Sinopfbd, 2025; 10(2): 474-496 https://doi.org/10.33484/sinopfbd.1688898

Sinop Üniversitesi Fen Bilimleri Dergisi
Sinop Uni J Nat Sci
E-ISSN: 2564-7873

https://dergipark.org.tr/tr/pub/sinopfbd

Book Rating Analysis Based on Feature Selection and Apriori Algorithm

Merve KÖLE[1] and Emin BORANDAĞ[2]

How to cite: Köle, M., & Borandağ, E. (2025). Book rating analysis based on feature selection and Apriori algorithm. *Sinop Üniversitesi Fen Bilimleri Dergisi, 10*(2), 474-496. https://doi.org/10.33484/sinopfbd.1688898

Research Article

Corresponding Author
Emin BORANDAĞ
emin.borandag@cbu.edu.tr

ORCID of the Authors
M.K: 0000-0002-8999-6072
E.B: 0000-0001-5553-2707

Received: 03.05.2025
Accepted: 29.09.2025

Abstract

Digital systems record data on the books we read, items we purchase, or the products we enjoy, all within their own framework for the purpose of data mining. The information to be derived from these data through data mining methods is of great importance for these systems. Data mining aims to speed up the decision-making process by uncovering hidden relationships and information in data via mathematical and statistical techniques. In this context, data mining techniques were applied to the dataset containing book votes through the use of the Apriori algorithm in this study. In this way, the relationships within the data set were identified, rules were generated, and the connections between the generated rules and user preferences were disclosed. With the application of the Apriori algorithm, rules were generated that show books associated with other books, as well as authors often liked together with other authors. The study developed rules indicating which other books a reader might enjoy based on their interest in a particular book, as well as which authors might appeal to readers who like a given author. The rules produced in the study are intended to be used in book recommendation lists and sales strategies.

Keywords: Feature selection, data mining, association rules, apriori algorithm, book recommendations, machine learning

Apriori Algoritması ile Özellik Seçimi Algoritması Tabanlı Kitap Oylama Analizi

[1]EÜAŞ - The Electricity Generation Corporation, Department of Information Technologies, Ankara, Türkiye

[2]Celal Bayar University, Faculty of Technology, Department of Software Engineering, Manisa, Türkiye

This work is licensed under a Creative Commons Attribution 4.0 International License

Öz

Dijital sistemler, veri madenciliği amacıyla, okuduğumuz kitapların, satın aldığımız ürünlerin veya beğendiğimiz ürünlerin verilerini kendi çerçeveleri içinde kaydederler. Bu verilerden veri madenciliği yöntemleriyle elde edilecek bilgiler, dijital sistemler için büyük önem taşımaktadır. Veri madenciliği, matematiksel ve istatistiksel tekniklerle verilerdeki gizli ilişkileri ve bilgileri ortaya çıkararak karar alma sürecini hızlandırmayı amaçlamaktadır. Bu bağlamda, bu çalışmada Apriori algoritması kullanılarak kitap oylarını içeren veri kümesine veri madenciliği teknikleri uygulanmıştır. Bu şekilde, veri kümesi içindeki ilişkiler belirlenmiş, kurallar üretilmiş ve üretilen kurallar ile kullanıcı tercihleri arasındaki bağlantılar ortaya çıkarılmıştır. Apriori algoritmasının uygulanmasıyla, beğenilen kitapları diğer kitaplarla ve beğenilen yazarları diğer yazarlarla birlikte gösteren kurallar üretilmiştir. Çalışmanın sonucunda, belirli bir kitabı beğenen bir kişinin hangi diğer kitapları beğenebileceği ve belirli bir yazarı beğenen bir kişinin hangi diğer yazarları beğenebileceği konusunda kurallar geliştirilmiştir. Çalışmada üretilen kuralların, kitap öneri listelerinde ve satış stratejilerinde kullanılması amaçlanmaktadır.

Anahtar Kelimeler: Özellik seçimi, veri madenciliği, ilişkilendirme kuralları, Apriori algoritması, kitap önerileri, , makine öğrenmesi

Introduction

Nowadays, predictions for the future have gained great importance. Computer technologies are typically used for such predictions by employing recorded data [1]. Practically valuable knowledge should be discovered by applying techniques and rules to these recorded data. The process of extracting information from data refers to data mining [2, 3]. One of the most widely used models in data mining is association rule mining, particularly the Apriori algorithm. Basket analysis, also known as market basket analysis, identifies associations between products purchased together [4]. Besides, several algorithms have been developed for the association rules technique [5]. One of these algorithms is the Apriori algorithm introduced by Agrawal and Srikant in 1994 [6]. The Apriori algorithm has been the subject of numerous studies in the literature. In the study, the preprocessing and data analytics stages were carried out on the BookCrossing dataset as in the preprocessing and data analytics stages in the ICTs-based architecture [7]. This study employed data mining approaches on the dataset of book ratings using the Apriori algorithm in this manner, and discovered the relationships in the dataset, generated the rules, and revealed the reader preferences with the generated rules. Accordingly, the applied Apriori algorithm generated rules that characterize interrelated preferences among books and authors in the dataset. As a result, the findings of the study established rules about which books individuals would prefer if they liked a particular book and which author they would choose if they preferred a particular author. It is expected that these rules will also apply to book recommendation lists and sales strategies. The main contributions of the study to the literature are enumerated below; 1. Studies in the literature have generally selected datasets related to health and sales. There are studies conducted on situations in which only the users are in the datasets. In contrast, in our study, the Apriori algorithm was run through a survey dataset related to the user's own book ratings, which involve the user's own likes. 2. The data set in the study, which contains voting, was adapted to the general structure of the Apriori algorithm, thus demonstrating that the Apriori algorithm can also run on survey data sets. 3. In the studies in the literature, association rules are inadequate because the liking information of the products purchased together is not included in the dataset. In this study, the association rule was established based on the books both purchased and liked. 4. Another difference in our study from the examples in the literature is that not only the Apriori algorithm was run on the dataset. Considering the structure of the Apriori algorithm based on the frequency perspective, the effect of the ratings in the dataset on the density was ascertained by feature selection, and association rules were extracted based on the most effective ratings in determining the density. The rest of the paper is organized as follows: Section-2 presents the Related Works, Section-3 presents the Materials and Methods and clarifies the Proposed Methods, and finally, Section-4 presents the Conclusion.

