USE OF ASSOCIATION RULES TO PERFORM THE MINING OF A MARKETING DATABASE COMERCIAL

Priscilla Leão de Lima
prilima2413@gmail.com
Graduate department of the university center - FAMETRO, Manaus - AM, Brazil

Richardyson Nobrega da Fonseca
richardyson.rndf@gmail.com
Graduate department of the university center - FAMETRO, Manaus - AM, Brazil

Rilmar Pereira Gomes
rilmargomes@hotmail.com
Graduate department of the university center - FAMETRO, Manaus - AM, Brazil

David Barbosa de Alencar
david002870@hotmail.com
Graduate department of the university center - FAMETRO, Manaus - AM, Brazil

Abstract

Currently, companies seek to find methods to analyze their customer behaviors and profiles. From this, a case study was carried out in a drugstore, in which the data mining technique and the Apriori algorithm were applied, for a better understanding. of your sales. First, a bibliographic survey on data mining and its tasks was carried out. Soon the sales made and their respective products were examined. Finally, marketing standards were presented according to the data analyzed, in order to assist the organizational management and marketing of the examined company.

Keywords: Data Mining; Standards; Commercial marketing.

1. 1. INTRODUCTION

Organizational management is one of the pillars for the evolution of a company, as it is related to planning, indicators, goals and organizations, carrying out strategies to obtain positive returns for the institution.

One of the aspects that evolve the management of organizations is the growth of technology, it is easier to assist in this administration, making new discoveries in a large amount of data in a short period of time producing content and reports with advantageous data for the business for decision making.

Therefore, to improve these aspects within an organization, the information technology sectors use the data mining technique (Data Mining), which consists of a study to create standards and relate variables in a

database. Companies that use this method will be ahead of competitors, as it is a differential that has the purpose of forecasting trends, improving the marketing of products according to the customer's profiles.
However, this mining is composed by algorithms, one of which is Apriori, which works with association rules based on correlations of elements in a database, with the intention of generating patterns, according to the verification of the acquisition of frequent products of the company.

According to the aforementioned facts, there are some institutions that still cannot understand their commercial plan, putting together a strategic scheme to increase their profits and with the help of technological resources can improve their rate of outflow of items offered to the consumer.

Therefore, in this work, it aims to assist decisions based on patterns discovered about products and sales according to a case study of a drugstore, verifying trends in monthly purchases, in order to contribute to the management of the company as in investments, administration marketing and identification of the most commercialized products with groupings of the items sold.

2. Bibliographic reference

In order to obtain data mining standards using the knowledge discovery method in the database, queries were used with all sales items using the Apriori algorithm and another with groups of product categories. We use (KDD), a non-trivial technique, in which it detects patterns that are relevant, true and current through data analysis with iterative performance because the result of the tasks depends on each other and it is interactive because the activities can be repeated. In which they are composed of several stages:

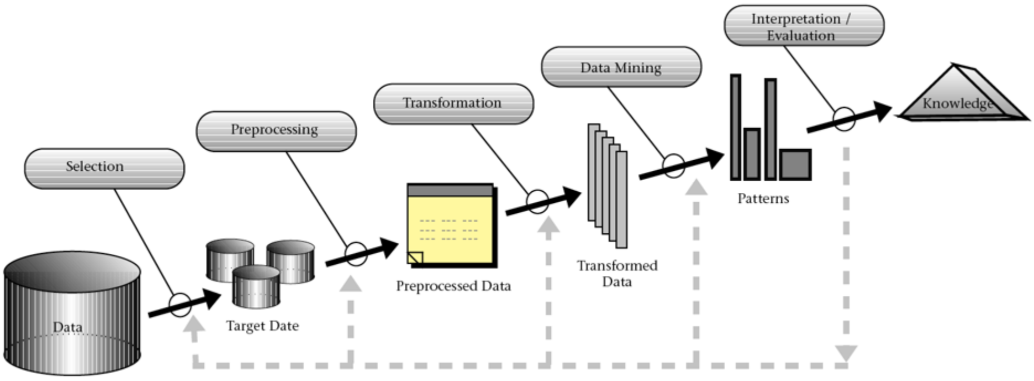


Figure 1- KDD process
Source: Fayyad et al, 1996.

