International Journal of Engineering & Technology, 7 (2.3) (2018) 68-72

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

C4.5 Classification Data Mining for Inventory Control

Robbi Rahim[1]*, Ilka Zufria[2], Nuning Kurniasih[3], Muhammad Yasin Simargolang[4], Abdurrozzaq Hasibuan[5], Dian Utami Sutiksno[6], Ricardo Freedom Nanuru[7], Jusuf Nikolas Anamofa[8], Ansari Saleh Ahmar[9], Achmad Daengs GS[10]

[1]School of Computer and Communication Engineering, Universiti Malaysia Perlis, Kubang Gajah, Malaysia
[2]Department of Information System, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia
[3]Faculty of Communication Science, Library and Information Science Program, Universitas Padjadjaran, Bandung, Indonesia
[4]Department of Informatics, Universitas Asahan, Kisaran, Indonesia
[5]Department of Industry Engineering, Universitas Islam Sumatera Utara, Medan, Indonesia
[6]Politeknik Negeri Ambon, Ambon, Indonesia
[7]Universitas Halmahera, Tobelo, Indonesia
[8]STKIP Gotong Royong, Masohi, Indonesia
[9]Department of Statistics, Universitas Negeri Makassar, Makassar, Indonesia
[10]Universitas 45 Surabaya, Surabaya, Indonesia
**Corresponding author E-mail: usurobbi85@zoho.com*

Abstract

Data Mining is a process of exploring against large data to find patterns in decision making. One of the techniques in decision-making is classification. Classification is a technique in data mining by applying decision tree method to form data, algorithm C4.5 is algorithm that can be used to classify data in tree form. The system has been built that shows the results of good performance and minimal error in view of the system that is able to distinguish the anomaly traffic with normal traffic. Data mining inventory system applications can facilitate the control of inventory in the company to reduce production costs.

Keywords: *Classification, C4.5 Algorithm, Data Mining*

1. Introduction

The inventory problem is one of the important issues that the company must solve. One effort to anticipate this inventory problem is to establish a control system on inventory to control the excess or shortage of inventory. If the company has excess inventory there are many risks that must be overcome such as the possibility of damage to goods, maintenance costs, as well as large capital. Conversely, if the company lack of inventory it will cause disappointment for customers, the loss of opportunities to gain profits and cause a sense of lack of trust from customers and the shift of customers to other manufacturers that will harm the company itself.

Problems faced by companies such as previous explanation can be overcome by applying data mining method[1]–[5]. Data mining is intended to provide real solutions for decision makers, to develop their business. And also, data mining can forecasting the data [6]–[9] Data mining methods that can be applied such as the classification C4.5 method[10]–[12]. Some research conducted through a database of sales transactions can be obtained various information about the habits or behavior of consumers. For example it can be known what products are often purchased and rarely purchased by consumers in every transaction, and known when there is an increase in purchases by consumers and also do defiance in the market with various promotions, if consumer behavior in decision-making can be known, then the company can design various strategies to controlling the inventory. Application of C4.5 algorithm as data classification process will simplify the process of grouping and searching process[13]–[16], it is expected that this algorithm can be a solution to process inventory data

2. Methodology

Data Mining is a multidisciplinary field of science, describing work areas that include database technology, machine learning, statistics, pattern recognition, information retrieval, artificial neural networks, knowledge-based systems, artificial intelligence, high-performance computing, and data visualization[17]–[20].

Data Mining is defined as data mining or an attempt to extract valuable and useful information on a very large database [21]. The most important thing in data mining techniques is the rule of finding the high frequency patterns among sets of item sets called the Association Rules function[12].

Classification Algorithm C4.5 or also known as decision tree algorithm is a method of classification and prediction is very strong and famous[2], [18], [22]–[24]. This algorithm includes the method of data mining, which is the process of finding patterns by sifting through a large amount of data using pattern recognition technology. In general, C4.5 Algorithm for building decision tree is as follows:

a. Select attribute as root.

b. Create a branch for each value.
c. Split case in the branch tree.
d. Repeat the process for each branch until all the cases on the branch have the same class.