Related Works

Apriori algorithm has been the subject of numerous studies in the literature. Özçakır et al. [8] developed software in the Visual Basic programming language on the Microsoft.NET platform using the pastry sales data of a patisserie company and employing the association rules method to perform the data mining process. The data retained the sales from seven sales branches comprising the range between the last four months of 2002 and the first two months of 2006. Accordingly, these data were analyzed by applying certain thresholds with the Apriori algorithm in the developed software. Different durations and sales locations were used as input values to identify which sales were associated with goods purchased together [8]. In 2008, Erdem et al. used the association rule method to analyze the data of 214.000 patients who applied to the emergency department and discovered several rules. The study's findings shed light on patient profiles and reasons for requesting emergency care, which varied based on geographical variations. It also contributed to structuring the emergency departments by offering distinctive perspectives [9]. Ay et al. [10] a data-oriented decision support system application within the supermarket's settlement layout conducted in Migros Türk A.Ş., one of the primary sales companies. The data set included 12.077 products and 558 product groups. The study operated the Apriori algorithm, one of the association rule algorithms, and multidimensional scaling techniques via preparing a relational database. As a result, the study revealed 518 rules for the layout system. Hong et al. [11] used the Apriori algorithm, which produces association rules, to analyze truck accidents on Korean highways. The content of the dataset in the study consisted of truck accidents that occurred on Korean highways from 2008 to 2017. Erpolat [12] analyzed the December 2009 invoices of its clients in an authorized car care service operating in the Turkish automotive sector, comprising a dataset with 50 different products and 40 invoices in total. The data analysis was executed by Apriori and FP-Growth algorithms utilizing Weka and Excel programs. The study surveyed which products the customers purchased in tandem. Accordingly, efforts were made to design campaigns and promotions aimed at increasing profits based on these findings. Bilgin and Acun [13] generated software employing the Apriori algorithm to analyze the error logs of the resource planning software designed for the metal, furniture, and retail industries. The Apriori algorithm aided in identifying the correlations between the factors that create errors via applying this software and resulted in developing software documenting such correlations. Accordingly, the study conclusively yielded one quintuple, six quartets, and 14 tripartite rules and identified the pattern that led to the most error amount. Doğrul et al. [14] employed the Apriori algorithm, which generates the association rules, to analyze the traffic accident data of Çankırı province. The study dataset retained traffic accidents that occurred in 2013, which Çankırı Gendarmerie and Police Teams reported. As a result, the analysis revealed the rules about the places and dates when accidents occurred, and thus these rules indicated that additional measures would avoid such accidents. Çelik et al. [15] found association rules by applying the Apriori algorithm to the sales data of a retail company in Kütahya province. After excluding the data in a form too small to affect the analysis, the study discovered the

products purchased in tandem, listed them as a rule, and identified customers' demographic features. In a study on global cyberspace security conducted by Li et al. [16], 181 rules were extracted from 56096 websites by applying the Apriori algorithm and association rules. In another study conducted by Wang et al. [17], association rules were used in the Apriori algorithm. In addition, a MapReduce model was proposed to facilitate the use of the Apriori algorithm. Table 1 shows the studies related to the Apriori algorithm.

Table 1. *Comparison of works regarding the Apriori algorithm.*

Author	Subject	Type of Dataset	Feature Selection	Algorithm Name
Özçakır et al. [8]	Cake Sales Data	Sales	No	Apriori
Erdem et al. [9]	Emergency Service	Health	No	Apriori
Ay et al. [10]	Market Sales	Sales	No	Apriori
Hong et al. [11]	Traffic	Health	No	Apriori
Erpolat et al. [12]	Automotive Sector	Sales	No	Apriori, FP-Growth
Acun et al. [13]	Software Error Logs	Software	No	Apriori
Doğrul et al. [14]	Traffic	Health	No	Apriori
Çelik et al. [15]	Retail Company	Sales	No	Apriori
Li et al. [16]	Cyberspace Security	Software	No	Apriori
Wang et al. [17]	Commodity Hardware	Software	No	Apriori
The present study	BookCrossing	Rating	Yes	Apriori

Materials and Methods

This section provides data for the principles of data mining, association rules, the Apriori algorithm, and feature selection.

Data Mining

Data mining employs various programs and algorithms on databases to forecast future trends, thereby extracting patterns that describe the dataset [18].

Association Rules

Agrawal et al. [19] first proposed the concept of association rules. Association rules are a method developed to explore interesting relationships between variables in large databases. These rules are one of the commonly used techniques in data mining. The association rules technique can be used to analyze and reveal relationships within the data [20, 21]. Market basket analysis is one of the most popular applications of association rules [22].

Mathematical Model

I = {i1, i2...,ip} refers to the cluster (set) of objects. p denotes the number of elements. D, however, stands for the set of operations. Each transaction in database D is a set of items defined as F $\subset$ I. F

symbolizes each transaction, whereas the k-factor set designates the cluster with k items. However, FID is the only separator that belongs to each transaction. FID is a unique number for each transaction [23]. The association rule can be defined as follows; A1, A2, ..., Ap → B1, B2, ….., Bn, or A→B. This rule displays whether object A exists when object B is in the same event or action. Association rules consist of two halves; while the left side of the rule (A) is referred to as the antecedent, the right side (B) is the consequent [7].

Definitions

Association rules consisting of relations between objects in association analysis are calculated with support and confidence criteria [24]. Association rules are expressed as A→B, where A is the antecedent and B is the consequent. Their strength is measured by support Equation 1 [6] and confidence Equation 2 [25].

S(A, B) = (Number of lines with A and B together)/(Total number of rows) (1)

Confidence(A,B) = (Number of rows containing A and B) / (Number of rows containing A) (2)

Minimum Support and Minimum Confidence

These values – Minimum Support and Minimum Confidence– are generated by the developer. The validity of the established association rules depends on provisions of the Equations 3 and 4 given below [26].

$$\text{support}(A \rightarrow B) \geq \min(\text{support}) \quad (3)$$

$$\text{confidence}(A \rightarrow B) \geq \min(\text{confidence}) \quad (4)$$

Lift

It is another calculation of values used in association rules. The equation is given in (5). It specifies a negative correlation between A and B when the value is less than one and a positive correlation when it is higher than one. If the value is exactly equal to 1, it signifies that A and B are independent of each other [5].

$$\text{Lift}(A \rightarrow B) = (\text{confidence}(A \rightarrow B))/(\text{support}(B)) \quad (5)$$

Apriori Algorithm

The Apriori algorithm was created in 1994 by Agrawal and Srikant [6]. It is employed to specify the association rules of the data [27]. This algorithm takes advantage of the previous step by using a priori knowledge of frequently recurring objects [6]. The name of the Apriori Algorithm originates from the word 'Apriori' referring to the term 'prior' as it constantly takes information from the previous step (prior) [28]. Initially, the database is scanned to find the total number of each item, and an item set is created by combining items providing minimal support. Then, the acquired set is employed to locate subsets. This process resumes until finding all associations. It requires scanning the database to find each subset of associations. During these scans, the Apriori algorithm uses the Apriori property, which improves

efficiency and reduces the scanning area [28]. If a set of k-items satisfies the minimum support criterion, then the subsets of this set also meet the minimum support criterion, according to the characteristics of the Apriori algorithm [1]. Table 2 below lists the annotations used for the Apriori algorithm.

Table 2. *Notations*

Symbol	Description
E	Minimum support (threshold) value
k-itemset	an item set with k items
Mk	Frequent k item set
Ck	Candidate element set with K item (potentially frequent item sets)
T	Set of operations
D	The set of database-related transactions
C	Confidence level (value)
M	Frequent item set

The Apriori algorithm consists of two phases: joining and pruning steps. In the joining step, each candidate item set (Ck) with k elements is available by combining the Mk-1 among themselves to find Mk. A database scan is required to calculate Mk's to identify the number of candidates in Ck. Extensive computations are required when the number of Ck candidates is large [29]. The Apriori algorithm's second phase is required to reduce the Ck number. In the pruning stage, any infrequent (k-1) candidate item set cannot be a subset of a frequent k candidate set. If any (k-1) subset of a candidate k-item set is not present in Mk-1, the candidate is removed from Ck since it is an infrequent candidate set [6]. The classic summary code of the Apriori algorithm is displayed below [7].

Feature Selection

Feature selection refers to a selection process of the best g attributes among u attributes in the dataset by evaluating those features according to the applied algorithm [30]. Thanks to the feature selection, it eliminates the superfluous features in the data set and reduces the running time of the learning algorithm. This situation makes it possible to develop a better classification model [31].

