Journal of Scientometric Research, 2025; 14(1):32-45.
https://www.jscires.org

Research Article

Personalized Library Book Recommendations Using K-Means Clustering and Association Rules

Faris Mushlihul Amin*, Evi Fatimatur Rusydiyah, Anisa Nur Azizah

UIN Sunan Ampel, INDONESIA.

ABSTRACT

The library serves as the primary information source and has a significant impact on raising educational standards. This study aims to improve library service quality by developing a personalized book recommendation system. The personalization of the recommendation system is the design of a prediction model for books that each user will borrow based on the user's interests, behavior and other related information. The novelty of this study lies in the combination of K-Means Clustering and Association Rule techniques to create a more accurate recommendation model tailored to each user's preferences, behavior and other relevant information. One common issue in building recommendation systems is the Cold Start Problem, which refers to the challenge of making accurate recommendations for new users or items with little to no historical data. Therefore, K-Means Clustering is utilized to segment users based on their borrowing patterns, which helps address the Cold Start Problem by recommending popular books within each cluster. User information from book loan transactions is used as input for the clustering model. Next, a recommendation model for each cluster will be made using the Apriori algorithm. Apriori was chosen for its simplicity, low computational cost and ability to efficiently identify frequent patterns, making it ideal for large datasets. This study trials the number of clusters between k=3, 4, 5 and 6. The best results for the recommendation model used a combination of 6 clusters and the Apriori algorithm. The application of the clustering method can improve the recommendation model with a difference of 3.61% average accuracy, 0.67% average precision, 3.61% average recall and 0.49% average F1 score compared to the recommendation model without clustering method.

Keywords: Book Recommendation, Library, Cold Start Problem, K-Means Clustering, User Behavior, Association Rule.

Correspondence:
Faris Mushlihul Amin
UIN Sunan Ampel, INDONESIA.
Email: faris@uinsby.ac.id
ORCID: 0000-0001-5414-8850

Received: 31-10-2023;
Revised: 03-03-2024;
Accepted: 20-11-2024.

INTRODUCTION

Libraries have an important role in tertiary institutions to improve the quality of education through reading literacy.[1] The library contains not only books but also articles, novels, magazines, final projects and other printed media. The large variety of books in the library can make it time-consuming for some readers to find their desired book.[2] Additionally, many books are poorly recommended, leading to inefficient use of library resources.[3] A personalized recommendation system helps optimize library resources by reducing waste, streamlining the user search process and enhancing marketing efforts by effectively targeting user preferences.[4] These systems are designed to predict the likelihood of books being borrowed by readers based on their interests, behavior and other relevant information, thereby fostering a greater interest in reading among all members of the academic community.[3] The implementation of personalization in recommendation systems can provide significant benefits for libraries. Previous research on the implementation of book recommendation systems in libraries using association rules based on relationships between book titles yielded promising results, with a precision value of 70%.[5] Other studies have used gender classification to enhance the effectiveness of book recommendation models for users.[6] Based on these results, user preferences are considered to be able to improve the recommendation model.

The library recommendation system can also be implemented using the K-Means method. Information about user preferences is utilized as input for the clustering model. In the implementation of the K-Means method in a recommendation system, as discussed in previous research,[7] K-Means was used to classify library visitors by age and to recommend books that are age-appropriate. Clustering can be performed using methods like K-Means, which is a clustering algorithm that groups datasets into k clusters based on the similarity of the data.[8] The similarity within a cluster is

DOI: 10.5530/jscires.20251005

Copyright Information :
Copyright Author (s) 2025 Distributed under Creative Commons CC-BY 4.0

Publishing Partner : Manuscript Technomedia. [www.mstechnomedia.com]

measured by the proximity of the data points to the centroid or center of the cluster.[9]

K-Means demonstrates strong performance in grouping large datasets with fast and efficient computation times.[10] Another approach to understanding user needs is by analyzing item purchasing patterns. Insights into book borrowing patterns can be gained by applying Association Rules to library borrowing data over a certain period. The Apriori method is one of the most commonly used techniques for extracting information from borrowing data.[11] Agrawal and Srikant introduced the Apriori algorithm in 1994 as a fundamental method for identifying frequent itemsets for Boolean association rules.[12] The Apriori algorithm is a simple yet effective Association Rule technique used to discover high-frequency patterns through association rules. To identify these patterns, itemsets with frequency or support values that exceed the minimum support threshold (MinSup) are identified. The high-frequency patterns obtained are then used to construct association rules.[13,14]

METHODOLOGY

The recommendation model is highly useful for users as it shortens the time needed to search for books, thereby optimizing library services. Personalized book recommendations also benefit librarians by offering insights into the optimal number of books needed in the library. Additionally, recommendation systems can serve as an indicator of users' reading interests by analyzing the books that are most frequently borrowed or viewed. By examining such behavioral patterns, these systems can generate recommendations that more accurately align with individual reading preferences.

This study utilized primary data from a campus library in Indonesia. The data included student book loan transactions, book information and user information. The book borrowing history data comprised several attributes such as Borrowing Date, Student Name, Book Title and Return Date. The book information data consisted of nine attributes, including Book Title, Author's Name, Publisher, ISBN, Year, Quantity, Price and Study Program. Meanwhile, the user information data included attributes such as User ID, Gender, Member Type, Instance Name, Email, among others. This study recommends the TOP 10 books for each user. The recommendation model employs the Association Rule method using the Apriori Algorithm. Additionally, this study performs clustering based on user preferences to enhance recommendation results. The research steps are detailed in Figure 1.

Understanding and clarifying the methodological aspects of the machine learning technique is crucial for achieving reproducible results.[15] In this study there are several steps implemented in the methodology. The initial step is data preprocessing, which consists of several stages. Data cleaning aims to remove unused or duplicate data in the process of building a recommendation model. Additionally, to address data inconsistencies, various issues were identified that needed correction, such as typographical errors, mispronounced names and inconsistent data types. These issues were rectified and missing data was filled in with a value of 0.

The next stage is to find user preferences that represent in more detail user information. In this stage, data on book borrowing transactions is utilized as a reference for determining user preferences. In addition to Gender and Instance Name information, each user is associated with a preferred Publisher, Location, Collection Type and Favorite Book Topic, based on the most frequently borrowed items. This stage produces 6 variables that are used as input for cluster learning using the K-Means Clustering method.

In the next stage, users are grouped into several clusters using unsupervised learning, resulting in k clusters based on user interests. This study aims to determine the optimal k value by conducting an elbow method test.[16] For each cluster, a book recommendation model is then developed using the Association Rule method with the Apriori Algorithm.

The results of the model will recommend books based on the TOP 10 items with the highest confidence value for each user. Some of the recommendations from each cluster will be combined and the overall evaluation of the recommendation model will be calculated. In this study, accuracy, precision and recall are considered in determining the best recommendation model. The primary objective of this study is to develop a book recommendation model that aligns with the reading interests and needs of users, particularly students. The final recommendations consist of a selection of the TOP 10 books, based on other students' loan transactions and criteria that are relevant to the users.

