J. Int. Intl. Sys. Soft, Vol. 1, No. 1, 2022
ejournal.akakom.ac.id/index.php/JISS/index

JISS Journal of Intelligent Software Systems

Journal of Intelligent Software System – JISS

The Prediction on the Students' Graduation Timeliness Using Naive Bayes Classification and K-Nearest Neighbor

Original Article doi: 10.26798/jiss.v1l1.597

submit: *2022-06-14*, **Accepted:** *2022-07-21*, **Publish:** *2022-07-23*

Anwarudin[1,2*], Widyastuti Andriyani[3†], Bambang Purnomosidi DP[3‡], and Domy Kristomo[3§]

1 Student, Master in Information Technology Universitas Teknologi Digital Indonesia, Yogyakarta, Indonesia

2 STIKES Guna Bangsa ,Yogyakarta

3 Master in Information Technology Universitas Teknologi Digital Indonesia, Yogyakarta, Indonesia

* *E-mail: student.anwarudin@mti.akakom.ac.id*
† *E-mail: widya@utdi.ac.id*
‡ *E-mail: bpdp@utdi.ac.id*
§ *E-mail: domy@utdi.ac.id*

Abstract: The college quality can be seen from the level of punctuality of student graduation. The Prediction on students' graduation timelines can be used as one of the supporting decisions to evaluate students' performance. Currently, the Medical Laboratory Technology study program of STIKES Guna Bangsa Yogyakarta does not have tools to predict the level of students' graduation punctuality early yet. The purpose of this study is to evaluate the application of the Naive Bayes Classification and K-Nearest Neighbor algorithms with predictive modeling of student graduation period. This study applied the academic data from students of the Medical Laboratory Technology study program for the Academic Year (TA) 2015/2016 to 2018/2019. This study utilized an experimental approach by comparing the methods of the Naive Bayes Classification (NBC) and K-Nearest Neighbor (KNN) algorithms. The validation model uses 5-fold Cross Validation, while the evaluation model uses a Confusion Matrix. The results illustrated that the prediction with NBC in this case obtained an accuracy of 96.11%, precision of 82.11% and Recall of 100.00%. Meanwhile, predictions using KNN obtained accuracy of 97.68%, precision of 100.00% and Recall of 86.11%. Thus, KNN is an algorithm with an enhanced level of accuracy to solve the case of predicting the timeliness of students' graduation of the Medical Laboratory Technology Study Program STIKES Guna Bangsa Yogyakarta

Keywords: Prediction on the students' graduation timelines, Naive Bayes Classification, k-Nearest Neighbor, n-Folds Cross Validation

1. Introduction

The increasing number of colleges (universities) with various disciplines lead in the increasing of competition level among colleges as well. Therefore, colleges have to keep improving to be the best and remain in existence among the community.) the quality of a university can be seen from the student success rate and low student failure rate [1]. Based on statistical data for college in 2018, dropout students came from the medical program group with a ratio of 1.74 of the total students [2]. Problems that often occur in the Medical Laboratory Technology

study program of STIKES Guna Bangsa Yogyakarta as follows: 1) students do not graduate on time (the length of study ¿ 3 years), and 2) students graduate on time but with a minimum Grade Point Average (GPA) (GPA=2.50). Thus, graduate. Thus, graduate students who have a minimum GPA are not able to compete in the work field especially as a Civil Servant (PNS). Therefore, to improve the students' quality and success, predictions are required as early as possible to minimize the student failure, it is by predicting the timeliness of student graduation in the future.

Data Mining is a powerful technique for various fields including education [3]. DM is now considered as one of the promising educational technologies so it is called education data mining (EDM) [4]. In the case of prediction, it is needed a model that is able to produce data classification patterns with the ultimate goal of forecasting. The technique or method that will be applied to identify this information is by digging the student academic data. The applied classification algorithms to predict used by previous researchers such as: C.45 [5], Random Forest, Bayesian Network [6], Decision Tree, Naïve Bayes [3]; [5] dan SVM [5]. This research was conducted to find the best algorithm in the case of predicting the timeliness of student graduation. The algorithms that will be applied are Nave Bayes Classification (NBC) and K-Nearest Neighbor (KNN).

