An android-based application to predict student with extraordinary academic achievement

I G I Sudipga*, I N S W Wijaya, M L Radhitya, I M A Mahawan and I N A Arsana
Program Studi Teknik Informatika, STMIK STIKOM Indonesia, Jl. Tukad Pakerisan No.97 Bali, Indonesia

*iwansudipa@stiki-indonesia.ac.id

Abstract. The study period of students who pass the time limit and high numbers of dropout in a college can affect the value of campus accreditation. The anticipation of that possibility, the college must make predictions about potential students don’t graduate on time. This study aims to build a system capable of predicting students who have the potential. If students with unpredictable graduation risks can be identified in the early stages, then the indication of dropout rates may be reduced by providing special appeals to students at risk. Prediction analysis applies the K-Nearest Neighbors method to dig up the trace data stack and look for the proximity of the data with the new data. The test data used student class of 2011 with 100 students as sample data. This method of classification is based on several attributes, namely the evaluation of the 1st semester to semester 6th, the number of GPA, credits that have been taken each semester, number of credits passed, and the number of credits that didn’t pass. The result of classification becomes the output of the system which is then entered into the testing phase. This stage compares the output with the original data with 70.73% accuracy result.

1. Introduction
Dropout (DO) students and past study period affect the assessment of Borang accreditation. The higher the number of students DO and pass the study period the lower value of Borang standard 3 of Higher Education National Accreditation Board. Borang is a form for data and information used to assess the quality of study programs or tertiary institutions. Higher education quality assessment standards are determined by BAN-PT.

The number of students who experience DO or pass the study period needs to be considered. Anticipation to reduce DO students and students through the study period is to make predictions. Predictions aim to find out students who have a risk of not passing on time. If it can be known earlier, the student DO level can be reduced by giving special treatment. Prediction analysis is carried out by classifying students based on GPA variables for 3 years, SKS taken, SKS graduating, SKS which do not pass. Determination of these variables is based on the interview data of a private campus in Bali. The interview was conducted with the head of the Study Program on the campus.

Predictions are done by classifying students according to the previous variables. The method used to group students is K Nearest Neighbor (KNN). Related studies make predictions to determine teaching action plans to suppress dropout [1]. The analysis in this study classifies students who DO based on the learning history taken by the KNN method. In addition, KNN is used to classify student failures in online
courses [2]. Two previous studies used KNN to predict a case by grouping students who have a high risk of failure. The results indicate that KNN can accurately predict student performance. The two studies became the basis for predicting students who had the opportunity to drop out and through the study period using KNN. In this study, the training data used were 18 students graduating through the study period and 82 students graduated from the study period.

2. Literature review
One of the main problems in lectures is the study period and dropout rate. In a large data set, it is certainly very difficult to analyze or predict students who have the potential to drop out. At first, Burgos et al propose knowledge discovery techniques to analyze the data history of students who have the potential to pass the study period in a course [3]. Another study that analyzed dropouts at the University of Nariño, Pasto, Colombia based on a student data repository from 2004-2016 proposed a decision-tree to classify students who could potentially drop out [4]. Another technique that can be used for classification is KNN. KNN (K-Nearest Neighbors) is a classification method that is very efficiently applied in lithology identification, reservoir type recognition, and flow unit classification and so on. However, this method has limitations that are very determined final results in the initial center selection. This method does not work consistently because of inherent limitations, such as initial center selection and center shift during grouping [5]. In [6] which predicts trust between Facebook users based on interaction data, useful information, and the amount of trust between users compared to 3 methods to predict trust, namely MLP, KNN, and SVM. The results obtained by KNN were 73% and better than the results obtained by SVM which was 71% but lower than MLP which reached 83% with a value of k = 12. KNN gets better results than SVM because KNN calculates the output class based on the nearest k-neighbor from the sample data that is considered. The KNN results are very dependent on the value of k, therefore, the value of k must be chosen appropriately.

3. Research data and method
This study uses data obtained from STMIK STIKOM Indonesia. This data is the alumni-student data for the year 2011. At the initial stage, the researchers sorted data from all students in the year 2011. Only 100 students were used in this study. Whereas, 18 students are students who have passed the study period and the rest are students who have not passed the study period. Next, we grouped the data. Then start counting the pivot points in the group that contains students through the study period. This pivot point value is the average value for each parameter of each student. This pivot point will be a benchmark when calculating data distance. The pivot point illustration is shown in Fig. 1.

