Market basket analysis with association rules in the retail sector using Orange. Case Study: Appliances Sales Company

Marcos Martinez, Belén Escobar,
Universidad Nacional de Asunción,
Asunción, Paraguay
{*marcosmartinez,mescobar*}*@pol.una.py*

and

María E. García-Díaz, Diego P. Pinto-Roa
Universidad Nacional de Asunción,
Asunción, Paraguay
{*mgarcia,dpinto*}*@pol.una.py*

Abstract

This research analyzes the shopping basket by using association rules in the retail area, specifically in a home goods sales company such as appliances, computer items, furniture, and sporting goods. With the rise of globalization and the advancement of technology, retail companies are constantly struggling to maintain and raise their profits and offer the products and services that the customer wants to obtain. In this sense, they need a new approach to identify different objectives to be more competitive and successful, looking for new decision-making strategies. By providing large amounts of data collected in business transactions, the need arises to intelligently analyze such data to extract valuable knowledge that will support decision-making and understand the association patterns that occur in sales-customer behavior. Predicting which product will make the most profit, products sold together, this type of information is of great value for storing products in the inventory. Knowing when a product is out of fashion can support inventory management effectively. In this sense, this work presents the rules of association of products obtained by analyzing the data with the FPGrowth algorithm using the Orange tool.

Keywords: Data Mining (DM), Market Basket Analysis (MBA), Association Rules, Orange Canvas, FP-Growth, Retail, Business Intelligence (BI), Knowledge Discovery in Databases (KDD).

1 Introduction

Data mining is an essential tool for collecting information from different data sets in almost every industry and business in the retail sector. Its great importance in decision-making has made it a key component in the retail sector.

Inventory management requires having the acquisition of items very well planned because the high cost of storage and location of the products is essential to energize the company's resources. This way, knowing the indications of customers' purchasing patterns based on the associations between several outcomes gives to the stock managing a significant value.

Small and medium-sized enterprises (SMEs) and large corporations generate substantial data sets, primarily stored thanks to the development of robust data collection and storage tools. But, traditional storing data are no longer enough to succeed in business and efficiently manage businesses. Business Intelligence (BI) and Data Mining (DM) paradigms provide new ways to analyze data [1].

This work will analyze a commercially powerful retail company nationwide with several decades of sales experience and offers various products classified into various categories, determining functional purchasing

patterns from customer purchase history to help customer-owners make better and timely decisions.

Predictive Analytics (PA) of data enables companies in the sector to have their products adequately available to customers, on-site, and at the right time. And as is well known, this area is still in the research phase due to the multiple factors that can influence the result; therefore, it is not a minor topic [2].

The company has expanded branches in years and records a large amount of sales data management is becoming increasingly complex. Accordingly, it is important to obtain accurate and reliable information from customers' data, thus making decisions that help maintain and improve the services provided, gain advantages over other companies in the field, and achieve scalability. With the PA, it is possible to analyze large amounts of data and project likely future trends.

Sales transaction records obtained over ten years were processed and transformed into the Market Basket Analysis (MBA) form using the Orange tool [3]. The conclusion of this investigation presents the results of this study.

The organization of this work is as follows. Section 2 discusses the leading research of the research area. Section 3 details the theoretical concepts. Section 4 describes the materials used and the methods performed. Section 5 presents the results of computational experiments and discussions. Finally, section 6 concludes this document and discusses possible future research.

2 Literature Review

The literature reports several similar works applied in various environments. For example, Sagin and Ayvaz [4] find association rules, using the shopping cart technique, with the purchasing records of a branch of a hardware company using the WEKA tool [5]. Using the FP-Grow algorithm, Chandwani [6] tracks the different association rules that make up a basket by applying them to data collected from various retail and wholesale stores to predict future trends and behaviors. Sivri and Kasapbas [7] use data-mining association analysis methods such as the Apriori and FP-Growth algorithms. The primary goal of figuring out which sets of items are purchased together in the original transaction database of an online store of a Turkish retail company, using customer age and gender data for analysis.

In the same sense, Gaikwad et al. [8] present the analysis of several Apriori algorithms, and their experimentation occurred in a retail market transactional database. They observe that the support and confidence assessment for the traditional Apriori algorithm consume more time, space and generates the number of candidate sets expected. Meanwhile, Yeuksel Unvann [9] analyzes sales data from any supermarket received from Vancouver Island University website using WEKA software with FP-Growth algorithm, with 225 products, it obtained the top 10 rules according to the conviction value.

Musalem et al. [10] in their research present an approach to identifying the relationships between product categories used to divide a retailer's business into subsets of categories. Web browser data reveal dependencies between item categories. The result produces an intuitive graphical representation of these dependencies using data analysis techniques such as multidimensional scaling and clustering, with four-item category groups acquired by customers. Their output found that retailers can benefit from switching to a customer management approach that identifies relationships between item categories rather than the traditional category management approach, where they manage item categories separately.

Kaur and Kang [11] sought in their research to examine customer behavior and increase sales in retail companies by improving marketing. The primary goal of the marketing MBA is to provide the information to the retailer to understand the buyer's purchasing behavior. It can help the retailer in making the right decisions. In their work, they used different types of mining, such as classification rules, grouping, fusion rule mining, rule induction technique, and the Apriori algorithm and other techniques. Finally, they provide a new algorithm that can be useful for examining customer behavior and helping increase sales.

Moodley et al. [12] have carried out the analysis of MBA data using the UK retail grocery sector as a case study. They mention the importance of research in the area as competition has intensified, shopping habits and demographics have changed and price sensitivity has increased. Uninorms were used as an alternative measure to add support and confidence in analysis. The experiments were conducted with data from consumer panels to compare uninorm with three other popular measures Jaccard, cosine and conviction. It was found that uninorm measure was much better other models in terms of their adherence to the fundamental monotonicity property of support in the MBA analysis.

According to Raja et al. [13] in the current market scenario, most supermarkets and showrooms of retailers are unaware of the status of their stock, the percentage of sales in each period, and the frequency with which the products are sold. They propose a large data platform to overcome the problem by analyzing the average sales frequency of products and the least-sold product, using Business Intelligence and the MBA for their experiments. The FP-Grow algorithm is used to search products that have already expired and thus avoid their sale. They also use other tools such as the Apache Hive, Hadoop cluster. For their part,

Hashem et al. [14] have focused on generating minimum hierarchical rules that provide complete information by proposing an algorithm to derive minimum multi-level association rules and inter-level association rules using a closed network-based approach of item sets. The experimental results express the mixed relationship between the widespread and specialized view of transaction item sets.

Gangurde et al. [15] in their research propose a prediction model for MBA using data cleansing and neural network approach to help improve the quality of the input dataset by eliminating all kinds of errors. They use the machine learning-based MBA model without supervision, based on the artificial neural network design. Apriori's existing algorithm is modified using the neural network method to optimize prediction results. Practical results show that the predictive model for MBA exceeds the previous method of the Apriori algorithm.

Ratore et al. [16] perform the analysis of the sales data of a grocery store. He uses the Apriori algorithm in Weka and R programming for their work. Based on the results of the experiment, a comparative analysis is done between R and Weka.

