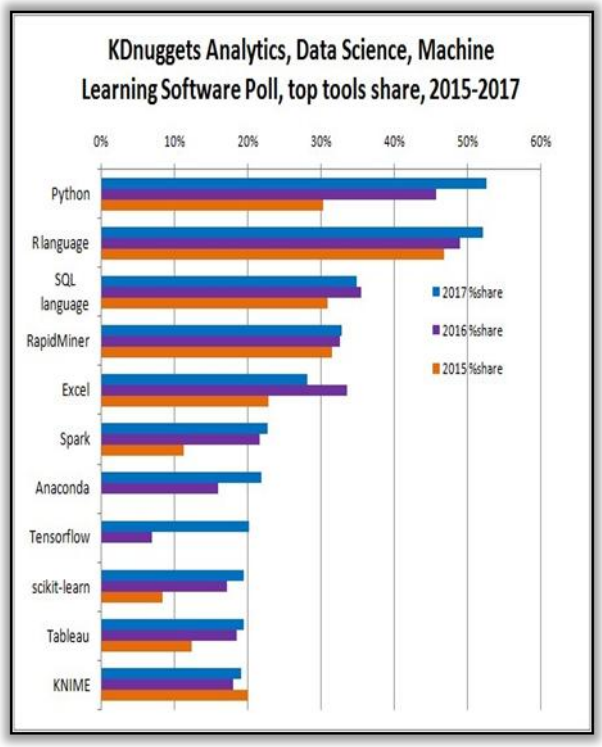


Fig. 1 KDnuggets Analytics/Data Science 2017 Software Poll

This paper focuses on the three tools – R, Rapid Miner, KNIME in terms of data classification. Most important data classification method- decision tree is used as a technique for comparison of the tools. The main idea behind selecting the tools is their popularity among Big Data users and their implementation in general data mining tasks.

The paper is organized as follows: Section II summarizes the related work done in the area. Methodology of the experiments conducted is explained Section III. Section IV presents the experimental results along with the experimental setup. Section V presents the Result Evaluation. Conclusion is presented in Section VI.

II. RELATED WORK

The Nurdatillah Hasim et.al [3] had studied various open source data mining tools for forecasting. Important features and functionalities of Weka, RapidMiner, KEEL, Orange and Tanagra were presented in the paper. It was concluded that Weka and RapidMiner were most flexible and functional tools.

Revised Manuscript Received on November 30, 2019.
***** Correspondence Author
Hemlata*, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, India. Email: hemlatachahal@gmail.com ORCID: 0000-0002-6105-7399
Preeti Gulia, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, India. Email: preeti.gulia81@gmail.com ORCID: 0000-0001-8535-4016

Magdalena Graczyk et.al [4] had compared various machine learning algorithms for building models and implemented them in KEEL, RapidMiner and Weka. Some common methods like decision trees, neural network and support vector machine were exercised on real datasets. Differences were found between different models after using various performance measures. Evaluation of the performance of four open source data mining tools- KNIME, RapidMiner, Weka and Orange was done by Luís C. Borges et.al [5]. The main aim was to find the best tool for classification. The characteristics, platform used, advantages and disadvantages of the most important open source data mining tools - KNIME, R, RapidMiner were presented by Hemlata et.al [6]. Analysis had shown that R is the best tool among the three tools.

Angela Lausch et.al [7] presented an outline of the data mining techniques and tools. Analysis was done by example implementation and concluded that Rapidminer and KNIME tools were best suited for the analysts who do not have much experience in programming. Linked Open Data (LOD) approach was presented as a new possibility for data mining analysis. The characteristics of frequently used free software tools were described by Alan Jovic et.al [8]. Analysis was done by implementing various algorithms in different data mining areas. According to analysis, RapidMiner, R, Weka, and KNIME were the best data mining and analytical tools.

Ahmad Al-Khoder et.al [9] evaluated four open source data mining tools – R, RapidMiner, Weka and KNIME for choosing the tool in research and analytics. The performance was evaluated by using three classification algorithms – Naive Bayes (NB), Decision Tree (DT) and K- nearest Neighbour (KNN). It was concluded that R is the best tool in terms of visualisation and formats of input and output. Weka was best in terms of accuracy and performance. Neha Chauhan et.al [10] had presented the pros and cons and comparison of various data mining tools – RapidMiner, KNIME and Weka. The comparison was done on the basis of different parameters of the tools. The study concluded that RapidMiner is best in all respects.

A recommendation system and automated estimation process to plan software development using R- scripts had been proposed by Jaideep Jagadeeshwar et.al [11]. The R-scripts can be reused for future prediction. Hilda Kosorus et.al [12] presented a comparison of data mining tools – R, Weka and RapidMiner by using them for time series analysis for sensor data in health field.

Table- I: Comparison of the tools used in the study

Tools	R	RapidMiner	KNIME
Logo			
Features	• Statistical package for statistical computing. • Freeware substitution for SPSS. • Suitable for analysis, graphical and software development activities.	• Provides machine learning environment • Compatible with SPSS, Excel, MySQL and other database softwares. • Specially used for predictive analysis and statistical computing.	• Data flow or pipeline is visually createdfor users. • Cross validation functionality • Specifically used for Business Intelligence and Analytics
Advantages	• Vast statistical library. • Data can be imported and exported from Excel spreadsheet. • Better numerical programming.	• Many procedures for attribute selection are offered. • Cross validation and independent validation are used for evaluating various models. • Very flexible.	• Easy to use. • New nodes are created by dragging and dropping. • Provides interface for visualisation and analysis.
Disadvantages	• Suitable for persons with basic knowledge of programming. • Does not have much control over the details. • Less specialised with data mining.	• Limited partitioning capability. • Suitable for people who can work on database file	• Not fit for complex workflows. • Dataset partitioning is limited. • Error calculation ability is limited.

III. METHODOLOGY USED IN THE STUDY

The present study follows three steps:

a) Dataset selection.

b) To select the Open Source Data Mining tools
c) To select the evaluation technique

The study is done by the combination of the datasets, techniques of evaluation and algorithms. The tools are compared with the help of accuracy metric.

A. Data Sets

The datasets were downloaded from online UCI Machine Learning Repository [13] for the purpose of performance analysis of three most famous open source data mining tools: R, KNIME and RapidMiner. Table II shows the details of the datasets: name, variable data type, number of instances, number of attributes and number of class attribute values.

Table- II: Dataset Details

Dataset Name	Variable Data Type	# Instances	# Attributes	# Class Values
Iris	Real	150	4	3
Census	Integer, Polynomial	31978	13	2

B. Tools for Analysis

Three most popular data mining tools have been selected- R, KNIME and RapidMiner for the analysis. These tools are selected for analysis because these are the most popular tools among users and are frequently used for analysis of Big Data. User friendliness is also taken into consideration while selecting them.

C. Evaluation Technique

Data classification is selected as the technique of evaluation of various selected tools using the selected datasets. The process of data classification is:

i. The training step
ii. The testing step

In the training step, a new model is created with the help of existing or actual value of class variable. In testing step, the model is tested to predict the value of class after which the predicted values are compared with the actual values.

D. Classification Technique

Mostly classification is done by supervised learning. In the study decision tree algorithm is used in all the three open source data mining tools by using both the datsets.

E. Performance Evaluation

The decision tree is evaluated by accuracy metric. Accuracy metric is constructed by comparing the predicted with the actual. The number of instances correctly classified is divided by the total instances for calculation of accuracy metric.

IV. EXPERIMENTAL RESULTS

In this section the experimental setup of the tests conducted and its results are explained. Various tools(R, KNIME, RapidMiner) are evaluated by constructing decision trees using the datasets given in Table II.

A. Experimental Setup

The experiments are conducted on a computer running Windows 2007 with an Intel Core i3 M350 processor and 4 GB of main memory. For conducting experiments three open source softwares were installed on the machine:-

- R version 3.4.0 (2017-04-21) Copyright (C) 2017 The R Foundation for Statistical Computing. Platform: x86_64-w64-mingw32/x64 (64-bit). It can be freely downloaded from [14].
- RapidMiner Studio Free 7.5.001 Version 7.5 Copyright (C) 2001-2017 RapidMiner GmbH. It can be freely downloaded from [15].
- KNIME Analytics Platform version 3.1.0 Copyright (C) KNIME GmbH, Konstanz, Germany. It can be freely downloaded from [16].

In the tests random sampling with partition mode is used. While conducting the tests, the algorithm is used with its default parameter values except when it is possible to ignore the missing values present in the datasets.

