UNIVERSIDADE CATÓLICA PORTUGUESA

Segmentação e definição do perfil do Cliente do Continente *Online*

www.continente.pt

Trabalho Final na modalidade de Relatório de Estágio
apresentado à Universidade Católica Portuguesa
para obtenção do grau de mestre em Business Economics

por

Vasco Teixeira Esteves da Silva Pereira

sob orientação de
Professor Mário Ferreira

Faculdade de Economia e Gestão da Universidade Católica Portuguesa
Fevereiro de 2015

Agradecimentos

Quero, em primeiro lugar, agradecer ao Professor Doutor Mário Ferreira por todo o acompanhamento e orientação prestados durante a elaboração deste trabalho.

Quero também agradecer ao Professor Ricardo Ribeiro e a toda a equipa de Business Intelligence da Modelo.com, em especial ao Ricardo Alves, ao Tiago Ventura, ao Tiago Branco e à Paula Brochado.

Agradeço aos meus pais, irmãos e amigos, todo o apoio dado, durante este período e em especial, à Sofia Álvares Ribeiro, pela companhia e ajuda, durante a elaboração deste trabalho.

Sumário Executivo

O objetivo deste trabalho é a segmentação e a definição do perfil do cliente do Continente Online. A pesquisa para este trabalho incluiu uma revisão de literatura sobre o E-Commerce, sobre a segmentação, sobre a definição de perfis e sobre o Data Mining.

Os principais resultados indicam que, a segmentação e a definição de perfis com base em atributos comportamentais, permitem obter um conhecimento detalhado sobre os hábitos de consumo dos clientes.

A definição dos perfis dos clientes é importante para o Continente Online e recomenda-se que sejam elaborados novos estudos, utilizando simultaneamente atributos fatuais e comportamentais.

A estrutura deste sumário executivo é explicada na tabela seguinte:

O objetivo deste trabalho é a segmentação e a definição do perfil do cliente do Continente Online.	Objetivo
A pesquisa para este trabalho incluiu uma revisão de literatura sobre o E-Commerce, sobre a segmentação, sobre a definição de perfis e sobre o Data Mining.	Método
Os principais resultados indicam que, a segmentação e a definição de perfis com base em atributos comportamentais, permitem obter um conhecimento detalhado sobre os hábitos de consumo dos clientes.	Resultados
A definição dos perfis dos clientes é importante para o Continente Online e recomenda-se que sejam elaborados novos estudos, utilizando simultaneamente atributos fatuais e comportamentais.	Conclusões

Índice

Índice de Figuras

Índice de Tabelas

1. Introdução

1.1. Introdução

O presente Relatório foi desenvolvido no decurso de um estágio na empresa Modelo.com da Sonae MC, no período de 15 de Setembro de 2014 a 23 de Janeiro de 2015, no contexto de Trabalho Final de Mestrado (TFM) do Master in Business Economics.

A Modelo.com, também denominada por Continente *Online* (COL), é uma empresa pertencente à Sonae MC, que se dedica à comercialização de produtos, através duma plataforma *online*, ou seja, é a empresa responsável pelo *e-commerce* da Sonae MC.

O *e-commerce* pode ser definido, de entre muitas definições, como a compra e venda de informação, produtos ou serviços através da Internet (Kalakota & Whinston, 1997:3).

Em Portugal, assistiu-se nos últimos anos a um crescimento do *e-commerce*, com o volume de negócio a crescer 52% entre 2009 e 2012 e com uma expectativa de crescimento de 70% entre 2014 e 2017. O número de utilizadores de Internet também aumentou nos últimos anos, o que potenciou o crescimento do *e-commerce* (ACEPI, 2013).

Especificamente o *e-commerce* dos produtos do Retalho Geral ou, na terminologia anglo-saxónica do *Fast Moving Consumer Goods* (FMCG), representa ainda uma pequena percentagem do total transacionado neste mercado, cerca de 0,9% (Kantar Worldpanel, 2014). Apesar deste mercado apresentar uma cota baixa, as três maiores empresas a operar em Portugal apostam no seu crescimento e acreditam no seu potencial. Para 2016 é expectável que esta cota atinja os 1,4%, ficando ainda muito abaixo do *benchamark* de referência, o Reino Unido, onde este mercado solidifica uma cota de 5%.

Para uma empresa que venda produtos ou serviços é imprescindível conhecer o cliente e atender às suas necessidades e valores (Pickton and Broderick, 2005: 373). Também no setor do *e-commerce* deve-se ter isto em conta (Nielsen, 2014), daí que a Modelo.com tenha identificado que existia a necessidade de aprofundar o conhecimento para estar mais próxima dos seus clientes.

Uma das formas de conhecer melhor os clientes é através da sua segmentação, ou seja, através da identificação de subdivisões ou segmentos de clientes em grupos distintos que possam representar valor ou potencial para a empresa (Kotler, 1994).

A análise *à posteriori* destes segmentos permite identificar perfis de clientes (Jansen, 2007) e ajustar as estratégias de *marketing* de uma forma mais eficiente, nomeadamente através de diferentes técnicas de *Marketing Mix* (Armstrong & Kotler, 2005: 54).

O objetivo da construção de perfis é a possibilidade de prever o comportamento dos clientes com base na informação de cada um (Verhoef, Spring, Hoekstra, Lee, 2002).

O desafio lançado pela Modelo.com, ao qual este relatório pretende responder, foi o seguinte:

"**Segmentação do Cliente COL - Como definir / traçar o perfil do cliente *e-commerce*?** "

Para responder ao que foi proposto, foram explorados dois métodos de segmentação e a partir dos segmentos encontrados foram gerados dois perfis, recorrendo a dois tipos de atributos, fatuais (Beane, Ennis, 1987), (Lilien, Rangaswamy, 2003) e comportamentais (Yankelovich, Meer, 2006).

A resposta ao desafio culminou com a identificação do perfil ou dos perfis que permitem identificar os clientes com maior importância para a empresa, procurar prever o seu comportamento e potenciar eventuais campanhas de Marketing (Leung, 2009) a realizar pelo Modelo.com.

1.2. Motivação e objetivos

A compra de produtos e serviços, com recurso a uma plataforma *online* é, nos dias de hoje, uma realidade em todo o mundo representando uma importante fatia das transações do comércio global.

Acompanhando a evolução das tecnologias de informação, o *e-commerce* tem crescido possibilitando a empresas como a Amazon, Alibaba, eBay e Rakuten, se tornarem gigantes no comércio global.

No entanto, no setor dos FMGC o *e-commerce* tardou a atingir a mesma expressividade, muito por causa da tipologia de produtos transacionados e das dificuldades acrescidas na logística associada à sua entrega ao consumidor. Contudo, em países como os Estados Unidos da América, Reino Unido e Alemanha, o comércio *online* destes bens é já uma realidade com uma expressão considerável. Nestes países surgiram nomes de referência como a Amazon, a Tesco, o Walmart e o Allyouneed.

No caso da Walmart e Tesco, grandes *retailers* tradicionais, que adicionaram modelos de *e-commerce*. Outras empresas apresentam um percurso genuinamente *online*, como é o caso da Amazon e Allyouneed.

Em Portugal, o *e-commerce* no FMGC, é dominado por três grandes retalhistas tradicionais, o Continente, o Jumbo e o Corte-Inglês. Neste setor só o Continente possuiu cerca de 80% da cota de mercado (Fonte da Modelo.com).

A Modelo.com conta já com 15 anos de atividade oferecendo uma vasta gama de produtos, cerca de 30.000. Na sua plataforma estão à disposição

grande parte dos bens transacionados nos supermercados e hipermercados da Sonae MC, bem como campanhas temporárias, nomeadamente, a dos manuais escolares e produtos de farmácia das lojas Well's.

Neste trabalho, apenas as categorias de FMCG, foram alvo de estudo, já que são as que apresentam maior expressividade a nível de vendas.

A empresa fatura em média 100.000 € por dia, tendo cerca de 1.000 clientes a efetuarem compras diariamente (Fonte da Modelo.com). No ano de 2013, a empresa contava com 274 funcionários e obteve um lucro de 1.270.297€ (Fonte da base de dados SABI).

Tendo em conta a quantidade de transações (cerca de 30.000 artigos vendidos por dia) e o número de clientes, foi um desafio analisar toda a informação disponível sobre os clientes e transforma-la em conhecimento útil que apoie a tomada de decisão.

Abraçar este desafio foi por si só um fator motivante sabendo que esta é a maior empresa de *e-commerce* no setor do Retalho Geral no País (Fonte da Modelo.com).

1.3. Originalidade e contribuição para o conhecimento

Acreditamos que o tema é original pois permite gerar um novo conhecimento sobre os diferentes tipos de clientes que utilizam o Continente *Online* e auxiliar na tomada decisão dos *Marketers* e dos Gestores da Organização.

Na construção deste relatório foram tidos em conta os vários estudos sobre a segmentação e sobre o perfil do cliente do Continente realizados pela Modelo.com nomeadamente o "O nosso cliente em 7 segmentos" e o "Segmentação Valor COL".

No estudo "O nosso cliente em 7 segmentos" utilizaram-se atributos fatuais (idade, género e outros) e comportamentais que serviram para a construção da segmentação. Recorreu-se ao *Data Mining* e a algoritmos de *Clustering* para encontrar os segmentos e, posteriormente, foram construídos 7 perfis de clientes. Contudo, este estudo não foi direcionado apenas para o cliente do *e-commerce* pelo que não permitiu identificar qual o seu perfil e, consequentemente prever o seu comportamento.

O segundo estudo, já se focou no cliente do *e-commerce*, ou seja, no cliente do Continente *Online*. Os atributos utilizados foram comportamentais, mais propriamente, o gasto médio por visita por cliente (*Monetary*), o número de visitas por cliente (*Frequency*) e o número de dias desde a última compra (*Recency*), sendo este os três atributos, que compõem a RFM. Neste trabalho, também se recorreu ao *Data Mining* e a algoritmos de *Clustering* para realizar a segmentação sendo utilizada a corrente *Data Driven*. Através do algoritmo *k-means*, também utilizado no trabalho anterior, foram definidos quatro *Clusters* que posteriormente permitiram a definição dos perfis com base nos três atributos selecionados.

A originalidade deste trabalho, acreditamos, provêm da metodologia utilizada, bem como das fontes de dados e do tipo de atributos selecionados.

Ao contrário dos dois estudos anteriores, neste trabalho procurou-se encontrar a melhor forma de segmentar os clientes e de definir o seu perfil, comparando os perfis resultantes dos dois tipos de segmentação, a segmentação *Market Driven* e a segmentação *Data Driven*.

Através da segmentação e da posterior definição dos perfis dos clientes acreditamos será possível melhorar o relacionamento com o cliente e criar uma

comunicação mais eficiente, que será reconhecida pelo cliente, podendo resultar na sua fidelização (Banon, 2004), (Smith, 1956) e (Amat, 2002).

Os *Marketers* e outros intervenientes da Modelo.com poderão tomar melhores decisões sabendo que vão ao encontro das necessidades dos seus clientes (Chen, 2007), (Dickson, 1982), bem como prever o seu comportamento e criar ações de *Marketing* direcionadas para os perfis corretos (Jansen, 2007).

Além disso, vai ser possível identificar os consumidores com maior importância para a empresa e criar um sistema de sugestão de produtos, para clientes que partilham determinados atributos (Leung, 2009).

2. Revisão de Literatura

2.1. Definição dos Conceitos Chave

2.1.1. e-commerce

O *e-commerce* ou comércio eletrónico abrange todo o tipo de negócio ou atividade que tenha em vista a transação de produtos, serviços ou informação através duma plataforma eletrónica (Rosen, 2000), como é o caso do modelo de negócio da Modelo.com.

Pode existir uma ideia errada sobre a dimensão do *e-commerce*, já que é vulgarmente definido como o ato de comprar ou vender através a Internet. Mas a realidade é que se estende para o universo das transações de informação e para qualquer transação, que envolva a passagem de direitos de propriedade via Internet (Andam, 2003).

2.1.2. Segmentação

A segmentação é o ato de dividir em partes ou grupos menores, denominados por segmentos. Entender o conceito da segmentação, foi fulcral para a realização deste relatório, já que, sem a segmentação não seria possível a construção dos perfis (Jansen, 2007).

Quando aplicada a um conceito de mercado à segmentação pode ser definida como o processo de dividir um grupo heterógeno de clientes em subgrupos homogéneos ou segmentos que contêm clientes que partilham atributos comuns, necessidades, hábitos ou gostos (Giha, Singh, Ewe, 2003).

2.1.3. Perfil

Apesar da denominação "perfil" ser utilizada em vários contextos no dia-a-dia, é importante explicar a sua natureza no contexto de mercado. O perfil pode ser descrito como um determinado padrão ou uma representação de um conjunto de clientes que partilham um conjunto de atributos ou características que os levam a agir de uma maneira semelhante (Ahola, Runsala, 2001).

2.1.4. Data Mining

Data Mining pode-se definir como um conjunto de ferramentas ou processos de exploração e análise dados, com o objetivo específico de descobrir padrões e regras que sejam importantes e significativos, convertendo a informação em bruto em conhecimento auxiliando a tomada de decisão (Road, Jose & Getta, 2003).

Também é possível definir como um conjunto de ferramentas estatísticas, que permitem analisar os dados e aplicar algoritmos de descoberta recorrendo,

para isso, a mecanismos computacionais com a finalidade de produzir modelos de previsão (Ruxandra Petre, 2013).

2.1.5 *Clustering*

É uma classe de técnicas estatísticas que, quando aplicadas a dados que exibam heterogeneidade, permitem identificar e analisar segmentos ou *clusters* que contêm observações e partilham similaridades (Venkatesan, 2007).

2.2. Enquadramento e contexto teórico para a investigação

2.2.1. O *e-commerce* no Mundo

O *e-commerce* é, nos dias de hoje, uma realidade importante demonstrando sinais de um crescimento contínuo e sustentável, atraindo novas empresas e penetrando em novos mercados.

Não existe uma única definição para o *e-commerce* (Golubova, 2012), contudo para a introdução ao tema, aceitou-se a seguinte definição (Kalakota, Whinston, 1997): o *e-commerce* é a compra e venda de produtos, serviços e informação, através da Internet.

Um estudo realizado (eMarketer, 2014), demonstra que até ao 3º trimestre do ano de 2014, existiam 1.21 biliões de compradores *online* em todo o mundo, estimando-se um volume de negócios de cerca de 1.455 triliões de dólares em vendas no *e-commerce* apenas na componente B2C (Business to Consumer).

Estes números representam um aumento de 19.2% comparando com o ano de 2013. Prevê-se (eMarketer, 2014) que em 2018 as vendas no *e-commerce* atinjam os 2.356 triliões de dólares em todo o mundo.

No relatório anual, "*e-commerce: Evolution or Revolution in the Fast Moving Consumer Goods*" (Nielsen, 2014), foram levantadas algumas tendências dos últimos anos que demonstram como o comportamento de compra *online* evolui favorecendo assim o crescimento do *e-commerce*. Nos últimos três anos, a intenção de compra *online* duplicou (Nielsen, 2014). Este facto associado a uma maior penetração da Internet levou a um aumento das compras. Tanto o *Online Browsing* como o *Online Buying* aumentaram. Esta mudança deve-se ao aumento do uso de dispositivos móveis com acesso à Internet, cada vez mais comuns, já que representam uma alternativa mais económica, face aos computadores.

Com o aparecimento de aplicações de compra para os dispositivos móveis, que permitem comprar *online* e guardar as listas de compras, tornou-se ainda mais fácil e económico comprar *online*, incentivando ainda mais o crescimento do *e-commerce*.

Em "*Understanding how US online shoppers are reshaping the retail experience*" (Price Waterhouse Coopers, 2013) foram analisados os comportamentos de compra de 7005 consumidores em oito territórios, em 2011: Estados Unidos da América, China, Hong-Kong, Alemanha, França, Reino Unido, Holanda e Suíça.

Concluiu-se que mais de 40% da população da Europa Ocidental comprava produtos através do *e-commerce* com destaque para a Alemanha, a Suíça e a França, com 50% dos consumidores, a utilizarem plataformas *online* para realizarem as suas compras. As previsões para Europa, obtidas através deste estudo, indicam que a média Europeia para consumidores que compram *online* em 2015 será de 49% dos consumidores com países a destacarem-se como o Reino Unido e a Holanda, em que a previsão aponta para os 70%. Contudo, o seu crescimento não é idêntico em todas as categorias de produtos.

As categorias mais transacionáveis via *e-commerce* são os produtos duradouros e produtos ou serviços não consumíveis relacionados com o entretenimento (Nielsen, 2014). Os destaques identificados são a roupa, as

viagens de avião e as reservas de hotéis, representando a maior fatia. Num segundo nível encontram-se *e-books*, bilhetes de eventos, artigos de desporto e por fim brinquedos.

A intenção de compra *online* (*purchase intention rates*) quando elevada, é associada a uma relação de proximidade entre procurar e comprar *online*, como é o caso destas categorias, que apresentam taxas de *browsing/buying* entre 35% a 46%. Outra importante conclusão retirada deste estudo (Nielsen, 2014), é que, existem categorias que estão mais relacionados com o *online browsing* do que com o *online buying*. É o caso dos equipamentos eletrónicos, dos telemóveis, dos artigos informáticos, dos carros e das motas, onde o rácio entre a procura e a compra, é mais baixo, sendo a procura superior à compra em 7% (em média).

De todas as 22 categorias analisadas (Nielsen, 2014) apenas uma apresenta um rácio entre procura e compra superior a 1 - os bilhetes de avião e as reservas de hotéis, em que a procura equivale a 40% e a compra 48%.

2.2.2. O *e-commerce* e os FMCG

Passemos agora ao contexto do *e-commerce* nos produtos FMCG. Esta nomenclatura abrange os produtos que são vendidos a um ritmo mais elevado e consumidos numa base regular (Çelen, Erdogan, Taymaz, 2005), também vulgarmente chamados de produtos de consumo.

Dentro dos FMCG, podemos ter produtos perecíveis como alimentos frescos, nomeadamente, verduras, carne e peixe e produtos não perecíveis como detergentes, produtos de higiene e produtos farmacêuticos.

Apesar do comércio *online* dos FMCG ser, no *e-commerce*, relativamente pequeno, começa a demonstrar sinais de crescimento relevantes o que se deve à elevada frequência com que compramos estes produtos, comparativamente a

outras categorias, como artigos informáticos, artigos desportivos ou carros (Srinivasu, 2014).

Os produtos de consumo, apesar de acarretarem uma intenção de compra *online* mais baixa do que outros produtos, apresentam rácios de procura/compra próximos de 1, nomeadamente nos cosméticos, com 33% de procura e 31% de compra e no retalho alimentar, com 30% de procura e 27% de compra, ambos demonstrando uma forte correlação entre procura e compra (Nielsen, 2014).

Os FMCG vendidos através do *e-commerce* em 2013 representavam apenas 3,7% das vendas do *e-commerce* no mercado Global. (Kantar Worldpanel, 2014) defende que este número tenderá a aumentar, com a possibilidade de crescimento até 5% em 2016.

Em 2013, as vendas de produtos alimentares via *e-commerce* aumentaram cerca 31% em todo o Mundo face a 2012, com países como a França atingir um crescimento de 54% e a China com um crescimento de 45%.

A França e o Reino Unido, são os países Europeus onde o consumo de FMCG via *e-commerce* tem a maior quota de mercado. As vendas *online* representam 3.9% do mercado FMCG no caso da França e 4,7% no Reino Unido, a mais alta da Europa. Até 2016, prevê-se, que a França ultrapasse o Reino Unido (Kantar Worldpanel, 2014).

Outros indicadores de crescimentos foram analisados, nomeadamente o "*average penetration rate*", o "*spend per trip*" e a "*frequency*". O "*average penetration rate*" traduz a percentagem de agregados familiares que faz compras *online* pelo menos uma vez por ano, sendo a média Mundial de 20%, no Reino Unido de 22,1%, em França 30,1% e em Portugal 8,4%. O "*spend per trip*" traduz a relação entre as compras *online* e *offline* de FMCG, sendo a média mundial igual 3, ou seja, o consumidor gasta em média três vezes mais numa experiência no comércio *online* do que no offline. No Reino Unido este indicador assume o valor de 4,8, em França de 2 e em Portugal de 2,4. Por último a "*frequency*" que

representa o número de vezes que o consumidor utiliza o *E-commerce* para a compra de FMCG, sendo a média Mundial de 4 vezes por ano, no Reino Unido de 11,5 vezes, na França de 5,4 e em Portugal de 2,7 vezes por ano.

A compra de FMCG através duma plataforma *online* acarreta dificuldades extras, porque toda a logística de processamento, de transporte e de entrega das encomendas tem que ser estruturada tendo em conta que estamos, em alguns casos, a lidar com produtos perecíveis. No caso dos alimentos frescos, este acarretam ainda, múltiplas normas sobre o manuseamento e acondicionamento, tornando ainda mais complexa, a sua venda *online* (Hays, Keskinocak, López, 2005).

Apesar de existirem inúmeras vantagens na utilização do *e-commerce* e destas serem reconhecidas pelos consumidores nomeadamente, a poupança de tempo, a conveniência, o controlo de custos e a diminuição da compra impulsiva, existem ainda muitas ideias pré-concebidas e barreiras culturais (Golubova, 2012). É nos alimentos frescos ou perecíveis que existe uma maior desconfiança por parte dos consumidores. Preconceitos como a qualidade na seleção de alimentos efetuada pelo cliente na loja física, é superior à que o funcionário da empresa faz quando realiza o picking dos produtos ou, ainda que, os prazos dos alimentos entregues por este serviço são de menor validade que os produtos da loja (Bachl, Koll, 2013).

Todas estas ideias representam barreiras para a expansão do *E-commerce* neste sector.

2.2.3. O *e-commerce* em Portugal

Segundo um estudo realizado pela Associação do Comércio Eletrónico e publicidade interativa, denominado por "Estudo da Economia Digital 2012-2017", em 2012 havia 6,7 milhões de internautas (utilizadores de internet), que

representavam cerca de 64% da população portuguesa, 2,5 milhões de compradores *online*, 1,4 milhões de domínios registados e 2,7 milhões de dispositivos móveis, dos quais 96% com acesso à internet.

Outro estudo realizado pelo Instituto Nacional de Estatística, (INE 2013), encontrou dados muito semelhantes, referentes ao ano 2013, onde 67% das famílias dispunham de um computador, 62% tinham acesso à Internet e apenas 15% efetuaram compras *online*.

