Student's Skills Competency Test Prediction Using C4.5 Algorithm

Ultach Enri [1], Jajam Haerul Jaman [2], Muhammad Rizky Ananda [3]
{ultach@staff.unsika.ac.id [1], jajam.haeruljaman@staff.unsika.ac.id[2],
1441177004221@student.unsika.ac.id [3]}

Informatics Engineering Program Faculty of Computer Science Universitas Singaperbangsa Karawang[1,2,3]

Abstract. Competency Skills (UKK) is part of government intervention in ensuring the quality of education in the Secondary Vocational School aims to measure the achievement of a certain level of competency in students' appropriate competency skills. The purpose of the research is to find out how competent the students' in their vocation as well as new strategies for educators in providing more effective learning by using C4.5 algorithm combining with feature selection. Overall the results of the validation of the model experiments that have the best influence are supplied with a set of test accuracy values of 96.875%, and 13 optimal attributes are selected. Therefore, in this study, the C4.5 algorithm with feature selection can provide good and effective results.

Keywords: C4.5 Algorithm, Data Mining, Feature Selection.

1 Introduction

Indonesia's competitiveness in facing competition and free trades among countries is determined by the outcome of human resources development. One of the state’s efforts in fulfillment of high-level human resources is the vocational education training. According to the law number 20 the year 2003 article 15, vocational education is a secondary education that prepares students primarily to work in a specific field. Uji Kompetensi Keahlian (UKK) or Skills Competency Test is an annual activity conducted by every vocational high school and it is government intervention in ensuring the quality of education. The implementation of UKK aims to measure the achievement of students‘ competence at certain levels according to the competency of expertise taken during the learning period at the vocation school. The UKK consists of practice exams which are generally held before the implementation of the national exam and theory exam which is part of the national exam implementation series [6]. The expertise competency test is conducted to measure how competent the students are to implement their skills during the learning process based on their major.

The decision tree is a common method used for the classification of data mining. This method is popular because it can classify and demonstrate relationships between attributes. Many algorithms can be used to construct a decision tree, one of them is C4.5 algorithm [10], which is part of the classification algorithm in machine learning and data mining. In making the decision tree, each algorithm implements the size of the selection of different attributes. The size of the attribute selection is the size used in determining the best criteria for grouping tuples [4].

MSCEIS 2019, October 12, Bandung, Indonesia

DOI 10.4108/eai.12-10-2019.2296548

Features selection is one of the commonly used terms in data mining to reduce input according to size managed on preprocessing and analysis [7]. The features selection technique aims to reduce irrelevant features and feature dimensions on the data [8]. the features used in the decision tree are the features that are considered relevant in determining the target class of a data object [1]. Features selection can make good classifications more efficient and effective by reducing the amount of data analyzed, identifying features appropriate for consideration in the learning process [11], one method for features selection is CorrelationAttributeEval which can evaluate the features that are highly correlated with the class target, but not correlated to each other[12].

In previous research conducted by David Hartanto Kamagi and Seng Hansun [5] the implementation of data mining with C4.5 Algorithm to predict student graduation rate can be concluded that the C4.5 algorithm can be implemented to predict the graduation rate of students with four categories which are pass fast, pass right, pass late and drop out. The most influential attribute as the result is the grade in the sixth semester and the model has good performance by gaining an accurate rate of 87.5%.

One of the vocational schools which holding the UKK annually is SMK TI Muhammadiyah Cikampek. Student's UKK graduation level in computer and network engineering in SMK TI Muhammadiyah Cikampek in three years can be seen in Table 1.

Table 1. Computer and Network Engineering Competency Test Graduation data.

No.	Year	(%)	
		Passed UKK	Did not pass UKK
1	2016	37%	63%
2	2017	76%	24%
3	2018	74%	26%

Based on Table 1. there were only 37% of students who passed the UKK in 2016, and there was a big jump in the next year, where there was 76% of students passed the UKK and decreased 2% in 2018. Despite a significant increased in the year 2017, but the percentage for students who did not pass the UKK is not a little.

In this research students, skills competency test data from the department of computer and network engineering will be used, by using the selection feature that is CorrelationAttributeEval that serves to find the attributes that have a relationship with the class target, to produce a decision tree with optimal attributes. This research aims to predict the student competency skills test majoring in computer and network engineering using C4.5 algorithm, so we can find out how competent the computer and network engineering students can pass in competency skills test and it can be used as a benchmark for educators as well as a new strategy for educators in delivering more effective learning and can reduce the percentage of undergraduate of students.

