Received: 25 December 2024 | Revised: 24 June 2025 | Accepted: 31 August 2025 | Published online: 13 October 2025

RESEARCH ARTICLE

Medinformatics
2026, Vol. 3(2) 185–193
DOI: 10.47852/bonviewMEDIN52025088

BON VIEW PUBLISHING

Data Mining in the COVID-19: Application of Association Rules to Analyze Epidemiological Data of Patients in Mexico

Bogart Yail Marquez[1,*], Raul Barutch Pimienta-Gallardo[2] and Arturo Realyvásquez-Vargas[3]

[1]*Department of Systems and Computing, Tecnológico Nacional de México, Mexico*

[2]*Department of Economic and Administrative Science, Tecnológico Nacional de México, Mexico*

[3]*Department of Industrial Engineering, Tecnológico Nacional de México, Mexico*

Abstract: This research uses data mining methodology to extract information from a large database containing 1,048,576 records of COVID-19 patients in Mexico during 2023. The Apriori algorithm is used to find the association rules and analyze the clinical and demographic variables. The target is to find out the chance of intubation by the factor interaction. After preprocessing the data, the Apriori algorithm employs a minimum support of 21.6% and a minimum confidence of 60% to meet the market basket needs. In all, 833 association rules are produced, which establish connections between variables like lab findings, intensive care unit admission, chronic obstructive pulmonary disease or asthma diagnosis, and intubation requirements. Certain rules show very dependable correlations with confidence levels of 99%. The results enable proactive medical interventions and resource management by offering crucial information for the early identification of high-risk individuals. The study demonstrates how data mining techniques may be used to retrieve important information from huge epidemiological databases. To find hidden relationships in a large database of COVID-19 patients in Mexico, this work effectively uses the Apriori algorithm, producing vital data for public health decision-making.

Keywords: COVID, disease management, data mining, association rules, epidemiological data

1. Introduction

This research delves into the analysis of massive COVID-19 data in Mexico for the year 2023, utilizing advanced data mining techniques, focusing on applying Apriori association rules [1–4]. Crucial elements that highlight the dynamic and changing character of the pandemic provide the foundation for the continued need to do COVID-19 analyses. Continuous data analysis is still essential as we negotiate the intricacies of a global health emergency. As new SARS-CoV-2 viral variations keep appearing, careful observation and investigation are required. Examining these variations' traits, transmissibility, and possible effects is essential for modifying public health tactics and guaranteeing vaccine efficacy [5]. Since COVID-19 vaccines have been widely used, it is crucial to do continuous research to evaluate their long-term effectiveness, spot any possible decline in immunity, and ascertain whether booster shots are necessary [6–8]. Population immunity is maintained, and vaccination tactics are informed by this analysis [9]. In addition, recent work has shown that association rule mining can reveal hidden patterns in COVID-19 mortality data [10].

Figure 1 shows a map of Mexico with the number of COVID-19 cases in each state indicated by the color intensity. The more examples there are in a given state, the more intense the hue.

Figure 2 displays Mexico's daily cases, cumulative instances, and confirmed accumulations for both men and women. The data was sourced from the Mexican government. Comparable applications of Apriori to public health data have been reported in Latin America, such as the analysis of symptom patterns in Brazil [11].

Data analysis provides public health officials with up-to-date information on the virus's spread, enabling evidence-based decisions such as implementing or adjusting preventative measures, allocating resources, and designing targeted interventions [12]. It is also essential for evaluating the effectiveness of therapies and strategies, as ongoing research continues to improve patient outcomes [13]. Beyond the clinical dimension, analyzing COVID-19 data is crucial for understanding the pandemic's broader societal and economic implications. This includes identifying vulnerable groups, assessing how well financial assistance programs are working, and guiding the creation of policies to mitigate long-term effects [14]. Moreover, international collaboration in COVID-19 data analysis fosters a shared understanding of the virus's behavior and the effectiveness of response strategies, encouraging the exchange of best practices and coordinated action [15]. By encouraging the sharing of best

*Corresponding author: Bogart Yail Márquez, Department of Systems and Computing, Tecnológico Nacional de México. Email: bogart@tectijuana.edu.mx

