Research Paper

Classification of Survivors and Non-survivors of the Latest Epidemic Using Association Rules Algorithm

Nasrin Talkhi[1, 2], Nooshin Akbari sharak[1], Zahra Pasdar[3], Maryam Salari[4], Seyed Masoud Sadati[5], Mohammad Taghi Shakeri[6*]

1. Department of Biostatistics, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
2. Department of Biostatistics, School of Health, Mashhad University of Medical Sciences, Mashhad, Iran.
3. Institute of Applied Health Sciences, School of Medicine, Medical Sciences and Nutrition, Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, Scotland.
4. Expert Management and Information Technology, Mashhad University of Medical Sciences, Mashhad, Iran.
5. Center of Statistics and Information Technology Management, Imam Reza Hospital, Mashhad University of Medical Sciences, Mashhad, Iran.
6. Social Determinants of Health Research Center, Mashhad University of Medical Sciences, Mashhad, Iran.

Use your device to scan and read the article online

Citation Talkhi N, Akbari sharak N, Pasdar Z, Salari M, Sadati SM, Shakeri MT. Classification of Survivors and Non-survivors of the Latest Epidemic Using Association Rules Algorithm. Iranian Journal of Health Sciences. 2025; 13(2):155-166. http://dx.doi.org/10.32598/ijhs.13.2.1095.1

http://dx.doi.org/10.32598/ijhs.13.2.1095.1

ABSTRACT

Background and Purpose: Association rule mining can discover hidden patterns and relationships between variables that may not be apparent through other data analysis techniques. We aimed to find practical patterns in COVID-19 data and predict patient survivor status using association rules.

Materials and Methods: In this cross-sectional study, clinical data of 51460 hospitalized patients tested by polymerase chain reaction (PCR) were collected from February 20, 2020, to September 12, 2021, in Khorasan Razavi Province, Iran. An Apriori algorithm was used to extract association rules or patterns in data.

Results: Most participants (51.0%) were male; their Mean±SD age was 54.55±22.15 years. Fever (37%), cough (38.4%), respiratory distress (56%), PO_2 level less than 93% (52.9%), muscular pain (19.1%) and decreased consciousness (8.9%) were common symptoms. Based on the association rules, if a patient was older than 75 years, had respiratory distress, reduced consciousness and PO_2 level <93%, then this patient is who has died. The PCR test result of a male who used drugs was positive. Vomit and diarrhea lead to positive PCR test results, too. The most common symptom seen in men was respiratory distress, while the most common symptom in women was hypertension. Muscular pain due to COVID-19 is more common in women than men. Furthermore, the accuracy and area under the receiver operating characteristics curve were obtained as 92.28 and 86.80 on the testing dataset, respectively.

Conclusion: Simple methods such as association rules mining and complex methods could be helpful and give valuable results, and predicting death using association rules provides high accuracy.

Keywords: Apriori algorithm, Association rules mining, Associative classifiers, Classification-based association rule (CBA) algorithm, SARS-CoV-2

Article info:
Received: 03 Jul 2024
Accepted: 26 Mar 2025
Available Online: 01 Apr 2025

*** Corresponding Author:**
Mohammad Taghi Shakeri, Professor.
Address: *Social Determinants of Health Research Center, Mashhad University of Medical Sciences, Mashhad, Iran.*
Tel: *+98 (915) 5151799*
E-mail: *ShakeriMT@mums.ac.ir*

Introduction

The novel coronavirus disease 2019 (COVID-19), caused by the SARS-CoV-2 virus, was declared a pandemic by the World Health Organization (WHO) on March 11, 2020 [1, 2]. Due to its rapid spread worldwide, COVID-19 has posed a significant global threat, overwhelming health care systems and severely impacting economies. According to WHO reports, the epidemiological trends of COVID-19 vary across different regions, and the fatality rate differs between countries [3]. Moreover, the clinical and epidemiological characteristics of COVID-19 are heterogeneous and differ between survivors and non-survivors [4, 5]. Studies indicate that non-survivors tend to be older and predominantly male compared to survivors [3, 6]. The most common symptoms of COVID-19 include fever, dry cough, headache, myalgia, sore throat, loss of taste or smell, nausea and diarrhea [7-10]. Patients with a weakened immune system and underlying conditions, such as diabetes, hypertension, and cardiovascular diseases, are at a significantly higher risk of severe disease and mortality [11]. Given the diverse risk profiles among different patient groups, early diagnosis and effective management are crucial, particularly for those with a high probability of mortality [12].

Mathematical, statistical and machine learning methods have played a pivotal role in analyzing COVID-19 patient data and improving risk prediction. Data mining, in particular, provides various techniques, including classification, clustering, regression, association rule mining, and correlation analysis, that have been widely applied in health care research [13-15]. These techniques enable predictive modeling, identifying high-risk groups, optimizing resource allocation and enhancing treatment protocols. For instance, machine learning models have been employed to predict the spread of the virus, assess the effectiveness of interventions, and identify vulnerable populations. Additionally, natural language processing has been used to mine medical literature and clinical data, uncovering potential treatments, while clustering techniques have helped track viral mutations [16].

Among data mining techniques, association rule mining is beneficial for discovering hidden patterns in large datasets. The well-known rule-based Apriori algorithm identifies frequent item sets and extracts significant associations without requiring complex assumptions or extensive parameter tuning. Several studies have successfully applied Apriori to provide health care insights. For instance, Biilah et al. [17] Utilized the Apriori algorithm on a dataset of 539 students to identify key features associated with depression and stress, extracting 8 significant rules. Similarly, Ilayaraja and Meyyappan [18] applied Apriori to analyze disease frequency within specific geographical regions. Abdullah et al. [19] investigated similarities between medical and purchase bills using this approach, while Jena and Kamila [20] employed it to identify association rules in depression-related data.

While various machine learning and statistical models have been used to predict COVID-19 patient outcomes, many of these techniques rely on complex computational frameworks that require extensive data preprocessing, high computational resources, and specialized expertise. These methods, particularly deep learning and ensemble-based machine learning models, often function as “black-box” models, lacking interpretability for clinicians. A significant research gap exists in the lack of simple, rule-based methods to extract transparent, interpretable relationships between clinical features and patient outcomes.

The Apriori algorithm addresses this gap by generating easily understandable association rules from large datasets, making it an accessible tool for clinicians. Unlike regression models or deep learning algorithms that require statistical expertise for interpretation, Apriori produces human-readable decision rules that can directly inform medical decision-making and patient management.