As previsões para 2017, revelam uma maior penetração da internet e uma maior propensão para comprar *online*. (ACEPI, 2013) Os números apontados para 2017 são de 8,4 milhões de utilizadores de internet (80% da população estimada para 2017), de 3,5 milhões de compradores *online* e de 4,9 milhões de dispositivos móveis com acesso à internet.

O valor do gasto médio por ano, no *e-commerce* em Portugal, evoluiu de 909€ em 2012, para 988€ em 2014. Espera-se que sejam alcançados 1.089€ em 2017, o que ainda fica muito longe, da média atual do Reino Unido, que ronda 2.500€.

O volume de negócios no comércio eletrónico *Business to Consumer* (B2C), tem seguido um aumento constante, apresentando um aumento de 1.557 milhões de euros em 2009 para 2.939 milhões de euros em 2014. Em 2017, prevê-se que seja atingido o valor de 4.028 milhões de euros.

Este crescimento representa um aumento de 52% entre 2009 e 2012 e um aumento expectável de 70% até 2017. O rácio entre o *e-commerce* e o comércio total em Portugal, tem seguido um crescimento constante aumentando de 18% em 2009 para 27,1% em 2011 (ACEPI, 2013).

Apesar do crescimento em Portugal, a realidade do *e-commerce*, continua a evidenciar barreiras que demovem os portugueses de comprar mais ativamente, nomeadamente, os elevados custos de transporte e a concorrência direta de agentes ou lojas físicas (Couceiro, 2012).

É possível, assim, concluir que o *e-commerce* em Portugal continuará a crescer nos próximos anos aliado uma maior penetração da internet no território Nacional e às mudanças nos hábitos de consumo dos Portugueses.

2.2.4. *E-commerce* dos FMCG em Portugal

Com o crescimento do *e-commerce* registado nos últimos anos, as grandes empresas do Retalho a operar no mercado português começaram a encarar as vendas *online*, não apenas como um canal de vendas adicional mas, como um mercado promissor e que justifica investimento.

Segundo a Associação Portuguesa das Empresas de Distribuição, no 1º semestre de 2014, as vendas de FMCG em lojas física, não apresentaram crescimento, face ao ano anterior. As vendas realizadas nos três maiores *sites* de Retalho, o site do Continente, o site do El Corte Inglês e o site do Jumbo registaram, em média, aumentos na faturação de 0,4%.

Um estudo conduzido pela Marktest identificou algumas tendências no consumo de FMCG através do *e-commerce*, no ano de 2012. Do total dos inquiridos, apenas 3% tinha comprado bens alimentares e de retalho geral nos últimos doze meses (Marktest, 2012).

Cerca de 0,9% dos FMCG, transacionados no mercado nacional, foram comprados através de plataformas *online* (Kantar Worldpanel, 2014). Para os próximos anos é expectável que os portugueses recorram ao *e-commerce* com maior frequência para a comprar este tipo de produtos. A estimativa para 2016, é que a compra destes produtos represente 1,4%. O cenário em Portugal é ainda um pouco tímido comparativamente com o Reino Unido em que o peso do *e-commerce* na venda de FMCG atinge os 5%. Contudo, é evidente que os consumidores portugueses começam a recorrer com maior frequência ao *e-commerce* para comprarem este tipo de produtos.

No caso do Continente, a plataforma foi criada em 2001, registando numa fase inicial cerca de mil utilizadores. Em 2014, evolui para cerca de meio milhão de utilizadores registados. A empresa El Corte Inglês, lançou a sua plataforma um pouco mais tarde que o Continente, em 2004, angariando cerca de cinco mil utilizadores nos primeiros anos de atividade. Em 2014, a plataforma tinha cerca de cem mil utilizadores registados. O Jumbo foi o último a criar a sua plataforma no ano 2007, começando com cerca de trinta mil clientes registados, tendo atingido no ano 2014 cerca de duzentos e cinquenta mil clientes.

Ainda que o aumento dos clientes registados não signifique diretamente um aumento das vendas e dos resultados, é evidente que a compra de FMCG no *e-commerce* em Portugal, segue a tendência Europeia, continuando a aumentar e a quebrar barreiras nos hábitos de consumo.

2.2.5. A Sonae SGPS

Comecemos por explicar como é constituído o Grupo Sonae, em termos da sua estrutura acionista e por áreas de negócio, para contextualizarmos melhor a empresa Modelo.com.

A estrutura acionista do Grupo Sonae evidencia uma forte presença da empresa Efanor, detentora de 52,65% das ações da Sonae, seguida pelo BPI com 8,9%, pelo BESTINVER, que detém 4,99%, pela Fundação Berardo com 2,5% e por último o Norges Bank com uma participação de 2,0%. Os restantes 28,96%, pertencem a múltiplos investidores (www.sonae.pt).

O grupo Sonae subdivide-se em três empresas, a Sonae, a Sonae Indústria e a Sonae Capital, mantendo-se a sua estrutura acionista, praticamente inalterada. Por exemplo no caso da Efanor que detém 52,65% do Grupo, possui 53% da Empresa Sonae, 51% da Sonae Industria e 56% da Sonae Capital.

A Sonae MC pertence à empresa Sonae, juntamente com a Sonae SR, com a Sonae Sierra, com a Sonaecom, com a Sonae RP e por fim com a Investment Management (www.sonae.pt).

A Sonae MC e a Sonae SR são as empresas cujo Core Business é o Retalho e ambas contam com uma participação de 100% da empresa Sonae. A área de negócio da Sonae MC dedica-se predominantemente ao retalho de produtos alimentares, enquanto a Sonae SR ao retalho de todos os produtos não alimentares.

As restantes empresas do Grupo atuam em distintas áreas de negócio. A Sonae Sierra opera na área da criação e gestão de centros comerciais, sendo detida em 50% pela Sonae e 50% pela Grosvenor (www.sonae.pt).

A Sonaecom atua na área das Telecomunicações, *Software*, Sistemas de Informação e Media, sendo detida em 55% pela Sonae. A Sonae RP tem como área de negócio a gestão do património, com participação de 100% da empresa Sonae. A Investment Management, como o próprio nome indica, faz a gestão de investimentos da empresa Sonae que detêm 100% do seu capital (www.sonae.pt).

A Sonae MC, como já mencionado anteriormente, é uma empresa que pertence ao Grupo Sonae, tendo conseguido alcançar a posição de líder de mercado nacional, no retalho de produtos alimentares (www.sonae.pt).

Esta empresa liderou as mudanças que ocorreram nos hábitos de consumo dos portugueses, através da criação do primeiro hipermercado em Portugal, em 1985 (Continente de Matosinhos), dando origem a uma nova tendência de consumo, o consumo nos hipermercados, que se tornou na prática mais comum de consumo dos portugueses (www.sonae.pt).

A empresa tem no seu *portfolio*, um conjunto de marcas, que oferecem uma vasta gama de produtos, apostando na qualidade superior e nos preços mais competitivos (www.sonae.pt).

A Sonae MC detém as seguintes marcas, que são uma referência a nível nacional: Continente e Continente Modelo (hipermercados), Continente Bom Dia (supermercados de conveniência), Bom Bocado (cafetaria e restaurantes), Note! (livraria/papelaria) e Well´s (saúde, bem estar e ótica) (www.sonae.pt).

2.2.6. A Modelo.com

A Modelo.com é a empresa responsável pela gestão do canal de *e-commerce* da Sonae MC. Esta esteve incumbida do lançamento da plataforma *online*, denominada Continente Online ou COL e gere todas as operações inerentes a um modelo de *e-commerce* de Retalho.

Algumas das atividades realizadas pela empresa, englobam o funcionamento do *site*, nomeadamente a gestão de conteúdos, a gestão de espaços publicitários e de ações de campanha, bem como a gestão dos clientes e dos processos de compra. A empresa é responsável pelo processo de seleção de produtos, denominado por *picking*, pela gestão de encomendas e pelo envio ao domicílio, bem como das restantes operações necessárias para o negócio, incluindo a gestão do CRM (*Customer Relationship Management*), onde se processam todas as interações com os clientes e o serviço pós-venda (Fonte Modelo.com).

A plataforma inicial do COL foi lançada no ano de 2000 e opera em quinze lojas físicas, estrategicamente escolhidas pela sua importância, em termos populacionais e de potencial económico. As lojas estão distribuídas pelas áreas Metropolitanas de Lisboa e Porto e estão presentes em cidades como Portimão, Coimbra, Covilhã, Leiria, Guimarães e Viseu (Fonte Modelo.com).

Foram obtidas algumas informações sobre a atividade da Modelo.com desde 2004 a 2013 através da base de dados SABI (Sistema de Análises de Balanços

Ibéricos). Nas próximas tabelas, estão presentes os valores referentes, aos quatro anos mais recentes.

Proveitos operacionais	12.592.419 €
Lucro após impostos	1.270.297 €
Total do ativo	17.921.809 €
Número de funcionários	274

Tabela 1-Resumo de dados sobre a Modelo.com

Entre os anos de 2004 e 2013, a empresa passou por um período com resultados negativos, mais concretamente nos anos 2004 e 2005.

	31/12/2013	31/12/2012	31/12/2011	31/12/2010
Fluxos de caixa	2.263.560,57 €	1.819.126,09 €	1.806.138,31 €	4.363.843,90 €
Valor acrescentado bruto	6.284.608,80 €	5.846.731,14 €	5.899.012,86 €	9.043.220,25 €
EBIT	844.403,78 €	987.084,67 €	976.952,40 €	4.233.946,27 €
EBITDA	1.837.667,82 €	1.831.175,25 €	1.819.733,29 €	5.285.001,68 €

Tabela 2-Resultados da Modelo.com em 2010 e 2013

No ano de 2006 a empresa registou resultados positivos e um forte crescimento face aos anos anteriores. Até 2010 os resultados foram sempre positivos, atingindo o seu auge nesse mesmo ano.

	31/12/2013	31/12/2012	31/12/2011	31/12/2010
Proveitos Operacionais	12.592.418,55 €	11.682.601,33 €	11.153.442,00 €	13.070.656,64 €
Resultados Correntes	1.264.112,59 €	1.266.427,68 €	1.238.226,06 €	4.484.799,01 €
"Resultado Líquido do Exercício	1.270.296,53 €	975.035,51 €	963.357,42 €	3.312.788,49 €
Total Activo	17.921.808,83 €	19.047.340,00 €	20.016.853,53 €	17.102.858,42 €
Capital Próprio	14.041.361,36 €	13.746.100,34 €	13.734.422,25 €	12.775.614,74 €

Tabela 3-Resumo de dados sobre a Modelo.com

A partir de 2010, a empresa apresentou resultados inferiores, notando-se alguma melhora em 2013.

2.3. Análise e discussão das principais teorias disponíveis para a área de estudo

Para a construção deste Relatório de Estágio, foi necessária uma extensiva pesquisa nas áreas da Segmentação, de *Profiling*, de *Data Mining* e de Algoritmos de *Clustering*, recorrendo não apenas a Trabalhos Académicos, mas também a Livros Especializados sobre os diversos temas.

Nos dias de hoje, uma empresa tem que girar em torno do cliente o que se traduz num mercado que é "*Customer Driven*". Os clientes esperam que lhes seja fornecido o que procuram e com adaptações ou customização (Galbraith, 2005).

As empresas que pretendam ser competitivas são levadas a mudar a sua metodologia ou a ajustar as suas atividades, do ponto de vista do vendedor, para o ponto de vista do comprador. Em consequência os produtos e serviços oferecidos podem, em alguns casos, passar dum formato *standard*, para um formato customizado e com variantes individuais (Forza, Salvador, 2007).

As novas tecnologias vieram fornecer as plataformas necessárias para que ocorresse a passagem para um mundo de produtos adaptados às necessidades e gostos de cada cliente, um fato comum dentro da realidade do *e-commerce* (Reichwald, Seifert, Walcher, Piller, 2004).

Segundo (Kolter, 1994) o mercado pode ser definido como o conjunto de todos os potenciais clientes que partilham uma necessidade e que estão dispostos a uma possível troca para a satisfazer. Sob esta definição de mercado, é possível considerar que qualquer utilizador da Internet pode representar um potencial cliente.

Mas quem é o cliente do universo *e-commerce* segundo a literatura? Este cliente pode ser definido como aquele que realiza transações financeiras com o vendedor *online* de bens e serviços. Mas também todos aqueles que consomem a informação sobre os produtos e serviços vendidos pelas empresas (Lawrence, Corbitt, Fisher, Tidwell, 2000:167).

Uma das métricas fundamentais para uma empresa conseguir obter vantagens competitivas face aos seus concorrentes é conhecer o seu mercado e os seus clientes. Para as alcançar a empresa deve estar alinhada com as necessidades e vontades dos clientes (McDonald, Dunbar, 1998).

Assumir que as necessidades dos clientes são todas iguais pode ser um erro crasso já que dentro do mercado, existem diferentes perfis e grupos de clientes.

(Levitt, 1986) recordou que quando não se está a pensar em segmentos não se está a pensar de todo, ou seja, não se está a ter em conta a existência de diferentes vontades e escolhas partilhadas por grupos de clientes.

No processo de conhecimento dos clientes, a segmentação é o primeiro passo a ser trabalhado e que permitirá classificar os clientes de acordo com os segmentos encontrados (Jansen, 2007), abrindo portas para a definição de perfis, permitindo construção de modelos comportamentais dos clientes (Verhoef, Spring, Hoekstra, Lee, 2002).

A orientação para o cliente ao nível individual leva a que a construção destes modelos contemple cada vez mais atributos que representem preferências de compra e o historial de compras do indivíduo, em vez de atributos como variáveis demográficas e geográficas (Jansen, 2007).

Manter o nível de satisfação do cliente ajuda a fidelizar o cliente, sendo esta uma estratégia mais lucrativa e estável do que a captação de novos clientes.

Este deve ser um dos objetivos fulcrais da empresa (Reichheld, Sasser, 1990:105) por isso é da maior importância saber quem é o cliente e atender às suas necessidades.

2.3.1. Segmentação

Comecemos por analisar o que é a Segmentação e o seu papel no processo de definição de perfis. (Banon, 2004) afirma que o sucesso do *Marketing* começa com uma boa segmentação. Para (Smith, 1956) a segmentação é uma importante ferramenta de *Marketing* que ajuda os *Marketers* a definir as necessidades e vontades dos clientes, estabelecer os objetivos e alocar recursos de uma forma

mais e eficiente e permite realizar avaliações de performance com maior exatidão.

A Segmentação, aplicada num contexto de *Marketing*, é o termo utilizado para descrever o processo de dividir os clientes em grupos homogéneos ou segmentos que partilham atributos comuns como necessidades, hábitos ou gostos (Giha, Singh, Ewe, 2003).

Um segmento é, portanto, um subgrupo de clientes que partilham uma ou mais características que os levam a ter necessidades semelhantes (Leung, 2009).

Através deste processo, é possível realizar uma comunicação mais orientada com os clientes da empresa. Para tal é necessário descrever as características dos segmentos ou *Clusters*, encontrados durante o processo (Verhoef, Spring, Hoekstra, Lee, 2002).

A segmentação pode ser dividida em duas metodologias ou correntes, a Segmentação *Market Driven* e a Segmentação *Data Driven*.

No caso da primeira, os atributos são definidos pelos *Marketers* ou Gestores, estes identificam quais sãos as características que permitem definir os segmentos. Ou seja, existe uma pré-seleção dos atributos, tendo em conta o objetivo que se pretende atingir (Rud, Parr, 2001).

Uma Segmentação *Data Driven* ocorre quando a segmentação é criada através da utilização de dados dos clientes da empresa sem que, à partida, se decidam os atributos a utilizar (Scridon, 2008).

É através dum conjunto de ferramentas estatísticas, denominado por *Data Mining*, que se descobrem os padrões nos dados, permitindo encontrar os segmentos existentes nomeadamente a Técnicas de *Clustering* (Jansen, 2007), que será abordado em maior profundidade neste capítulo e na Metodologia.

Muitos autores defendem que a segmentação com base em processos analíticos, ou seja, numa perspetiva de *Data Driven*, onde os segmentos são

determinados via a utilização de processo de *Data Mining* é a mais eficiente e que gera melhor conhecimento (Breiman, Friedman, Stone, 1984).

Antes de se validar os segmentos encontrados estes devem preencher um conjunto de características sob o ponto de vista da sua importância para a empresa (Scrindon, 2008). Os segmentos encontrados devem ser relevantes para o objetivo da empresa e devem também ser facilmente distinguidos e caracterizados.

No processo de conhecimento dos clientes, a Segmentação é o primeiro passo a ser trabalhado e que permitirá classificar os clientes de acordo com os segmentos encontrados (Jansen, 2007), abrindo portas para a definição dos perfis.

2.3.2. Profiling

Abordemos agora o tema da definição de perfis, ou Profiling.

Após a Segmentação dos clientes, a cada segmento deve ser atribuído um perfil, construído com base em atributos ou descritores (Kotler, 1991), sejam eles demográficos, psicográficos, ou comportamentais.

O ato de determinar perfis implica que se descreva os clientes pelos seus atributos sejam eles a idade, o género, o rendimento ou o estilo de vida (Ahola, Runsala, 2001).

(Sindell, 2000), define o *Profiling* como o processo de construção do retrato do cliente recorrendo a bases de dados colaborativas que contêm toda a informação sobre o seu comportamento de compra, motivações e preferências por produtos ou serviços dos clientes.

É através da identificação destes atributos que é criada a base para os *Marketers* comunicarem com os clientes duma forma eficiente, satisfazendo-os e retende-os (Amat, 2002).

O *Profiling* permite, ainda, criar produtos e serviços, que vão ao encontro das características de cada perfil de cliente sendo por isso uma excelente ferramenta para os *Marketers* (Mazanec, 1992).

Com a criação de perfis é possível ajustar o produto ou serviços às necessidades de cada perfil de cliente, sejam estas relacionadas com o preço ou tamanho do produto (Kotler, 1991) ou relativas a preferências do cliente (Etzel, Woodside, 1992).

Utilizando as informações que foram obtidas previamente com a Segmentação, quer tenha sido gerada através de processos analíticos ou não, é possível estimar um perfil para cada segmento encontrado (Giha, Singh, Ewe, 2003).

Sabendo que o objetivo da definição de perfis é a construção de modelos de previsão do comportamento dos clientes, os atributos selecionados devem ir de encontro dos objetivos da empresa e do estudo a desenvolver (Jansen 2007).

Existem diferentes tipos de atributos que podem ser utilizados para definir os perfis e criar o modelo de previsão. Por exemplo, atributos geográficos culturais, étnicos, económicos, de estilo de vida entre outros. A seleção destes atributos depende das fontes e do tipo de dados disponíveis (Scrindon, 2008).

A seleção dos atributos é variada e ajusta-se às várias áreas de estudo que recorrem à segmentação e definição de perfis. A diversidade na sua seleção pode ser observada em vários trabalhos como, por exemplo, (Tremblay, Dunlap, 1978) e (Pickett,1993) que recorreram a atributos geográficos, (Anderson, 1974), atributos culturais, (Kinnear, 1974), atributos psicográficos (Samdahl, Robertson, 1989) e (Scott, Willis, 1994), que utilizaram atributos sociodemográficos.

Os atributos utilizados na construção dos perfis podem ser divididos em dois tipos: os atributos factuais (Beane, Ennis, 1987) e (Lilien, Rangaswamy, 2003), como por exemplo, atributos sociodemográficos e psicográficos e atributos

comportamentais. Segundo (Yankelovich, Meer, 2006) os atributos comportamentais contemplam informações como quais os produtos comprados pelos clientes e qual o montante gasto.

A construção de perfis, recorrendo principalmente ou unicamente a atributos sociodemográficos, foi durante muito tempo uma prática comum em muitas empresas fornecedoras de produtos e de serviços (McDonald, Dunbar, 1998). Este fato justifica-se pela abundância de fontes de dados com estes atributos e porque a construção de modelos de segmentação e de perfis, é mais simples, do que utilizando outras variáveis (Mansoorian, Myers, 1996).

(Pratt, 2011) e (Chang, K. Kim, H. Kim, 2002) utilizaram atributos factuais para a criação de perfis para consumidores de vinho, demonstrando a simplicidade do método.

Os atributos sociodemográficos continuam a ter um papel importante na segmentação e na criação de perfis sendo um meio de completar a informação sobre os segmentos e perfis encontrados (Wedel, Kamakura, 2000). Contudo, muitos autores defendem que a sua utilização, como o único atributo para o modelo, pode levar à definição de perfis enviesados (Frank, 1972), prejudicando o seu principal objetivo enquanto ferramenta de previsão comportamental.

A construção de perfis com atributos factuais, sociodemográficos e psicográficos pode ser mais simples de executar em termos de construção mas, quando o estudo em causa, envolve a compra frequente de bens consumíveis, o seu uso é desaconselhável (Frank, 1972) e (Guadugni, Litle, 1983).

Segundo (Franzak, 2001:1-2), os atributos que descrevem o comportamento pessoal do indivíduo, representam as suas vontades e necessidades, logo, estimar o seu comportamento com base nestas características, leva a uma maior satisfação do cliente.

A escolha de atributos comportamentais para construção de perfis, permite criar perfis que refletem um conhecimento profundo sobre os hábitos de

compra dos clientes, permitindo aumentar a sua fidelização, o seu valor e sua satisfação (Chen, 2007) e (Dickson, 1982).

Através da definição dos perfis, uma empresa pode descobrir quais os clientes mais valiosos e avaliar a forma como determinado perfil reage a uma promoção ou a qualquer outra atividade de *Marketing* (Leung, 2009). Pode permitir também, a construção dum sistema de recomendação ao cliente, baseado no comportamento de outros clientes, que partilham o mesmo perfil. É, assim, possível fornecer aos *Marketers*, melhores bases para a tomada de decisão, nomeadamente sob quais as melhores ações de *Marketing* a desenvolver para cada segmento e na alocação de recursos aplicados em cada ação e segmento (Jansen, 2007).

A construção do perfil do cliente, que consome produtos via *e-commerce*, já foi alvo de inúmeros estudos. Aqui como o foco de interesse, são os produtos de Retalho Geral ou FMCG, apenas foram selecionados, os estudos que cumprem este requisito.