Implicit Data

Building a recommendation model involves selecting the appropriate method based on the available dataset. The data represents user preferences, which will be used as a reference for future predictions. In a recommendation system, there are two types of data: explicit data and implicit data. Explicit data is information provided directly by the user, such as ratings, comments, or opinions about whether they like an item or not. On the other hand, implicit data is derived from user behavior and preferences.[17] Implicit data reflects how users interact with items, such as repeatedly playing a song, regularly purchasing certain products, or frequently watching movies of the same genre.[18]

Implicit data has both advantages and disadvantages. One advantage of implicit data is that it can be easily collected in large quantities.[19] However, a significant disadvantage is that it does not directly provide or infer a user's explicit preference for a particular item. Implicit data, often referred to as presumptive

feedback, reflects user behavior that is presumed to indicate preferences, rather than directly revealing them. As a result, utilizing implicit data requires more effort than explicit data, as it involves converting user behavior data into meaningful insights.[20]

Association Rule

Association rule is a data mining technique that is applied to obtain association rules from item combinations.[21] Association analysis produces patterns and implication rules that show the relationship between features in the data.[22] In the recommendation system, the association rule is used to look for relationships between items in user data to determine which items will be needed next. Association rule mining is an expression of the implications of the form X→Y, where X and Y are sets of antecedent and consequent items.[23] Association rule mining is used to find association relationships between itemsets X and Y, which meet the minimum support and confidence thresholds.[24]

Support: measures how frequently a particular itemset appears in the dataset. It is expressed as a percentage of the total transactions in the dataset that contain the itemset. The higher the support value, the more frequently the itemset appears in the dataset, indicating a more significant relationship.

Confidence: measures how often a rule is proven to be true. In the context of association rules, confidence measures how frequently itemset Y appears in transactions that also contain itemset X. It is expressed as a percentage of transactions containing itemset X that also contain itemset Y. The higher the confidence value, the stronger the relationship between itemset X and Y.

Association rules are determined based on support and confidence values that are more or equal to the specified minimum support and minimum confidence.[25] One of the methods in association rule mining is the Apriori Algorithm. Agrawal and Srikant introduced the Apriori algorithm in 1994 as a fundamental method for identifying frequent itemsets for Boolean association rules.[12] The Apriori algorithm is one of the simplest types of rule association methods that are applied to obtain high frequency patterns through association rules. To get high-frequency patterns, item patterns are searched for with frequency or support values that exceed the support threshold (MinSup). The high frequency patterns obtained are used to construct associative rules.[13] Frequent itemset is a collection of items that frequently appear together in a transaction dataset. In this context, "frequently" is measured by the support value, which is the percentage of transactions in the dataset that contain a particular itemset. In the frequent itemset there are several terms, such as itemset, support count (σ) and support. Support count (σ) is the frequency of occurrence of an itemset. Itemset is a collection of one or more items. To obtain the frequent itemset is illustrated in Figure 2.

Association rules are determined based on support and confidence values that are more or equal to the specified minimum support and minimum confidence. The rules for support and confidence values are shown in Equation (1) where N is a lot of data.

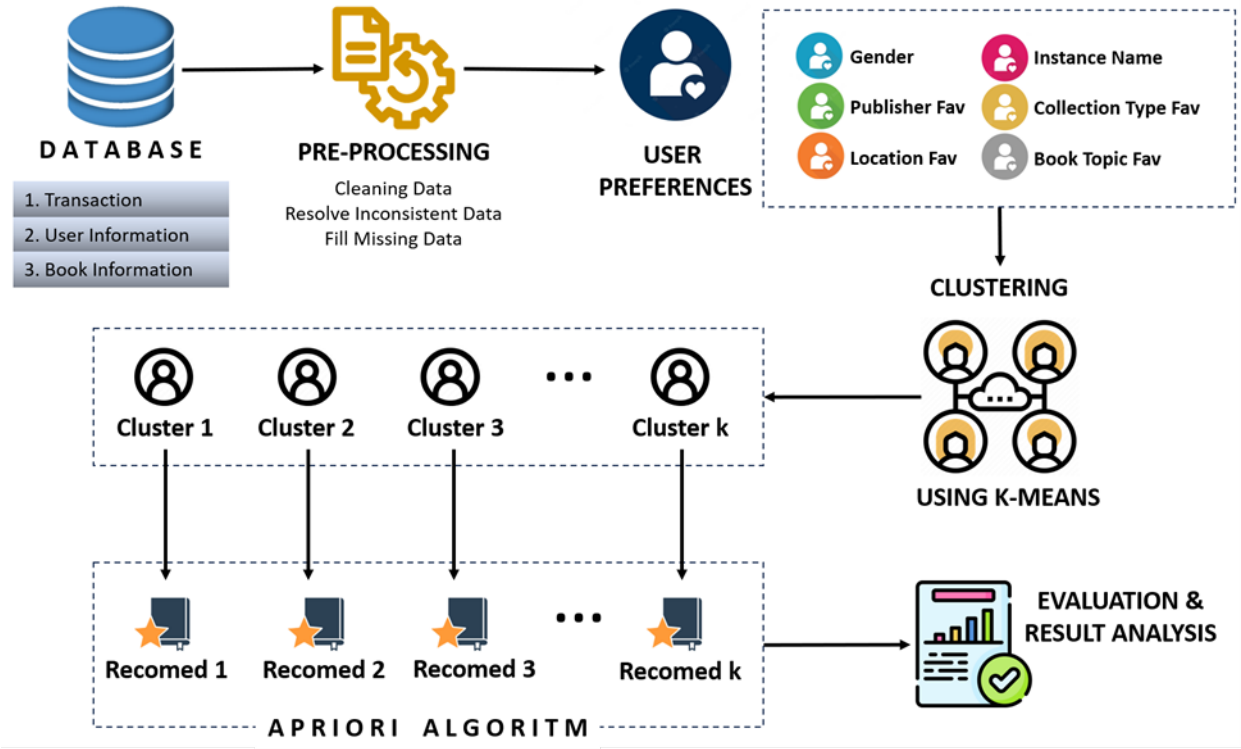

Figure 1: Flowchart in this study.

$$support, s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N} \qquad (1)$$

$$confidence, s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \qquad (2)$$

K-Means Clustering

Clustering is the process of grouping data into several clusters so that the data in a cluster have maximum similarity.[4] One method in clustering is K-Means. K-Means is a distance-based clustering method by dividing data into a number of Clusters.[27] The data will be divided into clusters based on the closest distance to the centroid. Centroid itself is the midpoint or center of clustering. Illustration of cluster formation in K-Means clustering can be seen in Figure 3.[28]

In measuring the distance from the data to the centroid, measurements are made using Euclidean.[29] Euclidean Distance (d) measurements can be found using Equation (3).

Where n is the number of data dimensions, x is the i-th data and y is the i-th centroid value.

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

RESULTS AND DISCUSSION

The initial stage of this study involves understanding the dataset through Exploratory Data Analysis (EDA). The research dataset comprises three types of files: book loan transaction files, user information files and book information files, all of which are stored in CSV format. Each of these files has distinct column features. Therefore, they will be detailed in tabular form as follows: Table 1 presents a sample of the transaction data, Table 2 displays a sample of the user information data and Table 3 shows a sample of the book information data.

