Predicting Student Interests Against Laptop Specifications Through Application of Data Mining Using C4.5 Algorithms

by Sufa atin

Submission date: 04-Mar-2024 01:17PM (UTC+0700)
Submission ID: 2311103850
File name: artikel_B.A.20.pdf (925.24K)
Word count: 2577
Character count: 12671

Predicting Student Interests Against Laptop Specifications Through Application of Data Mining Using C4.5 Algorithms

Predicting Student Interests Against Laptop Specifications Through Application of Data Mining Using C4.5 Algorithms

Y R Pratama[1], S Atin[2*], I Afrianto[2]
[1] Information System Department, Universitas Komputer Indonesia, Indonesia
[2] Informatics Engineering Department, Universitas Komputer Indonesia, Indonesia

*sufaatin@email.unikom.ac.id

Abstact. The purpose of this research was to provide information about the prediction of student interest in the laptop specifications through the application of data mining using the C4.5 algorithm. The method used in this research is a survey method by conducting interview. This research was conducted by conducting interviews with several students. The results of this research was a decision tree that describes the laptop specifications that is most in demand by students, so students who want to have a laptop can easily determine laptop specifications based on the number of enthusiasts on certain laptop specifications. The technique used in the application of data mining in this research is a classification technique. The conclusion in this research is that by implementing data mining, students don't need to look for various sources to find laptop specifications that are needed by students in meeting their college needs in long time.

1. Introduction

Laptop specifications are the most important thing to determine which laptops the user will choose. Laptop specifications are determined based on user requirements. In addition to user needs, laptop specifications are usually also determined by the purchasing power of the user. Of course in this case, students as users in this research will choose laptop specifications that are suitable for student needs and student purchasing power. To make it easier for students to determine laptop specifications, data mining is used. Data mining is a process of extracting or filtering data with a large enough data size through certain processes to find useful information from the large data [1]. One technique that is owned by data mining is classification. The technique consists of several methods and produces a decision tree [2]. The C4.5 algorithm is one of the algorithms used in classification techniques. With the implementation of data mining using the C4.5 algorithm, information about the specifications of the laptop that is most in demand by students can be known. Technological development makes data processing more dynamic [3]. Data mining can be used for various things, one of which is data mining is used to obtain information such as determining the specifications of laptops. For example, laptop specifications that are needed by students that are appropriate for the purchasing power of students. Variety of laptop specifications is certainly a major problem for students in determining the required laptop specifications. By utilizing the application of data mining using the C4.5 algorithm, the specifications of the laptop that students need will be drawn through the decision tree. The decision tree is an illustration of the decision procedure for determining the class of a specified variable [4][5]. The decision tree itself can help students determine which laptop specifications should be chosen. However, even though the decision tree has given students an overview of the specifications of the laptop that are most in demand by students, it is possible for students to choose other specifications according to the needs of the student. Therefore, the purpose of this research was to provide information about the prediction of student interest in laptop specifications through the application of data mining using the C4.5 algorithm.

Method

The method used in this research is a survey method by conducting interviews and questionnaires as a research tool. Interviews were conducted and questionnaires were distributed to students of the 7th semester Information System students of Universitas Komputer Indonesia as the population or sample to determine the relationship or influence of each variable tested [6]. The sampling technique used in this research is purposive sampling. Purposive sampling is one of the non-probability sampling techniques that is very effective when a research focuses on certain criteria [7].

3. Results and Discussion

To support the application of data mining using the C4.5 algorithm, of course the sample data is needed. Table 1 contains data from the results of interviews in this study. Sample data were obtained through interviews with several respondents who came from the 7th semester Information System students of Universitas Komputer Indonesia which can be seen in Table 1.

