Jurnal Mantik, 6 (3) (2022), ISSN 2685-4236 (Online)

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik

The implementation of machine learning for classifying eligible students for a scholarship at Budidarma University

Elsya Sabrina Asmita Simorangkir[1], Nirwan Yakub[2], Amran Manalu[3], Tarmizi[4], Rian Farta Wijaya[5]

[1,2,3,4,5]Magister Teknologi Informasi/Paskasarjana/Universitas Pembangunan Panca Budi, Indonesia

A R T I C L E I N F O

Article history:

Received Des 9, 2022
Revised Des 20, 2022
Accepted Jan 11, 2023

Keywords:

C4.5
Cart Algorithm
Data Mining
Classification
KIP

ABSTRACT

This study aimed to assist Universitas Budidarma in deciding the recipients of the Kartu Indonesia Pintar (KIP Kuliah) scholarship program. KIP Kuliah is managed by the Ministry of Education, Culture, Research, and Technology and aims to support academically talented students in furthering their education in higher education institutions. The study utilized a Decision Tree data mining classification method with the C4.5 and Cart algorithms. The results showed that if a potential student has the KIP Kuliah scholarship and a high test score, they will pass the verification and validation process. The accuracy of the C4.5 and Cart algorithms was 100% due to the use of matching data in the research. This study aims to make the selection process for KIP Kuliah recipients more efficient and targeted.

Corresponding Author:

Elsya Sabrina Asmita Simorangkir,
Paskasarjana/ Magister Teknologi Informasi
Universitas Pembangunan Panca Budi,
Address (Jl. Jend. Gatot Subroto Km 4,5 Sie Sikambing), Medan, Sumut, 20122, Indonesia.
Email: elsyabrinaas@gmail.com

1. INTRODUCTION

Education plays an important role in the development of a nation and has a big role in improving the quality of human life. In this regard, Indonesia realizes how important education is to society and wants to ensure that every citizen can get a quality and affordable education. However, the cost of education is often an obstacle for people, especially the lower middle class. Therefore, the government is focusing efforts on ensuring equitable distribution and affordability of education for the people. In this study, the main objective was to determine kip recipients studying at Budidarma University using the RapidMiner application and the C4.5 and Cart algorithm methods. By clustering student data variables, data mining can help determine which groups are eligible to receive KIP and those that do not.

To answer the formulation of this problem, this study will use the data of Budidarma University students as training and testing data. Using the RapidMiner application and the C4.5 and Cart algorithms, the system will cluster and create a prediction model for college KIP recipients. Once you have a prediction model, the system will be applied to the new data to predict college KIP recipients. The results of this study

are expected to help the government in determining the recipients of KIP lectures with more targeted and fairness.

The purpose of this study is to determine the recipients of KIP lectures using the RapidMiner application using the C4.5 algorithm and the Cart algorithm. The C4.5 algorithm and the Cart algorithm are data mining methods that will be used in this study to help determine the recipients of KIP lectures. By clustering data taken from student data, this method will help obtain targeted information in terms of the class of prospective KIP recipients and the group that does not receive KIP. Through this research, it is hoped that a better solution will be found to make it easier for the government to determine the recipients of KIP college.

To answer the formulation of the problem, it is necessary to carry out data analysis and modeling using the RapidMiner application and the two algorithms, namely C4.5 and Cart. The C4.5 and Cart algorithms will process data on students who meet the criteria for obtaining college KIP such as achievements, test scores, the amount of parents' income, and homeownership. The results of the analysis and modeling will provide information about prospective KIP recipients and groups that do not receive KIP.

By using data mining, the government can make policies that are more targeted and effective in realizing affordability and equitable distribution of education for the people of Indonesia. This research is very important for the government because the results of this research can help the government in determining the recipients of KIP lectures that are right on target and make better policies for the people of Indonesia.

The method that will be used in this study is to process the data of students who apply for KIP college. The data processed includes achievements, test scores, the amount of parents' income, and homeownership. In using the RapidMiner application, the data will be classified into two groups, namely the group of KIP recipients and the group that does not receive KIP. Then, in this study, the performance of the C4.5 and Cart algorithms will be compared in determining the recipients of KIP lectures.