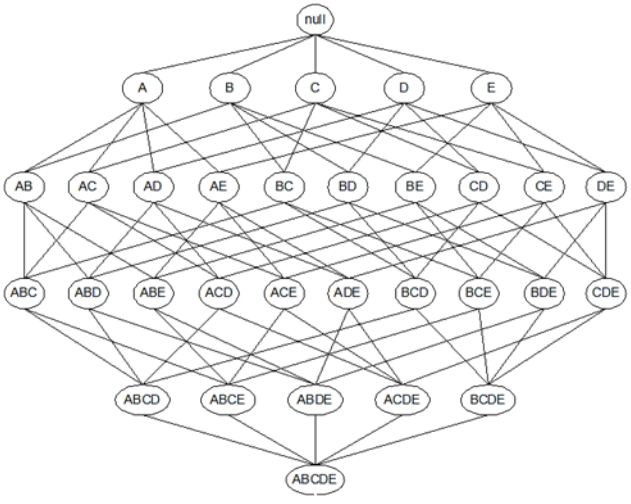

Figure 2: Frequent Itemset Formation.[26]

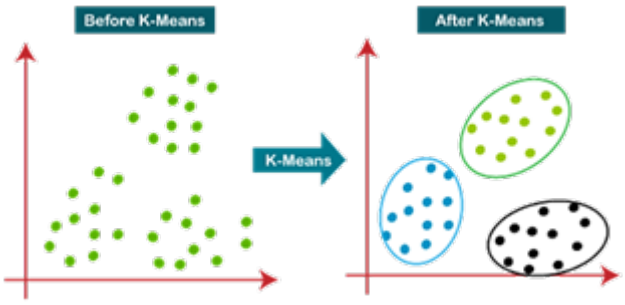

Figure 3: Illustration of Cluster Formation in K-Means Clustering.[28]

Tables 1-3 provide an overview of the transaction data, including various column names. Sample data is presented in Sample-1, Sample-2 and Sample-3. Each column contains a different amount of data, resulting in varying levels of completeness. Consequently, missing values were calculated for each column to determine which data are most suitable for model training. The transaction data represent a historical record of user book borrowing activities in the library. The dataset covers historical data from January 2, 2019, to December 20, 2022. This data includes information related to book borrowing, such as user ID, book ID, book title and other relevant details.

Based on Figure 4, the graph of book borrowing transactions at the library shows a decline in May 2020, indicated by the red arrow. This decline was due to the COVID-19 pandemic, which led to a lockdown and halted all academic activities, including those in the library and campus areas. As a result, students were required to study from home for several months and the library began to reopen in early 2022. Despite the reopening, the prolonged period of reduced activity caused a slow recovery in the number of students and library visitors, as reflected by the gradual increase in the graph until December 2022.

The COVID-19 pandemic has significantly impacted society, particularly in the way information is sought. Many users have turned to gadgets and online books for information. Consequently, in addition to providing book recommendations, this study aims to enhance reading interest, especially among students both within and outside the university. With the advancement of digital technology, libraries are expected not only to offer web-based services but also to provide effective reading recommendations tailored to user interests. This recommendation system operates based on implicit data, which lacks direct feedback from users about their preferences for specific books. Therefore, the feedback is derived solely from loan transaction data, which will be processed and analyzed to serve as a reference for user preferences.

Book data is a compilation of detailed information related to the books in the library. This data, collected from 2019 to 2022, includes a total of 4,420 book types. The dataset contains 28 attributes for each book, such as title, author, topic and publisher, as shown in Table 2. Based on the information in Table 2, not all

Table 1: Library Book Transaction Data Sample.

Columns Name	Total Data	Missing Data	Sample-1	Sample-2	Sample-3
loan_id	187192	0%	1099605	1099607	1099608
item_code	187192	0%	920156	2,02E+08	2,02E+08
biblio_id	187192	0%	3826	95747	92841
title	187192	0%	Organisasi sekolah dan pengelolaan kelas: Oleh: Haji Hadari Nawawi.	Asas-asas hukum pidana Islam	Sistem Pertanggung jawaban Pidana: Perkembangan dan penerapan/Hanafi Amrani dan Mahrus Ali.
gmd_name	187188	0%	Buku	Buku	Buku
language_name	125793	33%	NULL	Indonesia	
location_name	184534	1%	Perpustakaan UIN	Perpustakaan UIN	Perpustakaan UIN
Collection type name	187192	0%	Kol. Umum	Kol. Umum	Kol. Umum
member_id	187192	0%	19640529xxx	C9321xxx	C9321xxx
member_name	187192	0%	NOOR xxx	FAKHRI xxx	FAKHRI xxx
Member type name	185951	1%	Dosen	Mahasiswa	Mahasiswa
loan_date	187192	0%	03/02/2020	03/02/2020	03/02/2020
due_date	187192	0%	03/11/2020	03/09/2020	03/09/2020
renewed	187192	0%	0	0	0
is_lent	187192	0%	1	1	1
is_return	187192	0%	1	1	1
return_date	186304	0%	3/16/2020	06/11/2020	06/11/2020

columns are filled in for every book. Therefore, some features will be excluded from the recommendation model due to incomplete data.

User data is a compilation of detailed information about library members. This dataset includes 24 columns of user details with various uses and functions. The file contains information on 52,005 users, as detailed in Table 3. This data covers students from 2017 to 2021. However, the library also serves lecturers and staffs, so additional user files were included, bringing the total number of library users to 52,438.

The data is preprocessed through several steps: data cleaning, removing duplicates, resolving inconsistencies and transforming the data. The data cleaning stage involves selecting the features to be used as input for the recommendation model. Due to a significant amount of missing data, not all features are utilized in this study. Therefore, this study selects 9 features: transaction date, member ID, bibliographic ID, user gender, name of institution, library location, collection book name, book topic and book publisher. The next step involves resolving data inconsistencies by correcting the user gender data, library locations and names of institutions. Finally, the data is transformed by converting object data types into numerical formats to be processed as model input.

User Preferences

The next stage involves identifying user preferences in more detail. This requires analyzing book borrowing transaction data to determine user preferences. In addition to gender and institution information, each user has preferred attributes such as publisher, location, collection types and book topic. These preferences are determined based on borrowing history, where a higher number of borrowings indicates greater interest or preference. The criteria considered are not the book's bibliographic ID but rather attributes such as topic, publisher, location and collection type. These attributes serve as initial filters for assessing user interests. The results of user preferences are presented in Table 4.

Based on the user preferences results in Table 4, several analyses related to user interests were conducted. The first analysis concerns the book topics that are most popular among users, as illustrated in Figure 5. The graph displays the top 10 out of 143 book topics favored by users. The top 10 favorite topics include Economics, Law, Research, Education, Psychology, Accounting, Management, Islamic Education, Arabic and Fiqh. For instance, 318 users preferred book topics related to Economics, making it the most popular topic among users. Additionally, user institutions related to each book topic are detailed in Table 5.

Table 2: Sample Book Information Data in the Library.

Columns Name	TotalData	Missing Data	Sample-1	Sample-2	Sample-3
biblio id	4420	0%	98162	98163	98164
title	4420	0%	Trade policy protectionism and the third world.	Public Relations: Talents of PR	The Political economy of natural gas.
edition	3349	24%	NULL	NULL	NULL
isbn issn	3431	22%	978114000000000000	602-1232-51-4	978114000000000000
author	4317	2%	MIchael Davenport	Bambang Suratman-Siti Sri Wulandari	Ferdinand E Banks
topic	4377	1%	Trade - policy protectionism	Hubungan Masyarakat	Political Economy-Natural
gmd	4420	0%	Buku	Buku	Buku
publisher	4376	1%	Routledge	Salemba Humanika	Routedge
Publish place	4360	1%	London	Jakarta	London-New York
language	3519	2%	English		English
classification	4419	0%	382,753	659.2	338.2
location	3776	15%	Perpustakaan A. Yani	Perpustakaan Gunung Anyar	Perpustakaan A. Yani
publish year	4232	4%	2018	2017	2018
items	3776	15%	201900325	201900741-201900742-201900743-201900740-201900739	201803503
Collection types	3776	15%	Kol. Tandon	Kol. Umum	Kol. Tandon
call number	4415	0%	T 382.753 Mic t	U 659.2 Bam p	T 338.2 Fer p
opac hide	4420	0%	0	0	0
promoted	4420	0%	1	1	1
collation	4367	1%	i, 144 hlm, 24 cm	x, 102 hlm. 23 cm.	ii, 198 hlm, 24 cm
image	2797	37%	trade.jpg.jpg	public.jpg.jpg	NULL
input date	4420	0%	02/06/2019 14:00	02/06/2019 15:02	02/06/2019 15:08
last update	4420	0%	08/10/2021 9:23	02/07/2019 9:42	1/31/2020 14:52

Based on Table 5, it is evident that books on research topics are predominantly favored by users from the Bimbingan dan Konseling Islam (BKI) institution. Interestingly, books on Islamic Education are most popular among users from the Pendidikan Bahasa Inggris (PBI) institution. This observation aligns with the library's location at an Islamic university in Indonesia, which emphasizes not only academic teaching but also Islamic studies across all user programs. The analysis of user topic preferences also indicates that the Tarbiyah and Teacher Training Faculty would benefit from increased literacy in Management, Islamic Education, Fiqh and Education topics.

