Implementation of Data Mining Clustering Using the K-Means Method in Grouping Library Books

Maria Ulfah*[1)], Andi Sri Irtawty[2)]
[1,2] *Jurusan Teknik Elektro, Politeknik Negeri Balikpapan. Balikpapan. Indonesia*

E-mail: [*1)]maria.ulfah@poltekba.ac.id

Abstract: The K-Means algorithm is an iterative grouping algorithm in partitioning the data set into a number of clusters that have been set at the beginning. The formulation of the problem in this study is how to apply Data Mining with the K-Means clustering algorithm to find solutions in adding types of reading books that students are interested in. The application of the K-Means Algorithm aims to assist the librarian in classifying the data on borrowing books that students really like or are interested in. In the k-means algorithm, it can analyze more deeply by processing book lending transaction data with results that explain various information - information about the distribution of the intensity of borrowing a book so that the information from the processing results can help the library to determine the addition of a collection of books that are right on target. which is then processed using Rapid Miner. This provides benefits for the Politeknik Negeri Balikpapan Library in the plan to increase the collection of reading books in the library. The final results of the research are 107 book titles with a borrowing frequency of 231 times in the period 2019-June 2022. The test is carried out by finding the smallest value of the Davies Bouldin Index (DBI) where after the data is processed it is known that the smallest value is 0.407 with a total of 3 clusters. It was concluded that scores with the Highest, Medium and Low Clusters were obtained in the grouping of library books, namely the Highest Cluster with the number of book data being 2 titles, namely Teknologi Beton, teori dan Praktik Hotel Front Office. These two types of books are the most borrowed so that they can be recommended to be proposed in the procurement of library book collections. Medium cluster with 23 book titles and low cluster with 82 book titles.

Keywords: K-Means, Clustering, Davies Bouldin, Rapid miner

1. Introduction

The existence of the library cannot be separated from human culture. The high and low civilization of a nation can be seen from the condition of the library it has. In essence, the library is a cultural product in the form of an institution that collects, stores, organizes both printed and recorded works as a source of information and learning from generation to generation. In Indonesia there are five types of libraries and these five types of libraries are national libraries, public libraries, special libraries and university libraries.

The college library collections are held through a selection that refers to the needs of study programs that are organized and organized in such a way as to ensure the effectiveness and efficiency of services to the needs of the academic community. Even in every academic community, their needs cannot be equated because they have different needs in the literature. The literature used by students in each department is also different. With the organized grouping of each user in this case is a borrower from various majors, it can be seen what literature groups are most often borrowed by students.[1],[2],

Politeknik Negeri Balikpapan's Library aims to support information needs, the condition of library reading interest can be seen from the percentage of visitors, borrowers and borrowed books. Book borrowing data is used as a report. The library every year procures new book collections to be reproduced so that they must know the types of priority book collections to be reproduced. So it is necessary to analyze the data processing of book lending transactions that explain the distribution of the intensity of book lending, the processing information helps to add to the collection.

To find out which books are the most popular, a cluster technique is used using the K-Means method. The K-Means algorithm is an iterative clustering algorithm that partitions the data set into a number of K Clusters that have been set at the beginning [3]. Therefore, the authors in this study took the title "Application of Data Mining Clustering Using the K-Means Method in Grouping Balikpapan State Polytechnic Library Books With Davies Bouldin Index " to help classify loan book data of interest in supporting the Politeknik Negeri Balikpapan Library Manager in procuring book collections.

Data mining is a term used to describe the discovery of knowledge in databases. Data mining is a process that uses statistical techniques, mathematics, artificial intelligence, and machine learning to extract and identify

useful information and assembled knowledge from various large databases [4]. Basically, clustering is a method for finding and grouping data that have similar characteristics (Similarity) to one another. Clustering is a data mining method that is unsupervised, meaning that this method is applied without training and without a teacher and does not find the target output. In data mining there are two types of clustering methods used in grouping data, namely hierarchical clustering and non-hierarchical clustering [5] Hierarchical clustering is a method of grouping data that begins by grouping two or more objects that have the closest similarity. Then the process is passed to another object which has a second immediacy. This is the next process so that the cluster will form a kind of tree where there is a clear hierarchy (level) between objects, from the most similar to the least similar. Logically all objects in the end will only form a cluster. Dendograms are usually used to help clarify the hierarchical process [5].