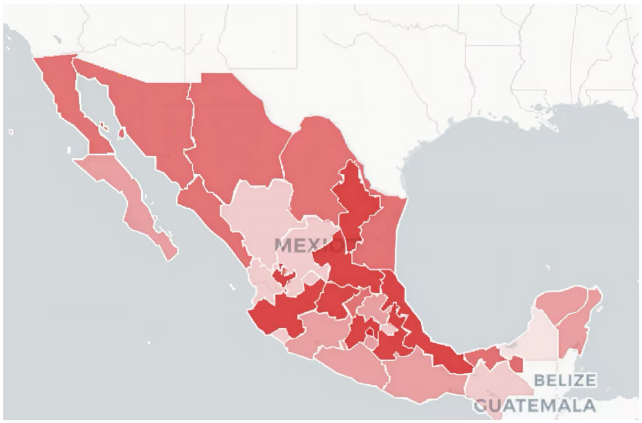

Figure 1. Mexico's COVID-19 cases visualized geographically

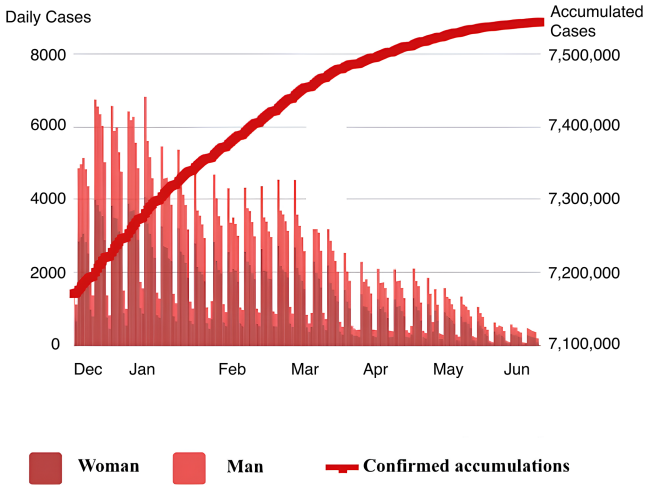

Figure 2. Evaluation of Mexico's daily and total COVID-19 cases

practices and lessons learned, collaborative analysis makes it possible to create a cohesive and well-coordinated strategy for handling the global health emergency. In addition to bolstering urgent responses, this strategy opens the door for long-term improvements in public health by combining resources, knowledge, and perspectives from other countries and organizations. Future pandemic preparedness and response plans can be strengthened with the help of the lessons learned from current COVID-19 improvements in global health policies and resource allocation; it is essential to comprehend transmission patterns, identify susceptible groups, and assess the efficacy of intervention efforts. This cooperative effort also supports the fair distribution of healthcare resources, strengthens early warning systems, and establishes strong surveillance systems. We can increase our ability to withstand new infectious diseases and lessen their effects on communities around the world by emphasizing data-driven decision-making and establishing international collaborations.

Figure 3 shows the kind of patients, differentiating between those who received outpatient care and those who were hospitalized; data was taken from the Government of Mexico [11]. Figure 4 shows the histogram of age and gender ranges for men and women.

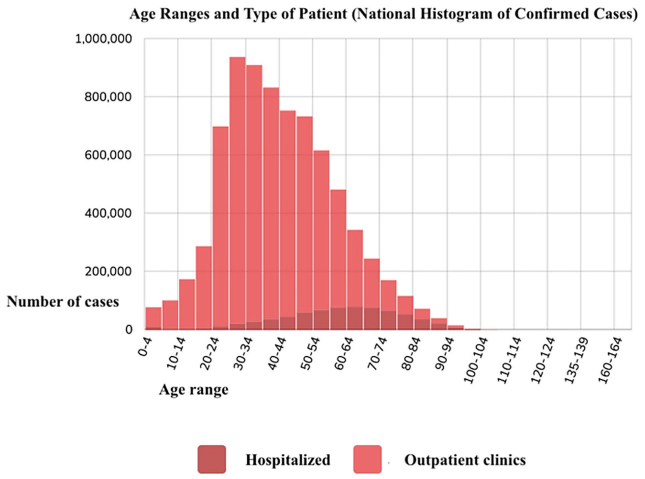

Figure 3. Using data analysis to develop flexible public health approaches

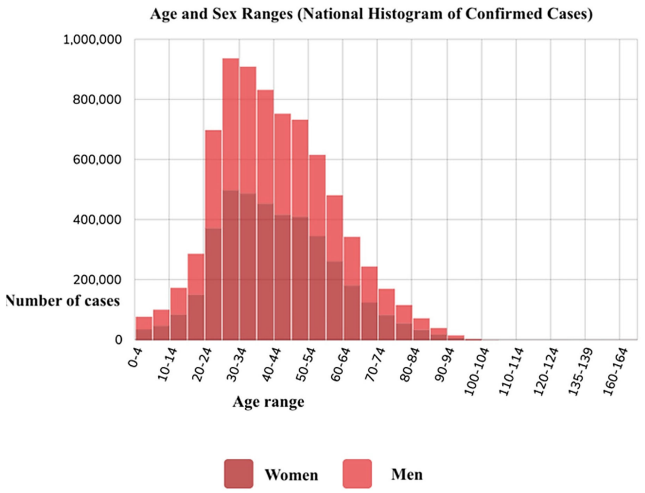

Figure 4. COVID-19 patient type and demographic analysis in Mexico

A persistent dedication to data analysis is required, considering the COVID-19 epidemic. This analytical method is essential for maximizing public health responses, adjusting policies to the virus's changing nature, and protecting the welfare of populations everywhere. The abundance of data produced by the pandemic offers a singular chance to investigate trends, connections, and patterns that are essential for well-informed public health decision-making. This study is based on a variety of factors, each of which clarifies a particular aspect of virus impact and spread.

The selected variables for analysis include:

1) INTUBATED: Identifies whether the patient required intubation.
2) PNEUMONIA: Indicates if the patient was diagnosed with pneumonia.
3) PREGNANCY: Identifies if the patient is pregnant.
4) DIABETES: Indicates if the patient has a diabetes diagnosis.
5) COPD (EPOC): Identifies if the patient is diagnosed with chronic obstructive pulmonary disease (COPD).
6) ASTHMA: Indicates if the patient has a diagnosis of asthma.
7) IMMUNOSUPPRESSION: Identifies if the patient presents immunosuppression.

8) HYPERTENSION: Indicates if the patient has a diagnosis of hypertension.
9) OTHER_COMORBIDITIES: Identifies whether the patient has been diagnosed with other diseases.
10) CARDIOVASCULAR: Indicates if the patient has a diagnosis of cardiovascular disease.
11) OBESITY: Identifies if the patient has a diagnosis of obesity.
12) CHRONIC_RENAL_FAILURE: Indicates if the patient has a diagnosis of chronic renal failure.
13) SMOKING: Identifies if the patient has a smoking habit.
14) OTHER_CASE: Indicates if the patient had contact with another case diagnosed with SARS-CoV-2.
15) LAB_SAMPLE: Identifies if a laboratory sample was taken from the patient.
16) ANTIGEN_SAMPLE: Indicates if the patient's antigen sample for SARS-CoV-2 was taken.
17) MIGRANT: Identifies if the patient is a migrant.
18) ICU (Intensive Care Unit): Indicates if the patient requires admission to an ICU.

