An Extensive Study of Data Analysis Tools (Rapid Miner, Weka, R Tool, Knime, Orange)

Venkateswarlu Pynam [1], R Roje Spanadna[2], Kolli Srikanth [3]
Assistant professors, Department of Information Technology, University College of engineering Vizianagaram
JNTUniversity KAKINADA, Andhara Pradesh :525003

Abstract

In today's data has been increasing in the concept of 3 v's (volume, velocity and variety) technology. Due to the large and Complex Collection of Datasets is difficult to process on traditional data processing applications. So that leads to arrive a new technology called Data Analytics. It is a science of exploring raw data and elicitation the useful information and hidden pattern. The main aim of data analysis is to use advance analytics techniques for huge and different datasets. The size of the dataset may vary from terabytes to zetta bytes and that can be structured or unstructured. The paper gives the comprehensive and theoretical analysis of five open source data analytics tools which are RapidMiner, Weka, R tool, KNIME and Orange. By employing the study the choice and selection of tools and be made easy, these tools are evaluated on basis of various parameters like amount of data used, response time, ease of use, price tag, analysis of algorithm and handling.

Keywords - *Data Analytics, Big Data, Data analytical tools, Visualization tools, Data mining.*

I. INTRODUCTION

Data is a collection of values in the form of raw data which is translated into forms that is easy to process. The data is been increasing exponentially in the digital form since last few decades. Data size has raise from gigabytes to terabytes. This explosive rate of data increment is growing day by day and estimations tell that the amount of information in world gets double almost every month. This type of massive amount of data in both structured and unstructured is called Bid Data. When handling and processing of data has become difficult with conventional databases and software techniques. There are different problems with big data [1] like processing of large data without solid analytical techniques become difficult which often leads to inaccurate result. Data Analytic is the science of analyzing data to convert information to useful knowledge. This knowledge could help us understand our world better and in many contexts enable us to make better decisions. The data analytics techniques are structured around of different category of data analytics namely descriptive, inferential, predictive and prescriptive analytics. With the increasing need of data analysis [7] some tools that are directly analyze the data and derive a conclusion are in demand.

There are thousands of Big Data tools out for data analysis at present. Data analysis is the process of inspecting the data, cleaning the data, transforming the data and modeling data with the goal of discovering useful information, Suggesting conclusions and supporting decision making. Data analysis in the areas of open source data tools, data visualization tools, sentimental tools, data extraction tools and databases. These tools are generating a report to summaries the conclusions and provide better visualizations and produce accurate results with minimum effort. There are different tools available for data analytics like RapidMiner, Weka, KNIME, R tool, Orange, OpenRefine, Solver, Julia, etc [5]. we have choose five tools among these for comparison which are RapidMiner, Weka, KNIME, R tool and Orange then we will find out most efficient tool among these on basis of few parameters.

II. DATA ANALYTICAL TOOLS

Open Source Data Tools Rapid Miner is a data science software platform which has been developed by Ralf Klinkenberg, Ingo Mierswa, and Simon Fischer at the Artificial Intelligence. RapidMiner [9] that provides an unified climate for data preparation, machine learning, deep learning, text mining, and predictive analytics and business analytics. RapidMiner is used for business, commercial applications, research, education, training, rapid prototyping and application development and supports all machine learning process including data preparation, results visualization, model validation and optimization [8].

RapidMiner uses a client or server model with the server offered as either as a premise or in social or separate cloud infrastructures. There is no scoping mechanism in RapidMiner processes therefore objects can be stored and retrieved at any nesting level. The parameter optimization schemes are also available in RapidMiner. Numerous clustering operators are

available in RapidMiner that generate a cluster attribute e.g. the K-Means operator. The macro is one of the advanced topics of RapidMiner. RapidMiner naturally calculate the type of attributes of particular dataset and all attributes have legitimate role. The type and proper role can be changed by using the comparable operators [3]. RapidMiner operators because writing scripts can be time engrossing and error prone [3]. RapidMiner keeps datasets in memory as long as possible. So if there is any memory left RapidMiner will not dispose of previous results of the process. The report described RapidMiner's strengths as a "platform that supports an extensive breadth and depth of functionality, and with that it comes quite close to the business market.

