Edulib 13(2) (2023) 176-185

JURNAL UPI

EDULIB

Journal of Library and Information Science

Journal homepage: http://ejournal.upi.edu/index.php/edulib/index

Library Book Data Clustering with The K-Means Algorithm To Enhance Book Loans

Denni Pratama, Sari Hermawan, Christina Juliane*

Faculty of Economics and Business, Universitas Brawijaya, Indonesia

*Correspondence: E-mail: khusaini@ub.ac.id

ABSTRACT

When compared to the year before the pandemic (2018–2019), the number of book borrowings at the Indonesian University of Education Library was not optimal as of June. Borrowing in 2022 can still be increased by arranging the most borrowed books in one group. The purpose of this research is to classify books more optimally, which will be applied in the arrangement of books. Optimal book arrangement allows library visitors to more efficiently find books based on the books that are borrowed most often, so that they are interested in borrowing other books that are in the same group. Data mining is a term used to describe knowledge in a database from a storage by finding patterns and trends in data through examination with statistical and mathematical techniques. Clustering is a method of data mining that can be used to determine clusters of data. One algorithm that can be used is K-Means. From the clustering obtained, there are two (2) clusters of groups. Book titles in cluster 0 contain book titles related to research methodology, statistics, measurement scales, assessments, and learning evaluations. While cluster 1 tends to contain psychology, counseling, guidance, religion, philosophy, management, economics, and history, This data can be used by librarians in prioritizing the purchase of a collection of books in the next procurement.

ARTICLE INFO

Article History:
Submitted/Received 28 Jul 2023
First Revised 05 Sep 2023
Accepted 12 Oct 2023
First Available online 16 Nov 2023
Publication Date 01 Nov 2023

Keyword:
Clustering;
Library;
K-Means;
Data Mining.

1. INTRODUCTION

Books are the second most sought-after reference source after electronic journal articles (Nurfadillah & Ardiansah, 2021). When the Covid-19 pandemic hit in 2020-2021, of the 1,000 most lent book titles in 2020, 25,040 were borrowed, while in 2021 there were only 1,926 borrowed. From January 2022 to June 2022 (New Normal Era / Post-pandemic), there have been 17,376 loan transactions. When compared to 2018 with 44,046 and 2019 with 67,760 loan transactions, of course book loan transactions in 2022 can still return. One way is by arranging the most lent books into one group (Karputri & Yustanti, 2022).

One of the clustering methods is K-Means (Sudarsono & Lestari, 2021). The K-Means algorithm divides data into several groups and can accept input data as data without class labels (Febrianto et al., 2021). This algorithm divides data into clusters so that data with the same characteristics are grouped into the same cluster and data with different characteristics are grouped into other groups (Nasir, 2020).

This study aims to optimize the arrangement of books so that visitors can efficiently find books based on the most frequently borrowed books so that they are interested in borrowing other books in the same group. The clustering pattern obtained can also be used by librarians in prioritizing the purchase of book collections in the next procurement.

Data mining is a process that uses statistical techniques, mathematics, artificial intelligence, and machine learning to extract and identify useful information and knowledge collected from databases of different sizes (Siregar, 2018). One of the functions of data mining is clustering (Mahmuda et al., 2017). Data mining is also called Knowledge Discovery in Database (KDD) which is the activity of collecting and using data to determine the regularity, patterns and relationships of large data sets (Yunita, 2018). Data Mining can be divided into four groups, namely Prediction Modeling, Cluster analysisAssociation analysis and anomaly detection (Fatmawati & Windarto, 2018)

Clustering is a technique for grouping records in a database based on certain criteria (Sibuea & Sapta, 2017). The clustering results are provided to end users to provide insight into what is happening in the database where there is a high level of similarity between objects in the cluster (Adhitama et al., 2020). There are many clustering methods whose use depends on the type of data being clustered and the purpose of the application (Silitonga & Morina, 2017).

K-Means is a data clustering method that combines data based on similar characteristics. The goal of data clustering is to minimize the objective function of each cluster, thereby maximizing the variation of data between clusters. The K-Means algorithm can divide data into several groups according to their similarities, so that data with similar qualities are grouped in one cluster and data with different characteristics are grouped in another cluster.

