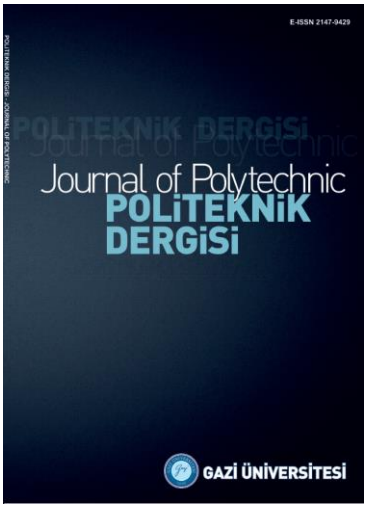


POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

Unlocking the multidisciplinary potential of data science: ınsights from apriori analysis

Veri biliminin çok disiplinli potansiyelinin kilidini açmak: apriori analizinden içgörüler

***Yazar(lar) (Author(s))**: Merve Nur BARUN[1], Emrah ÖNDER[2]*

ORCID[1]: 0000-0002-2545-9534

ORCID[2]: 0000-0002-0554-1290

To cite to this article: Barun M. N., Önder E., "Unlocking the multidisciplinary potential of data science: ınsights from apriori analysis", ***Journal of Polytechnic***, *(*): *, (*).

Bu makaleye şu şekilde atıfta bulunabilirsiniz: Barun M. N., Önder E., "Unlocking the multidisciplinary potential of data science: ınsights from apriori analysis", ***Politeknik Dergisi***, *(*): *, (*).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: 10.2339/politeknik.1432158

Unlocking the Multidisciplinary Potential of Data Science: Insights from Apriori Analysis

Highlights

- *Data science holds paramount significance for the progress of technology and science.*
- *Data science is pivotal in decision and policymaking, developing learning methods for educators.*
- *Data science is significant in fields such as cosmology and ecology.*
- *Data science is crucial in breast cancer treatment, and genetic science in the health domain.*

Graphical Abstract

This study identifies and analyzes other areas where researchers are working in the field of data science and provides guidance for future research work. Apriori analysis was applied to two different datasets using the R Studio program.

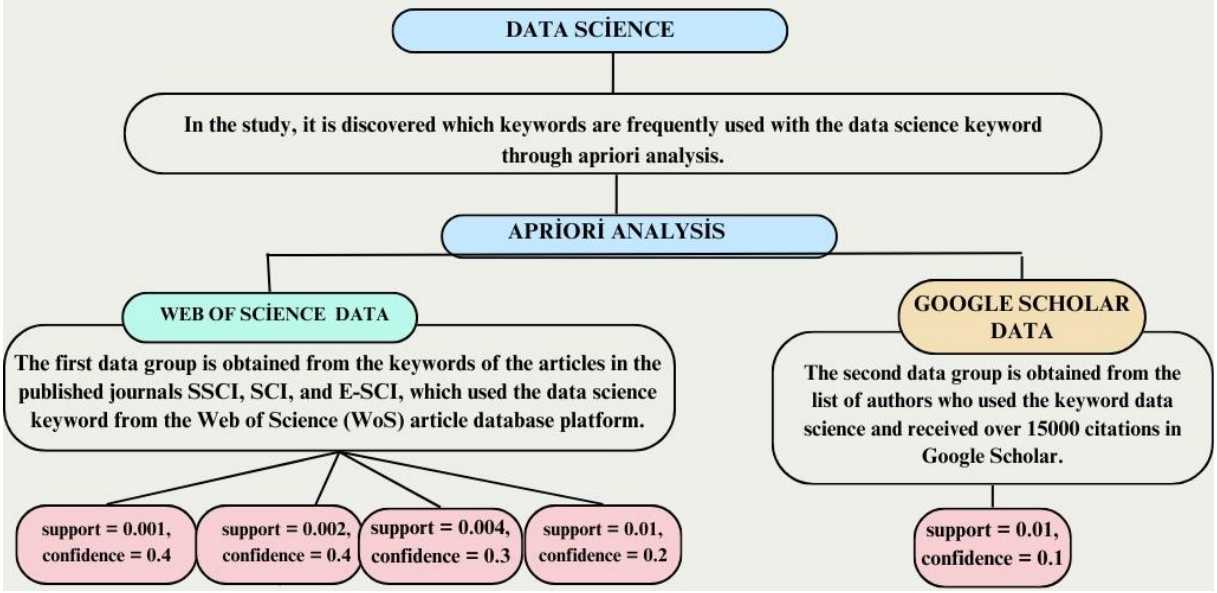


Figure a.

Aim

This study aims to identify, analyse other fields where researchers work in data science, and provide guidance for future research endeavours.

Design & Methodology

Apriori analysis is applied to two different data groups using the R Studio program.

Originality

Data science holds paramount significance for the progress of technology and science. It is important to discern the existing studies in data science and identify areas where research is deficient.

Findings

Data science is used in decision and policymaking, developing learning methods for educators, breast cancer treatment, and genetic science in the health domain.

Conclusion

The support of data scientists to people working in other fields will contribute significantly to the development of science because of the interdisciplinary nature of data science.

Declaration of Ethical Standards

This article do not require ethical committee permission and/or legal-special permission.

Unlocking the Multidisciplinary Potential of Data Science: Insights from Apriori Analysis

Araştırma Makalesi / Research Article

Merve Nur BARUN[1*], Emrah ÖNDER[2]

[1]Institute of Social Sciences, Faculty of Business Administration, Numerıcal Methods, Istanbul University, Istanbul, Türkiye
[2]Faculty of Business Administration, Numerıcal Methods, Istanbul University, Istanbul, Türkiye

(Received : 05.02.2024 ; Accepted : 09.09.2024 ; Early View : 12.09.2024)

ABSTRACT

Data science is of great importance for the advancement of technology and science. Therefore, it is imperative to distinguish existing work in the field of data science and identify areas where research is lacking. Therefore, this study aims to identify and analyze other areas where researchers are working in the field of data science and provide guidance for future research efforts. This paper describes the application of apriori analysis to two different datasets using R Studio. The first dataset consists of 2262 articles in SSCI, SCI and E-SCI indexed journals obtained from Web of Science database using the keyword "data science". The second data set was obtained from the list of more than 15,000 cited authors (316 authors) specializing in data science in Google Scholar. The study covers a total of 2262 articles and 316 authors. The articles cover 6533 unique keywords. The use of apriori analysis, a data mining method, on the obtained datasets involves the use of support, confidence and removal values to determine association rule outputs. The results of apriori analysis show that data science is very important in decision making and policy making, developing learning methods for educators, breast cancer treatment, and genetic science in the health field. Moreover, data science is also important in various fields such as cosmology and ecology. This result reaffirms the interdisciplinary nature of data science.

Keywords: Data Science, Web of Science, Association Rules Analysis, Apriori Algorithm, Data Mining.

Veri biliminin çok disiplinli potansiyelinin kilidini açmak: apriori analizinden içgörüler

ÖZ

Teknolojinin ve bilimin ilerlemesinde çok büyük öneme sahip olan veri bilimi alanında hangi çalışmaların gerçekleştirildiği ve hangi alanlarda çalışmaların eksik kaldığını belirlemek büyük önem taşımaktadır. Bu çalışma veri bilimi alanında çalışan araştırmacıların başka hangi alanlarda çalıştığını belirlemek ve analiz etmek için yapılmıştır. Bu kapsamda R ile iki farklı veri grubuna apriori analizi uygulanmıştır. İlk veri grubu Web of Science veri tabanınından data science anahtar kelimesini kullanmış olan SSCI, SCI, E-SCI yayınlanmış dergilerdeki makaleler elde edilmiştir. İkinci veri grubu Google Scholor'da Veri bilimi anahtar kelimesini kullanmış en çok atıf alan yazarların (316 yazar) listesinden seçilmiştir. Çalışmada toplam 2262 makale kullanılmıştır. Makalelerde 6533 tekil anahtar kelime olduğu gözlemlenmiştir. Elde edilen veri gruplarına R Studio programında veri madenciliği yöntemi olan Apriori analizi uygulanmıştır. Birliktelik kuralı çıktılarını belirlemek için destek, güven ve kaldıraç (ilginçlik) değerleri kullanılmıştır. Apriori analiz sonucuna göre veri bilimi ile kaldıraç değeri en yüksek konular karar ve politika belirlemesi, eğitimcilerin öğrenme yöntemlerini geliştirmesi, sağlık alanında meme kanseri tedavisi ve genetik bilimidir. Veri bilimi, evren bilimi (kozmoloji) ve ekoloji gibi daha birçok alanda önemli bir yere sahiptir. Bu durum veri biliminin multidisipliner bir alan olduğunu bir kez daha ortaya konmuştur.

Anahtar Kelimeler: Veri Bilimi, Web of Science, Birliktelik Kuralları Analizi, Apriori Algoritması, Veri Madenciliği.

1. INTRODUCTION

Although data science is a new concept, it has a long history. Basic data science methods have been used in statistics for many years. The development of statistical methods dates back to the 18th century [32]. While statistics has long worked with smaller data, data science works with more comprehensive data. Thanks to data science technology, large data sets can be processed and analyzed quickly [12]. The concept of data science has become popular in recent years with the development of computer technology and big data.

Statistics is a field of science that involves collecting, analyzing, interpreting, and inferring data [37]. Data science is a field of science used to analyze large data sets, perform exploratory data analysis, and make predictions using machine learning and artificial intelligence techniques [59].

Data science also includes some methods used in statistics. Basic statistical methods such as t-tests, ANOVA, regression analysis, clustering, and classification are also used in data science projects [7,8]. Therefore, data science and statistics are complementary disciplines.

Data science is a multidisciplinary field that combines statistics, mathematics, computer science, and domain knowledge to collect, process, analyze, and interpret data [42].