2. Literature Review

To clarify the position of this research, a review of similar research is prepared based on a review approach of the similarity of the object of discussion and the similarity of the methods applied. Discussed the comparison of methods for predicting student academic performance [3]. The data used contained student demographic information, previous academic records, and

family background information. In this study, GPA was chosen as the dependent parameter. The independent parameter comprise of nine attributes such as gender, race, hometown (ht), GPA, family income (fi), university mode entry (ume), and SPM grades in three subjects; Malay Language (bm), English (bi), and Mathematics (math). The algorithms used are: DT, NBC, and Rule Based classification. The experimental results illustrate that Rule Based classification is the best model among other classification techniques with the highest accuracy value of 71.3%.

Discussed the prediction of on-time graduation rates by determining the factors that affect student graduation rates. The algorithm used is C4.5. Attributes include: name, gender, age, student_status, thesis_taking status, IPS_Smt1, IPS_Smt2, IPS_Smt3, IPS_Smt4, IPS_Smt5, IPS_Smt6, IPS_Smt7, IPS_Smt8, GPA, and graduation status. Based on the experimental results, the accuracy of the application of C.45 is 95%, Precision is 91.89% and Recall is 93.00%.

Discussed a decision support system to determine the best university [7]. The method used is Simple Additive Weighting (SAW). The criteria applied are Academic achievement, Master Lecturer Extracurricular, and Accreditation Status Facilities Scholarship. The result of the research is a decision support system model with SAW to determine the best university based on the specified criteria.

Identified the student performance from three colleges to improve student performance and prevent Drop Out [6]. The data consists of socioeconomic, demographic such as student academic information. Datasets of student academic information were collected, tested and applied to four classification algorithms such as J48, PART, Random Forest and Bayes Network Classifiers. The classification results point up that Random Forest with twelve selected attributes has an accuracy rate of 99% while Random Forest without feature selection (using all attributes) has an accuracy rate of 84.33%. Thus, Random Forest with feature selection proved to have the best accuracy.

Discussed the student segmentation based on evidence of failure or high performance when

entering undergraduate programs by comparing 6 prediction algorithms such as random forests, decision trees, support vector machines, naive Bayes, bagged trees and boosted trees [5]. The attributes consist of 19 attributes originating from previous school background, psychometric factors, social factors and demographics. The results showed that Random Forest is the best algorithm for predicting the academic performance of college students.

Based on previous research, the difference between this study and the previous one is comparing the NBC and KNN algorithms using attributes from student demographics and student academic performance (compulsory course scores and local content course scores).

3. Method

Figure 1 is the existing stage model in the process of knowledge discovery in database (KDD), which are data selection, preprocessing. data mining and interpretation and evaluation [8].

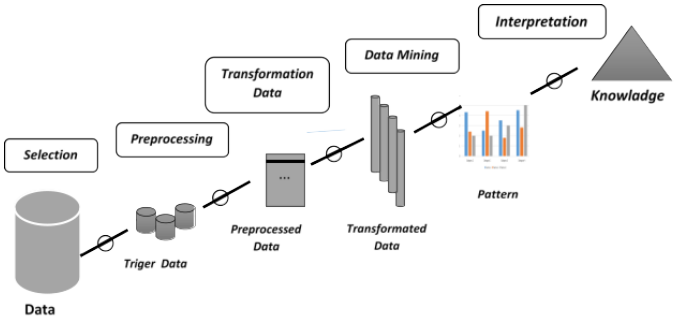

Figure 1. KDD process staged

3.1. Data

At the initial stage, the researchers conducted a Focus Group Discussion (FGD) with the management of the Diploma Degree (D3) Medical Laboratory Technology study program and the New Student Admissions (PMB) department to consider the attributes that affect student

graduation. Input attributes from the study program management include mandatory courses (TLM), local content courses (MLK), GPA, and gender. The input attributes from the PMB section include the origin of the department and the region. The source of this research data was taken from the Academic Information System (SIAKAD) of STIKES Guna Bangsa Yogyakarta, especially the data of D3 students of Medical Laboratory Technology study program Academic Year (TA) 2015/2016 – 2018/2019.