B. R Results

Iris dataset is already available in the R software but Census dataset is imported in R. Both are shown in fig 2 and fig 3. Iris dataset describes the characteristics of flowers. The parameters taken are sepal length, sepal width, petal length and petal width. On the basis of these parameters species of flowers are predicted i.e. setosa, virginica and versicolor. Census dataset is about the personal details of people like age, education, marital status etc. for predicting the salary of people.

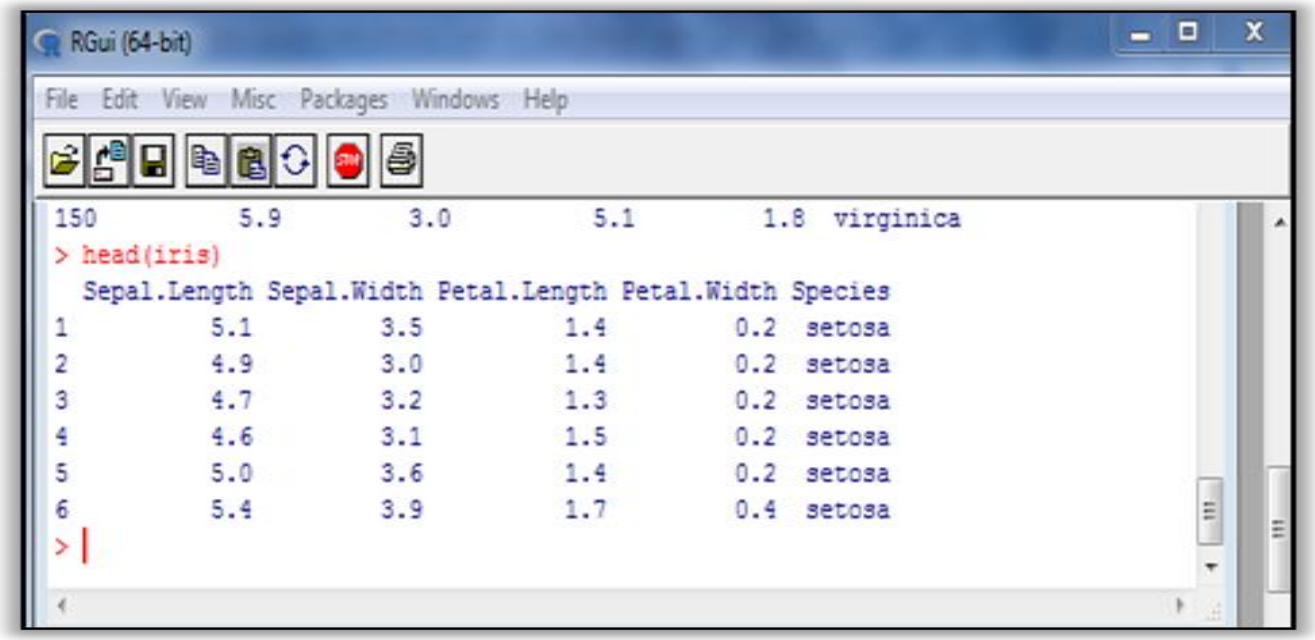


Fig. 2 Iris Dataset(in R)

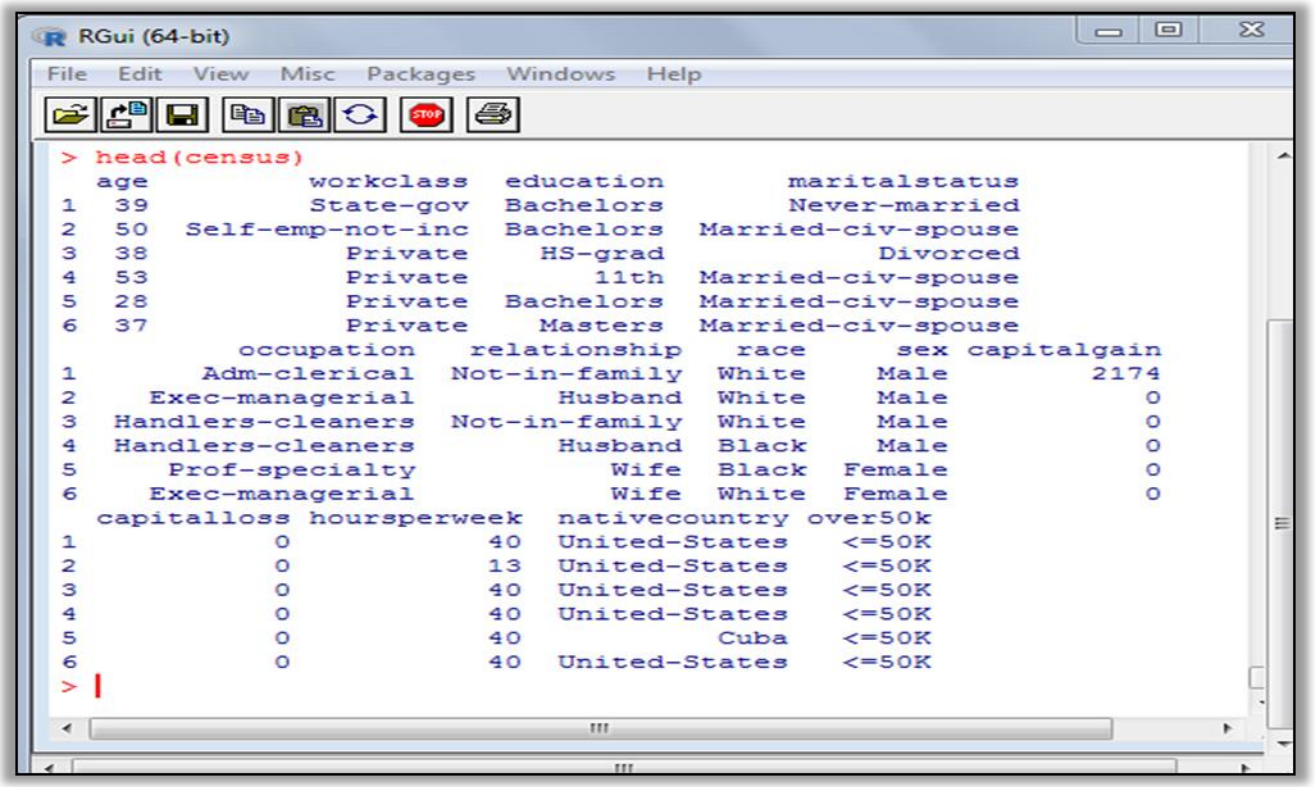


Fig. 3 Census dataset (in R)

A new model is created by taking training datasets of 105 instances from Iris and 19186 instances from Census. The models are created by installing C50 package and using its C5.0 Rfunction (implementation of C5.0 algorithm of creating decision tree in R). Fig 4 and Fig 5 show the creation of decision trees for both the datasets.

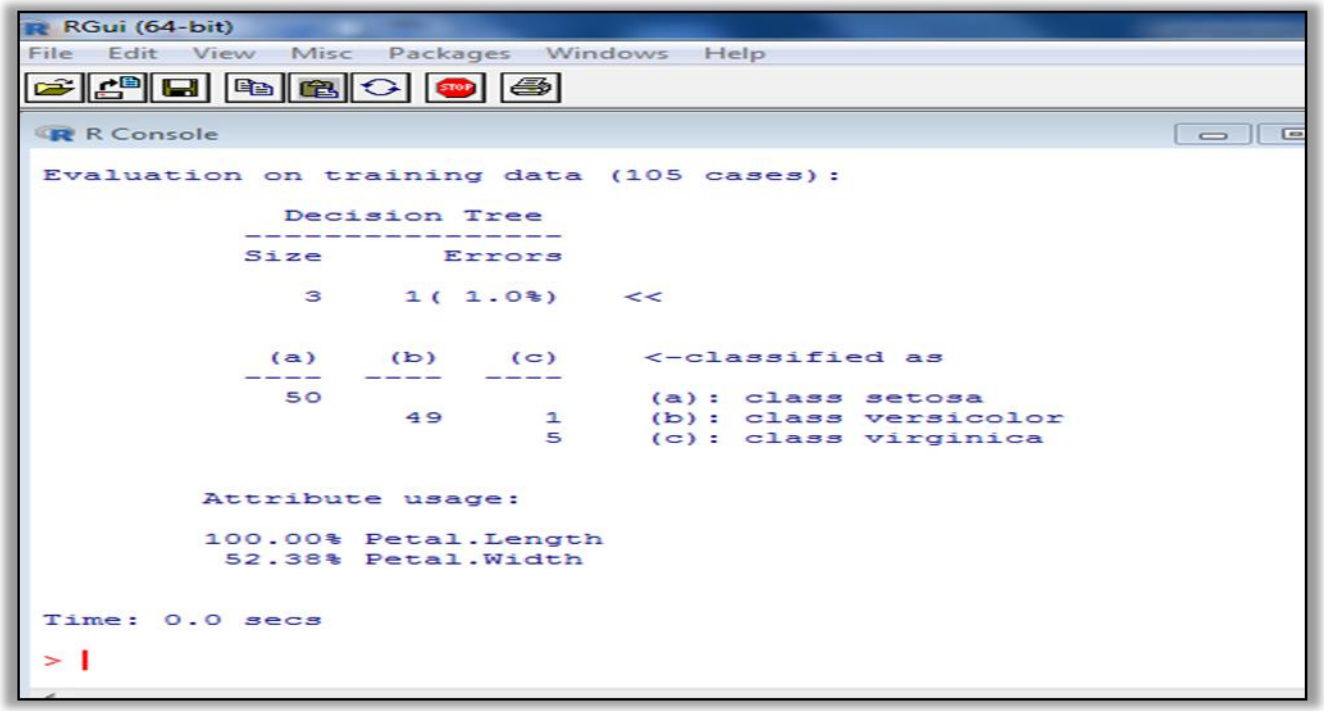


Fig. 4 Decision Tree of training dataset of Iris (in R)

