*Proceedings of the First Australian International Conference on Industrial Engineering and Operations Management, Sydney, Australia, December 20-21, 2022*

# A Product Network Analysis Using A Priori Algorithm for Extending the Market Basket in Retail

**Gonzalo Manuel Infante Caldas**
Universidad de Lima
Carrera de Ingeniería Industrial
Lima, Perú
20180936@aloe.ulima.edu.pe
(https://orcid.org/0000-0002-5315-042X)

**Xiomara Dayana Molina Rubio**
Universidad de Lima
Carrera de Ingeniería Industrial
Lima, Perú
20181210@aloe.ulima.edu.pe
(https://orcid.org/0000-0003-4524-0844)

**Yvan Jesus Garcia Lopez**
Universidad de Lima
Carrera de Ingeniería Industrial
Lima, Perú
ygarcia@ulima.edu.pe
(https://orcid.org/0000-0001-9577-4188)

**José Antonio Taquía Gutiérrez**
Universidad de Lima
Instituto de Investigación Científica
Lima, Perú
jtaquia@ulima.edu.pe
(https://orcid.org/0000-0002-1711-6603)

## Abstract

Market basket analysis provides an insight into customer consumption patterns and trends in the industry. These will be achieved by analyzing and studying the performance of the large datasets of transactions made by consumers held in retail stores. These commercial transactions will be analyzed using the Machine Learning technique called the A priori algorithm by establishing association rules and determining those groups of items in a market basket whose association could represent better economic benefits for companies. This study will analyze the historical sales data of the product groups, in order to identify relationships that al-low companies in the sector to generate patterns to propose the increase of their portfolio based on the products with the greatest purchasing trends. At the end of this investigation, commercial strategies will be proposed to improve sales, take advantage of spaces in stores and implement more effective strategic offers, based on the groups of articles with the best associations found.

## Keywords
Machine learning, basket analysis, portfolio extension, retail, a priori algorithm.

# 1. Introduction

The retail industry is facing great changes as a result of the abnormal fluctuation in demand and the change in consumer habits as a result of the COVID 19 pandemic. In addition, there is the constant struggle of companies in the sector to maintain and increase their profits and offer the products and services that the customer wants to obtain (Kim, 2020). This highlights the need for retailers to provide a more specialized service, offer a greater variety of products and increase their sales are the main problems to be faced in this generation of post-pandemic organizations. Likewise, the need to use data and sales analysis tools based on data regression algorithms is observed, these algorithms being part of Machine Learning methodologies.

Nowadays, with the increase of globalization and the advance of technology, retail companies are constantly struggling to maintain and increase their profits. Having large amounts of data collected from commercial transactions, gives rise to the need to analyze them to extract considerations of products and consumer purchases, this can occur through the application of Machine Learning techniques (Moldenhauer and Zwirnmann, 2019).

This research will seek to use the Machine Learning methodology in companies in the retail sector to forecast the demand for certain products, analyze the historical sales data and its variation, in order to evaluate the products with the greatest potential to expand the customer's market basket. Currently, there are many tools within the study methodology, however, in this report the use of the A priori Algorithms tool will be evaluated. By performing the analysis of the shopping basket and employing Apriori Algorithms to determine the associations of the different products, the portfolio extension will proceed and will allow retail companies to know consumer behavior and plan the sales and marketing strategies to be applied to have greater visibility and revenue (Moldenhauer and Zwirnmann, 2019).

## 1.1 Objectives

General objective: Analyze the correlation between increasing the product network and expanding the purchase network within companies in the retail sector, in order to generate healthy margins in the industry and increase the brands present in the cata-log of retail companies.
Specific objectives:

- Propose the increasement of the portfolio of new national and imported products in the study sector.
- Identify strategies that increase sales and generate healthy margins in companies in the retail sector.
- Interpret and compare the main statistical indicators of the A priori Algorithms

# 2. Literature Review

## 2.1 Retail

The retail industry is comprised of retail companies that sell consumer, non-food and pharmaceutical products. Currently, these companies are in a constant struggle to maintain and increase their profits and to offer the products and services that the customer (Martinez and Diaz-García, 2021). Many of these companies seek to stand out by means of a differential factor based on a price approach or by providing personalized service to the consumer and to achieve this they require an analysis of their buyers and the interaction they have with their products, in order to propose strategies that allow them to make effective and efficient commercial decisions.

