Analysis K-Means Clustering to Predicting Student Graduation

M Wati1*, W H Rahmah1, N Novirasari1, Haviluddin1, E Budiman1, Islamiyah1

1 Department of Informatics, Universitas Mulawarman, Samarinda City, East Kalimantan 75117, Indonesia
2 Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia

Email: masnawati.ssi@gmail.com*

Abstract. The prediction of students' graduation outcomes has been an important field for higher education institutions because it provides planning for them to develop and expand any strategic programs that can help to improve student academics performance. Data mining techniques can cluster student academics performance in predicting student graduation. The aim of this study is to analysis the performance of data mining techniques for predicting students' graduation using the K-Means clustering algorithm. The data pre-processing used for data cleaning, and data reducing using Principle Component Analysis to determine any variables that affect the graduation time. This algorithm processes datasets of student academics performance numbering 241 students with 16 variables. Based on the clustering using K-means, the highest accuracy rate is 78.42% in the 3-cluster model and the smallest accuracy rate is 16.60% in the 4-cluster model. The influential variable in predicting student graduation based on the value of the loading factor is the GPA total of the 1st to 6th semester.

1. Introduction

Education is the most important component of human life. Education can be in the form of theory, practice, even moral. In the Government Regulation of RI No.12/2012 concerning Higher Education, Chapter 1 Article 1 paragraph 1 explains "Education is a mindful planned effort to actualize the learning atmosphere and learning process to get up their religious-spiritual potential, self-control, personality strength, intelligence and character, and skills" [1], [2]. The prediction of students' graduation outcomes has been an important field for higher education institutions because it provides planning to develop and expand any strategic programs that can improve student academic performance. It can also affect the institutions' reputation in describing graduates' quality [3]–[6]. Most of the studies that have been done use the techniques of data mining or Multi-attribute Decision Making to predict students' completion. Some of the techniques were C4.5 Decision Tree [7], Naïve Bayes [6], MADM [8]–[10], and Support Vector Machine [3], [4].

In previous research, the prediction of student learning outcomes was carried out using the C4.5 and Naïve Bayesian methods. This study compares the performance of the two methods in classifying students' graduation times into 3 classes. The classification model formed shows that these two methods have an accuracy rate above while the Naïve Bayes Classifier's precision rate is 60%, and in the Tree C4.5 method is 58.82% [6]. In this study, it has discussed the problem of classifying student graduation times using data mining techniques. It is also necessary to analyze the data's characteristics to see the clustering of student graduation data distribution. The number of clusters formed can be used for predicting student graduation.
The data clustering aims to provide an agglomerate of similarity data records. Clustering is often confused as classification, but they both have different goals. In simple terms, the distance intra-clusters need to be minimized for better clustering results, as in the K-means algorithm \([11]\). K-Means is a non-hierarchical clustering method that tries to partition data set into some clusters. Hence, the data has the same characteristics collected into the same cluster, and others are collected into other clusters. K-Means method is the notable cluster analysis algorithm in data mining. Through some experiments clustering in the K-means method, the obtained that its cluster result varies along with the initial cluster central point \([12], [13]\). The advantages of K-means algorithm are quick convergence to distortion minimum and apprehending how many clusters in the dataset \([14], [15]\).

2. Methodology
This study's steps based on the cross-industry standard process for data mining (CRISP-DM). The steps are dataset collecting, data pre-processing, modeling, and evaluation of clustering presented in Figure 1.

In Figure 1, the data preprocessing step consists of data cleaning and then reducing the attributes using the Principle Component Analysis (PCA) method. In this step, the dataset has been generated to perform clustering analysis using the K-means method. The evaluation stage is carried out to see the K-means method's performance in clustering student graduation data. In predicting student graduation through K-Means clustering is used student academic performance data which consists of attributes which are presented in Table 1.