Data science is of great importance in almost every aspect of our lives. Thanks to technologies such as the internet and mobile devices, huge amounts of data are

* *Corresponding Author*
e-posta : barun.mervenur@gmail.com

constantly being generated. Big data processing with data science techniques and technologies provides information about people's online behavior, buying habits, social media usage, health status, travel plans and many more [19]. Data science is used in almost all fields including education, economy, health, agriculture, industry, transportation, security, energy, engineering, business, manufacturing, food and retail [3,4,11,22]. Data science technologies can optimize traffic flow, prevent security problems, and improve energy efficiency [57]. Crime data helps law enforcement agencies develop strategies to reduce crime rates [57]. Health data increases the accuracy of medical diagnoses and helps improve treatments [38]. Weather data can help organize travel plans [15].

In parallel with the developing technologies (internet of things, wearable technologies, cloud computing, artificial intelligence, blockchain, machine learning, big data, mobile applications, e-commerce, etc.), the importance of data science is increasing day by day. Companies utilize more data in their operational, managerial and strategic decision-making processes. For this reason, it is important to examine the academic fields of study related to data science. Thus, it will be possible to determine in which areas data science is academically studied and in which areas more limited studies are conducted. For this reason, in our study, we wanted to discover which keywords are frequently used with the keyword data science with apriori analysis. Association analysis is a basic technique frequently used in data mining. Apriori algorithms can create frequent itemsets by reducing (pruning) unnecessary sub-items [45].

The Apriori algorithm enables fast and efficient frequency discovery in large data sets. For this reason, apriori analysis was applied to two datasets using R programing language. The first dataset was obtained from the keywords of articles published in SSCI, SCI and E-SCI journals using the keyword data science from the Web of Science (WoS) article database platform. A total of 2262 articles were used in the first dataset. It was observed that there were 6533 unique keywords in the articles. The second dataset was obtained from the list of authors who used the keyword data science and received over 15000 citations in Google Scholar. The number of observations in the second data set is 316. As a result of the analysis, it was concluded that researchers working in the field of data science are also working in the fields of big data, artificial intelligence, visualization, information systems, machine learning, statistics, prediction, education, data mining and IoT (Internet of Things).

This research examines the historical development and current importance of data science, with a particular focus on the apriori algorithm, a fundamental technique in data mining. The paper explores the intersection of data science and statistics, emphasizing how these disciplines complement each other, especially in the context of large data sets.

The aim of this paper is twofold: first, to illuminate the historical development of data mining techniques and the evolution of the apriori algorithm; second, to apply this algorithm to large datasets to reveal meaningful relationships and trends in the field of data science.

In a world where large amounts of data are constantly being generated, understanding how data science techniques like a priori algorithms can process and analyze this information is crucial to optimizing decision-making processes at all levels of society. The study's findings provide insights into current research trends in data science and pave the way for future advances in the field.

2. LITERATURE

Association analysis is an essential technique in data mining. Many researchers have contributed to its development, which dates back to 1904 [47]. In 1904, British statistician and psychologist Charles Spearman used an approach similar to correlation analysis to measure the relationship between two variables using a technique known as the "Spearman correlation" [48].

Data mining and data analysis techniques form the basis of technologies such as artificial intelligence and machine learning, frequently used today [27]. These techniques have accelerated with the development of computers and data storage technologies [34]. The origins of data mining and data analysis techniques date back to the 1960s when computers were used. During this period, database management techniques were developed, and it became possible to store and process large amounts of data [26]. In the 1980s, developments in database management techniques advanced, and modern database systems, such as relational databases, began to be used [49]. After this period, data mining techniques were also developed, making exploring data and finding patterns easier [24]. In the 1990s, the development of data mining techniques accelerated, and techniques such as relationship analysis emerged. During this period, relationship analysis was used primarily in the retail sector. This technique analyzes customers' shopping behavior, and unique campaigns and discounts could be offered. In the 2000s, data mining techniques and data analysis methods were further developed, and as a result, technologies such as extensive data analysis and machine learning emerged. These technologies allow more significant data sets to be processed, and more complex patterns and relationships could be discovered.

The origins of the Apriori algorithm date back to the 1940s by relating to the concept of customer basket analysis. However, since it is not known who first used the concept of customer basket analysis, it is mentioned with Agrawal et al., who found the apriori algorithms [23]. Modern association analysis techniques were developed in the 1990s. IBM researchers such as Rakesh Agrawal, Tomasz Imielinski and Arun Swami have played a significant role in developing these techniques [51]. Association analysis is first developed in 1993 by Rakesh Agrawal, Tomasz Imielinski, and Arun Swami.

This technique is introduced in the article "Mining Association Rules Between Sets of Items in Large Databases" by Agrawal, Imielinski, and Swami [1]. The apriori algorithm, which emerged by Agrawal et al., is used to process data more quickly with data filtering methods. In the following years, algorithms such as FP-Growth algorithm, Eclat algorithm and SON (Savasere, Omiecinski, and Navathe) algorithm, H-Mine algorithm and SWIM algorithm were also developed for frequency discovery. SON algorithm is used to determine and investigate the relationships between elements in a data set [41]. Eclat algorithm is used to determine the connection between clusters [5, 54] FP-Growth algorithm is an algorithm used to rank frequently occurring elements [36]. H-Mine algorithm, which emerged in 2001, was developed for frequency discovery to be used in high-dimensional data [56]. SWIM algorithm emerged in 2008. This algorithm is used for frequency discovery in large data sets [6].

Thanks to developing technologies and large data sets, relationship analysis has become even more important in recent years. It is widely used, especially in the retail sector, to analyze customer behavior and determine customer preferences [45]. Relationship analysis is a method used to determine how often a product is purchased with another product by examining the purchasing behavior of customers for related products. In this way, retail companies learn what kind of products their customers are interested in and which products they prefer together. It allows them to offer special campaigns and services to customers by better understanding their customers' behaviors and preferences [18]. It is used in making critical business decisions such as sales forecasts, product recommendations, marketing strategies and customer segmentation [20]. Relationship analysis gained popularity in the 1990s and early 2000s, especially with its use in the retail sector. Companies working in this field could use relationship analysis to understand their customers' shopping behaviors and offer special discounts and campaigns [14]. In addition, Apriori algorithm and relationship analysis techniques are used in many fields such as marketing, finance, banking, food, health, education, engineering and agriculture.

Using association analysis, it is possible to determine which diseases coexist. This information will facilitate the diagnosis and treatment of diseases [46]. If the patient has symptoms related to a particular disease, which treatments may be more effective could be determined by analytical methods [60]. In addition, relationship analysis could be used to collect information about eating habits and food preferences. Using this analysis, it could be determined which foods are consumed together. In this way, recommendations could be created for a healthy diet [39]. With the help of relationship analysis, information could be collected about sports and exercise habits and which sports activities are done together could be determined. This information could also be used for personal training programs and gym marketing strategies [44]. Banks, just like retail companies, could analyze customer behavior and preferences using relationship analysis. By checking which financial products customers prefer together, they could provide better customer service and optimize their marketing strategies [9]. In addition, the banking and finance sector uses a priori algorithm to analyze customer purchasing behavior, manage credit risk, and provide customer-specific credit offers [33]. Association analysis is also used for weather forecasting. This analysis develops weather forecast models by determining which weather conditions occur together [30]. In addition, studies have been conducted in the education sector that enable students to make physically appropriate choices [29]. Such analyses are used to improve the quality of education. The behavior of web users could be analyzed using association analysis [40], systems could be created by collecting information such as which other products a user searches for or looks at while searching for a product through web mining [55]. The telecommunications industry uses the Apriori algorithm to analyse customer usage behaviour as in other marketing areas. These analyses are used to offer special tariffs and campaigns to the customer. These examples show that association analysis could be used in different fields. Association analysis could be used for many different purposes and to identify features occurring in many datasets.

Sertçelik and Önder conducted a similar study to the one in this article by using the apriori algorithm for studies using the keyword “Management Information Systems” [43]. The use of this method, which is frequently used in marketing and market basket analysis, to examine the associations between academic fields of study has not been found in other academic studies in the literature. The lack of such a study in the field of data science has been identified as a deficiency in the literature. Therefore, this study was carried out.

3. MATERIAL AND METHOD

Two different data groups are used in the study. The datasets consist of articles in journals that have used the keyword data science and a list of the most cited authors who have used the keyword data science. The first data set is obtained from the keywords of the articles in the published journals SSCI, SCI, SCI-E and E-SCI, which used the data science keyword from the Web of Science (WoS). At this stage, it is tried to ensure that all successful studies representing the relevant literature were included in the analysis. All studies using the keyword data science are obtained from the first search. In order to meet the success criterion, SSCI, SCI, SCI-E and E-SCI articles are included in the analysis. Other data are obtained from a list of academics who received more than 15000 citations from Google Scholar. The most cited authors in Google Scholar are selected in descending order. The second data set is created by looking at which other keywords the authors who used the Data Science keyword used. The Apriori algorithm is applied to both data groups using the R programming language and the R Studio development environment. In

the data set included in the first analysis, 2262 articles and 6533 unique keywords are used. Keywords of 316 authors are used in the data set included in the second analysis.

3.1. Apriori Analysis

Apriori algorithms are based on two primary criteria: "support" and "confidence". Support refers to how often a variable is used in the data set, while confidence refers to how often a variable is used together with all other variables. Association analysis is usually performed using frequency tables of data. Frequency tables show how often each variable is used. These tables allow the calculation of "support" and "confidence" criteria used in association analysis [25].

I = { i 1 , i 2 , …, i m } I show the dataset.