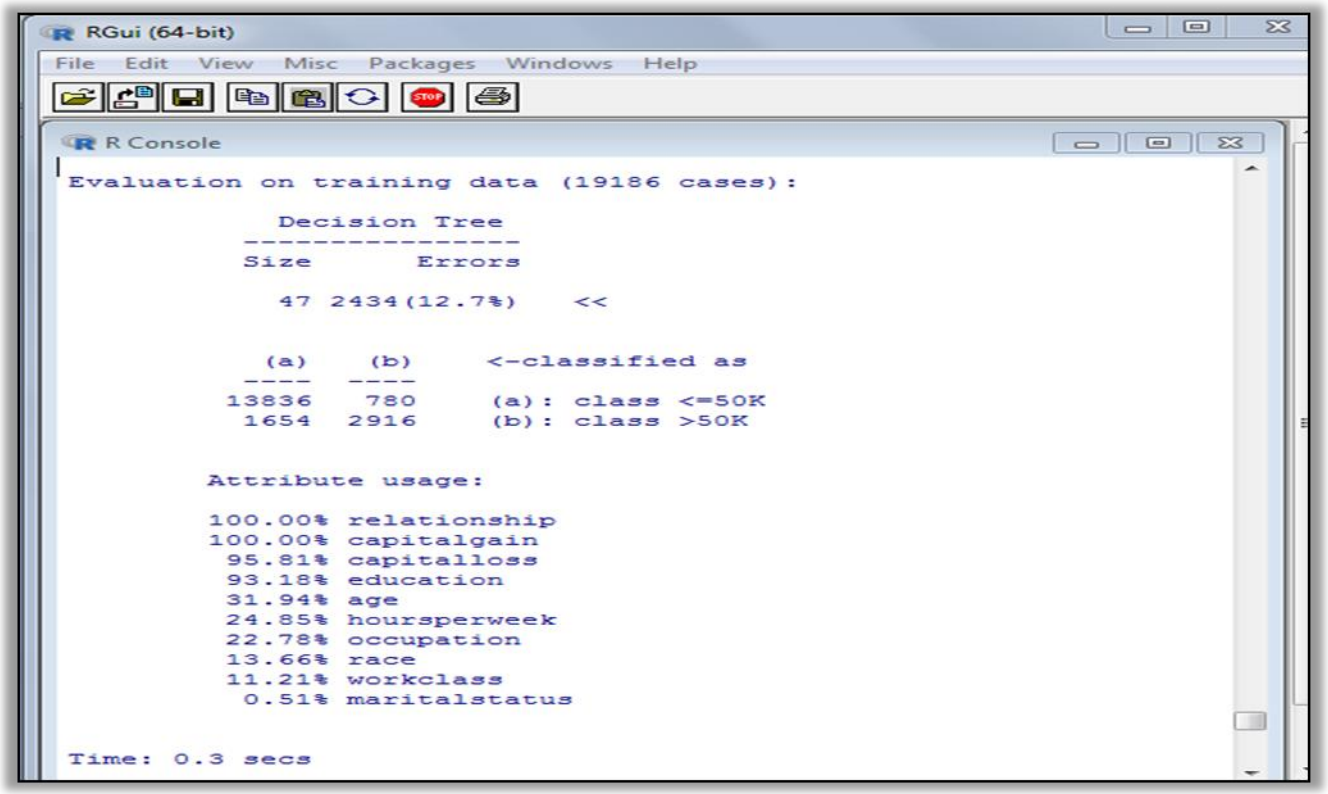


Fig. 5 Decision Tree of training dataset of Census (in R)

Results show that the error in Iris dataset is very small i.e. 1% but in Census it is considerable i.e. 12.5%. The size of the decision tree of Iris dataset is three i.e. the level of the tree comes out to be three whereas the size of tree of census dataset is 47. The tree here is complex as the levels of the tree is very high which is very difficult to analyse.

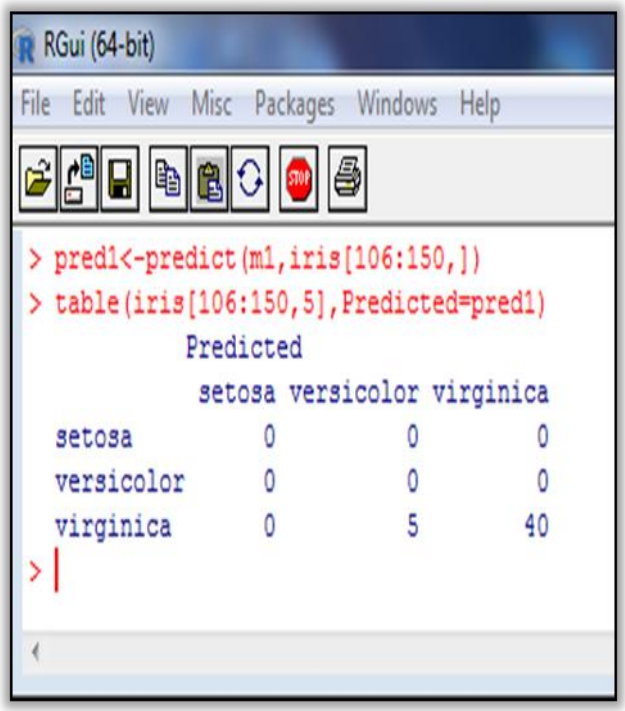


Fig. 6 Confusion matrix of Iris (in R)

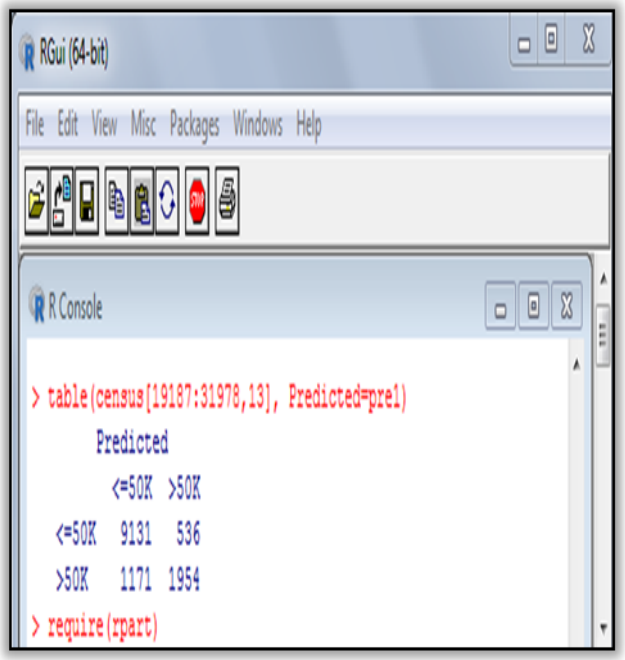


Fig. 7 Confusion matrix of Census (in R)

Confusion matrix (Fig 6) of Iris depicts that only 5 instances out of 45(test data) are misclassified. There are five instances whose actual species is virginica but is predicted as versicolor. Confusion matrix (Fig 7) of Census depicts that 1707 instances out of 12791(test cases) are misclassified. There are 1171 instances whose actual salary is >50K but predicted to be <=50K and 536 instances whose actual salary is <=50K and predicted to be >50K. So the accuracy metric in Iris comes out to be 89% and in Census it comes out to be 86.7%. Fig 8 and fig 9 shows the decision trees of both the datasets graphically. Iris dataset is depicted with only three leaf nodes and two inner nodes, whereas, Census dataset is depicted with five leaf nodes and four inner nodes.

