Prediction of Student Study Period Based on Admission Pathways Using Support Vector Machine Algorithm

Cut Maya Putri Audilla, Slamet Riyadi, and Asroni*

Universitas Muhammadiyah Yogyakarta, Jln.Brawijaya, Tamantirto, Kasihan, Bantul, Yogyakarta 55183, Indonesia
**Corresponding author: sroni@umy.ac.id*

Abstract

In Indonesia, the quality of a university is measured based on the accreditation by BAN-PT (National Accreditation Board for Higher Education). BAN-PT possesses several main standards in measuring the quality of a university, one of which is students and graduates. The accuracy of the student study period is a crucial issue because it is the basis for the effectiveness of a university. Prediction is a process of systematically estimating something most likely to happen in the future based on past and present information to minimize the error (difference between something that happens and the forecast results). One technique used to make predictions is data mining. Universitas Muhammadiyah Yogyakarta (UMY), as one of the best private universities in Indonesia, must maintain the quality of its students. Student admission at UMY is an internal selection carried out through several methods: student achievement and academic ability tests. The Support Vector Machine (SVM) method is part of the prediction method. Analysis of the SVM prediction utilized the historical data from alumni of the Faculty of Law of UMY in the graduation year of 2015-2019. The application of SVM has provided better accuracy, precision, and recall results. The best kernel accuracy level was the SVM RBF kernel with an optimum C value of 10 and a gamma value of 0.4 with an accuracy of 96.00%.

Keywords: Prediction, Study Period, Student, Support Vector Machine

1. Introduction

In Indonesia, the quality of a university is measured based on the accreditation carried out by BAN-PT (National Accreditation Board for Higher Education). BAN-PT [1] measures a university quality based on several main standards. One of which is students and graduates with several assessed components: the admission of new students and graduates (GPA and average study period). In other words, the quality of a university is determined by the admission of new students and the study period of its students [2].

One of the problems frequently occurring in a higher education institution is how to improve the quality of its students. To achieve the best quality, it is essential to predict the student study period to see whether the determining factor in the accuracy of graduation is based on admission. If the student study period can be predicted, the handling of students will be more effective. Prediction systematically estimates something most likely to happen in the future based on past information to minimize the error (the difference between something happening and the predicted results).

One technique commonly used to make predictions is data mining. Data mining (data development) is often referred to as knowledge discovery in databases (KDD), referring to the discovery of information originating from large data sets[3], [4]. Data mining is currently increasingly being employed for data in the education sector to obtain useful information. The information is generally used as a basis for improving the quality of higher education.

Being one of the best private universities in Indonesia, Universitas Muhammadiyah Yogyakarta (UMY) must maintain the quality of its students. Student admission at UMY is an internal selection carried out in various ways, such as student achievement and academic ability tests. The Faculty of Law is one of the best faculties at UMY. Unfortunately, it has not been able to predict the influence of the admission factor on the study period. This factor was selected to determine the student performance based on admission pathways, whether they impact their study period. Thus, information is highly required to improve the quality of the faculty.

In processing student data for predicting the study period, many methods have been previously applied, such as Naive Bayes, C4.5 algorithm, K-Nearest Neighbor (K-NN) algorithm, Neural Network, Random Forest, and Support Vector Machine (SVM) [5]–[7]. SVM is a method with fairly high accuracy in predicting the potential classification of academic data. It is proven that, in many implementations by several previous studies, SVM has provided better results than other algorithms.

Based on previous research, no one has predicted the study period based on admission pathways with the SVM algorithm.

2. Method

Research methods are steps and procedures carried out to achieve research goals and answer research problems. These steps and procedures are the embodiment of the research framework. In the case of this prediction, a method or model is required to produce a data classification pattern with the ultimate goal of prediction. The technique or method used to discover the information is how to process data mining on student academic data. Data mining consists of various techniques used to make predictions and classifications, where these techniques can estimate the possibilities by looking at some information and existing data patterns [8], [9].