Data mining is the extraction of information or patterns that are important or interesting from the data residing on large databases that have been unknown but have potential useful information. Data mining is done with a special tool, which executes data operations that have been defined based on the analysis model. Here are some advantages of C4.5 Classification algorithm or decision tree[2], [22], among others:

a. The results of the analysis of a tree diagram that is easy to understand.
b. Easy to build, and requires less experimental data than other classification algorithms.
c. Able to process nominal data and continue.
d. The result model can be easily understood.
e. Using statistical techniques that can be validated.
f. Computational time is relatively faster than other classification techniques.
g. The accuracy can match other classification techniques

The steps of the C4.5 Classification algorithm can be detailed as follows:

a. Create a decision tree from training data.
b. Convert the resulting tree to an equivalent set of rules. The number of rules is equivalent to the number of possible paths from the root to the leaf node.
c. Generate each rule by removing the initial conditions so as to increase classification accuracy.
d. Sort every rules according to their accuracy and use these rules in the same order when classifying the next data.

3. Results and Discussion

The C4.5 algorithm can be used to classify the number of product supplies. The classification results can be used to predict the consumption level of a product. The following experiment shows the process of C4.5 algorithm in classification.

Table 1: Details Product

Name	Type	Category	Qty
Oil	Jumbo	Lubrication	120
Brake fluid	Jumbo	Lubrication	118
Greases	Kyodoyushi	Lubrication	62
Greases	Cobra	Lubrication	26
Jack	Unigo	Sparepart	12
Radiator Coolant	Jumbo	Treatment	21
Jack	Akio	Sparepart	89
Drum	Jumbo	Lubrication	102
Oil	Idemitsu	Lubrication	68
Drum	Ultraline	Lubrication	108
Radiator Coolant	Ultraline	Treatment	20
Greases	Trane	Lubrication	70
Brake fluid	Fuso	Lubrication	76
Drum	JOT	Lubrication	6
Drum	Idemitsu	Lubrication	11
Oil Shock	Jumbo	Lubrication	130
Oil Shock	Kayaba	Lubrication	78
Cup Kit,Whl Cylinder	RCA	Sparepart	5
Cup Kit	RCA	Sparepart	30
Cup Kit,Whl Cylinder	RCA	Sparepart	10
Pin,Cylinder Slide	RCA	Sparepart	15
Cylinder Kit, FR Disc Brake	RCA	Sparepart	10
Shoe kit, brake, RR	RCA	Sparepart	20
Pin, Cylinder Slide	RCA	Sparepart	15
Insulator, Engine, FR	RCA	Sparepart	10
Cup Kit	RCA	Sparepart	20
Insulator, Engine, FR	RCA	Sparepart	10
Cover Assy,Clutch Fr	RCA	Sparepart	10
Cup Kit	RCA	Sparepart	10
Insulator, Engine, FR	RCA	Sparepart	10
Insulator, Engine, FR	RCA	Sparepart	15
Pump Assy, Fuel	RCA	Sparepart	30
Coil, Ignition	RCA	Sparepart	20
Cup kit	RCA	Sparepart	30
Cup kit	RCA	Sparepart	20
Cup kit	RCA	Sparepart	5
Bulb	RCA	Sparepart	10
Belt,V-Ribbed	RCA	Sparepart	10
Liner, FR fender, LH	RCA	Sparepart	10
Jar, Washer	RCA	Sparepart	10

Table 1 is a detail of product information that has been sold to consumers, the data is then processed weighting with the following criteria:

a. Qty Sold 0 - 59 Low value
b. Qty Sold 60 - 99 Middle value
c. Qty Sold ≥ 100 High Value

Based on these criteria then each item sold product is determined the level of consumption as in table 2.