(1) Support(A,B)= frequency (A,B)/N N: total number of observations

(2)
Confidence(A→B)=P(A/B)=support(A,B)/support(A)

(3)Lift(A→B)=P(B/A)/P(B)=support(A,B)/support(A)* support(B)

Association Analysis stages [43,58]:

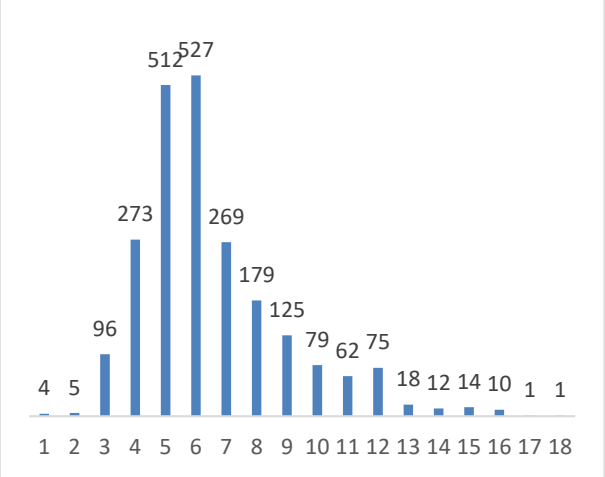
1. There must be a minimum support value.
2. A minimum confidence value should be obtained.
3. Sub-process clusters with a support value greater than the minimum support value should be identified.
4. The rules for the sub-process set with a confidence value greater than the minimum confidence value should be determined.
5. Rules should be ordered from largest to smallest according to the lift value.işımının direkt olarak gördüğü yutucu plaka yüzey alanıdır (m^2).

The confidence coefficient varies between 0 and 1. A high confidence value indicates that the value rule is correct and reliable. The lift value is between 0 and ∞ (infinity). If the lift value is equal to 1, the items have no relationship [17]. A lift value greater than 1 indicates the strong relationship [28].

From this point of view, apriori analysis is performed by changing the support and confidence values in the study. The analysis results are presented in the findings section and discussed in the conclusion and evaluation section.

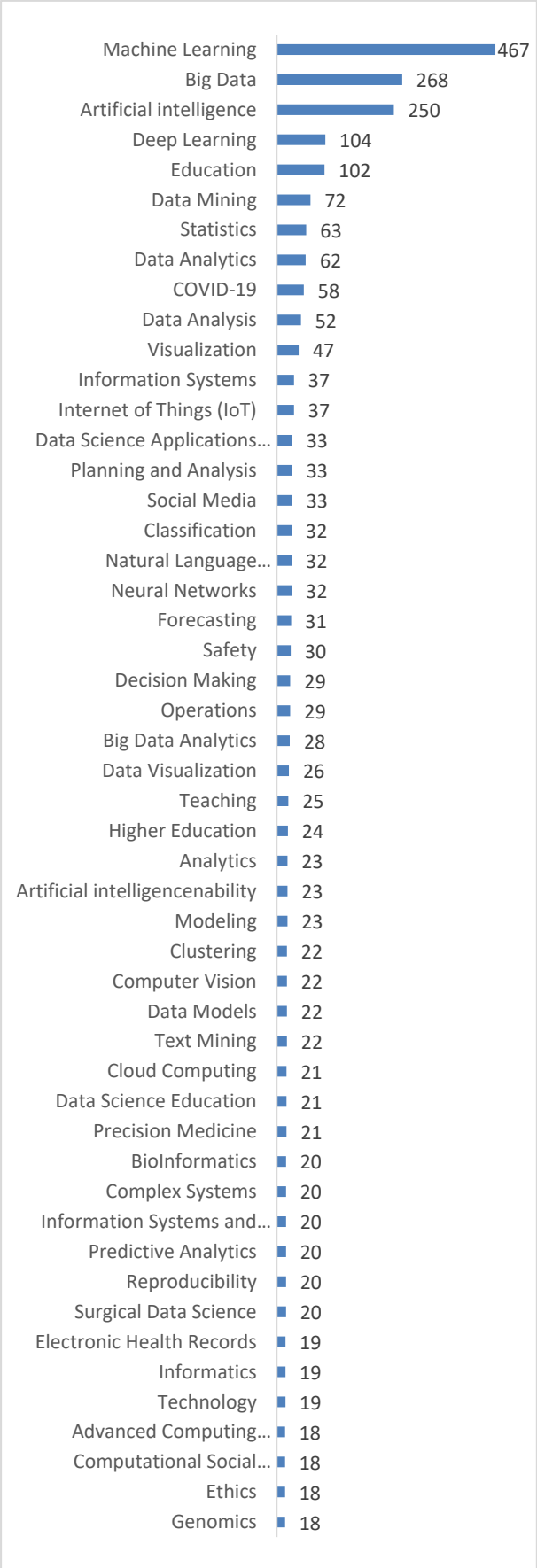
4. RESULTS AND DISCUSSION

Initially, the Apriori Algorithm is implemented using R Studio on the dataset obtained from the Web of Science (WoS) platform. To analyse articles in the field of data science, we started our study by examining the number of keywords found in each article. Graph 1 illustrates the distribution of articles based on the number of keywords. Each article contains a minimum of 1 keyword and a maximum of 18 keywords. Notably, 527 articles exhibit a maximum usage of 6 keywords.



Graph 1. Number of Articles by Keyword Quantity

Subsequently, we identified the most frequently used keywords in data science for all the articles we obtained in data science. As depicted in Graph 2, a total of 50 distinct academic research topics, each with a frequency exceeding 18, are identified. The graph highlights that the predominant focus in data science research lies in the domain of machine learning. Additionally, it has been observed that many studies have been carried out in the fields of big data and artificial intelligence. It has been observed that a relatively high number of studies have been carried out in the fields of deep learning and education. Although deep learning is an expected field of study within the scope of data science, the fact that it has been studied as frequently in education has revealed that data science is a crucial concept in education. Moreover, the analysis reveals that data science has been applied across various domains, from bioinformatics to health, and from social media to electronics.



Graph 2. Academic Research Topics

The apriori analysis results applied to the first data group are applied by changing the support and confidence values. Tables 1, 2, 3 and 4 are given according to the changing values. In Table 1, support=0.001, confidence=0.4 is applied. The lift value as high as 754 in Table 1 shows that it is very interesting to work with (Harun et al., 2017).

Looking at Table 1, it is concluded that those working in the field of data science, working in the fields of policy and organisation, see big data as an essential tool in decision-making. According to rule 39 of the output report, there is an association between breast cancer, gene expression, and mathematical oncology. Rule 40 indicates a relationship between data science, gene expression, and mathematical oncology, suggesting that data science is vital in cancer diagnosis and treatment. Rule 42 presents an exciting result, showing that researchers involved in data science competitions, national ecological observatory networks, and remote sensing also work in species classification.

All numerical values in Table 1 are identical, implying consistent quality across all rules. It is noteworthy that these rules are present only three times among the 2262 articles, yielding a support value of 0.00133 (3/2262).

Furthermore, the co-occurrence of "graph measures" and "symmetry" is observed only in three out of 2262 articles. The scarcity of rules appearing in numerous articles strengthens the study. A higher count will enhance the robustness of the study.

Table 1. Apriori Algorithm

rules = apriori(data = dataset, parameter = list(support = 0.001, confidence = 0.4))

rules = apriori(data = dataset, parameter = list(support = 0.001, confidence = 0.4))								
	lhs		rhs	support	confidence	coverage	lift	count
[1]	{Graph Measures}	=>	{Symmetry}	0,00133	1,00000	0,00133	754	3
[2]	{Symmetry}	=>	{Graph Measures}	0,00133	1,00000	0,00133	754	3
[3]	{Data For Decision Making}	=>	{Policy and Organization}	0,00133	1,00000	0,00133	754	3
[4]	{Policy and Organization}	=>	{Data For Decision Making}	0,00133	1,00000	0,00133	754	3
[5]	{National Ecological Observatory Network}	=>	{Species Classification}	0,00133	1,00000	0,00133	754	3
[6]	{Species Classification}	=>	{National Ecological Observatory Network}	0,00133	1,00000	0,00133	754	3
[7]	{Graph Measures, Quantitative Graph Theory}	=>	{Symmetry}	0,00133	1,00000	0,00133	754	3
[8]	{Quantitative Graph Theory, Symmetry}	=>	{Graph Measures}	0,00133	1,00000	0,00133	754	3
[9]	{Graph Measures, Graphs}	=>	{Symmetry}	0,00133	1,00000	0,00133	754	3
[10]	{Graphs, Symmetry}	=>	{Graph Measures}	0,00133	1,00000	0,00133	754	3
[11]	{Graph Measures, Networks}	=>	{Symmetry}	0,00133	1,00000	0,00133	754	3
[12]	{Networks, Symmetry}	=>	{Graph Measures}	0,00133	1,00000	0,00133	754	3
[13]	{Data Science, Graph Measures}	=>	{Symmetry}	0,00133	1,00000	0,00133	754	3
[14]	{Data Science, Symmetry}	=>	{Graph Measures}	0,00133	1,00000	0,00133	754	3
[15]	{Artificial intelligencenable Mobility, Public Transport}	=>	{Multimodality}	0,00133	1,00000	0,00133	754	3
[16]	{Gene Expression, Mathematical Oncology}	=>	{Receptor Status}	0,00133	1,00000	0,00133	754	3
[17]	{Breast Cancer, Gene Expression}	=>	{Receptor Status}	0,00133	1,00000	0,00133	754	3
[18]	{Data For Decision Making, Data Science}	=>	{Policy and Organization}	0,00133	1,00000	0,00133	754	3
[19]	{Data Science, Policy and Organization}	=>	{Data For Decision Making}	0,00133	1,00000	0,00133	754	3
[20]	{Data Science Competition, National Ecological Observatory Network}	=>	{Species Classification}	0,00133	1,00000	0,00133	754	3
[21]	{Data Science Competition, Species Classification}	=>	{National Ecological Observatory Network}	0,00133	1,00000	0,00133	754	3
[22]	{National Ecological Observatory Network, Remote Sensing}	=>	{Species Classification}	0,00133	1,00000	0,00133	754	3
[23]	{Remote Sensing, Species Classification}	=>	{National Ecological Observatory Network}	0,00133	1,00000	0,00133	754	3
[24]	{Data Science Competition, Remote Sensing}	=>	{National Ecological Observatory Network}	0,00133	1,00000	0,00133	754	3
[25]	{Data Science Competition, Remote Sensing}	=>	{Species Classification}	0,00133	1,00000	0,00133	754	3
[26]	{Graph Measures, Graphs, Quantitative Graph Theory}	=>	{Symmetry}	0,00133	1,00000	0,00133	754	3
[27]	{Graphs, Quantitative Graph Theory, Symmetry}	=>	{Graph Measures}	0,00133	1,00000	0,00133	754	3
[28]	{Graph Measures, Networks, Quantitative Graph Theory}	=>	{Symmetry}	0,00133	1,00000	0,00133	754	3
[29]	{Networks, Quantitative Graph Theory, Symmetry}	=>	{Graph Measures}	0,00133	1,00000	0,00133	754	3
[30]	{Data Science, Graph Measures, Quantitative Graph Theory}	=>	{Symmetry}	0,00133	1,00000	0,00133	754	3
[31]	{Data Science, Quantitative Graph Theory, Symmetry}	=>	{Graph Measures}	0,00133	1,00000	0,00133	754	3
[32]	{Graph Measures, Graphs, Networks}	=>	{Symmetry}	0,00133	1,00000	0,00133	754	3
[33]	{Graphs, Networks, Symmetry}	=>	{Graph Measures}	0,00133	1,00000	0,00133	754	3
[34]	{Data Science, Graph Measures, Graphs}	=>	{Symmetry}	0,00133	1,00000	0,00133	754	3
[35]	{Data Science, Graphs, Symmetry}	=>	{Graph Measures}	0,00133	1,00000	0,00133	754	3
[36]	{Data Science, Graph Measures, Networks}	=>	{Symmetry}	0,00133	1,00000	0,00133	754	3
[37]	{Data Science, Networks, Symmetry}	=>	{Graph Measures}	0,00133	1,00000	0,00133	754	3
[38]	{Artificial intelligencenable Mobility, Data Science, Public Transport}	=>	{Multimodality}	0,00133	1,00000	0,00133	754	3
[39]	{Breast Cancer, Gene Expression, Mathematical Oncology}	=>	{Receptor Status}	0,00133	1,00000	0,00133	754	3
[40]	{Data Science, Gene Expression, Mathematical Oncology}	=>	{Receptor Status}	0,00133	1,00000	0,00133	754	3
[41]	{Breast Cancer, Data Science, Gene Expression}	=>	{Receptor Status}	0,00133	1,00000	0,00133	754	3
[42]	{Data Science Competition, National Ecological Observatory Network, Remote Sensing}	=>	{Species Classification}	0,00133	1,00000	0,00133	754	3
[43]	{Data Science Competition, Remote Sensing, Species Classification}	=>	{National Ecological Observatory Network}	0,00133	1,00000	0,00133	754	3
[44]	{Graph Measures, Graphs, Networks, Quantitative Graph Theory}	=>	{Symmetry}	0,00133	1,00000	0,00133	754	3
[45]	{Graphs, Networks, Quantitative Graph Theory, Symmetry}	=>	{Graph Measures}	0,00133	1,00000	0,00133	754	3
[46]	{Data Science, Graph Measures, Graphs, Quantitative Graph Theory}	=>	{Symmetry}	0,00133	1,00000	0,00133	754	3
[47]	{Data Science, Graphs, Quantitative Graph Theory, Symmetry}	=>	{Graph Measures}	0,00133	1,00000	0,00133	754	3
[48]	{Data Science, Graph Measures, Networks, Quantitative Graph Theory}	=>	{Symmetry}	0,00133	1,00000	0,00133	754	3
[49]	{Data Science, Networks, Quantitative Graph Theory, Symmetry}	=>	{Graph Measures}	0,00133	1,00000	0,00133	754	3
[50]	{Data Science, Graph Measures, Graphs, Networks}	=>	{Symmetry}	0,00133	1,00000	0,00133	754	3