2 Methods

2.1 Features Selection

Several feature selection methods have been introduced in the machine learning domain. The main aim of these techniques is to remove irrelevant or redundant features from the dataset[8]. The features selection is one of the commonly used terms in data mining. It used to reduce inputs according to size managed on preprocessing and analysis [7]. Its technique is performed to reduce irrelevant features and reduce the feature dimensions on the data [9]. on the decision tree, the feature selection algorithm used for decision tree construction determines the level of accuracy of the decision tree. The features used in the decision tree are the features that are considered relevant in determining the class target of a data object [1]. Features selection can make the classifications both more efficient and effective by reducing the amount of data analyzed, or identifying the appropriate features for consideration in the learning process [11]. One of the selections of features in WEKA is the CorrelationAttributeEval.

2.2 CorrelationAttributeEval

CorrelationAttributeEval evaluates the features that are highly correlated with the class target but are not correlated to each other. Evaluate attribute values by measuring the correlation(Pearson) between attributes and classes. Nominal attributes are considered based on values with basic values and treat each value as an indicator. Overall correlation for nominal attributes through average weight [12]. Calculated by the Pearson correlation formulation, the following formula from Pearson correlation :

$$r = \frac{n\, \Sigma\, x.y - \Sigma\, x\, \Sigma\, y}{\sqrt{(n\, \Sigma\, x^2 - (\Sigma\, x)^2)(n\, \Sigma\, y^2 - (\Sigma\, y)^2)}} \tag{1}$$

Description :
r = Correlation Coefficient
n = Amount of data
x = Attribute
y = Class Target

2.3 Research Methodologies

In this research in predicting the students' competency skills test majoring in computer and network engineering using C4.5 algorithm using CRISP-DM *(Cross-Industry Standart Process for Data Mining)* methodology as can be seen in **Figure 1**.

In this research, students' competency skills test data majoring in computer and network engineering data will be used using C4.5 algorithm with 16 attributes i.e. safety tools, tools and materials, wiring, installation of network devices, cable network configuration, wireless network configuration, server router configuration, client network configuration, firewall configuration, cable network testing, wireless network testing, tool usage, work safety, working time and results at the class target. After that, both datasets were tested with implementing the C4.5 classification algorithm and then the algorithm model is validated using two ways of test options supported by WEKA i.e. use training set and supplied test set. In use training method WEKA will use the training data as data testing, and in a supplied test requires users to have separate training and testing data as can be seen in **Figure 2**.

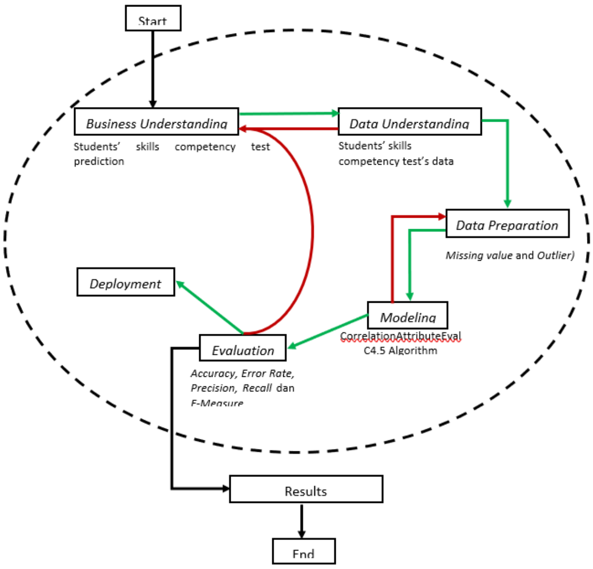
Start
Business Understanding
Students' skills competency test prediction
Data Understanding
Students' skills competency test's data
Data Preparation
Missing value and Outlier)
Deployment
Modeling
CorrelationAttributeEval
C4.5 Algorithm
Evaluation
Accuracy, Error Rate, Precision, Recall dan F-Measure
Results
End

Fig. 1. Research methodology.

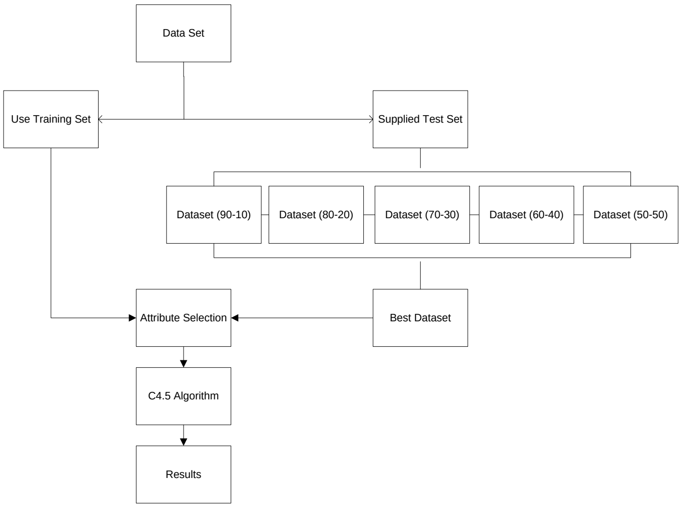

Fig. 2. Research scenario.