Table 2: Detail Product with range weight

Name	Type	Category	Range
Oil	Jumbo	Lubrication	High
Brake fluid	Jumbo	Lubrication	High
Greases	Kyodoyushi	Lubrication	Middle
Greases	Cobra	Lubrication	Low
Jack	Unigo	Sparepart	Low
Radiator Coolant	Jumbo	Treatment	Low
Jack	Akio	Sparepart	Middle
Drum	Jumbo	Lubrication	High
Oil	Idemitsu	Lubrication	Middle
Drum	Ultraline	Lubrication	High
Radiator Coolant	Ultraline	Treatment	Low
Greases	Trane	Lubrication	Middle
Brake fluid	Fuso	Lubrication	Middle
Drum	JOT	Lubrication	Low
Drum	Idemitsu	Lubrication	Low
Oil Shock	Jumbo	Lubrication	High
Oil Shock	Kayaba	Lubrication	Middle
Cup Kit,Whl Cylinder	RCA	Sparepart	Low
Cup Kit	RCA	Sparepart	Low
Cup Kit,Whl Cylinder	RCA	Sparepart	Low
Pin,Cylinder Slide	RCA	Sparepart	Low
Cylinder Kit, FR Disc Brake	RCA	Sparepart	Low
Shoe kit, brake, RR	RCA	Sparepart	Low
Pin, Cylinder Slide	RCA	Sparepart	Low
Insulator, Engine, FR	RCA	Sparepart	Low
Cup Kit	RCA	Sparepart	Low
Insulator, Engine, FR	RCA	Sparepart	Low
Cover Assy,Clutch Fr	RCA	Sparepart	Low
Cup Kit	RCA	Sparepart	Low
Insulator, Engine, FR	RCA	Sparepart	Low
Insulator, Engine, FR	RCA	Sparepart	Low
Pump Assy, Fuel	RCA	Sparepart	Low
Coil, Ignition	RCA	Sparepart	Low
Cup kit	RCA	Sparepart	Low
Cup kit	RCA	Sparepart	Low
Cup kit	RCA	Sparepart	Low
Bulb	RCA	Sparepart	Low
Belt,V-Ribbed	RCA	Sparepart	Low
Liner, FR fender, LH	RCA	Sparepart	Low
Jar, Washer	RCA	Sparepart	Low

Based on values in table 1 and table 2, a decision tree will be created with values for each variable.

A. Calculates entropy and gain value

Calculate the total number of cases, the number of cases for high, medium, and low consumption. Then do a calculation to find the Entropy value of each attribute value and the Gain value for each attribute.

Number of cases : 40
Number of cases with High result : 5
Number of cases with middle result : 6
Number of cases with low result : 29

Calculate the entropy value of the total cases:

$$E(X_1, X_2, X_3) = \frac{n1}{n}E(X_1) + \frac{n2}{n}E(X_2) + \frac{n3}{n}E(X_3)$$

$$Entropy(X)total = \left(\left(-\frac{total\ high}{total\ cases}\right) * \log_2\left(\frac{total\ high}{total\ cases}\right)\right) + \left(\left(-\frac{total\ middle}{total\ cases}\right) * \log_2\left(\frac{total\ middle}{total\ cases}\right)\right) + \left(\left(-\frac{total\ low}{total\ cases}\right) * \log_2\left(\frac{total\ low}{total\ cases}\right)\right)$$

$$Entropy(X)total = \left(\left(-\frac{5}{40}\right) * \log_2\left(\frac{5}{40}\right)\right) + \left(\left(-\frac{6}{40}\right) * \log_2\left(\frac{6}{40}\right)\right) + \left(\left(-\frac{29}{40}\right) * \log_2\left(\frac{29}{40}\right)\right)$$

$$= 0.375 + 0.41054483912493089 + 0.33636164732584795 = 1.1219064864507788$$

The next process is to calculate every entropy in table 2, here are some calculation results for the entropy sample in table 2.

$$Entropy(Type = Jumbo) = \left(\left(-\frac{2}{2}\right) * \log_2\left(\frac{2}{2}\right)\right) + \left(\left(-\frac{0}{2}\right) * \log_2\left(\frac{0}{2}\right)\right) + \left(\left(-\frac{0}{2}\right) * \log_2\left(\frac{0}{2}\right)\right) = 0$$

$$Entropy(Type = Kyodoyushi) = \left(\left(-\frac{0}{1}\right) * \log_2\left(\frac{0}{1}\right)\right) + \left(\left(-\frac{1}{1}\right) * \log_2\left(\frac{1}{1}\right)\right) + \left(\left(-\frac{0}{1}\right) * \log_2\left(\frac{0}{1}\right)\right) = 0$$

