Indonesian Journal of Informatics Education

Research Paper ISSN: 2549-0389

Implementation of the C4.5 Algorithm to Predict Student Achievement at SMK Negeri 6 Surakarta

Giovanni Anggiesta Putri
Program Studi Pendidikan Teknik Informatika dan Komputer
Universitas Sebelas Maret
giovannianggiesta@student.uns.ac.id

Dwi Maryono
Program Studi Pendidikan Teknik Informatika dan Komputer
Universitas Sebelas Maret

Febri Liantoni
Program Studi Pendidikan Teknik Informatika dan Komputer
Universitas Sebelas Maret

Abstract:

Data mining is a knowledge used to get information from multiple data. C.45 Algorithm is one of data mining algorithm to classify data to many categories. Implementation of data mining not only could be used in industrial section but it could be used to in educational section (educational data mining) to help teacher and student improve their learning quality. This research purposed to know the implementation of data mining to predict student achievement from many factors could be. The research use Knowledge Discovery in Database method and it would be analyzed by accuracy calculated from classify model that be form. Result of the research is the rules that formed by the decision tree and it could predict student achievement . Teacher could use it to give special treatment to student who got not good prediction.

Keyword : Data Mining, C4.5 Algorithm, Educational Data Mining

Introduction

Sekolah Menengah Kejuruan or Vocational High School is a vocational education in Indonesia. Vocational education is purposed to fulfill people's need for labor. According to Pavlova (2009) education and training activities at vocational high school are a process to prepare students who have understanding, skills, development, behavior, attitudes, work habits and appreciation of work needed by the business community / industry that supervised by the government and society. or in contracts with institutions as well as on a productive basis(Ponto, 2016).

Learning achievement is a change that is accepted by a human when studying (Purwanto, 2016). Learning achievement includes thinking, knowledge, attention, understanding, conceptualization and some affective reasoning related to feelings, emotions, levels of acceptance or rejection of an object, and psychomotor skills related to involving limbs (Bakar, 2018). Student performance differs depending on individual student conditions. Various factors such as family conditions, school, interests, motivation, and past achievements determine the achievement of learning outcomes.

Data mining is data mining is the process of filtering data according to relevant business interests from very large data sets using different techniques and algorithms (Umadevi & Marseline, 2018). Currently, data mining is being used intensively in the world of education which is called Educational Data Mining. Educational data mining is used to make educational decisions based on data obtained to improve the quality of learning (Aldowah, Al-Samarraie, & Fauzy, 2019). In addition, EDM is also used to predict student learning outcomes and inform efficient programs for better learning (Angeli, Howard, Ma, Yang, & Kirschner, 2017).

One of the most widely used algorithm methods is the classification algorithm method. The classification algorithm is a data mining group that is used to group data into certain classes or target category variables (Kusrini & Luthfi, 2009) . Several types of classification algorithms include decision trees, Neural Networks, Naïve Bayes and K-Nearest Neighbor (Hussain, Dahan, Ba-Alwib, & Ribata, 2018). The decision tree is a very powerful and well-known method of classification and prediction algorithms. This method has a tree structure where each internal node represents each attribute in the test. Then each class label is represented by each leaf node (Sharma & Kumar, 2016). The C4.5 algorithm is one of the algorithms in data mining which is often used for classification by forming a decision tree. According to Takhur and Markandaiah, the C4.5 algorithm is a development of the ID3 algorithm developed by Quinlan in 1993. Compared to ID3, there are several improvements in its development (Wang, Zhou, & Xu, 2019).

This research aim to know about the implementation of C4.5 algorithm to predict student achievement in Vocational High School. This paper contains with research method, result, analysis and conclusion.

Research Method

Method

The research method used the Knowledge Discovery in Database (KDD). KDD is one of the well-known methods of finding useful knowledge from data using statistical methodologies and a logical system of multiple controls (Guruvayur & Suchithra, 2018). KDD and data mining are two different concepts but they are related to one another. According to Fayyad (1996) in Guruvayur (2018), the KDD process can be broadly explained into the process of data selection, data cleaning / pre-processing, transformation (data transformation), data mining, and evaluation.