The necessity to address several clinical, demographic, and behavioral factors that could affect the disease's severity and spread served as the foundation for the selection of these variables. When analyzed using the Apriori association rules technique, these variables—which are thought to be essential for a thorough understanding of the dynamics of COVID-19 in Mexico—promise to reveal important relationships and patterns that will support well-informed decision-making and the creation of successful public health initiatives.

One well-liked data mining tool for finding patterns of relationships between variables in big datasets is the Apriori association technique [16]. It was first proposed by Agrawal, Imielinski, and Swami in 1993 and is frequently utilized in applications like product suggestion, market analysis, and, in our example, the analysis of COVID-19 epidemiological data [17]. Datasets with transactions are subjected to the Apriori approach. The technique looks for relationships between the things that make up each transaction

A support threshold is established prior to the application of Apriori. The lowest frequency at which an association must appear in the dataset in order to be deemed significant is indicated by the support threshold. Recent applications, such as Sinisterra-Sierra et al. [18], demonstrate how this parameter remains central even in advanced models applied to COVID-19 datasets. Rules that fall short of this cutoff are eliminated. Items that satisfy the predetermined support criterion are identified at this step. Iteratively, it creates larger item sets from individual item sets at first, eliminating those that don't satisfy the support criteria [19].

Association rules are created from the frequently occurring item sets. "A -> B," where A and B are sets of objects, is the form of an association rule. Using confidence and support, the association's strength is evaluated. Support is the relative frequency of the link in the dataset, and confidence is the conditional probability that B will occur given that A has occurred. Only association rules that satisfy a predetermined confidence threshold are included in the created rules. This guarantees the selection of only statistically significant rules. Until now, larger item sets that satisfy the support threshold can be produced; the process of creating frequent item sets and association rules is repeated [20].

Item sets may indicate if a patient has certain risk factors, symptoms, or preventive measures in the context of COVID-19. Examples of patterns that could be identified using association rules are "Patients with hypertension are more likely to develop pneumonia." Large datasets can include hidden patterns that can be found with the use of the Apriori association rule technique. It has been useful in finding important connections in several domains, such as epidemiological data analysis.

Based on set theory, the Apriori association rule technique finds patterns in transactional datasets by utilizing confidence and support measures. The Apriori approach can be explained mathematically as follows:

Assume for the moment that a set of transactions D is made up of n different transactions:

Let's assume we have a set of transactions D consisting of n distinct transactions:

$$D = \{T1, T2, \ldots, Tn\} \quad (1)$$

Each transaction Ti is a set of distinct items:

$$Ti = \{Ii1, Ii2, \ldots, Iim\} \quad (2)$$

I is the set of all distinct items in D.

A frequent item is a set of items that appears with a frequency equal to or greater than the defined support threshold.

Generation of 1-Itemsets:

Initially, 1-ia temset ($L1$) is generated by counting the frequency of each item in D.

$$L1 = \{all\ items\ I\ with\ frequency\ \geq support\ threshold\}$$

Iteratively, frequent itemsets of size (Lk) are generated from frequent size $k-1$ ($Lk-1$) itemsets. This is done using the Apriori principle, which states that if an item is not frequent, its subsets are also not frequent.

$$Lk = \{X \cup Y \mid X \in Lk-1\ and\ Y \in Lk-1\ and\ X \cap Y = \emptyset\}$$

The frequent itemsets generate association rules with a defined confidence threshold.

Association Rule: X→Y with confidence Conf(X→Y) = Support(X∪Y)/Support(X)

Support (Support (Support(X)): The relative frequency of an item

Support (X) = Number of transactions where X appears / Total number of transactions in

Support(X) = Total number of transactions in D

Confidence (Conf (Conf(X→Y)): The conditional probability that Y occurs given that X has occurred.

$$Conf(X \rightarrow Y) = Support(X \cup Y) / Support(X)$$

The Apriori approach is used until no more significant itemsets that satisfy the support threshold can be produced. A collection of association rules outlining significant patterns in the transactional data is the end result.

2. Materials and Methods

Gathering COVID-19 epidemiology data for Mexico in 2023 is the first step in the technique. The information comes solely from the Open Data sets made available by the Mexican government's Ministry of Health's Directorate General of Epidemiology (DGE). These datasets, which include comprehensive information about each patient, including factors like the requirement for intubation, the diagnosis of pneumonia, age, pregnancy, smoking, and behaviors, are the main source of information. The careful gathering of COVID-19 epidemiological data for Mexico in 2023

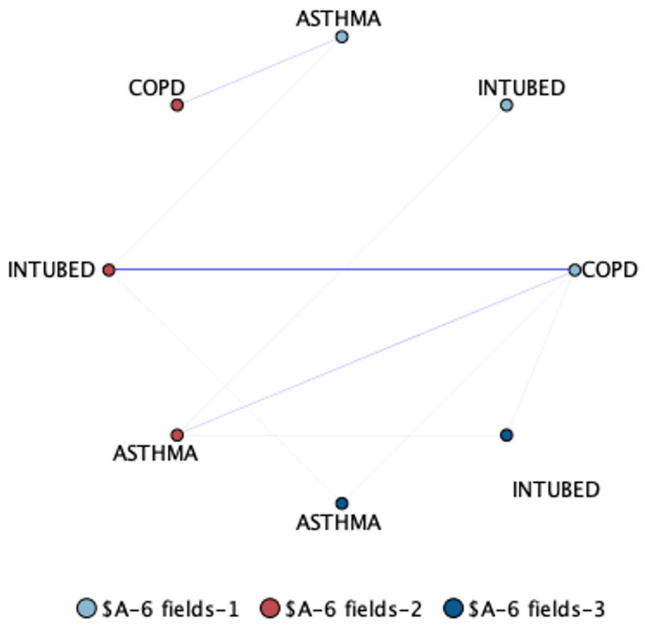

Figure 5. Investigating links between COPD and intubation using mesh plot insights

forms the basis of our technique. This method only entails retrieving and combining data from the Open Data sets that have been painstakingly supplied by the Ministry of Health of the Government of Mexico's DGE. As the authoritative body, the DGE provides clear and thorough datasets that capture the health status of a sizable population. With 40 characteristics carefully documented for each patient, the dataset is large, covering an astounding 1,048,576 patients, and is distinguished by its richness of information. These variables capture many facets of the demographics and status of patients. The characteristics mentioned in the introduction—such as the requirement for intubation, the diagnosis of pneumonia, age, pregnancy status, smoking habits, and other relevant factors—are the focus of our investigation. Figure 5 shows the mesh plot illustrates the relationship between Chronic Obstructive Pulmonary Disease (COPD) and intubation events, highlighting the frequency and co-occurrence patterns among patient records

