On-Time Graduation Prediction System Using Data Mining Classification Method

Suwitno[1], Arief Wibowo[2]
{suwit.ang1305@gmail.com[1], arief.wibowo@budiluhur.ac.id[2]}

Department of Information System, Faculty of Science and Technology, Universitas Buddhi Dharma, Indonesia[1], Department of Information System , Faculty of Information Technology, Universitas Budi Luhur, Indonesia[2]

Abstract. The collection of data on academic information system database of Higher Education is often not utilized maximally, whereas from data with data mining technique can give knowledge which not yet known before. The purpose of this research is to know how to form the prediction model of student's graduation rate on- time at Buddhi Dharma University of Tangerang through student passing data. Prediction of student graduation on-time using comparison of algorithm C4.5 and K-NN done with data selection stage, data transformation, data mining and interpretation. This study uses 300 training data and 90 data testing. Then the process of classification technique using decision tree method using C4.5 algorithm and Euclidean distance calculation using K-NN algorithm. Evaluation of classification performance is done to know how well the accuracy of a model is formed. Based on the research that has been done, the model is formed with the help of Rapid miner software, and calculated average value of k-fold cross validation on testing up to k = 10 for algorithm C4.5 and K-NN. Testing is done with Confusion Matrix and ROC curves. Accuracy results obtained prove that Algorithm C4.5 yields 90% accuracy percentage and K-NN yield 87% accuracy percentage. Thus the C4.5 algorithm has a higher accuracy value than K-NN. This C4.5 algorithm can be used as prototype predictions of students' graduation on-time at Buddhi Dharma University Tangerang.

Keywords: C4.5, K-NN, Data Mining, K-Fold Cross Validation, Confusion Matrix.

1 Introduction

The evolution of information technology is so advanced today, causing the level of accuracy of a data is needed in life. Any information that exists becomes an important thing to determine every decision in a particular situation. This leads to the provision of information into a means to be analyzed and summarized into a knowledge of useful data when making a decision.

Higher education institution is a college which is the organizer of academic education for students. Universities are required to provide quality education for students to produce intelligent, ethical, creative and competitive human resources. In the education system, the student is an important asset for an educational institution and for that matter the student's

WMA-1 2018, January 20-21, Indonesia

DOI 10.4108/eai.20-1-2018.2281900

graduation rate is on time. The percentage rise and fall of the students' ability to complete on-time graduation is one of the elements of university accreditation assessment.

1.1 Data Mining

Data mining is an activity of extraction to obtain important information that is implicit and previously unknown, from a data. Data mining is defined as the process of finding patterns in the data. This process is automatic or (usually) semi-automatic.

1.2 C4.5 Algorithm

The C4.5 algorithm is designed by J. Ross Quinlan, named C4.5 because it is descended from the ID3 approach to construct decision trees. C4.5 is a suitable algorithm used for classification problems in machine learning and data mining. C4.5 maps the attributes of the classes so they can be used to find predictions for data that have not yet appeared. In the decision tree the central node is an attribute of the tested data (tuple), the branch is the result of the attribute test, and the leaf is the class formed. The stages in the C4.5 algorithm framework are:

a. Note the label on the data, if it is all the same, then the leaves will be formed with the value of the entire data label.

b. Calculating the total value of the information (Entropy)

$$\mathrm{Entropy} = -\sum_{i=1}^{m} \mathrm{p_i} \log_2(\mathrm{p_i}) \qquad (1)$$

c. Calculate the info value of each attribute (Info)

$$\mathrm{Info_A(D)} = \sum_{j=1}^{v} \frac{|\mathrm{D_j}|}{|\mathrm{D}|} \times \mathrm{Info}(\mathrm{D_j}) \qquad (2)$$

d. Calculating the gain value of each attribute (Gain)

$$\mathrm{Gain(A)} = \mathrm{Entropy} - \mathrm{Info_A(D)} \qquad (3)$$

e. After the decision tree branch is formed, the calculation is performed again as in steps a through d. However, if the branch has reached the maximum allowable branches, the leaf will be formed with the majority value of the data value.

