Open Access **Baghdad Science Journal** **P-ISSN: 2078-8665**
Published Online First: March, 2023 **E-ISSN: 2411-7986**

DOI: https://dx.doi.org/10.21123/bsj.2023.7571

A Comparative Study on Association Rule Mining Algorithms on the Hospital Infection Control Dataset

Yahya Asmar Zakur [1*] ***Seyed Bagher Mirashrafi*** [1] ***Laith Rezouki Flaih*** [2]

[1] Department of Statistics, Faculty of Mathematical Science, University of Mazandaran, Mazandaran, Iran.
[2] Department of computer, College of Sciences, Cihan University, Erbil, Iraq.
*Corresponding author: y.zakur@stu.umz.ac.ir *
E-mails address: b.ashrafi@umz.ac.ir , Laith.flaih@cihanuniversity.edu.iq.

Received 19/6/2022, Revised 25/12/2022, Accepted 26/12/2022, Published Online First 20/3/2023

Abstract:

Administrative procedures in various organizations produce numerous crucial records and data. These records and data are also used in other processes like customer relationship management and accounting operations.It is incredibly challenging to use and extract valuable and meaningful information from these data and records because they are frequently enormous and continuously growing in size and complexity.Data mining is the act of sorting through large data sets to find patterns and relationships that might aid in the data analysis process of resolving business issues. Using data mining techniques, enterprises can forecast future trends and make better business decisions.The Apriori algorithm has been introduced to calculate the association rules between objects; the primary goal of this algorithm is to establish an association rule between various things. The association rule describes how two or more objects are related.We have employed the Apriori property and Apriori Mlxtend algorithms in this study and we applied them on the hospital database; and, by using python coding, the results showed that the performance of Apriori Mlxtend was faster, and it was 0.38622, and the Apriori property algorithm was 0.090909. That means the Apriori Mlxtend was better than the Apriori property algorithm.

Keywords: Apriori Mlxtend, Apriori Property, Association Rule Mining, Hospital Readmission, Machine learning, Performance of Algorithms.

Introduction:

This study's main objective was to compare the two types of this algorithm and identify the one that performed the best so that it could be suggested for use in future studies. Many studies have focused on the work of the algorithm and ways to develop it in a variety of fields, including the field of health. In addition to being utilized to examine patient data at the hospital being investigated. The main idea of the study is to discover the data in order to find new knowledge that helps in the process of reducing and controlling infection transmission in the hospital and determining the main cause for entering the hospital, through the use of data mining methods in building a system capable of controlling the health reality, also for determining the best algorithm that did this duty very well. The healthcare organizations as hospitals or governmental health care sectors or even health insurance companies possess rich data sources, such as electronic medical records, administrative reports, and other benchmarking results. Today, data mining in healthcare is used mainly for predicting various diseases, supporting biomedical and clinical diagnosis, advising doctors in making appropriate treatment decisions and avoiding accidental risky events from medical or organizational errors. Additionally, the advantages of data mining are much more than these, as it can provide question-based answers, anomaly-based discoveries, provide more informed decisions, probability measures, predictive modeling, and finally human decision support [1]. Using data mining, healthcare providers can be very effective in such fields as medical research, pharmaceuticals, medical devices, genetics, hospital management, and health care insurance, etc.

One of the most common themes in analyzing complex data is the classification, or categorization, of elements, the task is to classify a given data

instance into a pre-specified set of categories, the classification work by given a set of categories (subjects, topics) and a collection of text documents, the process of finding the correct topic (or topics) for each document [2]. Classification problem occurs when an object needs to be assigned into a predefined group or classified based number of observed attribute related to that object, many ideas have emerged over the years on how to achieve quality results from web classification systems, thus there are different approaches that can be used to a degree such as Clustering, Naïve Bays (NB) and Bayesian Networks, Neural Networks (NNs), Decision Trees (DTs), Support Vector Machines (SVM) etc.[3].

Data Mining for Hospital Readmission has formed a branch of applied artificial intelligence which allows a search of valuable information, especially in large volumes of data. The growing number of databases has created the need to have technologies that intelligently utilize the information and knowledge, thus making data mining an increasingly important research area [4] . Likewise, data mining has been extensively used in healthcare problems due to the increasing amount of data in healthcare systems, especially in this digital era. The interest in hospital readmission rates is growing worldwide, contributing to the growing research in hospital readmissions, such as identifying the risk factors or predictors which led to readmission and predicting the readmission risks based on various related areas through statistics, machine education, and data mining [5].

Association rule mining [6- 7], is a rule-based data mining technique [8- 9] for finding interesting relationships between things in large datasets. The method works by starting with frequently occurring item sets to find the rules. These rules are presented as (if → then) statements that provide information about item association likelihood. In market-basket analysis, for example, knowing that if a customer buys bread, they are likely to buy cheese is applicable [10]- [11]. Given an item's collections and transactions, each containing a subset of the items, determine the total number of items. The definition of an association rule is the implication A→ B, where A and B are the subsets and A∩B = 0. The antecedent left-hand side (LHS) and consequent right-hand side (RHS) of the rule are the sets of items for the short item sets (A, B). Association rule mining is the procedure of extracting rules from a given database that satisfy the stated minimal support and confidence requirements [12]. support is the frequency of the set (A∪B) in the dataset.

In contrast, Confidence is the conditional probability of finding B after discovering A and is calculated as (A∪B)/sup (A). The standard item set includes items that provide the bare minimum of Confidence and Support. The model that can determine the coronavirus infection as positive was constructed based on various pre-defined standard symptoms. The World Health Organization (WHO) and India's Ministry of Health and Family Welfare released guidelines for these symptoms. The system provided the severe symptoms of the diseases in this model. It allows users to talk about their symptoms, with the computer estimating a disease based on facts. To obtain the most accurate results, this factual information is subsequently evaluated using the ARM-based Apriori algorithm. Other traditional models, such as Support Vector Machine (SVM), Artificial Neural Networks (ANNs), and Random Forests (RF), have been studied and analyzed, and the suggested technique predicts a better accuracy score [13].