No contexto internacional, (Morganosky, Cude, 2000), definiram o perfil do cliente que compra FMCG via *e-commerce* nos Estados Unidos da América, através das respostas a um inquérito, realizado a 243 indivíduos. Nesta amostra, 32% dos inquiridos, apresentava uma idade próxima ou inferior a 34 anos, face aos 7% com 55 anos ou mais. Relativamente ao rendimento, 50% apresentava rendimentos de classe média alta, enquanto apenas 12% obtinha rendimentos de classe média baixa e baixa.

(Raijas, Tuunainen, 2001), identificaram que o perfil do cliente *e-commence* de FMCG tende para indivíduos mais jovens, com formação e rendimentos elevados e residentes em zonas com elevada densidade populacional.

Em Portugal, (Marktest, 2012) realizou um estudo, onde foram identificados, com base em atributos sociodemográficos, o perfil ou perfis, do cliente que compra produtos FMCG, através do *e-commerce*. Dentro da amostra, foi

identificado o perfil do cliente, que mais compra estes produtos numa plataforma *online*. Este cliente é do sexo feminino, com idade compreendida entre os 35 e os 44 anos, com área de residência na Grande Lisboa e que pertencem a classes sociais média alta e alta.

Em contrapartida, o perfil do cliente menos propenso a este de tipo de compra foi identificado como sendo do sexo feminino, com idade igual ou superior a 65 anos, residente na região interior do país e pertencente às classes socais média baixa e baixa (Marktest, 2012).

O que se pode concluir com estes estudos, é que existe um grupo de indivíduos que partilha um conjunto de atributos que definem o seu perfil. Será que o cliente do Continente Online também partilha este perfil? Esta é uma da questões que se pretende responder, com este relatório.

2.3.3. Data Mining

Passemos a próxima fase deste desafio, que leva à aplicação do conceito de *Data Mining*, pois trata-se do conceito introdutório à técnica de *Clustering*.

Este conceito, é referido por muitos autores, (Breiman, Friedman, Stone, 1984), (Jansen, 2007) e (Scrindon, 2008), que o identificam como sendo uma excelente via para processar grandes quantidade de dados e para segmentar os clientes.

De entre muitos trabalhos estudados referentes a este tema, (Ruxandra Petre, 2013), demonstrou ser o mais recente e relevante, já que explica o *Data Mining*, numa perspetiva empresarial, o que vai de encontro ao ambiente em que este relatório foi desenvolvido, ou seja, dentro da Modelo.Com.

O *Data Mining* é, nos dias de hoje, a maneira mais eficaz para lidar com elevados volumes de dados, como é o caso da Modelo.com (Ruxandra Petre, 2013). Esta quantidade de dados necessita de ser analisada de forma a

podermos extrair informações valiosas, ou seja, para produzir conhecimento (Bounsaythip, Rinta-Runsal, 2001).

Num ambiente empresarial, em constante evolução, é imposto às empresas que mantenham a sua competitividade e capacidade de adaptação a novos desafios. O conhecimento criado através destas ferramentas pode representar uma vantagem competitiva, projetando a organização para a frente dos seus concorrentes. Descobrir padrões nos dados pode ajudar a prever comportamentos, estando um passo à frente e permitindo reagir atempadamente.

O *Data Mining* pode ser descrito como o processo de exploração e análise de grandes quantidades de dados com o objetivo específico de descobrir padrões e regras que sejam importantes e significativos, extraindo conhecimento de dados brutos ou não trabalhados (Road, Jose & Getta, 2003).

Em termos mais técnicos este conjunto de ferramentas consiste na aplicação de análise de dados e de algoritmos de descoberta, recorrendo a mecanismos computacionais, produzindo modelos sobre os dados (Ruxandra Petre, 2013).

Os métodos utilizados para encontrar estes padrões ou modelos, podem ser divididos em duas categorias: Métodos Descritivos e Métodos Preditivos. Os Métodos Descritivos são aplicados quando é necessário interpretar os dados, expondo através da visualização as diversas relações entre os dados.

Os Métodos Preditivos são métodos que visam a construção automática de modelos comportamentais, ou seja, a construção de modelos que contemplam todas as possíveis interações entre os dados, permitindo prever possíveis valores para uma ou mais variáveis.

Dentro destas duas categorias existem várias técnicas que podem ser utilizadas num ampla variedade de dados. Na figura abaixo, podemos observar as seis técnicas utilizadas no *Data Mining*.

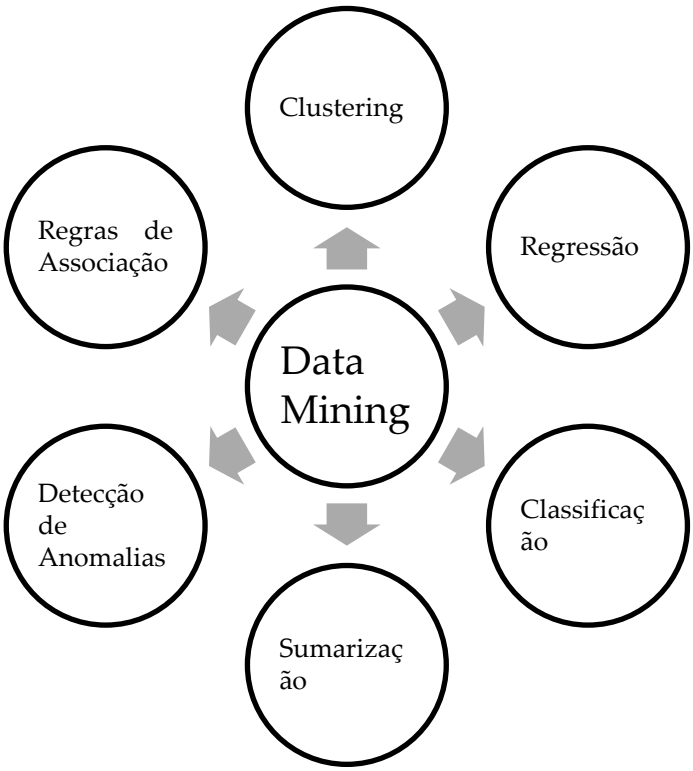

Ilustração 1-Ténicas de Data Mining (Ruxandra Petre, 2013)

Depois de avaliado o tema e analisados os dados disponíveis concluiu-se que para atingir o que era pretendido, a técnica mais apropriada era a técnica de *Clustering*. Esta decisão teve por base trabalhos realizados por outros autores e o aconselhamento por parte de membros da equipa de *Business Intelligence* da Modelo.com.

Como vamos explicar a técnica de *Clustering*, em maior detalhe mais à frente neste capítulo, peguemos numa simples definição utilizada por (Ruxandra Petre, 2013), esta é uma técnica descritiva que procura encontrar um conjunto finito de grupos ou *Clusters* para melhor descrever os dados. A autora refere também, quais as possíveis aplicações empresariais do *Data Mining*, fazendo referência específica ao setor do Retalho, além da possibilidade de aplicação em inúmeros setores como o da Banca e Seguros.

O *Data Mining* utilizado no setor de Retalho permite conhecer melhor os clientes, ajudando a identificar os seus comportamentos de compra, a aumentar

o consumo de produtos e a reduzir os custos do negócio, nomeadamente através de sistemas de distribuição e transporte mais eficazes.

As suas aplicações passam pela Segmentação dos Clientes, tema central deste relatório, pela Análise do Comportamento de Compra, pelo Aumento na Retenção de Clientes, ou seja, diminuindo o Cliente Churn e pela Análise do Impacto de Campanhas Publicitárias e de Vendas.

No gráfico seguinte podemos observar possíveis aplicações duma solução de *Data Mining*, num contexto empresarial.

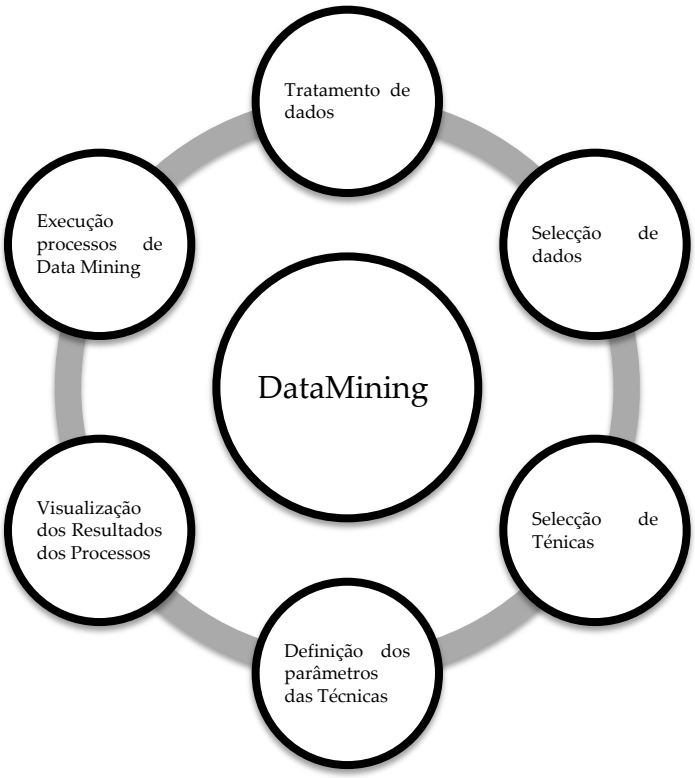

Ilustração 2-Aplicaçõs do Data Mining em ambiente empresarial (Ruxandra Petre, 2013)

O gráfico seguinte ilustra um exemplo duma solução de *Data Mining*.

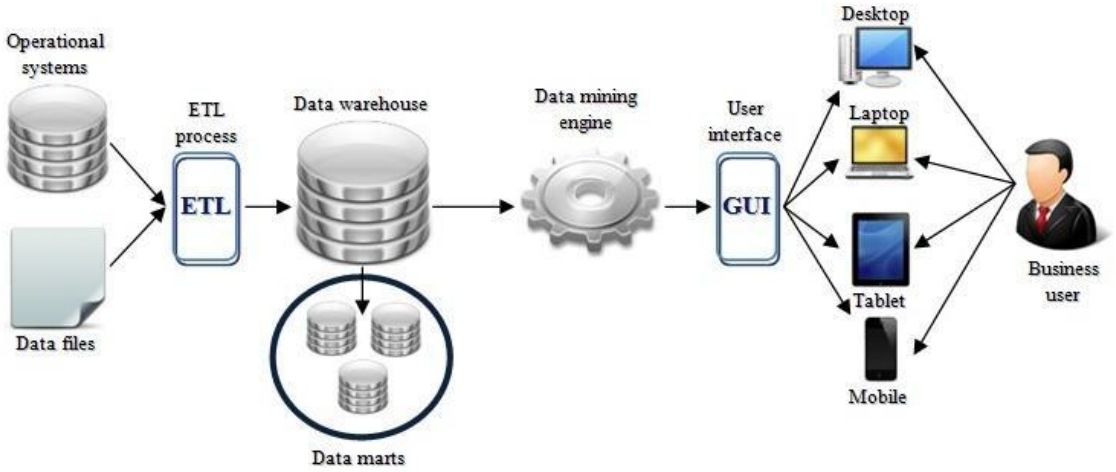

Ilustração 3- Arquitetura dum Processo de Data Mining numa Empresa (Ruxandra Petre, 2013)

Todo o processo tem início com a extração de dados através das fontes de dados, que podem ser sistemas transacionais ou base de dados. Este é o primeiro passo do processo denominado por *Extract, Transform and Load* (ETL). É através deste processo que as bases de dados são alimentadas, podendo conter dados brutos, dados sumários ou Metadados.

Depois de criadas as bases de dados, o próximo passo, é a utilização do *Data Mining Engine*, que permite realizar a análise dos dados e a execução de múltiplas aplicações de *Data Mining*, nomeadamente, a execução de processos ou técnicas de *Clustering*.

2.3.4. Clustering

A escolha do método de *Clustering* de entre os restantes métodos justifica-se pela sua utilização amplamente disseminada, em trabalhos académicos, que abordam a temática da segmentação do cliente e do estudo do comportamento de compra.

Segundo (Venkatesan, 2007), que define o método de *Clustering* como " uma classe de técnicas estatísticas, que podem ser aplicadas a dados que exibam agrupamentos naturais", para uma análise de segmentos ou agrupamentos, a

análise de Cluster é um método mais fiável, já que não se identifica quais as variáveis dependentes e independentes.

Todas as possíveis relações entre as variáveis são examinadas e agrupadas em grupos relativamente homogéneos. Os elementos pertencentes a um grupo são sempre mais semelhantes entre si comparativamente com elementos que pertençam a outros grupos.

Em (Venkatesan, 2007), estão definidos os principais passos para a construção de *Cluster*, que o autor enumera da seguinte forma:

1. Formulação do problema (Seleção das variáveis para análise).
2. Cálculo da distância entre as amostras ou clientes, ao longo das variáveis escolhidas.
3. Aplicação do método ou técnica de *Clustering*.
4. Decisão sobre o número de *Clusters*, no caso de se tratar dum algoritmo de *Clustering a priori*, por exemplo, *K-Means*.
5. Interpretação e Descrição dos *Clusters*.

O autor utilizou o mesmo algoritmo e o mesmo tipo de distância, utilizadas para a construção do modelo, exibido no presente trabalho, ou seja, o algoritmo *K-Means* e a Distância Euclidiana. Esta distância é para o autor, o método principal, para a construção de *Clusters*, já que mede a distância real entre duas observações que estão a ser agrupadas.

A utilização do Algoritmo K-Means deve-se ao facto da sua simplicidade e velocidade, em termos de processamento e cálculo. Além disso, (Venkatesan, 2007), salienta que é um dos algoritmos mais robustos. Este algoritmo tem a particularidade de ser capaz de lidar com grandes quantidades de dados e de ser menos sensível a outliers.

Os passos do Algoritmo K-Means, foram descritos pelo autor pela seguinte ordem:

1. Escolha do número de *Clusters*.
2. Criação de pontos (K), os centroides de cada *Clusters*.
3. Atribuição de cada ponto ou observação, ao centroide mais próximo.
4. Recálculo dos novos centroides de cada *Cluster*, em função das observações.
5. Repetição dos passos (4) e (5), até que seja encontrado um critério de convergência.

Uma das grandes falhas do Algoritmo K-Means (Venkatesan, 2007), é o facto de que o algoritmo, não ser capaz de dar uma estimativa do número de *Clusters* existentes nos dados. Tratam-se dum algoritmo em que o número de *Clusters* é definido *a priori*, existe a imposição dum número de *Clusters*, que pode enviesar a interpretação dos resultados.

Contudo, existe um método para determinar o número apropriado de *Clusters*, denominado por "*Elbow Criterion*". Segundo (Venkatesan, 2007), o que se pretende observar com este método, é se a variação dentro de cada *Cluster* e entre *Clusters*, se altera à medida que o número de *inputs* (K), é alterado.

Como o objetivo de uma Análise de *Clusters*, é minimizar a variação dentro de cada *Cluster* e aumentar a variação entre *Clusters*, se calcularmos o rácio entre os dois, para diferentes (K), por exemplo, K= (1,2,3,4,n), o número ideal para (K) é atingido, quando ao adicionarmos mais um (K), por exemplo K=n+1, n+2, n, não geramos informação significativa.

Outro método utilizado para assegurar a robustez dos resultados obtidos, (Venkatesan, 2007) é a utilização de diferentes pontos de partida, ou *Seeds*. Os *Seeds* são pontos aleatórios, ou seja, utilizando o algoritmo com diferentes *Seeds*, caso os resultados sejam semelhantes, confirma-se a robustez do modelo.

(Kronenberg, 2008), aborda a temática da seleção de *Seeds* para o algoritmo K-Means. Como o algoritmo exige que se introduza valores aleatórios de *Seeds*, a única forma de criar os melhores *Clusters* e reduzir a margem de erro, é a utilização de *Seeds* diferentes em vários testes, até se atingir os melhores resultados. Para o autor, este processo é pouco eficiente, pois requer múltiplos testes e como tal, é um processo demoroso.

O autor propõe seis métodos alternativos para determinar os *Seeds* iniciais e compara os resultados com os métodos mais utilizados, nomeadamente o teste com diferentes *Seeds*.

Em (Chen, 2012), um artigo sobre soluções de *Data Mining* para a indústria do Retalho Online, foi elaborado um modelo de segmentação do cliente utilizando uma solução *Data Mining*.

Segundo o estudo, recorrendo a um modelo de segmentação do cliente, é possível ter um conhecimento mais aprofundando sobre o seu comportamento de compra e com esse conhecimento melhorar os resultados da empresa.

Os dados utilizados em (Chen, 2012), são provenientes duma empresa de Retalho Online que opera na Europa e foram recolhidos ao longo do ano de 2011. O *Data Mining Engine* ou o programa utilizado, foi o SAS Entreprise Miner.

O modelo construído por Chen, utilizou o algoritmo de *Clustering K-Means* e a análise RFM (*Recency, Frequency, Monetary*). Os *Clusters* foram identificados com base nas três variáveis da análise RFM, ou seja, os *Clusters* encontrados, aglomeram os clientes que partilham valores mais próximos destas três variáveis.

O algoritmo foi aplicado em três cenários diferentes, K=3, K=4 e K=5, em que K representa o número de *inputs* iniciais ou *Seeds*. Os resultados do algoritmo com *input* de 5 *Clusters* foi o que permitiu uma melhor observação e interpretação dos dados, tendo sido o escolhido para a interpretação dos dados.

A análise das três variáveis foi realizada em cada um dos cinco *Clusters* o que permitiu identificar os *Clusters* com maior valor para a empresa e os com menos valor.

Dois trabalhos realizados por Vera Migueis, António Camanho e João Cunha, foram particularmente importantes para a elaboração do presente relatório. Recorreram em ambos os casos ao *Data Mining* e utilizaram o método de *Clustering* para alcançar os seus objetivos.

Embora os objetivos e as metodologias dos trabalhos sejam diferentes, ambos utilizaram como fonte de dados uma Empresa de Retalho, que opera na Europa, contendo toda a informação sobre os clientes, nomeadamente, o nome, a morada, o número de pessoas no agregado familiar e outras.

Esta informação provém de cartões de fidelidade ou de cliente que, para além de informações sobre o cliente, contém toda a informação sobre todas as suas transações. A estrutura mercadológica dos produtos comercializados pela empresa era a seguinte; Unidade de Negócio, sendo a mais abrangente, seguida de Categoria, Subcategoria, Produto e a Marca, sendo o nível mais fino.

Em "*Mining Customer Loyalty Card Programs*" de (Migueis, Camanho e Cunha, 2011), os autores utilizaram o Algoritmo K-Means para efetuar o *Clustering*. Estes defenderam a sua utilização, devido a duas grandes vantagens face a outros algoritmos, a sua facilidade de implementação e utilização e a sua velocidade em termos de processamento.

As variáveis utilizadas para o *Clustering* dos Clientes, foram a *Frequency* e a *Monetary* (FM). As variáveis traduzem, a número médio de compras por mês e o valor médio gasto por compra pelo cliente. Estas são duas das variáveis utilizadas na análise RFM (*Recency, Frequency and Monetary*). Os autores excluíram a variável *Recency*, já que o período de análise não era suficientemente grande para que houvesse diferenças suficientes entre os clientes nesta variável.

Cientes das limitações do Algoritmo de *Clustering* K-Means, os autores recorreram à "*Elbow Curve*", já anteriormente referida neste capítulo, e ao *índex* de Davies-Bouldien. Este *índex* é uma função do rácio entre o somatório de pontos dentro de um *Cluster (Intracluster)* e o somatório de pontos entre *Clusters* (*Intercluster*). Quanto mais baixo for o valor do *índex* associado ao número inicial de *Clusters* (K), melhor é o valor do número inicial de *Clusters*.

Segundo o *índex* de Davies-Bouldien, o melhor valor para número inicial de *Clusters*, estava entre K=[4,5]. O resultado da "*Elbow Curve*", K=5, conciliado com o valor obtido pelo *índex*, permitiu determinar o número ideal para K, ou seja, K=5.

A segmentação dos clientes com base nas duas variáveis FM, permitiu identificar os Clusters com maior valor para a empresa e os menos valiosos.

No segundo trabalho feito pelos autores (Migueis, Camanho e Cunha, 2012), denominado por "*Customer data Mining for lifestyle segmentation*", a metodologia foi diferente, já que as variáveis para o *Clustering*, não eram variáveis de compra, como as utilizadas na análise anterior.

As variáveis utilizadas para a construção do modelo, foram os produtos comprados pelos clientes, ou seja, o nível de produto da estrutura mercadológica.

O que os autores procuraram foi analisar as relações entre os produtos comprados pelos clientes, ou seja, foi feita uma análise aos *Shopping Baskets* dos Clientes. A segmentação pretendida teve apenas em conta o facto se o cliente comprou ou não determinado produto, independentemente da quantidade comprada.

O algoritmo utilizado, denominado por VARCLUS, analisa a correlação entre as variáveis e não a distância entre variáveis, como é o caso do K-Means. Foram encontrados 6 Clusters, que depois de uma análise aos produtos consumidos, permitiu classificar os clientes por estilo de vida.

Em seguida, foi feita uma comparação entre os produtos consumidos em cada *Cluster*, com os restantes níveis da estrutura mercadológica, nomeadamente a Unidade de Negócio, a Categoria e a Marca.

A análise de cada *Cluster* permitiu identificar quais os clientes que compram mais ativamente, determinado produto ou marca, bem como outras informações sobre o seu estilo de vida e informações sobre o seu poder de compra, através da sua afinidade por marcas. Todas estas teorias contribuíram para construção do modelo utilizado neste Relatório de Estágio.

2.4. Formulação e explicação de hipóteses

O tema deste trabalho, como já foi anteriormente mencionado, é a segmentação dos clientes do Continente *Online* e a posterior definição do Perfil, ou dos Perfis dos mesmos.

O objetivo da construção de perfis, é a possibilidade de prever o comportamento do cliente (Jansen 2007), o que para empresa representa uma vantagem competitiva, criando um canal de comunicação mais eficiente e direcionado para o seu cliente (McDonald, Dunbar, 1998).

Através deste conhecimento, é possível ir de encontro às necessidades do cliente, o que se traduz numa maior fidelização, num aumento do valor e numa maior satisfação do cliente (Reichheld, Sasser, 1990:105).

Analisando os dados disponíveis, foi necessário averiguar, qual ou quais os atributos, que levam à criação dos perfis, mais adequados ao cliente do Continente *Online* (Jansen 2007).