1.3 K-NN Algorithm

The K-Nearest Neighbor (K-NN) algorithm is a method for classifying objects based on learning data closest to the object. The K-NN algorithm uses a supervised algorithm. The difference between supervised learning and unsupervised learning is in supervised learning aims to find new patterns in data by linking existing data patterns with new data. Whereas in unsupervised learning, data does not have any pattern, and the purpose of unsupervised learning to find patterns in a data.

Nearest Neighbor is an approach to calculate the proximity between a new case and an old case, based on matching of a number of features. To define the distance between two points ie the point on the training data (x) and the point in the data testing (y) then used the Euclidean formula, with the equation:

$$D\ (x, y) = \sqrt{\sum_{i=1}^{n} \left(x_i - y_j\right)^2} \qquad (4)$$

Stages in the K-NN algorithm are:

a. Specifies the parameter k (the number of nearest neighbors).
b. Calculates the square of the Euclidean distance (query instance) of each object against the given training data.
c. Then sort those objects into groups that have the smallest Euclidean distance.
d. Gathering a new category k (classification of Nearest Neighbor)
e. By using the most Nearest Neighbor categories, the predicted value of the counted query instance can be predicted.

1.4 K-Fold Cross Validation

In this research, the method used to test the classification pattern is by k-fold cross validation method. In the k-fold cross validation data is divided into k section, D_1, D_2 ... D_k, and each D has the same amount of data. Testing with k = 5 or k = 10 can be used to estimate the error rate, because the training data on each fold is quite different from the original training data. Calculating the value of its accuracy can be done using equations:

$$Accuracy \ = \frac{Amount\ of\ right\ class}{Amount\ of\ data}\ x\ 100\ \% \qquad (5)$$

2 Study Overview

Table 1. Comparison of similar studies.

No	Research	Algorithm	Result
1.	Model Algoritma K-Nearest Neighbor (K-NN) Untuk Prediksi Kelulusan [1]	K-Nearest Neighbor (K-NN)	The highest accuracy value in the data cluster k = 5 is 85.15% and the AUC value is 0.888.
2.	Implementasi Teknik Data Mining Untuk Memprediksi Tingkat Kelulusan Mahasiswa Pada Universitas Bina Darma Palembang [2]	Decision Tree J48 (C4.5)	By using 3-fold cross validation test option. The average classification of the J48 algorithm reaches 90% accuracy.
3.	Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa [3]	Decision Tree J48 (C4.5)	Predicted accuracy of 87.5% of 60 training data and 40 data testing
4.	Perancangan Aplikasi Prediksi Kelulusan Tepat Waktu Bagi Mahasiswa Baru Dengan Teknik Data Mining [4]	K-Nearest Neighbor (K-NN)	Predicted accuracy of 83.36%
5.	Model Prediksi Tingkat Kelulusan Mahasiswa Dengan Teknik Data Mining Menggunakan Metode Decision Tree C4.5 [5]	Decision Tree C4.5	Testing resulted in accuracy of 61.22%.
6.	Penerapan K-Optimal Pada Algoritma K-NN untuk Prediksi Kelulusan Tepat Waktu Mahasiswa Program Studi Ilmu	K-Nearest Neighbor (K-NN) with k-fold	Results obtained from k = 5 with an accuracy of 80% applied as k-optimal

	Komputer Fmipa Unlam Berdasarkan IP Sampai Dengan Semester 4 [6]	cross validation	
7.	Identifikasi Lama Studi Berdasarkan Karakteristik Mahasiswa Menggunakan Algoritma C4.5 [7]	Decision Tree C4.5	Accuracy of classification reached 90%, test results totaling 40 test samples showed an accuracy of 67.5%
8.	Implemenasi Algoritma K-Nearest Neighbor Untuk Prediksi Waktu Kelulusan Mahasiswa [8]	K-Nearest Neighbor	Accuracy of 90%

3 On-Time Student Graduation Decision

Based on the number of students who graduated inappropriately in time at Buddhi Dharma University related to academic report of year 2016 about 28,36% increase 5,2% from previous year, hence existence of prototype prediction information system this is mandatory requirement and must be implemented for management of the university to conduct evaluation and monitoring every semester. Based on the policy of the Vice Rector I that the on-time graduation must be based on the academic guidebook where for the stratum level one is a maximum of 8 semesters or 4 years. The hypothesis of this research are:

a. Suspected algorithm C4.5 and K-NN is an algorithm that can be used to predict on-time graduation of students according to the availability of data at Buddhi Dharma University.

b. It is suspected that the most accurate classification algorithm for predicting student graduation on-time is the C4.5 algorithm.