In contrast to the hierarchical clustering method, the non-hierarchical clustering method begins by first determining the desired number of clusters (two clusters, three clusters, or so on). After the number of clusters is known, then the cluster process is carried out without following the hierarchical process. This method is commonly called K-Means Clustering [5][6]

The K-Means algorithm is a relatively simple algorithm for classifying or grouping a large number of objects with certain attributes into K clusters. In the k-Means algorithm, the number of K clusters has been determined beforehand. K-Means is a clustering algorithm for data mining that was created in the 70s and is useful for clustering elemental learning (unsupervised learning) in a data set based on certain parameters. K-Means is an algorithm for classifying or grouping objects (in this case data) based on certain parameters into a number of groups, so that it runs faster than hierartical clustering (if small) with a large number of variables and produces denser clusters [7].]. K-Means has the following properties: there are always K clusters, at least one data in each cluster, this cluster is non-hierarchical and there will be no overlap, and each member of a cluster is adjacent to another cluster because proximity does not always involve the center. of that cluster.

2. Methods

The following is the method of conducting the research:

Figure 1. Research Methods

1.Planning Phase

The first step in this study is to conduct a research plan. There are four activities in planning, namely determining research objectives, identifying problems, determining problem boundaries and literature studies.

2. Preprocessing Stage

In the preprocessing stage, the activity carried out is to type back the data obtained into Microsoft Excel to record all book lending transactions After the data is copied at this stage, data cleaning will also be carried out, namely the deletion of data that is not clearly written or data that cannot be read. This cleaning process is carried out in order to get the correct calculation results.

3. Data Processing Phase

Processing data with rapidminer software and also processing data based on identifying problems in the research. Using the K-Means method, and through the application of al-goritma k-means clustering, it is hoped that in processing this data, it is hoped that in processing this data, it will get good results to group books based on borrowing frequency. learning outcomes and generating new knowledge. Processing data with Rapidminer software

4. Analysis Stage

After all the data is collected, the analysis stage is carried out.

5. Documentation Stage

The documentation process is carried out from the beginning of the study to the end of the study. The final result of the documentation process is in the form of a progress report and a Final Research Report

Here's the research flow chart:

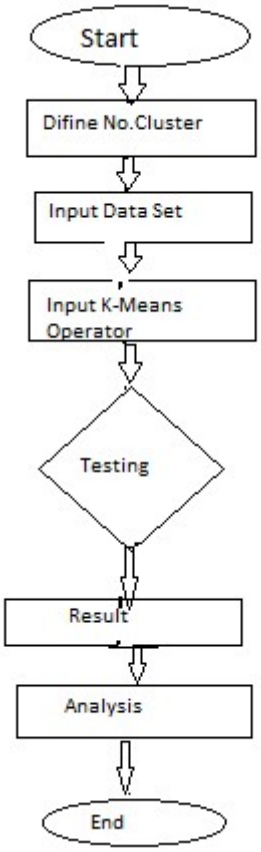

Figure 2. Flowchart

3. Result and Discussion

A. Data Set

The dataset used is book review data at the Balikpapan State Polytechnic Library from 2019-June 2022. The number of types of books recorded is 107 titles with a total borrowing frequency of 231.

B. Preposesing Data

At this stage, use the RapidMiner tool with the K-Means method. The New Process view is for preparing worksheets on RapidMiner. Importing the data to be tested contains a .xls or .xlsx format. next is the model for importing Mocrosoft Excel files. To enter the data to be run it is necessary to go through the right click -> Insert Operator -> Data Acess -> File -> Read Excel

From the existing dataset, data reprocessing is then carried out, one of which is to change the name of the book title attribute to a unique code (id) with the data type being a polynominal. For the Total attribute of borrowing each book each year the data type is integer. Having preposesed the data using Rapid Miner, there is no missing value for both attributes.

C. Modelling and Evaluation

At this stage, added the K-Means operator. via right click -> Insert Operator -> Modeling -> Segmentation -> K-Means.

Next, work on the settings on the K-Means Clustering Parameters menu, setting the k value, where k is the value that will be used to determine the number of clusters to be created. Here the number of clusters to be created is as many as 3 clusters (low, medium, high)

Next, work on the settings on the K-Means Clustering Parameters menu, setting the k value, where k is the value that will be used to determine the number of clusters to be created. Here the number of clusters to be created is as many as 3 clusters (low, medium, high)

The modeling form becomes as follows:

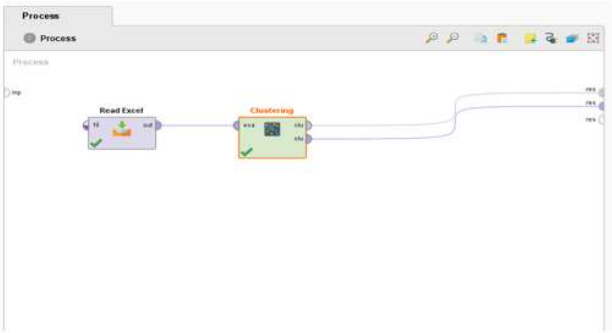

Figure 3. Clustering Modeling with the K-Means operator

In Figure 4, the data results are decomposed into 3 clusters, namely cluster 0, cluster 1, cluster 2 with each cluster pocketing the results of grouping data cluster_0 consists of 82 book titles, cluster_1 consists of 23 book titles, cluster_2 consists of 2 book titles.