As seguintes hipóteses pretendem averiguar, de entre os dados disponíveis, qual ou quais os atributos que levam à construção do melhor perfil:

H1: O perfil do cliente do Continente *Online* deve ser definido com base em atributos factuais, nomeadamente, o género, a idade e localização geográfica.

H2: O perfil do cliente do Continente *Online* deve ser definido com base em atributos comportamentais, mais concretamente, através dos produtos comprados.

Ambas as hipóteses pretendem chegar a um perfil, que permita determinar qual o comportamento futuro dos clientes com esse perfil.

O que difere entre as hipóteses são a segmentação utilizada e o tipo de atributos selecionados para a criação do perfil.

Para testar a hipótese H1, será construída uma segmentação *Market Driven* (Rud, Parr, 2001) e serão utilizados atributos fatuais, mais concretamente, a idade, o género e o local de residência. Estes dados provém do Google Analytics e permitirão segmentar os clientes, por atributos, numa abordagem semelhante, à desenvolvida em (Pratt, 2011) e (Chang, K. Kim, H. Kim, 2002).

Em seguida, será definido o perfil do cliente, que tem maior probabilidade de recorrer ao *e-commerce*, para comprar os seus produtos de Retalho Geral. Definido o perfil, este será comparado com o perfil encontrado no estudo conduzido pela Marktest (Marktest, 2012).

A hipótese H2, será construída recorrendo a uma segmentação *Data Driven* (Scridon, 2008), com a utilização do *Data Mining* (Breiman, Friedman, Stone, 1984) e com aplicação do algoritmo de *Clustering* K-Means (Chen, 2012) e, (Migueis, Camanho e Cunha, 2011).

Os atributos utilizados, serão comportamentais, mais concretamente, as categorias de produto comprados pelos clientes (Yankelovich, Meer, 2006), (Franzak, 2001:1-2). A definição do perfil ou dos perfis será realizada através da análise dos *Clusters* encontrados (Chen, 2007).

Por fim, o que determinará o perfil que melhor se adequa às necessidades da Modelo.com serão os seguintes objetivos:

- Prever o comportamento do cliente.
- Identificar os clientes, com maior relevância para a empresa.
- Apoiar os *Marketers* no desenvolvimento de campanhas promocionais e no desenvolvimento de um sistema de recomendação.

2.5. Apresentação e breve explicação do modelo de segmentação, construído com o algoritmo K-Means

Neste relatório, foram realizados, dois tipos de segmentação, uma *Market Driven* e uma *Data Driven*.

Para a construção da primeira, apenas foi necessário recorrer a técnicas estatísticas simples que não necessitam de uma explicação profunda.

Para a construção da segunda já foi necessário a utilização de ferramentas estatísticas complexas, denominadas por *Data Mining*, mais concretamente a técnica de *Clustering* com recurso ao algoritmo K-Means.

Por isso, a segunda segmentação será alvo de uma análise mais profunda e crítica, recorrendo a outros trabalhos realizados neste campo.

Esta segmentação baseou-se em atributos comportamentais focando-se nas categorias de produto compradas pelos clientes, numa abordagem semelhante à utilizada por (Migueis, Camanho e Cunha, 2012).

Esta metodologia utiliza um algoritmo ou técnica de *Clustering* para encontrar os grupos ou *Clusters* de clientes. Este algoritmo, denominado por K-Means, foi também utilizado num outro trabalho realizado por (Migueis, Camanho e Cunha, 2011) e por outros autores nomeadamente em (Chen, 2012).

Depois de identificado cada *Cluster* será analisado para identificar os produtos mais consumidos. Cada *Cluster* terá maior afinidade por determinados produtos, ou seja, os clientes, que pertencerem a esse Cluster, compram com maior frequência um certo tipo de produtos.

Por exemplo, num determinado Cluster legumes e frutas apresentam-se como os produtos mais consumidos. Isto permitirá atribuir um perfil aos clientes, com base nos produtos consumidos.

Neste caso, podemos inferir, que se trata dum cliente, com elevada probabilidade de consumir estes produtos.

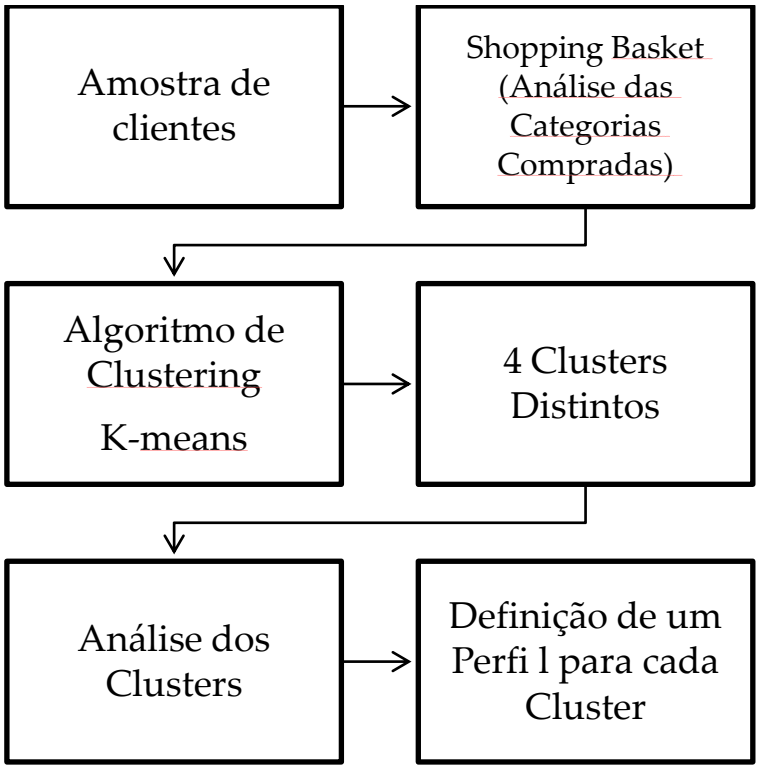

Ilustração 4-Estrutura da Segmentação Data Driven e da construção dos Perfis

3. Metodologia

3.1. Descrição das fontes dos dados utilizados e definição da amostra

As fontes dos dados utilizados, para construir as segmentações, tiveram duas proveniências diferentes. Da abordagem *Market Drive* (Rud, Parr, 2001), os dados foram obtidos, através do Google Analytics, tema que será explorado no próximo subcapítulo.

Relativamente à segmentação, *Data Driven*, os dados provém diretamente da plataforma do Continente *Online*, (Scridon, 2008), (Jansen, 2007) e (Scrindon, 2008).

A obtenção dos dados do cliente inicia-se quando este efetua o registo na plataforma, introduzindo os dados necessários para realizar as suas compras, nomeadamente, o nome, o NIF (Número de Identificação Fiscal), o correio eletrónico, o contacto telefónico e a morada de entrega.

O correio eletrónico do cliente é fundamental, pois permite identificar o cliente e serve de meio de contacto, para o caso de ser necessário prestar-lhe assistência. Para o presente relatório, este dado foi o meio escolhido, para identificar o cliente como único.

Todas as transações realizadas pelos clientes na plataforma *online* (www.continente.pt), são registadas num servidor operacional e armazenadas em bases de dados.

Para o estudo em questão, foi utilizado um servidor, que contém várias bases de dados, nomeadamente a utilizada para este relatório, onde se encontrava a informação necessária.

Dentro desta base de dados existem várias tabelas, entre as quais, as que contêm a informação sobre os clientes e as transações.

Antes da definição da amostra, foi analisada a informação disponível nas tabelas, para avaliar qual seria a melhor abordagem, tendo em conta o objetivo da segmentação dos clientes pelas suas compras.

Foi então necessário, escolher qual dos cinco níveis da estrutura mercadológica (Direção, Unidade de Negócio, Categoria, Subcategoria e Unidade base) utilizada pela empresa, faria mais sentido utilizar, tendo em conta a qualidade da informação de cada nível e a quantidade de dados existentes em cada nível.

Foi escolhido o nível de Categoria, que sendo um nível intermédio, ainda permite fazer a distinção entre, por exemplo, conservas e refrigerantes, não entrando num nível demasiado fino, por exemplo, conservas de milho, o que geraria uma enorme quantidade de dados, obrigando a reduzir o período de análise.

Nesta fase definiu-se quais seriam os dados a utilizar na construção do modelo, ou seja, o *email* e as categorias dos produtos comprados pelos clientes.

Recorrendo ao *SQL Server Management Studio*2014, foi possível extrair os dados distribuídos por várias tabelas da base de dados, agregá-los e tratá-los. Para tal, foi construído um *Query* (*Structured Query Language*), que permitiu obter os dados finais com a informação pretendida, o *email* do cliente e a contagem das categorias compradas por cada cliente.

A estrutura dos dados obtidos a partir do *Query*, apresentam a seguinte configuração:

Email	alimentação infantil	mel e compotas	bebidas quentes	cervejas	águas	cuidados do corpo
Cf45@hotmail.com	12	3	0	18	10	1
12345@gmail.com	6	9	4	6	20	3

Tabela 4-Estrutura dos dados obtidos a partir do Query

Este *Query*, foi utilizado em todos as transações dos clientes desde Fevereiro de 2014 até a Outubro de 2014. Não foram utilizadas transações ocorridas antes de Fevereiro de 2014, porque a estrutura mercadológica, sofreu alterações o que implicaria a construção de um novo Query.

Uma preocupação que se teve em conta na seleção da amostra foi a utilização de dados sobre as transações, que fossem o mais recentes possível.

Para a definição da amostra, também foram tidas em conta as limitações da ferramenta, o SQL Server Data Mining Add-ins. Tratando-se uma ferramenta que realiza o processamento de dados no Microsoft Office Excel 2010, existe limitação quanto ao número máximo de linhas, 1.048.576 de linhas e quanto ao número máximo de colunas, 16.384 colunas.

É de salientar, a importância de dois dados sobre o perfil do cliente da Modelo.com, obtidos através de vários trabalhos internos realizados na empresa. Estes dados foram fundamentais para a definição da amostra, sendo eles, o número médio de itens por compra e o número médio de dias entre cada compra por cliente.

O número médio de itens por compra é igual a 30 itens. Este valor ajudou na definição de um valor mínimo de itens por encomenda para a amostra, para excluir encomendas atípicas e que podiam enviesar os resultados.

Só foram contempladas para a amostra, encomendas ou transações, com um número de itens maior ou igual a 15 itens por encomenda. Assim foi possível limpar os dados e excluir possíveis *outliers*.

Esta prática foi também utilizada, por (Migueis, Camanho e Cunha, 2012) que para a definição da sua amostra de entre a população, apenas consideraram as transações, em que os clientes compraram pelo menos 10 produtos diferentes, eliminando assim possíveis *outliers*.

O número médio de dias entre cada compra por cliente, que demonstra a periocidade média de compra por cliente, assume o valor de 36 dias, ou seja, em média cada cliente realiza uma transação por mês, aproximadamente.

Sabendo que em média, cada cliente realiza uma compra a cada 36 dias, o período definido para a amostra, contempla três meses. Este período inclui pelo menos duas transações por cliente e evita possíveis impactos sazonais, o que poderia ocorrer, caso se escolhesse apenas um mês.

A amostra, que reúne todas as imposições mencionados, contem 19548 clientes que efetuaram transações durante o período de 01 de Agosto de 2014 a 01 de Novembro de 2014.

3.2. Dados provenientes do Google Analytics

A obtenção de dados para a segmentação *Market Driven*, foi feita com recurso ao Google Analytics.

Para efeitos comparativos, definiu-se o período de recolha de dados, igual ao período da amostra da segmentação *Data Driven*, ou seja, 01 de Agosto de 2014 a 01 de Novembro de 2014.

A informação fornecida pelo Google Analytics, não está associada às transações, apenas representa o perfil do visitante do site. O perfil do visitante é

construído, com base em informações recolhidas através de *cookies* (informação sobre a navegação gravada no dispositivo) e dos motores de busca (Fettman, 2014).

Existem algumas limitações na recolha das informações, dado que é possível eliminar as *cookies*, perdendo-se assim o histórico da informação de navegação. Outra limitação é o fato de vários indivíduos, utilizarem o mesmo dispositivo e a mesma conta de acesso, o que pode enviesar os resultados (Fettman, 2014).

Contudo os dados recolhidos, permitem encontrar os segmentos existentes de uma forma simples, conforme os atributos escolhidos, sendo estes, o género, a idade e localização geográfica.

Posteriormente é possível definir qual o perfil do cliente do Continente *Online*, analisando qual o conjunto de atributos, que define o cliente, com maior probabilidade de comprar através do *site*, numa abordagem semelhante à desenvolvida por (Morganosky, Cude, 2000), (Raijas, Tuunainen, 2001) e (Marktest, 2012).

3.3. Dados provenientes do Questionário

Foi realizado um questionário, com o intuito de aprofundar o conhecimento sobre o comportamento dos Portugueses no *e-commerce*. Além disso, pretendia-se comparar, os resultados obtidos através do Google Analytics, com os resultados deste questionário.

As questões iniciais abrangem o *e-commerce* em geral, evoluindo para o setor do Retalho Alimentar, numa tentativa de perceber o potencial e fragilidades desta área.

Este questionário foi publicado em quinze grupos da rede social *Facebook*, onde foi solicitado aos seus membros, que preenchessem este questionário com o maior rigor, explicando a importância da veracidade das respostas. Também foi comunicado a todos os membros, que este questionário seria utilizado na construção numa tese de Mestrado.

Estes grupos têm vários objetivos, desde divulgação de empresas e negócios, a motivos lúdicos e sociais. O número de membros pertencentes aos grupos varia entre 600 a 22.000 membros.

A limitação inerente a este meio de divulgação é que as respostas podem ter sido dadas por utilizadores que tenham empatia com o autor deste relatório o que pode enviesar os resultados.

Contudo, considerou-se que o canal utilizado para divulgar o questionário seria o mais adequado, quer pela facilidade na divulgação, quer pelo facto do perfil dos frequentadores das redes sociais se aproximar do perfil dos utilizadores do *e-commerce*.

No decorrido período obteve-se um total de 262 respostas que podem ser observadas em maior detalhe na secção dos Anexos.

3.4. Descrição estatística das variáveis do modelo de segmentação *Data Driven*

Na apresentação dos resultados da segmentação *Market Driven*, já será feita uma análise estatística, que implica a descrição das variáveis, pelo que neste subcapítulo, apenas serão descritas, as variáveis utilizadas na construção da segmentação *Data Driven*.

Comecemos pela análise da estrutura das variáveis, a estrutura mercadológica. Nesta, estão esquematizadas, todas as atividades desenvolvidas pela empresa, numa organização hierárquica, partindo do nível mais abrangente, a Direção Comercial (DC), para o nível mais fino, a Unidade Base (UB).

dc	un	cat	scat	ub
administrativa	charcutaria& queijos	cereais	vinagre	azeites bio
alimentar	mercearia salgada	bolachas	molhos mesa	mostarda
bazar	higiene e beleza	refeições	swf-temperos	amendoím
casa	limpeza do lar	aperitivos	swf-gord. liquidas	outros óleos
cons/gg e fardamento	takeaway	sobremesas	conservas peixe	swf-gord. liquidas

Tabela 5-Exemplo da Estrutura Mercadológica

A estrutura mercadológica é utilizada pela empresa Modelo.com e pela Sonae SGPS, pelo que, estão presentes direções comerciais de todo o Grupo.

Como o foco deste relatório, são os produtos do Retalho Geral ou FMCG, comercializados pela Modelo.com, apenas se selecionaram os níveis da estrutura, relacionados com esta área do Retalho.

Na seguinte tabela é possível observar a distribuição dos diferentes níveis e o seu número:

Tipologia	Descrição	Contagem	Exemplo
dc	Direção Comercial	15	Alimentar
un	Unidade de Negócio	81	mercearia salgada
cat	Categoria	438	gorduras liquidas
scat	Sub Categoria	1768	Azeite
ub	Unidade Base	6184	azeite dop

Tabela 6-Exemplo da distribuição dos diferentes níveis

Dentro das 438 categorias existentes na estrutura mercadológica, selecionou-se 127 Categorias, que representam os produtos de Retalho Geral.

Através da observação da tabela e do gráfico seguinte, é possível compreender o volume de dados que seria gerado, se fosse utilizado um nível mais baixo, que a estrutura mercadológica, por exemplo, a Unidade Base e que complicaria a interpretação dos resultados.

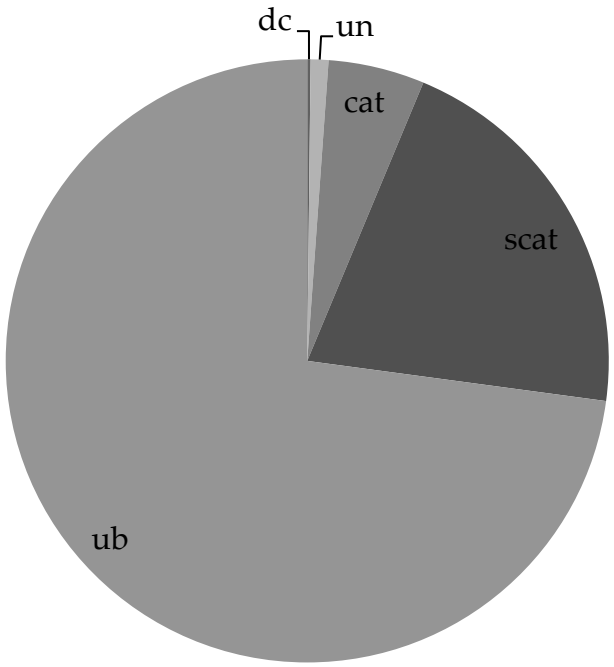

Ilustração 5**-Gráfico da distribuição da Estrutura Mercadológica**

Depois de definida a amostra, foram selecionados aleatoriamente, 60% dos clientes e as respetivas transações.

As transações que cada cliente realizou, foram agregadas ao nível de categoria. Ou seja, quando o cliente compra várias unidades de negócio dentro duma categoria, por exemplo, doze vezes bolachas integrais, é contabilizado que o cliente comprou 12 vezes essa categoria.

Email do Cliente	doçaria	águas	aperitivos	iogurtes	legumes
12345@netcabo.pt	2	14	23	2	5
abcdef@gamil.com	5	2	8	20	0

Tabela 7-Exemplo da contabilização das categorias compradas

Em seguida, foi necessário analisar qual seria a melhor metodologia para permitir uma leitura das variáveis, pelo algoritmo de *Clustering*. A solução encontrada foi, a transformação das variáveis em variáveis binárias, que permitem uma aplicação eficaz do algoritmo.

A transformação para variáveis binárias, teria que refletir o número de vezes que determinada categoria foi comprada por cada cliente e o peso dessa categoria no total de categorias compradas.

Para tal, aplicou-se um rácio entre a quantidade comprada de categoria X e a quantidade total de categorias compradas por cada cliente, de forma a identificar as categorias mais compradas.

Email do Cliente	frango	doçaria	águas	iogurtes	Legumes	Total
12345@netcabo.pt	15	2	14	2	5	38
Percentagem	0,39	0,05	0,37	0,05	0,13	1

Tabela 8-Exemplo do rácio, que demonstra o peso da categoria no total de categorias compradas

A interpretação dos valores deste rácio, implicou a decisão sobre o valor que seria transformado em 1, ou seja, representando uma categoria comprada com frequência e com um peso maior nas compras de cada cliente e qual o valor que assumiria 0, uma categoria não comprada com frequência suficiente.

Foram testados cinco valores, para a transformação das variáveis em binário, com o objetivo de encontra as categorias, com a maior relevância nas compras dos clientes.

Na tabela seguinte, podemos observar as diferenças no número de categorias, transformadas para binário:

Valores	100%	10%	5%	4%	3%	2%
Total Binário	722108	23653	89754	122662	169135	227756
Categorias/Total Categorias	1	0,03	0,12	0,17	0,23	0,32

Tabela 9- Exemplo dos valores da transformação para binário

A coluna que assume o valor 100% representa a situação em que todas as categorias são transformadas para binário, independentemente do valor do rácio. Analisando a tabela, é possível identificar a informação que cada valor acrescenta, partindo do valor 10%, até ao valor 2%.

Supondo que se utilizou, o valor 10%, o procedimento de conversão para uma variável binária, foi a seguinte:

Email do Cliente	frango	doçaria	águas	iogurtes	legumes	Total
12345@netcabo.pt	15	2	14	2	5	38
Percentagem	0,39	0,05	0,37	0,05	0,13	1
Conversão para Binário	1	0	1	0	1	4

A definição do valor do rácio utilizado para a transformação das variáveis em variáveis binárias, será demonstrado no capítulo dos Resultados, já que implicou a aplicação do algoritmo de *Clustering*.

3.5. Descrição do Software Utilizado

O SQL Server Management Studio (SSMS), é um software que permite aceder, configurar, gerir, administrar e desenvolver todos os componentes do SQL Server. O SSMS combina um grupo alargado de ferramentas gráficas com um layout acessível, fornece uma experiência mais rica e facilitada na utilização do SQL Server.

O SQL Server, um programa desenvolvido pela Microsoft, apresenta inúmeros serviços relacionados com o acesso, gestão e extração de informação, a partir de bases de dados. Na próxima tabela estão presentes algumas das funções do SQL Server:

Serviço do SQL Server	**Principal Função**
Database Engine	Armazenamento de dados OLTP (Processamento de Transações em Tempo Real)
Reporting Services	Relatório de dados
Analysis Services	Análise de dados OLAP (Processamento Analítico *Online*)
Integration Services	Fluxos de dados
Data Quality Services	Limpeza de dados
Master Data Services	Repositório único de dados
Replicação	Replicação de dados entre servidores

É de salientar a importância do Analysis Services do SQL Server, já que é esta a funcionalidade, que permite criar, gerir e explorar modelos de *Data Mining*.

É através duma ligação ao Analysis Services , que o SQL Server Data Mining Add-ins(Data Mining Client for Excel)opera, permitindo a criação e utilização de Modelos de Data Mining.

O SQL Server Data Mining Add-ins, está dotado de algoritmos de *Data Mining*, como algoritmos de Classificação, de Associação, de Clustering, permitindo realizar Segmentações de Clientes, Previsões de Vendas, Análise de *Baskets*, entre outras aplicações.

Existem algumas limitações em termos da quantidade de dados que podem ser processados, já que a ferramenta, está integrada no Microsoft Excel.