The dedication to using this official government-provided dataset guarantees the uniformity, quality, and validity of the information. It also emphasizes how crucial it is for the research method to be transparent so that the scientific community may examine and replicate the findings. Normalization and the imputation of missing values are two prevalent issues in large datasets that are addressed by a thorough preprocessing process. Furthermore, variables of relevance are chosen, including age, pneumonia diagnosis, intubation requirement, and other factors mentioned in the introduction. Notably, the dataset includes 40 variables and 1,048,576 patients; nevertheless, we will concentrate on the variables listed in the introduction.

Use of the Apriori methodology: Using the painstakingly gathered dataset, the study moves on to the use of the Apriori association rule methodology, a complex approach frequently used in data mining to find patterns and correlations in big datasets. Based on association rule mining, the Apriori method was created specifically for transactional datasets [21]. Every patient in our situation is a transaction, and the variables in these records are regarded as objects. Finding correlations between various sets of variables and comprehending the consequences of these correlations considering COVID-19 dynamics are the objectives.

The dataset is rigorously preprocessed before the Apriori method is applied. To provide a clear and organized input for the Apriori algorithm, this entails resolving missing values, standardizing data, and encoding categorical variables [22]. The Apriori algorithm operates according to certain parameters. A 10% support level is established in our analysis. This indicates that an item is deemed noteworthy for additional investigation if it appears in at least 10% of the patient data. In order to capture relationships that are generally common in the dataset, this threshold was chosen.

Furthermore, 60% is the minimum rule confidence level. The minimal degree of confidence needed for an association rule to be deemed legitimate is determined by this parameter. A 60% confidence level means that, given the occurrence of the antecedent variables, the consequent variable is likely to occur with at least a 60% chance. There can be no more than five antecedents. By restricting the number of variables that can contribute to a rule's antecedent, this parameter limits the complexity of the rules that are formed. This parameter's setting aids in the creation of clear and understandable regulations. Frequent item sets on the specified support threshold are iteratively generated by the Apriori algorithm [23]. It finds combinations of variables that occur with the designated frequency by gradually expanding from individual variables to larger sets.

Following the Apriori principle, which directs the creation of rules based on their antecedent and consequent relationships, the association rules are obtained from the frequent itemsets. Each rule's confidence and support offer numerical indicators of its importance. By carefully setting these settings, the Apriori algorithm is guaranteed to find strong and significant correlations in the COVID-19 dataset. A more nuanced understanding of the complex associations between various health factors is made possible by the resulting association rules, which are informed by the established thresholds and provide important information for public health decision-making. The Apriori technique generates frequent itemsets iteratively once the dataset has been preprocessed and parameter thresholds have been defined. This crucial stage entails finding variable combinations that appear frequently enough in the dataset to serve as the foundation for the creation of association rules later. The Apriori algorithm repeatedly iterates through the dataset, gradually investigating different variable combinations. It starts with individual variables, determining the percentage of patient records that include the variable and determining its support. Following the Apriori principle, which states that if a set is frequent, then all of its subsets must likewise be frequent, subsequent iterations combine variables to create larger sets.

A key factor in these rounds is the 10% support level that has been set. Only items that are deemed frequent and kept for additional examination are those that exceed this level. The resulting itemsets are guaranteed to represent combinations of variables that occur frequently enough to be statistically significant thanks to this selective filtering. The technique efficiently generates candidate itemsets by using a level-wise approach. Finding common itemsets of size one (individual variables) is the first step, and then more important itemsets are gradually investigated. To maximize computing efficiency, the approach eliminates candidate itemsets that don't satisfy the support criterion at every level. One of the main steps in this stage is transaction scanning, in which the algorithm looks through the whole dataset to find instances of potential itemsets. This data is essential for figuring out whether

each item's support exceeds the predetermined threshold. The closure property, which stipulates that item subsets must likewise be frequent if an itemset is, is utilized by the Apriori method. This characteristic ensures that the frequently created itemsets meet statistical significance standards by directing the algorithm's systematic itemset identification and exploration.

The procedure gradually increases the itemset's size through a sequence of iterations until either the maximum specified itemset size is reached or no more expansion is feasible. This iterative expansion offers a thorough understanding of the interactions between various variables by capturing associations of various intricacies. By utilizing the natural structure of frequently occurring itemsets, the Apriori method avoids laborious searches and drastically lowers computational expenses. This effectiveness is particularly important when working with big datasets, like the vast COVID-19 dataset that is being studied. By establishing item setseter thresholds, the Apriori algorithm generates frequent itemsets iteratively. This crucial stage entails finding variable combinations that appear frequently enough in the dataset to serve as the foundation for the creation of association rules later.

The Apriori method gradually explores variable combinations by iterating over the dataset several times [24]. It starts with individual variables, determining the percentage of patient records that include the particular variable and determining its support. Following the Apriori principle, which states that if a set is frequent, then all of its subsets must likewise be frequent, subsequent iterations combine variables to create larger sets. A key factor in these rounds is the 10% support level that has been set. Only items that are deemed frequent and kept for additional examination are those that exceed this level. The closure property, which stipulates that item subsets must likewise be common if an itemset is, is utilized by the Apriori method. This characteristic ensures that the often created itemset adheres to statistical significance by directing the algorithm in its systematic identification and exploration of itemsets. The procedure gradually increases the itemset's size through a sequence of iterations until either the maximum specified itemset size is reached or no more expansion is feasible. A thorough understanding of the interactions between various variables is provided by this iterative expansion, which captures associations of various complexities [24, 25].