III. ANALYSIS TECHNIQUES

There are various phases in each of the analysis process which are performed in order to get the output. These phases are performed in sequential order to achieve the desired goal effectively. The phases of analytics [10] are:

1. Identify the problem,
2. Designing data requirement,
3. Pre-Processing data,
4. Performing analytics over data and
5. Visualizing data.

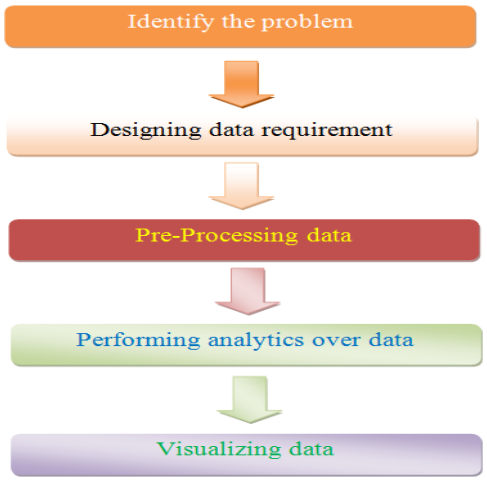

Figure : 1 The phases of analytics

A. *Identify the problem*

Now a day's analytics are performed on web datasets because if increasing the use of internet and growing business of organizations over internet. This leads to gradual increase of data size day by day. The organizations are wants to make predictions over the data to make desired decisions. The analytical applications must be scalable for collecting the datasets. Let us assume that there is an e-commerce website and wants to increase the business [4].

Identifying a wider variety of data sources may increase the probability of finding hidden patterns [2] and correlations. For example, to provide insight, it can be beneficial to identify as many types of related data sources as possible, especially when it is unclear exactly what to look for. Depending on the business scope of the analysis the nature of the business problems being addressed, the required datasets and their sources can be internal and/or external to the enterprise.

B. *Designing data requirement*

To perform the data analytics for a distinct problem, it needs datasets from associated domains. Based on the domain and problem specification, the data source can be determined and based on the problem definition the data characteristics of these datasets can be defined.

For example, if we are going to perform social media analytics like problem specification, we use the data source as Facebook or Twitter. For identifying the user characteristics, we need user profile information, likes, and posts as data attributes.

C. *Preprocessing data*

In data analytics, we don't use the duplicate data sources, data characteristics, data tools, and algorithms all of them will not need data in the duplicate configuration. This advantage is to the achievement of data operations, such as data cleansing, data aggregation, data augmentation, data sorting, and data formatting to furnish the data in a financed arrangement to the data tools and as well as algorithms that will be used in the data analytics.

Preprocessing is used to achieve data operation to decipher data into a fixed data arrangement previously furnished data to algorithms. The data analytics process will be proposed formatted data as the input. In Big Data, the datasets need to be formatted and transfer to Hadoop Distributed File System (HDFS) and used further by distinct nodes with Mappers and Reducers in Hadoop clusters.

D. *Performing analytics over data*

After data is available in the appropriate format for data analytics applications will be performed. The data analytics applications are achieved for determining essential knowledge from data to take improved decisions towards business in data mining concepts. It may use either descriptive or predictive analytics for business perception.

Analytics can be achieved with various machine learning and custom algorithmic concepts, such as data

regression, data classification, data clustering, etc. For Big Data, the equivalent algorithms can be converted in to MapReduce algorithms for working on Hadoop clusters by converting their data analytics logic to the MapReduce which is to be run over Hadoop clusters. These models need to be calculated and improved by discrete stages of machine learning concepts. The improved algorithms can provide better observation.

E. Visualizing Data

The capability to analyze large amounts of data and find useful judgment brings little value that can clarify the results are the analysts. The Data Visualization is committed to using data visualization approach to distinctly disseminate the analysis results for effective clarification by business users. Business users are able to understand the results in order to achieve value from the analysis. The results of completing the Data Visualization provide users with the ability to perform visual analysis [4].

IV. DATA ANALYTICS TOOLS

A. Rapidminer

Rapid Miner is applicable in both Free and open-source software and economic version and is a popular predictive analytic platform. Rapid Miner is helping activity enclose predictive analysis in their work processes with its user amicable, well-healed library of data science and machine learning algorithms through its all-in-one programming surrounding like Rapid Miner Studio. Likewise the basic data mining appearances like data cleansing, filtering, clustering, etc. The tool is also compatible with weak scripts. Rapid Miner is used for business or commercial applications, research and education.

