FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Exploring Fish Purchasing Behaviour using Data Analytics

Rodrigo Teodoro Passos

MSc in Eletrical and Computer Engineering

Supervisor: Prof. Gonçalo Figueira

Co-Supervisor: Prof.ª Vera Miguéis

June 24, 2019

Resumo

Nas últimas décadas têm ocorrido mudanças significativas no setor do retalho. As mudanças no setor do retalho também se aplicam ao setor do peixe fresco, que tem sido alvo do interesse de investigadores internacionais por razões políticas e económicas. Tendo em conta este ambiente competitivo, que valoriza a qualidade e o serviço fornecido ao consumidor assente em custos aceitáveis, é necessário a adoção de estratégias focadas no cliente.

Integrada no projeto ValorMar, esta dissertação realiza um estudo aprofundado do comportamento do consumidor de peixe fresco. Para tal, foram aplicadas técnicas de data mining a dados transacionais da SONAE, no sentido de criar modelos que descrevessem o comportamento do consumidor. Este comportamento é descrito por variáveis críticas, como a frequência e valor gasto, a adesão a promoções, o consumo relativo de diferentes espécies de peixe, o tipo de peixe (aquacultura e selvagem), e mesmo as relações entre os vários produtos do carrinho de compras.

Ter conhecimento das preferências do consumidor permite melhorar a sua experiência de compra, enquanto se otimiza a cadeia de valor como um todo. Nesta análise à cadeia de valor de peixe fresco, foram encontradas diversas oportunidades de criação de valor futuro. Na verdade, as diversas segmentações geradas suportam a ideia de um conjunto diversificado de consumidores de peixe fresco e, por isso, diferentes tipos de ações de marketing podem ser efetuadas. Fazendo um balanço entre os resultados esperados e obtidos, acredita-se que a utilização de dados transacionais permite colocar o foco no cliente e, desta maneira, desenvolver melhores decisões estratégicas, conduzindo a melhores resultados para a cadeia de valor.

A metodologia proposta por esta dissertação tem como objetivo não só a identificação de diferentes grupos de clientes recorrendo a técnicas de **clustering**, mas também a análise do **carrinho de compras** dos um clientes de peixe fresco. O trabalho desenvolvido mostra que a extração de conhecimento de grandes bases de dados permite melhorar as decisões estratégicas das empresas e a sua relação com os clientes.

Os resultados alcançados traduzem não só as diferenças claras entre grupos de clientes de peixe fresco, mas também as suas preferências. Em suma, este trabalho confirma a importância que o data mining tem na compreensão dos clientes e no modo como a análise a grande bases de dados podem ajudar as empresas em decisões estratégicas.

Abstract

In recent decades, there have been significant changes in the retail sector. The changes in the retail sector also apply to the fresh fish sector, which has been of interest to international researchers for political and economic reasons. Given this competitive environment, which values quality and customer service based on acceptable costs, it is necessary to adopt customer-focused strategies.

As part of the ValorMar project, this dissertation carries out an in-depth study of the consumer behaviour of fresh fish. To this end, data mining techniques were applied to transaction data from SONAE, in order to create models that describe consumer behaviour. This behaviour is described by critical variables, such as the frequency and value spent, the adherence to promotions, the relative consumption of different species of fish, the type of fish (farmed and wild), and even the relationships between the various products in the shopping cart.

Being aware of consumer preferences allows you to improve your shopping experience, while optimizing the value chain as a whole. In this analysis of the fresh fish value chain, several opportunities for future value creation were found. In fact, the various segmentations generated support the idea of a diversified set of consumers of fresh fish and, therefore, different types of marketing actions can be carried out. Balancing the expected and obtained results, it is believed that the use of transactional data allows the focus on the customer and, in this way, develop better strategic decisions, leading to better results for the value chain.

The methodology proposed by this dissertation aims not only to identify different groups of customers using **clustering** techniques, but also to analyse the **shopping cart** of a fresh fish customer. The work developed shows that the extraction of knowledge from large databases allows improving the strategic decisions of companies and their relationship with customers.

The results achieved reflect not only the clear differences between groups of fresh fish customers, but also their preferences. In short, this work confirms the importance that data mining has in understanding customers and in how the analysis of large databases can help companies in strategic decisions.

Acknowledgements

Esta dissertação marca o ponto final do meu percurso no Mestrado de Engenharia Eletroténica e Computadores na FEUP. Deste modo, encerro um capítulo da minha vida repleto de histórias, dificuldades ultrapassadas e de momentos inesquecíveis. Da FEUP levarei as memórias de um caminho feito de aprendizagem constante e de uma curiosidade insaciável. Serei eternamente grato à FEUP e a todas as pessoas que dela fazem parte.

Começo por agradecer à SONAE, não só por disponibilizar os dados em estudo, mas também pelo permanente interesse no desenvolvimento deste trabalho.

Um agradecimento especial e merecido aos meus orientadores. O professor Gonçalo Figueira e a professora Vera Miguéis receberam-me de braços abertos neste projeto e foram inexcedíveis em todos os momentos. Os dois mostraram uma disponibilidade total para ajudar e para discutir o melhor caminho para a resolução dos vários desafios que foram surgindo durante a dissertação.

Agradeço também a todos os meus amigos e colegas de curso, pelas horas que passei com eles e pelo papel que desempenharam para tornar este percurso mais fácil. A todos eles, obrigado.

Naturalmente, nada disto seria possível sem o apoio permanente que recebi da minha família. Do fundo do coração, agradeço aos meus avós, aos meus pais Paula e Victor e ao meu irmão Ricardo.

Por fim, agradeço à Inês. Pelos conselhos, pela paciência e porque sempre acreditou em mim.

Rodrigo Passos

"Where there is data smoke,
there is business fire"

Thomas Redman

Contents

List of Figures

List of Tables

Abreviaturas e Símbolos

EU	European Union
SKU	Stock Keeping Unit
KDD	Knowledge Discovery in Databases
DM	Data Mining
kg	Kilograms
MBA	Market-basket analysis
NI	Not Identified
RFM	Recency, Frequency, Monetary Value

Chapter 1

Introduction

1.1 General Context

In the last decades there have been significant changes in the retail sector resulting from globalization, the increased competitiveness and transformation on consumer's purchasing behaviour. Nowadays, the variety of choice and quick access to information allow consumers to make more informed decisions, and increase their sensitivity to multiple factors, such as quality and price. Economic and social factors, as well as increased competition and consumer's requirements, force companies to add value and enhance their processes and products to sustain and improve its market position.

Taking this competitive environment into account, which values the quality and the service provided to the customer at acceptable costs, it is necessary to adopt strategies focused on the customer. This is the main reason why retailers want to know their customers: who they are, what they buy, when they buy, why they buy. These questions can be answered based on transactional information companies collect. Those answers are able to understand consumer behaviours and consequently estimate demand. These estimates costs resulting from excessive inventory levels and reduces stockouts, leading to better service levels. In the end, the company is able to improve the overall value chain operations, making the company more competitive and able to lead the market.

The relevance of knowing customers also applies to the fish sector. Over the last years, consumer purchasing behaviour towards fish and seafood products has been capturing the interest of researchers internationally for political and economic reasons. These aspects are related to nutrition and diet, food safety, sustainability and business of the fish industry. As a matter of fact, this sector, in addition to all the above issues, still had to overcome other barriers. The food crisis in the 90's [5], the introduction of rules created by the European Union at the beginning of 2000 and the increased consumer concern to maintain a healthy and environmentally sustainable lifestyle are other difficulties that the fish sector has faced.

Fresh fish is one of the portuguese identities to the world and represents an important sector of the national economy. According to the Associação Fórum Empresarial da Economia do Mar

(AFEM), the economic value of the activities related to the sea considered in the Portuguese economy is currently about 2% of the national GDP and directly employs about 75 thousand people. Considering a broader vision, between direct and indirect effects, the total value should be around 5% to 6% of the GDP and give work to more than 100 thousand people [6].

Portugal is the largest consumer of fish per inhabitant (about 57 $kg/year/inhabitant$) at the European level and the third largest in the world [7]. Moreover, with an Exclusive Economic Zone exceeding $1,700,000\ km^2$ and 942 km coast line, the fishery and the consumption of fish products are of an extreme importance. With a sea coast abundant in fish, it is easy to understand that fish has quickly become a primordial element for food, and consequently a factor of population fixation. In fact, this sector has gained significant importance by the amount of people dependent on this activity in Portugal. In addition to the jobs in the sea, it is important to note the various employees on land such as trade, shipbuilding / repair, transport, administration and research. Therefore, a deeper knowledge of the preferences and patterns of seafood consumption by the Portuguese population as well as an assessment of the underlying dynamics are required.

This thesis is integrated in ValorMar project, which was born from the commitment of a broad spectrum of entities, from companies to research centers, which recognize the relevance of the sea economy. This mobilizing project aims the valorization and efficient use of natural endogenous resources and the development of value chains associated to marine resources. ValorMar's main objective is to develop a technological platform to support the traceability information of the fish and its availability to the final consumer, in-store or outside, in an integrated way, considering the fish value chain. The value chain of the company is linked to the value chains of suppliers, upstream, and the buyers, downstream. Both are subject of analysis within this project, although this thesis focuses on the latter. In the first case, the conditions of fishing, transport and storage of the fish are analyzed. In the second case, we analyze not only the characteristics associated with each product, such as the type, origin, production method, base price, but also demographic data of the consumer. Finally, within the framework of this project, this thesis will try to understand different purchasing patterns regarding the purchase of fresh fish. In order to understand consumer behavior, secondary data (transaction-related values) will be used.

This dissertation results are beneficial for the company for several reasons. By knowing the consumer's behaviour, the company will not only reduce waste, but also increase its offer and provide a better service level to customers, since it and will be able to understand their preferences. Furthermore, due to the relationship of trust and loyalty between the company and the consumer is established regarding the quality and cost of the product. In fact, having the customer as the focus of the company concerns, their aim is to ensure the ultimate satisfaction of the consumer.

1.2 Objectives and methodology

This section summarizes the specific research objectives. These objectives are linked to the chapters of the thesis and consequently, for each objective, the corresponding thesis chapter is indicated.

As mentioned previously, companies know that focusing on customers is the key to achieve the desired success. In this sense, companies seek to understand the market and, from there, create solutions capable of satisfying the customers. In the scope of this dissertation, one of the largest food retailers in Europe is used as a case study. This company stands out for its 245 stores throughout mainland Portugal. These stores differ by size and the variety of products offered. This company achieved great success with the introduction of the loyalty card in the market. This card provides multiple discounts to customers and information about the customers to the company.

This dissertation seeks to present descriptive models that faithfully characterize the behavior and shopping habits of fresh fish customers. For this purpose, segmentation and analysis of the shopping cart are performed, allowing a better understanding of the needs and expectations of customers. To make this study possible, the transactions that customers made with the loyalty card during one-year period were used. This consumer behavior study will be done using software such as Microsoft SQL for extraction of transactional data and R for data processing and application of data mining (DM) techniques.

In a first phase, the various criteria that will allow to understand the behavior of the consumer will be defined and, from these, groups of fresh fish customers with different behavioral patterns will be identified. Having obtained the customers' segments, these clusters will be characterized based on the variables used in their training and also on demographic and socio-economic characteristics of the customers. After the global analysis of customer behavior, the behavior of the customers in the different types of stores (from now on called Insignias) will also be studied. With this study, it will be possible to see if there are different behaviors and/or different types of customers in the various Insiginias of the company. This procedure allows companies to structure the knowledge concerning the customers, and to define the target market for specific marketing actions.

The research also aims to explore several forms of segmentation. The methodology proposed in this thesis aims to construct a segmentation model based on the relative consumption of the species and a segmentation model based on wild/farmed fish. This methodology intends to infer the preferred fresh fish of each consumer and, based on that, to attract customers with individualized marketing campaigns.

Lastly, this thesis aims to use market basket analysis to discover frequent associations between products. Through this methodology it will be possible, for example, to support the design of differentiated marketing actions to encourage customers from different segments to visit the stores more often and spend more on the visits to the stores . This involves promoting the purchase of different kinds of products, that are still not purchased by the customers despite their potential interest.

1.3 Thesis Outline

This thesis includes six chapters, which are summarized below.

Chapter 2 intends to provide theoretical background on the various topics and subjects addressed in this work. The characteristics and factors related to consumer's behaviour are enumerated and explained.

Chapter 3 presents a literature review of DM and the benefits this concept brings to the retail industry. In addition, it introduces the main techniques of DM that can be used in the context of the dissertation.

Chapter 4 describes the retail company used as a case study. It also contains a description of the company position in the market, store formats, organizational structure and products classification. This chapter also includes a characterization of the company's customers. This chapter also includes a characterization of the company's customers and transactions.

In chapter 5, the analysis of consumer behavior is performed. A segmentation based on consumer behavior is carried out and the obtained clusters are characterized using the available information about the customers. This chapter also proposes a model to characterize the customers' profile within each segment, by means of a decision tree. Other segmentation models are also developed. One is based on species consumption and the another is based on wild/farmed fish consumption. Finally, there is a model to identify product associations using a *priori* algorithm.

Chapter 6 presents the summary and conclusions of the research developed in this thesis. It also presents some directions for future research.

Chapter 2

Theoretical Background

2.1 Consumer Behaviour

In developed countries, consumers of food products are presented with a wide array of decisions in everyday life. To make the right choice, consumers use all the information they have available [8]. To understand consumer behavior, it is necessary to identify the factors that lead the consumer to make the right decisions. The identification and understanding of these factors are the first steps to understand the consumer's profile. Indeed, what determines these factors is the perception that the consumer acquires on them. According to several studies, the perception of price, quality and value are considered key factors in the behavior of the purchase and the choice of the product [9][10][11].

Throughout this chapter a theoretical and detailed analysis will be made on these factors and the way they influence the purchase of a product and the satisfaction of a consumer. In addition, a particular analysis will also be developed for the fish consumers and their behavior.

2.1.1 Quality Sensitivity

Perceived quality is defined as the consumer's judgment about the superiority or excellence of a product [12][13]. Due to the effect of globalization, the distance that a product travels from the producer to the consumer has increased. For this reason, maintaining products with safety and quality has become a major challenge.

The perception of quality has gained importance, because, over the last few decades, the credibility of the food industry has been called into question by numerous food crisis. The most serious health hazards appear, mostly, in contamined food products, such as methanol in wine, salmonella in eggs, lead in milk powder, benzene in mineral water, dioxins in chickens (bird flu), nitrofurans in the poultry sector and the illegal use of drugs, antibiotics and hormones in beef (mad cow disease). As a result of this crisis, consumers began to demand high quality in food with integrity, safety guarantees and transparency [14]. Thus, the idea of traceability – "the ability to trace the history, application or location of that which under consideration"[15] – has gained increasing preponderance in order to maintain the quality and food safety of a product.

With the aim of increasing the robustness of the rules on food and feed safety in the European Union, on 28 January 2002, Regulation CE nº178/2002 [16] was created. This document establishes the general principles and standards of food law and lays down procedures in matters of food safety. The legislation introduced and defined the concept of traceability as "the ability to trace and follow a food, feed, food-producing animal or substance intended to be, or expected to be incorporated into a food or feed, through all stages of production, processing and distribution" [16]. As a consequence of these regulations, various labelling schemes from producers and distributors are now put in practice for fish products. These aim to promote resource sustainability, distinction of quality and product safety [17].

In fisheries products, traceability is essential because it is possible to verify the origin, freshness and shelf life. These aspects are fundamental when deciding to purchase this type of product [18]. In addition, the importance of traceability increases as consumers are increasingly concerned about health, quality of life and longevity [19] and fish consumption is increasingly associated with such issues, as well as improvements in people lifestyles [20].

Consumers increasingly value quality, safety and environment friendly products, as well as having a transparent traceability and market [18]. In fact, the application of traceability to fish has many advantages both downstream in the product and upstream in the consumer. This allows a product to follow-up along the process from the production line as raw material to the final product to the consumer [18]. A good traceability system has the potential to reduce risks and costs associated with foodborne disease outbreaks [21]. For example, it could reduce their magnitude and possible health impact and consequently reduce or avoid medical costs [22]. Reducing health risks caused by products sold allows to credibilize a company and to improve its reputation and competitiveness in the market. Therefore, traceability has the ability to attract customers, due to the trust it establishes with the product and with the company. Other studies on the impact of traceability indicate that perceived safety, quality, high nutritional and availability value lead to a higher price [23] and that is worth it [19].

Finally, it is possible to affirm that the application of food traceability offers an increase in public health protection, improves consumer confidence in companies, allows a more transparent trade and a better sharing of responsibility along the supply chain [24]. In short, quality is one of the most decisive factors when choosing a product.

2.1.2 Price Sensitivity

Another of the most important criteria when deciding to buy a product is price [25][19]. Price is an important market stimulus and consumer consciousness and perception of price deserves attention. From the consumer's point of view, the concept of price is defined as the sacrifice to obtain a product [26][27][28] [29]. In fact, the scientific community states that there is a clear distinction between objective price (the actual price of a product) and perceived price (the price as encoded by consumer) [30]. Several studies indicate that consumers do not always remember the exact price associated with the product, but in a way that means something to them [11]. For

example, some consumers may note that the real price of a Coke is €1,20 and others can only associate and remember the price as being "cheap" or "expensive".

After defining the concept of price and its importance, attention must be given to how small changes in price are interpreted and perceived by the consumer. In a developed research, it is possible to verify that the consumer's perception modifies according to the magnitude of the change of the price [31]. Beyond this, a lot of the product variation in perception of price changes can be accounted for by the importance of the product in the consumers' budget, the price stability of the product, and the frequency of purchase. The sensitivity that a consumer has with price varies depending on the value given to a product.

