IMPLEMENTATION OF GAIN RATIO AND K-NEAREST NEIGHBOR FOR CLASSIFICATION OF STUDENT PERFORMANCE

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

Teknik Informatika[1]
STMIK Nusa Mandiri Jakarta[1]
http://nusamandiri.ac.id[1]
tyas.setiyorini@gmail.com[1]

Teknik Informatika[2]
STMIK Swadharma[2]
http://swadharma.ac.id[2]
rtriasmono@gmail.com[2]

Abstract— Predicting student performance is very useful in analyzing weak students and providing support to students who face difficulties. However, the work done by educators has not been effective enough in identifying factors that affect student performance. The main predictor factor is an informative student academic score, but that alone is not good enough in predicting student performance. Educators utilize Educational Data Mining (EDM) to predict student performance. KK-Nearest Neighbor is often used in classifying student performance because of its simplicity, but K-Nearest Neighbor has a weakness in terms of the high dimensional features. To overcome these weaknesses, Gain Ratio is used to reduce the high dimension of features. The experiment has been carried out 10 times with the value of k is 1 to 10 using the student performance dataset. The results of these experiments are obtained an average accuracy of 74.068 with the K-Nearest Neighbor, and obtained an average accuracy of 75.105 with the Gain Raatio and K-Nearest Neighbor. The experimental results show that Gain Ratio is able to reduce the high dimensions of features that are a weakness of K-Nearest Neighbor, so the implementation of Gain Ratio and K-Nearest Neighbor can increase the accuracy of the classification of student performance compared to using the K-Nearest Neighbor alone.

Keywords: Gain Ratio, K-Nearest Neighbor, Student Performance

Abstrak— *Memprediksi kinerja siswa sangat berguna dalam menganalisa siswa yang lemah dan memberikan dukungan pada siswa yang menghadapi kesulitan. Namun, pekerjaan yang dilakukan oleh pendidik belum cukup efektif dalam mengidentifikasi faktor-faktor yang mempengaruhi kinerja siswa. Faktor utama yaitu skor akademik yang informatif, tetapi itu saja tidak cukup untuk dijadikan faktor dalam memprediksi kinerja siswa. Pendidik memanfaatkan Educational Data Mining (EDM) untuk memprediksi kinerja siswa. K-Nearest Neighbor sering digunakan pada klasifikasi kinerja siswa karena kesederhanaannya, namun K-Nearest Neighbor memiliki kelemahan dalam hal tingginya dimensi fitur. Untuk mengatasi kelemahan tersebut digunakan Gain Ratio untuk mengurangi tingginya dimensi fitur. Percobaan telah dilakukan sebanyak 10 kali dengan nilai k yaitu 1 sampai dengan 10 dengan menggunakan dataset student performance. Hasil dari percobaan tersebut adalah didapatkan rata-rata akurasi sebesar 74,068 dengan K-Nearest Neighbor, serta didapatkan rata-rata akurasi sebesar 75,105 dengan Gain Ratio dan K-Nearest Neighbor. Hasil percobaan tersebut menunjukkan bahwa Gain Ratio mampu mengurangi tingginya dimensi fitur yang menjadi kelemahan K-Nearest Neighbor, sehingga penerapan Gain Ratio dan K-Nearest Neighbor dapat meningkatkan akurasi klasifikasi kinerja siswa dibanding dengan menggunakan K-Nearest Neighbor saja.*

Kata Kunci: *K-Nearest Neighbor, Gain Ratio, Kinerja Siswa*

INTRODUCTION

Predicting student performance at an early stage is very beneficial in figuring out weak students (Pandey & Taruna, 2016) and permits academic establishments to provide suitable support for students who face difficulties (Altujjar et al., 2016). Prediction models are used to detect trends and Detecting trends and patterns of behavior in learning problems can be identified using prediction methods (Villagrá-Arnedo et al., 2017). Many factors other than academic factors are taken into consideration in constructing student performance prediction models, such as psychological, social, and demographic factors (Altujjar et al., 2016). The main predictor factor is an informative student academic score, but that

DOI: 10.33480/pilar.v16i1.813

alone is not good enough in predicting student performance (Carnegie et al., 2012). Social, personal and academic elements also affect in predicting student performance (Fernandes et al., 2019). The work carried out by means of educators has no longer been quite effective in identifying which factors will improve student performance, how students can improve, and whether students have the potential to do better (Yang & Li, 2018).