In the second implementation, support = 0.002, confidence = 0.4. In Table 2, it is seen that researchers working in data science, human factors and pedestrians also work in the field of bicycles with 100% confidence. In addition, as seen in the 9th and 10th rules, data science applications in education are important in determining learning strategies. Rules 23, 32, 34, 42, 46, and 48 reveal that data science plays a significant role in crime detection and reliability.

Table 2. Apriori Algorithm
rules = apriori(data = dataset, parameter = list(support = 0.002, confidence = 0.4))

rules = apriori(data = dataset, parameter = list(support = 0.002, confidence = 0.4))								
	lhs		rhs	support	confidence	coverage	lift	count
[1]	{Pedestrians}	=>	{Bicycles}	0,00354	1,00000	0,00354	282,75	8
[2]	{Bicycles}	=>	{Pedestrians}	0,00354	1,00000	0,00354	282,75	8
[3]	{Human Factors, Pedestrians}	=>	{Bicycles}	0,00354	1,00000	0,00354	282,75	8
[4]	{Bicycles, Human Factors}	=>	{Pedestrians}	0,00354	1,00000	0,00354	282,75	8
[5]	{Data Science, Pedestrians}	=>	{Bicycles}	0,00354	1,00000	0,00354	282,75	8
[6]	{Bicycles, Data Science}	=>	{Pedestrians}	0,00354	1,00000	0,00354	282,75	8
[7]	{Data Science, Human Factors, Pedestrians}	=>	{Bicycles}	0,00354	1,00000	0,00354	282,75	8
[8]	{Bicycles, Data Science, Human Factors}	=>	{Pedestrians}	0,00354	1,00000	0,00354	282,75	8
[9]	{Data Science Applications in Education, Learning Strategies}	=>	{Teaching Strategies}	0,00221	0,71429	0,00309	269,29	5
[10]	{Data Science, Data Science Applications in Education, Learning Strategies}	=>	{Teaching Strategies}	0,00221	0,71429	0,00309	269,29	5
[11]	{Pedestrians}	=>	{Human Factors}	0,00354	1,00000	0,00354	251,33	8
[12]	{Bicycles}	=>	{Human Factors}	0,00354	1,00000	0,00354	251,33	8
[13]	{Bicycles, Pedestrians}	=>	{Human Factors}	0,00354	1,00000	0,00354	251,33	8
[14]	{Data Science, Pedestrians}	=>	{Human Factors}	0,00354	1,00000	0,00354	251,33	8
[15]	{Bicycles, Data Science}	=>	{Human Factors}	0,00354	1,00000	0,00354	251,33	8
[16]	{Bicycles, Data Science, Pedestrians}	=>	{Human Factors}	0,00354	1,00000	0,00354	251,33	8
[17]	{Human Factors}	=>	{Pedestrians}	0,00354	0,88889	0,00398	251,33	8
[18]	{Human Factors}	=>	{Bicycles}	0,00354	0,88889	0,00398	251,33	8
[19]	{Data Science, Human Factors}	=>	{Pedestrians}	0,00354	0,88889	0,00398	251,33	8
[20]	{Data Science, Human Factors}	=>	{Bicycles}	0,00354	0,88889	0,00398	251,33	8
[21]	{Crash Analysis, Safety}	=>	{Crash Prediction Models}	0,00221	0,62500	0,00354	235,63	5
[22]	{Information Systems, Urban Transportation Data and Information Systems}	=>	{Urban Transportation Data}	0,00221	0,62500	0,00354	235,63	5
[23]	{Crash Analysis, Data Science, Safety}	=>	{Crash Prediction Models}	0,00221	0,62500	0,00354	235,63	5
[24]	{Data Science, Information Systems, Urban Transportation Data and Information Systems}	=>	{Urban Transportation Data}	0,00221	0,62500	0,00354	235,63	5
[25]	{Graphs, Networks}	=>	{Quantitative Graph Theory}	0,00398	0,90000	0,00442	226,20	9
[26]	{Data Science, Graphs, Networks}	=>	{Quantitative Graph Theory}	0,00398	0,90000	0,00442	226,20	9
[27]	{Artificial intelligencenability, Artificial intelligencenability and Resilience}	=>	{Resilience}	0,00265	0,85714	0,00309	215,43	6
[28]	{Artificial intelligencenability, Artificial intelligencenability and Resilience, Data Science}	=>	{Resilience}	0,00265	0,85714	0,00309	215,43	6
[29]	{Crash Prediction Models}	=>	{Crash Analysis}	0,00221	0,83333	0,00265	209,44	5
[30]	{Crash Analysis}	=>	{Crash Prediction Models}	0,00221	0,55556	0,00398	209,44	5
[31]	{Crash Prediction Models, Safety}	=>	{Crash Analysis}	0,00221	0,83333	0,00265	209,44	5
[32]	{Crash Prediction Models, Data Science}	=>	{Crash Analysis}	0,00221	0,83333	0,00265	209,44	5
[33]	{Crash Analysis, Data Science}	=>	{Crash Prediction Models}	0,00221	0,55556	0,00398	209,44	5
[34]	{Crash Prediction Models, Data Science, Safety}	=>	{Crash Analysis}	0,00221	0,83333	0,00265	209,44	5
[35]	{Quantitative Graph Theory}	=>	{Graphs}	0,00398	1,00000	0,00398	188,50	9
[36]	{Graphs}	=>	{Quantitative Graph Theory}	0,00398	0,75000	0,00531	188,50	9
[37]	{Crash Data}	=>	{Crash Analysis}	0,00265	0,75000	0,00354	188,50	6
[38]	{Crash Analysis}	=>	{Crash Data}	0,00265	0,66667	0,00398	188,50	6
[39]	{Networks, Quantitative Graph Theory}	=>	{Graphs}	0,00398	1,00000	0,00398	188,50	9
[40]	{Data Science, Quantitative Graph Theory}	=>	{Graphs}	0,00398	1,00000	0,00398	188,50	9
[41]	{Data Science, Graphs}	=>	{Quantitative Graph Theory}	0,00398	0,75000	0,00531	188,50	9
[42]	{Crash Data, Data Science}	=>	{Crash Analysis}	0,00265	0,75000	0,00354	188,50	6
[43]	{Crash Analysis, Data Science}	=>	{Crash Data}	0,00265	0,66667	0,00398	188,50	6
[44]	{Data Science, Networks, Quantitative Graph Theory}	=>	{Graphs}	0,00398	1,00000	0,00398	188,50	9
[45]	{Crash Data, Safety}	=>	{Crash Analysis}	0,00221	0,71429	0,00309	179,52	5
[46]	{Crash Data, Data Science, Safety}	=>	{Crash Analysis}	0,00221	0,71429	0,00309	179,52	5
[47]	{Crash Analysis, Safety}	=>	{Crash Data}	0,00221	0,62500	0,00354	176,72	5
[48]	{Crash Analysis, Data Science, Safety}	=>	{Crash Data}	0,00221	0,62500	0,00354	176,72	5
[49]	{Teaching Strategies}	=>	{Learning Strategies}	0,00265	1,00000	0,00265	174,00	6
[50]	{Learning Strategies}	=>	{Teaching Strategies}	0,00265	0,46154	0,00575	174,00	6

In the third table, support = 0.004 and confidence = 0.3 are applied.

