International Journal of Combinatorial Optimization Problems and Informatics, 13(4), Dec 2022, 83-102.
ISSN: 2007-1558.

www.editada.org

Analysis of Bacterial Association Patterns that trigger Bacterial Vaginosis

Freddy de la Cruz-Ruiz[a], Juana Canul-Reich[b], Erick de la Cruz-Hernández[c] and Rafael Rivera-Lopez[d]*

.[a]División Académica de Ciencias y Tcnologías de la Información. Universidad Juárez Autónoma de Tabasco, Km. 1 Carretera Cunduacán-Jalpa de Méndez, Col. La Esmeralda, Cunduacán, 86690, Tabasco, México.
E-mail: freddy.delacruz@ujat.mx
[b]División Académica de Ciencias y Tcnologías de la Información. Universidad Juárez Autónoma de Tabasco, Km. 1 Carretera Cunduacán-Jalpa de Méndez, Col. La Esmeralda, Cunduacán, 86690, Tabasco, México.
***Corresponding author. E-mail: juana.canul@ujat.mx*
[c]División Académica Multidisciplinaria de Comalcalco. Universidad Juárez Autónoma de Tabasco, Ranchería Sur 4ta. sección. Comalcalco, Tabasco, México.
E-mail: erick.delacruz@ujat.mx
[d]Departamento de Sistemas y Computación. Tecnológico Nacional de México. Instituto Tecnológico de Veracruz, Av. Miguel Angel de Quevedo 2779, Formando Hogar, 91897 Veracruz, Veracruz, México.
E-mail: rrivera@itver.edu.mx

Abstract. Bacterial Vaginosis (BV) is a dysbiosis of the normal flora residing in the patient's vaginal mucosa. Objective: Running the apriori algorithm to mine association rules in a dataset that holds records of patients diagnosed with BV+. Method: To select the rules created with statistical significance the functions is.redundant, is.significant, and is.maximal were used. Also, eight quality metrics were used. Results: The best percentage of support to find the frequent itemsets was 7%. The confidence percentage to create the rules was 90%. The best metric was Fisher's exact test. The algorithm reported 58 rules. After selection with the functions and metrics, 17 rules were reported. Biological validation reduced the rules to 5. Rules reported that Atopobium vaginae, Gardnerella vaginalis, Megasphaera phylotype 1, and Ureaplasma parvum interact with each other to develop BV+. Conclusion: Knowing the bacteria (patterns) involved in the development of BV supports in the diagnosis of BV+.

Keywords. Bacterial Vaginosis, Machine learning, Association rules, Quality metrics, ARules package functions.

Article Info
Received June 27, 2022
Accepted Ago 16, 2022

1 Introduction

Bacterial vaginosis is a dysbiosis of the normal flora residing in the patient's vaginal mucosa. Both the populations of Lactobacillus and the populations of Gram - bacteria [1], have altered their growth pattern in patients diagnosed with bacterial vaginosis [2, 3].

This bacterial infection is characterized by the following symptoms: grayish-white leucorrhea, odor (fishy), and vaginal irritation [4]. Classical diagnostic tests include the Amsel's criteria and the Nugent scale [5, 6]. These tests consider the clinical conditions shown by the patient such as the bacterial morphotypes reflected through Gram staining [7]. However, these tests are subjective, and their evaluation greatly depends on the experience of the bacteriologist. Alternatively, real-time qPCR is used [8] to diagnose bacterial vaginosis. PCR (Polymerase Chain Reaction) is characterized by measuring the density of the bacterial growth associated with vaginosis [9, 10].

Also, bacterial vaginosis is studied with machine learning techniques [11–14]. Association rule algorithm is one of the machine learning techniques used to study diseases such as HIV and Chagas disease [15, 16] or to recommendation of medicine need [17]. To the best of our knowledge bacterial vaginosis has not been studied with association rules.

An association rule is an expression of the form: if X ⇒ Y, where X is named the antecedent or LHS (left-hand side) and Y is named the consequent or RHS (right-hand side) [18]. The rule means that if we find all items in X in a transaction in the dataset, it is likely that the transaction also contains the items in Y. Often, rules are restricted to only a single item in the consequent to limit the large number of rules that are generated in the phase of combinatorial explosion of rules produced with the apriori algorithm [19]. Apriori generates rules based on the restriction of a minimum threshold for the support and confidence [20].

These are basic metrics that lead the apriori algorithm to generate the rules. The inconvenience with apriori algorithm is that it generates a large number of association rules, so the problem with quality and quantity is present. To solve this problem, quality measures can be used to filter or sort-discovered rules. There exist two types of metrics, objective and subjective. “Objective measures are said to be data-driven and consider only the data cardinalities. Subjective measures are user-driven in the sense that they consider the user’s apriori knowledge and goals” [21]. In this study, greater weight is given to the objective metrics such as hyper-confidence [22, 23], hyper-lift [22, 23], lift [23, 24], conviction [23, 24], cosine [23, 25], the Gini index [23, 25], Fisher’s exact test [22, 23], and rulew power factor [23, 26]. The performance of each of them is analyzed since there are no ideal metrics, only metrics that work according to the needs of the user.

The use of association rules can improve the obtaining of hidden knowledge. So in this study, we interpret all items in the antecedent of an association rule as the coexistent bacteria to trigger bacterial vaginosis (consequent of the rule). Therefore, the biological meaning of the association rules in this work is that they describe the bacteria that interact with each other to develop bacterial vaginosis. This study uses association rule algorithm to mine a real dataset that holds data of patients diagnosed with VB+. Also, quality metrics to evaluate those rules that model the interactions between bacteria associated with bacterial vaginosis are used to discard non-relevant rules.

This paper is organized as follows. Section 2 describes the related work. Section 3 describes materials and methods used in this research. Presents the experimental design of the study. Section 4 describes the results obtained and discussion derived from this research. Section 5 presents conclusions drawn from this research.

2 Related work

To study bacterial vaginosis, Baker et al [12] analyzed a dataset with 1601 instances and 418 attributes. Instances are divided into three subcategories: time series, clinical and medical data. The time series count the study time. The clinical data treat a questionnaire in which risk factors for vaginosis and Amsel’s criteria are investigated, and the medical data taxonomically identify the species of bacteria associated with vaginosis based on the rRNA sequence of 16 Svedberg.

They used attribute selection algorithms: CfsSubsetEval, ClassifierSubsetEval, ConsistencySubsetEval, FilteredSubsetEval, WrapperSubsetEval, and to these they added the following search methods: BestFirst, RankSearch, GeneticSearch, LinearForwardSelection, GreedySelection, SubsetSizeForwardSelectionwise.

Once the most significant subgroups of attributes were formed, they chose the classification algorithms: A1 (Bagging), A2 (RBFNetwork), A3 (J48), A4 (NaïveBayes), A5 (AdaBoost.M1), A6 (RandomForest), A7 (LogitBoost), A8 (Kstar (K *)), and A9 (FT).

Cross validation was repeated 10 times. All algorithms were run in the Weka tool environment. The authors used a combination of five feature selection algorithms, six search methods, and three classifier algorithms (used for wrapper methods) assembled to create 20 distinct feature selection sets. Based on the execution time (0: 00: 02), reduction of the number of attributes (14), and sensitivity of 92%; the authors conclude that the algorithm FS16 A9 (Attribute Evaluator: WrapperSubsetEval; Search Method: GreedyStepwise; Classifier (For Wrappers): NaïveBayes) is the best algorithm to investigate the problem of bacterial vaginosis. They also demonstrate the feasibility of studying bacterial vaginosis with machine learning techniques.

Continuing the same line of research, Baker et al [12] using only the clinical and medical data from the dataset (Ravel et al., 2011), they analyze the ability to predict the class based on clinical attributes (Amsel’s criteria) or medical attributes (OTU regions based on the 16 S rRNA sequence). To achieve their goals, they use the following attribute selection algorithm: WrapperSubsetEval that uses a classifier: oneR, Bagging, NaïveBayes, to determine the subset of attributes. Cross-validation is applied to approximate the precision of the learning scheme. They use the search methods: BestFirst, GeneticSearch, SubsetSizeForwardSelection, LinearForwardSelection.

The classification algorithms used were the following: Bagging, RandomForest, NaiveBayes, RBFNetwork. The metrics used were the following: accuracy (AC), precision (PR), sensitivity (RC) and measure F (FM). They used a combination of five feature selection algorithms, six search methods, and three classifier algorithms assembled to create 20 distinct feature selection sets. In addition, they selected nine classification algorithms for their experiments. Considering the accuracy 95.7527%, execution time (0:00:01), reduction in attributes, and sensitivity 0.847%, the authors conclude that the WNLR algorithm (Wrapper Subset Eval / Linear Forward Selection / NaïveBayes FSS with RBF Network) and the medical datasets are the best to understand the development of vaginosis bacterial.