Information Gain

Information Gain is an entropy-based feature selection method. Entropy is a term standing for describing the degree of ambiguity or unpredictability within a system [32]. The Information Gain score indicates the difference between the information required to identify the J feature and the knowledge necessary to recognize the J feature using the T feature. The entropy model is employed in the calculation of the Information Gain score. The entropy value ranges between 0 and 1. The greater the uncertainty, the closer this score is to 1. Equations 6, 7, and 8 are the formulas used to calculate information gain, which depicts how the entropy value of the J feature changes based on the T feature [33].

$$R(j) = \sum_{j \in J} p(j)\, log^2(p(j)) \tag{6}$$

$$R\left(\frac{J}{T}\right) = \sum_{t \in T} p(t) \sum_{j \in J} p\left(\frac{j}{t}\right) . log_2(p(j/t)) \quad (7)$$

$$Information\ Gain = R(J) - R(J/T) \quad (8)$$

Classifier Attribute Eval

It calculates the value of an attribute with the help of a classification algorithm selected by the user [34].

Gain Ratio

The Gain Ratio is an unsymmetrical measure. This criterion is employed to counteract the propensity to select attributes of information gain with a wide range of values. The Gain Ratio values range from 0 to 1. When the gain ratio is equal to 1, it explicitly states that the T information may accurately predict the J data. When the Gain Ratio is equal to 0, it indicates no correlation between J and T. Equation 9 illustrates the calculation of the Gain Ratio [35].

$$\text{Gain Ratio} = \text{Information Gain} / R(T) \quad (9)$$

One-R

One-R is a straightforward feature selection method proposed by Holte [36]. A rule is generated initially by the One-R algorithm for each attribute in the training dataset. Secondly, classification accuracy values are computed for every generated rule. Then, the rule's feature with the lowest error is chosen [35]. Each variable in the data set is given its model in this manner. For each model generated, the model is split by considering the different values the variable can take on behalf of the data set. After this step, the class that iterates more frequently for each divergent value is recognized as the base class, whereas samples from other classes are considered an error. The total error of the model is determined by summing the errors calculated for each unique value. The models are arranged based on total errors from the minor to the major. Then, the variable comprising the model below the predetermined threshold value is removed from the data set. A new data set is obtained in this manner [37].

Relief

Kira et al. [38] developed the relief algorithm in 1992 and defined it as a prospective feature selection method that weights the features according to their interrelationships. This method generates successful outcomes for datasets with two classes; however, it fails to function for datasets with multiple classes. Kononenko developed the ReliefF algorithm in 1994 to eliminate this constraint; as a result, it also performed well for datasets with more classes. The three fundamental components comprising the relief algorithm are as follows [39]. To define the corresponding feature value in the nearest instance with the same and different classes. To calculate the weight of the relevant attribute. To range the attributes based on their weights and select the top three features falling within a given threshold value.

Symmetrical Uncertainty

Symmetrical Uncertainty (SU) eliminates the adverse effects of information gain. It calculates the information gain by dividing it by the sum of the entropy values of J and T. The symmetric uncertainty coefficient value ranges between 0 and 1. It indicates that T information may predict J information if it is equal to 1. However, it signifies no correlation between J and T if it is equivalent to 0 [35]. Equation 10 below depicts the formula for the symmetric uncertainty coefficient.

$$SU = 2 * \text{Information Gain} / R(J) + R(T) \quad (10)$$

Results and Discussion

In this study, links between authors and between books were determined by using book rating data and the Apriori algorithm. Additionally, the study aimed to assess multiple books and authors preferred simultaneously based on readers' ratings. The dataset was analysed to determine the interlinks between authors and novels. As a result, these analyses led to the discovery of association rules.

Dataset

The Book-Crossing dataset employed in the study is a collection containing the ratings of the books by the readers. The data had previously been assembled by Cai-Nicolas Ziegler [40]. The Book-Crossing dataset consists of 3 components: the ratings of the books (BXBookRatings), the features of the books (BXBookDescription), and the attributes of the readers (BXUsers). Table 3 presents information about the dataset.

Table 3. *Contents of the book-crossing dataset.*

Related data	Numeric data
Total Number of Voting	1.149.779
Total Number of Users	278.858
Total Number of Books	271.065
Vote / Book	4.24
Vote / User	4.12
Book/User	0.97

Proposed Methods

The selected dataset helped inferencing the association rules that revealed the author-based relations and book-based links. The sub-headings "Association Rule Inference According to Author Ratings by Readers" and "Apriori Algorithm Results for Chosen Authors" explain the rule inference according to the authors, whereas the sub-headings "Association Rules Inferred from Users' Favourite Books" and "Association Rules Results for Chosen Books" exemplify the details and results of the rule inference studies according to the selected books. Association Rule Inference According to Author Ratings by Readers. The study findings revealed the rules whereby if readers evaluate author A, they may prefer

other authors. The relationships between authors derived from these rules are expected to be valuable for the book recommendation systems and suggested product lists. The pipeline of the workflow is given in Figure 1. There were five steps involved in the operations of the database used.

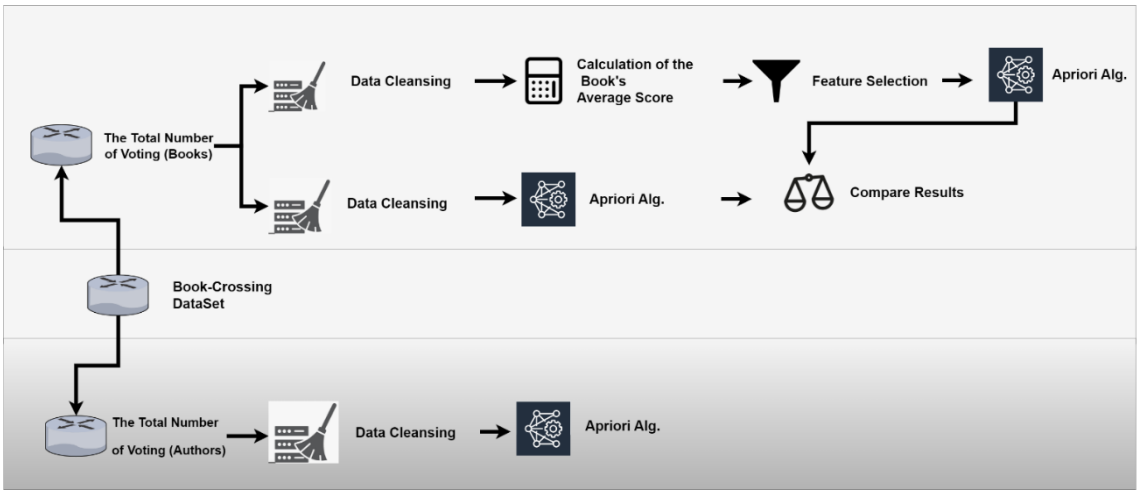


Figure 1. *The pipeline of the workflow.*

Data Generation

The SQL Server Management Studio was used to create the data used in the study [41] by utilizing the SQL files of the Book-Crossing dataset. Data entry into the database was performed using the T-SQL language. The dataset consisted of 1.149.779 ratings, 271.065 books, and 278.858 readers (users).

Data Cleansing

The ratings for the books available in the BXBookRatings table, albeit missing in the BXBooksDescription table, were excluded from the dataset. As a result, this process eliminated 118.604 ratings. SQL Server Management Studio in the T-SQL scripting language served the elimination process. After matching the BXBookRatings table comprising rating data and BXBooksDescription tables containing book descriptions, the process cleansed the unmatched ratings.

Data Integrity

The BXBookRatings table and the BXBooksDescription table are joined using the ISBN column to acquire the necessary UserID, ISBN, BookRating, and BookAuthor data columns. Data Selection: The data to be included in the study were chosen using four steps.

