ISSN 2347 - 3983
Volume 8. No. 6, June 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter60862020.pdf
https://doi.org/10.30534/ijeter/2020/60862020

Prediction of Students Study Period using K-Nearest Neighbor Algorithm

Thomas Asril[1], Sani M. Isa[2]
[1] Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480, thomas.asril001@binus.ac.id
[2] Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480, sani.m.isa@binus.ac.id

ABSTRACT

Currently, many higher educations are highly oriented to improve the quality of education and students' learning achievement. Student achievement determined based on the final grades on certain courses. This study is conducted to use a classification algorithm that applied to predict student study period based on student achievement of final grade. This study proposed K-Nearest Neighbor (K-NN) algorithm that classifies by estimate the distance of student grade. This study analyzes 1,989 computer science student grades data in BINUS University from 2016 to 2019 and the algorithm reaches accuracy of 93.2% in predicting study on-time status, 91.5% in predicting study total year and 75.63% in predicting study total semester. In addition, the accuracy was evaluated from 398 testing data. The conclusion of this study is K-NN Algorithm can be applied to predict student study period based on student grades and increase the graduation rates of students.

Key words: Data Mining, K-Nearest Neighbor, Study Period, Supervised Learning

1. INTRODUCTION

Higher education is a sector that is quite developing and developing rapidly, especially in East Asia. Higher education has a lot of interest in this country because the economy is based on export and the export sector requires skilled labor for its operations. Now, production has become knowledge-based and has an impact on increasing demand for highly qualified professionals. The university itself has produced knowledge in a creative knowledge production system. This is needed to create the morale of skilled workers in creating knowledge related to various situations. The transition from knowledge production to knowledge is very important for universities in developing countries [1].

Bina Nusantara University is one of the educational institutions in Indonesia that has huge contribution in activities that take place in the academic and non-academic fields. This is proven because Bina Nusantara University ranked #801-1000 Top 1000 World University Rankings 2020, which makes Bina Nusantara University is one of the best university in Indonesia [2]. This predicate makes Bina Nusantara University has large number of students. The number of Bina Nusantara University students has always increased from 36,076 people in 2017/2018 to 38,380 in 2018/2019. The number of active students majoring in Informatics Engineering are 6,452 people in 2018/2019 [3].

This condition will be compounded by student performance which is [4] claimed that it is common for most first-year students to perform poorly in the first semester and having bad grades. Students experience many changes due to the assessment of the education system and loss of self-esteem. Student anxiety levels are often high, which in turn, reduces the ability to think critically, learn, integrate, and of course, recall information learned in class during examinations or assessments. Student academic performances can be affected by several factors: a) personal: biopsychosocial, class attendance, emotional intelligence; b) institutional: information, orientation, courses and services, aspects related to teachers (evaluations, pedagogy, interpersonal relationships); and c) social: socioeconomic status, family [9]. Even more, the existence of the maximum study period which is the time needed by a student to complete studies in college and make the student has a big target. The study period is required to undertake a study program at Strata 1 (S-1) level. The study load in the S-1 Study Program is calculated with a Semester Credit Unit (SCU) of at least 144 SCU. The maximum length of study for a student is seven years (14 semesters) [5].

Given this lack of student performances, it will make poor student grades and will affect their study period. This is proven by most of the researchers in the world applied the student grades with the GPA to assess the performance of the student. They applied GPA (grade point average) to evaluate performance of the students in certain semesters [10, 11, 12]. Some other researcher, they measure student performance through the result of certain courses or the previous year result [13, 14, 15].

Seeing this opportunity, this study focuses in predicting study period based on student grades in certain courses in each semester. Metadata attributes that were courses from Bina Nusantara University internal databases were used as predictor variables. There are many new information can be discovered from grading data. Discussing new information discovery, data mining is the most popular approach. Data mining is study that applied to analysis and discover new information from large amount of data. The information is generated by using certain method and algorithm. Appropriate method must be defined based on information that will be gained. In data mining, supervised learning is a method that used to predict and estimate information based on past data. There are many algorithms included in supervised learning, such as Naïve Bayes, Decision Tree, Nearest Neighbor, Neural Network Support Vector Machine, etc [6].

Based on the dataset, nearest neighbor is appropriate algorithm because nearest neighbor has become a popular supervised learning algorithm and fits with numerical data and most of grading data are numerical. This study conducted nearest neighbor algorithm by define the number of k. This algorithm is called as K-Nearest Neighbor (K-NN). This study will predict the study period with K-NN and review the performances of the models and evaluate them using several metrics such as accuracy, precision, recall, and f1 score to determine which model has the best accuracy. Finally, this study aims to achieve better accuracy to classify student study period which is expected to be a solution to increase the graduation rates of its students.