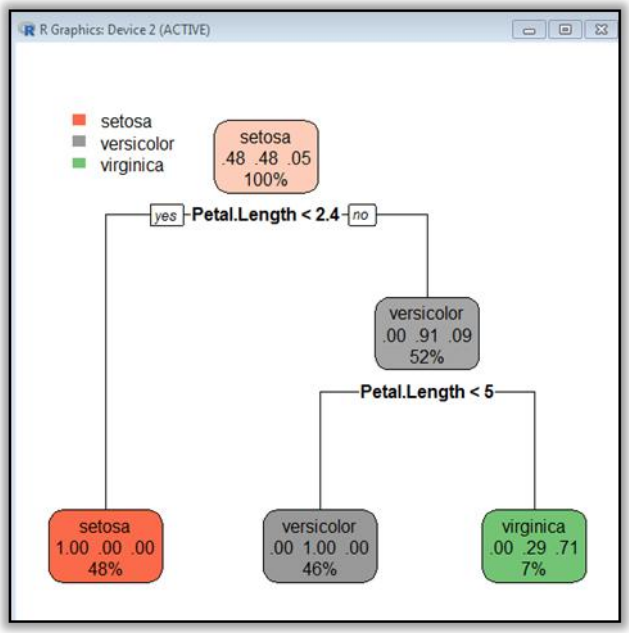


Fig. 8 Decision Tree of Iris (in R)

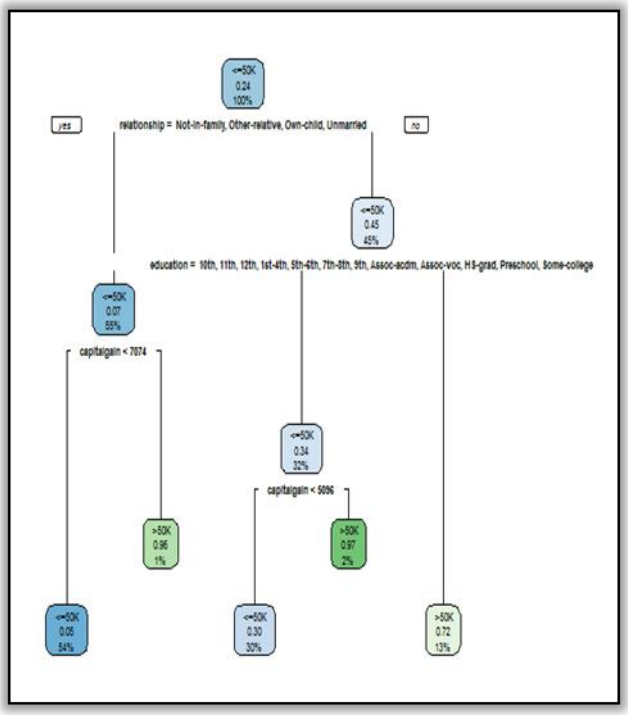


Fig. 9 Decision Tree of Census (in R)

C. RapidMiner Results

Iris dataset is already available in the sample data folder of RapidMiner and Census dataset is added in local repository by following steps to add data in local repository. Two operators- read csv file and split validation are added on the process area and the required input-output lines are connected. In split validation operator split ratio is set to be 0.9 and sampling type is chosen as stratified sampling. Fig. 10 shows the process of Iris and fig 11 shows the process of Census.

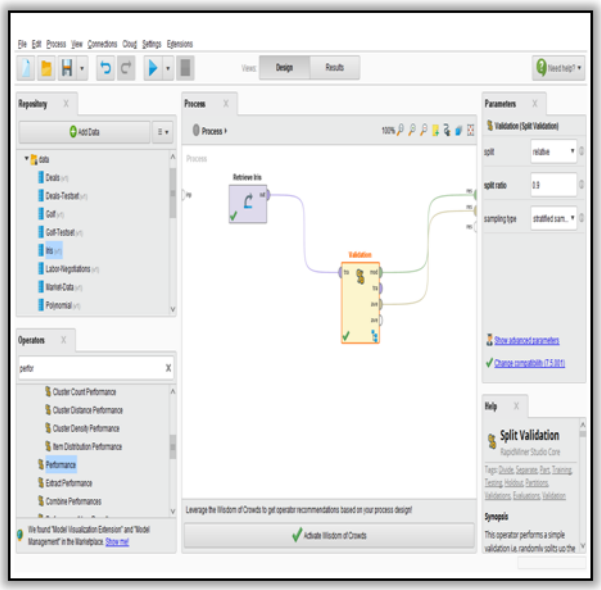


Fig. 10 Iris dataset Process (in RapidMiner)

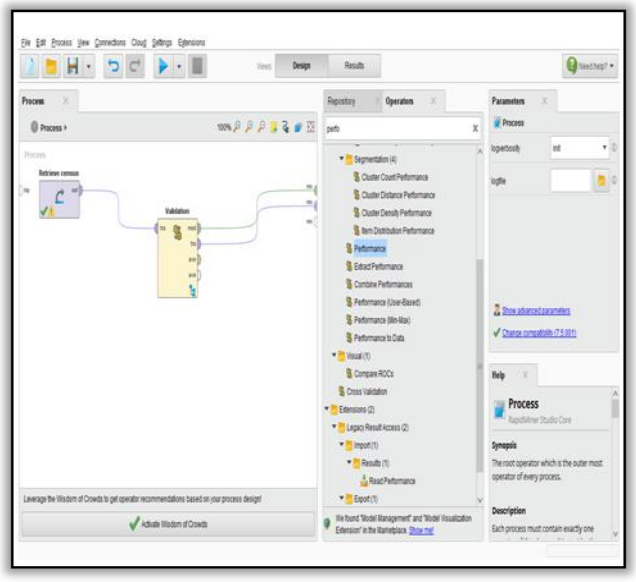


Fig. 11 Census Dataset Process (in RapidMiner)

Split Validation operator splits the datasets in two parts- training data and test data in 90:10 ratio. In train data classification operator decision tree is added and in test data Apply Model operator is added. For evaluating the performance of the model Performance operator is also added in test data. It creates the confusion matrix of the model. By running the process, a decision tree and confusion matrix is created which are shown in fig 12, fig13, fig 14 and fig 15. Iris dataset has five leaf nodes and four inner nodes. Census dataset has fifteen leaf nodes and fourteen inner nodes which, in number, is much large than R.

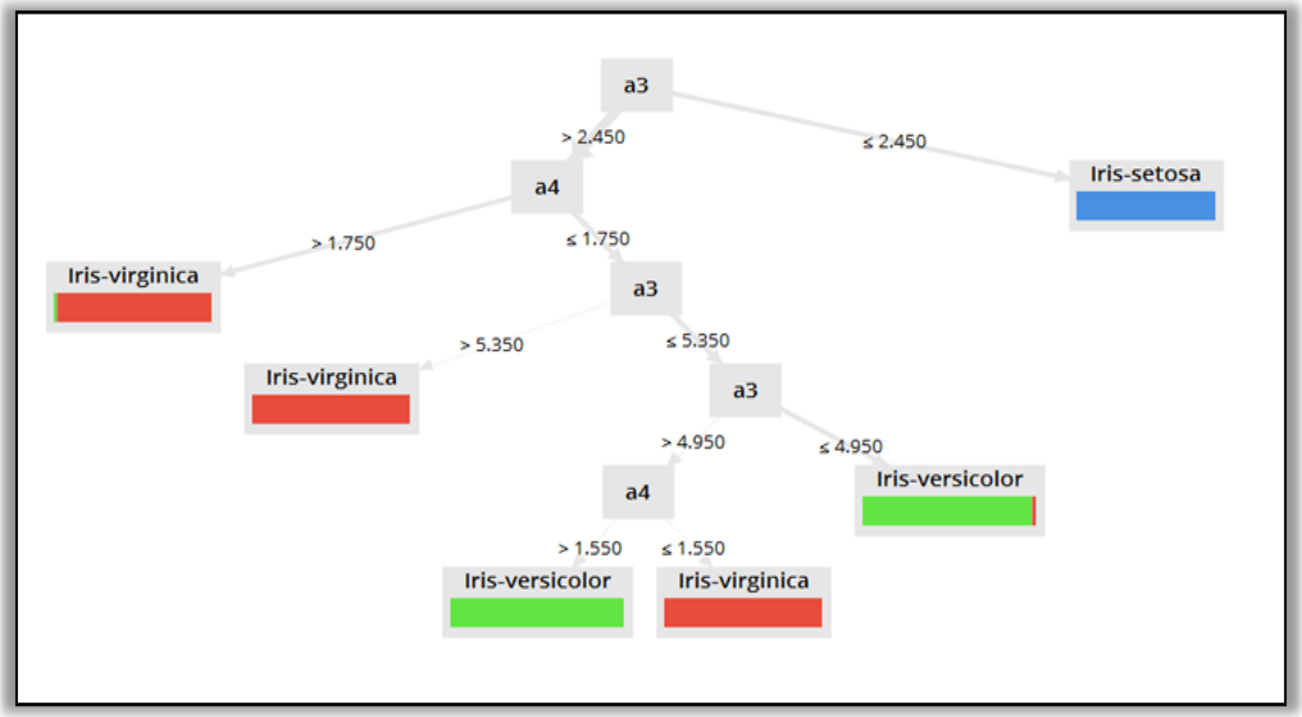


Fig. 12 Iris Decision Tree (in RapidMiner)