## 2.2 Market basket analysis

According to Kumar, Kashyap and Gayathri (2021), Market Basket Analysis or (MBA), also known as association rule learning or affinity analysis, is a data mining technique that can be used in many domains.

Particularly, in the study sector, MBA allows the retailer to understand the buyer's buying behavior, which can help them in proper decision making and have a new way to understand the variation in data (Kumar et al. 2021). Currently, this technology is in its infancy within the study sector, which denotes a whole new range of possibilities.

## 2.3 Rules of association

According to Christian et al. (2021), in a market basket analysis, items A and B have a frequency together, to find the most representative relationship, the criteria of support, confidence and elevation should be used to measure the association rules.

Support is used to express the proportion of each identified association (1), confidence refers to the probability that customers who receive product group "A" will also receive product group "B", measuring the strength of the relationship between the items (2) and elevation indicates the relationship between the support and the expected one if A and B are independent (3). In equation (1), N represents the amount of data.

$$\text{Support } (A \Longrightarrow B) = \frac{\text{frequency}(A, B)}{N} \qquad (\mathbf{1})$$

$$\text{Confidence } (A \Longrightarrow B) = \frac{\text{frequency}(A, B)}{\text{frecuency}(A)} \qquad (\mathbf{2})$$

$$\text{Lift } (A \Longrightarrow B) = \frac{\text{confidence}(A, B)}{\text{support } (B)} \qquad (\mathbf{3})$$

### 2.4 Apriori Algorithm

One of the data regression approaches is based on employing Machine Learning methodology, which consists in the use of a priori algorithms based on finding time series regressions within the databases of different fields, since it can be applied to many sectors, such as: industrial, nuclear, medicine and many others (Pavlyshenko, 2019).
According to Maske and Joglekar (2018), the apriori algorithm allows generating sets conformed by recurrent elements, thus expanding from set to set until it no longer finds, more valid and successful extensions.

## 3. Methods

The following research will be quantitative explanatory applied because it will seek to analyze the data obtained from retail companies and provide a description of their status before, during and after the use of apriori algorithms to optimize sales and identify products with purchase potential. It is also applied, since it seeks to identify products with potential to increase the portfolio of groups of articles or products, considering the number of associations obtained. Python software will be used as tools for the simulation and the application of Machine Learning techniques. The variables used in this research are the items sold, the transactions carried out and the groups of items.

With respect to the data, it should be noted that to carry out a Market Basket Analysis, the transactions made by customers should be collected, which allows identifying which products are the most demanded when analyzing purchasing patterns through the identification and formation of essential associations between products (Christian et al, 2021).

The hypotheses proposed considering the information collected are the following:
H1: The Machine Learning methodology will allow the identification of products through the association rule, which will lead to the detection of new related products.
H2: The analysis of historical consumer data will facilitate the proposal of successful commercial strategies.
H3: The application of Machine Learning methodology will optimize the analysis of products and trends in the retail sector.

## 4. Data Collection

Considering the above, the data used for the research are transactional data from a supermarket chain of the retail sector recognized in Latin America that has chains of hypermarkets and supermarkets; specifically, the article focused on the purchase records of specific stores located in Lima, Peru. For this study the original name of the company will be omitted for confidentiality reasons, therefore the name Company Alpha will be used. The transactions extracted as data are from a store representative of the location where the company has operations. It should be noted that we worked with data from the year 2021.

The data first went through the collection stage, then it went through a series of validation processes and data cleaning to be used within the algorithms proposed in this report.

Entering in greater detail to the information used for running the algorithms, this went through a statistical sieve, to eliminate values that were not related to the products sold within the company, since within the transactions there was evidence of services that would not be useful for the study as ticket rounding, discounts, etc.

Also, to simplify the variables of the report, the different presentations and weights of the products were conglomerated in groups of articles that would allow better results in the association rules. Consider that the process through which the data passes is represented in Figure 1.