Alternatively, Beck and Foster [14], to classify the microbial communities in the BV+ and BV- categories, they used three machine learning techniques including genetic programming (GP), random forests (RF), and logistic regression (LR). The interest of the authors in the classification model is to analyze the accuracy of the classification since that will determine how well the samples are classified in the mentioned categories. The above-mentioned techniques were applied to the dataset published by Ravel et al. in 2011, and consists of 396 patients of which 97 were BV+ using the definition of the Nugent scale. These authors classified the microbial communities by amplifying and sequencing the variable region V1 - V2 of the 16S rRNA gene. They also used the Srinivasan et al. dataset, which consists of 220 patients, 97 of whom were BV+ using Amsel's criteria. These authors classified the microbial communities by amplifying and sequencing the variable region V3 - V4 of the 16 S rRNA gene. The authors found that RF and LR classify the vaginosis class with an accuracy between 90% and 95% and mainly when it comes to the dataset in which the diagnosis of BV+ was made with the Nugent scale. They are also faster in terms of execution time compared to GP. The authors argue that this study demonstrates the feasibility of using classification models to identify important microbial communities related to BV.

The association rules have also been used to study diseases such as Chagas disease, and HIV. Marchan et al [15] used association rules to investigate risk factors for transmission of Chagas caused by Trypanosoma cruzi. They used the cross-industry standard data mining process CRISP-DM and the Arules library of the R statistical package with its apriori function. This machine learning technique creates rules based on the constraint exerted by support and confidence metrics, and in this way it reveals hidden knowledge patterns in transactional databases.

The dataset that they analyze is made up of data from 293 families according to epidemiological characteristics. They performed a random serological diagnosis on 88 individuals and determined the presence of seropositivity or absence of seronegativity of total IgM, IgA, and IgG anti-Trypanosoma antibodies.

By applying the apriori function of the Arules package, it is possible to predict and associate in 93% and 100% multiple risk factors for a serology positive and negative, respectively. Only 2 factors were determined by the conventional Chi-square method. They conclude that the association rules can show hidden relationships between some elements of the variables. The use of association rules can improve the obtaining of hidden knowledge as compared to the classic use of variable selection techniques.

To study patients with HIV/AIDS, Fernandez et al [16], also used association rules. They performed the extraction of association rules using the apriori algorithm and the KDD process. They analyze a database of clinical and administrative records of HIV-infected patients from July 1980 to March 2006. The database is 155 MB in size and consists of 111 tables containing information on 6277 patients. When preprocessing the mentioned tables, a table with 6277 transactions with 17 items per transaction was obtained. They used the free software tool ARView developed specifically for extracting association rules, programmed in Java.

In the first analysis, with the default configuration (minimum confidence of 80%, coverage between 10% and 100%, number of items between 1 and 5) and without imposing any restriction, 14,203 association rules were obtained. To reduce the number of rules, they varied the parameters and incorporated restrictions.

The authors argue that this study is an approach to the problem of extraction of associations between variables. They report how the use of data mining techniques can lead to the extraction of patterns, confirming in some cases the knowledge we have about the diseases and opening possible pathways of medical research.

3 Materials and Methods

3.1. Dataset details

We used the dataset provided by [8]. This study was conducted between August 2016 and October 2018 in Tabasco, a state in the southeastern region of Mexico. The population under study comprised sexually active women aged from 18 to 50 years that underwent their annual gynecological inspection routine at the Laboratory of Research in Metabolic and Infectious Diseases, Universidad Juarez Autonoma de Tabasco.

The dataset is made up of 201 observations and 58 variables. All variables in the dataset are numeric except the variable ID and Cytology which are categorical. There exist three classes in the dataset: the class for positive vaginosis cases (51), the class for negative vaginosis cases (134), and the class for indeterminate vaginosis cases (16).

3.2. Apriori algorithm

Apriori algorithm allows for mining frequent itemsets, maximal frequent itemsets, closed frequent itemsets, and association rules. This algorithm is one of the most widely used and has two stages [29]:
First, to identify all itemsets that occur with a frequency above a certain limit (frequent itemsets). The mentioned limit is the support threshold that must be investigated from successive runs of the apriori algorithm on the dataset under study. Settled down the appropriate support threshold to investigate the frequent itemsets, the apriori function returns the frequent itemsets, Table 1, represents 8 examples of frequent itemsets.

Table 1 represents the frequent itemsets mined with the minimum support threshold of 7%. The first stage of the apriori algorithm consists in investigating these sets. The frequent itemsets are all those that exceed the minimum threshold established by the support.

Table 1. Frequent itemsets that exceed the 7% support threshold

itemsets †
[1]{*GardnerellaPos, gasseriUdetectable, UreaplasParPos*}
[2]{*GardnerellaPos, gasseriUdetectable, MycoplasHomiNeg*}
[3]{*crisLowGrowthDensity, jensHighGrowthDensity, UreaplasUreaNeg*}
[4]{*crisLowGrowthDensity, jensHighGrowthDensity, MycoplasmaGeniNeg*}
[5]{*crisLowGrowthDensity, GardnerellaPos, MegasphaeraNeg*}
[6]{*crisHighGrowthDensity, GardnerellaPos, inersHighGrowthDensity*}
[7]{*AtopobiumPos, MegasphaeraNeg, UreaplasParNeg*}
[8]{*AtopobiumPos, inersLowGrowthDensity, MycoplasmaGeniNeg*}

†First stage of the *apriori* algorithm.

Second, to convert those frequent itemsets into association rules [18]. Similarly to the support threshold, the appropriate confidence threshold was investigated from successive runs of the apriori algorithm to extract rules. Once the appropriate confidence threshold is established, the algorithm returns the set of rules that describe the patterns of behaviors among the items.

3.3. Association rule mining

The problem of association rule mining is defined as:
Let I = {i1 , i2 , ..., in} be a set of n binary attributes called items. Let D = {t1 , t2, ..., tm} be a set of transactions called the database. Each transaction in D has a unique transaction ID and contains a subset of items in I. A rule is defined as an implication of the shape: $X \Rightarrow Y$, where X, Y, $\subseteq$ I and $X \cap Y = \emptyset$[18]. An association rule is made up of two sets of items joined by an implication (if ⇒ then). The set to the left of the arrow is named antecedent (left hand side) and the set to the right of the arrow is named consequent (right hand side). Below, is a rule-shaped model that describes the association between antecedent (lhs) and consequent (rhs).

[1] {*AtopobiumPos, GardnerellaPos, UreaplasParPos*} ⇒ {*VaginosisPos*}

To mine the association rules, we used the programming language R version 3.6.3 on GNU/Linux OS distro openSuSe Leap version 15.2 with the ARules library that implements the apriori algorithm of the statistical package version 1.6.6. The KDD process [27] was implemented for extracting association rules from our Vaginosis dataset. To analyze the extracted rules and graphically represent them, we used the package arulesViz version 1.3.3 [28].

3.4. Data Preprocessing

In section 3.1, it was mentioned that all variables of the dataset under study were numeric. Therefore, all variables were transformed from numerical variables into categorical variables, since it is the type of variable suitable for extracting association rules. For example, originally our dataset comes with two crispatus variables, one quantitative and the other qualitative. Both variables were transformed as follows:

- Crispatus qualitative variable: original name Lactobacilluscrispatus<20 with values as 1 for the presence and 2 for absence. The variable was renamed to crispatus with categorical values as crispatusPresent instead of 1, and crispatusAbsent instead of 2.
- Crispatus quantitative variable: original name crispatusCq with continuous values starting with 0.0. It was divided into three categories according to the quantitative value of the variable. For the Cq value equal to 0.0, we used the variable named as undetectable, for the value Cq < 25 we used the variable named as high growth density and for the value of Cq > 25 we used the variable named as low growth density.

Variables in the dataset that were not associated with vaginosis were discarded, and only those variables fulfilling this condition were retained and are shown in Table 2.

Table 2. Variables selected for the mining process of association rules from the dataset under study

Variable	Description
AGE30	Age divided into < 30 and > 30
1. *Megasphaera,* 2. *Atopobium,* and 3. *Gardnerella*	1. *Magasphaera type 1*, 2. *Atopobium vaginae,* and 3. *Gardnerella vaginalis* Classified as positive or negative during diagnosis.
VBPCR	Vaginosis diagnosis by PCR (Polymerase Chain Reaction). Classified as positive, negative or indeterminate vaginosis.
1. MH, 2. MG, 3. UP, and 4. UU	1. *Mycoplasma hominis*, 2. *Mycoplasma genitalium*, 3. *Ureaplasma parvum*, and 4. *Ureaplsma urealyticum*. Classified as positive or negative during diagnosis.
CrsipatusCqRange, GasseriCqRange, JenseniiCqRange, and InersCqRange	Cq (Cycle quantification in thermocycler) value for which growth density is detected. Classified as undetectable, low, and high growth density.