The rest of the paper is organized as follows. Section 2 highlights previous works related to the paper. Section 3 presents the proposed method used in the paper. Section 4 gives details of the evaluation results and discussion based on the results shown. Lastly, Section 5 provides the conclusion of the paper.

2. PREVIOUS WORKS

There are many studies have been applied data mining in education, especially in grading system. The study topic used as a comparison in this study the main criteria using student grading data. The amount of study on student performance predictions is quite large, but study on prediction of study period is only slightly. Study conducted by M. Ihsan Zul make a prediction of student grades by using supervised learning classifications include, among others, using K-Nearest Neighbor as an algorithm and K-fold of 10 as a parameter, creating a system that can predict student grades based on total absenteeism, quiz, homework, report and midterm as features. The accuracy of prediction results achieved 70.51% [6].

Other study conducted by Kabakchieva that predicted student performance based on their university-performance, pre university and personal characteristics. The study applied four data mining algorithms such as K-Nearest Neighbor, Decision Tree, OneR Rule Learner, and Neural Network. The study can achieve the highest accuracy on the Neural Network (73.59%) and followed by the K-NN algorithm (70.49%) [23].

In 2013, Kabakchieva made comparison between several algorithms such as Decision Tree (J48), Bayesian (Naïve Bayes and BayesNet), K-Nearest Neighbor classifier and Rule learners (OneR and JRip). The data provided run many transformation so some parameters were deleted. The results obtained are 66.59% accuracy for the decision tree (J48) classifier. Bayesian classifier gets 59% result which is not very good result. Similarly, the accuracy of the K-NN classifier, get around 60%, JRip classifier get 63% and OneR get 54-55%. From the related study results, JRip and J48 classifier get the highest accuracy value which is around 64-66% [7].

On the other hand, Alfere and Maghari enhanced the accuracy using modified KNN Classifier. The study combines the Euclidian, Cosine, Minkowski with its classifier and can scored 94.1% accuracy and takes less time than other algorithms with 0.97885 sec using the Weighted KNN Algorithm [22].

Another method proposed by Amra and Maghari use personal data to help the ministry of education to improve the performance due to early prediction of student performance. Their results showed that KNN classification method gives better accuracy (approx 83.65%) when compared to Naïve Bayes classifier which gives the accuracy of approx 75.77% [17].

In other study, Cherry, Enrique and Alvin made system to predict students' employability using Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), Naïve Bayes (NB), K-Nearest Neighbor (KNN), and Random Forest (RF). From the learning algorithms, KNN obtained 79.90% for accuracy and SVM has the highest accuracy with 91.22% [25].

Ahuja and Kankane perform algorithms, that are Naive Bayes, K-nearest neighbors, and Ctree to build classification models to predict the student's probability of getting his/her graduation degree. The accuracy for Naïve Bayes can reach 84.3%, Ctree get 86.56%, and K-NN has the lowest accuracy that is 76.45% [24].

Some study work by H. Al-Shehri, et al. presented two prediction models for the estimation of student's performance in final examination. That study applied K-Nearest Neighbor (K-NN) algorithm and Support Vector Machine (SVM) algorithm to predict the student's grade. The study outcome showed that Support Vector Machine (SVM) achieved a 1% better results with correlation coefficient of 96%, while the K-Nearest Neighbor (K-NN) achieved 95% [8].

3. PROPOSED METHOD

This study conducted by applied some steps. Each step described in following parts:

3.1 Data Preparation

The dataset used in this study was obtained from internal databases that is stored in Microsoft SQL Server. The data are

BINUS computer science students who have graduated from 2016 and 2019 with a total of 1,989 students. In addition to student data, this study will use student grades consisting of 48,414 records, 4 tables, 15 columns. After the dataset is collected, pre-processing process such as data cleaning and data transformation. The creation and selection of features aim to provide a good quality dataset that can improve the performance of the study period prediction.

3.2 Data Processing

At this stage includes all activities to prepare data to be entered into a modeling tool that will be applied at a later stage. In the data pre-processing phase, this process is carried out in detail as follows:

Data Cleaning. In collecting data, the authors work with internal data division as the owner of data info, grades, status and study period of a student. At this stage the IT Data Center collect the data by carrying out several activities such as filling in missing values, handling outlier data, correcting inconsistent data, and resolving redundancy problems. The internal data division also combine four tables which are student personal information table, course table, student score table, and study period table into data as excel. The feature list can be seen in Table 1.