Now make sure to highlight the repository so that the folders end up in right place. Now create a folder named ‘data’

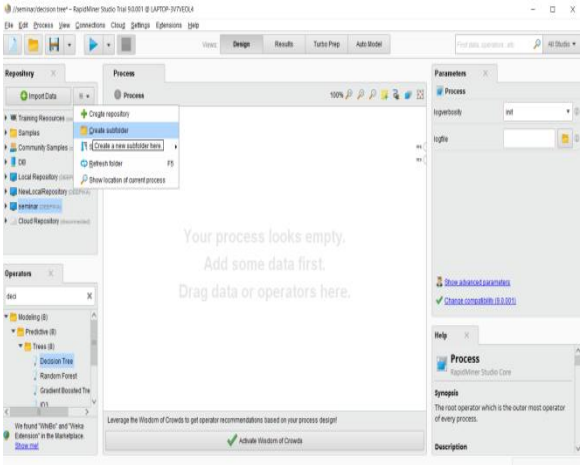

Figure: 2 import the new data

Now to load our data we can simply select the button: ’import data’. Click on the button ‘import data’

Step 1: After locating the file click ‘next’.

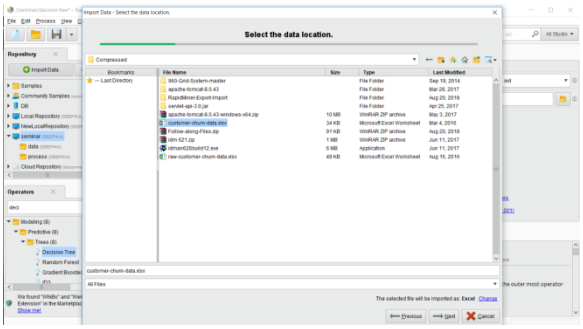

Figure: 3 loading the data

Step 2: Loads in the data and displays much like a spreadsheet.

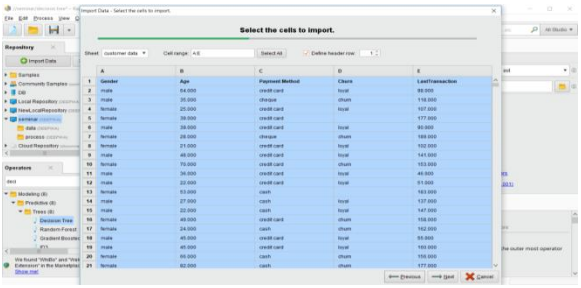

Figure: 4 Loads data in spreadsheet

Step 3: In this window we can decide if we want to exclude any certain column by selecting the ‘exclude column’ entry. Further you can change the ‘name’, ’role’ or ‘type’ of an attribute. Since the default for each column for loading is ‘general attribute’ in this case we need to change the role of our ‘churn’-attribute.

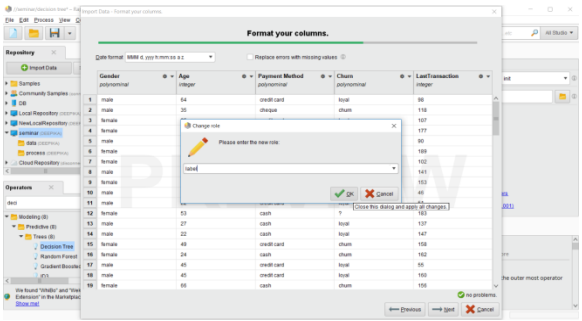

Figure: 5 load a general attribute

BUILDING A DECISION TREE

To create a decision tree first we have to import a dataset. Here, we are using a dataset about customer churn. After downloading the dataset and importing

into the rapid miner tool, we have to retrieve the data from our repository. Now click on the process directory, highlight your customer data and drag it over.

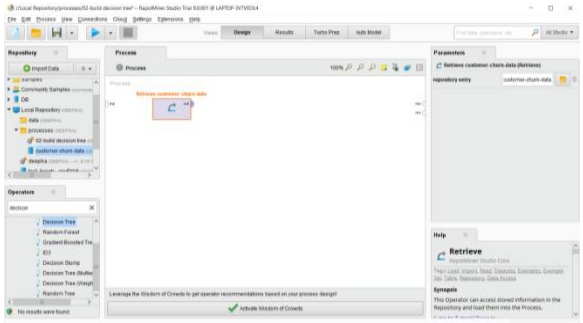

Figure: 6 process directory

Before we actually build a model we have to inspect our data for issues and see if we need to do any further preparation. so click on the 'output ' port of the operator and drag a connection on to the 'results' port of the process panel.