Selection of Required Columns

The UserID, BookAuthor, and BookRating columns were chosen from the data set to infer the association rules. The reason to select them is that they produce the rows and columns of the matrix used as input for the UserID and BookAuthor in the Apriori algorithm. However, the purpose of selecting the

BookRating column was to calculate the average score of the authors and use this score to select the dataset.

Calculation of the Author's Average Score

Estimation of the author's score was done based on the chosen columns. Any ratings of the readers were considered directly for any author's book; however, in cases where the readers rated the same author's book more than once, the average author score was considered for analysis.

Data Selection with Medium and High Scores

Since the inference of the association rules based on the authors rated by the users, medium- or higher-scored ratings were chosen from the dataset. Accordingly, the data set consisted entirely of the authors rated by the users. There were 797.908 ratings in the dataset before the elimination. This number dropped to 279.915 ratings in the dataset after selecting the medium or higher author average scores.

Cleansing of Sparse Data Below Average

Since the generation of an association rule is unfeasible for books with a total of one rating in the dataset, the books with multiple ratings were considered to generate these rules. As a result, the availability ratios of all authors in the dataset were calculated. The authors below the average score were disqualified based on the average ratings of the author's books in the data set. Table 4 contains numerical information about the chosen data.

Table 4. *Numerical information of selected data*

Related Data	Numeric Data
Total Number of Voting	11.622
Total Number of Authors	3.067
Total Number of Users	4.240

Data Conversion

The dataset was converted into a matrix format compatible with the Apriori algorithm (1, 0 or True, False) [42]. Other processes related to data transformation are specified below. All author names were coded in uppercase format, removing all punctuations to make the features unique. The dataset was transformed into a matrix format. There was an encoding process with the following scorings: If the user rated the author above five or more, the matrix scored it as 1 in the dataset; however, if the author had no rating at all, the matrix retained it as 0.

Apriori Algorithm Results for Chosen Authors

This section entitles the frequent item sets in the results of the Apriori algorithm, the rules generated by the Apriori algorithm, and the visualization of these rules. The Apriori algorithm initially discovers the frequent item sets by an elimination process. The total number of these sets is 1.229. The number of

frequent-item sets with one item is 599, according to the Apriori algorithm results. Table 5 illustrates algorithm results.

Table 5. *Number of frequent item clusters belonging to the authors*

Related Data	Numeric Data
M1 frequency number of 1-item clusters	599
M2 frequency number of 2-item clusters	529
M3 frequency number of 3-item clusters	89
M4 frequency number of 4-item clusters	12

Table 5 below depicts the frequency algorithm results for the M4 item set as an exemplary. Additionally, Table 6 below displays the information about the authors to further clarify the frequent item sets and rules and Table 7 illustrates the information about the authors [43].

Table 6. *The frequency of the 4-item set belonging to the authors*

No	Author 1	Author 2	Author 3	Author 4
1	JOHANNA LINDSEY	AMANDA QUICK	JULIE GARWOOD	JUDE DEVERAUX
2	JULIA QUINN	NORA ROBERTS	JULIE GARWOOD	CATHERINE ANDERSON
3	LAVYRLE SPENCER	NORA ROBERTS	CATHERINE ANDERSON	SUSAN E. PHILLIPS
4	JULIE GARWOOD	JD ROBB	NORA ROBERTS	CATHERINE COULTER
5	JULIE GARWOOD	JOHANNA LINDSEY	NORA ROBERTS	CATHERINE COULTER
6	JULIE GARWOOD	NORA ROBERTS	CATHERINE COULTER	NICHOLAS SPARKS
7	JAMES PATTERSON	JOHN GRISHAM	PATRICIA SCORNWELL	JOHN SANDFORD
8	JANET EVANOVICH	NORA ROBERTS	JULIE GARWOOD	JENNIFER CRUISE
9	JD ROBB	NORA ROBERTS	JULIE GARWOOD	JUDE DEVERAUX
10	JUDITH MCNAUGHT	JULIA QUINN	JULIE GARWOOD	SUSAN E. PHILLIPS
11	JUDITH MCNAUGHT	LAVYRLE SPENCER	NORA ROBERTS	JULIE GARWOOD
12	JUDITH MCNAUGHT	NORA ROBERTS	JULIE GARWOOD	SUSAN E. PHILLIPS

Table 7. *Information about the authors*

No	Author 1	Subject	Related Field
1	JOHANNA LINDSEY	Romance, Historical Love	For Table 6.
2	AMANDA QUICK	Contemporary Romanticism	For Table 6.
3	JULIE GARWOOD	Romance, Historical Love	For Table 6.
4	JUDE DEVERAUX	Romance, Historical Love	For Table 6.
5	BRIAN LUMLEY	Horror, Fiction	For Figure 1
6	ANNE RICE	Horror, Paranormal	For Figure 1
7	DEAN KONTZ	Horror, Thriller	For Figure 1
8	STEPHEN KING	HORROR, THRILLER	FOR FIGURE 1

Considering the frequency of the M4 item cluster, Johanna Lindsey, who is famous for her historical romance novels, ranked as the first item, while Amanda Quick, known for her contemporary romantic suspense novels, ranked as the second, Julie Garwood, who writes American romance novels, was the third, and Jude Deveraux, known for his romance novels, was the fourth. Table 8 illustrates the numerical information of the rules generated by the Apriori algorithm, whereas Table 9 lists author association rules and confidence values.

Table 8. *Numerical information on the rules of the relationship between authors*

RULE CONFIDENCE VALUE	NUMERIC DATA
rules with 100% confidence	96
rules with 90% and more confidence	96
rules with 80% and more confidence	174
rules with 70% and more confidence	187

Table 9. *Author association rules with 100% confidence*

NO	ANTECEDENT	CONSEQUENT	C
1	JAMIE SUZANNE	FRANCINE PASCAL	1
2	JOSEPH CONRAD	THOMAS HARDY	1
3	MARY JO PUTNEY	JULIE GARWOOD	1
4	AMANDA QUICK, JOHANNA LINDSEY	JUDE DEVERAUX	1
5	AMANDA QUICK, JOHANNA LINDSEY	JULIE GARWOOD	1
6	AMANDA QUICK, JUDE DEVERAUX	JULIE GARWOOD	1
7	ANNE MCCAFFERY, MERCEDES LACKEY	MARION Z. BRADLEY	1
8	CATHERINE ANDERSON, CHRISTINE FEEHAN	JULIA QUINN	1
9	CATHERINE ANDERSON, JULIE GARWOOD	JULIA QUINN	1
10	JULIA QUINN, NORA ROBERTS	CATHERINE ANDERSON	1
11	JAMIE SUZANNE	FRANCINE PASCAL	1
12	JOSEPH CONRAD	THOMAS HARDY	1

The first rule in Table 9 states the following: A user (reader) who likes the books of the author Jamie Suzanne also rates the books of the author Francine Pascal. The confidence level of this rule was 100%. People who prefer Jamie Suzanne also prefer Francine Pascal as an author. However, the thirdly ranked rule remarks as follows. the user who rates Mary Jo Putney also gives ratings to Julie Garwood. The confidence level of this rule is also 100%. Table 9 discusses the association rules mentioned above regarding confidence. lift values, another rule evaluation metric, are presented in Table 10. There is no inter-author association if this value is equal to 1. In other words, author a and author b are unrelated. However, a value greater than or equal to 1 indicates a positive correlation. According to rule 1, the lift value between Jamie Suzanne and Francine pascal was 32. This Figure denotes a positive correlation. there is no lift value equal to 1 among the top 10 rules, suggesting that all the rules are interrelated.