A. Data Collection

This study employed secondary data obtained from the Bureau of Information Systems (BSI) of UMY containing data on alumni of the Faculty of Law in the 2015 to 2019 graduation years, amounting to 1,220 with 16 attributes: student number, class, admission pathways, name, gender, origin regency, province, year of entry, yudisium, graduation, the number of semesters taken, first semester GPA, second semester GPA, third semester GPA, fourth semester GPA, and GPA.

B. Preprocessing

After collecting and creating the dataset, data mining preprocessing was performed with the following stages:

1. Data selection refers to the process of analyzing relevant data from the database because not all data are necessary for the data mining process. Therefore, data selection is highly needed to determine the required data. The data obtained from

the BSI were in the form of a database of student data of the Faculty of Law with attributes: student number, class, admission pathways, name, gender, origin, regency, province, year of entry, yudisium, graduation, number of semesters taken, GPA of the first to the fourth semester, and GPA. This study utilized several related attributes filtered and selected, encompassing gender, admission pathways, GPA of the first to the fourth semester, GPA, number of semesters taken, and adding the APPROPRIATE and INAPPROPRIATE description attribute. The authors decided to use the attribute number of semesters taken, starting from the seventh semester, because it is the ideal time for students to start graduating. Specifically, APPROPRIATE is for students with a study period of seven to eight semesters, and INAPPROPRIATE is for students who took nine to 12 semesters to graduate.

2. Data cleaning is the process of analyzing the data quality by changing, correcting, or deleting incorrect, incomplete, and inaccurate data or those with the wrong format in the database to produce high-quality data [10]. After processing the data, the data cleaning process was conducted, requiring a fairly long process. To obtain the best results, it is necessary to clean data that can later interfere with the prediction process. The authors utilized Microsoft Excel to perform data cleaning. The data cleaned were data noise or empty and outlier.
3. Data transformation is the stage to change the data with the appropriate form to be processed in data mining. In this study, the data were processed from Microsoft Excel 2016 and used for processing on the RapidMiner software. The authors performed the transformation and initialization of data to make this research run well. The data initialized encompassed the attributes of GPA of the first to the fourth semester and GPA.

C. Research Trial

Research trial is the primary stage in implementing data mining in finding information from the data. The technique used in data mining was prediction using the Support Vector Machine (SVM) algorithm and adding the n-Folds Cross-Validation technique as a cross-validation method. This method divided the data into two parts: training data and testing data. Furthermore, after testing the data, a cross-processing was carried out where the testing data were then used as training data and vice versa. The previous training data were converted into testing data. The next stage was to determine the kernel value through several kernel functions. The kernel functions comprised Poly, RBF, and Sigmoid. The difference between the three kernel functions lies in the mapping into feature space, and each has advantages and disadvantages in each case. Hence, it is necessary to conduct experiments in finding the best kernel function to use in a case [11].

Each kernel function was adjusted to each parameter value required and determined by the researchers. In adding parameter values, it is necessary to note that the value of C must be greater than zero to make the range of parameter C in the interval $(0,\infty)$. In this study, only three values were taken from that interval. For the parameter, the values should also be greater than zero to make the range in the interval $(0,\infty)$. Therefore, this study utilized only three values taken from the interval and each parameter value in each kernel function was the interval value.

The determination of parameter values for each kernel function to obtain optimal results is displayed in Table 1.

Table 1. Used SVM hyperparameters for the research study.

Used hyperparameter	Parameters and its value
Poly	$\gamma = 1$ $\mathrm{C} = 1$
RBF	$\gamma = 0.4$ $\mathrm{C} = 10$
Sigmoid	$\gamma = 0.2$ $\mathrm{C} = 100$

The determination of the values of C and γ above was based on previous studies: the research of Huang, Hung, Lee, Li, and Jiang using C = {10, 50, 100}, γ = {2.4, 5, 10} and C = {5, 10, 50}, γ = {0.08, 4, 11}, Erfanifard, Behnia and Moosavi's research using C = {100, 200, 300} and γ = {0.2, 0.3, 0.4}, and Rusydina's research using C = {0.25, 0.50, 0.75, 1, 2, 3, 4} and γ = {0.005, 0.05, 0.1, 0.15}. The researchers made a slight adjustment in selecting the possible values used in this study.