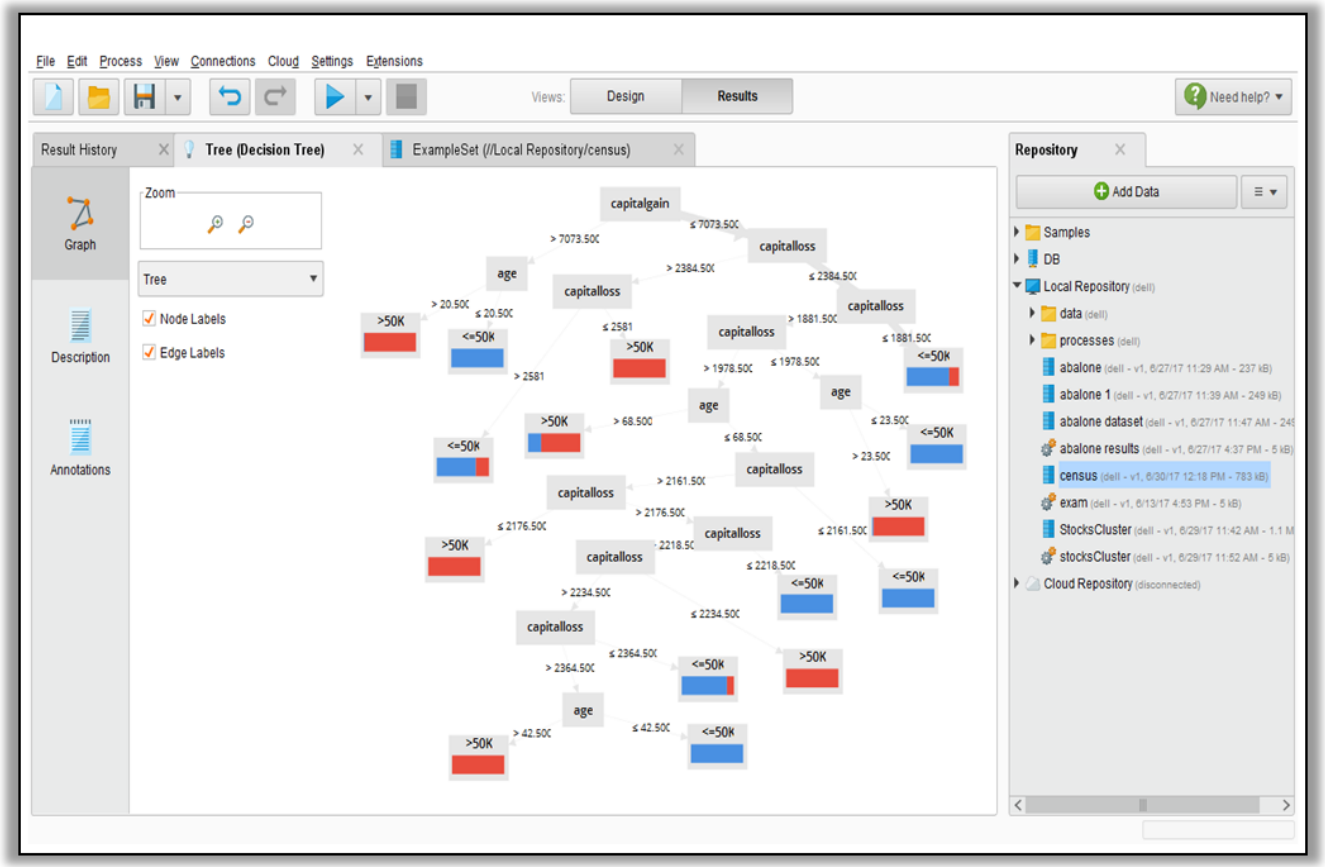


Fig. 13 Census Decision Tree (in RapidMiner)

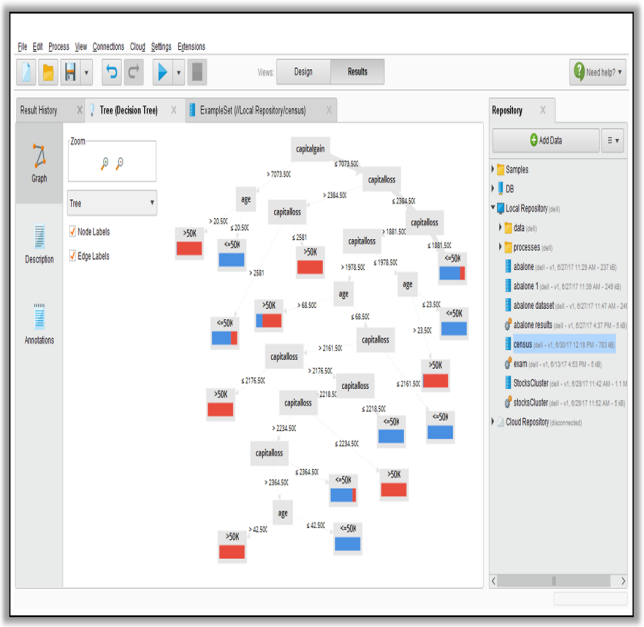


Fig. 14 Iris Confusion Matrix (in RapidMiner)

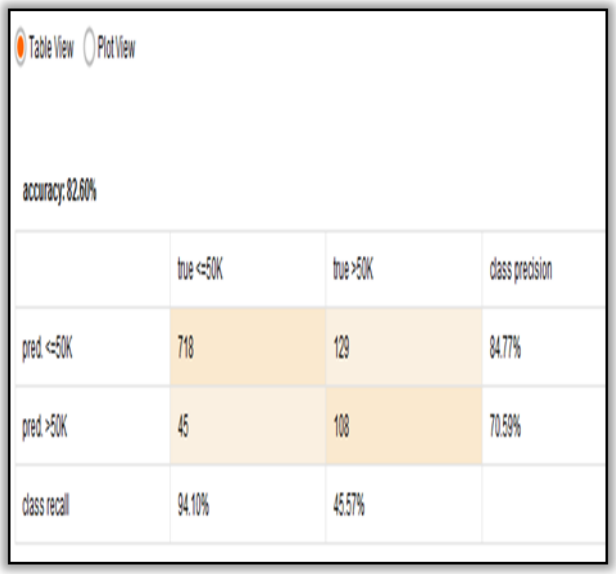
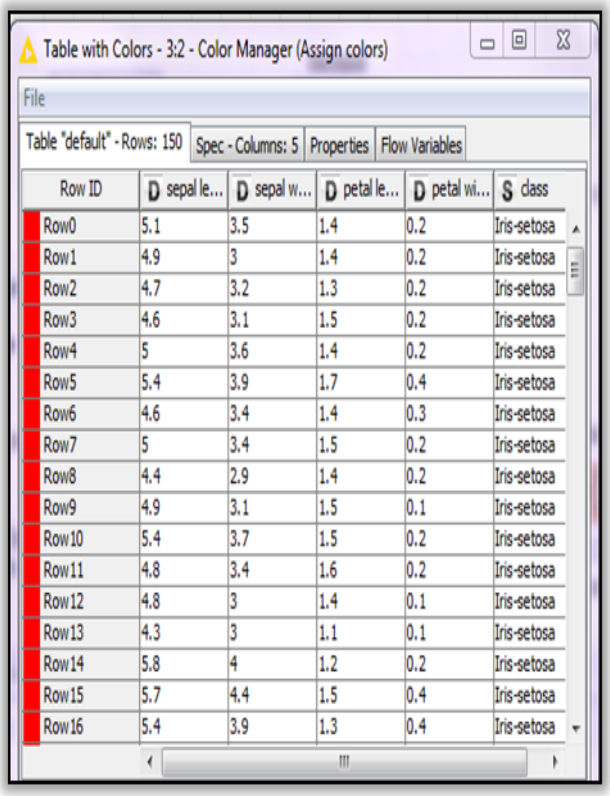


Fig. 15 Census Confusion Matrix (in RapidMiner)

Results show that accuracy of Iris dataset is 100% but accuracy percentage of Census dataset is 82.6% which means that as the dataset increases the accuracy level decreases.

