IMPLEMENTATION OF K-NEAREST NEIGHBOR AND GINI INDEX METHOD IN CLASSIFICATION OF STUDENT PERFORMANCE

Tyas Setiyorini[1]; Rizky Tri Asmono[2]

Informatics Engineering [1]
STMIK Nusa Mandiri Jakarta, Indonesia[1];
http://nusamandiri.ac.id[1]
tyas.setiyorini@gmail.com[1]

Informatics Engineering [2]
STMIK Swadharma, Jakarta, Indonesia[2]
http://swadharma.ac.id[2]
rtriasmono@gmail.com[2]

The work is distributed under the Creative Commons Attribution-Non-Commercial International 4.0 License.

Abstract**—***Predicting student academic performance is one of the important applications in data mining in education. However, existing work is not enough to identify which factors will affect student performance. Information on academic values or progress on student learning is not enough to be a factor in predicting student performance and helps students and educators to make improvements in learning and teaching. K-Nearest Neighbor is a simple method for classifying student performance, but K-Nearest Neighbor has problems in terms of high feature dimensions. To solve this problem, we need a method of selecting the Gini Index feature in reducing the high feature dimensions. Several experiments were conducted to obtain an optimal architecture and produce accurate classifications. The results of 10 experiments with values of k (1 to 10) in the student performance dataset with the K-Nearest Neighbor method showed the highest average accuracy of 74.068 while the K-Nearest Neighbor and Gini Index methods showed the highest average accuracy of 76.516. From the results of these tests it can be concluded that the Gini Index is able to overcome the problem of high feature dimensions in K-Nearest Neighbor, so the application of the K-Nearest Neighbor and Gini Index can improve the accuracy of student performance classification better than using the K-Nearest Neighbor method.*

Keywords: *K-Nearest Neighbor, Gini Index, Student Performance*

Intisari—Memprediksi kinerja akademik siswa adalah salah satu aplikasi penting dalam data mining bidang pendidikan. Namun, pekerjaan yang ada tidak cukup untuk mengidentifikasi faktor mana yang akan mempengaruhi kinerja siswa. Informasi nilai akademik atau kemajuan pembelajaran siswa saja tidak cukup untuk dijadikan faktor dalam memprediksi kinerja siswa serta membantu para siswa dan pendidik untuk melakukan perbaikan dalam pembelajaran dan pengajaran. K-Nearest Neighbor merupakan metode yang sederhana untuk klasifikasi kinerja siswa, namun K-Nearest Neighbor memiliki masalah dalam hal dimensi fitur yang tinggi. Untuk menyelesaikan masalah tersebut diperlukan metode seleksi fitur Gini Index dalam mengurangi dimensi fitur yang tinggi. Beberapa percobaan dilakukan untuk mendapatkan arsitektur yang optimal dan menghasilkan klasifikasi yang akurat. Hasil dari 10 percobaan dengan nilai k (1 sampai dengan 10) pada dataset *student performance* dengan metode K-Nearest Neighbor didapatkan rata-rata akurasi terbesar yaitu 74,068 sedangkan dengan metode K-Nearest Neighbor dan Gini Index didapatkan rata-rata akurasi terbesar yaitu 76,516. Dari hasil pengujian tersebut maka dapat disimpulkan bahwa Gini Index mampu mengatasi masalah dimensi fitur yang tinggi pada K-Nearest Neighbor, sehingga penerapan K-Nearest Neighbor dan Gini Index dapat meningkatkan akurasi klasifikasi kinerja siswa yang lebih baik dibanding dengan menggunakan metode K-Nearest Neighbor saja.

Kata Kunci: K-Nearest Neighbor, Gini Index, Kinerja Siswa

INTRODUCTION

Predicting student academic performance is one of the important applications in data mining in education (Altujjar, Altamimi, Al-Turaiki, & Al-Razgan, 2016). Student performance prediction systems at an early stage can be very useful to guide student learning. Predicting student performance can help identify weak students (Pandey & Taruna, 2016) and enable academic institutions to provide appropriate support for students who face difficulties(Altujjar et al., 2016). In order for the prediction model to be truly useful as an effective aid for learning, the prediction model must provide a tool to interpret progress adequately, to detect trends and patterns of behavior and to identify the causes of learning problems (Villagrá-Arnedo et al., 2017). However, there is not enough work to identify which factors will influence its performance, in which ways students can make progress, and whether students have the potential to do better (Yang & Li, 2018). Student academic information becomes one of the factors of an educator's assessment in predicting student performance, but the academic value factor alone is not sufficient in predicting student performance. Information on student learning progress is not enough as an indicator of students and educators to make improvements in teaching and learning (Yang & Li, 2018). Social, personal and academic factors also influence in predicting student performance in schools (Fernandes et al., 2019).