This study proposes a classification-based association rule (CBA) approach, integrating Apriori-based rules to predict survivor and non-survivor cases in COVID-19 patients. To our knowledge, this specific application has not been previously investigated.

The findings of this study have significant implications for public health and clinical decision-making. By identifying key clinical and demographic factors associated with COVID-19 mortality through an interpretable, rule-based approach, health care providers can:

1) Develop early warning systems to identify high-risk patients; 2) Optimize hospital resource allocation by prioritizing critical cases; 3) Enhance public health surveillance through transparent, data-driven decision-making; 4) Improve treatment protocols and preventive strategies by recognizing high-risk patient profiles; 6) By integrating association rule mining into clinical practice, this study contributes to evidence-based decision-making and strengthens community health policies to reduce COVID-19-related mortality.

Materials and Methods

The relevant data were collected prospectively from hospitals located in Khorasan Razavi Province, Iran, from February 20 to September 12, 2021 (Figure 1). The necessary data and information for the subjects were extracted from questionnaires administered by trained nurses. These questionnaires were designed to gather detailed clinical and demographic information, including patient symptoms, medical history and test results, and were registered in the medical care monitoring center (MCMC) database. This study's inclusion criteria included all individuals referred to hospitals with COVID-19 symptoms, diagnosed with COVID-19 by a physician and required hospitalization. Additionally, all participants had to have been tested by the polymerase chain reaction (PCR).

The province's significance is that it is a major pilgrimage site for people from across Iran and some regional Muslim countries. So, it is an ideal location for obtaining a diverse and representative sample. For data collection, the nurses followed a standardized procedure to ensure the consistency and reliability of the data. Before analysis, we performed a thorough data-cleaning process to identify and remove inaccuracies, irrelevant entries, missing information and incomplete records.

Our dataset included age, sex, presence of PCR test results, fever, cough, muscular pain, respiratory distress, decreased consciousness, decreased sense of smell, reduced sense of taste, convulsions, headache, confusion, chest pain, skin inflammation, stomachache, nausea, vomit, diarrhea, anorexia, smoking status, drug use, PO_2, cancer, chronic liver diseases, diabetes, chronic blood diseases, immunodeficiency, chronic heart diseases, chronic kidney diseases, asthma, chronic neurological disorders, hypertension. The data cleaning phase was completed as part of the preprocessing phase before analysis and modeling. All collected data were carefully reviewed during the data cleaning to ensure the values were reasonable and consistent. Categorical variables were appropriately coded and transformed into binary variables as needed. Additionally, missing data were removed to maintain the integrity of the dataset. Following the data cleaning phase, the final dataset comprised 33 variables and 51460 cases.

Association rule mining

Association rule (or pattern) mining is common and one of the most popular data mining techniques, also known as association rule learning, frequent item set analysis, or association analysis. This technique attempts to find an association between many attributes (items or baskets that are also called transactions) or identify the items that often occur together [21-23]. Data analysts are interested in determining the frequency of item sets (or customer transactions) containing a particular set of items in large databases or predicting the behavior of customers [22, 24].

Association rule mining is based on the "market-basket" data model. It is usually performed on transaction data from online stores, yet it has been employed using various contexts of data, such as medical data [25]. Transaction data are usually provided as tuples form [transaction ID, item ID, item ID, ...]. A set of items forms

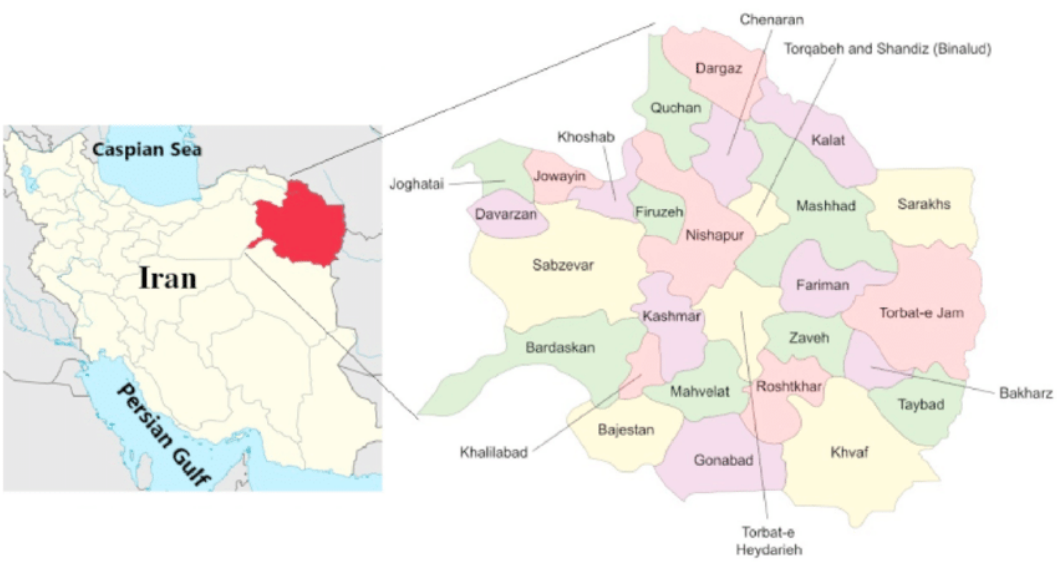


Figure 1. Location of Razavi Khorasan province and its cities in Iran

a basket or transaction (In our study, we considered each individual's symptoms as a single transaction), and there are some assumptions in the market-basket model. The number of items in a basket is assumed to be small and usually smaller than the total number of all items. The total number of baskets is assumed to be very large [22].

The basic formula of association rules is presented as [X -> Y]. The [X] is an "if statement" and consists of a set of items, and [Y] is a "then statement" that includes another set of items. The [X] is called antecedent or left-hand-side (LHS) and [Y] is called consequent or right-hand-side (RHS) [22]. To distinguish the best rule among the generated rules (strength of association rules), some metrics such as support, confidence and lift are calculated as Equation 1:

$$Support(X) = \frac{frequency\ (X)}{n}$$

1. $$Confidence\ (X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)} = P(Y|X)$$

$$Lift = \frac{Support(X \cup Y)}{Support(X) \times Support(Y)}$$

, where n is the number of all transactions or records in the data set, and the confidence expresses how likely an item Y has occurred given that item X occurred. The best rule is one with higher support and confidence.

