Decision Making: Applications in Management and Engineering
Vol. 6, Issue 1, 2023, pp. 365-378.
ISSN: 2560-6018
eISSN: 2620-0104
DOI: https://doi.org/10.31181/dmame0317102022r

ASSOCIATION RULE MINING FOR PREDICTION OF COVID-19

Vishnu Kumar Rai[1], Santonab Chakraborty[2] and Shankar Chakraborty[1*]

[1] Department of Production Engineering, Jadavpur University, Kolkata, West Bengal, India
[2] Industrial Engineering and Management Department, Maulana Abul Kalam Azad University of Technology, West Bengal, India

Received: 13 August 2021;
Accepted: 26 September 2022;
Available online: 17 October 2022.

Original scientific paper

Abstract: *COVID-19 is a raging pandemic that has created havoc with its impact ranging from loss of millions of human lives to social and economic disruptions of the entire world. Therefore, error-free prediction, quick diagnosis, disease identification, isolation and treatment of a COVID patient have become extremely important. Nowadays, mining knowledge and providing scientific decision making for diagnosis of diseases from clinical datasets has found wide-ranging applications in healthcare sector. In this direction, among different data mining tools, association rule mining has already emerged out as a popular technique to extract invaluable information and develop important knowledge-base to help in intelligent diagnosis of distinct diseases quickly and automatically. In this paper, based on 5434 records of COVID cases collected from a popular data science community and using Rapid Miner Studio software, an attempt is put forward to develop a predictive model based on frequent pattern growth algorithm of association rule mining to determine the likelihood of COVID-19 in a patient. It identifies breathing problem, fever, dry cough, sore throat, abroad travel and attended large gathering as the main indicators of COVID-19. Employing the same clinical dataset, a linear regression model is also proposed having a moderately high coefficient of determination of 0.739 in accurately predicting the occurrence of COVID-19. A decision support system can also be developed using the association rules to ease out and automate early detection of other diseases.*

Key words: *COVID-19, Association rule mining, Frequent pattern growth, Prediction, Regression.*

* Corresponding author.
E-mail addresses: vishnurrr40@gmail.com (V. K. Rai), santonabchakraborty@gmail.com (S. Chakraborty), s_chakraborty00@yahoo.co.in (S. Chakraborty),

1. Introduction

Coronavirus disease 2019 (COVID-19) is mainly caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) which is contagious in nature. The transmission of COVID-19 occurs when a person inhales virus-containing respiratory droplets and airborne particles from an infected patient. The first known case of COVID was from Wuhan, China. It has now become a raging pandemic creating havoc with its impact ranging from loss of millions of lives to social and economic disruptions around the globe. The impact of COVID in India, being the second populous country, is threatening. In India, the first few cases were from Kerala resulting in the first wave. Total lockdown was imposed from 25th March, 2020 and the number of active cases began to drop from September, 2020. A larger and much powerful second wave hit India on March 2021. Presently, India has the largest number of COVID cases in Asia. As of 12 June 2021, it has the second-highest number of confirmed cases in the world (after the United States) with 29.3 million reported cases of COVID-19 infection. It has also the third-highest number of COVID-19 deaths (after the United States and Brazil) with 367,081 deaths. India became the first country to report over 400,000 new cases in a 24 hour period on 30 April 2021. As of 30 June 2021, the total number of confirmed cases in India is 3,03,62,848 with total number of deaths as 3,98,454. The rapid increase of cases in the peak of both first and second waves of COVID-19 had put tremendous pressure on the medical infrastructure leading to shortage of hospital beds, oxygen cylinders, vaccines and other medicines in the country. There was also a great chance that the primary health workers were being infected by the same disease while treating the COVID patients.

The standard diagnostic procedure for COVID is to detect the presence of coronavirus's nucleic acid in human body which is usually performed by real-time reverse transcription polymerase chain reaction (rRT-PCR), transcription-mediated amplification (TMA) or by reverse transcription loop-mediated isothermal amplification (RT-LAMP) test from a nasopharyngeal swab. Each of these testing procedures for COVID-19 is time consuming requiring plenty of resources. At the peak of waves, when there are millions of daily cases, it has become crucial for having a more quick and efficient approach to determine whether a person has COVID-19 or not. While detecting this disease and treating the infected patients, the concerned healthcare sector is generating huge volume of valuable information which can be effectively deployed for rapid diagnosis, identification and treatment of an individual. In the present day pandemic scenario, mining knowledge and providing scientific decision making for diagnosis of COVID-19 from the clinical dataset has turned out to be extremely important. With rapid development of computational facilities, data mining technology has gained increasing attention to discover interesting knowledge in the form of useful patterns, changes, associations, anomalies and structures from large volume of data stored in databases, data warehouses or other data repositories. Association rule mining is an effective data mining tool mainly deployed to extract association relationships or correlations/co-occurrences among a given set of items. Due to its simplicity in framing rules based on 'If-Then' statements, association rule mining is now being extensively used to explain patterns from seemingly independent repositories, like transactional databases, relational databases or clinical databases (Kaur & Madan, 2015; Sabthami et al. 2016). The developed rules can assist the physicians in diagnosing patients based on the conditional probability while comparing the symptom relationships in the data from the past cases (Anand Hareendran & Vinod Chandra, 2017; Cheng &

Wang, 2017). In this paper, an attempt is put forward to employ association rule mining as an effective predictive tool for diagnosing COVID-19 patients based on frequent pattern (FP) growing algorithm. Using a large clinical database containing 5434 records of COVID cases, breathing problem, fever, dry cough, sore throat, abroad travel and attended large gathering are identified as the most important COVID-19 indicators. With the help of Rapid Miner Studio software, the corresponding association rules are framed for prediction of COVID-19 disease in a patient. The developed regression model would also aid in COVID-19 diagnosis correlating the symptoms and likelihood of this disease. Its application would save a lot of time and resources in the case of huge influx of patients. Using this predictive model, the patients themselves can envisage whether they have the disease or not and can start taking necessary precautions (Stilou et al., 2001).