As indicated in Table 3, distance education and online education exhibit a collaborative relationship with data science applications in education, demonstrating a confidence level of 100%. As seen in the other rules, information systems of data science are used together with artificial intelligence, machine learning, big data, deep learning, decision trees, and random forest areas. This is not statistically surprising either, with the interestingness coefficient decreasing.

Table 3. Apriori Algorithm
rules = apriori(data = dataset, parameter = list(support = 0.004, confidence = 0.3))

rules = apriori(data = dataset, parameter = list(support = 0.004, confidence = 0.3))								
	lhs		rhs	support	confidence	coverage	lift	count
[1]	{Graphs}	=>	{Networks}	0,00442	0,83333	0,00531	145,00	10
[2]	{Networks}	=>	{Graphs}	0,00442	0,76923	0,00575	145,00	10
[3]	{Data Science, Graphs}	=>	{Networks}	0,00442	0,83333	0,00531	145,00	10
[4]	{Data Science, Networks}	=>	{Graphs}	0,00442	0,76923	0,00575	145,00	10
[5]	{Distance Education and Online Learning, Education}	=>	{Data Science Applications in Education}	0,00486	1,00000	0,00486	68,55	11
[6]	{Data Science, Distance Education and Online Learning, Education}	=>	{Data Science Applications in Education}	0,00486	1,00000	0,00486	68,55	11
[7]	{Data Science Applications in Education, Education}	=>	{Distance Education and Online Learning}	0,00486	0,39286	0,01238	59,24	11
[8]	{Data Science, Data Science Applications in Education, Education}	=>	{Distance Education and Online Learning}	0,00486	0,39286	0,01238	59,24	11
[9]	{Data Science, Data Science Applications in Education}	=>	{Distance Education and Online Learning}	0,00486	0,34375	0,01415	51,84	11
[10]	{Distance Education and Online Learning}	=>	{Data Science Applications in Education}	0,00486	0,73333	0,00663	50,27	11
[11]	{Data Science Applications in Education}	=>	{Distance Education and Online Learning}	0,00486	0,33333	0,01459	50,27	11
[12]	{Data Science, Distance Education and Online Learning}	=>	{Data Science Applications in Education}	0,00486	0,73333	0,00663	50,27	11
[13]	{Information Systems}	=>	{Information Systems and Technology}	0,00619	0,37838	0,01636	42,79	14
[14]	{Data Science, Information Systems}	=>	{Information Systems and Technology}	0,00619	0,37838	0,01636	42,79	14
[15]	{Information Systems and Technology}	=>	{Information Systems}	0,00619	0,70000	0,00884	42,79	14
[16]	{Data Science, Information Systems and Technology}	=>	{Information Systems}	0,00619	0,70000	0,00884	42,79	14
[17]	{Data Science Applications in Education, Distance Education and Online Learning}	=>	{Education}	0,00486	1,00000	0,00486	22,18	11
[18]	{Data Science, Data Science Applications in Education, Distance Education and Online Learning}	=>	{Education}	0,00486	1,00000	0,00486	22,18	11
[19]	{Data Science, Data Science Applications in Education}	=>	{Education}	0,01238	0,87500	0,01415	19,40	28
[20]	{Data Science Applications in Education}	=>	{Education}	0,01238	0,84848	0,01459	18,82	28
[21]	{Distance Education and Online Learning}	=>	{Education}	0,00486	0,73333	0,00663	16,26	11
[22]	{Data Science, Distance Education and Online Learning}	=>	{Education}	0,00486	0,73333	0,00663	16,26	11
[23]	{Data Science, Data Science Education}	=>	{Education}	0,00531	0,60000	0,00884	13,31	12
[24]	{Data Science Education}	=>	{Education}	0,00531	0,57143	0,00928	12,67	12
[25]	{Advanced Computing Applications}	=>	{Artificial intelligence}	0,00796	1,00000	0,00796	9,05	18
[26]	{Advanced Computing Applications, Machine Learning}	=>	{Artificial intelligence}	0,00442	1,00000	0,00442	9,05	10
[27]	{Advanced Computing Applications, Data Science}	=>	{Artificial intelligence}	0,00796	1,00000	0,00796	9,05	18
[28]	{Advanced Computing Applications, Data Science, Machine Learning}	=>	{Artificial intelligence}	0,00442	1,00000	0,00442	9,05	10
[29]	{Neural Networks}	=>	{Deep Learning}	0,00486	0,34375	0,01415	7,48	11
[30]	{Data Science, Neural Networks}	=>	{Deep Learning}	0,00486	0,34375	0,01415	7,48	11
[31]	{Neural Networks}	=>	{Artificial intelligence}	0,00840	0,59375	0,01415	5,37	19
[32]	{Data Science, Neural Networks}	=>	{Artificial intelligence}	0,00840	0,59375	0,01415	5,37	19
[33]	{Big Data, Machine Learning}	=>	{Artificial intelligence}	0,01370	0,51667	0,02653	4,67	31
[34]	{Big Data, Data Science, Machine Learning}	=>	{Artificial intelligence}	0,01370	0,51667	0,02653	4,67	31
[35]	{Decision Trees}	=>	{Machine Learning}	0,00486	0,91667	0,00531	4,44	11
[36]	{Data Science, Decision Trees}	=>	{Machine Learning}	0,00486	0,91667	0,00531	4,44	11
[37]	{Analytics}	=>	{Big Data}	0,00531	0,52174	0,01017	4,44	12
[38]	{Analytics, Data Science}	=>	{Big Data}	0,00531	0,52174	0,01017	4,44	12
[39]	{Deep Learning, Machine Learning}	=>	{Artificial intelligence}	0,00973	0,47826	0,02034	4,33	22
[40]	{Data Science, Deep Learning, Machine Learning}	=>	{Artificial intelligence}	0,00973	0,47826	0,02034	4,33	22
[41]	{Big Data, Deep Learning}	=>	{Machine Learning}	0,00707	0,76190	0,00928	3,69	16
[42]	{Big Data, Data Science, Deep Learning}	=>	{Machine Learning}	0,00707	0,76190	0,00928	3,69	16
[43]	{Safety}	=>	{Artificial intelligence}	0,00531	0,40000	0,01326	3,62	12
[44]	{Data Science, Safety}	=>	{Artificial intelligence}	0,00531	0,40000	0,01326	3,62	12
[45]	{Random Forest}	=>	{Machine Learning}	0,00575	0,72222	0,00796	3,50	13
[46]	{Data Science, Random Forest}	=>	{Machine Learning}	0,00575	0,72222	0,00796	3,50	13
[47]	{Artificial intelligence, Big Data}	=>	{Machine Learning}	0,01370	0,72093	0,01901	3,50	31
[48]	{Artificial intelligence, Big Data, Data Science}	=>	{Machine Learning}	0,01370	0,72093	0,01901	3,50	31
[49]	{Deep Learning}	=>	{Artificial intelligence}	0,01636	0,35577	0,04598	3,22	37
[50]	{Data Science, Deep Learning}	=>	{Artificial intelligence}	0,01636	0,35577	0,04598	3,22	37

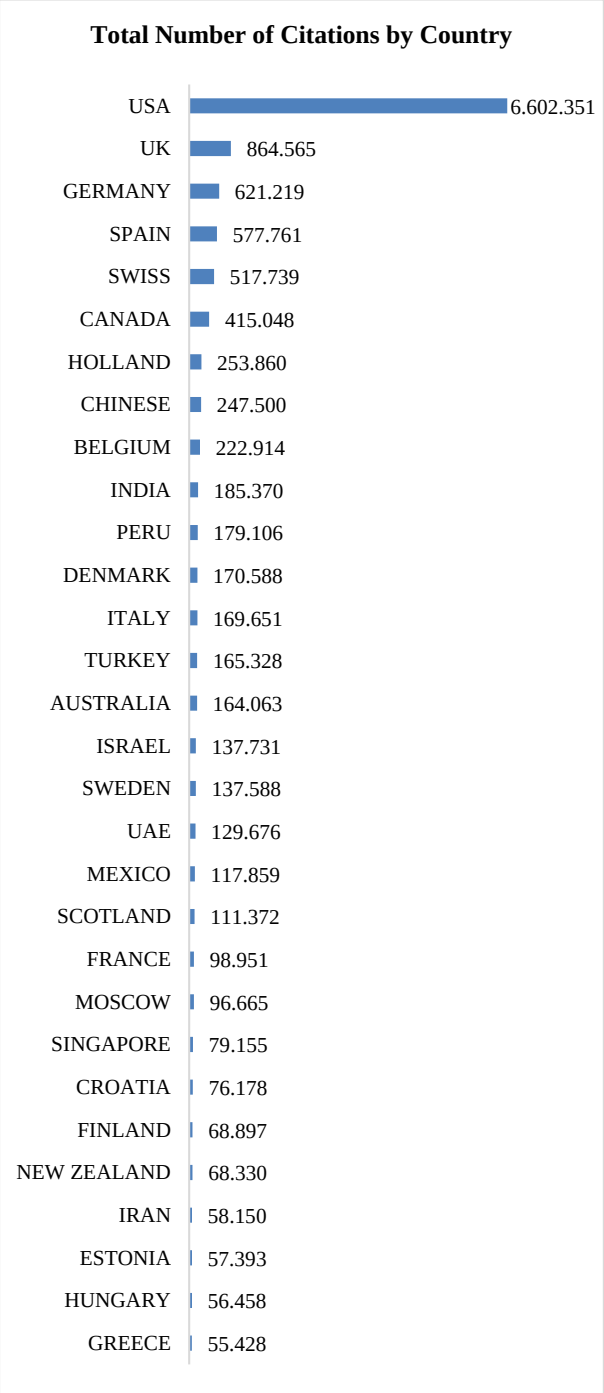
In order to create an alternative model, support = 0.01 and confidence = 0.2 are determined for Table 4. The keywords "Data Science" and "Data Science Application in Education", with "Education" keywords have the highest lift value. The most interesting relationship is between these keywords. Authors working in the field of Data Science and Data Science in Education with a probability of 87% also worked in the field of education. Researchers in data science have also carried out studies in big data, artificial intelligence, visualisation, information systems, machine learning, statistics, forecast, education, data mining and IOT with 100% probability.

