Journal of Applied Engineering and Technological Science
Vol 1(2) 2020 : 162-168

UTILIZATION OF DATA MINING TO CLASSIFY THE LOCATIONS OF NEW STREET FOOD BUSINESSES ATTRACTED AND POTENTIALLY BIG PROFITS IN THE CITY OF SURAKARTA

Ismail setiawan[1], Eko Purbiyanto[2]*
Program Studi Manajemen Informatika, AMIK Harapan Bangsa Surakarta
ismailsetiawan@amikhb.ac.id
ekopurbiyanto@amikhb.ac.id*

* *Corresponding Author*

ABSTRACT

This study uses C4.5 algorithm to classify potentially large-profit businesses in the city of Surakarta. The data used are street vendors who are divided into 4 types of merchandise namely snack market, snack market, heavy foods and drinks. The locations that are targeted for classification are Car Free Day. The division of the Car Free Day zone was carried out to find out which areas had more influence on certain foods to get big profits. Car free day zone is divided into 4 parts, namely Purwosari area - rumah makan Diamond, rumah makan Diamond - toko buku Gramedia, toko buku Gramedia - Ngarsopuro and the last one is Ngarsopuro – bundaran Gladakg. Based on the results of the study, the most profitable area to sell is the toko buku Gramedia - Ngarsopuro. Besides this research also classifies based on the ability of the production of raw materials, namely medium and large. The best business category that requires this type of medium-sized raw material is selling at the toko buku Gramedia - Ngarsopuro area, while for the best raw material the best area is the same among Purwosari - Rumah makan Diamonds, Gramedia toko buku - Ngarsopuro and Ngarsopuro – bundaran Gladak.

Keywords : *Classification, C4.5, Business Location, Big Profit, Car Free Day.*

1. INTRODUCTION

The number of street vendors in the city of Surakarta, especially when car free day makes many people want to follow in the footsteps of a sword that has long been selling. Many of them apparently do not know what the most profitable business to be sold at the event. The author sees that there are 4 types of businesses that are commonly found during car free day events.

This research tries to classify which business is the most profitable and which location is the best. The C4.5 algorithm is well known for classifying problems in the form of decision trees. This study tries this approach to be able to provide prospective street vendors in determining what business is good and the right location for them to sell.

The location where data is collected is divided into 4 zoning. For the first zone, which is between Purwosari Station and Diamond Restaurant (Purwosari - Diamond Restaurant). The second zone is located between the Diamond restaurant to Gramedia bookstore (Diamond restaurant - Gramedia bookstore). For the third location starting from Gramedia bookstore to Ngarsopuro night market (Gramedia bookshop - Ngarsopuro). Whereas the last location is the night market of Ngarsopuro until the Gladak roundabout (Ngarsopuro - Gladak roundabout).

The types of businesses that are researched are snack market, snack market, heavy meals and drinks. Snack is a trader who sells typical snack market which are usually only found on car free day. Snack market are a type of snack market but can be easily found even though not in events such as car free days such as chiki bread and candy. For heavy food is food that is usually eaten in the morning and is filling. As for drinks, all kinds of drinks are sold by going around or staying still in their place

2. LITERATURE REVIEW

Research on data mining is mostly done to find out or make decisions regarding the election schedule for the graduation train departure of students choosing the right time to play soccer the choice of soccer team members and so forth. To conduct research on data mining requires a large amount of data. Research on good business locations for selling is still scarce in indexed journals.

Some researchers have conducted research utilizing data mining as has been done by Nugroho in 2014, the prediction of student graduation is done using the c45 algorithm. From these results it was found that the main predictors of student graduation were determined by the majors when they were still in high school. Majoring in Natural Sciences makes it easier for them to graduate on time and has a very satisfying predicate. In addition, research conducted by (Santoso, 2011) regarding the application of the c45 method to predict customer loyalty was found that the main factor customers re-order is the excellent service from the store sales.

When viewed from the two studies above, the use of c45 is able to predict things that can be calculated or quantitative data. In contrast to what (Fakhrurrifqi & Wardoyo, 2013) did where they compared the performance of the nearest neighbor c45 and LVQ algorithms to classify students' abilities. Each algorithm has its weaknesses and strengths while the intellectual ability factor or IQ c45 algorithm is better at making predictions. While the researchers will do the authors still use a tree algorithm like c45 but the writer will choose the id3 algorithm.