The K-Means algorithm is carried out in several steps, including the following steps (I) Determine the value of k or the number of clusters you want to form from the dataset. (ii) Determine the center value (centroid), this value is determined randomly. (iii) Calculate the distance between the centroid point and the point of each object using Euclidean Distance. Euclidean Distance is the straight line distance between two points in Euclidean space. (iv) Group objects based on the distance to the nearest centroid. (v) Repeat steps 2 to 4, iterate until the centroid is optimal.The Euclidean formula is as follows .

$$De = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2}$$

Description :

DOI: https://doi.org/10.17509/edulib.v13i1.50003
p- ISSN 2528-1410 e- ISSN 2527-8045

De : Euclidean Distance
i: Number of objects
(x, y) : Object coordinates
(s, t) : Centroid coordinates

2. METHODS

Research Stages

The stages in this research begin with preparing test data using KDD. Next, determine the number of K tests to be tested. Then test the data using the K-Means algorithm based on K. Continued by evaluating the cluster using DBI, determining the optimal number of clusters and drawing information from the grouping results. The stages of this research are illustrated in Figure 1

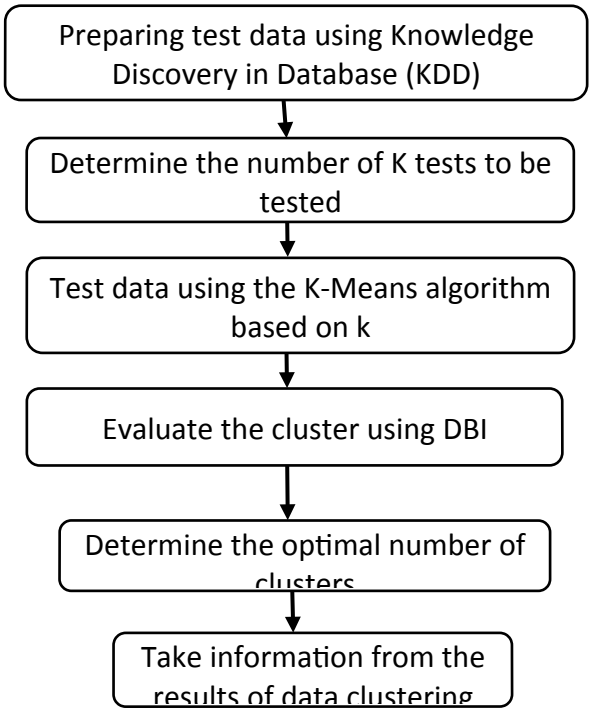

Figure 1 Research Framework

Conducting Test Data Preparation

Research begins with data preparation for testing. The data prepared must be in accordance with the steps in KDD. Starting with data selection, data cleaning process (Pre-Processing), data transformation process , data mining process and searching for patterns or information from selected data, and the last step is interpretation and evaluation that leads to new useful information (Interpretation/Evaluation) (Firdaus et al., 2021).

DOI:https://doi.org/10.17509/edulib.v13i1.50003
p- ISSN 2089-6549 e- ISSN 2528-2182