3.6. Definição do Método Principal: Algoritmo K-Means do software SQL Server Data Mining Add-ins

Tendo já sido definido e explicado o Algoritmo K-Means, é necessário demonstrar o funcionamento do algoritmo do *Software* em causa, visto ter algumas diferenças.

Como já foi referido, o Algoritmo K-Means, é um método de *Clustering* bem conhecido e utilizado em situações, onde se pretende agrupar um conjunto de observações duma amostra num grupo ou *Cluster*.

O algoritmo agrupa observações semelhantes em grupos, procurando minimizar as diferenças entre observações do mesmo *Cluster* e ao mesmo tempo, maximizar a distância entre *Clusters*.

A palavra "*Means*" refere-se ao centróide do *Cluster* que é um ponto inicial de dados escolhidos de forma arbitrária. Durante o processo de construção dos *Clusters*, o centroide irá sofrer alterações e correções até representar a verdadeira média das observações presentes no *Cluster*.

O " K " na expressão K-Means refere-se ao número arbitrário de pontos, que são utilizados para iniciar o processo de agrupamento das observações em *Clusters*.

O algoritmo calcula as Distâncias Euclidianas entre as observações dentro de um *Cluster* e o vetor que representa a média desse *Cluster*. Quando o valor das distâncias e do vetor são mínimos é porque se atingiu o número final de (K) *Clusters*.

Uma das grandes vantagens do K-Means, é que atribui uma observação a um só *Cluster*, excluindo qualquer hipótese dessa observação não pertencer ao *Cluster*. O que determina se a observação pertence ao *Cluster* é a distância entre a observação e o centroide.

A grande diferença entre Algoritmo K-Means existente no *Software* utilizado e o Algoritmo K-Means referido na literatura estudada, está na definição dos (K) ou *Clusters*, ou seja, nas condições *a priori*.

O "CLUSTER_COUNT", parâmetro utilizado neste algoritmo para definir o número inicial de *Clusters* difere do tradicional (K) porque, ainda que se especifique o número de *Clusters* pretendido, não é uma imposição rígida. Caso o algoritmo não seja capaz de construir a partir dos dados, o número de *Clusters* pretendido, este encontra o número ideal de *Clusters* com os dados disponíveis.

Outra característica deste algoritmo é que permite que não se utilize qualquer valor no parâmetro "CLUSTER_COUNT". Ao atribuirmos um valor igual a {0} a este parâmetro o algoritmo é capaz de determinar o melhor número de *Clusters* a construir, recorrendo a métodos heurísticos.

Este algoritmo, corrige um dos principais problemas apontados pela literatura, nomeadamente por (Venkatesan, 2007), na construção de modelos de *Clustering* recorrendo ao K-Means.

Ainda assim, é recomendada a utilização de um método semelhante à " *Elbow Curve*" para avaliar qual o melhor número de *input* de *Clusters* a utilizar.

O método consiste em utilizar vários valores e utilizar a média dos *Clusters* criados. O *input* que leva a um número final de *Clusters* criados mais próxima da média aproxima-se do valor ideal.

O outro parâmetro necessário para a utilização do algoritmo K-Means são as *Seeds*, definidas neste *Software* como "CLUSTER_SEED". Aqui não existe diferença face ao tradicional algoritmo K-Means.

Especifica-se o valor do parâmetro, ou seja, o número de *Seeds* que posteriormente será utilizado para gerar aleatoriamente os *Clusters*, na etapa inicial da construção dos modelos.

O método de teste do número de *Seeds* a utilizar é igual ao proposto pela literatura (Venkatesan, 2007), ou seja, alterando os valores utilizados, o que muda a forma inicial como os *Clusters* são criados.

Se os *Clusters* não se sofrem grandes modificações, quando se altera as *Seeds*, é possível considerar o modelo estável e robusto.

4. Resultados

4.1. Análise descritiva dos dados com base em estatísticas básicas

Tendo em conta que foram elaborados dois tipos de segmentação a demonstração dos resultados obtidos será realizada em dois subcapítulos separados.

4.1.1. Segmentação *Market Driven*

Neste, o tema central, é o resultado da segmentação, com a utilização de atributos fatuais.

Os dados obtidos através do Google Analytics, permitiram concluir que existem grupos de clientes com maior propensão, para realizar compras de produtos FMCG, na plataforma do Continente *Online*. Para uma discussão dos resultados, mais completa, foram comparados os resultados obtidos através do Google Analytics e dos resultados do questionário.

Atribubos	**Percentagem**
Género	
Feminino	67,93%
Masculino	32,07%
Idade	
18-24	12,55%
25-34	38,10%
35-44	28,54%
45-54	10,07%
55-64	7,66%
65+	3,09%
Localidade	
Lisbon	48,15%
Porto	20,22%
Vila Nova de Gaia	5,91%
Coimbra	5,48%
Amadora	5,26%
Maia	4,30%
Braga	3,73%
Sintra	3,55%
Oeiras	3,40%

Tabela 10-Resultados do Google Analytics em função dos atributos

A segmentação está implícita na observação dos atributos, por exemplo, existem dois segmentos dentro do atributo Género, que partilham as mesmas características, um grupo é composto por clientes do sexo feminino e o outro por clientes do sexo masculino.

Para a definição do perfil, procurou-se em cada atributo, qual o segmento, que exibia a maior percentagem.

No caso do género, o segmento feminino, apresenta o valor mais elevado, com cerca de 68%, face aos 32% do segmento masculino.

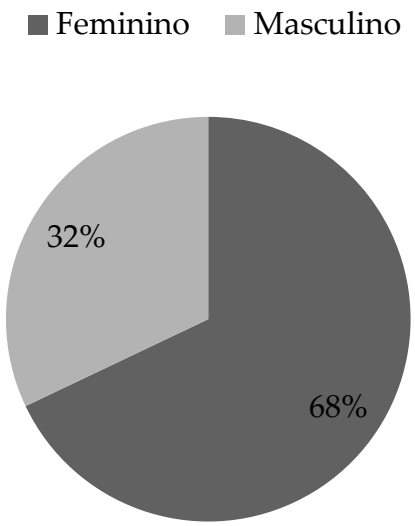

Ilustração 6-Distribuição do número de Sessões por Género

Nos resultados do questionário, o segmento feminino, também apresenta a maior percentagem, com 60%, que se aproxima dos 68% obtidos através do Google Analytics.

No caso da Idade, existem dois segmentos, que se destacam dos restantes. O segmento em que a idade está compreendida entre 25 e 34 anos e o segmento em que a idade está compreendida, entre 35 e 44 anos. Seguinte a lógica, foi escolhido o segmento com o maior peso, neste caso onde a idade, está entre os 25 e os 34 anos.

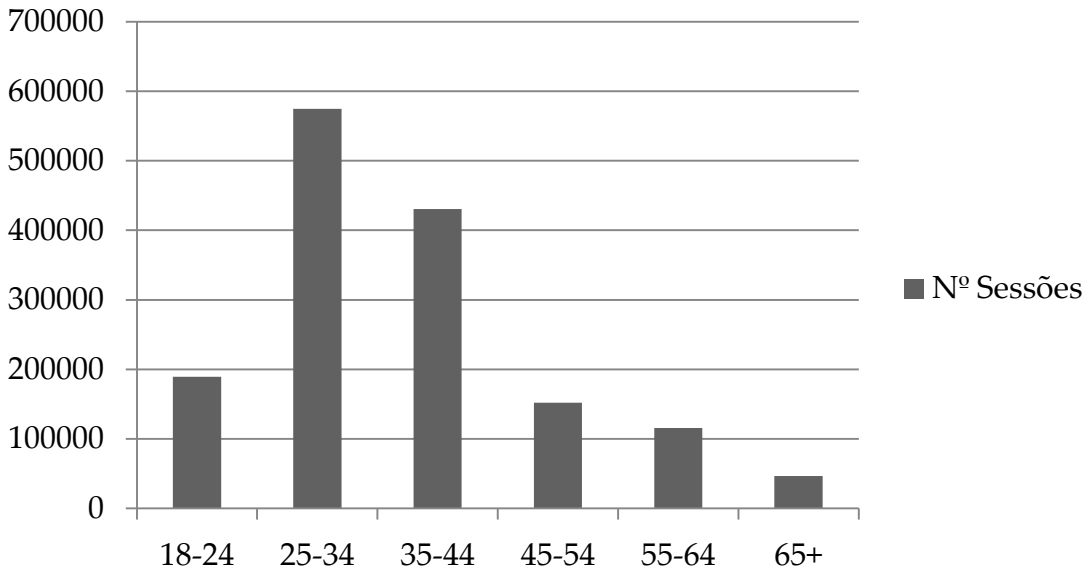

Ilustração 7-Distribuição do número de sessões por Idade

Também nos resultados do questionário, são as camadas mais jovens, que exibem maior propensão ao consumo. Contudo, o segmento que se destaca dos restantes, é o segmento com idade compreendida entre, 18 e 25 anos, seguido do segmento, com 25 a 35 anos. Este fato pode ser resultado, duma das limitações do meio utilizado para divulgação.

Por último, no atributo Localidade, é o segmento correspondente à cidade de Lisboa, que se destaca dos restantes com 48%, seguido da cidade do Porto, com 20%.

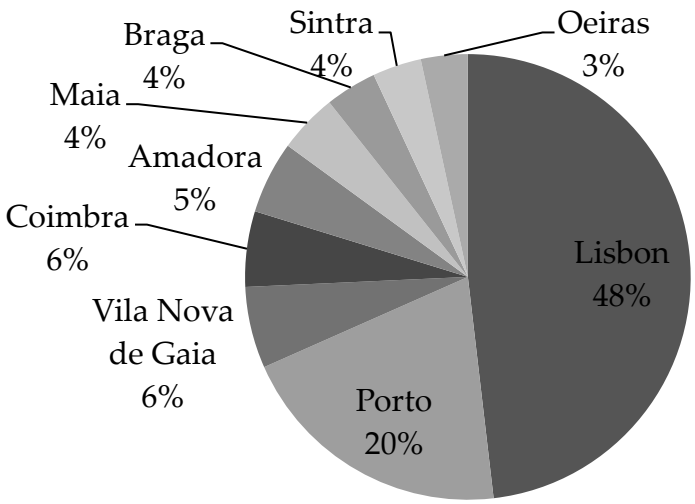

Ilustração 8-Distribuição do número de sessões por Localidade

Os resultados do questionário relativamente a este atributo, não foram parametrizados da mesma forma, sendo apenas dividido entre zonas, por exemplo, Norte, Centro e Sul. Contudo, o segmento com maior peso, foi o relativo à zona Norte do País, com 45%. Mais uma vez, este resultado pode dever-se às limitações do meio de divulgação.

Após a análise dos segmentos, concluiu-se que, o perfil do cliente do Continente *Online*, segundo o Google Analytics, pode ser definido, como um

cliente de sexo feminino, com idade compreendida entre os 25 e os 34 anos e residente em Lisboa.

4.2. Análise dos resultados principais via *Clustering*

4.2.1. Segmentação *Data Driven*

Depois de definida a metodologia e as variáveis a utilizar, deu-se início à construção do modelo de segmentação, recorrendo à técnica de *Clustering*, com o Algoritmo K-Means.

No capítulo da Descrição das Variáveis, foi atingido um ponto onde era necessário decidir sobre qual o valor do rácio das Categorias, para transformar em binário. Para tal era necessário aplicar o algoritmo para os cinco valores diferentes {0,02;0,03;0,04;0,05;0,1}.

Utilizando os mesmos parâmetros (Seed e Count Cluster), pretendia-se comparar os *Clusters* gerados, a partir dos cinco valores diferentes. Os resultados foram os seguintes:

Valor do Rácio	0,02	0,03	0,04	0,05	0,1
Clusters Gerados	9	9	10	6	9

Tabela 11- Demonstração do número de Clusters com valores do rácio

A ferramenta utilizada, gera imagens que demonstram a distribuição espacial dos *Clusters* gerados e as ligações entre si, que facilita a sua análise. Os conjuntos de *Clusters* criados foram os seguintes:

