NOVA IMS Information Management School

Master's Degree Program in

# MDSAA

**Data Science and Advanced Analytics**

## Business Cases with Data Science

Case 3: Recommendation System

Ana Caleiro, 20240696
Duarte Marques, 20240522
Moeko Mitani, 20240670
Oumayma Ben Hfaiedh, 20240699
Sarah Leuthner, 20240581

Group Q

# 1. INTRODUCTION

In an effort to increase their market share, businesses must create a personalized experience for their clients. In e-commerce, the way to achieve it is through recommendation systems. These systems will analyze the customer's historical data, and suggest products based on previous purchases, on what similar customers are buying or just show recommendations based on the items themselves.

However, to correctly recommend, it is also important to have good customer segmentation. Without it, the recommendations will not match what the client might want to try, and instead of increasing our profits, we will be losing money implementing a system with no returns.

In this project, we tackled both problems: create a sound customer segmentation and a recommendation system that covers all possible scenarios. Using the CRISP-DM framework, K-Means, Collaborative Filtering, Apriori and Association Rules, and a combination of functions of our own creation, we aim to take the first steps for the possible implementation of this project, conceive interfaces of what the systems should look like and finally address the next steps of this endeavor.

# 2. BUSINESS UNDERSTANDING

## 2.1. BACKGROUND

With a legacy of over 50 years, we, RECHEIO, operate as a supermarket tailored specifically to serve business clients, offering solutions beyond the traditional retail customer.

Our value lies not just in product variety, but in ensuring high-quality, consistency, and specialized goods that precisely meet our clients' needs. Unlike traditional wholesalers, RECHEIO emphasizes service excellence, reliability, and sustainable growth.

While RECHEIO operates physical stores, a significant part of our model is built around delivery services. However, this creates challenges, particularly in logistics and maintaining strong customer relationships. To address these, we are actively undergoing a digital transformation to streamline operations and deliver a more responsive experience.

## 2.2. BUSINESS OBJECTIVES

Our main goal is to increase RECHEIO's share of total purchases of our clients. We want customers to choose us for more of their procurement needs. To do this, we aim to develop a recommendation system that enhances the relevance of our offerings and encourages greater product adoption across all channels. Two critical questions we want to answer are:

1. **How can we enrich our customer data with meaningful features to improve recommendations?**
2. **How can we deliver relevant, personalized recommendations through the right channels?**

## 2.3. BUSINESS SUCCESS CRITERIA

The success of this project will be evaluated based on the following criteria:

1. Accurately identify customer segments through improved clustering techniques.
2. Develop a recommendation system that leads to increased customer purchases.

3. Strengthen customer engagement and retention.
4. Increase the percentage of total purchases made through our platform.
5. Ultimately, this project aims to enhance customer satisfaction and loyalty while driving revenue growth through intelligent, data-driven recommendations.

# 3. METHODOLOGY

In this project, the CRISP-DM model is utilized to understand business problems, to deploy the model, and to ensure the successful execution of this project.

## 3.1. DATA UNDERSTANDING

The dataset comprises three tables: **Clients**, **Products**, and **Transactions**. It includes information on 6,462 customers and 4,583 unique products. Over the course of 2022, a total of 884,099 transactions were recorded. The following chapters will explore various aspects of this data in detail.

### 3.1.1. Key Findings and Trends

Table 1 shows the key findings and trends of the customers in the dataset per unique feature.

*Table 1: Key Findings and Trends*

| Feature | Description |
|---|---|
| ZIP Code | The clients are from 261 different ZIP Code areas. 14,35% of the customers are from the ZIP Code 4050. |
| Client Type | The clients are divided into 27 types. Most customers (if Client Type available) are from Hospitality, Institutions/Canteens or Portuguese Restaurants. |
| Production Description | At RECHEIO there are 4,583 individual products available (4,573 unique) |
| Product Category | The products can be divided into 268 categories. |
| Transactions Date | There are 884,099 transactions within the year 2022. Most products are sold in summer and fall with the peak at October (60,000 – 80,000) as well as the start and the end of the week. The least products are sold in January, February and December as well as the middle of the week, especially Thursdays. Around 1,000 customers buy every month at RECHEIO. |

### 3.1.2. Data Anomalies

In the datasets, all columns have the correct metadata and data type. In table 2 are the key findings regarding anomalies in the dataset.

*Table 2: Key Findings and Trends*

| Duplicates | |
|---|---|
| Product Description | Possibility of duplicates as there are 4,583 products, but only 4,573 unique values. This is due to similar products sharing the same description in the Product set; however, the origin or packaging differ. |

| | |
|---|---|
| ZIP Code - Client Type | Duplicates in the table client only show that clients are from the same area with the same type of establishment (no duplicates in index). |
| Duplicate Rows in Table Transactions | Duplicate in this case are customers buying the products more than once. |
| **Missing Values** | |
| ID Client Type | 5,782 missing values (~90%) |

As there are no major anomalies, no treatments of such are necessary. The column ID Client Type kept, even though of most missing values due to information loss. They are used to validate clustering. There is also no need of outlier treatment, since we are using real transactional data. Removing clients that buy more frequently should not be removed, since they are the best customers.