Now, click the port to establish the connection and come over to your 'run process' button and run it.

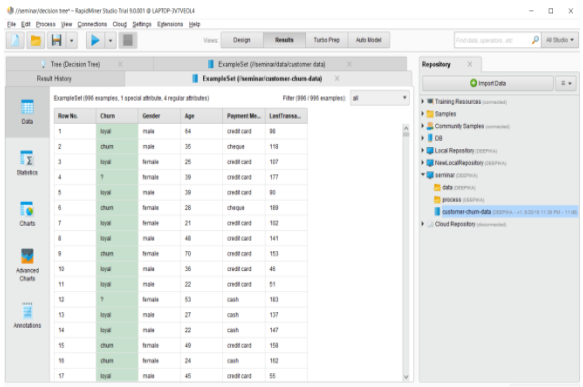

Figure: 7 running the dataset

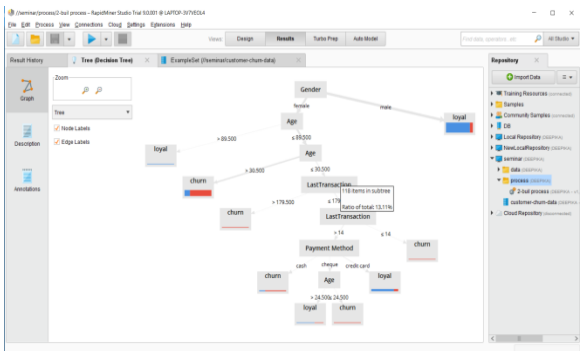

Figure: 8 decision tree for customer churn data

B. Knime

Knime is a data mining tool that can be used gaining approximately any kind of analysis. We explored how to visualise a dataset and retrive essential appearance. Predictive modelling was using a linear regression predictor to evaluation sales for each item accordingly [6]. Finally, we refine out the appropriate columns and exported it to a .csv file

1. File reader

The most familiar way to store nearly small amounts of data is static a text file. Among text files, the most familiar pattern has been so far the CSV (Comma Separated Version) format. The "comma" in the CSV phrase is just one of the available characters to separate data inner the file. Semicolon, colon, dot, tab, and many other signs are uniformly sufficient. A more rigid clarification of the file structure cause of course for quick reading. However, occasionally you need a more malleable definition of the file structure to get to a result, even if it desires a bit of a longer composition time.

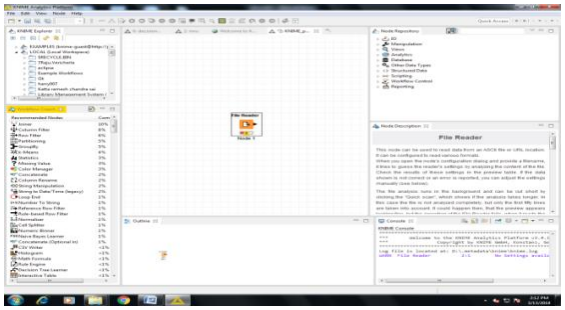

Figure: 9 read data from an ASCII file or URL location

2. Partitioning

The input table is division into two partitions (i.e. row-wise), e.g. train and test data. The two separations are accessible at the two output ports.

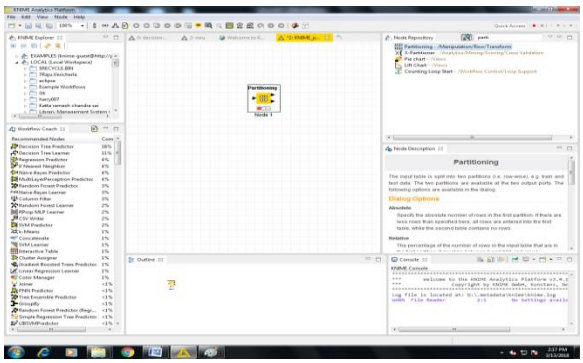

Figure: 10 partitioning the data.

3. Decision Tree

After the data is partitioned into train and test data, a Decision Tree Model is trained and applied. The Decision Tree learner node is important for the guidance of a decision tree model. Here is a abrupt

description of the basic environment available in its configuration window.

