The effect of mining data k-means clustering toward students profile model drop out potential

Windania Purba*, Saut Tamba, Jepronel Saragih

Faculty of Technology and Computer Science, Universitas Prima Indonesia, Indonesia

E-mail: *winda.nia04@gmail.com.

Abstract. The high of student success and the low of student failure can reflect the quality of a college. One of the factors of fail students was drop out. To solve the problem, so mining data with K-means Clustering was applied. K-Means Clustering method would be implemented to clustering the drop out students potentially. Firstly the result data would be clustering to get the information of all students condition. Based on the model taken was found that students who potentially drop out because of the unexciting students in learning, unsupported parents, diffident students and less of students behavior time. The result of process of K-Means Clustering could known that students who more potentially drop out were in Cluster 1 caused Credit Total System, Quality Total, and the lowest Grade Point Average (GPA) compared between cluster 2 and 3.

1. Introduction

The highest of successful and the lowest of unsuccessful students were reflecting from the quality of the university. The university becomes the main social in scoring successful or unsuccessful in high education. The successful or students achievement study could be seen as successful and superior of university. Vise versa, the unsuccessful or the lowest students study was unable of university in doing learning process in high education. One of the problem is students drop out. Drop out was a process of students quit from students status, that caused many things that depend on university itself. The highest number of drop out students in university can be minimize with the university policy to guide and prevent students from drop out to detect students in first education was very important to keep students from drop out. It maybe department organizer education to give guidelines to the students needed [1].

The students understanding information who potentially drop out was important to known and understood. The understanding could be done by find the data to showing the data which had and then clustering to the result of data so showing the pattern or students drop out clustering. This education could be used in helping university to knowing the students situation and become education early in processing of taking a decision to preventive in anticipation of drop out students, to increase the achievement students, to increase curriculum, to increase process of learning and teaching activity and there were another advantages that could be gotten from the result of getting data.

2. Review Literature

Mining data is the using automatic of analysis technique to knowing the previous relationship cannot detected between the item data. Mining data is a process to finding the same data from database/data...
set in different area likely, financial, retail industry, science, statistic, medicine, intelliegece, neurology. The size of data is increase quickly, so the new technology and high speed and algorithm that needed to collect and process of data [2-3]. Clustering is one of data mining that unsupervised. Clustering is one of data mining that unsupervised. Clustering is a process divided data to the cluster based on the similarity. K-Means is a cluster technique data that the point data of existence in cluster is depend on degree of the member [4].

Steps to do clustering with K-Means is: 1) Determining how many cluster that want to shape, where K value is the number of cluster; 2) Determining the centroid cluster first. The begining Centroid was depend on randomly from the data and number of centroid same with the number of cluster; 3) After determined the begining centroid, so very data will find nearby centroid is count every distance of each centroid by using correlations formula between two objects of Euclidean Distance; 4) After counting the distance of data the centroid, so the next steps is clustering of data based on minimum distance. A data will become a member of cluster that has the nearest (small) from the cluster; 5) Based on the clustering, the next is finding the new centroid based on membership from each cluster is counting the cluster average; 6) Back to step 3; 7) Iteration stop means if there is no any moving data. Determine centroid in iteration wil use the formula below: 
\[ \bar{v}_{ij} = \frac{1}{N_i} = \sum_{k=0}^{N_i} x_{ki} \]

(1). Where: 1) \( V_{ij} \) Centroid of average cluster k-i to k-j variable; 2) \( N_i \) The number of members of cluster; 3) k-i, i, k index from cluster; 4) j Index from variable; 5) \( X_{kj} \) The value of data k variable ke-j in the cluster. Determine the correlation between two objects is using Euclidean Distance below: 
\[ d_{\text{Euclidean}}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \]

(2) Where: 1) \( d_{xy} \) the distance of x to cluster y; 2) \( x_i \) The data i in data attribute to n; 3) \( y_i \) The data to-j in data attribute to n. That is the formula which is used to determine the correlation between two objects with Euclidean Distance formula [5-8].

3. Result and Discussion

The mechanism of test would be done was:
To determining the cluster based on data that had been available, it needed a flowchart to easier in determine the counting plot as a plot to find the result of cluster implementation to processing the data. There were some steps. They were:
Figure 1. Flowchart Algorithm K-Means Clustering.

There was variable that used in clustering was Semester Credit System, quality and Grade Predicate Academic (GPA), the students data that belong to drop out in last four years, so their score could be seen in table 1.
Table 1. Student Data.