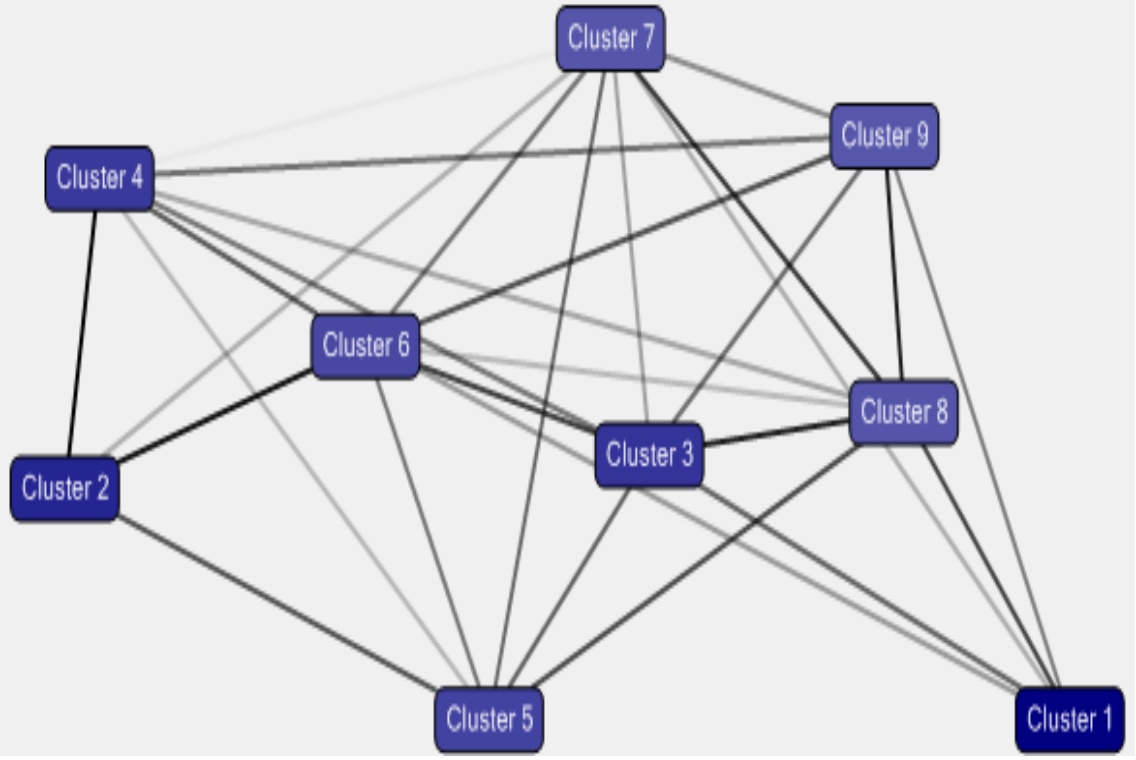

Ilustração 9-Disposição dos Clusters criados com o valor {0,02}

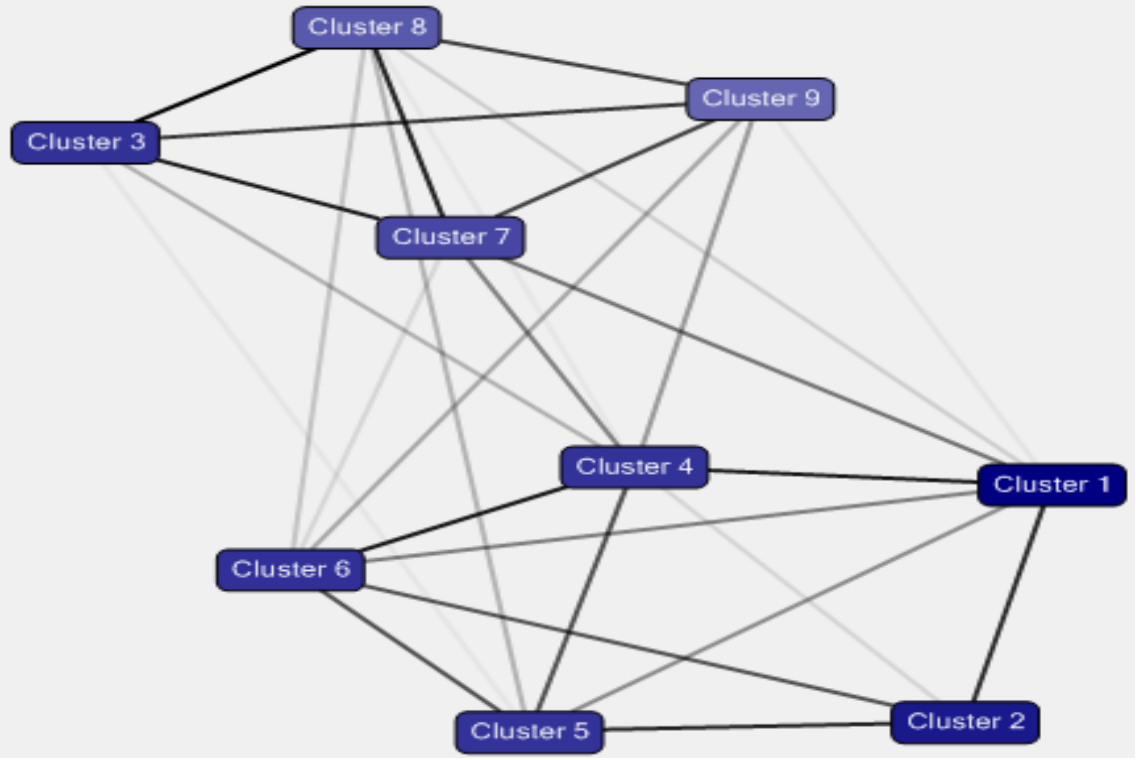

Ilustração 10-Disposição dos Clusters criados com o valor {0,03}

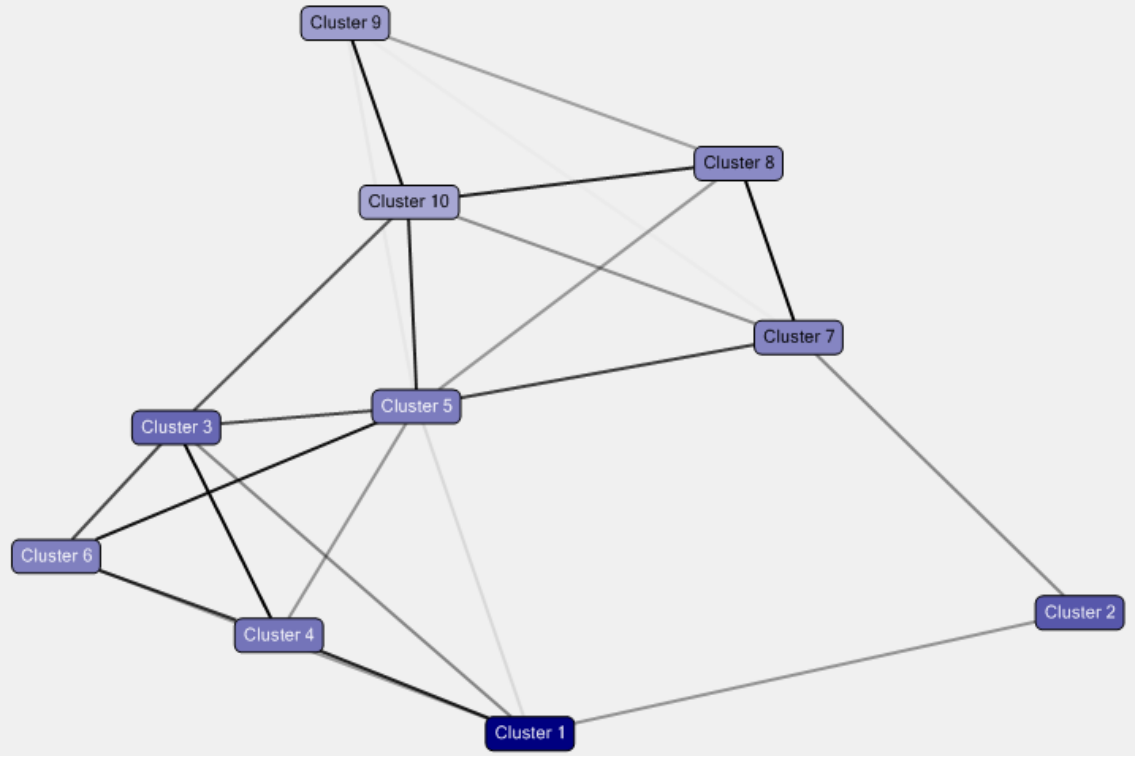

Ilustração 11-**Disposição dos Clusters criados com o valor {0,04}**

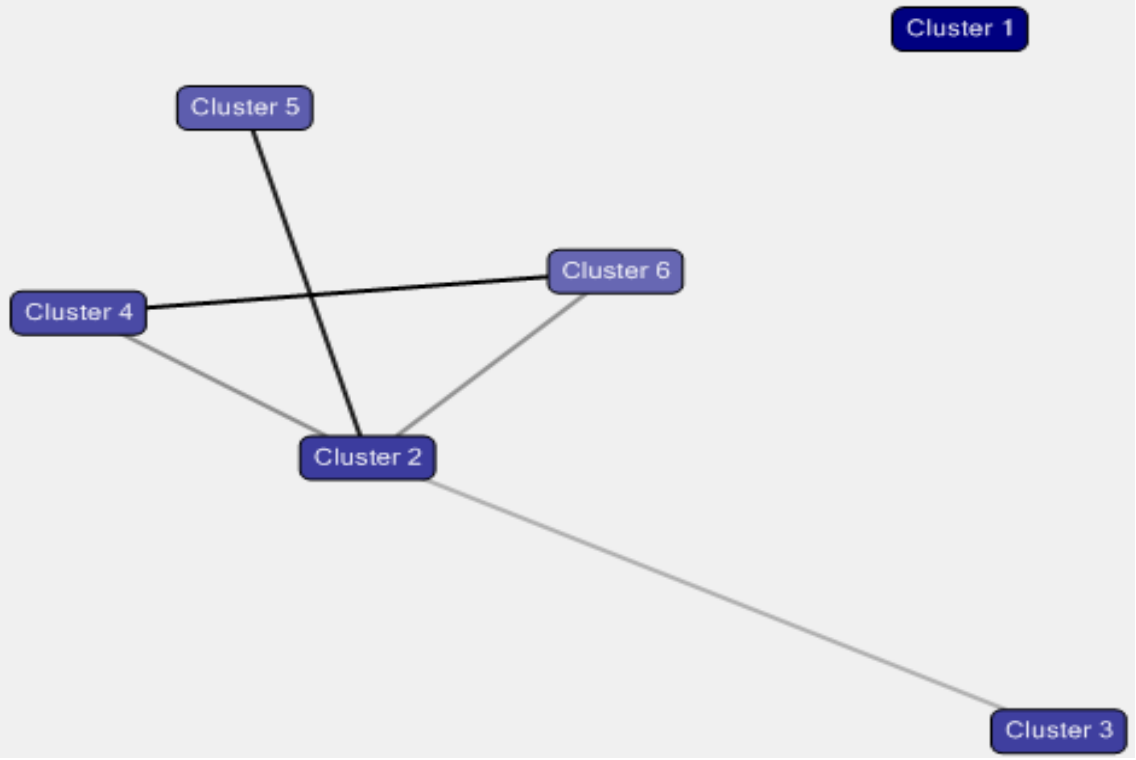

Ilustração 12-Disposição dos Clusters criados com o valor {0,05}

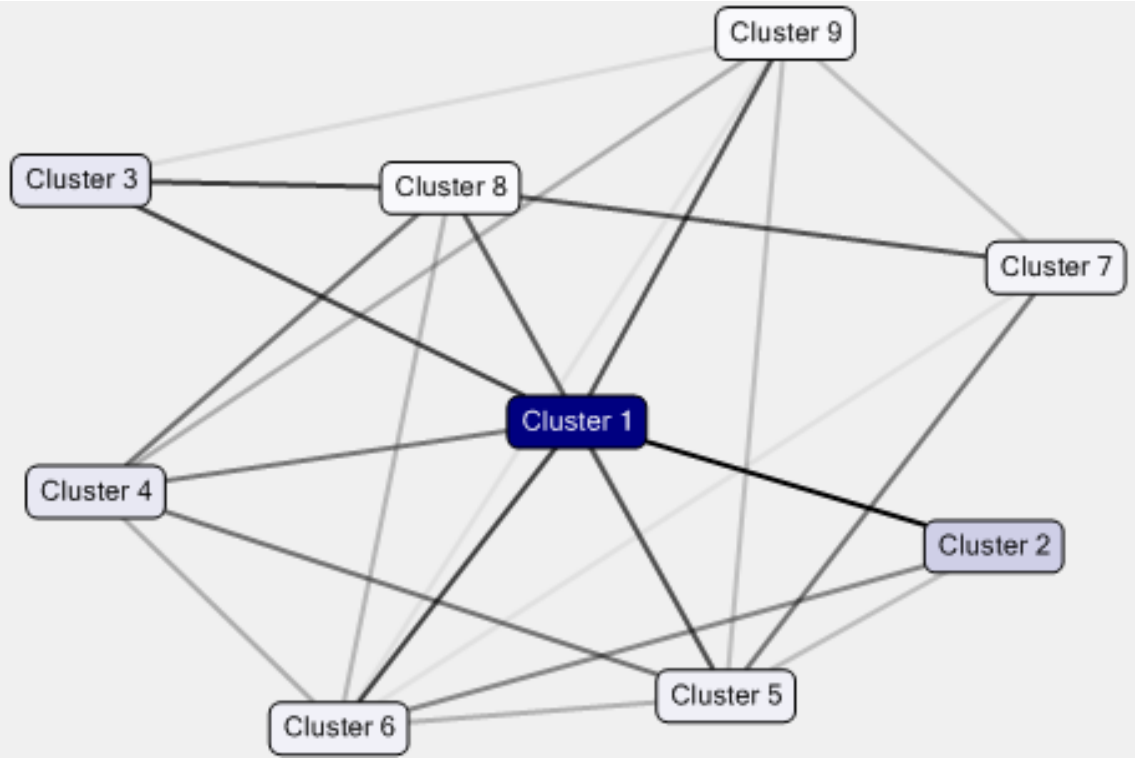

Ilustração 13-**Disposição dos Clusters criados com o valor {0,1}**

A resposta ao valor a utilizar provém das técnicas de Análise de *Clusters*. Estas remetem para uma procura de *Clusters* afastados uns dos outros, em termos espaciais, maximizando a distância entre *Clusters* e através da análise das ligações entre eles.

Através da observação das ligações entre os *Clusters*, ou seja, dos segmentos de reta que unem as estruturas onde estão estes inseridos é possível inferir sobre a distribuição das variáveis pelos *Clusters*.

As ligações que apresentam uma espessura e cor mais fortes conduzem a *Clusters* que estão mais relacionados entre si. Como o que se procurava, era descobrir *Clusters* que permitissem identificar grupos de Clientes pelas categorias compradas, uma distribuição heterogénea das variáveis pelos *Clusters* é essencial. Para tal, o cenário com menos ligações e apresentando ligações mais fracas, representa a melhor opção possível.

Tendo em conta estas duas técnicas e comparando os cinco cenários, é através da utilização do valor {0,05} que se obtém os melhores resultados. Em termos de disposição espacial são-nos apresentados dois *Clusters* que se

encontram afastados dos restantes. Na sua grande maioria, as ligações entre os *Clusters* são fracas e inexistentes em alguns casos.

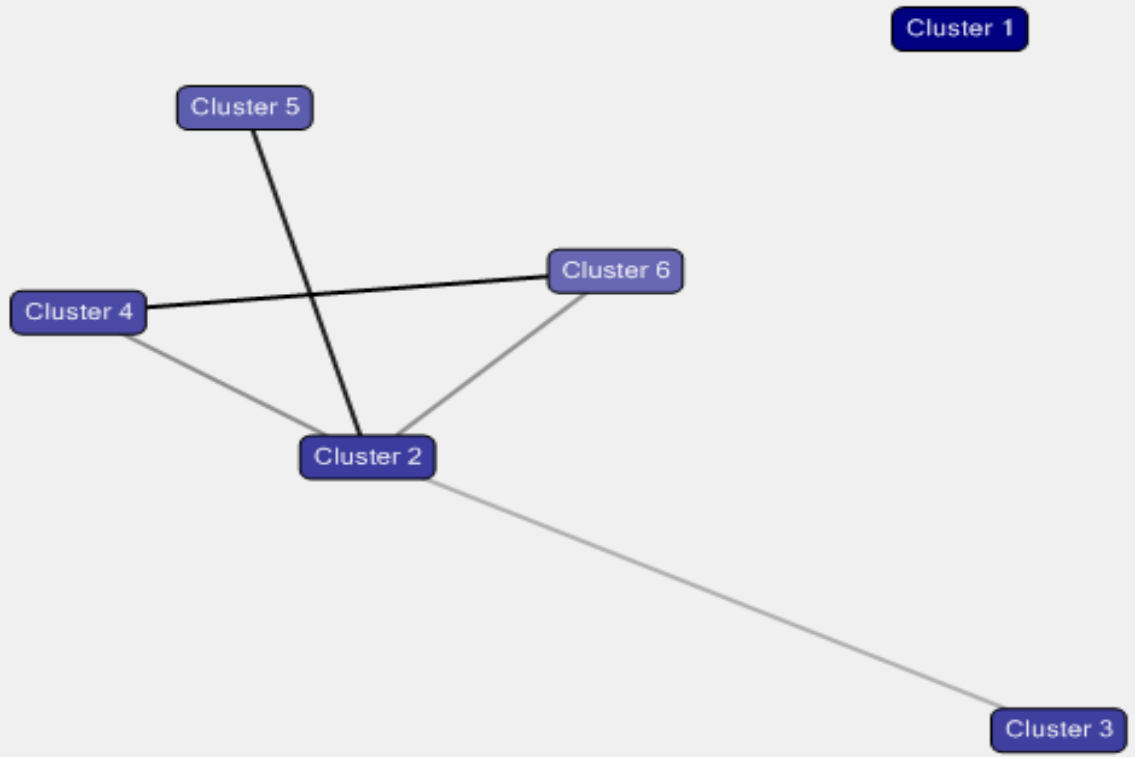

Ilustração 14-Disposição do valor escolhido {0,05}

Comparativamente, estes são os melhores *Clusters* e como tal o valor para conversão para binário é o mais adequado.

Para tornar mais robusta a escolha deste valor foi realizada uma comparação entre a distribuição das variáveis, ou seja, das categorias, pelos cinco cenários. O que se procurava eram *Clusters* com categorias "chave", categorias mais consumidas em determinado *Cluster*, que refletissem diferentes preferências de consumo.

A tabela seguinte representa uma parte da distribuição das categorias compradas, quando o valor utilizado para a conversão em binário é {0,05}. Os valores em percentagem representam a probabilidade da variável ou categoria ocorrer no *Cluster*.

Categoria	Total	Cl.1	Cl.2	Cl.3	Cl.4	Cl.5	Cl.6
águas	1780	9 %	5 %	0 %	14 %	1 %	100 %
alimentação infantil	643	7 %	6 %	9 %	5 %	7 %	7 %
aperitivos	575	4 %	7 %	7 %	5 %	10 %	6 %
auto	25	0 %	0 %	0 %	1 %	0 %	1 %
bacalhau	14	0 %	0 %	0 %	0 %	0 %	0 %
bagagens	1	0 %	0 %	0 %	0 %	0 %	0 %
banho	10	0 %	0 %	0 %	0 %	0 %	0 %

Tabela 12-Distribuição das categorias compradas com valor binário de {0,05}

Através da análise da tabela, é notória a existência de *Clusters* (Cl), com grande afinidade com determinada Categoria, como o caso do *Cluster* 6, onde o consumo de "águas", tem 100% de probabilidade de ocorrer neste *Cluster*. O mesmo acontece com outros *Clusters*, sendo este o tipo de distribuição que se procurava.

Se compararmos a distribuição das categorias com o cenário onde o valor é {0,04}, sendo este o segundo melhor cenário, em termos de distribuição espacial e ligações, é óbvio que a distribuição das variáveis é mais homogénea, não permitindo detetar as categorias "chave", identificando perfis de clientes distintos.

Categoria	Total	Cl.1	Cl.2	Cl.3	Cl.4	Cl.5	Cl.6	Cl.7	Cl.8	Cl.9	Cl.10
águas	2505	16 %	4 %	11 %	7 %	48 %	16 %	15 %	23 %	99 %	81 %
alimentação infantil	822	7 %	7 %	12 %	8 %	11 %	7 %	5 %	6 %	7 %	19 %
aperitivos	925	4 %	12 %	10 %	12 %	7 %	20 %	6 %	8 %	16 %	7 %
auto	58	0 %	1 %	0 %	0 %	0 %	1 %	0 %	2 %	2 %	1 %
bacalhau	50	0 %	1 %	0 %	0 %	0 %	0 %	0 %	1 %	2 %	1 %
bagagens	4	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
banho	18	0 %	0 %	0 %	0 %	0 %	0 %	1 %	0 %	0 %	0 %

Tabela 13-Tabela 12-Distribuição das categorias compradas com valor binário de {0,04}

Estando definido o valor a utilizar para conversão para variável binária, seguiu-se a definição dos valores dos parâmetros Seed e Count Clusters, segundo as melhores práticas possíveis.

O algoritmo utilizado é capaz de determinar o melhor número de *Clusters* dentro da amostra, através de métodos heurísticos. Para ativar esta opção basta estabelecer o valor para o parâmetro "Count Cluster" igual a zero.

Em seguida procurou determinar-se qual o melhor valor para o segundo parâmetro, as "*Seeds*".

Count Cluster	Seed	Nº Clusters	Média	NºClusters/Média	1-Média	Mediana	NºClusters/Mediana	1-Mediana
0	0	4	4,142	0,969	0,03	4	1	0
0	1	3	4,142	0,727	0,272	4	0,75	0,25
0	3	5	4,142	1,212	-0,212	4	1,25	-0,25
0	4	4	4,142	0,969	0,03	4	1	0
0	5	5	4,142	1,212	-0,212	4	1,25	-0,25
0	7	6	4,142	1,454	-0,454	4	1,5	-0,5
0	9	3	4,142	0,727	0,272	4	0,75	0,25
0	10	3	4,142	0,727	0,272	4	0,75	0,25

Tabela 14-Determinação do parâmetro Seed

Foram testados oito valores para o parâmetro "Seed", mantendo o "Count Cluster" constante. Ao analisar as Médias e as Medianas é possível identificar quais os valores das *Seeds*, que vão de encontro ao que se procura, ou seja, uma baixa variação quanto à Média e à Mediana. Os valores que apresentam melhores resultados estão coloridos a vermelho. Foi escolhido para comparação, um terceiro valor, representado a amarelo.

Depois de identificados os dois melhores resultados, procedeu-se a uma análise dos *Clusters*, idêntica à utilizada no passo anterior. Primeiro uma análise à disposição espacial e às ligações entre *Clusters* e em seguida, uma análise da distribuição das variáveis, ou das categorias.

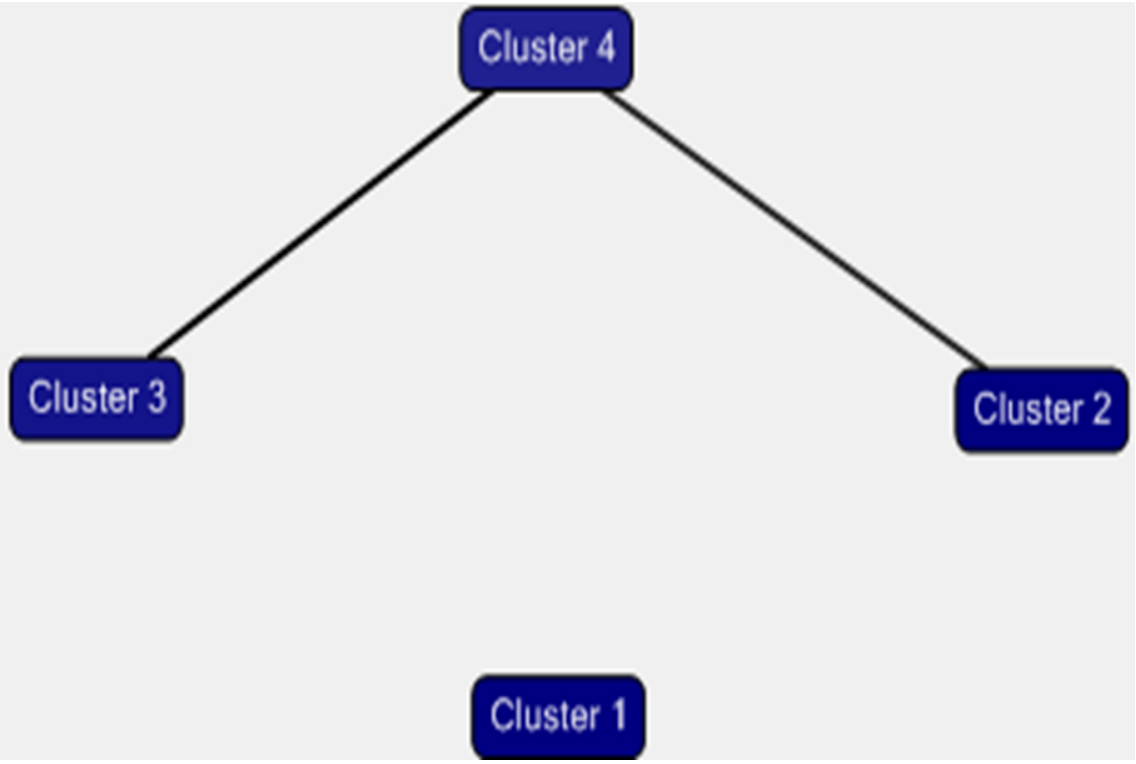

Ilustração 15-Disposição Espacial dos Clusters Count=0 e Seed= 0

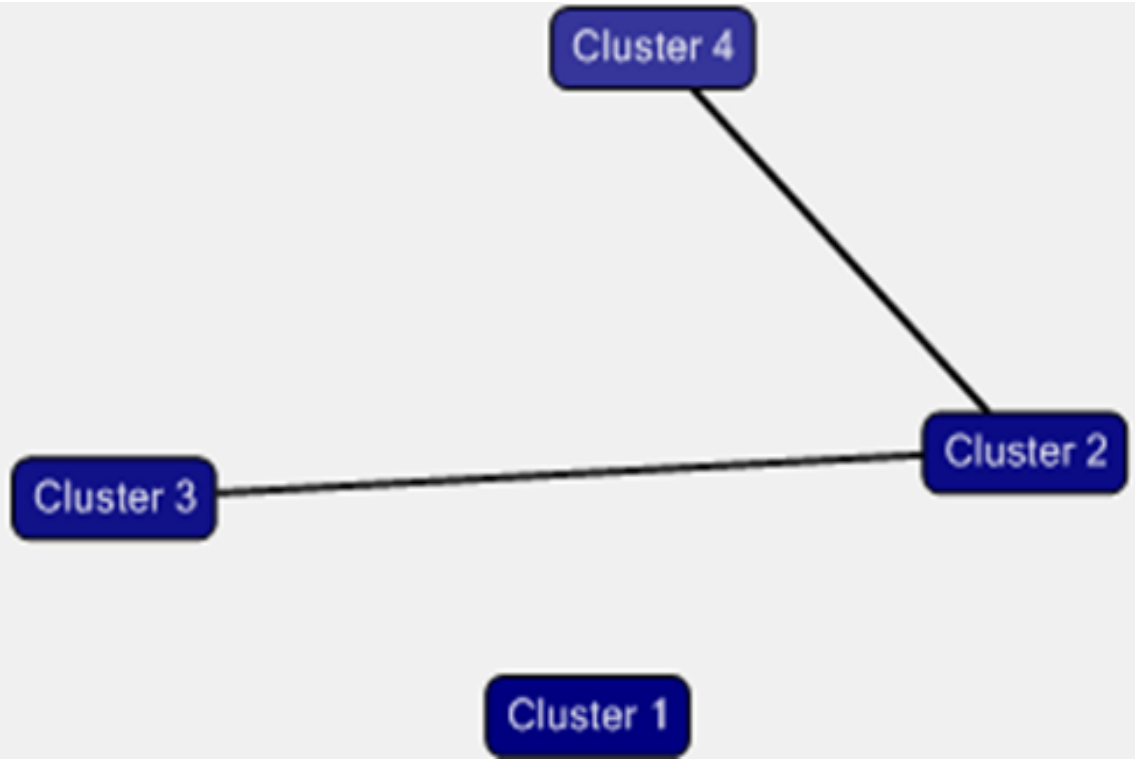

Ilustração 16-Disposição Espacial dos Clusters Count=0 e Seed= 4

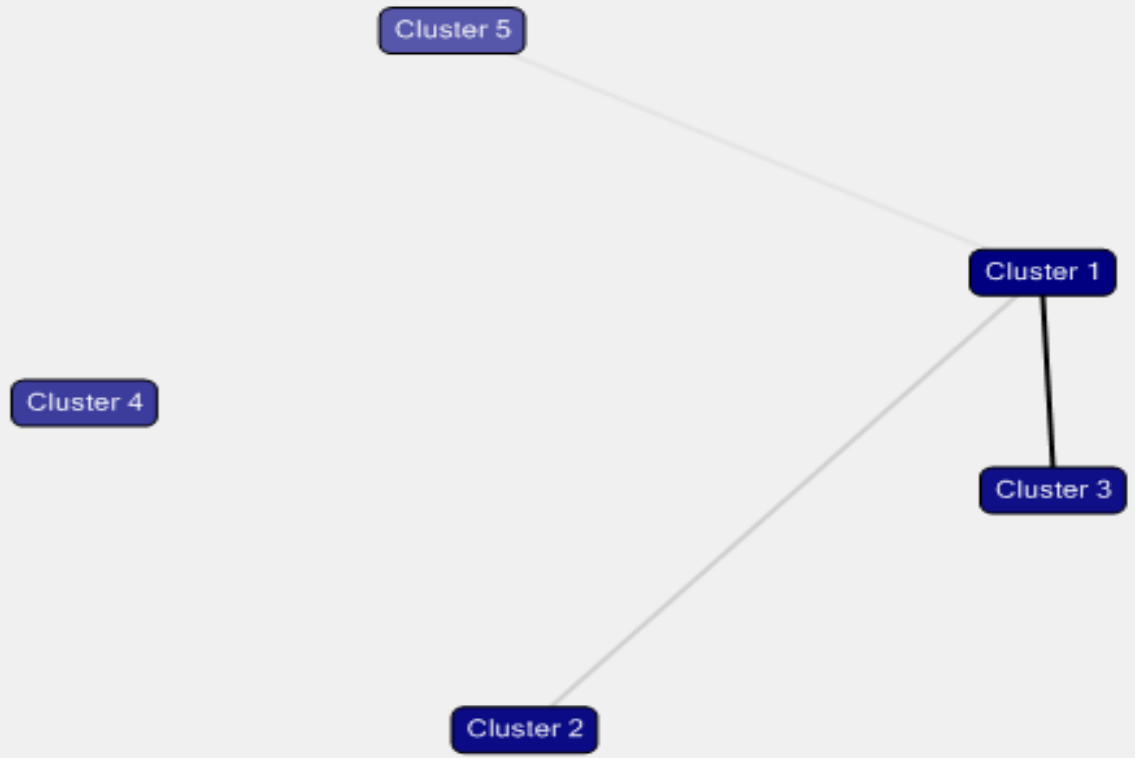

Ilustração 17-Disposição Espacial dos Clusters Count=0 e Seed= 5

Através da comparação entre os três cenários, é possível excluir o que apresentava o pior resultado. Entre outros dois a escolha foi baseada nas ligações já que o conjunto de *Clusters* com os parâmetros Count=0 e Seed= 4,

apresenta uma das duas ligações mais fina do que as duas ligações do outro conjunto de *Clusters* e com uma análise da distribuição das variáveis que apoiou a decisão.

A análise da distribuição das variáveis revelou o que se procurava; uma distribuição heterógena de categorias pelos *Clusters* e categorias "chave" em cada *Cluster*. Assim foi possível identificar perfis de consumo únicos em cada *Cluster*.

Definidos os parâmetros a utilizar, o próximo passo realizado foi a aplicação do algoritmo. Até aqui foi utilizado um conjunto aleatório de observações com 65% do tamanho da amostra.

A construção do próximo modelo utilizou uma parcela de 35% da amostra, também contendo um conjunto aleatório de dados.

Esta prática evitou que se construísse o modelo sempre sobre os mesmos dados o que poderia enviesar os resultados. Além disso, com uma quantidade menor de dados, o tempo de processamento é menor.