The most popular technique for predicting student performance is data mining(Shahiri, Husain, & Rashid, 2015). Classification is a technique that is widely used to predict student performance (Altujjar et al., 2016) Several studies with classification techniques have been carried out, such as Artificial Neural Networks (Alkhasawneh & Hobson, 2011), Regression (Conijn, Snijders, Kleingeld, & Matzat, 2017), Support Vector Machine (Al-Shehri et al., 2017), Decision Tree (Lopez Guarin, Guzman, & Gonzalez, 2015), Naive Bayes (Lopez Guarin et al., 2015), dan K-Nearest Neighbor (Pandey & Taruna, 2016).

The K-Nearest Neighbor classification is a well-known pattern recognition method that has been used extensively in several applications (Cover & Hart, 1967) and has attracted wide interest in the research community(Gou et al., 2014) (Lin, Li, Lin, & Chen, 2014)(Lin et al., 2014). K-Nearest Neighbor is a method that is able to solve classification problems, has significant advantages and often produces competitive results from several other data mining methods(Adeniyi, Wei, & Yongquan, 2016). The simplicity of the K-Nearest Neighbor is its main virtue, which allows the classification of two or more patterns based on fairly simple rules(Han, Kamber, & Pei, 2012).

K-Nearest Neighbor is a simple method but because of its simplicity, the k-NN method has several problems that must be faced, the main problem is related to the high dimension of features (López & Maldonado, 2018). K-Nearest Neighbor also has several drawbacks, namely the complexity of computing the similarity of large data. To reduce the complexity of K-Nearest Neighbor can be done by one method, namely by reducing the dimensions of high features (de Vries, Mamoulis, Nes, & Kersten, 2003). High feature dimensions are not permitted for many learning algorithms (Shang et al., 2007). Dimension reduction is very important in pattern formation (López & Maldonado, 2018).

The main problem in classification is that high feature dimensions can be overcome by feature selection methods namely Gini Index (Shang et al., 2007) Using the right feature selection method can improve classification performance (Wang, Li, Song, Wei, & Li, 2011) and improve accuracy (Xu, Peng, & Cheng, 2012). Feature selection performs high feature reduction by removing irrelevant attributes (Koncz & Paralic, 2011). The Gini Index is used to separate attributes and get better classification accuracy (Shang et al., 2007). Gini Index is applied for feature selection and weight adjustment (Shankar & Karypis, 2000). Compared to other feature selection methods, the Gini Index shows better classification performance (Shang et al., 2007).

This explanation explained that the Gini Index has good potential in reducing the high dimension of features. Therefore in this study will use a combination of the two methods namely K-Nearest Neighbor and Gini Index to improve accuracy in the classification of student performance.

MATERIALS AND METHODS

Material

The student performance dataset obtained from the UCI Machine Learning Repository was used in this study. The student performance dataset consists of 30 attributes and 1 class. Table 1 shows the attributes and their description. Table 2 shows the attributes, data, and description of the data.

Table 1. Attributes and Descriptions on the Student Performance Dataset

No	Atribut	Information
1	Result	Graduation Result. (Is a class attribute)
2	School	School name
3	Sex	Gender
4	Age	Age
5	Address	Address
6	Famsize	Number of family members
7	Pstatus	Status of living with parents or not
8	Medu	Mother's education

No	Atribut	Information
9	Fedu	Father's education
10	Mjob	Mother's job
11	Fjob	Father's occupation
12	Reason	Reasons for choosing a school
13	Guardian	Student Guardians
14	Traveltime	Travel time from home to school
15	Studytime	Study time in a week
16	Failures	Amount of failure
17	Schoolsup	Additional educational support
18	Famsup	Family education support
19	Paid	additional tutoring
20	Activities	Extracurricular activities
21	Nursery	
22	Higher	Want to take higher education
23	Internet	Internet access at home
24	Romantic	Having a boyfriend or not
25	Famrel	Quality of family relationships
26	Freetime	Free time after school
27	Goout	Go with friends
28	Dalc	Consuming alcohol on weekdays
29	Walc	Consuming alcohol on weekends
30	Health	Current health status
31	Absences	Number of absences

Source: (Cortez & Silva, 2008)