3.2. Data selection

Before the data is processed, a data selection process is carried out (data selection), in which the data are sorted by its relevancy to the analytical task taken from the SIAKAD database to generate a student academic dataset. At this stage, the selection of academic data consists of predictor variables and one target variable. Attributes comprise to 15 predictor variables and one target variable. Predictor variables include: gender, regional origin, department origin, IPS 1, IPS 2, IPS 3, IPS 4, TLM 1, TLM 2, TLM 3, TLM 4, MLK 1, MLK 2, MLK 3, and MLK 4 (Table 1), while the target variable is graduation status (Table 2).

Table 1. Variabel Input

No.	Variable	Description
1.	Gender	The differences in each individual are divided into men and women forms
2.	Origin (island)	The area where the student lived before taking the Diploma Degree (D3)
3.	Department Origin	Department or major based on student interest before taking D3 studies
4.	IPS 1	The 1st semester student achievement index score
5.	IPS 2	The 2nd semester student achievement index score
6.	IPS 3	The 3rd semester student achievement index score
7.	IPS 4	The 4th semester student achievement index score
8.	TLM 1	Compulsory course grades in semester 1
9.	TLM 2	Compulsory course grades in semester 2
10.	TLM 3	Compulsory course grades in semester 3
11.	TLM 4	Compulsory course grades in semester 4
12.	MLK 1	Value of local content courses semester 1
13.	MLK 2	Value of local content courses semester 2
14.	MLK 3	Value of local content courses semester 3
15.	MLK 4	Value of local content courses semester 4
16.	Graduation Status	Graduation status based on GPA and study period

Table 2. Variabel Output

No.	Variable	Classification
1.	Graduation Status	Punctual
		Belated

At this stage, feature selection is also carried out using 15 attributes, and then weighted using the Chi-Square method as equation 1.

$$X_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{1}$$

Description:

X = a chi-square statistic, O = the observed frequency, E = the expected frequency.

Based on the results of feature selection with Chi-square, it was obtained 10 attributes that are potential or have a high weight (Table 3). In this study, all attributes are used for classification.

Table 3. Attribute Weighting Results

Attribute	Weight
Gender	2.605
Origin (Island)	6.457
Department Origin	33.230
MLK1	72.265
MLK2	97.729
MLK3	101.701
MLK4	145.130
IPS1	156.487
IPS2	166.370
TLM1	170.037
TLM2	170.197
TLM3	184.134
TLM4	185.534
IPS3	189.017
IPS4	214.935

3.3. Pre-processing

This stage is data cleaning if there is missing data (missing value), contains noise, errors or outliers, and double data as in Table 1. In handling the empty data, if there is more than one field empty, then it will be deleted. However, if there is only one empty field, the value will be filled in by taking the mean value.