The minimum support and minimum confidence must be determined earlier to find strong rules. When a rule's desired support and confidence surpasses the predetermined values of the minimum support and minimum confidence, it is considered a strong rule [24, 26].

Another introduced metric is Lift. This metric controls how popular items Y and X are and tells us how likely Y will occur when X occurs. The lift value is interpreted in three forms. A lift value greater than one implies that Y is expected to occur if X has occurred, while a lift value smaller than one shows that Y is unlikely to occur if X has occurred. The third form is when the lift value equals one, which means no association exists between desired items [22, 26]. Various algorithms are applied for exploring association rules from data [21, 22, 27]. In this study, the apriori algorithm was considered.

Apriori algorithm

The Apriori algorithm is a data mining and association rule mining algorithm. This algorithm uses prior knowledge about the properties of a frequent item set. For this reason, it is called the Apriori algorithm. This algorithm attempts to proceed by finding item sets with high frequency and provides a comprehensive analysis of hidden association rules in data [21, 26]. Moreover, the redundant rules are filtered, and statistically significant rules are identified using the Fisher exact test [28]. A rule [X->Y] is defined as a redundant rule if , where conf is confidence score [29].

The CBA algorithm

After mining the association rules, we are interested in predicting death status using association rules. We can access the associative classification approach by combining the classification and association rule mining. This approach is a reliable classification strategy with an improved accuracy rate [30]. CBA is a classifier based on association rules using ranking, pruning, and classification strategies introduced by Liu et al. (2021). The CBA has a rule generator and a classifier builder. To find the correlation between attributes, CBA uses an Apriori algorithm. Furthermore, a classifier was built based on extracted association rules. The results of previous studies show that CBA performed remarkably on real-world data [31]. Further details on the structure and performance of the CBA algorithm have been previously published [32]. Thus, in this study, the M1 approach was considered. The confusion matrix and extracted indices from the classifier's performance, such as sensitivity, specificity, accuracy and area under curve (AUC) of the receiver operating characteristics (ROC) curve, were used to evaluate the classifier's performance.

Results

A total of 51460 records were analyzed. In this sample, 26232(51%) of the subjects were male. The Mean±SD age among non-survivors was 51.91±22.06 years and 67.94±17.17 years among survivors (P<0.001), while it was 54.55±22.15 years in the total population. Further information regarding the symptoms, underlying diseases, and demographic characteristics of the study population is detailed in Table 1. Respiratory distress (56% or 2801 cases) was the most common symptom, followed by a PO_2<93% with 52.9% or 27200 cases. Other common symptoms were cough (19767 cases/38.4%) and fever (19018 cases/37%).

Table 1. Descriptive statistics for characteristics of the study population

Variables [Reference]	Level	No. (%)		
		Total (n=51460)	Survivor (n=42976)	Non-survivor (n=8484)
Age [Age <18]	18≤Age≤45	13379(26.0)	12667(29.5)	703(8.3)
	45≤Age≤65	15841(30.8)	13595(31.6)	2246(26.5)
	65≤Age≤75	7964(15.5)	6051(14.1)	1913(22.5)
	Age group 5 [x≥75]	11018(21.4)	7523(17.5)	3495(41.2)

Variables	No. (%)			Variables	No. (%)		
	Total	Survivor	Non-survivor		Total	Survivor	Non-survivor
Fever [no]	19018(37.0)	16280(37.9)	2738(32.3)	Anorexia [no]	3778(7.3)	3158(7.3)	620(7.3)
Cough [no]	19767(38.4)	17282 (40.2)	2485(29.3)	Smoking status [no]	1243(2.4)	989(2.3)	254(3.0)
Muscular pain [no]	9834(19.1)	8854(20.6)	980(11.6)	Drug use [no]	1892(3.7)	1436(3.3)	456(5.4)
Distress [no]	2880 (56.0)	22451(52.2)	6350(74.8)	PCR [negative]	509(2.3)	2247(44.7)	<0.001
Consciousness [no]	4594(8.9)	2367(5.5)	2227(26.2)	PO_2 [>93%]	27200(52.9)	20006(46.6)	7194(84.8)
Decreased sense of smell [no]	539(1.0)	511(1.2)	28(0.3)	Cancer [no]	944(1.8)	571(1.3)	373(4.4)
Decreased sense of taste [no]	253(0.5)	230(0.5)	23(0.3)	Liver disease [no]	325(0.6)	228(0.5)	97(1.1)
Convulsions [no]	363(0.7)	314(0.7)	49(0.6)	Diabetes [no]	6384(12.4)	4694(10.9)	1690(19.9)
Headache [no]	3647(7.1)	3471(8.1)	176(2.1)	Blood diseases [no]	269(0.5)	184(0.4)	85(1.0)
Confusion [No]	1098(2.1)	984(2.3)	114(1.3)	Immunodeficiency [no]	81(0.2)	58(0.1)	23(0.3)
Chest pain [no]	1773(3.4)	1581(3.7)	192(2.3)	Heart disease [no]	5714(11.1)	4099(9.5)	1615(19.0)
Skin inflammation [no]	71(0.1)	65(0.2)	6(0.1)	Kidney disease [no]	1023(2.0)	669(1.6)	354(4.2)
Stomachache [no]	1393(2.7)	1276(3.0)	117(1.4)	Asthma [no]	1052(2.0)	866(2.0)	186(2.2)
Nausea [no]	3045(5.9)	2713(6.3)	332(3.9)	Neurological disease [no]	889(1.7)	598(1.4)	291(3.4)
Vomit [no]	2156(4.2)	1948(4.5)	208(2.5)	HTN [no]	9243(18.0)	6991(16.3)	2252(26.5)
Diarrhea [no]	1834(3.6)	1683(3.9)	151(1.8)	Sex [Female]	26232(51.0)	21426(49.9)	4806(56.6)

HTN: Hypertension; distress, respiratory distress.

[.]: Reference level.

All included variables were binary and indicated the presence of a desired symptom. For example, the PO_2 symbol indicates individuals with PO_2<93%. Figure 2 displays the absolute item frequency.