2.1.3 Value Sensitivity

The idea of value has different concepts associated with it, since what constitutes value is highly subjective and idiosyncratic. According to a developed study by Uhl and Brown [31], it was possible to create four definitions of value for the consumer.

1. *Value is to have low price*: This concept of value is used when a product is on promotion;
2. *Value is having what you want in a product*: This idea of value is directly related to the concept of utility, that is, the satisfaction we get from the use of a product.
3. *Value is having quality to the price that was paid*: This definition is applied when a trade-off between product, price and quality is made. Another study reinforces this premise by explaining that among many other perceptions, it is essential that consumers have the feeling that they paid a fair price for the quality offered [32].
4. *Value is what is received for what is given*: value as a ratio of attributes weighted by their evaluations, divided by price weighted by its evaluation [33].

Within the scope of this dissertation, the value of fresh fish may be related to the high nutritional value and consequent health benefits.

2.2 Fish Consumer Behaviour

The fish industry is currently an international business with a global production estimated at US$ 232 billion in 2012 [34] . The fish industry and all the extra activities that surround it provide livelihoods and income to hundreds of millions of people. Fish and seafood are widely accepted to be an essential component of a balanced and healthy diet because they have a low fat content and provide high quality proteins as well as many micronutrients such as vitamins and minerals [35]. Therefore, since public health authorities are interested in promoting fish and seafood consumption in order to improve public health, it is important to learn which are the main factors influencing consumers' behaviour towards these food products. Indeed, world per capita fish consumption has increased from an average of 9.9 kg (live weight equivalent), in the 1960s, to 19.2 kg, in 2012

[34]. This trend is having a relevant negative ecological impact because the increasing fishing pressure (overfishing) is leading to an important decline of natural fish resources and becoming unsustainable for several species [36].

The reasons that have been discussed so far have led to a highlight that in a demand-driven market a better understanding of consumer purchasing behaviour towards fish products is paramount to developing more effective marketing and policy strategies. As already mentioned, there are several factors that influence consumer decision-making. In the analysis of fish consumer literature, the following factors were identified:

- Sensory perception: sensory characteristics of fish such as taste, smell and texture are expected to be key determinants of fish consumption and they are also extremely important to evaluate freshness.

- Health beliefs: concerning health benefits, several studies demonstrate that fish and seafood are widely perceived by consumers as healthy food with a number of specific health and nutritional benefits mainly associated with the high content in proteins and Omega-3 fatty acids together with a low fat content [37] [38] [39] [40] [41] [42][43] [44] [45].

- Convenience perception: similarly to other food products, fish consumption is expected to be influenced by consumers' needs for convenience, i.e. the desire to save time and effort in food preparation. Regarding to this factor, consumer behaviour should be different towards fresh and processed fish products. The former is expected to be perceived as difficult to prepare, while the last could be perceived as a quick and easy meal option [36].

- Country of origin: Several studies highlight that country of origin is one of the most important fish attributes of consumers' choice [39] [46] [42] [47] [43] [48]. Moreover, these studies show convergent patterns towards a clear preference for domestic fish products which are perceived as being superior to imported fish in terms of quality, safety and freshness.

- Production method (wild vs farmed): several studies carried out in different countries [49][42] [50] [47] [48] [51] [52] [39] [53] show that wild fish is perceived as being superior to farmed fish by the majority of consumers in terms of taste, safety, healthiness and nutritional value. This perception seems to be accentuated in older consumers with more traditional eating habits and in people living in coastal areas where stronger wild fish consumption habits and better availability of caught fish in terms of variety and freshness are typical [50].

- Fish availability: fish consumption may be also strongly affected by the availability of fish assortment. This happens because the preferred fish products are not available, the available alternative fish products may appear to be weak substitutes and thus consumers may decide not to buy any fish products.

In particular, this review identifies and discusses the main barriers of fish consumption as well as consumers' preferences about the most relevant attributes of fish and seafood products.

To understand which factors have the greatest impact on consumer behaviour of fresh fish, DM techniques that recognize consumption patterns can be used.

2.3 Data Mining for Consumer Behaviour

Retail is one of the main economic axes in the national and world economy. This sector stands out for the great competitiveness between companies. The quality (2.1.1), price (2.1.2) and the value (2.1.3) of a product are the distinguishing factors among the various companies. For this reason, detecting similarities and differences among customers, predicting their behaviours, proposing better options and opportunities to customers became very important for customer-company engagement. The purpose of customer profiling is to target valued customers for special treatment based on their anticipated future profitability to the stores.

To help solve these challenges, companies have begun to turn to Knowledge Discovery Databases. The decision to use retail-oriented DM techniques is a smart strategy, because of the amount of transactional data each business has yet to explore. DM techniques are suitable for profiling the customers due to their proven ability to recognize and track patterns within the set of data [54]. Data for knowledge extraction can be collected in different ways. A simple way to collect data is a questionnaire. However, most of the important data in store data can be collected from each transaction of goods that are recorded in databases. The data collected in a database can be sales, demographic data or items information. Each customer is identified by his identity in a loyalty card. The method of data collection would handle all necessary time levels such as years, seasons, and months, simultaneously, and can extract both independent time level patterns and interrelationship patterns among the time levels used.

Briefly, there are several issues in the retail sector that can be raised and resolved using DM techniques. Retail companies that invest in this knowledge source will be able to gain a competitive advantage over competitors and provide a better product and service to the consumer. However, the use of DM in retail is still incipient and most companies still use mass strategies to instigate customers loyalty.

Since the research reported in this thesis focuses on a descriptive analysis and the understanding of consumer behavior, this involved a study of DM techniques. This study aimed at exploring the techniques and getting insight into the advantages and disadvantages of each one, in order to develop models that could meet the objectives of the company. Nowadays, the identification of customer groups with similar behaviour patterns is often done in ad-hoc way.

2.4 Summary

As discussed, the consumer has decisive factors that lead him to make decisions when buying a product. Combining these three variables – quality, price and value – is essential when a retail company wants to understand and predict the behavior of a consumer in its stores. In the scope of this dissertation, we should understand the impact that each variable has on the fresh fish customer.

The variable quality may have an impact on the customer's preference for fresh fish for certain species or their origin (wild or farmed). The sensitivity to the variable price may vary depending on the age and/or social condition of the client. The variable value may have an influence on the quantity of fish purchased in relation to other products.

In the context of the dissertation, the analysis to the transactional data can allow a better understanding of the practical effect of these factors on consumer behavior. For a faster and more efficient analysis DM techniques will be used.

Chapter 3

Introduction to Data Mining

The application of DM has been gaining importance in recent years, so this chapter contextualizes and reviews the concept. As this thesis requires the use of various DM techniques, this chapter aims to summarize some essential techniques in the retail context.

3.1 Context

In the current competitive market, successful business organization need to be able to react rapidly to the changing market demands both locally and globally. In that sense, having greater and faster access to sources of knowledge about what might occur in the future is a source of inestimable value. This is possible by using DM techniques that can extract potentially useful information from vast resources of raw data.

The development of the technology in the last decades allowed to increase the capacity of storage and analysis of large amounts of data. Actually, it is estimated that every 20 months the amount of data stored in all databases in the world tends to double [55]. The ability to access large volumes of data enables the creation of knowledge and value for a company in areas such as decision support, prediction, forecasting and estimation. This will help companies to make important business decisions, which can give a particular business the competitive edge. However, transforming data into useful knowledge is a slow and complex process - described as "data rich but information poor situation" [56]. For this reason, it has become imperative to create a process with well-defined phases and oriented to the production of knowledge - Knowledge Discovery in Databases.

Knowledge Discovery in Databases (KDD) is a process that seeks to establish links, relationships and patterns through available data. It includes a set of phases, ranging from the preparation of the data to the validation of the results obtained. KDD uses tools from a variety of areas of expertise such as statistics, artificial intelligence, machine learning, information theory and computing. As previously mentioned and can be observed in Figure 3.1, the KDD has several phases such as data selection, data preprocessing, data transformation, DM and interpretation/ evaluation of the obtained results.

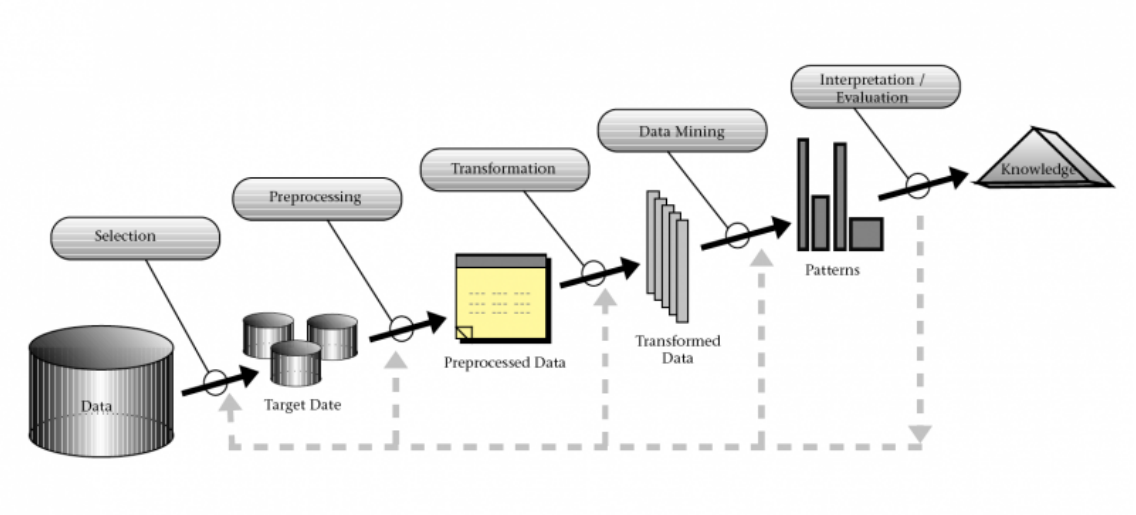

Figure 3.1: Overview of the Steps That Compose the KDD Process [1].

All KDD phases and their description are presented below:

1. Data Selection: At this stage the objectives are to understand the domain of the application and identify the goal of the KDD process. In addition, the dataset for which the analysis is to be made will be chosen.

2. Data Preprocessing: this phase consists of simple tasks, such as: remove the noise, collect the necessary information to model or account for noise and decide on strategies for handling missing data fields;

3. Data Transformation : where the data is transformed or consolidated into forms appropriate for mining algorithms (data normalization, aggregation or discretization);

4. Data Mining: an essential process where intelligent methods are applied in order to extract data patterns or relationships.

5. Interpretation/Evaluation: last stage of the process, where it should be identified interesting patterns or relationships. These should be useful and add value.

The extraction of hidden predictive information from large databases is a powerful tool with great potential to help organizations to define the information market needs of tomorrow. DM tools predict future trends and behaviors, allowing businesses to make knowledge-driven decisions that will affect the company, both short and long term. The automated prospective analysis offered by DM tools of today is much more effective than the analysis provided by tools in the past. DM answers business questions that traditionally were too time-consuming to resolve. DM tools search databases for hidden patterns, finding predictive information that experts may miss because it was outside their expectations.

The DM techniques can also be categorized in supervised learning and unsupervised learning. In *supervised learning*, the user provides the algorithm with pairs of inputs and desired outputs,

and the algorithm finds a way to produce the desired output given an input. In particular, the algorithm is able to create an output for an input it has never seen before without any help from a human. Supervised learning builds a model that makes predictions based on evidence in the presence of uncertainty. A supervised learning algorithm takes a known set of input data and known responses to the data (output) and trains a model to generate reasonable predictions for the response to new data. Supervised learning is used whenever we want to predict a certain outcome from a given input, and we have examples of input/output pairs. Supervised learning uses classification and regression techniques to develop predictive models. An easy way to distinguish between classification and regression tasks is to ask whether there is some kind of continuity in the output. The main difference between them is that the output variable in regression is numerical (or continuous) while for classification is categorical (or discrete).

In *unsupervised learning*, only the input data is known, and unknown output data is given to the algorithm. While there are many successful applications of these methods, they are usually harder to understand and evaluate. The next sections will explore more deeply methods from those DM categories, with the main focus being the clustering techniques that are necessary for a consumer segmentation approach. Unsupervised learning subsumes all kinds of machine learning where there is no known output, no teacher to instruct the learning algorithm. In unsupervised learning, the learning algorithm is just shown the input data and asked to extract knowledge from this data. Unsupervised learning finds hidden patterns or intrinsic structures in data. It is used to draw inferences from datasets consisting of input data without labeled responses.

Since the research reported in this thesis focuses on customers identification, attraction, development and retention, this involved a study of clustering, classification and association DM techniques. This study aimed at exploring the techniques and getting insight into the advantages and disadvantages of each one, in order to develop models that could meet the objectives of the company. The next sections explore more deeply clustering, classification and association DM techniques.

3.2 Clustering

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters [3]. Clustering is the most common unsupervised learning technique. Cluster analysis techniques have been quite developed in an attempt to enable the analysis of ever larger datasets. The increasing need to turn large amounts of data into useful information has made this type of analysis very common in its exploitation. It is used for exploratory data analysis to find hidden patterns or groupings in data. By clustering, one can identify dense and sparse regions and, therefore, discover overall distribution patterns and interesting correlations among data attributes. The application areas of cluster analysis are vast, including biology, geography, document classification, image processing, market research, among many others.

In this work, cluster analysis is used to determine groups of customers in the fresh fish retail sector according to the transactions they make with loyalty cards. This type of analysis is useful for marketing purposes since it allows to identify the group to which a customer belongs and consequently its profile of segmentation, allowing to offer a more personalized treatment to meet their interests. Market segmentation is, in fact, an area where cluster analysis is widely applied [57].

In general, we intend to divide the set of objects into groups. These objects are characterized by variables whose nature (quantitative or qualitative) will play a key role in the choice of these measures.

In the analysis of clusters it is necessary to make decisions and take into account aspects that depend on each particular problem. However, it is possible to indicate a sequence of steps which are requirements in most analyzes. The various steps are aimed to answer the questions that are usually posed during the analysis, of the following stand out [58]:

- What objects do we intend to group?
- What variables should we consider to characterize these objects?
- Are there missing values, or erroneous values that can be fixed? How to integrate information on data collected from different sources? Are all variables relevant to the analysis? Should they be transformed in any way? In summary, what kind of preprocessing should be applied to the data before grouping it?
- From the long list of methods of cluster formation, which one should I choose in view of my particular problem?
- What distinguishes one cluster from the other? What is the clearest and most succinct way of summarizing my results, and how do I validate them?

In this section we will briefly discuss the possible and most common answers to each of these questions.

Common algorithms to perform clustering includes *k-means*, *k-medoids*, hierarchical clustering, Gaussian mixture models, hidden Markov models, self-organizing maps, fuzzy c-means clustering, and subtractive clustering. There is a much greater variety of algorithms with the emergence of new approaches. In the scope of this dissertation, only two of the best known algorithms will be presented in detail, representatives of **hierarchical algorithm** categories and **partition algorithms** based on mean squared error.

3.2.1 Data Preprocessing

Before applying the method for analyzing the data, it is necessary to work them in such a way that the method of exploitation can reveal interesting structures that exist in the sample. This set of procedures are called data preprocessing. During data preprocessing several procedures can be

performed [56]. Some aspects to be taken into account at this stage of the exploratory data analysis are listed below:

1. **Data Cleaning:** it can be done to remove noise or correct inconsistencies. Through data cleansing procedures, we try to complete the missing values, attenuate the noise by detecting potential outliers and correct inconsistencies in the data.

2. **Data Integration:** data from various sources are integrated into a coherent set. There are some issues to consider when combining data, for example matching objects from multiple sources or the perception that two variables with different names actually coincide. It is important to use the available information about each variable, such as name, domain, type or meaning, to avoid errors at this stage.

3. **Data Transformation:** Variable transformations are performed, for example, standardization, which can improve the performance of the algorithms involved or simply reveal patterns of interest in the data that are not visible from the original space, where the study variables were defined. When scaling variables, the data can be transformed as follow:

$$\frac{x_i - mean(x)}{sd(x)} \tag{3.1}$$

Where $mean(X)$ is the mean of x values, and $sd(x)$ is the standard deviation (SD).

3.2.2 Methods of cluster formation

Once the dissimilarity between objects is defined, it is necessary to choose the method of clusters to be used, from a wide variety of existing proposals. The choice of clusters depends on the type of data available and the purpose of the analysis to be performed. Often there are several candidate methods to apply to the data set under study and there are, *apriori*, insufficient arguments to restrict the choice of the most appropriate method [59].

The most discussed distinction among different types of clustering is if a set of clusters is hierarchical or partitional. A **partitional clustering** is a division of the set of data object into clusters so that each data object is in exactly one cluster. If clusters are allowed to have subclusters, then we have a **hierarchical clustering**, which is a set of nested clusters that are organized as a tree.

Although there is a wide variety of clustering methods and algorithms in the literature, this paper only refers to partitioning methods and hierarchical methods. Hierarchical methods will be referred to in general terms, as their application becomes impractical for large datasets and, for that reason, it does not make sense to be used in this problem.