4 Design of Predicting Information System

4.1 Research Methodology

In this research will use dataset which is used as training data and data testing as much as 390. Training data will be used for formation of algorithm pattern C4.5 and K-NN. Training data will be used for the formation of algorithm pattern C4.5 and K-NN. Then the data testing is used to test the algorithm pattern that has been formed. In algorithm C4.5, the process determines the category on-time or not a student graduated by using the steps that have been discussed in the stages of algorithm C4.5 where sought total entropy overall training data.

Next process are calculate the gain attribute calculation to determine the node of the tree and subsequent branch, the rule that has been formed will be tested into the data testing. While in the K-NN algorithm, the process of determining the category on-time or whether a student graduated by calculating proximity distance with the entire training data using euclidean distance equation, the smallest value of the overall calculation of training data is a category defined as a prediction class of data testing. In the development will be made a prototype web-based information system with PHP and MySQL. The result of this research is the implementation of C4.5 algorithm into the prediction information system on-time prediction of students' graduation at Buddhi Dharma University.

4.2 Sample Selection Method

The data used in this study is secondary data, the data used is taken based on the data of graduation of Buddhi Dharma University students 2010, 2011 and 2012 with the stratum level 1 of the entire study program.

4.3 Research Steps

The steps used in this study adopted several steps contained in CRISP-DM model (Cross Standard Industries Process for Data Mining), in this research there are 6 stages:

a. Phase of business understanding process
The purpose of the research will be to predict whether students graduate on-time or not at Buddhi Dharma University. The data that will be used in this study is the data of students who graduated from 2010 to 2012 as many as 13 attributes and 390 records with details of 300 records for training data and 90 records for data testing.

b. Phase of data understanding
Alumni data for training data and test data collected has 390 records and 13 attributes. All these attributes are collected and analyzed to view dominant data patterns and data types to assist in the process of selecting appropriate data mining methods and algorithms.

Table 2. Attribute and Description.

No.	Attribute	Description
1.	Waktu_kuliah	Time Session
2.	Jenis_kelamin	Gender
3.	Prodi	Study Program
4.	Ips1	Grade Point (GP 1)
5.	Ips2	Grade Point (GP 2)
6.	Ips3	Grade Point (GP 3)
7.	Ips4	Grade Point (GP 4)
8.	Ipk_4	Grade Point Average (GPA 4)
9.	Total_ SKS_ Lulus4	Amount of SKS that has passed until the 4th semester
10.	Jur_Asl_Sekolah	Major of school
11.	Status_Asal_Sklh	Graduated school status
12.	Status_Pek_Ortu	Parent’s job status
13.	Cuti	Leave of absence amount

c. Phase of data processing
This research was conducted from the previous phase results. Alumni data collected in the previous stage after selection there are 13 attributes that will be managed for the modeling stage. To enter the next stage then made observations of the data and found some data discrepancies and require preprocessing stages.

d. Phase of modeling
The modeling process is done by testing the data mining methods to be compared ie C4.5 and K-NN against 13 existing attributes. In the process of modeling will be seen the accuracy of each method. At this stage also conducted experiments on the attributes of

alumni data in the form of modifications or delete attributes that have no significant effect. This is done to increase the value of accuracy. The testing process of each method will be carried out using the 10-folds cross validation testing technique that produces the test statistic values of accuracy, precision, recall and f-measrure. The method with the best accuracy value will be implemented on the prototype to be designed. Testing methods will be done with the help of tools rapidminer 5.3.

e. Phase of Evaluation
The method with the best accuracy value will be implemented on the prototype to be designed. The evaluation phase will be performed by testing the prototype using data other than the training data that has been used during the modeling process, to determine the error rate in the model used. The data used amounted to 90 records will be predicted graduation students on-time or not based on reality data.

f. Deployment
After forming the model and analyzing and measuring in the previous stage, then at this stage also applied the most accurate model for the determination of the classification of students' predictions on-time graduation.