Epidemics continue to be a public health problem across the world. Even with technological advancements, there are still challenges to predicting infections.

To undertake effective surveillance and investigation of water-borne illnesses through social media with next-generation data, they proposed FREEDOM (Effective Surveillance and Investigation of Water-borne Diseases from Data-centric Networking Using Machine Learning). They acquired data from Twitter, preprocessed the tweet content, performed hierarchical spectral clustering, and created the frequent word set from each cluster to use the apriori process in the suggested model. Eventually, human behavior was implemented to derive inferences from the frequent word set. The support and confidence values of the outcome derived from the Apriori algorithm exhibited the different water-borne diseases that were not listed in the WHO (World Health Organization), and the surveillance of those diseases with percentage ranking and was achieved using data-centric networking, according to the experimental results. Again, it was matched with the real data. This form of study enables doctors and health organizations (government sector) to maintain track of water-borne infections and associated symptoms for early detection and safe recovery, decreasing the mortality toll significantly [14].

Also, proposed a method for assessing the behavior of groups of people and demonstrated how to forecast person location for the next months[15]. In this work, clustering technique was used. The challenge of discovering associative rules was also investigated. To discover the optimal rules, they employed scalable Apriori algorithms. We utilized the standard mlxtend package to aggregate data by cluster, user login, and time for analysis. They

looked at the apriori and k-means clustering methods to generate a template for user pattern recognition. They investigated the issue of finding associative rules that could detect and describing patterns in huge amounts of information. The best rules were determined through scalable Apriori algorithms. They used the standard mlxtend package to aggregate data by cluster, user login, and time for analysis. They were confronted with the problem of data inaccuracy and consistency with real-world settings while working, and were obliged to limit the minimal support for associative rules [16]. The Apriori approach used to create association rules using data obtained during emergency department visits, either through triage or electronic health records. The purpose of this study was to discover the link between various major complaints or illnesses stated by patients during triage examination, as well as the relationship between different main complaints and prior medical history or outpatient medication. The initial set of results was assessed and presented. In this work, the association rules and Apriori algorithm were created using the Python 3.7.4 programming language and the Mlxtend library. Extracting meaningful information from medical data was a difficult but beneficial process. In this study, they looked at the association between the many primary complaints provided by patients, the relationship between prior and current illnesses, and the relationship between reported complaints and outpatient medicines. The Apriori technique was used to derive the association rules from a real dataset of visits to emergency rooms. This study offered critical info on the correlations between various data elements from a large quantity of patient data, allowing medical workers to look for other related concerns that the patient may not have reported[16]. This study intended to investigate disease association for other patient categories in the future, as well as incorporating other factors. The mining association rules procedure consists of two sections: All standard item sets must be made from scratch to begin. Creating robust association rules utilizing the collections of commonly used items. The most extensively used machine learning algorithm for mining association rules is Apriori. In this paper, two types of Apriori have been used and compared to find the best one in the context of performance[17- 18].

Materials and Methods:

1- Apriori Property Algorithm:

The Apriori property [19] all non-empty subsets of the frequent itemset are frequent as the algorithm's central assumption. This attribute is referred to as the Apriori property. Thus, if an item set is rare, all of its supersets would also be rare. Due to the database's extensive quantity of unique items, the search space for all item sets is exponential. Consequently, the fundamental procedure for constructing and calculating the frequency of all item sets across the database will be unnecessarily time-consuming.

Moreover, the database in issue could be huge, including millions of transactions, making frequency counting a challenging operation in and of itself. So, it will be using the Apriori property [20]. Every non-empty subset of the frequent itemset must likewise be regular [21]. The Apriori property depends on the following notation. If an itemset G does not satisfy the minimum support threshold, min sup, then G is not frequent, that is, P(G) < min-sup. If item A is added to the itemset G, then the resulting itemset (i.e., G∪A) cannot occur more frequently than G. Therefore, G ∪A is not frequent also, P (G ∪A) < min sup. This property belongs to the anti-monotonicity category of qualities [22]; Likewise, if a set fails a test, all its supersets will also yield the same test. Due to the monotony of the property in the context of failing [23], It is known as anti-monotony. A two-step strategy, including the join and prune operations, is utilized to comprehend how the algorithm operates.

In the join step, L_{k-1} joins with itself if the two item sets of L_{k-1} have the same initial (k-2) items and the first itemset's (k-1) th thing is lexicographically less than the relevant item of the second itemset. It uses the Apriori property. The pruning step reduces the size of C_k. Using the Apriori property, the prune step minimizes the size of C_k. Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset, according to the Apriori property [24]. The Apriori approach requires a lot of processing to join and prune item sets and check a portion of each transaction against candidates. Many candidates also need a large amount of memory throughout the algorithm's execution. Hence, an effective data structure is required, reducing processing costs while efficiently organizing candidate item sets in memory space. The Apriori property's main benefit is that it minimizes the number of candidate item sets, lowering the algorithm's length and time complexity. On the other hand, the number of potential item sets generated may be too vast for the main memory to process it [25]. Also, there are some indicators like as below [26]:

1- "Support represents the chance that the next attribute value X of a dimension appears in every record (R), as illustrated in the formula below."

"Support(X) = (Transactions containing X) / (Total Transactions)":

"Support (X)= count(X)/count (R) = P(X)".

In general, the support is a percentage of the item's appearance out of the total items.

2- As demonstrated in the following formula, Confidence represents the likelihood that attributes X and Y appear concurrently in all records R.

Confidence (X→ Y) = Support (X∪Y)/ Support(Y)=P(Y\X)

Interpreted as: How often items in Y appear in transactions that contain Y only.