In contrast to hierarchical clustering methods, **partitional clustering** aims successive clusters using some iterative processes. Partitional clustering assigns a set of data points into k-clusters by using iterative processes. In these processes, n data are classified into k-clusters. The predefined

criterion function E assigns the data into k^{th} number set according to the maximization and minimization calculation in k sets. The most popular and commonly used partitioning methods are *k-means* and *k-medoids*.

k-Means

The *k-means* is by far the most popular clustering algorithm. Originally known as Forgy's method [60], *k-means* has been used in various fields including data mining, statistical data analysis and other business applications. The *k-means* algorithm was later developed by MacQueen and it was suggested the term *k-means* for describing an algorithm that assigns each item to the cluster with the nearest centroid (mean) [61].

k-means uses the Euclidean distance measure and iteratively assigns each record in the derived clusters. This algorithm requires the definition of the initial seeds (initial items defined as the clusters mean) in the first iteration of the algorithm. After classifying a new item, it is calculated a new mean for the corresponding cluster and the process continues. This algorithm involves several iterations which differ concerning the initial seeds. The process is finished when the partitioning criterion function, usually the square-error, converges to a value close to the minimum.

$$E(C) = \sum_{j=1} \sum_{x_i \in C_k} (x_i - C_j)^2 \tag{3.2}$$

Based on the concepts above, the computing process for *k-means* is presented as follows:

Algorithm 1 *k-means* algorithm

1. select K points as initial centroids
2. **repeat**;
3. (re)assing each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
4. update the cluster means, that is, calculate the mean value of the objects for each cluster;
5. **until** centroids do not change

The *k-means* procedure is graphically illustrated in Figure 3.2. For a better interpretation the graph refers to three input clustering fields and a two-dimensional clustering solution.

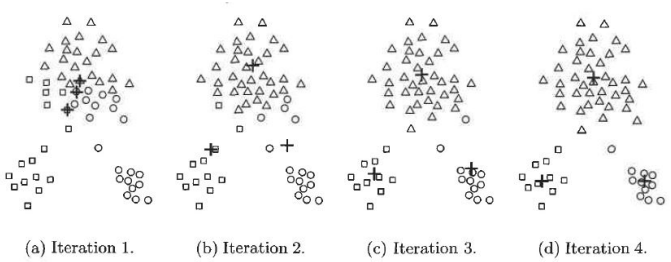

Figure 3.2: Using the *k-means* algorithm to find three clusters in the sample data [2].

k-means advantages include its speed and scalability: it is one of the fastest clustering models and it can efficiently handle long and wide datasets with many records and many input clustering fields. *k-means* has the disadvantages of requiring the prior specification of the number of groups (common to all partitioning techniques), and depending heavily on the initial seeds. Selecting different initial seeds may generate differences in clustering results, especially when the target dataset contains many outliers.

k-Medoids

The *k-medoids* algorithm was introduced by Kaufman and Rousseeuw (1990) and it is very similar to *k-means* algorithm. The basic strategy of *k-medoids* clustering algorithms is to find k cluster in n objects by first arbitrarily finding a representative object (the medoid) for each cluster.

Unlike to *k-means*, instead of taking the mean value of the items in a cluster as the reference point, this algorithm considers the medoid as the most representative item. The method used to choose the medoid may vary, but the most standard procedure is to pick the item which has the lowest average distance to all other items. Since this is an operation computationally demanding, sometimes the search for the medoid is made only on a sample of items. The advantage of *k-medoids* over *k-means* is that [62].

Algorithm 2 *k-medoids* algorithm

1. select k points as initial medoids
2. **repeat**;
3. Create k clusters, assigning each object to the nearest mediatoid;
4. For each medoid m_i, $i = 1,\ldots k$ and
5. For each of the remaining objects x_i, $j \in \{\ 1,2,\ldots, i-1, i+1,\ldots,n\ \}$
6. Replace m_i with x_j and calculate the total cost of the configuration;
7. Select the lowest cost configuration
8. **until** medoids do not change

Comparison between *k-means* and *k-medoids*

One of the most important characteristics of clusters methods is to deal with large data sets, composed of objects described by variables of various types [63]. However, classical algorithms can either handle large datasets efficiently but are restricted to quantitative variables (such as the *k-means* method), or they can handle many types of variables but they are not very efficient when the data set is large (as is the case with the *k-medoids* method).

The similarities and differences between the two methods are following summarized. These two methods have several similarities:

- They are partitioning methods and assign each object to a single cluster;
- They establish clusters based on prototypes (centroids for *k-means* and medoids for *k-medoids*).
- They have difficulty dealing with non-spherical clusters with different sizes. They also perform poorly when clusters have very different dispersions.
- They consider all variables not taking into account the existence of possible clusters involving only a subset of the variables.
- They produce different sets of clusters after each execution of the algorithm when the initialization of the centroids or medoids is made of random form.
- They need to specify *a priori* the number of clusters.

Despite the many similarities, the two methods also have some differences:

- The *k-means* method assigns all objects to some cluster. On other hand, *k-medoids* considers some objects to be noise and therefore not belonging to any of the clusters.
- The *k-means* method is restricted to data for which the notion of centroid makes sense while the *k-medoids* algorithm allows a greater variability in the data type, since it only requires that there is a measure of dissimilarity between each pair of objects.
- The *k-means* method is more sensitive to outliers, which can distort the data distribution. This effect is particularly aggravated because the objective function is based on the mean square error. Thus, the *k-medoids* algorithm is more robust since its objective function is based on absolute error.
- The *k-means* method has a computational complexity of $O(n)$ while the computational complexity of the *k-medoids* method is $O(n^2)$.

3.2.3 Validation

The validation of the obtained results is one of the most important questions of the analysis of clusters and the most difficult steps of the exploratory analysis, being therefore often neglected by the analysts. Through the evaluation of the results, evidence is sought that the partition obtained effectively captures the actual structure of the data. In general terms, there are three approaches to cluster validation, based on the following criteria [58]::

- **Internal criteria**: which uses the internal information of the clustering process to evaluate the goodness of a clustering structure without reference to external information. It can be also used for estimating the number of clusters and the appropriate clustering algorithm without any external data.
- **External criteria**: which consists in comparing the results of a cluster analysis to an externally known result, such as externally provided class labels. It measures the extent to which cluster labels match externally supplied class labels. Since we know the "true" cluster number in advance, this approach is mainly used for selecting the right clustering algorithm for a specific data set.
- **Relative criteria**: which evaluates the clustering structure by varying different parameter values for the same algorithm (e.g.,– varying the number of clusters k). It's generally used to determine the optimal number of clusters.

In order to support the choice of the number of clusters, there are different metrics that aim to evaluate the quality of the clustering result (see Tibshirani et al., 2001, for a revision). The theoretical considerations have led to the development of computer tools for the practical implementation of the proposed ways to study stability. In this thesis, three practical tools of package of the program R were used. These are part of the **external criteria** and are described below.

Package clv

The concept of stability by Ben-Hur and Guyon (2003) is based on the finding that if the clustering properly represents the structure in the data, it should be stable with respect to small changes in the data set. They proposed two measures of stability: a measure based on the index of similarity between two partitions (implemented by the *cls.stab.sim.ind* function) and a measure based on the pattern-wise agreement concept (the *cls.stab.opt.assign* function).

The algorithm of ***cls.stab.sim.ind*** function can be described in the following steps:

1. Cluster the original data set in order to obtain the reference partition
2. Select a random sub-sample of observations from the original data set and group the objects from this sub-sample.
3. Calculate the stability between the reference partition and the partition of the sub-sample using the index of similarity between the two partitions (e.g.: the Rand index).

4. Repeat the procedure several times

5. Repeat the procedure for different values of k (the number of groups).

The ***cls.stab.opt.assign*** function is based on the idea of pattern-wise agreement and pattern-wise stability. Given two groupings L_1 and L_2, the pattern-wise agreement can be defined as follows:

$$\delta_\sigma(i) = \begin{cases} 1, \sigma(L_1(i)) = L_2(i) \\ 0, \sigma(L_1(i)) \neq L_2(i) \end{cases} \tag{3.3}$$

where σ: $\{1,...,k_1\} \Rightarrow \{1,...,k_2\}$

Pattern-wise stability is defined as the fraction of sub-sampled partitions where the sub-sampled labelling of pattern i agrees with that of the reference labelling, by averaging the pattern-wise agreement:

$$n(i) = \frac{1}{N(i)} \sum \delta_\sigma(i) \tag{3.4}$$

where N_i – the number of sub-samples where pattern i appears.

The stability of group j in the reference partition is the average of pattern-wise stability:

$$c(j) = \frac{1}{L_1 = j} \sum_{i \in (L_1 = j)} n(i) \tag{3.5}$$

The stability of the reference partition into k groups is defined as:

$$S_k = \min_j c(j) \tag{3.6}$$

Package fpc

The package fpc includes two functions for measuring stability: *clusterboot* and ***nselectboot***. In this thesis, only the ***nselectboot*** function was used. The ***nselectboot*** function is based on the work of Fang and Wang (2012). The authors focus on the concept of stability as robustness to randomness present in the sample. Drawing on the work of Wang (2010), they formulate the concept of stability in the following way: if one draws samples from the population and applies a selected clustering algorithm, the results of grouping should not be very different.

Briefly, the ***nselectboot*** function is based on the following general idea: several times two bootstrap samples are drawn from the data and the number of clusters is chosen by optimising an instability estimation from these pairs.

Finally, the stability criterion is becoming an increasingly popular method for the selection of parameters of clustering methods, especially for determining the number of groups k. If the taxonomy method is selected correctly and the parameters of this method are also selected correctly

(e.g.: the number of groups, the distance metric), then clustering should provide results that are not very different from each other, i.e. the results should be stable. The methods presented in this thesis are just some proposed ways for measurement of stability, but not the only ones that can be found in the literature. There are other methods proposed which can be found, for example, in the works of: Granichin et al. (2015), Hosein et al. (2011), Koepke, Clarke (2013) and Ryazanov (2016).

3.3 Association

This section presents a methodology known as association analysis, which is a DM technique that has received great attention from researchers. Association rule mining searches for interesting relationships or correlation relationships among items in a given dataset. A common example of the application of this technique is the discovery of patterns, namely of products that are bought together.

Based on the concept of strong rules, association rules were introduced by Agrawal et al. to discover regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets. For example, the famous rule $\{\text{diapers}\} \Rightarrow \{\text{beer}\}$ found in the sales data of a supermarket would indicate that if a customer buys *diapers*, they are likely to also buy *beer*. This is the standard application for association techniques and is called market-basket analysis (MBA) [64]. This kind of analysis is valuable for direct marketing, sales promotions, and for discovering business trends. MBA can also be used effectively for store layout, catalog design, and cross-sell.

In order to select interesting rules from the set of all possible rules, constraints on various measures of significance and interest are used. The best-known constraints are minimum thresholds on support and confidence. Let X, Y be itemsets, $X \Rightarrow Y$ an association rule and T a set of transactions of a given database.

The **support** $supp(X)$ is an indication of how frequently the itemset X appears in the dataset. A high support value means that the rule involve a great part of database. The support of X with respect to T is defined as the proportion of transactions t in the dataset which contain the itemset X.

The **confidence** is an indication of how often the rule has been found to be true. The confidence value of a rule, $X \Rightarrow Y$, with respect to a set of transactions T, is the proportion of the transactions that contains X which also contains Y. The confidence of the association rule is defined as:

$$\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} \tag{3.7}$$

Another measure that characterizes an association rule is the **lift**. This measure evaluates the level of dependency between the elements of an association rule. It is obtained by dividing the support of X and Y, $s(X,Y)$, representing the percentage of occurrences of X and Y in the same

database, by the product of the support of X and Y considered separately, as shown in expression (3.27).

$$\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \times \text{supp}(Y)} \tag{3.8}$$

The lift represents the tendency to buy the product sets X and Y together. If the lift is equal to 1, there is independence between the occurrence of sales of product sets X and Y. If the lift is greater than 1, the products tend to be bought together, and if it is lower than 1, they tend to be bought separately. Rules presenting a lift less or equal than 1 are usually disregarded.

Several association algorithms exist. However, in the context of this dissertation, only the apriori algorithm will be analyzed and used.

3.3.1 Apriori algorithm

Apriori algorithm, a classic algorithm, is useful for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact the algorithm uses *prior knowledge* of frequent itemset properties. Apriori employs an iterative approach known as level-wise search, where k-itemsets are used to explore (k+1) itemsets. For an easy and intuitive explanation of the operation of this algorithm, consider the transaction database 3.1 and minimum support count is 2.

Table 3.1: Transaction database example.

Transaction ID	I1	I2	I3	I4	I5
t_1	1	1	0	0	1
t_2	0	1	0	1	0
t_3	0	1	1	0	0
t_4	1	1	0	1	0
t_5	1	0	1	0	0
t_6	0	1	1	0	0
t_7	1	0	1	0	0
t_8	1	1	1	0	1
t_9	1	1	1	0	0

The first step of the algorithm consists of calculating the support of each individual product and and selecting those that have equal or greater support to the minimum support defined. This gives us itemset L_1.

Table 3.2: Itemset L_1.

Itemset	Support Count
I1	6
I2	7
I3	6
I4	2
I5	2

The second step is to generate a candidate set C2 using L_1 (this is called join step). The condition of joining L_{k-1} and L_{k-1} is that it should have (K-2) elements in common. It is necessary to check all subsets of an itemset are frequent or not and if not frequent remove that itemset. Finally, find support count of these itemsets by searching in dataset and select those that have equal or greater support to the minimum support defined. This give us itemset L_2

Table 3.3: Itemset L_2.

Itemset	Support Count
{I1, I2}	4
{I1, I3}	4
{I1, I5}	2
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2

This process should be repeated until frequent items can not be found. The last itemset found was L_3.

Table 3.4: Itemset L_3.

Itemset	Support Count
{I1, I2, I3}	2
{I1, I2, I5}	2

Once all the frequent itemsets have been discovered, it is necessary to calculate the trust for each rule. For this the calculation formula 3.7 is applied.

3.4 Classification

Classification, one of the most common DM tasks, seems to be a human imperative. In order to understand and communicate about the world, we are constantly classifying, categorizing and grading. Classification consists on examining the features of a newly presented object and assigning it to one of a predefined set of classes. In this dissertation, the classification intends to construct a model to predict consumer behavior through database records into a number of predefined classes based on certain criteria [65] [66].

Data classification is a two-step process as it can be seen in the Figure 3.3. In the first one, i.e the learning step, the classification algorithms build the classifier 3.3 (a). The classifier is built from the training set made up of database tuples and their associated class labels. Each tuple that constitutes the training set is referred to as a category or class. These tuples can also be referred to a sample, an object or data points. In the second one, the classifier is used for classification 3.3 (b). Here the test data is used to estimate the accuracy of classification rules. The classification rules can be applied to the new data tuples if the accuracy is considered acceptable.

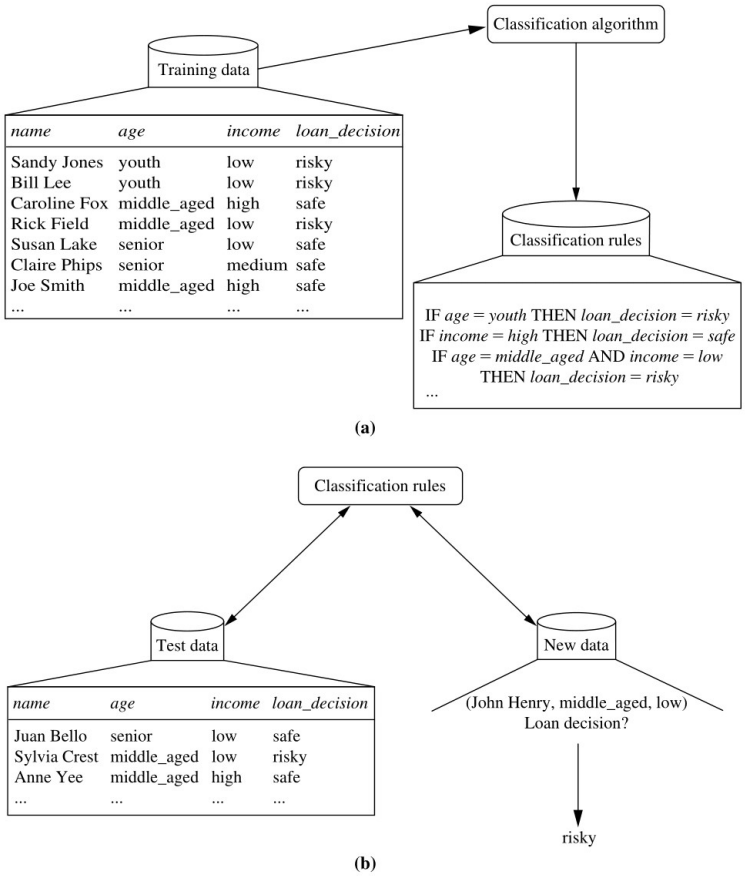

Figure 3.3: Data Classification Process [3].

Common algorithms for performing classification include boosted and bagged decision trees, random forests, Naïve Bayes, k-nearest neighbor, logistic regression, support vector machine (SVM) and neural networks. In the next section will be explained the decision tree algorithm, because it was the only algorithm necessary for the scope of this dissertation.

3.4.1 Decision Tree

Decision tree classifiers are a well-known technique of classification which allows to easily obtain the classification rules. The aim is to create a model that predicts the value of a target variable based on several input variables.

Decision trees are a combination of computational and mathematical techniques to aid the representation, generalization and categorization of a given set of data. A decision tree is defined as a classification procedure that recursively partitions a data set into smaller subdivisions on the basis of a set of tests defined at each branch (or node) in the tree (Figure 3.4). The tree is composed of a root node (formed from all of the data), a set of internal nodes (splits), and a set of terminal nodes (leaves). Each node in a decision tree has only one parent node and two or more descendant nodes.

In Figure 3.4 you can see that each box is a node at which tests (T) are applied to recursively split the data into successively smaller groups. The labels (A, B, C) at each leaf node refer to the class label assigned to each observation.

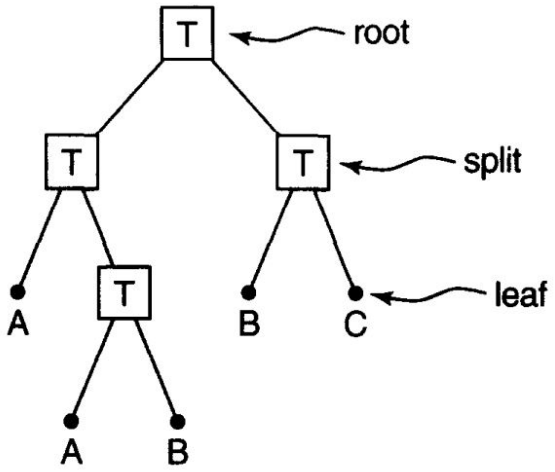

Figure 3.4: Decision tree classifier [4].