Specifically, the Apriori technique is used in our study to find association rules where the consequence "consecuent" variable icons consequencetubation. Because of this strategic emphasis, we may allocate resources and implement targeted clinical interventions by calculating the likelihood that a patient would be intubated based on the interaction of other variables.

"EPO" (chronic obstructive pulmonary disease) is the only antecedent in an illustrated association rule, which results in "Intubado." This criterion indicates that there is a high probability that a patient may need to be intubated if they are diagnosed with EPOC, a prevalent chronic lung illness marked by airflow restriction and breathing difficulty. The rule has a 97.57% support rate, meaning that almost all of the examined patient data show this correlation. Additionally, the 65.76% confidence level indicates that 65.76% of patients with an EPOC diagnosis underwent intubation. These measurements offer a quantitative basis for comprehending the connection between the likelihood of intubation and EPOC.

By concentrating on this correlation, the analysis emphasizes how crucial it is to identify and treat EPOC as soon as possible in order to lower the chance of serious respiratory issues that could require intubation. Furthermore, the findings highlight how data-driven methodologies such as Apriori can be used to uncover important risk variables in patient groups.

This rule also makes it possible to look more closely at how other factors that may affect the chance of intubation, like age, gender, comorbidities, and environmental factors, interact with EPOC. Patient outcomes and therapeutic decision-making can be greatly improved by recognizing and comprehending such complex connections. In the end, this analytical method advances our knowledge of the causes of intubation and opens the door to proactive, individualized treatment approaches.

3. Results

3.1. Generation of association rules

In order to derive association rules—a critical step in gleaning valuable insights from the COVID-19 dataset—frequent item sets must first be generated. In order to reveal possible patterns and dependencies, this step entails creating rules that record the interactions between various variables. The large dataset used in this investigation had 1,048,576 valid transactions (patient records), yielding 833 association rules. These guidelines offer a thorough understanding of the relationships in the context of COVID-19 in Mexico by illustrating the complex relationships between different health indicators.

The frequency with which a specific association occurs in the dataset is reflected in the support measure. The minimal observed support in this instance is 58.48%, which indicates that about 58.5% of patient records include even the least common link. However, the greatest support of 99.05% indicates that the relationships are highly widespread. These numbers offer helpful information about the frequency of various rules in the data. Confidence gauges how accurately one can forecast a rule's result when its condition is satisfied. The outcome happens approximately six times out of ten when the condition is present, according to the minimal confidence level of 62.94%. The highest confidence level, on the other hand, is 99.79%, indicating extremely trustworthy rule predictions. Finally, lift assesses the likelihood that a rule's result will occur given the circumstance as opposed to occurring by chance. The outcome's chance is somewhat increased, as indicated by the lowest lift value of 0.977. When the condition is satisfied, the odds show a slight improvement, as indicated by the highest lift of 1.063.

The usefulness of the discovered rules is further explained by deployment metrics. The corresponding regulation may not have a significant effect on the deployment of resources or interventions, as indicated by the minimum deployability value of 0.124%. In contrast, this rule is far more effective when evaluating deployment maximizing rules, with a maximum deployability of 35.28%. The lowest frequency of each rule is 58.33%, which is the minimum rule support in this case. This demonstrates how a significant portion of the dataset still contains even the least supported rules. However, because the most popular rules are the most widely accepted, the maximum rule support is 98.22%. These metrics provide a more thorough understanding of the relationships that the Apriori algorithm found. Rules show significant trends in how COVID-19-related factors interact, providing a useful basis for well-informed decision-making in clinical and public health settings.

The model was further adjusted to increase its predicted accuracy and dependability because the initial analysis's margin of error was considerable. This improved method's main goal is to ascertain the likelihood that a patient will need to be intubated, with the ultimate objective being to pinpoint important factors that

Table 1. Predictive models for intubation risk

Consequent	Antecedent	Support %	Confidence %
INTUBED	LAB_RESULT and ICU	53.02066803	99
INTUBED	ANTIGEN_RESULT and LAB_RESULT and ICU	22.10254669	99
INTUBED	LAB_RESULT and ASTHMA and ICU	51.98707581	99
INTUBED	LAB_RESULT and COPD and ICU	52.81820297	99
INTUBED	ANTIGEN_RESULT and LAB_RESULT and ASTHMA and ICU	21.67778015	99
INTUBED	ANTIGEN_RESULT and LAB_RESULT and COPD and ICU	22.01223373	99
INTUBED	LAB_RESULT and ASTHMA and COPD and ICU	51.84774399	99
INTUBED	ANTIGEN_RESULT and LAB_RESULT and ASTHMA and COPD and ICU	21.60320282	99
INTUBED	ANTIGEN_RESULT and LAB_RESULT and ASTHMA and COPD	21.83609009	98.93347542
INTUBED	ANTIGEN_RESULT and LAB_RESULT and COPD	22.25055695	98.92891125
INTUBED	ANTIGEN_RESULT and LAB_RESULT and ASTHMA	21.91772461	98.90524923
INTUBED	ANTIGEN_RESULT and LAB_RESULT	22.34916687	98.89651288
INTUBED	LAB_RESULT and ASTHMA and COPD	52.47612	98.80254865
INTUBED	LAB_RESULT and COPD	53.45983505	98.79978664
INTUBED	LAB_RESULT and ASTHMA	52.63767242	98.76400952
INTUBED	LAB_RESULT	53.68719101	98.75850651
INTUBED	ANTIGEN_RESULT and COPD and ICU	28.05624008	87.22122703
INTUBED	ANTIGEN_RESULT and ASTHMA and COPD and ICU	27.5642395	87.04780094
INTUBED	ANTIGEN_RESULT and ICU	28.46469879	86.33783847
INTUBED	ANTIGEN_RESULT and ASTHMA and ICU	27.93750763	86.19233579
INTUBED	ANTIGEN_RESULT and COPD	28.67183685	85.34854946
INTUBED	ANTIGEN_RESULT and ASTHMA and COPD	28.16467285	85.19205764
INTUBED	ANTIGEN_RESULT	29.10633087	84.4345712
INTUBED	ANTIGEN_RESULT and ASTHMA	28.56111526	84.31039952

require more focus. By being aware of these crucial elements, medical professionals can take prompt action to avoid intubation and deliver more efficient, proactive treatment.