In their research, Setiabudi et al. [17] apply the MBA theory and use the Apriori algorithm by searching for sets of common items that customers purchase in a Minimarket. Item pairs that exceed minimum support will be included in the frequent itemsets that are selected. Frequent itemsets that exceed minimum support will generate association rules after decoding. A set of usual items can develop association rules and find confidence using hybrid dimension association rules. Test results show that its application can generate information about what type of product is frequently purchased at the same time by customers based on the criteria of hybrid dimension association rules. These results also show a correlation between the data, including support and confidence, that can be analyzed.

Castelo-Branco et al. [18] present relevant aspects of the data mining application to optimize sales in retail stores, such as market basket analysis, association rules, and cross-selling and up-selling. Some typical Business Intelligence applications and examples of data extraction success applied to retail sales are presented to briefly describe the CRISP-DM (Cross Industry Standard Process for Data Mining) model.

3 Theoretical concepts

3.1 Business Intelligence (BI)

Globalization, automation, and new technologies have made enormous amounts of data available in the world whose manipulation, management, and analysis have become extraordinarily complex, mainly in making high-management decisions in companies and industries. Business Intelligence (BI) approaches these problems. According to Gang et al. [19], BI is an extensive set of collection, consolidation, analysis, and access to information capabilities for a solution, such as data warehouse, for querying and reporting data, analyzing multidimensional data, mining, and other technologies.

Over the past few decades in business communities and even academics, Business Intelligence and Analysis (BI&A) and the field of analytics of large amounts of data have become increasingly important. As early as the 1990s, BI became a popular term in the Business and Communities of Communication and Information Technologies (ICT). In the late 2000s is introduced business analysis to represent the critical analytical component in BI [20] where Data Warehouse and Data Mining are crucial technologies in this process.

3.2 Knowledge Discovery in Databases (KDD)

Han et al. [21] defines KDD as a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable data patterns. It is possible to appreciate the KDD process in Figure 1. This process consists of an iterative sequence as detailed below [21]:

1) Data Selection: the relevant data to the analyst order is retrieved from the database

2) Data Pre-Processing: It consists of preparing the data eliminating noise and inconsistent data. It is the most laborious phase for data to be reliable.

3) Data Transformation: Data is transformed or consolidated into appropriate forms for mining by performing summary or aggregation operations.

4) Data Mining: The essential process to extract patterns from data.

5) Pattern Assessment and Presentation: To identify the interesting patterns representing knowledge based on some measures of interest, and knowledge visualization and representation techniques are used to present the extracted knowledge to the user.

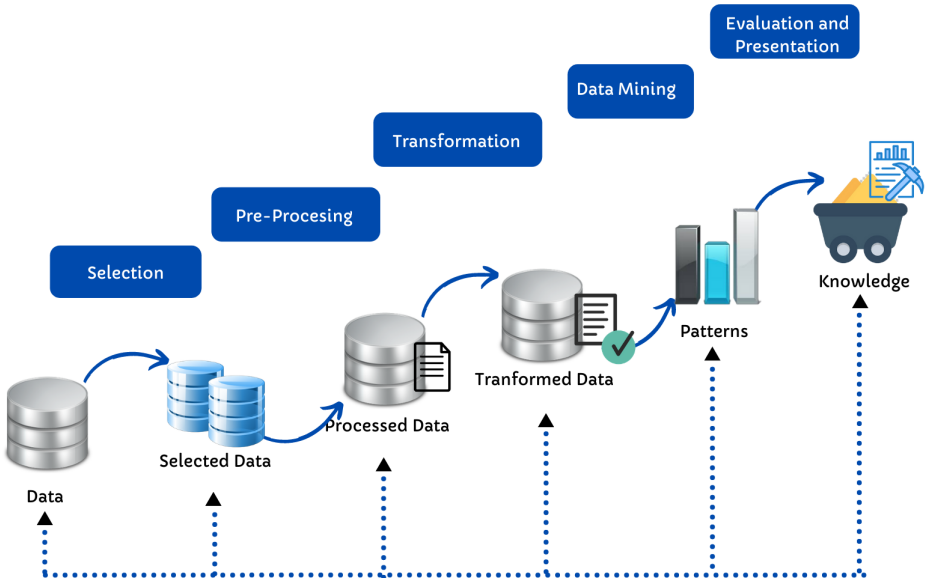


Figure 1: KDD Process

Therefore, we take a broad view of data mining functionality: it is defined as the process of discovering interesting patterns and knowledge from large amounts of data. Data sources can be varied; they could include databases, data stores, electronic forms, web, other repositories of information or data that are dynamically transmitted to the system. For his part, Barry et al. [22] define data mining as the process of finding and predicting hidden database information. It is a powerful technology with great potential to help organizations focus on the most accurate data in their data warehouses.

The retail industry collects vast amounts of data on sales, customer purchase history, freight transport, consumption, and services is, therefore, a suitable application area for data mining. The growing number of data gathered continues to increase rapidly due to the increasing availability, ease, and the popularity of web businesses. Sales data mining helps identify customer-purchasing behaviors, discover customer-purchasing patterns and trends, improve customer service quality, achieve better retention and customer satisfaction, improve goods consumption ratios, design more poignant goods transportation and distribution policies and reduce the cost of business processes.

Until recently, companies analyzed historical databases to obtain statistics in an objective way, drawing conclusions from past actions with which to defend some strategic decision, determined with some degree of confidence that supported those positions. This process is called Descriptive Analytics. Today, however, other types of analyses called Predictive and Prescriptive Analytics are performed, using statistical algorithms and mathematical rules, which seek to predict hypothetical and future scenarios, to have the ability to modify the indicators in that scenario and see their direct and indirect impact on the decisions to be made. According to Lafuente [2], the following are defined:

- **Predictive Analytics**: It is the use of advanced mathematical techniques to predict missing data. It focuses on predicting relevant information but does not automate any processes.
- **Prescriptive Analytics**: It is the use of algorithms and business rule management systems to automate decision-making. The priority objective is to optimize resources and increase the company's operational efficiency.

3.3 Association Rules

According to Fayyad et al. [23] in Data Mining, two types of tasks are distinguished: the descriptive tasks that comprise classification and regression tasks; and predictive tasks with association and grouping tasks. In predictive analytics, obtaining association rules looks for the most common patterns that frequently appear in a dataset that is called Dataset and define it as "A dataset corresponding to the contents of a single database table or data array. Each column represents a particular variable, and each row represents a particular dataset member in question." Itemsets are also sets of articles that frequently appear together

in a transaction data collection. Finding frequent patterns plays an essential role in mining associations, correlations, and other interesting relationships between data [24].

By the pattern search process, it is gaining insights through large amounts of valuable data to the company, which makes it possible to know the consumer behavior, establish marketing strategies, and thus be able to increase sales and, therefore, profits. This technique allows it to obtain information about customers to detect trends, determine potential market directions, and see potential sales opportunities.

3.4 Market Basket Analysis

Market Basket Analysis (MBA) is a methodology that investigates customers' shopping habits by finding associations between the different items that customers place in their shopping baskets. The discovery of these partnerships can help retailers develop marketing strategies by gaining assistance on which items are frequently purchased together [25]. The ultimate idea is always to reduce costs and improve the profits of companies.

In their research work on MBA, Chen et al. [26] explain that existing methods do not discover important purchasing patterns in a multi-store environment. This fact is due to the implicit assumption that the products considered are always on the shelves in all stores and propose a new method to overcome this weakness. Products mixing changes rapidly over time, and a more significant number of stores and periods are considered.