$$Entropy(Type = Cobra) = \left(\left(-\frac{0}{1}\right) * \log_2\left(\frac{0}{1}\right)\right) + \left(\left(-\frac{0}{1}\right) * \log_2\left(\frac{0}{1}\right)\right) + \left(\left(-\frac{1}{1}\right) * \log_2\left(\frac{1}{1}\right)\right) = 0$$

$$Entropy(Type = Unigo) = \left(\left(-\frac{0}{1}\right) * \log_2\left(\frac{0}{1}\right)\right) + \left(\left(-\frac{0}{1}\right) * \log_2\left(\frac{0}{1}\right)\right) + \left(\left(-\frac{1}{1}\right) * \log_2\left(\frac{1}{1}\right)\right) = 0$$

$$Entropy(Type = Jumbo) = \left(\left(-\frac{0}{2}\right) * \log_2\left(\frac{0}{2}\right)\right) + \left(\left(-\frac{0}{2}\right) * \log_2\left(\frac{0}{2}\right)\right) + \left(\left(-\frac{2}{2}\right) * \log_2\left(\frac{2}{2}\right)\right) = 0$$

The calculation process is done until all entropy values are obtained, then Gain ratio calculation process is done as follows:

$$Gainratio(a) = Entropy(X) - \sum_{j=1}^{k} \frac{|X_i|}{|X|} * Entropy(X_i)$$

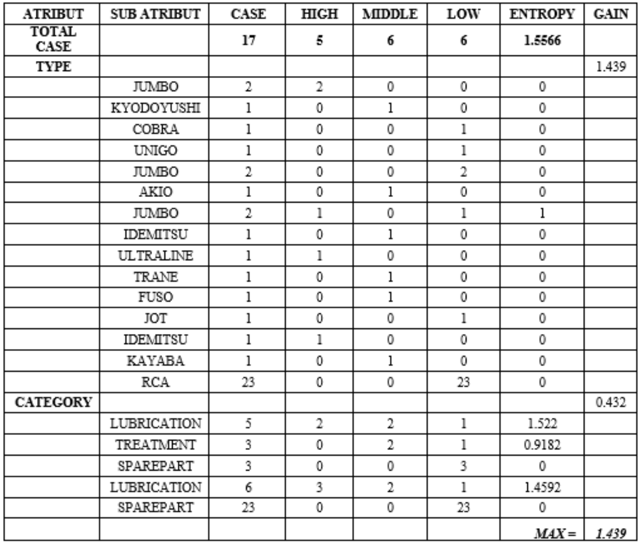
$$\begin{aligned} Gain(Type) &= 1.1219064864507788 \\ &- \left(\left(\frac{2}{40} * 0\right) + \left(\frac{1}{40} * 0\right) + \left(\frac{1}{40} * 0\right) + \left(\frac{1}{40} * 0\right) + \left(\frac{2}{40} * 0\right) \right. \\ &+ \left(\frac{1}{40} * 0\right) + \left(\frac{2}{40} * 1\right) + \left(\frac{1}{40} * 0\right) + \left(\frac{1}{40} * 0\right) + \left(\frac{1}{40} * 0\right) \\ &\left. + \left(\frac{1}{40} * 0\right) + \left(\frac{1}{40} * 0\right) + \left(\frac{1}{40} * 0\right) + \left(\frac{1}{40} * 0\right) + \left(\frac{23}{40} * 0\right)\right) \\ &= 1.4389929330205086882352941176471 \end{aligned}$$

$$Gainratio(a) = Entropy(X) - \sum_{j=1}^{k} \frac{|X_i|}{|X|} * Entropy(X_i)$$

$$\begin{aligned} &Gain\ (Category) = 1.1219064864507788 \\ &- \left(\left(\frac{5}{40} * 1.5219280948873621\right) \right. \\ &\qquad + \left(\frac{3}{40} * 0.91826249711435848\right) \\ &\qquad + \left(\frac{3}{40} * 0\right) + \left(\frac{6}{40} * 1.4591812431473434\right) \\ &\qquad \left. + \left(\frac{23}{40} * 0\right)\right) \\ &-(0.44762591026098885294117647058824 \\ &+ 0.16204632302018090823529411764706\ + 0 \\ &+ 0.51500514464023884705882352941176 + 0) \\ &= 0.43196261392262949176470588235324 \end{aligned}$$

Based on the entropy formula and the above gain obtained results as in table 3 below.