Model Classification Analysis

The data analysis uses the C4.5 algorithm to find out the decision tree model and the accuracy test, recall test and precision test used to determine the accuracy of the decision tree model and the actual data. This test uses confusion matrix which is often used to test the performance of a classification data mining algorithm.

Based on the number of classes, this analysis method uses the binary class test which only consists of two categories, which are assumed to be P for positive classes and N for negative classes. This analysis model processes the classification model that is formed to predict the true positive class from an unknown number of samples (Tharwat, 2018).

a. Accuracy Test

Accuracy test is used to predict the ratio of true positive predictions compared to the overall positive predicted results.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN})$$

b. Precision Test

Precision test is the ratio of a positive true prediction to an overall positive predicted result.

$$\text{Precision} = (\text{TP}) / (\text{TP} + \text{FP})$$

c. Recall Test

Recall test is the ratio of true positive predictions to the overall true positive data.

$$\text{Recall} = (\text{TP}) / (\text{TP} + \text{FN})$$

Result and Analysis

Data Description

The data used in based on questionnaire from class XI Department of Software Engineering, Film and Television Production Program (PFPT), and Multimedia (MM) with a total of 100 people. Student data will be grouped into "Good" and "Not Good" categories..

The attributes that exist in the data obtained include the name of the student, student class, gender, year of birth, national exam scores for junior high school in Indonesian, national exam scores for junior high school subjects, national exam scores for SMP level in English subjects, National examination for science subjects for junior high school level, average score for junior high school level national exams, motivation to register, reasons for registering, choice of programs when entering SMK Negeri 6 Surakarta, parental support, residence status, distance household, first transportation to school, laptop facilities for studying at home, smartphone facilities for studying at home, adequate internet network at home, frequency of studying at home, frequency of playing games, frequency of watching TV / YouTube, participation in community organizations, participation in sports / community arts, father's education, mother's education, father's job, mother's work education, father's income, mother's income and rank. The rank attribute will become the category class. The attributes will be shown in Table 1.

Table 1. The Attributes

Attribute’s Name	Value
Department	Television and Movie Production
	Multimedia
	Software Engineering
Gender	Male
	Female
Birth year	2002
	2003
	2004
Motivation	Own Interest
	Recommendation from parents
	Other reasons
Reason to register	Interest with the department/major
	Recommendation from family
	Not interest with other department

Attribute's Name	Value
Department's Choice	First
	Second
	Third
Parents' support	Yes
	No
Living	With parents
	With other family
	Boarding house or dormitory
Distance school from home	Less than 2 Km
	2 – 6 Km
	More than 6 Km
Transportation	With parents
	Walk
	Motor cycle or by cycle
	Public Transportation
Breakfast before school	Yes
	No
Laptop at home	Yes
	No
Have smartphone	Yes
	No
Good internet connection at home	Yes
	No
Time for study	Less than 30 minutes
	30 minutes till 1 hour
	More than 1 hour
Time for gaming	Less than 30 minutes
	30 minutes till 1 hour
	More than 1 hour
Time for social media	Less than 1 hour
	1 till 2 hours
	More than 2 hours
Time for watching	Less than 1 hour
	1 till 2 hours
	More than 2 hours
Join some organization community	Yes

Attribute’s Name	Value
	No
Join some community outside school	Yes
	No
Frequency of other’s activity	Often (3-4 times a week)
	Rarely (1-2 times a week)
	Not joining
Father’s income	Rp. 0
	Less than Rp.1.000.000,00
	Rp. 1.000.000,00 – Rp. 1.999.999,00
	Rp. 2.000.000,00 – Rp. 4.999.9999
	Rp. 5.000.000,00 – Rp. 20.000.000,00
Mother’s income	Rp. 0
	Less than Rp.1.000.000,00
	Rp. 1.000.000,00 – Rp. 1.999.999,00
	Rp. 2.000.000,00 – Rp. 4.999.9999
	Rp. 5.000.000,00 – Rp. 20.000.000,00
Father’s education	Not known
	Primary school
	Junior High School
	Senior High School
	Vocational High School
	Diploma
	Bachelor
Mother’s education	Not known
	Primary school
	Junior High School
	Senior High School
	Vocational High School
	Diploma
	Bachelor
Father’s work	Not Working
	Government officer /Soldier/Police
	Entrepreneur
	Farmer
	Employee
	Others

Attribute’s Name	Value
	Pass away
Mother’s work	Not Working
	Government officer /Soldier/Police
	Entrepreneur
	Farmer
	Employee
	Others
	Pass away

Data Pre-processing

1. Data Cleaning

The data cleaning step is done by cleaning the data from duplication of data and empty data so that the data is not used as training data. The results of checking blank data that are found are 16 cases of blank data on Father's Work and 19 cases of blank data on Father's income.