2. RESEARCH METHOD

In this research method, we will use Data Mining and Data Classification techniques. Data Mining is the process of processing data to transform it into a form that is easier to analyze and useful in making decisions and predicting the future. The C4.5 algorithm will be used as one of the data mining classification methods that will help explore the data and find the relationship between the input variable and the target variable. The first step in this process is to calculate the entropy to assist in the creation of the decision tree. This process uses mathematical techniques, statistics, and Artificial Intelligence technology to ensure accurate results.

In the C4.5 algorithm, the first step after preparing the data is to train the selection of attributes that can be calculated using the concept of entropy. Entropy states input a collection of objects. Here is the entropy calculation formula:

$$Entropy(S) = \sum_{i=n}^{n} - pi * log_2(pi)$$

Keterangan:
S = Himpunan Kasus
n = Jumlah partisi S
p_i = probabilitas yang didapat dari jumlah kelas dibagi total kasus

After calculating the entropy value in the C4.5 algorithm the selection of attributes is carried out using Information Gain. To calculate the gain, which can be calculated by the following formula:

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si)$$

Keterangan :
S = Himpunan kasus
A = Atribut
n = Jumlah atribut
|Si| = Jumlah partisi ke -i
|S| = jumlah kasus dalam S

Cart algorithm (Classification and Regression Tree) one of the methods or algorithms of one of the data exploration techniques is the decision tree technique. CART was developed to perform classification analysis.

If a data set D contains examples from n classes, gini index (D) is defined as

$$gini\ (D) = 1 - \sum_{j=1}^{n} p_j^2$$

Where p_j the relative frequency of class j in D

If a data set D is split on A into two subsets D_1 and D_2, the gini index gini (D) is defened as

$$gini_{\ A}(D) = \frac{|D_1|}{|D|}\ gini\ (D_1) + \frac{|D_2|}{|D|}\ gini\ (D_2)$$

Reduction in Impurity:

$$\Delta gini\ (A) = gini(D) - gini_{\ A}(D)$$

The attribute provides the smallest gini $_{split}$ (D) (or the largest redution in impurity) is chosen to split the node (need to enumerate all the possible spliting for each attibute)

3. RESULTS AND DISCUSSIONS

The results of this study are expected to determine the accuracy value of the C4.5 Algorithm and the Cart algorithm in the classification of prospective KIP recipients who meet the criteria, by processing the data and selecting the necessary attributes, then testing the data with manual calculations and RapidMiner software. The results of this study are in the form of a calculation process based on the C4.5 algorithm and the Cart algorithm. The following is a sample of the data used:

Table 1. Data on KIP College Candidates

No	Name/Alias	Prestasi	Nilai Ujian	Jumlah Penghasilan Orang Tua	Punya KIP/Sejenisnya	Kepemilikan Rumah	Status
1	Calon 1	Ada	55	3,5 JT	Ya	Sewa	Tidak Lulus
2	Calon 2	Tidak	75	2,5 Jt	Tidak	Sendiri	Tidak Lulus
3	Calon 3	Ada	77	2,2 Jt	Ya	Sewa	Lulus
4	Calon 4	Ada	50	2,3 Jt	Tidak	Sendiri	Tidak Lulus
5	Calon 5	Ada	91	2,0 Jt	Ya	Sendiri	Lulus
6	Calon 6	Ada	80	2,5 Jt	Ya	Sendiri	Lulus
7	Calon 7	Tidak Ada	45	1,2 Jt	Tidak	Sendiri	Tidak Lulus
8	Calon 8	Tidak Ada	60	2,9 Jt	Tidak	Sewa	Tidak Lulus
9	Calon 9	Tidak Ada	65	3,1 Jt	Tidak	Sendiri	Tidak Lulus
10	Calon 10	Tidak Ada	80	3,4 Jt	Tidak	Sendiri	Tidak Lulus