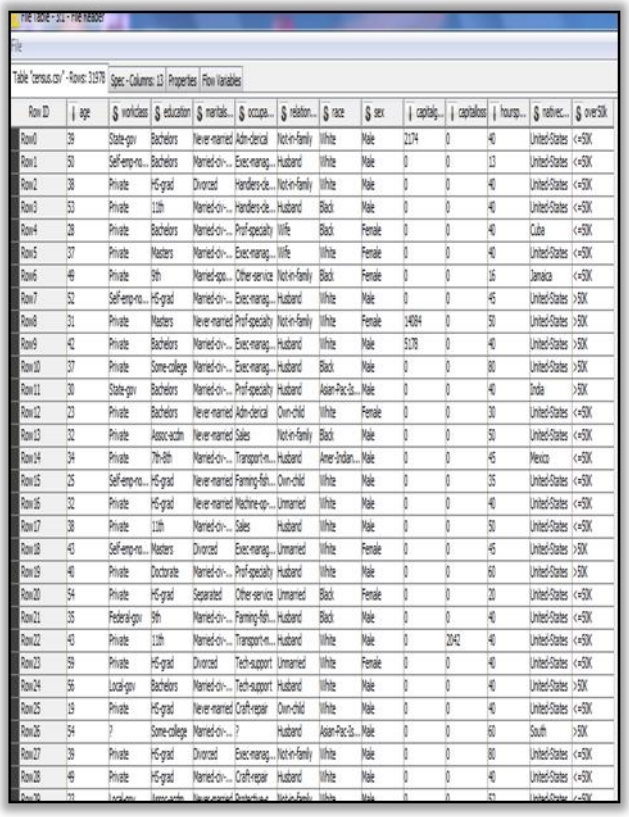
D. KNIME Results

Iris dataset is already available in the sample data folder of KNIME and Census dataset is imported through read .csv file operator. The data tables in both Iris and Census datasets are presented in fig. 16 and 17.



Row ID	D sepal le...	D sepal w...	D petal le...	D petal wi...	S class
Row0	5.1	3.5	1.4	0.2	Iris-setosa
Row1	4.9	3	1.4	0.2	Iris-setosa
Row2	4.7	3.2	1.3	0.2	Iris-setosa
Row3	4.6	3.1	1.5	0.2	Iris-setosa
Row4	5	3.6	1.4	0.2	Iris-setosa
Row5	5.4	3.9	1.7	0.4	Iris-setosa
Row6	4.6	3.4	1.4	0.3	Iris-setosa
Row7	5	3.4	1.5	0.2	Iris-setosa
Row8	4.4	2.9	1.4	0.2	Iris-setosa
Row9	4.9	3.1	1.5	0.1	Iris-setosa
Row10	5.4	3.7	1.5	0.2	Iris-setosa
Row11	4.8	3.4	1.6	0.2	Iris-setosa
Row12	4.8	3	1.4	0.1	Iris-setosa
Row13	4.3	3	1.1	0.1	Iris-setosa
Row14	5.8	4	1.2	0.2	Iris-setosa
Row15	5.7	4.4	1.5	0.4	Iris-setosa
Row16	5.4	3.9	1.3	0.4	Iris-setosa

Fig. 16 Iris Dataset (in KNIME)



Row ID	age	workclass	education	marital...	occupa...	relation...	race	sex	capitalg...	capitalloss	hoursp...	native...	over50k
Row0	39	State-gov	Bachelors	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
Row1	50	Self-emp-no...	Bachelors	Married-civ...	Exec-manag...	Husband	White	Male	0	0	13	United-States	<=50K
Row2	38	Private	HS-grad	Divorced	Handlers-cle...	Not-in-family	White	Male	0	0	40	United-States	<=50K
Row3	53	Private	11th	Married-civ...	Handlers-cle...	Husband	Black	Male	0	0	40	United-States	<=50K
Row4	28	Private	Bachelors	Married-civ...	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
Row5	37	Private	Masters	Married-civ...	Exec-manag...	Wife	White	Female	0	0	40	United-States	<=50K
Row6	49	Private	9th	Married-spo...	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
Row7	52	Self-emp-no...	HS-grad	Married-civ...	Exec-manag...	Husband	White	Male	0	0	45	United-States	>50K
Row8	31	Private	Masters	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
Row9	42	Private	Bachelors	Married-civ...	Exec-manag...	Husband	White	Male	5178	0	40	United-States	>50K
Row10	37	Private	Some-college	Married-civ...	Exec-manag...	Husband	Black	Male	0	0	80	United-States	>50K
Row11	30	State-gov	Bachelors	Married-civ...	Prof-specialty	Husband	Asian-Pac-Is...	Male	0	0	40	India	>50K
Row12	23	Private	Bachelors	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
Row13	32	Private	Assoc-acdm	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
Row14	34	Private	7th-8th	Married-civ...	Transport-m...	Husband	Amer-Indian...	Male	0	0	45	Mexico	<=50K
Row15	25	Self-emp-no...	HS-grad	Never-married	Farming-fish...	Own-child	White	Male	0	0	35	United-States	<=50K
Row16	32	Private	HS-grad	Never-married	Machine-op...	Unmarried	White	Male	0	0	40	United-States	<=50K
Row17	38	Private	11th	Married-civ...	Sales	Husband	White	Male	0	0	50	United-States	<=50K
Row18	43	Self-emp-no...	Masters	Divorced	Exec-manag...	Unmarried	White	Female	0	0	45	United-States	>50K
Row19	40	Private	Doctorate	Married-civ...	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
Row20	54	Private	HS-grad	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
Row21	35	Federal-gov	9th	Married-civ...	Farming-fish...	Husband	Black	Male	0	0	40	United-States	<=50K
Row22	43	Private	11th	Married-civ...	Transport-m...	Husband	White	Male	0	2042	40	United-States	<=50K
Row23	59	Private	HS-grad	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K
Row24	56	Local-gov	Bachelors	Married-civ...	Tech-support	Husband	White	Male	0	0	40	United-States	>50K
Row25	19	Private	HS-grad	Never-married	Craft-repair	Own-child	White	Male	0	0	40	United-States	<=50K
Row26	54	?	Some-college	Married-civ...	?	Husband	Asian-Pac-Is...	Male	0	0	60	South	>50K
Row27	39	Private	HS-grad	Divorced	Exec-manag...	Not-in-family	White	Male	0	0	80	United-States	<=50K
Row28	49	Private	HS-grad	Married-civ...	Craft-repair	Husband	White	Male	0	0	40	United-States	<=50K

Fig. 17 Census Dataset (in KNIME)

Different operators were used like Color manager, Partitioning, Decision Tree Learner, Decision Tree Predictor, Scatter Plot and Scorer for analysis of the data. Partitioning operator splits the data in 90-10 ratio. 90% data is used as training dataset in decision tree learner operator and remaining 10% data is used as test data by decision tree predictor. The processes of both datasets are displayed in fig 18 and 19.

