JURNAL RISET INFORMATIKA **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 2, No. 3 June 2020 DOI: https://doi.org/10.34288/jri.v2i3.75

IMPLEMENTATION OF C4.5 METHOD TO DETERMINE THE FACTOR OF BEING LATE FOR COMING TO SCHOOL

Cendana Puspitasari

Information Systems Study Program
STMIK Nusa Mandiri Jakarta, Indonesia
www.nusamandiri.ac.id
cendanapuspitasari834@gmail.com

Abstract

Being late to school is a deviant act that violates the rules or regulations in the school, both written and unwritten. The discipline of students coming to school is the first to see. Some common factors that cause delays include distance to school, wake-up, departure hours, travel conditions, and vehicles used. In this study, the authors used the Classification algorithm with the C4.5 and Forward selection methods. The sample used was questionnaire data for class VIII (eight) State Junior High School students 271, totaling 270 students. Using training data, certain attributes are determined to form a classifier model. The results of this study are the results of the accuracy of the C4.5 method of 60.74%, with the results of the tree showing congestion is a factor of school delay and the results of the accuracy of 65.93% for Forward selection and the three best attributes.

Keywords: C4.5 Method, Data Mining, Forward selection, Delay, Classification

Abstrak

Terlambat hadir ke Sekolah merupakan tindakan menyimpang yang menyalahi aturan atau tata tertib yang ada di sekolah baik tertulis maupun tidak tertulis. Kedisiplinan siswa datang ke sekolah menjadi yang pertama yang dilihat, beberapa faktor umum yang terjadi keterlambatan bisa terjadi antara lain: jarak ke sekolah, jam bangun, jam berangkat, kondisi perjalanan, dan kendaraan yang dipakai. Dalam penelitian ini, penulis menggunakan algoritma Klasifikasi dengan metode C4.5 dan metode Forward selection .Sampel yang digunakan adalah data kuesioner siswa-siswa kelas VIII (delapan) di Sekolah Menengah Pertama Negeri 271 sebanyak 270 siswa. Penggunaan data training, ditentukan atribut-atribut tertentu untuk membentuk model classifer. Hasil dari penelitian ini adalah hasil akurasi metode C4.5 sebesar 60,74% dengan hasil tree menunjukkan kemacetan merupakan faktor keterlambatan sekolah dan hasil akurasi 65.93% untuk Forward selection dan mendapatkan 3 atribut terbaik.

Kata Kunci : Metode C4.5, Data Mining, Forward selection, Keterlambatan, Klasifikasi

INTRODUCTION

Being late in coming to school is a deviant act that violates the rules or regulations in the school both written and unwritten. Teaching discipline to school-age children or students is very important to make students understand the rules that exist in the school. However, the facts on the ground show that there are still many students who often arrive late to school. Being late coming to the school is not without reason (Sulistiyono 2018), various reasons are expressed by students who are often late. Violations of the rules are often found in schools (Akhmad Rizkon 2019) generally done by students. Violation is the act of violating the rules by someone intentionally (Insyiroh and Naqiyah 2017)

There are several reasons why students come late to school, late at night, watch movies or TV too late, the habit of getting up late, the distance from home to school (Pramono et al. 2018), and travel conditions and vehicles used (Amirulloh and Taufiqurrochman 2017)

The algorithm used is the C4.5 method and the type of feature selection used is Forward selection, which is managing variables that have been selected through feature selection and increasing prediction accuracy (Saleh 2017).

The purpose of this research is to help overcome the problem of students coming to school late, being able to understand the factors of students being late in coming to school, and

helping the school in tackling the factors that occur in late coming to school.

RESEARCH METHODS

In this writing using qualitative research methods. The qualitative method in the form of making observations, direct interviews, and getting the results of respondents directly from the field through a questionnaire to all class VIII SMP.