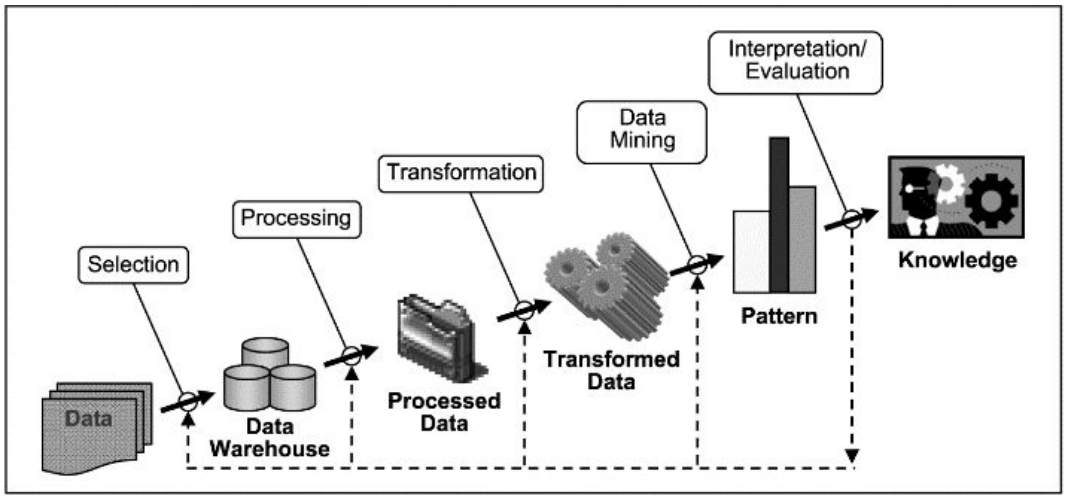

Figure 2 Knowledge Discovery of Database Stages

Determining the Number of K-Means

After the test data passes the KDD stage process, the data is ready to be used for the next stage, namely determining the number of K tests to be tested.

K-Means Algorithm Testing

After the number of K tests is determined, the data is tested with the K-Means algorithm based on the K value. The K-Means algorithm is a data analysis method that groups data with a partition system. The K-Means algorithm is a model that uses the center to produce clusters, where the centroid is the middle part of a cluster.

Cluster Evaluation Using the Davies Bouldin Index (DBI)

After going through the K-Means algorithm process, the data clustering results for each K will be confirmed by looking at the DBI value results. DBI is a scale function of the number of cluster dispersions to divide the cluster. The approach to measuring DBI is to maximize the distance between clusters and minimize the distance within the cluster. The smaller the DBI value, the more optimal the cluster scheme.

3. RESULTS AND DISCUSSION

Data Selection

This study uses data obtained from the report of the 1,000 most borrowed books from January 2018 to June 2022 at the Library of the Indonesian National Education University. The parameters used are book title, author name, number of loan transactions in one year and year of borrowing. The clustering process is carried out to find samples of the most borrowed titles as shown in Table 1.

Table 1 Book Borrowing Transaction Data

No	Author	Book Title	Loan Amount	Year
1	Sardiman, A.M.	Interaksi Dan Motivasi Belajar Mengajar	141	2022
2	Arikunto, Suharsimi	Prosedur Penelitian	139	2022
3	Sugiyono	Metode Penelitian Kuantitatif, Kualitatif dan R & D	130	2022

DOI: https://doi.org/10.17509/edulib.v13i1.50003
p- ISSN 2528-1410 e- ISSN 2527-8045

No	Author	Book Title	Loan Amount	Year
4897	Teguh Amor Putria	Telusur Bandung	24	2018
4898	Perkins, Margaret	Becoming a Teacher of Reading	24	2018
4899	Geldard, Kathryn	Konseling Keluarga	24	2018

In Rapidminer use the Read Excel operator as shown in figure 3.

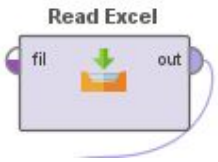

Figure 3 Data Selection

Data Cleaning

At the data cleaning stage, attribute construction is carried out (Heri Cahyana & Sasmito Aribowo, 2018). This is done because there is incomplete author and title data. After the data is repaired, when the data is imported into Rapidminer, no missing or error data is found as shown in Figure 4.

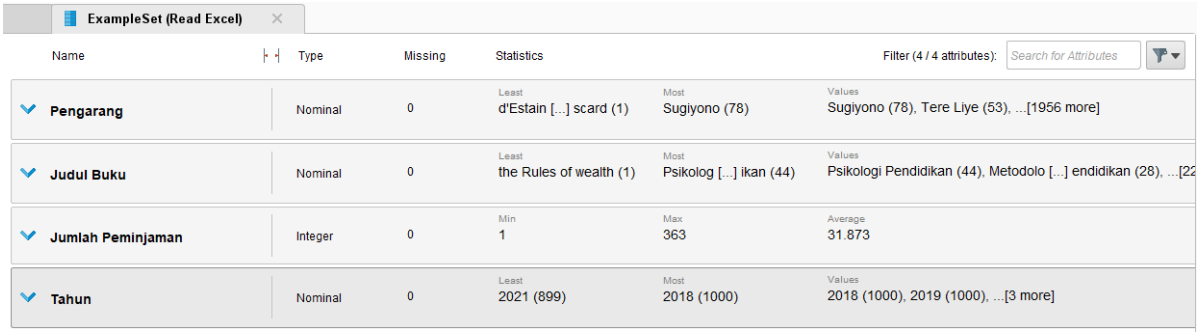

Figure 4 No Missing Data

Data Transformation

Data transformation is carried out using the Nominal to Numerical operation in Rapidminer as shown in Figure 5. The attributes transformed are the author, book title and year attributes.