Definidos os parâmetros para a construção do modelo, o passo seguinte foi a aplicação do algoritmo e a análise dos *Clusters* gerados. Os resultados foram muito satisfatórios, tendo o algoritmo encontrado o mesmo número de *Clusters* no novo conjunto de dados utilizados, ou seja, quatro *Clusters*.

Em termos de distribuição espacial e ligações, os dois resultados são muito semelhantes, como podemos observar na seguinte figura, o que suporta a robustez do modelo.

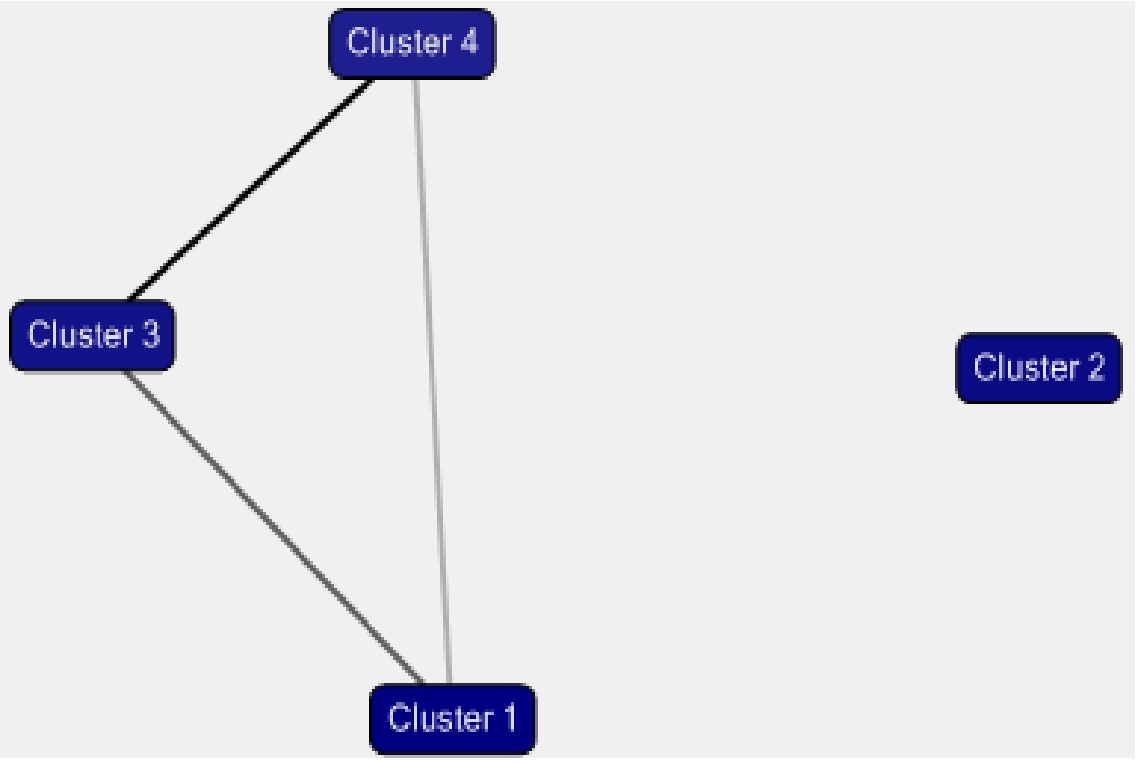

Ilustração 18-Distribuição Espacial do Clusters com Amostra 35% (Resultado Final)

Realizou-se uma análise à distribuição das variáveis, das categorias de produto, pelos quatro *Clusters*. Tal como na amostra anterior, identifica-se uma distribuição heterogénea das categorias com categorias "chave" em cada *Cluster*.

Através da análise dos resultados, foi possível encontrar as categorias mais compradas em cada *Cluster*, ou seja, que têm maior probabilidade de ocorrer, o que permitiu identificar diferentes tipos de consumo e consequentemente iniciar a definição dos perfis.

Categoria	Cluster 1	Cluster 2	Cluster 3	Cluster 4
águas	20 %	11 %	29 %	12 %
alimentação infantil	7 %	7 %	6 %	5 %
aperitivos	7 %	4 %	7 %	6 %
auto	0 %	0 %	0 %	0 %

Tabela 15-Exemplo da distribuição da probabilidade das Categorias ocorrerem, em cada Cluster

Analisando a distribuição das variáveis *Cluster* a *Cluster* identificou-se as categorias com maior probabilidade de ocorrerem em cada *Cluster*, denominadas como categorias "chave".

Dentro destas categorias, apenas foram selecionas para definir o perfil do *Cluster* as que apresentaram uma probabilidade superior a 15%.

A análise destas categorias, que podem ser classificados como os atributos ou descritores do perfil, possibilitaram a definição do perfil do cliente em cada *Cluster* (Kotler, 1991).

Categoria	**Probabilidade**	**Atributo**
iogurtes e sobrem_	75 %	Define
leite e bebidas soja	42 %	Define
bolachas	39 %	Define
conservas	32 %	Define
ingredientes básicos	25 %	Define
sumos e nectares	25 %	Define
refrigerantes	22 %	Define
águas	20 %	Define
limpeza geral	16 %	Define
refeições congeladas	16 %	Define
frutas	15 %	Não Define
cereais	15 %	Não Define
especialidades	14 %	Não Define
prod papel e consum_	13 %	Não Define
vegetais e frut cong	11 %	Não Define
cuidados do corpo	11 %	Não Define
petcare	10 %	Não Define
bebidas quentes	8 %	Não Define
limp_ e tratam roupa	8 %	Não Define
hig e protecção bebé	8 %	Não Define

Tabela 16- Definição das categorias "chave" do Cluster 1

No *Cluster* 1 destaca-se o consumo de categorias que não exigem confeção, nomeadamente iogurtes, leite, bolachas e conservas. A este *Cluster*, foi atribuído o perfil com a nomenclatura de "Práticos". Estamos perante um perfil de cliente, que consome uma ampla gama de produtos, necessários para o dia-a-dia, na sua grande maioria, produtos pré-fabricados, por exemplo iogurtes, bolachas e refeições congeladas.

O *Cluster* 2 recebeu o nome de "Saudáveis", dado que as categorias mais consumidas são os legumes e as frutas. Além disso, este é o único *Cluster*, que consome a categoria "especialidade f&l", com uma probabilidade superior a 15%. Nesta categoria, estão presentes os produtos *Gourmet* e produtos alimentares e, nomeadamente, frutas exóticas e legumes biológicos. Contrariamente ao *Cluster* 1 este perfil de cliente, demonstra a procura por uma alimentação equilibrada, evidencia gosto pela confeção e preza produtos requintados e distintos.

Categoria	Probabilidade	Atributo
legumes	90 %	Define
frutas	69 %	Define
iogurtes e sobrem_	51 %	Define
ingredientes básicos	29 %	Define
especialidades f&l	26 %	Define
conservas	25 %	Define
leite e bebidas soja	22 %	Define
bolachas	17 %	Define
prod papel e consum_	17 %	Define
especialidades	16%	Define

Tabela 17- Definição das categorias "chave" do Cluster 2

O *Cluster* 3 foi denominado por "Cuidadosos", consumindo em maior quantidade, artigos de limpeza geral e limpeza para a casa. Na definição dos perfis dos consumidores, também foram tidas em conta, categorias que só são consumidas dentro do *Cluster*. É o caso deste *Cluster*, já que é o único onde o

consumo de produtos para o corpo, ocorre com uma probabilidade superior a 15%.

Categoria	Probabilidade	Atributo
prod papel e consum_	88 %	Define
ingredientes básicos	63 %	Define
limpeza geral	55 %	Define
leite e bebidas soja	43 %	Define
conservas	34 %	Define
iogurtes e sobrem_	32 %	Define
limp_ e tratam roupa	31 %	Define
águas	29 %	Define
cuidados do corpo	23 %	Define
bolachas	21 %	Define
temperos	18 %	Define
refrigerantes	17 %	Define
limpeza da cozinha	16 %	Define

Tabela 18- Definição das categorias "chave" do Cluster 3

O *Cluster* 4 foi denominado por "Básicos" sendo o *Cluster* que mais compra ingredientes básicos, temperos e conservas. Este perfil pode ser descrito como o cliente que compra um pouco de tudo.

Categoria	Probabilidade	Atributo
ingredientes básicos	82 %	Define
temperos	82 %	Define
conservas	68 %	Define
leite e bebidas soja	36 %	Define
iogurtes e sobrem_	32 %	Define
bolachas	28 %	Define
legumes	26 %	Define
prod papel e consum_	23 %	Define
limpeza geral	19 %	Define
refrigerantes	18 %	Define
vegetais e frut cong	15,24%	Define

Tabela 19- Definição das categorias "chave" do Cluster 4

4.3. Discussão dos resultados com testes das hipóteses propostas na revisão de literatura

A escolha entre as duas hipóteses apresentadas dependia da avaliação dos resultados obtidos em cada uma, tendo em conta as métricas de sucesso estabelecidas. Esses objetivos eram os seguintes:

1) Prever o comportamento do cliente.
2) Identificar os clientes, com maior relevância para a empresa.
3) Apoiar os Marketers no desenvolvimento de campanhas promocionais e no desenvolvimento de um sistema de recomendação.

A hipótese H1 gerou um perfil de cliente, através duma segmentação *Market Driven* e com base em atributos fatuais, género, idade e localidade, definindo o cliente do Continente *Online*. O perfil identificado, foi o de um cliente do género feminino, com idade compreendida entre os 25 e 34 anos e residente na Cidade de Lisboa.

Ao comparamos este resultados, com os resultados da Marktest (Marktest, 2012), é notória a semelhança entre os dois. Este estudo definiu o perfil do

cliente que consome FMCG via *e-commerce* em Portugal, como sendo um cliente do género feminino, com idade compreendida entre 35 e 44 anos e residente na área da Grande Lisboa.

Apesar de existir uma diferença no atributo referente à idade, ambos os resultados são coerentes, com a definição do perfil deste tipo de clientes, proposta em (Raijas, Tuunainen, 2001) e em (Morganosky, Cude, 2000).

Analisando, esta hipótese relativamente às métricas impostas existem limitações que reduzem a sua eficácia, quer na previsão do comportamento dos clientes, quer na identificação dos clientes mais importantes e também na possibilidade de apoiar a tomada de decisão e desenvolvimento de campanhas de *Marketing* direcionadas.

As limitações desta hipótese, já tinham sido identificadas na revisão de literatura, (Frank, 1972) e (Guadugni, Litle, 1983). A utilização de atributos fatuais pode levar à definição de perfis enviesados, comprometendo a sua capacidade de previsão do comportamento do cliente.

Com o perfil encontrado, é difícil prever, o que é que o cliente vai comprar, já que o tipo de produtos transacionados são bens consumíveis do dia-a-dia e com uma frequência elevada.

Com este perfil, será muito difícil, criar campanhas de *Marketing* direcionadas, nomeadamente, através de um sistema de recomendação.

Ainda que, este perfil possa ajudar os *Marketers*, no desenvolvimento duma comunicação mais eficaz, melhorando a relação com o cliente, por si só, o perfil criado com esta hipótese, não alcança resultados satisfatórios, comparativamente com a hipótese H2.

A hipótese H2, criada com a construção duma segmentação *Data Driven* e com a utilização de atributos comportamentais, permitiu a definição de quatro perfis de clientes, baseando-se nas categorias compradas.

Conforme a literatura apresentada, a utilização destes atributos, permite criar perfis, que refletem as vontades e as necessidades do cliente. A previsão do seu comportamento será mais exata e trará um conhecimento profundo, sobre os seus hábitos de compra, permitindo aumentar a sua satisfação (Franzak, 2001:1-2), (Chen, 2007) e (Dickson, 1982).

O perfil definido em cada *Cluster* permite determinar com precisão qual a probabilidade dum cliente consumir determinada categoria, dentro do *Cluster* a que pertence.

Este tipo de perfil permite prever o tipo de produtos que o cliente irá comprar, prever como o cliente vai reagir a determinada promoção ou campanha, identificar quais os perfis com maior relevância e permite o desenvolvimento de um sistema de recomendação (Leung, 2009).

Vamos relembrar as duas hipóteses:

H1: O perfil do cliente do Continente *Online* deve ser definido com base em atributos factuais, nomeadamente, o género, a idade e localização geográfica.

H2: O perfil do cliente do Continente *Online* deve ser definido com base em atributos comportamentais, mais concretamente, através dos produtos comprados.

Comparando as duas hipóteses apresentadas, claramente que a hipótese H2 alcança uma performance superior nas métricas impostas. Como tal, esta é a hipótese que melhor respondeu à questão de investigação e às necessidades da Modelo.com.

5. Conclusões

5.1. Discussão dos principais resultados e aplicabilidade para *Marketers*, gestores e outros intervenientes

Os perfis encontrados com a hipótese H2 preenchem os requisitos dum perfil eficiente na classificação de clientes.

Estes permitem, descrever os clientes com base nos atributos selecionados (Ahola, Runsala, 2001), permitem reconhecer o comportamento de compra, as motivações e as preferências dos clientes por determinados produtos (Sindell, 2001) e permitem prever o seu comportamento, com base nas suas vontades e nas suas necessidades (Franzak, 2001:1-2), (Chen, 2007) e (Dickson, 1982).

Para além destes requisitos, existem outros que podem ser determinantes, para que se atinja o sucesso na definição de um perfil (Leung, 2009) e (Jansen, 2007).

A definição de um perfil deve permitir identificar os clientes mais valiosos da empresa e descrever o seu comportamento de compra o que também, cremos, foi alcançado com a construção desta hipótese.

Para tal, recorreu-se à análise dos resultados da segmentação, que permitiram hierarquizar os *Clusters*, pelo seu peso no total e analisar algumas variáveis do comportamento de compra de cada perfil.

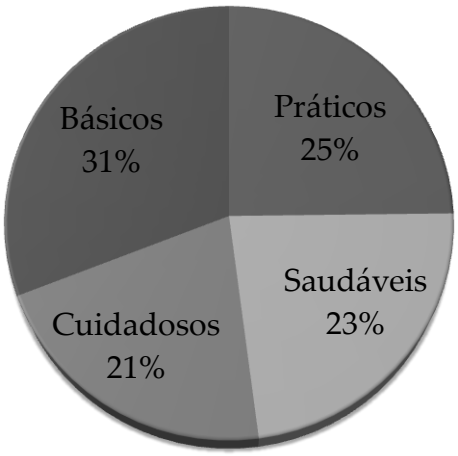

Ilustração 19-Distribuição dos Clientes pelos Clusters

Os clientes com o perfil "Básico" são mais numerosos, seguidos dos Práticos, dos Saudáveis e por último, pelos Cuidadosos. Conciliando esta informação com o levantamento de variáveis de compra, é possível identificar o perfil com mais clientes, o perfil em que os clientes compram mais e o perfil, em que os

clientes gastam mais dinheiro. Estas são algumas das possibilidades para os *Marketers*.

Cluster	Valor Total Gasto por Cluster	Valor Gasto por Cliente	Total de artigos Comprados por Cluster	Número Médio de Itens por Cliente
Práticos	3º	3º	23,86%	3º
Saudáveis	1º	1º	29,15%	2º
Cuidadosos	4º	2º	17,33%	4º
Básicos	2º	4º	29,66%	1º

Tabela 20-Análise de variáveis de compra dos Clusters

Em termos de valor económico, o perfil "Saudáveis" é o que gasta mais, quer no conjunto, quer individualmente. Em contrapartida, o perfil "Básicos", remete para um cliente que gasta menos, mas que compra mais produtos.

Daqui podemos concluir, que os produtos comprados pelos clientes "Saudáveis" apresentam preços superiores, aos produtos comprados pelos "Básico".

São este tipo de informações, que ajudam os *Marketers*, na sua tomada de decisão e no desenvolvimento de ações de *Marketing* (Jansen, 2007).

Para aumentar as vendas, é possível recorrer a comunicação direcionada e ajustada a cada perfil. O seu meio de divulgação pode ser, através de um sistema de voucher ou descontos via correio electrónico ou através de um sistema de recomendação na plataforma do Continente *Online*.

Este sistema pode ser implementado na empresa e é referido na literatura, como uma das métricas, para uma definição de perfis eficiente (Leung, 2009).

Através dos perfis encontrados, foi criado um sistema de recomendação, que sugere categorias a cliente que pertençam a um determinado perfil. Um cliente que demonstre determinado perfil, mas que não compra uma categoria,

receberá a sugestão de compra dessa mesma categoria, possivelmente com um desconto atribuído.

Este sistema é suportado, pela própria definição de perfil, ou seja, clientes que partilhem o mesmo perfil, também partilham vontades, necessidades e gostos por produtos (Franzak, 2001:1-2).

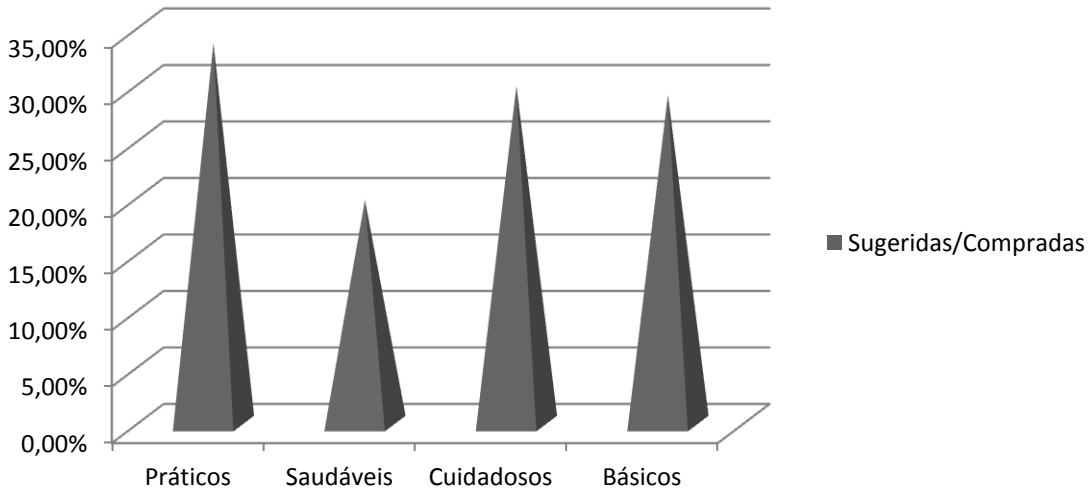

Ilustração 20-Margem de aumento de vendas em cada Perfil

Partindo deste princípio, foi possível calcular a margem de aumento de vendas, recorrendo ao sistema de recomendação. Os valores obtidos, representam o aumento máximo de vendas, dentro de cada *Cluster*, aceitando que os clientes, compram todas as sugestões.

Como foi demonstrado, quando um perfil é criado com base no comportamento do cliente, é possível obter informações fulcrais sobre o cliente, que ajudam os intervenientes na sua tomada de decisão, podendo representar uma vantagem competitiva, face à concorrência.

5.2. Limitações da investigação

As principais limitações deste relatório são inerentes à construção da hipótese H1.A obtenção de dados, através do Google Analytics, acarreta um conjunto de características próprias da ferramenta que reduzem a credibilidade dos dados obtidos. A recolha de dados, é baseada em informações fornecidas

através de *cookies* (informação sobre a navegação gravada no dispositivo) e através dos motores de busca. A eliminação desta informação leva à criação de um novo perfil de utilizador, podendo por isso influenciar os resultados.

Outra limitação é o fato de vários indivíduos utilizarem o mesmo dispositivo e a mesma conta de acesso para comprar no Continente *Online*, o que também influencia os resultados.

Para dar maior credibilidade à construção da hipótese H1, os resultados do Google Analytics foram comparados com os resultados do questionário.

Também na construção do questionário existem limitações. Como o meio de divulgação utilizado foi a rede social Facebook é provável que o universo das respostas contemple, na sua grande maioria, indivíduos com empatia pelo autor do mesmo. Estes estariam mais dispostos a realizar a tarefa pedida, o que também pode ter afetado os resultados obtidos.

5.3. Recomendações para a investigação futura

Na realização de trabalhos futuros, que tenham como tema a segmentação e definição de perfis, é aconselhado que se utilize não só atributos fatuais como atributos comportamentais.

A utilização conjunta dos tipos de atributos para a segmentação e definição de perfis dado que as vantagens de um dos tipos compensam as limitações do outro e vice-versa (Currim, 1981).

Esta combinação permite criar perfis mais precisos, dotar os intervenientes, *Marketers* e gestores, dum conhecimento mais profundo sobre os seus clientes, deixando-os mais satisfeitos, fidelizando-os.

6. Bibliografia

ACEPI, A. do C. E. e P. I. 2013. ***Economia Digital em Portugal 2009-2017***: 24.

Ahola, J., & Rinta-runsala, E. 2001. ***Data Mining Case Studies in Customer Profiling***. 24.

A. Mansoorian and G. M. Myers. 1996. Private Sector Versus Public Sector Externalities. ***Regional Science and Urban Economics***. 26. 543-555.

Amat, J. 2002. Using reporting and data mining techniques to improve knowledge of subscribers ; applications to customer profiling and fraud management. ***Journal of Telecommunications and Information Technology***, 1–5.

Andam, Z. R. 2003. *e-commerce* and e-Business. ***E-ASEAN Task Force***, 1–47.

Anderson T.W., Henion K.E., Cox E.P. 1974. ***Socially vs. ecologically concerned consumers***. AMA Combined Conference Proceedings. Chicago, IL: American Marketing Association, p. 304–11.

Armstrong, Gary & Kotler, Philip. 2005. ***Marketing: An Introduction***. Upper Saddle River, N.J. Prentice Hall, (7 Ed).

Bachl, T., & Koll, O. 2013. *E-COMMERCE* – THE POTENTIAL FOR RETAILING? ***ECR Conference Brussels***, (May): 1–65.

Bannon, D. P. 2004. Marketing Segmentation and Political Marketing. ***Political studies association***.

Beane, T. P., & Ennis, D. M. 1987. Market Segmentation: A Review. ***European Journal of Marketing***.

Catherine Bounsaythip, E. R.-R. 2001. ***Overview of Data Mining for Customer Behavior Modeling***, 1–59.

Celen, A., Erdogan, T., & Taymaz, E. 2005. ***Fast Moving Consumer Goods: Competitive Conditions and Policies***: 1–56.

Chang, T., Lansing, E., Kim, M., & Lansing, E. 2002. ***PROFILE OF WINERY VISITORS OF MICHGAN***: 7.