Educators utilize Educational Data Mining (EDM) to predict student performance (Altujjar et al., 2016). EDM makes use of a database of the education system to investigate students and their learning styles more comprehensively in order to design instructional policies as a way to enhance their academic performance and reduce the failure charge at the end of every college year (Fernandes et al., 2019). The method that is widely used in EDM to predict student performance is classification (Altujjar et al., 2016). Some classifications of student performance research had been conducted, such as K-Nearest Neighbor (Pandey & Taruna, 2016), Decision Tree (Lopez Guarin et al., 2015), dan Naive Bayes (Lopez Guarin et al., 2015).

K-Nearest Neighbor has attracted great interest for researchers (Gou et al., 2014) (Lin et al., 2014)(Lin et al., 2014). From the three research studied, K-Nearest Neighbor is able to provide performance with the best accuracy (Shahiri et al., 2015). Efficiently K-Nearest Neighbor is able to identify student performance as slow students, average students, good students and excellent students (Minaei-Bidgoli & Kashy, 2003)(Mayilvaganan & Kalpanadevi, 2015). K-Nearest Neighbor provides excellent accuracy in predicting styles for student development in better education (Gray et al., 2014). The advantage of K-Nearest Neighbor is its simplicity, which allows the classification of two or more patterns using fairly simple rules (Han et al., 2012).

The simplicity of K-Nearest Neighbor also raises several problems, the main problem being related to the high dimensional features (López & Maldonado, 2018). Another disadvantage of K-Nearest Neighbor is the complexity of computing big data similarities. One way to reduce the complexity of K-Nearest Neighbor is to reduce the high dimensional features (de Vries et al., 2003). The high dimension of features is also a major problem in classification so it is not permitted for many learning algorithms (Shang et al., 2007).

High data dimensions can complicate testing and training in classification. Selection of a subset of features is very important in data mining (Karegowda & Manjunath, 2010). Dimension reduction is very important in pattern formation (López & Maldonado, 2018). The filter approach is the Gain Ratio that has been used for the selection of the most important features in the classification (Karegowda & Manjunath, 2010). Gain Ratio is used as an attribute selection criteria in algorithms such as C4.5 (Dai & Xu, 2013). Attributes that are not relevant to class variables can be deleted using Gain Ratio. Gain Ratio can effectively and efficiently assess the relationship between attributes and class. (Chen et al., 2008). Gain Ratio is one of the attribute selection methods that can significantly improve classification accuracy (Snousy et al., 2011),

Gain Ratio has good potential in reducing the high dimension of features which is a problem in K-Nearest Neighbor. Therefore this research will use two methods are Gain Ratio and K-Nearest Neighbor to increase accuracy in the classification of student performance.

DATA AND METHODOLOGY

Data

The student performance dataset was used in this study. This dataset is obtained from the Machine Learning Repository, UCI. The student performance dataset consists of 30 attributes and 1 class. Table 1 shows the attributes and information in the student performance dataset.

Table 1. Attributes and Descriptions on the Student Performance Dataset

Number	Attributes	Description
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
9	Fedu	Father's education
10	Mjob	Mother's job
11	Fjob	Father's job
12	Reason	Reasons for choosing 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	Early childhood education
22	Higher	Want to take higher

Number	Attributes	Description
		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 shows the attributes, data, and description of the data in the student performance dataset.

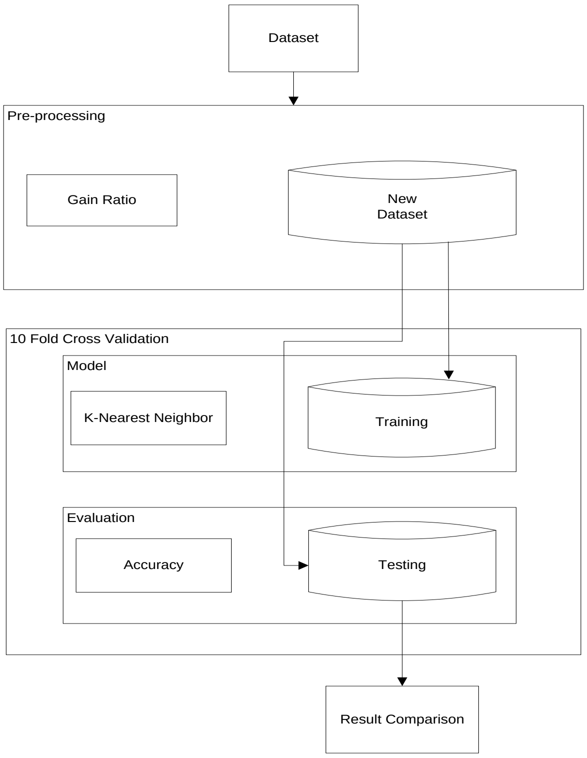
Table 2. Attributes, Data and Data Description on the Student Performance Dataset