## 3.2. DATA PREPARATION

For the objectives, the datasets are merged into a single dataset based on the specific purpose. For clustering, all datasets are merged with the client table (where one row represents a single customer). In contrast, for the recommendation system, the transactions table serves as the primary dataset, to which the other data is merged.

### 3.2.1. Data Engineering

Table 3 shows the new features created with their explanation.

*Table 3: New Feature Description*

| New Feature | Description |
|---|---|
| sub_category | More general categories, by taking only the first word of product categories. |
| main_category | Overall general product categories, grouping of product categories. |
| HoReCa_category | Product categories inspired by RECHEIO website HoReCa categories. |
| HoReCa_sub_ category | Product subcategories inspired by RECHEIO website HoReCa categories. |
| transaction_count | Total number of distinct purchase days for each client during the study period. |
| product_count | Total number of products bought by the client across all transactions. |
| unique_products | Count of different products purchased by the client. |
| recency_days | Number of days since the client's last purchase, relative to the study end date. |
| first_purchase_days | Days between the client's first purchase and the start of the study period. |
| last_purchase_days | Days between the client's most recent purchase and the start of the study period. |
| avg_products_per_transaction | Average number of products per transaction for each client. |
| avg_days_between_ transactions | Divides the total days in the year (364) by the number of transactions to estimate how often (in days) the customer makes a purchase on average. |
| top_category | Top categories based on the HoReCa category with the highest value (e.g. sales, transactions, or some other metric) among the specified columns. |
| num_categories | How many unique HoReCa categories a client bought from. |
| top_subcategories | Top subcategories based on the HoReCa category with the highest value among the specified columns. |
| num_subcategories | How many unique HoReCa subcategories a client bought from. |
| repeate_rate | It shows how much of a client's purchasing is repeated. |

| | |
|---|---|
| percentage_cat_ | Proportion of total products each client has purchased from a specific HoReCa subcategories |

### 3.2.2. Scaling

To scale the data the "MinMaxScaler" was used because the data is not normally distributed.

### 3.2.3. Feature Selection

To enhance the dataset quality and minimize redundancy, we identified and addressed highly correlated numerical features using a correlation **threshold of |0.8|**. At the end, 12 features were used, including original and newly created features for clustering.

# 4. RESEARCH QUESTION 1 – DATA ENRICHMENT BY CLUSTERING

The first research question, **"How can we enrich our customer dataset with relevant information to perform better recommendations?"**, is addressed and concluded in this chapter with the results of the clustering process. However, earlier chapters, such as **Feature Engineering**, also contribute significantly to this outcome.

## 4.1. MODELING AND MODEL SELECTION

To improve the quality of the clusters, we explored three different algorithms: K-Means, SOM, and DBSCAN. We chose not to apply Principal Component Analysis (PCA) due to its impact on interpretability, which is essential for understanding and explaining our clustering results.

At the end, K-Means is chosen because of its high $R^2$ value. Furthermore, **six clusters** are identified as the optimal number for the clusters with this dataset.

## 4.2. RESULTS AND EVALUATION

The model is aligned with the company's objective of increasing customer purchases through a recommendation system. By generating meaningful customer clusters, the model uncovers insights into behavior, preferences, and loyalty patterns. These insights support more personalized and effective recommendations, helping to drive higher engagement and repeat purchases. Continuous refinement of the clusters ensures the system remains responsive and data-driven, ultimately contributing to business growth through smarter targeting and improved customer experience.

### 4.2.1. Customer Profiling

With K-Means algorithm, we identified six customer clusters. Table 4 shows the profiles of customers in each cluster. With their profiles, we identified their characteristics as follows:

1. **Cluster 0:** 160 clients (Approximately 7.5%)
2. **Cluster 1:** 381 clients (Approximately 24.9%)
3. **Cluster 2**: 600 clients (Approximately 39.2%)
4. **Cluster 3**: 93 clients (Approximately 6.1%)
5. **Cluster 4**: 180 clients (Approximately 11.8%)
6. **Cluster 5:** 115 clients (Approximately 10.5%)

*Table 4: Customer Profiling*

| # | % | Key Behavior | Product Preference | Suggested Persona |
|---|---|---|---|---|
| 0 | 10.5% | • High product count<br>• Low variety, low repeat rate<br>• Moderate recency<br>• Narrow shopping behavior | • High: Drinks, Beverages, and Spirits<br>• Medium: Produce | Drink-focused bar/café owners |
| 1 | 24.9% | • Very high product count<br>• Narrow product selection<br>• Early-year starters<br>• Low frequency | • High: Produce<br>• Medium: Grocery, Fresh | Loyal produce-heavy restaurant buyers |
| 2 | 39.2% | • Low engagement throughout the year<br>• Very high product count<br>• Infrequent bulk buyers<br>• Low variety, very focused | • High: Produce<br>• Medium: Grocery, Fresh | Bulk-buying kitchens (e.g., hotels, caterers) or wholesalers (very clear and focused purchasing patterns) |
| 3 | 6.1% | • Moderate activity (moderate engagement)<br>• Low diversity<br>• Mid-year engagement | • High: Fresh, Produce | Local small restaurants with fresh focus (they probably buy only what they need) |
| 4 | 11.8% | • Low total activity<br>• Early-year shoppers<br>• Buy infrequently and in small amounts | • High: Produce, Grocery<br>• Avoids processed and perishable goods | Small vendors or seasonal corner shops |
| 5 | 7.5% | • Low total activity (low engagement)<br>• Narrow purchases<br>• Focus on ready-to-use items | • High: Non-Food, Prepared Meals | Cafeterias, convenience stores (they are probably with alternative suppliers for produce or perishables) |