Association rule mining, logistic regression, discriminant analysis etc. are different types of machine learning approaches. In association rule mining, simple 'If…Then' clauses are framed to discover the existent relationships between independent relational databases, transactional databases, and other forms of data repositories, requiring simple counting to observe frequently occurring patterns, and similarities and dissimilarities among different objects. On the other hand, logistic regression employs binary variables where the response records either success or failure for a given event. It can also be extended to combine more than one independent continuous or categorical variable. Discriminant analysis develops a set of prediction equations based on independent variables for classifying new items and interpreting the relationship among the considered variables. The application of discriminant analysis assumes that the data is normally distributed and each attribute has the same variance. Among these machine learning approaches, association rule mining is the simplest one, requiring no assumption about the underlying distribution of the initial dataset, while framing easy to understand rules. With the help of different parameters, like support, confidence and lift, it can clearly identify the strongest rule penetrating more inside the problem.

2. Association rule mining

Association rule mining is one of the techniques of data mining to extract interesting relations (dependencies) and patterns/links among variables from large seemingly independent datasets in order to draw useful inferences and decisions for practical use. Application of association rule mining helps in generating simple 'If-Then' statements to analyze frequently occurring patterns in a dataset and/or identify the inherent relationships between independent and dependant variables in the dataset. It can also frame useful rules from qualitative and categorical datasets which are often difficult to interpret (Ordonez et al., 2006). An association rule consists of two components, i.e. an antecedent (If) and a consequent (Then). An antecedent is an item found within the dataset and a consequent is an item observed in combination with the antecedent (Freeda & Florence, 2017). Thus, in a rule, $X \rightarrow Y$, X is the antecedent (If) and Y is the consequent (Then). In a clinical database, the rule {Symptom1, Symptom2}→{Disease1} signifies that a patient having Symptom1 and Symptom2 would tend to have Disease1. For example, there is a set of symptoms $A = \{a_1, a_2, \ldots, a_n\}$ and B indicates entries of multiple patients in a clinical database $B = \{b_1, b_2, \ldots, b_n\}$. Each patient contains a subset of elements in $A-B \subseteq A$, and the corresponding association rule is the implication $X \rightarrow Y$, where $X \subseteq A$, $Y \subseteq A$ and $X \cap Y = \emptyset$. In a clinical database, an antecedent is a specific symptom or combination of

symptoms and a consequent can be a disease caused due to occurrence of the antecedent. The generated rules would thus help the concerned physicians in making faster decisions with correct diagnosis of a disease (Kulkarni & Mundhe, 2017; Lakshmi & Vadivu, 2017).

The effectiveness of the developed association rules is usually validated using three parameters, i.e. support, confidence and lift. Support measures the percentage of items in the given dataset following a particular rule, i.e. how often a rule occurs in the dataset.

$\text{Support}(X \rightarrow Y) = P(X \cup Y)$

Confidence evaluates the probability of inclusion of item X also leading to the inclusion of Y. It is the conditional probability of how often a rule is found out to be true.

$\text{Confidence}(X \rightarrow Y) = P(Y|X) = \text{Support}(X \rightarrow Y)/\text{Support}(X)$

Lift finally measures the performance of an association rule.

$\text{Lift}(X \rightarrow Y) = \text{Confidence}(X \rightarrow Y)/\text{Support}(X \rightarrow Y)$

Higher value of confidence signifies higher strength of a particular rule. On the other hand, higher value of lift symbolizes that having X and Y together is not accidental, but due to the presence of a meaningful relationship between them.

Lift = 1 signifies that the probability of occurrence of antecedent and consequent is not dependant on each other.

Lift > 1 determines the degree to which antecedent and consequent are dependent to each other.

Lift < 1 signifies that one item has a negative effect on the other, i.e. one item is a substitute for the other item present in the rule.

There are three popular algorithms, i.e. apriori algorithm, ECLAT (equivalence class clustering and bottom-up Lattice transversal) algorithm and frequent pattern (FP) growth algorithm deployed for framing the relevant association rules from a given dataset (Prithiviraj & Porkodi, 2015). In apriori algorithm (which is an array-based algorithm), frequent itemsets are used for generation of the association rules. It employs a breadth-first search and hash tree to efficiently identify the frequent itemsets in a transactional database. But, its application is time-consuming and the corresponding runtime may increase exponentially (Jain & Gautam, 2013; Sambasiva Rao & Uma Devi, 2017). The ECLAT adopts a depth-first search technique to find out the frequent itemsets in a relational database. It has less execution time than apriori algorithm. The FP growth is a tree-based algorithm which employs depth-first search to compress the dataset to form an FP-tree. It is faster than the other two algorithms and its runtime increases linearly. But for large FP-tree, it may not fit in the memory space, thus being expensive to build (Thamer, et al., 2020). As in this paper, FP growth algorithm is employed for development of the association rules for effective prediction of COVID-19, its procedural steps are detailed out here-in-under.