Table 1: Feature list

Type of Data	Feature Name	Feature Type
Student Personal Data	ADMIT_TERM	Nominal
	MASKED_NIM	Nominal
	MASKED_BINUSIANID	Nominal
	MASKED_NAME	Nominal
	BIRTH_PLACE	Nominal
	BIRTH_YEAR	Nominal
Course Data	COURSE_CODE	Nominal
	COURSE_NAME	Nominal
Student Grade Data	FINAL_SCORE	Numeric
	FINAL_GRADE	Nominal
Study Period Data	REGISTER_DATE	Nominal
	STATUS	Nominal
	GRADUATION_DATE	Nominal

Most features (12) are nominal variables that have varying amounts of values, and only one feature is numeric variable named FINAL_SCORE. This feature will be the predictor variable in the next step. The example of dataset after using data cleaning can be seen in Figure 1.

From the example, we can see the student number, binusian id and name are masked by the internal data division to maintain student privacy. There are several personal data such as birthplace, and birth year. Moreover, there are course outline data such as course code and course name. In the next step, the course code will be used for the feature name.

Data Transformation: At this stage, the features that will be used in the data will be only 4 of 13 features. In addition, the authors transpose the data and makes certain course code as the predictor variables. COMP6047 (Algorithm and Programming), COMP6048 (Data Structure), COMP6056 (Program Design Method) and COMP6100 (Software Engineering) are selected based on the highest total credits (SCU) and as requirements to pass in the semester [21]. Predicted variables are created by calculating the difference between the students register and graduation date. From the process, 3 predicted variables are generated such as "OnTimeStatus", "TotalYear" and "TotalSemester". After that, the authors change the data in the form of excel into Command Separated Values (CSV). The summary of selected features of dataset can be seen in Table 2.

MASKED_NIM	MASKED_BINUSIANID	MASKED_NAME	BIRTH_PLACE	BIRTH_YEAR	COURSE_CODE	COURSE_NAME	FINAL_SCORE	FINAL_GRADE	REGISTER_DATE	STATUS	GRADUATION_DATE
16XXXXX001	XXXXXX001	AxxxxN	JAKARTA	1994	T0016	Algorithm and Programming	58	D	17/Sep/2012	Completed Program	26/Nov/2016
16XXXXX001	XXXXXX001	AxxxxN	JAKARTA	1994	CB412	CB: Self Development	82	B	17/Sep/2012	Completed Program	26/Nov/2016
16XXXXX001	XXXXXX001	AxxxxN	JAKARTA	1994	T0026	Data Structures	69	C	17/Sep/2012	Completed Program	26/Nov/2016
16XXXXX001	XXXXXX001	AxxxxN	JAKARTA	1994	T0016	Algorithm and Programming	75	B	17/Sep/2012	Completed Program	26/Nov/2016
16XXXXX001	XXXXXX001	AxxxxN	JAKARTA	1994	T0104	Program Design Methods	86	A	17/Sep/2012	Completed Program	26/Nov/2016
16XXXXX001	XXXXXX001	AxxxxN	JAKARTA	1994	T1414	Software Engineering	91	A	17/Sep/2012	Completed Program	26/Nov/2016
16XXXXX001	XXXXXX001	AxxxxN	JAKARTA	1994	EN002	Entrepreneurship II	88	A	17/Sep/2012	Completed Program	26/Nov/2016
16XXXXX001	XXXXXX001	AxxxxN	JAKARTA	1994	T0123	OO Software Engineering	62	D	17/Sep/2012	Completed Program	26/Nov/2016
16XXXXX001	XXXXXX001	AxxxxN	JAKARTA	1994	T0144	Advanced Topics in Software	65	C	17/Sep/2012	Completed Program	26/Nov/2016
16XXXXX001	XXXXXX001	TxxxxL	BEKASI	1995	T0016	Algorithm and Programming	68	C	17/Sep/2012	Completed Program	26/Nov/2016
16XXXXX001	XXXXXX001	TxxxxL	BEKASI	1995	CB412	CB: Self Development	85	A	17/Sep/2012	Completed Program	26/Nov/2016
16XXXXX001	XXXXXX001	TxxxxL	BEKASI	1995	T0026	Data Structures	88	A	17/Sep/2012	Completed Program	26/Nov/2016
16XXXXX001	XXXXXX001	TxxxxL	BEKASI	1995	T0104	Program Design Methods	65	C	17/Sep/2012	Completed Program	26/Nov/2016
16XXXXX001	XXXXXX001	TxxxxL	BEKASI	1995	T1414	Software Engineering	72	C	17/Sep/2012	Completed Program	26/Nov/2016