## 5. RESEARCH QUESTION 2 – RECOMMENDATION SYSTEM

Secondly, the research question **"How can we provide relevant recommendations to our customers, based on the channels we have?"** is addressed. Currently, RECHEIO operates through two channels: Stores and Delivery. Customers can make purchases in one of three ways:

- On their own, using the Website or Customer App
- In-store, using the POS or with assistance from an employee
- Through a visiting salesman

The goal is to provide personalized recommendations in each purchasing context, leveraging both the customer's purchase history and behavioral data from similar customers. This approach aims to increase RECHEIO's share of a customer's total purchases by offering more as well as timely suggestions.

## 5.1. MODELING

The suggested recommendations are composed of four key components:

- **Usual Basket/Last Purchase:** At the beginning of the purchasing process, customers are presented with items based on their historical behavior. They can quickly add frequently purchased items or replicate their last order to start a new one.
- **Smart Basket Recommendation:** During checkout, customers receive recommendations of N additional products that they have never purchased before. These suggestions are based on the customer's behavior and their similarity to others within the same behavioral cluster.
- **Substitute RECHEIO Products:** To promote RECHEIO's own brand, customers are prompted to replace certain items in their cart with RECHEIO-branded alternatives.
- **Single Item Recommendation:** Customers are shown similar or complementary products to encourage discovery and upselling.

### 5.1.1. Usual Basket / Last Purchase

To streamline the ordering process, the "**Usual Basket**" function recommends a list of frequently purchased items based on the customer's historical behavior. The number of recommended items is determined by the customer's average number of products per transaction.

Alternatively, customers can choose to repeat their exact **last purchase**, providing a quick and familiar option for reordering.

From these personalized recommendations, customers have full flexibility - they can select all, remove items, or individually choose products to add to their basket.

In case of new customers without any transaction history and no assignment of a cluster, the system recommends the top 15 most popular products across the entire customer base as a starting point.

### 5.1.2. Smart Basket Recommendation

A Smart Basket Recommendation is generated for each customer based on the cluster they belong to. Each cluster is analyzed to identify its top products, those that are frequently purchased by similar customers. Any of these products that the customer has not yet purchased will be recommended as potential additions to their basket. To generate these recommendations, several models and techniques have been tested:

- **Simple Cluster-Based Recommendation**: based on the most popular items bought by other clients in the same cluster, excluding the products the client has already added to their shopping cart.
- **Apriori and Association Rules Recommendation:** frequent item sets are discovered, and association rules are generated (e.g. customers who buy X, also buy Y).
- **Hybrid (user-based and item-based) Collaborative Filtering with Clustering:** combining the methods to benefit from a broader understanding of preferences, improving the relevance and diversity of the recommendations within a cluster.

    - **User-based Filtering** recommends products based on what similar customers (those with similar purchasing patterns) have bought.
    - **Item-based Filtering** recommends products similar to the ones the customer has already bought.
- **Hybrid (user-based and item-based) Collaborative Filtering <u>without</u> Clustering:** same method but using all users (global similarity) for a broader, more diverse product exposure.

### 5.1.3. Substitute RECHEIO Products

The approach to identify substitute RECHEIO Products leverages two text mining techniques: N-Grams and the Jaccard coefficient.

Initially, bigrams (sequence of the first two words) from non-RECHEIO product names are compared with the RECHEIO-branded products within the same product category to identify potential matches. However, one key limitation of the N-Gram approach is its inability to effectively capture long-distance dependencies within text.

To address this limitation, a second comparison is performed using the Jaccard similarity metric, which measure the similarity between two sets and is defined as the size of the intersection divided by the size of the union of the two sets [1]:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

If multiple RECHEIO products result in positive matches, the product with the highest popularity (i.e., the highest proportion of transactions) is selected as the recommended substitute. If no RECHEIO products are matched, the product with the overall highest popularity in that product category is recommended.