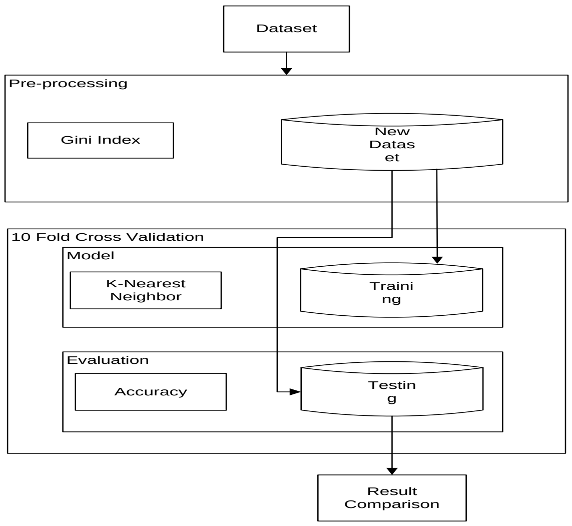
Table 2. Attributes, Data and Data Description on the Student Performance Dataset

No	Atribut	Data	Data Description
1	Result	Fail/ pass	Failed / passed
2	School	MS/ GP	MS: Mousinho da Silveira GP: Gabriel Pereira
3	Sex	M/ F	Male/ Female
4	Age	15-22	
5	Address	R/U	R: rural, U: urban
6	Famsize	LE3/GT3	LE3: <=3 GT: >3
7	Pstatus	A/T	A: separate Q: With parents
8	Medu	0/ 1/ 2/ 3/ 4	0: tidak ada 1: SD 2: SMP 3: SMA 4: pendidikan yang lebih tinggi
9	Fedu	0/ 1/ 2/ 3/ 4	0: tidak ada 1: SD 2: SMP 3: SMA 4: pendidikan yang lebih tinggi
10	Mjob	Techer/ health/ services/ at home/ other	Teacher: teacher Health: in the health sector Services: PNS At home: at home Other: other
11	Fjob	Techer/ health/ services/ at home/ other	Teacher: teacher Health: in the health sector Services: PNS At home: at home Other: other
12	Reason	Home/ reputation/ course/ other	Home: close to home Reputation: the reputation of the school Course: subjects
13	Guardian	Mother/ father/ other	Father / Mother / Others
14	Traveltime	1/ 2/ 3/ 4	1: <15 minutes 2: 15-30 minutes 3: 30 minutes - 1 hour 4:> 1 hour
15	Studytime	1/ 2/ 3/ 4	1: <2 hours 2: 2-5 hours 3: 5-10 hours 4:> 10 hours
16	Failures	1/ 2/ 3/ 4	1: 1 time 2: 2 times 3: 3 times 4:> 3 times
17	Schoolsup	Yes/ no	
18	Famsup	Yes/ no	
19	Paid	Yes/ no	
20	Activities	Yes/ no	
21	Nursery	Yes/ no	
22	Higher	Yes/ no	
23	Internet	Yes/ no	
24	Romantic	Yes/ no	
25	Famrel	1/ 2/ 3/ 4/ 5	1: very bad 2: bad 3: normal 4: good 5: very good
26	Freetime	1/ 2/ 3/ 4/ 5	1: very bad 2: bad 3: normal 4: good 5: very good
27	Goout	1/ 2/ 3/ 4/ 5	1: very bad 2: bad 3: normal 4: good 5: very good
28	Dalc	1/ 2/ 3/ 4/ 5	1: very bad 2: bad 3: normal 4: good 5: very good

29	Walc	1/ 2/ 3/ 4/ 5	1: very bad 2: bad 3: normal 4: good 5: very good
30	Health	1/ 2/ 3/ 4/ 5	1: very bad 2: bad 3: normal 4: good 5: very good
31	Absences	0-75	

Source: (Cortez & Silva, 2008)

Metode



Source: (Setiyorini & Asmono, 2019)
Figure 1. Application of the K-Nearest Neighbor and Gini Index Method

Figure 1 shows the proposed K-Nearest Neighbor and Gini Index methods in this study. At the pre-processing stage, feature selection is performed using the Gini Index method so that it produces a new dataset with the most optimal attributes. Then the new dataset is divided into training data and testing data with the 10 Fold Cross Validation method. Then the training data is classified using the K-Nearest Neighbor method. The final step of testing data is tested by looking at performance accuracy.