Figure 1: Example of dataset after using data cleaning

Table 2: Summary table of selected features of dataset

Feature Name	Description	Feature Type	Values
COMP6047	Student score in Algorithm and Programming	Num	0 – 100
COMP6048	Student score in Structure score	Num	0 – 100
COMP6056	Student score in Program Design Method	Num	0 – 100
COMP6100	Student score in Software Engineering	Num	0 – 100
OnTimeStatus	As on-time / not on-time	Nom	1 = on-time (89,6%), 0 = not on-time (10,4%)
TotalYear	Total year of study period	Nom	4 values (4, 5, 6, 7)
TotalSemester	Total semester of study period	Nom	7 values (7, 8, 9, 10, 11, 12, 13)

4 features are nominal variables while the rest of features are numeric variables. But the final dataset are imbalance, because there are 89,6% student that are on-time in their study, and only 10,4% are not on-time. This problem may affect to the prediction accuracy especially for predicting the students total year or the total semester in their study, because the 10,4% not on-time status are distributed to 3 values for total year which are '5', '6' and '7' years. The 10,4% not on-time are grouped to 5 values for total semester which are 9', '10', '11', '12' and '13' semesters. The example of dataset after using data cleaning can be seen in Figure 2.

COMP6047	COMP6048	COMP6056	COMP6100	OnTimeStatus	TotalYear	TotalSemester
51	71	71	82	0	5	9
83	86	59	83	0	5	9
76	76	85	84	0	6	11
68	56	86	74	0	5	9
58	66	71	73	0	5	10
59	79	76	76	0	5	9
60	58	61	80	0	5	10
79	67	72	78	0	5	9
56	86	56	80	0	5	9
75	85	60	71	0	5	9
76	53	58	78	0	6	12
68	74	87	66	0	5	9
0	0	0	0	0	7	13
67	53	69	71	0	5	10

Figure 2: Example of dataset after using data transformation

3.3 Model Selection

Nearest Neighbor is a classification algorithm that studies neighboring classes as an approach and uses the K value to determine the number of nearest neighbors as one of the guidelines for sample data points [18]. At runtime, the sample data points must be in memory which makes this algorithm also called a memory-based technique. This algorithm is called KNN because the classification is very dependent on the number of uses of the closest neighbors that can determine the class of a data point. Some researchers [18] improved KNN according to their distances from new sample data point. But memory requirement and computational complexity remain the main concern always. To further improve the dataset, some data points can be also eliminated from data set, and those data points don't affect the result. For example, a testing data fill a certain position in the Cartesian graph, the labeled data will surround the testing data. In this study, label refer to on-time status, total year and total semester are indicated as prediction label selected from closest neighbor to the testing data.

Measuring distances on data can be done with several approaches. Euclidean distance and Manhattan distance are most popular approaches. In this study, authors chose Euclidean Distance described in the equation (1) [16].

$$dist(x_1 x_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \qquad (1)$$

In the equation (1), distance (x1, x2) is distance between x1 and x2. x1 is attribute value in training data, and x2 is in testing data value. The selected attributes are used to measure the distance. The shortest would be chosen to predict the testing data label/class defined in k value. KNN nearest neighbor classification algorithm can be expressed as the following pseudo code can be seen in Figure 3 [17].

K← the number of nearest neighbors

For each object Z do

Calculate the distance between every object x and z in the training set d(x, z)

Neighborhood←the k neighbors, closest to z in the training set

Zclass←select class (according to neighborhood)End for

Figure 3: pseudo code of KNN nearest neighbor classification algorithm [17]

The next example shows that there are three classes X, Y and Z as shown in Figure 4. P is the data whose label want to identify with the K value is 5. The KNN would use the Euclidean distance to calculate each sample pair that is closest to the P value. In the figure 4, the four nearest neighbor samples are contained in class label X, while single sample belongs to class label Z. So the P data can be classified as a class label Z [20].

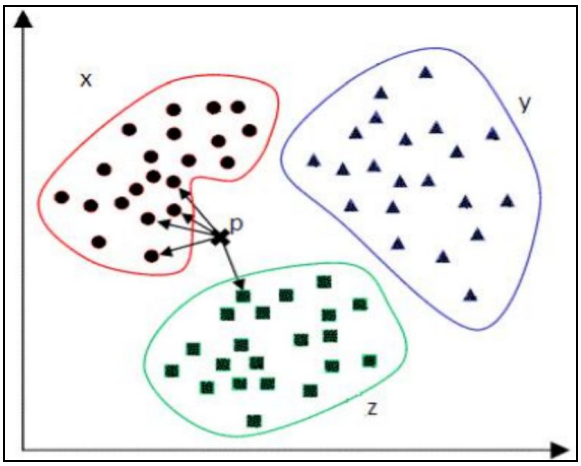

Figure 4: A visual example of K-NN classifier [20]

The advantages using KNN are fast training, ease in comprehension and implementation, strong in noise training data and good for multiclass classification [19]. The KNN algorithm has disadvantages, such as the advanced determination of the optimal k and the requirement to define the value of k in selecting fit attributes. This will reduce noise effects on the classification and making limits between the classification bias. If the k value is too low, the accuracy will be decreased too [26].