Table 3: The calculation results for entropy and gain values



ATRIBUT	SUB ATRIBUT	CASE	HIGH	MIDDLE	LOW	ENTROPY	GAIN
TOTAL CASE		**17**	**5**	**6**	**6**	**1.5566**	
TYPE							1.439
	JUMBO	2	2	0	0	0	
	KYODOYUSHI	1	0	1	0	0	
	COBRA	1	0	0	1	0	
	UNIGO	1	0	0	1	0	
	JUMBO	2	0	0	2	0	
	AKIO	1	0	1	0	0	
	JUMBO	2	1	0	1	1	
	IDEMITSU	1	0	1	0	0	
	ULTRALINE	1	1	0	0	0	
	TRANE	1	0	1	0	0	
	FUSO	1	0	1	0	0	
	JOT	1	0	0	1	0	
	IDEMITSU	1	1	0	0	0	
	KAYABA	1	0	1	0	0	
	RCA	23	0	0	23	0	
CATEGORY							0.432
	LUBRICATION	5	2	2	1	1.522	
	TREATMENT	3	0	2	1	0.9182	
	SPAREPART	3	0	0	3	0	
	LUBRICATION	6	3	2	1	1.4592	
	SPAREPART	23	0	0	23	0	
						MAX =	***1.439***

B. Determining the Root Node

1) From the calculation results in the above table, it is known that the greatest Gain value is on the Brand attribute of 1.439, so the Brand attribute becomes the root node.
2) At attribute Brand. There are 14 attribute values, namely Jumbo, Kyodoyushi, Cobra, Unigo, Jumbo, Akio, Jumbo, Idemitsu, Ultraline, Trane, Fuso, Jot, Idemitsu and Kayaba. Almost all attribute values have already classified the case to 1 so no further calculations are required.
3) While the value of Jumbo attribute has not classified the case into a single decision so it needs to be done again.

From the process can be generated temporary tree as follows.

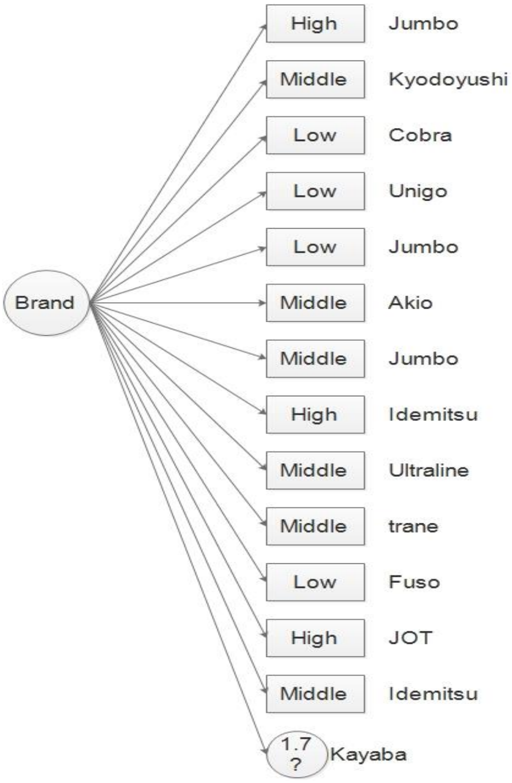


Fig. 1: Root Node

C. Branch nodes process

The calculation is performed to find the branch node of the JUMBO attribute. The calculation is done by finding the value of the attribute other than the root node (Type), i.e. by finding the number of cases for High, Medium and Low result, and Entropy value of all cases when Type = JUMBO. And then do the calculation to find the Gain value, and attribute with the biggest Gain value, it will be the branch node of the JUMBO attribute value.