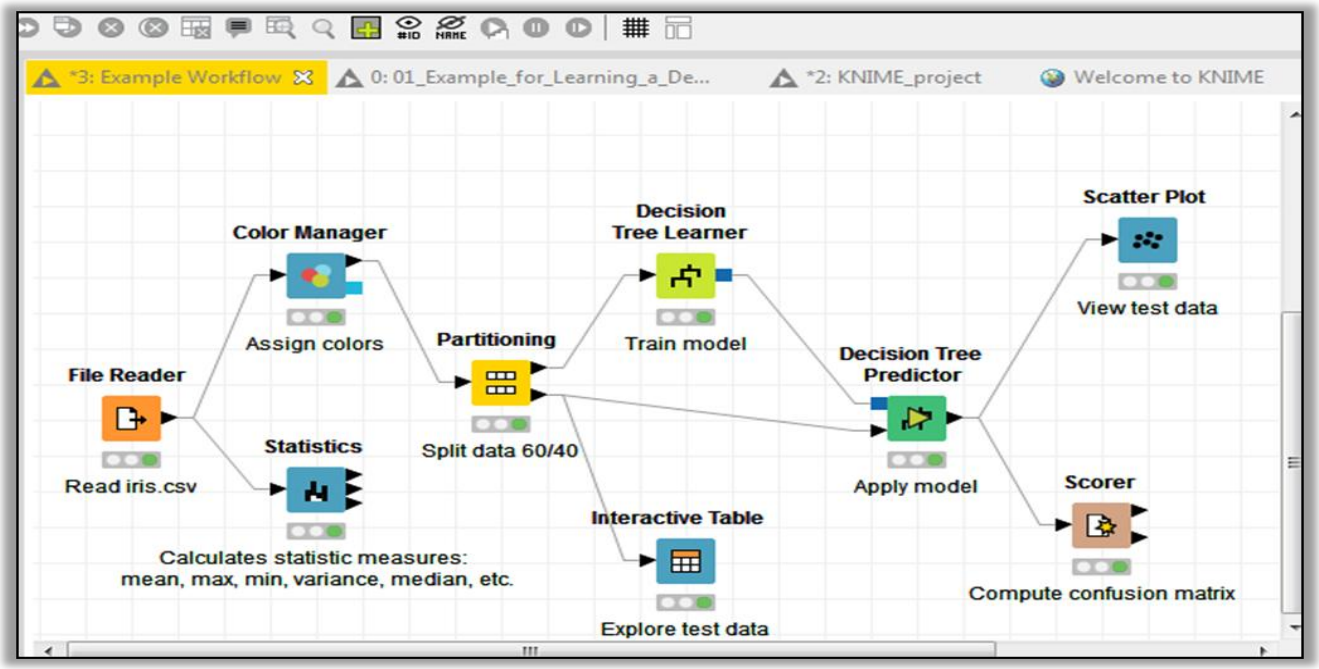


Fig. 18 Iris Dataset Process (in KNIME)

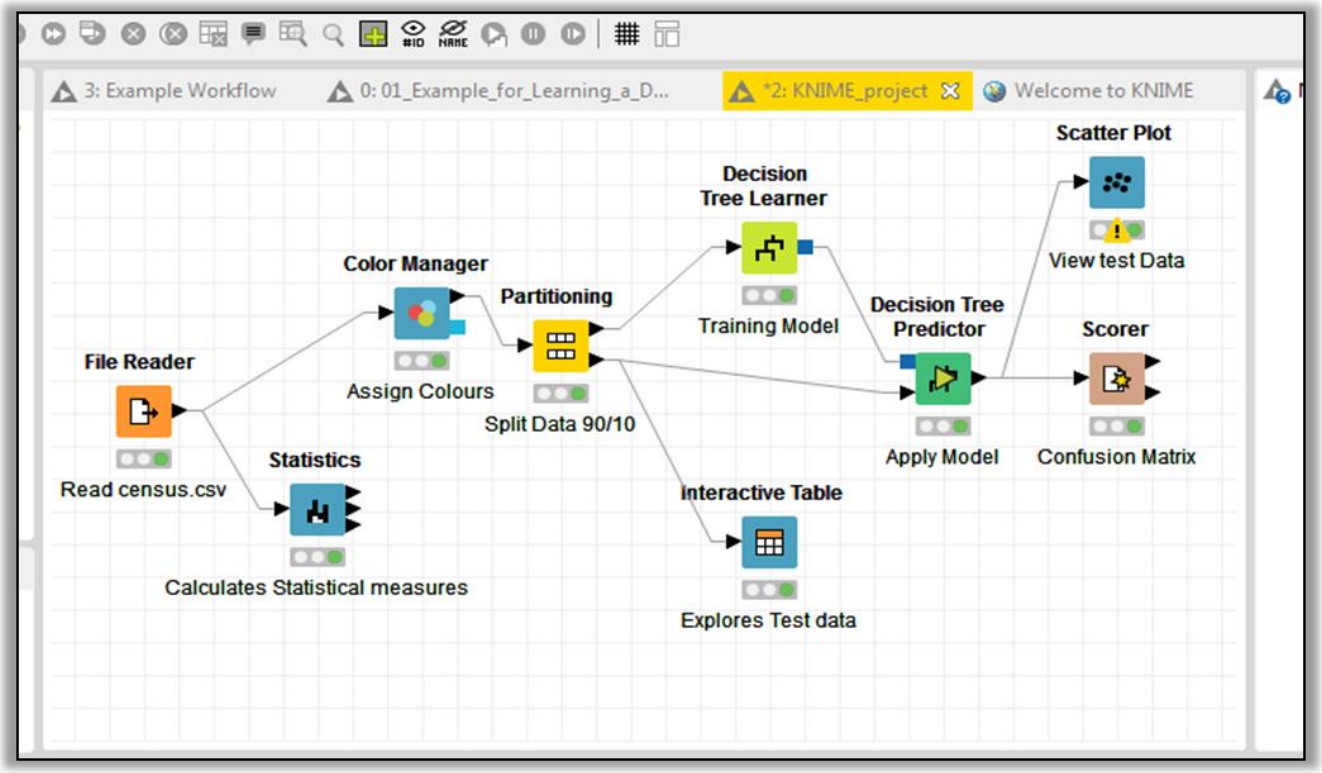


Fig. 19 Census Dataset Process (in KNIME)

Decision Tree learner creates a decision tree of 90% data and is displayed in fig 20 and fig 21. The tree of Iris dataset is much smaller than the tree of Census dataset because the instances of Iris are less as compared to Census.

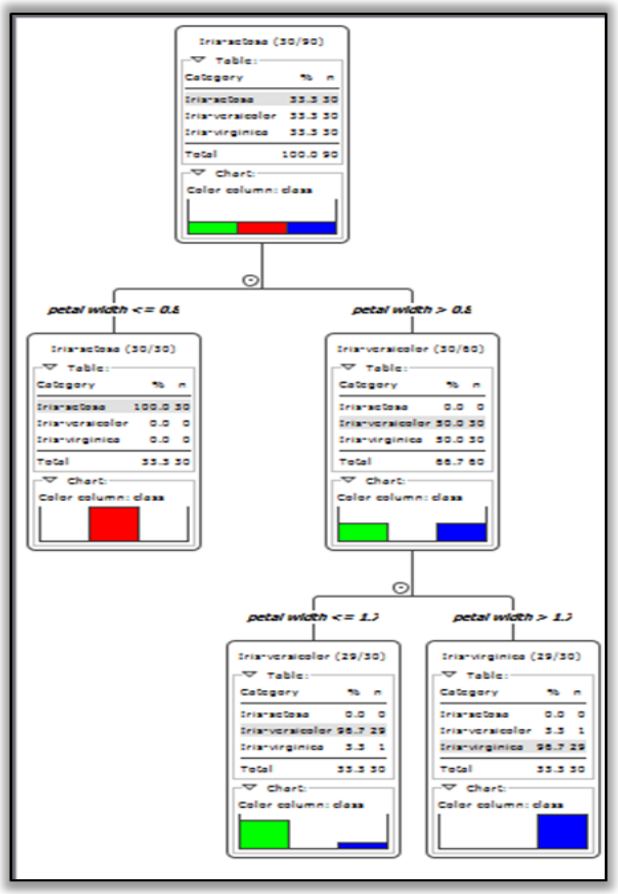


Fig. 20 Iris Dataset Decision Tree (in KNIME)

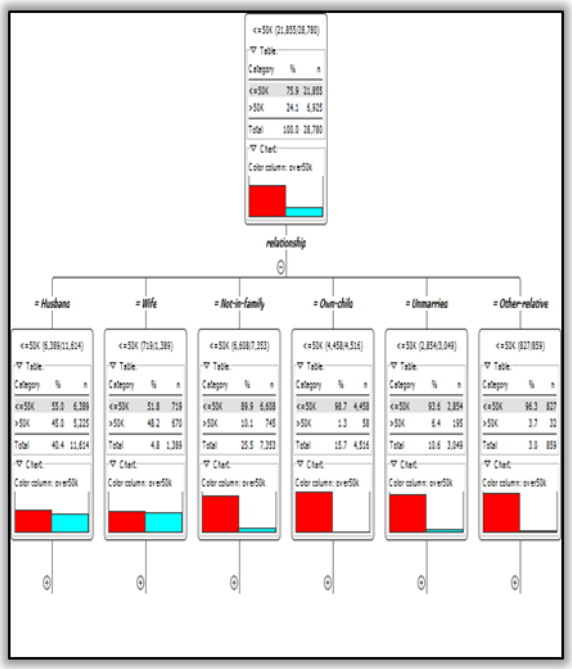


Fig. 21 Census Dataset Decision Tree (in KNIME)

Confusion Matrix is generated with the Scorer operator in KNIME. Confusion matrix of Iris dataset (Fig 22) shows that the accuracy level is 93.33% and Cohen's kappa value is 0.9.Four instances are misclassified i.e. four instances which are actually of virginica species and are predicted as versicolor species. Whereas confusion matrix of Census dataset (Fig 23) shows the accuracy level as 85.116% and Cohen's Kappa value is 0.55. Again as the dataset increases accuracy level decreases.