3.5. Basic metrics

3.5.1. Support

Support is used to measure significance (importance) of an itemset. Since it uses the count of transactions, it is called a frequency constraint. An itemset with support greater than a set minimum support threshold, supp (X) > σ, is called a frequent or large itemset [20, 23].

$$Supp(X) = \frac{|\{t \in D; X \subseteq t\}|}{|D|} = \frac{cX}{|D|} = P(X)$$

where X = frequent itemsets, cX = represents the number of transactions that contain all items in x, D = dataset that contains each transaction (t), and P = is the probability that a certain itemset will occur.
Range: [0, 1]
The value of the support will depend on the dataset under study, for example in unbalanced datasets the support is set with low values if it is the minority class.

3.5.2. Confidence

Sets how useful X is to predict the presence of Y. Confidence is direct and gives different values for the rules X⇒Y and Y⇒X. Association rules satisfy a minimum confidence constraint, conf (X⇒Y) > γ [20, 23].

$$Conf(X \Rightarrow Y) = \frac{supp(X \Rightarrow Y)}{supp(X)} = \frac{supp(X \cup Y)}{supp(X)} = \frac{c_{XY}}{c_X} = \frac{P(X \cap Y)}{P(X)} = P(Y \mid X)$$

where X = itemset of the left-hand side (antecedent), Y = itemset of the right-hand side (consequent), cXY, cX = is the event in that a transaction contains items X and Y, and P = estimate of the conditional probability of Y given X.
Range: [0, 1]
Since confidence reflects the strength of the rule, it is recommended that it be set to a value close to 1.

3.6. Metrics of quality

One problem with the process of generating rules is the large amount of them that the apriori algorithm produces. To solve this problem, we used metrics of quality to filter out those considered to be of low-quality or said the other way around, to keep those of high-quality.

3.6.1. Hyper-Confidence

The confidence level for observations with too high or low counts for rule X⇒Y it uses the hypergeometric model. Since the counts are drawn from a hypergeometric distribution (represented by the random variable CXY with known parameters given by the counts cX and cY); we can calculate a confidence interval for the observed counts cXY stemming from the distribution [22, 23]. Hyper-confidence reports the confidence level as

$$hyper - conf(X \Rightarrow Y) = 1 - P[C_{XY} \geqslant c_{XY} \mid c_x, c_Y]$$

where X = itemset of the left-hand side (antecedent), Y = itemset of the right-hand side (consequent), CXY = a random variable representing a hypergeometric distribution, and cX and cY = represent the count of each item.
Range: [0, 1]
A confidence level of, e.g., > 0.95 indicates that there is only a 5% chance that the high count for the rule has occurred randomly. Note that hyper-confidence is equivalent to the statistic used to calculate the p-value in Fisher's exact test.

3.6.2. Hyper-Lift

The adaptation of the lift measure is more robust for low counts using a hypergeometric count model [22, 23]. Hyper-lift is defined as

$$hyper - lift_{\delta}(X \Rightarrow Y) = \frac{c_{XY}}{Q_{\delta[C_{XY}]}}$$

where X = itemset of the left-hand side (antecedent), Y = itemset of the right-hand side (consequent), cXY = is the number of transactions containing X and Y, and Qδ[CXY] is the quantile of the hypergeometric distribution with parameters cX and cY given by δ (typically the 99 or 95% quantile).

Range: [0, ∞] (1 indicates independence)

Due to the adjustment of this metric, it reports values close to 1, which represents acceptable values regardless of the characteristics of the dataset.

3.6.3. Lift

The lift measures how many times more X and Y occur together than expected if they were statistically independent [23, 24]. Lift is defined as

$$lift(X \Rightarrow Y) = lift(Y \Rightarrow X) = \frac{conf(X \Rightarrow Y)}{supp(Y)} = \frac{conf(Y \Rightarrow X)}{supp(X)} = \frac{P(X \cap Y)}{P(X)P(Y)}$$

where X = itemset of the left-hand side (antecedent), Y = itemset of the right-hand side (consequent), Supp (X) = frequent itemsets, Supp (Y) = frequent itemsets, P(X ∩ Y) = probability of occurrence of transactions containing X and Y, P(X) = probability of occurrence of transactions containing X, and P(Y) = probability of occurrence of transactions containing Y.
A lift value of 1 indicates independence between X and Y.

Rare itemsets with low counts (low probability), which by chance occur a few times (or only once) can produce enormous lift values.

Range: [0, ∞] (1 means independence)

If the lift is > 1, that lets us know the degree to which those two occurrences are dependent on one another, and makes those rules potentially useful for predicting the consequent in future data sets.
If the lift is < 1, that lets us know the items are substitute to each other. This means that the presence of one item has a negative effect on the presence of the other items and viceversa.

3.6.4. Conviction

The conviction is a measure that evaluates the degree to which the antecedent term influences the occurrence of the consequent term of an association rule [23, 24]. It is defined as

$$conviction(X \Rightarrow Y) = \frac{1 - supp(Y)}{1 - conf(X \Rightarrow Y)} = \frac{P(X)P(\bar{Y})}{P(X \cap \bar{Y})}$$

where X = itemset of the left-hand side (antecedent), Y = itemset of the right-hand side (consequent), Ȳ = E¬Y is the event that Y does not appear in a transaction, and P(X) = probability of occurrence of X.
Range: [0, ∞] (1 indicates independence; rules that always hold have ∞)
A high conviction value means that the consequent is highly dependent on the antecedent. For instance, in the case of a perfect confidence score, the denominator becomes 0 (due to 1 - 1) for which the conviction score is defined as
“inf”. Similar to lift, if items are independent, the conviction is 1.

3.6.5. Cosine

Cosine is the geometric mean between interest factor (I) and the support measure (s), which is a widely-used similarity measure for vector-space models. It is used to measure the similarity between LHS and RHS [23, 25]. It is defined as

$$cosine(X \Rightarrow Y) = \frac{supp(X \cup Y)}{\sqrt{supp(X)supp(Y)}} = \frac{P(X \cap Y)}{\sqrt{P(X)P(Y)}} = \sqrt{P(X \mid Y)P(Y \mid X)}$$

where X = itemset of the left-hand side (antecedent), Y = itemset of the right-hand side (consequent), Supp (X ∪ Y) = computes the support of the combined itemset, √supp(X)supp(Y) = square root of the multiplication of the support of the antecedent with the support of the consequent, P(X ∩ Y) = probability of occurrence of transactions containing X and Y, √P(X)P(Y) = square root of the multiplication of the probabilities of occurrence of the antecedent with that of the consequent, and √P(X|Y) P(Y|X) = square root of the multiplication of the conditional probabilities of X given Y and Y given X.
Valid values lie in the range [0, 1], where a value from 0.0 to 0.5 means no correlation, and from 0.51 to 1 means existing correlation.

3.6.6. Gini index

Gini index or Gini impurity measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen [23, 25]. It is defined as

$$gini(X \Rightarrow Y) = P(X)[P(Y \mid X)^2 + P(\bar{Y} \mid X)^2] + P(\bar{X})[P(Y \mid \bar{X})^2 + P(\bar{Y} \mid \bar{X})^2] - P(Y)^2 - P(\bar{Y})^2$$

where X = itemset of the left-hand side (antecedent), Y = itemset of the right-hand side (consequent).
This metric is defined in terms of the probabilities estimated from a 2 * 2 contingency table.
P(X) = probability of occurrence of X, P(Y|X)2 = probability of occurrence of Y given X, P(Ȳ|X)2 = probability that event Y will not appear in a transaction given X, P(X̄) = probability of occurrence that event X will not appear in the transaction, P(Y|X̄)2 = probability of occurrence that event Y will appear in the transaction given that even X will not appear in the transaction, P(Ȳ|X̄)2 = probability of no occurrence of Y given that even X will not appear in the transaction, and P(Ȳ) = probability of no occurrence of Y.

Range: [0,1]

0 means that the rule does not provide any information for the dataset.

3.6.7. Fisher's exact test

Fisher's exact test (significance test to identify if the rules represent real patterns) computes the p-value from a contingency table of 2 * 2. It returns the p-value associated with the probability of observing the rule only by random [22, 23].

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

where a, b, c, and d = items a, b, c, and d which are the exact frequencies in a 2 * 2 contingency table.
(a + b)! = factorial of the sum of a + b, n! = is the factorial of the number of elements in the contingency table, and a! = factorial of a and so on for each element.

Range: [0,1] (p-value scale)

The closer the p-value is to zero, the lower the probability of observing the rule only by random.

3.6.8. Rule Power Factor (RPF)

RPF focuses on the importance (weights the confidence of a rule by its support) of the association between the antecedent and the consequent of rules. RPF works well even where confidence fails [23, 26]. It is defined as

$$rpf(X \Rightarrow Y) = supp(X \cup Y) * conf(X \cup Y)$$

where X = itemset of the left-hand side (antecedent), Y = itemset of the right-hand side (consequent), supp(X ∪ Y) = computes the support of the combined itemset, and conf (X ∪ Y) = confidence of the rule.
Range: [0, 1]
RPF is more informative regarding the importance of rules. When the antecedent and consequent association increases, rule importance increases.