2. Data Transformation

This step uses the concept hierarchy to simplify the attribute data of the national exam scores, the average national exam scores, father's education, mother's education, father's income, and mother's income. National examination score data for each subject and the average are filled in with a number format so that it needs to be categorized so as not to form a regression.

The process of simplifying the data using the concept hierarchy is done by sorting the data from the smallest value to the largest value. Conversion table of father's income and mother's income is in table 2 and the conversion table of father and mother's education is in table 3. The conversion table for the national exam subject scores is in table 4.3 and the conversion table for the average national exam is in table 4.

Table 2 Conversion Table of Income Attribute Value

Original Value	Conversion Value
Rp. 0	Low
Less than Rp.1.000.000,00	Low
Rp. 1.000.000,00 – Rp. 1.999.999,00	Low
Rp. 2.000.000,00 – Rp. 4.999.9999	Average
Rp. 5.000.000,00 – Rp. 20.000.000,00	High

Table 3 Conversion Table of Educational Value

Original Value	Conversion Value
Not known	Not known
Primary school	Primary-Junior High School
Junior High School	Primary-Junior High School
Senior High School	Senior/Vocational High School
Vocational High School	Senior/Vocational High School
Diploma	Diploma/Bachelor
Bachelor	Diploma/Bachelor

Table 4 Conversion Table of National Exam Value

Original Value	Conversion Value
81-100	Really Good
61-80	Good
41-60	Enough
<41	Low

Table 5 Conversion Table of Average National Exam Score

Original Value	Conversion Value
81-100	Good
51-80	Enough
<51	Low

Classification and Analysis

Classification model was formed based on the different parameters in Weka. The difference between the parameters that is done is by changing the split, unpruned to true or false and changing the minNumObj. The results of data testing were formed 4 classification models which were differentiated based on different parameters.

1. Classification model with pruned parameter, minNumbObj = 2 and split 50%

   Testing the classification model with parameters pruned, split 50% and minNumbObj = 2 has a high accuracy of 88.2%. The decision tree that is formed but only has one leaf node which makes the classification rules that should be formed unknown. The results of the accuracy calculation matrix will be shown in table 6.

Table 6. Results of Classification Model Data 1

		Prediction Class		Total
		Good	Not Good	
Actual Class	Good	29	0	29
	Not Good	5	0	5
	Total	34	0	34

Based on table 4.5, 29 data are classified as true positive, in order to obtain data calculation analysis:

a. Accuracy Test

$$Accuracy = \frac{29}{29+5} \times 100\ \% = \frac{29}{34} \times 100\ \% = 88.2\ \%$$

b. Precision Test

$$Precision = \frac{29}{29+5} = \frac{29}{34} = 0.882$$

c. Recall Test

$$Recall = \frac{29}{29+5} = \frac{29}{34} = 0.882$$

However, the decision tree is formed only has 1 node because the decision tree is trimmed to maximize accuracy. This causes no rules to be formed in this classification model, so that this decision tree is not ideal to use even though the results of the accuracy test show high results..

2. Classification model with unpruned parameter, minNumbObj = 2 and split 50%

This decision tree model has 82% accuracy test results with unpruned parameters and%, and minNumObj = 2 with a 50% split dataset. The matrix formed for the calculation of the test for accuracy, sensitivity and precision is shown in table 7.