The identified rules are strong and clinically relevant because the new model has been calibrated to satisfy a minimum confidence threshold of 80%. By making this modification, the model may better concentrate on association rules, with the subsequent variable "intubate" denoting whether a patient will need to be intubated. These guidelines facilitate better clinical decision-making by offering insightful information on the connections among different medical conditions, patient characteristics, and the risk of intubation.

A number of association rules that show the strongest connections between particular antecedent factors and the consequent variable of intubation are highlighted by the results in Table 1. These guidelines highlight the significance of comprehending multiple impacts on intubation results and are based on actual patient data. This analysis emphasizes the significance of classifying patients according to their risk profiles, as factors like age, comorbidities like diabetes or cardiovascular diseases, and pre-existing respiratory conditions (like COPD or asthma) frequently emerge as significant contributors to the probability of intubation. Targeted interventions, including improved monitoring, early respiratory support delivery, or preventive medications, might be given priority to high-risk individuals indicated by the model. By doing this, healthcare systems may more effectively manage resources and lower the frequency of critical care interventions like intubation.

More predicted accuracy is attained by the improved model, which also lays the groundwork for future improvements. To further improve forecasts, future iterations can include extra variables like socioeconomic determinants, environmental influences, or genetic predispositions. The dynamic character of data-driven healthcare and the possibility of ongoing predictive model improvement to promote improved patient outcomes are highlighted by this iterative approach. The analysis's conclusions give medical professionals and administrators practical advice and a better knowledge of the variables influencing intubation risk. The medical community may significantly improve the quality of treatment for patients at risk of serious respiratory problems and lessen the need for invasive interventions by utilizing these findings.

This shows that the rules investigate the connection between laboratory findings (LAB_RESULT) and the requirement for ICU admission as antecedent or prior conditions. This indicates that whether the patient was intubated (INTUBED)—support of 53.02%—is the final or ensuing outcome of interest in these criteria. The 53.02% support rate means that 53.02% of the cases in the dataset meet the set of requirements (LAB_RESULT and ICU as antecedents) and the consequent (INTUBED). As stated otherwise, this set of circumstances is comparatively typical. The 99% confidence level indicates that there is a substantial likelihood (99%) that the consequent (INTUBED) will also be satisfied when the conditions (LAB_RESULT and ICU) are met. This implies that the antecedent conditions and the desired outcome have a solid and trustworthy link. According to these association rules, there is a direct correlation between test findings and the necessity for ICU admission, and the likelihood of a patient being intubated is strongly correlated with this combination of diseases.

An investigation of the connections between laboratory results, COPD, and the need for ICU admission is indicated by the combination of LAB_RESULT, COPD, and ICU as antecedents. Whether the patient was intubated (INTUBATED) is the ensuing focus in these regulations. This combination of criteria occurs in a significant percentage of cases, with a support level of 52.81%. Intubation is almost 99% likely to occur when LAB_RESULT, COPD, and ICU criteria are met, according to the confidence level of nearly 99%, which denotes a very trustworthy link.

An important relationship between certain clinical circumstances and the risk of intubating a patient is revealed by the Apriori association rule. Antecedents in this relationship include the necessity for admission to the ICU, the presence of laboratory findings (LAB_RESULT), and the presence of asthma (ASTHMA). The regularity of this association is highlighted by the fact that this combination of conditions is seen in more than half of the dataset's cases, with a support of 51.99%. Additionally, the 99% confidence level indicates that there is a very strong correlation between these antecedents and the outcome of intubation (INTUBE). To identify patients at risk of clinical deterioration early and enable proactive and individualized medical therapies, consequence development is essential. In complex scenarios, these correlations offer valuable insights to maximize medical resources and improve care for patients in critical condition.

The degree of support slightly decreases when the COPD variable —which determines whether the patient has been diagnosed with COPD—is included. This suggests that the combination of conditions, including the presence of COPD, is present in a slightly lower percentage of cases in the dataset. The trust in the connection between the antecedent and subsequent situations does not change in spite of this decrease in support. This result implies that while COPD might be linked to intubation requirements, its influence on the total dataset is less pronounced. The confidence level's stability suggests that the variables' association endures, offering important information about the precise impact of COPD on the risk of intubation in COVID-19 patients.

According to the Apriori association rule, there is a 99% chance that a patient will be intubated (INTUBED) in roughly 22.10% of cases with positive antigen results, specific laboratory results, and a need for admission to the ICU. The significance of these symptoms as important markers of the severity of the disease in COVID-19 patients is highlighted by this observation. The high degree of confidence in this association implies that the likelihood of intubation is closely related to the co-occurrence of positive antigen results, certain laboratory results, and the requirement for intensive care unitization.

Data noise was addressed through a thorough preprocessing pipeline that included imputation of missing values, normalization of categorical variables, and exclusion of inconsistent records. These steps reduced unexplained variance and improved the accuracy of the generated rules. To optimize the computational efficiency of the Apriori algorithm, strategic thresholds for support (10%) and confidence (60–80%) were established, which limited the generation of redundant itemsets. Additionally, early pruning based on the Apriori principle was applied to prevent the exploration of infrequent item combinations, significantly reducing computational load in a dataset with over one million records. These measures enhanced both the integrity of the analysis and the operational scalability of the model, making it suitable for deployment in clinical and public health environments.