As shown in Figure 1, the first step is the selection consisting of data that will be used in the analysis, which can be set or subsets of data and have different formats, such as structured, unstructured and semi-structured. Where data on the sale item, product and product group tables was selected.

Second, it is the pre-processing, in which the format to be followed will be molded to guarantee the correct delivery, checking the inconsistencies and anomalies that may occur during the process, thus avoiding duplication of data. Where data divergent from the reality of the environment have been eliminated.

Third, Transformation, which techniques are used to improve analysis such as normalization, aggregations, checking if there is a need for new attributes.

Fourth, Data Mining was performed by the WEKA tool, which uses standard algorithms, with significant findings from the data analyzed. The main steps of this process, is the identification of the problem, verifying the goals that must be met, according to decrease data that are not relevant to the case studied, third is to remove insignificant points for the first stage, being necessary to create parameters to verify the usefulness of the information, fourth, would be cleaning as duplicate or erroneous data. Mining tasks are divided in two that aim to specify what will be sought for information, the predictive ones in which the value (dependent or target) of a data based on other values (independent) will be assumed and the descriptive ones which is the pattern analysis, in which checks the relationship of the data.

Therefore, to make data discovery, association rules are used, in this technique, items that are correlated are listed, thus forming a subset of data, such as examining customers who purchased an "X" product, but that there is also a probability of being sold together, the product "Y" from the number of times of transactions. in which it has two parameters: The support is calculated through the ratio between the number of records with a given antecedent to the total number of records (SANTOS, 2004).

$$\mathrm{Is} = \frac{\mathrm{Ra}}{\mathrm{Tr}}$$

Is is the support index
Ra is the number of records with antecedents; e
Tr is the total number of records

And the confidence that is calculated through the ratio between the number of records with a given antecedent and consequent by the total number of records with the respective consequent. (SANTOS, 2004)

$$\mathrm{Ic} = \frac{\mathrm{Rac}}{\mathrm{Ra}}$$

Ic is the reliability index;
Rac is the number of records with antecedent and consequent;
and *Ra* is the number of records with antecedents

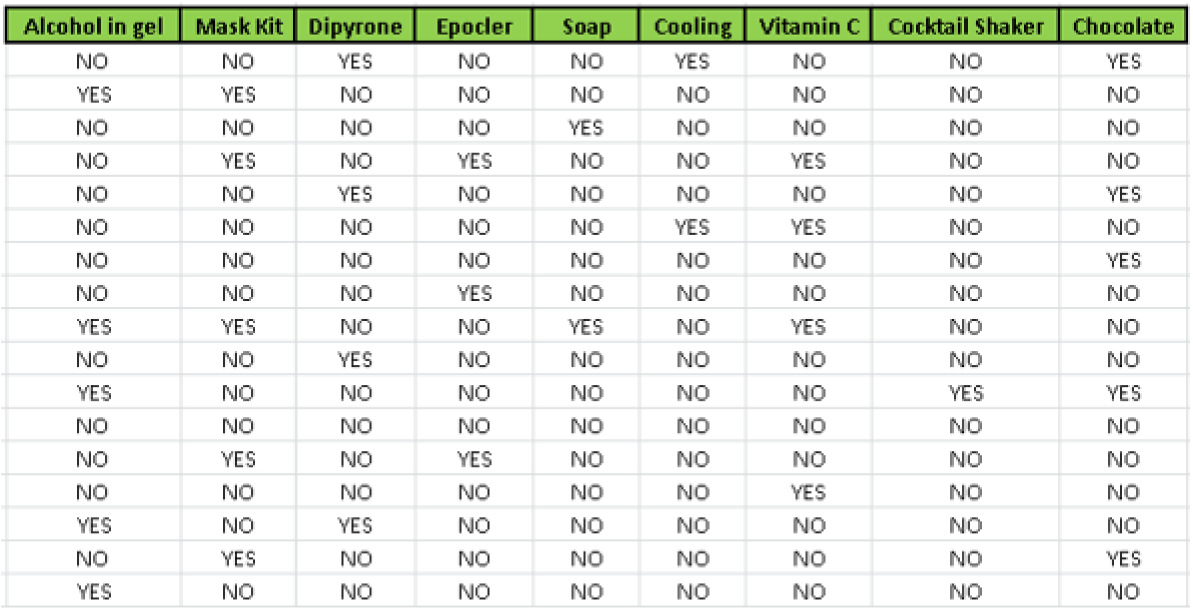
And the last phase of this process is the evaluation, a step that is verified if the analysis has valid results for the problem studied, explaining to the person in charge of the business that will make decisions based on the fruits exposed by the WEKA tool, a free software, which consists of a grouping of algorithms to perform data mining, connecting to a Database Management System through another program, such as the Workbench, offering the production, execution and optimization of queries in SQL language.