Figure 1. Data grouping illustration.

Figure 1 illustrates how data is grouped. The minus sign (-) denotes student data that does not exceed the study period. While the plus sign (+) denotes the data of students who passed the study period. A
red area is a group of students who passed through the study period. Data groups that are above the red area state students who have not passed the study period.

3.1. KNN
The results of the previous sorting were used for the KNN process. In the initial stage, for example, there are X that indicate new student data to be predicted. Prediction starts by calculating the distance X to the pivot point using Euclidean Distance. The Euclidean Distance calculation formula is as follows:

\[ D(X, P_v) = \sqrt{\sum_{i=1}^{n} (X_i - P_{vi})^2} \]  

(1)

Where, the distance is denoted by \( D(X, P_v) \). The variable X is the student data to be predicted and \( P_v \) is the pivot point. Each variable \( X \) and \( P_v \) has the same number of parameters as n. Distance calculation starts with calculating the difference from each parameter. Then calculate the square value of each difference that has been obtained. Finally, the value of \( D(X, P_v) \) is obtained by summing all the results of the square value.

Pivot point calculation is obtained from the average value of student data that has passed the study period added with the standard deviation. This pivot point is the threshold to determine the fiction. By looking at the distance from \( X \) to the pivot point, \( X \) can be categorized to students who pass the study period or not. If \( X \) is smaller than the pivot point, prediction \( X \) is the student who passed the study period.

4. Result and Discussion
The testing phase uses the same data. This is done to see the suitability of the method that implemented in the case. The use of the right parameters will be seen in this test. Based on the testing, we get 72.22% of students who have passed the study period and 70.73% for students who have not passed the study period. These results are obtained based on the comparison of predictive data to the original data.

The number and type of parameters that used can affect the value of accuracy. Especially if you use the right parameters. In this study, the prediction accuracy of students who have passed the study period has exceeded 70%. Accuracy measurement is done by calculating prediction accuracy. The exact number of predictions is compared with the amount of original data. Another thing that can affect accuracy is the amount of data. The more data available, the more accurate.

5. Conclusion
According to the calculation processing done, these conclusion as follows:
- KNN method can be used and implemented in prediction system for who exceed the study period
- The disadvantage of the KNN method is that the level of accuracy is determined by the amount of training data. Large data training data requires more calculations that affect the decision process.
- Future work is addition of attributes to the KNN method calculation and adding training data to obtain better prediction accuracy.

We would like thanks to STMIK STIKOM Indonesia, especially the head of study program majoring in informatics engineering and computer system for their contributed in providing comment and explanation to the assessment attributes used in this research.

References
[1] Burgos C, Campanario M L, Peña D de la, Lara J A, Lizcano D and Martinez M A 2018 Data mining for modeling students’ performance: A tutoring action plan to prevent academic dropout Comput Electr Eng. 66 pp 541–556
[2] Tanner T and Toivonen H 2010 Predicting and preventing student failure &ndash; using the k-nearest neighbour method to predict student performance in an online course environment Int J Learn Technol. 5 4 p 356
[3] Burgos C, Campanario M L, De D, Lara J A, Lizcano D and Martínez M A 2017 *Action plan to prevent academic dropout* pp 1–16

[4] Timaran Pereira R and Caicedo Zambrano J 2018 Application of decision trees for detection of student dropout profiles *Proc - 16th IEEE Int Conf Mach Learn Appl ICMLA 2017* pp 528–531

[5] Wang X, Yang S, Zhao Y and Wang Y 2018 *Journal of Petroleum Science and Engineering* Lithology identification using an optimized KNN clustering method based on entropy-weighted cosine distance in Mesozoic strata of Gaoqing field, Jiyang depression *J Pet Sci Eng.* Elsevier Ltd 166 pp 157–174

[6] Khadangi E and Bagheri A 2013 Comparing MLP, SVM and KNN for predicting trust between users in Facebook *Proc 3rd Int Conf Comput Knowl Eng ICCKE 2013* pp 466–470