The following table 5 illustrates an example of a substitution recommendation for rice:

*Table 5: Example of Recommended Substitutions*

| ID Product | Product Description | ID Substitute | RECHEIO Substitute | Product Category |
|---|---|---|---|---|
| 38781 | ARROZ CAÇAROLA CAROLINO 1KG | 521206 | ARROZ CAROLINO MASTERCHEF 1KG | ARROZ |
| 84761 | ARROZ BASMATI ATLANTIC 5KG | 850533 | ARROZ BASMATI MASTERCHEF 5KG | ARROZ |
| 903221 | ARROZ VAPORIZADO AMANHECER 1KG | 903223 | ARROZ VAPORIZADO MASTERCHEF 5KG | ARROZ |

### 5.1.4. Single Item Recommendation

For customers browsing individual products, RECHEIO should provide item-based recommendations inspired by Amazon's "Customers who bought this item also bought…" feature. This method relies on Association Rules to identify products frequently purchased together. It uses three key metrics:

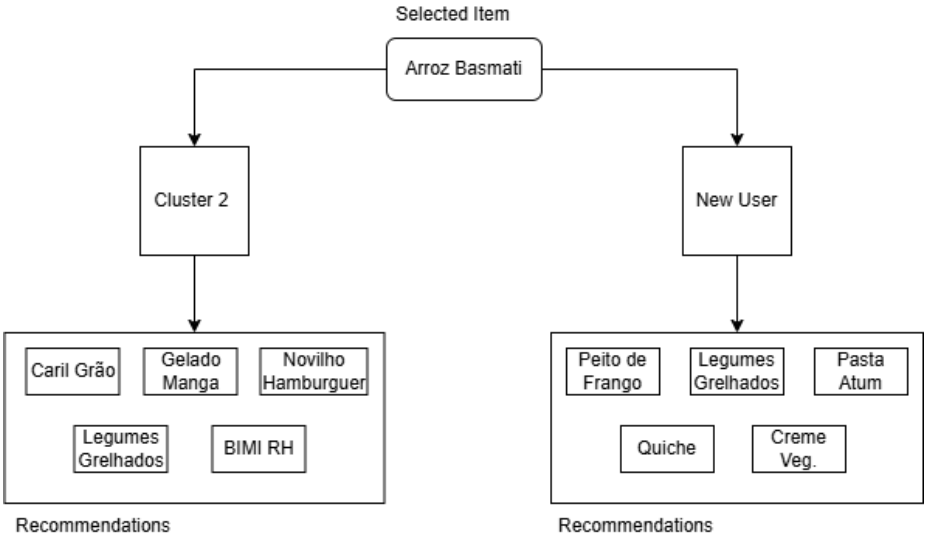
- **Support**: how often a product appears in transactions (ratio of product's frequency to the total number of transactions).
- **Confidence**: The likelihood of purchasing item B when item A is bought.

[1] https://mayurdhvajsinhjadeja.medium.com/jaccard-similarity-34e2c15fb524

- **Lift**: Whether two items are bought together more often than by chance (comparing the observed frequency of two items being bought together with what would be expected if they were independent).

To build these associations, we compared using product categories versus product description. As product descriptions produce more intuitive and customer-friendly results, this feature was chosen. For example, “SUMO SUNQUICK” and “Coca-Cola” are linked based on descriptions, whereas category-based associations would suggest “Concentrados” and “Carbonatados,” which are less specific and intuitive.

*Figure 1: Example of An Item Recommendation for Cluster 2 and A New User*




The recommendation process follows three stages:

1. **Cluster-Based Recommendations**: For known users, recommendations are generated from products with high lift values within their cluster.
2. **Global Rules Fallback**: If cluster-specific associations are weak or unavailable, the system uses the strongest product associations across all clusters.
3. **New User Recommendations**: For users without transaction history, global rules are applied directly.

This approach ensures that the customer always receives relevant and meaningful recommendations.

## 5.2. RESULT AND EVALUATION

In this step, potential models for the Smart Basket Recommendation are evaluated. Additionally, the results are illustrated through mockups of possible interface scenarios to demonstrate how the recommendations could be presented to users.

### 5.2.1. Monte Carlo Simulation

To determine which of the potential models is most suitable, **Monte Carlo Cross-Validation** is applied. This method involves conducting multiple random test scenarios, where users and products are split

in different ways, to estimate the performance of the recommendation system. In each iteration, users are randomly sampled, and the system's ability to recommend products that the customer ultimately purchases (but were hidden during the recommendation generation process) is evaluated.[2]

This technique provides robust estimates of system performance without needing to test the system in real-world conditions, due to its randomized evaluations. Precision and Recall (formulas shown below) are used as quality metrics to assess the recommendation accuracy.[3]

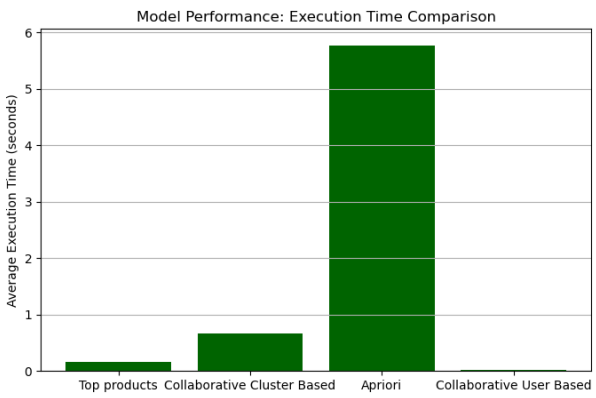
$$\text{Precision} = \frac{True\ Positives}{True\ Positives\ +\ False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives\ +\ False\ Negatives}$$