Research Stages

At this writing, there are stages of research that have been carried out, namely in Figure 1 below:

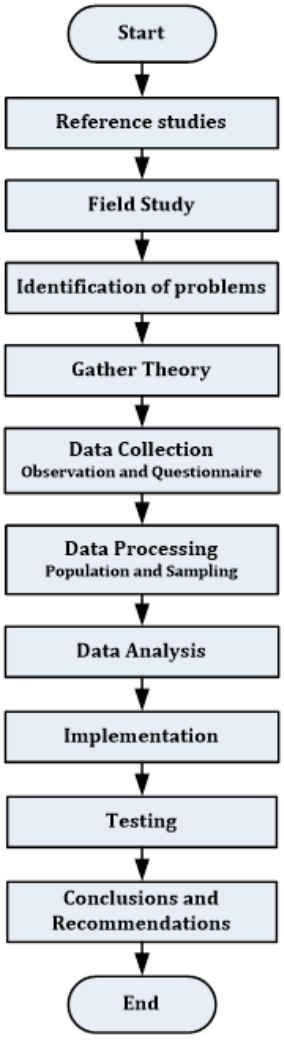


Figure 1. Research Stages

The stages of research in this study are as follows:

1. Reference studies, using reference references in the form of journals and several other sources;
2. Field Study, visit various schools and see the problems that are most often found. Problem Identification, after seeing the problem the author starts looking for problems related to the problem you want to examine.
3. Gather Theory, gather several theories to solve the problem you want to study.
4. Data Collection, at this stage direct observation or ask the school what stages have been carried out and conduct questionnaires directly to students.
5. Data Processing, this stage is the collection of the entire population and take samples to be tested.
6. Data Analysis, this technique uses the CRIPS-DM method.
7. Implementation, this technique uses Rapid Miner with the C4.5 method and Forward selection.
8. Testing, at this stage using Cross-Validation as a comparison / to know the consistency in using the C4.5 method and Forward selection or without Forward selection.

Data Collection Stages

The author collected references about the classification algorithm and made direct observations at SMP Negeri 271 in the morning. After that, the authors conducted interview sessions with several schools regarding the case of students who arrived late. The author also conducted a questionnaire to all students of class VIII.

Population and Sample

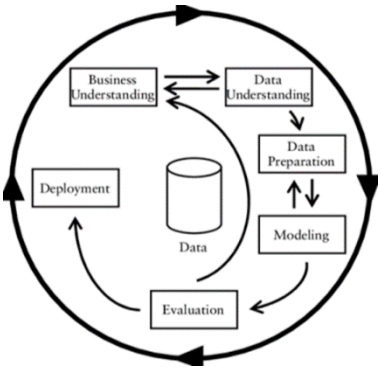
The population in this study is the whole of the students of class VIII in State Junior High School 271 as many as 277 students.

The size or number of samples is very dependent on the magnitude of the level of accuracy or error desired by the researcher. The greater the error rate, the smaller the number of samples, and the greater the number of samples approaching the population, the smaller the chance of generalization errors and vice versa. The sample formula used is the Slovin Formula. In this study using 1%.

$$n = \frac{N}{1+Ne^2} \quad \text{..... (1)}$$

Data analysis

In this study there are data analyzes that have been carried out, which are as follows:



Source: (Budiman, Prahasto, and Christyono 2012)

Figure 2. CRISP-DM methodology

The data analysis method in this study uses the CRISP-DM process steps. CRISP-DM stands for Cross-Industry Standard Process for Data Mining and has 6 stages, as follows: Business Understanding; Data Understanding; Data Preparation; Modeling; and Evaluation.

Data Set

The data set from the survey results were 270 students in SMPN 271. Consisting of the respondent's identity and the variables to be tested. The respondent's identity is in the form of Name, Gender, NIS, Age. While the variables used for testing are Departure Hours, Wake Hours, Distance, Vehicles, Congestion, Red Lights, and Delay.

Training Data

Training data is sample data used by classification algorithms (decision trees). In the training data, certain attributes are determined to form the classifier model. The following training data such as table 1 below.