5 Result

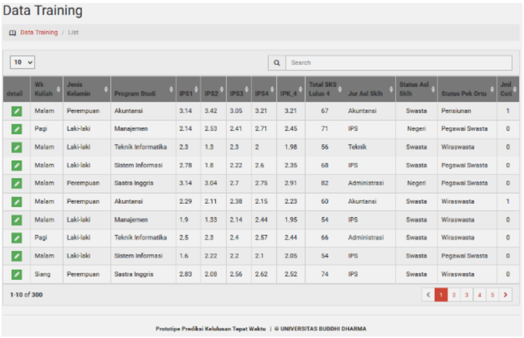
The resulting model will be comparative to find the best level of accuracy that will be used to determine the pattern of the ability of students who have the ability to pass on-time or not. In this research, the validation process is done to find, and convert the data to be used in data mining algorithm method and get good accuracy and performance. In the dataset to be used this, the validation of data used is to delete incomplete or empty data that has no value (null). After that, attribute selection is done to select which attributes are needed from the dataset used in the process of analyzing the student's graduation on-time at Buddhi Dharma University.

Table 3. Comparison of accuracy and AUC.

Prediction	C4.5 Algorithm	K-NN Algorithm
Success prediction on-time	171	169
Success prediction not on-time	99	93
Level of Accuracy	90.0%	87.3%
AUC	0.874	0.500

By looking at the comparison of accuracy and AUC, it can be seen that the C4.5 algorithm has the best accuracy and performance, so the rule generated by C4.5 algorithm serve as the rule for prototype making which can facilitate the prediction of the student's on-time graduation.

6 Implementation



Data Training

Data Training / List

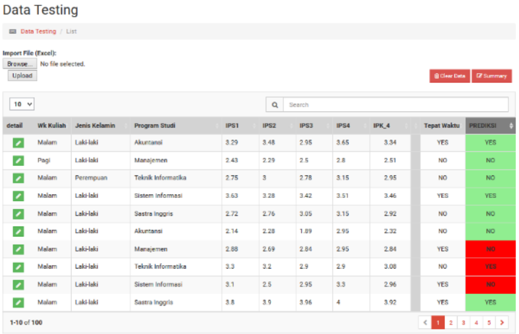
detail	Wk Kuliah	Jenis Kelamin	Program Studi	IPS1	IPS2	IPS3	IPS4	IPK_4	Total SKS Lulus 4	Jur Asl Sklh	Status Asl Sklh	Status Pek Ortu	Jml Cuti
✎	Malam	Perempuan	Akuntansi	3.14	3.42	3.05	3.21	3.21	67	Akuntansi	Swasta	Pensiunan	1
✎	Pagi	Laki-laki	Manajemen	2.14	2.53	2.41	2.71	2.45	71	IPS	Negeri	Pegawai Swasta	0
✎	Malam	Laki-laki	Teknik Informatika	2.3	1.3	2.3	2	1.98	56	Teknik	Swasta	Wiraswasta	0
✎	Malam	Laki-laki	Sistem Informasi	2.78	1.8	2.22	2.6	2.35	68	IPS	Swasta	Pegawai Swasta	0
✎	Malam	Perempuan	Sastra Inggris	3.14	3.04	2.7	2.75	2.91	82	Administrasi	Negeri	Pegawai Swasta	0
✎	Malam	Perempuan	Akuntansi	2.29	2.11	2.38	2.15	2.23	60	Akuntansi	Swasta	Wiraswasta	1
✎	Malam	Laki-laki	Manajemen	1.9	1.33	2.14	2.44	1.95	54	IPS	Swasta	Wiraswasta	0
✎	Pagi	Laki-laki	Teknik Informatika	2.5	2.3	2.4	2.57	2.44	66	Administrasi	Swasta	Wiraswasta	0
✎	Malam	Laki-laki	Sistem Informasi	1.6	2.22	2.2	2.1	2.05	54	IPS	Swasta	Pegawai Swasta	0
✎	Siang	Perempuan	Sastra Inggris	2.83	2.08	2.56	2.62	2.52	74	IPS	Swasta	Wiraswasta	0

1-10 of 300

Prototipe Prediksi Kelulusan Tepat Waktu | © UNIVERSITAS BUDDHI DHARMA

Fig. 1. Data Training Form.