Table 7. Results of Classification Model Data 2

		Prediction Class		Total
		Good	Not Good	
Actual Class	Good	37	6	43
	Not Good	3	4	7
	Total	40	10	50

Based on table 7, 37 data were obtained with true good classification and 4 correct data were not good), so that the data calculation analysis was obtained:

a. Accuracy Test

$$Accuracy = \frac{37+4}{37+4+6+3} \times 100\,\% = \frac{41}{50} \times 100\,\% = 82\,\%$$

b. Precision Test

$$Precision = \frac{37}{37+3} = \frac{37}{40} = 0.925$$

c. Recall

$$Recall = \frac{37}{37+6} = \frac{37}{43} = 0.86$$

The rules formed from this classification model are :

a. If students do not have parental support, learning achievement is not good.
b. If students do not have parental support, the frequency of playing games is less than 30 minutes and does not participate in the community, the learning achievement is good.
c. If students have parental support, the frequency of playing games is more than 1 hour and the learning frequency is less than 30 minutes, the learning achievement is not good.
d. If students have parental support, the frequency of playing games is more than 1 hour, the learning frequency is more than 1 hour, the transportation used is using public transportation, the learning achievement is not good.
e. If students have parental support, the frequency of playing games is 30 minutes to 1 hour, the signal at home is adequate, and the father's job falls into other categories, then learning achievement is good.
f. If students have parental support, the frequency of playing games is more than 1 hour, and the learning frequency is 30 minutes to 1 hour, the learning achievement is good.

3. Classification model with pruned parameter, minNumbObj = 4 and split 70%

The test results that have pruned parameters, 70% split dataset, and minNumObj = 4 get an accuracy of 90%. The matrix formed for the calculation of the test for accuracy, sensitivity and precision is shown in table 8.

Table 8. Results of Classification Model Data 3.

		Prediction Class		Total
		Good	Not Good	
Actual Class	Good	27	0	27
	Not Good	3	0	3
	Total	30	0	30

Based on table 4.11, 27 data with true positive classification were obtained, so that the data calculation analysis was obtained:

a. Accuracy Test

$$Accuracy = \frac{27}{27+3} \times 100\ \% = \frac{27}{30} \times 100\ \% = 90\ \%$$

b. Precision Test

$$Precision = \frac{27}{27+3} = \frac{27}{30} = 0.90$$

c. Recall

$$Recall = \frac{27}{27+0} = \frac{27}{27} = 1.000$$

Just like the first decision tree classification model, the resulting decision tree only has 1 node because the decision tree is trimmed to maximize accuracy. This causes no rules or rules to be formed in this classification model, so that this decision tree is less than ideal to use even though the results of the accuracy test show high results..

4. Classification model with unpruned parameter, minNumbObj = 4 and split 70%

This decision tree model has unpruned parameters, the split dataset is 70 %%, and minNumObj = 4 has an accuracy of 86.67%. The matrix formed for the calculation of the accuracy, sensitivity and precision test is shown in Table 4.8.

Table 9 Results of Classification Model Data 4.

		Prediction Class		Total
		Good	**Not Good**	
Actual Class	**Good**	26	1	26
	Not Good	3	0	4
	Total	29	1	30

Based on table 4:12, 27 data are obtained with true positive classification, in order to obtain data calculation analysis:

a. Accuracy Test

$$Accuracy = \frac{27}{27+3} \times 100\ \% = \frac{27}{30} \times 100\ \% = 90\ \%$$

b. Precision Test

$$Precision = \frac{27}{27+3} = \frac{27}{30} = 0.90$$

c. Recall Test

$$Recall = \frac{27}{27+0} = \frac{27}{27} = 1.000$$

The rules formed from this classification model are :

a. If the student has a male gender, the learning frequency is less than 30 minutes, then the learning achievement is good.

b. If the student has a male gender, the learning frequency is 30 minutes to 1 hour and the value of Indonesian is in the sufficient category, the learning achievement is not good.

c. If the student has a female gender, born in 2004 or 2002, the learning achievement is good.

d. If the student is female, born in 2003, often has breakfast before leaving for school, and has a low science score, the learning achievement is not good.

Based on the four decision tree classification models that have been analysed, 3 of the 4 classification models have high accuracy test results but have true negative results which are 0. The 1st classification model with 82% accuracy, the 3rd test result classification model with accuracy 90%, and the fourth test result classification model with an accuracy of 86.67%. Classification models that have a true negative value of 0 are less than ideal to be used as a classification model because the results of good categories are fake positive such as those in the 1st classification model matrix and the 3rd classification model.