3. Materials and methods

Through a questioning about the commercial plan of companies, we applied the methodology of a case study of a drugstore in the month of February 2021, producing a large number of significant variables relevant to a phenomenon, in order to develop knowledge about marketing, using stages the KDD process, such as pre-processing, in which the drugstore base was analyzed, verifying the following tables, customer, product, sale, item sale and product group with approximately fifty thousand sales and one hundred thousand items of sale, carried out by the institution, using the Workbench tool, which manages the Mysql database.

Another step of this process is Data Mining and for its execution, the database was connected to a program that groups algorithms for data mining, generating graphs and discoveries of easy patterns, called Weka, being necessary the use of queries , in order to obtain the tendencies of the enterprise, having algorithms, however the most suitable for our study is the a priori algorithm, which contains value of the transactions, the percentage and the confidence as a parameter.

After making this connection, the product associations were first analyzed according to the drugstore's database, as shown in the following image:



Alcohol in gel	Mask Kit	Dipyrone	Epocler	Soap	Cooling	Vitamin C	Cocktail Shaker	Chocolate
NO	NO	YES	NO	NO	YES	NO	NO	YES
YES	YES	NO	NO	NO	NO	NO	NO	NO
NO	NO	NO	NO	YES	NO	NO	NO	NO
NO	YES	NO	YES	NO	NO	YES	NO	NO
NO	NO	YES	NO	NO	NO	NO	NO	YES
NO	NO	NO	NO	NO	YES	YES	NO	NO
NO	NO	NO	NO	NO	NO	NO	NO	YES
NO	NO	NO	YES	NO	NO	NO	NO	NO
YES	YES	NO	NO	YES	NO	YES	NO	NO
NO	NO	YES	NO	NO	NO	NO	NO	NO
YES	NO	NO	NO	NO	NO	NO	YES	YES
NO	NO	NO	NO	NO	NO	NO	NO	NO
NO	YES	NO	YES	NO	NO	NO	NO	NO
NO	NO	NO	NO	NO	NO	YES	NO	NO
YES	NO	YES	NO	NO	NO	NO	NO	NO
NO	YES	NO	NO	NO	NO	NO	NO	YES
YES	NO	NO	NO	NO	NO	NO	NO	NO

Figure 2- Associations of some database attributes

Source: Own authorship, (2021).

Figure 2 shows the transactions that were carried out during the establishment's purchases, with their respective products, soon we were able to carry out a study on the rules of associations and generate standards that assist in decision making aiming to improve commercial performance.