Chen, D., Sain, S. L., & Guo, K. 2012. Data mining for the *online* retail industry: A case study of RFM model-based customer segmentation using data mining. ***Journal of Database Marketing & Customer Strategy Management***, 19(3): 12.

Couceiro, C. 2012. ***Aproximar à distância – A fidelização no E-commerce : Relatório de estágio como Assistant Business Developer da Europa do Sul na Pixmania***.

Dickson, P. R. 1982. Person-Situation: Segmentation's Missing Link. ***Journal of Marketing***, 46(4): 56–64.

eMArketer. 2014. ***Worldwide Ecommerce Sales to Increase Nearly 20% in 2014.*** Disponível em http://www.emarketer.com/Article/Worldwide-Ecommerce-Sales-Increase-Nearly-20-2014/1011039#sthash.WkHAMT5F.dpuf.

Etzel, M. J. & Woodside, A. G. 1982. Segmentation vacation markets: The case of distant and near-home travelers. ***Journal of Travel Research***. 20(4). 10-14.

Fettman, E. 2014. Google Analytics Universal Guide. ***Google Analytics Universal Guide***, (March): 45.

Frank, R.E, Massy WF, Wind Y. 1972. ***Market segmentation Englewood Cliffs***: Prentice-Hall.

Franzak, F, Pitta, D & Fritsche, S. 2001. *Online* relationships and the consumer's right to privacy. ***Journal of Consumer Marketing.*** Vol. 18.Issue 7. p. 631-642.

Galbraith, J. 2005. ***Designing the Customer- Centric Organization***. (A Wiley Imprint, Ed.) (First Ed): 204. 989 Market Street, San Francisco, CA 94103-1741: Jossey-Bass.

Giha FE, Singh YP, Ewe HT. 2003. ***Customer Profiling and Segmentation based on Association Rule Mining Technique***. Proc. Softw. Engin.and Appl. 397.

Golubova, S. 2012. ***E-commerce Adoption and Implementation Strategy for a High-Tech Firm***.

Guadagni, Peter & John D.C. Little .1983. ***A Logit Model of Brand Choice Calibrated on Scanner Data***. Marketing Science. 2 (Summer). 203-238.

Hays, T., Keskinocak, P., & De Lopez, V. M. 2005. Strategies and challenges of internet grocery retailing logistics. ***Applications of Supply Chain Management and E-commerce Research***: 217–252. Springer.

INE, Instituto Nacional de Estatística. 2013. ***Sociedade da Informação e do Conhecimento, Inquérito à Utilização de Tecnologias da Informação e da Comunicação pelas Famílias***. do Instituto Nacional de Estatística

Jansen, S. M. H. 2007. ***Customer Segmentation and Customer Profiling for a Mobile Telecommunications Company Based on Usage Behavior Acknowledgments***. https://dke.maastrichtuniversity.nl/westra/PhDMaBa-teaching/GraduationStudents/StephanJansen2007/Stephan_Jansen2007.pdf.

Kantar Worldpannel. 2014. ***Accelerating the growth of community-led conservation***.

Kalakota, R. & Whinston, A. 1997. **Electronic Commerce: A Managers Guide**. Reading MA: Addison-Wesley.

Kinnear T.C., Taylor J.R., Ahmed S.A. 1974. ***Ecologically concerned consumers: who are they?*** J Mark.

Kotler, P. 1994. Reconceptualizing marketing: An interview with Philip Kotler. ***European Management Journal***.

Kotler, P., & Armstrong, G. 2001. Principles of Marketing. (V. Wong, J. Saunders, & G. Armstrong, Eds.)***World Wide Web Internet And Web Information Systems***, vol. 42: 785. Prentice Hall.

Kronenberg, D. 2008. ***Investigating Seed Values for the K-means Clustering Algorithm David Kronenberg***: 6.

Lawrence, Corbitt, Fisher, Lawrence & Tidwell. 2000. ***Internet Commerce Digital Models For Business***. (2nd Ed). Wiley & Son. Singapore.

Leung, C.-H. 2009. An Inductive Learning Approach to Market Segmentation based on Customer Profile Attributes. ***Asian Journal of Marketing***, Vol. 3(Issue 3): p65.

Levitt, T. 1986. The Marketing Imagination. ***Journal of Advertising***.

Lilien, G. L., & Rangaswamy, A. 2006. Marketing decision support models: The marketing engineering approach. In R. Grover, M. Vriens, R. (Ed. . Grover, & M. (Ed. . Vriens (Eds.), ***The handbook of marketing research: Uses, misuses, and future advances.***: 230–254. Sage Publications, Inc.

Lilien, Gary L. and Arvind Rangaswamy, ***"Marketing Management and Strategy: Marketing Engineering Applications"***, 1/1/2003, (2nd Ed).

Marktest. (2012). Retrieved 28 de Fevereiro de 2014, from Grupo Marte http://mpt.marktest.pt/mpt/

Mazanec, J.A. 1992. Classifying tourists into market segments: a neural network approach: ***Journal of travel and tourism marketing***. (First ed). 39-59.

Mcdonald, M., & Dunbar, I. 2010. ***Market Segmentation***: 492. Woodeaton, Oxford: Published by Goodfellow Publishers Limited.

Miguéis, V. L., Camanho, a. S., & Cunha, J. F. E. 2011. Mining customer loyalty card programs: The improvement of service levels enabled by innovative segmentation and promotions design. ***Lecture Notes in Business Information Processing***, 82 LNBIP: 16.

Miguéis, V. L., Camanho, A. S., & Falcão e Cunha, J. 2012. Customer data mining for lifestyle segmentation. ***Expert Systems with Applications***, vol. 39: 8.

Morganosky, Michelle A. & Cude, Brenda J. (2000). Consumer Response to *Online* Grocery Shopping. ***International Journal of Retail & Distribution Management.*** *28*(1). 17-26.

MK, Euro Info Correspondence Centre (Belgrade, Serbia). 2002. ***E-commerce Factor of Economic Growth***. Disponível em http://www.eicc.co.yu/newspro/viewnews.cgi?newsstart3end5, Outubro de 2014.

Modelo.com. 2014. *O nosso cliente em 7 segmentos*. PowerPoint

Modelo.com. 2014. *Segmentação Valor COL*. PowerPoint

M. Wedel & W.A. Kamakura. 2000. ***Market Segmentation: Conceptual and Methodological Foundations.*** (2nd ed). Norwell. MA: Kluwer Academic Publishers 382 pp.

Nielsen. 2014. ***E-commerce: Evolution or revolution in the fast-moving consumer goods world?*** 1–21.

Parr Rud, O. 2001. ***Data Mining Cookbook Modeling Data for Marketing, Risk, and Customer Relationship Management***: 430.

Petre, R. 2013. ***Data Mining Solutions for the Business Environment***, IV(4): 21–29.

Pickett G.M., Kangun N., Grove SJ. 1993. ***Is there a general conserving consumer?*** A public policy concern. Public Policy Mark 1993;12(2):234– 43.

Pickton, D., & Broderick, A. 2005. Setting Objectives, Determining Strategy and Tactics. ***Integrated Marketing Communications***: 412–442. Prentice Hall.

Pickton, David & Broderick, Amanda . 2005. Chapter 17: ***Identifying target audiences and profiling target markets.*** (2nd ed).In Pickton, David & Broderick, Amanda: Integrated marketing communication :371-398

Pratt, M. 2011. Profiling wine tourists, more than just demographics. ***6th AWBR International Conference | Bordeaux Management School, 9-10 June 2011***, (June): 9–10.

PWC. 2013. ***Understanding how US online shoppers are reshaping the retail experience***: 16.

Raijas, Anu & Tuunainen, Virpi Kristiina. (2001). Critical Factors in Electronic Grocery Shopping. *The International Review of Retail.* ***Distribution and Consumer Research***. *11*(3). 255-265.

Reichwald, R., Seifert, S., & Walcher, D. 2004. Customers as part of value Webs: towards a framework for webbed customer innovation tools. ***37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the***.

Road, H., Jose, S., & GETTA, J. R. 2003. ***Fast Algorithms for Mining Association***, 487–499.

Rosen A. 2000. ***The E-commerce Question and Answer Book***.(5nd ed). USA. American Management Association.

SABI. 2014. *Resultados Não Consolidados entre 2004 e 2013 da Modelo.com*. Disponível em https://sabi.bvdinfo.com

Salvador, F., Forza, C. 2007. Principles for efficient and effective sales configuration design. ***International Journal of Mass Customizatio.*** 2(1/2). 114-127.

Samdahl, DM., Robertson R. 1989. ***Social determinants of environmental concern: specification and test of the model***. Environ Behav .21(1): 57–81.

Scott D., Willits F.K. 1994. ***Environmental attitudes and behavior: a Pennsylvania survey***. Environ Behav. 26(2):239– 60.

Scridon, M. A. 2008. Understanding Customers - Profiling and Segmentation. ***Management & Markeing***: 10.

Sindell, K. 2000. ***Loyalty marketing for the Internet Age: How to identify, attract, serve, and retain customers in an e-commerce environment***. Chicago: Dearborn Trade.

Smith, W. R. 1956. Product Differentiation and Market Segmentation as Alternative Marketing Strategies. ***Journal of Marketing***, 21(1): 3.

Sonae. 2014. ***História***. Disponível em www.sonae.pt/pt/sonae/historia (2014/11/20;15H 30M).

Sonae. 2014. ***Marcas***. Disponível em www.sonae.pt/pt/marcas/ (2014/11/20;15H 50M).

Sonae. 2014. ***Investidores***. Disponível em www.sonae.pt/pt/investidores/areas-de-negocio/ (2014/11/20;16H 00M).

Srinivasu, R. 2014. Fast Moving Consumer Goods Retail Market , Growth Prospect , Market Overview and Food Inflation in Indian Market – an Overview. ***International Journal of Innovative Research in Science, Engineering and Technology***, 3(1): 9.

Tremblay, K. R., Jr. and R. E. Dunlap. 1978. ***Rural-urban residence and concern with environmental quality***: a replication and extension. Rural Sociology.

Venkatesan, R. 2007. ***Cluster Analysis For Segmentation***.

Verhoef, P. C., Spring, P. N., Hoekstra, J. C., & Leeflang, P. S. H. 2002. The commercial use of segmentation and predictive modeling techniques for database marketing in the Netherlands. ***Decision Support Systems***, 34(4): 471–481.

Yankelovich, D., & Meer, D. 2006. Rediscovering market segmentation. ***Harvard Business Review***.

Zhang, G. 2007. Customer segmentation based on survival character. ***2007 International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM 2007***, 3386–3391.

7. ANEXOS

7.1. Glossário

Algoritmo - Um algoritmo, é um conjunto específico de instruções ou comandos, bem definidos e encadeados, com o objetivo da realização de um procedimento ou para a resolução de um problema. Uma das exigências para a construção de um algoritmo, é que exista um fim, ou seja, o algoritmo tem que estar definido para ser executado, durante um período de tempo finito. Alguns algoritmos mais específicos são denominados por método, procedimento ou técnica, como é o caso dos algoritmos de *Clustering*, também denominados por Método de *Clustering*. A implementação de um algoritmo, por ser realizada por qualquer autómato, nomeadamente um computador ou por uma pessoa.

Árvore de Decisão – Uma Árvore de Decisão é um diagrama em forma de árvore, que é utilizado para determinar, um curso de ação ou para demostrar

uma probabilidade estatística. Cada ramo da árvore de decisão, representa uma possível escolha ou ocorrência. A sua estrutura em árvore, mostra como uma escolha leva ao próximo passo e a utilização de ramos, indica que cada opção é mutuamente exclusiva.

Base de Dados – Uma Base de Dados (BD), é uma coleção de informações, que estão devidamente organizadas, possibilitando um acesso fácil, bem como, uma gestão eficiente dos dados armazenados e a sua atualização. As Bases de Dados, podem ser classificadas de acordo com tipos de dados armazenados, nomeadamente, bibliográficos, numéricos ou de imagens.

Business Intelligence – O *Business Intelligence* representa um conjunto de ferramentas e sistemas, que desempenham um papel fulcral no planeamento estratégico da empresa. Estes sistemas permitem que uma empresa consiga reunir, armazenar, aceder e analisar, toda a informação necessária e transformá-la em conhecimento, o que permitirá mais eficácia na Tomada de Decisão.

Computação – A Computação define-se como sendo, qualquer método de cálculo ou implementação de um algoritmo realizado por um computador.

Churn – O *Churn*, também denominado por *Customer Attriction*, é a perda de clientes ao longo de um determinado período.

Cluster- O *Cluster* é um grupo de objetos ou de observações, que partilham mais características e que são mais semelhantes, comparativamente a outros grupos de objetos ou observações.

CRM – O *Customer Relationship Management*, é um sistema de gestão de interações da empresa com os seus clientes. Este sistema normalmente recorre ao uso de sistemas informáticos e de informação.

Data Mining – O *Data Mining* é um processo analítico concebido para a exploração de dados, geralmente grandes quantidades de dados, denominados como *Big Data.* A exploração destes dados, procurar identificar padrões consistentes ou relacionamentos sistemáticos entre variáveis. O procedimento

seguinte é a validação dos resultados, aplicando os padrões encontrados, a novos subconjuntos de dados. O objetivo final é a previsão do comportamento das variáveis.

Distância Euclidiana – A Distancia Euclidiana, é a diferença entre as raízes dos quadrados das coordenadas de um par de objetos.

E-commerce – O *E-commerce* ou em Português, o comércio eletrônico é, um termo que se aplica a qualquer tipo de negócio ou transação comercial, que envolva a transferência de informações, através da Internet.

K-Means – O K-Means é um algoritmo de *Clustering* ou um método de *Clustering* não hierárquico e À-Priori. O Algoritmo começa por converter o número de observações ou de componentes, no número final de Clusters exigido. Daí ser um algoritmo, onde o número de Clusters (K) é determinado previamente. Neste primeiro passo, o número final de Clusters é escolhido, tendo em conta que os pontos dentro dos Clusters, encontram-se o mais afastados possível entre si. Em seguida, o algoritmo examina cada observação ou componente, atribuindo-o a um determinado Cluster dependendo da sua distância mínima ao centroíde. A posição do centróide, o ponto central do *Cluster*, é recalculada, cada vez que um novo componente é adicionado. Este ajuste da posição do centróide, só termina quando todos os componentes estão agrupados no número final *Clusters*.

Machine Learning – O *Machine Learning* é uma área científica que se dedica ao estudo da inteligência artificial (IA), conseguida através da construção de algoritmos, que dotam os computadores da capacidade aprendizagem, mesmo sem serem explicitamente programados para tal.

Market Basket – O Market Basket é o nome dado ao conjunto de produtos comprados por um determinado grupo de indivíduos ou indivíduo, durante um determinado período de tempo.

Online Browsing - Traduz a navegação e a procura por informação *online.*

Online Buying – Ato de comprar *online* .

RFM – A RFM (*Recency, Frequency, Monetary*), é uma análise ou técnica de *Marketing*, utilizada para determinar quais são os melhores clientes para uma determinada empresa, através da análise e comparação de três variáveis num determinado período de tempo. A *Recency*, representa a última compra efetuada pelo cliente, a *Frequency*, o número de vezes que o cliente compra durante o período de análise e a *Monetary*, que representa o montante gasto pelo cliente.

Segmentação – A Segmentação é uma subdivisão do mercado em grupos distintos de consumidores do ponto de vista do seu valor para a empresa ou de características partilhadas pelos grupos, podendo estas ser, demográficas, socioculturais, comportamentais, geográficas, culturais ou psicográficas.

7.2. Respostas do questionário enviado via Facebook

Pergunta 1: Sexo?

Masculino	**105**	**40%**
Feminino	**157**	60%

Pergunta 2: Idade?

0-18	**17**	**6%**
18-25	**91**	35%
25-35	**51**	19%
35-65	**89**	34%
+65	**14**	5%

Pergunta 3: Zona de Residência?

Norte	**117**	**45%**
Centro	**72**	27%
Sul	**22**	8%
Açores	**10**	4%
Madeira	**11**	4%
Estrangeiro	**30**	11%

Pergunta 4: Estado Civil?

Solteiro	**166**	**63%**
Casado	**96**	37%

Pergunta 5: Agregado Familiar?

1	**28**	**11%**
2	**43**	16%
3	**50**	19%
4	**60**	23%
5	**48**	18%
+5	**33**	13%

Pergunta 6: Já alguma vez fez compras *Online*?

Sim	**245**	**94%**
Não	**17**	6%

Pergunta 7: Com que frequência faz compras *online*?

Todos os dias	**5**	**2%**
Uma vez por semana	**9**	3%
Duas vezes por mês	**21**	8%
Todos os meses	**52**	20%
De três em três meses	**41**	16%
De seis em seis meses	**38**	15%
+ de seis meses	**32**	12%
Uma vez por ano	**48**	18%
Nunca	**16**	6%

Pergunta 8: Que valor costuma gastar?

0€	**21**	**8%**
0-50€	**64**	24%
50-100€	**81**	31%
100-250€	**56**	21%
250-500€	**28**	11%
+ de 500€	**12**	5%

Pergunta 9: Que dispositivo utiliza para realizar as suas compras?

Computador	**239**	**91%**

Telemóvel	**50**	19%
Tablet	**33**	13%
Consola	**1**	0%
Outro	**0**	0%
Nenhum	**13**	5%

Pergunta 10: Que tipo de produtos, considera mais simples de serem comprados através deste tipo de serviço?

Eletrónica	**93**	**35%**
Informática	**75**	29%
Petcare	**39**	15%
Artigos Desportivos	**72**	27%
Livros	**130**	50%
Material Escolar	**58**	22%
Vinhos	**41**	16%
Outro	**76**	29%

Pergunta 11:Que tipo de produtos não consideraria comprar, através de um serviço *online*?

Carne	**172**	**66%**
Peixe	**198**	76%
Legumes	**124**	47%
Fruta	**125**	48%
Pão	**108**	41%
Roupa	**33**	13%
Massa	**37**	14%
Enlatados	**31**	12%
Água	**43**	16%
Outro	**16**	6%

Pergunta 12: Identifica-se com algum destes grupos?

Vegetais, Frutas e Especialidades&Groumet	**48**	**18%**
Iogurtes&Sobremesas, Leite&Bebidas de Soja e Bolachas	**28**	11%
Conservas, Temperos e Ingredientes Básicos	**25**	10%
Limpeza Geral, Limpeza&Tratamento de roupa e Papel&Consumíveis	**18**	7%
Não	**97**	37%

Pergunta 13: Considere um serviço de compras *online*, que vende produtos alimentares. Recorreria a este serviço?

Sim	**179**	**68%**
Não	**83**	32%

Pergunta 14: Que razões não o levariam a comprar um produto perecível *online*?

Pode chegar danificado	**75**	**29%**
Não confio na seleção do produto, prefiro ser eu a escolher	**164**	63%
O produto da loja é mais fresco e com prazos de validade maiores	**53**	20%
Nenhuma. Compro com regularidade produtos perecíveis	**38**	15%

Pergunta 15: Que razões o levariam a recorrer a este tipo de serviço?

Poupança de tempo na escolha e compra dos produtos	**102**	**39%**
Evitar deslocações, que implicam tempo e gastos	**179**	68%
Possibilidade de criar listas, que podem ser encomendadas várias vezes	**85**	32%
Possibilidade de consultar preços e promoções com rapidez	**81**	31%
Não vejo vantagens	**21**	8%

Pergunta 16: Se existisse uma loja de conveniência ao lado da sua residência, continuaria a utilizar este serviço?

Sim	**147**	**56%**
Não	**115**	44%

Pergunta 17: Considere que recorria a este serviço. Ao comprar *online* estaria a substituir as suas idas ... aos estabelecimentos?

Diárias	**19**	**7%**
Semanais	**88**	34%
Mensais	**68**	26%
Pontuais	**63**	24%
Nenhuma das demais	**20**	8%

Pergunta 18: Que característica considera fundamentais para este tipo de serviço?

Rapidez de entrega	**151**	**58%**
Facilidade na compra	**132**	50%
Fiabilidade (Produtos em condições)	**174**	66%
Fiabilidade (Encomendas Completas)	**121**	46%
Serviço de atendimento ao cliente eficaz	**83**	32%
Serviço de Entrega Grátis	**120**	46%

Pergunta 19: Qual das opções de entrega utilizaria?

Entrega em casa	**257**	**98%**
Levantar na Loja	**5**	2%

Pergunta 20: Considere que os serviços de entrega são pagos. Reconhece que este serviço tem vantagens e está disposto a pagar para o utilizar?

Sim	**166**	**63%**
Não	**96**	37%

Pergunta 21: Suponha que encomendou um produto através deste serviço, que situação o deixaria mais descontente?

Produto chegou danificado ou em mau estado	**167**	**64%**
Produto difere da informação de compra	**111**	42%
Produto nunca chegou	**148**	56%
Produto foi trocado por um semelhante	**77**	29%
Nenhuma, são todas compreensíveis	**16**	6%

Pergunta 22: Alguma destas situações, o levariam a deixar de utilizar este serviço?

Produto chegou danificado	**101**	**39%**
Produto difere da informação de compra	**77**	29%
Produto nunca chegou	**179**	68%
Produto foi trocado por um semelhante	**42**	16%
Nenhuma, são todas compreensíveis	**6**	2%

Pergunta 23: Caso o produto não existisse em *stock*, ficaria contente se fosse substituído por um semelhante?

Sim	**92**	**35%**

Não	**170**	65%

Pergunta 24: Suponha agora que, em vez de comprar apenas um produto, realizou uma encomenda com vários produtos. Dessa encomenda, um deles não foi entregue e ficou bastante chateado. Se fosse outro produto não, mas este não podia mesmo faltar. Porquê?

A encomenda era uma receita e este produto era mesmo necessário	**89**	**34%**
Consumo estes produtos todos os dias	**90**	34%
O produto estava em promoção e foi a razão por ter feito a encomenda	**67**	26%
Outro	**16**	6%

Pergunta 25: Perante uma situação destas, desistiria logo na primeira vez que isto acontecesse?

Sim	**118**	**45%**
Não	**144**	55%

Pergunta 26: Supondo que nunca mais utilizou este serviço, devido à situação anterior, voltaria a utilizar o serviço se?

Me fosse oferecido um vale ou cupão de desconto em compras	**79**	**30 %**
Houvesse algum contacto da parte do serviço de assistência a pedir desculpa	**130**	50 %
Não voltaria	**53**	20 %