3.7. Arules function

3.7.1. The is.redundant function

Each metric described in Section 3.6 along with the is.redundant function filter out redundant rules. A rule is redundant if a more general rule with the same or a higher confidence exists. That is, a more specific rule is redundant if it is only equally or even less predictive than a more general rule [30]. This function receives as an argument the set of rules and a quality metric, and from these two arguments it determines a subset of rule not redundant.

3.7.2. The is.significant function

The is.significant function evaluates the statistical significance of the rules created. This function uses Fisher's method with $\alpha = 0.01$ and the Bonferroni adjust [22]. This function receives as arguments the set of rules, the transactional database and the statistical methods and returns a subset of statistically significant rules.

3.7.3. The is.maximal function

The is.maximal function report only maximal rules [18]. A frequent itemset is maximal if none other frequent itemsets are its superset. An association rule is defined as maximal if it was generated with a maximal itemset. This function reports a subset in which it determines which rule is maximal and which is not.

3.7.4. Subset function

It return subsets of vectors, matrices or data frames that meet a logical condition [18], for example, the score for one specific metric of a particular rule or a rule of interest. It receives as an argument the set of rules and the subset to choose.

3.7.5. Pearson's correlation coefficient

To analyze the linear correlation of the variables included in the dataset under study and contrast with the rules created, Pearson's correlation coefficient was computed. This function receives as an argument the numeric type database and returns an array with the significance of the association between variables.

3.7.6. Graph-based visualization with items and rules as vertices

This representation focuses on how the rules are composed of individual items and shows which rules share items [28]. In a rule, all the items that are part of the antecedent point to the circle labeled with the rule number and from this circle the consequent of the rule is pointed to, an example of this visualizationcan be seen in Figure 2.

3.8. Experimental Studies

Once the dataset has been preprocessed, it was loaded into memory as a transaction type object with the read.transaction function of the arules package as required by this package. The dataset format was specified as basket, header equal to true and the separator was the comma character since it is a csv file. The development of the experimental process is described in this subsection and Figure 1 shows the sequence of each experiment performed during the research pipeline.

3.8.1. Determination of support threshold

To create the association rules, it is necessary for the apriori algorithm to find the frequent itemsets. To do so it must assign the appropriate support threshold. To obtain the appropriate support threshold, percentages are chosen in any specific range and assigned to the apriori algorithm to search for the frequent itemsets. The objective is to find frequent itemsets that are not redundant or in other words that are specific (measured through quality metrics).

Of all the possible results of the experimentation with the analyzed support percentages, it was chosen a range specific to create the rules. For this research the percentages chosen for the support were as follows: positive class 5%, 10%, 15%; negative class 20%, 35%, 50%, and indeterminate class 1%, 2%, 4%, respectively.

3.8.2. Determination of confidence threshold

Similarly to support metric, the appropriate percentage for confidence was also investigated. For example, 85% confidence was assigned and the algorithm was run to create the rules. The rules created for each confidence percentage (80%, 85%, 90%, and 95%) were analyzed and from this analysis the percentage with the highest performance was chosen (measured through quality metrics). The percentage with the highest performance was 90% and this was determined from the specificity of the rules created with each percentage. The confidence was set at 90% for all the classes and runs of the algorithm.

3.8.3. Evaluation of rules using selected metrics

The apriori algorithm produces a large number of rules, many of them uninteresting or redundant. To select the rules of interest and discard redundant rules, the is.redundant function of the arules package is used and as an argument, it receives the set of rules created and a quality metric.

What the function does is, assuming it is given the cosine metric, it iterates through each rule and selects only those rules that have a value 0.51 to 1. The rules that have a value below that interval are discarded.

3.8.4. Statistical validation

Statistical validation was performed using the is.significant function from the arules package and takes as an argument the subset of rules, the transactional database, the Fisher method, α of 0.01, and the Bonferroni adjustment.

From the subset subjected to statistical validation, only those rules that are statistically significant was selected; those that are not were discarded.

3.8.5. Maximal rules

The subset of rules that passed the statistical validation was scrutinized by the is.maximal function and took the mentioned subset as an argument. The function iterates over the subset to select only those maximal rules.

An itemset is maximal in a set if no proper superset of the itemset is contained in the set. We define here maximal rules, as the rules generated by maximal itemsets.

3.8.6. Biological validation

The subset resulting from the selection process with the quality metrics and the functions of the arules package was subjected to biological validation. Biological validation depends on the experience of the expert (bacteriologist) to determine whether the patterns modeled by the rules represent real patterns associated to vaginosis, Figure 1.

The selected rules are only those that meet three restrictions: the restriction imposed by each quality metric, the restriction imposed by statistical validation, and the restriction imposed by biological validation.

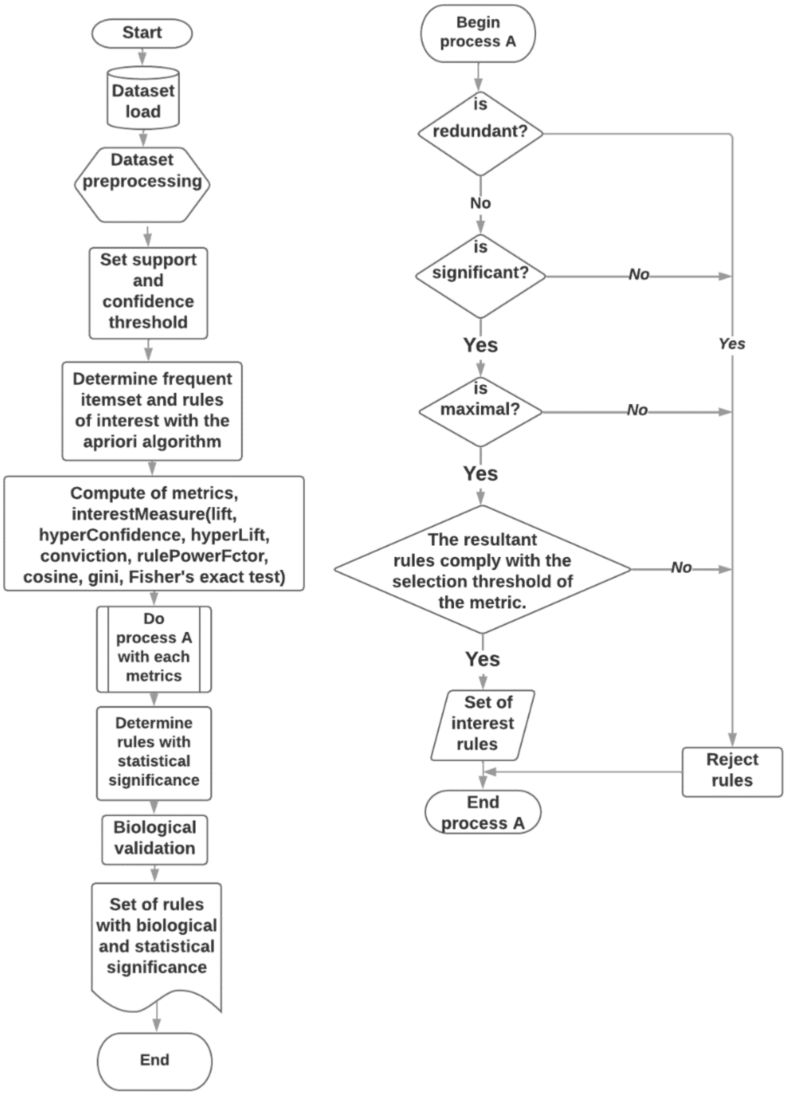
Start
Dataset load
Dataset preprocessing
Set support and confidence threshold
Determine frequent itemset and rules of interest with the apriori algorithm
Compute of metrics, interestMeasure(lift, hyperConfidence, hyperLift, conviction, rulePowerFctor, cosine, gini, Fisher's exact test)
Do process A with each metrics
Determine rules with statistical significance
Biological validation
Set of rules with biological and statistical significance
End
Begin process A
is redundant?
No
is significant?
No
Yes
Yes
is maximal?
No
Yes
The resultant rules comply with the selection threshold of the metric.
No
Yes
Set of interest rules
Reject rules
End process A

Figure 1. Experimental process.

4 Results and discussions

4.1. Determination of support and confidence threshold

The apriori algorithm was run to investigate the appropriate support threshold to create the frequent itemsets (section 3.2 describes what a frequent itemset is). Support intervals were established with the objective to investigate the best percentage. For example, the 5% support was assigned for the class positive and the frequent itemsets reported by the algorithm were analyzed. From these analysis of the results reported by the algorithm, the support percentages chosen to create the frequent itemsets were as follows: for the positive class these were 5%, 10%, and 15%, for the negative class these were 20%, 35%, and 50%, and for the indeterminate class these were 1%, 2%, and 4%, respectively.