Therefore, we created an SQL, as expressed below, with the objective of improving marketing, according to sales transactions, such as, for example, checking the best-selling items.

```
SQL01: Bestsellers itemSELECT
IF(max(CASE WHEN itvprocodigo = 2 THEN 1 ELSE  0 END)=1,'SIM','NAO')AS    Álcool_emgel,
IF(max(CASE WHEN itvprocodigo = 5 THEN 1 ELSE 0 END)=1,'SIM','NAO')AS VitaminaC,
IF(max(CASE WHEN itvprocodigo = 9 THEN 1 ELSE 0 END)=1,'SIM','NAO')AS Kit_máscara
FROM itemvenda
```

According to SQL01, we selected the products that had the most outlets during the month analyzed from the item-sale table. And from the best sellers, we also run a query to check products that have a low level of sales for other items in the drugstore.

SQL02: Items with low sales index

```
SELECT
IF(max(CASE WHEN itvprocodigo = 14 THEN 1 ELSE      0 END)=1,'SIM','NAO')AS Sabonete,
IF(max(CASE WHEN itvprocodigo = 23 THEN 1 ELSE 0 END)=1,'SIM','NAO')AS Refrigerante,
IF(max(CASE WHEN itvprocodigo = 20 THEN 1 ELSE 0 END)=1,'SIM','NAO')AS Coqueteleira
  FROM itemvenda
```

In the SQL02 presented, we selected the items that contain low sales rates from the item-sale table. And the third element analyzed, are associations of a product referring to other objects, how many exits were made according to the grouping of frequent combinations, to perform this comparison , the following query was done:

SQL03: Product standards

```
SELECT
IF(max(CASE WHEN itvprocodigo =     2    THEN    1 ELSE     0 END)=1,'SIM','NAO')
AS Alcool_emgel_Giovana_Baby_02,
IF(max(CASE WHEN itvprocodigo = 1 THEN 1 ELSE  0 END)=1,'SIM','NAO')
AS Dipirona_500mg_01,
IF(max(CASE WHEN itvprocodigo = 10 THEN 1 ELSE 0 END)=1,'SIM','NAO')
AS dvd_10,
IF(max(CASE WHEN itvprocodigo = 3 THEN 1 ELSE 0 END)=1,'SIM','NAO')
as Protetor_Solar_Vichy_03,
IF(max(CASE WHEN itvprocodigo = 9 THEN 1 ELSE 0 END)=1,'SIM','NAO')
AS Kit_de_máscara_facil_09,
IF(max(CASE    WHEN    itvprocodigo    =8    THEN 1 ELSE     0 END)=1,'SIM','NAO')
AS Perfume_Alfazema_08,
IF(max(CASE WHEN itvprocodigo =4 THEN 1 ELSE 0 END)=1,'SIM','NAO')
AS Fralda_Huggies_M_04,
IF(max(CASE WHEN itvprocodigo = 5 THEN 1 ELSE  0 END)=1,'SIM','NAO')
AS Vitamina_C_05,
IF(max(CASE WHEN itvprocodigo = 13 THEN 1 ELSE 0 END)=1,'SIM','NAO')
AS Epocler_13,
IF(max(CASE WHEN itvprocodigo = 15 THEN 1 ELSE 0 END)=1,'SIM','NAO')
AS Umidificador_de_ar_15 FROM itemvenda
where    itvprocodigo in (2,1,10,3,9,8,4,5,13,15)
GROUP BY itvvencodigo
```

In SQL03 above, we checked which patterns will emerge from the registered products according to the item sale table in the database, which are sold more frequently from the item groupings. And to improve the organization of the products, the following query was created:

SQL04: Product categories

```
SELECT
MAX(CASE WHEN grupoproduto.grpcodigo=1 THEN 'SIM' ELSE 'NAO' END) AS
MEDICAMENTO_01 ,
MAX(CASE WHEN grupoproduto.grpcodigo=2 THEN 'SIM' ELSE 'NAO' END) AS
HIGIENE_02 ,MAX(CASE WHEN grupoproduto.grpcodigo=3 THEN 'SIM' ELSE 'NAO' END) AS
HHOSPITALAR_03 ,
MAX(CASE WHEN grupoproduto.grpcodigo=4 THEN 'SIM' ELSE 'NAO' END) AS COVINIENCIA_04
FROM produto,itemvenda,grupoproduto WHERE
itemvenda.itvprocodigo=produto.prodcodigo and      prodgrpcodigo=grupoproduto.grpcodigo
GROUP BY itvvencodigo
```

And finally, the SQL04 mentioned, performs an analysis to separate the products into categories, from the product table, sale item and product group in order to obtain trends from the grouped elements .