3- C_K denotes the candidate item set, which is the item set retrieved via downward merging. Where (K) represents the number of elements

Frequent itemsets are those having Support greater than or equal to the stipulated minimum support, represented by L_K. Any non-empty subset of a frequent itemset is, in fact, a frequent itemset.

2- Mlxtend Apriori Algorithm Implementation [27]:

Using the Apriori algorithm, which is dependent on the Mlxtend library, the following will be applied in general:

The First Step is to Find the Itemset.

The second step is to create a K-Itemset.

The third step is to check the Support.

The fourth step is to look for items that are used frequently.

The 5th Step: The K-1 Frequent Item Set is #?

If "No," proceed to Step 2.

Otherwise, proceed to the next step.

The sixth step is to come to a complete stop (End).

Join step: C_k generating by doing a joining for L_{k-1} with itself.

Prune step: Any of the (k-1)-item sets that are not frequent are not eligible to be a subset of frequent (k-itemset)

- C_k: Size of k candidate itemset
- L_k Size of k frequent itemset

The phases of the Apriori algorithm will then be as follows in Fig.1:

L_1= (frequent itemset),
For ($k=1:l_k!=\emptyset;k++$) do start
C_{k+1}= candidates generated from the L_k;
For each transaction t in the database do
Increment the count of all of the candidates in the C_{k+1}
That are contained in the t
L_{k+1}= Candidates in the C_{k+1} with the min-support
End
Back to $\cup_k L_k$;

Figure 1. The phases of the Apriori algorithm.

- Support (A → C) = Support (A ∪ C), the range will be [0, 1]

Instead of association rules, the support measure is applied to item sets. The association rule mining approach produces a table with three support measurements: "antecedent support," "consequent support," and "support."

The proportion of transactions including the antecedent (A) is calculated as antecedent support, whereas the Support for the itemset of the consequent is calculated as consequent Support (C). The Support for the combined item sets (AC) is then computed using the support measure. The Support is contingent on "antecedent support and subsequent support" as measured by minimum support levels (antecedent support, consequent Support). Typically, Support is used to estimate the quantity or frequency of database entries (which is often considered relevant or essential). It calls itemset a "frequent itemset" if you support more than a minimum-support criterion. It's worth noting that, according to the downward closure property, all subsets of a frequent itemset are also frequent.

- Confidence: $A \rightarrow C = \frac{Support\ \ A \rightarrow C}{Support\ A}$, the range [0,1].

The probability of seeing the consequent in a transaction if it contains the antecedent is the Confidence of rule A->C. It is worth noting that the metric is not symmetric or directed; for example, the Confidence for A->C differs from that for C->A. If the consequent and antecedent always occur together, the Confidence is 1 (maximal) for a rule A-> C.

- Lift $A \rightarrow C = \frac{(\text{Support } A \rightarrow C)}{(\text{Support } C)}$, the range = [0, ∞].

The lift measure is used for determining how often the antecedent and consequence of a rule A→C occur together than we would expect if they were statistically independent. If A and C are independent, the Lift score will be exactly 1[28].

Baghdad Science Journal
P-ISSN: 2078-8665
E-ISSN: 2411-7986

Patients and Control

This section tries to identify which of the two algorithms implemented in this study provides the best performance and examines the medical dataset in search of factors that will aid physicians in making the best decision possible. The dataset used in this study was gathered in 2021 from Al-Kut Hospital in Wasit Governorate, Iraq, and is an accurate data set licensed only for scientific research. The dataset contained 10,160 patient samples.

Results:

The First Method: Apriori Property Algorithm:

It is possible to determine the Apriori property's performance; the Support for the Apriori property algorithm was 0.091, that shows the performance of this algorithm is always same over the time, as shown in Fig.2:

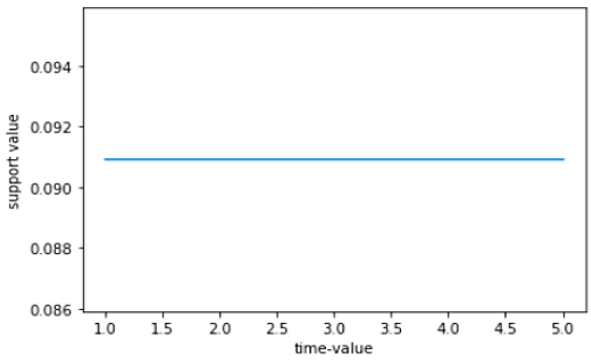

Figure 2. The performance of the Apriori property algorithm.

The Second Method: Apriori Mlxtend:

The performance of the Mlxtend-based apriori algorithm will be determined by computing the method's support values over time. As depicted in Fig.3, the minimum Support was 0.25, while the maximum Support was approximately 0.6., which means the performance continues improved over the time.

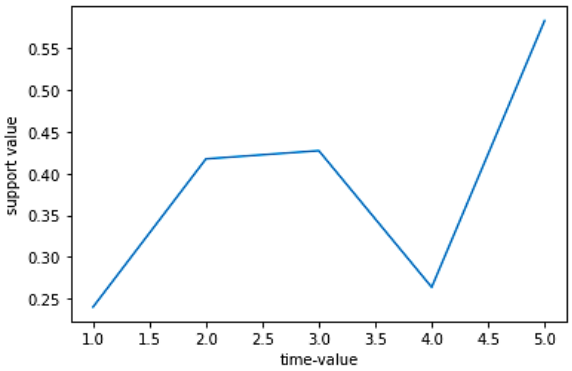

Figure 3. The performance of the Apriori - Mlxtend algorithm.

In general, and for giving more Support and power to the studying, and by taking the average for the Support, Confidence, and lift, the results will be divided as below in Table. 1, which shows the compartion of performance between Apriori Property and Apriori Mixtend, based on some factors like (Average of support, Average of Confidence, Average of Lift and Time of Running), and calculating the average by using the Eq below:

$$\text{The average} = \frac{\text{Summation of the items}}{\text{The total number of the items}}$$

Table 1. The comparison between the algorithms.