Table 4. Missing Value

No	G	DO	O	IPS SMT 1	TLM	MLK1201	MLK1202	MLK	IPS
				TLM1201	TLM1202	TLM1203	TLM1204	TLM1205	TLM1206
				2	2	2	2	2	2
				3	4	4	3	3	3
1	F	IPA	OJ	4	4	4	3	2	3
2	F	IPA	OJ	3	4	4	4	4	4
3	M	IPS	J	3	4	4	3	3	3
4	F	SMK KES	OJ	3	4	4	3	4	4
5	M	SMK	OJ	4	4	4	3	3	3
6	F	IPS	OJ	4	4		4	4	
7	F	IPA	OJ	3	4	4	3	3	3
8	M	IPS	OJ	4	4	3	4	4	
9	F	IPS	OJ	3	4	4	4	3	3
10	F	IPA	OJ	3	4	4	3	3	3
11	F	SMK KES	OJ	3	4	3	4	3	4
12	M	IPA	J	2	4	4	3	4	4
13	F	IPA	OJ	4	4	3	4	3	3
14	M	SMK KES	J	4	4	4	3	4	4
15	F	SMK	OJ	4	3	3	4	2	3
16	F	SMK KES	OJ	4	4	4	4	4	4
17	M	SMK KES	OJ	4	0	0	0	2	2
18		IPS	OJ	4	4	4	4	3	3
19	P	IPS	OJ	4	4	4	4	4	3

Note : G: Gender; DO: Department Origin; O: Origin; J: Java ;OJ: outside of Java

3.4. Data Transformation

Data transformation is used to change data in a suitable form in the data mining process. In this stage, normalization, attribute selection and discretization are carried out Data.

3.5. Data Mining

The academic dataset used is 257 instances (records). The dataset consists of 15 input variables and 1 output variable. The data mining technique utilizes classification with the NBC and KNN algorithms.

3.6. Cross Validation

Cross validation is a technique to assess or validate the accuracy of a model based on certain datasets. This study uses 5-fold cross validation for training and testing data. This indicates that this study divides the training data into five equal parts and then carries out the learning process five times. Each experiment, take another part of the dataset for testing and use the remaining four parts for learning [9].

3.7. Evaluation Model

The measurement of the performance evaluation of the classification technique in this study uses the Confusion Matrix. This model provides information on the comparison of the classification results carried out by the model with the actual classification results, Confusion Matrix [6]. formed based on four binary classification results. In binary classification, usually the dataset has two labels positive (P) and negative (N). The result is True Positive (TP) which is a acceptable positive prediction, True Negative (TN) is a acceptable negative prediction, False Positive (FP) is a false positive prediction and False Negative (FN) is a false negative prediction. The calculation of the performance measurement using the confusion matrix includes: Accuracy (equation 2), Precision (equation 3), and Recall (equation 4) [10].

Accuracy is the sum of all correct classifications divided by the number of cases. Thus,

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FN + FP)} \tag{2}$$

Precision is the number of acceptable positive classifications divided by the total number of positive classifications. Thus,

$$Precision = \frac{(TP)}{(TP + FP)} \tag{3}$$

Sensitivity (Recall) is the number of acceptable classifications divided by the total number of positives. Thus,

$$Recall = \frac{(TP)}{(TP + FN)} + \frac{TP}{P} \tag{4}$$

4. Algorithms

4.1. NBC Model Prediction

The following are the stages of applying the NBC method in case of predicting graduation on students of the D3 Medical Laboratory Technology Study Program STIKES Guna Bangsa.

- First, calculate the mean and standard deviation of each variable in continuous value. Attributes with continuous values include IPS 1, IPS 2, IPS 3 and IPS 4, TLM 1, TLM 2, TLM 3, TLM 4, MLK 1, MLK 2, MLK 3, and MLK 4 in each category.

- Second, calculate the Prior Probability (P) of the discrete variable graduation status categories (gender, department origin, regional origin) and the probability for each category itself.

- Third, calculate the probability of each category given certain inputs. At this stage, we can use the Gaussian density function with equation 5 for this purpose.

$$P(X_i = x_i \mid Y_i = y_i) = \frac{1}{\sqrt{2\pi}(\sigma_{ij})} e^{\frac{-(x_i - \sigma_{ij})^2}{2\sigma^2}} \tag{5}$$

P : Opportunity, X_i : i-th attribute, x_i : i-th attribute value, Y_i: The class sought, y_i: The sub class sought.