3. RESEARCH METHODS

The stages in this study were carried out following the stages in the CRIPS-DM method. As shown in Figure 1. The initial stage starts with understanding the problem that you want to solve (Fakhrurrifqi & Wardoyo, 2013). The problem in this research is the location of the most profitable business when used to sell. The next step is to understand the data that will be used in research (Nugroho, 2014). At this stage the researchers took data from street vendors in the car free day area of the city of Surakarta 6 times by taking data from 150 traders for one time. So that there are 900 merchant data taken.

The data obtained is still mixed and it will be difficult to do the grouping therefore data that has not been included in the discrete category will be changed for computing purposes. The next step is sharing 70:30 data, of which 70 are for model data and 30 for testing data (Andriani, 2012). The results of the formation of the model will be tested with testing data to see the performance of the model being built.

To understand the research method that I use can be seen in Figure 1 as follows

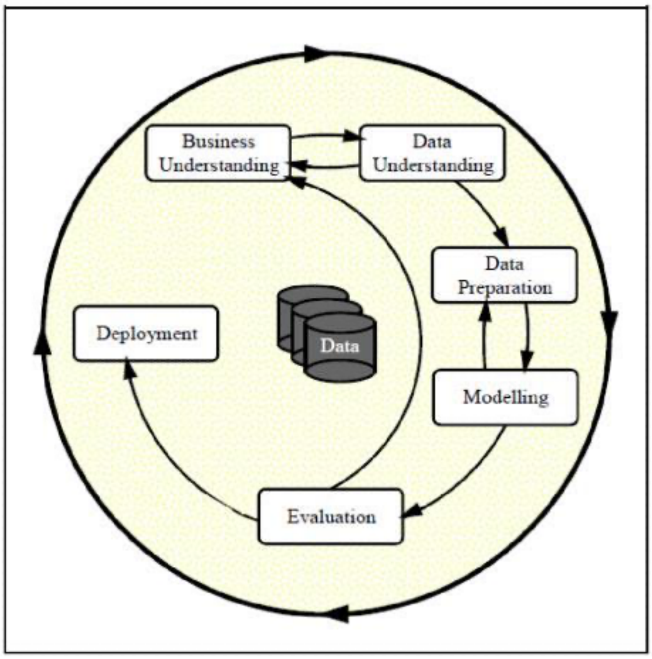

Figure 1. Crips- DM (Source: Azevedo & Santos, 2008)

4. RESULTS AND DISCUSSIONS

A. Business understanding

The approach taken to look at cases comprehensively needs to be seen from a variety of perspectives. The data collected must be able to provide the right information in decision making. This study takes data with the following categories.

1) Trader's name
2) Types of food businesses
3) Ease of getting raw materials
4) Production capacity
5) Initial capital
6) Number of products sold in one day
7) Turnover
8) Maximum trading hours
9) Location of sale
10) Benefits
11) Profits / hour
12) Advantages / products
13) Advantages / (products * products sold)
14) Remarks

For information carried out based on the distribution of turnover per day with initial capital multiplied by 10%, if greater than income per product multiplied by hours of sales, then enter in a large profit so that the location of the place where the tomb is a reference for new traders.

B. Data understanding

The computes process depends on the amount of data inputted. The more data provided, the more expensive the computational process will be. Therefore, this study conducted the selection of predictor attributes and results. Predictor attributes are 9 and result attributes are 1. Predictor attributes selected are as follows:

1) Trader's name

2) Types of food businesses
3) Ease of getting raw materials
4) Production capacity
5) Initial capital
6) Number of products sold in one day
7) Turnover
8) Maximum trading hours
9) Location of sale

While the result attribute is information that has the value "Yes" or: no ".

C. Data preparation

Data preparation is carried out to obtain a complete value so that the computing process does not experience interference (Riwayati, 2014). Based on the results of data collection, no blank data was found. The research object provides complete data from the questionnaire distributed.

The following are sample data collected in this study.