Evaluation of algorithm performance can be done with certain data mining techniques. Confusion matrix the most popular one. It maps the detail result of prediction described in matrix and give difference column between actual and prediction label that can be seen in Table 3.

Table 3: Confusion Matrix (Two Labels)

		Prediction	
	Label / Label	X	Y
Actual	X		
	Y		

The confusion matrix result is the accuracy of prediction based on training and testing data. From the confusion matrix, we can also get several performance metrics, such as precision, recall and F1 Score.

4. RESULT AND DISCUSSION

In this section, the experimental evaluation aspects of classification methods using the K-Nearest Neighbor. All experiments were implemented with scikit-learn library on a workstation with Core I7 CPU and 8 GB RAM. The dataset are divided into two sets using with the ratio is 80:20. There are 1,591 data used as training data. This section also uses 3 predicted values which are 'on-time status', 'total year' and 'total semester'. All of training data can be seen in Table 4, Table 5, Table 6.

Table 4: Example of training data in predicting on-time status

C47	C48	C56	C61	OT
90	97	95	92	1
75	59	56	65	1
66	84	79	80	1
85	81	87	86	1
…	…	…	…	…
66	70	70	79	1

Table 5: Example of training data in predicting total year

C47	C48	C56	C61	TY
70	85	81	92	4
53	37	66	76	6
93	94	98	90	4
71	85	77	86	5
…	…	…	…	…
66	70	70	79	4

Table 6: Example of training data in predicting total semester

C47	C48	C56	C61	TS
57	73	79	82	8
78	75	79	76	8
76	77	65	83	7
74	71	75	71	7
…	…	…	…	…
66	70	70	79	7

The features are separated in two groups and named as predictor variable and predicted variable. This study defines COMP6047 (C47), COMP6048 (C48), COMP6056 (C56) and COMP6100 (C61) as predictor variables. While On-time (OT), Total Year (TY) and Total Semester (TS) are used as predicted variables. Then, data must be preprocessed to gain the best accuracy. This preprocessing step was done by measuring accuracy. Table shown the sample of data that is used as training data. There are 398 data used as testing data. Table 7, Table 8, Table 9 shown generated testing data by using sci-kit learn library train_test_split method with the same random_state = 4. This table consists of columns used for the prediction result and would be compared to the actual column to compute the accuracy. The algorithm is used to produce prediction of student on-time status, total year, total semester based on training and testing data. The example of selected testing data with the prediction result can be seen in Table 7, 8, 9.

Table 7: Example of testing data with the prediction result in predicting on-time status

C47	C48	C56	C61	OT (Prediction)	OT (Actual)
84	88	69	83	1	1
66	73	85	89	1	1
73	80	76	90	1	1
90	91	81	85	1	1
66	55	43	83	0	0

Table 8: Example of testing data with the prediction result in predicting total year

C47	C48	C56	C61	TY (Prediction)	TY (Actual)
87	93	82	78	4	4
79	94	89	92	4	4
66	70	69	77	4	4
75	88	81	74	4	4
90	91	89	85	4	4

Table 9: Example of testing data with the prediction result in predicting total semester

C47	C48	C56	C61	TS (Prediction)	TS (Actual)
81	77	79	85	7	7
68	74	87	66	7	9
84	87	81	85	7	7
73	77	69	92	7	7
78	69	76	81	7	7

In the example of testing data in predicting total semester, incorrect prediction will be highlighted red because the prediction and the actual value is not same. It will prove the testing accuracy will not as high as the other variables. Final testing accuracy result has done by analyzing the best k value of K-NN algorithm. The accuracy comparison for on-time status can be seen in Table 10.

Table 10: Testing accuracy comparison of K-NN Algorithm in predicting on-time status

Test Number	K (of K-NN)	Accuracy (%)
1	2	86.93
2	3	83.92
…	…	…
8	9	93.2
…	…	…
41	40	90.2

Based on k evaluation, the best k on testing result is highlighted green which is 9 with an accuracy of 93.2%. This table is visualized in Figure 5.