Table 1. Pieces of Training Data

No	Jam Berangkat	Jam Bangun	Jarak (KM)	Kendaraan	Kemacetan	Lampu Merah	Keterlambatan
1	Siang	Pagi	Dekat	Motor	Sedang	Ada	Terlambat
2	Siang	Cukup Siang	Dekat	Jalan Kaki	Tidak	Tidak Ada	Tidak
3	Siang	Pagi	Dekat	Jalan Kaki	Tidak	Tidak Ada	Tidak
4	Siang	Cukup Siang	Dekat	Motor	Sedang	Ada	Terlambat
5	Pagi	Cukup Siang	Dekat	Motor	Sedang	Ada	Tidak
6	Siang	Pagi	Dekat	Motor	Tidak	Tidak Ada	Tidak
7	Siang	Pagi	Dekat	Motor	Tidak	Tidak Ada	Tidak
8	Pagi	Cukup Siang	Dekat	Motor	Sedang	Ada	Tidak
9	Siang	Cukup Siang	Dekat	Jalan Kaki	Tidak	Tidak Ada	Tidak
10	Siang	Pagi	Dekat	Motor	Tidak	Tidak Ada	Tidak

Testing

Cross-validation testing is performed to determine the consistency of the performance of the classification system with the most invariant feature extraction method for rotation (Nurhasan, Hikmah, and Utami 2018).

RESULTS AND DISCUSSION

Manual Process C4.5

In this study the data set used consisted of 270 records with 6 regular attributes and 1 special attribute as a target, following an example of training data used in this study. = ((-104)/270)*(log)$_2$ (104/270)+((-166)/270) *(log)$_2$ (166/270)= 0,9616.

Table 2. Calculation of Entropy and Gain Node 1

Node	Atribut	Value	SUM(Score)	Belated	Not belated	Entropy	Gain
1	JamBerangkat	Pagi	63	21	42	0,91829	
		Cukup Siang	206	82	124	0,9698	
		Siang	1	1	0	0	
							1,4872
	Jam Bangun	Pagi	178	70	108	0,9668	
		Cukup Pagi	91	33	58	0,9448	
		Siang	1	1	0	0	
							0,6426
	Jarak	Dekat	241	88	153	0,9468	
		Sedang	20	12	8	0,9709	
		Jauh	9	4	5	0,9910	
							0,2214
	Kendaraan	Motor	222	91	131	0,9764	
		Jalan Kaki	37	8	29	0,7531	
		Sepeda	2	0	2	0	

Node	Atribut	Value	SUM(Score)	Belated	Not belated	Entropy	Gain
		Umum	9	5	4	0,9910	
							0,2950
	Kemacetan	Macet	5	1	4	0,7219	
		Sedang	116	62	54	0,9965	
		Tidak	149	41	108	0,8487	
							1,8448
	Lampu Merah	Tidak Ada	114	50	64	0,9890	
		Ada	156	54	102	0,9305	
		Tidak Ada	114	50	64	0,9890	
							1,0816

Rapidminer Process C4.5

Following are the steps in the process of making a Decision tree, as follows::

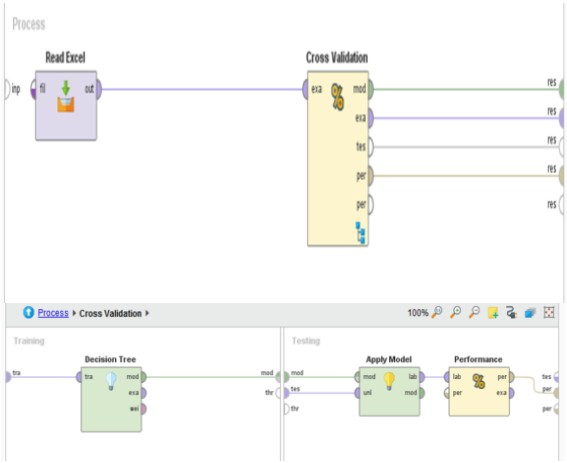


Figure 2 Decision tree process in Rapidminer

Next in Figure 2 is the Decision Tree process in Rapidminer using Cross-Validation (without forward selection)