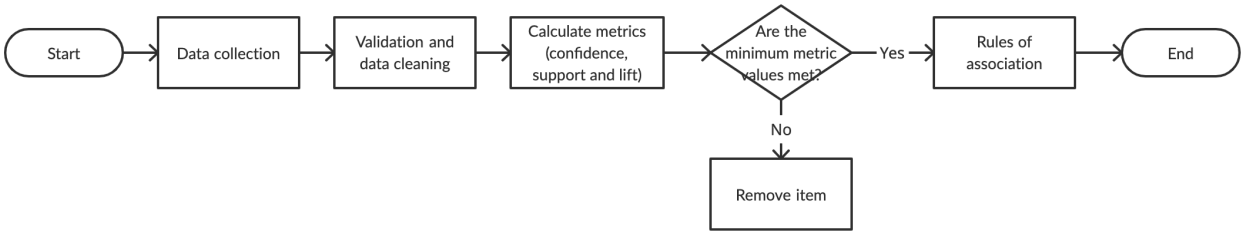


Figure 1. Processing data

## 5. Results and Discussion

### 5.1 Numerical Results

According to the information collected and in the first graphs of the algorithm, it is possible to observe the products with the greatest support in sales within the period of time that comprise the historical data available, they are grouped in baskets of up to three products. In particular, in Table 1 it is possible to validate that the products with the greatest purchase support are basic necessities such as water and vegetables. Particularly, this table is confirmed by baskets of combinations of a single product.

Taking into account the results in Table 2, it can be seen that the combinations of items with the highest purchase support are those that take the individual products shown in Table 1.In greater detail, Table 2 takes into consideration baskets of combinations of up to two products.

Table 1. Sales support per item

| Support | Itemset |
|---|---|
| 0.3359 | {'PLASTIC BAGS'} |
| 0.1220 | {'PACKAGED BROWN EGG'} |
| 0.1135 | {'PREPARED VEGETABLES'} |
| 0.0989 | {'GRASSES AND LEAVES'} |
| 0.0986 | {'STILL WATER'} |

Table 2. Sales support by item groups

| Support | Itemset |
|---|---|
| 0.0446 | {'VEGETABLE SOUP/ SALAD', 'HERBS AND LEAVES'} |
| 0.0446 | {'PACKAGED BROWN EGG', 'PLASTIC BAGS'} |
| 0.0441 | {'VEGETABLE SOUP/ SALAD', 'PREPARED VEGETABLES'} |
| 0.0419 | {'HERBS AND LEAVES', 'PREPARED VEGETABLES'} |
| 0.0404 | {'PLASTIC BAGS', 'PREPARED VEGETABLES'} |

### 5.2 Graphical Results

Continuing with the graphical part of the results achieved, you can see in Figure 2 all the associations made by the algorithm throughout the entire period of time covered by this investigation and the supports and trusts of each of them are taken as bases. these.

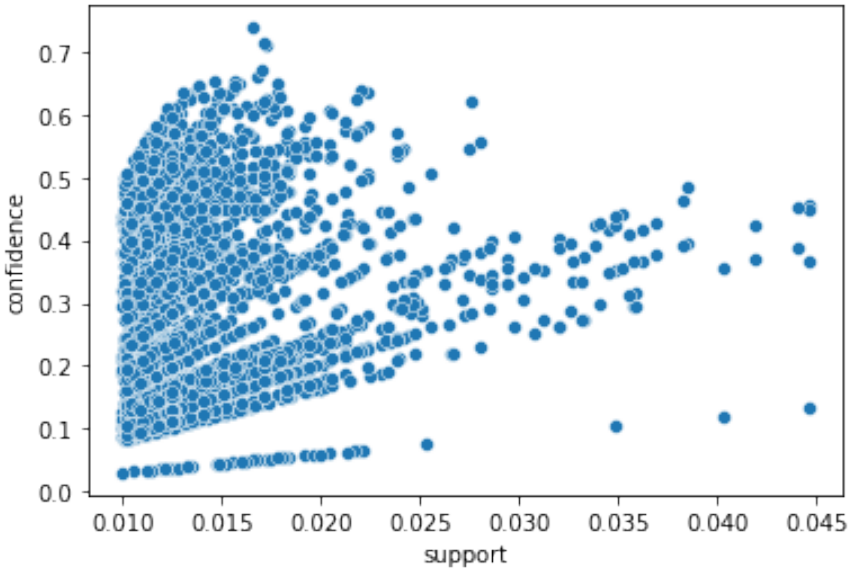


Figure 2. Total association rules

In order to identify the associations with the highest probabilities and repetitions within the entire algorithm, filters were applied, which will be mentioned in the following section. The graphical results achieved can be seen in Figure 3. The Figure shows graphically those associations that are most likely to arise or with the highest rates of chances to occur.

