INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




Sistema de Recomendação baseado em Data Mining

BRUNO PATRÍCIO FERREIRA
Julho de 2016

POLITÉCNICO
DO PORTO

Instituto Superior de Engenharia do Porto

Sistema de Recomendação

baseado em Data Mining

Bruno Patrício Ferreira

Dissertação para obtenção do Grau de Mestre em Engenharia Informática, Área de Especialização em Sistemas de Informação e Conhecimento

Orientador: Doutora Maria de Fátima Coutinho Rodrigues

Porto, Julho 2016

Dedico este trabalho aos meus pais,
Luís Martinho e Isabel Cristina

Resumo

A elevada competitividade que existe no mundo empresarial (provinda dos mais diversos fatores) força as empresas a estarem em constante evolução e/ou adaptação às tendências e mudanças que o mercado apresenta. Além do fator concorrência e outros demais, todas as empresas ambicionam, certamente, poderem melhorar e aperfeiçoar todo o âmbito do seu negócio.

Para tal, as empresas precisam de conhecer bem o seu negócio. Conhecer bem um negócio é por vezes uma tarefa complexa. Não é suficiente conhecer todos os números, processos e fluxos, é necessário também que se assimile verdadeiro conhecimento sobre a empresa.

Adquirir conhecimento é um processo que envolve análises a quantidades significativas de dados e informação. Com o conhecimento, provêm melhores decisões uma melhor gestão. Para auxiliar este processo recorre-se a técnicas de *Data Mining*. Este termo tem estado cada vez mais presente nas empresas pela sua enorme potencialidade na descoberta de padrões implícitos nos dados que permitem conhecer verdadeiramente a informação que está presente nos sistemas que gerem a empresa, e consequentemente, advém o conhecimento profundo do negócio.

Esta dissertação tem como propósito elaborar um projeto, que visa auxiliar a cadeia de perfumarias MASS neste processo de aquisição de conhecimento relativamente ao seu negócio. Para este efeito, será realizada uma aplicação que irá gerar análises gráficas, classificações de clientes, análises de associação e recomendação de produtos. Pretende-se assim desenvolver e descrever análises que permitam que os gestores desta empresa conheçam melhor os seus clientes, os seus produtos e as suas vendas, ou seja, que aprofundem o conhecimento sobre o seu negócio.

Palavras-chave: Inteligência Empresarial, Mineração de Dados, Classificação, Regras de Associação, Sistema de Recomendação

Abstract

The high competitive edge that exists in the business world (provided by various reasons) leads companies to be constantly evolving and adapting to the trends and changes that the market presents. In addition to the competition and other factors, all companies certainly aspire to improve the entire scope of their business.

To address these facts, companies need to know well with their entire business. Knowing well a business is often a complex task. It is not enough to know all the numbers, processes and work flows, it is also necessary to acquire knowledge about the company.

Acquiring knowledge is a process that involves analysis to significant amounts of data and information. With knowledge comes better decisions for better management. To assist this process *Data Mining* techniques can be very helpful. This term has been increasingly present in companies for its potential in the discovery of patterns implicit in the data that allow managers to truly know the information that is present in the systems that manage the company, and with that, comes the deep knowledge about the business.

The main purpose of this thesis is a project aims to assist MASS perfumeries in this process of knowledge acquisition about their business. For this purpose, it will be made an application that performs graphical analyses, customer clustering, association analysis and product recommendation. Is intended to develop and describe analyses that allow this company to know better their customers, their products and their sales, that is, to deepen their knowledge about the company.

Keywords: Business Intelligence, Data Mining, Classification, Association Rules, Recommendation System

Agradecimentos

O meu maior obrigado aos meus pais, Luís Martinho e Isabel Cristina por serem os maiores responsáveis por todo o meu percurso académico e profissional. Não tenho como agradecer o suficiente toda a motivação, conselhos, confiança, paciência, sacrifício pessoal e financeiro.

Agradeço à Doutora Maria de Fátima Coutinho Rodrigues pela enorme disponibilidade, dedicação e a imprescindível ajuda prestada neste projeto. Agradeço também todos os conhecimentos transmitidos que serão importantes na minha futura carreira profissional.