Table 1. Data Collection 1

No	Name of Merchant	Type of Business	Food Ease of Getting Raw Materials	Today's Production Capacity	Initial capital	Sold Products	Turnover	Maximum Trading Hours	Location of sale
8	Traders 8	heavy meal	easy	91	Rp 90.000	88	Rp 308.000	3	CFD (Gramedia up to Ngarsopuro)
9	Traders 9	snack market	easy	84	Rp 90.000	74	Rp 259.000	2	CFD (Diamond up to Gramedia)
13	Traders 13	Drink	easy	96	Rp 50.000	87	Rp 217.500	2	CFD (Gramedia up to Ngarsopuro)
14	Traders 14	snacks	easy	68	Rp 90.000	66	Rp 297.000	3	CFD (Gramedia up to Ngarsopuro)
17	Traders 17	Drink	easy	52	Rp 50.000	28	Rp 70.000	2	CFD (Purwosari up to Diamond)
18	Traders 18	Drink	easy	68	Rp 50.000	65	Rp 162.500	2	CFD (Gramedia up to Ngarsopuro)
22	Traders 22	snack market	easy	50	Rp 90.000	47	Rp 164.500	2	CFD (Gramedia up to Ngarsopuro)
23	Traders 23	snack market	easy	98	Rp 90.000	91	Rp 318.500	2	CFD (Ngarsopuro up to Gladak)
27	Traders 27	snacks	easy	89	Rp 90.000	85	Rp 382.500	3	CFD (Purwosari up to Diamond)
28	Traders 28	snacks	easy	77	Rp 90.000	66	Rp 297.000	3	CFD (Ngarsopuro up to Gladak)

29	Traders 29	snack market	easy	70	Rp 90.000	68	Rp 238.000	2	CFD (Ngarsopuro up to Gladak)
30	Traders 30	snack market	easy	91	Rp 90.000	85	Rp 297.500	2	CFD (Diamond up to Gramedia)
31	Traders 31	Drink	easy	90	Rp 50.000	90	Rp 225.000	2	CFD (Gramedia up to Ngarsopuro)
33	Traders 33	Drink	easy	68	Rp 50.000	68	Rp 170.000	2	CFD (Ngarsopuro up to Gladak)

D. Modelling

The next step is to do the root calculation by finding the entropy of each node (Akter & Wamba, 2016). The formula for finding nodes with the C4.5 algorithm is as follows (Wu & Kumar, 2009):

$$Entropy\ (S) = \sum\nolimits_{j}^{k} = 1 P_j Log_2 P_j \quad \text{...} (1)$$

Information :
- S is the case dataset
- k is the number of S partitions
- pj is the probability obtained from Sum (Yes) divided by Total Cases.

$$gain\ ratio\ (a) = \frac{gain\ (a)}{gain\ (b)} \quad \text{..} (2)$$

Where:
a = attribute.
gain (a) = information gain on attribute a
Split (a) = split information on attribute a

The attribute with the highest Gain Ratio value is chosen as the test attribute for the node (Santoso, 2011). With gain is information gain. This approach applies normalization to information gain by using what is called split information. Split Info expresses entropy or potential information with the formula (Holmes & Jain, 2012):

$$SplitInfo\,(S,A) = -\sum\nolimits_{i=1}^{n} \frac{S_i}{s}\ log_2 \frac{S_i}{s} \quad \text{...} (3)$$

Where:
S = space (data) sample used for training.
A = attribute.
Si = number of samples for attribute i
where Xi represents the i-th subset in sample X.

$$Gain\ (A) = Entropi\ (S) - \sum\nolimits_{i=1}^{k} \frac{|S_i|}{|S|}\ X\ Entropi\ (S_i) \quad \text{..} (4)$$

Where:
S = space (data) sample used for training.
A = attribute.
| Si | = number of samples for V.
| S | = the sum of all sample data.
Entropy (Si) = entropy for samples that have a value of i

The reason for using the gain ratio (a) at C4.5 (not gain (a)) as a criterion in the selection of attributes is that the gain turns out to be biased towards attributes that have many unique values (Sunjana, 2010).