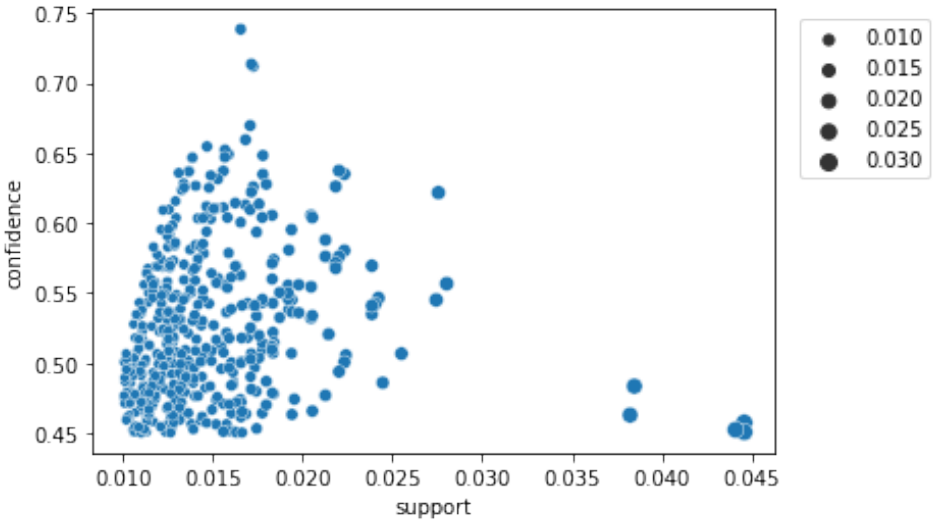


Figure 3. Filtered Association Rules.

Figure 4 is a network graphic, created in the NetworkX library program, used to identify the top rules taking the lift of each of these as the weighting variable. This graph creates a set of antecedents and consequences after sorting the total number of rules. Considering that a priori models are unsupervised techniques, it is very important to simplify

the analyst's interpretation experience to make better decisions. In the graph below lift value appear between items connections.

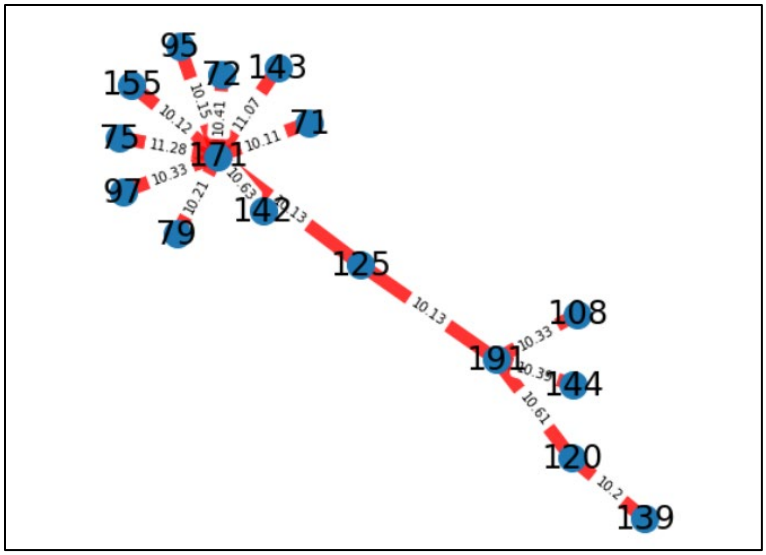


Figure 4. Network graph with highest lift values

### 5.3 Proposed Improvements

To carry out the selection of the associations with the best results, filters were applied, where the support of the antecedents and consequents are greater than 0.02 and 0.01, respectively, as well as the confidence and elevations of the associations must be greater than 0.45 and 1, respectively. These modifications are represented in Figure 1 and the best association rules are represented in Table 3.