For shopping cart analysis, a transaction consists of every single item that customer purchases in a single purchase. That is, in the dataset, each record contains all items purchased by a customer. A process groupes Individual items to be analyzed by an identifier representing the transaction or occasion, such as the sales ID or invoice.

3.5 Orange Canvas Tool

Orange Canvas is a data mining tool for both beginners and data scientists. Thanks to its friendly interface, users can focus on data analysis rather than laborious coding, simplifying the construction of complex data analysis pipelines [3].

Orange performs data analysis by stacking components in workflows. Each component called a widget, incorporates some data retrieval, pre-processing, visualization, modeling, or evaluation tasks. Combining different widgets in a workflow allows it to create complete data analysis schemas as it goes. The Association rule package has two widgets: one for association rules and one for frequent itemsets [3].

The association rules widget uses the FP-Growth algorithm, which is the most outstanding improvement over Apriori that eliminates candidate generation. This algorithm adopt the divide-and-conquer strategy by compressing the database representing frequent elements into a structure called FP-tree. This data structure retains all essential information and divides the compressed database into a set of conditional databases, each associated with a set of common elements and can mine each separately [4]. Then scan the database only twice:

- In the first scan, all frequent items and their support counts (frequencies) are derived and sorted in each transaction's descending support count order.
- In the second scan, items in each transaction are merged into a tree, and the elements (nodes) that appear in common in different transactions are counted.

Each node is associated with an item and its count. Nodes with the same tag are linked by a pointer called node binding. Because items are sorted in descending frequency order, nodes closest to the root of the tree are shared by more transactions, resulting in a very compact representation that stores all necessary information.

FP-Growth algorithm works in the tree by choosing an element in the increasing frequency order and extracting sets of frequent elements that contain the chosen element when called recursively in the conditional tree, that is, the tree conditioned on a chosen element. FP-Growth algorithm is approximately an order of magnitude faster than the original Apriori algorithm [27].

4 Materials and methods

4.1 Scope of Study

The company that provides data has 67 years of presence in the market and has constituted one of the most respected brands in Paraguay. The automatization of this company started in 1990, and therefore large

amounts of information have been collected. Because in 2008, the company expanded its product catalog incorporating furniture and others. Therefore, this study considers the transactions made from that year. For this work, the data source is extracted from a database in Informix, in an out-of-date version. It has made it challenging to obtain data in a readable format for Orange Canvas, causing additional effort for data cleansing before analysis.

4.2 Data collection

A database copy was mounted on a personal computer, hosted on an Informix server to obtain the data. Data available for various sales periods are necessary to get meaningful information about customer behavior. Here, multi-year data have been obtained from 2008 to 2017 to analyze seasonal trends and patterns. For this reason, the analysis performed excludes the 2018 and 2019 data. As there is no database dedicated to extracting intelligent information, we set up a transactional-like database considering the movements corresponding to appliances and furniture sales.

This work implemented a SQL script performed to obtain product sales reports. This script process the data and ready to be used in the Data Mining phase. It gets the minable view with essential fields of selected tables (Items-Customers-Sales) from 2008 to 2017. A total of 1565415 records indicate sales transactions made by customers. The same procedure extracts all sales transactions from 2017 and December 2008–2017, obtaining three reports of three periods for further analysis.

4.3 Preprocessing

In preprocessing phase, the following entities were analyzed:

- *Sellers*: Out of a total of 1292 seller registrations, this phase left 780 with movements in the study period. The gender of the name was considering to setup sex. With an effectiveness of 90%, a process applied an individual check until 100% of the records had set sex. For the age variable, seller data have been crossed with human resource records, identifying 132 sellers [28].

- *Articles*: It has the products information marketed, classified in "family", "line" and "group" and distributed in 13 items. A total of 37413 records, 26189 were in the "Discontinued" state, i.e. they did not have a valid classification for the study, which is why they should be recategorized, with 20270 records remaining for the sample. Considering the description of the article, a new classification was assigned to 14074 records, with 20009 records valid for study [28].

- *Customers*: Of 454737 records, 182449 records have been excluded, for not owning movements during the study period, with 272288 remaining. At the end of the entire cleaning process, 272265 optimal records were used for testing [28].

- *Sales*: Records for this entity have mainly been valid.

In Table 1, you can see the minable views obtained from different combinations, where each contains sales records in one detail for each item sold [28].

4.4 Dataset transformation

For data mining analysis, it was necessary to format each report obtained from SQL queries to analyze each of the three periods. Since the Orange tool performs its analysis using the MBA methodology, requiring that data be imported be in the basket format, each record corresponds to the set of items sold on an invoice. The items purchased by a customer in a single transaction and columns represent all products available in the store.

Table 2 shows an example of how the data are initially found after being extracted from the SQL report. The "id" sales and product name columns were selected as these would be the fields that use for data mining.

The solution is to traverse the data records using a cursor, so it decides to import the dataset to the Oracle database server. This process is because this server can implement cursors to perform the journey and iteration of each record. The basket format supported by Orange requires that each field in each record be composed of "?" (question mark) if the item does not exist in the sales record and "1" if the item exists in the sales record. Then, in the process, it gets products grouping by each invoice id into a single record.

To obtain this format proceeded as follows:

- The report was imported from csv to Oracle database based on SQL Developer tool into a helper table named "VentasRetail".

Table 1: Mineable Views List

Entities	Mineable Views	
Products	1. Family 2. Line	3. Group
Products/Sellers	1. Gender/Group 2. SellerType/Group 3. Gender/Family 4. SellerType/Family 5. Gender/Line 6. SellerType/Line	7. Gender/SellerType/Group 8. Gender/SellerType/Family 9. Gender/SellerType/Line 10. Gender/Products 11. SellerType/Products 12. Gender/SellerType/Products
Products/Customers	1. Age/Group 2. Age/Family 3. Age/Line 4. Gender/Group 5. Gender/Family 6. Gender/Line	7. Personality/Group 8. Personality/Family 9. Personality/Line 10. Age/Gender/Group 11. Age/Gender/Family 12. Age/Gender/Line
Products/Time	1. Weekday/Group 2. Month / Group 3. Sales' Peak / Group	4. Trimester/Group 5. Semester/Group 6. Season/Group
Products/Time/Customers	1. Age/Weekday/Group 2. Age/Season/Group 3. Gender/Weekday/Group	4. Gender/Season/Group 5. Gender/Age/Weekday/Group 6. Gender/Age/Season/Group
Products/Customers/Sellers	1. Age/GenderSeller/Group 2. Age/SellerType/Group 3. Personality/SellerType/Group	4.Personality/GenderSeller/Group 5. GenderCustomer/GenderSeller/Group 6. GenderCustomer/SellerType/Group

- An empty table called "MatrizProductosRetail" was also created with one column for the Sales ID and 370 columns corresponding to all products (analysis variables).
- All sales IDs other than "VentasRetail" table were inserted into "MatrizProductosRetail" table in the "Id" field, and the other fields (products) are null.
- An index (Oracle Index) was assigned to the sales "Id" column of the "MatrizProductosRetail" table, which would serve to speed up access and query for each record in the table.
- Then, an SQL query was executed using a cursor that contains two fields in the "VentasRetail" table to traverse each record in that table sequentially. This process assigns a value of "1" in the field that corresponds to the product name that contains the cursor in that iteration. In the same name in the "MatrizProductosRetail" table on the condition, the mapping must be entered in the record that corresponds to the same sales id.

