INFOTECH journal
ISSN : 2460-1861 (Print), 2615-4250 (Online)
Vol. 11 No. 2 Desember 2025, pp. 402-407

ANALYSIS OF MULTIVARIABLE RELATIONSHIP PATTERNS IN SYNTHETIC HEALTH DATA USING THE APRIORI ALGORITHM

Nur Azizah Harahap[1], Fadhlan Ihsan Lubis[2], Andika Syahdewa[3], Hengki Gunawan[4], Marsini Sibuea[5], Darma Juang[6], Muhammad Syahputra Novelan[7]
[1,2,3,4,5,6,7]Master Of Information Technology Study Program, Faculty Of Postgraduate Study
University Pembangunan Pancabudi, Medan, Indonesia
Correspondence: nazizahhrp@gmail.com

ABSTRACT
This study aims to explore multivariable relationship patterns in synthetic healthcare data using the Apriori algorithm, a popular technique in the field of data mining. In the context of increasingly data-driven healthcare systems, identifying meaningful associations among patient attributes—such as medical conditions, age groups, admission types, and hospitalization cost categories—is essential for supporting efficient service delivery. The dataset used in this study underwent several preprocessing stages, including data cleaning, categorization of age and cost, and one-hot encoding of categorical features to facilitate association rule mining. Through the implementation of the Apriori algorithm, several association rules were discovered, accompanied by key metrics such as support, confidence, and lift. Notable patterns include associations between Diabetes and Urgent admissions, as well as between Middle Age patients and High hospitalization costs. Visualizations, including bar charts and network graphs, were employed to enhance interpretability and present item relationships more intuitively. Although the resulting lift values ranged between 1.01 and 1.03—indicating relatively weak correlations—the findings remain relevant for initial segmentation strategies and data-driven decision-making in hospital or clinical settings. This research demonstrates that the Apriori algorithm can effectively extract.

Keywords: *Apriori Algorithm, Association Rule Mining, Data Mining, Multivariable Pattern Analysis.*

Article History :
Received date : 12-10-2025
Revised date : 06-11-2025
Published date : 01-12-2025

DOI :
https://doi.org/10.31949/infotech.v11i2.16585

1. INTRODUCTION

In the rapidly evolving digital era, the transformation of information technology has had a significant impact across various sectors, including the health sector (Harahap et al., 2025). One of the implications of this development is the emergence of vast and complex volumes of health data. Therefore, an appropriate approach is required to process and analyze these data effectively, so they can be optimally utilized to support decision-making processes, improve service quality, and facilitate strategic planning within healthcare institutions (Putera Utama Siahaan et al., 2025).

One of the widely used approaches for uncovering hidden patterns in large datasets is data mining, which refers to the process of discovering meaningful information from massive amounts of data. Among the popular techniques in data mining is the Apriori algorithm, which is utilized to identify association rules between items within a dataset. In the healthcare context, the Apriori algorithm holds great potential for extracting new knowledge from patient data, such as relationships between specific medical conditions and variables like hospitalization costs, patient age, and types of healthcare services received. Several previous studies have highlighted the advantages of the Apriori algorithm in this context (Al-Sarayrah, 2024; Ardiansyah et al., 2023; Febrianny Ulfha et al., 2020; Novianti et al., 2020; Parinduri et al., 2024; Saputra et al., 2020).

For instance, a study conducted by Darmawan et al. (2022) employed the Apriori algorithm to identify patterns among groups of individuals experiencing social welfare issues. Similarly, Hadinata and Kurniawan (n.d.) applied the same algorithm in the context of snack product purchase analysis, demonstrating its effectiveness in discovering relationships between items. Nawangsih and Pratama (2023) also utilized the Apriori algorithm to analyze customer behavior in mini markets.

In addition, Ritha et al. (Nola Ritha et al., 2021) successfully identified patient visit patterns in healthcare facilities. All of these studies demonstrate that the Apriori algorithm possesses high flexibility and capability in exploring multivariate data.

However, there remains a research gap in studies that explicitly explore multivariable relationships among attributes such as medical conditions, age, patient admission types, and hospitalization cost categories using the Apriori algorithm.

In this study, synthetic healthcare data were used to ensure privacy, eliminate ethical risks, and allow methodological experimentation in a controlled environment. The objective is not to generate clinically conclusive interpretations, but to evaluate the capability of the Apriori algorithm in identifying multivariable relationship patterns in healthcare-structured datasets. In addition, this study interprets the resulting association rules using metrics such as support, confidence, and lift. Finally, the study presents the findings through visualizations, including bar charts and network graphs, to provide a clearer picture of the interrelationships among items.