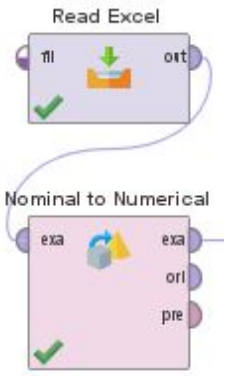

Figure 5. Operators Nominal to Numerical

DOI:https://doi.org/10.17509/edulib.v13i1.50003
p- ISSN 2089-6549 e- ISSN 2528-2182

Data Mining Process

The data mining process is carried out using clustering operators and performance operators with settings as shown in table 2 (Clustering K-Means) and table 3 (Cluster Distance Performance).

Table 2 K-Means Clustering Operator Settings

Parameter	Options	Description.
K	2 -6 cluster	Number of specified clusters
Max runs	10	Maximum number of K-Means runs (default)
Measure types	Numerical Measures	Type of numeric measurement
Numerical measure	Euclidean Distance	Distance calculation from 2 points
Max optimization steps	100	Maximum number of iterations (default)

Table 3 *Cluster Distance Performance Operator Settings*

Operator	Parameter	Options	Description.
Cluster Distance Performance	Main criterion	Davies Bouldin	Criteria selected for clustering evaluation
	Normalize	Checked	Results normalized
	Maximize	Checked	Results maximized

Figure 6 shows the data mining process carried out on Rapidminer. It begins with input dataset, nominal to numerical conversion, using multiply, performing clustering K 2 to K 6, and finally performing performance to find the DBI value of each K.

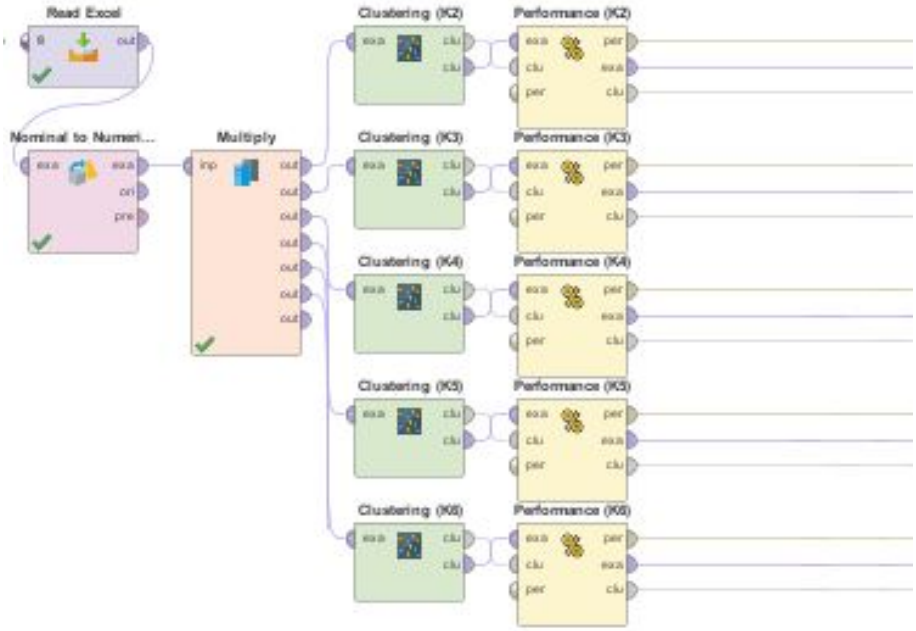

Figure 6 Data Mining Process on Rapidminer

The results of the DBI test on the process model with experiments K = 2 to K = 6 can be seen in table 4.