2.4 Evaluation and Validation

In a classification it is important to specify the costs associated with correct or incorrect classification, by doing that it can be a valuable when the cost of different misclassification varies significantly [2]. Using training data to derive a classifier and then to estimate the accuracy of the resulting learned model can result in misleading overoptimistic estimates due to overspecialization of the learning algorithm to the data. The confusion matrix is a useful tool for analyzing how well the classifier can recognize tuples of different classes [3].

3 Results and Discussion

3.1 Use Training Set

Results made using the use training set validation model Use Training Set with overall 511 data by C4.5 algorithm by using features selection can be seen in the following table:

Table 2. Use Training Set result.

Numbers of Attributes	Accuracy	Precision	Recall	F-Measure
15	**95,8904%**	0,9497	0,9541	0,952
14	95,1076%	0,951	0,951	0,951
13	95,1076%	0,951	0,951	0,951
12	93,7378%	0,938	0,937	0,937
11	94,3249%	0,943	0,943	0,943
10	93,7378%	0,938	0,937	0,937
9	93,7378%	0,938	0,937	0,937
8	93,7378%	0,938	0,937	0,937
7	94,3249%	0,944	0,943	0,943
6	94,3249%	0,944	0,943	0,943
5	93,7378%	0,938	0,937	0,937
4	93,7378%	0,938	0,937	0,937
3	90,411%	0,907	0,904	0,903
2	86,6928%	0,879	0,867	0,868
1	86,6928%	0,879	0,867	0,868

The overall dataset of the data validation model used is the use training set and feature selection has done to find the optimal attribute using algorithm has been modeled that is C4.5 and the best accuracy value is 95.8904% with 15 attributes.

3.2 Supplied Test Set

The result of the supplied test by using data that has been divided generates a variety of accuracy value-based in the distribution of processed datasets. Accuracy value obtained as shown in Table 3. The best accuracy value will be selected for the process of feature selection.

Table 3. The comparison result of dataset distribution supplied test set.

Data Training	*Data Testing*	*Accuracy*	*Error Rate*	*Precision*	*Recall*	*F-Measure*
460	51	94.1176%	5.8824%	0,941	0,941	0,941
409	102	95.098%	4.902%	0.951	0.951	0.951
358	153	94.7712%	5.2288%	0.948	0.948	0.948
307	204	95.5882%	4.4118%	0.956	0.956	0.956
255	256	**96.875%**	3.125%	0.969	0.969	0.969

Based on Table 3, the datasets distribution is 90%-10%, 80%-20%, 70%-30%, 60%-40% dan 50%-50%. The comparison has been done to find the best accuracy value after the feature selection has been done when the best accuracy of the five data distribution mode is selected the highest accuracy value is the amount of 96.875% in the 50-50 datasets distribution. Features selection with C4.5 algorithm is performed with the supplied test set model validation so that the following values are obtaine.

Table 4. shows the accuracy value obtained after performing features selection in each supplied test set with 50-50 datasets, where there are three highest accuracy values of both 15 attributes, 14 attributes and 13 attributes with each having a value that is 96.875%. Of these three attributes, values are selected one of the optimal attributes with 13 attributes, according to the function of the feature selection is to reduce the input according to the size that will be managed on processing and analysis. The decision tree is shown in **Figure 3**.

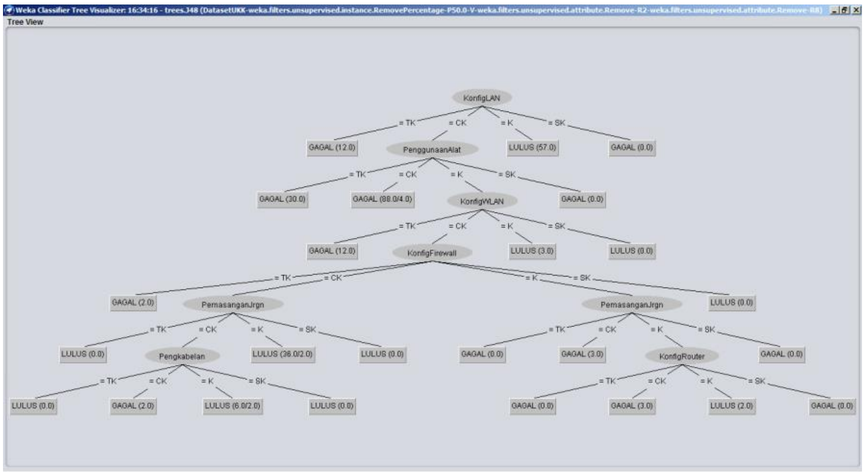

Fig. 3. Decision Tree.