Similarly to support metric, the appropriate percentage for confidence was also investigated. For example, 85% confidence was assigned and the algorithm was run to create the rules. The rules created for each confidence percentage (80%, 85%, 90%, and 95%) were analyzed and from this analysis the percentage with the highest performance was chosen. The percentage with the highest performance was 90% and this was determined from the specificity of the rules created with each percentage. The confidence was set at 90% for all the classes and runs of the algorithm.

Functionality of each quality metric was investigated. The quality metrics Hyper-Confidence, Lift, Hyper-Lift, Conviction, RPF, Cosine, Gini, and Fisher's exact test was chosen. These metrics were chosen according to the characteristics of the dataset (imbalanced). For example, the positive class has a low count in its instances with respect to the negative class, this causes behaviors out of range in metrics as the lift. Therefore, the metrics that were chosen from the theoric revision consider this characteristic.

4.2. Positive class

This section report the number of rules created for each setted support % and the number of rules that get over the restriction imposed by each quality metric. The apriori algorithm was run with support for 5% and reported 1477 rules. After making the selection with the quality metrics (hyper-confidence, lift, hyper-lift, conviction, cosine, rpf, gini, and Fisher's exact test) and functions of the ARules package (is.redundant, is.significant, and is.maximal) only 117 rules got over the imposed restrictions, see Table 3.

It can be seen in Table 3 that the rules reported after the selection do not meet the restriction of all metrics since they report a low score, for example, the cosine metric reported a score below 0.4. It can also be seen that the lift value is high since report a score of almost 4. Based on the result of the metrics, it can be said the rules created with the support of 5% are redundant, and so it is a percentage not suitable to create the rules.

For 10%, 148 rules were reported, after making the selection, 18 rules remained. For this percentage of support, the selected rules comply with the restriction of all metrics, even the lift tends to decrease. For the 15% support only one rule is reported. After making the selection, the same rule continues. All metrics except lift in this support percentage report acceptable values. The lift tends to increase in value, Table 3.

Table 3. Interest rule mining based on support-confidence percentages and resulting rules after selection

Set of unfiltered rules created for the positive class											
Supp/Conf	Set of rules	Supp§	Confꬱ	HyperConf♠	HyperLift ♣	Liftæ	Convictionȣ	RPF⇈	Cosine‡	Gini✡	Fisher's exact test⸙
0.05/0.9	1477	0.07147	0.9787	1	1.861	3.857	10.565	0.06987	0.5205	0.08321	1.143^{-07}
0.1/0.9	148	0.1187	0.9623	1	2.117	3.793	11.239	0.11432	0.6702	0.1417	1.091^{-13}
0.15/0.9	1	0.1542	1	1	2.385	3.941	NaN	0.1542	0.7796	0.2031	2.919^{-23}
Maximal rules, not redundant (Fisher'sexacttest), and significant by Fisher's method (α = 0.01), and Bonferronia djusts											
0.05/0.9	117	0.06064	0.9770	1	1.795	3.851	9.504	0.05923	0.4811	0.06943	2.412^{-07}
0.1/0.9	18	0.1081	0.9555	1	2.051	3.766	9.664	0.10325	0.6376	0.1256	3.965^{-13}
0.15/0.9	1	0.1542	1	1	2.385	3.941	NaN	0.1542	0.7796	0.2031	2.919^{-23}

§ Measures the frequency of a set of items. Range[0,1]; ꬱ Sets the utility of X to predict the presence of Y. Range[0,1]; ♠ Evaluates the confidence level for observations with too high or low counts. Range[0,1]; ♣ Adaptation of the lift metric, it is more robust for low or rare counts. Range[0, ∞]; æ Measures how often X and Y occur together than expected if they were statistically independent. Range[0, ∞]; ȣ Evaluates the degree to which the antecedent influences the occurrence of the consequent. Range[0, ∞]; ⇈ It focuses on the importance of the association between the antecedent and the consequent of the rule. Range[0,1]; ‡ Measures the similarity between the LHS and RHS. Values from 0.0 to 0.5 means no correlation. Range[0,1]; ✡ Measures the probability that a particular variable will be misclassified when chosen at random. Range[0,1]. ⸙ Significance test to identify whether a rule represents a true pattern. Range[0,1].

4.3. Negative class

For the 20% support the algorithm reported 1478 rules, after making the selection with the quality metrics (hyper-confidence, lift, hyper-lift, conviction, cosine, rpf, gini, and Fisher's exact test) and functions of the ARules package (is.redundant, is.significant, and is.maximal) the number of rules decreased to 96, Table 4. All quality metrics report ideal values, even the lift stabilizes. The explanation for this behavior is the percentage of support used. The percentage used depends on the occurrences of the class, since a high count in the instances of the class allows to use a greater support.

Higher percentages of support imply more specific the rules created; however, high percentages cannot be used with classes that represent low counts in its occurrences.

The 35% support reported 239 rules, and these decreases to 20 rules after making the selection. For the 50% support, 56 rules were created and after making the selection with the quality metrics (hyper-confidence, lift, hyper-lift, conviction, cosine, rpf, gini, and Fisher's exact test) and functions of the ARules package (is.redundant, is.significant, and is.maximal) 7 rules remain. It can be observed in Table 4 that for these support percentages the quality metrics stabilize at the maximum.

When the dataset under study is balanced, all the metrics work properly. However, when the dataset is unbalanced due to the low count in the instances, metrics such as lift they report out-of-range values.

Table 4. Interest rule mining based on support-confidence percentages and resulting rules after selection

Set of unfiltered rules created for the positive class											
Supp/Conf	Set of rules	Supp§	Confꭆ	HyperConf♠	HyperLift ♣	Liftæ	Convictionȣ	RPF⥣	Cosine‡	Gini✡	Fisher's exact test⸙
0.20/0.9	1478	0.2788	0.9650	1	1.225	1.447	8.782	0.2691	0.6293	0.07762	1.156^{-06}
0.35/0.9	239	0.4350	0.9646	1	1.283	1.447	10.812	0.4197	0.7905	0.15373	5.731^{-11}
0.50/0.9	56	0.5477	0.9680	1	1.318	1.452	12.502	0.5300	0.8913	0.2384	7.927^{-22}
Maximal rules, not redundant (Fisher'sexacttest), and significant by Fisher's method (α = 0.01), and Bonferronia djusts											
0.20/0.9	96	0.2290	0.9478	1	1.184	1.422	7.675	0.2167	0.5689	0.05070	1.767^{-06}
0.35/0.9	20	0.3915	0.9509	1	1.255	1.426	12.372	0.3729	0.7453	0.11974	3.033^{-10}
0.50/0.9	7	0.5330	0.9568	1	1.302	1.435	10.505	0.5103	0.8744	0.2141	4.774^{-21}

§ Measures the frequency of a set of items. Range[0,1]; ꭆ Sets the utility of X to predict the presence of Y. Range[0,1]; ♠ Evaluates the confidence level for observations with too high or low counts. Range[0,1]; ♣ Adaptation of the lift metric, it is more robust for low or rare counts. Range[0, ∞]; æ Measures how often X and Y occur together than expected if they were statistically independent. Range[0, ∞]; ȣ Evaluates the degree to which the antecedent influences the occurrence of the consequent. Range[0, ∞]; ⥣ It focuses on the importance of the association between the antecedent and the consequent of the rule. Range[0,1]; ‡ Measures the similarity between the LHS and RHS. Values from 0.0 to 0.5 means no correlation. Range[0,1]; ✡ Measures the probability that a particular variable will be misclassified when chosen at random. Range[0,1]. ⸙ Significance test to identify whether a rule represents a true pattern. Range[0,1].

4.4. Indeterminate class

For the 1% support, 2028 rules were created, of which only 1 remained in the set after selection. The is.redundant function, when iterating over the set of rules, it discards them all since of the 2028 rules produced by the algorithm only 1 remains. This shows us that low support creates highly redundant rules. The lift is skyrocket to the maximum, see Table 5. For 2% support 16 rules were created and only one remained after selection. For the support of 4%, 0 rules were created. The low count in the observations for this class means that for this percentage of support no itemsets was found since the frequency with which they are present in the transactional database is much lower than the minimum threshold established.