Table 10. *Evaluation metrics for the top 10 rules*

NO	ANTECEDENT	CONSEQUENT	C	LIFT
1	JAMIE SUZANNE	FRANCINE PASCAL	1	326
2	JOSEPH CONRAD	THOMAS HARDY	1	706
3	MARY JO PUTNEY	JULIE GARWOOD	1	151
4	AMANDA QUICK, JOHANNA LINDSEY	JUDE DEVERAUX	1	157
5	AMANDA QUICK, JOHANNA LINDSEY	JULIE GARWOOD	1	151
6	AMANDA QUICK, JUDE DEVERAUX	JULIE GARWOOD	1	151
7	MERCEDES LACKEY, ANNE MCCAFFREY	MARION Z. BRADLEY	1	146
8	CHRISTINE FEEHAN, CATHERINE ANDERSON	JULIA QUINN	1	471
9	JULIE GARWOOD, CATHERINE ANDERSON	JULIA QUINN	1	471
10	NORA ROBERTS, JULIA QUINN	CATHERINE ANDERSON	1	605

Association Rules Inferred from Users' Favourite Books: This study aimed to identify the associations between books by using information about the books that readers rated and the ratings they gave to these books. The analysis will reveal which other possible books readers who prefer book a might also choose through association rules. The associations identified among the books that might uncover by these rules will enable the development of book recommendation systems. The related study steps are as follows: data generation, data selection, and application of the Apriori algorithm together with the data mining process. Table 11, show that numerical information on book voting.

Table 11. *Numerical information on book voting*

RELATED DATA	NUMERIC DATA
Total number of voting	1.149.779
Total number of users who vote	105.282
Total number of books voted	269.862
Post-qualification voting	1.031.175
Number of users after elimination	92.107
Number of books after elimination	269.862

Data Generation: The dataset consisted of 1.149.779 ratings, 271.065 of which were books and 278.858 were users (readers). the study employed the BXBOOKRATINGS table containing reader ratings of books. The book ratings, not present in the BXBooksDescription table, were eliminated from the dataset this process removed 118.604 ratingsfrom the dataset. the data selection process for the study involved four steps.

Selection of Ratings Greater Than or Equal To 5: Since the study would generate rules based on the books that users preferred, the rating scores greater than or equal to 5 were selected from the data set. There were 1.031.175 ratings in the cleansed dataset before the removal process. after this elimination, however, the total number of ratings lowered to 367.261.

Calculation of Rating Intensities of Books and Grouping: If Any book is referenced once in the dataset, this book is unsuitable for rule formation. hence, each book in the dataset was subject to the calculation for the rating status.

Density Group Selection Based on User/Book Ratio: For each density group, the user/book ratio was calculated, and thus the selection was grounded on the user/book ratio. The number of columns indicates books, whereas the number of rows displays unique users in the matrix inserted into the algorithm in the study. Ratings are likely to overlap when the user/book ratio is close or higher. Hence, it is possible to deduce the association rule from here since the rating ratio for the same book will increase. The user/book average of the divided groups based on their density in the data set is 96. As a result, groups between 6 and 16, whose user/book ratio was above average and equal to average and close to values between 48 and 192, respectively, were chosen from the data set.

Table 12. *Sparse-looking matrix.*

User ID	B1	B2	B3	B4	B5	B6	B7	B8
1	1	0	0	0	1	0	0	0
2	0	1	0	0	0	0	0	1
USERID	B9	B10	B11	B12	B13	B14	B15	B16
1	0	1	1	1	0	0	0	0
2	0	0	0	0	0	1	0	0

Table 13. *Dense-looking matrix*

User ID	B1	B2	B3	B4	B5	B6	B7	B8
1	1	0	0	0	1	0	0	0
2	0	1	0	0	0	0	0	1
USER ID	B9	B10	B11	B12	B13	B14	B15	B16
1	0	1	1	1	0	0	0	0
2	0	0	0	0	0	1	0	0

Table 12 illustrates the matrix with a high number of books and a low number of users. It appears sparse because the number of books is greater than the number of users. Table 13 displays a matrix with a denser appearance. Since the user/book ratio is larger than the matrix in Table 12, it gives a denser matrix appearance.

Rating Choice by Feature Selection Based on Score Density: Apriori algorithm was run to infer rules for the chosen data. Despite the lowest minimum support value, There were merely ten rules generated. Among the 18 groups, those listed between 6 to 16 were chosen for analysis in the preceding steps. since narrowing the data selection to more groups would result in book loss, the density distribution of each book within the score types was examined. The score column with the best impact on the density status was identified and the scores kept in this score column were used in the data series by comparing the weights between the group class showing the density status in the generic book dataset and the 5, 6, 7,

8, 9 and 10 rating scores. The colored sections in Figure 2 correspond to the following: orange denotes a book in group 15 (extreme density), green denotes a book in group 9 (moderate density), and blue refers to a book in group 5 (dense). The score group with the best density impact is identified by comparing the weights of these books over 5, 6, 7, 8, and 9 rating scores with the density groups in the generic dataset, and the other score groups are removed from the data set. when the result indicates that 8 and 10 ratings are effective on density, an association rule is generated based on rating scores of 8 and 10.



Figure 2. *The diagrams follow the same format.*

This process is accomplished by feature selection. The group columns indicate the density class in the generic data set. The score-based rating weights for this book are bookrating5, bookrating6, bookrating7, bookrating8, bookrating9, and bookrating10. Book a, the first in rank, received the following scores and frequencies: 5-point once, 6-point once, 7-point once, 8-point eight times, 9-point three times, and 10-point five times. Therefore, the generic density status of this book was signified by 6 (in number) in the group column. The score that best predicts the density group will receive the score column using feature selection algorithms on this dataset. The Apriori algorithm will be employed to conduct ratings belonging to this score column. Information gain, classifier attribute evaluator, gain ratio, one-r, relief, and symmetrical uncertainty attribute evaluator were the six algorithms used for the feature selection. Table 14 lists the results of the feature selection algorithms as sorted based on the effect of score types in determining denser groups. Application of the Apriori algorithm together with the data mining: this section primarily covers the running of the Apriori algorithm on the ratings of the score types in the chosen column in Table 14. The best rule-scoring types will prevail based on the data outcomes, then the results and rules of these score types will be evaluated accordingly.

Table 14. *Results of feature selection algorithms.*

METHOD	ATTRIBUTE RANKING	SELECTED SCORE TYPES
UNFILTERED	10, 8, 9, 7, 6, 5	8, 10
INFO GAIN	8, 9, 7, 10, 6, 5	8, 9
CLASSIFIER	10, 9, 6, 7, 8, 5	10, 9
GAIN RATIO	8, 7, 9, 6, 5	8, 7
ONE R	8, 9, 5 ,7, 6, 10	8, 9
RELIEF	9, 8, 10, 7, 5, 6	9, 8
SYMMETRICAL UNCERTAINTY	8, 9, 7, 10, 6, 5	8, 9

Table 15 displays the outcomes of the feature selection procedure, the firstly-chosen two points, the total number of ratings, the support value, and the support*ratings results, respectively. The unfiltered scores, in other words, the most iterative score types were used in the process. In the output performance process, there will be a comparison between the rule outcomes of the ratings derived from the algorithm-chosen score types and the rule outcomes generated by directly taking the ratings of the most redundant score type without utilizing feature selection. This comparison reveals how the feature selection procedure outperforms the Apriori algorithm. The attribute ranking in Table 15 illustrates the ranking of the feature selection algorithms based on the score types. The Apriori algorithm was fed by the data of the score types in the selected column.