In addition, the 1st classification model and the 3rd classification model only have 1 node because the pruned parameters are executed which causes these 2 classification models to have no rules or rules that are formed to classify data into good and not good categories.

For the 4th classification model besides having a true negative result with a value of 0 and having a fake positive result of 3, this classification model also has a fake positive value of 1.Fake positive means that data is assumed to be in the positive category which is actually negative. The absence of data with true negative values makes this classification model doubtful as an ideal classification model even though there are rules or regulations that are formed.

So the second classification model which has a 50% split dataset parameter and an unpruned tree is more ideal for use in decision making. This classification model has a true positive value of 37 and a true negative of 4. Even though it has a fake positive = 3 and a fake negative = 6.

For the rules that are formed, the attribute "Parental Support" has the greatest influence and dominance on decision making in the formed decision model. It can be seen that students who do not have the support of their parents have poor learning achievement. "The frequency of playing games" is the next attribute that has an influence on the decision making of students with good or poor achievement. After that, the attributes of "Signal", "Community" and "Learning Frequency" become the next determinants. Students who play the game for 30 minutes to 1 hour but the signal available at their home is not good have a categorized good learning achievement. However, if it has a good signal, the "Father's Job" and "Major" attributes will determine the next category. For those who have a "Father's Job" in addition to the "Entrepreneurial" category, it will be in the good category, but for those whose father has a job as an entrepreneur, the attribute "Department" will determine the student's learning achievement..

Students who have “Game Play Frequency” less than 30 minutes and participate in certain communities but have never participated in activities outside of school have not good learning achievement. Students who play games for more than 1 hour but have a learning frequency of less than 30 minutes have not good learning achievement the same as students who have a frequency of more than 1 hour but depart by public transportation also have not good performance.

Conclusion

Based on the research results, the c4.5 algorithm, which is a data mining algorithm for classification, shows that this algorithm has a good performance in making a classification model that is used to predict student learning achievement. The accuracy obtained from this study shows that one of the resulting classification models has a result of 82% and forms rules that can be used in predicting student achievement.

References

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics*. https://doi.org/10.1016/j.tele.2019.01.007

Angeli, C., Howard, S. K., Ma, J., Yang, J., & Kirschner, P. A. (2017). Data mining in educational technology classroom research: Can it make a contribution? *Computers and Education*. https://doi.org/10.1016/j.compedu.2017.05.021

Bakar, R. (2018). The influence of professional teachers on Padang vocational school students' achievement. *Kasetsart Journal of Social Sciences*. https://doi.org/10.1016/j.kjss.2017.12.017

Guruvayur, S. R., & Suchithra, R. (2018). A detailed study on machine learning techniques for data mining. *Proceedings - International Conference on Trends in Electronics and Informatics, ICEI 2017*, *2018-Janua*, 1187–1192. https://doi.org/10.1109/ICOEI.2017.8300900

Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science*, *9*(2), 447–459. https://doi.org/10.11591/ijeecs.v9.i2.pp447-459

Kusrini, & Luthfi, E. T. (2009). *Algoritma Data Mining*. (T. A. Prabawati, Ed.). Yogyakarta: ANDI.

Ponto, H. (2016). *Evaluasi Pembelajaran Pendidikan Kejuruan*. Yogyakarta: Deepublish.

Purwanto. (2016). *Evaluasi Hasil Belajar* (VII). Yogyakarta: Pustaka Pelajar.

Tharwat, A. (2018). Classification assessment methods. *Applied Computing and Informatics*. https://doi.org/10.1016/j.aci.2018.08.003

Umadevi, S., & Marseline, K. S. J. (2018). A survey on data mining classification algorithms. In *Proceedings of IEEE International Conference on Signal Processing and Communication, ICSPC 2017* (Vol. 2018-Janua, pp. 264–268). https://doi.org/10.1109/CSPC.2017.8305851

Wang, X., Zhou, C., & Xu, X. (2019). Application of C4.5 decision tree for scholarship evaluations. *Procedia Computer Science*. https://doi.org/10.1016/j.procs.2019.04.027