Table 1. Sample Data

No	Respondents	Brand	Processor	VGA	RAM	Operating System	Price	Results
1	Responden 1	Asus	Intel	On Board	2 GB	DOS	Cheap	Purchased
2	Responden 2	Dell	Intel	On Board	2 GB	DOS	Cheap	Not Purchased
3	Responden 3	Axioo	Intel	On Board	2 GB	Windows	Cheap	Not Purchased
4	Responden 4	Acer	Intel	Nvidia	4 GB	Windows	Middle	Purchased
5	Responden 5	Asus	AMD	AMD	4 GB	DOS	Middle	Purchased
6	Responden 6	Asus	Intel	Nvidia	8 GB	Windows	Expensive	Not Purchased
7	Responden 7	HP	Intel	Nvidia	4 GB	Windows	Middle	Not Purchased
8	Responden 8	Lenovo	AMD	AMD	4 GB	DOS	Cheap	Not Purchased
9	Responden 9	Lenovo	AMD	AMD	4 GB	Windows	Middle	Purchased
10	Responden 10	Asus	AMD	On Board	2 GB	DOS	Cheap	Not Purchased
11	Responden 11	HP	Intel	Nvidia	2 GB	Windows	Middle	Purchased
12	Responden 12	Acer	Intel	Nvidia	2 GB	Windows	Middle	Purchased
13	Responden 13	Acer	AMD	AMD	8 GB	Windows	Expensive	Not Purchased
14	Responden 14	Apple	Intel	On Board	2 GB	Mac OS	Cheap	Not Purchased
15	Responden 15	Apple	Intel	On Board	4 GB	Mac OS	Middle	Purchased
16	Responden 16	Asus	Intel	Nvidia	4 GB	Windows	Middle	Purchased
17	Responden 17	Apple	Intel	On Board	4 GB	Mac OS	Cheap	Purchased
18	Responden 18	Axioo	Intel	On Board	2 GB	Windows	Cheap	Not Purchased
19	Responden 19	Dell	Intel	On Board	2 GB	Windows	Cheap	Purchased

No	Respondents	Brand	Processor	VGA	RAM	Operating System	Price	Results
20	Responden 20	Asus	AMD	On Board	2 GB	Windows	Cheap	Purchased

The C4.5 algorithm starts with the process of selecting the highest gain attribute as the root of the tree, then creates a branch for each value, then divides the case in branches, then repeats the process for each branch until all cases in the branch have the same class. To facilitate the application of methodology and system design, the flow of analysis and design is made as shown in Figure 1.

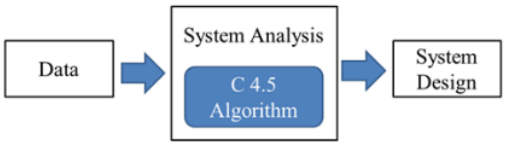

Figure 1. Flow of design and analysis

The flow chart is used to describe the classification process using the C4.5 algorithm can be seen in Figure 2.

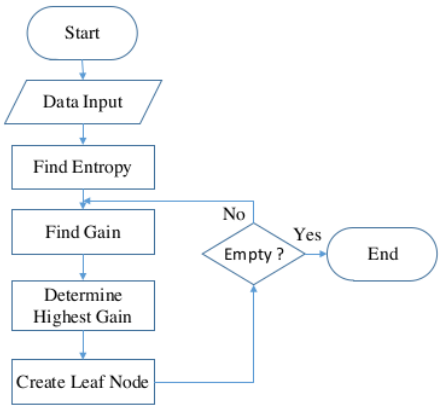

Figure 2. C 4.5 algorithms flowcart

To get accurate calculation results, calculations used Entropy and Gain for each variable [8]. Eentropy measures uncertainty between random variables in a data [9]. The high Entropy value will affect the classification process [10]. The equation used to calculate Entropy and Gain, as follows:

$$Entropi\ (S) = \sum_{j=1}^{k} - p_j\ \log_2\ p_j$$

S : Case set
k : Number of S partition
Pj : Probability obtained from the total (Yes / No) divided by the total case

$$\text{Gain (S,A)} = \text{Entrophy (s)} - \sum_{i=1}^{n} \frac{|si|}{|s|} * \text{Entrophy (si)}$$

S : Case set

A : Attribute
n : Number of A attribute partition
|Si| : Number of cases on the i partition
|S| : Number of S partition

In Table 2, Entropy and Gain calculations have been performed. The gain obtained will affect whether or not the next node will occur. Each internal node is a test node and corresponds to an attribute; the edges leaving a node correspond to the possible values taken on by that attribute [11]. When we calculate Gain in one of the attributes and get the result Gain is the largest of the other attributes, mark the attribute. Next we need to pay attention to the biggest attribute value of the biggest gain attribute to be used as a key on the next node. When the node occurs, the largest gain value is 1, then the node calculation ends. Gain with a value of 1 can be taken the largest attribute value from the number of "Purchased" in the attribute. Entrophy and gain calculation results can be seen in Table 2.