The cursor would make comparisons and insertions of value "1" very slowly if it did not have the index created. It happens because it must compare each cursor iteration with more than one million sales "id" other than the array.

After the above steps, the products are grouped by each invoice "id" in a single record, leaving as shown in Table 3.

Finally, the table obtained is exported without the "id" field to a CSV to update the table´s NULL fields to "?" (corresponding to products unsold in a transaction). The file is imported into the Jalatext [29] text editor, a tool that supports many records. Changes are made, leaving the final format as shown in Table 4.

4.5 Induction to Association Rules

Association rule is a data mining technique that seeks to find items that appear together in transactions in each dataset, in which metrics support and confidence help measure the strength of the association rule [30].

a. Support: is a measure that counts the frequency at which the terms of an association rule are in data, that is, the number of transactions ($|X|$) in which items present in a rule occur together in data relative to the total number of transactions ($|D|$) [30].

$$Sop(x) = \frac{|X|}{|D|} \tag{1}$$

Table 2: Initial format of data sales

Id Sale	Product Name
0001	Product B
0001	Product C
0002	Product A
0002	Product B
0003	Product B
0003	Product C
0003	Product D
0004	Product A
0004	Product C
0004	Product D
0005	Product A
0005	Product D

Table 3: Result of Data Transforming

Sale Id	Product A	Product B	Product C	Product D
0001	NULL	1	1	NULL
0002	1	1	NULL	NULL
0003	NULL	1	1	1
0004	1	NULL	1	1
0005	1	NULL	NULL	1

Minimum support thresholds of 1%, 0.5%, 0.1%, 0.05%, and 0.01% were used to obtain a considerable number of common items for generating association rules.

b. Confidence: measure refers to a correspondence value between the items that make up a rule, i.e., the measure denotes the percentage of the transaction that together contains the term antecedent and consequential term about the number of transactions included in the antecedent part [30]. Confidence can be obtained as denoted below:

$$Conf(x \Rightarrow y) = \frac{Sop(x \cup y)}{Sop(x)} = \frac{|x \cup y|}{|x|} \tag{2}$$

The confidence index used was 1%, so as not to rule out possible interesting rules.

c. Lift: is a measure used to assess the degree of dependence on the terms of a rule. The Lift value of an association rule is the relationship between rule confidence and the expected confidence of the rule [31]. It is defined as follows:

$$Lift(x \Rightarrow y) = \frac{Conf(x \Rightarrow y)}{Sop(y)} \tag{3}$$

Evaluation of an association rule can be done as follows [32]:

- If $Lift(x \Rightarrow y) = 1$, then items occurrence of "y" is independent of items occurrence of "x", and vice versa.
- If $Lift(x \Rightarrow y) > 1$, then items occurrence of "y" influences the probability items occurrence of "x".
- If $Lift(x \Rightarrow y) < 1$, then items occurrence of "y" influences the probability the not items occurrence of "x".

4.6 Data Mining on Orange Canvas

Orange Canvas [33] is a free software tool used for the Association Rules discovery. This software has a friendly and practical graphical interface and uses the following widgets, leaving the workflow as shown in Figure 2.

Table 4: Shopping basket Orange format

Product A	Product B	Product C	Product D
?	1	1	?
?	1	1	?
?	1	?	?
1	?	1	1
1	?	?	1

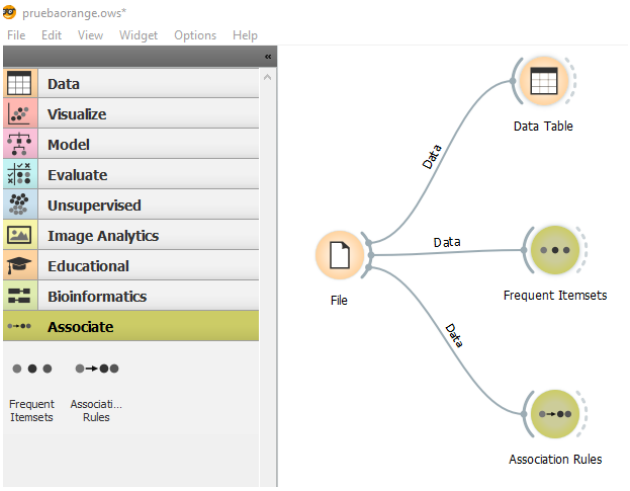


Figure 2: Association Rules Workflow

4.6.1 Widget File

Having the file "matrix-_productos-_10anhos.csv" in market basket format for Orange, use the file widget used to read data from files, select the mentioned file and import it to use.

4.6.2 Widget Data Table

It receives one or more datasets in its input and presents them as a spreadsheet. As shown in Figure 3, the columns are the item groups, and the rows are the transactions made by the customers, where the symbol "?" indicates that the item was not acquired in that transaction and the number "1" suggests that it was obtained.

4.6.3 Widget Frequent Itemsets

Finds common items in a dataset based on a support measure for rule [34]. Figure 4 displays sets of items, where the minimum support of 0.5% was selected, thus finding 58 item sets.

4.6.4 Widget Association Rules

It implements the FP-Growth frequent pattern-mining algorithm, which makes rule search relatively fast and efficient. Figure 5 shows the 18 rules obtained by selecting the minimum support of 0.05% and minimum confidence of 1%.

5 Results and Discussion

5.1 Rules Obtained in the Experiment

Using the article variable "Group" and performing the analysis in three periods, the following rules considered more relevant and detailed below have been obtained.

The Tables 5, 6 and 7 show the top 10 rules according to the conviction value, corresponding to sales transactions of the three periods studied. Note that in all periods analyzed, the best rule accordingly: bed sets {Headers, Mattresses, Mattress Bases, Night Tables, Pillows} have a strong association with each other,

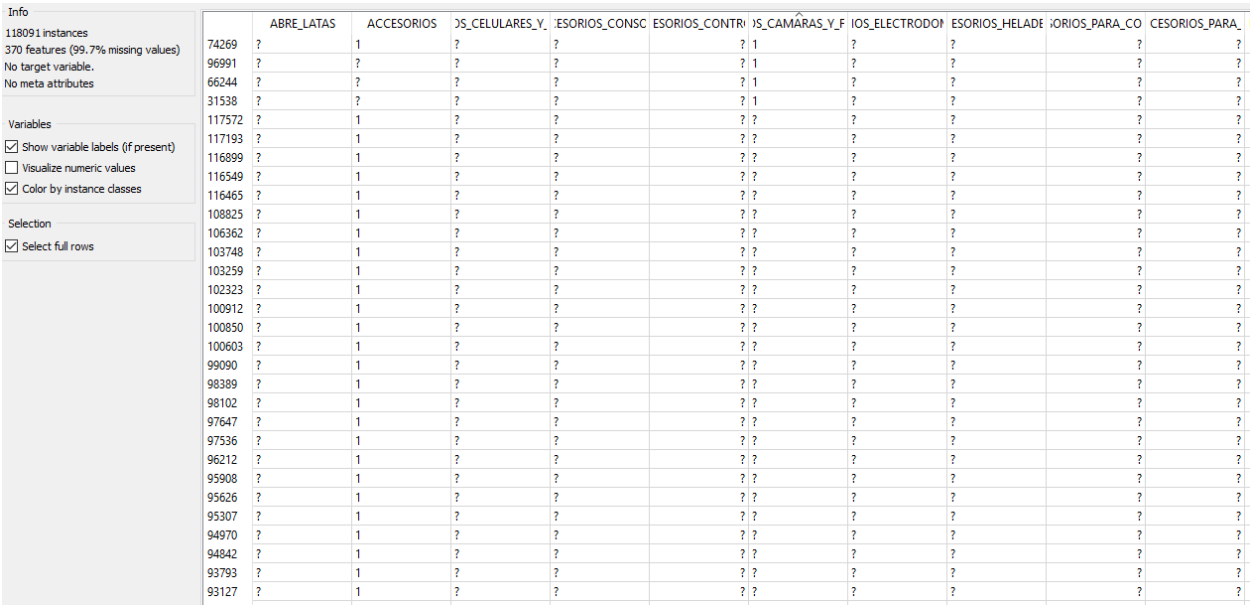


Figure 3: Data Table View (tags in Spanish)