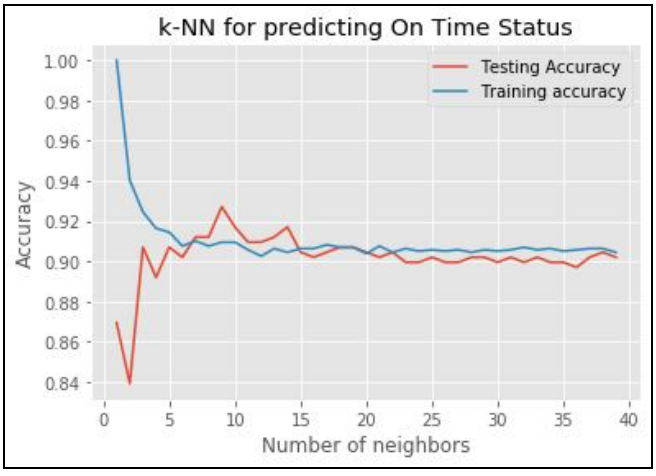

Figure 5: Comparison of training and testing accuracy graph in predicting on-time status

The testing accuracy had decreased in the second k and rose significantly to reach 93.2% in the ninth k, after that k decreased and stable at 90%. Testing accuracy comparison for student total year can be seen in Table 11.

Table 11: Testing accuracy comparison of K-NN Algorithm in predicting total year

Test Number	K (of K-NN)	Accuracy (%)
1	2	85.18
2	3	89.2
…	…	…
8	9	91.5
…	…	…
41	40	89.7

Based on k evaluation, the best k on testing result is highlighted green which is 9 with an accuracy of 91.5%. This table is visualized in Figure 6.

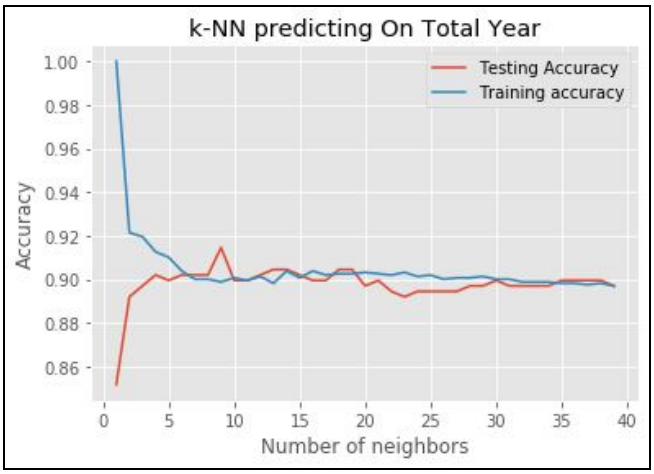

Figure 6: Comparison of training and testing accuracy graph in predicting total year

Different from the previous graph, in the process of k in predicting the total year, testing accuracy has increased to the ninth k with an accuracy of 91.5%. And the last testing accuracy comparison for student total semester can be seen in Table 12.

Table 12: Testing accuracy comparison of K-NN Algorithm in predicting total semester

Test Number	K (of K-NN)	Accuracy (%)
1	2	60.55
2	3	68.6
3	4	70.35
4	5	69.1
…	…	…
41	40	75.63

Based on k evaluation, the best k on testing result is highlighted green which is 40 with an accuracy of 75.63%. This table is visualized in Figure 7.

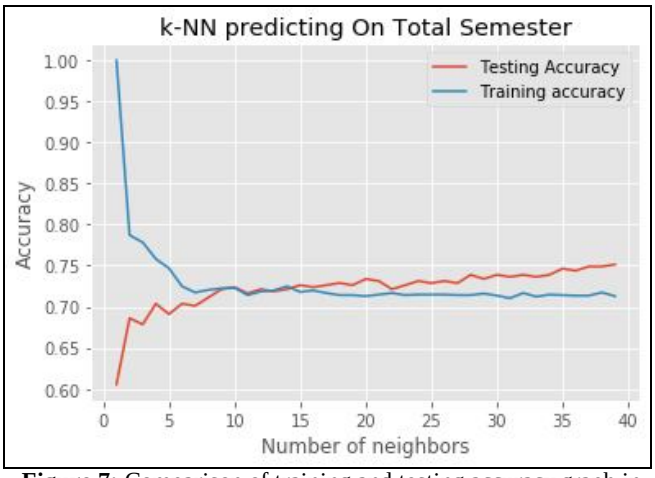

Figure 7: Comparison of training and testing accuracy graph in predicting total semester

This graph is the most different of all graph, because the testing accuracy rose stably and experiencing a peak at k=40. This condition is influenced by the number of predicted variables. The imbalance data is also the factor that graph can distinct with others. After calculating the accuracy based on k, this study also calculates the performance metrics using confusion matrix, precision, recall, F1 score for student on-time status that can be seen in Table 13 and Table 14.