K-Nearest Neighbor

K-Nearest Neighbor is an effective, intuitive and simple method (Gou et al., 2014)(Lin et al., 2014). In pattern recognition, the K-Nearest Neighbor algorithm is a non-parametric method that is useful for grouping objects based on close features. The K-Nearest Neighbor concept is a label or class determined by the majority vote of its neighbors (Won Yoon & Friel, 2015). The working principle of K-Nearest Neighbor is to find the closest distance between the data that is evaluated with k nearest neighbors in the training data. The calculation equation to find Euclidean with d is distance and p is the data dimension, namely:

$$d_i = \sqrt{\sum_{i=1}^{p}(x_{1i} - x_{2i})^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

Where:
x1: sample test data d: distance
x2: test data p: a dimension of data

Gini Index

Gini Index is the probability of two randomly selected data that have different classes. The Gini Index is used by Breiman (Breiman, 2001) to produce a classification tree in the decision tree. Suppose S is 1 set of number s data. This data has a number of different m classes (Ci, i = 1, ..., m). Based on the class, we can divide S into a number of m subsets (Si, i = 1, ..., m), for example, Si is a dataset incorporated in the Ci class, si is the amount of data from Si, then the Gini Index can be formulated as follows:

$$Gini\ Index\ (S) = 1 - \sum_{i=1}^{m}\left(\frac{S_i}{S}\right)^2 \ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

RESULTS AND DISCUSSION

Table 3 shows the results of the experiment, which is the comparison of the accuracy of the K-Nearest Neighbor method with the K-Nearest Neighbor and Gini Index on the classification of student performance using the student performance dataset. Table 3 shows the K-Nearest Neighbor method obtained the largest average accuracy is 74.068 while the K-Nearest Neighbor method and Gini Index obtained the largest average accuracy is 76.516.

Table 3 Comparison of Accuracy with K-Nearest Neighbor with K-Nearest Neighbor and Gini Index

	Accuracy	
Experiment (k)	**K-Nearest Neighbor**	**K-Nearest Neighbor dan Gini Index**
1	68,96	72,41
2	62,55	67,24
3	75	75,96
4	72,6	74,62
5	76,34	77,78
6	76,34	78,07
7	77,58	79,12
8	77,11	79,22
9	77,1	80,75
10	77,1	79,99
Average	74,068	76,516

Source: (Setiyorini & Asmono, 2019)

The results of these experiments indicate that the Gini Index is able to overcome the problem of high feature dimensions in K-Nearest Neighbor so that the accuracy of the classification of student performance is better than using the K-Nearest Neighbor method alone. This proves the research of Shang et al. that the Gini Index is able to reduce high dimensional dimensions so that it gets better classification accuracy (Shang et al., 2007). The results also prove the research of Setiyorini and Asmono (Setiyorini & Asmono, 2017), that the Gini Index is an effective method to improve the performance of K-Nearest Neighbor so as to improve the accuracy of the cognitive level classification of problems in Bloom's Taxonomy.

CONCLUSION

The results of 10 experiments with a value of k (1 to 10) in the student performance dataset with the K-Nearest Neighbor method obtained the greatest average accuracy is 74.068 while the K-Nearest Neighbor and Gini Index methods obtained the largest average accuracy is 76.516. From the results of these experiments, it can be concluded that feature selection with the Gini Index is able to reduce the feature dimensions which are high so that the application of K-Nearest Neighbor and Gini Index can improve the classification accuracy of student performance better than using the K-Nearest Neighbor method alone.

REFERENCE

Adeniyi, D. A., Wei, Z., & Yongquan, Y. (2016). Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. *Applied Computing and Informatics*, *12*(1), 90–108. https://doi.org/10.1016/j.aci.2014.10.001

Al-Shehri, H., Al-Qarni, A., Al-Saati, L., Batoaq, A., Badukhen, H., Alrashed, S., … Olatunji, S. O. (2017). Student performance prediction using Support Vector Machine and K-Nearest Neighbor. *Canadian Conference on Electrical and Computer Engineering*, 17–20. https://doi.org/10.1109/CCECE.2017.7946847

Alkhasawneh, R., & Hobson, R. (2011). Modeling student retention in science and engineering disciplines using neural networks. In *2011 IEEE Global Engineering Education Conference, EDUCON 2011* (pp. 660–663). https://doi.org/10.1109/EDUCON.2011.5773209

Altujjar, Y., Altamimi, W., Al-Turaiki, I., & Al-Razgan, M. (2016). Predicting Critical Courses Affecting Students Performance: A Case Study. *Procedia Computer Science*, *82*(March), 65–71. https://doi.org/10.1016/j.procs.2016.04.010

Breiman, L. (2001). *Classification and regression tree*.

Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting student performance from LMS data: A comparison of 17 blended courses using moodle LMS. *IEEE Transactions on Learning Technologies*, *10*(1), 17–29. https://doi.org/10.1109/TLT.2016.2616312

Cortez, P., & Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. *In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)*, 5–12.

Cover, T. M., & Hart, P. E. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

de Vries, A. P., Mamoulis, N., Nes, N., & Kersten, M. (2003). Efficient k-NN search on vertically decomposed data (p. 322). https://doi.org/10.1145/564728.564729

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Erven, G. Van. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*, *94*(February), 335–343. https://doi.org/10.1016/j.jbusres.2018.02.012

Gou, J., Zhan, Y., Rao, Y., Shen, X., Wang, X., & He, W. (2014). Improved pseudo nearest neighbor classification. *Knowledge-Based Systems*, *70*, 361–375. https://doi.org/10.1016/j.knosys.2014.07.020

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques*. *Data Mining*. https://doi.org/10.1016/b978-0-12-381479-1.00001-0

Koncz, P., & Paralic, J. (2011). An approach to feature selection for sentiment analysis. In *INES 2011 - 15th International Conference on Intelligent Engineering Systems, Proceedings* (pp. 357–362). https://doi.org/10.1109/INES.2011.5954773

Lin, Y., Li, J., Lin, M., & Chen, J. (2014). A new nearest neighbor classifier via fusing neighborhood information. *Neurocomputing*, *143*, 164–169. https://doi.org/10.1016/j.neucom.2014.06.009

Lopez Guarin, C. E., Guzman, E. L., & Gonzalez, F. A. (2015). A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining. *Revista Iberoamericana de Tecnologias Del Aprendizaje*, *10*(3), 119–125. https://doi.org/10.1109/RITA.2015.2452632

López, J., & Maldonado, S. (2018). Redefining nearest neighbor classification in high-dimensional settings. *Pattern Recognition Letters*, *110*, 36–43. https://doi.org/10.1016/j.patrec.2018.03.023

Pandey, M., & Taruna, S. (2016). Towards the integration of multiple classifier pertaining to the Student's performance prediction. *Perspectives in Science*, *8*, 364–366. https://doi.org/10.1016/j.pisc.2016.04.076

Setiyorini, T., & Asmono, R. T. (2017). Penerapan Gini Index dan K-Nearest Neighbor untuk Klasifikasi Tingkat Kognitif Soal pada Taksonomi Bloom. *Jurnal Pilar Nusa Mandiri*, *13*(2), 209–216.

Setiyorini, T., & Asmono, R. T. (2019). *Laporan Akhir Penelitian Mandiri*.

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, *72*, 414–422. https://doi.org/10.1016/j.procs.2015.12.157

Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., & Wang, Z. (2007). A novel feature selection algorithm for text categorization. *Expert Systems with Applications*, *33*(1), 1–5. https://doi.org/10.1016/j.eswa.2006.04.001

Shankar, S., & Karypis, G. (2000). A Feature Weight Adjustment Algorithm for Document Categorization.

Villagrá-Arnedo, C. J., Gallego-Durán, F. J., Llorens-Largo, F., Compañ-Rosique, P., Satorre-Cuerda, R., & Molina-Carmona, R. (2017). Improving the expressiveness of black-box models for predicting student performance. *Computers in Human Behavior*, *72*, 621–631. https://doi.org/10.1016/j.chb.2016.09.001

Wang, S., Li, D., Song, X., Wei, Y., & Li, H. (2011). A feature selection method based on improved fisher's discriminant ratio for text sentiment classification. *Expert Systems with Applications*, *38*(7), 8696–8702. https://doi.org/10.1016/j.eswa.2011.01.077

Won Yoon, J., & Friel, N. (2015). Efficient model selection for probabilistic K nearest neighbour classification. *Neurocomputing*, *149*(PB), 1098–1108. https://doi.org/10.1016/j.neucom.2014.07.023

Xu, T., Peng, Q., & Cheng, Y. (2012). Identifying the semantic orientation of terms using S-HAL for sentiment analysis. *Knowledge-Based Systems*, *35*, 279–289. https://doi.org/10.1016/j.knosys.2012.04.011

Yang, F., & Li, F. W. B. (2018). Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Computers and Education*, *123*(October 2017), 97–108. https://doi.org/10.1016/j.compedu.2018.04.006