An analysis on adult datset in a decision tree using rapid miner tool

D. Kerana Hanirex*, K.P. Thooyamani,
Department of CSE, Bharath University, Chennai-73.
*Corresponding author:E-Mail:keranarobinson@gmail.com

ABSTRACT

Data mining is the method of analyzing information from totally different views and summarizing it into helpful information. Data processing package is one among variety of analytical tools for analyzing information. Rapid Miner can be a package platform that was developed by the company of an identical name stated as Rapid Miner that has associate integrated atmosphere for machine learning, processing, text mining, data analytics and business analytics. It is used for business and industrial applications nevertheless as for analysis, education, training, speedy prototyping, and application development and supports all steps of the data mining methodology at the side of results image, validation and improvement.

KEY WORDS: Rapid Miner, Data mining, Decision tree, Algorithm, Adult Dataset.

1. INTRODUCTION

The decision tree model is that the model of computation or communication within which associate algorithmic program or communication method is taken into account to be essentially a choice tree. i.e., A sequence of branching operations supported with comparisons of some quantities.

C4.5 is associate algorithmic program accustomed generate a choice tree developed by Ross Quinlan. C4.5 is associate extension of Quinlan's earlier ID3 algorithmic program. The choice trees generated by C4.5 is used for classification, and for this reason, C4.5 is commonly cited as an applied mathematics classifier. In call tree learning, ID3 (Iterative Dichotomiser 3) is associate algorithmic program developed by Ross Quinlan usually generate a choice tree from a dataset. ID3 is that the precursor to the C4.5 algorithmic program, and is usually utilized in the machine learning and linguistic communication process domains.

2. MATERIALS AND METHODS

Rapid miner tool: Rapid Miner uses a client or server model with the server offered as package as a service or on cloud infrastructures. Rapid Miner provides an advanced analytical resolution through template-based frameworks that speed delivery and decrease errors by nearly eliminating the necessity to put in writing down code. Rapid Miner provides processing and machine learning procedures including: data loading and transformation (Extract, transform, load (ETL)), information preprocessing image, math modeling, evaluation, and preparation. Rapid Miner is written using the Java programming language. RapidMiner provides a graphical computer program to execute analytical workflows. RapidMiner utility typically extended with additional plugins. This paper is enforced with Rapid Miner 6.0.

Dataset details: For experimental analysis the datset Adult is selected from the UCI dataset repository. The Adult Dataset is used here with the following attributes namely Age, work class, final weight, education, education no, marital status, occupation, relationship, race, gender, capital gain, capital loss, hours, native. The row describes the attributes, the column describes the sample dataset. Total number of instances in the adult dataset is 48842, total number of attributes in adult dataset is 14.

Implementation details: The adult dataset was implemented through Rapid miner and the decision was build using various attributes as labels, and the graph was build between the those labels.

3. RESULT AND DISCUSSIONS

Tree build details:
RELATIONSHIP = Husband: Male { Male=70, Female=0}
RELATIONSHIP = Not-in-family
| FNLWGT > 50057.500
| | AGE > 26.500
| | | AGE > 28.500
| | | | EDUCATION NO > 13.500: Female { Male=0, Female=2}
| | | | EDUCATION NO ≤ 13.500
| | | | | HOURS > 38
| | | | | | HOURS > 41.500: Male { Male=9, Female=0}
| | | | | | HOURS ≤ 41.500
| | | | | | | AGE > 41: Female { Male=0, Female=3}
| | | | | | | AGE ≤ 41
| | | | | | | | FNLWGT > 206914.500: Male { Male=3, Female=0}
| | | | | | | | FNLWGT ≤ 206914.500