Table 3. Best Association Rules

| Antecedents | Consequents | Ant support | Cons support | Support | Confidence | Lift |
|---|---|---|---|---|---|---|
| {'GRASSES AND LEAVES', 'CARROTS/GRAINS'} | {'VEGETABLE SOUP/ SALAD'} | 0.022 | 0.097 | 0.016 | 0.738 | 7.574 |
| {'TUBERCLES AND ROOTS', 'CARROTS/GRAINS'} | {'VEGETABLE SOUP/ SALAD'} | 0.024 | 0.097 | 0.017 | 0.713 | 7.318 |
| {'CARROTS/GRAINS', 'TOMATOES AND PEPPERS'} | {'VEGETABLE SOUP/ SALAD'} | 0.024 | 0.097 | 0.017 | 0.712 | 7.303 |
| {'VEGETABLE SOUP/ SALAD', 'ONIONS AND PEPPER'} | {'TUBERCLES AND ROOTS'} | 0.025 | 0.082 | 0.017 | 0.669 | 5.107 |

The table can be interpreted as the association rules that are most likely to be raised. This can be seen due to their high confidence indices between combinations of groups of items. Based on these results, future decisions could be made that affect strategies against the competition.

### 5.4 Source ode

Next, the source code used in the Python software will be detailed in order to analyze the collected data. As can be seen, the code was used to analyze only baskets of combinations of 3 groups of items, likewise, a minimum support of 0.006 was taken.

```
frequent_itemsets = apriori(df,
                            min_support = 0.006,
                            max_len = 3,
                            use_colnames = True)
frequent_itemsets.head(500)
```

Due to the amount of data and to present a table detailing the associations with a support greater than 0.01, the following code was proposed.

```
rules = association_rules(frequent_itemsets,
                            metric = 'support',
                            min_threshold=0.01)
```

Finally, with the data already displayed, it was possible to identify the best associations through the application of filters in the support of the antecedents and consequent, as well as in the confidence of each association identified by the algorithm.

```
filtered_rules = rules[(rules['antecedent support'] > 0.02)and
                        (rules['consequent support'] >0.01) and
                        (rules['confidence'] > 0.01) and
                        (rules['lift'] > 1.0)]
```

### 5.5 Discussion

For future research, it is proposed to test the effectiveness of the portfolio extension and commercial strategies taken based on the information obtained from the MBA such as customized promotions, cross-selling and space management.

### 5.1 Portfolio extension

As can be seen in Table 3, the products with the strongest associations are those of first necessity, emphasizing vegetables. However, as a retail company, CompanyAlpha sells its vegetables as a single unit and generally does not have brands on them, so it would not be of much benefit for the company to increase the brands of this group.

As an alternative measure, the data was filtered again, emphasizing the products that are characterized by having commercial brands, whether they are own, national or imported brands, this new data can be seen in Table 4.

Table 4. Best groups of products with brands per support

| Support | Itemset |
|---|---|
| 0.3359 | {'PLASTIC BAGS'} |
| 0.1220 | {'PACKAGED BROWN EGG'} |
| 0.0986 | {'STILL WATER'} |
| 0.0629 | {'CANNED FISH.TUNA'} |
| 0.0624 | {'PACKAGED HAM'} |

| | |
|---|---|
| 0.0538 | {'FAMILY YOGURT FUNC'} |
| 0.0533 | {'FAMILY YOGURT REG'} |
| 0.0529 | {'REGULAR BUTTER'} |

## 5.2 Commercial strategies

*Customized promotions*

According to Moldenhauer and Zwirnmann (2019), analyzing the shopping basket of each buyer allows segmenting customers according to their demographic traits, life stage or place of residence. Additionally, it is possible to identify and design new promotional discounts or pricing.

Accordingly, knowing the associations of groups of items allows companies to customize their promotional and advertising actions depending on the target groups.