Algorithms	The factors			
	Average of Support	**Average of Confidence**	**Average of Lift**	**Time of Running**
Apriori Property	0.090909	0.775362	6.137681	Medium
Apriori Mlxtend	0.38622	-	-	Fast

In addition, Fig.4, which displays the performance of algorithms, can be observed by averaging the support values for the Apriori property and Apriori Mlxtend algorithms.

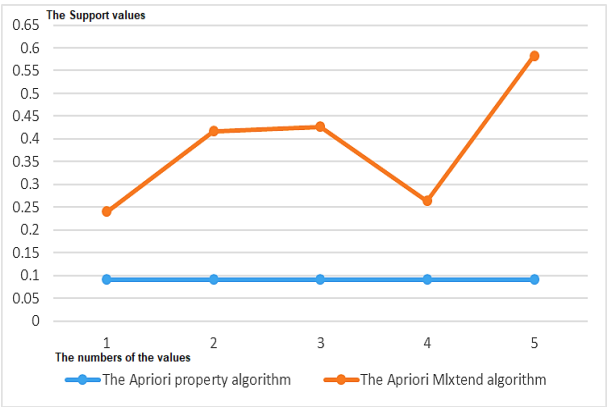

Figure 4. The Support for algorithms (Apriori property and Apriori Mlxtend).

Baghdad Science Journal
P-ISSN: 2078-8665
E-ISSN: 2411-7986

The Fig. 5, which displays the employing of the support values averages from the two methods, as below:

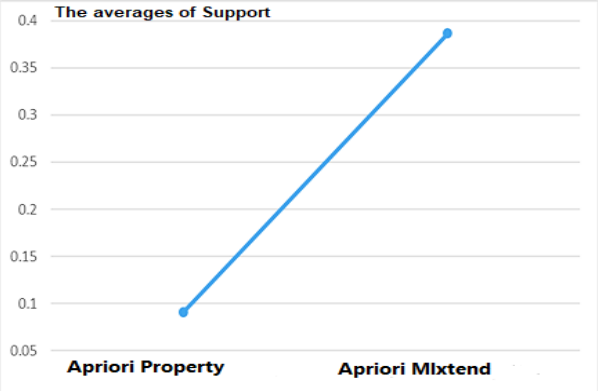

Fig 5. The averages of Support for algorithms (Apriori property and Apriori Mlxtend).

Some of the Medical Indicators:

Effectiveness of the Diseases:

Applying the Data mining techniques to discover the reasons for entering the hospital, it is clear that: The pandemic of covid 19 was the leading cause of hospitalization, followed by pneumonia and the remainder of the disorders, as the Fig.6.

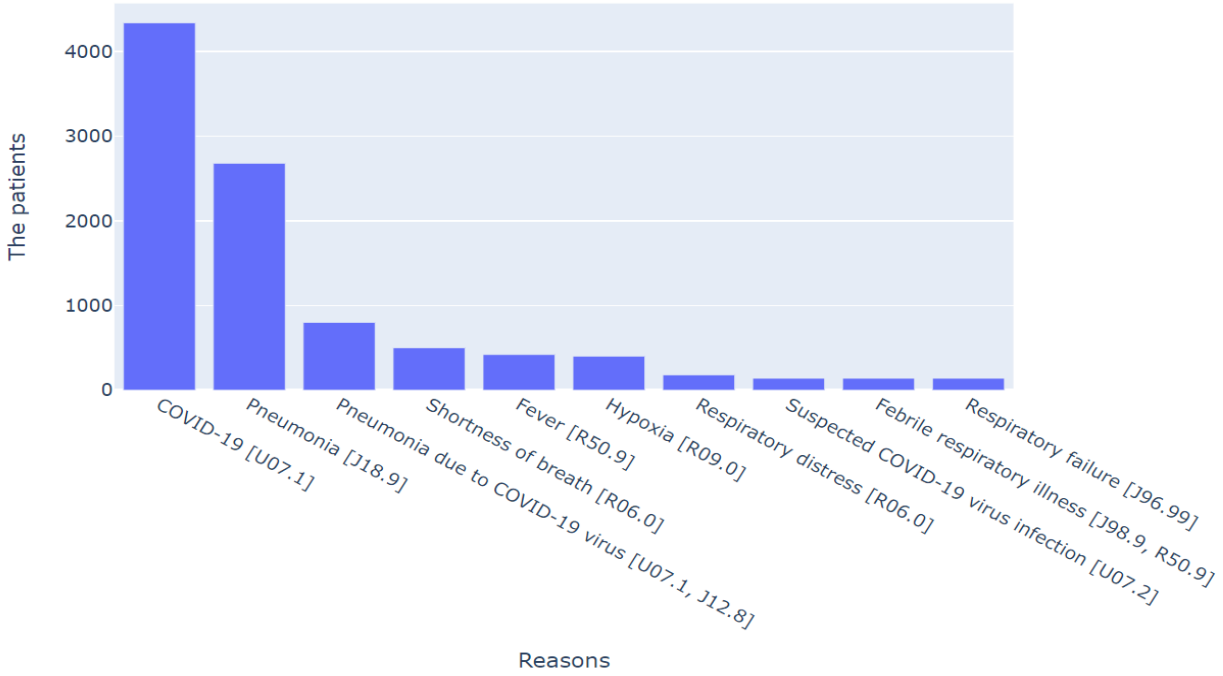

Figure 6. Some of the reasons that cause hospital Readmission.

As for the ages of the hospital patients, it is evident that they were between 20 and 100 years old when they were admitted. The median was 60 years, and the average was 60 years, as in Fig.7.