DOI: https://doi.org/10.17509/edulib.v13i1.50003
p- ISSN 2528-1410 e- ISSN 2527-8045

Table 4 DBI Test Results

***Cluster* (K)**	**DBI**	**Jumlah Anggota *Cluster***
2	0.61	C 0: 2882 C 1: 2017 Total keseluruhan: 4899
3	0.73	C 0: 1145 C 1: 1624 C 2: 2130 Total keseluruhan: 4899
4	0.71	C 0: 2073 C 1: 373 C 2: 1307 C 3: 1146 Total keseluruhan: 4899
5	0.69	C 0: 958 C 1: 1431 C 2: 329 C 3: 968 C 4: 1213 Total keseluruhan: 4899
6	0.73	C 0: 729 C 1: 1373 C 2: 791 C 3: 325 C 4: 1086 C 5: 595 Total keseluruhan: 4899

Evaluation

Based on table 4 shows that each cluster has a different DBI value. DBI testing from the number of K = 2 to K = 6 based on the process model used obtained varying DBI values. In accordance with its function, namely DBI can measure the validity of a cluster, the cluster results are more optimal if the DBI value is smaller or closer to 0. In this study, the smallest DBI value in the Davies Bouldin Performance test was 0.614. This value was obtained at K = 2 and the members of cluster 0 were 2,882 and the members of cluster 1 were 2,017 members. Figure 7 shows some members of each cluster.

DOI:https://doi.org/10.17509/edulib.v13i1.50003
p- ISSN 2089-6549 e- ISSN 2528-2182

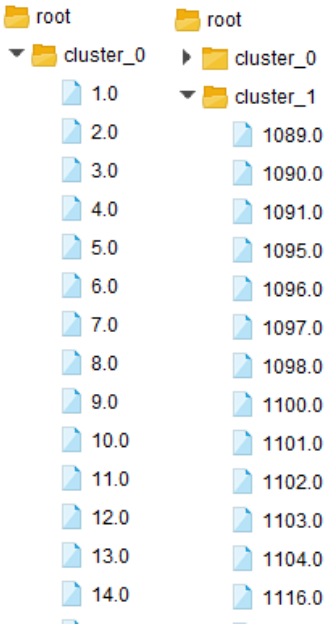

Figure 7 Members of Each Cluster

Knowledge

From Figure 7, the book data that has been clustered into 2 (two) clusters can be seen. If depicted in a scatter or bubble plot. Cluster 0 is dominated by light blue gradated with dark blue while cluster 1 is dominated by cream gradated with a color that tends to orange. This scatter plot shows the cluster against the number of book loans with the points on the plot representing the book title.

The book titles in cluster 0 contain book titles related to research methodology, statistics, measurement scales, assessment, and learning evaluation. While in cluster 1 tends to contain psychology, counseling guidance, religion, philosophy, management, economics and history as shown in the scatter plot in figure 8.

Figure 8 Distribution Of Cluster K 2

DOI: https://doi.org/10.17509/edulib.v13i1.50003
p- ISSN 2528-1410 e- ISSN 2527-8045

4. CONCLUSION

From the library book lending data, the clustering results using the K-Means algorithm are divided into 2 (two) clusters. Cluster 0 mostly contains books related to research and learning methodology, while cluster 1 contains books on psychology, religion, philosophy, economic management and history. The results of this clustering can be used as a reference for librarians in arranging books so as to attract visitors to borrow other similar books that interest them.

5. REFERENCES

Adhitama, R., Burhanuddin, A., & Ananda, R. (2020). PENENTUAN JUMLAH CLUSTER IDEAL SMK DI JAWA TENGAH DENGAN METODE X-MEANS CLUSTERING DAN K-MEANS CLUSTERING DETERMINING VOCATIONAL IDEAL CLUSTER NUMBER IN CENTRAL JAVA WITH X-MEANS CLUSTERING AND K-MEANS CLUSTERING METHODS. *Jurnal Informatika Dan Komputer) Akreditasi KEMENRISTEKDIKTI, 3*(1). https://doi.org/10.33387/jiko

Fatmawati, K., & Windarto, A. P. (2018). DATA MINING: PENERAPAN RAPIDMINER DENGAN K-MEANS CLUSTER PADA DAERAH TERJANGKIT DEMAM BERDARAH DENGUE (DBD) BERDASARKAN PROVINSI. *CESS (Journal of Computer Engineering System and Science), 3*(2), 173–178. .