Table 15. *Numerical information of selected data*

METHOD	ATTRIBUTE RANKING	FIRSTLY - CHOSEN ONES	THE TOTAL NUMBER OF RATINGS	SUPPORT	SUPPORT* RATING
UNFILTERED	10, 8, 9, 7, 6, 5	8, 10	19.869	0.0004	8
INFO GAIN	8, 9, 7, 10, 6, 5	8, 9	19.394	0.0004	8
CLASSIFIER	10, 9, 6, 7, 8, 5	10, 9	16.829	0.0005	8
GAIN RATIO	8, 7, 9, 6, 5	8, 7	18.482	0.0004	8
ONE R	8, 9, 5, 7, 6, 10	8, 9	19.394	0.0004	8
RELIEF	9, 8, 10, 7, 5, 6	9, 8	19.394	0.0004	8
SYMMETRICAL UNCERTAINTY	8, 9, 7, 10, 6, 5	8, 9	19.394	0.0004	8

Apriori rule inference results based on feature selection: Table 15 illustrates the ratings of the score types employed in running the Apriori algorithm. Table 16, however, presents the findings of these score types. When the first two score types are analysed directly in the dataset without using feature selection algorithms, the number of rules with 100% confidence is 78, according to the available dataset. However, the number of rules notably escalates to 786 when taking into account the outcomes of the feature selection algorithms. The scores that were explicitly the most recurring were 8 and 10 and using feature selection techniques instead of the logic of density in this example allowed for more rule inference. Analysis of the Table indicated that score types 8 and 9 produced the highest number of reliable rules. Ranking of the score types 8 and 9 was identified by information gain, One-r, Relief, and Symmetrical Uncertainty attribute evaluator algorithms.

Table 16. *Apriori algorithm results of selected score types*

METHOD	ATTRIBUTE RANKING	CHOSEN ONES	NUMBER OF RULES WITH A CONFIDENCE VALUE OF 10% OR MORE	NUMBER OF RULES WITH A CONFIDENCE VALUE OF 50% OR MORE	NUMBER OF RULES WITH A CONFIDENCE VALUE OF 90% OR MORE	NUMBER OF RULES WITH A CONFIDENCE VALUE OF 100% OR MORE
UNFILTERED	10,8,9,7,6,5	8,10	1515	593	78	78
INFO GAIN	8,9,7,10,6,5	8,9	6941	4215	809	786
CLASSIFIER	10,9,6,7,8,5	10,9	3365	1630	258	239
GAIN RATIO	8,7,9,6,5	8,7	779	263	18	18
ONE R	8,9,5,7,6,10	8,9	6941	4215	809	786
RELIEF	9,8,10,7,5,6	9,8	6941	4215	809	786
SYMMETRICAL UNCERTAINTY	8,9,7,10,6,5	8,9	6941	4215	809	786

Table 17. *Association rules with 100% confidence between books.*

NO	ANTECEDENT	CONSEQUENT	CONFIDENCE	LIFT
1	ONE FOR THE MONEY, HARD EIGHT: A STEPHANIE PLUM NOVEL	TWO FOR THE DOUGH	1	170
2	HEART OF THE SEA, JEWELS OF THE SUN	TEARS OF THE MOON	1	330
3	I IS FOR INNOCENT, G IS FOR GUMSHOE	E IS FOR EVIDENCE: A KINSEY MILLHONE MYSTERY	1	364
4	HARRY POTTER AND THE SORCERER'S STONE, HARRY POTTER AND THE ORDER OF THE PHOENIX	HARRY POTTER AND THE PRISONER OF AZKABAN	1	270
5	A IS FOR ALIBI, G IS FOR GUMSHOE	D IS FOR DEADBEAT	1	320

Evaluation of the 4th rule indicated that the readers who rated the books "Harry Potter and The Sorcerer's stone" and "harry potter and the order of the phoenix" also preferred the book titled "Harry Potter and The Prisoner of Azkaban". Assessment of the rules in terms of lift value revealed that there was none equal to 1. The rules with a lift value of one consist of independent units. On the other hand, there is a positive relationship with rule values greater than 1. The dataset also included book series, such as "the harry potter and lord of the rings". Some of these series were proposed as association rules. These rules were not the postulation of an assumption that the readers of the first book in the series would also read the second book. Contrarily if the users who enjoyed the first book in the series also enjoyed the second book, this data led to rule generation. The Apriori algorithm also stopped generating rules for a series if all users scored the second book in the series less than 5. If the algorithm includes the first book of the series in the analysis, it seems to omit the other books in the series while evaluating the association criteria. it is not appropriate to endorse the series in light of these findings. It is possible to state that the suggestion systems may generate a rule presuming that a user who has read the first book in the series will also read the second; however, evidence is necessary to support such a rule. as a result, based on user rating data, our study revealed the significance of the accuracy while generating such a rule. In essence, it is not an assumption that the user reading the first book of "the lord of the rings" series will

also read the second book, but a data grounding on the rating scores in the dataset. Therefore, according to the confidence value of this rule, these data serve as proof for the book series. An advantage of the study's design was its rating-based structure. Figure 2 displays the confidence rates of the association rules in three colors: green 70% or higher, blue 69% to 65%, and red less than 65%.

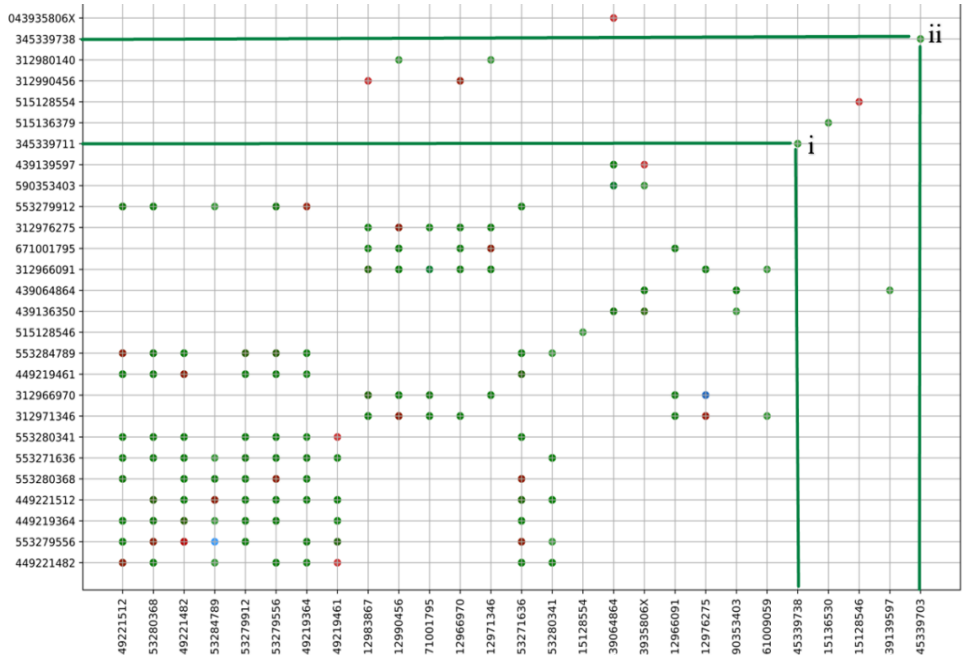


Figure 3. *Association rules with a confidence value of 60% and above.*

Figure 3 visualizes some book-related association rules. These rules retain a confidence level of 60% or more. The first green section (i) in Figure 3 proves that readers who liked the first book titled “The fellowship of the ring” (0345339703), the first book of the lord of the rings series, also liked the third book of this series, titled “the return of the king” (0345339738). The other green-lined section (ii) in Figure 3 displays that the readers who preferred the third book of “the return of the king” (0345339738) also enjoyed the second book titled “the two towers” (0345339711) in the “lord of the rings” series. It is clear that all users enjoyed reading the entire series since the first book. Therefore, it is rational to utilize it in suggestion systems since all these books in the series are included in the rules for the popular books, indicating that it is a popular series. The rules that emerged from these books in this series also proved the accuracy of recommending this series. This information may serve as the accuracy test while suggesting these series. Among 511 books in the dataset analysed, 20 books, exclusively belonging to three book series, were eliminated from the dataset, and the ratings of the 491 books were rerun on the Apriori algorithm. Here, the purpose was to demonstrate the rules generated among other books instead of some prominent book series. For a clearer presentation of the rules for the books, supplementary information is set out in Table 18, and the rules for these books are set out in Table 19 [43].