Association rule mining using FP growth algorithm mainly consists of two steps, i.e. a) generation of frequent itemsets and b) formation of association rules from the frequent itemsets. In order to demonstrate generation of frequent itemsets employing FP growth algorithm, let us consider a clinical dataset containing different symptoms for nine infected patients. In this database, P and S represent Patient and Symptom respectively.

P1 = (S1, S2, S5); P2 = (S2, S4); P3 = (S2, S3); P4 = (S1, S2, S4); P5 = (S1, S3); P6 = (S2, S3); P7 = (S1, S3); P8 = (S1, S2, S3, S5) and P9 = (S1, S2, S3)

Now, the corresponding FP-tree is developed based on the following steps:

a) Scan the dataset to determine the support count of each symptom. Remove the less-frequent symptom(s) and sort the frequent symptoms in descending order of their occurrence.
b) Scan the dataset of one patient at a time resulting in formation of the FP-tree. For each transaction,
   i) If it has a set of unique symptoms, form a new path and set the counter for each node to 1.
   ii) If it shares a set of common symptoms, increase the common symptom itemset node counters and create new nodes, if needed.
c) This process needs to be continued until each patient case is mapped into the tree.

This algorithm scans the database only twice while directly compressing it into the corresponding FP-tree. In this algorithm, minimum support (basically acts as a cut-off) can be used to classify the frequent and less-frequent itemsets in a database. The less-frequent items are ignored while developing the FP-tree. Identification of the most appropriate cut-off for subsequent FP-tree generation is a critical task. Lower cut-off with minimum support may include many itemsets resulting in less significant results. On the other hand, higher cut-off may result in finding out zero itemsets with no generation of FP-tree. In this illustrative example, the support count of each symptom is determined as given below (in descending order): S2 = 7, S1 = 6, S3 = 6, S4 = 2 and S5 = 2. Now, the patient datasets are rearranged according to the descending order of support count of different symptoms.

P1 = (S2, S1, S5); P2 = (S2, S4); P3 = (S2, S3); P4 = (S2, S1, S4); P5 = (S1, S3); P6 = (S2, S3); P7 = (S1, S3); P8 = (S2, S1, S3, S5) and P9 = (S2, S1, S3). Based on the dataset with nine patient cases and five symptoms in the illustrative example, the following FP-tree of Figure 1 is developed. This FP-tree is generated while considering null as the root node. The count of each symptom for each patient case is highlighted in parenthesis at each node.

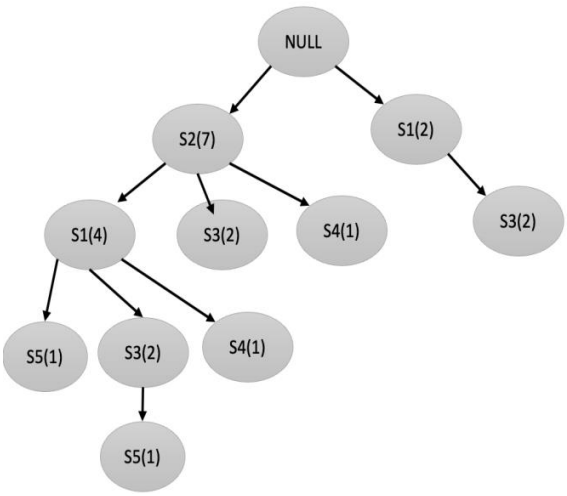

Figure 1. FP-tree for the illustrative example

Next, the developed FP-tree is mined. The lowest node of the tree is checked first along with its links. The lowest node represents the symptom with minimum support count. From the lowest node, traverse the path in the FP-tree to the null node. Each such path is termed as conditional pattern base. The conditional FP-tree is formed

while counting the symptoms in the path. The symptoms meeting the minimum support of 2 are considered here for subsequent generation of frequent itemsets, as exhibited in Table 1. In this table, six 2-frequent and two 3-frequent itemsets are generated.

Table 1. Generation of frequent itemsets for the illustrative example

Symptom	Conditional pattern base	Conditional FP-tree	Frequent itemsets
S5	{{S2,S1:1},{S2,S1,S3:1}}	[S2:2, S1:2]	{S2,S5:2},{S1,S5:2},{S2,S1,S5:2}
S4	{{S2,S1:1},{S2:1}}	[S2:2]	{S2,S4:2}
S3	{{S2,S1:1},{S2:2},{S1:2}}	[S2:4, S1:2], [S1:2]	{S2,S3:4},{S1,S3:4},{S2,S1,S3:2}
S1	{{S2:4}	[S2:4]	{S2,S1:4}

Based on the frequent itemsets in Table 1, the corresponding association rules are thereby generated using the following steps:

a) Generate all non-empty subsets of each frequent itemset U.

b) For every non-empty subset F of U, formulate the rule :

F→(U–F) if (support_count(U)/support_count(F) ≥ minimum_confidence where minimum_confidence is the threshold confidence level.