Figure 6 illustrates the top 10 out of 70 publishers that are most favored by users. Publisher information can provide insights into the quality of books, including paper quality, design, color and topic coverage, as each publisher has its own standards for printed books. In this study, the most preferred publisher among users is Rajawali Pers, with 1,323 users selecting books from this publisher. Following Rajawali Pers, other popular publishers include Ar-Ruzz Media, Bumi Aksara, Salemba Empat, UIN Maliki Press, UIN Sunan Ampel Press, Mandar Maju, Hikmah Press, Kencana and Raja Grafindo Persada.

The results of these user preferences can aid in making decisions about book procurement for libraries. Librarians often face challenges in selecting books due to the vast array of options available. They must provide books that meet users' needs and the preference results outlined above can serve as a valuable tool for decision support. By utilizing these insights, librarians can narrow down their search for books to stock in the library. Consequently, books on Economics and those published by Rajawali Pers should be prioritized for increased stock. This approach helps prevent long wait times for book loans and enhances library service efficiency.

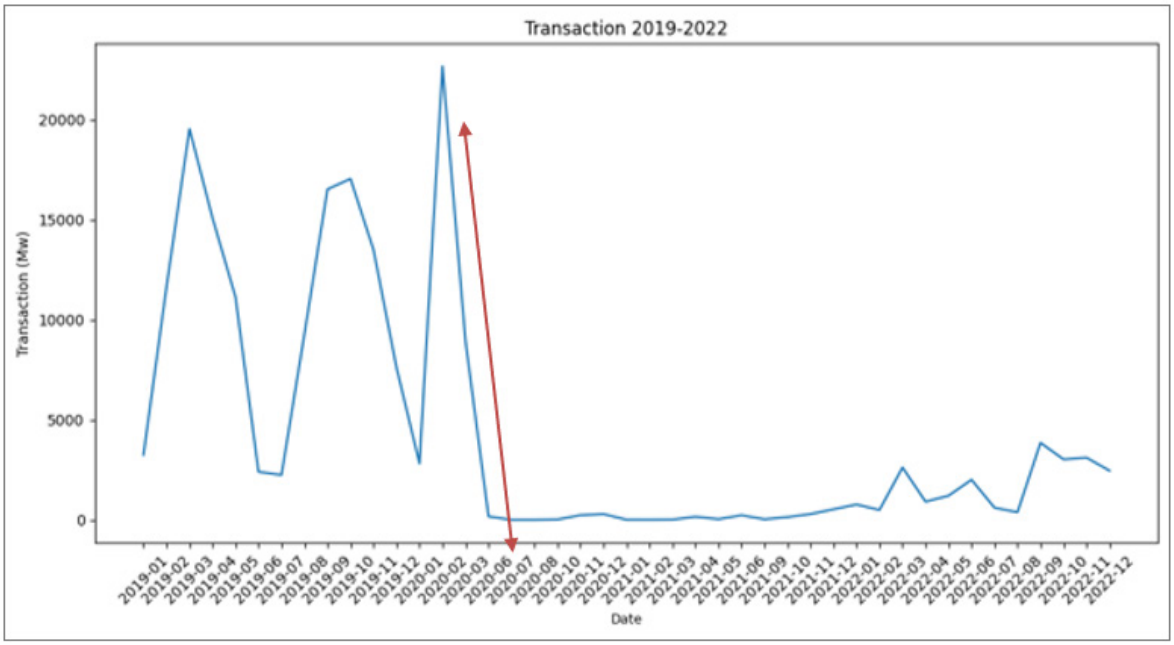

Figure 4: Library Book Lending Transaction Chart.

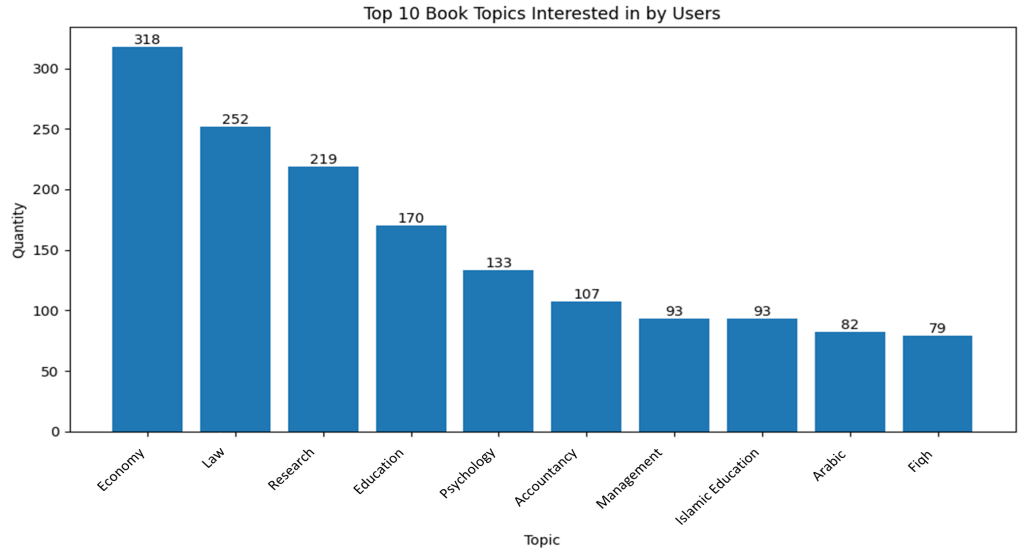

Figure 5: Comparison Chart TOP10 Top Book Topics Interested by Users.

User Cluster Using K-Means Clustering

The purpose of using clustering techniques is to group users into distinct reading interest categories, which in turn influences the creation of a recommendation model, especially given the lack of direct user feedback. This study employs the K-Means Clustering method to identify user groups based on implicit user information derived from transaction history. The inputs for creating the cluster model include user preferences (such as favorite topics, publishers, locations and collection types), user gender and user institutions. Before clustering, it is essential to determine the value of *k* (the number of clusters to be formed). To achieve this, this study utilizes the Elbow Method evaluation technique to identify the optimal *k* value. The results of the Elbow Method evaluation are illustrated in Figure 7.

The optimal number of clusters is identified by the "elbow" point on the graph, where there is a significant decrease in the evaluation metric (inertia or WCSS). This point represents the optimal number of clusters that balances data division with cluster quality. According to the Elbow Method results shown in Figure 7, an elbow is observed around k=5, where the line's change in slope becomes evident. Determining the exact location of this elbow is crucial as it signifies the number of clusters with minimal variability or error. The Elbow Method calculates inertia by summing the squared distances between data points and their nearest cluster center. However, the Elbow Method alone does not always provide a definitive optimal *k* value. To complement this, the study also evaluates the optimal number of clusters using the Silhouette Score, which assesses the quality of clustering based on

Table 3: Sample User Information Data in the Library.