Table 4. Apriori Algorithm
rules = apriori(data = dataset, parameter = list(support = 0.01, confidence = 0.2))

rules = apriori(data = dataset, parameter = list(support = 0.01, confidence = 0.2))								
	lhs		rhs	support	confidence	coverage	lift	count
[1]	{Data Science, Data Science Applications in Education}	=>	{Education}	0,01238	0,87500	0,01415	19,40	28
[2]	{Data Science, Education}	=>	{Data Science Applications in Education}	0,01238	0,28000	0,04421	19,19	28
[3]	{Data Science Applications in Education}	=>	{Education}	0,01238	0,84848	0,01459	18,82	28
[4]	{Education}	=>	{Data Science Applications in Education}	0,01238	0,27451	0,04509	18,82	28
[5]	{Big Data, Machine Learning}	=>	{Artificial intelligence}	0,01370	0,51667	0,02653	4,67	31
[6]	{Big Data, Data Science, Machine Learning}	=>	{Artificial intelligence}	0,01370	0,51667	0,02653	4,67	31
[7]	{Artificial intelligence, Big Data}	=>	{Machine Learning}	0,01370	0,72093	0,01901	3,49	31
[8]	{Artificial intelligence, Big Data, Data Science}	=>	{Machine Learning}	0,01370	0,72093	0,01901	3,49	31
[9]	{Deep Learning}	=>	{Artificial intelligence}	0,01636	0,35577	0,04598	3,22	37
[10]	{Data Science, Deep Learning}	=>	{Artificial intelligence}	0,01636	0,35577	0,04598	3,22	37
[11]	{Artificial intelligence, Data Science, Machine Learning}	=>	{Big Data}	0,01370	0,27928	0,04907	2,36	31
[12]	{Artificial intelligence, Machine Learning}	=>	{Big Data}	0,01370	0,27434	0,04996	2,32	31
[13]	{Artificial intelligence}	=>	{Machine Learning}	0,04996	0,45200	0,11052	2,19	113
[14]	{Machine Learning}	=>	{Artificial intelligence}	0,04996	0,24197	0,20645	2,19	113
[15]	{Data Science, Machine Learning}	=>	{Artificial intelligence}	0,04907	0,23974	0,20469	2,17	111
[16]	{Artificial intelligence, Data Science}	=>	{Machine Learning}	0,04907	0,44758	0,10964	2,17	111
[17]	{Deep Learning}	=>	{Machine Learning}	0,02034	0,44231	0,04598	2,14	46
[18]	{Data Science, Deep Learning}	=>	{Machine Learning}	0,02034	0,44231	0,04598	2,14	46
[19]	{Big Data, Data Science}	=>	{Machine Learning}	0,02653	0,22556	0,11760	1,09	60
[20]	{Big Data}	=>	{Machine Learning}	0,02653	0,22388	0,11848	1,08	60
[21]	{Analytics}	=>	{Data Science}	0,01017	1,00000	0,01017	1,01	23
[22]	{Teaching}	=>	{Data Science}	0,01105	1,00000	0,01105	1,01	25
[23]	{Data Visualization}	=>	{Data Science}	0,01149	1,00000	0,01149	1,01	26
[24]	{Modeling}	=>	{Data Science}	0,01017	1,00000	0,01017	1,01	23
[25]	{Artificial intelligencenability}	=>	{Data Science}	0,01017	1,00000	0,01017	1,01	23
[26]	{Big Data Analytics}	=>	{Data Science}	0,01238	1,00000	0,01238	1,01	28
[27]	{Forecasting}	=>	{Data Science}	0,01370	1,00000	0,01370	1,01	31
[28]	{Neural Networks}	=>	{Data Science}	0,01415	1,00000	0,01415	1,01	32
[29]	{Operations}	=>	{Data Science}	0,01282	1,00000	0,01282	1,01	29
[30]	{IOT}	=>	{Data Science}	0,01636	1,00000	0,01636	1,01	37
[31]	{Safety}	=>	{Data Science}	0,01326	1,00000	0,01326	1,01	30
[32]	{Decision Making}	=>	{Data Science}	0,01282	1,00000	0,01282	1,01	29
[33]	{Classification}	=>	{Data Science}	0,01415	1,00000	0,01415	1,01	32
[34]	{Planning and Analysis}	=>	{Data Science}	0,01459	1,00000	0,01459	1,01	33
[35]	{Visualization}	=>	{Data Science}	0,02078	1,00000	0,02078	1,01	47
[36]	{Information Systems}	=>	{Data Science}	0,01636	1,00000	0,01636	1,01	37
[37]	{COVID-19}	=>	{Data Science}	0,02564	1,00000	0,02564	1,01	58
[38]	{Statistics}	=>	{Data Science}	0,02785	1,00000	0,02785	1,01	63
[39]	{Data Analytics}	=>	{Data Science}	0,02741	1,00000	0,02741	1,01	62
[40]	{Data Mining}	=>	{Data Science}	0,03183	1,00000	0,03183	1,01	72
[41]	{Deep Learning}	=>	{Data Science}	0,04598	1,00000	0,04598	1,01	104
[42]	{Data Science Applications in Education, Education}	=>	{Data Science}	0,01238	1,00000	0,01238	1,01	28
[43]	{Artificial intelligence, Deep Learning}	=>	{Data Science}	0,01636	1,00000	0,01636	1,01	37
[44]	{Deep Learning, Machine Learning}	=>	{Data Science}	0,02034	1,00000	0,02034	1,01	46
[45]	{Artificial intelligence, Big Data}	=>	{Data Science}	0,01901	1,00000	0,01901	1,01	43
[46]	{Big Data, Machine Learning}	=>	{Data Science}	0,02653	1,00000	0,02653	1,01	60
[47]	{Artificial intelligence, Big Data, Machine Learning}	=>	{Data Science}	0,01370	1,00000	0,01370	1,01	31
[48]	{Big Data}	=>	{Data Science}	0,11760	0,99254	0,11848	1,00	266

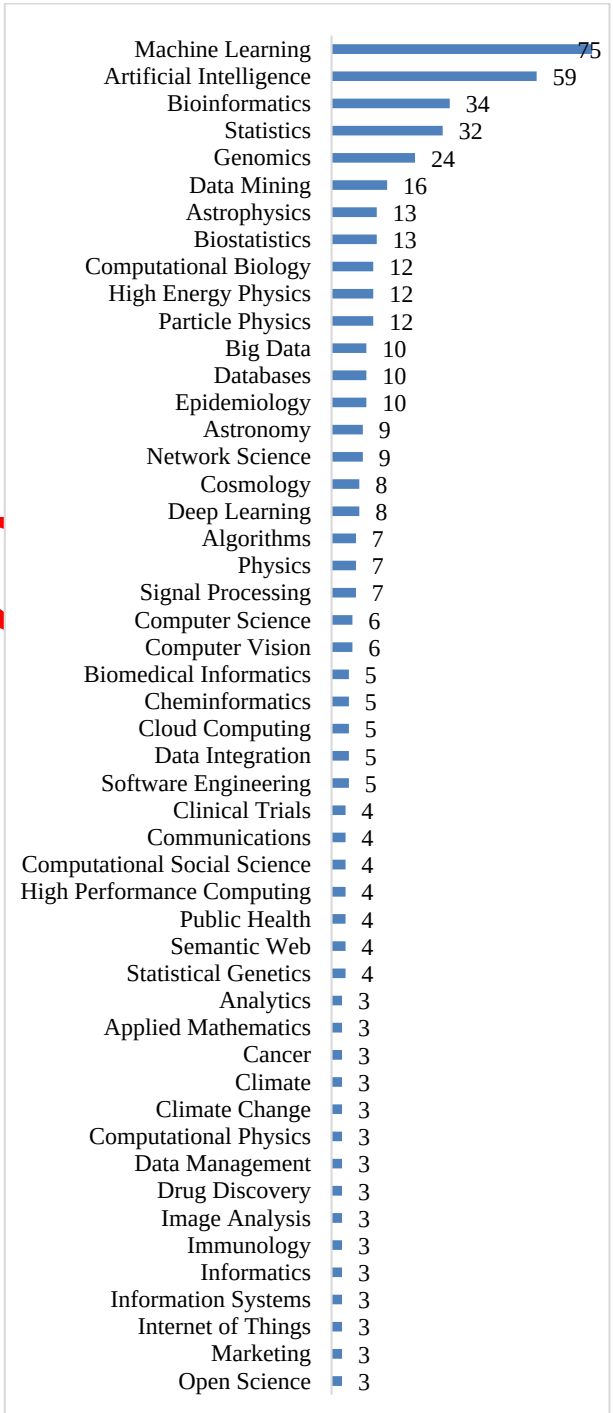
As a result of the application made to the list of the most cited authors obtained from Google Scholar as the second data group, graph 3, graph 4, table 5 and table 6 are obtained.

Graph 3 gives the number of citations made to the researchers' countries. The USA received the most citations, followed by the United Kingdom, Germany, Spain, Sweden and Canada. In the citation ranking, Turkey is located just after Italy and before Australia.