The COPD variable, which determines whether a patient has been diagnosed with COPD, shows a more noticeable color intensity in the mesh plot, which visually highlights relationships between variables through color intensity. Compared to other variables, this intensity color has a better correlation with the intubation state (INTUBATED). In essence, it suggests that the likelihood that a patient may need to be intubated is more strongly correlated with the existence of characteristics linked to COPD. Although the color intensity of the association between the ASTHMA and COPD variables is slightly lower than that of COPD, it is still noticeable. Patterns and correlations in the data can be quickly seen because of this color's obvious visual representation of important associations between variables. This competence is essential for making well-informed clinical decisions and helps to provide a thorough awareness of the variables affecting a patient's requirement for intubation when they have medical problems.

4. Conclusion

By finding multiple association rules in the enormous database of COVID-19 patients, the Apriori algorithm proved to be a useful tool in revealing the basic relationships between symptoms, comorbidities, and clinical outcomes. In particular, correlations between conditions like abnormal laboratory test results, ICU admission, diagnosis of asthma or COPD, and the need for intubation in COVID-19 patients were highlighted with an exceptionally high level of confidence (99%). Since these findings allow for the early identification of high-risk patients and highlight the potential need for intubation and intensive care, their practical significance cannot be disputed. This study demonstrates how useful machine learning and data mining methods are for drawing significant insights from massive datasets during a pandemic.

The association rules show complex connections between COVID-19 patients' clinical circumstances and their risk of intubation. With a 53.02% incidence probability and 99% confidence in predicting intubation, the study identifies substantial and extremely reliable correlations, such as abnormal laboratory findings and ICU admission. Furthermore, adding COPD preserves a steady and trustworthy correlation with intubation, albeit marginally reducing support. With a 51.99% support level and a 99% confidence level in predicting intubation, the guidelines also highlight the importance of particular clinical circumstances, such as asthma and ICU admission. In addition, there is a very high 99% chance of intubation in about 22.10% of cases with positive antigen results, particular laboratory findings, and ICU admission, highlighting their crucial function as important markers of illness severity.

The findings directly impact clinical practice by offering a sound basis for decision-making and underlining the use of this data in developing public health initiatives and distributing resources on a broad scale. The methods and procedures employed are generalizable, providing a strong strategy for evaluating COVID-19 epidemiological data globally, despite the focus being on patient data in Mexico. In addition to successfully extracting useful information from the medical records of COVID-19 patients in Mexico, this work emphasizes the critical role that data plays in medical decision-making by utilizing sophisticated data mining techniques, as well as the creation of public health regulations.

5. Future Work

As part of future work, we intend to integrate specific scenarios where the discovered association rules could inform clinical decision-making, such as in the triage of patients with comorbidities (e.g., COPD, asthma) within intensive care settings. Furthermore, we aim to apply these rules to real-world contexts, such as regional outbreaks or hospital resource crises, to demonstrate how combinations of variables (e.g., lab results + ICU admission) effectively predict the need for intubation. This contextualization would enhance the persuasiveness and practical utility of the proposed model, facilitating its adoption in healthcare protocols and resource management strategies.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors. The research exclusively used secondary epidemiological data publicly available through the official datasets of the Mexican Ministry of Health.

Conflicts of Interest

The authors declare that they have no conflicts of interest related to this work.

Data Availability Statement

The epidemiological dataset that supports the findings of this study is openly available at the Mexican Government's DGE repository: https://datos.covid-19.conacyt.mx/.

Author Contribution Statement

Bogart Yail Marquez: Conceptualization, Software, Validation, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Raul Barutch Pimienta-Gallardo:** Validation, Formal analysis, Resources, Writing – review & editing, Visualization, Funding acquisition. **Arturo Realyvásquez-Vargas:** Methodology, Validation, Investigation, Resources, Writing – original draft, Visualization.

References

[1] Dinov, I. D. (2023). *Data science and predictive analytics: Biomedical and health applications using R* (2nd ed.). Switzerland: Springer. https://doi.org/10.1007/978-3-031-17483-4

[2] Aljehani, S., & Alotaibi, Y. (2025). Preserving privacy in association rule mining using multi-threshold particle swarm optimization. *Information Sciences*, *692*, 121673. https://doi.org/10.1016/j.ins.2024.121673

[3] Kumar, S., Katiyar, V., & Katiyar, D. (2025). A review on data mining and techniques of clustering algorithms. In S. Mishra, A. Kumar Singh, & P. Prajapati (Eds.), *Challenges and opportunities for innovation in India* (pp. 223–227). CRC Press. https://doi.org/10.1201/9781003606260-41

[4] 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. https://doi.org/10.1016/j.compbiomed.2021.104249

[5] Malik, J. A., Ahmed, S., Mir, A., Shinde, M., Bender, O., Alshammari, F., ..., & Anwar, S. (2022). The SARS-CoV-2 mutations versus vaccine effectiveness: New opportunities to new challenges. *Journal of Infection and Public Health*, *15*(2), 228–240. https://doi.org/10.1016/j.jiph.2021.12.014

[6] Chi, W.-Y., Li, Y.-D., Huang, H.-C., Chan, T. E. H., Chow, S.-Y., Su, J.-H., Ferrall, L., ..., & Wu, T.-C. (2022). COVID-19 vaccine update: vaccine effectiveness, SARS-CoV-2 variants, boosters, adverse effects, and immune correlates of protection. *Journal of Biomedical Science*, *29*(1), 82. https://doi.org/10.1186/s12929-022-00853-8