Columns Name	Total Data	Missing Data	Sample-1	Sample-2	Sample-3
member_id	52005	0%	5010221004	5010221005	5010221006
Member name	52005	0%	DODIK xxx	EVI xxx	FEBBY xxx
gender	52005	0%	1	0	0
birth_date	52005	0%	12/24/2002	12/17/2002	10/18/2003
Member type id	52005	0%	Mahasiswa	Mahasiswa	Mahasiswa
postal_code	11955	77%	NULL	NULL	NULL
inst_name	51790	1%	Fakultas Syari'ah dan Hukum-Hukum Ekonomi Syari'ah (Muamalah).	Fakultas Syari'ah dan Hukum-Hukum Ekonomi Syari'ah (Muamalah).	Fakultas Syari'ah dan Hukum-Hukum Ekonomi Syari'ah (Muamalah).
is_new	40004	23%	0	0	0
Member since date	52005	0%	9/8/2021	9/8/2021	9/8/2021
register_date	52005	0%	9/8/2021	9/8/2021	9/8/2021
expire_date	52005	0%	9/8/2022	9/8/2022	9/8/2022
is_pending	52005	0%	0	0	0
last_login	26667	49%	NULL	NULL	NULL
last_login_ip	26667	49%	NULL	NULL	NULL
input_date	44237	15%	9/8/2021	9/8/2021	9/8/2021
last_update	51896	0%	9/8/2021	9/8/2021	9/8/2021

negative values. The results of the Silhouette Score measurements are presented in Table 6.

A negative Silhouette Score value indicates that data points are more similar to clusters other than their assigned cluster, suggesting suboptimal cluster separation. Therefore, the best *k* value is associated with the smallest or no negative Silhouette Score values. According to the results presented in Table 6, *k* values of 3, 4, 5 and 6 each have a negative Silhouette Score value of 1. Based on the analyses in Figure 7 and Table 6, this study evaluates the recommendation model using various user cluster values of *k* (3, 4, 5 and 6). In addition to assessing cluster quality through the Silhouette Score, the study also employs correlation analysis to evaluate the relationship between cluster labels and each user preference feature.

Based on Figure 8, the correlation plots for each *k* cluster trial reveal that the relationship between features is strongest in the *k*=3 and *k*=4 cluster configurations. For *k*=3, the features most highly correlated with the cluster labels are Collection Type, Location and User Gender, with correlation values of 0.81, 0.81 and -0.66, respectively. For *k*=4, the highly correlated features are Collection Type, Location and User Gender, with correlation values of 0.83, 0.83 and -0.49. In contrast, for *k*=5 and *k*=6, the correlation values for these features fall below 0.3. This indicates that as the number of clusters increases, the correlation between the cluster labels and the features decreases.

The Results of The Book Recommendation Model

The aim of this study is to develop a personalized book recommendation model for libraries. To ensure optimal results, the first step involves dividing the book loan transaction data into training and testing datasets. Data from January 2, 2019, to November 30, 2022, is used for training, while data from December 1, 2022, to December 30, 2022, serves as the testing set. The training data is then utilized to create the book recommendation model using the Apriori Method.

In the Apriori Method, the initialization of minimum support and minimum confidence is tailored to the data requirements of

Table 4: Sample User Preferences Information Data in the Library.

Member id	Gender	Instance Name	Fav. Collection Type	Fav. Location	Fav. Book Topic	Fav. Publisher
A92217053	0	Fakultas Adab dan Humaniora-Sejarah Peradaban Islam (SPI).	Kol. Umum	Perpustakaan UIN	Tokoh Pendidikan Islam	Ar-Ruzz Media
E03217047	0	Fakultas Ushuluddin dan Filsafat-Ilmu Al Qur'an dan Tafsir.	Kol. Umum	Perpustakaan UIN	Ekonomi	Bumi Aksara
D92217028	0	Fakultas Tarbiyah dan Keguruan-Pendidikan Bahasa Arab (PBA)	Kol. Umum	Perpustakaan UIN	Belajar	PT. Revka Petra Media
D04217010	0	Fakultas Tarbiyah dan Keguruan-Pendidikan Matematika (PMT).	Kol. Umum	Perpustakaan UIN	Novel	Karya Media
B92219131	0	Fakultas Dakwah dan Komunikasi-Pengembangan Masyarakat Islam (PMI).	Kol. Umum	Perpustakaan UIN	Islam-Indonesia	PT. Mizan Pustaka
I72219039	0	Fakultas Ilmu Sosial dan Ilmu Politik-Hubungan Internasional.	Kol. Umum	Perpustakaan UIN	Kejahatan Hak Asasi	Rajawali Pers
C01219045	0	Fakultas Syari'ah dan Hukum-Hukum Keluarga Islam (AS).	Kol. Umum	Perpustakaan UIN	Ekonomi	Bumi Aksara
11020121089	1	Fakultas Psikologi dan kesehatan-Psikologi.	Kol. Umum	Perpustakaan UIN	Psikologi	Salemba Empat
⋮	⋮	⋮	⋮	⋮	⋮	⋮
3020220055	1	Fakultas Adab dan Humaniora-Sejarah Peradaban Islam (SPI).	Kol. Umum	Perpustakaan UIN	Sejarah Islam	Madani Media

this study. A minimum support of 0.005 was chosen to balance the discovery of significant rules without being overly restrictive. This support threshold indicates the frequency with which an itemset must appear in the dataset to be deemed significant. Additionally, a minimum confidence of 0.001 was set to ensure that the generated rules are sufficiently strong and reliable. This confidence value represents the likelihood that a rule is accurate based on the dataset. By selecting these parameters, the study aims to generate relevant rules with an adequate level of confidence for the recommendation system.

To determine the best recommendation model, evaluation metrics are used to assess its performance. Accuracy measures the overall quality of the recommendation model across all user preferences. Precision gauges the percentage of recommended items that are truly relevant to the user. Recall assesses the model's ability to identify all the books that should be recommended. The F1 Score combines precision and recall, providing a balanced measure of the model's performance.

In the context of library book borrowing, where users rarely borrow more than one book at a time, the effectiveness of the recommendation model is particularly dependent on recall. For instance, if a model suggests 10 books but only one is suitable, the model is considered highly effective if that one relevant book meets the user's needs. Thus, a higher recall value is prioritized over other evaluation scores to ensure that the recommendation model effectively identifies and suggests suitable books for users.

This study aims to enhance the recommendation model's performance by experimenting with various numbers of user clusters. Additionally, it addresses challenges such as the Cold Start Problem by recommending popular books based on popularity scores. Books that are popular within each user cluster are recommended to new users in the same cluster. The comparative results of these experiments are summarized in Table 7.

According to the evaluation results presented in Table 7, incorporating user clusters significantly improves the average evaluation metrics compared to using the entire dataset without

clustering. Specifically, the recommendation model with 6 user clusters outperforms other configurations, achieving an average accuracy of 4.329%, precision of 0.848%, recall of 4.326% and F1 Score of 0.764%. In contrast, the 3-cluster configuration shows lower accuracy and recall values compared to the 1st and 2nd clusters. The experimental results suggest that increasing the number of clusters generally leads to better predictions and more accurate book recommendations for users.

Figure 9 illustrates the conclusions drawn from the experiments summarized in Table 7. The graph shows that as the number of clusters increases, the evaluation scores improve, although the differences are relatively modest. Notably, the accuracy and recall values are nearly identical across the different cluster configurations, whereas precision and F1 Score values are lower. This indicates that while the book recommendation model performs well in terms of overall accuracy and the completeness of recommendations, it struggles with the relevance of the recommended books to user preferences. This discrepancy might be due to the fact that users do not necessarily borrow all recommended books (out of 10) within a short timeframe, which affects the precision and relevance of the recommendations.