Number of cases with Brand = JUMBO: 2

Number of cases with Brand = JUMBO which HIGH consumption level: 1

Number of cases with Brand = JUMBO which consumption MEDIUM level: 0

Number of cases with Brand = JUMBO which LOW consumption rate: 1

Calculate the entropy value of the total cases:

$$E(X_1, X_2, X_3) = \frac{n1}{n}E(X_1) + \frac{n2}{n}E(X_2) + \frac{n3}{n}E(X_3)$$

$$Entropy(X)total = \left(\left(-\frac{total\ High}{total\ Case}\right) * \log_2\left(\frac{total\ high}{total\ case}\right)\right) + \left(\left(-\frac{total\ middle}{total\ case}\right) * \log_2\left(\frac{total\ middle}{total\ case}\right)\right) + \left(\left(-\frac{total\ low}{total\ case}\right) * \log_2\left(\frac{total\ low}{total\ case}\right)\right)$$

$$Entropy(X)total = \left(\left(-\frac{1}{2}\right) * \log_2\left(\frac{1}{2}\right)\right) + \left(\left(-\frac{0}{2}\right) * \log_2\left(\frac{0}{2}\right)\right) + \left(\left(-\frac{1}{2}\right) * \log_2\left(\frac{1}{2}\right)\right) = 0.5 + 0 + 0.5 = 1$$

From the process can be generated decision tree as follows:

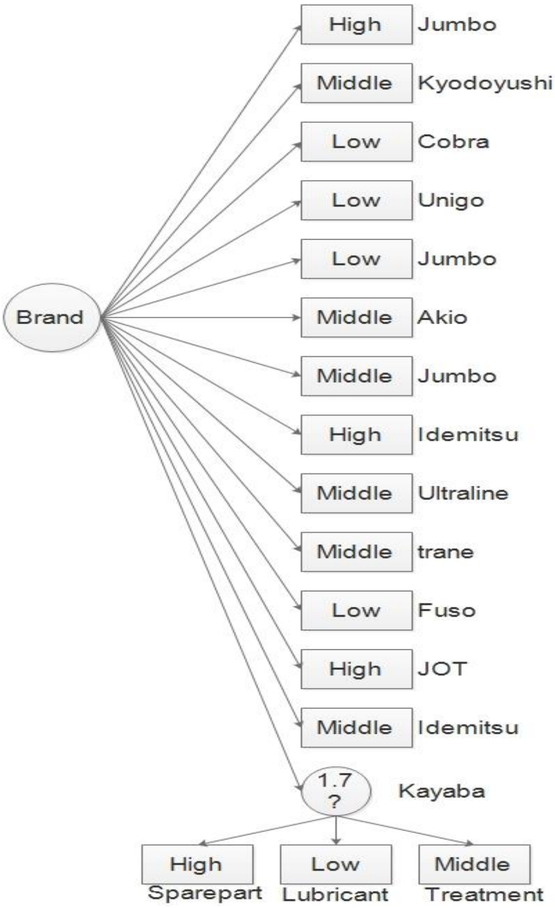


Fig. 2: Branch Node

Figure 2 is the end result of the process of using the C4.5 algorithm to classify products that can be execute.

4. Conclusion

The performed experiment can produce information predicted consumer consumption level based on data at a certain period, the calculation of entropy, gain and node can be used to identify consumer behavior by the company so that company can take policy related to product inventory.

References

[1] D. Siregar, D. Arisandi, A. Usman, D. Irwan, and R. Rahim, “Research of Simple Multi-Attribute Rating Technique for Decision Support,” *J. Phys. Conf. Ser.*, vol. 930, no. 1, p. 012015, Dec. 2017.

[2] E. Buulolo, N. Silalahi, Fadlina, and R. Rahim, “C4.5 Algorithm To Predict the Impact of the Earthquake,” *Int. J. Eng. Res. Technol.*, vol. 6, no. 2, pp. 10–15, 2017.

[3] J. Suyono, A. Sukoco, M. I. Setiawan, S. Suhermin, and R. Rahim, “Impact of GDP Information Technology in Developing of Regional Central Business (Case 50 Airports IT City Development in Indonesia),” in *Journal of Physics: Conference Series*, 2017, vol. 930, no. 1.

[4] Ep. E.P. and S. R, “Big data management with machine learning inscribed by domain knowledge for health care,” *Int. J. Eng. Technol.*, vol. 6, no. 4, p. 98, Sep. 2017.