Number	Attribute	Data	Description Data
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 T: with parents
8	Medu	0/ 1/ 2/ 3/ 4	0: nothing 1: elementary school 2: middle school 3: high school 4: higher education
9	Fedu	0/ 1/ 2/ 3/ 4	0: nothing 1: elementary school 2: middle school 3: high school 4: higher education
10	Mjob	Techer/ health/ services/ at home/ other	
11	Fjob	Techer/ health/ services/ at home/ other	
12	Reason	Home/ reputation/ course/ other	
13	Guardian	Mother/ father/ other	
14	Traveltime	1/ 2/ 3/ 4	1: <15 minute 2: 15-30 minute 3: 30 minute - 1 hour 4: > 1 hour
15	Studytime	1/ 2/ 3/ 4	1: < 2 hour 2: 2-5 hour 3: 5-10 hour 4: > 10 hour
16	Failures	1/ 2/ 3/ 4	1: once 2: twice 3: three 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	

Sumber: (Cortez & Silva, 2008)

Methodology

K-Nearest Neighbor and Gain Ratio proposed in this research are shown in Figure 1.



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

At the pre-processing stage, feature selection is performed using the Gain Ratio method so as to produce a new dataset with the most optimal attributes. Then the new dataset is divided into training data and testing data using the 10 Fold Cross Validation method. Then the training data is classified using the K-Nearest Neighbor. The final step of testing data is tested by looking at performance accuracy.

K-Nearest Neighbor

K-Nearest Neighbor is a famous method for classification, which has proven successful in many applications (Buttrey & Karo, 2002). K-Nearest has frequent and significant advantages in producing competitive results (Adeniyi et al., 2016). K-Nearest Neighbor is powerful, intuitive, effective, and simple (Gou et al., 2014)(Lin et al., 2014). Pattern recognition on the K-Nearest Neighbor is done by grouping objects based on close features. The class is determined by the voice of the majority of its neighbors is the K-Nearest Neighbor concept (Won Yoon & Friel, 2015). The working principle of K-Nearest Neighbor is to find the closest distance between the data evaluated with k nearest neighbors in the training data. The calculation equation for finding Euclidean with d is distance and p is the dimension of data is:

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

where: x1 = sample test data, x2: test data, d: distance, p: dimension of data.

Gain Ratio

Gain Ratio calculations is very measurable and efficient for big datasets with many examples (Chen et al., 2008). Gain Ratio can calculate attribute weights, which can be combined with other methods to achieve better performance (Zhang & Sheng, 2004). In most dataset and classification methods, attribute selection with Gain Ratio slightly increases the accuracy of classification (Snousy et al., 2011).

Gain Ratio was introduced by Quinlan, which was originally used to select attributes among a set of attributes that can classify well in the C4.5 algorithm (Quinlan, 1993). The C4.5 algorithm step is to determine the most predictive attribute and separate vertices based on that attribute. To calculate Gain Ratio, split information is needed. Split information can be calculated as follows:

$$SplitInfo_A(D) = \sum_{j=1}^{y} \frac{|D_j|}{|D|} \times log(\frac{|D_j|}{|D|}) \text{ .... } (2)$$

Where Dj to Dy is a subset of y resulting from solving D using attribute A which has as many as y values.

Next the Gain Ratio is calculated by:

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)} \text{ .... } (3)$$

Where: Gain (A) = information gain on attribute A, SplitInfoA (D) = split information on attribute A. The attribute with maximum Gain Ratio is chosen as the best separation attribute.

RESULTS AND DISCUSSION

Comparison of the accuracy of K-Nearest Neighbor with Gain Ratio and K-Nearest Neighbor in the classification of student performance using the student performance dataset is shown in Table 3. The experiment has been carried out 10 times with the value of k is 1 to 10 in Table 3 show that the average obtained accuracy of 74.068 using the K-Nearest Neighbor, and obtained an average accuracy of 75.105 by using the Gain Ratio and K-Nearest Neighbor.