Febrianto, A., Achmadi, S., & Sasmito, A. P. (2021). PENERAPAN METODE K-MEANS UNTUK CLUSTERING PENGUNJUNG PERPUSTAKAAN ITN MALANG. *Jurnal Mahasiswa Teknik Informatika), 5*(1).

Firdaus, E. A., Maulani, S., & Dharmawan, A. B. (2021). PENGUKURAN MINAT BACA MAHASISWA DENGAN METODE CLUSTERING DI PERPUSTAKAAN AKADEMI KEPERAWATAN RS. DUSTIRA CIMAHI MENGGUNAKAN DATA MINING. *JURNAL NUANSA INFORMATIKA, 15*(1).

Heri Cahyana, N., & Sasmito Aribowo, A. (2018). *Metode Data Mining K-Means Untuk Klasterisasi Data Penanganan Dan Pelayanan Kesehatan Masyarakat*. Seminar Nasional Informatika Medis (Snimed) 2018, 24–31.

Karputri, D. L., & Yustanti, W. (2022). Analisis Klastering Buku sebagai Evaluasi untuk Peningkatan Minat Baca Perpusatakaan SMAN 1 Grogol. *Journal of Emerging Information Systems and Business Intelligence, 3*(3).

Mahmuda, F., Armys Roma Sitorus, M., Widyastuti, H., & Ely Kurniawan, D. (2017). Clustering Profil Pengunjung Perpustakaan (Studi Kasus Perpustakaan BP Batam). *Journal of Applied Informatics and Computing (JAIC), 1*(1).

Nasir, J. (2020). Penerapan Data Mining Clustering Dalam Mengelompokan Buku Dengan Metode K-Means. *Jurnal SIMETRIS, 11*(2).

Nurfadillah, M., & Ardiansah, A. (2021). PERILAKU PENCARIAN INFORMASI MAHASISWA DALAM MEMENUHI KEBUTUHAN INFORMASI SEBELUM DAN SAAT PANDEMI COVID-19. *Fihris: Jurnal Ilmu Perpustakaan Dan Informasi, 16*(1), 21.

Sibuea, F. L., & Sapta, A. (2017). PEMETAAN SISWA BERPRESTASI MENGGUNAKAN METODE K-MEANS CLUSTERING. *JURTEKSI (Jurnal Teknologi Dan Sistem Informasi), IV*(1), 85–92.

DOI:https://doi.org/10.17509/edulib.v13i1.50003
p- ISSN 2089-6549 e- ISSN 2528-2182

Silitonga, P. D. P., & Morina, I. S. (2017). Klusterisasi Pola Penyebaran Penyakit Pasien Berdasarkan Usia Pasien Dengan Menggunakan K-Means Clustering. *Jurnal TIMES, VI*(2), 22–25.

Siregar, M. H. (2018). KLASTERISASI PENJUALAN ALAT-ALAT BANGUNAN MENGGUNAKAN METODE K-MEANS (STUDI KASUS DI TOKO ADI BANGUNAN). *JURNAL TEKNOLOGI DAN OPEN SOURCE, 1*(2), 83–91.

Sudarsono, B. G., & Lestari, S. P. (2021). Clustering Penerima Beasiswa Yayasan Untuk Mahasiswa Menggunakan Metode K-Means. *JURNAL MEDIA INFORMATIKA BUDIDARMA, 5*(1), 258.

Yunita, F. (2018). PENERAPAN DATA MINING MENGGUNKAN ALGORITMA K-MEANS CLUSTRING PADA PENERIMAAN MAHASISWA BARU (STUDI KASUS : UNIVERSITAS ISLAM INDRAGIRI). *Jurnal SISTEMASI, 7*(3), 238–249.

DOI: https://doi.org/10.17509/edulib.v13i1.50003
p- ISSN 2528-1410 e- ISSN 2527-8045