Table 18. *Information about the books.*

BOOK NAME	SUBJECT
THE STREET LAWYER	LAWYERS, LEGAL FICTION
THE TESTAMENT	LAWYERS, LEGAL FICTION, PROBATE
THE CHAMBER	LAWYERS, LEGAL FICTION, LAW FIRM
POINT OF ORIGIN	THRILLER, MYSTERY
CRUEL AND UNUSUAL	THRILLER, MYSTERY
UNNATURAL EXPOSURE	THRILLER, MURDER

Table 19. *Association rules with a confidence value of 50% or more between books.*

NO	ANTECEDENT	CONSEQUENT	CONFIDENCE (OUT OF 100%)
1	THE CHAMBER, THE STREET LAWYER	THE TESTAMENT	80
2	POINT OF ORIGIN, UNNATURAL EXPOSURE	CRUEL AND UNUSUAL	80
3	THE CHAMBER, THE CLIENT	THE PELICAN BRIEF	50
4	VIOLETS ARE BLUE, 1ST TO DIE: A NOVEL	ROSES ARE RED	57
5	POSTMORTEM, UNNATURAL EXPOSURE	FROM POTTERS FIELD	57
6	THE TESTAMENT, THE CHAMBER	THE RAIN MAKER	80

Association Rules Results for Chosen Books: The Apriori algorithm generated 786 rules with the highest confidence level. These rules are anticipated to allow us to predict inter-book relations of the readers (users) who chose a particular book. Table 17 describes some of these 786-generated rules. according to the first criterion, individuals who give ratings for the book “one for the money and hard eight” will likely prefer “two for the dough”. The confidence level of this rule is 100%. As revealed in Table 17, the rule in the first row depicts that users choosing “The chamber and the street lawyer” books might also prefer “The testament”. The confidence level of this rule is 80%, indicating that this rule holds for 80 people out of 100 who chose “The chamber and the street lawyer” books and selected the testament” book, too. According to data provided in Table 18, all three books fall under the category of legal fiction novels. The book street lawyer is about a lawyer, the chamber is about a law firm, and the testament is about three persons experiencing unanticipated incidents due to a testimonial. The rule in the second line of Table 19 states that users who liked the books “point of origin” and “unnatural exposure” also enjoyed two other books titled “cruel and unusual”. The confidence level of the rule is 80%. The subject of all the books falling under this rule is suspense.

Conclusion

It is crucial to collect and appropriately interpret the opinions regarding the books. Using the rules of the association, it is possible to study the relationships between the authors or books that individuals enjoy. The results of the research can be used to generate person-based recommendation lists. Using people's book ratings, the Apriori algorithm was utilized in this study to construct rules based on the books and authors of the books. The relationships between the authors of the books were identified in

the dataset. In order to create a rule for voting on sparse books, the Apriori algorithm was applied to the votes belonging to the most intense score type, utilizing the methods of grouping according to the density of the books, elimination based on the statistical data of this grouping, and feature selection. In addition, as the dataset consisted of voting data and not shopping data, rules were established by considering both the purchase status and the popularity of a book series. In addition, popular book series were identified in the study. If there are users who have read the first book in a series and disliked it, subsequent books in the series should not be recommended. However, such a rule can only be generated when shopping datasets are used. If there are rules with a confidence value of 70% or above for an entire book series, the model will propose that series. As a result, rules were established to predict which additional books or authors an individual might prefer. It is anticipated that the rules derived from this research will prove useful for book recommendation systems and sales strategies.

Acknowledgements-

Funding/Financial Disclosure The authors have no received any financial support for the research, authorship, or publication of this study.

Ethics Committee Approval and Permissions The work does not require ethics committee approval and any private permission.

Conflict of Interests The authors stated that there are no conflict of interest in this article.

Authors Contribution Authors contributed equally to the study. All authors read and approved the final manuscript.

References

[1] Alaeddinoğlu, M., Aydın, T., & Dal, D. (2014). Birliktelik kurallari ile mekânsal-zamansal veri madenciliği. *Erzincan University Journal of Science and Technology, 5*(2), 191-212.

[2] Xiao, W., Jing, L., Xu, Y., Zheng, S., Gan, Y., & Wen, C. (2021). Different data mining approaches based medical text data. *Journal of Healthcare Engineering, 2021*(1), 1285167. https://doi.org/10.1155/2021/1285167

[3] Pinheiro, L. I. C. C., Pereira, M. L. D., Fernandez, M. P., Filho, F. M. V., de Abreu, W. J. C. P., & Pinheiro, P. G. C. D. (2021). Application of data mining algorithms for dementia in people with HIV/AIDS. *Computational and Mathematical Methods in Medicine, 2021*(1), 4602465. https://doi.org/10.1155/2021/4602465

[4] Hong, J., Tamakloe, R., & Park, D. (2020). Discovering insightful rules among truck crash characteristics using apriori algorithm. *Journal of advanced transportation, 2020*(1), 4323816. https://doi.org/10.1155/2020/4323816

[5] Alan, M. A., & Yeşilyurt, C. (2019). Birliktelik kuralları madenciliği ile yatan hasta profilinin çıkarılması. *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 23*(4), 1917-1926.

[6] Agrawal, R., & Srikant, R. (1994, September 12-15). *Fast algorithms for mining association rules* [Conference presentation]. Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), Santiago, Chile. https://www.vldb.org/conf/1994/P487.PDF

[7] Wu, J., Guo, S., Huang, H., Liu, W., & Xiang, Y. (2018). Information and communications technologies for sustainable development goals: state-of-the-art, needs and perspectives. *IEEE Communications Surveys & Tutorials, 20*(3), 2389-2406. https://doi.org/10.1109/COMST.2018.2812301

[8] Özçakır, F. C., & Çamurcu, A. Y. (2007). Birliktelik kuralı yöntemi için bir veri madenciliği yazılımı tasarımı ve uygulaması. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 6*(12), 21-37.

[9] Erdem, S., & Özdağoğlu, G. (2008). Ege Bölgesi'ndeki bir araştırma ve uygulama hastanesinin acil hasta verilerinin veri madenciliği ile analiz edilmesi. *Anadolu University Journal of Sciences & Technology, 9*(2).

[10] Ay, D., & Çil, İ. (2010). Migros Türk AŞ. de birliktelik kurallarının yerleşim düzeni planlamada kullanılması. *Endüstri Mühendisliği, 21*(2), 14-29.

[11] Hong, J., Tamakloe, R., & Park, D. (2020). Discovering insightful rules among truck crash characteristics using Apriori algorithm. *Journal of Advanced Transportation, 2020*(1), 4323816. https://doi.org/10.1155/2020/4323816

[12] Erpolat, S. (2012). Otomobil yetkili servislerinde birliktelik kurallarının belirlenmesinde Apriori ve Fp-Growth algoritmalarının karşılaştırılması. *Anadolu University Journal of Social Sciences, 12*(2), 137–146.