*Table 6: Results Monte Carlo Evaluation of Recommendation Models*

| Recommendation Model | Precision | Recall | Speed |
|---|---|---|---|
| **Simple Cluster-Based** | 0.012 | 0.120 | 0.128 Seconds |
| **Apriori and Association Rules** | 0.230 | 0.029 | 8.222 Seconds |
| **Hybrid Collaborative filtering with Clustering** | 0.085 | 0.420 | 0.414 Seconds |
| **Hybrid Collaborative filtering without Clustering** | 0.200 | 0.130 | 0.021 Seconds |

*Figure 2: Model Performance Time Comparison Plot*




Based on the comparison of recommendation models, the Hybrid Collaborative Filtering with Clustering model provides a strong balance between coverage and performance, with a high recall (0.420), moderate precision (0.085), and reasonable speed (0.414 seconds). The Hybrid Collaborative Filtering without Clustering model offers higher precision (0.2) and the fastest response time (0.021

[2] https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b/

[3] https://medium.com/@piyushkashyap045/understanding-precision-recall-and-f1-score-metrics-ea219b908093

seconds), but with lower recall (0.130), it retrieves fewer relevant items. The Apriori and Association Rules model shows the highest precision (0.230) but has very low recall (0.029) and the slowest speed (8.222 seconds), making it less suitable for real-time systems. The Simple Cluster-Based model is quick (0.128 seconds) but underperforms in both precision (0.012) and recall (0.12), limiting its effectiveness. Overall, incorporating clustering in hybrid models enhances recommendation coverage, though it requires a modest trade-off in speed. For this reason, we suggest moving to A/B testing with collaborative cluster-based model and collaborative user-based model to test the efficiency of using clusters in issuing recommendations.

### 5.2.2. Qualitative Assessment

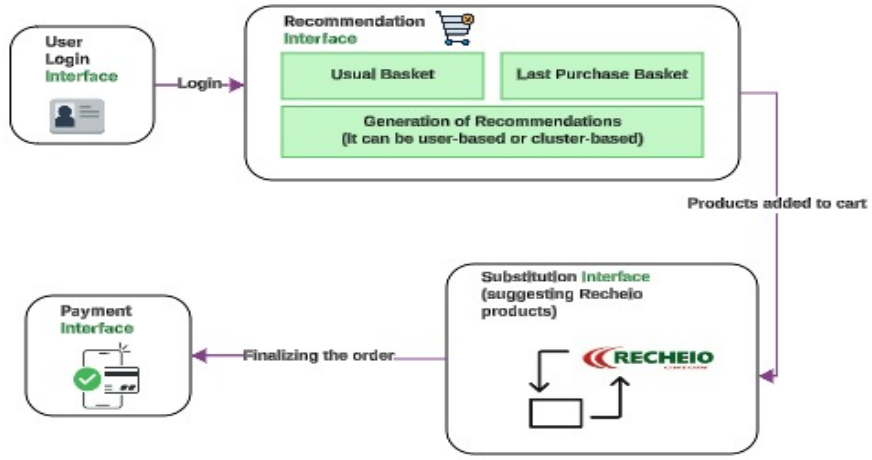
In addition to quantitative evaluation, we conducted a qualitative analysis to ensure our recommendations make sense in a business context and are not just statistically relevant.

We tested a few clients and products to check if the suggestions were appropriate. For example, a client buying White Wine from Douro received recommendations for other types of Wine and similar drinks, which aligns with common buying patterns. In the future, creating sample personas with different shopping behaviors could help further assess the quality of recommendations.

### 5.2.3. Scenarios with Results

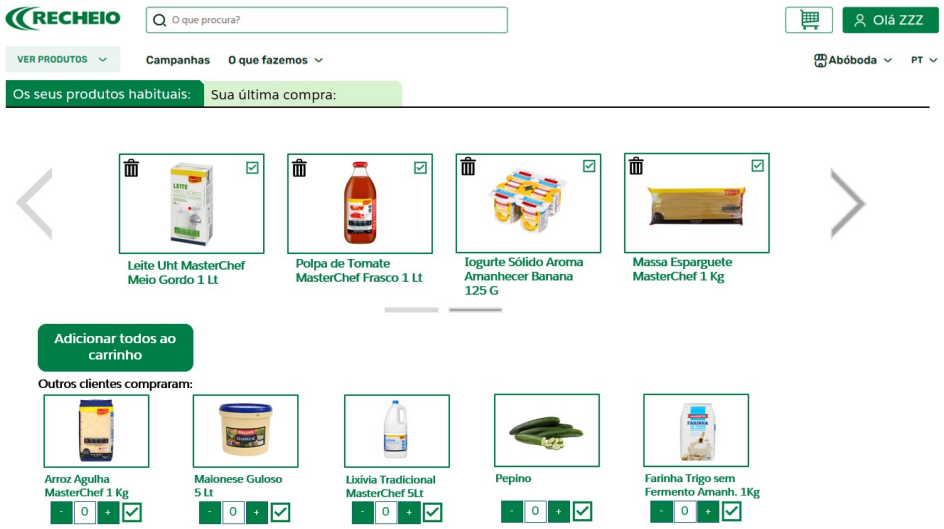
To illustrate how the algorithms are expected to work in live environment, we envisioned the website pages for all the approaches detail in chapter 5.1.