This study is expected to contribute to the development of data-driven decision support systems in healthcare services. The findings of this research are also anticipated to assist in patient segmentation and the more efficient management of healthcare services.

2. METHOD

2.1 Research Design and Approach

This study employs an exploratory quantitative approach using data mining methods, specifically the Apriori algorithm, to uncover hidden association patterns within healthcare data. The primary objective of this research is to identify multivariable relationships among patient medical conditions, age, hospital admission types, and hospitalization cost categories. Since synthetic data are utilized, this study poses no risk to patient data privacy while remaining representative within the context of analytical exploration.

2.2 Data Sources and Preparation

The data used in this study consist of synthetic healthcare data in CSV format, containing more than 50,000 records. The dataset includes the following attributes: Medical Condition, Age, Admission Type (Entry type: Elective, Emergency, Urgent), Billing Amount, Billing Category (Low, Medium, High).

Before the analysis was conducted, the data underwent several preprocessing stages, including the numerical attribute Billing Amount was categorized into Billing Category using a quartile-based discretization method, the data were converted into a boolean format using the one-hot encoding technique as preparation for input into the Apriori algorithm and the data were cleaned to remove any duplicates or missing values.

2.3 Analysis Procedure: Apriori and Association

The main analysis was conducted using the Apriori algorithm, one of the unsupervised learning techniques in data mining. The objective was to extract relevant association rules based on combinations among the variables.

The main procedure is as follows:

1. Frequent Itemsets Generation: Identifying item combinations with high occurrence frequency. (≥ min-support).
2. Rule Generation: Generating rules from the identified itemsets based on a minimum confidence threshold.
3. Evaluation Metrics: Evaluating the strength of the rules using support, confidence, and lift. In this context, support represents the percentage of occurrences of an item

combination within the database, while confidence measures the strength of the relationship between items in an association rule (Pendidikan et al., 2017). Meanwhile, the lift ratio is a parameter used to determine the strength of the generated association rules based on their support and confidence values. The lift ratio is typically used to assess whether an association rule is valid or invalid (Rahmi et al., 2021).

The following formulas were used:

Support:

$$Support(X \rightarrow Y) = \frac{|X \cup Y|}{N} \quad (1)$$

Confidence:

$$Confidence(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)} \quad (2)$$

Lift:

$$Lift(X \rightarrow Y) = \frac{Confidence(X \rightarrow Y)}{Support(Y)} \quad (3)$$

Where N represents the total number of transactions or patient records in the dataset.

2.4 Tools and Analysis Implementation

This study was implemented using the Python programming language, utilizing several supporting libraries such as pandas and numpy for data manipulation, mlxtend for executing the Apriori algorithm and generating association rules, and matplotlib, seaborn, and networkx for visualizing the results in the form of bar charts and network graphs.

The parameters applied in the analysis consisted of a Minimum Support of 0.05 (5%), Minimum Confidence of 0.3 (30%), and Minimum Lift greater than 1, which indicates a positive association. The generated rules were subsequently sorted based on the highest lift values to highlight the strongest levels of association among the variables.

3. DISCUSSION

This chapter presents the results of the analysis of multivariable relationship patterns in synthetic healthcare data using the Apriori algorithm. The study aims to explore the interrelationships among key attributes such as Medical Condition, Billing Amount, Age Group, and Admission Type. Through data preprocessing, association rule extraction, and data visualization processes, new insights were obtained regarding patterns that hold potential for supporting decision-making within healthcare information systems.

3.1. Overview of Analysis Results

This chapter presents the results of applying the Apriori algorithm to synthetic healthcare data to identify relationship patterns among attributes such as medical condition, hospitalization cost category, age group, and admission type. The results are presented in the form of association rule tables and visualizations to support interpretation. The main focus of this analysis is to uncover multivariate relationships that can be utilized to support data-driven decision-making.

3.2. Summary of Data Preprocessing

Before the analysis process, the data underwent a preprocessing stage to ensure quality and compatibility with the Apriori algorithm.

This stage involved removing missing or incomplete data, converting numerical attributes into categorical variables through discretization (such as grouping ages and categorizing billing levels), and encoding categorical attributes into a transactional format to support the mining process.

The detailed preprocessing procedures are described in Chapter 2; however, in general, this stage aimed to transform the data into a binary format to align with the input requirements of the Apriori algorithm.

