Jurnal Mantik, 5 (2) (2021) 1424-1428

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik/index

Clusterization Using K-Means Clustering Algorithm In Predicting Student Graduation Time

Syaiful Zuhri Harahap[1], Masrizal[2]

[12]Information System, Faculty of Science & Technology
[12] Labuhanbatu University, Jl. S.M. Raja No.126 A Aek Tapa Rantauprapat, Kab. Labuhanbatu Sumatera Utara, Indonesia

E-mail: syaifulzuhriharahap@gmail.com[1], masrizal120405@gmail.com[2]

ARTICLEINFO

Article history:

Received: July 19, 2021
Revised: August 05, 2021
Accepted: August30, 2021

Keywords:

Data Mining,
K-Means Clustering,
Predicting Student Graduation

ABSTRACT

Education at the college level is a suggestion that students can get a degree in order to have knowledge in the field of computer science. Taking a decision from a BIG DATA for Predicting student graduation time is useful to provide a means of knowing the estimated time of a student's graduation by seeing which students fall into a certain cluster based on the parameters of the Cumulative Achievement Index (GPA) and attendance. It is hoped that it can help the campus and students to predict the graduation rate on time and to improve the reputation for the campus itself and timely graduation for students so that their graduation is not late, besides that the campus can do things that need to be done if they are predicted pass not on time like by making motivation and other things.

Copyright © 2021 Jurnal Mantik.
All rights reserved.

1. Introduction

For students in the information systems study program so that in carrying out the lectures they live they can carry out their studies on time. if students are not notified from now on then students who have information systems study programs at the Labuhanbatu University so that they undergo lectures who are less enthusiastic in lectures which can make a target Achievement Index (IP) for the semester and they may be lazy and have an effect on them not doing their job. study on time and make the study program no longer in demand because students graduate in a long time. to analyze the K-means Clustering algorithm for the perception of students of the Labuhanbatu University information systems study program in undergoing a good lecture process so that they graduate on time.

2. Method

The flow of data and information grew significantly in various sizes and media, which was then referred to as Big Data. Big data has been identified since 1944. Big data was identified by a librarian named Fremont Rider from Westleyan University in the United States. In that year he has estimated that libraries in American universities will grow to 200 million volumes by 2020. [1]

Student graduation is a moment that is eagerly awaited by every student and graduating on time is something that is expected.[3] Not only that, each campus also wants its students to graduate on time, because this is useful for the campus to improve its reputation and accreditation. But in reality the students can't predict it, they are just exploring the course of the lecture, so the conclusion is that their graduation time is too late. This can be seen from the imbalance in the percentage of student graduation from year to year as well as the comparison of the number who have graduated and have not graduated in some batches. One of the methods used to predict the time of graduation for these students is to use the K-Means Clustering

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

method. Cluster analysis is a multivariate method that has the main goal of grouping objects based on their characteristics. Cluster analysis classifies objects so that each object that is very close in similarity to other objects is located in the same cluster. Cluster analysis is one of the useful analysis tools such as summarizing information. In summarizing this information, we can try the method of grouping objects based on certain similarities between the objects to be studied.

2.1 Knowledge Discovery In Database (KDD)

KDD is a method used to obtain knowledge from existing databases. The results of the knowledge obtained can be used for a knowledge base that is used in making decisions. In more detail, the KDD process is as shown in the following picture which was adopted from.[5]

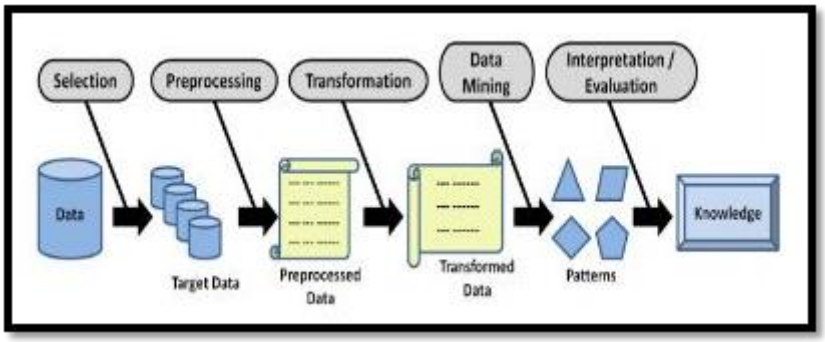

Fig 1. Process Knowledge Discovery In Database (KDD)

a. Selection

Selection is used to determine the variables to be taken so that there are no similarities and unnecessary repetition occurs in data mining processing.

b. Preprocessing

In preprocessing there are two stages, namely as follows:

1) Data Cleaning Eliminate unnecessary data such as handling missing values, noise data and handling inconsistent and relevant data.
2) Data Integration Performed on attributes that identify unique entities.

c. Transformation

Changing the data according to the appropriate extension format in data mining processing because some methods in data mining require a special format before they can be processed in data mining.

d. Data mining

The main process is the method applied to gain new knowledge from the processed data. In this study, a clustering technique was applied, namely the K-Means Clustering method.

e. Evaluation/ Interpretation

Identify interesting patterns into the identified knowledge base. At this stage, generate typical patterns and predictive models that are evaluated to assess existing studies that have met the desired target.

f. Knowledge

The resulting patterns will be presented to the user. At this stage the new knowledge generated can be understood by everyone who will be used as a reference for decision making.[5]

2.2 The K-means algorithm process is as follows

a. Determine k so how many clusters are formed

b. Raise k Centroids as the center point of the initial cluster randomly

Determination of the initial centroid is done randomly from the available objects as many as k clusters.

$$v = \frac{\sum_{i=1}^{n} x_i}{n} \quad ;i = 1,2,3,\ldots n$$

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Where
V: centroid in cluster Xi: object i
n : the number of objects / number of objects that are members of the cluster

c. Calculate the distance of each object to each centroid of each cluster.[6]

$$d(x,y) = \|x - y\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad ;i = 1,2,3,\ldots,n$$

Where
Xi: i-th object Yi: i-th y power
n: number of objects

d. Make the object into the nearest centroid.

In general, during iteration there are two methods for assigning objects to each cluster, namely hard k-means, where each object needs to be specifically declared to be part of the cluster by calculating its proximity to the cluster core.

e. Building iterations, the next step is to determine the new centroid by making equations

f. Create a loop in step 3 if the position of the new centroid is not the same

Do this by checking and comparing the group assignment matrix of the previous iteration with the group assignment matrix of the current iteration.

3. Result and Discussion

3.1 Implementasi RapidMiner 9.10.0

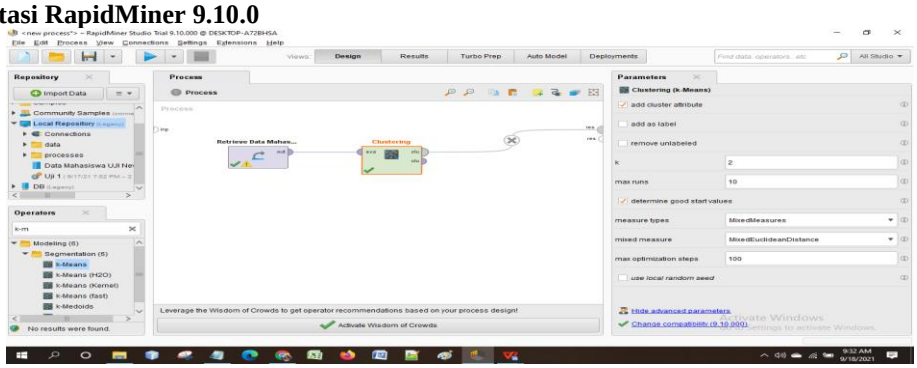

Fig 2. Implementation of RapidMiner Student Data for the 2019/2020 Information System

From Figure 1 above, we can see the operator needs that we will use to manage student data, information systems that will predict student graduation. First, we need a Retrieve that functions as a place for the data we input, in this case the researcher uses student data in the information systems study program, faculty of science and technology, Labuhanbatu University 2019/2020, the file extension is in excel. Second, we use the K-Means Clustering operator which functions to predict student graduation, in this case the researchers set the cluster as much as 2, namely: cumulative achievement index and attendance to predict student graduation so that students know how to graduate on time, if later categorized to pass on time or pass not on time.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

3.2 Claster Model

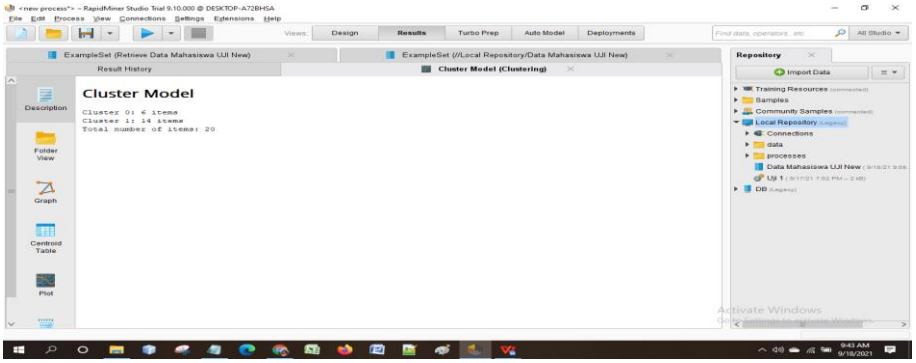

Fig 3. Cluster Model

In Figure 3 it is explained that the number of data/items that we input is 20 students, cluster 0 is students who are predicted to graduate not on time totaling 6 people, cluster 1 is students who are predicted to graduate on time totaling 14 people.

