Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 5, Number 1, October 2020
DOI : https://doi.org/10.33395/sinkron.v5i1.10621

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Period Study Accuracy Prediction using Sequential Minimal Optimization Algorithm

Hendri Noviyanto[1)*], Bayu Mukti[2)]
[1)2)]Surakarta University, Indonesia
[1)]hendrinoviyantoo@gmail.com, [2)]bayu.unsa@gmail.com

Submitted : Sep 9, 2020 | **Accepted** : Oct 15, 2020 | **Published** : Oct 15, 2020

Abstract: The study period is quite influential in the assessment of a university. The imbalance in the ratio of students to lecturers causes the quality of teaching and learning to decline, this is because one lecturer has to manage many students. Acquisition of accreditation scores and society's assumptions about higher education are also strongly influenced by the number of student graduations on time. Therefore, the prediction of the accuracy of the study period is needed as consideration for related parties to solve the problem of student learning delay. Sources of data in this study were taken from a database stored at the University of Surakarta, namely the Temporary Achievement Index with data of 209 instances and 5 attributes. The proposed method in this study is the Sequential Minimal Optimization algorithm. The validation method uses k-fold Cross-Validation with a value of K = 10. This method is compared with other methods such as naive Bayes, KNN, and Decision Tree. The results of this study, the proposed method can predict the accuracy of the study period quite well with the acquisition of accuracy of 88.52%. However, several other methods such as NaiveBayes obtained better accuracy of 90.91%, KNN of 91.86%, and Decision Tree of 96.65%. From the results of the comparison of these methods, the Decision Tree obtained the highest accuracy value. In future studies, researchers aim to enrich features in the prediction process. These features are related to student activities, such as student backgrounds, social activities, additional activities on campus and off-campus, and other aspects.

Keywords: classification; data mining; period study; prediction; smo;

INTRODUCTION

Academic data of higher education stores various kinds of data that can be processed into important information. One of the information that can be obtained from the dataset is the student's study period. To obtain information, it is necessary to process academic data with data mining techniques knowledge discovery of that data. Data mining is a series of data extraction processes by looking for patterns recognition that draw from a fairly large amount of data that must fulfill (Witten et al., 2011).

The student's study period is quite influential on the credibility of a College. According to a circular from (BAN-PT, 2019) regarding accreditation, it is stated that one of the elements of higher education assessment is the ideal educational efficiency figure. Therefore, predicting the study period of students is very important to maintain the stability of the teaching and learning process in a university. The period study student's that is taken too long can cause students to drop out and bad image for the institution and students.

Prediction is a classification technique used to find models that describe and differentiate data classes or concepts that aim to be used to predict the class of objects whose class labels are unknown (Agarwal, 2014). The process of predicting the accuracy of a student's study period can be used to evaluate student conditions. So, the higher education institution can anticipate students who have the potential to be late in completing their studies.

Research related to the field of classification for the prediction process has been widely carried out. This process aims to obtain information about the accuracy of the student's study period. Several classification methods have been proposed such as Naïve Bayes (Mauriza & Nugroho, 2014; Salmu & Solichin, 2017), K-Nearest Neighbor (Saputra & Primadasa, 2018; Susanto & Fatta, 2018), Decision Tree (Mashlahah, 2013; Rohmawan, 2018), and Apriori algorithm (Kurniawan & Nurjoko, 2016) to get high accuracy values. (Kabakchieva, 2013)The classification process uses numerical values to determine the superiority of the method used. The measured values are like the results of accuracy.

* Corresponding Author

LITERATURE REVIEW

Research (Mutiara, 2015) applies K-Optimal to the KNN algorithm to predict graduation on time. According to Mutiara et al, the k value in the KNN algorithm greatly affects the accuracy value that will be obtained. By using the value of K = 5, an accuracy of 80% has been obtained. K-Nearest Neighbor has the advantage of being able to classify objects with the closest distance. Research (Prasetyo et al., 2016) using the KNN algorithm gets the best value at K = 100 with an accuracy value obtained reaching 97.90%. Research (Susanto & Fatta, 2018) can obtain accuracy values of 98.46% in predicting student graduation. Susanto et al. Used the value of K = 14 and k-fold = 5 so that they were able to produce a fairly good performance. KNN is known to have a fairly good performance, in research (Saputra & Primadasa, 2018) predicting the graduation of STMIK Bina Nusantara students with semester V student data using the parameter K = 5. According to Saputra and Primadasa, the KNN method is strongly influenced by the amount of training data used.