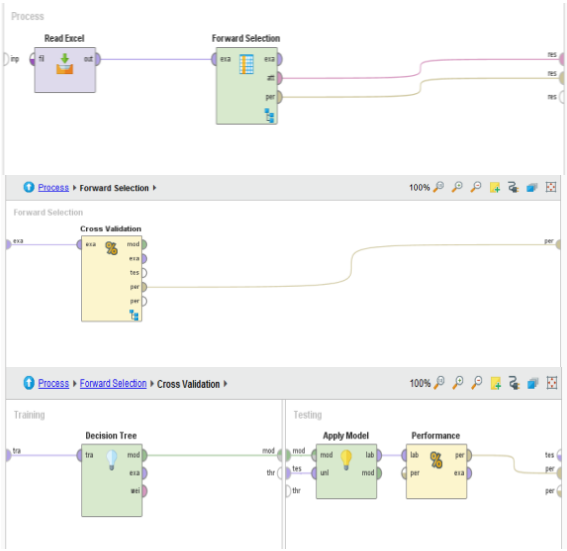


Figure 3 Decision tree process with Forward selection in Rapidminer

Next in Figure 3 is the Decision Tree process by using the Forward selection on the Rapidminer as a comparison material with no Forward selection

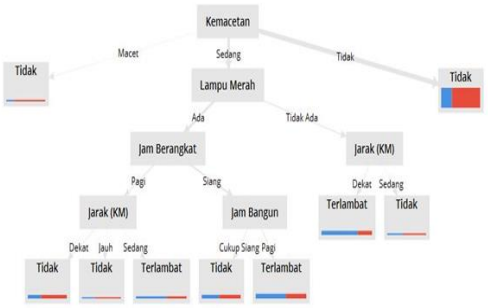


Figure 4. Implementation of Decision Tree in Rapidminer

Next in Figure 4 is the result of the Decision Tree, namely "Kemacetan" is above or the first problem factor in the delay in coming to school.

accuracy: 60.74% +/- 7.44% (mikro: 60.74%)

	true Terlambat	true Tidak	class precision
pred. Terlambat	40	42	48.78%
pred. Tidak	64	124	65.96%
class recall	38.46%	74.70%	

Figure 5. Decision Tree Accuracy Results

Next in Figure 5 is the accuracy of the Decision Tree 60,74%.

accuracy: 65.93% +/- 9.91% (mikro: 65.93%)

	true Terlambat	true Tidak	class precision
pred. Terlambat	44	32	57.89%
pred. Tidak	60	134	69.07%
class recall	42.31%	80.72%	

Figure 6. Decision tree accuracy results with Forward selection

The following figure 6, is the result of accuracy with added Forward selection 65,93%.

Table 3. Decision tree results and Forward Selection

Atribute	Weight
Jam Berangkat	1
Jam Bangun	0
Jarak (KM)	0
Kendaraan	1
Kemacetan	1
Lampu Merah	0

Based on table 3 above, the results of evaluation and validation produce the 3 best attributes or those with a value of 1 and 3 with a value of 0. It is found that the attributes of "Jam Beragkat", "Kendaraan" and "Kemacetan" are the best attributes.

Here are the results of a comparison of Rapidminer evaluations using Forward selection and without using Forward selection

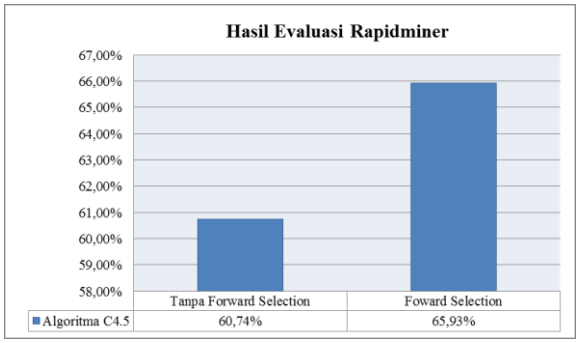


Figure 7 Comparison of Results Evaluation RapidMiner

In figure 7, based on the results of the study, the results are the use of forward selection by 65.93% and without using forward selection is 60.74%.