Table 1. Attributes of student academics performance

| Attributes              | Parameters                                      |
|-------------------------|-------------------------------------------------|
| Gender                  | Male, Female                                    |
| Birthdate               | In years                                        |
| Age                     | Ages when registered                            |
| Birthplace              | In city, out of town                            |
| High school Status      | Public, private                                 |
| Types of school         | Vocational high school, senior high school       |
| Hometown                | In city, out of town                            |
| Enrollment paths        | SNMPTN, SBMPTN, SMMPTN                          |
| GPA 1st Semester        | \(1 = \text{GPA} \leq 1.5\)                     |

Figure 1. Steps of clustering to predicting students’ graduation
2.1. Data preprocessing
Data cleaning is useful for cleaning data sets that have a missing value in which data that has missing value attributes are removed from the data set. Data reduction process using Principle Component Analysis (PCA). PCA used to attributes reduction. It used to eliminate irrelevant attributes in predicting students' graduation. The selecting relevant and non-correlated attributes without affecting the information in the initial data set, then the predictor is developed using the K-means method to clustering data set [16], [17].

2.2. Modeling
The K-means clustering method utilized to predict student graduation in this study. The algorithm of K-means clustering method shown in figure 2.
Figure 2. Flowchart of K-Means algorithm

2.3. Evaluation
The evaluation step is done for measuring the performance of K-Means in clustering the students' graduation using the in the confusion matrix [18]–[20].

3. Result and Discussion
The data pre-processing step carried out data cleaning and reduction data with the PCA method, so we obtained the relevant attributes to clustering the dataset using the K-means algorithm. The value of each PCA loading factor variables is presented in Table 2.

Table 2. Value of PCA loading factor for each variable

|                   | Gender | Birthdate | Age | SS  | TS  | Hometown | Enroll | GPA  | IP1  | IP2  | IP3  | IP4  | IP5  | IP6  | SKS2 | ANG |
|-------------------|--------|-----------|-----|-----|-----|----------|--------|------|------|------|------|------|------|------|------|-----|
| Gender            | 1      | -0.12     | -0.134 | -0.015 | -0.174 | 0.089 | 0.263 | 0.411 | 0.37  | 0.342 | 0.338 | 0.331 | 0.304 | 0.208 | 0.024 | 0.175 |
| Birthdate         | 0.12   | 1         | 0.041 | -0.058 | -0.15  | 0.678 | -0.055 | 0.006 | 0.086 | -0.017 | 0.01  | -0.069 | -0.069 | 0.019 | 0.04  | 0.03  |
| Age               | -0.134 | 0.041     | 1    | 0.194 | -0.022 | -0.079 | -0.205 | -0.316 | -0.181 | -0.288 | -0.236 | -0.36  | -0.281 | -0.202 | -0.121 | -0.352 |
| SS                | -0.015 | -0.058    | 0.194 | 1    | 0.257 | -0.113 | -0.023 | -0.087 | -0.058 | -0.027 | -0.081 | -0.065 | -0.13  | 0.004 | 0.014 | 0.05  |
| TS                | -0.174 | -0.15     | -0.022 | 0.257 | 1    | -0.248 | -0.013 | -0.104 | -0.135 | -0.125 | -0.125 | -0.044 | 0.054  | 0.047 | 0.001 | -0.033 |
| Hometown          | 0.089  | 0.678     | -0.079 | -0.113 | -0.248 | 1    | 0.026 | 0.084 | 0.117 | 0.028 | 0.097 | 0.041 | 0.084  | 0.009 | -0.039 | 0.004 |
| Enroll            | 0.263  | -0.055    | -0.205 | -0.023 | -0.013 | 0.026 | 1    | 0.523 | 0.417 | 0.412 | 0.348 | 0.335  | 0.438  | 0.376 | 0.237 | -0.533 |
| GPA               | 0.411  | 0.006     | -0.316 | -0.087 | -0.104 | 0.084 | 0.523 | 1    | 0.683 | 0.756 | 0.776 | 0.748  | 0.756  | 0.688 | 0.257 | 0.487 |
| IP1               | 0.37   | 0.086     | -0.181 | -0.085 | -0.135 | 0.117 | 0.417 | 0.683 | 1    | 0.552 | 0.465 | 0.46   | 0.364  | 0.399 | 0.091 | 0.375 |
| IP2               | 0.342  | -0.017    | -0.288 | -0.027 | -0.125 | 0.028 | 0.412 | 0.756 | 0.552 | 1    | 0.626 | 0.497  | 0.446  | 0.407 | 0.304 | 0.447 |
| IP3               | 0.382  | 0.01      | -0.236 | -0.081 | -0.125 | 0.097 | 0.348 | 0.776 | 0.465 | 0.626 | 1    | 0.634  | 0.573  | 0.416 | 0.043 | 0.206 |
| IP4               | 0.331  | -0.069    | -0.36  | -0.065 | -0.044 | 0.041 | 0.335 | 0.748 | 0.46  | 0.497 | 0.634 | 1    | 0.608  | 0.511 | 0.038 | -0.236 |
| IP5               | 0.304  | -0.069    | -0.281 | -0.13 | -0.054 | 0.084 | 0.438 | 0.756 | 0.364 | 0.446 | 0.573 | 0.608 | 1    | 0.414  | 0.269 | 0.405 |
Based on the result of PCA in Table 2, there are 5 variables loading factor which has a negative loading factor value which are birthplace = -0.055, ages = 0.205, school status = -0.023, type of school = -0.013, and class of students = -0.533. They were removed because it did not have a significant influence on the determination of graduation time. The relevant attributes for use in predicting students’ graduation are:

- Gender
- Hometown
- GPA in 1st semester
- GPA in 2nd semester
- GPA in 3rd semester
- GPA in 4th semester
- GPA in 5th semester
- GPA in 6th semester
- GPA in 1st to 6th semester
- Credits in 2nd semester
- Credits in 3rd semester
- Credits in 4th semester
- Activeness in student organizations
- Parent’s earnings
- Parent’s education level

The number of students in the dataset used for the clustering process was 241 students. In clustering this dataset, the start work in the K-means method is to determine the number of clusters to be reviewed. In this study, 3 experimental models were carried out, namely an experiment with 2 clusters, 3 clusters, and 4 clusters. These three experiments were then evaluated to see which cluster model was best used in predicting student graduation.

In K-means clustering, data selected randomly to be the center of the initial cluster according to the number of clusters determined. In the 3-cluster model, data are selected randomly through the average, minimum, and maximum values of the GPA attribute. Clusters center is formed, namely the 13th data, the 91st data, and the 202nd data in the dataset.

### Table 3. Initial cluster centre of K-means algorithm

| Student | Gender | HT | IP1 | IP2 | IP3 | IP4 | IP5 | IP6 | IPK | SKS2 | SKS3 | SKS4 | AO | PI | PE |
|---------|--------|----|-----|-----|-----|-----|-----|-----|-----|------|------|------|----|----|----|
| 13      | 1      | 1  | 1   | 1   | 1   | 2   | 3.3 | 3.1 | 3.7 | 4.0  | 3.5  | 3.6  | 3.46| 2  | 2  |
| 91      | 1      | 2  | 2   | 1   | 2   | 2   | 3.3 | 1.3 | 2.1 | 3.2  | 3.3  | 3.3  | 2.69| 3  | 2  |
| 202     | 1      | 2  | 2   | 1   | 2   | 2   | 3.9 | 3.9 | 3.8 | 4.0  | 4.0  | 3.9  | 3  | 3  |

Based on the initial cluster center in Table 3, the next step determines the centroid distance for each data to each cluster center using Euclidian distance on equation (1).

\[
d(x, y) = |x - y| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}
\]  

(1)

Based on equation (1), the distance between each data and the center of the centroid is obtained as follows:

Centroid distance for centroid-1 to data-1:
Centroid distance for centroid-2 to data-1:

$$\sqrt{\frac{1}{1} \sum (c1 - d1)^2}$$

Centroid distance for centroid-3 to data-1:

$$\sqrt{\frac{1}{1} \sum (c1 - d1)^2}$$

Each data is grouped based on the distance of the closest centroid. Comparison of a confusion matrix for 3 models of clusters is presented in Table 4.

| Table 4. Comparison of Confusion matrix 3 cluster models |
|---------------------------------------------------------|
| 2-Cluster | 3-Cluster | 4-Cluster |
| Accuracy   | 61.41%    | 78.42%    | 16.60%    |
| Error Rate | 38.59%    | 21.58%    | 83.40%    |
| Recall     | 59.60%    | 84.91%    | 49.66%    |
| Specificity| 69.77%    | 65.56%    | 63.54%    |
| Precision  | 69.77%    | 68.10%    | 32.52%    |

Based on Table 4 confusion matrix of K-Means clustering, the accuracy, error, recall, specificity, and precision rate of each experiment for 2-cluster, 3-cluster, and 4-cluster models is presented in Figure 3.