4. Results and discussions

As previously mentioned, the analysis was carried out on a drugstore basis with the objective of ascertaining some standards that may be relevant to the institution, assisting the team responsible for the layout of the items, in which it will be essential to change the places of these products, placing them close and visible to customers.

So, tests were developed that obtained satisfactory results, in which the company's administrators could have a notion of their sales flow, such as, for example, the products that had more outlets in February 2021.

Figure 3: Best selling products on display at Weka
Source: Own authorship, (2021).

In figure 3, the three items that were most sold were presented, which are Gel alcohol, Vitamin C and Mask kit according to the analyzed data, from that the stock must be checked so that they always contain these products available for sale. sale. However, an analysis was also made of products that have a low output rate.

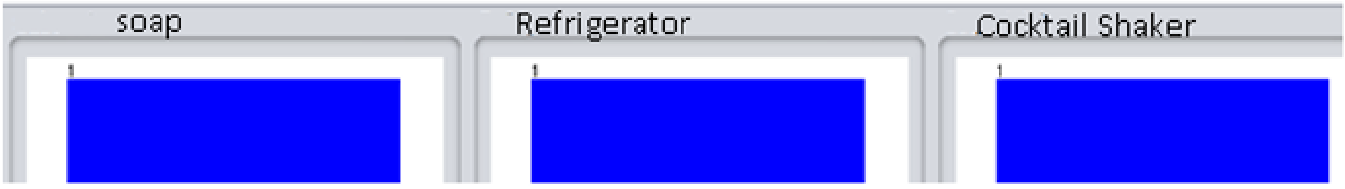
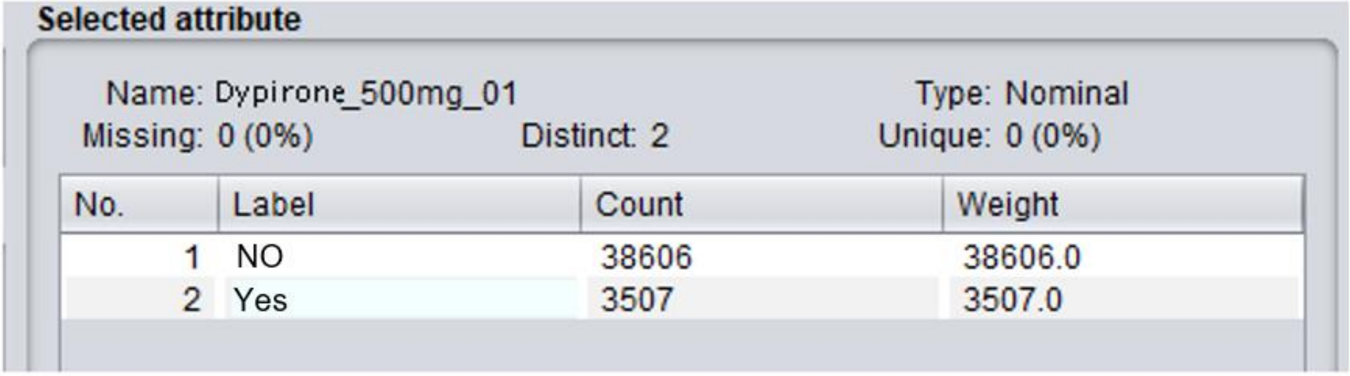


Figure 4: Less sold products exposed on Weka
Source: Own authorship, (2021).

Therefore, those products that were less sold according to Figure 4, which are soft drinks, soap and cocktail shakers, should be placed in more strategic places in view of customers, so that the percentage of sales will increase more and more.