*Figure 3: Diagram of The Recommendation Interface When Opening The Shopping Cart*



**Usual Basket / Last Purchase and Smart Basket Recommendation:**

*Figure 4: Example Usual Basket and Smart Basket Recommendation*

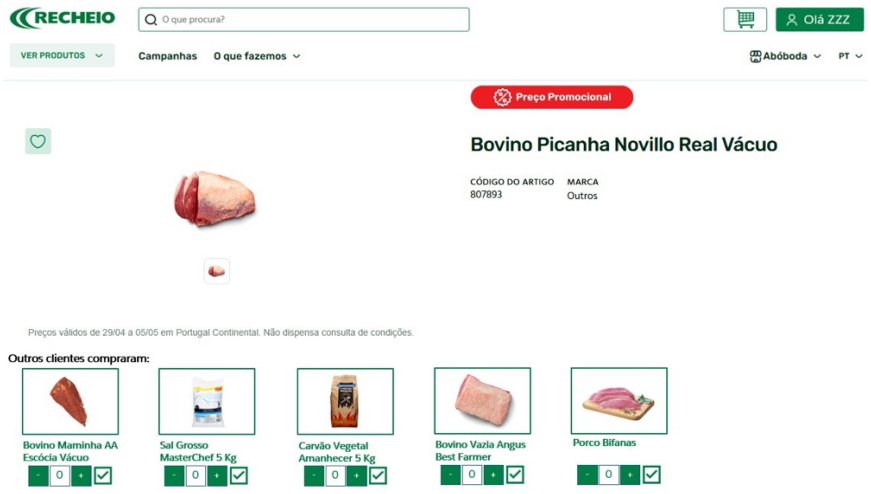



When opening their shopping cart, the user will have the opportunity to select between the items they usually buy or repeat their last purchase. In an effort to increase readability, the items are displayed horizontally with only a select few showing at the same time (the user can side scroll to see the remaining). The user can add/remove individually from the list, or they can add all to their shopping cart.

Since we want to increase our sales, we then recommend items based on their “Smart Basket”, giving them the opportunity to add the desired quantity and add to cart.

**Single Item Recommendation:**

*Figure 5: Example Single Item Recommendation*




When a client decides to browse the online store, the page of any given item will suggest products that are associated with it.