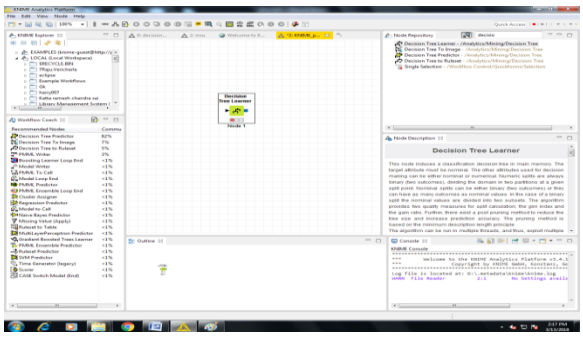

Figure: 11 Decision Tree learner node

4. Decision tree image

Decision tree aspect on an image are presently supported image type is PNG. The data input is optional. It can be used to provide a column with color information. This color information is needed for the chart in the nodes of the decision tree.

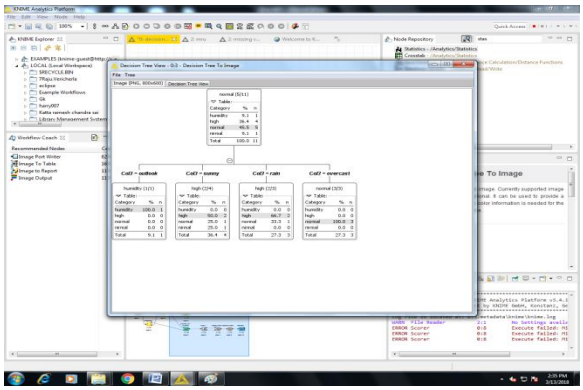

Figure: 12 Nodes of the decision tree

5. Decision tree predictor

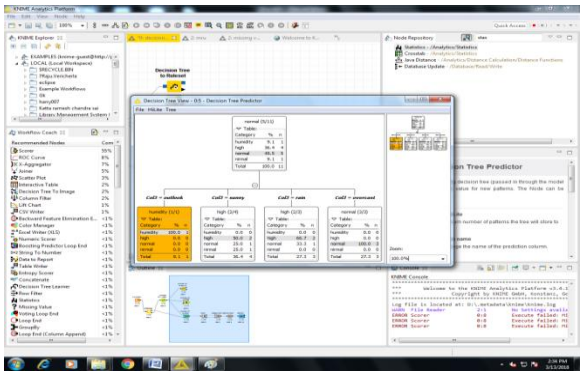

Figure: 13 predictors of the decision tree

6. Scorer

Compares two columns by their attribute value pairs and shows the confusion matrix, i.e. how many rows of which attribute and their classification match. Additionally, it is possible to hilight cells of this matrix to determine the underlying rows. The dialog allows you to select two columns for comparison; the values from the first selected column are represented in the confusion matrix's rows and the values from the second column by the confusion matrix's columns. The output of the node is the confusion matrix with the number of matches in each cell. Additionally, the second out-port reports a number of accuracy statistics such as True-Positives, False-Positives, True-Negatives, False-Negatives, Recall, Precision, Sensitivity, Specificity, F-measure, as well as the overall accuracy and Cohen's kappa.

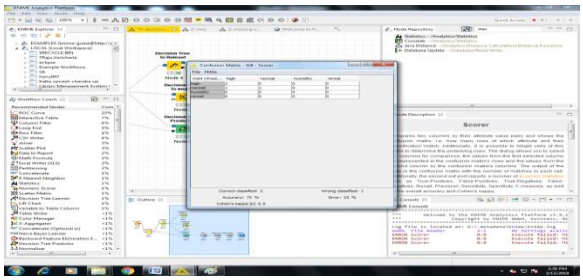

Figure: 14 confusion matrix of the node

7. Entropy scorer

Scorer for clustering results given a reference clustering. Connect the table containing the reference clustering to the first input port (the table should contain a column with the cluster IDs) and the table with the clustering results to the second input port (it should also contain a column with some cluster IDs). Select the respective columns in both tables from the dialog. After successful execution, the view will show entropy values (the smaller the better) and some quality value (in [0,1] - with 1 being the best possible value, as used in Fuzzy Clustering in Parallel Universes , section 6: "Experimental results").