There were also some queries that showed results that helped in the analysis of the sales items, for example, we analyzed the products and one of them is Dipirona.



Selected attribute

Name: Dypirone_500mg_01 Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	NO	38606	38606.0
2	Yes	3507	3507.0

Figure 5: Sales on the product Dipirona exposed in Weka
Source: Own authorship, (2021).

As shown in figure 3, 3507 sales containing Dipirona were presented and 38606 did not have this item in the product outlets, from the analysis, it was observed that there was little sale, in which a planning must be made to increase the outflow of this product. product, such as a promotion. After this verification, we use the Apriori Algorithm to generate patterns:

1. Kit_de_máscara_facil_09=NAO Fralda_Huggies_M_04=SIM Epocler_13=NAO Umidificador_de_ar_15=NAO 2845 ==> Alcool_emgel_Giovana_Baby_02=NAO 2690 <conf:(0.95)> lift:(1.09) lev:(0.01) [227] conv:(2.45)
2. dvd_10=NAO Fralda_Huggies_M_04=SIM Epocler_13=NAO Umidificador_de_ar_15=NAO 2849 ==> Alcool_emgel_Giovana_Baby_02=NAO 2690 <conf:(0.94)> lift:(1.09) lev:(0.01) [224] conv:(2.4)
3. Perfume_Alfazema_08=NAO Fralda_Huggies_M_04=SIM Epocler_13=NAO Umidificador_de_ar_15=NAO 2850 ==> Alcool_emgel_Giovana_Baby_02=NAO 2690 <conf:(0.94)> lift:(1.09) lev:(0.01) [223] conv:(2.38)
4. Alcool_emgel_Giovana_Baby_02=SIM Kit_de_máscara_facil_09=NAO Epocler_13=NAO

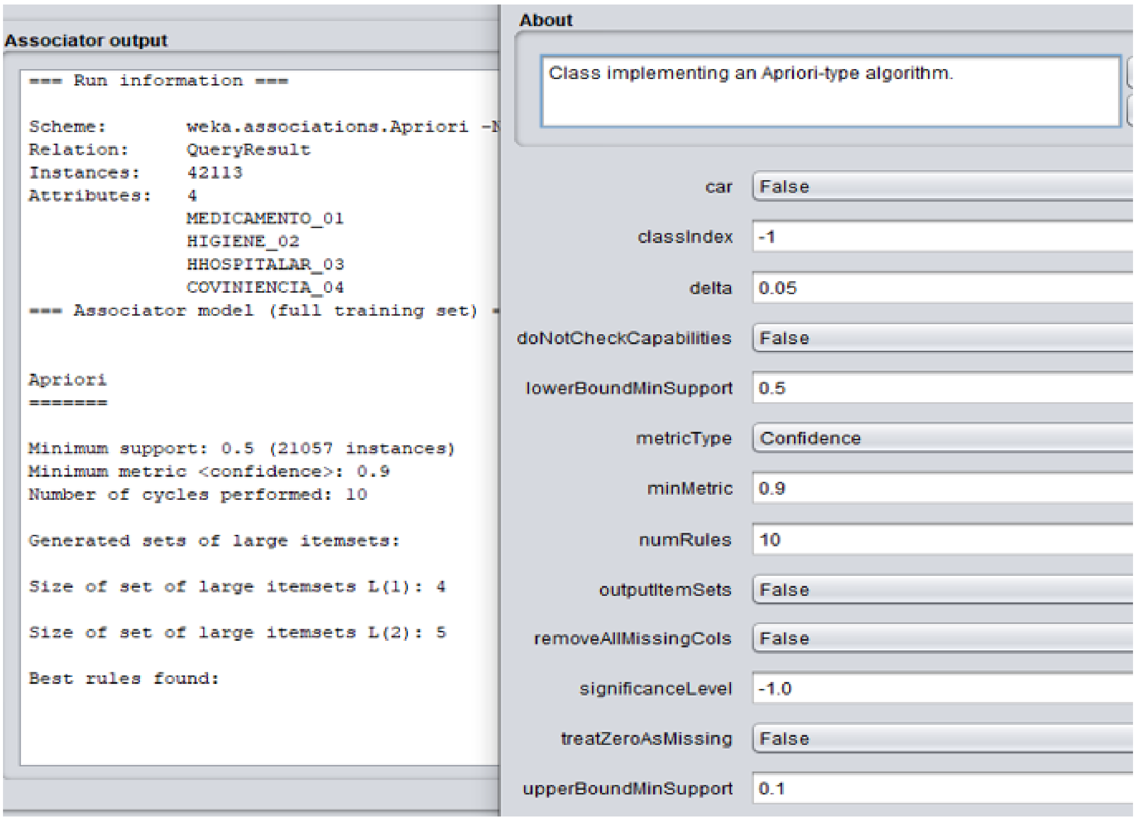
Umidificador_de_ar_15=NAO 2753 ==> Fralda_Huggies_M_04=NAO 2598 <conf:(0.94)> lift:(1.09) lev:(0.01) [218] conv:(2.39)
5. Fralda_Huggies_M_04=SIM Epocler_13=NAO Umidificador_de_ar_15=NAO 3053 ==> Alcool_emgel_Giovana_Baby_02=NAO 2881 <conf:(0.94)> lift:(1.09)