For example, consider the first frequent itemset U = (S2,S1,S5). Generate all the non-empty subsets of U as F: {S1},{S2},{S5},{S1,S2},{S2,S5},{S1,S5},{S1,S2,S5}. For every non-empty subset F of U, the corresponding association rules are framed, as given in Table 2. In this table, support_count is the number of occurrences of all the elements in a set (U or F) together in the dataset, Confidence Calculated = (support_count(U)/support_count(F))×100, Support = ((support_count(U))/N)×100, Lift = Confidence Calculated/Support and N is the total number of patient cases in the example. Here, the threshold confidence value is arbitrarily taken as 80%. It can be noticed from this table that among the generated rules, only rules 3, 5 and 6 are accepted with their confidence levels greater than or equal to the set threshold value. Thus, for this example, the following association rules are developed: S5→(S2,S1); (S1,S5)→S2 and (S2,S5)→S1.

Table 2. Association rules for the illustrative example

Rule No.	Association Rules	support_ count(U)	support_ count(F)	Confidence Calculated	Threshold Confidence	N	Support	Lift	Accepted/ Rejected
1	S1→(S2,S5)	2	6	33.33	80	9	22.22	1.50	Rejected
2	S2→(S1,S5)	2	7	28.57	80	9	22.22	1.29	Rejected
3	S5→(S2,S1)	2	2	100.00	80	9	22.22	4.50	Accepted
4	(S1,S2)→S5	2	4	50.00	80	9	22.22	2.25	Rejected
5	(S1,S5)→S2	2	2	100.00	80	9	22.22	4.50	Accepted
6	(S2,S5)→S1	2	2	100.00	80	9	22.22	4.50	Accepted
7	(S1,S2,S5) →(Null)								Rejected

It has been noticed that the apriori algorithm of association rule mining has already been successfully deployed for prediction/diagnosis of heart diseases (Said et al., 2015; Domadiya & Rao, 2018; Jamsheela, 2021), dengue (Jahangir et al., 2018), brain tumor (Sengupta et al., 2013), chronic kidney disease (Alaiad et al., 2020), infectious diseases (Brossette et al., 1998), pandemic diseases (Burvin & Dhanalakshmi, 2018; Aiswarya et al., 2020), COVID-19 (Çelik, 2020; Shawkat et al., 2021; Tandan et al., 2021), pediatric primary care (Downs & Wallace, 2000),

treatment of patients in an emergency department (Sarıyer & Taşar, 2020) etc. In this paper, based on a huge dataset of COVID-19 patients and using the FP growth algorithm of association rule mining, an attempt is put forward to discover COVID-19 symptom patterns and rules which would support the initial identification of severe COVID-19 cases for early treatment and isolation. Based on the most frequent symptoms, a first-order regression model is also developed to assist prediction of COVID-19.

3. Data collection

In order to predict COVID-19 based on development of the corresponding association rules, the related data is collected from Kaggle.com which is the world's largest online data science community. The data consists of the symptoms and other factors responsible for COVID-19 infection. They are based on the guidelines provided by the World Health Organization (WHO) (www.who.int) and the Ministry of Health and Family Welfare, India (https://main.mohfw.gov.in). The data is in 'Yes' and 'No' format, where 'Yes' represents the presence of a particular symptom and 'No' denotes absence of it. Based on the given guidelines, the considered factors for COVID-19 infection are as follows: a) breathing problem, b) fever, c) dry cough, d) sore throat, e) running nose, f) asthma, g) chronic lung disease, h) headache, i) heart disease, j) diabetes, k) hyper tension, l) fatigue, m) gastrointestinal, n) abroad travel, o) contact with other COVID patient, p) attended large gathering, q) visited public exposed places, r) family working in public exposed places and s) wearing masks at all times. In this database, COVID-19 is treated as the decisional (target) variable. The clinical dataset is in tabular form containing 5434 records of infected COVID cases. The snapshot of a small portion of the considered dataset is shown in Figure 2.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Breathing Problem	Fever	Dry Cough	Sore throat	Running Nose	Asthma	Chronic Lung Disease	Headache	Heart Disease	Diabetes	Hyper Tension	Fatigue	Gastroin testinal	Abroad travel	Contact with COVID Patient	Attended Large Gathering	Visited Public Exposed Places	Family working in Public Exposed Places	Wearing Masks	COVID-19
2	Yes	Yes	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No	Yes
3	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	No	No	Yes	No	No	No	Yes	Yes	No	No	Yes
4	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes	No	No	No	No	No	Yes
5	Yes	Yes	Yes	No	No	Yes	No	No	Yes	Yes	No	No	No	Yes	No	Yes	Yes	No	No	Yes
6	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	Yes	No	No	Yes
7	Yes	Yes	Yes	No	No	No	No	No	Yes	No	Yes	No	No	No	No	No	No	No	No	Yes
8	Yes	Yes	Yes	No	No	No	Yes	No	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes
9	Yes	Yes	Yes	No	Yes	No	No	No	No	Yes	Yes	No	Yes	Yes	No	No	Yes	No	No	Yes
10	Yes	Yes	Yes	No	Yes	No	Yes	No	No	Yes	No	Yes	No	Yes	Yes	Yes	No	No	No	Yes
11	Yes	Yes	Yes	No	No	Yes	No	No	No	Yes	Yes	Yes	No	No	No	No	Yes	No	No	Yes
12	Yes	Yes	Yes	No	No	No	Yes	No	Yes	No	Yes	No	No	Yes	No	Yes	No	No	No	Yes
13	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes	No	Yes
14	Yes	Yes	Yes	Yes	No	Yes	Yes	No	No	No	Yes	No	Yes	Yes	No	No	Yes	Yes	No	Yes
15	Yes	Yes	Yes	Yes	No	Yes	No	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes
16	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	No	Yes
17	Yes	Yes	Yes	No	Yes	No	No	Yes	Yes	No	No	No	No	No	No	No	No	No	No	Yes
18	Yes	Yes	Yes	No	Yes	No	No	Yes	Yes	No	No	No	Yes	No	Yes	No	No	No	No	Yes
19	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes	No	No	Yes	No	No	Yes
20	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	No	Yes	No	No	No	No	No	Yes	No	Yes
21	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No	Yes
22	Yes	Yes	Yes	No	Yes	No	No	Yes	No	Yes	Yes	Yes	No	Yes	No	No	No	No	No	Yes
23	Yes	Yes	Yes	No	Yes	Yes	Yes	No	No	No	Yes	Yes	No	Yes	Yes	No	No	No	No	Yes
24	Yes	Yes	Yes	No	Yes	No	Yes	No	No	Yes	No	Yes	No	No	No	No	No	Yes	No	Yes
25	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No	No	Yes	Yes	No	No	No	Yes	Yes	No	No	Yes
26	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No	Yes	No	Yes	Yes	Yes	Yes	No	No	Yes
27	Yes	Yes	Yes	No	No	No	No	No	Yes	Yes	Yes	Yes	No	No	Yes	No	Yes	No	No	Yes
28	Yes	Yes	Yes	No	No	No	No	No	Yes	No	Yes	No	Yes	Yes	No	No	Yes	Yes	No	Yes
29	Yes	Yes	Yes	No	Yes	Yes	Yes	No	No	No	No	No	No	No	No	Yes	No	No	No	Yes
30	Yes	Yes	Yes	No	No	Yes	No	Yes	Yes	Yes	No	No	No	Yes	No	No	No	No	No	Yes
31	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No	No	Yes	No	No	No	Yes
32	Yes	Yes	Yes	No	No	No	No	Yes	No	Yes	No	No	Yes	No	Yes	No	No	No	No	Yes
33	Yes	Yes	Yes	No	No	No	No	Yes	No	No	Yes	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes

Figure 2. A portion of the COVID-19 dataset

4. Rule mining for COVID-19 prediction

In this paper, using COVID-19 symptom dataset and employing FP growth algorithm, the corresponding association rules are extracted for early detection of

this disease based on the application of Rapid Miner Studio Educational 9.9.000 Software. In this software, there are options to select different operators which can perform varying functions ranging from data input to data analysis. Each operator has an input node and an output node through which the data is processed. These operators are combined together to perform a specific task. In order to extract the association rules using this software, the following steps are adopted:

a) Data input: The data is fed into the software through the Read CSV (comma-separated values) Operator. The output node of Read CSV operator is connected to the input node of FP-Growth Operator.
b) Finding the frequent itemsets: The FP-Growth Operator is utilized to find out the frequent itemsets in the dataset. The output of this operator is then connected to the input node of Create Association Rule Operator.
c) Extraction of the association rules: The Create Association Rule Operator extracts the corresponding association rules while considering the input in the form of frequent itemsets from FP-Growth Operator.
d) Displaying the results: The Create Association Rule Operator has two output nodes, i.e. ite-frequent item sets obtained and rul-association rules extracted. These nodes are finally connected to two result nodes to display both the frequent itemsets and association rules.

Figure 3 portrays the flow diagram representing the positioning of different operators in a logical way to provide the intended results. The FP growth algorithm thus generates frequent itemsets with size ranging from 1 to 5. The frequent itemsets of sizes 3, 4 and 5 with their corresponding support values for COVID-19 prediction are depicted in Tables 3-5 respectively. In Table 5, there is a frequent itemset of size 5, i.e. {Covid-19, Dry cough, Fever, Sore throat, Breathing problem} which signifies that frequent occurrence of these four symptoms would lead to COVID-19. A support value of 0.374 symbolizes that there are 37.4% patient cases having dry cough, fever, sore throat and breathing problem resulting in COVID-19 infection. The corresponding association rules developed by this software are provided in Table 6. In this table, 'Premises' signifies the antecedent and 'Conclusion' signifies the consequent of the association rule generated. Thus, the occurrence of any of these rules would lead to increased likelihood of this disease in a patient. In order to generate these rules, the values of minimum support and confidence are considered as 30% and 1 respectively. The minimum support value of 30% symbolizes that at least 30% of the patient database contains any of these nine rules. On the other hand, lift > 1 indicates the existence of meaningful relationships between the symptoms and COVID-19 prediction. The formation of these rules and interrelations among them are pictorially exhibited in Figure 4. In this figure, all the variables (symptoms) and rules are shown in different blocks. The block for each rule has the format: Rule X (Support of the rule X/Confidence of rule X) where X is the corresponding association rule number. It provides a visual information of the unique symptoms/factors for COVID-19 and inter-connections among the generated rules. Now, based on all these association rules, it can be concluded that there are six most important factors, i.e. breathing problem, fever, dry cough, sore throat, abroad travel and attended large gathering responsible for infection of this disease in a person.