Table 5: The 10 most relevant rules from 2008 to 2017

Confidence	Antecedent	Consequent
0.99	Headboards, Mattresses, Nightstands	Mattres Bases
0.99	PC Table, Software, Webcam	Displays
0.98	Headboards, Mattresses	Mattres Bases
0.97	Pillows, Mattres Bases	Mattresses
0.95	Webcam	Display
0.60	Nightstands	Bedhead
0.47	Abdominal Bench	Treadmills
0.42	Pots and Pans	Induction Cooking
0.42	Kitchen Scales	Ovens
0.38	Blu Ray	LCD and Plasma TV

considering the maximum confidence % found in the relationships of these products, indicating that they have been maintained over the years.

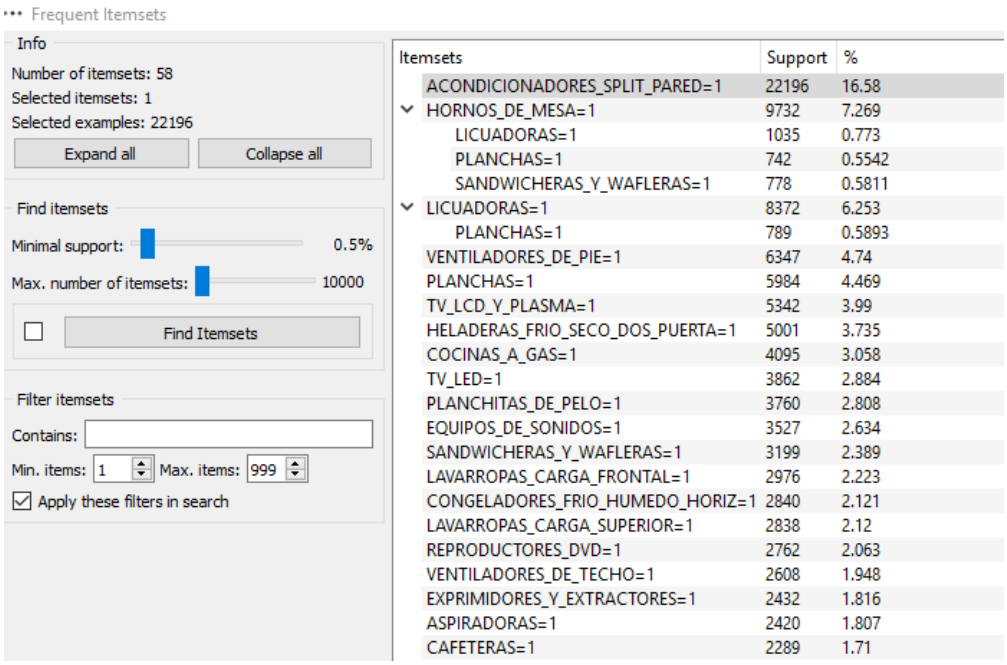
In the period 2017 and period December 2008–2017, we can see the relationship {Split Conditioners, Mattress Bases - mattress} and absent according to the confidence % in the period 2008–2017, indicating that these products began to be sold more often together recently. This relationship can produce year-end promotions in our country (summer) by offering bed and split "set" and boosting the sale of both products.

In December, rules were obtained where they included the most used appliance products in that season, such as blenders, fans, and juicers. It can help the company develop offers, mass marketing of these products in that month, and others.

The information obtained can apply for planning where to place products and, which to place together or nearby to influence customers.

The FP-Grow algorithm used in Orange has three parameters for verifying its results: support, confidence, and Lift. Support describes all products' frequency occurrence in the rule: how frequent all products in the union of a given item A1 with another A2 item are purchased together. Confidence describes the rule conditional probability; therefore, it is a measure of the conditional probability that a customer will purchase all A2 products since they already have A1 products in their basket. Furthermore, finally, Lift gives an estimate of the association measure between articles A1 and articles A2; that is, it shows the extent to which the purchase of all items in A1 depends on the purchase of all items in A2 [35]. The values of these parameters for the 2017 period can be seen in Figure 6.

Table 8 shows the number of rules and the runtime founded, which was fast because of the FP-Growth implementation used in the Orange association rules component. The algorithm is more efficient because by



Itemsets	Support	%
ACONDICIONADORES_SPLIT_PARED=1	22196	16.58
HORNOS_DE_MESA=1	9732	7.269
LICUADORAS=1	1035	0.773
PLANCHAS=1	742	0.5542
SANDWICHERAS_Y_WAFLERAS=1	778	0.5811
LICUADORAS=1	8372	6.253
PLANCHAS=1	789	0.5893
VENTILADORES_DE_PIE=1	6347	4.74
PLANCHAS=1	5984	4.469
TV_LCD_Y_PLASMA=1	5342	3.99
HELADERAS_FRIO_SECO_DOS_PUERTA=1	5001	3.735
COCINAS_A_GAS=1	4095	3.058
TV_LED=1	3862	2.884
PLANCHITAS_DE_PELO=1	3760	2.808
EQUIPOS_DE_SONIDOS=1	3527	2.634
SANDWICHERAS_Y_WAFLERAS=1	3199	2.389
LAVARROPAS_CARGA_FRONTAL=1	2976	2.223
CONGELADORES_FRIO_HUMEDO_HORIZ=1	2840	2.121
LAVARROPAS_CARGA_SUPERIOR=1	2838	2.12
REPRODUCTORES_DVD=1	2762	2.063
VENTILADORES_DE_TECHO=1	2608	1.948
EXPRIMIDORES_Y_EXTRACTORES=1	2432	1.816
ASPIRADORAS=1	2420	1.807
CAFETERAS=1	2289	1.71

Figure 4: Frequent Itemsets View (tags in Spanish)

using a common item tree, it can be processed faster than the data structure used in Apriori [36].