**Substitute RECHEIO Products:**

*Figure 6: Example Single Item Recommendation*

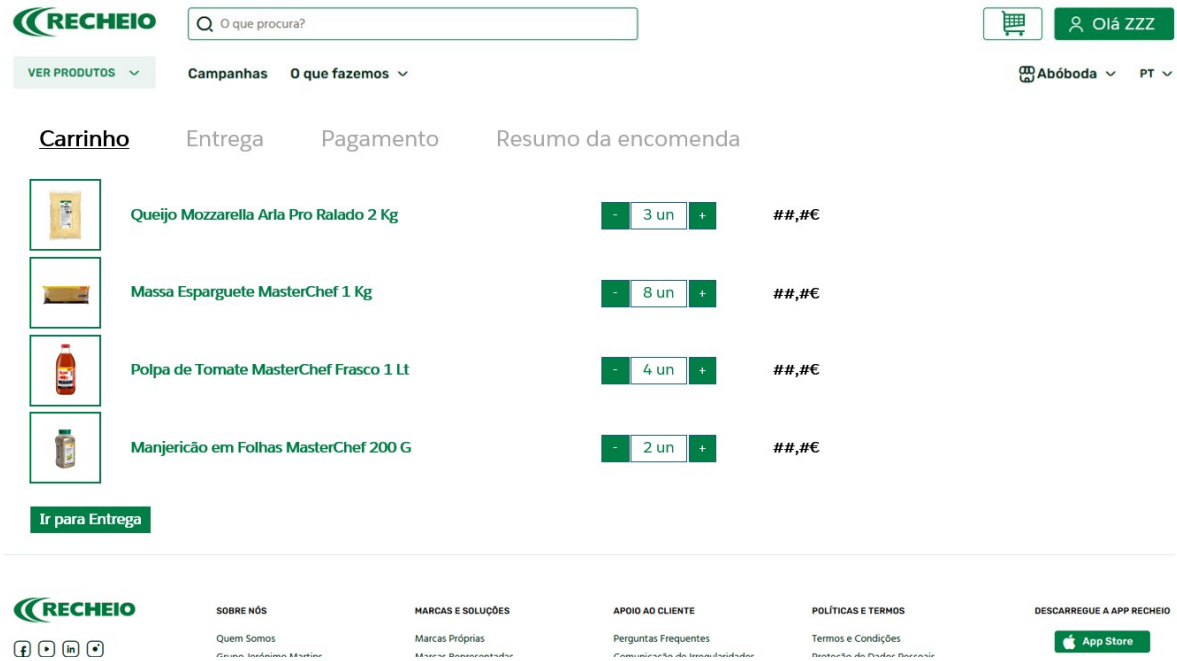
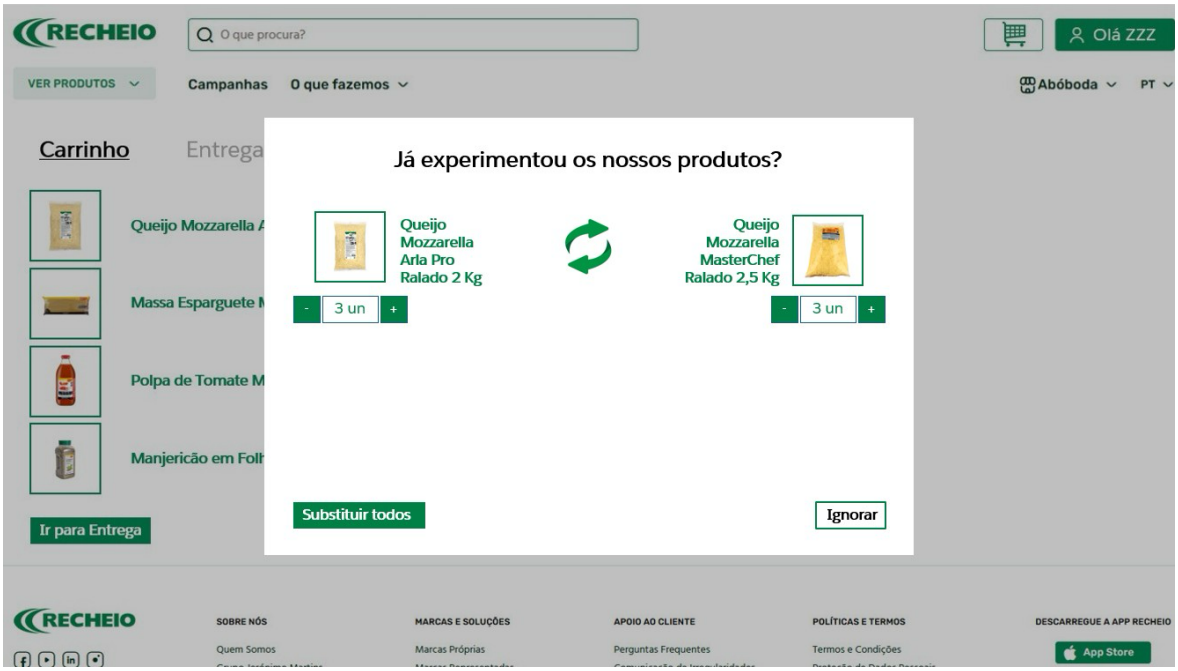
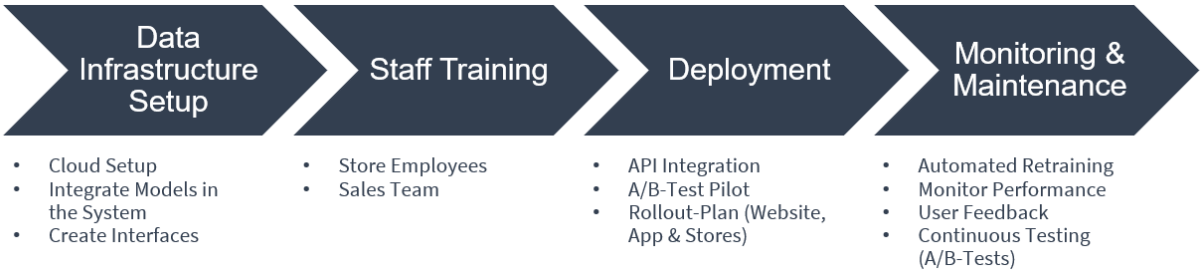



*Figure 7: Example Substitute RECHEIO Product Recommendation*




When finalizing the purchase, the system will suggest replacing the products that are not from "RECHEIO" or "MasterChef" brands with products of the aforementioned brands.

## 6. DEVELOPMENT AND MAINTENANCE PLANS




To implement this recommendation system, we recommend RECHEIO to follow four essential steps: Data Infrastructure Setup, Staff Training, Deployment and Monitoring & Maintenance.

If not already in place, the first step is to select and establish a cloud-based infrastructure. Common industry platforms include Google Cloud Platforms (GCP), Microsoft Azure and Amazon Web Services (AWS) with the latter being the most widely adopted. Major retailers like Walmart and Aldi already use AWS for backend operations. Then, the recommendation model should be deployed to the cloud and integrated within existing systems (e.g., CRM, POS), ensuring a smooth and secure data pipeline. Based on the proposed mockups, the interfaces must be developed and embedded within the relevant customer channels.

The second step will focus on training the staff. Store staff and the sales team should understand, support and promote the use of the recommendation system. This will enhance customer service and cross-selling. Additionally, marketing teams can leverage recommendations in personalized campaigns, product bundling and customer engagement strategies.