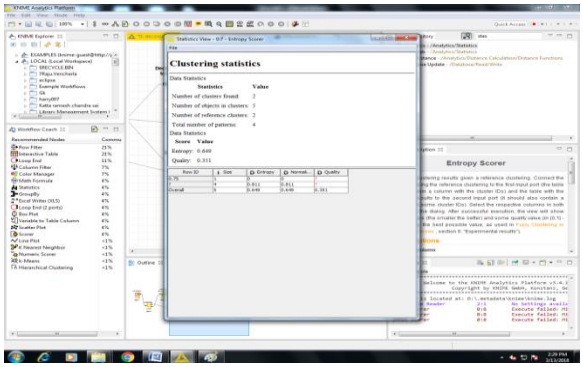

Figure: 15 clustering results

8. Numeric scorer

This node computes certain statistics between the a numeric column's values (ri) and predicted (pi)

values. It computes R^2=1-SSres/SStot=1-Σ(pi-ri)²/Σ(ri-1/n*Σri)² (can be negative!), mean absolute error (1/n*Σ|pi-ri|), mean squared error (1/n*Σ(pi-ri)²), root mean squared error (sqrt(1/n*Σ(pi-ri)²)), and mean signed difference (1/n*Σ(pi-ri)). The computed values can be inspected in the node's view and/or further processed using the output table.

Statistics:

This node calculates statistical moments such as minimum, maximum, mean, standard deviation, variance, median, overall sum, number of missing values and row count across all numeric columns, and counts all nominal values together with their occurrences. The dialog offers two options for choosing the median and/or nominal values calculations:

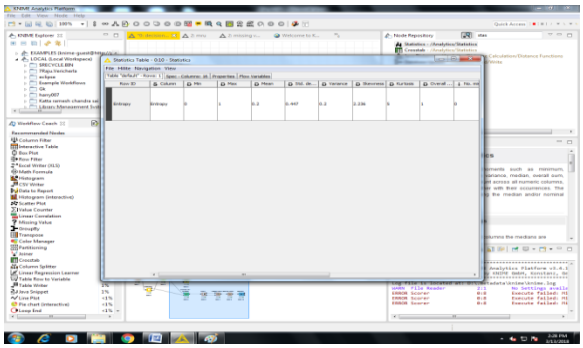

Figure: 16 statistical moments' calculations

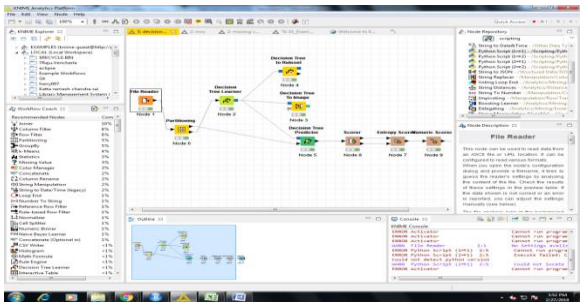

Figure: 17 decision tree for data set

C. Weka

Initially after starting the weka explorer the following window will be appeared where we can perform various operations using different datasets available [4]. To load the required dataset simply click on the button open file and choose the path C:/weka-3.8/data

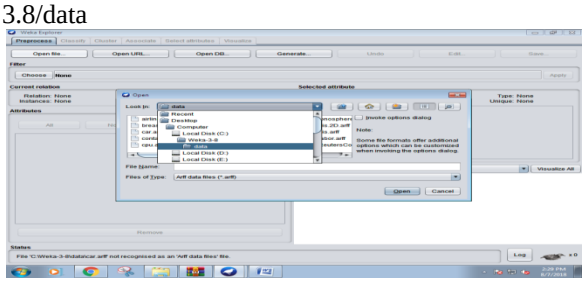

Figure: 18 load the thyroid dataset

Select the file hypothyroid.arff from the given datasets and click on open button

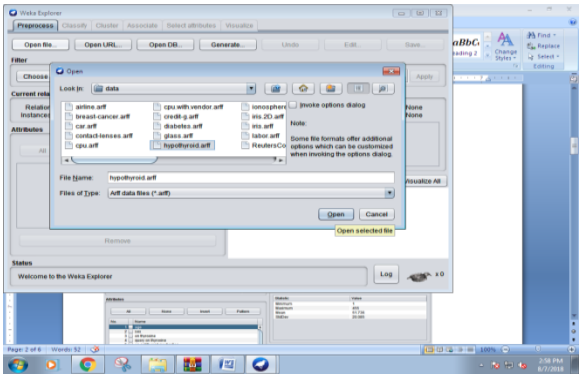

Figure: 19 select .arff file from datasets

With the Selected dataset Preprocessing is perfomed and the respective graph is shown based on the class and data items selected as shown below.