Research (Kurniawan & Nurjoko, 2016) uses a priori algorithm to predict student graduation rates. In this study, the graduation category was measured using study time and GPA by looking for support and confidence values. Research (Rohmawan, 2018) compared accuracy with the Decision Tree and Artificial Neural Network methods. The accuracy value obtained in the Decision Tree algorithm is 74.51% while the Neural Network is 79.74%. Research (Putri & Waspada, 2018) applies the C4.5 algorithm to predict student graduation. This research resulted in an accuracy value of 60.52%. After the decision tree is cut using error-based pruning the accuracy value becomes61.57% and when cut using a confidence value the accuracy value becomes 62.44%.

In several studies that have been conducted by previous researchers, the KNN algorithm has proven to have a fairly good performance in classifying information about student graduation. In this study, researchers proposed a Sequential Minimal Optimization (SMO) algorithm for the predictionstudy period. SMO is an algorithm that works by performing a problem-solving process (dual form) only contains a value of α which is different from other algorithms. By using analytic quadratic programming as its inner loop, this algorithm can solve problems with how to use two pieces of data in each process. So, the process of finding the optimal solution can be complete.

METHOD

In this research, there are some steps taken, starting from the process of preprocessing the dataset used as input, determining the training and testing data, the classification process, and the analysis of the results. The research flow diagram can be seen in Fig 1 below.

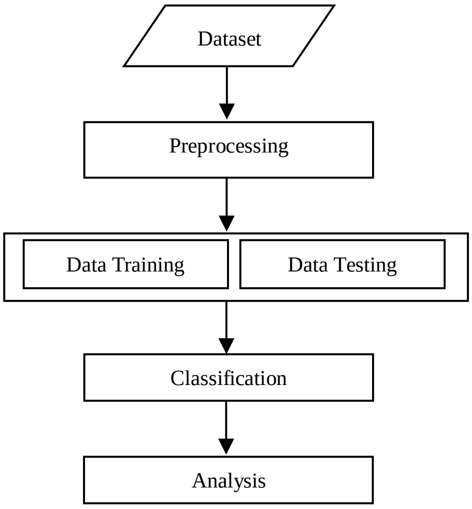

Fig.1 .Research flow

Based on Fig 1above, the research flow can be explained as follows:

Dataset

This study uses some datasets obtained from the University database. The number of instances is 209 and 4 attributes.

Preprocessing

At the preprocessing stage, the dataset is processed to get data that is clean from noise and by the input format so that it can be used as training and testing data. At this stage, the step taken is to separate the data and give class labels to the data that has been obtained by the provisions.

Classification

The classification process is carried out 3 times. This is done to get the value of each method applied. The first process is to classify the dataset using 10 fold Cross-validation. The second process is to classify the initial dataset (training data) with new data (testing data). The proposed model can be seen in Fig 2. The classification process is carried out using the proposed SMO method and the comparison method NB, Decision Tree, and KNN.

Analysis

This process is carried out after the entire testing process is complete to gain knowledge of whether the proposed method can solve the problem of predicting the accuracy of the student's study period.

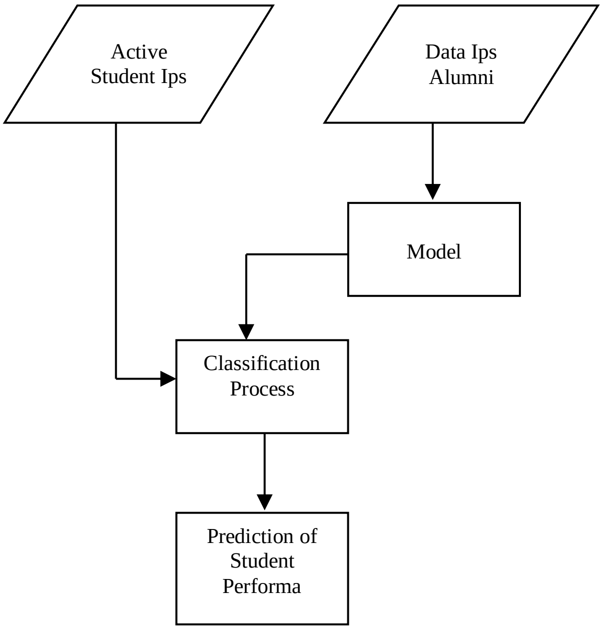

Fig.2 .Proposed Model

Sequential Minimal Optimization Algorithm

Sequential Minimal Optimization (SMO) is an optimization of the Support Vector Machine (SVM) algorithm. SMO works simply and solves the quadratic programming (QP) problem into the quadratic programming sub-problem.SMO fixes the shortcomings of SVM by solving QP problems in the SVM algorithm without using additional matrix space and without repeating the same numeric value for each subproblem. The way SMO works is not like the predecessor method, SMO will solve small problems that are possible to solve at every step. SMO involves two Lagrange multipliers for optimization on QP problems. SMO is a decomposition method that works based on the principle of a two-element working set and works analytically, these results in a very large number of iterations, but because each iteration is quite small, the total time required is also shorter.(Santosa, 2011). The working set principle works by changing the multiplier α_i in a certain number at each iteration. The thing that must be considered in the decomposition method so that it runs faster is the selection of a working set which causes global problems to be quick achieved.