A decision tree can be used to predict the value of the class for all items, by starting at the root of the tree and moving through it until a leaf node, which provides the classification value. The full description of the algorithm is given below.

Algorithm 3 Decision tree algorithm

1. Place the best attribute of the dataset at the root of the tree.
2. Split the training set into subsets. Subsets should be made in such a way that each subset contains data with the same value for an attribute.
3. Repeat step 1 and step 2 on each subset until you find leaf nodes in all the branches of the tree.

The advantages of decision trees include the ability to handle data measured on different scales, the lack of any assumptions concerning the frequency distributions of the data in each class, the flexibility, and the ability to handle non-linear relationships between attributes and classes [4]. Finally, the analyst can easily interpret a decision tree. One of the main drawbacks of the decision trees is the instability, which means that different training datasets from a given problem domain will produce very different trees [67].

3.5 Conclusion

DM is a very powerful tool that should be used with utmost care to increase customer satisfaction, providing good, safe and useful products at reasonable and economical prices. This should

be used to make the business more competitive and profitable. DM should be used in any way that protects the privacy of common man, so that the confidentiality and individuality of human being is preserved. It should not be used in any way that may cause undue hardship, financial loss or emotional setback.

Chapter 4

Case study: description of the retail company

This chapter describes the main characteristics of the company used as case study. Section 4.1 describes the current position and importance of the company in the domestic market, the types/formats of the existing stores, the classification of its products and the organizational structure. Section 4.2 aims to describe the data used in this dissertation. Section 4.3 seeks to characterize the company's customers, in particular as regards the consumption of fresh fish.

4.1 Company's description

The company used as case study is one of the largest food retailers in Europe. This company is a reference in the retail market, having started a revolution in the consumption habits and commercial patterns in the country where it is based, with the implementation of the first hypermarket in the 80s. The company's strategy consists of consolidating its leadership position in the market and expanding the frontiers of business, taking advantage of resources and skills development. These goals derive from the company vision of leadership in the business.

There has been an increase in competition in the retail industry in Portugal, and there is a need to redefine strategies in order to obtain the best possible results in all areas of the business. Effectively, in the fresh fish business unit there is a growing concern with the sales volume in recent years. This concern is due to a loss of market share in the fresh fish business area to the competition for no apparent reason. In this sense, in order to design strategies for customer acquisition and customer retention, it became necessary to understand the consumer and his/her behavior in fresh fish sector.

What makes it possible to understand consumer behavior is the data stored through the loyalty card. In fact, one of the most successful ideas that allowed to reinforce the importance of this company in the national panorama was the creation of the loyalty card in 2007. This card can be used in all of the company's stores and gives access to multiple promotional campaigns. In addition to all the advantages it brings to the customer, this card provides the company with the

customer's data. The loyalty program enabled collecting data on each customer profile, i.e. name, address, date of birth, gender, number of people in the household, the telephone number and the number of one identification document. This data is collected when customers join the loyalty program by filling out a form. The use of the loyalty card by customers has also enabled collecting data regarding customer transactions, i.e. date, time, store, products and prices. The above data set, is critical to attempt to understand consumer behavior.

4.2 Exploratory Analysis of Different Stores

In order to support the development of the thesis research, the company made available **two distinct databases**. The first database contains aggregated (non-transactional) data and will be used for a first general analysis. The second database contains transactional data from the stores selected through the analysis of the first database. The second database is the source of the various analyzes carried out in this dissertation. After this summary presentation of the set of databases presented and used in this dissertation, we will proceed to a more detailed and in-depth analysis of each of them.

The first database the company contains sales records for the year of 2017 (01/01/2017-31/12/2017) of 270 stores in mainland Portugal and islands. In total, there are 8.575.208 records of 3 different categories: fresh fish and seafood, codfish and, finally, frozen fish and seafood. In the context of this dissertation, only the 2.618.440 records of fresh fish were analyzed.

At an early stage, it is necessary to understand the major data structures present. First of all, the product structure deserves attention as illustrated in Figure 4.1. In fact, in the scope of this dissertation the category that stands out most is Fresh Fish, consisting of 8 subcategories, 22 base unit and 391 SKU.

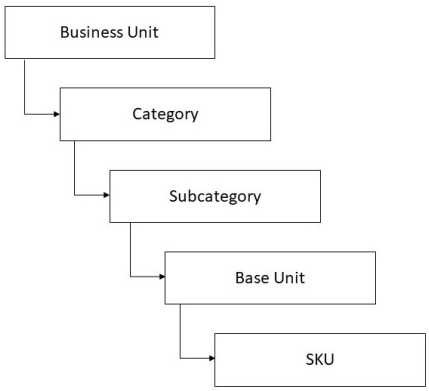

Figure 4.1: Retailer's Product Structure.

The store structure is another that also deserves attention from the analyst. In fact, as mentioned before, this company has a chain of foodbased stores, i.e. hypermarkets, large supermarkets and small supermarkets, denominated of Insignias. These formats differ essentially by the sales

area of the store and by the range and price of products offered. The stores that will integrate the analysis are of the following formats:

- **Continente** Large stores with a wide range of products, located in areas of high population density and customer traffic.

- **Modelo** Medium sized stores, usually outside major cities.

- **Continente Bom Dia** Smaller stores located to serve specific population centres.

Only the stores of mainland Portugal will be accounted for, as well as only these three main store formats, which totals 245 establishments. In Figure 4.2 we can see represented the store structure.

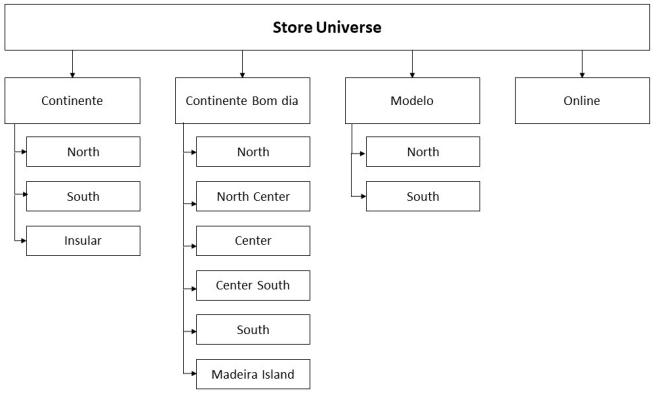

Figure 4.2: Retailer's Store Structure.

After describing the main data structures and the amount of data available, it became imperative to choose a store and a set of products that represent the national consumption for each Insignia. The selection of stores and products to be considered were made according to the two retailer's hierarchical structures - a product structure and a store structure, which were detailed in Figures 4.1 and 4.2.

At first, a comprehensive analysis of the consumption and revenue of the various stores was carried out at national level for the year 2017. This first approach provided an overview and a taste of consumer preferences for fresh fish. As a result of this analysis, it was easy to understand that a small set of species accounted for the majority of consumption and income.

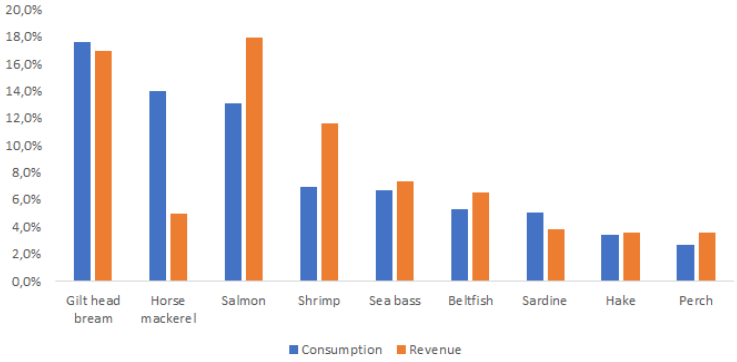

Figure 4.3: Consumption and revenues.

In Figure 4.3 it is possible to see the species that represent about 80% of the consumptions and the revenues. In fact, 9 of the 150 species sold by SONAE stores represent around 80% of the consumptions and the revenues.

After this first approach, it was decided to reduce the spectrum of species and stores to be analyzed in this dissertation to facilitate data pre-processing and processing.

For each Insignia, the store that best approximates the overall consumption profile of that Insignia is selected. Let c_{ij} be the relative consumption of species i in store j, with $i \in \{1,...,N\}$ and $j \in \{1,...,M\}$. The most representative store j^* is selected according to:

$$j^* = \underset{j}{\operatorname{argmin}} \left(\sum_{i=1}^{N} [c_{ij} - \frac{1}{M} \cdot \sum_{j'=1}^{M} (c_{ij'})] \right) \tag{4.1}$$

Through equation 4.1 the results presented in Figures 4.4, 4.6, 4.5 were obtained. In Table 4.1 presented below, it is possible to see the final results of the application of equation 4.1 for each Insignia and the species that represent 80% of the consumption of each Insignia.

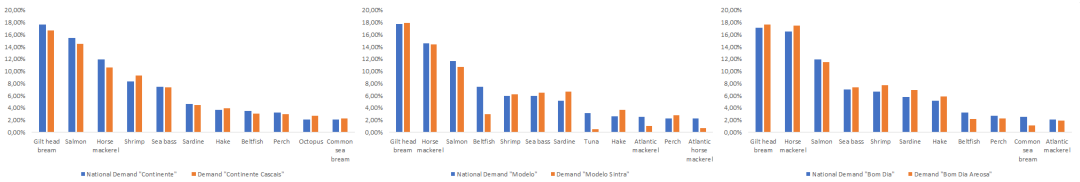

Figure 4.4: Species consumption of the "Continente" representative and of national demand.

Figure 4.5: Species consumption of the "Modelo" representative and of national demand.

Figure 4.6: Species consumption of the "Bom Dia" representative and of national demand.

From this first segregation by Insignia, some differences can be observed regarding consumption. In general, the species consumed in the different Insignias are the same. In fact, there are nine species (gilt head bream, salmon, horse mackerel, shrimp, sea bass, sardine, hake, sword and nile perch) that are repeated in the three Insignias and represent a significant consumption. In a

comparative analysis, the Insignia "Continente" and "Bom Dia" present very similar relative intakes. On the other hand, the Insignia "Modelo" is the one that shows higher differences in both the relative consumption and the variety of species consumed (i.e., twelve different species).

Table 4.1: Stores chosen for each Insignia.

Insignia	**Store**
Continente	Continente Cascais
Bom Dia	Bom Dia Areosa
Modelo	Modelo Sintra

After selecting the set of stores that will be the target of study, we started to explore each of the available variables.

4.3 Exploratory Analysis of Transactional Data

The second database this contains transactional data of the selected stores (see Table 4.1) for the last 2 years (01/01/2017 - 12/31/2018). Through the organization's loyalty program it is possible to link customers to specific transactions and to track each customer's purchase trail. The data sample under study corresponds only to data referring to the year 2018. We believe because the more recent the data, the most past is more representative of what may happen in the future is the data. Each transaction is characterized by two subsets of variables, which describe: (1) the transaction, (2) the customer performing the transaction. These variables are described below.

Variables with information related to the transaction:

- **TID**: it is a retailer transaction ID number.
- **Customer ID**: It is the variable that identifies the customer. it consists of a code with at least four different digits for each customer.
- **Time Key**: is an eight-digit number with the format YYYYMMDD;
- **Location**: name of the store where the purchase was made. Stores are organised in a hierarchical structure, aggregated by format and geographical distribution, as in Figure 4.2;
- **SKU**: it refers to a unique identifier or code that refers to the particular stock keeping unit. Products are also organized hierarchically in a product structure, already outlined in Figure 4.1.
- **Quantity**: continuous quantitative variable for each transaction. In the case of fresh fish, this variable represents the kilograms (kg) traded. Negative values represent the kg returned from product.

- **Gross Sales**: it is a continuous quantitative variable that indicates the value in euros (excluding discounts) for each transaction. Negative values correspond to cash returns.

- **Promotional information**: information from promotional campaigns is available at the SKU/ Day/ Store Format level (only available for fresh fish category)

Variables with customer information:

- **Customer ID**: it is the variable that identifies the customer. It consists of a code with at least four different digits for each customer;

- **Gender**: it is a binary variable that indicates the gender of the customer, being the masculine gender represented by M and the feminine by F;

- **District**: is the variable that identifies the district where the customer is resident;

- **Age Range**: it indicates the age range of the customer. The possible range are: [0:18[,]18:25[,]25:35[,]35:45[,]45:55[,]55:65[, >65.

- **Family Typology**: differentiates the various family groups that may exist, such as: active adults, family supporters, family with young adult, family with kids, senior.

After gathering all the data, it is necessary to choose the set of customers that should be analyzed. As mentioned previously (4.3), only transactions completed in 2018 will be considered. In addition, since this analysis is focused on fresh fish, it should only be considered customers that:

- Bought, at least once, fresh fish;

- Spent, in general, a value higher than 0 €;

- Spent, in fresh fish, a value higher than 0 €;

All customers with outliers in any of the above criteria should also be removed. In this dissertation, the removal of outliers was performed conservatively. In fact, 22 clients were removed, as they had spent too much on fresh fish ($\geq$ 77€) in a single transaction (usually carried out on holidays). This will have to be done, since they are not representative of the typical consumer that is to be studied. After the application of these filters, the characterization and exploitation of transactional and socio-demographic data was conducted.

The number of customers of fresh fish under study is **46,330**. In fact, as can be seen in Figure 4.7, fresh fish customers represent about 23% of the company's customers. Figure 4.7 shows and reflects that most of the company's customers (77%) did not buy fresh fish during the year 2018. This information increases the importance of this study and the various analyses, since the fresh fish sector is a market that deserves to be explored and has a considerable growth margin.

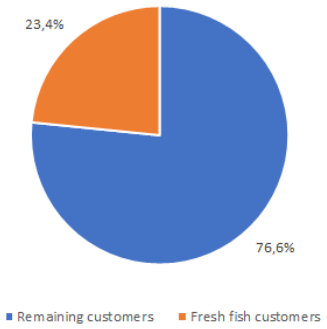

Figure 4.7: Representativeness of fresh fish customers.

Of the fresh fish customers analysed, 61.9% are female, 35.1% are male and 3.1% are not identified (NI). The distribution of their ages is represented in Figure 4.8.

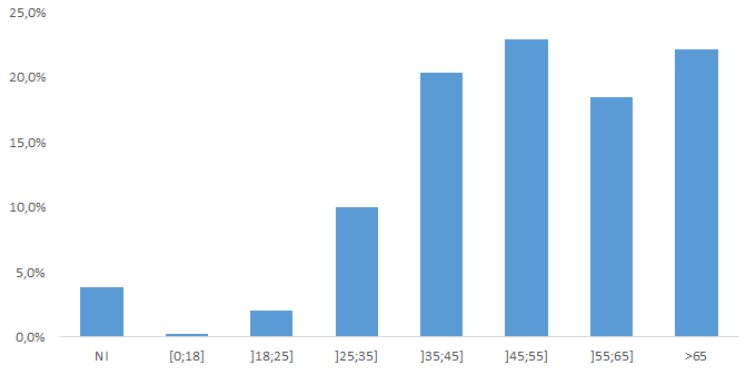

Figure 4.8: Bar chart of the Age Range variable.

In Figure 4.8, it can be observed that the age of the customers has an asymmetric distribution, evidencing that they belong, mostly, to higher age groups. Table 4.8 shows that 80% of the customers are 35 years of age or more, which highlights the asymmetry of this distribution. On the other hand, younger customers do shopping in SONAE stores less frequently. This may be justified either by buying large quantities at once or by their parents making the purchases or they may not be loyal to the company.

In order to perceive the family constitution of the company's customers, the **family typology** variable was analyzed. From the Table 4.2, it can be seen that most customers (63,4%) are active or senior adults.

Table 4.2: Relative frequency of Family Typology variable.

Family Tipology	**Relative Frequency**
NI	1,2 %
Active adults	30,7 %
Family supporters	4,1%
Family w/ young adult	9,6%
Family w/ kids	21,7%
Senior	32,7%

With the analysis of the variables gender, age range and family typology, it is possible state that most of the customers of this company are characterized as being female, over 65 and an active / senior adult.

After the sociodemographic analysis of the target customers of this study, it was decided to perceive their consumption habits. In this sense, a global analysis of consumption will be carried out in the first phase and then a more detailed analysis is carried out on the consumption of fresh fish.

In a first approach to the analysis of the impact and importance of fresh fish consumption in this company, the number of transactions in the fresh fish sector should be compared to all others. The results of this study can be seen in Figure 4.9.

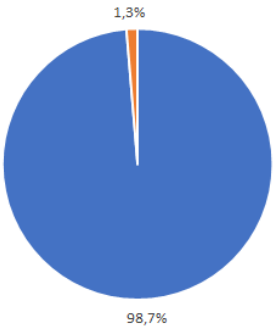

Figure 4.9: Distribution of transactions in 2018.

Table 4.3: Transactions ranking by Category.

Order	Category	Relative value
1	Fruit	6,7%
2	Vegetables	5,7%
3	Yogurt, Dessert	4,3%
4	Bread	3,4%
5	Cheese	3,4%
...	...	...
24	Fresh Fish	1,3%

The results obtained indicate that fresh fish is a category with very few transactions performed (1.3%) and, consequently, is not one of the categories with the highest number of transactions. These results reinforce the importance of this study, since the understanding of consumer behavior will allow increasing the number of transactions performed.

After analyzing the relative consumption frequency by Category, It was decided to understand the sales and spending values, first at a global level and then in the fresh fish sector.