Table 5. **KNN Performance Results**

	True.Punctual	True.Late
Pred. Punctual	214	6
True.Late	0	37
Class Recall	100.00%	86.05%
	Accurancy: 97.68%	

Table 6. **Comparison of NBC and KNN Algorithm Performance**

Model	Accuracy	Precision	Recall
NBC	96,11%	82,11%	100,00%
KNN	97,68%	100,00%	86,11%

5. Conclusions

The results of the study concluded that there are fifteen attributes which affect student graduation, namely gender, department origin, regional origin, TLM 1, TLM 2, TLM 3, TLM 4, MLK 1, MLK 2, MLK 3, MLK 4, GPA 1, GPA 2, GPA 3, and GPA 4. The implementation of NBC in the case of predicting the punctuality of student graduation obtained an accuracy rate of 96.11% with a precision level of 82.11% and recall of 100.00%. Therefore, prediction using the KNN algorithm obtained an accuracy of 97.68% with a precision level of 100.00% and a recall of 86.11%. Thus, the algorithm with the best level of accuracy and precision to solve the case of predicting the punctuality of students' graduation of Medical Laboratory Technology STIKES Guna Bangsa Yogyakarta is the K-Nearest Neighbor (KNN). Further research can enhance other data mining classification algorithms besides NBC and KNN as a comparison method.

References

[1] W. Purba, S. Tamba, and J. Saragih, "The effect of mining data k-means clustering toward students profile model drop out potential The effect of mining data k-means clus-

tering toward students profile model drop out potential," J. Phys. Conf. Ser., 2018, doi: 10.1088/1742-6596/1007/1/012049.

[2] Kementerian Ristekdikti, "Statistik Pendidikan Tinggi," Pus. Data dan Inf. Ilmu Pengetahuan, Teknol. dan Pendidik. Tinggi, 2018.

[3] F. Ahmad, N. H. Ismail, and A. A. Aziz, "The prediction of students' academic performance using classification data mining techniques," Appl. Math. Sci., vol. 9, pp. 6415–6426, 2015, doi: 10.12988/ams.2015.53289.

[4] M. Wook, Z. M. Yusof, and M. Z. A. Nazri, "The Acceptance of Educational Data Mining Technology among Students in Public Institutions of Higher Learning in Malaysia," Int. J. Futur. Comput. Commun., vol. 4, no. 2, pp. 112–117, Apr. 2015, doi: 10.7763/ijfcc.2015.v4.367.

[5] V. L. Miguéis, A. Freitas, P. J. V Garcia, and A. Silva, "Early segmentation of students according to their academic performance: A predictive modelling approach," Decis. Support Syst., vol. 115, pp. 36–51, Nov. 2018, doi: 10.1016/j.dss.2018.09.001.

[6] S. Hussain, F. M. Dahan, Neama Abdulaziz, Ba-Alwib, and N. Ribata, "Educational Data Mining and Analysis of Students ' Academic Performance Educational Data Mining and Analysis of Students ' Academic Performance Using WEKA," Indones. J. Electr. Eng. Comput. Sci., vol. 9, no. February, pp. 447–459, 2018, doi: 10.11591/ijeecs.v9.i2.pp447-459.

[7] N. Aminudin, M. Huda, A. Kilani, W. Hassan, W. Embong, and A. M. Mohamed, "Higher education selection using simple additive weighting," Int. J. Eng. Technol., vol. 7, no. 2, pp. 211–217, 2018.

[8] V. N. Vapnik, "Statistics The Elements of Statistical Learning," Math. Intell., vol. 27, no. 2, pp. 83–85, 2009, [Online]. Available: http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf.

[9] Piotr Kokoszka and M. Reimherr, Introduction to Functional Data Analysis. 2017.

[10] A. J. Larner, The 2x2 Matrix: Contingency, Confusion and the Metrics of Binary Classification. 2021.