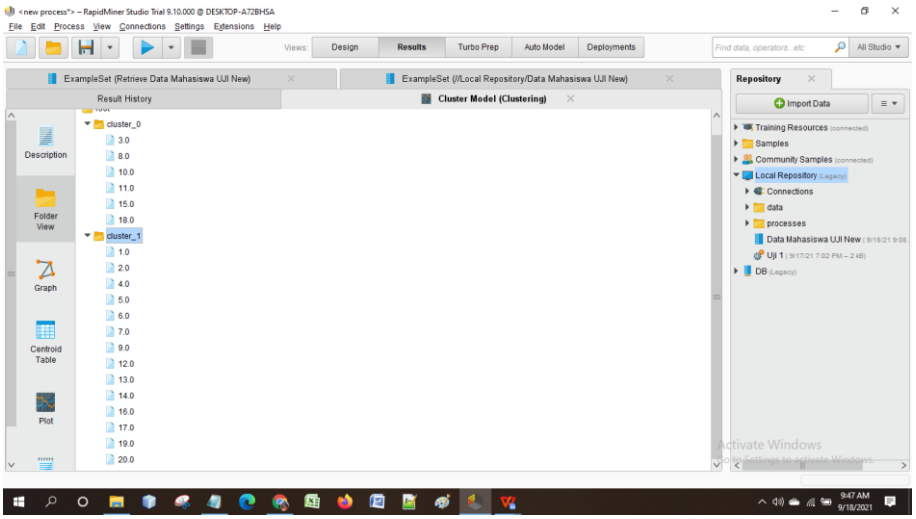

Fig 4. Cluster 0 and Cluster 1

4. Conclution

Based on the results obtained from this study with the K-Means clustering method. The result is to get 2 categories in predicting student graduation not graduating on time and graduating on time . This is evidenced by the calculation of the closest distance based on the determination of the centroid value randomly using the Euclidean Distance formula, on the number of students of the information systems study program, faculty of science and technology, Labuhanbatu University and student achievement and attendance index. So there are 6 students of the information system study program who are predicted to graduate not on time, 14 people graduate on time, that's the result of the implementation of K-Means Clustering on the RapidMiner9.10.0 software.

5. References

[1] R. P. S. Putri and I. Waspada, "Penerapan Algoritma C4.5 pada Aplikasi Prediksi Kelulusan Mahasiswa Prodi Informatika," Khazanah Inform. J. Ilmu Komput. dan Inform., vol. 4, no. 1, pp. 1–7, 2018.

[2] A. P. Narendra, "Data Besar, Data Analisis, dan Pengembangan Kompetensi Pustakawan," Rec. Libr. J., vol. 1, no. 2, pp. 83–93, 2015.

[3] H. Priyatman, F. Sajid, and D. Haldivany, "Klasterisasi Menggunakan Algoritma K-Means Clustering untuk Memprediksi Waktu Kelulusan," JEPIN (Jurnal Edukasi dan Penelit. Inform., vol. 5, no. 1, pp. 62–66, 2019.

Jurnal Mantik is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

[4] M. Linda, "penerapan datamining dalam mengelompokkan kunjungan wisatawan ke objek wisata unggulan di prov Dki jakarta dengan k-means," jiska (Jurnal Inform. Sunan Kalijaga), vol. 2, no. 3, pp. 167–174, 2018.
[5] Gustientiedinaa, M. H. Adiyaa, and Y. Desnelitab, "Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan Pada RSUD Pekanbaru," J. Nas. Teknol. dan Sist. Inf. Univ. Andalas, vol. 5, no. 1, pp. 17–24, 2019.
[6] T. Rismawan and S. Kusumadewi, "Aplikasi K-Means Untuk Pengelompokkan Mahasiswa Berdasarkan Nilai Body Mass Index (Bmi) & Ukuran Kerangka," Proc. Semin. Nas. Apl. Teknol. Inf., vol. 1, no. 1, pp. 43–48, 2008.
[7] F. Nur, M. Zarlis, and B. B. Nasution, "Penerapan Algoritma K-Means Pada Siswa Baru Sekolahmenengah Kejuruan Untuk Clustering Jurusan," infotekjar j. Nas. Inform. Dan teknol. Jar., vol. 1, no. 2, pp. 100–105, 2015.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Jurnal Mantik is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).