Graph 3. Total Number of Citations by Country

The most frequently used keywords in data science are as follows: Machine Learning (75), Artificial Intelligence (59), Bioinformatics (34), Statistics (32), and Genomics (24). As depicted in Graph 4, 50 unique keywords with a frequency of more than 3 are shown together with Data Science. As expected, most research in data science has focused on machine learning (75) and artificial intelligence (59). Interestingly, authors who have made significant contributions to the field have also studied bioinformatics (34) and genomics (24). The analysis indicates a close relationship between the fields of statistics and data science, as well as bioinformatics.



Graph 4. Academic Research Topics

Table 5. Descriptive Statistics of Number of Citations

Min	Q1	Median	Mean	Q3	Max
1	3	4	4.114	5	5

The number of keywords in the articles studied in the field of data science is also examined in the second data group. A minimum of 1 keyword and a maximum of 11 keywords are observed in an article. As observed in 157 articles, a maximum of 5 keywords are used.

Number of articles with 1 keyword: 11

Number of articles with 2 keywords: 16

Number of articles with 3 keywords: 56

Number of articles with 4 keywords: 76

Number of articles with 5 keywords: 157

In analysing the data obtained from Google Scholar, the support value is determined to be 0.01 and the confidence value to be 0.1.

As shown in Table 6, as a result of the analysis, it is seen that those who work together in the field of data science and genomics also work in the field of bioinformatics with a probability of 54%.

Those working in cosmology and data science, as well as those working only in the field of cosmology, work in machine learning with 62.5% confidence.

Table 6. Apriori Algorithm

rules = apriori(data = dataset, parameter = list(support = 0.01, confidence = 0.1))

rules = apriori(data = dataset, parameter = list(support = 0.01, confidence = 0.1))							
	lhs		rhs	support	confidence	coverage	lift
[1]	{Genomics}	=>	{Bioinformatics}	0,04114	0,54167	0,07595	5,03
[2]	{Bioinformatics}	=>	{Genomics}	0,04114	0,38235	0,10759	5,03
[3]	{Data Science, Genomics}	=>	{Bioinformatics}	0,04114	0,54167	0,07595	5,03
[4]	{Bioinformatics, Data Science}	=>	{Genomics}	0,04114	0,38235	0,10759	5,03
[5]	{Biostatistics}	=>	{Genomics}	0,01266	0,30769	0,04114	4,05
[6]	{Biostatistics, Data Science}	=>	{Genomics}	0,01266	0,30769	0,04114	4,05
[7]	{Genomics}	=>	{Biostatistics}	0,01266	0,16667	0,07595	4,05
[8]	{Data Science, Genomics}	=>	{Biostatistics}	0,01266	0,16667	0,07595	4,05
[9]	{Bioinformatics}	=>	{Computational Biology}	0,01266	0,11765	0,10759	3,10
[10]	{Bioinformatics, Data Science}	=>	{Computational Biology}	0,01266	0,11765	0,10759	3,10
[11]	{Computational Biology}	=>	{Bioinformatics}	0,01266	0,33333	0,03797	3,10
[12]	{Computational Biology, Data Science}	=>	{Bioinformatics}	0,01266	0,33333	0,03797	3,10
[13]	{Computer Vision}	=>	{Machine Learning}	0,01266	0,66667	0,01899	2,81
[14]	{Computer Vision, Data Science}	=>	{Machine Learning}	0,01266	0,66667	0,01899	2,81
[15]	{Cosmology}	=>	{Machine Learning}	0,01582	0,62500	0,02532	2,63
[16]	{Cosmology, Data Science}	=>	{Machine Learning}	0,01582	0,62500	0,02532	2,63
[17]	{AI}	=>	{Machine Learning}	0,08544	0,45763	0,18671	1,93
[18]	{AI, Data Science}	=>	{Machine Learning}	0,08544	0,45763	0,18671	1,93
[19]	{Machine Learning}	=>	{AI}	0,08544	0,36000	0,23734	1,93
[20]	{Data Science, Machine Learning}	=>	{AI}	0,08544	0,36000	0,23734	1,93
[21]	{Data Mining}	=>	{AI}	0,01582	0,31250	0,05063	1,67
[22]	{Data Mining, Data Science}	=>	{AI}	0,01582	0,31250	0,05063	1,67
[23]	{High Energy Physics}	=>	{Machine Learning}	0,01266	0,33333	0,03797	1,40
[24]	{Data Science, High Energy Physics}	=>	{Machine Learning}	0,01266	0,33333	0,03797	1,40
[25]	{Data Mining}	=>	{Machine Learning}	0,01582	0,31250	0,05063	1,32

6. CONCLUSION

This study aims to explore various areas in the field of data science and provide guidance for future research efforts. Therefore, data was obtained from WoS and Google Scholars to examine the works in the field of Data Science. Apriori analysis with R programming language was applied to these two separate datasets to examine the authors and their works working in the field of data science. The analysis of both datasets confirms that the majority of studies in this field are concentrated in the fields of machine learning, artificial intelligence, big data and deep learning, in line with expectations.

Based on the first set of data, it is concluded that decision makers predominantly use data science as a key tool to increase policy effectiveness and drive improvements [2,13]. Despite the importance of data science in policy making, there is limited research on how data science can be better integrated into smaller-scale policy decisions or local government contexts [10]. Apriori analysis of the same dataset also reveals that data science plays a crucial role in the diagnosis, treatment and gene detection of breast cancer. Data science is already used in the diagnosis of many diseases to provide early and accurate diagnosis and treatment. The use of genomic information in medical applications is strongly linked to advances in genomic technologies and sciences. Data science is being used to decipher the genetic makeup of cancer cells and the genome of cancer cells to develop personalized treatment approaches based on genetic characteristics [50]. However, the results of the analysis show that most studies focus on the diagnosis and treatment of breast cancer. More research is needed on how data science can contribute to less studied areas in healthcare, such as rare diseases or mental health.

It is also noteworthy that researchers involved in data science competitions, national networks of ecological observatories and remote sensing are also working on species classification. The review by Passnacht et al. shows that research in this area has increased significantly in recent years [16]. Despite this growth, some ecological niches or underrepresented species are less explored, highlighting a potential gap in current research.

Another result of the same dataset is that researchers working in data science, human factors and pedestrians can work in the field of cycling with 100% confidence. The increase in the number of recorded traffic accidents involving cyclists in recent years and the high rate of fatal traffic accidents is a global challenge for public health, urban development and sustainability [52]. While the intersection of data science and transportation safety is well documented, there is a need for more in-depth studies on how data science can proactively prevent accidents rather than just analyzing post-accident data.

As a result of the analysis, it was revealed that data science applications in education are important in determining learning strategies. Data science applications will provide students with more effective and efficient education. There is a relationship between data science applications in education and distance education and online education at 100% confidence level. Distance education will also become more efficient thanks to data science applications. However, research is still lacking on how data science can address challenges in underserved educational settings such as rural areas or schools with limited resources.

The second data set was obtained from the fields of study of the most cited authors from Google Scholar. Apriori analysis was applied to the second data set in R studio program. As a result of the analysis, it is seen that those who work together in data science and genomics are 54% likely to work in bioinformatics. Bioinformatics is a field that has gained importance with the development of data science tools and technology. Since the size of the human genome can reach up to 200 GB, Big Data Analytics in Bioinformatics is very important [35]. Bioinformatics research is analyzed with voluminous datasets and complex data analytics methods [21].

As a result of the apriori analysis applied to the data obtained from Google Scholar, it is seen that those working in the field of cosmology and data science and those working only in the field of cosmology also work in the field of machine learning. Human beings have been trying to understand the universe since the day it existed and have carried out many studies in this way. With the development of data science and machine learning methods, these methods have been used to understand the universe. Machine learning and data science applications have significantly improved the way cosmologists interpret data [31].

The fact that those working in the field of data science work in so many different fields proves once again that the field of data science is a multidisciplinary field. In this context, the support of data scientists to people working in other fields will contribute significantly to the development of science.

DECLARATION OF ETHICAL STANDARDS)

This article do not require ethical committee permission and/or legal-special permission.

AUTHORS' CONTRIBUTIONS

Merve Nur BARUN: Performed the experiments and analyse the results. Wrote the manuscript.

Emrah ÖNDER: Performed the experiments and analyse the results. Wrote the manuscript.

CONFLICT OF INTEREST

There is no conflict of interest in this study.

REFERENCES

[1] Agrawal, R., Imieliński, T., & Swami, A., "Mining association rules between sets of items in large databases", In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216., (1993).

[2] Anthopoulos, L., & Kazantzi, V., "Urban energy efficiency assessment models from an AI and big data

perspective: Tools for policy makers", ***Sustainable Cities and Society,*** 76, 10349, (2022).

[3] Ataş, K., Kaya, A., & Myderrizi, I., "Yapay Sinir Ağı Tabanlı Model ile X-ray Görüntülerinden Covid-19 Teşhisi", ***Politeknik Dergisi,*** 26(2), 541-551, (2023).

[4] Balcı, F., & Yılmaz, S., "Faster R-CNN Structure for Computer Vision-based Road Pavement Distress Detection", ***Politeknik Dergisi,*** 26(2), 701-710, (2023).

[5] Bayardo Jr, R. J., "Efficiently mining long patterns from databases", In Proceedings of the 1998 ACM SIGMOD international conference on Management of data, 85-93, (1998).

[6] Bellinger, C., Sharma, S., Japkowicz, N., & Zaïane, O. R., "Framework for extreme imbalance classification: SWIM—sampling with the majority class", ***Knowledge and Information Systems, 62***, 841-866, (2020).

[7] Chen, L. P., "Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python: by Peter Bruce, Andrew Bruce, and Peter Gedeck", O'Reilly Media Inc., Boston, United States, 272-273, (2021).

[8] Chen, L.P., "Model-based Clustering and Classification for Data Science: With Application in R by Harles Bouveyron, Gilles Celeus, T. Bredan Murphy and Adrian E. Raftery (2019)," ***Biometrical Journal***, 62, 1120–1121, (2020).

[9] Chiang, D. A., Wang, Y. F., Lee, S. L., & Lin, C. J., "Goal-oriented sequential pattern for network banking churn analysis", ***Expert systems with applications,*** 25(3), 293-302, (2003).