When the Apriori algorithm was performed on the total population, we discovered 34 significant rules with minimum support of 0.1 and minimum confidence of 0.5 (Table 2). The minimum support and confidence thresholds for Apriori were determined through a grid search approach. The items or characteristics, PO_2 and respiratory distress, sex were the most common consequent or RHS. Three rules expressed that having fever (support=0.192), having PO_2<93% (support=0.276),

Table 2. Frequent pattern apriori algorithm-based association rules

Total						Non-survivor					
LHS	**RHS**	**Support**	**Confidence**	**Lift**	**Count**	**LHS**	**RHS**	**Support**	**Confidence**	**Lift**	**Count**
65≤Age<75	PO_2	0.102	0.659	1.246	5246	Consciousness, PO_2	Non-survivor}	0.036	0.553	3.355	1857
65≤Age<75	Distress	0.102	0.657	1.175	5236	Distress, consciousness	Non-survivor}	0.026	0.520	3.152	1348
HTN	Female	0.105	0.585	1.192	5403	distress, consciousness, PO_2	Non-survivor	0.023	0.557	3.381	1189
HTN	PO_2	0.120	0.670	1.267	6190	male, consciousness, PO_2	Non-survivor	0.021	0.557	3.377	1055
HTN	Distress	0.120	0.671	1.198	6200	age≥75, consciousness	Non-survivor	0.020	0.546	3.312	1007
Age≥75	Male	0.113	0.528	1.036	5821	age≥75, consciousness, PO_2	Non-survivor	0.016	0.592	3.591	842
Age≥75	PO_2	0.153	0.717	1.356	7899	female, consciousness, PO_2	Non-survivor	0.016	0.549	3.327	802
Age≥75	Distress	0.145	0.676	1.208	7453	male, distress, consciousness	Non-survivor	0.014	0.521	3.160	741
45≤Age<65	PO_2	0.170	0.552	1.044	8739	male, distress, consciousness, PO_2	Non-survivor	0.013	0.556	3.370	654
45≤Age<65	Distress	0.183	0.594	1.061	9408	age≥75, distress, consciousness	Non-survivor	0.012	0.581	3.526	632
Fever	Male	0.192	0.520	1.020	9892	female, distress, consciousness	Non-survivor	0.012	0.518	3.141	607
Male	PO_2	0.276	0.542	1.026	14224	age≥75, male, consciousness	Non-survivor	0.011	0.556	3.371	562
PO_2	Male	0.276	0.523	1.026	14224	age≥75, distress, consciousness, PO_2	Non-survivor	0.011	0.611	3.705	559
Male	Distress	0.290	0.569	1.016	14915	female, distress, consciousness, PO_2	Non-survivor	0.010	0.560	3.394	535
Distress	Male	0.290	0.518	1.016	14915		Survivor				
PO_2	Distress	0.390	0.738	1.319	20082	Female	Survivor	0.419	0.854	1.023	21550
Distress	PO_2	0.390	0.697	1.319	20082	Cough	Survivor	0.336	0.874	1.047	17282
Age≥75, PO2	Distress	0.116	0.756	1.351	5971	Fever	Survivor	0.316	0.856	1.025	16280
Age≥75, distress	PO_2	0.116	0.801	1.516	5971	PCR+	Survivor	0.285	0.851	1.019	14679
45≤Age<65, PO2	Distress	0.127	0.748	1.336	6533	45≤Age<65	Survivor	0.264	0.858	1.028	13595
45≤Age<65, distress	PO_2	0.127	0.694	1.314	6533	18≤Age<45	Survivor	0.246	0.947	1.135	12676
PCR+PO_2	Distress	0.118	0.724	1.293	6056	Muscular pain	Survivor	0.172	0.900	1.078	8854
PCR+distress	PO_2	0.118	0.677	1.280	6056	Female, cough	Survivor	0.169	0.892	1.069	8718
Male, fever	PO_2	0.105	0.548	1.036	5416	Male, cough	Survivor	0.166	0.857	1.026	8564
Fever, PO_2	Male	0.105	0.533	1.046	5416	Female, fever	Survivor	0.155	0.875	1.047	7982
Fever, PO_2	Distress	0.135	0.686	1.226	6968	PCR+, female	Survivor	0.141	0.861	1.032	7274
Fever, distress	PO_2	0.135	0.755	1.428	6968	Fever, cough	Survivor	0.138	0.855	1.024	7125
Cough, PO_2	Distress	0.142	0.709	1.267	7311	45≤Age<65, female	Survivor	0.133	0.882	1.057	6822
Cough, distress	PO_2	0.142	0.714	1.350	7311	18≤Age<45, male	Survivor	0.124	0.937	1.122	6396
Female, PO_2	Distress	0.186	0.736	1.315	9551	18≤Age<45, female	Survivor	0.122	0.958	1.148	6280

Total						Non-survivor					
LHS	RHS	Support	Confidence	Lift	Count	LHS	RHS	Support	Confidence	Lift	Count
Female, distress	PO_2	0.186	0.688	1.301	9551	18≤Age<45, cough	Survivor	0.116	0.965	1.156	5992
Male, PO_2	Distress	0.205	0.740	1.323	10531	45≤Age<65, cough	Survivor	0.115	0.893	1.069	5919
Male, distress	PO_2	0.205	0.706	1.336	10531	18≤Age<45, distress	Survivor	0.108	0.919	1.101	5539
Distress, PO_2	Male	0.205	0.524	1.029	10531	PCR+, fever	Survivor	0.106	0.877	1.051	5468

LHS: Left-hand-side or antecedent; RHS: Right-hand-side or consequent.