After obtaining the results from the n-Folds and the best kernel functions, the best kernel results were evaluated using classification evaluation measurements encompassing recall, precision, and accuracy measurements. Tests were carried out using the Rapidminer Studio platform.

3. Results and Discussion

A. Data Collection Results

In this study, the 2015 to 2019 alumni data totaling 1,220 were obtained from the BSI of UMY. However, the data used were only 990 divided into four datasets based on the admission pathways: CBT, PBT, PMDK, and PNUAN. Each dataset had a large amount of data, as illustrated Figure 1.

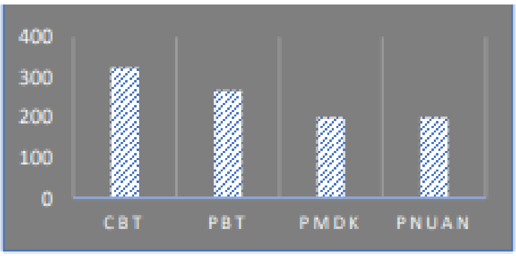

Figure 1. The number of university applicant based on admission types.

The CBT dataset consisted of 324 data, the PBT dataset comprised 267 data, the PMDK dataset encompassed 200 data, and PNUAN contained 200 data.

B. Preprocessing Results

1. The data selection stage was carried out by selecting data following the research variables determined in the research method. The information from the selected attributes has represented the information required to be an indicator of research. The data used comprised gender, admission pathways, GPA of the first to the fourth semester, GPA, number of

semesters taken, and information containing APPROPRIATE and INAPPROPRIATE.

2. At the data cleaning stage, the irrelevant and inconsistent data such as outlier or noise and incomplete were removed. The attribute value was lacking or null.
3. At the data transformation stage, several data were converted into a form suitable for processing in data mining. The data changed were semester GPA and GPA and adjusted to the semester grade standards and final grades of UMY.

C. Research Trial Results

The training and testing aim to obtain the most optimal kernel function to be applied in predictive modeling of the study period based on the admission pathways in this study. The kernel functions used in this test scenario included the Poly, Radial Basis Function (RBF), and Sigmoid. The test results for calculating the accuracy of calcification were based on the accuracy calculation of each kernel function using the calculation of 5-fold cross validity. Table 2 displays the results of calculating the percentage accuracy of the SVM implementation trial with its hyperparameters.

Table 2. Prediction accuracy in different SVM hyperparameters.

Admission type	Used Hyperparameter		
	Poly	RBF	Sigmoid
CBT	92.60%	94.45%	91.37%
PBT	92.10%	94.36%	88.00%
PMDK	87.50%	86.00%	83.00%
PNUAN	95.50%	96.00%	93.50%

Table 2 depicts that the prediction of the study period based on the admission pathways with SVM on various kernel functions had fairly good accuracy. From the accuracy calculation of the 5-fold cross-validation, the best accuracy result was in the PNUAN dataset. These three datasets of kernel functions had an excellent level of accuracy: Poly (95.50%), RBF (96.00%), and Sigmoid (93.50%).

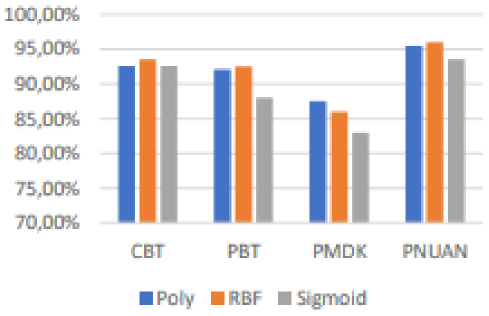

Figure 2. Prediction accuracy in bar graphic.