In the 2-cluster model, the accuracy rate obtained is 61.41%, the error rate is 38.59%, and the precision rate is 69.77%. In the 3-cluster model, the accuracy rate is 78.42%, the error rate is 21.58%, and the precision rate amounting to 68.10%. In contrast, in the 4-cluster model, the accuracy rate is 16.60%, the error rate is 83.40%, and the precision rate is 32.52%. The highest accuracy rate of K-Means clustering modeling with 3-cluster models to predict student graduation use the student academic dataset in Table 1. This shows that the most appropriate clustering of data to predict student graduation uses a 3-cluster model. The graph can illustrate the results of the study in Figure 4.
Figure 4. Effect of independent variables on the dependent variable chart.

In Figure 4, Gender influences students’ graduation time, with the male gender study period is faster than the female gender. GPA 6th semester, Total credits 4th semester, and Grade Point Average total affect student graduation time. Students who have a higher 6th semester GPA and Grade Point Average from the first semester to the 6th semester can have passed on time than students with lower both of GPA 6th semester and GPA total. Students who have the greatest credits in the 4th semester have to pass on time than students who have lower credits. Students who have higher parents’ educational levels do not result in students having a faster study period than others. Organizational activity influences student graduation, which is students who are active in organizational membership has a slower study period than students who are not active in organizational membership.

4. Conclusion
As the results of this study obtained that the attributes are birthplace, ages, school status, type of school, and a class of students all have negative loading factor value in the PCA method, which means it all has a small correlation value to the prediction of the student graduation. Variables affect student graduation time are gender, which is male students graduate faster than females, GPA in 6th semester and GPA 1st to 6th semester that students who have higher grades graduate faster than others, students who have higher credits in 4th-semester is faster than other, and students who are active in the organization late to graduate than students who are not active in the organization. Based on 3 cluster models is obtained that the 3-cluster model is the best clustering with an accuracy rate obtained is 78.42%, an error rate is 21.58%, and a precision rate is 68.10%. As future work, it is necessary to carry out a cluster analysis using other methods, resulting in better cluster performance.

5. References
[1] P. Pannen, A. Wirakartakusumah, and H. Subhan, “Autonomous 5 higher education institutions in Indonesia,” Gov. Manag. Univ. Asia Glob. Influ. local responses, p. 56, 2019.
[2] P. Indonesia, Government Regulation No. 12/2012 about Higher Education. Indonesia, 2012.
[3] Y. Pang, N. Judd, J. O’Brien, and M. Ben-Avie, “Predicting students’ graduation outcomes through support vector machines,” Proc. - Front. Educ. Conf. FIE, vol. 2017-Octob, pp. 1–8, 2017, doi: 10.1109/FIE.2017.8190666.
[4] D. Konar, R. Pradhan, T. Dey, T. Sapkota, and P. Rai, “Predicting Students’ Grades Using CART, ID3, and Multiclass SVM Optimized by the Genetic Algorithm (GA): A Case Study,” Recent Adv. Hybrid Metaheuristics Data Clust., pp. 85–99, 2020, doi: 10.1002/9781119551621.ch5.
[5] C. D. Casuat and E. D. Festijo, “Predicting Students’ Employability using Machine Learning Approach,” ICETAS 2019 - 2019 6th IEEE Int. Conf. Eng. Technol. Appl. Sci., 2019, doi: 10.1109/ICETAS48360.2019.9117338.
[6] M. Wati, Haeruddin, and W. Indrawan, “Predicting degree-completion time with data mining,”
[7] E. Budiman and N. Dengan, “Performance of Decision Tree C4.5 Algorithm in Student Academic Evaluation,” Lect. Notes Electr. Eng., vol. 488, pp. 380–389, 2018, doi: 10.1007/978-981-10-8276-4.