The value entered into the working set is the variable that most violates the KTT condition; the following is the equation of SMO (3-1) and KTT condition (3-2).

$$max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_, K(x_i, x_j) \alpha_i \alpha_j, \qquad (3\text{-}1)$$

Must fulfill :

$$K = (\vec{x}_i, \vec{x}_j) = \Phi(\vec{x}_i).\Phi(\vec{x}_j)$$

$$\sum_{i=1}^{n} y_i \alpha_i = 0$$

$$\begin{gathered} \alpha_i = 0 \Longrightarrow y_i f(\vec{x}_i) \geq 1, \\ 0 < \ \alpha_i < C \Longrightarrow y_i f(\vec{x}_i) = 1, \\ \alpha_i = C \Longrightarrow y_i f(\vec{x}_i) \leq 1, \end{gathered} \qquad (3\text{-}2)$$

RESULT

This study uses the WEKA machine learning application (Garner, 1995) to predict the accuracy of student learning. The test scheme in this study is to test the dataset using the proposed method, namely Sequential Minimal Optimization. The validation method used is cross-validation with a value of K = 10. The validation method is used to share training data and testing data. The dataset used comes from the Surakarta University database with a total of 209 instances and 4 attributes, namely the Temporary Achievement Index from semester 1 (one) to semester 4 (four). Testing the proposed method will be compared with other methods such as Naïve Bayes (NB), K-Nearest Neighbor (KNN), and Decision Tree (J48).

Table 1
Classification Result

	Algorithm			
	SMO	***NB***	***KNN***	***J48***
Accuracy	88.52%	90.91%	91.86%	96.65%

In Table 1 above, it can be seen that the accuracy value obtained by using the proposed method is 88.52%, Naïve Bayes 90.91%, K-Nearest Neighbor 91.86%, and 96.65% Decision Tree. Seeing the accuracy results that have been shown in Table 1, the Decision Tree can predict better than other methods.

In Table II, the assessment process using the Confusion Matrix shows the results of predicting new data (testing) using the test set technique. This technique uses a model that has been obtained from the previous prediction process. The new data provided consists of 3 data, namely data with 19 instances and 4 attributes, data with 4 instances and 4 attributes, and data with 1 instances with 4 attributes.

Table 2
Confusion Matrix

20 Instance							
SMO		NB		J48		KNN	
a	b	a	b	a	b	a	b
11	0	11	0	11	0	11	0
0	8	0	8	0	8	0	8
4 Instance							
2	0	2	0	2	0	2	0
1	1	1	1	1	1	1	1
1 Instance							
1	0	1	0	1	0	1	0
0	0	0	0	0	0	0	0

The results shown in Table 2 indicate that all the methods used can predict data well. However, the results will be different when the data usage becomes larger because it is related to the quality of the data inputted.

Table 3
Dataset

IPs-1	IPs-2	IPs-3	IPs-4	Prediction
2.9	2.9	2.9	3	Yes
3.0	3.31	3.23	3.4	No

Table 3 is an example of a dataset processed by machine learning. The researcher tries to make an input class on the dataset with the wrong value. The results given are quite good, namely, the model that has been created can be used to predict data well by distinguishing between correct and incorrect inputs.

DISCUSSIONS

In this study, we used the Sequential Minimal Optimization method as a proposed method. This method is compared with other methods that have been used for the prediction process such as NaiveBayes, K-Nearest Neighbor, and Decision Tree. In this study, the accuracy value obtained by the Sequential Minimal Optimization method obtained 88.52% (see table 1). The accuracy value when compared with other methods is the lowest. However, when we try to evaluate the model we get with less new data. All models are able to provide fairly good performance. Table 2 shows the ability of several of these methods in predicting new input data and pouring it into a confusion matrix table. The result is that all methods are able to work optimally. This is supported by other studies that are shown in table 3. We deliberately make the input data have errors to see whether the model that has been created is able to evaluate properly. From this research, it can be concluded that the model works well as evidenced by the correct prediction process results.