Um obrigado especial à minha namorada Fátima Pereira. A ela agradeço toda a compreensão, tempo, carinho e motivação dados incondicionalmente nos momentos mais complicados.

Agradeço ao Engenheiro Agostinho Vieira e à MASS Perfumarias pela oportunidade que me facultaram em realizar este projeto e de me fornecerem todas as componentes que foram necessárias.

Não posso deixar de agradecer também aos meus amigos, Eng.º André Ribeiro, Eng.º Ricardo Moreira, Eng.º João Oliveira pela motivação e conhecimentos partilhados.

A todos, muito obrigado.

Índice

1 Introdução 1

1.1 Contexto e Problema 1

1.2 Abordagem Preconizada 2

1.3 Análise de Valor 3

1.4 Estrutura da dissertação 4

1.5 Resultados Atingidos 5

2 Contexto e Estado da Arte 7

2.1 Análise e Contexto 7

2.2 Análise de Valor 9

2.3 Estado da Arte 14

2.4 Tecnologia Relevante 18

2.4.1 C5.0 18

2.4.2 Medidas de avaliação de modelos de classificação 19

2.4.3 Apriori 21

2.4.4 Medidas de avaliação de regras de associação 22

2.5 Exemplos de Aplicações 23

3 Avaliação de Soluções e Abordagem 29

3.1 Tecnologias 29

3.1.1 SQL Server 29

3.1.2 C# 29

3.1.3 R 30

3.2 Abordagem 31

4 Projeto 33

4.1 Design 33

4.2 Arquitetura 34

5 Dados 37

5.1 Análise 37

5.2 Fonte de dados 39

5.3 Staging Area 40

5.4 Processo ETL 45

5.4.1 Clientes 47

5.4.2 Produtos 49

5.4.3 Vendas 52

5.5 Exploração 54

5.6 Avaliação de Resultados 56

6 Classificação 57

6.1 Definição 58

6.2 Implementação 59
6.2.1 Preparação de Dados para Classificação 59
6.2.2 Conjuntos Candidatos 60
6.2.3 Simplificação de Modelos 64
6.2.4 Comparação de Modelos 69
6.2.5 Resultados 73

6.3 Avaliação de Resultados 74

6.4 Funcionamento 75

7 Regras de Associação 77

7.1 Definição 78

7.2 Implementação 79

7.3 Avaliação de Resultados 82

7.4 Funcionamento 83

8 Recomendação de Produtos 85

8.1 Definição 86

8.2 Implementação 87

8.3 Avaliação de Resultados 89

8.4 Funcionamento 91

9 Conclusões 93

9.1 Objetivos e Trabalho Futuro 93

9.2 Apreciação Final 95

Referências 97

Anexos 102

Anexo I – Product recommendation based on shared customer's Behaviour 102

Anexo II – Ficheiro Excel Processo ETL: Clientes 114

Anexo III – Ficheiro Excel Processo ETL: Artigos 126

Anexo IV – Ficheiro Excel Processo ETL: Vendas 143

Lista de Figuras

Figura 1 - Processo de descoberta de conhecimento Fonte: [10] 15
Figura 2 - Matriz de Confusão Fonte: [18] 19
Figura 3 - Exemplo R COM 30
Figura 4 - Exemplo R.NET 30
Figura 5 – Diagrama de Componentes 34
Figura 6 - Esquema do Processo ETL 47
Figura 7 – Pirâmide de valor de cliente Fonte: [44] 57
Figura 8 - Assinatura da função de simplificação de modelos 64
Figura 9 - Fluxograma de ModelAdjustment 65
Figura 10 - Output da função *ModelAdjustment* 66
Figura 11 - Output de percentagem mínima de 20% 66
Figura 12 - Exemplo de Curva de Aprendizagem 66
Figura 13 - Assinatura da função *ModelComparison* 69
Figura 14 - Fluxograma da função *ModelComparison* 70
Figura 15 – Parte do *Output* de *ModelComparison* 71
Figura 16 – Parte do Output de *ModelComparison* 71
Figura 17 - Exemplo Curva ROC 71
Figura 18 - Árvore de decisão do modelo de classificação 73
Figura 19 - Interface de classifcação de clientes 75
Figura 20 - Apresentação da classificação de clientes 76
Figura 21 - Volume anual de vendas da amostra 77
Figura 22 – Algoritmo de geração de associações entre artigos 81
Figura 23 – Interface de regras de associação 83
Figura 24 – *Output* das regras de associação 83
Figura 25 – Algoritmo de recomendação 88
Figura 26 - Interface de recomendação 91
Figura 27 - *Output* de recomendação 91