| No  | NIM          | Name                  | Courses Program         | Credits | Sks  | Quality Total | GPA |
|-----|--------------|-----------------------|-------------------------|---------|------|---------------|-----|
| M1  | 8330303014   | Vera                  | S1 - Teknik Informatika | 14      | 47   | 3.36          |
| M2  | 8330303015   | Caroline Nata Wijaya  | S1 - Teknik Informatika | 14      | 56   | 4.00          |
| M3  | 8330303016   | Silvia                | S1 - Teknik Informatika | 3       | 12   | 4.00          |
| M4  | 8330303017   | Dewi Sandro Maria     | S1 - Teknik Informatika | 14      | 24   | 1.71          |
| M5  | 9330303001   | Dearni Rapita Saragih | S1 - Teknik Informatika | 57      | 154  | 2.70          |
| M6  | 9330303002   | Hamonangan Nasution   | S1 - Teknik Informatika | 88      | 106  | 1.20          |
| M7  | 9330303003   | Dadang Suhenda        | S1 - Teknik Informatika | 61      | 157  | 2.57          |
| M8  | 9330303004   | Paulus P.H Nababan    | S1 - Sistem Infomasi    | 48      | 154  | 3.21          |
| M9  | 9330303006   | Muhammad Tarmizi      | S1 - Sistem Infomasi    | 48      | 144  | 3.00          |
| M10 | 9330303007   | Bunga Andriani Lubis  | S1 - Sistem Infomasi    | 57      | 151  | 2.65          |
| M11 | 9330303008   | Muhammad Fachri       | S1 - Sistem Infomasi    | 48      | 148  | 3.08          |
| M12 | 9330303009   | Rizkina Fitri Fauziah | S1 - Sistem Infomasi    | 48      | 121  | 2.52          |
| M13 | 9330303011   | Melda Riski Harahap   | S1 - Sistem Infomasi    | 57      | 181  | 3.18          |
| M14 | 9330303013   | Eka Fitri Siregar     | S1 - Sistem Infomasi    | 57      | 165  | 2.89          |
| M15 | 9330303015   | Siskia Lasmana Togatorop | S1 - Sistem Infomasi | 57      | 160  | 2.81          |
| M16 | 9330303016   | Rohani Tinambunan     | S1 - Sistem Infomasi    | 62      | 138  | 2.23          |
| M17 | 9330303018   | Hanova Margareth      | S1 - Sistem Infomasi    | 57      | 168  | 2.95          |
| M18 | 9330303019   | Sherly Minarti Nainggolan | S1 - Sistem Infomasi | 23      | 0    | 0.00          |
| M19 | 9330303020   | Musmulyadi            | S1 - Sistem Infomasi    | 57      | 164  | 2.88          |
| M20 | 9330303022   | Robin                 | S1 - Sistem Infomasi    | 57      | 163  | 2.86          |
| M21 | 9330303024   | Eddy                  | S1 - Sistem Infomasi    | 57      | 176  | 3.09          |
| M22 | 9330303025   | Feriyanto Tanwijaya   | S1 - Sistem Infomasi    | 57      | 162  | 2.84          |
| M23 | 9330303027   | Melianty Sitompul     | S1 - Sistem Infomasi    | 65      | 173  | 2.66          |
| M24 | 9330303028   | Shahani Singh         | S1 - Sistem Infomasi    | 71      | 163  | 2.30          |
| M25 | 9330303031   | Jovin                 | S1 - Sistem Infomasi    | 54      | 168  | 3.11          |
| M26 | 9330303032   | Yoko Junior           | S1 - Sistem Infomasi    | 57      | 136  | 2.39          |
| M27 | 9330303037   | Ferry Wijaya          | S1 - Sistem Infomasi    | 54      | 156  | 2.89          |
| M28 | 9330303041   | Muhammad Ramadhan Lubis | S1 - Sistem Infomasi | 54      | 149  | 2.76          |
| M29 | 9330303048   | Radius Hia            | S1 - Sistem Infomasi    | 47      | 128  | 2.72          |
| M30 | 9330301302   | Winter Pandiangan     | S1 - Sistem Infomasi    | 14      | 0    | 0.00          |
| M31 | 10330303001  | Abdul Ghani Abadi     | S1 - Teknik Informatika | 98      | 322  | 3.29          |
| M32 | 10330303004  | Armando Pernando Sitorus | S1 - Teknik Informatika | 21      | 0    | 0.00          |
| M33 | 10330303005  | Bayakta Sebayang      | S1 - Teknik Informatika | 159     | 200  | 1.26          |
| M34 | 10330303008  | Indra Putra           | S1 - Teknik Informatika | 104     | 291  | 2.80          |
| M35 | 10330303009  | Jeffri Aritonang      | S1 - Teknik Informatika | 101     | 302  | 2.99          |
| M36 | 10330303010  | Jeffry Suherman       | S1 - Teknik Informatika | 98      | 338  | 3.45          |

To do the cluster data become cluster it was done some steps, there were:
1. Determine the specific cluster. In this research the data would be clustering become three clusters. In cluster 1 had a characteristic data. That was students that potentially in drop out.
2. In cluster 2 belonged to good characteristic data. In cluster 3 was gotten to characteristic very good students.

Determined the centroid cluster begining. In this research the string point was determined randomly and gotten in the starting point every cluster. It could be seen in table 4.3. where cluster point that taken.

Table 2. The Central Point Of The Initial Cluster.

| 1st Data | Cluster | Credit Total | Quality Total | GPA |
|----------|---------|--------------|---------------|-----|
| 4        | C1      | 14           | 24            | 1.71|
After determining the centroid, so every data would found the near centroid by counting the distance every data to each centroid by using correlation formula between two objects of Euclidean Distance. There was a count of manual centroid. It had just count 36 data only.