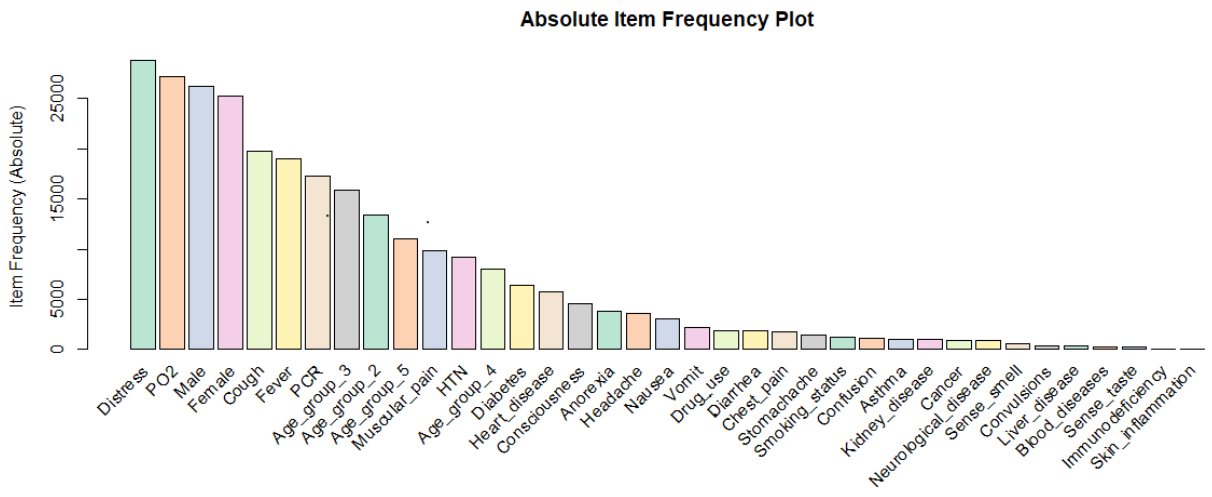


Figure 2. Item frequency plot for the total population

and respiratory distress (support=0.290) was common in males compared to females. Also, the simultaneous presence of fever and PO_2<93% (support=0.105) and respiratory distress and PO_2<93% (support=0.205) were seen in men. Further extracted rules are detailed in Table 2. These 34 extracted rules were visualized in Figure 3, too. This network graph focuses on the association between individual items (characteristics) through rule sets. As a brief demonstration of this plot, the rules are indicated by circles. The characteristics or items shown in the boxes (LHS items) are connected to rules, which are connected to RHS items using arrows. Furthermore, the rules with larger circles represent the rules with higher support values, while circles with a red color imply higher lift values. Fever and being male containing (arrow from fever to male) create rule 33. The fever characteristics are the LHS or antecedent and being male is the consequent RHS.

Furthermore, we studied extracted rules in male, female, survivor and non-survivor patients and patients with positive PCR test results. In non-survivor patients, the algorithm explored 14 significant rules. The strongest rule was (Consciousness, PO_2) with the largest support value of 0.036, and it says to us if the consciousness in a patient decreases and PO_2<93%, then this patient is who has died. Another rule indicated that if a patient was larger than 75 years old, had respiratory distress, and had decreased Consciousness and PO_2<93%, then the patient has died. Nineteen rules were generated in survivor's patients. Being female led to survival with a support value equal to 0.419, which had the largest support value, so it was the top rule. In patients with positive PCR test results, 12 rules were generated, reported in Table 3. The second generated rule showed the PCR test result of a male who used drugs was positive with a support value of 0.013.

When patients were disaggregated by female sex, 10 significant rules were extracted. Most of the females had hypertension. This result is taken from the first rule with a support value of 0.105. Also, the muscular pain due to COVID-19 is more common in women. Also, the simultaneous presence of a PO_2<93% and hypertension was more common in females compared to males. Other generated rules can be seen in detail in Table 3.

Table 3. Frequent pattern apriori algorithm-based association rules

PCR+						Female					
LHS	**RHS**	**Support**	**Confidence**	**Lift**	**Count**	**LHS**	**RHS**	**Support**	**Confidence**	**Lift**	**Count**
Drug use	PCR+	0.019	0.530	1.582	1003	Diabetes	Female	0.070	0.561	1.145	3584
Male, drug use	PCR+	0.013	0.533	1.591	684	Distress, HTN	Female	0.070	0.577	1.178	3579
Distress, drug use	PCR+	0.011	0.510	1.520	589	Heart diseases	Female	0.057	0.516	1.052	2946
Cancer	PCR+	0.009	0.515	1.536	486	65≤age<75, PO_2	Female	0.054	0.529	1.079	2775
Vomit, diarrhea	PCR+	0.008	0.579	1.727	400	65≤age<75, distress	Female	0.054	0.529	1.080	2771
Male, distress, drug use	PCR+	0.008	0.503	1.501	387	Distress,PO_2, HTN	Female	0.053	0.576	1.176	2724
45≤Age<65, drug use	PCR+	0.007	0.510	1.523	345			Male			
Fever, drug use	PCR+	0.007	0.523	1.561	340	Distress	Male	0.290	0.518	1.016	14915
18≤Age<45, consciousness	PCR+	0.006	0.559	1.666	329	PO_2	Male	0.276	0.523	1.026	14224
Chest pain, heart diseases	PCR+	0.006	0.504	1.503	320	Distress, PO_2	Male	0.205	0.524	1.029	10531
Distress, cancer	PCR+	0.006	0.512	1.528	319	Age≥75	Male	0.113	0.528	1.036	5821
Female, drug use	PCR+	0.006	0.524	1.563	319	Fever, PO_2	Male	0.105	0.533	1.046	5416
		Female				Fever, distress	Male	0.096	0.534	1.048	4930
HTN	Female	0.105	0.585	1.192	5403	Fever, distress, PO_2	Male	0.073	0.536	1.052	3735
Muscular pain	Female	0.098	0.514	1.049	5059	18≤age<45, distress	Male	0.063	0.542	1.063	3264
65≤Age<75	Female	0.083	0.534	1.088	4249	18≤age<45, PO_2	Male	0.053	0.569	1.116	2704
PO_2, HTN	Female	0.070	0.584	1.191	3614	Consciousness	Male	0.050	0.562	1.103	2584

LHS: Left-hand-side or antecedent; RHS: Right-hand-side or consequent; PCR: Polymerase chain reaction.

Similarly, 10 potential rules were identified for male patients. The most common symptom seen in men was respiratory distress, while the most common symptom in women was hypertension. Interestingly, the simultaneous presence of a PO_2<93% and respiratory distress were more common in males than females. More details regarding identified rules are shown in Table 3.

In the second step, we predict survivors and non-survivors using the CBA algorithm. Using 90% of the data, the model was trained based on extracted association rules using the total population. On the remaining 10%, the model was tested. The results of model performance are reported in Table 4.

The model predicts survivors and non-survivors with 93.50% and 92.28% accuracy on training and testing data, respectively. Of the 6669 non-survivors, 5647 were correctly identified as non-survivors (sensitivity=73.95% on the training data). On the other hand, from 773 non-survivors, 612 were correctly identified as non-survivors (sensitivity=72.17% on the testing data). This algorithm identifies persons' survivor status by specificity values of 97.36% and 96.25% in the training and test phases, respectively. The AUC value and ROC curve are shown in Figure 4.

Discussion

Association rules mining is an active research field in the data mining community and has previously been used when investigating health care problems [33]. Different algorithms have been proposed to discover patterns or hidden relationships between symptoms and diseases [31, 34]. Discovering the relationship between attributes is critical to understanding a disease and its biomarkers and helps decision-makers and researchers [35].

