IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue, 5, March 2019 E-ISSN: 2347-2693

An Analysis on the Performance of Rapid Miner and R Programming Language as Data Pre-processing Tools for Unsupervised Form of Insurance Claim Dataset

Surya Susan Thomas[1*], Ananthi Sheshasaayee[2],

[1,2]PG & Research Department of Computer Science, Quaid -E- Millath Government College for Women, Chennai, India

[*]*Corresponding Author: susann.research@gmail.com, 9940439667*

DOI: https://doi.org/10.26438/ijcse/v7si5.14 | Available online at: www.ijcseonline.org

Abstract— Data Science has emerged as a super science in almost all the sectors of analytics. Data Mining is the key runner and the pillar stone of data analytics. The analysis and study of any form of data has become so relevant in todays' scenario and the output from these studies give great societal contributions and hence are of great value. Data analytics involves many steps and one of the primary and the most important one is data pre-processing stage. Raw data has to be cleaned, stabilized and processed to a new form to make the analysis easier and correct. Many pre-processing tools are available but this paper specifically deals with the comparative study of two tools such as Rapid Miner and R programming language which are predominantly used by data analysts. The output of the paper gives an insight into the weightage of the particular tool which can be recommended for better data pre-processing.

Keywords- Data analytics, data pre-processing, noise removal, clean data, Rapid Miner, R programming

I. INTRODUCTION

Data Pre-processing is a vital step in data analytics through machine learning.[1] The presence of duplicates, missing values ,noisy data, unwanted and unnormalized values makes the data dirty and correct evaluations cannot be made from it. [2] To eliminate these obstacles and clean data for further analysis is a tedious and time-consuming procedure. [3]It is estimated that nearly 60% of the time is spend for data pre-processing. Hence it makes it clear how important is this part. [4]With the rise of big data, data cleaning has become an imperative step in machine learning.

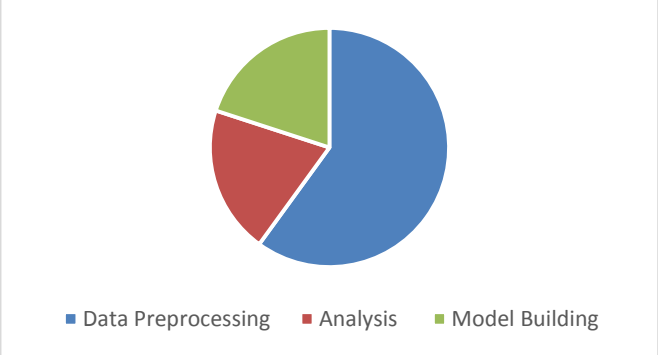

Figure 1: Chart representing the time consumption analysis of machine learning

Data pre-processing involves mainly four steps. [5][6]

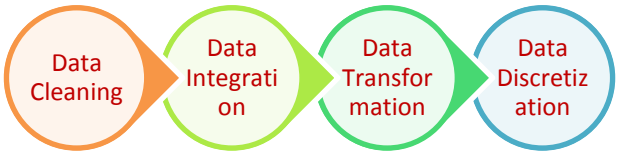

Figure 2: Illustrates the main steps involving in data pre-processing

II. MATERIALS AND METHODS

A. Tools for Data pre-processing

Raw data is procured and for the analysis of the observations it has to be pre-processed. The result after pre-processing data is the training set. Every dataset has to be segregated into train set and test set. Normally 60% of the dataset will be training set and the rest will be the test set. Noise reduction is one of the challenging part of pre-processing.[7] The removed values may also have large deviating values that have too many null feature values. These extensively

deviating features are known as outliers. Another pre-processing challenging point is handling missing data.
Many tools are [8] available for pre-processing but this paper mainly focusses on the comparative study of just two tools Rapid Miner and R programming language. These tools are extensively used by researchers to clean and process data. In this study, raw data or unsupervised data is taken and pre-processing is done with the help of R and Rapid Miner. Dataset used in this study is an unsupervised form of insurance claim dataset.