Baghdad Science Journal
P-ISSN: 2078-8665
E-ISSN: 2411-7986

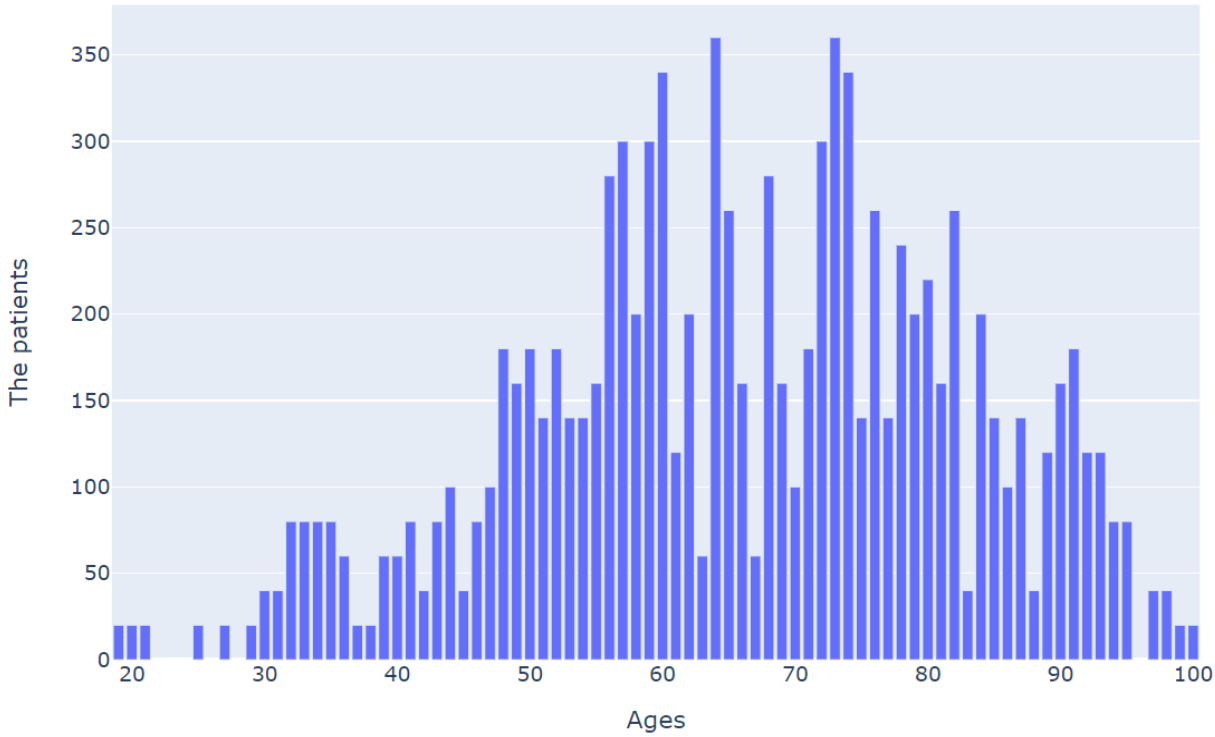

Figure 7. The ages of patients in the hospital.

The Genders of the Patients and the Old Visit:

The genders of the paitants can be decleared as below: the number of males admitted to the hospital is more than females. as shown in Fig.8.

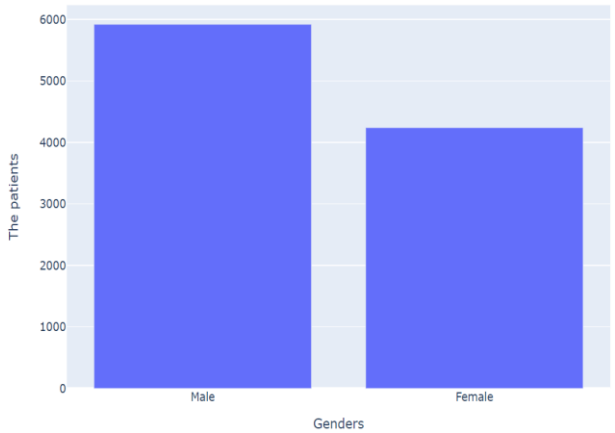

Figure 8. The genders of the patients in the hospital.

The number of patients who have never been admitted is more significant than the number of patients who have previously been recognized, as below in Fig.9.

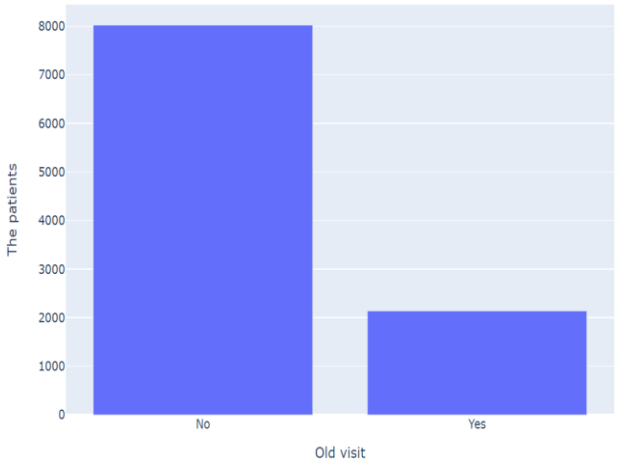

Figure 9. The past –old- visit for the patients in the hospital.

The Types of Admission:

More patients were admitted to the ward than the intensive care unit, as Fig.10.