In Fig 1, it shows a form data training which is the whole training data used in this research. Data Training on this menu is read only.



Data Testing

Data Testing / List

Import File (Excel):

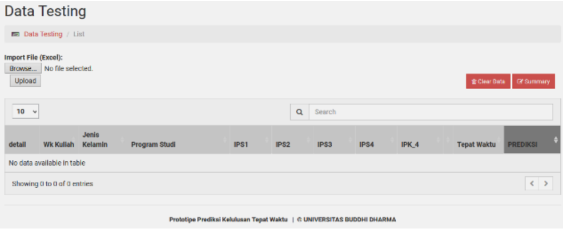
No file selected.

detail	Wk Kuliah	Jenis Kelamin	Program Studi	IPS1	IPS2	IPS3	IPS4	IPK_4		Tepat Waktu	PREDIKSI
✎	Malam	Laki-laki	Akuntansi	3.29	3.48	2.95	3.65	3.34		YES	🟩 YES
✎	Pagi	Laki-laki	Manajemen	2.43	2.29	2.5	2.8	2.51		NO	🟩 NO
✎	Malam	Perempuan	Teknik Informatika	2.75	3	2.78	3.15	2.95		NO	🟩 NO
✎	Malam	Laki-laki	Sistem Informasi	3.63	3.28	3.42	3.51	3.46		YES	🟩 YES
✎	Malam	Laki-laki	Sastra Inggris	2.72	2.76	3.05	3.15	2.92		NO	🟩 NO
✎	Malam	Laki-laki	Akuntansi	2.14	2.28	1.89	2.95	2.32		NO	🟩 NO
✎	Malam	Laki-laki	Manajemen	2.88	2.69	2.84	2.95	2.84		YES	🟥 NO
✎	Malam	Laki-laki	Teknik Informatika	3.3	3.2	2.9	2.9	3.08		NO	🟥 YES
✎	Malam	Laki-laki	Sistem Informasi	3.1	2.5	2.95	3.3	2.96		YES	🟥 NO
✎	Malam	Laki-laki	Sastra Inggris	3.8	3.9	3.96	4	3.92		YES	🟩 YES

1-10 of 100

Fig. 2. Data Testing Form.

In Fig 2, it shows a form data testing that’s can be imported from excel file that already exist according to the format of the column used.



Data Testing

Data Testing / List

Import File (Excel):

No file selected.

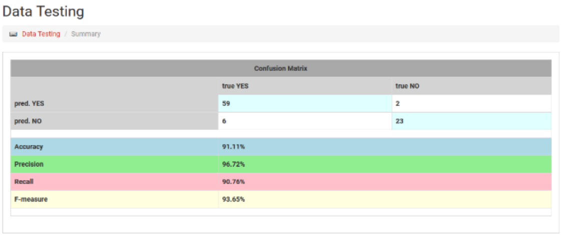
detail	Wk Kuliah	Jenis Kelamin	Program Studi	IPS1	IPS2	IPS3	IPS4	IPK_4		Tepat Waktu	PREDIKSI
No data available in table											

Showing 0 to 0 of 0 entries

Prototipe Prediksi Kelulusan Tepat Waktu | © UNIVERSITAS BUDDHI DHARMA

Fig. 3. Data testing form that has been imported.

In Fig 3, it shows a data testing form that has been imported. It’s a predicted system with color, where green is indicated that the prediction is done correctly and red is indicated that the prediction is wrong, making it easier for the user to evaluate. To view the summary of testing data, click summary button and the display will be as shown in Fig 4.



Data Testing

Confusion Matrix		
	true YES	true NO
pred. YES	59	2
pred. NO	6	23

Accuracy	91.11%
Precision	96.72%
Recall	90.76%
F-measure	93.65%

Fig. 4. Summary data testing form.