Table 4. Supplied Test Set Dataset 50% result.

Numbers of Attributes	Accuracy	Precision	Recall	F-Measure
15	96,875%	0,969	0,969	0,969
14	96,875%	0,969	0,969	0,969
13	96,875%	0,969	0,969	0,969
12	94,9219%	0,949	0,949	0,949
11	94,9219%	0,949	0,949	0,949
10	94,9219%	0,949	0,949	0,949
9	94,9219%	0,949	0,949	0,949
8	94,9219%	0,949	0,949	0,949
7	94,9219%	0,949	0,949	0,949
6	94,9219%	0,949	0,949	0,949
5	94,1406%	0,941	0,941	0,941
4	94,1406%	0,941	0,941	0,941
3	89,0625%	0,893	0,891	0,889
2	87,1094%	0,883	0,871	0,872
1	87,1094%	0,883	0,871	0,872

4 Conclusion

Overall a comparison of the validation model that has the best influence is the supplies test with the datasets 50% with the accuracy value 96,875% as the result, the *precision* of 0.969, *recall* of 0,969 and *f-measure* of 0,969 and the optimal attributes chosen is as many as 13 attributes. Where appropriate with the utilization of the attribute selection is to reduce the input size to be managed on processing and analysis. Therefore, in this research C4.5 algorithm by using feature selection can provide a good and effective model.

References

[1] Delki, A.: Perbandingan Algoritma Feature Selection Information Gain dan Symmetrical Uncertainty Pada Data Ketahanan Pangan. Skripsi, Institut Pertanian Bogor. (2013)

[2] Gorunescu, F.: Data Mining, Concepts, Models, and Techniques, Springer. (2011)

[3] Han, J & Kamber, M.: Data Mining Concepts and Techniques, Elsevier. (2006)

[4] Iriadi, N., & Nuraeni, N.: Kajian penerapan metode klasifikasi data mining algoritma C4.5 untuk prediksi kelayakan kredit pada Bank Mayapada Jakarta. Jurnal Teknik Komputer BsI, 1-6 (2016)

[5] Kamagi, D. H., & Hansun, S.: Implementasi data mining dengan algortima C4.5 untuk memprediksi tingkat kelulusan mahasiswa. ULTIMATICS, 1-6. (2014)

[6] KEMENDIKBUD_RI.: Konsep Pembelajaran di Sekolah Menengah Kejuruan | Direktorat Pembinaan SMK. Diambil kembali dari Kementerian Pendidikan dan Kebudayaan RI: https://psmk.kemdikbud.go.id/konten/1869/konsep-pembelajaran-di-sekolah-menengah-kejuruan (2016)

[7] Mongkareng, D., Setiawan, N. A., & Permanasari, A. E.: Implementasi data mining dengan seleksi fitur untuk klasifikasi serangan pada intrusion detection system (IDS). CITEE, 1-8. (2017)

[8] Pinar Yildirim,: Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease. Interntional Journal of Machine Learning and Computing, Vol. 5, No. 4. (2015)

[9] Sari, B. N.: Implementasi teknik seleksi fitur information gain pada algoritma klasifikasi machine learning untuk prediksi performa akademik siswa. Seminar Nasional Teknologi Informasi dan Multimedia, 1-7. (2016)

[10] Supriyanti, W., Kusrini, & Amborowati, A.: Perbandingan kinerja algortima C4.5 dan Naive Bayes untuk ketepatan pemilihan konsentrasi mahasiswa. Jurnal INFORMA Politeknik Indonusa Surakarta, 1-7. (2016)

[11] Utami, L. D., & Wahono, R. S.: Integrasi metode information gain untuk seleksi fitur dan adaboost untuk mengurangi bias pada analisis sentimen review restoran menggunakan algoritma naive bayes. Journal of Intelligent Systems, 1-7. (2015)

[12] Vika, S.: Perbandingan Algoritma Naive Bayes dan K-Nearest Neighbor Untuk Memprediksi Luas Lahan Panen Tanaman Padi Di Kabupaten Karawang. Karawang: Universitas Singaperbangsa Karawang (2017)