The graph in Figure 2 explains that the kernel function performance test implemented in predicting the student study period with the admission pathways using the SVM obtained the most optimal results using the RBF kernel with a C parameter value = 10 and

γ = 0.4. The test results could be used as evaluation material for the study period for students of Law Science of the Faculty of Law of UMY.

4. Conclusions

Following the analysis results, the prediction model using the SVM method with several input data variables could produce predictions for the study period based on the student admission pathways. The CBT pathway tended to graduate more on time than others. In terms of prediction accuracy, more student data had positive prediction errors (false positive), and only a few students had negative prediction errors (false negative). In a nutshell, the system built with the SVM method was relatively adequate. The most optimal kernel function used as a companion to the SVM method that could produce the best accuracy was the RBF, with a parameter value of C of 10 and γ of 0.4, with an accuracy of 96%.

5. Suggestions

Future researchers are expected to discover the biggest problems causing students not to graduate on time to obtain solutions to improve study accuracy. It is necessary to compare other optimization methods as a comparison in determining the most appropriate optimization method.

References

[1] I. G. P. Purnaba and Dwiwahju Sasongko, *BUKU III BORANG AKREDITASI "BOOK III ACCREDITATION FORM GUIDANCE."* 2017.

[2] J. F. Ulysses, "Data Mining Classification Untuk Prediksi Lama Masa Studi Mahasiswa Berdasarkan Jalur Penerimaan Dengan Metode Naïve Bayes 'Data Mining Classification for Predicting the Length of Student Study Period Based on the Admission Path Using the Naïve Bayes Method,'" *Magister Teknik Informatika Universitas Atma Jaya Yogyakarta*, 2012.

[3] S. Guha and S. Kumar, "Emergence of big data research in operations management, information systems, and healthcare: Past contributions and future roadmap," *Production and Operations Management*, vol. 27, no. 9, pp. 1724–1735, 2018.

[4] M. S. Islam, M. M. Hasan, X. Wang, and H. D. Germack, "A systematic review on healthcare analytics: application and theoretical perspective of data mining," in *Healthcare*, 2018, vol. 6, no. 2, p. 54.

[5] H.-C. Chen *et al.*, "Pulse-line intersection method with unboxed artificial intelligence for hesitant pulse wave classification," *Information Processing & Management*, vol. 59, no. 2, p. 102855, Mar. 2022, doi: 10.1016/j.ipm.2021.102855.

[6] K. R. Pradeep and N. C. Naveen, "Lung cancer survivability prediction based on performance using classification techniques of support vector machines, C4. 5 and Naive Bayes algorithms for healthcare analytics," *Procedia computer science*, vol. 132, pp. 412–420, 2018.

[7] M. Jupri and R. Sarno, "Taxpayer compliance classification using C4. 5, SVM, KNN, Naive Bayes and MLP," in *2018 International Conference on Information and Communications Technology (ICOIACT)*, 2018, pp. 297–303.

[8] T. Ahmad and H. Chen, "Short and medium-term forecasting of cooling and heating load demand in building environment with data-mining based approaches," *Energy and Buildings*, vol. 166, pp. 460–476, 2018.

[9] K. Meenakshi, G. Maragatham, N. Agarwal, and I. Ghosh, "A Data mining Technique for Analyzing and Predicting the success of Movie," in *Journal of Physics: Conference Series*, 2018, vol. 1000, no. 1, p. 012100.

[10] H. A. Sulistyo, T. F. Kusumasari, and E. N. Alam, "Implementation of Data Cleansing Pattern Module for Data Quality Management Application using Open Source Tools," in *2020 3rd International Conference on Computer and Informatics Engineering (IC2IE)*, 2020, pp. 7–12.

[11] A. Holzinger *et al.*, "Interactive machine learning: experimental evidence for the human in the algorithmic loop," *Applied Intelligence*, vol. 49, no. 7, pp. 2401–2414, 2019.