3.3. Results of Association Rule Extraction Using the Apriori Algorithm

After preprocessing, the process of identifying frequent itemsets was carried out using the Apriori algorithm with a minimum support value of 0.05. Subsequently, a minimum confidence threshold of 0.3 was applied to generate the association rules.

The following table presents the top 15 association rules, sorted according to the highest lift values:

Tabel 1. Top 15 Association Rules Sorted by Highest Lift Values

Antecedents	Consequents	Support	Confidence	Lift
(Medical Condition_Diabetes)	(Admission Type_Urgent)	0.058180	0.347055	1.036905
(Medical Condition_Hypertension)	(Admission Type_Elective)	0.058036	0.348405	1.036529
(Medical Condition_Obesity)	(Admission Type_Emergency)	0.056324	0.338642	1.028770
(Medical Condition_Obesity)	(Billing catagory_Medium)	0.057009	0.342758	1.028274
(Medical Condition_Cancer)	(Billing catagory_Low)	0.056685	0.340956	1.022868
(Age Group_Middle Age)	(Admission Type_Elective)	0.076721	0.342917	1.020203

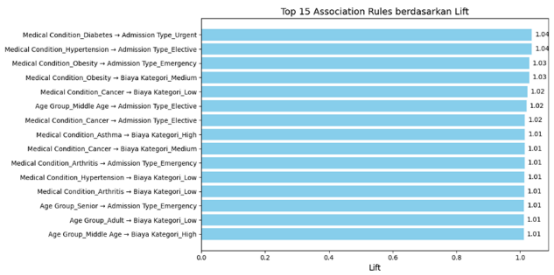
(Medical Condition_Cancer)	(Admission Type_Elective)	0.056721	0.341173	1.015014
(Medical Condition_Asthma)	(Billing catagory_High)	0.055982	0.338269	1.014807
(Medical Condition_Cancer)	(Billing catagory_Medium)	0.056216	0.338138	1.014414
(Medical Condition_Arthritis)	(Admission Type_Emergency)	0.056000	0.333906	1.014385
(Medical Condition_Asthma)	(Admission Type_Elective)	0.054321	0.332100	1.013452
(Medical Condition_Hypertension)	(Billing catagory_Low)	0.057876	0.330512	1.012384
(Age Group_Senior)	(Admission Type_Urgent)	0.055910	0.329800	1.011920
(Medical Condition_Diabetes)	(Billing category_Medium)	0.057210	0.328900	1.011102
(Medical Condition_Arthritis)	(Billing catagory_Low)	0.054910	0.327500	1.010805

These rules reveal potentially important patterns. For instance, patients with diabetes tend to be admitted through Admission Type: Urgent, while obese patients show a strong correlation with Medium billing categories and Admission Type: Emergency. The Middle Age group exhibits a significant relationship with Elective admissions. Moreover, conditions such as asthma, arthritis, and cancer also demonstrate consistent associations with specific types of care and cost levels.

A lift value > 1 in all these rules indicates the presence of a positive association between the antecedents and consequents. However, since the values are relatively close to 1, the strength of the associations can be considered weak to moderate. Therefore, these rules are more suitable for use as initial insights or segmentation mapping, rather than for automated predictive modelling.

3.4. Visualization of Association Rules

To clarify the relationships among items, the association rules were visualized in two formats:

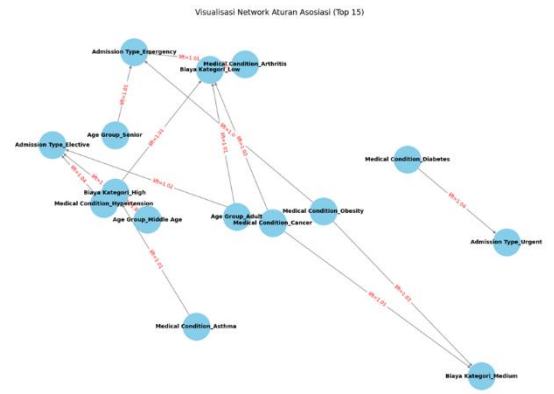


Gambar 3.1: Bar Chart of the Top 15 Association Rules Based on Lift

This bar chart displays the top 15 association rules generated by the Apriori algorithm, sorted according to their lift values. The X-axis represents the association rule labels, typically formatted as a combination of the antecedent and consequent (e.g., Medical Condition_Obesity → Admission Type_Emergency), while the Y-axis represents the lift value of each rule.