The summary of the testing data in Fig. 4 consists of a confusion matrix that represents the test results of imported testing data and calculates the accuracy, precision, recall and f-measure values. In Figure 6 it can be seen that with a student count of 90, the fifty nine students are predicted accurately able to pass on-time and twenty three students are predicted to pass right not on-time. So from a total of 90 data successfully predicted as many as 82 data, so the accuracy value obtained as follows:

$$\text{Accuracy} = \frac{(\text{T on time+T not on time})}{(\text{T on time} + \text{F on time} + \text{T not on time} + \text{F not on tine})}$$

$$= \frac{(59+23)}{(59+2+23+6)}$$

$$= \frac{82}{90} = 0.9111$$

$$= 91{,}11\%$$

7 Summary

7.1 Conclusion

From the measurement of performance and performance that has been done on two methods of classification algorithm, the result of this research can be concluded that:

a. Data mining classification method is appropriate to be implemented into the prototype of student predictions information system on-time.

b. The C4.5 algorithm has the best accuracy between the two classification algorithms. This algorithm will be implemented into the prototype predictions of graduation students on-time. It can be seen that C4.5 algorithm has an accuracy value of 90% and AUC value of 0.874 which belongs to the category of good classification

7.2 Suggestion

a. Using other classification algorithms contained in data mining, such as Naïve Bayes, ID3, CART, Random Forest, Linear Discriminant Analysis and Support Vector Machine algorithms.

b. Adds optimization algorithms such as PSO (Particle Swarm Optimization), Backward Elimination, Forward Selection, and GA (Genetic Algorithm)

References

[1]A. Rohman, “Model Algoritma K-Nearest Neighbor (K-NN) Untuk Prediksi Kelulusan Mahasiswa,” Neo Tek., 2015.

[2]Andri, Y. N. Kunang, and S. Murniati, “Implementasi Teknik Data Mining Untuk Memprediksi Tingkat Kelulusan Mahasiswa Pada Universitas Bina Darma Palembang,” Semin. Nas. Inform. 2013 (semnasIF 2013), vol. 2013, no. A, pp. 56–63, 2013.

[3]D. H. Kamagi and S. Hansun, “Implementasi Data Mining dengan Algoritma C4 . 5 untuk Memprediksi Tingkat Kelulusan Mahasiswa,” Ultim. J., vol. VI, no. 1, pp. 15–20, 2014.

[4]M. S. Mustafa and I. W. Simpen, “Perancangan Aplikasi Prediksi Kelulusan Tepat Waktu Bagi Mahasiswa Baru Dengan Teknik Data Mining (Studi Kasus : Data Akademik Mahasiswa STMIK Dipanegara Makassar),” Creat. Inf. Technol. J., vol. Vol. 1, No, pp. 270–281, 2014.

[5]D. D. A. Saputra and N. Insani, “Model Prediksi Tingkat Kelulusan Mahasiswa Dengan Teknik Data Mining Menggunakan Metode Decision Tree C4.5,” E-Journal Univ. Negeri Yogyakarta, vol. 1, 2014.

[6]M. A. Banjarsari, I. Budiman, and A. Farmadi, “Penerapan K-Optimal Pada Algoritma KNN Untuk Prediksi Kelulusan Tepat Waktu Mahasiswa Program Studi Ilmu Komputer FMIPA Unlam Berdasarkan IP Sampai Dengan Semester 4,” Kumpul. J. Ilmu Komput., vol. 2, no. 2, pp. 159–173, 2016.

[7]B. Swarasmaradhana, M. A. Mukid, and A. Rusgiyono, “Identifikasi Lama Studi Berdasarkan Karakteristik Mahasiswa Menggunakan Algoritma C4.5,” J. Gaussian, vol. 3, no. 2010, pp. 1–11, 2014.

[8]I. Budiman, D. T. Nugrahadi, and R. A. Nugroho, “Implementasi Algoritma K-Nearest Neighbour Untuk Prediksi Waktu Kelulusan Mahasiswa,” in Prosiding Seminar Nasional Riset Terapan 2016, 2016, vol. 5662, pp. 9–10.