B. Rapid Miner

Rapid Miner formerly YALE (Yet Another Learning Environment) is a platform for aiding data mining and machine learning procedures such as data importing, data pre-processing, data transformation, data visualization, modelling, evaluation and deployment. RapidMiner is written in Java programming language.[9]The GUI of RapidMiner is the most attractive part where designing a process becomes much easier and presentable. Data mining processes are made up of arbitrarily interconnected operators described in XML files. Learning methods and attribute evaluators of Weka machine learning are incorporated in this tool. Moreover, certain statistical modelling features of R-project is also integrated which makes it a powerful tool than its predecessors. Text mining, multimedia mining, feature engineering, data stream mining etc can be implemented by RapidMiner. It is widely used in electronics, banking, insurance IT industry, market research ,pharma industry etc.[10]

C. R Project

R is a programming language mainly used for statistical computing and graphics. It is very similar to the S language, more precisely a different implementation of S. [11]It is a free software. Mathematical symbols and formulas can be produced easily with well-designed publication quality plots. R can be extended easily via packages which is very supportive for a developer and more packages can be added from the CRAN family which covers a wide range of modern statistics. Linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering etc can be performed with the help of R.

D. Data Pre-processing using Rapid Miner

The dataset was pre-processed employing both the tools. As the dataset was unsupervised, data cleaning, noise reduction, transformation and discretization and outlier detection was performed on the data by various methods and procedures available in the tools. Input data had many missing values and was an obstacle in processing the data smoothly.

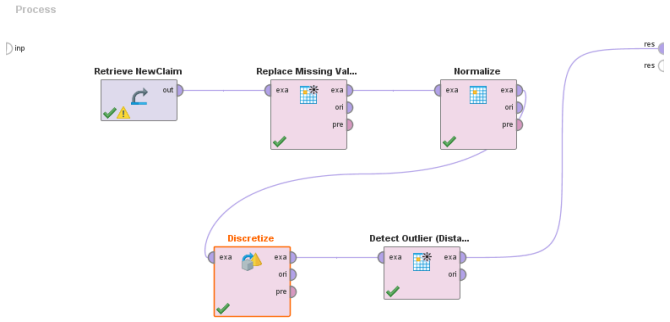

Figure 3: Illustrating the Pre-processing design in RapidMiner

E. Data preprocessing with R programming language[3]

tgtLines <- sapply (1:76, function(i) d[grep(paste0("^", i),d)[1]]) head(tgtLines,2)
[1] "1 id: patient identification number" ## [2] "2 ccf: social security number (I replaced this with a dummy value of 0)"
Throwing the IDs out...
nms <- str_split_fixed (tgtLines," ",2)[,2] head(nms,2)
[1] "id: patient identification number" ## [2] "ccf: social security number (I replaced this with a dummy value)

III. RESULTS AND DISCUSSION

A. Data Visualization after pre-processing in RapidMiner

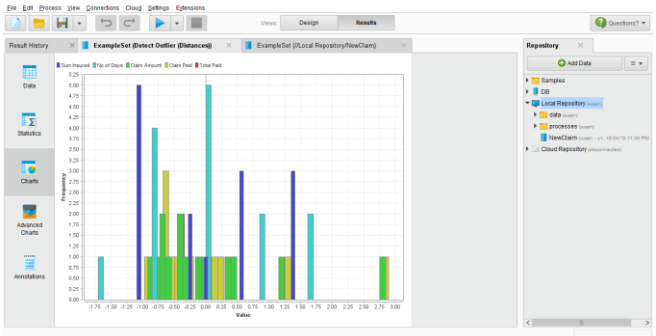

Figure 4: Visualization in RapidMiner

B. Imputing missing variables using R Project

Table 1: Sample data before pre-processing

1	50000	100%	50000	1	Paid	Paid
2	35000	85.70%	30000	1	Paid	Paid
3	40000	100%	40000	1	Paid	Paid
4	100000	100%	100000	1	Paid	Paid
5	150000	100%	150000	1	Paid	Paid
6	NA	88.88%	40000	1	Paid	Paid

7	50000	86%	43000	1	Paid	Paid
8	NA	86.52%	64890	1	Paid	Paid
9	58000	100%	58000	1	Paid	Paid
10	76500	68.40%	52356	1	Paid	Paid
11	54500	82.97%	NA	1	Paid	Paid
12	67980	100%	67980	1	Paid	Paid