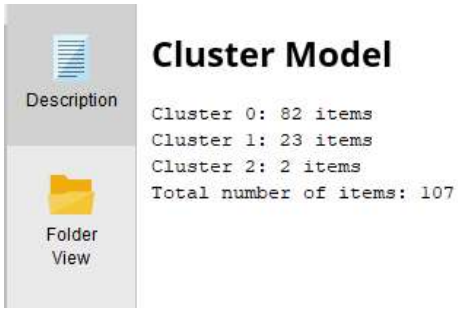

Figure 4. Cluster Model

E-ISSN: 2798-4664

Here are the members of Cluster 0

Figure 5. Cluster 0 Members

Following members of Cluster 1

Figure 6. Cluster 1 Members

Following members of Cluster 2

Figure 7. Cluster 2 Members

From the results of clustering with K = 3, it can be seen that Cluster 2 (high) is the most popular group of books consisting of 2 book titles, namely Concrete Technology and Front Office hotel Theory and Practice books, the titles of the books that are most in demand for borrowing during the period 2019 to June 2022 as many as 14 and 13 times the frequency of borrowing.

Cluster 1 (medium) consists of 23 book titles which are books with a moderate frequency of borrowing in other words categorized as in demand for borrowing. While Cluster 0 (low) consists of 82 book titles with the frequency of borrowing their books in the rare category in other words less attractive to borrow.

D.Perfomance

At this stage, add the K-Means operator. via right click -> Insert Operator -> Modeling -> Segmentation -> K-Means. For clustering performance with K=3 using the Davies Bouldin Index (DBI) value. Evaluation using the Davies Bouldin Index has an evaluation scheme from the internal cluster, where the good or bad results of the cluster are seen from the quantity and proximity of the clustered data.

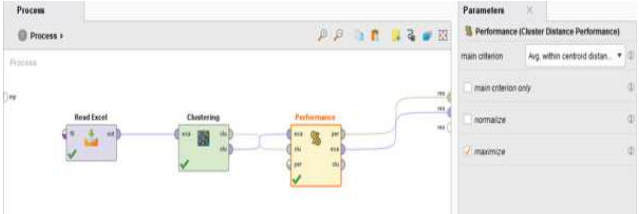

Figure 8.. Perfomance Model

E-ISSN: 2798-4664

From the results of the perfomantion, the Davies Bouldin index value was obtained at 0.407 as figured 9

Figure 9. Davies Bouldin Index Value

4. Conclusions

From the results of data processing through the Rapid Miner Application with K = 3, the results of cluster_0 (low) were obtained consisting of 82 book titles with the frequency of borrowing their books in the rare category in other words less in demand for borrowing, cluster_1 (medium) consisting of 23 book titles which are books with a moderate borrowing frequency in other words categorized as being in demand for borrowing, cluster_2 (high) consists of 2 book titles is the most popular group of books consisting of 2 book titles, namely Concrete Technology and the Front Office hotel Theory and Practice books. The perfomantion result obtained the Davies Bouldin index value of 0.407. This grouping of book data can be an input for library managers in the procurement of book collections based on the frequency of borrowing books.

Acknowledgement

The researchers would like to thank to Pusat Penelitian dan Pengabdian Masyarakat (P3M) Politeknik Negeri Balikpapan for funding this research.

Reference

[1] Intan fitri andyni. (2013). Grouping of book borrowers using the k-means method at the Central Library of Upn "veterans" East Java
[2] Deka Dwinavinta. (2014). Klasterisasi judul buku dengan Metode K-Means. Universitas Islam Indonesia
[3] Parlina, Iin. (2018). Memanfaatkan Algoritma K-Means Dalam Menentukan Pegawai Yang Layak Mengikuti Asessment Center. CESS (Journal of Computer Engineering System and Science), 3(1), 87–93.
[4] Kusrini and Lutfi, ET 2009. Data Mining Algorithms. Yogyakarta: Andi Offset
[5] Santosa, Budi. (2007). Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis. Yogyakarta: Graha Ilmu.
[6] Sulaiman. (2020). Analisis pola belanja konsumen menggunakan algoritma k-means dan apriori pada haura swalayan. UIN Sultan Syarif Kasim Riau
[7] Arai, K., dan Barakbah, A. R. (2007). Hierarchical k-means: an algorithm for centroids initialization for k-means. Reports of the Faculty of Science and Engineering, 36(1), 25–31