| | | | | | | | | | AGE > 30.500: Female { Male=0, Female=2}
| | | | | | | | | | AGE ≤ 30.500: Male { Male=2, Female=0}
| | | | | | HOURS ≤ 38: Female { Male=0, Female=2}
| | | AGE ≤ 28.500: Female { Male=0, Female=3}
| | AGE ≤ 26.500: Male { Male=6, Female=0}
| FNLWGT ≤ 50057.500: Female { Male=0, Female=3}
RELATIONSHIP = Other-relative: Male { Male=4, Female=1}
RELATIONSHIP = Own-child
| AGE > 18.500
| | FNLWGT > 102186.500
| | | FNLWGT > 145283: Male { Male=12, Female=5}
| | | FNLWGT ≤ 145283: Female { Male=0, Female=3}
| | FNLWGT ≤ 102186.500: Male { Male=4, Female=0}
| AGE ≤ 18.500: Female { Male=0, Female=3}
RELATIONSHIP = Unmarried: Female { Male=3, Female=13}
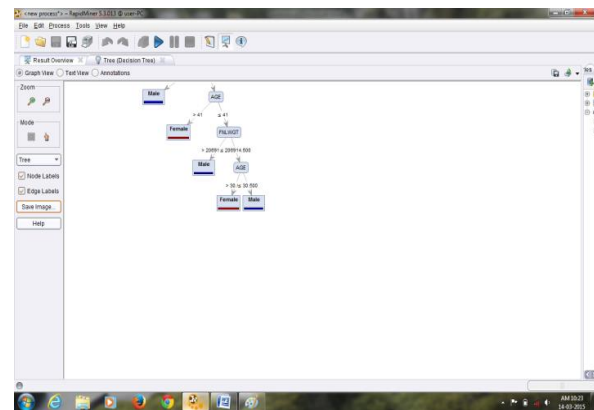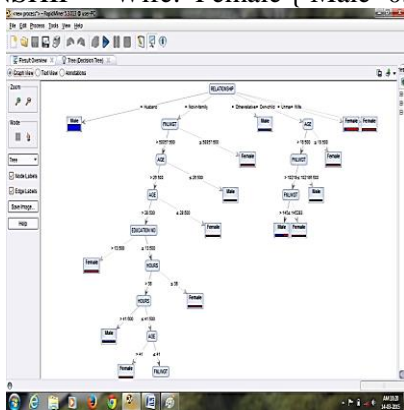RELATIONSHIP = Wife: Female { Male=0, Female=11}

Figure.1(i).Decision Tree for Label1:Relationship
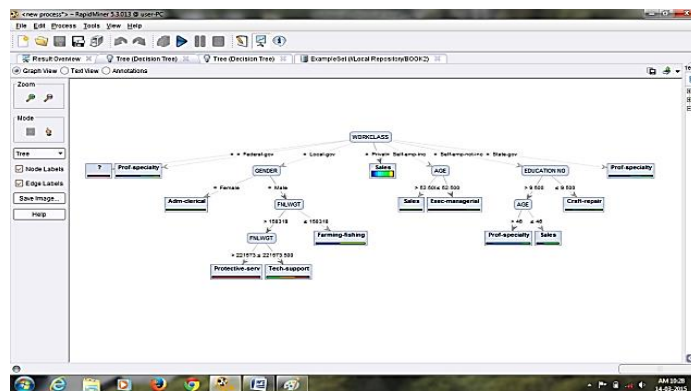

Fig:1(ii)Decision Tree for Label1:Relationship

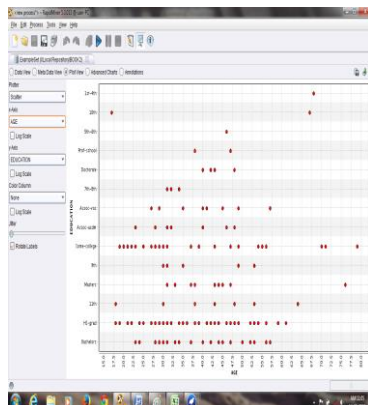**Figure.2. Decision tree for Label2as Workshop**

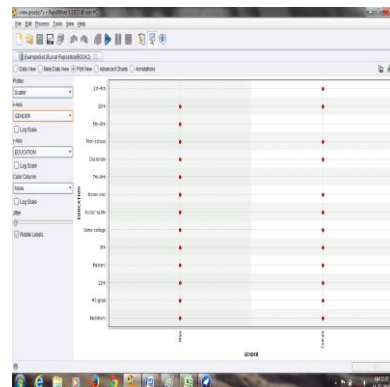**Figure.3. Scatter Graph between Age and Education**

Figure.4. Scatter Graph between Gender and Education

Output: In this paper the adult dataset was analysed with the Rapid miner tool in order to find out the time taken to build the decision tree and the graph is plotted between various attributes.

Table.1. Decision Tree Results

S.No	Labels	Time Taken For Execution
1	Gender	0.5S
2	Education	0.5S
3	Occupation	0.6S

From the table we can understand that to generate decision tree, the time taken by Gender as label is 0.5s. Time taken by Education as label is 0.5s. Time taken by Occupation as label is 0.6s.

4. CONCLUSION

Here for adult dataset we are using the Rapid miner tool in order to build the Decision tree based on various labels and the Graph has been plotted between various attributes and time taken in decision tree for Gender is 0.5 sec, for Education 0.5sec, for Occupation it is 0.6 sec, So finally we can conclude that the time varies for different labels but the decision tree can be built with 100 percent accuracy if there is no missing attribute.