Table 5. Interest rule mining based on support-confidence percentages and resulting rules after selection

Set of unfiltered rules created for the positive class											
Supp/Conf	Set of rules	Supp§	Confꭆ	HyperConf♠	HyperLift ♣	Liftæ	Convictionȣ	RPF⇈	Cosine‡	Gini✡	Fisher's exact test⸙
0.01/0.9	2028	0.01554	1	0.9996	1.562	12.56	NaN	0.01554	0.4413	0.02676	3.741^{-04}
0.02/0.9	16	0.02488	1	1	2.5	12.56	NaN	0.02488	0.559	0.04322	1.68^{-06}
0.04/0.9	0	0	0	0	0	0	0	0	0	0	0
Maximal rules, not redundant (Fisher'sexacttest), and significant by Fisher's method (α = 0.01), and Bonferronia djusts											
0.01/0.9	1	0.02488	1	1	2.5	12.56	NaN	0.02488	0.559	0.04322	1.68^{-06}
0.02/0.9	1	0.02488	1	1	2.5	12.56	NaN	0.02488	0.559	0.04322	1.68^{-06}
0.04/0.9	0	0	0	0	0	0	0	0	0	0	0

§ Measures the frequency of a set of items. Range[0,1]; ꭆ Sets the utility of X to predict the presence of Y. Range[0,1]; ♠ Evaluates the confidence level for observations with too high or low counts. Range[0,1]; ♣ Adaptation of the lift metric, it is more robust for low or rare counts. Range[0, ∞]; æ Measures how often X and Y occur together than expected if they were statistically independent. Range[0, ∞]; ȣ Evaluates the degree to which the antecedent influences the occurrence of the consequent. Range[0, ∞]; ⇈ It focuses on the importance of the association between the antecedent and the consequent of the rule. Range[0,1]; ‡ Measures the similarity between the LHS and RHS. Values from 0.0 to 0.5 means no correlation. Range[0,1]; ✡ Measures the probability that a particular variable will be misclassified when chosen at random. Range[0,1]. ⸙ Significance test to identify whether a rule represents a true pattern. Range[0,1].

4.5. Performance of quality metrics

To study the performance of each metric; it was done with the is.redundant function. The function compares the redundancy of the rule based on the value calculated for the metric and reports a vector of logical scores in which there are two possible values: true for a redundant rule and false for a non-redundant rule.

Hyper-confidence, rule power factor, cosine, and gini metrics showed the same performance, see Table 6. The hyper-lift metric only varies with respect to hyper-confidence, rule power factor, cosine, and gini in the rules reported for the positive class (26, 14, and 1 rules), in the other classes they have an identical performance, see Table 6.

The performance of the lift metric is different in terms of the number of rules reported for the positive and negative class, however, for the indeterminate class the performance is identical to the other metrics, see Table 6.

The conviction metric is the only one that presents a performance in which it does not match the performance of the other metrics, for example, do not report any rules for the indeterminate class after making the selection, see Table 6.

Table 6. Performance of the metrics of interest for each vaginosis class.

Metrics	Supp/Conf	Class	Unfiltered rules set	Filtered rules
Hyper-confidence	0.05, 0.1, 0.15/0.9	Vaginosis positive	1477, 148, 1	16, 8, 1
Hyper-confidence	0.20, 0.35, 0.50/0.9	Vaginosis negative	1478, 239, 56	18, 6, 3
Hyper-confidence	0.01, 0.02, 0.04/0.9	Indeterminate	2028, 16, 0	1, 1, 0
Hyper-lift	0.05, 0.1, 0.15/0.9	Vaginosis positive	1477, 148, 1	26, 14, 1
Hyper-lift	0.20, 0.35, 0.50/0.9	Vaginosis negative	1478, 239, 56	18, 6, 3
Hyper-lift	0.01, 0.02, 0.04/0.9	Indeterminate	2028, 16, 0	1, 1, 0
Lift	0.05, 0.1, 0.15/0.9	Vaginosis positive	1477, 148, 1	24, 10, 1
Lift	0.20, 0.35, 0.50/0.9	Vaginosis negative	1478, 239, 56	33, 9, 5
Lift	0.01, 0.02, 0.04/0.9	Indeterminate	2028, 16, 0	1, 1, 0
Conviction	0.05, 0.1, 0.15/0.9	Vaginosis positive	1477, 148, 1	20, 7, 0

Conviction	0.20, 0.35, 0.50/0.9	Vaginosis negative	1478, 239, 56	24, 8, 4
Conviction	0.01, 0.02, 0.04/0.9	Indeterminate	2028, 16, 0	0, 0, 0
rulePowerFactor	0.05, 0.1, 0.15/0.9	Vaginosis positive	1477, 148, 1	16, 8, 1
rulePowerFactor	0.20, 0.35, 0.50/0.9	Vaginosis negative	1478, 239, 56	18, 6, 3
rulePowerFactor	0.01, 0.02, 0.04/0.9	Indeterminate	2028, 16, 0	1, 1, 0
Cosine	0.05, 0.1, 0.15/0.9	Vaginosis positive	1477, 148, 1	16, 8, 1
Cosine	0.20, 0.35, 0.50/0.9	Vaginosis negative	1478, 239, 56	18, 6, 3
Cosine	0.01, 0.02, 0.04/0.9	Indeterminate	2028, 16, 0	1, 1, 0
Gini index	0.05, 0.1, 0.15/0.9	Vaginosis positive	1477, 148, 1	16, 8, 1
Gini index	0.20, 0.35, 0.50/0.9	Vaginosis negative	1478, 239, 56	18, 6, 3
Gini index	0.01, 0.02, 0.04/0.9	Indeterminate	2028, 16, 0	1, 1, 0
Fisher’s exact test	0.05, 0.1, 0.15/0.9	Vaginosis positive	1477, 148, 1	117, 18, 1
Fisher’s exact test	0.20, 0.35, 0.50/0.9	Vaginosis negative	1478, 239, 56	96, 20, 7
Fisher’s exact test	0.01, 0.02, 0.04/0.9	Indeterminate	2028, 16, 0	1, 1, 0

Of all the metrics, the Fisher’s exact test is the one that reports the largest number of rules. This metric calculates the exact probability of a specific set of frequencies in 2 * 2 contingency tables. The is.redundant function iterates over each rule of the set submitted to selection, and each rule is evaluated in the contingency table used by the Fisher’s exact test to select the rules. This characteristic allows that this metric to be stable and to have a greater scope when determining whether a rule is redundant or not, see Table 6. Of the metrics studied, Fisher’s exact test is the one that showed the best performance.

4.6. Patterns involved in the development of vaginosis

In this section only the positive class is analyzed since it is our interest to know the patterns that develop it. To mine the patterns involved in the development of vaginosis, the 7% support was chosen from the interval investigated in section 4.2. We chose this percentage for the support to represent the rules by observing that they represented the best patterns of the experiment (specific and non-redundant rules).

A subset of the transactional database was extracted preserving all variables except for the Lactobacillus variable with the qualitative approach, Lactobacillus variable with the quantitative approach was also preserved. The reason why the Lactobacillus variable with the quantitative approach was preserved was the precision with which the thermocycler during the PCR reaction calculates the growth density of these bacteria.

Since Fisher’s exact test metric was the one that had the best performance, as was described in section 4.5, it was chosen to make the selection of the association rules. To accept a rule as a real pattern, it must overcome three restrictions: get over the minimum threshold of each calculated quality metric, statistical validation, and biological validation.

Loaded the dataset under study in memory as a transaction type object, the apriori algorithm read 201 transactions and 29 items. When creating the rules with the support of 7% and confidence of 90%, the minimum number of items equals 2 in the antecedent and restricted the consequent to the item VaginosisPos, the algorithm created 58 rules.

After the selection with the functions of the ARules package (is.redundant, is.significant, and is.maximal) and the Fihser’s exact test metric, only 17 rules remained in the set, see Table 7.

Table 7. Rules with statistical significance and frequency of each rule in the dataset. Created with the 7% support, 90% confidence and selected with Fisher's exact test metric.

LHS‡	RHS	Freq
[1]{*AtopobiumPos,MegasphaeraPos,MycoplasGeniNeg,UreaplasUreaNeg*} ⇒	{VaginosisPos}	25
[2]{*AtopobiumPos,inersHighGrowthDensity,MegasphaeraPos,UreaplasUreaNeg*} ⇒	{VaginosisPos}	16
[3]{*AtopobiumPos,GardnerellaPos,MegasphaeraNeg,MycoplasHomiNeg*} ⇒	{VaginosisPos}	16
[4]{*AtopobiumPos,GardnerellaPos,MegasphaeraNeg,UreaplasUreaNeg*} ⇒	{VaginosisPos}	18
[5]{*AtopobiumPos,gasseriUndetectable,MycoplasGeniNeg,UreaplasUreaNeg*} ⇒	{VaginosisPos}	22
[6]{*AtopobiumPos,GardnerellaNeg,MegasphaeraPos*} ⇒	{VaginosisPos}	17
[7]{*AtopobiumPos,GardnerellaPos,inersHighGrowthDensity*} ⇒	{VaginosisPos}	17
[8]{*AtopobiumPos,GardnerellaPos,MycoplasGeniNeg,UreaplasParPos*} ⇒	{VaginosisPos}	17
[9]{*AtopobiumPos,GardnerellaPos,MycoplasGeniNeg,MycoplasHomiNeg,UreaplasUreaNeg*} ⇒	{VaginosisPos}	17
[10]{*AtopobiumPos,MegasphaeraPos,UreaplasParPos*} ⇒	{VaginosisPos}	15
[11]{*AtopobiumPos,gasseriUndetectable,MegasphaeraPos*} ⇒	{VaginosisPos}	15
[12]{*AtopobiumPos,GardnerellaPos,gasseriUndetectable*} ⇒	{VaginosisPos}	15
[13]{*AtopobiumPos,GardnerellaPos,UreaplasParPos,UreaplasUreaNeg*} ⇒	{VaginosisPos}	15
[14]{*AtopobiumPos,crisLowGrowthDensity,GardnerellaPos,UreaplasUreaNeg*} ⇒	{VaginosisPos}	15
[15]{*AtopobiumPos,MycoplasHomiPos,UreaplasUreaNeg*} ⇒	{VaginosisPos}	18
[16]{*AtopobiumPos,inersHighGrowthDensity,UreaplasParPos*} ⇒	{VaginosisPos}	16
[17]{*AtopobiumPos,inersHighGrowthDensity,MycoplasHomiNeg*} ⇒	{VaginosisPos}	16

† Pattern bacterial that trigger bacterial vaginosis with statistical significance
‡A higher number of bacteria (≥ 3) in LHS increases the accuracy of the diagnosis.