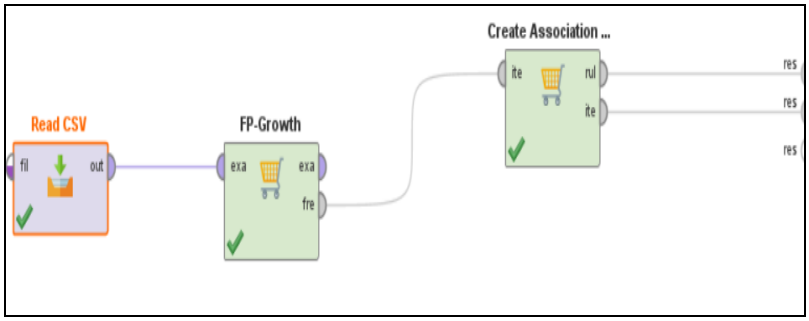

Figure 3. Flow diagram for extraction of association rules

Table 3. Frequent itemset of size 3

SIZE	SUPPORT	ITEM1	ITEM2	ITEM3
3	0.611	Covid-19	Dry cough	Fever
3	0.593	Covid-19	Dry cough	Breathing problem
3	0.530	Covid-19	Dry cough	Running nose
3	0.392	Covid-19	Dry cough	Fatigue
3	0.367	Covid-19	Dry cough	Visited public exposed places
3	0.399	Covid-19	Dry cough	Headache
3	0.349	Covid-19	Dry cough	Contact with COVID patient
3	0.415	Covid-19	Dry cough	Hyper tension
3	0.376	Covid-19	Dry cough	Asthma
3	0.353	Covid-19	Dry cough	Attended large gathering
3	0.424	Covid-19	Dry cough	Abroad travel
3	0.596	Covid-19	Fever	Sore throat
3	0.510	Covid-19	Fever	Breathing problem
3	0.384	Covid-19	Fever	Running nose
3	0.351	Covid-19	Fever	Fatigue
3	0.377	Covid-19	Fever	Visited public exposed places
3	0.411	Covid-19	Fever	Contact with COVID patient
3	0.360	Covid-19	Fever	Hyper tension
3	0.378	Covid-19	Fever	Attended large gathering
3	0.381	Covid-19	Fever	Abroad travel
3	0.526	Covid-19	Sore throat	Breathing problem
3	0.370	Covid-19	Sore throat	Running nose
3	0.376	Covid-19	Sore throat	Visited public exposed places
3	0.401	Covid-19	Sore throat	Contact with COVID patient
3	0.384	Covid-19	Sore throat	Attended large gathering
3	0.374	Covid-19	Sore throat	Abroad travel
3	0.376	Covid-19	Breathing problem	Contact with COVID patient
3	0.355	Covid-19	Breathing problem	Attended large gathering

Table 4. Frequent itemsets of size 4

SIZE	SUPPORT	ITEM1	ITEM2	ITEM3	ITEM4
4	0.519	Covid-19	Dry cough	Fever	Sore throat
4	0.433	Covid-19	Dry cough	Fever	Breathing problem
4	0.363	Covid-19	Dry cough	Fever	Contact with COVID patient
4	0.354	Covid-19	Dry cough	Fever	Abroad travel
4	0.447	Covid-19	Dry cough	Sore throat	Breathing problem
4	0.352	Covid-19	Dry cough	Sore throat	Contact with COVID patient
4	0.448	Covid-19	Fever	Sore throat	Breathing problem
4	0.362	Covid-19	Fever	Sore throat	Contact with COVID patient

Table 5. Frequent itemsets of size 5

SIZE	SUPPORT	ITEM1	ITEM2	ITEM3	ITEM4	ITEM5
5	0.374	Covid-19	Dry cough	Fever	Sore throat	Breathing problem

Table 6. Association rules generated for COVID-19 prediction

RULE NO.	PREMISES	CONCLUSION	SUPPORT	CONFIDENCE	LIFT
1	Abroad travel	Covid-19	0.451	1	1.24
2	Dry cough, Attended large gathering	Covid-19	0.390	1	1.24
3	Dry cough, Abroad travel	Covid-19	0.424	1	1.24
4	Fever, Attended large gathering	Covid-19	0.378	1	1.24
5	Fever, Abroad travel	Covid-19	0.381	1	1.24
6	Sore throat, Attended large gathering	Covid-19	0.384	1	1.24
7	Sore throat, Abroad travel	Covid-19	0.374	1	1.24
8	Breathing problem, Attended large gathering	Covid-19	0.355	1	1.24
9	Dr cough, Fever, Abroad travel		0.354	1	1.24

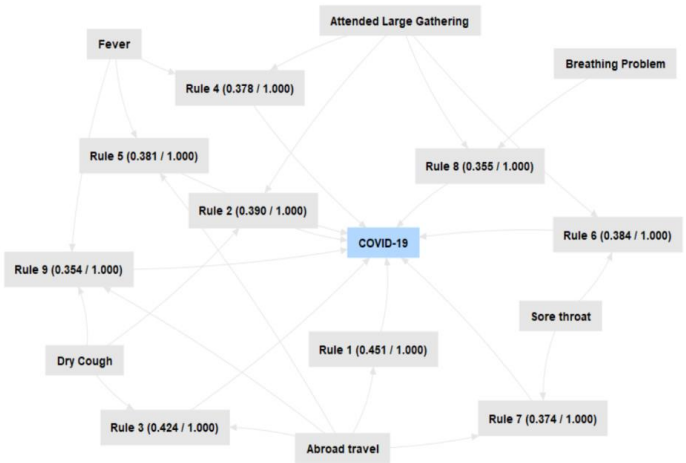

Figure 4. Formation of association rules

Considering the initial dataset, and breathing problem, fever, dry cough, sore throat, abroad travel and attended large gathering as the major factors for COVID-19 infection, a linear regression model is developed using the following steps:

a) Data input: The relevant data is fed into Rapid Miner software through Read CSV operator with COVID-19 as the decisional (dependent) variable.
b) Data preprocessing: The Replace Operator is employed to convert the data from 'yes' and 'no' to 1 and 0 respectively. The Guess Types Operator is used to transform all the variable data into numerical data.