11	Calon 11	Ada	78	2,8 Jt	Tidak	Sendiri	Tidak Lulus
12	Calon 12	Tidak Ada	40	2,1 Jt	Ya	Sendiri	Tidak Lulus
13	Calon 13	Tidak Ada	75	1,1 JT	Ya	Sendiri	Lulus
14	Calon 14	Tidak Ada	95	1,5 Jt	Ya	Sendiri	Lulus
15	Calon 15	Ada	80	2,2 Jt	Ya	Sendiri	Lulus
16	Calon 16	Tidak Ada	58	2,5 Jt	Tidak	Menumpang	Tidak Lulus
17	Calon 17	Tidak Ada	70	2,6 Jt	Tidak	Sendiri	Tidak Lulus
18	Calon 18	Tidak Ada	92	1,8 Jt	Ya	Sendiri	Lulus
19	Calon 19	Tidak Ada	47	1,7 Jt	Ya	Sendiri	Tidak Lulus
20	Calon 20	Tidak Ada	59	2,9 Jt	Tidak	Menumpang	Tidak Lulus
21	Calon 21	Tidak Ada	88	3,8 Jt	Tidak	Sendiri	Tidak Lulus
22	Calon 22	Tidak Ada	87	3,9 Jt	Tidak	Sendiri	Tidak Lulus

Cart Algorithm Calculation (Data Preprocessing) some data are numerical in form, therefore, it is changed into categories to facilitate the calculation process made in the form of groupings, namely Test Scores and Parental Income.

Table 2. Test Scores

Nilai Ujian	Kategori
<60	Rendah
60 s.d 80	Sedang
>80	Tinggi

Table 3. Number of Parents' Income

Jumlah Penghasilan Orang Tua	Kategori
Dibawah atau sama dengan 2 Jt	<=2 Jt
Diatas 2 Jt	>2 Jt

Data tersebut dilakukan proses normalisasi sehingga datanya menjadi seperti table berikut :

Tabel 4. Praprosessing Data (Normalisasi)

No	Nama/Alias	Prestasi	Nilai Ujian	Jumlah Penghasilan Orang Tua	Punya KIP/Sejenisnya	Kepemilikan Rumah	Status
1	Calon 1	Ada	Rendah	>2 Jt	Ya	Sewa	Tidak Lulus
2	Calon 2	Tidak Ada	Sedang	>2 Jt	Tidak	Sendiri	Tidak Lulus
3	Calon 3	Ada	Sedang	>2 Jt	Ya	Sewa	Lulus
4	Calon 4	Ada	Rendah	>2 Jt	Tidak	Sendiri	Tidak Lulus
5	Calon 5	Ada	Tinggi	<=2 Jt	Ya	Sendiri	Lulus
6	Calon 6	Ada	Sedang	>2 Jt	Ya	Sendiri	Lulus
7	Calon 7	Tidak Ada	Rendah	<=2 Jt	Tidak	Sendiri	Tidak Lulus
8	Calon 8	Tidak Ada	Sedang	>2 Jt	Tidak	Sewa	Tidak Lulus
9	Calon 9	Tidak Ada	Sedang	>2 Jt	Tidak	Sendiri	Tidak Lulus
10	Calon 10	Tidak Ada	Sedang	>2 Jt	Tidak	Sendiri	Tidak Lulus
11	Calon 11	Ada	Sedang	>2 Jt	Tidak	Sendiri	Tidak Lulus
12	Calon 12	Tidak Ada	Rendah	>2 Jt	Ya	Sendiri	Tidak Lulus
13	Calon 13	Tidak Ada	Sedang	<=2 Jt	Ya	Sendiri	Lulus
14	Calon 14	Tidak Ada	Tinggi	<=2 Jt	Ya	Sendiri	Lulus
15	Calon 15	Ada	Sedang	>2 Jt	Ya	Sendiri	Lulus
16	Calon 16	Tidak Ada	Rendah	>2 Jt	Tidak	Menumpang	Tidak Lulus
17	Calon 17	Tidak Ada	Sedang	>2 Jt	Tidak	Sendiri	Tidak Lulus
18	Calon 18	Tidak Ada	Tinggi	<=2 Jt	Ya	Sendiri	Lulus
19	Calon 19	Tidak Ada	Rendah	<=2 Jt	Ya	Sendiri	Tidak Lulus
20	Calon 20	Tidak Ada	Rendah	>2 Jt	Tidak	Menumpang	Tidak Lulus
21	Calon 21	Tidak Ada	Tinggi	>2 Jt	Tidak	Sendiri	Tidak Lulus
22	Calon 22	Tidak Ada	Tinggi	>2 Jt	Tidak	Sendiri	Tidak Lulus