[5] G. Nivedhitha and N. Rupavathy, “Data mining in personalized service of digital library,” vol. 7, no. 1.7, pp. 51–53, 2018.

[6] A. S. Ahmar, “A Comparison of α-Sutte Indicator and ARIMA Methods in Renewable Energy Forecasting in Indonesia,” *Int. J. Eng. Technol.*, vol. 7, no. 1.6, pp. 9–11, 2018.

[7] A. S. Ahmar *et al.*, “Modeling Data Containing Outliers using ARIMA Additive Outlier (ARIMA-AO),” *J. Phys. Conf. Ser.*, vol. 954, no. 1, 2018.

[8] A. S. Ahmar, A. Rahman, A. N. M. Arifin, and A. A. Ahmar, “Predicting movement of stock of ‘Y’ using sutte indicator,” *Cogent Econ. Financ.*, vol. 5, no. 1, 2017.

[9] A. Rahman and A. S. Ahmar, “Forecasting of primary energy consumption data in the United States: A comparison between ARIMA and Holter-Winters models,” in *AIP Conference Proceedings*, 2017, vol. 1885.

[10] P. He, L. Chen, and X. H. Xu, “Fast C4.5,” in *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, ICMLC 2007*, 2007, vol. 5, pp. 2841–2846.

[11] B. HSSINA, A. MERBOUHA, H. EZZIKOURI, and M. ERRITALI, “A comparative study of decision tree ID3 and C4.5,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 4, no. 2, 2014.

[12] S. Sharma and S. Bhatia, “A study of frequent itemset mining techniques,” *Int. J. Eng. Technol.*, vol. 6, no. 4, p. 141, Oct. 2017.

[13] R. Rahim, Nurjamiyah, and A. R. Dewi, “Data Collision Prevention with Overflow Hashing Technique in Closed Hash Searching Process,” *J. Phys. Conf. Ser.*, vol. 930, no. 1, p. 012012, Dec. 2017.

[14] P. harliana and R. Rahim, “Comparative Analysis of Membership Function on Mamdani Fuzzy Inference System for Decision Making,” *J. Phys. Conf. Ser.*, vol. 930, no. 1, p. 012029, Dec. 2017.

[15] R. Rahim, I. Zulkarnain, and H. Jaya, “Double hashing technique in closed hashing search process,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 237, no. 1, p. 012027, Sep. 2017.

[16] R. Rahim, I. Zulkarnain, and H. Jaya, “A review: search visualization with Knuth Morris Pratt algorithm,” in *IOP Conference Series: Materials Science and Engineering*, 2017, vol. 237, no. 1, p. 012026.

[17] B. Singh and H. K. Singh, “Web Data Mining Research : a Survey,” *Computer (Long. Beach. Calif).*, vol. 2, no. 1, pp. 1–10, 2010.

[18] Y. U. Zheng, “Trajectory Data Mining : An Overview,” *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 3, pp. 1–41, 2015.

[19] M. J. C. M. Belinda, R. Umamaheswari, and S. A. David, “Study of high yielding crops cultivation in India using data mining techniques,” vol. 7, no. 1.7, pp. 121–124, 2018.

[20] J. H. Ku and Y. S. Jeong, “A study on social big data analysis using text clustering,” vol. 7, no. 2, pp. 1–4, 2018.

[21] K. Madadipouya, “A Survey on Data Mining Algorithms and Techniques in Medicine,” *JOIV Int. J. Informatics Vis.*, vol. 1, no. 3, pp. 61–71, 2017.

[22] Y. Mardi, “Klasifikasi Menggunakan Algoritma C4.5,” *J. Edik Inform.*, vol. 2, no. 2, pp. 213–219, 2017.

[23] A. P. Muniyandi, R. Rajeswari, and R. Rajaram, “Network anomaly detection by cascading k-Means clustering and C4.5 decision tree algorithm,” in *Procedia Engineering*, 2012, vol. 30, pp. 174–182.

[24] K. Sreenivasa Rao, N. Swapna, and P. Praveen Kumar, “Educational data mining for student placement prediction using machine learning algorithms,” vol. 7, no. 1.2, pp. 43–46, 2018.