According to the standards generated above, there are some standards, as if there are no sales containing products, such as item 1, which does not include the Mask Kit, Epocler and Air Humidifier, in one sale, there are 95% of Huggies Diapers being sold. To improve the output of these items, administrators must improve their planogram, improving promotions at strategic points.

In another experiment, we selected the product category groups, helping the customer to choose items from his purchase, a clearer way to visualize the products, not leaving him disoriented and needing help, such as the need to ask someone employee, the location of a particular object. And for this procedure, in knowing new patterns, we use the Apriori algorithm:

1. MEDICAMENTO_01=NAO HIGIENE_02=NAO HHOSPITALAR_03=NAO 12974 ==> COVINIENCIA_04=SIM 12974 <conf:(1)> lift:(1.37) lev:(0.08) [3535] conv:(3535.17)
2. MEDICAMENTO_01=NAO HHOSPITALAR_03=NAO 16201 ==> COVINIENCIA_04=SIM 14875 <conf:(0.92)> lift:(1.26) lev:(0.07) [3088] conv:(3.33)

For a better view of the customer, the products must be separated by categories according to the standard generated above, some standards were verified in which, if there is no purchase of medicine, then there is no hygiene or hospital sales, but it contains convenience products with 100% of trust. Another example is that if it does not contain medicines and hospital supplies, then there is convenience item output with 92% reliability. And some changes were also made to the settings to analyze the result.



Associator output

```
=== Run information ===

Scheme:       weka.associations.Apriori -N
Relation:     QueryResult
Instances:    42113
Attributes:   4
              MEDICAMENTO_01
              HIGIENE_02
              HHOSPITALAR_03
              COVINIENCIA_04
=== Associator model (full training set) =


Apriori
=======

Minimum support: 0.5 (21057 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 10

Generated sets of large itemsets:

Size of set of large itemsets L(1): 4

Size of set of large itemsets L(2): 5

Best rules found:
```

About

Class implementing an Apriori-type algorithm.

car	False
classIndex	-1
delta	0.05
doNotCheckCapabilities	False
lowerBoundMinSupport	0.5
metricType	Confidence
minMetric	0.9
numRules	10
outputItemSets	False
removeAllMissingCols	False
significanceLevel	-1.0
treatZeroAsMissing	False
upperBoundMinSupport	0.1

Figure 6: Configuration changes in product category defaults displayed on Weka
Source: Own authorship, (2021).

After executing the changes in the lower limit of the support for 0.5 and upper for 0.1, there was no pattern found according to the frequency of the data, in which there is confidence greater than 90% and support of 50%, according to analysis of 42113 instances.