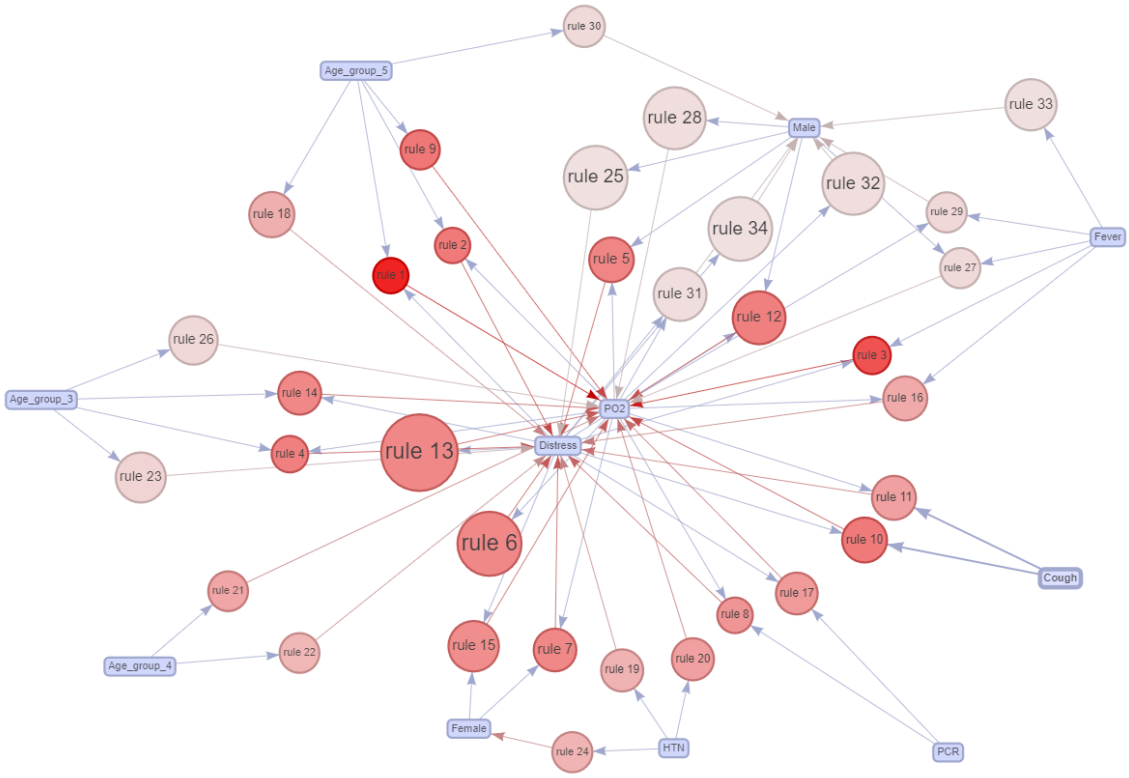


Figure 3. Graph-based visualization to visualize extracted rules for the total population

To our knowledge, very few studies have been published that apply association rules mining on COVID-19 data. However, Tandan et al. used rule-based machine learning approaches (association rules mining). They identified frequent symptoms and found patterns in the rules extracted. Of note is that the sample size in this study was comparatively small (1560 patients included). In contrast, in the current study, removing missing data yielded a total sample of 51460 subjects. Tandan et al. found that the presence of a fever, cough, body sore-

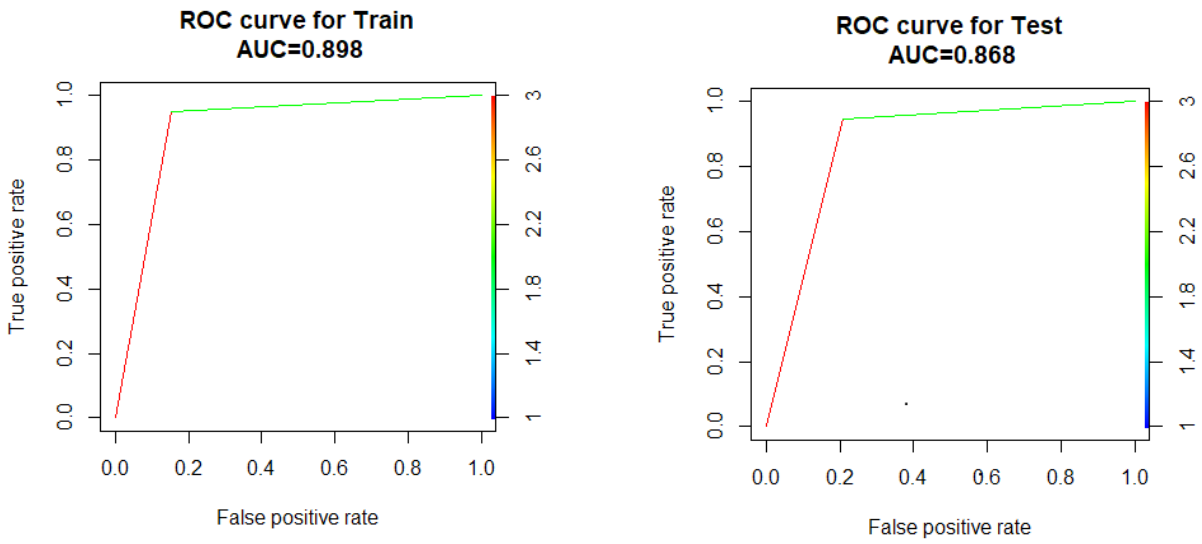
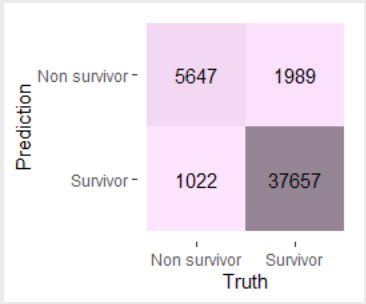
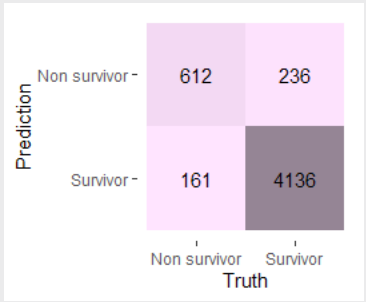


Figure 4. The presentation of the ROC curve

AUC: Area under curve.