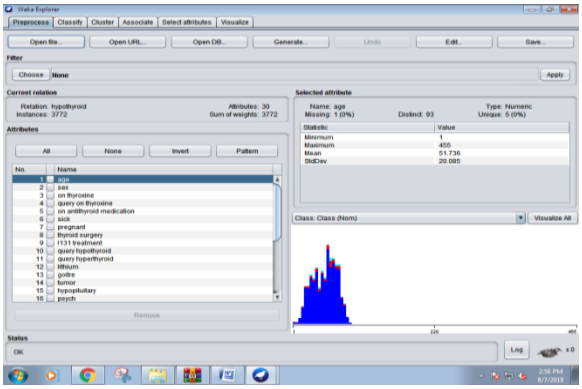

Figure: 20 preprocessing the data

Here we are classifying the dataset based on percentage split with 65% which yields 95.97% for correctly classified instances. To Show the output screen simply click on start button.

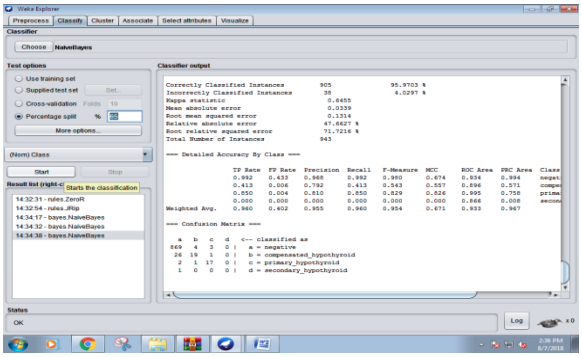

Figure: 21 split the data

To generate a Decision tree simply click on the folder Trees and select the algorithm to generate a decision tree. Here we are selecting J48 algorithim to generate

based on the test option "Use Training set". Now the following decision tree is generated based on the classified data items.

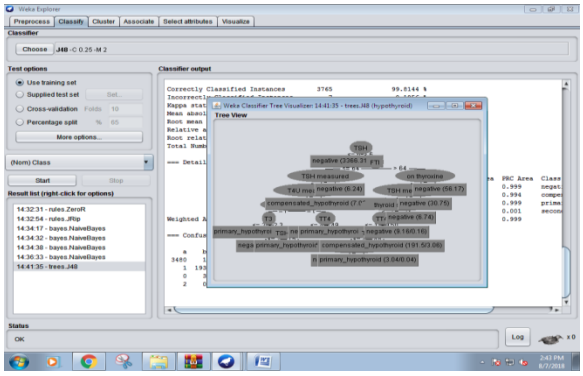

Figure: 22 generate a Decision tree

D. Orange

Orange provides data visualisation and data analysis for novice and expert, through interactive workflows. The File widget will now read the famous data set on iris flower dataset, and send it to the workflow. The changes will proliferate through the workflow updating its appliance.

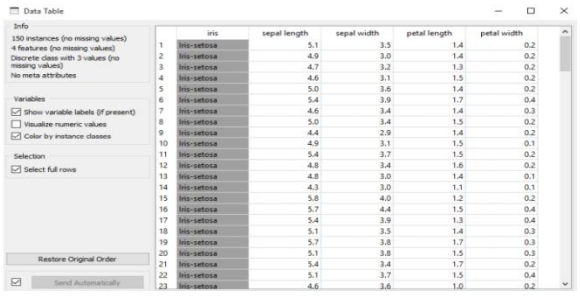

Figure: 23 read the data on iris flower dataset

Our aim is to inspect different types of animals, classification of them. Field colander design on the canvas and attach it to the File appliance.

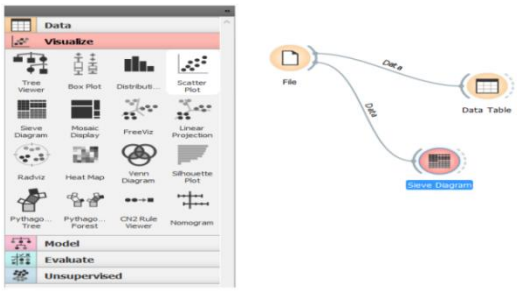

Figure: 24 split the data

We can visualize the pre-processed data in the form of simple graphs. The above pre-processed data can be visualized by using the box plot graph.