[13] Bilgin, T., & Acun, G. (2015). Yazılım hata logları kullanılarak veri madenciliği uygulaması gerçekleştirilmesi. *Marmara Journal of Pure and Applied Sciences, 27*(1), 14–20. https://doi.org/10.7240/mufbed.83214

[14] Doğrul, G., Akay, D., & Kur, M. (2015). Trafik kazalarının birliktelik kuralları ile analizi. *Gazi Journal of Engineering Sciences, 1*(2), 265–283.

[15] Çelik, B., & Aytekin, P. (2019, 1-4 Mayıs). *Perakende sektöründe pazar sepeti analizi için Apriori algoritması ile birliktelik kurallarının oluşturulması* [Conference presentation]. PPAD Pazarlama Kongresi - MMRA Marketing Congress Bildiri Kitabı, Kuşadası, Türkiye. https://www.researchgate.net/publication/344538987_Perakende_Sektorunde_Pazar_Sepeti_Analizi_icin_Apriori_Algoritmasi_ile_Birliktelik_Kurallarinin_Olusturulmasi

[16] Li, Z., Li, X., Tang, R., & Zhang, L. (2021). Apriori algorithm for the data mining of global cyberspace security issues for human participatory based on association rules. *Frontiers in Psychology, 11,* 582480. https://doi.org/10.3389/fpsyg.2020.582480

[17] Wang, H. B., & Gao, Y. J. (2021). Research on parallelization of Apriori algorithm in association rule mining. *Procedia Computer Science, 183,* 641–647. https://doi.org/10.1016/j.procs.2021.02.109

[18] Hussain, S., & Hazarika, G. C. (2014). Educational data mining model using Rattle. *International Journal of Advanced Computer Science and Applications, 5*(6). https://doi.org/10.14569/IJACSA.2014.050605

[19] Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In Proceedings of the ACM SIGMOD international conference on Management of data, 207-216.

[20] Hussain, S. (2017). Survey on current trends and techniques of data mining research. *London Journal of Research in Computer Science and Technology, 17*(1), 11.

[21] Xu, R., & Luo, F. (2021). Risk prediction and early warning for air traffic controllers' unsafe acts using association rule mining and random forest. *Safety Science, 135,* 105125. https://doi.org/10.1016/j.ssci.2020.105125

[22] Birant, D., Kut, A., Ventura, M., Altınok, H., Altınok, B., Altınok, E., & Ihlamur, M. (2010). İş zekası çözümleri için çok boyutlu birliktelik kuralları analizi. *Akademik Bilişim, 10,* 256.

[23] Akpınar, H. (2000). Veri tabanlarında bilgi keşfi ve veri madenciliği. *İstanbul Üniversitesi İşletme Fakültesi Dergisi, 29*(1), 1-22.

[24] Liu, Y., Hu, X., Luo, X., Zhou, Y., Wang, D., & Farah, S. (2020). Identifying the most significant input parameters for predicting district heating load using an association rule algorithm. *Journal of Cleaner Production, 275,* 122984. https://doi.org/10.1016/j.jclepro.2020.122984

[25] Altay, E. V., & Alataş, B. (2021). Differential evolution and sine cosine algorithm based novel hybrid multi-objective approaches for numerical association rule mining. *Information Sciences, 554,* 198–221. https://doi.org/10.1016/j.ins.2020.12.055

[26] Dong, X., Hao, F., Zhao, L., & Xu, T. (2020). An efficient method for pruning redundant negative and positive association rules. *Neurocomputing, 393,* 245–258. https://doi.org/10.1016/j.neucom.2018.09.108

[27] Fernandez-Basso, C., Ruiz, M. D., & Martin-Bautista, M. J. (2021). Spark solutions for discovering fuzzy association rules in Big Data. *International Journal of Approximate Reasoning, 137,* 94–112. https://doi.org/10.1016/j.ijar.2021.07.004

[28] Ayberkin, D., & Özen, Ü. (2019). Apriori algoritmasının kullanılmasına yönelik bir yazılım tasarımı ve uygulaması: İŞKUR verilerinin değerlendirilmesi üzerine bir örnek çalışma. *Journal of Business in the Digital Age, 2*(2), 95–102.

[29] Han, J., Kamber, M., & Pei, J. (2000). *Data mining: Concepts and techniques.* Morgan Kaufmann.

[30] Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research, 3,* 1289–1305.

[31] Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis, 1*(1-4), 131-156. https://doi.org/10.3233/IDA-1997-1302

[32] High, R., Eyres, G. T., Bremer, P., & Kebede, B. (2021). Characterization of blue cheese volatiles using fingerprinting, self-organizing maps, and entropy-based feature selection. *Food Chemistry, 347,* 128955. https://doi.org/10.1016/j.foodchem.2020.129011

[33] Omuya, E. O., Okeyo, G. O., & Kimwele, M. W. (2021). Feature selection for classification using principal component analysis and information gain. *Expert Systems with Applications, 174,* 114765. https://doi.org/10.1016/j.eswa.2021.114765

[34] Hakim, M. F., & Saputra, S. (2025). Analisis prediksi kelulusan mahasiswa universitas dinamika bangsa menggunakan metode naïve bayes. *Jurnal Informatika Dan Rekayasa Komputer (JAKAKOM), 5*(1), 1264-1273. https://doi.org/10.33998/jakakom.2025.5.1.1999

[35] Novakovic, J., Strbac, P., & Bulatovic, D. (2011). Toward optimal feature selection using ranking methods and classification algorithms. *Yugoslav Journal of Operations Research, 21*(1), 119–135. https://doi.org/10.2298/YJOR1101119N

[36] Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning, 11*(1), 63–90. https://doi.org/10.1023/A:1022631118932

[37] Kaynar, O., Arslan, H., Görmez, Y., & Işık, Y. E. (2018). Makine öğrenmesi ve öznitelik seçim yöntemleriyle saldırı tespiti. *Bilişim Teknolojileri Dergisi, 11*(2), 175–185. https://doi.org/10.17671/gazibtd.368583

[38] Kira, K., & Rendell, L.A. (1992). The Feature Selection Problem: Traditional Methods and a New Algorithm. In: Proceedings of the 10th National Conference on Artificial Intelligence, AAAI Press, Atlanta, 129-134.

[39] Kononenko, I. (1994, 2 April). *Estimating attributes: Analysis and extensions of RELIEF.* [Conference presentation]. European conference on machine learning. Berlin, Heidelberg: Springer https://doi.org/10.1007/3-540-57868-4_57

[40] Ziegler, C. N., McNee, S. M., Konstan, J. A., & Lausen, G. (2005, 3 May). *Improving recommendation lists through topic diversification.* Proceedings of the 14th International Conference on World Wide Web). ACM. https://doi.org/10.1145/1060745.1060754

[41] Dewson, R. (2014). *SQL Server Management Studio*. In Beginning SQL Server for Developers: Fourth Edition (pp. 25-42). Berkeley, CA: Apress.

[42] Liu, H., & Wang, B. (2007). An association rule mining algorithm based on a Boolean matrix. *Data Science Journal, 6,* 559-565. https://doi.org/10.2481/dsj.6.S559

[43] Bartley, P. (2009). Book tagging on Library Thing: how, why, and what are in the tags? *Proceedings of the American Society for Information Science and Technology, 46*(1), 1-22. https://doi.org/10.1002/meet.2009.1450460228