A lift value greater than 1 generally indicates that the antecedent and consequent co-occur more frequently than would be expected by random chance. However, in this study, the lift values fall within a narrow range of 1.01–1.03, which reflects extremely weak associations. In practical terms, a lift of 1.03 suggests that the co-occurrence is only about 3% higher than random expectation, which is typically considered statistically insignificant in medical or clinical data mining contexts. Therefore, these rules should not be interpreted as evidence of meaningful or actionable clinical relationships.

Nonetheless, visualizing these rules remains useful for exploratory and methodological purposes, particularly within the controlled environment of synthetic data. The ranking of lift values highlights which combinations appear slightly more frequently than others and helps illustrate how the Apriori algorithm behaves when applied to a structured healthcare dataset. These insights may serve as a preliminary reference for segmentation or pattern-mapping, while recognizing that stronger and clinically reliable associations would require validation using real-world patient data.



Gambar 3.2: Network Graph Association Rules

This graph depicts the association rules in the form of a network, where nodes represent individual items, such as Medical Condition_Cancer or Billing Category_Medium, and the edges indicate the direction of the association from the antecedent to the consequent.

The size of each node reflects the frequency of the item within the dataset, meaning that larger nodes correspond to items that occur more frequently. The level of connectivity, or degree, demonstrates the central role an item holds within the discovered rules, as items that appear across numerous rules will display many incoming and outgoing arrows.

This information illustrates the structure of interrelationships among attributes in a visual and intuitive manner, revealing key items such as Medical Condition_Cancer or Obesity that are associated with multiple other attributes.

These central attributes may serve as strategic starting points in formulating clinical or administrative decisions.

3.5. Discussion

The results of this analysis align with the study title, "Analysis of Multivariable Relationship Patterns in Synthetic Healthcare Data Using the Apriori Algorithm," as it successfully identified multivariate relationship patterns among healthcare attributes through data mining techniques. The findings indicate that certain medical conditions, particularly Diabetes, Hypertension, and Obesity, tend to be associated with specific care pathways such as Emergency and Elective admissions.

The analysis also reveals a tendency for patients with particular conditions to fall within a consistent range of care costs; for instance, patients with Asthma commonly fall into the high-cost category. Additionally, age appears to influence the type of healthcare services received, where the Middle Age group is more frequently associated with Elective procedures.

The visualizations further demonstrate that several attributes hold a central role in forming relationships among items, offering potential value as a foundation for medical and administrative decision-making.

However, the lift values obtained were relatively close to 1, indicating that the discovered associations possess weak to moderate predictive strength. To generate more meaningful insights, future research could explore alternative approaches, such as employing the FP-Growth algorithm or incorporating larger and more diverse real-world datasets.

3.6. Preliminary Conclusions

This chapter demonstrates that the Apriori algorithm is capable of effectively exploring relationships among attributes in healthcare data. Visualizations help strengthen the understanding of the discovered rules. These findings can serve as a foundation for the development of decision support systems or patient segmentation in data-driven hospital management.

4. CONCLUSION

Based on the results obtained, this study demonstrates that the Apriori algorithm is capable of identifying significant association patterns among variables in the healthcare dataset. The analysis shows that certain medical conditions, such as Diabetes, Hypertension, and Obesity, tend to be associated with specific types of hospital admissions, including Urgent, Elective, and Emergency categories. Hospitalization cost groups Low, Medium, and High also exhibit consistent relationships with several medical conditions as well as particular age groups, such as Middle Age and Senior patients. The lift values generated in this study range from 1.01 to 1.03, indicating positive but relatively weak correlations. In addition, visualizations in the form of bar charts and network graphs help clarify the relationships among items and enhance understanding of the underlying data structure.

5. GENERAL PROVISION

This study provides several recommendations for further development. First, it is recommended to apply the Apriori algorithm to real-world healthcare datasets to ensure that the analysis results are more applicable and can be directly implemented. Future research will validate the Apriori-generated patterns using actual hospital datasets, subject to ethical approval and anonymization procedures. In addition, experimenting with different support and confidence values can be used to generate rules that are more selective and specific. Future studies may also compare the Apriori algorithm with FP-Growth or Eclat to evaluate their effectiveness and efficiency. Furthermore, the analysis results can be integrated into decision support systems to assist with patient segmentation, medical needs prediction, and budget management. Finally, enriching the variables—such as length of stay, gender, or insurance type—can provide deeper insights into healthcare data exploration.

By considering these recommendations, it is expected that research in healthcare data analytics using association algorithms can advance and provide a tangible impact on improving the quality of healthcare services.