5.2 Runtime importing data

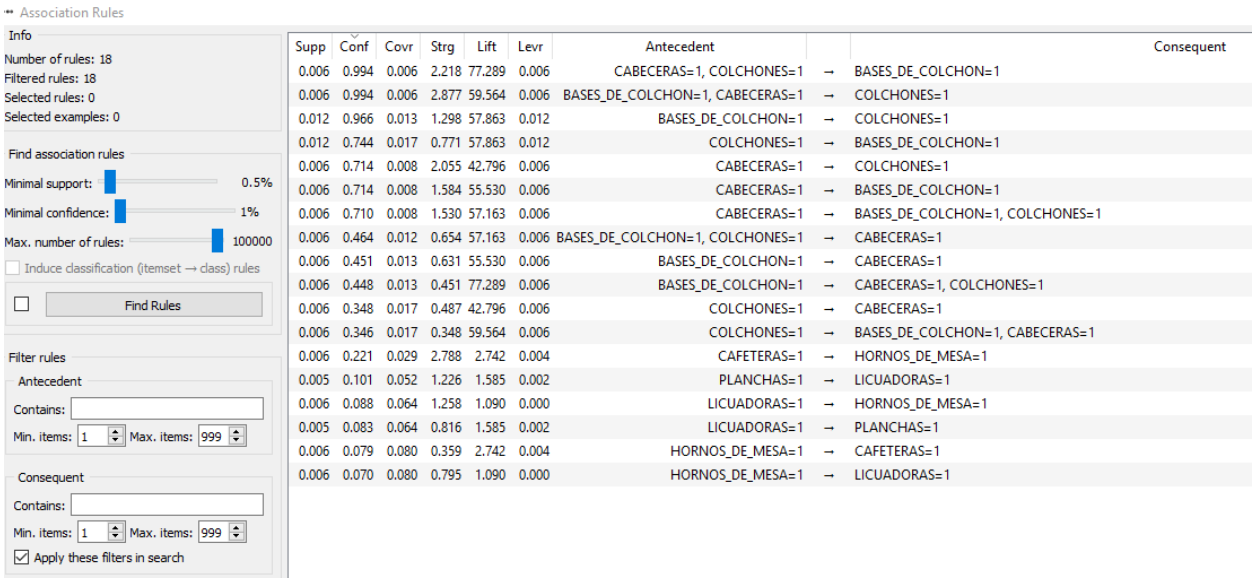
When importing the 2008–2017 dataset into a computer with an Intel I5 processor and 4GB RAM, the import takes a minimum of four hours to complete the process. This time is due to many records in the data table. Subsequently, using a computer with AMD Ryzen 5 PRO 2400G processor and 16GB RAM, data were imported for approximately three minutes.

For the analysis of this period, with more than 1124000 records, Orange requests a reload of the import data each time the program is executed, except data from the other periods studied since they had a maximum of 185000 records. Based on these observations, mining analysis of large amounts of data requires a powerful machine with a good processor and RAM to execute imports and queries in the shortest possible time.

5.3 Statistic analysis

Figure 7 shows the statistical graph indicating the number of best-selling products in the period 2017, and Figure 8 shows the best-selling products in December. Figure 9 below shows days with the most sales in the 2017 period and best-selling items in those days. Widgets used are from the Orange Visualize package.

Orange has several options for viewing analyzed data, as well as helping make it more interactive. The tool can save all charts selected for viewing data and patterns found, such as images in png format.



Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.006	0.994	0.006	2.218	77.289	0.006	CABECERAS=1, COLCHONES=1	→	BASES_DE_COLCHON=1
0.006	0.994	0.006	2.877	59.564	0.006	BASES_DE_COLCHON=1, CABECERAS=1	→	COLCHONES=1
0.012	0.966	0.013	1.298	57.863	0.012	BASES_DE_COLCHON=1	→	COLCHONES=1
0.012	0.744	0.017	0.771	57.863	0.012	COLCHONES=1	→	BASES_DE_COLCHON=1
0.006	0.714	0.008	2.055	42.796	0.006	CABECERAS=1	→	COLCHONES=1
0.006	0.714	0.008	1.584	55.530	0.006	CABECERAS=1	→	BASES_DE_COLCHON=1
0.006	0.710	0.008	1.530	57.163	0.006	CABECERAS=1	→	BASES_DE_COLCHON=1, COLCHONES=1
0.006	0.464	0.012	0.654	57.163	0.006	BASES_DE_COLCHON=1, COLCHONES=1	→	CABECERAS=1
0.006	0.451	0.013	0.631	55.530	0.006	BASES_DE_COLCHON=1	→	CABECERAS=1
0.006	0.448	0.013	0.451	77.289	0.006	BASES_DE_COLCHON=1	→	CABECERAS=1, COLCHONES=1
0.006	0.348	0.017	0.487	42.796	0.006	COLCHONES=1	→	CABECERAS=1
0.006	0.346	0.017	0.348	59.564	0.006	COLCHONES=1	→	BASES_DE_COLCHON=1, CABECERAS=1
0.006	0.221	0.029	2.788	2.742	0.004	CAFETERAS=1	→	HORNOS_DE_MESA=1
0.005	0.101	0.052	1.226	1.585	0.002	PLANCHAS=1	→	LICUADORAS=1
0.006	0.088	0.064	1.258	1.090	0.000	LICUADORAS=1	→	HORNOS_DE_MESA=1
0.005	0.083	0.064	0.816	1.585	0.002	LICUADORAS=1	→	PLANCHAS=1
0.006	0.079	0.080	0.359	2.742	0.004	HORNOS_DE_MESA=1	→	CAFETERAS=1
0.006	0.070	0.080	0.795	1.090	0.000	HORNOS_DE_MESA=1	→	LICUADORAS=1

Figure 5: Association Rules View (tags in Spanish)

Table 6: Most relevant rules of 2017

Confidence	Antecedent	Consequent
1.00	Pillows, Mattres Bases	Mattresses
0.99	Mattres Bases, Bedhead, Nightstands	Mattresses
0.99	Split Wall Conditioners, Mattres Bases	Mattresses
0.94	Pillows, Mattresses	Mattres Bases
0.85	Toys	Swimming pools
0.75	Cookware	Blenders
0.71	Pot Set, Pots and Pans	Cookware
0.70	Keyboard and Mouse, Display	Webcam
0.64	Abdominal Bench	Treadmills
0.42	Juicers and Extractors, Ovens	Mixers

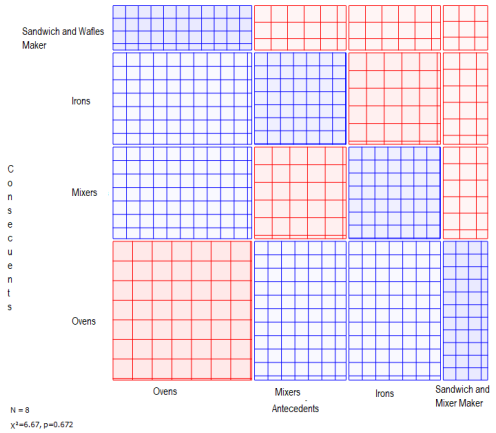


Figure 7: Rules's Sieve Diagram from 2008-2017 with minimum support 0.5%

Table 7: The 10 most relevant rules for December from 2008 to 2017

Confidence	Antecedent	Consequent
0.99	Mattres Bases, Bedhead, Nightstands	Mattresses
0.99	Webcam	Display
0.99	Wall Split Conditioners, Mattres Bases	Mattresses
0.94	Pillows, Mattresses	Mattres Bases
0.87	Toys	Swimming Pools
0.79	Iron, Sandwich Makers, Standing Fans	Mixers
0.76	Display, Motherboard	PC Table
0.76	Digital Cameras	Camera and filmmaker accessoriess
0.64	Keyboard and Mouse, Display	Webcam
0.56	Computer Memory	Digital Cameras
0.56	Juicers and Extractors, Ovens	Mixers