CONCLUSION

The process of predicting the accuracy of the student's study period in this study will be developed using student personal data. In future studies, researchers aim to enrich features in the prediction process. These features are related to student activities, such as student backgrounds, social activities, additional activities on campus and off-campus, and other aspects. Researchers are aware that many aspects can affect the performance of students to achieve the target of graduating on time with good grades. With a variety of features to be used, it is hoped that the results will be better.

ACKNOWLEDGMENT

Thanks to Surakarta University for providing student data and LP2M for facilitating it. Thanks to the Directorate of Research and Community Service for funding this research.

REFERENCES

Agarwal, S. (2014). Data mining: Data mining concepts and techniques. In *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*. https://doi.org/10.1109/ICMIRA.2013.45

BAN-PT. (2019). *Instrumen Akreditasi Perguruan Tinggi*. *April*, 7–9.

Garner, S. R. (1995). WEKA: The Waikato Environment for Knowledge Analysis. *Proc New Zealand Computer Science Research Students Conference*, 57–64.

Hamsa, H., Indiradevi, S., & Kizhakkethottam, J. J. (2016). Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm. *Procedia Technology*, *25*, 326–332. https://doi.org/10.1016/j.protcy.2016.08.114

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies*, *13*(1), 61–72. https://doi.org/10.2478/cait-2013-0006

Kurniawan, H., & Nurjoko. (2016). *Aplikasi Datamining Untuk Memprediksi Tingkat Kelulusan Mahasiswa Menggunakan Algoritma Apriori Di Ibi Darmajaya Bandar Lampung*. *2*(01), 79–93.

Mashlahah, S. (2013). *Prediksi Kelulusan Mahasiswa Menggunakan Metode Decision Tree Dengan Penerapan Algoritma C4.5*.

Mauriza, A. F., & Nugroho, Y. S. (2014). *Implementasi Data Mining untuk Memprediksi Kelulusan Mahasiswa Fakultas Komunikasi dan Informatika UMS Menggunakan Metode Naive Bayes* [Universitas Muhammadiyah Surakarta]. https://doi.org/10.1038/132817a0

Mutiara, I. dan A. (2015). Penerapan K-Optimal Pada Algoritma Knn Untuk Prediksi Kelulusan Tepat Waktu Mahasiswa Program Studi Ilmu Komputer Fmipa Unlam Berdasarkan Ip Sampai Dengan Semester 4. *Klik - Kumpulan Jurnal Ilmu Komputer*, *2*(2), 159–173. https://doi.org/10.20527/KLIK.V2I2.26

Prasetyo, T. F., Susandi, D., & Widianingrum, I. S. (2016). Prediksi Kelulusan Mahasiswa Pada Perguruan Tinggi Kabupaten Majalengka Berbasis Knowledge Based System. *Seminar Nasional Telekomunikasi Dan Informatika | Vol: | Issue : | 2016*, *May*.

Putri, R. P. S., & Waspada, I. (2018). Penerapan Algoritma C4.5 pada Aplikasi Prediksi Kelulusan Mahasiswa Prodi Informatika. *Khazanah Informatika: Jurnal Ilmu Komputer Dan Informatika*, *4*(1), 1. https://doi.org/10.23917/khif.v4i1.5975

Rohmawan, E. (2018). Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Desicion Tree Dan Artificial Neural Network. *Jurnal Ilmiah Matrik*, *20*(1), 21–30.

Salmu, S., & Solichin, A. (2017). *Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan Naïve Bayes : Studi Kasus UIN Syarif Hidayatullah Jakarta Prediction of Timeliness Graduation of Students Using Naïve Bayes : A Case Study at Islamic State University Syarif Hidayatullah Jakarta*. *April*, 701–709.

Santosa, B. (2011). *Tutorial Support Vector Machine 1 Ide Dasar Support Vector Machine*. 1–23. https://doi.org/10.1016/S0924-0136(01)00706-3

Saputra, A. Y., & Primadasa, Y. (2018). Penerapan Teknik Klasifikasi Untuk Prediksi Kelulusan Mahasiswa

Menggunakan Algoritma K-Nearest Neighbor. *Techno.Com*, *17*(4), 395–403. https://doi.org/10.33633/tc.v17i4.1864

Susanto, E. S., & Fatta, H. Al. (2018). Informatika Universitas Amikom Yogyakarta Menggunakan Metode K-Nearest Neighbor. *Jurnal Teknologi Informasi*, *XIII*, 67–72.

Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining. In *Encyclopedia of Ecology, Five-Volume Set* (Third Edit). Elsevier. https://doi.org/10.1016/B978-008045405-4.00153-1