The data mentioned above is processed by entering the following formula:
First: Value the value of Entroy by the formula:

$$Entropy(S) = \sum_{i=n}^{n} - pi * log_2(pi)$$

Second: Look for Gain Value
Results on Node 1/Table below:

Table 5. Node 1 Results

Node		Nilai Attribute	Jumlah Kasus	Lulus	Tidak Lulus	Entropy	Gain
1	Total		22	7	15	0,902	
	Prestasi						0,135818182
		Ada	7	5	2	0,863	
		Tidak Ada	15	3	12	0,721	
	Nilai Ujian						0,240636364
		Rendah	7	0	7	0	
		Sedang	10	4	6	0,97	
		Tinggi	5	3	2	0,97	
	Jumlah Penghasilan Orang Tua						0,145454545
		<=2 Jt	6	4	2	0,918	
		>2 Jt	16	3	13	0,696	
	Punya KIP/Sejenisnya						0,501545455
		Ya	10	7	3	0,881	
		Tidak	12	0	12	0	
	Kepemilikan Rumah						0,053545455
		Menumpang	2	0	2	0	
		Sewa	3	1	2	0,918	
		Sendiri	17	6	11	0,936	

From the gain calculation, it is found that the highest gain is having KIP / Similar so that the root node is Have KIP / Similar as in the following figure:

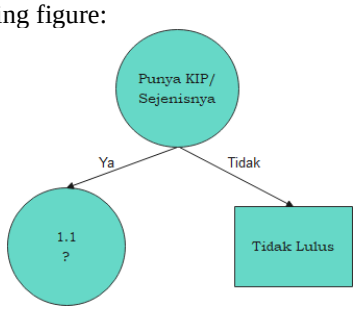

Figure 1. Root Nodes Have KIP

From the decision tree, there is still a branch that does not have a decision, namely the "Yes" branch. Therefore, the entropy and gain value searches are carried out the same as before, except that the amount of data used is less, only the "Yes" value on the Has KIP / Similar attribute. Then the calculation process is as follows:

Table 6. Node 1.1 Calculation

c		Nilai Attribute	Jumlah Kasus	Lulus	Tidak Lulus	Entropy	Gain
1.1	Ya-Prestasi		10	7	3	0,881	
	Prestasi						0,0355
		Ada	5	4	1	0,721	

	Tidak	5	3	2	0,97	
Nilai Ujian						0,881
	Rendah	3	0	3	0	
	Sedang	4	4	0	0	
	Tinggi	3	3	0	0	
Jumlah Penghasilan Orang Tua						0,0355
	<=2 Jt	5	4	1	1	
	>2 Jt	5	3	2	0,97	
Kepemilikan Rumah						-0,0238
	Menumpang	0	0	0	0	
	Sewa	2	1	1	1	
	Sendiri	8	6	2	0,811	

In this calculation, the highest gain is the test score. All branches on the entropy test score are 0 so the test score branch already has a decision.

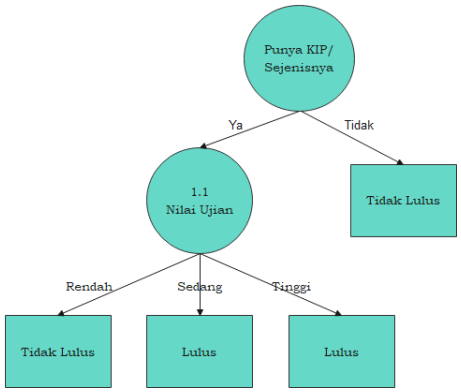

Figure 2. Test Score Root Node

In this calculation, the highest gain is the test score. All branches on the entropy test score are 0 so the test score branch already has a decision.