[10] Dawes, S. S., "The evolution and continuing challenges of e-governance", ***Public administration review***, 68, 86-102, (2008).

[11] Değer, K., Özkaya, M. G., & Boran, F. E., "Modelling and Analysis of Future Energy Scenarios on the Sustainability Axis", ***Politeknik Dergisi,*** 26(2), 665-678, (2023).

[12] Donoho, D., "50 years of data science", ***Journal of Computational and Graphical Statistics***, 26(4), 745-766, (2017).

[13] Durmuş Şenyapar, H. N., Cetinkaya, U., & Bayındır, R., "Renewable Energy Incentives and Future Implications for Turkey: A Comparative Bibliometric Analysis", ***Politeknik Dergisi,*** 27(1), 329-342, (2024).

[14] Edastama, P., Bist, A. S., & Prambudi, A., "Implementation of data mining on glasses sales using the apriori algorithm", ***International Journal of Cyber and IT Service Management,*** 1(2), 159-172, (2021).

[15] Fathi, M., Haghi Kashani, M., Jameii, S. M., & Mahdipour, E., "Big data analytics in weather forecasting: A systematic review", ***Archives of Computational Methods in Engineering,*** 29(2), 1247-1275, (2022).

[16] Fassnacht, F. E., Latifi, H., Stereńczak, K., Modzelewska, A., Lefsky, M., Waser, L. T., ... & Ghosh, A., "Review of studies on tree species classification from remotely sensed data", ***Remote sensing of environment***, 186, 64-87, (2016).

[17] Harun, N. A., Makhtar, M., Abd Aziz, A., Zakaria, Z. A., & Syed, F., "The application of apriori algorithm in predicting flood areas", ***management,*** 17, 18, (2017).

[18] Hegland, M., "The apriori algorithm–a tutorial", ***Mathematics and computation in imaging science and information processing,*** 209-262, (2007).

[19] Javaid, M., Haleem, A., Singh, R. P., Rab, S., & Suman, R., "Internet of Behaviours (IoB) and its role in customer services", ***Sensors International,*** 2, (2021).

[20] Ji, L., Zhang, B., & Li, J., "A new improvement on apriori algorithm", In 2006 International Conference on Computational Intelligence and Security, 1, 840-844, (2006).

[21] Kashyap, H., Ahmed, H. A., Hoque, N., Roy, S., & Bhattacharyya, D. K., "Big data analytics in bioinformatics: A machine learning perspective", arXiv preprint arXiv:1506.05101, (2015).

[22] Korkmaz, Ş., & Alkan, M., "Derin Öğrenme Algoritmalarını Kullanarak Deepfake Video Tespiti", ***Politeknik Dergisi,*** 26(2), 855-862, (2023).

[23] Korschun, D., & Welker, G., "We are Market Basket: The story of the unlikely grassroots movement that saved a beloved business", Amacom, (2015).

[24] Kunnathuvalappil Hariharan, N., "Applications of Data Mining in Finance", ***Naveen International Journal of Innovations in Engineering Research and Technology,*** 5(2), 72-77, (2018).

[25] Li, Z., Li, X., Tang, R., & Zhang, L., "Apriori algorithm for the data mining of global cyberspace security issues for human participatory based on association rules", ***Frontiers in Psychology***, 11, (2021).

[26] Mannila, H., "Theoretical frameworks for data mining", ***ACM SIGKDD Explorations Newsletter***, 1(2), 30-32, (2000).

[27] Mikut, R., & Reischl, M., "Data mining tools", ***Wiley interdisciplinary reviews: data mining and knowledge discovery,*** 1(5), 431-443, (2011).

[28] Mohapatra, D., Tripathy, J., Mohanty, K. K., & Nayak, D. S. K., "Interpretation of optimized hyper parameters in associative rule learning using eclat and apriori", In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 879-882, (2021).

[29] Nan, S., & Chen, M., "An apriori-algorithm-based analysis method on physical fitness test data for college students", Easychair Working Paper, (2020).

[30] Nandagopal, S., Karthik, S., & Arunachalam, V. P., "Mining of meteorological data using modified apriori algorithm" ***European Journal of Scientific Research,*** 47(2), 295-308, (2010).

[31] Ntampaka, M., Avestruz, C., Boada, S., Caldeira, J., Cisewski-Kehe, J., Di Stefano, R., ... & Wandelt, B., "The role of machine learning in the next decade of cosmology", arXiv preprint arXiv:1902.10159, (2019).

[32] O'Hagan, A., "The Bayesian approach to statistics", ***Handbook of probability: Theory and applications***, 85-100, (2008).

[33] Olodude, O. O., & Oladejo, B. F., "Enhanced customer-based knowledge management system for products generation in banking system", ***Computer Science Series,*** 11(1), 129-137, (2013).

[34] Osman, A. S., "Data mining techniques", ***Data Science and Networking,*** 2(1), (2019).

[35] Patel, D. T., "Big data analytics in bioinformatics", ***In Biotechnology: Concepts, Methodologies, Tools, and Applications,***1967-1984, (2019).

[36] Pei, J., Mao, R., Hu, K., & Zhu, H., "Towards data mining benchmarking: a test bed for performance study of frequent pattern mining", In Proceedings of the 2000 ACM SIGMOD international conference on Management of data, 592, (2000).

[37] Pfannkuch, M., & Wild, C., "Towards an understanding of statistical thinking", ***The challenge of developing statistical literacy, reasoning and thinking***, 17-46, (2004).

[38] Raghupathi, W., & Raghupathi, V., "Big data analytics in healthcare: promise and potential", ***Health information science and systems,*** 2, 1-10, (2014).

[39] Rong, C., Liu, Z., Huo, N., & Sun, H., "Exploring Chinese dietary habits using recipes extracted from websites", IEEE Access, 7, 24354-24361, (2019).

[40] Sathya, M., & Devi, P. I., "Apriori algorithm on web logs for mining frequent link. In 2017 IEEE International Conference on Intelligent Techniques in Control", ***Optimization and Signal Processing (INCOS),*** 1-5, (2017).

[41] Savasere, A., Omiecinski, E. R., & Navathe, S. B., "An efficient algorithm for mining association rules in large databases", ***Georgia Institute of Technology,*** (1995).

[42] Semeler, A. R., Pinto, A. L., & Rozados, H. B. F., "Data science in data librarianship: Core competencies of a data librarian", ***Journal of Librarianship and Information Science,*** 51(3), 771-780, (2019).

[43] Sertçelik, Ş., & Önder, E., "Yönetim Bilişim Sistemleri Kapsamında Akademik Araştırma Alanlarının İncelenmesi: Apriori Algoritması ile Bir Analiz", ***Gümüşhane Üniversitesi Sosyal Bilimler Dergisi,*** 14(2), 680-690, (2023).

[44] Shao, L., "Research on sports training decision support system based on improved association rules algorithm", ***Security and Communication Networks,*** 1-6, (2021).

[45] Singh, J., Ram, H., & Sodhi, D. J., "Improving efficiency of apriori algorithm using transaction reduction", ***International Journal of Scientific and Research Publications***, 3(1), 1-4, (2013).

[46] Sornalakshmi, M., Balamurali, S., Venkatesulu, M., Krishnan, M. N., Ramasamy, L. K., Kadry, S., & Lim, S., "An efficient apriori algorithm for frequent pattern mining using mapreduce in healthcare data", ***Bulletin of Electrical Engineering and Informatics,*** 10(1), 390-403, (2021).

[47] Spearman, C., "The proof and measurement of association between two things", ***The American Journal of Psychology,*** 15(1), 72–101, (1904).

[48] Spearman, C., "Footrule for measuring correlation", ***British Journal of Psychology,*** 2(1), 89, (1906).

[49] Sumiran, K., "An overview of data mining techniques and their application in industrial engineering", ***Asian Journal of Applied Science and Technology,*** 2(2), 947-953, (2018).

[50] Suwinski, P., Ong, C., Ling, M. H., Poh, Y. M., Khan, A. M., & Ong, H. S., "Advancing personalized medicine through the application of whole exome sequencing and big data analytics", ***Frontiers in genetics,*** 10, 49, (2019).

[51] Ullah, I., "Logıcal Reasonıng and Data Mınıng Algorıthms", ***Recent Advances In Statıstıcs***, 103, (2011).

[52] Useche, S., Montoro, L., Alonso, F., & Oviedo-Trespalacios, O., "Infrastructural and human factors affecting safety outcomes of cyclists", ***Sustainability,*** 10(2), 299, (2018).

[53] Usha, D., Niveditha, V. R., Kirubadevi, T., & Thamizhikkavi, P., "Use of predictive analytical algorithm by crime investigation team: An analysis", ***International Journal of Advances Science and Technology,*** 29, 2986-2992, (2020).

[54] Uysal, M., Acharya, A., & Saltz, J., "Structure and performance of decision support algorithms on active disks", University of Maryland, (1998).

[55] Veeramalai, S., Jaisankar, N., & Kannan, A., "Efficient web log mining using enhanced Apriori algorithm with hash tree and fuzzy", ***International journal of computer science & information Technology (IJCSIT),*** 2, 1-15, (2010).

[56] Wang, Y., "Categorization of Association Rule Mining Algorithms", In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, (2003).

[57] Wu, W., Lin, W., Hsu, C. H., & He, L., "Energy-efficient hadoop for big data analytics and computing: A systematic review and research insights", ***Future Generation Computer Systems,*** 86, 1351-1, (2018).

[58] Yuan, X., "An Improved Apriori Algorithm for Mining Association Rules", In AIP Conference Proceedings, 1820 (1), 080005, (2017).

[59] Yücel, M., Osmanca, M. S., & Mercimek, İ. F., "Machine Learning Algorithm Estimation and Comparison of Live Network Values of the Inputs Which Have the Most Effect on the FEC Parameter in DWDM Systems", ***Politeknik Dergisi,*** 27(1), 133-138, (2024).

[60] Zhang, W., Ma, D., & Yao, W., "Medical diagnosis data mining based on improved Apriori algorithm", ***Journal of Networks,*** 9(5), 1339, (2014).