Table 2: Sample data after pre-processing

1	50000.00	100%	50000	1	Paid	Paid
2	35000.00	85.70%	30000	1	Paid	Paid
3	40000.00	100%	40000	1	Paid	Paid
4	100000.00	100%	100000	1	Paid	Paid
5	150000.00	100%	150000	1	Paid	Paid
6	45000.00	88.88%	40000	1	Paid	Paid
7	50000.00	86%	43000	1	Paid	Paid
8	75000.00	86.52%	64890	1	Paid	Paid
9	58000.00	100%	58000	1	Paid	Paid
10	76500.00	68.40%	52356	1	Paid	Paid
11	54500.00	82.97%	NA	1	Paid	Paid
12	67980.00	100%	67980	1	Paid	Paid

C. Data Visualization after preprocessing in R Project

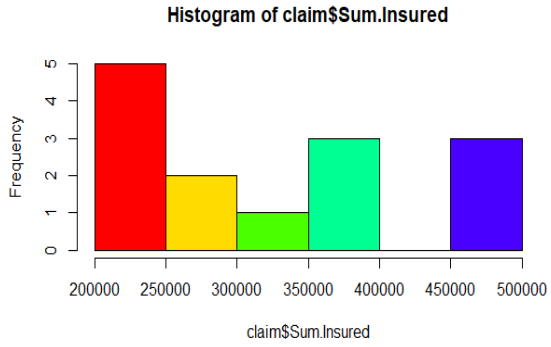

Figure 5

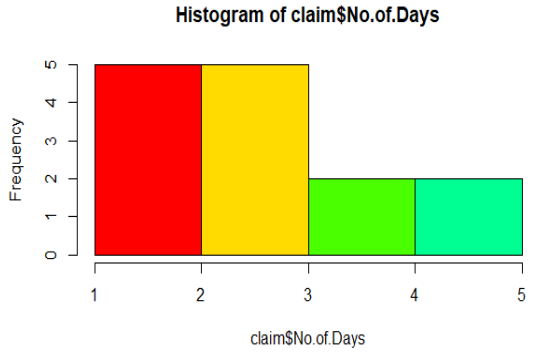

Figure 6

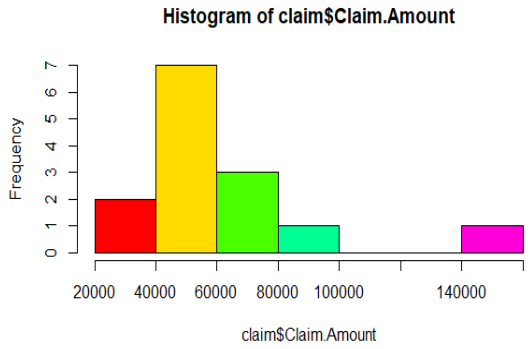

Figure 7

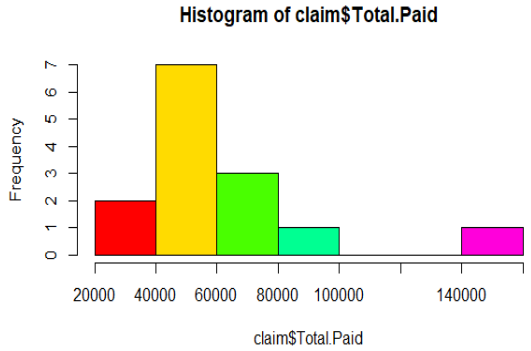

Figure 8

Figure 5, Figure 6, Figure 7, Figure 8 illustrates the graphical representation of processed data in R.

D. Discussion

Dataset was cleaned ,noise reduced and transformed into a fine set so the study and analysis of the data could be done more sleek and smooth.

Data Pre-processing in RapidMiner was more flexible on the design part as all the functions was dragged and dropped and function values and range were altered accordingly to the need of the process structure.If the connections are not correct error messages are popped out instantly so it can be rectified promptly.The coding part in RapidMiner is very minimal and so the chance of making errors.Since it is a drag and drop concept here ,time taken to build a process is relatively less.Result ,Statistical and graphical representation of the result can be viewed straight away and desired type of chart for the output can be selected from the list which makes preprocessing effortless. The need for deep learning methods and some of the more advanced specific machine learning algorithms (e.g. extremely randomized trees, various inductive logic programming algorithms) is currently limited.[12]

Data Pre-processing in R project is less flexible than Rapidminer as R is a programming language.Each process build needs to be coded with the help of installed packages in R which is the backbone of this tool.The prescence of packages and by writing codes makes the tool more dynamic,robust and flexible.Big data can be well handled by R than RapidMiner.Statistical and graphical representation is very minute and core.It is illustrated in Table 3. It takes more time to build a process in R than RapidMiner.Moreover occurrence of errors while coding is higher.