[7] Krause, P. R., Fleming, T. R., Peto, R., Longini, I. M., Figueroa, J. P., Sterne, J. A. C., ..., & Boutron, I. (2021). Considerations in boosting COVID-19 vaccine immune responses. *The Lancet*, *398*(10308), 1377–1380. https://doi.org/10.1016/S0140-6736(21)02046-8

[8] Zhuang, C., Liu, X., Chen, Q., Sun, Y., Su, Y., Huang, S., ..., & Xia, N. (2022). Protection duration of COVID-19 vaccines: waning effectiveness and future perspective. *Frontiers in Microbiology*, *13*, 828806. https://doi.org/10.3389/fmicb.2022.828806

[9] Solante, R., Alvarez-Moreno, C., Burhan, E., Chariyalertsak, S., Chiu, N.-C., Chuenkitmongkol, S., ..., & Kiertiburanakul, S. (2023). Expert review of global real-world data on COVID-19 vaccine booster effectiveness and safety during the omicron-dominant phase of the pandemic. *Expert Review of Vaccines*, *22*(1), 1–16. https://doi.org/10.1080/14760584.2023.2143347

[10] Qian, X., Zuo, Z., Xu, D., He, S., Zhou, C., Wang, Z., ..., & Qian, Z. (2024). Demystifying COVID-19 mortality causes with interpretable data mining. *Scientific Reports*, *14*(1), 10076. https://doi.org/10.1038/s41598-024-60841-w

[11] Marques, J. G., Carvalho, B. M., Guedes, L. A., & da Costa-Abreu, M. (2024). Using association rules to obtain sets of prevalent symptoms throughout the COVID-19 pandemic: An analysis of similarities between cases of COVID-19 and unspecified SARS in São Paulo, Brazil. *International Journal of Environmental Research and Public Health*, *21*(9), 1164. https://doi.org/10.3390/ijerph21091164

[12] Gulzar, K., Ayoob Memon, M., Mohsin, S. M., Aslam, S., Akber, S. M. A., & Nadeem, M. A. (2023). An efficient healthcare data mining approach using Apriori algorithm: A case study of eye disorders in young adults. *Information*, *14*(4), 203. https://doi.org/10.3390/info14040203

[13] Vallée, A. (2023). Geoepidemiological perspective on COVID-19 pandemic review, an insight into the global impact. *Frontiers in Public Health*, *11*, 1242891. https://doi.org/10.3389/fpubh.2023.1242891

[14] Bambra, C., Riordan, R., Ford, J., & Matthews, F. (2020). The COVID-19 pandemic and health inequalities. *Journal of Epidemiology and Community Health*, *74*(11), 964–968. https://doi.org/10.1136/jech-2020-214401

[15] Abdalla, W., Renukappa, S., & Suresh, S. (2023). Managing COVID-19-related knowledge: A smart cities perspective. *Knowledge and Process Management*, *30*(1), 87–109. https://doi.org/10.1002/kpm.1706

[16] Xo'Jayev, O. Q., & Jumanazarov, A. D. (2023). Analysis of associative rules of sensor data based on the Apriori algorithm. *Academic Research in Educational Sciences*, *4*(5), 272–277.

[17] Poli, N. S., & Sikder, A. S. (2023). Predictive analysis of sales using the Apriori algorithm: A comprehensive study on sales forecasting and business strategies in the retail industry. *International Journal of Imminent Science & Technology*, *1*(1), 1–15. https://doi.org/10.70774/ijist.v1i1.1

[18] Sinisterra-Sierra, S., Godoy-Calderón, S., & Pescador-Rojas, M. (2023). COVID-19 data analysis with a multi-objective evolutionary algorithm for causal association rule mining. *Mathematical and Computational Applications*, *28*(1), 12. https://doi.org/10.3390/mca28010012

[19] Huang, D., Liang, T., Hu, S., Loughney, S., & Wang, J. (2023). Characteristics analysis of intercontinental sea accidents using weighted association rule mining: Evidence from the Mediterranean Sea and Black Sea. *Ocean Engineering*, *287*, 115839. https://doi.org/10.1016/j.oceaneng.2023.115839

[20] Xia, X. (2023). Learning behavior mining and decision recommendation based on association rules in interactive learning environment. *Interactive Learning Environments*, *31*(2), 593–608. https://doi.org/10.1080/10494820.2020.1799028

[21] Bhargava, M., & Selwal, A. (2013). Association rule mining using Apriori algorithm: A review. *International Journal of Advanced Research in Computer Science*, *4*(2), 327–330.

[22] Rogers, M. P., Janjua, H. M., Walczak, S., Baker, M., Read, M., Cios, K., . . . , & Kuo, P. C. (2024). Artificial intelligence in surgical research: Accomplishments and future directions. *The American Journal of Surgery*, *230*, 82–90. https://doi.org/10.1016/j.amjsurg.2023.10.045

[23] Vivekanandan, S. J., & Gunasekaran, G. (2023). A computation of frequent itemset using matrix based Apriori algorithm. *International Journal of Experimental Research and Review*, *30*, 247–256. https://doi.org/10.52756/ijerr.2023.v30.022

[24] Shu, X., & Ye, Y. (2023). Knowledge discovery: Methods from data mining and machine learning. *Social Science Research*, *110*, 102817. https://doi.org/10.1016/j.ssresearch.2022.102817

[25] Sigahi, T. F., & Sznelwar, L. I. (2024). Which complexity? A review of typologies and a framework proposal for characterizing complexity-based approaches. *Kybernetes*, *53*(4), 1374–1394. https://doi.org/10.1108/K-11-2022-1507

How to Cite: Yail Marquez, B., Pimienta-Gallardo, R. B., & Realyvásquez-Vargas, A. (2026). Data Mining in the COVID-19: Application of Association Rules to Analyze Epidemiological Data of Patients in Mexico. *Medinformatics*, *3*(2), 185–193. https://doi.org/10.47852/bonviewMEDIN52025088