Based on the results of these experiments indicate that Gain Ratio is able to correct weaknesses in the dimensions of high features that are a problem in K-Nearest Neighbor, so Gain Ratio can improve the classification of student performance compared to using the K-Nearest Neighbor method alone. This proves the research of Snousy et al. That selection of the Gain Ratio attribute significantly increases the highest classification accuracy in most datasets and classification methods (Snousy et al., 2011). These results also prove Dai & Xu's research that the accuracy of the Gain Ratio algorithm classification is higher than other gain-based algorithms (Dai & Xu, 2013).

Table 3. Comparison of K-Nearest Neighbor with Gain Ratio and K-Nearest Neighbor Accuracy

	Accuracy	
Experiment (k)	**K-Nearest Neighbor**	**Gain Ratio and K-Nearest Neighbor**
1	68,96	72,22
2	62,55	67,05
3	75	75,28
4	72,6	74,23
5	76,34	76,53
6	76,34	74,71
7	77,58	77,4
8	77,11	77,11
9	77,1	78,26
10	77,1	78,26
Average	74,068	75,105

Source: (Setiyorini & Asmono, 2019)

CONCLUSION

The experiment has been carried out 10 times with the value of k is 1 to 10 using the student performance dataset. The results of these experiments are obtained an average accuracy of 74.068 with the K-Nearest Neighbor, and obtained an average accuracy of 75.105 with the Gain Ratio and K-Nearest Neighbor.

The experimental results show that Gain Ratio is able to reduce the dimensions of high features that are a weakness in K-Nearest Neighbor, so that the implementation of Gain Ratio and K-Nearest Neighbor can increase the accuracy of the classification of student performance compared to using K-Nearest Neighbor only.

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

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

Buttrey, S. E., & Karo, C. (2002). Using k-nearest-neighbor classification in the leaves of a tree. *Computational Statistics & Data Analysis*, *40*(1), 27–37. https://www.sciencedirect.com/science/article/pii/S0167947301000986

Carnegie, D. A., Watterson, C., Andreae, P., & Browne, W. N. (2012). Prediction of success in engineering study. *IEEE Global Engineering Education Conference, EDUCON*. https://doi.org/10.1109/EDUCON.2012.6201020

Chen, J., Huang, H., Tian, F., & Tian, S. (2008). A selective Bayes Classifier for classifying incomplete data based on gain ratio. *Knowledge-Based Systems*, *21*(7), 530–534. https://doi.org/10.1016/j.knosys.2008.03.013

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.

Dai, J., & Xu, Q. (2013). Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. *Applied Soft Computing Journal*, *13*(1), 211–221. https://doi.org/10.1016/j.asoc.2012.07.029

de Vries, A. P., Mamoulis, N., Nes, N., & Kersten, M. (2003). *Efficient k-NN search on vertically decomposed data*. 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

DOI: 10.33480/pilar.v16i1.813

Gray, G., McGuinness, C., & Owende, P. (2014). An application of classification models to predict learner progression in tertiary education. *Souvenir of the 2014 IEEE International Advance Computing Conference, IACC 2014*, 549–554. https://doi.org/10.1109/IAdCC.2014.6779384

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

Karegowda, A. G., & Manjunath, A. S. (2010). COMPARATIVE STUDY OF ATTRIBUTE SELECTION USING GAIN RATIO AND CORRELATION BASED FEATURE SELECTION. *International Journal of Information Technology and Knowledge Management*, *2*(2), 271–277. http://csjournals.com/IJITKM/PDF 3-1/19.pdf

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

Mayilvaganan, M., & Kalpanadevi, D. (2015). Comparison of classification techniques for predicting the performance of students academic environment. *2014 International Conference on Communication and Network Technologies, ICCNT 2014*, *2015-March*, 113–118. https://doi.org/10.1109/CNT.2014.7062736

Minaei-Bidgoli, B., & Kashy, D. (2003). Predicting student performance: an application of data mining methods with the educational web-based system LON-CAPA. *Frontiers in Education, 2003*, *1*, 1–6.

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

Quinlan, J. R. (1993). *{C4}.5 - Programs for Machine Learning*.

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

Snousy, M. B. Al, El-Deeb, H. M., Badran, K., & Khlil, I. A. Al. (2011). Suite of decision tree-based classification algorithms on cancer gene expression data. *Egyptian Informatics Journal*, *12*(2), 73–82. https://doi.org/10.1016/j.eij.2011.04.003

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

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

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

Zhang, H., & Sheng, S. (2004). Learning weighted naive bayes with accurate ranking. *Proceedings - Fourth IEEE International Conference on Data Mining, ICDM 2004*, 567–570. https://doi.org/10.1109/ICDM.2004.10030