Baghdad Science Journal
P-ISSN: 2078-8665
E-ISSN: 2411-7986

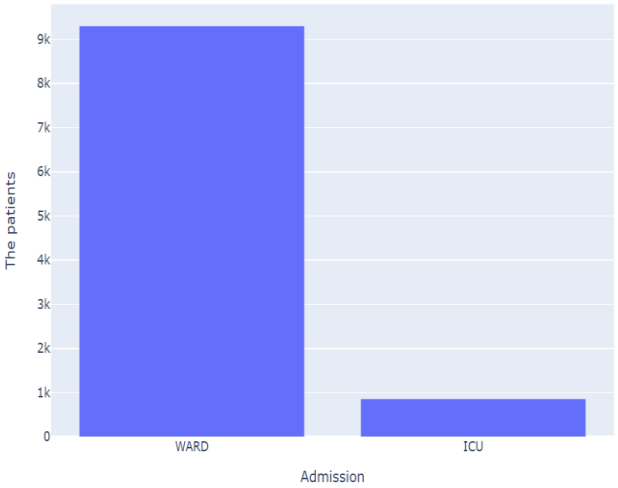

Figure 10. The types of admission for patients in the hospital.

More patient's heart rates ranged between 60 -160, with an average of 110 and a median of 110, as Fig.11.

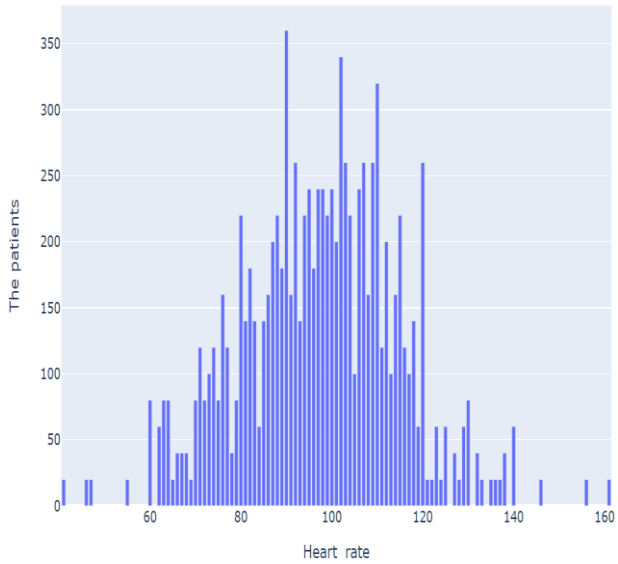

Figure 11. The heart rate for patients in the hospital.

The Respiratory Rate for Patients:

The patient's respiration rate ranges between 15 and 55 breaths per minute. The average oxygen rate was 35 breaths per minute, the median was 35 breaths per minute, see Fig.12.

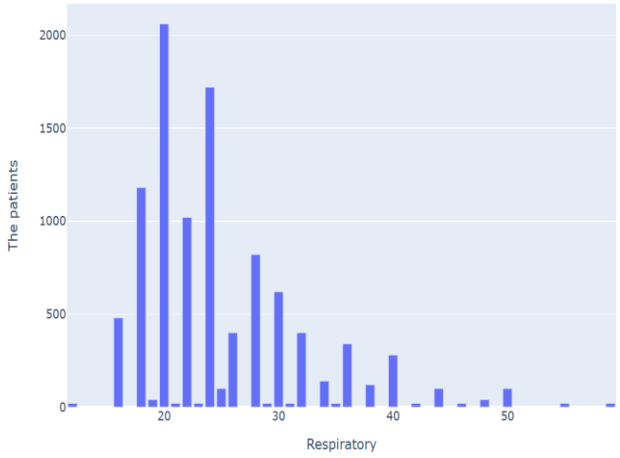

Figure 12. The respiratory rate for patients in the hospital.

The patients had oxygen rates ranging from 50 to 100 breaths per minute. The average rate per minute was 75. And the median was 75 per minute, see Fig.13.

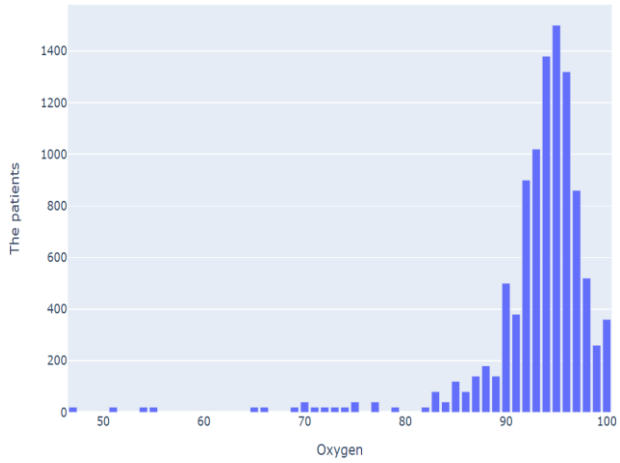

Figure 13. The oxygen rates for patients in the hospital.

The Intubated Patients:

The non-intubated patients are more than intubated. See Fig.14.

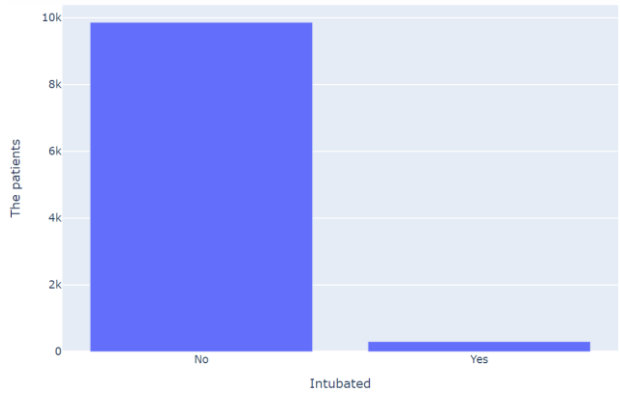

Figure 14. The intubated rate for patients in the hospital.