Table 3. Clustering Result.

| Cluster | Member of Cluster | Number of members |
|---------|------------------|------------------|
| 1       | [1, 2, 3, 4, 18, 30, 32] | 7 member         |
| 2       | [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 33] | 25 member        |
| 3       | [31, 34, 35, 36] | 4 member         |

From the conclusion above it can be shown that students that have potentially dropped out more belong to cluster 1. It was caused of total od System Credit Semester, Quality Total and the lowest of GPA. It could be compared between cluster 2 and 3.

4. Conclusion
Based on the the discussion and result of test, so it could be concluded:
1. Based on the result of the test so the clustering was taken from students position that potentially drop out was 3 clusters. They were students who very drop out potentially, drop out potentially, and not belong to drop out.
2. Clustering result could be used to conclude in determine students who potentially drop out.
3. To predict the students who potentially drop out, the data which was analyzed consist of scoring students data, quality and GPA.
4. The result of counting for all datas (36 records) based on the test that showed thet cluster 0 consists of 7 members that shown the students who potentially drop out. Then cluster 1 that consist of 25 members shown the well students cluster. Meanwhile cluster 2 shown the very good students with 4 members.

References
[1] Dekker, G.W., 2009, Prediction student drop out: A case study, USA, Academic Press, 2nd international Conference On Educational Data Mining, Cordoba, Spain.
[2] Ramesh Singh Yadava and P.K.Mishra. (2012). “Performance Analysis of High Performance k-Mean Data Mining Algorithm for Multicore Heterogeneous Compute Cluster.” Vol. 2 No.4.
[3] Tutik Khotimah 2014. “Pengelompokan Surat Dalam Al Qur’an Menggunakan Algoritma K-Means.” Ed. Jurnal Simetris, Vol 5 No 1.
[4] Deka Dwawinta 2014. “Klasterisasi Judul Buku dengan Menggunakan Metode K-Means.” Ed. Seminar Nasional Aplikasi Teknologi Informasi (SNATT).
[5] Bondu Venkateswarlu and Prof G.S.V.Prasad Ratu. 2013. “Mine Blood Donors Information through Improved K-Means Clustering.” International Journal of Computational Science and Information Technology (IJCSTITY) Vol.1, No.3.
[6] Johan Oscar Ong. 2013. “Implementasi Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing Presiden University.” Ed. Jurnal Ilmiah Teknik Industri, Vol.12, No.1.
[7] Afrisawati. 2013. “Implementasi Data Mining Pemilihan Pelanggan Potensial Menggunakan Algoritma K-Means.” Ed. Pelita Informatika Budi Darma, Volume : V,No. 3.
[8] Budanis Dwi Meilani dan Nofi Susanti, 2014, “Aplikasi Data Mining Untuk Menghasilkan Pola Kehilusan Siswa Dengan Metode Naive Bayes.” Ed. Jurnal LINK Vol 21/No. 2.
[9] Dr. M.P.S. Bhatia and Deepika Khurana. 2013. “Experimental study of Data clustering using k-Means and modified algorithms.” International Journal of Data Mining & Knowledge Management Process (IJDMP). Vol. 3 No.3.

[10] Er. Nikhil Chaturvedi and Er. Anand Rajavat. 2013. “An Improvement in K-mean Clustering Algorithm Using Better Time and Accuracy.” International Journal of Programming Language and Aplications (IJPLA) Vol. 3 No.4.

[11] Fadlina. 2014. “Data Mining untuk Analisa Tingkat Kejahatan Jalan Dengan Algoritma Association Rule Metode Apriori (Studi Kasus Di Polsekta Medan Sunggal).” Volume III No.1.

[12] Goldie Gunadi and Dana Indra Sensuse. 2012. “Penerapan Metode Data Mining Market Basket Analysis Terhadap Data Penjualan Produk Buku Dengan Menggunakan Algoritma Apriori dan Frequent Pattern Growth (FP-Growth) : Studi Kasus Percetakan PT. Gramedia.” Ed. Jurnal Telematika MKom Vol.4 No.1.

[13] Lindawati. 2008. “Data Mining Dengan Teknik Clustering Dalam Pengklasifikasian Data Mahasiswa Studi Kasus Prediksi Lama Studi Mahasiswa Universitas Bina Nusantara.” Ed. Seminar Nasional Informatika.

[14] Mujib Ridwan, 2013. “Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naïve Bayes Classifier.” Ed. Jurnal EECCIS Vol.7, No.1.

[15] Oyelade. O. J, 2010. “Application of k-Means Clustering algorithm for prediction of Students’ Academic Performance.” International Journal of Computer Science and Information Security. Vol. 7 No.1.

[16] Rajashree, 2010. “A hybridized K-means clustering approach for high dimensional dataset.” International Journal of Engineering, Science and Technology Vol. 2, No. 2.

[17] Ramzi A. Haraty, 2015. “An Enhanced k-Means Clustering Algorithm for Pattern Discovery in Healthcare Data.” Hindawi Publishing Corporation International Journal of Distributed Sensor Networks Volume 2015, Article ID 615740.