Table 2. Calculation of *Entropy* and *Gain*

Node	Atributs	Value	Number of Cases	Purchased	Not Purchased	Entrophy	Gain
1	**Total**		**20**	**11**	**9**	**0,993**	
	Brand						**0,142**
		Asus	6	4	2	0,918	
		Dell	2	1	1	1,000	
		Axioo	2	0	2	0,000	
		Acer	3	2	1	0,918	
		HP	2	1	1	1,000	
		Apple	3	2	1	0,918	
		Lenovo	2	1	1	1,000	
	Processor						0,003
		Intel	14	8	6	0,985	
		AMD	6	3	3	1,000	
	VGA						0,017
		On Board	10	5	5	1,000	
		Nvidia	6	4	2	0,918	
		AMD	4	2	2	1,000	
	RAM						0,168
		2 GB	10	5	5	1,000	
		4 GB	8	6	2	0,811	
		8 GB	2	0	2	0,000	
	Operating System						0,024
		Windows	12	7	5	0,980	
		DOS	5	2	3	0,971	
		Mac OS	3	2	1	0,918	
	Price						0,086
		Cheap	8	3	5	0,954	
		Middle	8	6	2	0,811	
		Expensive	4	2	2	1,000	
1.1	**RAM : 2GB**		**10**	**5**	**5**	**1,000**	
	Brand						**0,525**
		Asus	3	2	1	0,918	
		Dell	2	1	1	1,000	
		Axioo	2	0	2	0,000	
		Acer	1	1	0	0,000	
		HP	1	1	0	0,000	
		Apple	1	0	1	0,000	
		Lenovo	0	0	0	0,000	

Node	Atributs	Value	Number of Cases	Purchased	Not Purchased	Entrophy	Gain
	Processor						0,000
		Intel	8	4	4	1,000	
		AMD	2	1	1	1,000	
	VGA						0,236
		On Board	8	3	5	0,954	
		Nvidia	2	2	0	0,000	
		AMD	0	0	0	0,000	
	Operating System						0,174
		Windows	6	4	2	0,918	
		DOS	3	1	2	0,918	
		Mac OS	1	0	1	0,000	
	Price						0,035
		Cheap	7	3	4	0,985	
		Middle	3	2	1	0,918	
		Expensive	0	0	0	0,000	

Figure 1 illustrates the decision tree that is the result of Entropy and Gain calculations. The characteristics of decision trees consist of internal nodes, edges and leaf nodes [12]. Internal nodes are usually also called decision nodes namely nodes that represent a variable or a part of a variable. Edges are labels that explain the value or distance of values of a variable. Leaf nodes represent the results in decision making [13]. These three characteristics are inseparable entities. This decision tree can be used as a step to make decisions in choosing a laptop, can be seen ini Figure 3.

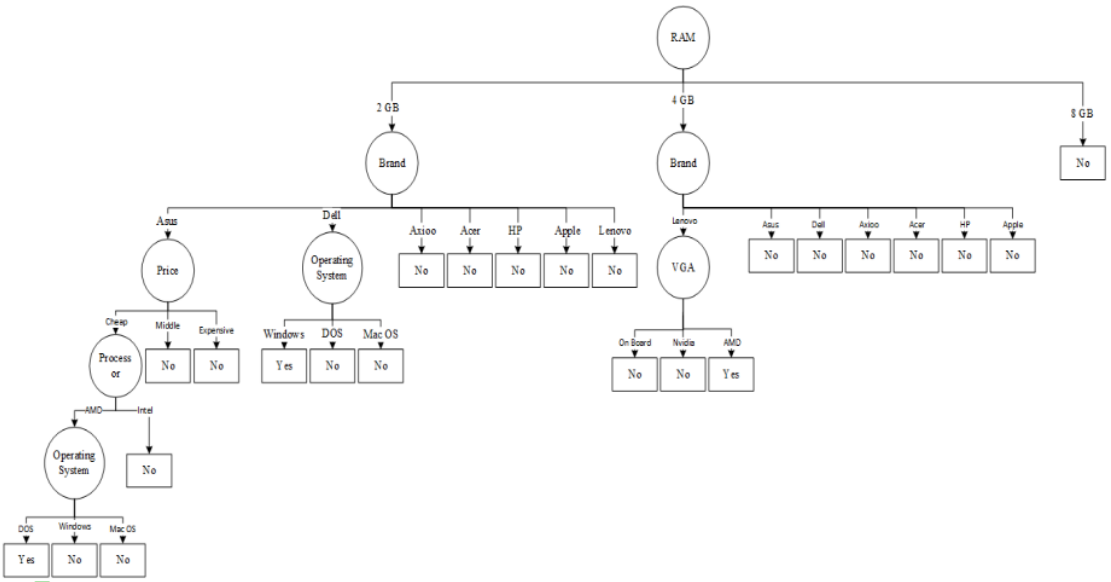

Figure 3. Decision tree result

4. Conclusion

Predicting something through the application of data mining using the C4.5 algorithm makes it easy for students, especially in determining the choice of laptop specifications that are most desirable for students to meet student needs and in accordance with the purchasing power of students. Students no longer need to look for various sources to find laptop specifications that are needed by students in meeting the needs