Discussion:

Data mining techniques find relationships, patterns, and rules, allowing professionals and specialists to discover unexpected and significant data. Consequently, students will have a deeper comprehension of systems and procedures. This data can then be used to construct new processes and judgments based on existing Knowledge, as well as modules and expert techniques that can manage and optimize systems in diverse sectors of life, including medicine and healthcare. Professionals can also leverage these modules to optimize performance and resource utilization. The strategies described in this study may also be of benefit. The Apriori algorithm was applied using two ways in this study, and the following conclusions were reached: The Apriori property was utilized, with a support average of 0.090909, a confidence average of 0.775362, a lift average of 6.137681, and a running time of slower. The Apriori based on the Mlxtend library was also employed, with a support average of 0.38622 and a faster run time. The Apriori property did poorly overall, but the Apriori based on Mlxtend performed exceptionally well; hence the Apriori Mlxtend method was the best. Significant hospital infection-related traits, such as (reasons for admission, age, race, gender of the patient, previous visits, types of entry, heart rate, the breathing rate of patients, oxygen rate, and intubation for patients).

Conclusion:

Establishing information systems in the healthcare sectors to incorporate significant urgent cases, such as the emergence of critical and unanticipated circumstances of covid-19, will improve hospital infection management. Related to covid-19 and other disorders will assist lessen the impact of uncontrolled emergency cases. For instance, this study revealed that a more significant number of patients were admitted to the hospital due to covid-19. In addition, the patient's age, gender, past-visit status (yes or no), type of admission (ICU or ward), heart rate, breathing rate, oxygen saturation level, and intubation must be provided by the system. This discovery will assist professionals and physicians in making appropriate decisions based on this knowledge, which may reduce the incidence of hospital-acquired infections since these criteria will be critical in the future for decision-making by doctors and medical professionals during the comparison of the patient's diverse outcomes (prior and posterior results). Future work will benefit from incorporating critical situations, such as the unanticipated advent of covid-19, into the design of healthcare information systems to provide effective hospital infection management in connection to covid-19 and other diseases. In addition, they were employing another large dataset with a different data mining algorithm in the future to conduct additional research and comparisons.

Acknowledgment:

Many thanks to everyone who supported this work.

Authors' Declaration:

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript ours. Besides, the Figures and images, which are not ours, have been given the permission for re-publication attached with the manuscript.
- Authors sign on ethical consideration's approval.
- Ethical Clearance: The project was approved by the local ethical committee in University of Mazandaran.

Authors' Contributions Statement:

Y. A. Z. Contributed to design and implementation of the research, to the analysis of the results and to the writing of the manuscript. S. B. revised and proofread the manusprit, and L. R. F.contributed in the revision and proofreading.

References:

1. Koutsojannis, C. Medical Knowledge Extraction: Particular Difficulties And Obligations. 1st edition. Chap 3, NOVA Publishers .2020; 42-43.
2. Feldman R, Sanger J. The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge university press. 2019. 34(1): 125-127
3. Muhammed M, Flaih L. Online Web Page Classification. Computer Engineering and Intelligent Systems. 2014; 5(1): 14-15. https://core.ac.uk/reader/234644747.
4. Silva M, Rodrigues O. Risk factors for surgical site infection: challenges to public health. J Microbiol Exp. 2019; 10(1): 1-8. https://doi.org/10.15406/jmen.2022.10.00345.
5. Kocakoç D, Türkölmez B. Using Data Mining Techniques for Designing Patient-Friendly Hospitals. In Advances in Econometrics, Operational Research, Data Science and Actuarial Studies. Springer, Cham. 2022; 321-343. https://doi.org/10.1007/978-3-030-85254-2_20.
6. Chen B, Huang Z, Liu C, Wu Z. Spatio-temporal data mining method for joint cracks in concrete dam based on association rules. Struct Control Health Monit. 2022; 29: 4-5. https://doi.org/10.1002/stc.2848.
7. Wardani D. Measuring Positive and Negative Association of Apriori Alg3orithm with Cosine Correlation Analysis. Baghdad Sci J. 2021; 18(3): 0554. https://doi.org/10.21123/bsj.2021.18.3.0554.