Perhitungan Algoritma Cart

Table 7. Iteration Data 1

		l	r	Class	l	r	Q(s|t)	2*P*Pr	Φ	
Prestasi	Ada	7	15	Lulus	4	3	0.7428571	0.4339	0.32231405	
		0.31818182	0.681818		0.57143	0.2				
				Tidak Lulus	3	12				
					0.42857	0.8			Φ	
Nilai Ujian	Rendah	7	15	Lulus	0	7	0.9333333	0.4339	0.40495868	
		0.31818182	0.681818		0	0.46667				
				Tidak Lulus	7	8				
					1	0.53333			Φ	
	Sedang	10	12	Lulus	4	3	0.3	0.4959	0.14876033	
		0.45454545	0.545455		0.4	0.25				
				Tidak Lulus	6	9				
					0.6	0.75			Φ	
	Tinggi	5	17	Lulus	3	4	0.7294118	0.3512	0.25619835	
		0.22727273	0.772727		0.6	0.23529				
				Tidak Lulus	2	13				
					0.4	0.76471			Φ	
Jumlah Penghasilan Orang	<=2 Jt	6	16	Lulus	4	3	0.9583333	0.3967	0.38016529	
		0.27272727	0.727273		0.66667	0.1875				
				Tidak Lulus	2	13				
					0.33333	0.8125			Φ	
Punya KIP/Sejenisnya	Ya	10	12	Lulus	7	0	1.4	0.4959	0.69421488	Akar
		0.45454545	0.545455		0.7	0				
				Tidak Lulus	3	12				
					0.3	1			Φ	
Kepemilikan Rumah	Sewa	3	19	Lulus	1	6	0.0350877	0.2355	0.00826446	
		0.13636364	0.863636		0.33333	0.31579				
				Tidak Lulus	2	13				
					0.66667	0.68421			Φ	
	Menumpang	2	20	Lulus	2	7	1.3	0.1653	0.21487603	
		0.09090909	0.909091		1	0.35				
				Tidak Lulus	0	13				
					0	0.65			Φ	
	Sendiri	17	4	Lulus	6	0	0.7058824	0.281	0.19834711	
		0.77272727	0.181818		0.35294	0				
				Tidak Lulus	11	4				
					0.64706	1				

Table 8. Iteration Data 2

No	Nama	Prestasi	Nilai Ujian	Jumlah Penghasilan Orang Tua	Punya KIP/ Sejenisnya	Kepemilikan Rumah	Status
5	Agus Minta Riang Zega	Ada	Tinggi	<=2 Jt	Ya	Sendiri	Lulus
14	Amran Saleh Harahap	Tidak Ada	Tinggi	<=2 Jt	Ya	Sendiri	Lulus
18	Desika Marbun	Tidak Ada	Tinggi	<=2 Jt	Ya	Sendiri	Lulus

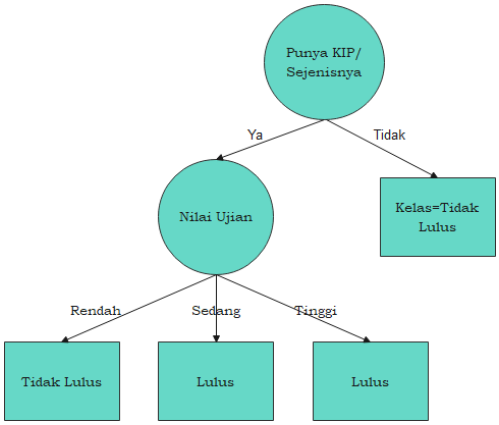

Figure 3. Agoritma Cart

RapidMiner is an open-source software. Furthermore, with the same data, testing was carried out with the Rapidminer application to test the results of manual calculations of the cart algorithm and the calculation results as shown in figure 4 below:

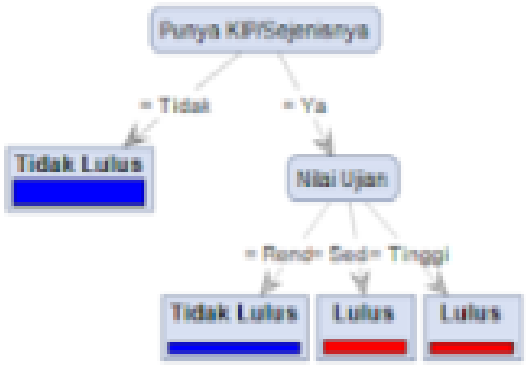

Figure 4. Decision Tree with RapidMiner

4. CONCLUSION

Based on the testing results, it can be concluded that the implementation of the C4.5 and Cart algorithms for classifying eligible college students for the KIP scholarship begins with analyzing the data needs, then conducting research and testing on several sample data to obtain the accuracy of the predictions made by the system created. Based on the testing results, the conclusion is: if they don't have KIP/similar then KIP scholarship has not passed, if they have KIP scholarship and low test scores then not passed, if they have KIP scholarship and average test scores then passed, if they have KIP scholarship and high test scores then passed. Therefore, the accuracy level of the implementation of the C4.5 and Cart algorithms is 100% because the data used in this research has the same results.