There are 335 instances whose salary was >50K are predicted to be the instances whose salary is <=50K and 141 instances whose actual salary is <=50K but is predicted to >50K.

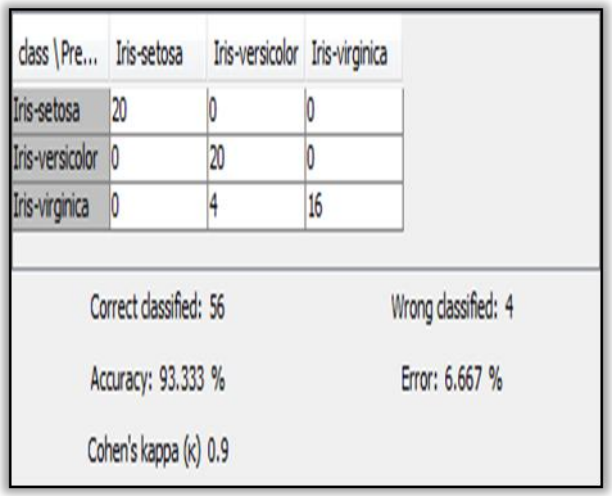


Fig. 22 Iris Confusion Matrix (in KNIME)

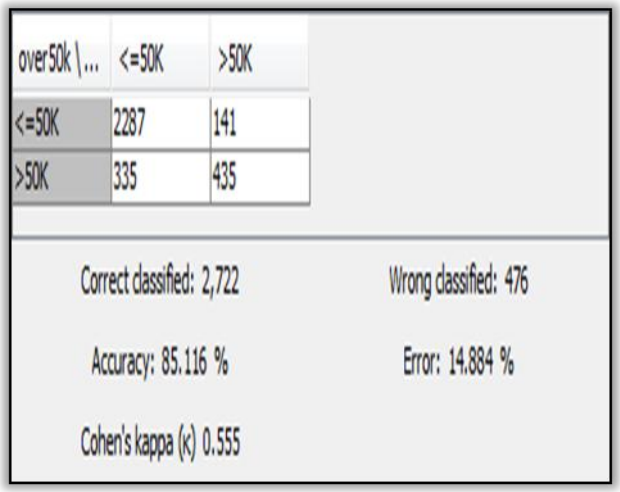


Fig. 23 Census Confusion Matrix (in KNIME)

V. EXPERIMENTAL RESULT EVALUATION

The analysis of the results can be performed by the datasets and the tools used. The results are calculated by simple arithmetic calculations. By considering the datasets shown in Table I, no clear relation could be drawn on the basis of type of variables, number of instances, number of values for the class attribute.

Table- III: Accuracy by Dataset

Dataset	Average Accuracy level
Iris	94.1%
Census	84.34%

Table III shows the accuracy level of Iris dataset is much higher than the Census dataset. Iris dataset has only 150 objects of 5 variables and Census dataset has 31978 objects of 13 variables. So it can be concluded that as the dataset increases accuracy level decreases.

Table- IV: Accuracy by Tool

Tool	Average Accuracy level
R	87.85%
RapidMiner	90.6%
KNIME	89.223%

On the basis of building decision trees, best accuracy is obtained by RapidMiner software as depicted in Table IV. Although all the three tools performed equally good but R shows the least accuracy level.

Table- V: Confusion Matrix of Iris Dataset

Tool	Instances Correctly Classified	Instances Wrong Classified	Error %
R	40	5	11.11%
RapidMiner	15	0	0%
KNIME	56	4	6.667%

Table- VI: Confusion Matrix of Census Dataset

Tool	Instances Correctly Classified	Instances Wrong Classified	Error %
R	11085	1707	13.34%
RapidMiner	812	188	18.8%
KNIME	2722	476	14.884%

Table V shows the confusion matrix of Iris dataset. It is concluded that RapidMiner tool is the best as it has 0% error. But according to Census dataset in table VI RapidMiner has performed least with highest error% i.e. 18.8% among the three tools. So, it is concluded that a no single tool is best for all datasets and in all situations. Hence, performance and accuracy changes with the quantity and quality of the dataset.

VI. CONCLUSION

Data Mining is the most important field of study and application in this vast and ever growing data world. The era of Big Data has a number of Open Source Big Data Mining Tools to analyse and predict the exploding data. In the paper, three most frequently used tools have been studied and compared so that future analysts can take help in selecting the tool to use in their analysis. For experiments two real dataset are taken and supervised learning method is selected. A decision tree is created for the datasets and the tools are analysed on the basis of accuracy and confusion matrix. The study depicts that Rapid Miner is best tool for a small database and R is best for a large database. But in terms of accuracy Rapid Miner excels among the tools. So, it is concluded that RapidMiner is better than others. Many other test substitutes have been left out for future work- test different algorithms using different parameters, test other tools, test more datasets, test other classification technique like K- nearest neighbour, regression, genetic algorithms etc.

REFERENCES

1. D. Pyle, Data Preparation for Data Mining, San Diego: Academic Press, 1999.
2. http://www.kdnuggets.com/2017/05/poll-analytics-data-science-machine-learning-software-leaders.html
3. Nurdatillah Hasim, Norhaidah Abu Haris "A Study of Open-Source Data Mining Tools for Forecasting", in the proceedings of IMCOM'15, January 08 – 10, 2015, ACM 2015.
4. Magdalena Graczyk, Tadeusz Lasota, and Bogdan Trawiński, "Comparative Analysis of Premises Valuation Models Using KEEL, RapidMiner, and WEKA", in the proceedings of ICCI 2009, Springer-Verlag Berlin Heidelberg 2009.
5. Luís C. Borges, Viriato M. Marques and Jorge Bernardino, "Comparison of Data Mining Techniques and Tools for Data Classification", in the proceedings of C3S2E13, Jul 10-12 2013, Portugal, ACM, 2013.
6. Hemlata, Dr. Preeti Gulia, "Comprehensive Study of Open- Source Big Data Mining Tools", International Journal of Artificial Intelligence and Knowledge Discovery, e-ISSN: 2231-0312, Vol. 6, Issue 1, January, 2016.
7. Angela Lausch, Andreas Schmidt, Lutz Tischendorf, "Data mining and linked open data – New perspectives for data analysis in environmental research", Ecological Modelling, 0304-3800, Elsevier, Science Direct, 2014.
8. A. Jović*, K. Brkić* and N. Bogunović, "An overview of free software tools for general data mining", in the proceedings of 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), IEEE, 2014.
9. Ahmad Al-Khoder, Hazar Harmouch, "Evaluating four of the most popular Open Source and Free Data Mining Tools", International Journal of Academic Scientific Research (272-6446), Volume 3, Issue 1, PP 13-23.
10. Neha Chauhan, Nisha Gautam, "Parametric Comparison of Data Mining Tools", International Journal of Advanced Technology in Engineering and Science, Vol. No. 3, Issue 11, November 2015.
11. Jaideep Jagadeeshwar Rao, Rakesh Kelappan and Paul Pallath, "Recommendation System to Enhance Planning of Software Development using R", in the proceedings of RSSE'14 June 3, 2014, Hyderabad, ACM 2014.
12. Hilda Kosorus, J¨urgen H¨onigl, Josef K¨ung, "Using R, WEKA and RapidMiner in Time Series Analysis of Sensor Data for Structural Health Monitoring", in the proceedings of 22nd International Workshop on Database and Expert Systems Applications, IEEE, 2011.
13. https://archive.ics.uci.edu/ml/datasets/Iris
14. https://www.rstudio.com
15. http://rapidminer.com
16. https://www.knime.org

AUTHORS PROFILE

Hemlata is currently working as Lecturer, Computer Engg., Government Polytechnic, Sanghi, Rohtak from 2007. She has published more than 12 research papers in various National/International journals and conferences including Springer, SCOPUS and UGC. Her research areas include Data Mining and Big Data.

Dr. Preeti Gulia is currently working as Assistant Professor at Department of Computer Science & Applications, M.D.University, Rohtak, India. She is serving the Department since 2009. She earned her doctoral degree in 2013. She has published more than 65 research papers and articles in journal and conferences of National/ International repute including Springer, ACM, Scopus. Her area of research includes Data Mining, Big Data, Machine Learning, Deep Learning, IoT, Software Engineering. She is an active professional member of IAENG, CSI and ACM. She is also serving as Editorial Board Member Active Reviewer of International/ National Journals. She has guided one research scholar as well as guiding four Ph.D. research scholars from various research areas.