Table 4. Performance of classification based on association rules on total population

Total of CF (n=51460)	
Training: n=46315	Testing: n=5145
	
Sensitivity: 73.95%	Sensitivity: 72.17%
Specificity: 97.36%	Specificity: 96.25%
Accuracy: 93.50%	Accuracy: 92.28%
95% CI for accuracy: (93.27, 93.72)	95% CI for accuracy: (91.52, 93.00)

ness, pneumonia, and sore throat were the symptoms that occurred most frequently. Using an Apriori algorithm, they reported the top 10 significant rules to be the presence of a cough, septic shock and respiratory distress syndrome as frequent consequences. Moreover, they studied significant rules by age in categories of <20, 20-45, 45-65 and >65 years, as well as by sex and the presence of chronic condition(s). Our study discovered significant rules in the total population and among survival and non-survival individuals with positive PCR test results and male and female subgroups. To the best of our knowledge, this is the first study identifying the important symptoms of COVID-19 by association rules mining according to survival status and then predicting survivors versus non-survivors using extracted association rules.

In another work, Mohammadi et al. [36] found common symptoms and the presence of underlying disease in 750 confirmed COVID-19 cases in Iran. Hypertension, diabetes, chronic obstructive pulmonary disease, and coronary heart disease were identified as the most common underlying diseases using neural network and logistic regression methods. Fever, cough, shortness of breath, fatigue, chills and headache were common symptoms.

This study has several limitations. First, access to comprehensive hospital and clinical data was restricted due to privacy concerns, which may have impacted the completeness of our dataset. Additionally, data quality issues, including missing values, potential recording errors, and data collection inconsistencies, could affect our findings' accuracy. The dataset was also imbalanced, with specific patient groups being underrepresented, which may have influenced the model's predictive performance. Furthermore, the generalizability of our results is limited, as the study was conducted within a specific geographic region and may not fully apply to other populations with different demographic or envi-

ronmental factors. Another limitation is using a particular data mining method classification-based association rule (CBA), whereas other advanced machine learning or statistical approaches might have yielded better predictive outcomes. Moreover, the study requires further validation through independent datasets or clinical trials to ensure the robustness of the findings. Finally, some confounding variables, such as genetic, environmental, or socioeconomic factors, were not fully controlled, potentially introducing bias into the results.

Conclusion

Simple methods, such as association rules mining, and complex methods can provide valuable information and rules. On the other hand, predicting death using association rules and the CBA algorithm provides high-accuracy predictions and can help diagnose death or survival. Therefore, in finding a more accurate model, this algorithm can be tested along with other machine learning algorithms to finally achieve an optimal and high-precision model for treating, preventing, or controlling COVID-19.

Ethical Considerations

Compliance with ethical guidelines

There were no ethical considerations to be considered in this research.

Funding

This research did not receive any grant from funding agencies in the public, commercial, or non-profit sectors.

Authors contributions

Conceptualization, methodology, formal analysis, and software: Nasrin Talkhi; Investigation: Nooshin Akbari Sharak; Data collection: Maryam Salari; Data management: Seyed Masoud Sadati; Writing the original draft: Nasrin Talkhi and Nooshin Akbari Sharak; Review and editing: Nasrin Talkhi and Nooshin Akbari Sharak; Supervision and validation: Mohammad Taghi Shakeri.

Conflict of interest

The authors declared no competing interests.

Acknowledgements

The authors would like to acknowledge and thank all the experts who participated in the study.

References

[1] WHO. Coronavirus disease 2019 (COVID-19): Situation report-51. Geneva: WHO; 2020. [Link]

[2] Crossfield SSR, Chaddock NJM, Iles MM, Pujades-Rodriguez M, Morgan AW. Interplay between demographic, clinical and polygenic risk factors for severe COVID-19. International Journal of Epidemiology. 2022; 51(5):1384-95. [DOI:10.1093/ije/dyac137] [PMID]

[3] Hu C, Liu Z, Jiang Y, Zhang X, Shi O, Xu K, et al. Early prediction of mortality risk among severe COVID-19 patients using machine learning. International Journal of Epidemiology. 2021; 49(6):1918-29. [DOI:10.1093/ije/dyaa171] [PMID]

[4] Chen T, Wu D, Chen H, Yan W, Yang D, Chen G, et al. Clinical characteristics of 113 deceased patients with coronavirus disease 2019: Retrospective study. BMJ. 2020; 368:m1295. [DOI:10.1136/bmj.m1295] [PMID]

[5] Yang X, Yu Y, Xu J, Shu H, Xia J, Liu H, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. The Lancet Respiratory Medicine. 2020; 8(5):475-81. [DOI:10.1016/S2213-2600(20)30079-5] [PMID]

[6] CDC COVID-19 Response Team. Severe outcomes among patients with coronavirus disease 2019 (COVID-19)-United States, February 12-March 16, 2020. MMWR. Morbidity and Mortality Weekly Report. 2020; 69(12):343-6. [DOI:10.15585/mmwr.mm6912e2] [PMID]

[7] Chechet GD, Kwaga JKP, Yahaya J, Noyes H, MacLeod A, Adamson WE. SARS-CoV-2 seroprevalence at urban and rural sites in Kaduna State, Nigeria, during October/November 2021, immediately prior to detection of the Omicron variant. International Journal of Epidemiology. 2022; 51(5):1361-70. [DOI:10.1093/ije/dyac141] [PMID]

[8] Talkhi N, Sharak NA, Rajabzadeh Z, Salari M, Sadati SM, Shakeri MT. Identification symptoms and underlying diseases related to COVID-19 and prediction of death status using artificial neural network and logistic regression: A data mining approach. Iranian Journal of Epidemiology. 2022; 18(3):244-54. [Link]

[9] Talkhi N, Akbari Sharak N, Pasdar Z, Salari M, Sadati SM, Shakeri MT. Potential pathological, clinical, and symptomatic findings of covid-19 to predict mortality in positive pcr individuals using data mining. Journal of Patient Safety & Quality Improvement. 2023; 11(1):13-21. [DOI:10.22038/psj.2023.70741.1390]

[10] Talkhi N, Akbari Sharak N, Yousefi R, Salari M, Sadati SM, Shakeri MT. Predicting COVID-19 Mortality and Identifying Clinical Symptom Patterns in Hospitalized Patients: A Machine-learning Study. Iranian Journal of Health Sciences. 2024; 12(1):39-48. [DOI:10.32598/ijhs.12.1.952.1]