of students, because the laptop specifications from the results of the data mining application have provided the most desirable specifications of laptops. Based on the results of the data processing process through the application of data mining using the C4.5 algorithm, the laptop specification category is of interest to the 7th semester Information System students of Universitas Komputer Indonesia as follows: (1) RAM 2 GB, Asus, Cheap, AMD Processor, and have a Windows Operating System; (2) RAM 2 GB, Dell, dan have an Windows Operating System; (3) RAM 4 GB, Lenovo, and VGA Card AMD.

Acknowledgements
Researcher is grateful to the team of entrepreneurs who provide insight and guidance that are very helpful in completing this research. This research is expected to be useful for students in determining laptop specifications.

References

[1] Sulastri H and Gufroni A I 2017 Penerapan Data Mining Dalam Pengelompokan Penderita Thalassaemia. *Jurnal Teknologi dan Sistem Informasi*, **3**(2), 299-305.
[2] Haryati S, Sudarsono A and Suryana E 2015 implementasi data mining untuk memprediksi masa studi mahasiswa menggunakan algoritma c4. 5 (studi kasus: universitas dehasen bengkulu). *Jurnal Media Infotama*, **11**(2).
[3] Soegoto E S 2013 *Entrepreneurship Menjadi Pebisnis Ulung*. Elex Media Komputindo.
[4] Utgoff P E 1989 Incremental induction of decision trees. *Machine learning*, **4**(2), 161-186.
[5] Swastina L 2013 Penerapan Algoritma C4. 5 Untuk Penentuan Jurusan Mahasiswa.
[6] Sugiyono 2012 *Metode Penelitian Kuantitatif Kualitatif & R&D*. Bandung : Alfabeta.
[7] Tongco M D C 2007 Purposive sampling as a tool for informant selection. *Ethnobotany Research and applications*, **5**, 147-158.
[8] Korting T S 2006 C4. 5 algorithm and multivariate decision trees. *Image Processing Division, National Institute for Space Research–INPE Sao Jose dos Campos–SP*, Brazil.
[9] Mazid M M, Ali S and Tickle K S 2010 Improved C4. 5 algorithm for rule based classification. In *Proceedings of the 9th WSEAS international conference on Artificial intelligence, knowledge engineering and data bases* (pp. 296-301). World Scientific and Engineering Academy and Society (WSEAS).
[10] Adhatra K, Gaykar A, Dhawan A, Jha R, and Honrao V 2013 Predicting students' performance using ID3 and C4. 5 classification algorithms. *arXiv preprint arXiv*:1310.2071.
[11] Lindell Y and Pinkas B 2000 Privacy preserving data mining. In *Annual International Cryptology Conference* (pp. 36-54). Springer, Berlin, Heidelberg.
[12] Dai W and Ji W 2014 A mapreduce implementation of C4. 5 decision tree algorithm. *International journal of database theory and application*, **7**(1), 49-60.
[13] Magerman D M 1995 Statistical decision-tree models for parsing. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics* (pp. 276-283). Association for Computational Linguistics.

Predicting Student Interests Against Laptop Specifications Through Application of Data Mining Using C4.5 Algorithms

ORIGINALITY REPORT

21%	21%	17%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.coursehero.com Internet Source	5%
2	ifory.id Internet Source	4%
3	Peng Chen. "The Application of an Improved C4.5 Decision Tree", 2021 7th Annual International Conference on Network and Information Systems for Computers (ICNISC), 2021 Publication	4%
4	www.scribd.com Internet Source	2%
5	www.ijnrd.org Internet Source	2%
6	Ruidan Liu, Yu Dong. "Fault Diagnosis of Jointless Track Circuit Based on ReliefF-C4.5 Decision Tree", Journal of Physics: Conference Series, 2022 Publication	2%

Exclude quotes	On	Exclude matches	< 2%
Exclude bibliography	On		