In the actual deployment, the API integration will allow front-end applications (website and app) to retrieve and display the recommendations. A pilot phase using A/B testing is advised to evaluate the performance of the final two model candidates in real time. After a set testing period, the two groups with different models should be compared using metrics like: click-through rate (CTR), conversion rate and revenue lift. The model with better results should be rolled out fully. Overall, deployment should be gradual, expanding across all digital platforms and physical stores.

Finally, monitoring & maintenance is an ongoing critical process to keep the system effective and aligned with user behavior and business goals. The model should be retrained daily (best practice: overnight), using the latest customer transactions to adapt to changing preferences and product availability. By regularly monitoring the performance with KPIs (e.g. engagement rate, latency, recommendation quality) and user feedback from customers as well as employees to model should be updated to maintain relevance and enhance user trust. For any optimized or new recommendation models, new A/B-Testing phases should be conducted.

This deployment strategy is only a recommendation. The next steps should be discussed and further refined in collaboration with RECHEIO.

## 7. CONCLUSION

To enrich RECHEIO's dataset and enable personalized recommendations, data exploration and preprocessing were conducted as a first step. Following this, feature engineering was applied to extract meaningful patterns and attributes from the raw data. Multiple clustering models were tested, including K-Means, Self-Organizing Maps (SOMs), and DBSCAN, and evaluated using $R^2$ as the performance metric. The best-performing model was selected, resulting in the identification of six

distinct customer clusters. Each cluster offers valuable insights into customer behavior and transaction patterns, enabling more targeted strategies and deeper customer engagement.

Based on these clusters, several types of recommendation methods were developed: Usual Basket/Last Purchase, Smart Basket, Substitute RECHEIO Products and Single Item Recommendation. For the Smart Basket Recommendation, various approaches were tested using Monte Carlo Cross-Validation, with Precision and Recall. Finally, mockups of possible interface designs were created to demonstrate how these recommendations could be implemented across different customer interaction channels.

## 7.1. BUSINESS IMPLICATIONS

By leveraging recommendation system, RECHEIO can unlock significant business advantages:

- **Increased Sales Through Personalization:** Personalized product recommendations, such as cross-selling and upselling, encourage customers to purchase more, leading to higher average basket sizes and increased overall revenue.
- **Improved Customer Retention and Loyalty**: Tailored shopping experiences foster stronger customer relationships, resulting in more repeat visits and greater long-term loyalty.
- **More Efficient Marketing and Promotions**: Recommendation engines enable targeted marketing by matching the right promotions to the right users, thereby improving the return on investment (ROI) of marketing campaigns.
- **Optimized Inventory and Reduced Waste**: Recommendations influence demand in predictable ways, helping to forecast product turnover and avoid overstocking or understocking, particularly important for perishable goods.

## 7.2. CONSIDERATIONS FOR MODEL IMPROVEMENT

To further improve the performance and robustness of the recommendation continuous model evaluation and iterative refinement are essential. Several key opportunities for optimization have been identified. First, testing more advanced models like more complex hybrid models, machine learning models or deep learning models like hierarchical softmax and multi-task learning (MTL) - can significantly enhance predictive capabilities. Furthermore, incorporating temporal elements such as seasonality, time decay, and/or purchase cycle detection would also allow the system to better reflect customers' real-world buying behavior. Giving additional weight to recent transactions can improve, relevance, particularly for fast-moving or trend-sensitive products.

Another important focus is improving the quality and contextual depth of the data. Integrating factors such as time of day, device type (purchase type: online, store, app) or customer demographics can make recommendations more precise and context aware. This involves also ensuring accurate and up-to-date customer profiles, either through staff training or customer self-service updates. This way, we could assign greater importance to the Client Type, allowing for more tailored recommendations.

Beyond transaction data, the system can be strengthened by incorporating implicit signals such as click behavior, scroll depth, watch time, and add-to-cart actions. These behavioral cues offer valuable insights into customer interest, even when no purchase is made.

## 8. REFERENCES


Jadeja, M. (2020, March 31). *Jaccard similarity*. Medium. https://mayurdhvajsinhjadeja.medium.com/jaccard-similarity-34e2c15fb524

Kashyap, P. (2023, March 17). *Understanding precision, recall, and F1-score metrics*. Medium. https://medium.com/@piyushkashyap045/understanding-precision-recall-and-f1-score-metrics-ea219b908093

Ravindran, A. (2020, August 10). *Cross-validation: K-Fold vs Monte Carlo*. Towards Data Science. https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b/

Rong, Y., Wen, X., & Cheng, H. (2014). A Monte Carlo algorithm for cold start recommendation. *Proceedings of the 8th ACM Conference on Recommender Systems*, 273–280. https://dl.acm.org/doi/pdf/10.1145/2566486.2567978

Yarlagadda, H. (2019, May 26). *Creating a grocery product recommender for Instacart*. Towards Data Science. https://towardsdatascience.com/creating-a-grocery-product-recommender-for-instacart-c1b6bdf5ae13