Based on the results obtained from the apriori algorithm and using the lift or better known as elevation as the main indicator. By obtaining a lift greater than 1, it can be affirmed that there is a strong association between the antecedent and the consequent. A strategy that should be applied is to apply discounts in the groups of items that present a higher lift, as an example, according to Table 5, would 'EXTRA-GDO 1 RICE, 'CONSERVA PESCAD.TUNA', 'WHITE BREAD WITH CRUST', among others.

Table 5. Association Rules.

| Antecedents | Consequents | Ant support | Cons support | Support | Confidence | Lift |
|---|---|---|---|---|---|---|
| {'RICE EXTRA-GDE 1'} | {'BLOND SUGAR'} | 0.043 | 0.031 | 0.010 | 0.240 | 7.636 |
| {'CANNED FISH.TUNA'} | {'LONG NOODLES'} | 0.063 | 0.036 | 0.011 | 0.179 | 4.909 |
| {'RICE EXTRA-GDE 1'} | {'ONION OILS AND CHILI PEPPERS'} | 0.043 | 0.050 | 0.010 | 0.237 | 4.710 |
| {'CANNED FISH.TUNA'} | {'RICE EXTRA-GDE 1'} | 0.063 | 0.043 | 0.012 | 0.196 | 4.586 |
| {'WHITE BREAD W/CRUSTS'} | {'PACKED HAMS'} | 0.051 | 0.062 | 0.013 | 0.251 | 4.017 |
| {'CANNED FISH.TUNA'} | {'CRACKERS'} | 0.063 | 0.047 | 0.011 | 0.174 | 3.711 |
| {'VEGETALES OILS'} | {'PACKED BROWN EGG'} | 0.028 | 0.122 | 0.012 | 0.416 | 3.410 |

*Cross selling*

According to Istrat, V., and Lalić, N. (2017), the identification of associations allows evaluating the implementation of cross-selling or cross-selling.

Applying a cross-selling strategy allows companies to reduce high inventories, these can be between the same brand or different brands.

Based on the results obtained in the table 5, companies could implement cross-selling between 'RICE EXTRA-GDO 1' and 'BLONDE SUGAR', 'CANNED FISH.TUNA' and 'LONG NOODLES', 'CANNED FISH.TUNA' and 'SALT BISCUITS', among others.

*Space management*

According to Musalem et al (2018), the analysis of shopping baskets allows identifying guidelines for the design of the distribution of store spaces and when the objective is to reduce the turnover time of products in the store or minimize the time that the customer takes to make purchases, products should be located based on the probability of joint purchase of customers.

Retail companies must consider the strongest associations of the products presented by their customers so that the latter have a pleasant experience and drive the purchase of the products together.

Using the data in the Table 5, the following are proposed to be placed on nearby shelves, considering the lift they present:

- 'WHITE BREAD WITH RINKS' and 'PACKAGED HAM'
- 'VEGETABLE OILS' and 'PACKAGED BROWN EGG'
- 'BLOND SUGAR' and 'PACKED BROWN EGG'
- 'CANNED FISH.TUNA' and 'LONG NOODLES'

## 6. Conclusion

**O1:** *Propose the increase of the portfolio of new national and imported products in the study sector.*

It can be concluded that, although the products with turnover rates and sales flow are the products of first need, however, many of these are not characterized by being marketed under a specific brand, but rather under individual units.

**O2:** *Identify strategies that increase sales and generate healthy margins in companies in the retail sector.*

It is highly important to recognize the usefulness of the algorithm to automate the process of identifying potential sales, since this allows not only to speed up this process, but also to make commercial decisions for the elaboration of new effective strategies.

**O3:** *Interpret and compare the main statistical indicators of the A priori Algorithms*

It can be concluded that, although it is true that support represents a key indicator in the structuring of all association rules, this is not a factor with which just having it is enough to make trading decisions.

## References

Chandwani, M. Market basket analysis using association rule. International Journal of Advance Research, Ideas and Innovations in Technology, 3-4, 2018.

Christian, M. A., Nathanael, N., Mauliani, A., Indrawati, A., Manik, L. P., and Akbar, Z. Re-al Market Basket Analysis using Apriori and Frequent Pattern Tree Algorithm. The 2021 International Conference on Computer, Control, Informatics and Its Applications,161- 165, 2021.