REFERENCES

Al-Sarayrah, A. (2024). RECENT ADVANCES AND APPLICATIONS OF APRIORI ALGORITHM IN EXPLORING INSIGHTS FROM

HEALTHCARE DATA PATTERNS. PatternIQ Mining, 1(2). doi: 10.70023/piqm24123

Ardiansyah, A., Zy, A. T., & Nugroho, A. (2023). Ciptaan disebarluaskan di bawah Lisensi Creative Commons Atribusi 4.0 Internasional. IMPLEMENTASI DATA MINING ALGORITMA APRIORI PADA SISTEM PERSEDIAAN OBAT (STUDI KASUS KLINIK PRATAMA KELUARGA KESEHATAN). Journal of Information System, Applied, Management, Accounting and Research, 7(3), 2598–8700. doi: 10.52362/jisamar.v7i3.1163

Darmawan, I. A., Randy, M. F., Yunianto, I., Mutoffar, M. M., & Salis, M. T. P. (2022). PENERAPAN DATA MINING MENGGUNAKAN ALGORITMA APRIORI UNTUK MENENTUKAN POLA GOLONGAN PENYANDANG MASALAH KESEJAHTERAAN SOSIAL. Sebatik, 26(1), 223–230. doi: 10.46984/sebatik.v26i1.1622

Febrianny Ulfha, N., & Amin, R. (2020). IMPLEMENTASI DATA MINING UNTUK MENGETAHUI POLA PEMBELIAN OBAT MENGGUNAKAN ALGORITMA APRIORI. 17(2), 396–402. Retrieved from https://journal.unpak.ac.id/index.php/komputasi

Hadinata, N., Ilmu Komputer Universitas Bina Darma, F., Jenderal Yani No, J. A., & Selatan, S. (n.d.). Analisis Pola Pembelian Produk Makanan Ringan Menggunakan Algoritma Apriori. 09. doi: 10.32736/sisfokom.V9.I1.623

Harahap, N. A., Novelan, M. S., Rambe, S. M., Syahri, R., & Datin, M. V. (2025). Analisis Performa Algoritma A* untuk Optimasi Penjadwalan Janji Temu Dokter di Rumah Sakit. Jurnal Manajemen Informatika, Sistem Informasi Dan Teknologi Komputer (JUMISTIK), 4(1), 462–468. doi: 10.70247/jumistik.v4i1.157

Nawangsih, I., & Purnamasari, P. (2023). Analisis Pola Pembelian Produk Kecantikan Menggunakan Algoritma Apriori. Jurnal Teknologi Informatika Dan Komputer, 9(1), 537–546. doi: 10.37012/jtik.v9i1.1614

Nola Ritha, Suswaini, E., & Pebriadi, W. (2021). Penerapan Association Rule Menggunakan Algoritma Apriori Pada Poliklinik Penyakit Dalam (Studi Kasus: Rumah Sakit Umum Daerah Bintan). Jurnal Sains Dan Informatika, 7(2), 222–230. doi: 10.34128/jsi.v7i2.329

Novianti, A., & Elisa, E. (2020). Penentuan Aturan Asosiasi Pola Pembelian Pada Minimarket Dengan Algoritma Apriori. Technology and Science (BITS, 2(1).

Parinduri, R. D., Defit, S., & Nurcahyo, G. W. (2024). Implementasi Algoritma Apriori dalam Data Mining untuk Optimalisasi Stok Obat di Apotik. Jurnal KomtekInfo, 89–97. doi: 10.35134/komtekinfo.v11i3.544

Pendidikan, J., & Konseling, D. (2017). Analisis Data Transaksi Penjualan Menggunakan Algoritma Apriori untuk Menentukan Paket Variasi Mobil (Studi Kasus: Bengkel Mobil Victory) (Vol. 4).

Putera Utama Siahaan, A., Azizah Harahap, N., Yuni Simanullang, R., & Wanny, P. (2025). Analysis of Inpatient Data Using Cluster Analysis on Simulation Dataset. Bulletin of Information Technology (BIT), 6(1), 33–39. doi: 10.47065/bit.v5i2.1830

Rahmi, A. N., & Mikola, A. (2021). IMPLEMENTASI ALGORITMA APRIORI UNTUK MENENTUKAN POLA PEMBELIAN PADA CUSTOMER (STUDI KASUS : TOKO BAKOEL SEMBAKO).

Saputra, R., & Sibarani, A. J. P. (2020). Implementasi Data Mining Menggunakan Algoritma Apriori Untuk Meningkatkan Pola Penjualan Obat (Vol. 7, Issue 2). Retrieved from http://jurnal.mdp.ac.id