Regarding each metric, Table 8, these show acceptable values except for the lift metric, which reports a high value. However, the hyper-lift metric corrects for bias of the lift and reports more realistic values.

Table 8. Metrics of the rules with statistical significance. Created with the 7% support, 90% confidence and selected with Fisher's exact test metric.

No	Supp	Conf	Lift	Hyperconf	Hyperlift	Conviction	RPF	Cosine	Gini	Fisher's exact test
1	0.12437811	1.0000000	3.941176	1	2.272727	NA	0.12437811	0.7001400	0.15821503	4.801688-18
2	0.07960199	1.0000000	3.941176	1	2.000000	NA	0.07960199	0.5601120	0.09633157	3.903820-11
3	0.07960199	1.0000000	3.941176	1	2.000000	NA	0.07960199	0.5601120	0.09633157	3.903820-11
4	0.08955224	1.0000000	3.941176	1	2.000000	NA	0.08955224	0.5940885	0.10955742	1.364731-12
5	0.10945274	0.9166667	3.612745	1	2.000000	8.955224	0.10033167	0.6288281	0.11918190	2.432086-13
6	0.08457711	1.0000000	3.941176	1	1.888889	NA	0.08457711	0.5773503	0.10290856	7.385605-12
7	0.08457711	1.0000000	3.941176	1	1.888889	NA	0.08457711	0.5773503	0.10290856	7.385605-12
8	0.08457711	1.0000000	3.941176	1	1.888889	NA	0.08457711	0.5773503	0.10290856	7.385605-12
9	0.08457711	1.0000000	3.941176	1	1.888889	NA	0.08457711	0.5773503	0.10290856	7.385605-12
10	0.07462687	1.0000000	3.941176	1	1.875000	NA	0.07462687	0.5423261	0.08982531	2.016973-10

11	0.07462687	1.0000000	3.941176	1	1.875000	NA	0.07462687	0.5423261	0.08982531	2.016973-10
12	0.07462687	1.0000000	3.941176	1	1.875000	NA	0.07462687	0.5423261	0.08982531	2.016973-10
13	0.07462687	1.0000000	3.941176	1	1.875000	NA	0.07462687	0.5423261	0.08982531	2.016973-10
14	0.07462687	1.0000000	3.941176	1	1.875000	NA	0.07462687	0.5423261	0.08982531	2.016973-10
15	0.08955224	0.9000000	3.547059	1	1.800000	7.462687	0.08059701	0.5636019	0.09230125	1.781025-10
16	0.07960199	0.9411765	3.709343	1	1.777778	12.686567	0.07491952	0.5433885	0.08732471	5.454796-10
17	0.07960199	0.9411765	3.709343	1	1.777778	12.686567	0.07491952	0.5433885	0.08732471	5.454796-10

Biological validation of the set of rules of interest resulted in 5 rules only, see Table 9. For the biological validation, the expert (biologist) analyzed the patterns to determine that they represented the patterns observed in the clinic.

Table 9. Rules with statistical and biological significance and frequency of each rule in the dataset.†

LHS‡	RHS	Freq
[1]{*AtopobiumPos, GardnerellaPos, inersHighGrowthDensity*} ⇒	{VaginosisPos}	17
[2]{*AtopobiumPos, MegasphaeraPos, UreaplasmaParPos*} ⇒	{VaginosisPos}	15
[3]{*AtopobiumPos, gasseriUndetectable, MegasphaeraPos*} ⇒	{VaginosisPos}	15
[4]{*AtopobiumPos, GardnerellaPos, gasseriUndetectable*} ⇒	{VaginosisPos}	15
[5]{*AtopobiumPos, inersHighGrowthDensity, UreaplasmaParPos*} ⇒	{VaginosisPos}	16

† Pattern bacterial that trigger bacterial vaginosis with statistical and biological significance
‡A higher number of bacteria (≥ 3) in LHS increases the accuracy of the diagnosis.

What does each rule represent? Rule number one tells us that Atopobium vaginae (AtopobiumPos) and Gardnerella vaginalis (GardnerellaPos) are required to be present in the vaginal mucosa of the patient along with Lactobacillus iners (inersHighGrowthDensity) at a high growth density to detonate VB+.

Lactobacillus iners (inersHighGrowthDensity) have been reported in the medical literature [8, 9] associated with altered normal flora, therefore it predisposes to the development of vaginosis. Rule number two tells us that Atopobium vaginae (AtopobiumPos), Megasphaera phylotype 1 (MegasphaeraPos), and Ureaplasma parvum (UreaplasParPos) must be present to develop bacterial vaginosis, see Table 9. Coexistence of Atopobium vaginae (AtopobiumPos) and Megasphaera phylotype 1 (MegasphaeraPos) causes the development of bacterial vaginosis as long as Lactobacillus gasseri (gasseriUndetectable) is absent, as can be seen in rule number three. It can coexist Atopobium vaginae (AtopobiumPos) and Gardnerella vaginalis (GardnerellaPos) to develop bacterial vaginosis as long as Lactobacillus gasseri (gasseriUndetectable) is undetectable in the vaginal mucosa, rule number four. Atopobium vaginae (AtopobiumPos) and Ureaplasma parvum (UreaplasParPos) must be present in the vaginal mucosa of the patient, Lactobacillus iners (inersHighGrowthDensity) must also be present in a high growth density to develop bacterial vaginosis as show in rule five, see Table 9.

Pearson correlation coefficient reported which Atopobium vaginae (AtopobiumPos) is one of the most significant bacteria during the diagnosis of vaginosis, see Table 10. Table 9 registers 5 rules, in the 5 rules Atopobium vaginae (AtopobiumPos) is present. This bacterium is important in the development of bacterial vaginosis from the biological stanpoint.

Table 10. Linear association between bacteria associated with vaginosis and the diagnosis of vaginosis.

Score	Description	Pearson	Bacteria
0.00 - 0.09	Null correlation.	1.- 0.059347656, 2.- 0.055255975, 3.- 0.082835156.	1.- Mycoplasma genitalium, 2.- Ureaplasma parvum, 3.- Ureaplasma urealyticum.
0.10 - 0.19	Very weak correlation		
0.20 - 0.49	Weak correlation	4.- 0.344004222, 5.- 0.357351802, 6.- 0.362987094	1.- Gardnerella vaginalis, 2.- Mycoplasma hominis, 3.- Megasphaera phylotype 1.

0.50 - 0.69	Moderate correlation		
0.70 - 0.84	Significant correlation	7.- 0.750937221	7.- Atopobium vaginae
0.85 - 0.95	Strong correlation		
0.96 - 1.0	Perfect correlation		

4.7. Graph-based visualization

Graph-based visualization with items and rules as vertices were used to represent the rules listed in Table 9 in the graphic format. In a rule, all the items that are part of the antecedent point to the circle labeled with the rule number and from this circle the consequent of the rule is pointed to, as shown in Figure 2.

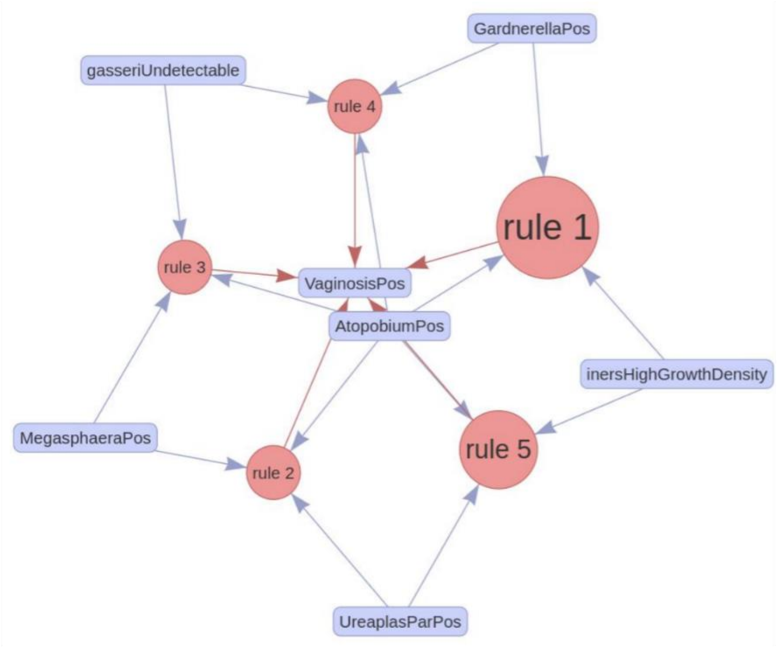

Figure 2. Rules obtained with the minimum threshold for the 7% support and confidence of 90% for the vaginosis class positive.