c) Data processing: The Set Role Operator changes COVID-19 variable as a special attribute. It would help in developing the corresponding regression model considering COVID-19 as the dependant variable.
d) Model development: The Linear Regression Operator finally generates the regression equation from the given dataset.

The corresponding flow diagram, as exhibited in Figure 5, develops the regression equation correlating infection of COVID-19 and the main medical factors in the following form:

Y = 0.030 + (0.208×Breathing problem) + (0.175×Fever) + (0.243×Dry cough) + (0.193×Sore throat) + (0.189×Abroad travel) + (0.177×Attended large gathering)

where Y is the target variable (presence of COVID-19). A value of Y less than 0.5 signifies less likelihood of COVID-19 in a patient; on the other hand, a value greater than or equal to 0.5 identifies more likelihood of COVID-19 infection in a patient. A moderately high coefficient of determination (R^2) value as 0.739 provides an indication of acceptable accuracy of the developed predictive model. It indicates that almost 73.9% variation of the dependent variable (presence/absence of COVID-19) can be explained by the independent variables (symptoms/factors).

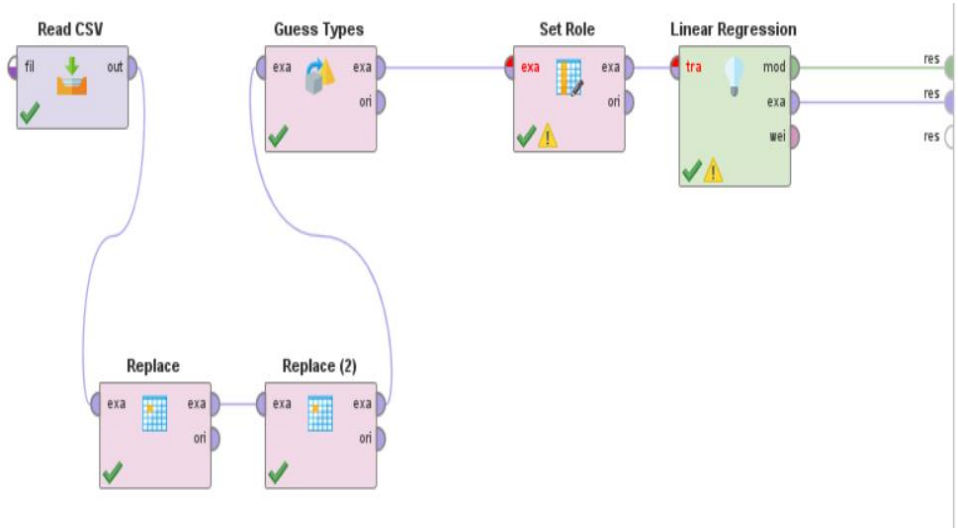

Figure 5. Flow diagram for regression equation

5. Conclusion

Keeping in mind the requirements of early detection of COVID-19 for faster isolation and treatment of an infected patient, this paper proposes the application of FP growth algorithm to find out the frequent itemsets and extract the association rues with their corresponding confidence and support values. It is noticed that six factors, i.e. breathing problem, fever, dry cough, sore throat, abroad travel and attended large gathering are mainly responsible for COVID-19 infection. A linear regression equation-based predictive model is also developed to correlate those factors and presence of COVID-19 in a patient. A moderately high coefficient of determination value suggests that almost 73.9% variation of the dependent variable (presence/absence of COVID-19) can be explained by the independent variables (symptoms/factors). It would help in early prediction of this disease, thus saving valuable time and resources. But, if a patient is asymptotic, this model would not

provide accurate results. Among the existing machine learning techniques, association rule mining has several advantages, like capability of dealing with different forms of data repositories, development of easy to understand clauses, no assumption about the underlying distribution of the data, application of support, confidence and lift parameters for developing the strongest rule etc. As a future scope, it is suggested to develop and integrate association rule mining in a decision support system for early diagnosis of COVID-19 and other severe diseases, like kidney related problems, brain tumor etc. With more real-time clinical dataset, those diseases can be diagnosed much faster while evaluating coexistence of the symptoms.

Author Contributions: V.K.R.: Data collection, software, analysis; S.C.: Draft preparation, review, technical writing; S.C.: Technical writing, editing. All authors have read and agreed to the published version of the manuscript.

Funding*:* This research received no external funding.

Data Availability Statement: The related data is collected from Kaggle.com which is the world's largest online data science community.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aiswarya, P., Bhanu Sridhar, M., & Kavitha, L. (2020). Detection and prediction of frequent diseases in India through association technique using apriori algorithm and random forest regression. International Journal of Engineering Research & Technology, 9(3), 386-393.