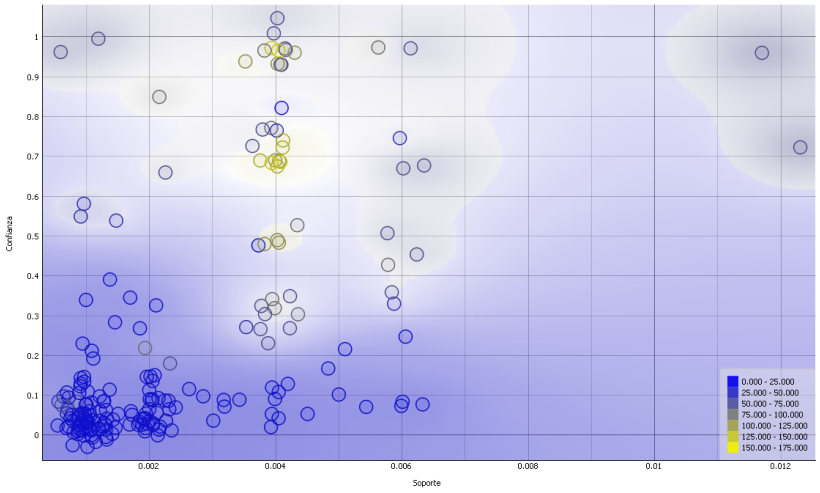


Figure 6: Dispersion rules with measures of support, confidence and lift in Orange

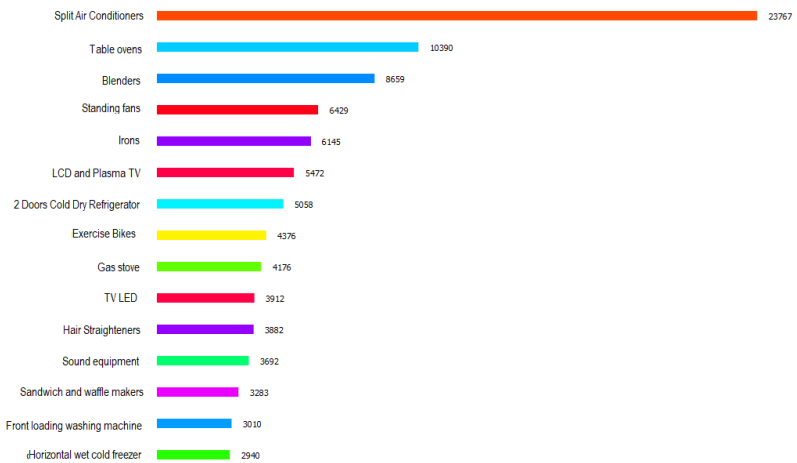


Figure 8: Best-selling products in December

Table 8: Rules and Execution Time Results

Description	Registry	% Support	% Confidence	Runtime (sec)	Number of Rules
Sales of the 2017	118091	0,01	1	3,58	5018
		0,05	1	2,28	457
		0,1	1	1,93	208
		0,5	1	1,83	18
		1	1	1,73	2
Sales of December	133890	0,01	1	3,94	12983
		0,05	1	2,65	835
		0,1	1	2,10	265
		0,5	1	2,05	8
		1	1	0	0
Sales of 2008-2017	1124500	0,01	1	15,52	8799
		0,05	1	14,25	773
		0,1	1	14,28	290
		0,5	1	13,83	8
		1	1	0	0

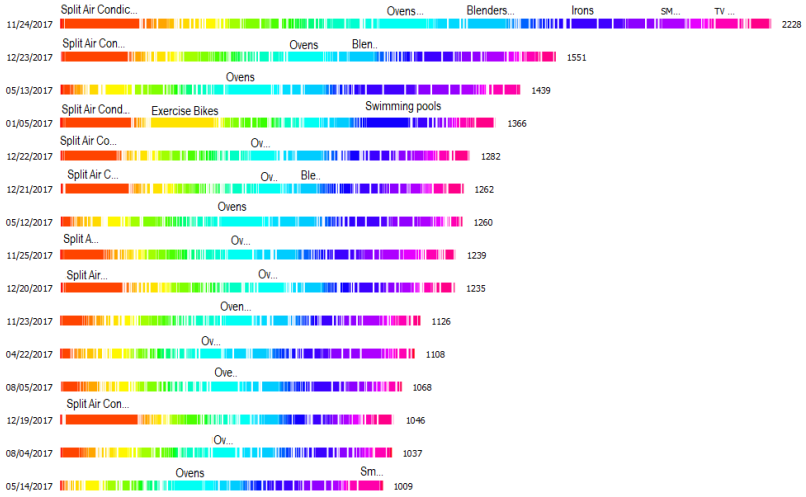


Figure 9: Days with the most sales in 2017

6 Conclusions and Future Work

This study finds the effectiveness of association rules techniques with Orange Canvas in transactional databases. These results help retail identify which products can increase sales. Also, it prepares marketing strategies to promote specific products at promotional prices, perform combos considering the associated patterns of items sold, or where to place products within the commercial lounge, and others.

The next significant contribution is how to prepare the data in the shopping basket format that the Orange tool accepts, generating promising results. Simulation tested several alternatives until found the one that would accurately obtain the format required by Orange. This fact occurred because of little specific information about the data files handled by the tool. It helps future research to refine how to convert this format faster and more efficiently. Finally, note that the company has set up a special area in BI in which two of the authors of this project were hired to implement and giving continuity to work started.

This work can be used as a reference in search new patterns, using other variables to obtain the transaction reports to be analyzed, such as the type of payment (counted or credit), the branch in which the sale occurred, the time when the sale was made, among others. This way would obtain new results that can help in decision-making and an overview of customer behavior based on selected data transformed. The company expert can analyze and evaluate the results obtained. Based on that analysis, identify opportunities or threats to the company in an analyzed context and answer why these events happen?

Also, the behavior history can be analyzed when a new product is introduced to the market based on

region and time variables. This analysis helps know whether it was successful or not and make correct decisions regarding including a new product for sale.

This tool will be applied to the context of e-Health, specifically searching patterns in Pediatric Hematoncology treatment data and in the search for patterns in academic performance at the University.

References

[1] U. Adnan, R. Raz, T. Ahmed, and S. Islam, "Application and analysis of retail inventory using data mining techniques," *Global Journal of Computer Science and Technology: G Interdisciplinary*, vol. 20, no. 2, pp. 27–33, 2020, "Online ISSN: 0975-4172 & Print ISSN: 0975-4350".

[2] J. Lafuente, "Analítica predictiva y analítica prescriptiva," *http://www.decidesoluciones.es/analitica-predictiva-y-analitica-prescriptiva/*, 2013, ultimo acceso 31/10/2019.

[3] Orange, "Orange - data mining fruitful and fun," 2020, [Web;Ultimo acceso el 09-01-2020]. [Online]. Available: URL{https://orange.biolab.si/}

[4] A. N. Sagin and B. Ayvaz, "Determination of association rules with market basket analysis: Application in the retail sector," *Southeast Europe Journal of Soft Computing*, vol. 7, no. 1, pp. 10 – 19, 2018. [Online]. Available: http://scjournal.ius.edu.ba/index.php/scjournal/article/view/149

[5] M. L. G. at the University of Waikato, "The workbench for machine learning," 2019, last access: November 2020. [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/index.html

[6] M. H. Chandwani, "Market basket analysis using association rule," *International Journal of Advance Research, Ideas and Innovations in Technology*, vol. 4, pp. 744–747, 2018.