Gross sales are an essential indicator for understanding the importance of Fresh Fish. After analyzing the previous transaction results, gross sales of fresh fish are expected to be low relative to other products. Figures 4.10 and 4.11 show that fresh fish is one of the categories with the highest relative gross sales value, showing the importance of this sector for the company.

Figure 4.10: Distribution of gross sales in 2018.

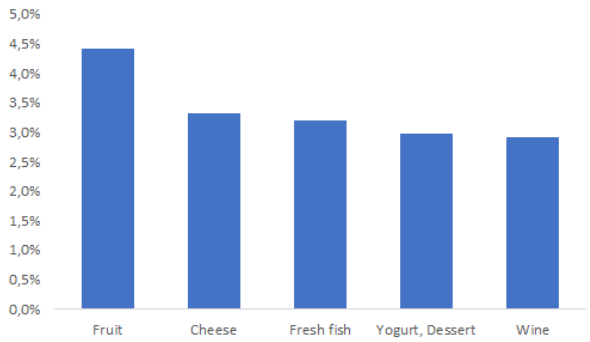

Figure 4.11: Gross sales by Category.

Then, the average amount spent by each customer during 2018 is studied. The results obtained are shown in Figures 4.12 and Table 4.4. This allowed to conclude that the average amount of money spent per visit is about 47,65€. The standard deviation is about 38,43€, which reflects the heterogeneity among customers in terms of value spent (see Figure 4.12).

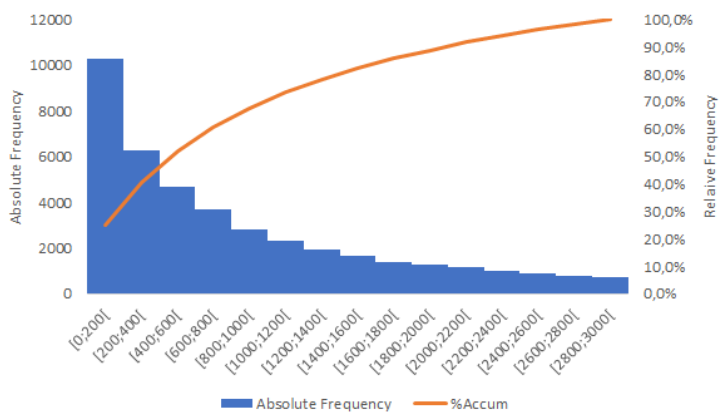

Figure 4.12: Amount spent in 2018.

Table 4.4: Global consumption data.

	Minimum	1st Quarter	Median	3rd Quarter	Maximum
Total number of store visits	1	7	19	42	360
Amount spent	1,12	235,54	699,49	1754,02	55919,33

With the knowledge of the values spent at the global level, it is essential to know what the values are spent on fresh fish. In this sense, the Figure 4.13 gives an idea of the distribution of the value spent on fresh fish. This allowed to conclude that the average amount of money spent per visit is about 10,42€.

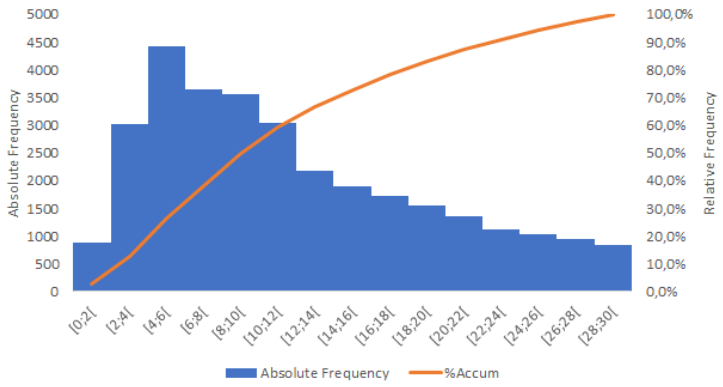

Figure 4.13: Amount spent on fresh fish in 2018.

Table 4.5: Fresh fish consumption data.

	Minimum	1st Quarter	Median	3rd Quarter	Maximum
Total number of store visits	1	1	2	4	154
Amount spent	0,2	7,79	16,61	40,88	6761,34

The standard deviation is about 8,44€, which reflects the heterogeneity among customers in terms of value spent (see Figure 4.12).

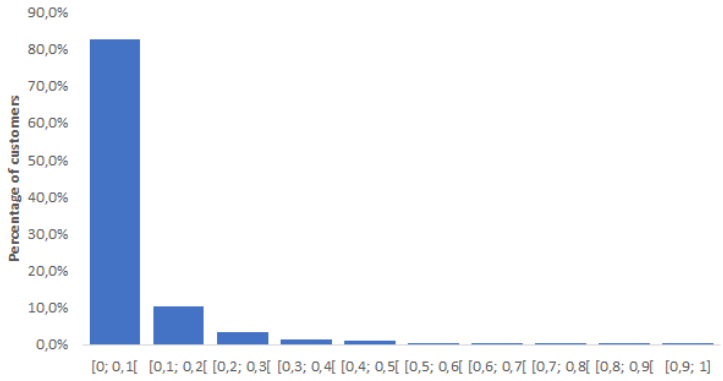

Figure 4.14: Ratio of the amount spent on fish to the amount spent in general.

Through Figure 4.14, it is possible to understand that most customers (82.5%) only spend 10% of the total value on fresh fish, thus demonstrating the reduced share of wallet of fresh fish, and hence the need for this study.

Having analyzed the amounts spent in global and specific terms, it is essential to realize how often a customer buys fresh fish during the year 2018.

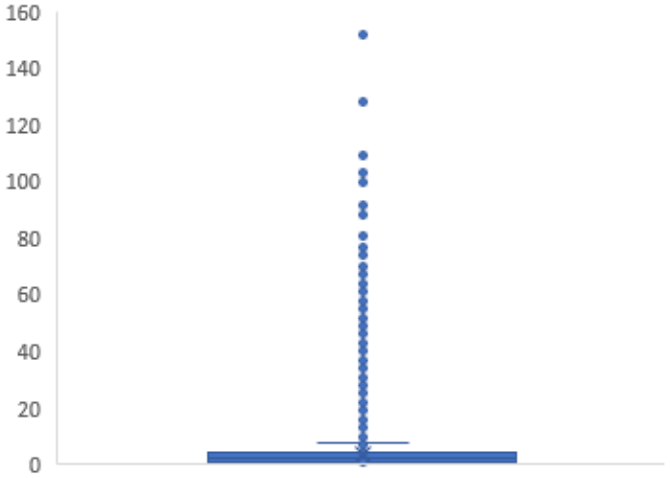

Figure 4.15: Number of trips to the store to buy fresh fish.

Table 4.6: Fish purchase results.

Statistics	Results
Minimum	1,0
Mean	3,9
Standard deviation	6,1
Maximum	154

In fact, the annual fish purchase frequency results are surprising since they show that most customers rarely buy fresh fish. In addition, customers buy fresh fish about 4 times per year, on average, with a standard deviation of 6 times per year.

It is interesting to note that according to the business experts, those people who are at the lower extreme of the mean time between purchases histogram are elder and/or retired people, whose visit to the store is part of a routine to avoid loneliness.

Another concern that should be taken into account in this study is the understanding of how often people go to the store and buy fish. In Figure 4.16 you can see that most consumers when going to the mall rarely buy fresh fish. In fact, about 35% of customers only buy fresh fish at 10% of the purchases they make.

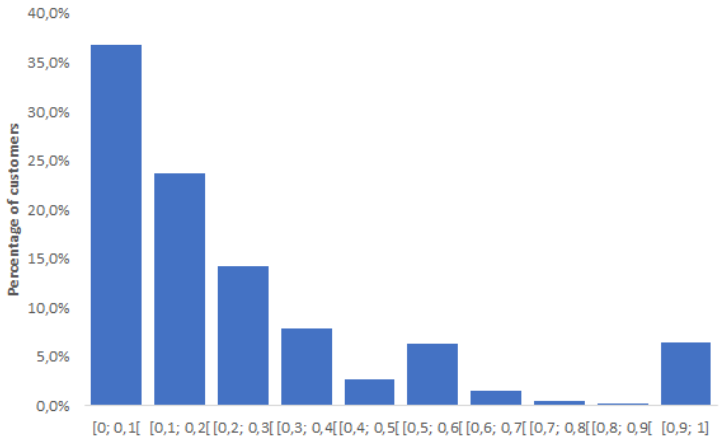

Figure 4.16: Fresh fish purchases ratio.

Finally, one of the most important factors influencing decision making, as seen in Chapter 2, is the reduction of the price of the product through promotion. In this sense, in order to understand the effects of the promotion on the purchase of fresh fish, a study was carried out on the number

of fresh fish purchases that result from promotions (ratio of fresh fish transactions on promotion of total fresh fish transactions). Figure 4.17 displays the results achieved.

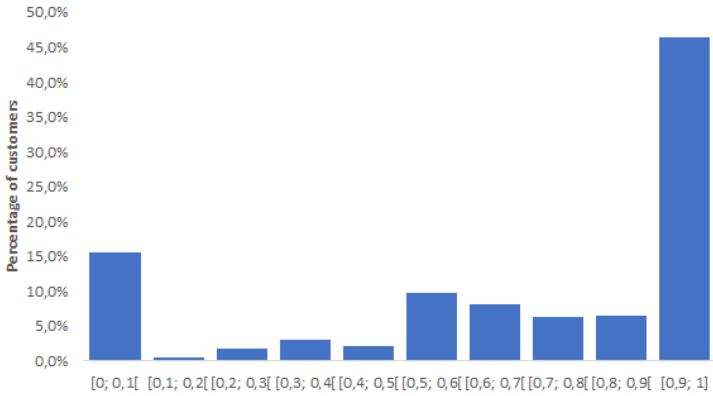

Figure 4.17: Ratio of transactions made in promotion.

The results obtained are clear, demonstrating that around 46.5% of consumers made the majority of their purchases using promotions. This result also shows the importance that promotion can have on consumer decision making and therefore must be a variable to take into account in consumer behavior modelling.

From the data obtained it is already possible to make a first diagnosis to the target customers of this study. They stand out because of the low number of times they buy fish when they go to the store, suggesting that they are either not fish consumers or buying fish at other retailers. In addition, it can be seen that the vast majority of fresh fish transactions are made on sale, showing that the price decrease of a product positively affects the quantities sold. Despite this initial perception of consumer behavior, with the segmentation that will be made other conclusions can be obtained. After completing this analysis of global consumption patterns, it is necessary to understand if there are significant differences between the three Insignias.

Chapter 5

Consumer behavior analytics

In order to decide and apply efficient and effective marketing strategies, a thorough knowledge of customer behavior is required. This way it will be possible to meet the real preferences and needs of customers. The characterization of customers can be achieved through market segmentation [68].

Market segmentation creates subsets of a market based on demographics, needs, priorities, common interests, and other psychographic or behavioral criteria used to better understand the target audience. Through the characterization and understanding of the various market segments it will be possible to leverage this segmentation in product, sales and marketing strategies.

Nowadays, with the high offer of products and services available, there is also a high diversity of customer needs and buying behavior. The current market segmentation models are based on consumer behavior deducted through the transactional record. Transactional data analysis is possible using DM techniques, such as cluster analysis (see section 3.2).

Another of the studies that can be carried out is the analysis of purchasing trends, such as products that are often bought together. Market basket analysis, supported by association DM techniques, is widely used for extracting product association rules. These rules can then be used by companies to propose potential purchases to customers, which are often associated to discounts as an incentive to buy. Market basket analysis is often used to support the design of promotions for all company customers in a massive scale, or to support decisions of product assortment within stores.

Finally, this chapter will present in detail the methodology, approach and results of each of the four analyses carried out. At the end of this chapter it will be possible to have an accurate view of the consumer behaviour of fresh fish and to obtain differentiated strategies to attract customers.

5.1 Methodology

A detailed analysis of SONAE's transactional data using DM techniques allows a deeper knowledge of the market and consequently allow designing to outline differentiated strategies,

taking into account the preferences of each customer. Thus, a set of studies and analyses will be carried out with the main objective of understanding the behavior of the fresh fish consumer.

The methodology proposed by this study aims not only to identify groups of similar customers, using segmentation techniques, but also to analyse the shopping cart of a fresh fish customer. These analyses will show the potential that the extraction of knowledge from large databases to improve strategic decisions of companies and their relationship with customers.

The database used for the above methodology corresponds to the database 2 presented and described in Chapter 4. For the extraction and selection of data the software used was Microsoft SQL Server. The pre-processing and data analysis phase was performed using Microsoft Excel. In the data processing task, the software used is R.

For segmentation purposes, the algorithm used to segment customers was the *k-means* algorithm, due to the speed of processing, efficiency and ease of application to the database under study. This segmentation algorithm requires, a *priori*, the definition of the number of clusters (k). To support the choice of the appropriate number of clusters, a curve was constructed that translates the reduction of the *Sum of Squared Errors* with the increase of the k value (*elbow curve*) and three different statistical methods were also used.

In addition to customer segmentation, based on their behavior relative to the purchase of fish, an analysis of the shopping cart was developed in order to determine rules of association between purchased goods, using the a *priori* algorithm. The choice of this algorithm is due to the fact that it is not computationally very demanding and to the fact that it is easy to implement.

This chapter presents in detail the methodology, approach and results of each of the four analyses carried out. At the end it will be possible to have an accurate view of the consumer behaviour of fresh fish and to obtain differentiated strategies to attract customers.

5.2 Behavioral market segmentation

5.2.1 Approach

The segmentation model introduced in this section consists of grouping customers based on their buying habits. Customer segmentation allows companies to efficiently and effectively target promotional campaigns, which are an essential means of attracting and retaining customers. This section presents a model that segments fresh fish consumers according to behavioural factors.

Concerning segmentation approaches informed by transaction records stored in databases, the RFM model introduced by a catalog company in the 1920's is an example of a widespread approach for segmenting customers by means of clustering techniques [69]. This model explores the information on the date of the last purchase (recency), on how often the customer makes purchases (frequency) and on the amount spent (monetary), extracted from the transactional database. Recent segmentation studies using the RFM model, whose objective was to specify segments in the hardware retail market [70].

In this case, after a literature analysis, it was decided that customer segmentation would be based on five variables:

- **Monthly Frequency**: variable that seeks to understand how many times per month a customer goes to the store to buy fresh fish. Frequency is a measure of the strength of the customer relationship with the fish sector. Loyal customers, by definition, purchase more often than disloyal customers [71].
- **Frequency Ratio**: variable that results from the division between the number of times the customer goes to the store to buy fresh fish and the number of times the customer goes to the store to buy any product.
- **Value Spent**: variable that represents the average amount spent on each fish transaction. The monetary value of each customer's past purchase can be an important predictor of future behavior [72].
- **Value Spent Ratio**: variable resulting from dividing the amount spent on fish by the amount spent on all product, within the whole period. With this variable, we can verify the impact that fresh fish has on the customer's shopping basket and the importance that this product can have for the customer.
- **Promotions Transactions Ratio**: variable resulting from the division of the number of fresh fish transactions carried out in the promotional period and the total number of fresh fish transactions. This variable represents the impact that promotions can have on the decision to buy fresh fish.

These five indicators represent *proxies* of customers' purchasing habits. After segmenting customers, a classification algorithm was used to characterize the profile of consumers. In order to obtain a finer analysis of the results obtained, a socio-demographic characterization of the set of customers of each cluster was carried out. Finally, the consumption of species for each cluster will be explored. In this way, it will be possible to understand if there are different preferences and if this influences the buying behavior.

In this section a segmentation based on Insignia's behaviour will also be made. Thus, the company will be able to understand the various consumer behaviors in the various Insignia.

5.2.2 Behavioral segments

The data set considered consists of 46,330 objects characterized by five variables. Each object corresponds to a customer.

Each customer is assigned to a cluster, according to the values it presents in each of the 5 variables mentioned previously. The percentage of customers belonging to each cluster is important to guide the marketing actions necessary to meet the company's objectives. That is, the higher the percentage of customers in a cluster, the greater its importance in strategic decision making. As

an example, if the number of customers in the clusters with more visits to the fish market is small, the company might want to launch marketing campaigns in order to motivate customers to go to the fish sector more frequently.

Regarding the number of market segments in the fresh fish sector, from the elbow curve, it was concluded that the most appropriate option would be the adoption of **three clusters**. This option was reinforced with the results obtained through other statistical methods, introduced in Table 5.1.

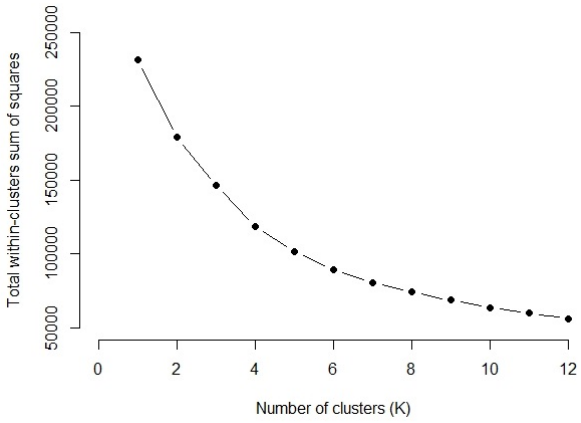

Figure 5.1: Elbow curve.

Table 5.1: Statistical results obtained for the number k of clusters.

Methods	Number of clusters (k)			
	3	4	5	6
sim.ind	**0,997**	0,993	0,992	0,741
opt.assign	**0,803**	0,490	0,363	0,402
nselectboot	0,057	0,041	**0,024**	0,062

As specified, three clusters were obtained. The distribution of the number of customers among the various clusters can be seen in the following Figure 5.2.