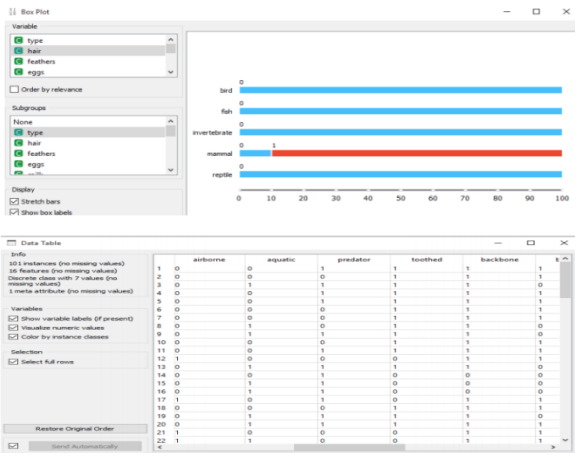

Figure: 25 preprocessing the data

A decision tree is a architecture that includes a root node, branches, and leaf nodes. Each subjective node stand for a test on an attribute, each branch stand for the outcome of a test, and each leaf node holds a class label. The uppermost node in the tree is the root node.

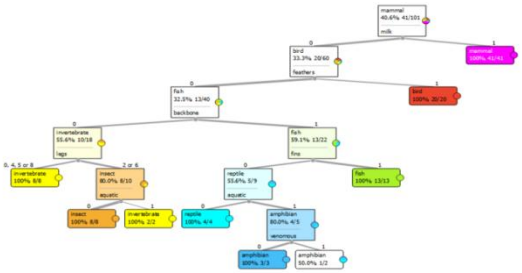

Figure: 26 decision tree for dataset

V. CONCLUSION

Depends on the analysis, Weka would be studied a very close to KNIME because of its many inherent appearance that require no coding knowledge. RapidMiner would be considered appropriate for experts, particularly those in the hard sciences, because of the additional programming skills that are needed, and the limited visualization support that is provided. RapidMiner has good and simple to use graphical efficiency, so it can be simply used and achieve on any system, furthermore it integrates superlative algorithms of other specified tools. R is the leading tool in visualization but it is a bit harder to create pretty graphs. R promotes reproducible research. R commands contribute an identical record of how an analysis was done. Commands can be alter, rerun, clarify, shared, etc. It can be concluded from information that though data analytics is the basic concept to all tool yet, In comparison, Orange offers

tools that seem to be targeted primarily at people with probably less need for custom applications into their own software but a distant accessible time with user communication, its written in python and origin is available, user preservatives are supported.

REFERENCES

[1] Lekha R. Nair , Sujala D. Shetty. "Research in Big Data and Analytics: An Overview" presented at International Journal of Computer Applications, Volume 108 – No 14, 2014.

[2] Mike Barlow. Real-Time Big Data Analytics: Emerging Architecture. Sebastopol, CA: O'Reilly Media, 2013, pp. 3.

[3] Sanjay Rathee. "Big Data and Hadoop with components like Flume, Pig, Hive and Jaql," presented at International Conference on Cloud, Big Data and Trust 2013, RGPV, 2015.

[4] Swasti Singhal, Monika Jena. "A Study on WEKA Tool for Data Preprocessing, Classification and Clustering" presented at International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-2, Issue-6,2013

[5] Kalpana Rangra, Dr. K. L. Bansal. "Comparative Study of Data Mining Tools", presented at International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 6, 2014.

[6] Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias K¨otter, Thorsten Meinl, Peter Ohl, Kilian Thiel and Bernd Wiswedel. "KNIME – The Konstanz Information Miner" presented at University of Konstanz Nycomed Chair for Bioinformatics and Information Mining, Germany.

[7] http://bigdata-madesimple.com/top-30-big-data-tools-data-analysis/

[8] http://opensourceforu.com/2017/03/top-10-open-source-data-mining-tools/

[9] https://rapidminer.com/wp-content/uploads/2014/10/RapidMiner-5-Operator-Reference.pdf

[10] http://pingax.com/understanding-data-analytics-project-life-cycle/