[11] Wadhwa P, Aishwarya, Tripathi A, Singh P, Diwakar M, Kumar N. Predicting the time period of extension of lockdown due to increase in rate of COVID-19 cases in India using machine learning. Materials Today. Proceedings. 2021; 37:2617-22. [DOI:10.1016/j.matpr.2020.08.509] [PMID]

[12] Chowdhury ME, Rahman T, Khandakar A, Al-Madeed S, Zughaier SM, Hassen H, et al. An early warning tool for predicting mortality risk of COVID-19 patients using machine learning. Cognitive Computation. 2021; 1-16. [DOI:10.1007/s12559-020-09812-7] [PMID]

[13] Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Magazine. 1996; 17(3):37-57. [DOI:10.1609/aimag.v17i3.1230]

[14] Cios KJ, Moore GW. Uniqueness of medical data mining. Artificial Intelligence in Medicine. 2002; 26(1-2):1-24. [DOI:10.1016/S0933-3657(02)00049-0] [PMID]

[15] Zaki MJ. Scalable algorithms for association mining. IEEE transactions on knowledge and data engineering. 2000; 12(3):372-90. [DOI:10.1109/69.846291]

[16] Albahri AS, Hamid RA, Alwan Jk, Al-qays ZT, Zaidan AA, Zaidan BB, et al. Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review. Journal of Medical Systems. 2020; 44(7):122. [DOI:10.1007/s10916-020-01582-x] [PMID]

[17] Biilah MAM, Raihan M, Akter T, Alvi N, Bristy NJ, Rehana H. Human depression prediction using association rule mining technique. In: Khanna A, Gupta D, Bhattacharyya S, Hassanien AE, Anand S, Jaiswal A, editords. International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing, vol 1388. Springer: Singapore; 2022. [DOI:10.1007/978-981-16-2597-8_19]

[18] Ilayaraja M, Meyyappan T. Mining medical data to identify frequent diseases using Apriori algorithm. Paper presented at: 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering. 21-22 February 2013; Salem, India. [DOI:10.1109/ICPRIME.2013.6496471]

[19] Abdullah U, Ahmad J, Ahmed A. Analysis of effectiveness of apriori algorithm in medical billing data mining. Paper presented at: 2008 International Conference on Emerging Technologies. 18-19 October, 2008. Rawalpindi, Pakistan. [Link]

[20] Jena L, Kamila NK. A model for prediction of human depression using Apriori Algorithm. 2014 International Conference on Information Technology. 22-24 December 2014; Bhubaneswar, India. [DOI:10.1109/ICIT.2014.65]

[21] Lakshmi KS, Vadivu G. Extracting association rules from medical health records using multi-criteria decision analysis. Procedia Computer Science. 2017; 115:290-5. [DOI:10.1016/j.procs.2017.09.137]

[22] Hahsler M. An R companion for introduction to data mining. Journal of Open Source Education. 2024; 7(82):223. [Link]

[23] Zakariya SM, Yaseen A, Khan IA. A Study of Association Rule Mining for Artificial Immune System-Based Classification. In: Luhach, AK, Poonia RC, GaoXZ, Singh Jat D, editord. Second International Conference on Sustainable Technologies for Computational Intelligence. Advances in Intelligent Systems and Computing, vol 1235. Springer: Singapore; 2022. [Link]

[24] Lu PH, Keng JL, Tsai FM, Lu PH, Kuo CY. An apriori algorithm-based association rule analysis to identify acupoint combinations for treating diabetic gastroparesis. Evidence-Based Complementary and Alternative Medicine: eCAM. 2021; 2021:6649331. [DOI:10.1155/2021/6649331] [PMID]

[25] Prabhakaran S. Association mining (market basket analysis). r-statistics.co; 2017. [Link]

[26] Makhtar M, Harun NA, Aziz AA, Zakaria ZA, Abdullah FS, Jusoh JA. An association rule mining approach in predicting flood areas. In: Herawan T, Ghazali R, Nawi NM, Deris MM, editors. Recent Advances on Soft Computing and Data Mining. SCDM 2016. Advances in Intelligent Systems and Computing, vol 549. Cham: Springer; 2017. [DOI:10.1007/978-3-319-51281-5_44]

[27] Altaf W, Shahbaz M, Guergachi A. Applications of association rule mining in health informatics: A survey. Artificial Intelligence Review. 2017; 47(3):313-40. [DOI:10.1007/s10462-016-9483-9]

[28] Webb GI. Discovering Significant Patterns. Machine Learning. 2007; 68(1):1-33. [DOI:10.1007/s10994-008-5045-y]

[29] Bayardo RJ, Agrawal R, Gunopulos D. Constraint-Based Rule Mining in Large, Dense Databases. Data Mining and Knowledge Discovery. 2000; 4(2):217-40. [DOI:10.1023/A:1009895914772]

[30] Mittal K. An approach towards enhancement of classification accuracy rate using efficient pruning methods with associative classifiers. International Journal of Information Technology. 2021; 14:1525-33. [DOI:10.1007/s41870-021-00673-3]

[31] Liu X, Niu X, Fournier-Viger P. Fast top-k association rule mining using rule generation property pruning. Applied Intelligence. 2021; 51:2077-93. [DOI:10.1007/s10489-020-01994-9]

[32] Kumi S, Lim C, Lee SG. Malicious URL Detection Based on Associative Classification. Entropy. 2021; 23(2):182. [DOI:10.3390/e23020182] [PMID]

[33] Czibula G, Czibula IG, Miholca DL, Crivei LM. A novel concurrent relational association rule mining approach. Expert Systems with Applications. 2019; 125:142-56. [DOI:10.1016/j.eswa.2019.01.082]

[34] Aqra I, Abdul Ghani N, Maple C, Machado J, Sohrabi Safa N. Incremental algorithm for association rule mining under dynamic threshold. Applied Sciences. 2019; 9(24):5398. [DOI:10.3390/app9245398]

[35] Nahar J, Imam T, Tickle KS, Chen YPP. Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Applications. 2013; 40(4):1086-93. [DOI:10.1016/j.eswa.2012.08.028]

[36] Mohammadi F, Pourzamani H, Karimi H, Mohammadi M, Mohammadi M, Ardalan N, et al. Artificial neural network and logistic regression modelling to characterize COVID-19 infected patients in local areas of Iran. Biomedical Journal. 2021; 44(3):304-16. [DOI:10.1016/j.bj.2021.02.006]