[7] E. Şafak Sivri and M. C. Kasapbasi, "Extracting association rules of turkish retail company from online transactions. case study," pp. 1176–1186, 2019.

[8] P. Gaikwad, S. Kamble, N. V. Thakur, and A. S. Patharkar, "Evaluation of apriori algorithm on retail market transactional database to get frequent itemsets," in *RICE*, 2017, p. 187–192, ACSIS, Vol. 10 ISSN 2300-5963.

[9] Y. A. Ünvan, "Market basket analysis with association rules," *Communications in Statistics - Theory and Methods*, vol. 0, no. 0, pp. 1–14, 2020. [Online]. Available: https://doi.org/10.1080/03610926.2020.1716255

[10] A. Musalem, L. Aburto, and M. Bosch, "Market basket analysis insights to support category management," *European Journal of Marketin*, vol. 52, no. 7/8, pp. 1–14, 2018. [Online]. Available: https://doi.org/10.1108/EJM-06-2017-0367

[11] M. Kaur and S. Kang, "Market basket analysis: Identify the changing trends of market data using association rule mining," *Procedia Computer Science*, vol. 85, pp. 78 – 85, 2016, international Conference on Computational Modelling and Security (CMS 2016). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050916305208

[12] R. Moodley, F. Chiclanav, F. Caraffini, and J. Carter, "Application of uninorms to market basket analysis," *International Journal of Intelligent Systems*, vol. 24, no. 1, pp. 39–49, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/int.22039

[13] B. Raja, J. Pamina, P. Madhavan, and A. S. Kumar, "Market behavior analysis using descriptive approach," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 7, pp. 171–175, 2019, available at SSRN: https://ssrn.com/abstract=3330017. [Online]. Available: http://dx.doi.org/10.2139/ssrn.3330017

[14] T. Hashem, C. F. Ahmed, M. Samiullah, S. Akther, B.-S. Jeong, and S. Jeon, "An efficient approach for mining cross-level closed itemsets and minimal association rules using closed itemset lattices," *Expert Systems with Applications*, vol. 41, no. 6, pp. 2914 – 2938, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417413008415

[15] R. Gangurde, B. Kumar, and S. D. Gore, "Optimized predictive model using artificial neural network for market basket analysis," *Computer Science and Electronics Journals*, vol. 9, no. 1, pp. 42 – 52, 2017. [Online]. Available: http://www.csjournals.com/

[16] M. Rathore and S. Gupta, "A comparative analysis of r and weka for market basket analysis using apriori algorithm," *Proceedings of International Conference on Communication and Computational Technologies. Algorithms for Intelligent Systems.*, pp. 212 – 219, 2020. [Online]. Available: https://doi.org/10.1007/978-981-15-5077-5_19

[17] D. H. Setiabudi, G. S. Budhi, I. W. J. Purnama, and A. Noertjahyana, "Data mining market basket analysis' using hybrid-dimension association rules, case study in minimarket x," in *2011 International Conference on Uncertainty Reasoning and Knowledge Engineering*, vol. 1, 2011, pp. 196–199, https://ieeexplore.ieee.org/document/6007796.

[18] F.Castelo-Branco, J. L. Reis, J. C. Vieira, and R. Cayolla, "Business intelligence and data mining to support sales in retail," *Marketing and Smart Technologies. Smart Innovation, Systems and Technologies*, vol. 167, pp. 406 – 419, 2019.

[19] T. Gang, C. Kai, and S. Bei, "The research application of business intelligence system in retail industry," in *2008 IEEE International Conference on Automation and Logistics*, 2008, pp. 87–91.

[20] H. Chen, R. H. L. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," *MIS Quarterly*, vol. 36, no. 4, pp. 1165–1188, 2012. [Online]. Available: http://www.jstor.org/stable/41703503

[21] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques.* Massachusetts, USA: Morgan Kaufmann, 2012, vol. 3.

[22] M. J. A. Berry and G. Linoff, *Data mining techniques for marketing, sales and customer support.* New Jersey, USA: John Wiley and Sons, 2011.

[23] U. Fayyad, G. Piatetsky-shapiro, P. Smyth, and T. Widener, "The kdd process for extracting useful knowledge from volumes of data," *Communications of the ACM*, vol. 39, pp. 27–34, 1996.

[24] D. T. Larose, *DISCOVERING KNOWLEDGE IN DATA: An Introduction to Data Mining.* New Jersey, USA: John Wiley and Sons, 2005.

[25] M. Kaur and S. Kang, "Market basket analysis: Identify the changing trends of market data using association rule mining," *Procedia Computer Science*, vol. 85, pp. 78–85, 2016.

[26] Y.-L. Chen, K. Tang, R.-J. Shen, and Y.-H. Hu, "Market basket analysis in a multiple store environment," *Decision Support Systems*, vol. 40, no. 2, pp. 339 – 354, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167923604000685

[27] X. Wu and V. Kumar, *The Top Ten Algorithms in Data Mining.* Minnesota, USA: Champan and Hall, 2009.

[28] J. Báez, C. Paredes, G. Sosa, and M. E. García-Díaz, "Descubriendo reglas de asociación en bases de datos del sector retail usando r," *In XXIV Congreso Argentino de Ciencias de la Computación (La Plata)*, pp. 432–441, 2018.

[29] C. R. Salvatella, "The java large text files editor," 2019. [Online]. Available: http://jalatext.com/

[30] J. P. Lucas, "Métodos de clasificación basados en asociación aplicados a sistemas de recomendación," Ph.D. dissertation, Departamento de Informática y Automática. Universidad de Salamanca, Oct 2010.

[31] J. L. Domínguez, "Una propuesta determinista para la obtención de reglas en problemas de minería de datos," 2019. [Online]. Available: http://hdl.handle.net/10272/16246

[32] L. A. Aburto Lafourcade, "Machine learning methods to support category management decisions in the retail industry," 2019. [Online]. Available: http://repositorio.uchile.cl/handle/2250/174696

[33] R. Khandelwal, Divyasharma, and H. Kanwar, "Analysing customer's purchasing pattern by market basket analysis," *i-manager's Journal on Computer Science*, 2019, visto 04/08/2019.

[34] Orange, "Orange3-associate 1: Frequent itemsets," 2016, [Web;Ultimo acceso el 07-01-2020]. [Online]. Available: URL{https://orange3-associate.readthedocs.io/en/latest/widgets/frequentitemsets.html}

[35] M. C. and Z. H., "Basket analysis in practice: Mathematical models and applications in offline retail," *Performance Management in Retail and the Consumer Goods Industry*, 2019. [Online]. Available: https://doi-org.ezproxy-cicco.conacyt.gov.py/10.1007/978-3-030-12730-5_24

[36] J. Torres and C. L. Abad, "Análisis comparativo de mecanismos de minería de datos para la generación de reglas de asociación aplicables a caches de grandes datos," *Revista Tecnologica ESPOL*, 2015, ultimo acceso 06/11/2019.