Table 5: User Institutions on Each Book Topic.

Topic	User Institutions
Economy	Fakultas Ekonomi dan Bisnis Islam-Ekonomi Syariah.
Law	Fakultas Syari'ah dan Hukum-Hukum Keluarga Islam (AS).
Research	Fakultas Dakwah dan Komunikasi-Bimbingan dan Konseling Islam (BKI).
Education	Fakultas Tarbiyah dan Keguruan-Pendidikan Agama Islam (PAI).
Psychology	Fakultas Psikologi dan kesehatan-Psikologi.
Accountancy	Fakultas Ekonomi dan Bisnis Islam-Akuntansi.
Management	Fakultas Tarbiyah dan Keguruan-Manajemen Pendidikan Islam (MPI).
Islamic Education	Fakultas Tarbiyah dan Keguruan-Pendidikan Bahasa Inggris (PBI).
Arabic	Fakultas Adab dan Humaniora-Bahasa dan Sastra Arab (BSA).
Fiqh	Fakultas Tarbiyah dan Keguruan-Pendidikan Agama Islam (PAI).

In this data testing, which spans one month, it is estimated that a user typically borrows 2 to 3 books per month. However, the model's recommendations generate a list of 10 book items. To better assess the effectiveness of the recommendation model, it is necessary to employ more relevant evaluation metrics, such as Mean Average Precision (MAP), which could provide a more comprehensive assessment of recommendation quality. The model's current performance metrics-average accuracy of 4.329%, average precision of 0.848%, average recall of 4.326% and average F1 Score of 0.764%-indicate significant room for improvement. Future research could explore advanced recommendation methods such as Collaborative Filtering Method[30] or Content-Based Filtering Method,[31,32] which are known to handle implicit data more effectively.

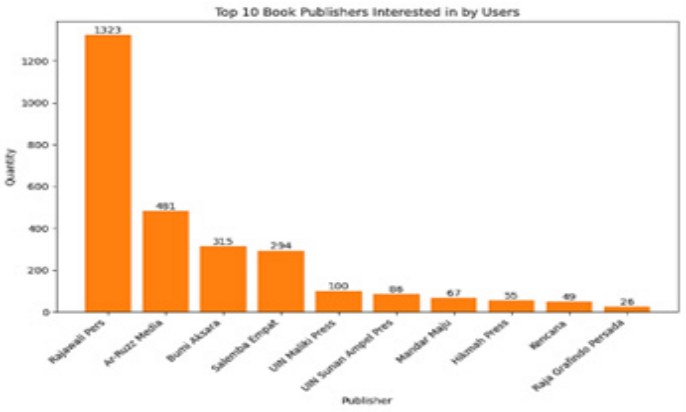

Figure 6: Comparison Chart TOP10 Top Book Publishers Interested by Users.

Table 6: The Result of the Silhouette Score.

k-Value	Number of Negative	Silhouette Score	k-Value	Number of Negative	Silhouette Score
2	0	0.533	9	329	0.7336
3	1	0.668	10	71	0.741
4	1	0.709	11	71	0.743
5	1	0.717	12	101	0.740
6	1	0.725	13	101	0.739
7	171	0.729	14	112	0.741
8	329	0.728	15	0	0.751

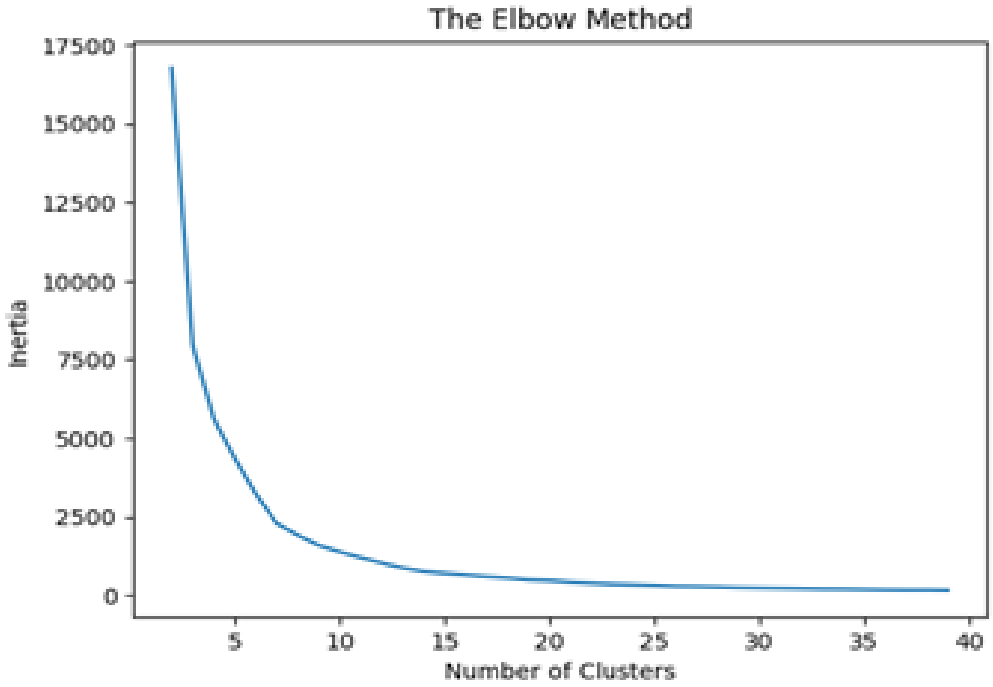

Figure 7: Graph of Determining k-Cluster Values Using the Elbow Method.

Table 7: Comparative Evaluation Values for Each Research Experiment.

Experiment	Accuracy	Precision	Recall	F1 Score
All Data or No Cluster	0.007150	0.001786	0.007150	0.002667
Cluster 1	0.027818	0.003571	0.027818	0.006094
Cluster 2	0.041279	0.001027	0.041279	0.001985
Cluster 3	0.012628	0.001416	0.012628	0.002447
Average 3 Cluster	0.027242	0.002005	0.027242	0.003509
Cluster 1	0.027818	0.003571	0.027818	0.006094
Cluster 2	0.041279	0.001027	0.041279	0.001985
Cluster 3	0.012370	0.001094	0.012370	0.001956
Cluster 4	0.015455	0.002098	0.015455	0.003633
Average 4 Cluster	0.024231	0.001948	0.024231	0.003417
Cluster 1	0.012370	0.001094	0.012370	0.001956
Cluster 2	0.015763	0.037037	0.015763	0.021399
Cluster 3	0.041279	0.001027	0.041279	0.001985
Cluster 4	0.051136	0.001020	0.050930	0.001977
Cluster 5	0.015455	0.002098	0.015455	0.003633
Average 5 Cluster	0.027201	0.008455	0.027159	0.006190
Cluster 1	0.090913	0.009534	0.090913	0.016690
Cluster 2	0.015763	0.037037	0.015763	0.021399
Cluster 3	0.012370	0.001094	0.012370	0.001956
Cluster 4	0.074114	0.000105	0.074114	0.000210
Cluster 5	0.015455	0.002098	0.015455	0.003633
Cluster 6	0.051136	0.001020	0.050930	0.001977
Average 6 Cluster	0.043292	0.008481	0.043257	0.007644