Anyway, according to the analyzed data, the transactions that were carried out during purchases of the establishment, with their respective products, were soon able to carry out a study on the rules of associations and generate standards that assist in decision making, aiming to improve performance company commercial.

5. Conclusion

Commercial marketing aims to please the customer, assisting in their choices, promoting better services or products, encompassing a set of techniques to increase the company's economic performance.

Due to the aforementioned facts, to help in commercialization, in which there is a deficit in commercial marketing, one must invest in qualified professionals. And with the support of technology, in which it has several techniques to assist commercial performance, in which data mining was applied and the Apriori algorithm was used to analyze the profiles of customers and sales, in which patterns of agreements with the products were generated and sales item and product groups, also intensifying the grouping of items by categories through a better way to expose products at points of sale, improving the space of the drugstore and enhancing stock management.

From the SQL queries presented, in which it was observed satisfactory results for the company, assisting in decision making, in which the most sold products were verified, which are Alcohol gel, Vitamin C and

Mask kit, therefore to improve it sales logistics, more orders must be placed with suppliers to increase the volume in storage. And the least sold products are soft drinks, soap and cocktail shakers, according to the data collected, these should be placed at more strategic points to force more exits as well as the products that are sold grouped. According to Faria, Andrade and Facó (2017, p2) a well-known and discussed example is a case of Walmart in the United States, which used data mining tools in its sales and observed a great correlation between the sale of diapers and beer.

Finally, as a probable future work, it is possible to indicate the most detailed analysis of all days of the week, checking the items that leave daily, to create strategies such as promotions according to the customer's shopping trend, with the purpose of leverage sales

6. References

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. Mining association rules between sets of itens in large databases. ACM SIGMOD Conference Management of Data, Washington, 1993. Proceedings... Disponível em: http://www.rakesh.agrawal-family.com/papers/sigmod93assoc.pdf. Acesso em: 10 nov. 2020.

BAÇÃO, F.; PAINHO, M. Aspectos metodológicos da utilização do data mining no âmbito da geografia. Finisterra, v. 38, n. 75, p. 135-147, 2016.

BEGON, M.; TOWNSEND, C. R.; HARPER, J. L. Ecology: from individuals to ecosystems. 4 ed. Oxford: Blackwell, 2016.

BEKKER, R. M. et al. Long term datasets: from descriptive to predictive data using ecoinformatics. Journal of Vegetation Science, v. 18, n. 4, p. 457-462, 2007. Disponível em: http://dx.doi.org/10.1111/j.1654-1103.2007.tb02559.x/ abstract. Acesso em: 10 nov. 2020.

BLACKBURN, T. M. Method in macroecology. Disponível em: http://wolfweb.unr.edu/~ldyer/classes/blackburn.pdf. Acesso em: 20 jan. 2021.

CHARIF, R. A.; WAACK, A. M.; STRICKMAN, L. M. Raven Pro 1.4 user`s manual. Cornell Lab of Ornithology, Ithaca, NY. 2010. Disponível em: http://www.birds.cornell.edu/brp/raven/RavenOverview.html. Acesso em: 14 fev. 2021.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. (1996); From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence.

SANTOS, C. B.; CARVALHO, D. R.; VAZ, M. S. G. Data Mining em Banco de Dados.
Educacionais: Aproveitamento Acadêmico dos Cursos de Informática da UEPG.
ERI - Escola Regional de Informática, Encontro de Inteligência Artificial. Guarapuava,
2004.

GIMENES, E. Data Mining - Data Warehouse: A Importância da Mineração de Dados em Tomadas de Decisões. Monografia (Graduação em Tecnologia em Processamento de Dados), Centro Estadual de Educação Tecnológica “Paula Souza” Faculdade de Tecnologia de Taquaritinga, 51 p., Taquaritinga, 2000.

FARIA V., ANDRADE A. E FACÓ J. - Mineração de Dados para Análise de Bancos de Dados Empresariais – 2017.

Copyright Disclaimer

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.
This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).