Istrat, V., and Lalić, N. Creating a Decision-Making Model Using Association Rules. Applied Artificial Intelligence, 538–553, 2017.

Kim, R. Y. The impact of COVID-19 on consumers: Preparing for digital sales. IEEE Engineering Management Review, 2-11, 2021.

Kumar, D., Kashyap, R., and Gayathri, N. Market basket analysis: Identify the changing trends of market data using association rule mining. Annals of the Romanian Society for Cell Biology, 51-59, 2021.

Marcos Martinez, B., and María García-Díaz, D. Market basket analysis with association rules in the retail sector using Orange. Case Study: Appliances Sales Company. CLEI Electronic Journal, COLOCAR PAGINAS, 2021

Maske, A., and Joglekar, B. Survey on Frequent Item-Set Mining Approaches in Market Basket Analysis. 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 1-7, 2018.

Moldenhauer, C., and Zwirnmann, H. Basket Analysis in Practice: Mathematical Models and Applications in Offline Retail. In Performance Management in Retail and the Consumer Goods Industry, 369-384, 2019.

Musalem, A., Aburto, L., and Bosch, M. Market basket analysis insights to support category management. European Journal of Marketing, 4-20, 2018.

## Biographies

**Gonzalo Manuel Infante Caldas** is a candidate to receive the title of industrial engineer from the Faculty of Engineering and Architecture of the University of Lima, Lima, Peru.

**Xiomara Dayana Molina Rubio** is a candidate to receive the title of industrial engineer from the Faculty of Engineering and Architecture of the University of Lima, Lima, Peru.

**Garcia-Lopez Yvan Jesus is PhD** (c) in Engineering and Environmental Science, UNALM, “Master of Business Administration” from Maastricht School of Management, Holland, and master’s in strategic business administration from Pontificia Universidad Católica del Perú. "Master of Science" in Computer Science, Aerospace Technical Center

- Technological Institute of Aeronautic, Brazil. Stage in Optimization of Processes and Technologies, University of Missouri-Rolla, USA, and Chemical Engineer from the National University of Callao. Specialization Study in Digital Transformation, by Massachusetts Institute of Technology, Business Analytics, Wharton School of Management, Data Science by University of California, Berkeley, Big Data and Data Scientist by MITPro, USA Postgraduate Professor: Specialized Master from IT, MBA Centrum Católica, MBA from Calgary, Canada, and Centrum Católica. Principal Consultant DSB Mobile, Executive Director of Optimiza BG, advisor to the Office of Electronic Government and Information Technology (ONGEI) - PCM, Managing Director of Tekconsulting LATAM, Executive Director of Optimiza Business Group, Ex- Vice Dean of Information Engineering of the Universidad del Pacifico, Former Information Technology Manager of "MINERA CHINALCO PERU" Subsidiary of the Transnational Aluminum Corporation of China, Beijing, China. Former Manager of Systems and Communications of Maple Energy PLC, Director of Information Technology of Doe Run Peru SRL, Project Manager in implementation of ERP SAP, EBusiness Suite - Oracle Financial and PeopleSoft. Process Analyst in transnational companies Fluor Daniel Corporation-USA, PETROBRAS-Brasil, Petróleos del Perú. He has more than 25 years of extensive experience in the management of investment projects, execution, and commissioning in Peru, Colombia, USA, Brazil, China.

**José Antonio Taquía** is a Doctoral Researcher from Universidad Nacional Mayor de San Marcos and holds a Master of Science degree in Industrial Engineering from University of Lima. He is a member of the School of Engineering and Architecture teaching courses on quantitative methods, predictive analytics, and research methodology. He has a vast experience on applied technology related to machine learning and industry 4.0 disrupting applications. In the private sector he was part of several implementations of technical projects including roles as an expert user and in the leading deployment side. He worked as a senior corporate demand planner with emphasis on the statistical field for a multinational Peruvian company in the beauty and personal care industry with operations in Europe and Latin America. Mr. Taquía has a strong background in supply chain analytics and operations modeling applied at different sectors of the industry. He is also a member of the Scientific Research Institute at the Universidad de Lima being part of the exponential technology and circular economy groups. His main research interests are on statistical learning, predictive analytics, and industry 4.0.