5 Conclusions

A dataset with records (201) of bacterial vaginosis was analyzed. The analysis began with the preprocessing of the dataset so that the apriori algorithm could create the frequent itemsets and from these the association rules.

Since one of the problems with the apriori algorithm when creating rules is the large number it creates, the is.redundant, is.significant, and is.maximal functions were used. The performance of the following quality metrics was also investigated: Hyper-confidence, Hyper-lift, Lift, Conviction, rulePowerFactor, Cosine, Gini index, and Fisher's exact test to select only the rules with statistical significance.

From the experimental process we have determined that the adequate support and confidence percentages for the apriori algorithm to generate the association rules were 7% and 90% respectively. Fisher's exact test was the metric that allowed us to select the rules of interest created with greater precision. The rules accepted as real patterns that trigger BV were only those that got over three constraints: the constraints of each quality metric, statistical validation, and biological validation. In other words, all coexistent bacteria reported in the antecedents of all five rules conform to date fundamental biological knowledge of bacterial vaginosis. Knowing the patterns involved in the development of BV allows making pertinent decisions regarding treatment to improve of the patient's health.

Association rules are an efficient machine learning technique to study bacterial vaginosis (polymicrobial syndrome), since they lack the subjectivity that classical techniques present and provide accurate information to the physician.

References

[1] Onderdonk, A. B., Delaney, M. L., and Fichorova, R. N.: The human microbiome during bacterial vaginosis. Clinical microbiology reviews, 29(2), 223-238 (2016)
[2] Ravel, J., Gajer, P., Abdo, Z., Schneider, G. M., Koenig, S. S., McCulle, S. L., ... and Forney, L. J.: Vaginal microbiome of reproductive-age women. Proceedings of the National Academy of Sciences, 108(Supplement 1), 4680-4687 (2011)
[3] Morris M, Nicoll A, Simms I, Wilson J, Catchpole M. Bacterial vaginosis: a public health review. BJOG. 2001 May;108(5):439-50. doi: 10.1111/j.1471-0528.2001.00124.x. PMID: 11368127.
[4] Bagnall, P., Rizzolo, D.: Bacterial vaginosis: a practical review. Journal of the American Academy of PAs, 30(12), 15-21 (2017)
[5] Beverly, E. S., Chen, H. Y., Wang, Q. J., Zariffard, M. R., Cohen, M. H., and Spear, G. T.: Utility of Amsel criteria, Nugent score, and quantitative PCR for Gardnerella vaginalis, Mycoplasma hominis, and Lactobacillus spp. for diagnosis of bacterial vaginosis in human immunodeficiency virus-infected women. Journal of clinical microbiology, 43(9), 4607-4612 (2005)
[6] Gajer, P., Brotman, R. M., Bai, G., Sakamoto, J., Schütte, U. M., Zhong, X., ... and Ravel, J.: Temporal dynamics of the human vaginal microbiota. Science translational medicine, 4(132), 132ra52-132ra52 (2012)
[7] Gad1, G. F., El-Adawy, A. R., Mohammed, M. S., Ahmed, A. F., and Mohamed1, H. A.: Evaluation of different diagnostic methods of bacterial vaginosis (2014)
[8] Sanchez-Garcia, E. K., Contreras-Paredes, A., Martinez-Abundis, E., Garcia-Chan, D., Lizano, M., and de la cruz-Hernandez, E.: Molecular epidemiology of bacterial vaginosis and its association with genital micro-organisms in asymptomatic women. Journal of medical microbiology, 68(9), 1373-1382 (2019)
[9] Kusters, J. G., Reuland, E. A., Bouter, S., Koenig, P., and Dorigo-Zetsma, J. W.: A multiplex real-time PCR assay for routine diagnosis of bacterial vaginosis. European Journal of Clinical Microbiology and Infectious Diseases, 34(9), 1779-1785 (2015)
[10] Zariffard, M. R., Saifuddin, M., Sha, B. E., and Spear, G. T.: Detection of bacterial vaginosis-related organisms by real-time PCR for Lactobacilli, Gardnerella vaginalis and Mycoplasma hominis. FEMS Immunology and Medical Microbiology, 34(4), 277-281 (2002)
[11] Baker, Y. S., Agrawal, R., Foster, J. A., Beck, D., and Dozier, G.: Detecting bacterial vaginosis using machine learning. In Proceedings of the 2014 ACM Southeast Regional Conference (pp. 1-4) (2014, March)
[12] Baker, Y. S., Agrawal, R., Foster, J. A., Beck, D., and Dozier, G.: Applying machine learning techniques in detecting Bacterial Vaginosis. In 2014 International Conference on Machine Learning and Cybernetics (Vol. 1, pp. 241-246). IEEE (2014, July)
[13] Beck, D., and Foster, J. A. Machine learning classifiers provide insight into the relationship between microbial communities and bacterial vaginosis. BioData mining, 8(1), 1-9 (2015)
[14] Beck, D., and Foster, J. A.: Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics. PloS one, 9(2), e87830 (2014)
[15] Marchán, E., Salcedo, J., Aza, T., Figuera, L., de Pisón, F. M., and Guillén, P.: Reglas de asociación para determinar factores de riesgo epidemiológico de transmisión de la enfermedad de Chagas. Ciencia e Ingeniería, 32(2), 55-60 (2011)
[16] Chausa Fernández, P., Gómez Aguilera, E. J., Cáceres Taladriz, C., García Alcaide, F., and Gatell Artigas, J. M. : Extracción de reglas de asociación en una base de datos clínicos de pacientes con VIH/SIDA (2006)
[17] Harahap, M., Husein, A. M., Aisyah, S., Lubis, F. R., and Wijaya, B. A.: Mining association rule based on the diseases population for recommendation of medicine need. In Journal of Physics: Conference Series (Vol. 1007, No. 1, p. 012017). IOP Publishing (2018, April)
[18] Hornik, K., Grün, B., and Hahsler, M.: arules-A computational environment for mining association rules and frequent item sets. Journal of statistical software, 14(15), 1-25 (2005)
[19] Zhan, F., Zhu, X., Zhang, L., Wang, X., Wang, L., and Liu, C.: Summary of Association Rules. In IOP Conference Series: Earth and Environmental Science (Vol. 252, No. 3, p. 032219). IOP Publishing (2019, April)
[20] Agrawal, R., Imielinski, T., and Swami, A.: Mining associations between sets of items in large databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data (pp. 207-216) (1993)
[21] Lenca, P., Vaillant, B., Meyer, P., and Lallich, S.: Association rule interestingness measures: Experimental and theoretical studies. In Quality Measures in Data Mining (pp. 51-76). Springer, Berlin, Heidelberg (2007)
[22] Hahsler, M., and Hornik, K.: New probabilistic interest measures for association rules. Intelligent Data Analysis, 11(5), 437-455 (2007)
[23] Hahsler, M.: A probabilistic comparison of commonly used interest measures for association rules. United States. Southern Methodist University (2015)

[24] Brin, S., Motwani, R., Ullman, J. D., and Tsur, S.: Dynamic itemset counting and implication rules for market basket data. In Proceedings of the 1997 ACM SIGMOD international conference on Management of data (pp. 255-264) (1997, June)
[25] Tan, P. N., Kumar, V., and Srivastava, J.: Selecting the right objective measure for association analysis. Information Systems, 29(4), 293-313 (2004)
[26] Kumar, S., and Joshi, N.: Rule power factor: a new interest measure in associative classification. Procedia Computer Science, 93, 12-18 (2016)
[27] Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, A., and Alvarado-Pérez, J. C.: El proceso de descubrimiento de conocimiento en bases de datos. Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional, 63-86 (2016)
[28] Hahsler, M., and Chelluboina, S.: Visualizing association rules: Introduction to the R-extension package arulesViz. R project module, 223-238 (2011)
[29] Hahsler, M., Buchta, C., Gruen, B., Hornik, K., Johnson, I., Borgelt, C., and Hahsler, M. M.: Package ‘arules’ (2021)
[30] Bayardo, R. J., Agrawal, R., and Gunopulos, D.: Constraint-based rule mining in large, dense databases. Data mining and knowledge discovery, 4(2), 217-240 (2000)