8. Ibrahim W, Abdullaev S, Alkattan H, Adelaja A, Subhi A. Development of a Model Using Data Mining Technique to Test, Predict and Obtain Knowledge from the Academics Results of Information Technology Students. Data. 2022; 1-2. https://doi.org/10.3390/data7050067.
9. Prasanthi N, Rao MVP. A Comprehensive Assessment of Privacy Preserving Data Mining Techniques. 2nd Int Conf Sustain Expert Syst.. 2022; 833- 842. https://doi.org/10.1007/978-981-16-7657-4_67.
10. Sastrt J, Suresh V. A Survey Paper on Frequent Itemset Mining. J Phys: Conf Sers. 2019; 1228(1): 1-7.
11. Abdulmajeed A, Tawfeeq T, Al-jawaherry A. Constructing a Software Tool for Detecting Face Mask-wearing by Machine Learning. Baghdad Sci J. 2022; 19(3): 0642. https://doi.org/10.21123/bsj.2022.19.3.
12. Dehbozorgi MR. Rastegar M, Sami A. Data mining-based cause identification of momentary outages in power distribution systems. Sustain Cities Soc. 2022; 77: 103587. https://doi.org/10.1016/j.scs.2021.103587.
13. Babu A, Raj, Varalatchoumy M, Gopila M, Justin F. Novel Approach for Predicting COVID-19 Symptoms using ARM based APRIORI Algorithm. International Conference on Computing Methodologies and Communication (ICCMC). IEEE. 2022; 1577-1580.https://doi.org/10.1109/ICCMC53470.2022.9753987.
14. Pradeepa S. Jaisaiarun P, Srinivasan P, Subbulakshmi V, Tarik A. FREEDOM Effective Surveillance and Investigation of Water-borne Diseases from Data-centric Networking Using Machine Learning Techniques. Int J Artif Intell Tools. 2022; 2250004. https://doi.org/10.1142/S021821302250004X.
15. Boyko, N, Komarnytska H, Kryvenchuk Y, Malynovskyy Y. Clustering Algorithms for Economic and Psychological Analysis of Human Behavior. In CMiGIN. 2019; 614-626.
16. Ivascu T, Cincar K, Carunta A. Extracting Association Rules from Emergency Department Data. The 7th IEEE International Conference on E-Health and Bioengineering. IEEE. 2019: 3-4.
17. Wang Y, Wu Y, Li Y, Yao F, Fournier-Viger P, Wu X. Self-adaptive nonoverlapping sequential pattern mining. Appl Intell. 2022; 52(6): 6646-6661. https://doi.org/10.1007/s10489-021-02763-y.
18. Raj S, Ramesh D, Sethi K. A Spark-based Apriori algorithm with reduced shuffle overhead. J Supercomput. 2021; *77*(1): 133-151. https://doi.org/10.1007/s11227-020-03253-7.
19. Datta S, Mali K, Das S, Kundu S, Harh S. Rhythmus periodic frequent pattern mining without periodicity threshold. J Ambient Intell Humaniz Comput. 2022; 1-13. https://doi.org/10.1007/s12652-021-03617-8.
20. Dinov D. (2023). Qualitative Learning Methods-Text Mining, Natural Language Processing, and Apriori Association Rules Learning. In Data Science and Predictive Analytics: Biomedical and Health Applications using R. Cham: Springer International Publishing. 2023; 385-437. https://Doi:10.1007/978-3-031-17483-4_7.
21. Tharini J, Shivakumar L. Computational Intelligence and Data Sciences. 1st edition. Florida, USA: CRC Press. Chap 11, High-Utility Itemset Mining: Fundamentals, Properties, Techniques and Research Scope. 2022; 195-210.
22. Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules. Proc. of the 20th Int Conf. on VLDB: Santiago, Chile. 1994; 487–499. https://doi/10.5555/645920.672836.
23. Zhao Z, Jian Z, Gaba S, Alroobaea R, Masud M, Rubaiee S. An improved association rule mining algorithm for large data. Int J Intell Syst. 2021; 30(1): 750-762. https://doi.org/10.1515/jisys-2020-0121.
24. Wu Y, Yuan Z, Li Y, Guo L, Fournier-Viger P, Wu X. NWP-Miner: Nonoverlapping weak-gap sequential pattern mining. Inf Sci. 2022; 588: 124-141. https://doi.org/10.1016/j.ins.2021.12.064.
25. Wang Y. Internet Medical Privacy Disclosure Mining and Prediction Model Construction Based on Association Rules. Teh. Vjesn. 2022; 29:1: 231-238. https://doi.org/10.17559/TV-20210903062509.
26. Das A, Jana S, Ganguly P, Chakraborty N. Application of Association Rule: Apriori Algorithm in E-Commerce. In 2021 Innovations in Energy Management and Renewable Resources (IEMRE). 2021; 1-7. https://doi.org/10.1109/IEMRE52042.2021.9386737.
27. Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. In Proc. of the 1993 ACM SIGMOD Int Conf. on Management of data. 1993; 207-216. https://doi.org/10.1145/170035.170072.
28. Ma H, Ding J, Liu M, Liu Y. Connections between Various Disorders: Combination Pattern Mining Using Apriori Algorithm Based on Diagnosis Information from Electronic Medical Records. Biomed Res Int. 2022; 2022: 1-16. https://doi.org/10.1155/2022/2199317.

Open Access **Baghdad Science Journal** **P-ISSN: 2078-8665**
Published Online First: March, 2023 **E-ISSN: 2411-7986**

دراسة مقارنة حول خوارزميات التعدين في قواعد الرابطة في مجموعة بيانات مكافحة العدوى في المستشفى

يحيى اسمر زاكور[1] **سيد باقر ميراشرفي[1]** **ليث رزوقي فليح[2]**

[1] قسم الاحصاء، كلية العلوم الرياضية، جامعة مازندران، مازندران، ايران.
[2] قسم الحاسوب، كلية العلوم، جامعة جيهان، أربيل، العراق.

الخلاصة:

إن العمليات الادارية في مختلف المؤسسات الحديثة ينتج عنها العديد من البيانات والمعلومات المهمة وباشكال مختلفة، ويمكن لهذه البيانات ان تستخدم في عمليات اخرى كالعمليات المحاسبية وادارة العلاقات مع الزبائن. إن استخلاص المعلومات المفيدة وذات الصلة يعد من التحديات الكبيرة خاصة مع ضخامة هذه البيانات وتعقيدها واشكالها المختلفة ونموها المستمر. لحل تلك المشكلة يمكن استخدام التنقيب عن البيانات (Data Mining)، وهي عمليات معقدة تجري على البيانات الكبيرة لغرض استخلاص معلومات وانماط مفيدة والكشف عن العلاقات بين تلك البيانات، والتي تساعد على حل مشكلات الاعمال من خلال تحليل البيانات، ومساعدة المنظمات على التنبؤ بالاتجاهات المستقبلية وبالتالي اتخاذ القرارات الافضل. لقد تم تقديم خوارزمية Apriori لغرض حساب قواعد الترابط بين الاشياء، إن الهدف الاساسي من هذه الخوارزمية هو تاسيس قواعد ترابط بين الاشياء المختلفة والتي تستخدم لوصف كيف ان شيئين او اكثر يرتبطون ببعضهم البعض.في هذه الدراسة تم توظيف نوعين من خوارزميات Apriori، وهي Apriori Property و Apriori Mlxtend، وتم تطبيقهم على قاعدة بيانات المستشفى، وباستخدام لغة بايثون تم الوصول الى ان سرعة خوارزمية Apriori Mlxtend كان 0.38622، بينما سرعة تنفيذ خوارزمية Apriori Property كان 0.090909، اي ان سرعة واداء Apriori Mlxtend كان افضل من سرعة واداء Apriori Property.

الكلمات المفتاحية: ابريوري ام ال اكس تيند، خاصية ابريوري، تعدين قواعد الرابطة، إعادة الدخول الى المستشفى، التعلم الألي، أداء الخوارزميات.