Table 13: Confusion matrix in predicting on-time status

Prediction / True	0	1	All
0	8	34	42
1	4	362	356
All	12	386	398

Table 14: Precision, Recall, F1 Score, Training and Testing Accuracy in predicting on-time status

Label / Class	Precision (%)	Recall (%)	F1 Score (%)	Train Acc (%)	Test Acc (%)
0	67	19	30	90.45 (k= 34)	93.2 (k = 9)
1	91	99	95		

Based on Table 13 and Table 14, it can be examined that label '1' has high precision (91%) , recall (99%) and F1 score (95%) because there is 89,6% data that leans to its label while label '0' has low recall and F1 score percentage.

Table 15: Confusion matrix in predicting total year

Prediction / True	4	5	6	7	All
4	356	0	0	0	356
5	29	5	0	0	34
6	4	0	0	0	4
7	3	1	0	0	4
All	392	6	0	0	398

Table 16: Precision, Recall, F1 Score, Training and Testing Accuracy in predicting total year

Label / Class	Precision (%)	Recall (%)	F1 Score (%)	Train Acc (%)	Test Acc (%)
4	91	100	95	90.06 (k= 23)	91.5 (k = 9)
5	83	15	25		
6	0	0	0		
7	0	0	0		

In Table 15 and Table 16 has the same problem where the class '6 years' and '7 years' can't be predicted because lack of imbalance data. But K-NN still can predict the label '4 years' correctly. Even the recall, precision and F1 score for label '4 years' is similar with the label '1' in the previous predicted variable.

Table 17: Confusion matrix in predicting total semester

Prediction / True	7	8	9	10	11	12	13	All
7	259	15	0	0	0	0	0	274
8	44	38	0	0	0	0	0	82
9	8	18	2	0	0	0	0	28
10	0	6	0	0	0	0	0	6
11	1	2	0	0	0	0	0	3
12	0	1	0	0	0	0	0	1
13	0	3	1	0	0	0	0	4
All	312	83	3	0	0	0	0	398

Table 18: Precision, Recall, F1 Score, Training and Testing Accuracy in predicting total semester

Label / Class	Precision (%)	Recall (%)	F1 Score (%)	Train Acc (%)	Test Acc (%)
7	91	100	95	70.71 (k= 36)	75.63 (k=40)
8	83	15	25		
9	0	0	0		
10	0	0	0		
11	0	0	0		
12	0	0	0		
13	0	0	0		

In Table 17 and Table 18, algorithm only can predict 3 of 7 labels, and only the label '7 semester' has the same precision, recall and F1 score as label '4 year'. It definitely affects the training and testing accuracy. K-NN only can give 70.71% for training accuracy and 75.63% for the testing accuracy.

5. CONCLUSION

This study proposed important variables to predict computer science student study period. The variables were

selected based on student basic course score. Selected variables that used to predict student study period are Algorithm and Programming (COMP6047) score, Data Structures (COMP6048) score, Program Design Method (COMP6056) score and Software Engineering (COMP6100) score. This study also proposed three predicted variables such as, (1) study on-time status, (2) study total year and (3) study total semester. These variables are combined with K-NN algorithm. Based on study analysis, the accuracy of the algorithm for first predicted variable reached 93.2 % with the best k is 9. For the second predicted variable, K-NN can reach the accuracy of 91.5% with the best k is 9. For the third predicted variable, K-NN only can reach the accuracy of 75.63% with the best k is 40. The biggest challenge in this study is the imbalance dataset between student that graduate on-time and not on-time. For future work, this study will add the dataset of not on-time student to improve the variation of dataset which may help in elevating the study further. We can also use the combination of other supervised learning algorithm to further enhance the quality of this study. We will also develop a prediction study period system based on basic course score and the prediction result can be used as early reminder system to increase the graduation rates of its students.

ACKNOWLEDGEMENT

The authors are grateful to the IT Enterprise Data Management Team of Bina Nusantara University for providing the student data that was used in this study free of cost. The authors would also like to thank the anonymous reviewers for their constructive comments and for improving this study.