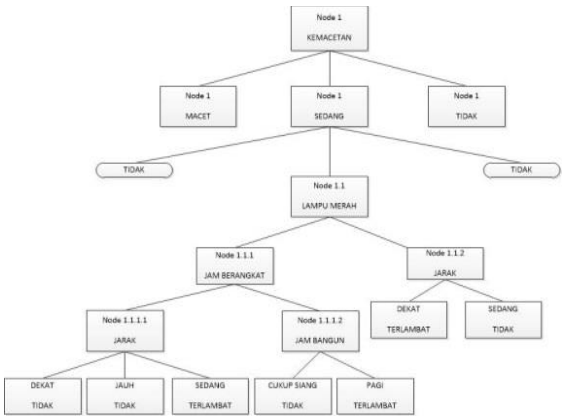


Figure 8 Results of the Decision Tree Manual

In Figure 8 above the results of the manual Decision Tree made by the author to prove the truth in this study. The first or top is Kemacetan (Congestion)

CONCLUSIONS AND SUGGESTIONS

Conclusion

The results of research using the C4.5 algorithm get an accuracy value of 60.74% and with an added Forward Selection an accuracy value of 65.93%. The "Kemacetan" attribute is the first root generated from the calculation and the Rapidminer and Forward selection software can also determine the relevant attributes namely the "Jam Berangkat", "Kendaraan" and "Kemacetan" attributes that can affect the accuracy value.

Suggestion

It is better if the SMP Negeri 271 schools provide counseling to parents about children's activities at home and monitor children's behavior. Further research should be able to use other methods in determining the delay factor so that it can produce better determination patterns and better accuracy results, and researchers can then create applications that can process the Classification algorithm.

REFERENCE

Akhmad Rizkon. 2019. "Pengaruh Metode Islah Mubasyir Terhadap Kedisiplinan Santri Pondok Pesantren Al-Basyariyah Kabupaten Bandung." *Jurnal Pendidikan Islam Indonesia* 4(1):23–29.

Amirulloh, Imam and Taufiqurrochman. 2017. "Komparasi Model Klasifikasi Algoritma Keterlambatan Siswa Masuk Sekolah." *Seminar Nasional Sains Dan Teknologi Fakultas Teknik Universitas Muhammadiyah Jakarta* (November):1–4.

Budiman, Irwan, Toni Prahasto, and Yuli Christyono. 2012. "Data Clustering Menggunakan Metodologi Crisp-DM Untuk Pengenalan Pola Proporsi Pelaksanaan Tridharma." Pp. 1–6 in *Seminar Nasional Aplikasi Teknologi Informasi (SNATI) 2012*. Yogyakarta: Universitas Islam Indonesia.

Insyiroh, Lailatul and Najlatun Naqiyah. 2017. "Studi Tentang Penanganan Siswa Yang Terlambat Tiba Di Sekolah Oleh Guru BK SMA Negeri 1 Gresik." *Jurnal BK UNESA*

7(1):1–8.

Nurhasan, Fuad, Noer Hikmah, and Dwi Yuni Utami. 2018. "Perbandingan Algoritma C4.5, KNN, Dan Naive Bayes Untuk Penentuan Model Klasifikasi Penanggung Jawab BSI Entrepreneur Center." *Jurnal Pilar Nusa Mandiri* 14(2):169–74.

Pramono, Fajar, Suwanda Aditya Saputra, Burhanuddin, and Kusuma Ade. 2018. "Komparasi Klasifikasi Penentuan Keterlambatan Siswa SMA Datang Upacara Menggunakan Algoritma C4.5." *Seminar Nasional Teknologi Informasi Dan Komunikasi* 2018(Sentika):80–86.

Saleh, Hamsir. 2017. "Prediksi Kebangkrutan Perusahaan Menggunakan Algoritma C4.5 Berbasis Forward Selection." *ILKOM Jurnal Ilmiah* 9(2):173–80.

Sulistiyono, Joko. 2018. "Peningkatan Kedisiplinan Masuk Sekolah Jam Pelajaran Pertama Melalui Konseling Kelompok Client Centered." *Jurnal Penelitian Pendidikan Indonesia* 3(2):1–8.