Lista de Tabelas

Tabela 1 - Modelo Canvas 12
Tabela 2 – *Staging Area*: Clientes Fidelidade 42
Tabela 3 – *Staging Area*: Produtos 43
Tabela 4 – *Staging Area*: Valores Auxiliares 43
Tabela 5 – *Staging Area*: Vendas de Clientes Fidelidade 44
Tabela 6 - Produtos e caraterizações por família 51
Tabela 7 - Resumo do conjunto candidato 1 60
Tabela 8- Matriz de confusão do modelo gerado com o conjunto candidato 1 61
Tabela 9 - Tabela resumo do conjunto candidato 2 61
Tabela 10 - Matriz de confusão do modelo gerado com o conjunto candidato 2 62
Tabela 11 - Resumo do conjunto candidato 3 62
Tabela 12 - Matriz de confusão do modelo gerado com o conjunto candidato 3 63
Tabela 13 - Valores resultantes da simplificação do conjunto 1 67
Tabela 14 - Valores resultantes da simplificação do conjunto 2 68
Tabela 15 - Valores resultantes da simplificação do conjunto 3 68
Tabela 16 – Avaliação modelo de classificação 74
Tabela 16 – Avaliação regras de associação 82
Tabela 17 - Matriz de demonstração da função: C x I → R 86
Tabela 18 – Avaliação 1 sistema de recomendação 89
Tabela 19 – Avaliação 2 sistema de recomendação 90

Acrónimos e Símbolos

Lista de Acrónimos

- ERP – Enterprise Resource Planning
- ETL – Extração Transformação e Carregamento
- TNR – Taxa de acerto na classe Negativa
- FPR – Taxa de falsos positivos
- TP – Positivos verdadeiros
- FN – Falsos negativos
- FP – Falsos positivos
- TN – Negativos verdadeiros
- AED – Análise exploratória de dados
- LHS – Lado esquerdo da regra
- RHS – Lado direito da regra

1 Introdução

1.1 Contexto e Problema

Desde que as empresas recorrem a computadores e a programas informáticos para a sua gestão, têm vindo a armazenar muitos e valiosos dados. Estes dados possuem inúmeros tipos de informações, dependendo do tipo de organização. Por exemplo: dados sobre artigos, produção, funcionários, clientes, fornecedores, vendas, distribuição, logística, transações, etc. Com o decorrer dos anos as empresas tendem a armazenar grandes volumes de dados. Este armazenamento já proporciona por si só uma grande vantagem, na medida em que se recorre cada vez menos ao uso do papel e a informação passa a estar guardada informaticamente e de forma estruturada.

No mundo empresarial tão competitivo como o que vivemos atualmente, as empresas têm uma necessidade cada vez maior de se destacar no mercado onde se encontram. Como consequência da concorrência, os clientes são cada vez mais exigentes e procuram soluções de melhor qualidade, com um melhor retorno financeiro. Então, optam por melhorar os seus produtos e serviços, contratar colaboradores mais qualificados, investir em melhores equipamentos e em novas tecnologias. No entanto, por vezes não aproveitam um recurso vital que têm ao seu dispor armazenado nos seus sistemas, que é o conhecimento que os sistemas de armazenamento têm presente nos seus dados.

Os dados armazenados tem muito mais utilidade do que para fins meramente histórico ou de consulta. Os dados além de simples informação armazenada com o decorrer dos anos, contêm também conhecimento sobre a empresa. Este conhecimento é superior ao conhecimento adquirido ou empírico que os gestores possam ter sobre a empresa. Encontra-se enraizado nos dados, mas dificilmente acessível a quem precisa dele. Existem empresas que atualmente armazenam *terabytes* de dados diariamente, o que se reflete em muito conhecimento que pode ser extraído, analisado e canalizado. Para extrair dos sistemas de armazenamento o conhecimento que lhe é intrínseco recorre-se a técnicas de *Data Mining*.