Alaiad, A., Najadat, H., Mohsen, B., & Balhaf, K. (2020). Classification and association rule mining technique for predicting chronic kidney disease. Journal of Information & Knowledge Management, 19(1), 2040015.

Anand Hareendran, S., & Vinod Chandra, S.S. (2017). Association rule mining in healthcare analytics. In: Data Mining and Big Data. Tan, Y. et al. (eds), Springer International Publishing, 31-39.

Brossette, S.E., Sprague, A.P., Hardin, J.M., Waites, K.B., Jones, W.T., & Moser, S.A. (1998). Association rules and data mining in hospital infection control and public health surveillance. Journal of the American Medical Informatics Association, 5(4), 373-381.

Burvin, J.S., & Dhanalakshmi, K. (2018). Pandemic disease detection and prevention system using mining with graph-based approach. International Journal of Pure and Applied Mathematics, 118(20), 4355-4360.

Çelik, A. (2020). Using apriori data mining method in COVID-19 diagnosis. Journal of Engineering Technology and Applied Sciences, 5(3), 121-131.

Cheng, C-W., & Wang, M.D. (2017). Healthcare data mining, association rule mining, and applications. In: Health Informatics Data Analysis. Health Information Science, Xu, D., Wang, M., Zhou, F., & Cai, Y. (eds), Springer, Cham, 201-210.

Domadiya, N., & Rao, U.P. (2018). Privacy-preserving association rule mining for horizontally partitioned healthcare data: a case study on the heart diseases. Sadhana, 43: 127.

Downs, S., & Wallace, M. (2000). Mining association rules from a pediatric primary care decision support system. In: Proceeding of the Annual Symposium of American Medical Informatics Association, Los Angeles, USA, 200-204.

Freeda, D.S. & Florence, M.L. (2017). An overview of disease analysis using association rule mining. International Journal of Scientific & Engineering Research, 8(4), 113-117.

Jahangir, I., Abdul, B., Hannan, A., & Javed, S. (2018). Prediction of dengue disease through data mining by using modified apriori algorithm. In: Proceedings of the 4th ACM International Conference of Computing for Engineering and Sciences, Kuala Lumpur, 1-4.

Jain, D., & Gautam, S. (2013). Implementation of apriori algorithm in health care sector: A survey. International Journal of Computer Science and Communication Engineering, 2(4), 26-32.

Jamsheela, O. (2021). Analysis of association among various attributes in medical data of heart patients by using data mining methods. International Journal of Applied Science and Engineering, 18(2), 2020215.

Kaur, J., &, Madan, N. (2015). Association rule mining: A survey. International Journal of Hybrid Information Technology, 8(7), 239-242.

Kulkarni, A.R., & Mundhe, S.D. (2017). Data mining technique: An implementation of association rule mining in healthcare. International Advanced Research Journal in Science, Engineering and Technology, 4(7), 62-65.

Lakshmi, K.S., & Vadivu, G. (2017). Extracting association rules from medical health records using multi-criteria decision analysis. Procedia Computer Science, 115, 290-295.

Ordonez, C., Ezquerra, N., & Santana, C.A. (2006). Constraining and summarizing association rules in medical data. Knowledge and Information Systems, 9(3), 259-283.

Prithiviraj, P., & Porkodi, R. (2015). A comparative analysis of association rule mining algorithms in data mining: A study. American Journal of Computer Science and Engineering Survey, 3(1), 1-10.

Sabthami, J., Thirumoorthy, K., & Muneeswaran, K. (2016). Mining association rules for early diagnosis of diseases from electronic health records. Middle-East Journal of Scientific Research, 24, 248-253.

Said, I.U., Adam, A.H., & Garko, A.B. (2015). Association rule mining on medical data to predict heart disease. International Journal of Science Technology and Management, 4(8), 26-35.

Sambasiva Rao, P., & Uma Devi, T. (2017). Applicability of apriori based association rules on medical data. International Journal of Applied Engineering Research, 12(20), 9451-9458.

Sarıyer, G., & Taşar, C. Ö. (2020). Highlighting the rules between diagnosis types and laboratory diagnostic tests for patients of an emergency department: Use of association rule mining. Health Informatics Journal, 26(2), 1177-1193.

Sengupta, D., Sood, M., Vijayvargia, P., Hota, S., & Naik, P.K. (2013). Association rule mining based study for identification of clinical parameters akin to occurrence of brain tumor. Bioinformation, 9(1), 555-559.

Shawkat, M., Badawy, M., & Eldesouky, A.I. (2021). A novel approach of frequent itemsets mining for Coronavirus disease (COVID-19). European Journal of Electrical Engineering and Computer Science, 5(2), 5-12.

Stilou, S., Bamidis, P.D., Maglaveras, N., & Pappas, C. (2001). Mining association rules from clinical databases: An intelligent diagnostic process in healthcare. In: Studies in Health Technology and Informatics, IOP Press, 84, 1399-1403.

Tandan, M., Acharya, Y., Pokharel., S., & Timilsina, M. (2021). Discovering symptom patterns of COVID-19 patients using association rule mining. Computers in Biology and Medicine, 131, 104249**.**

Thamer, M., El-Sappagh, S., & El-Shishtawy, T. (2020). A semantic approach for extracting medical association rules. International Journal of Intelligent Engineering & Systems, 13(3), 280-292.