REFERENCES

Abdullah H. Wahbeh, Qasem A. Al-Radaideh, Mohammed N. Al-Kabi, and Emad M. Al-Shawakfa, A Comparison Study between Data Mining Tools over some Classification Methods, IJACSA, 2013, 18-26.

Achudhan M, Prem Jayakumar M, Mathematical modeling and control of an electrically-heated catalyst, International Journal of Applied Engineering Research, 9 (23), 2014, 23013.

Gopalakrishnan K, Sundeep Aanand J, Udayakumar R, Electrical properties of doped azopolyester, Middle - East Journal of Scientific Research, 20 (11), 2014, 1402-1412.

Gopinath S, Sundararaj M, Elangovan S, Ratha krishnan E, Mixing characteristics of elliptical and rectangular subsonic jets with swirling co-flow, International Journal of Turbo and Jet Engines, 32 (1), 2015, 73-83.

Ilayaraja K, Ambica A, Spatial distribution of groundwater quality between injambakkam-thiruvanmyiur areas, south east coast of India, Nature Environment and Pollution Technology, 14 (4), 2015, 771-776.

Judith Gelernter, Data Mining of Maps and their Automatic Region—Time—Theme Classification, The SIGSPATIAL Special, 1 (1), 2009, 39-44.

Kerana Hanirex D, Kaliyamurthie KP, Kumaravel A, Analysis of improved tdtr algorithm for mining frequent itemsets using dengue virus type 1 dataset: A combined approach, International Journal of Pharma and Bio Sciences, 6 (2), 2015, 288-295.

Lingeswaran K, Prasad Karamcheti SS, Gopikrishnan M, Ramu G, Preparation and characterization of chemical bath deposited cds thin film for solar cell, Middle - East Journal of Scientific Research, 20 (7), 2014, 812-814, 2014.

Milan Kumari, Sunila Godara, Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction, IJCST, 2(2), 2011, 304-308.

Premkumar S, Ramu G, Gunasekaran S, Baskar D, Solar industrial process heating associated with thermal energy storage for feed water heating, Middle - East Journal of Scientific Research, 20 (11), 2014, 1686-1688.

Qasem A. Al-Radaideh, Eman Al Nagi, Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance, IJACSA, 3 (2), 2012, 144-151.

Rajeswari K, Vaithiyanathan V, Shailaja V Pede, Feature Selection for Classification in Medical Data Mining, IJETTCS, 2(2), 2013, 492-497.

Ritika, Research on data mining classification, IJARCSSE, 4 (3), 2014, 329- 332.

Sundar Raj M, Saravanan T, Srinivasan V, Design of silicon-carbide based cascaded multilevel inverter, Middle - East Journal of Scientific Research, 20 (12), 2014, 1785-1791.

Thair Nu Phyu, Survey of Classification Techniques in Data Mining, Proceedings of the International Multi Conference of Engineers and Computer Scientists, IMECS, 1 (1), 2009.

Thooyamani KP, Khanaa V, Udayakumar R, Application of pattern recognition for farsi license plate recognition, Middle - East Journal of Scientific Research, 18 (12), 2013, 1768-1774.

Thooyamani KP, Khanaa V, Udayakumar R, Efficiently measuring denial of service attacks using appropriate metrics, Middle - East Journal of Scientific Research, 20 (12), 2014, 2464-2470.

Thooyamani KP, Khanaa V, Udayakumar R, Partial encryption and partial inference control based disclosure in effective cost cloud, Middle - East Journal of Scientific Research, 20 (12), 2014, 2456-2459.

Thooyamani KP, Khanaa V, Udayakumar R, Using integrated circuits with low power multi bit flip-flops in different approch, Middle - East Journal of Scientific Research, 20 (12), 2014, 2586-2593.

Thooyamani KP, Khanaa V, Udayakumar R, Virtual instrumentation based process of agriculture by automation, Middle - East Journal of Scientific Research, 20 (12), 2014, 2604-2612.

Thooyamani KP, Khanaa V, Udayakumar R, Wide area wireless networks-IETF, Middle - East Journal of Scientific Research, 20 (12), 2014, 2042-2046.

Udayakumar R, Kaliyamurthie KP, Khanaa, Thooyamani KP, Data mining a boon: Predictive system for university topper women in academia, World Applied Sciences Journal, 29 (14), 2014, 86-90.