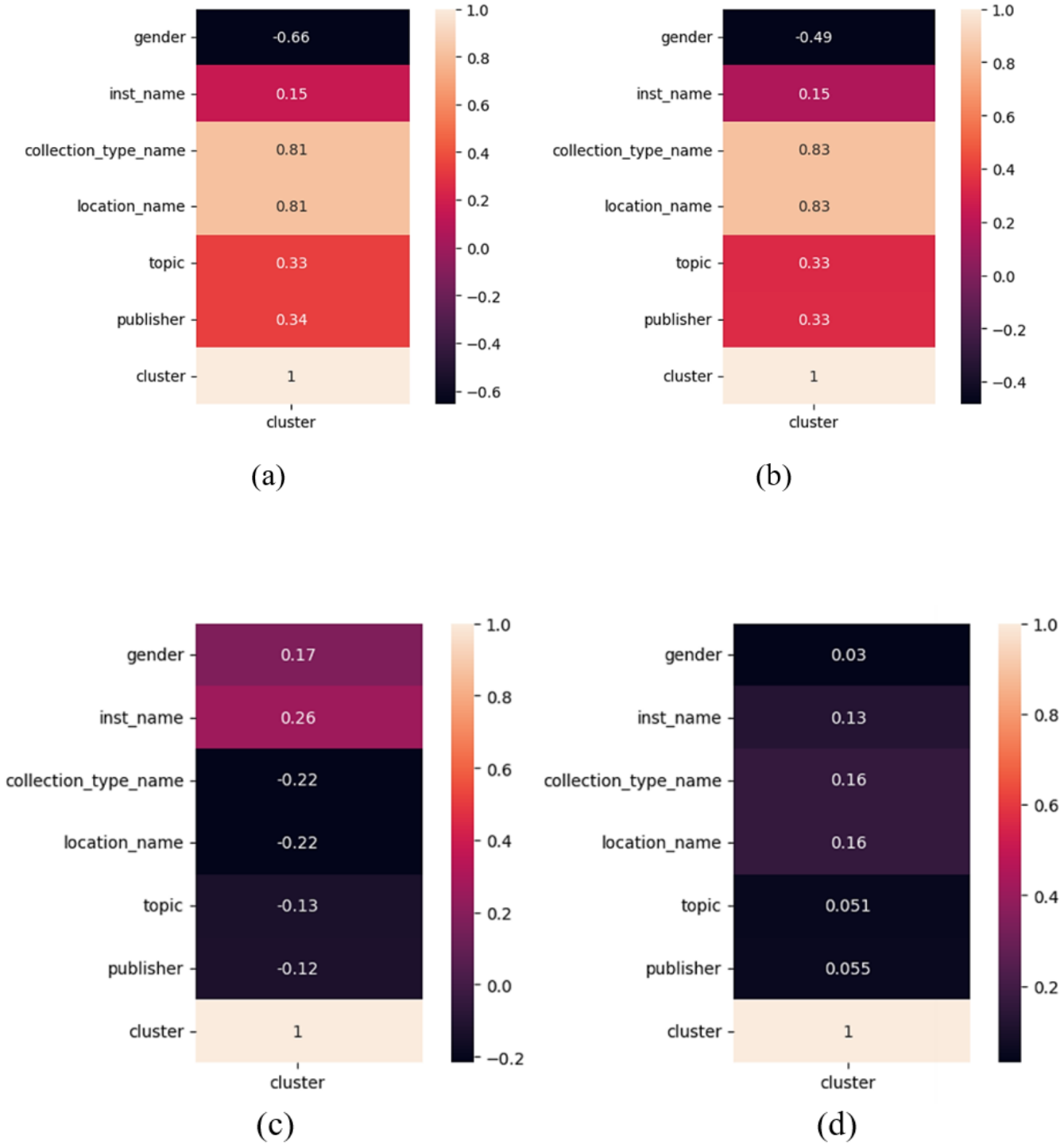

Figure 8: Correlation Value of Preference Features to (a) 3 Cluster Users, (b) 4 Cluster Users, (c) 5 Cluster Users, (d) 6 Cluster Users.

Additionally, the dataset's quality was compromised during the COVID-19 pandemic, which affected its representativeness. The study was based on a limited set of user preferences, focusing on favorite topics, publishers, locations and collection types. This narrow scope may only partially capture the diverse and nuanced preferences of users. Moreover, the study exclusively used K-Means Clustering and the Apriori algorithm without exploring other potential methods. Future research should address these limitations by utilizing a more diverse dataset, incorporating additional user preferences, experimenting with alternative clustering and recommendation techniques and refining the evaluation metrics.

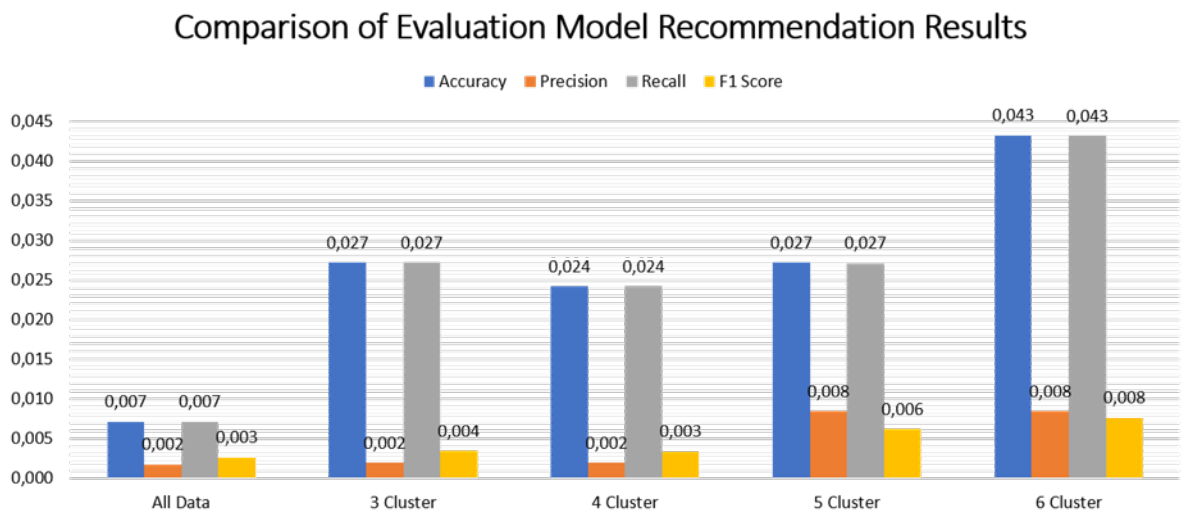

Figure 9: Comparison Chart TOP10 Top Book Publishers Interested by Users.

CONCLUSION

This study aims to develop a personalized book recommendation model by predicting books for each user based on their interests, behavior and other relevant information. The recommendation system was constructed using a combination of K-Means Clustering and Association Rule methods. The clustering method generated additional user cluster features, which enhanced the recommendation model. Furthermore, to address the Cold Start Problem, the study recommended popular books within each user cluster.