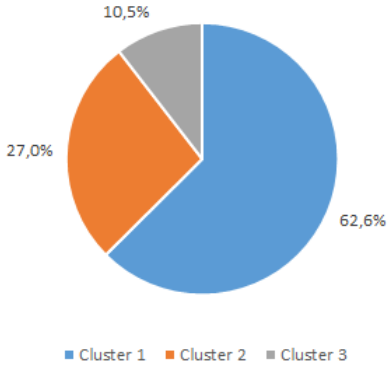

Figure 5.2: Pie chart of the relative distribution of customers in each cluster.

It can be seen that the distribution of customers across clusters is not balanced. Cluster 1 is the largest with 28,980 customers and represents approximately 62% of all customers under study. Clusters 2 and 3 then appear with 12,494 (22%) and 4,856 (10.5%) customers, respectively.

Table 5.2: Average values, globally and per cluster, of each customer's variables

	Monthly Frequency	Frequency Ratio	Value Spent	Value Spent Ratio	Promotions Transactions Ratio
Global	0,33	0,23	7,51	0,07	0,69
Cluster 1	0,39	0,17	**7,09**	0,04	**0,90**
Cluster 2	0,21	0,15	8,17	0,04	0,21
Cluster 3	0,26	**0,83**	8,35	0,28	0,71

In addition to the Table 5.2, another way to understand the results obtained is by using a parallel coordinate graph. With this representation, it is possible to observe the main differences and similarities between clusters. The following Figure 5.3 facilitates and clarifies the results obtained.

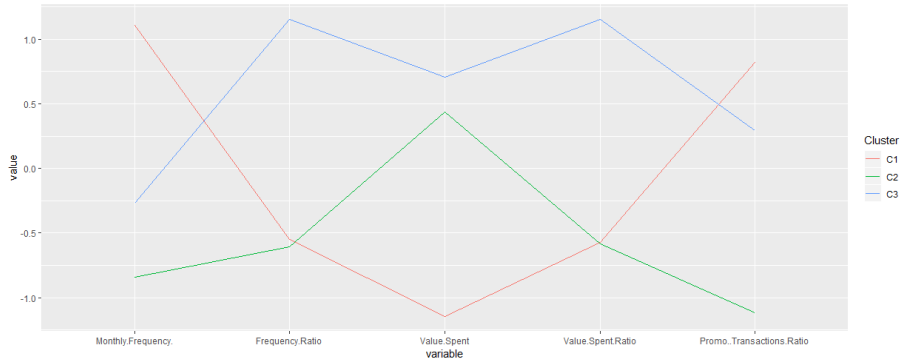

Figure 5.3: Coordinate plot for categorical results.

Looking at the results achieved, it is possible to characterize the customers that integrate each segment. A summary of the particularities of these customers is provided below.

- **Cluster 1**: this is the largest cluster, comprising 28,980 customers (62.6% of the total sample). This cluster is composed of people who go to the store most frequently to buy fresh fish, presenting figures considerably above the global average. However, this represents only 17% of the times that they make purchases in this company. These customers are even more distinguished in the **Value Spent** variable, which is the cluster that spends the least money on fresh fish. Despite this, these values can be justified by the high number of transactions carried out under promotion (**Promotions Transactions Ratio**). In an analysis of the shopping cart, only 4% of the total value is intended for fresh fish.

- **Cluster 2**: this cluster differs from the other two by the low values presented in the **Monthly Frequency** and **Promotions Transactions Ratio** variables. In fact, this group of customers

is not only the one who buys fresh fish less often, but also the one who buys less fresh fish when going to the commercial surface (only 15%, the lowest value among all clusters). In the analysis of the shopping cart, the fish represents only 4% of consumption, being a clearly low value. Despite this, the average value spent on fresh fish is considerable, which shows that there is a high consumption of other products. In fact, this group of customers does not have a great affinity with the company, being a likely customer of a competitor retailer. Despite this, it is a cluster that seems to be from a high social stratum, since both the amounts spent on fresh fish and in other products are high and the sensitivity to promotions is low.

- **Cluster 3**: this is the least representative cluster of the group of customers analysed. However, it is a group of customers that deserves attention for its distinctive consumption characteristics. In a generic analysis, in 4 of the 5 variables studied, the values obtained are higher than the global average. This group of customers, despite having a low Monthly Frequency of purchases of fresh fish, buy fresh fish 83% of the time they buy from a commercial surface. This shows that this cluster is composed of a group of loyal customers with a preference for fresh fish from SONAE. In addition, this group of consumers spends the most on fresh fish and makes a large number of promotional purchases. Finally, this cluster has the highest proportion of value spent on fresh fish in relation to what was spent on all products. This result proves, once again, the loyalty and preference of this group of consumers for SONAE.

After a generic characterization, a decision tree was constructed to understand the influence and importance of the 5 variables in the formation of each cluster.

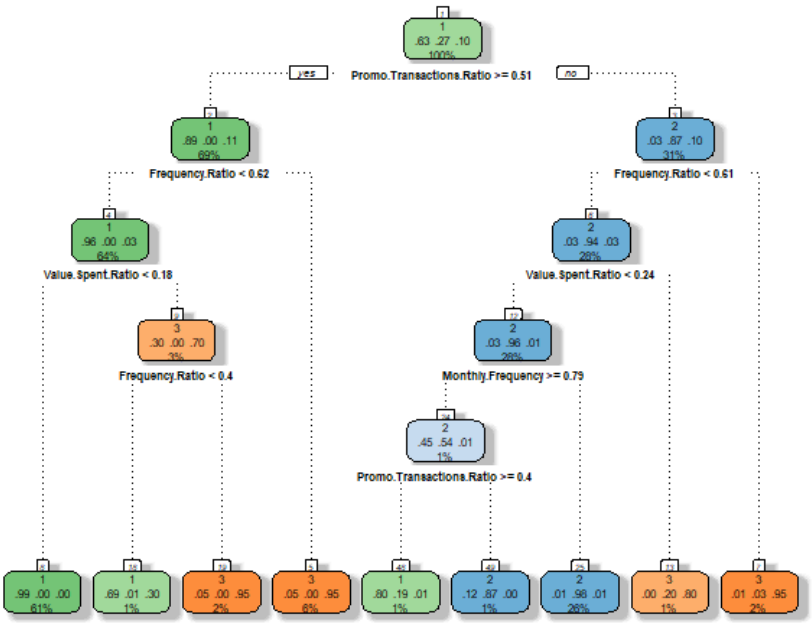

Figure 5.4: Decision tree.

Using the classification process, a performance of 98% was obtained, revealing a high discriminatory capacity of the clusters assigned based on the rules established by the decision tree. The constructed tree enables an easy assignment of new customers to the segments, based on some transactions, without requiring a new development of a segmentation analysis .The other conclusion that can be drawn from the tree is the fact that the **Promotions Transactions Ratio** variable is the one that most distinguishes the customer segments.

It is important to characterize each of the clusters obtained with the new information, in order to understand if there is a relationship between these characteristics and the behavioral variables.

This analysis begins by evaluating the distribution of customers belonging to each cluster in the variables **Gender** and **Age Group**.

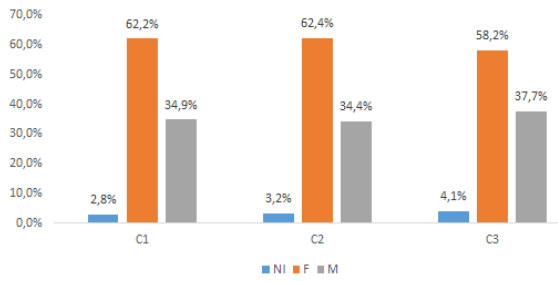

Figure 5.5: Gender distribution per cluster.

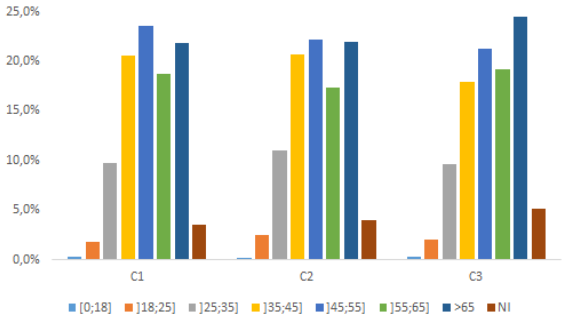

Figure 5.6: Age range per cluster.

Regarding gender, there are no significant differences in the distribution of customers of each gender in the different clusters. Women are the predominant gender in all clusters. As for the **Age Group**, the results obtained suggest that there are significant similarities between clusters 1 and 2 and, on the other hand, there are slight differences between these two clusters and cluster 3. In fact, clusters 1 and 2 have a higher relative percentage of people aged between 35 and 55 than cluster 3. Cluster 3 shows a higher relative percentage of people aged over 65 years. These results are in line with those presented below, which mirror the distribution of customers by *lifecycle* segment, defined by SONAE, for each of the segments related to the purchasing behaviour of fish.

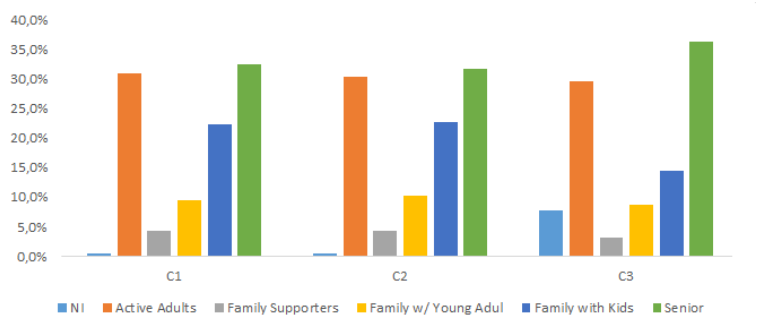

Figure 5.7: Family typology per cluster.

Briefly, from the sociodemographic analysis performed by cluster, it is possible to conclude that there are no relevant differences between clusters.

The results obtained can also be complemented with the information regarding the preferential consumption for each cluster. In this sense, the relative consumption of each species for each cluster was explored. By temporal restrictions, only species that represent 80% of the consumption are analyzed.

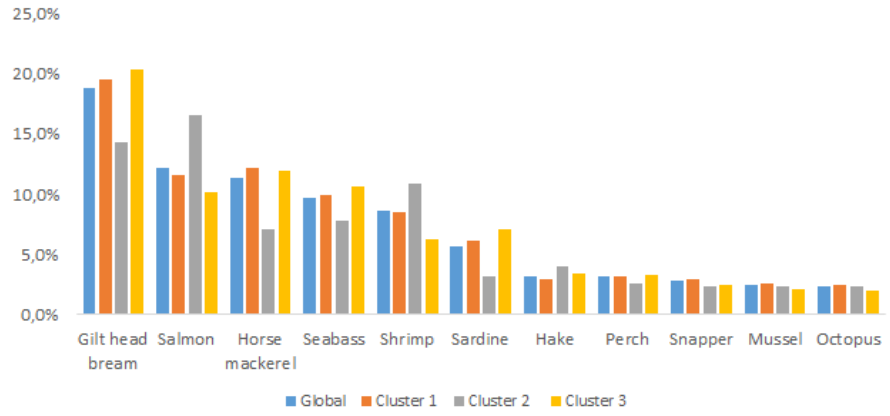

Figure 5.8: Species consumption per cluster.

From the results obtained, it is possible to observe differences in fish consumption between the clusters. Cluster 1 and 3 show a consumption very similar to the global average consumption. Cluster 2 shows significant differences in consumption. In fact, this cluster stands out not only for its low consumption in **gilt head bream** and **horse mackerel**, but also for its high consumption of **salmon** and **shrimp**. It can be seen that cluster 2 has a preference for more expensive fish, while the other clusters have a higher sensitivity to lower value species. After an in-depth analysis of each cluster, it is possible to make a reliable and realistic characterization of each set of customers.

Firstly, cluster 3, which was distinguished by the loyalty and preference for fresh fish from this company, is composed of more senior customers, something that can have several interpretations. The first is that, according to the specialists, the older customers have a greater concern with health and nutrition, which is why they consume greater quantities of fresh fish. The second reason is that senior customers visit shopping centres more often because of the loneliness they face at that stage of their lives.

The other two clusters (1 and 2) have similar social and demographic characteristics, but with differing behaviours. We hypothesize that, they appear to be from **different social strata**. Indeed, cluster 1 makes most of its purchases of fresh fish using promotion, on the other hand cluster 2, despite the high average value spent on fresh fish, makes little use of promotion. Another factor to take into account is related to **loyalty to the company**. Cluster 1 has the group of customers that most often attends the fresh fish sector, while cluster 2 in both the **Monthly Frequency** variable and the **Frequency Ratio** variable shows low values. Thus, these observations indicate

that cluster 1 is more loyal and sensitive to promotions and cluster 2 purchases from competing retail companies.

Finally, in order to deepen knowledge, it was decided to explore consumer behavior at the level of each Insignia. With this motivation, the process done previously at a global level is repeated at the level of the 3 selected stores (described in Chapter 4). The results obtained in the analysis of each Insignia allowed us to conclude that **there are no substantive differences between different Insignia**. In this way, consumer behaviour is reflected in the same way in the various stores of this company. The results obtained can be consulted in the appendix of this dissertation.

5.3 Segmentation based on relative consumption

5.3.1 Approach

Although some differences were seen in the previous clusters in terms of species consumption, a new clustering was attempted with this second perspective as a starting point. This new approach may help to create promotional campaigns at the product level, aimed at certain customers. In order to make this analysis successful, it was first chosen the species that represent 80% of national consumption. The relative consumption of each of these species in 2018 is then exploited on a customer-by-customer basis. Finally, using the *k-means* method, consumer segmentation is carried out based on this relative consumption of fish species.

5.3.2 Relative species consumption

In order to know the preferences of consumers of fresh fish, it is necessary to have a classification of the species most consumed, as well as their absolute consumption and relative consumption. Due to time constraints and data processing, only species representing about 80% of relative consumption at national level are considered for this study.

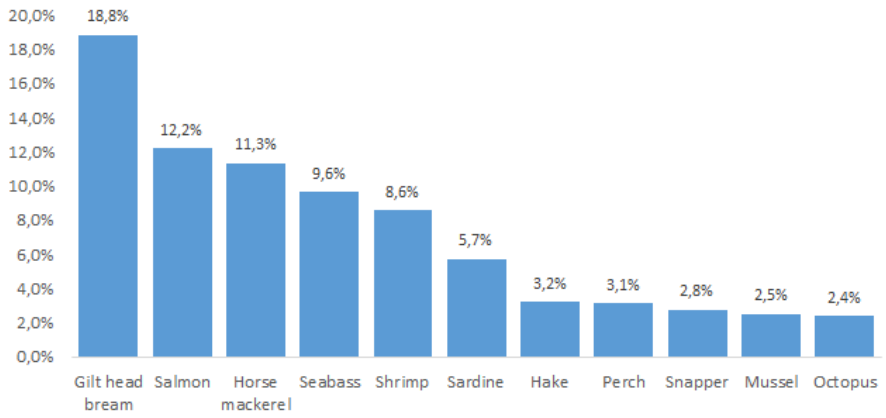

Figure 5.9: Most consumed species.

Through the previous Figure 5.9, it can be verified the set of species that represent about 80% of the national relative consumption. For each of these species, the relative consumption of each customer was analyzed. The definition of the number of segments to be obtained was made through the metrics described in section 3.2.3, culminating in the choice of 5 clusters.

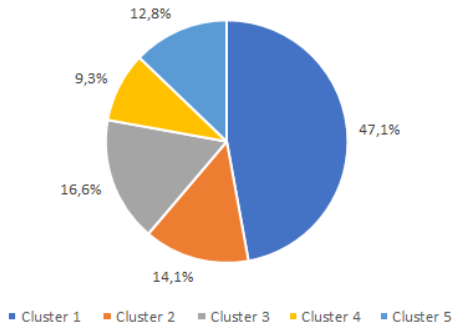

Figure 5.10: Cluster distribution.

Cluster 1 is the largest with 21833 customers, representing about half of them (47.1%). Clusters 2, 3 and 5 have similar customer numbers: 6539, 7698 and 5953, respectively. Finally, the smallest cluster is 4, which is made up of 4307 customers (9.3%).

Table 5.3: Relative Consumption.

	Global	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Gilt head bream	16,9 %	**7,8 %**	1,0%	**73,6%**	5,1%	2,9%
Salmon	15,7%	6,4 %	1,9%	6,5%	3,0%	**85,5%**
Horse mackered	9,2%	4,3%	0,4%	2,8%	**70,6%**	0,7%
Sea bass	8,3%	**14,8%**	0,6%	5,3%	2,0%	1,6%
Shrimp	16,0%	4,8%	**90,6%**	2,5%	2,0%	2,4%
Sardine	4,5%	**8,0%**	0,3%	1,4%	4,7%	0,5%
Hake	2,3%	4,5%	0,2%	0,5%	0,8%	0,4%
Nile perch	3,0%	5,1%	0,4%	1,2%	1,0%	1,8%
Snapper	2,5%	4,4%	0,2%	1,0%	1,3%	0,4%
Mussed	2,2%	4,4%	0,4%	0,3%	0,5%	0,2%
Octopus	1,6%	3,0%	0,1%	0,3%	0,4%	0,2%

Through the results achieved, conclusions can be drawn out to support strategic decisions for the future of the fresh fish sector at SONAE. In general, each cluster, with the exception of cluster 1, shows a predominant consumption of a single specie. Cluster 1 shows a very balanced consumption among the various species considered. In fact, the highest value is 14.8% and the lowest value is 3.0%, showing that there is not a great disparity of values. For this reason, it can be seen that this cluster is made up of a group of customers who consume a large variety of fresh fish. Cluster 2 shows quite different characteristics. This cluster shows a clear preference for shrimp consumption (90.6%). Cluster 3 is characterized by a preference for sea bream, which is the fresh fish most consumed in absolute terms. Cluster 4 has a tendency towards horse mackerel

and cluster 5 has a clear preference (85,5%) for salmon. Once the preferences of each cluster have been characterized, marketing strategies should be devised to attract customers with their preferred product.