REFERENCES

1. H. Verheul, **Higher Education Reform in Indonesia**, in *The Theory and Practice of Institutional Transplantation: Experiences with the Transfer of Policy Institutions*, 2003, pp. 185-198. https://doi.org/10.1007/978-94-011-0001-4_12
2. "**QS Top Universities**," 2019. [Online].
3. "**SRV5 PDDIKTI: Pangkalan Data Pendidikan Tinggi,**" 2019. [Online].
4. H. A. Virkler, **A Christian's guide to critical thinking**, 2005.
5. D. Sametko, H. A. Syafrudie and Sutrisno, **Kecerendungan Lama Studi dan Prestasi Belajar Mahasiswa Jalur Reguler dan Non-Reguler Program Studi Pendidikan Teknik Bangunan**, *TEKNOLOGI DAN KEJURUAN*, vol. 37, pp. 153-166, 2014.
6. M. Ihsan Zul, **Predicition of Student Final Grade by using K-Nearest Neigbor Algorithm**, in *The 2nd International Conference on Science and Technology For Sustainability 2016*, Pekanbaru, 2016.
7. D. Kabakchieva, **Predicting Student Performance by Using Data Mining Methods for Classification**, *Cybernetics and Information Technologies*, 2013. https://doi.org/10.2478/cait-2013-0006
8. H. Al-Shehri, A. Al-Qarni, L. Al-Saati, A. Batoaq and H. Badukhen, **Student performance prediction using Support Vector Machine and K-Nearest Neighbor**, *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering*, no. IEEE, 2017.
9. L. Rojas, A. Loria, N. T. Weng, J. Morera and L. Jiménez, **Factors that Affect the Academic Performance of Pharmacy Students at the University of Costa Rica**, *American Journal of Educational Research*, vol. 6, no. 6, pp. 729-734, 2018.
10. L. J. Stephen and L. A. Schaben, **The Effect of Interscholastic Sports Participation on Academic Achievement of Middle Level School Students**, *NASSP Bulletin*, vol. 86, pp. 34-41, 2002. https://doi.org/10.1177/019263650208663005
11. S. Galiher, **Understanding the effect of extracurricular involvement**, Indiana, 2006.
12. N. Darling, L. L. Caldwell and R. Smith, **Participation in School-Based Extracurricular Activities and Adolescent Adjustment**, *Journal of Leisure Research*, vol. 37, no. 1, pp. 51-73, 2005.
13. S. T. Hijazi and S. R. Naqvi, **Factors affecting students' performance**, *Bangladesh e-Journal of Sociology*, vol. 3, no. 1, 2006.
14. L. M. Tho, **Some evidence on the determinants of student performance in the University of Malaya introductory accounting course**, *Accounting Education*, vol. 3, no. 4, pp. 331-340, 1994. https://doi.org/10.1080/09639289400000031
15. R. Hake, **Interactive-Engagement Versus Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses**, *American Journal of Physics*, vol. 66, no. 1, 1998.
16. R. R. Sani, J. Zeniarja and A. Luthfiarta, **Penerapan Algoritma K-Nearest Neighbor pada Information Retrieval dalam Penentuan Topik Referensi Tugas Akhir**, *Journal of Applied Intelligent System*, pp. 123-133, 2016.
17. I. A. Amra and A. Y. A. Maghari, **Students Performance Prediction Using KNN and Naïve Bayesian**, in *The 8th International Conference on Information Technology (ICIT)*, Jordan, 2017.
18. T. N. Phyu**, Survey of Classification Techniques in Data Mining**, in *Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS)*, Hongkong, 2009.
19. S. D. Jadhav and H. P. Channe, **Comparative Study of K-NN , Naive Bayes and Decision Tree Classification Techniques**, *International Journal of Science and Research (IJSR)*, vol. 5, no. 1, pp. 1842-1845, 2016.

https://doi.org/10.21275/v5i1.NOV153131

20. B. Patankar and V. Chavda, **A Comparative Study of Decision Tree, Naive Bayesian and K-NN Classifiers in Data Mining**, *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, no. 12, 2014.

21. "**Computer Science Catalog in BINUS Curriculum**," 2012. [Online].

22. S. S. Alfere and A. Y. Maghari, **Prediction of Student's Performance Using Modified KNN Classifiers**, in *1st International Conference on Engineering & Future Technology (ICEFT 2018)*, Gaza, 2018.

23. D. Kabakchieva, **Student Performance Prediction by Using Data Mining Classification Algorithms**, *International Journal of Computer Science and Management Research* , vol. 1, no. 4, pp. 686-690, 2012.

24. R. Ahuja and Y. Kankane, **Predicting the Probability of Student's Degree Completion by Using Different Data Mining Techniques**, in *Fourth International Conference on Image Information Processing (ICIIP)*, India, 2017. https://doi.org/10.1109/ICIIP.2017.8313763

25. C. D. Casuat, E. D. Festijo and A. S. Alon, **Predicting Students' Employability using Support Vector Machine: A SMOTE-Optimized Machine Learning System**, in *International Journal of Emerging Trends in Engineering Research (IJETER)*, vol. 8, no. 5, pp. 2101-2106, 2020. https://doi.org/10.30534/ijeter/2020/102852020

26. D. A. Anggoro and N. D. Kurnia, **Comparison of Accuracy Level of Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) Algorithms in Predicting Heart Disease**, in *International Journal of Emerging Trends in Engineering Research (IJETER)*, vol. 8, no. 5, pp. 1689-1694, 2020. https://doi.org/10.30534/ijeter/2020/32852020