Experiments were conducted to determine the optimal number of user clusters. The silhouette score indicated that the best performance was achieved with k=6, yielding a score of 0.725. The evaluation metrics for the recommendation model with 6 user clusters were superior to those with other cluster numbers, showing an average accuracy of 4.329%, average precision of 0.848%, average recall of 4.326% and average F1 Score of 0.764%. This book recommendation model is effective in improving user reading experiences and enhancing library services.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

1. Rafiq M, Batool SH, Ali AF, Ullah M. University libraries response to COVID-19 pandemic: A developing country perspective. J Acad Librariansh. 2021;47(1):102280. doi: 10.1016/j.acalib.2020.102280.
2. Kortemeyer G, Dröschler S. A user-transaction-based recommendation strategy for an educational digital library. Int J Digit Libr. 2021;22(2):147-57. doi: 10.1007/s00799-021-00298-8.
3. Tian Y, Zheng B, Wang Y, Zhang Y, Wu Q. College library personalized recommendation system based on hybrid recommendation algorithm. Procedia CIRP. 2019;83:490-4. doi: 10.1016/j.procir.2019.04.126.
4. Makawana M, Mehta RG. Discovering search space using M-distance clustering of semantic relatedness based weighted network for the content-based recommender system. J Scientometr Res. 2023;12(2):243-53. doi: 10.5530/jscires.12.2.024.
5. Adistia LD, Akhriza TM, Jatmiko S. Sistem Rekomendasi Buku untuk Perpustakaan Perguruan Tinggi Berbasis Association Rule. J Resti (Rekayasa Sist dan Teknol Informasi). Vol. 3; 2019 Aug(2 SE-Artikel Rekayasa Sistem Informasi).
6. Saraswat M, Srishti. Leveraging genre classification with RNN for Book recommendation. Int J Inf Technol. 2022;14(7):3751-6. doi: 10.1007/s41870-022-00937-6.
7. Putra IM, Indrawan G, Aryanto KY. Sistem Rekomendasi Berdasarkan Data Transaksi Perpustakaan Daerah Tabanan dengan menggunakan K-Means Clustering. J Ilmu Komput Indones. 2018;3(1):18-22.
8. Kumari P, Kumar R. Clustering scientometrics of computer science journals for subarea decomposition. J Scientometr Res. 2023;12(2):383-94. doi: 10.5530/jscires.12.2.034.
9. Ahmar AS, Napitupulu D, Rahim R, Hidayat R, Sonatha Y, Azmi M. Using K-means clustering to cluster provinces in Indonesia. J Phys Conf S. 2018; 1028:12006. doi: 10.1088/1742-6596/1028/1/012006.
10. Syakur MA, Khotimah BK, Rochman EM, Satoto BD. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. IOP Conf S Mater Sci Eng. 2018;336:12017. doi: 10.1088/1757-899X/336/1/012017.
11. Sutisnawati Y, Reski M. Looking for transaction data pattern using apriori algorithm with association rule method. IOP Conf S Mater Sci Eng. 2019;662(2):22078. doi: 10.1088/1757-899X/662/2/022078.
12. Davagdorj K, Ryu KH. Association Rule Mining on Head and Neck squamous cell carcinoma Cancer using FP Growth algorithm. In: Proceedings of the international conference on information, system and convergence applications; 2018.
13. Harahap M, Husein AM, Aisyah S, Lubis FR, Wijaya BA. Mining association rule based on the diseases population for recommendation of medicine need. J Phys Conf Ser. 2018; 1007:12017. doi: 10.1088/1742-6596/1007/1/012017.
14. Bandyopadhyay S, Thakur SS, Mandal JK. Product recommendation for e-commerce data using association rule and apriori algorithm. In: Modelling and simulation in science, technology and engineering mathematics. Proceedings of the international conference on modelling and simulation (MS-17). Springer; 2019. p. 585-93. doi: 10.1007/978-3-319-74808-5_51.
15. Uludag K. Hyperparameters and tuning methods for random forest using python Sklearn package relevant to psychology studies. In: Clinical practice and unmet challenges in AI-enhanced healthcare systems. IGI Global; 2024. p. 204-19. doi: 10.4018/979-8-3693-2703-6.ch011.
16. Cui M. Introduction to the k-means clustering algorithm based on the elbow method. J Acc Audit Finan. 2020;1(1):5-8.
17. Chen S, Peng Y. Matrix factorization for recommendation with explicit and implicit feedback. Knowl Based Syst. 2018;158:109-17. doi: 10.1016/j.knosys.2018.05.040.
18. Ding J, Yu G, He X, Quan Y, Li Y, Chua TS, *et al.* Improving implicit recommender systems with view data. In: IJCAI; 2018. p. 3343-9. doi: 10.24963/ijcai.2018/464.
19. Li C, Zhou B, Lin W, Tang Z, Tang Y, Zhang Y, *et al.* A personalized explainable learner implicit friend recommendation method. Data Sci Eng. 2023;8(1):23-35. doi: 10.1007/s41019-023-00204-z.
20. Baltrunas L, Amatriain X. Towards time-dependent recommendation based on implicit feedback. In: Workshop on context-aware recommender systems (CARS'09). Citeseer; 2009. p. 25-30.
21. Cantabella M, Martínez-España R, Ayuso B, Yáñez JA, Muñoz A. Analysis of student behavior in learning management systems through a Big Data framework. Futur Gener Comput Syst. 2019;90:262-72. doi: 10.1016/j.future.2018.08.003.

22. Wang F, Li K, Duić N, Mi Z, Hodge BM, Shafie-khah M, *et al.* Association rule mining based quantitative analysis approach of household characteristics impacts on residential electricity consumption patterns. Energy Convers Manag. 2018;171:839-54. doi: 10.1016/j.enconman.2018.06.017.
23. Naresh P, Suguna R. Association rule mining algorithms on large and small datasets: A comparative study. In: International Conference on Intelligent Computing and Control Systems (ICCS). Vol. 2019. IEEE Publications; 2019. p. 587-92. doi: 10.1109/ICCS45141.2019.9065836.
24. Borah A, Nath B. Identifying risk factors for adverse diseases using dynamic rare association rule mining. Expert Syst Appl. 2018;113:233-63. doi: 10.1016/j.eswa.2018.07.010.
25. He Q, He W, Song Y, Wu J, Yin C, Mou Y. The impact of urban growth patterns on urban vitality in newly built-up areas based on an association rules analysis using geographical 'big data.' Land Use Policy. 2018;78:726-38. doi: 10.1016/j.landusepol.2018.07.020.
26. Hernandez C. JB, Garcia-Medina A, Porro V MA. Study of the behavior of cryptocurrencies in turbulent times using association rules. Mathematics. 2021;9(14):1620.
27. Istiawan D. Poverty mapping in Central Java Province using K-means algorithm. J Intell Comput Heal Inform. 2020;1(1):1-4. doi: 10.26714/jichi.v1i1.5380.
28. Sharma PK. Means clustering simplified in python [Internet]; 2022. Analytics Vidhya [cited Sep 18 2022]. Available from: https://www.analyticsvidhya.com/blog/2021/04/k-means-clustering-simplified-in-python/.
29. Putra AA, Nasucha M, Hermawan H. K-means clustering algorithm in web-based applications for grouping data on scholarship selection results. In: International Symposium on Electronics and Smart Devices (ISESD). Vol. 2021. IEEE Publications; 2021. p. 1-6.
30. Yannam VR, Kumar J, Babu KS, Sahoo B. Improving group recommendation using deep collaborative filtering approach. Int J Inf Technol. 2023;15(3):1489-97. doi: 10.1007/s41870-023-01205-x.
31. Sharma S, Rana V, Malhotra M. Automatic recommendation system based on hybrid filtering algorithm. Educ Inf Technol. 2022;27(2):1523-38. doi: 10.1007/s10639-021-10643-8.
32. Kafi Al A, Banshal SK, Sultana N, Gupta V. Source recommendation system using context-based classification: empirical study on multi-level ensemble methods. 2024;13(2):475-84.

Cite this article: Amin FM, Rusydiyah EF, Azizah AN. Personalized Library Book Recommendations Using K-Means Clustering and Association Rules. J Scientometric Res. 2025;14(1):32-45.