[8] M. Wati, N. Novirasari, E. Budiman, and Haeruddin, “Multi-Criteria Decision-Making for Evaluation of Student Academic Performance Based on Objective Weights,” in the Third International Conference on Informatics and Computer, 2019, no. 11, pp. 1–5, doi: 10.1109/ica.2018.8780421.

[9] M. Wati, F. M. Lubis, and A. Tejawati, “Penentuan Prioritas Kesejahteraan Keluarga Menggunakan Metode the Extended Promethee II,” Ilk. J. Ilm., vol. 12, no. 1, pp. 71–80, 2020, doi: 10.33096/ilkom.v12i1.528.71-80.

[10] M. Wati, H. S. Pakpahan, and N. Novirasari, “Comparative Analysis of Multi-Criteria Decision Making for Student Degree Completion Time based on Entropy Weighted,” Proc. ICAITI 2018 - 1st Int. Conf. Appl. Inf. Technol. Innov. Towar. A New Paradig. Des. Assist. Technol. Smart Home Care, pp. 56–61, 2019, doi: 10.1109/ICAITI.2018.8686746.

[11] M. E. Hiswati, A. F. O. Gaffar, Rihartanto, and Haviluddin, “Minimum wage prediction based on K-Mean clustering using neural based optimized Minkowski Distance Weighting,” Int. J. Eng. Technol., vol. 7, no. 2, pp. 90–93, 2018, doi: 10.14419/ijet.v7i2.12741.

[12] A. Sarker, S. M. Shamim, M. Shahiduz, Z. M. Rahman, M. Shahiduz Zama, and M. Rahman, “Employee’s Performance Analysis and Prediction using K-Means Clustering & Decision Tree Algorithm Mawlana Bhashani Science and Technology University Employee’s Performance Analysis and Prediction using K-Means Clustering & Decision Tree Algorithm,” Type Double Blind Peer Rev. Int. Res. J. Softw. Data Eng. Glob. J. Comput. Sci. Technol. C, vol. 18, no. 1, 2018.

[13] A. P. Windarto, “Implementation of Data Mining on Rice Imports by Major Country of Origin Using Algorithm Using K-Means Clustering Method,” Int. J. Artif. Intell. Res., vol. 1, no. 2, p. 26, 2017, doi: 10.29099/ijair.v1i2.17.

[14] A. Torrente and J. Romo, “Initializing k-means Clustering by Bootstrap and Data Depth,” J. Classif., 2020, doi: 10.1007/s00357-020-09372-3.

[15] A. E. M. Celestino, D. A. M. Cruz, E. M. O. Sánchez, F. G. Reyes, and D. V. Soto, “Groundwater quality assessment: An improved approach to K-means clustering, principal component analysis and spatial analysis: A case study,” Water (Switzerland), vol. 10, no. 4, pp. 1–21, 2018, doi: 10.3390/w10040437.

[16] M. Li, “Application of CART decision tree combined with PCA algorithm in intrusion detection,” Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS, vol. 2017-Novem, pp. 38–41, 2018, doi: 10.1109/ICSESS.2017.8342859.

[17] M. Z. F. Nasution, O. S. Sitompul, and M. Ramli, “PCA based feature reduction to improve the accuracy of decision tree c4.5 classification,” J. Phys. Conf. Ser., vol. 978, no. 1, 2018, doi: 10.1088/1742-6596/978/1/012058.

[18] A. Lawi and F. Aziz, “Comparison of Classification Algorithms of the Autism Spectrum Disorder Diagnosis,” Proc. - 2nd East Indones. Conf. Comput. Inf. Technol. Internet Things Ind. EICONCIT 2018, no. 1, pp. 218–222, 2018, doi: 10.1109/EICONCIT.2018.8878593.

[19] A. Lawi and F. Aziz, “Classification of credit card default clients using LS-SVM ensemble,” Proc. 3rd Int. Conf. Informatics Comput. ICIC 2018, pp. 1–4, 2018, doi: 10.1109/IAC.2018.8780427.

[20] K. Kusrini, E. T. Luthfi, M. Muqorobin, and R. W. Abdullah, “Comparison of naive bayes and K-NN method on tuition fee payment overdue prediction,” 2019 4th Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE 2019, vol. 6, pp. 125–130, 2019, doi: 10.1109/ICITISEE48480.2019.9003782.