As a second step, the need and importance of cross-referencing these results with the results obtained in relation to consumer behaviour was realized.

Table 5.4: Intersection of the two segmentations.

			Behaviour segmentation clusters			Total Row
			Cluster 1	Cluster 2	Cluster 3	
Relative consumption clusters	Cluster 1	Num. of clients	13544	6021	2268	21833
		% of total	29,2%	12,9%	4,9%	47,1%
		% by row	62,0%	27,6%	10,4%	100,0%
		% by column	46,7%	48,2%	46,7%	-
	Cluster 2	Num. of clients	4093	1671	775	6539
		% of total	8,8%	3,6%	1,7%	14,1%
		% by row	62,6%	25,6%	11,9%	100,0%
		% by column	14,1%	13,4%	15,6%	-
	Cluster 3	Num. of clients	5184	1648	866	7698
		% of total	11,2%	3,6%	1,9%	16,6%
		% by row	67,3%	21,4%	11,2%	100,0%
		% by column	17,9%	13,2%	17,8%	-
	Cluster 4	Num. of clients	2957	917	433	4307
		% of total	6,4%	2,0%	0,9%	9,3%
		% by row	68,7%	21,3%	10,0%	100,0%
		% by column	10,2%	7,3%	8,9%	-
	Cluster 5	Num. of clients	3202	2237	514	5953
		% of total	6,9%	4,8%	1,1%	12%
		% by row	53,8%	37,6%	8,6%	100,0%
		% by column	11,0%	17,9%	10,6%	-
Total Column			28980	12494	4856	46330
			62,5%	26,9%	10,4%	100,0%
			100,0%	100,0%	100,0%	-

From the analysis of this crossing between segmentations several ideas can be drawn to design marketing actions. These take into account not only the customer's consumption patterns, but also their favorite species.

Most of the customers in the 3 clusters based on consumer behavior are in cluster 1 of relative consumption. In fact, the preference for a wide variety of species (cluster 1) is transversal to all clusters.

In an analysis of the relative values per column, a set of strategies can be outlined and thought out. For the largest cluster resulting from behavior based segmentation (cluster 1), there is also a large set of customers who prefer shrimp (cluster 2 of segmentation based on relative consumption) and others by gilt head bream (cluster 3 of segmentation based on relative consumption). In cluster 2 resulting from behavior based segmentation, there is a considerable number of customers who have a preference for salmon. Cluster 3 resulting from behaviour-based segmentation has the

various customers distributed homogeneously across the various relative consumption clusters. For cluster 3 resulting from behaviour-based segmentation, promotional campaigns for gilt head bream should be conducted. Another fact that deserves attention is the residual number of customers in cluster 3 that has a preference for horse mackerel. In this sense, a promotional campaign in this group of customers should not bring great benefits to the company.

From the results obtained, future strategies should be considered to improve the results obtained in the fresh fish sector. In fact, the first strategy is to have a greater focus on species with high relative consumption (gilt head bream, salmon, horse mackered, sea bass, shrimp and sardine). Subsequently, for each cluster present differentiated strategies for the most consumed species.

5.4 Segmentation based on wild/farmed fish

5.4.1 Approach

As already mentioned, in order to add value to SONAE we were concerned to know and understand better the customer of the fresh fish sector. For this reason, the need arose to understand the consumers' preference for fresh fish or aquaculture. This problem is important not only for SONAE, as a retailer, but also for the knowledge of the entire fresh fish value chain. Fresh fish supply chain management is more complex as it is less controllable. For example, there may be more stock losses in wild fish (e.g. for reasons of fishermen's strike, adverse sea conditions, etc.) than in farmed fish.

For the above reasons the replacement of wild fish by farmed fish should be considered. Thus, customers of fresh fish were segmented to understand their preferences at category level (wild, aquaculture or other). In order to obtain a detailed and accurate analysis of the consumer, a cross-check of data between the clusters obtained and the clusters based on consumer behaviour was carried out (section 5.2).

5.4.2 Relative consumption by origin of fresh fish

In order to know the preferences of all fresh fish customers, we explored the relative consumption, customer by customer, of each of the three categories: farmed, wild and others. The results in Table 5.11 show the existence of 3 groups of consumers.

For this segmentation, the distribution among customers is very similar. Cluster 2 is the largest of the clusters, comprising 19,257 customers (41.6%). Cluster 1 is the smallest group, with 11,870 customers. Finally, cluster 3 has 15,203 customers.

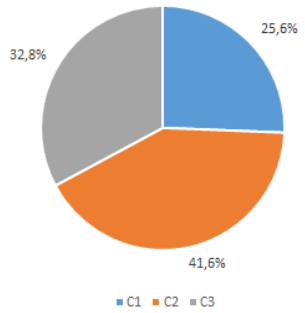

Figure 5.11: Cluster distribution.

The results achieved for the consumption of each of the categories are presented in the following Table 5.5.

Table 5.5: Relative consumption by Category.

	Farmed	Wild	Other
Cluster 1	7,0 %	**89,7%**	3,3%
Cluster 2	**94,6%**	3,4%	2,0%
Cluster 3	32,3%	26,7%	**41%**

The results obtained are interesting and allow comparison with the study carried out and presented in Chapter 2. In fact, according to this analysis, customers consider wild fish better than farmed fish in terms of taste, safety, healthiness and nutritional value. However, the results obtained support another argument. In fact, the largest cluster obtained (cluster 2) is characterized by a high consumption of farmed fish. In this sense, the price of a product is more important in decision making than its quality and health benefits. In other words, the value of the product (trade-off between quality and price) is an important and essential factor for the decision making of the fresh fish client.

In order to have a more comprehensive knowledge of consumer preferences, a cross-check was made between the clusters obtained and the clusters resulting from segmentation based on consumer behaviour. Table 5.6 shows the results of this crossover.

The results obtained allow the company to understand if the origin of fresh fish has different impacts for different customers (clusters) and, from there, to define differentiated strategies.

The first fact that stands out is that the preference for farmed fish is transversal to all clusters. An analysis by column shows that around one third of cluster 1 and cluster 2 (resulting from behavior based segmentation) customers have a preference for other fresh fish origins. In cluster 3, there is not much difference in preference between wild fish and other origins.

In short, the customers of the various clusters have a clear preference for farmed fish, something that is justified by price sensitivity. In this sense, it can be concluded that price is the critical variable for all clusters in choosing the origin of fresh fish.

Table 5.6: Intersection of the two segmentations.

			Behaviour segmentation clusters			Total Row
			Cluster 1	Cluster 2	Cluster 3	
Relative origin consumption	Cluster 1	Num.of clients	6956	3542	1372	11870
		% of total	15,0%	7,6%	2,9%	25,6%
		% by row	58,6%	29,8%	11,6%	100,0%
		% by column	24,0%	28,3%	28,3%	-
	Cluster 2	Num. of clients	12343	4844	2070	19257
		% of total	26,6%	10,5%	4,5%	41,6%
		% by row	64,1%	25,1%	10,7%	100,0%
		% by column	42,6%	38,8%	42,6%	-
	Cluster 3	Num. of clients	9681	4108	1414	15203
		% of total	20,9%	8,9%	3,0%	32,8%
		% by row	63,7%	27,0%	9,3%	100,0%
		% by column	33,4%	32,9%	29,1%	-
Total Column			28980	12494	4856	46330
			62,5%	26,9%	10,5%	100,0%
			100,0%	100,0%	100,0%	-

5.5 Market basket analysis

5.5.1 Approach

With regard to a strategy of differentiated promotions, an analysis of the shopping cart was carried out to identify product associations.

For the different product levels (Category and Subcategory), different product association conditions were considered. However, for both cases, only associations for a lift greater than 1 (to ensure the usefulness of the rules) were considered. For the purpose of this analysis, the shopping cart represents the set of all products purchased in each transaction during the analysis period (2018).

In a preliminary phase, an analysis of the shopping basket was made at the category level. After this exploratory phase and focusing on this problem, the analysis of the shopping cart was made at the subcategory level, since the objective was to discover the type of products potentially interesting to SONAE customers.

5.5.2 Mining Frequent Patterns, Associations, and Correlations

Initially, it is imperative to understand which categories are the target of most consumption by the customers who are being targeted by this study. As seen in Figure 4.9 and Table 4.3, fresh fish (represented by number 1201) is not one of the most consumed categories. The categories were considered associated if they met the following criteria: lift greater than 1, confidence greater than or equal to 30% (to ensure the reliability of the rules) and support greater than or equal to 2.5% (ensure that the rules were relatively frequent, given the number of shopping carts analyzed). Table 5.7 presents the results obtained.

Table 5.7: Association rules at Category level.

Antecedent (x)	**Consequent (x)**	Lift	Conf.	Supp (x,y)
Fresh fish, Cheese, Vegetables spec.	Vegetables	2,27	74%	2,5%
Fresh fish, Vegetables spec.	Vegetables	2,26	73%	2,8%
Fresh fish, Charcuterie	Cheese	2,25	69%	2,8%
Fresh fish, Cheese	Charcuterie	2,25	57%	2,8%
Fresh fish, Vegetables	Vegetables spec.	2,09	43%	2,8%
Fresh fish, Fruits	Vegetables	2,06	67%	4,7%
Fresh fish, Milk & Cream	Vegetables	2,03	66%	2,6%
Fresh fish, Basic ingredients	Vegetables	2,01	65%	2,5%
Fresh fish, Yoghurts & Desserts	Vegetables	2,00	65%	2,8%
Fresh fish, Charcutarie	Vegetables	1,98	64%	2,6%

With the criteria defined above, 46 associations involving fresh fish were obtained. In the previous Table 5.7, the 10 results with the highest lift are presented in descending order. In this first analysis, we observe, as expected, that a large number of rules include the most purchased products, such as milk or vegetables. In this sense, in order to understand in greater depth and detail the products that are purchased together with fresh fish, the shopping cart was analyzed at the subcategory level.

For this case, the subcategories were considered associated if they met the following criteria: lift greater than 1, confidence greater than or equal to 30% and support greater than or equal to 0.25%. This resulted in the identification of 69 association rules involving fresh fish.

Some rules contain products of the same category, such as rice and pasta. However, some relationships between products of different sections were also identified. For purposes of illustration, the following Table 5.8 shows the top ten product associations, when ordered by lift value.

Table 5.8: Association rules at Subcategory level.

Antecedent (x)	**Consequent (x)**	Lift	Conf.	Supp (x,y)
Sea bass	Gilt head bream	10,58	31%	0,5%
Sardine	Salad Vegetables	5,76	31%	0,28%
Gilt head bream, Pasta	Rice	5,46	53%	0,3%
Salmon, Soup vegetable	Cabbage	5,42	39%	0,3%
Gilt head bream, Soup vegetables	Cabbage	5,24	38%	0,3%
Gilt head bream, Rice	Pasta	5,16	54%	0,3%
Salmon, Pasta	Rice	5,15	50%	0,26%
Salmon, Rice	Pasta	5,00	52%	0,26%
Gilt head bream, Cabbage	Soup vegetables	4,53	61%	0,3%
Salmon, Cabbage	Soup vegetables	4,45	59%	0,3%

From the results of the previous Table 5.8 can be taken some ideas for future marketing actions. In this case, the company may for example provide a discount on the product resulting from the membership rule, if it has not recently been purchased by the consumer who purchased the corresponding previous product. This procedure draws customers' attention to a product of their

interest/preference. For example, for this particular case, the company may suggest a Golden Gift/discount voucher to customers who have bought Sea bass and have not bought Gilt head bream for more than one month. This type of action can lead customers to go to the store more often, but also increase the diversity of products purchased by customers.

5.6 Conclusion

The results obtained in this chapter confirm the importance that the DM can and should have in understanding customers and helping companies make decisions with large databases. In fact, using the transactions carried out with the business loyalty card in the year 2018, the DM allowed to obtain knowledge about fresh fish customers.

To analyze consumer behavior, the segmentation was based on 5 criteria, such as Monthly Frequency, Frequency Ratio, Value Spent, Value Spent Ratio, Promotions Transactions Ratio. Using a partitioning cluster analysis technique, customers were grouped into three clusters according to their shopping habits. The analysis also involved the construction of a decision tree in order to extract the rules underlying customer segmentation. Following this procedure, it was possible to draw a profile for each segment that can be used for customers classification with high precision.

In addition, consumption at species and category level was also explored. With this study and the results obtained, it becomes easier to meet and satisfy customer preferences.

Finally, significant product association rules were also identified, taking into account customers' market baskets. These rules allow the creation of differentiated promotions, which can be crucial to motivate customers to consume and to remain loyal to the company.

Chapter 6

Conclusion

This chapter provides not only a summary of the study developed, but also the conclusions that can be drawn. In addition, it identifies the main contributions of this dissertation to the ValorMar project and to the company being studied. Finally, it presents the direction for future research in this project.

6.1 Summary and conclusions

With the increase in competitiveness in the retail sector, it is increasingly important to know the consumer in order to meet their preferences. The development of methods to understand customers' consumption habits allows companies to increase their sales volume. The ValorMar project seeks to understand the value chain of fresh fish. As part of this project, this dissertation aims to carry out an in-depth study of the consumer behaviour of fresh fish.

In the course of that idea, this dissertation began by trying to understand the main factors that could influence decision making in the purchase of fresh fish. In fact, the consumer of fresh fish presents differentiating factors that lead them to make decisions when buying a product. The variables quality, price and value of the product are factors to take into account in the decision making of a fresh fish customer. In this sense, the retail company must focus on these variables in order to understand consumer behaviour.

Considering that this dissertation seeks to understand consumer behaviour using data analysis, the knowledge discovery in database process was presented and the main focus was given to DM. For this reason, this dissertation describes and uses DM tool such as association, classification, clustering and visualization. For each DM tool used, the techniques used in this dissertation are described, as well as their advantages and disadvantages. For the choice of the most appropriate techniques, the main criterion was the processing speed.

As a basis for this work, the transactions of the customers of a large retail company were explored. In order to determine and understand the behaviour and consumption habits of fresh fish customers, a number of analyses were carried out: (i) segmentation based on consumer behaviour,

(ii) segmentation based on relative consumption of species, (iii) segmentation based on relative consumption of categories and (iv) analysis of the shopping cart.

In short, the major contributions of the thesis are summarized as follows:

- the application of different DM techniques to large datasets obtained from the use of a loyalty card;
- the identification of market segments based on customers purchasing behavior inferred from 5 criteria and socio-demographic characterization of each cluster resulting from segmentation;
- obtaining market segments based on relative consumption of species and based on relative consumption of categories;
- the design of differentiated marketing promotions based on market basket analysis .

In a final assessment of the results obtained, it can be concluded that the expected objectives have been achieved. In this way, a deep and detailed analysis of transactional data will allow better decisions to be made, which will lead to better results for the company.

6.2 Directions for future research

Being aware of consumer preferences allows improving customers' shopping experience while optimizing the value chain as a whole. However, since the analysis of the fresh fish value chain is incipient, there is still a long way to go. In fact, there is a set of aspects and analyses that should and can be complementary developed.

One of the analyses that can enrich customer knowledge is the comparison of behaviour and consumption between different periods (e.g. weekdays and weekends). This study may show differences that may allow differentiated promotions and marketing actions.

Another analysis that can add value to this project is the consumption forecast for each of the clusters obtained and for each of the Insignias. Sales forecasting is essential in the fresh fish sector, not only because of the short shelf life of this product category, but also because of the need to control stock levels in order to avoid excessive stock costs and simultaneously the loss of customers due to stock failures.

It may also be interesting to include the geographical question of the stores, i.e., to verify how the location of the store can condition the consumer's behavior and its analysis. This refers to another question, concerning the area in which the stores are within their area of influence, i.e. how purchasing power and consumption preferences may vary from area to area.

Finally, and although it would require substantial computational resources, it would be interesting to extend this study to all the company's stores. In this way it is possible to have a deep global knowledge and the result of customer segmentation can be richer in knowledge.

The study carried out in this dissertation can be applied to other areas of business, such as finance and telecommunications.

Appendix A

Additional Results

A.1 Insignia Analysis

Another concern of the company was to see if there are different types of customers and behaviors through the Insignia where consumption occurred. In this sense, this dissertation sought to explore this problem and, summarily, will present the results obtained. The process was exactly the same as it was previously done for the overall analysis. It should be remembered that 3 stores have been chosen (see Table 4.1) whose data set has already been described (see Section 4.3).

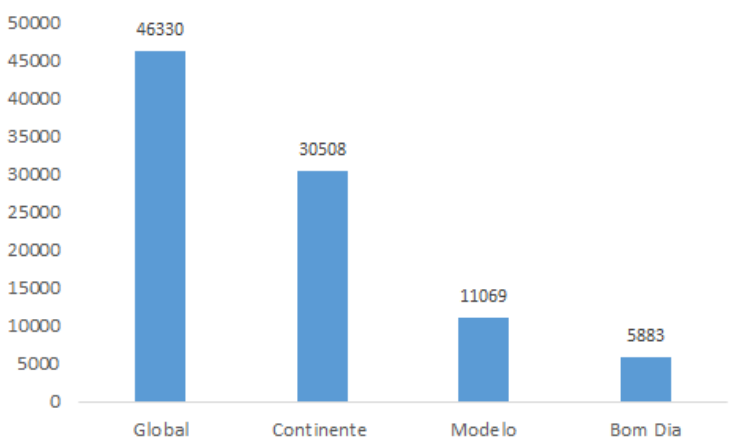

Figure A.1: Customers in each Insignia.

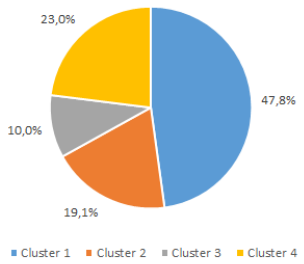

Figure A.2: Cluster distribution Continente.