IV. CONCLUSION

Results obtained from the output and from the discussions above clinch some points to learn.From the study it is evident that for data preprocessing ,RapidMiner will be more reccommendable than R project as the former does the work more placidly and consuming less time while for analysis, model building and validation especially for a large dataset, R project would be more fitting and judicious.

ACKNOWLEDGMENT

I extend my sincere gratitude to my research guide Dr. Ananthi Sheshasaayee for her support and guidance in writing this research paper.

REFERENCES

[1] D. Thornton, G. Van Capelleveen, M. Poel, J. Van Hillegersberg, and R. M. Mueller, "Outlier-based Health Insurance Fraud Detection for U.S. Medicaid Data," *Proc. 16th Int. Conf. Enterp. Inf. Syst.*, pp. 684–694, 2014.

[2] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data preprocessing for supervised leaning," *Int. J. Comput. Sci.*, vol. 1, no. 1, pp. 111–117, 2006.

[3] D. P. Methods, "Data_Preprocessing."

[4] T. H. Zolt´an Prekopcs´ak, G´abor Makrai and C. G´asp´ar-Papanek∗, "Radoop: Analyzing Big Data with RapidMiner and Hadoop." http://www.iasri.res.in/ebook/win_school_aa/notes/Data_Preprocessing.pdf, "No Title." .

[6] A. FAMILI, W. SHEN, R. WEBER, and E. SIMOUDIS, "Data preprocessing and intelligent data analysis," *Intell. Data Anal.*, vol. 1, no. 1–4, pp. 3–23, 1997.

[7] C. M. Teng, "Correcting noisy data," *Proc 16th Int. Conf Mach. Learn.*, pp. 239–248, 1999.

[8] K. Rangra and K. L. Bansal, "Comparative Study of Data Mining Tools," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 4, no. 6, pp. 2277–128, 2014.

[9] https://en.wikipedia.org/wiki/RapidMiner, "No Title." .

[10] Y. Ramamohan, K. Vasantharao, C. K. Chakravarti, and A. S. K. Ratnam, "A Study of Data Mining Tools in Knowledge Discovery Process," *Int. J. Soft Comput. Eng.*, 2012.

[11] Https://www.r-project.org/about.html, "No Title." .

[12] A. Jović, K. Brkić, and N. Bogunović, "An overview of free software tools for general data mining," *2014 37th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2014 - Proc.*, no. May, pp. 1112–1117, 2014.

Authors Profile

Dr.(Mrs.) Ananthi Sheshasaayee pursued Master of Computer Application from the University of Madras. She is currently the Head & Associate Professor, PG & Research Department of Computer Science,Quaid-E-Millath Government College for Women,Chennai,India.She secured her doctorate in Computer Science from the University of Madras in the year 2009. She is a member of IEEE & IEEE computer society. She has authored many computer related books and published more than 200 research papers in reputed international journals (SCI & Web of Science) and conferences including IEEE and it's also available online. She was the chairperson for several national and international conferences.Her main research work focuses on Cryptography Algorithms, Data Mining, IoT and Computational Intelligence based education. She has 27 years of teaching experience and 10 years of Research Experience.

Mrs Surya Susan Thomas pursed Master of Computer Application from Mahatma Gandhi University,Kerala,India in the year 2006.She is currently pursuing Ph.D. in the PG & Research Department of Computer Science,Quaid-E-Millath Government College for Women,Chennai,India.She has published nearly 10 research papers in reputed international journals(Scopus Indexed) and conferences including IEEE and it's also available online. Her main research work focuses on Data Mining, IoT and Anomaly Detection in Business Domain and Data Analytics. She has 2 years of teaching experience and 3 years of Research Experience.