REFERENCES

Alverina, D., Chrismanto, A. R., & Santosa, R. G. (2018). Perbandingan Algoritma C4.5 dan CART dalam Memprediksi Kategori Indeks Prestasi Mahasiswa. *Jurnal Teknologi Dan Sistem Komputer*, *6*(2), 76–83. https://doi.org/10.14710/jtsiskom.6.2.2018.76-83

Amin, M. F. (2017). Penerapan Algoritma Cart Untuk Memprediksi Status Kelulusan Mahasiswa. *Jutisi: Jurnal Ilmiah Teknik Informatika Dan Sistem Informasi*, *5*(3).

Eman, D., & Emanuel, A. W. R. (2019). *Machine Learning Classifiers for Autism Spectrum Disorder*.

Gusriani, N., & Parmikanti, K. (2015). Klasifikasi Ketepatan Masa Studi Mahasiswa FMIPA Unpad Angkatan 2001-2006 dengan Menggunakan Metode Classification and Regression Trees (CART). *Jurnal Matematika Integratif*, *11*(1), 7–14.

Heni Sulistiani, Y. T. U. (2018). Penerapan Algoritma Klasifikasi Sebagai Pendukung Keputusan Pemberian Beasiswa Mahasiswa. *Snti*, *April*, 300–305. https://doi.org/10.31227/osf.io/yuavj

Kamagi, D. H. (2014). *Implementasi data mining dengan algoritma c4. 5 untuk memprediksi tingkat kelulusan mahasiswa (studi kasus: program studi teknik informatika universitas multimedia nusantara)*. Universitas Multimedia Nusantara.

Kamagi, D. H., & Hansun, S. (2014). Implementasi Data Mining dengan Algoritma C4. 5 untuk Memprediksi Tingkat Kelulusan Mahasiswa. *Ultimatics: Jurnal Teknik Informatika*, *6*(1), 15–20.

Khotimah, K. (2021). Teknik Data Mining menggunakan Algoritma Decision Tree (C4. 5) untuk Prediksi Seleksi Beasiswa Jalur KIP pada Universitas Muhammadiyah Kotabumi. *Jurnal SIMADA (Sistem Informasi Dan Manajemen Basis Data)*, *4*(2), 145–152.

Rahmayuni, I. (2014). Perbandingan performansi algoritma c4. 5 dan cart dalam klasifiksi data nilai mahasiswa prodi teknik komputer politeknik negeri padang. *Jurnal Teknoif Teknik Informatika Institut Teknologi Padang*, *2*(1), 40–46.

Sinambela, Y. E. S. (n.d.). *Penerapan metode klasifikasi dengan algoritma cart pada data status daerah kabupaten di Indonesia*.

Susetyoko, R., Yuwono, W., & Purwantini, E. (2022). Model Klasifikasi Pada Seleksi Mahasiswa Baru Penerima KIP Kuliah Menggunakan Regresi Logistik Biner. *Jurnal Informatika Polinema*, *8*(4), 31–40.

TL, D. I., Widowati, A. I., & Surjawati, S. (2017). Faktor-faktor yang mempengaruhi prestasi akademik: Studi kasus pada mahasiswa Program Studi Akuntansi Universitas Semarang. *Jurnal Dinamika Sosial Budaya*, *18*(1), 39–48.

Untari, D., Hastuti, K., Hidayat, E. Y., Untari, D., Limão, N., & Gaol, N. Y. L. (2014). Data Mining Untuk Menganalisa Prediksi Mahasiswa Berpotensi Non-Aktif Menggunakan Metode Decision Tree C4. 5. *Program Studi Teknik Informatika-Fakultas Ilmu Komputer-Universitas Dian Nuswantoro*.

Widagdo, K. A. (2010). Pembentukan Pohon Klasifikasi Biner dengan Algoritma CART (Classification And Regression Trees)(Studi Kasus Penyakit Diabetes Suku Pima Indian). *Skripsi. Universitas Diponegoro: Semarang*.