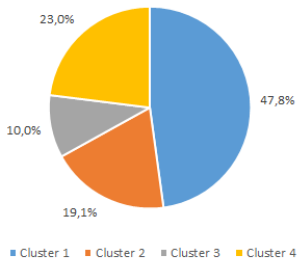

Figure A.3: Cluster distribution Modelo.

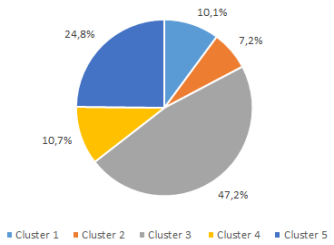

Figure A.4: Cluster distribution Bom Dia.

Table A.1: Average global number and average number per cluster, of each customer's variables Continente

	Monthly Frequency	Frequency Ratio	Value Spent	Value Spent Ratio	Promotions Transactions Ratio
Global	0,30	0,24	8,17	0,06	0,68
Cluster 1	0,22	0,18	**23,57**	0,09	0,59
Cluster 2	0,19	**0,86**	8,18	**0,26**	0,71
Cluster 3	0,20	**0,15**	7,35	0,04	**0,19**
Cluster 4	0,36	0,18	7,08	0,04	0,90

Table A.2: Average global number and average number per cluster, of each customer's variables Modelo

	Monthly Frequency	Frequency Ratio	Value Spent	Value Spent Ratio	Promotions Transactions Ratio
Global	0,37	0,26	7,19	0,07	0,72
Cluster 1	0,28	0,17	6,90	0,04	**0,91**
Cluster 2	0,23	0,16	**7,64**	0,04	0,21
Cluster 3	0,18	**0,87**	**7,97**	**0,30**	0,75
Cluster 4	1,94	0,35	6,81	0,09	0,74

Table A.3: Average global number and average number per cluster, of each customer's variables Bom Dia

	Monthly Frequency	Frequency Ratio	Value Spent	Value Spent Ratio	Promotions Transactions Ratio
Global	0,43	0,25	4,85	0,09	0,69
Cluster 1	0,23	**0,85**	4,99	**0,38**	0,67
Cluster 2	**2,33**	0,36	4,10	0,12	0,69
Cluster 3	0,31	0,16	3,81	0,05	**0,91**
Cluster 4	0,26	0,17	**11,05**	0,10	0,82
Cluster 5	0,27	0,17	4,33	0,05	**0,22**

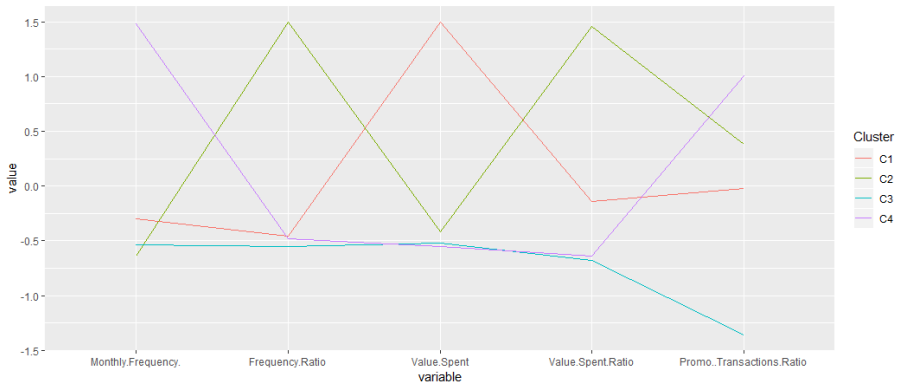

Figure A.5: Coordinate plot for categorical results from Continente Store

Figure A.6: Coordinate plot for categorical results from Modelo Store

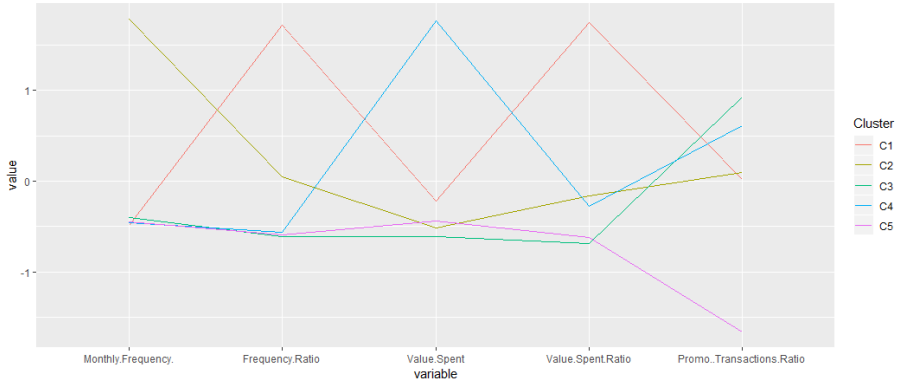

Figure A.7: Coordinate chart for categorical results from Bom Dia Store

References

[1] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

[2] Pang-Ning Tan. *Introduction to data mining*. Pearson Education India, 2018.

[3] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[4] Mark A Friedl and Carla E Brodley. Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment*, 61(3):399–409, 1997.

[5] Barreto Dias. Análise dos riscos na cadeia alimentar. *Evolução Europeia E Nacional. Segurança E Qualidade Alimentar*, 1:16–18, 2006.

[6] Banco BPI. A economia do mar em portugal: a estratégia e a realidade, num retrato doméstico e comunitário. URL: https://www.bancobpi.pt/content/conn/UCM/uuid/dDocName:PR_WCS01_UCM01011175 [last accessed 19.06.2019].

[7] Fishery and aquaculture country profiles. URL: http://www.fao.org/fishery/facp/PRT/en [last accessed 19.06.2019].

[8] Cameron R Peterson and Lee Roy Beach. Man as an intuitive statistician. *Psychological bulletin*, 68(1):29, 1967.

[9] Willard R Bishop Jr. Competitive intelligence. *Progressive Grocer*, 63(3):19–20, 1984.

[10] Jacob Jacoby and Jerry C Olson. Perceived quality. *Lexington, MA: Lexington Books*, 1985.

[11] Alan G Sawyer and Peter R Dickson. Psychological perspectives on consumer response to sales promotion. *Research on sales promotion: Collected papers*, pages 1–21, 1984.

[12] Valarie A Zeithaml. Consumer perceptions of price, quality, and value: a means-end model and synthesis of evidence. *The Journal of marketing*, pages 2–22, 1988.

[13] David Garvin. Quality on the line. *Harv. Bus. Rev.*, pages 65–75, 1983.

[14] Massimo Bertolini, Maurizio Bevilacqua, and Roberto Massini. Fmeca approach to product traceability in the food industry. *Food control*, 17(2):137–145, 2006.

[15] ISO 9000, definition of the term traceability: Iso 9000. https://www.keyence.com/ss/products/marking/traceability/law_basic.jsp. Accessed: 2019-01-29.

[16] Parlamento Europeu. Conselho. regulamento (ce) n. º 178/2002. *Jornal Oficial da União Europeia.(2002-02-28) Determina os princípios e normas gerais da legislação alimentar, cria a Autoridade Europeia para a Segurança dos Alimentos e estabelece procedimentos em matéria de segurança dos géneros alimentícios*, 2002.

[17] VM Moretti, GM Turchini, F Bellagamba, and F Caprino. Traceability issues in fishery and aquaculture products. *Veterinary research communications*, 27(1):497–505, 2003.

[18] Carmen Georgeta Nicolae, Nicoleta IȘFAN, Gratziela Victoria Bahaciu, Monica Paula Marin, and Liliana Mihaela Moga. Case study in traceability and consumer's choices on fish and fishery products. *Age*, 3:8–80, 2016.

[19] Erika Da Silva Maciel, Luciana Kimie Savay-da Silva, Julia Santos Vasconcelos, Jaqueline Girnos Sonati, Juliana Antunes Galvão, Leandro Kanamaru Franco De Lima, and Marília Oetterer. Relationship between the price of fish and its quality attributes: a study within a community at the university of são paulo, brazil. *Food Science and Technology*, 33(3):451–456, 2013.

[20] Torbjørn Trondsen, Joachim Scholderer, Eiliv Lund, and Anne E Eggen. Perceived barriers to consumption of fish among norwegian women. *Appetite*, 41(3):301–314, 2003.

[21] Jill E Hobbs et al. Traceability in meat supply chains. *CAFRI: Current Agriculture, Food and Resource Issues*, (04), 2003.

[22] Canada. Agriculture and Agri-Food Canada. *Costs of Traceability in Canada: Developing a Measurement Model*. Agriculture and Agri-Food Canada, 2007.

[23] Feng Wang, Zetian Fu, Weisong Mu, Liliana M Moga, Xiaoshuan Zhang, et al. Adoption of traceability system in chinese fishery process enterprises: Difficulties, incentives and performance. *Journal of Food, Agriculture and Environment*, 7(2):64–69, 2009.

[24] Nga Mai, Sigurdur Gretar Bogason, Sigurjon Arason, Sveinn Víkingur Árnason, and Thórólfur Geir Matthíasson. Benefits of traceability in fish supply chains–case studies. *British Food Journal*, 112(9):976–1002, 2010.

[25] Wim Verbeke and Isabelle Vackier. Individual determinants of fish consumption: application of the theory of planned behaviour. *Appetite*, 44(1):67–82, 2005.

[26] Olli T Ahtola. Price as a'give'component in an exchange theoretic multicomponent model. *ACR North American Advances*, 1984.

[27] Joseph David Chapman. *The impact of discounts on subjective product evaluations*. PhD thesis, Virginia Polytechnic Institute and State University, 1987.

[28] Tridik Mazumdar. Experimental investigation of the psychological determinants of buyers' price awareness and a comparative assessment of methodologies for retrieving price information from memory. *Virginia Polytechnic Institute and State University*, 1986.

[29] Kent B Monroe and Ram Krishnan. The effect of price on subjective product evaluations. *Perceived quality*, 1(1):209–2, 1985.

[30] Jacob Jacoby and Jerry C Olson. nconsumer response to price: An attitudinal, information processing perspective, oin moving ahead with attitude research, y. *Wind and P. Greenberg, eds. Chicago: American Marketing Association*, 73:86, 1977.

[31] Joseph N Uhl and Harold L Brown. Consumer perception of experimental retail food price changes. *Journal of Consumer Affairs*, 5(2):174–185, 1971.

[32] Philip Kotler. Marketing management, millennium ed. *Prentic Hall*, 2000.

[33] Peter R Dickson and Alan G Sawyer. *Point-of-purchase behavior and price perceptions of supermarket shoppers*. Marketing Science Institute, 1986.

[34] FAO. Rome: Food and agriculture organization of the united nations. *The state of world fisheries and aquaculture. Opportunities and challenges.*, 2014.

[35] Ann L Yaktine, Malden C Nesheim, et al. *Seafood choices: balancing benefits and risks*. National Academies Press, 2007.

[36] Domenico Carlucci, Giuseppe Nocella, Biagia De Devitiis, Rosaria Viscecchia, Francesco Bimbo, and Gianluca Nardone. Consumer purchasing behaviour towards fish and seafood products. patterns and insights from a sample of international studies. *Appetite*, 84:212–227, 2015.

[37] Themistoklis Altintzoglou, Filiep Vanhonacker, Wim Verbeke, and Joop Luten. Association of health involvement and attitudes towards eating fish on farmed and wild fish consumption in belgium, norway and spain. *Aquaculture International*, 19(3):475–488, 2011.

[38] Filiep Vanhonacker, Themistoklis Altintzoglou, Joop Luten, and Wim Verbeke. Does fish origin matter to european consumers? insights from a consumer survey in belgium, norway and spain. *British Food Journal*, 113(4):535–549, 2011.

[39] Dawn Birch and M Lawley. Buying seafood: Understanding barriers to purchase across consumption segments. *Food quality and preference*, 26(1):12–21, 2012.

[40] Wim Verbeke, Isabelle Sioen, Karen Brunsø, Stefaan De Henauw, and John Van Camp. Consumer perception versus scientific evidence of farmed and wild fish: exploratory insights from belgium. *Aquaculture International*, 15(2):121–136, 2007.

[41] Joanna Burger and Michael Gochfeld. Perceptions of the risks and benefits of fish consumption: Individual choices to reduce risk and increase health benefits. *Environmental research*, 109(3):343–349, 2009.

[42] Karen Brunsø, Wim Verbeke, Svein Ottar Olsen, and Lisbeth Fruensgaard Jeppesen. Motives, barriers and quality evaluation in fish consumption situations: Exploring and comparing heavy and light users in spain and belgium. *British Food Journal*, 111(7):699–716, 2009.

[43] Marta Cosmina, Eugenio Demartini, Anna Gaviglio, Christine Mauracher, Sonia Prestamburgo, and Giovanna Trevisan. Italian consumers' attitudes towards small pelagic fish. *New Medit*, 11(1):52–57, 2012.

[44] Jessica A Grieger, Michelle Miller, and Lynne Cobiac. Knowledge and barriers relating to fish consumption in older australians. *Appetite*, 59(2):456–463, 2012.

[45] Alexandra McManus, W Hunt, Janet Howieson, Beatriz Cuesta-Briand, Jennifer McManus, and Jessica Storey. Attitudes towards seafood and patterns of consumption in an australian coastal town. *Nutrition Bulletin*, 37(3):224–231, 2012.

[46] Dorothée Brécard, Boubaker Hlaimi, Sterenn Lucas, Yves Perraudeau, and Frédéric Salladarré. Determinants of demand for green products: An application to eco-label demand for fish in europe. *Ecological economics*, 69(1):115–125, 2009.

[47] Anna Claret, Luis Guerrero, Enaitz Aguirre, Laura Rincón, Mª Dolores Hernández, Inmaculada Martínez, José Benito Peleteiro, Amàlia Grau, and Carmen Rodríguez-Rodríguez. Consumer preferences for sea fish using conjoint analysis: Exploratory study of the importance of country of origin, obtaining method, storage conditions and purchasing price. *Food Quality and Preference*, 26(2):259–266, 2012.

[48] Shabbar Jaffry, Helen Pickering, Yaseen Ghulam, David Whitmarsh, and Prem Wattage. Consumer choices for quality and sustainability labelled seafood products in the uk. *Food Policy*, 29(3):215–228, 2004.

[49] IS Arvanitoyannis, A Krystallis, P Panagiotaki, and AJ Theodorou. A marketing survey on greek consumers' attitudes towards fish. *Aquaculture International*, 12(3):259–279, 2004.

[50] Carlos Cardoso, Helena Lourenço, Sara Costa, Susana Gonçalves, and Maria Leonor Nunes. Survey into the seafood consumption preferences and patterns in the portuguese population. gender and regional variability. *Appetite*, 64:20–31, 2013.

[51] Troy E Hall and Shannon M Amberg. Factors influencing consumption of farmed seafood products in the pacific northwest. *Appetite*, 66:1–9, 2013.

[52] Adriaan PW Kole, Themistoklis Altintzoglou, Rian AAM Schelvis-Smit, and Joop B Luten. The effects of different types of product information on the consumer product evaluation for fresh cod in real life settings. *Food Quality and Preference*, 20(3):187–194, 2009.

[53] Kolbrún Sveinsdóttir, Emilía Martinsdóttir, Ditte Green-Petersen, Grethe Hyldig, Rian Schelvis, and Conor Delahunty. Sensory characteristics of different cod products related to consumer preferences and attitudes. *Food Quality and Preference*, 20(2):120–132, 2009.

[54] Michael J Berry and Gordon Linoff. *Data mining techniques: for marketing, sales, and customer support*. John Wiley & Sons, Inc., 1997.

[55] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[56] Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Morgan Kaufmann, 2012.

[57] David J Hand. Data mining. *Encyclopedia of Environmetrics*, 2, 2006.

[58] João Branco. *Uma Introdução à Análise de Clusters*. Sociedade Portuguesa de Estatística, 2004.

[59] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[60] Edward W Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, 21:768–769, 1965.

[61] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[62] Osvaldo Gervasi, Marina L Gavrilova, Vipin Kumar, Antonio Laganà, Heow Pueh Lee, Youngsong Mun, David Taniar, and Chih Jeng Kenneth Tan. Computational science and its applications–iccsa 2005. In *Conference proceedings ICCSA*, page 112. Springer, 2005.

[63] Chih-Ping Wei, Yen-Hsien Lee, and Che-Ming Hsu. Empirical comparison of fast partitioning-based clustering algorithms for large data sets. *Expert Systems with applications*, 24(4):351–363, 2003.

[64] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM, 1993.

[65] Yen-Liang Chen, Chang-Ling Hsu, and Shih-Chieh Chou. Constructing a multi-valued and multi-labeled decision tree. *Expert Systems with Applications*, 25(2):199–209, 2003.

[66] Syed Riaz Ahmed. Applications of data mining in retail business. In *International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004.*, volume 2, pages 455–459. IEEE, 2004.

[67] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[68] Eric WT Ngai, Li Xiu, and Dorothy CK Chau. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, 36(2):2592–2602, 2009.

[69] Raymond Roel. Direct marketing's 50 big ideas. *Direct Marketing*, 50(May):45–62, 1988.

[70] Duen-Ren Liu and Ya-Yueh Shih. Integrating ahp and data mining for product recommendation based on customer lifetime value. *Information & Management*, 42(3):387–400, 2005.

[71] Wagner A Kamakura, Sridhar N Ramaswami, and Rajendra K Srivastava. Applying latent trait analysis in the evaluation of prospects for cross-selling of financial services. *international Journal of Research in Marketing*, 8(4):329–349, 1991.

[72] David C Schmittlein and Robert A Peterson. Customer base analysis: An industrial purchase process application. *Marketing Science*, 13(1):41–67, 1994.