E. Evaluation
In the first calculation result, the main root is production capacity.

Table 2. Early Root Determination

Node			Number of Cases	No	Yes	Entropi	Gain
1	total		470	104	366	0,762492	
	type of food business						0,010364
		Drink	109	30	79	0,848869	
		snacks	119	18	101	0,612986	
		heavy meal	119	24	95	0,725277	
		snack market	123	32	91	0,826992	
	production capacity today						**0,042798**
		Big Production	237	29	208	0,536115	
		Medium Production	233	75	158	0,906423	
	initial capital						0,00355
		Big Capital	361	74	287	0,731789	
		Low Capital	109	30	79	0,848869	
	maximum trading hours						0,00866
		3	238	42	196	0,672295	
		2	232	62	170	0,837478	
	selling locations						0,00245
		CFD (Ngarsopuro up to Gladak)	116	24	92	0,735509	
		CFD (Purwosari up to Diamond)	122	29	93	0,791203	
		CFD (Diamond up to Gramedia)	107	20	87	0,694975	
		CFD (Gramedia up to Ngarsopuro)	125	31	94	0,808093	

Furthermore, it can be seen in the following information:

Production capacity - large - large capital - Purwosari up to Diamond
Production capacity - large - small capital - Ngarsopuro until Gladak roundabouts
Production capacity - large - small capital - Gramedia bookstore to Ngarsopuro
Production capacity - medium - beverage - Purwosari to Diamond
Production capacity - medium - mild pampering - Ngarsopuro until Gladakg roundabout
Production capacity - medium - heavy food - Gramedia book store to Ngarsopuro
Production capacity - medium - hawker - Gramedia book store to Ngarsopuro

For prospective business actors who will sell at the car free day event can see the information above so they can determine what kind of food you want and where the location is good so that it can generate large profits.

5. CONCLUSION

This research shows that the right location for doing business and having big profits starts from the selection of raw materials with medium or large capacity. The right location for medium production capacity is the Gramedia bookstore to

Nagrsopuro. Whereas for large capacity locations, it is divided into 3 regions which share the same percentage of profits, namely Purwosari Station to Diamond Restaurants, Diamond Restaurants to Gramedia Bookstore and Gramedia Bookstore to Ngarsopuro Night Market.

References

Akter, S., & Wamba, S. F. (2016). Big data analytics in E-commerce: a systematic review and agenda for future research. *Electronic Markets*, *26*(2), 173–194. https://doi.org/10.1007/s12525-016-0219-0

Andriani, A. (2012). Penerapan Algoritma C4.5 Pada Program Klasifikasi Mahasiswa Dropout. *Seminar Nasional Matematika*, 139–147. Retrieved from http://demo.pohonkeputusan.com/files/PENERAPAN ALGORITMA C4.5 PADA PROGRAM KLASIFIKASI MAHASISWA DROPOUT.pdf?i=1

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *International Association for Development of the Information Society*, (January), 182–185. https://doi.org/ISBN: 978-972-8924-63-8

Fakhrurrifqi, M., & Wardoyo, R. (2013). Perbandingan Algoritma Nearest Neighbour, C4. 5 dan LVQ untuk Klasifikasi Kemampuan Mahasiswa. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *7*(2), 145–154.

Holmes, D. E., & Jain, L. C. (2012). *Data Mining : Foundation and Intelligent Paradigms Volume 2 : Statistical, Bayesian, Time Series and other Theoretical Aspects*. Berlin: Springer. https://doi.org/10.1007/978-3-642-23242-8

Nugroho, Y. S. (2014). Penerapan Algoritma C4.5 Untuk Klasifikasi Predikat Kelulusan Mahasiswa Fakultas Komunikasi Dan Informatika Universitas Muhammadiyah Surakarta. *Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) 2014*, (November), 1–6. https://doi.org/10.13140/RG.2.1.2734.8247

Riwayati, et al. (2014). Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) 2014 Yogyakarta, 15 November 2014 ISSN: 1979-911X. *Snast*, *3*(November), 211–216. https://doi.org/1979-911X

Santoso, teguh budi. (2011). ANALISA DAN PENERAPAN METODE C4.5 UNTUK PREDIKSI LOYALITAS PELANGGAN. *Jurnal Ilmiah Fakultas Teknik LIMIT'S*, *10*(1). https://doi.org/10.1080/01402390.2011.569130

Sunjana. (2010). Seminar Nasional Aplikasi Teknologi Informasi (SNATI). *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, *2010*(Snati), 24–29. Retrieved from http://journal.uii.ac.id/Snati/article/view/1857

Wu, X., & Kumar, V. (2009). *the top ten algorithms in data mining*. (V. Kumar, Ed.). london: crc press.