É neste âmbito, de fornecer conhecimento que já existe nos sistemas de informação, que visam auxiliar as empresas nas suas decisões que esta tese pretende atuar. Pretende-se desenvolver e explorar soluções que permitam à empresa MASS Perfumarias explorar o conhecimento implícito nos seus dados, para que esta possa também prosperar recorrendo à informação que está implícita nos seus sistemas.

1.2 Abordagem Preconizada

A MASS Perfumarias possui uma base de dados de tamanho elevado, com mais de 15 anos de idade e um conjunto enorme de dados por explorar. Estão contidos dados que permitem explorar o crescimento da empresa, avaliar o comportamento e preferências dos seus clientes, caracterizar os seus clientes e sobretudo permitem realizar análises preditivas e descritivas. Com este conjunto poderoso de dados, estão reunidas boas condições para que se possa aplicar técnicas de *Data Mining* e assim fornecer à empresa conhecimento que a permita crescer e obter melhores resultados.

O objetivo principal deste projeto é fornecer à MASS Perfumarias uma aplicação de recomendação de artigos, que lhe permita obter conhecimento relativo ao seu negócio e assim apoiar a tomada de melhores decisões no futuro, recorrendo a técnicas de *business intelligence*. Esta aplicação irá fornecer diversas análises (a clientes, produtos e vendas), relativas ao período que o utilizador pretenda, tendo em conta diversos parâmetros definíveis (tais como períodos de data, conjunto de clientes específicos, grupos de produtos, entre outros).

Esta aplicação terá também a capacidade de garantir que todos os dados necessários à execução das análises estão corretamente preparados, formatados e carregados. Isso permitirá à aplicação ter um processo totalmente automatizado de extração, transformação e carregamento das informações necessárias, de forma totalmente transparente para os utilizadores.

Relativamente às análises que serão efetuadas, estas podem ser divididas em três grandes grupos:

- Criar modelos preditivos de clientes, que permitam criar uma classificação de clientes.
- Levantar um conjunto de regras de associação que permitam à empresa saber quais os artigos que mais se vendem conjuntamente.
- Implementar uma área de recomendação de artigos para clientes tendo em conta a sua classificação.

Pretende-se também disponibilizar ao utilizador análises gráficas de teor estatístico referente a clientes, vendas e artigos. Estas análises tem o objetivo mostrar a empresa, de forma gráfica, alguns valores importantes referentes aos seus dados.

1.3 Análise de Valor

Este projeto consiste numa aplicação concebida para a empresa MASS perfumarias que visa apresentar aos gestores conhecimento implícito ao seu negócio, que está presente nos seus sistemas de informação. Esta empresa conta com 30 anos de existência no mercado com a venda de produtos na área de perfumaria, cosmética, maquilhagem e higiene corporal. Com as suas 9 lojas físicas e uma loja *online* tem vindo a acumular enormes volumes de dados, mais concretamente no seu ERP mySoftmais [1].

Com este grande volume de dados por explorar, esta aplicação objetiva auxiliar a empresa a tomar melhores decisões operacionais e estratégicas, com base nos dados que esta tem presente no seu ERP. Melhores decisões baseadas no conhecimento apresentado pela aplicação terão como consequência direta, o aumento do volume de faturação e o aumento de novos clientes recorrentes à loja.

Esta aplicação apresenta algumas características que a diferencia claramente das restantes soluções semelhantes:

- Adaptada às necessidades e características da empresa e dos seus gestores.
- Adaptada aos sistemas de informação da empresa.
- Comunicação direta entre os intervenientes.

Outras aplicações existentes no mercado não são dedicadas a nenhuma empresa em concreto, o que obriga a um grau de abstração quer ao tipo de dados disponíveis, quer ao tipo de informação possível de ser aplicada. Estas têm também como consequência um número de configurações e interações muito superior, o que leva à necessidade de o utilizador precisar de alguns conhecimentos técnicos e teóricos sobre B*usiness Intelligence*.

Com um sistema desta natureza a empresa poderá prosperar através de melhores decisões, tomadas com base no apoio de conhecimento extraído dos dados do seu ERP. A aplicação será capaz de extrair conhecimento que permite á empresa otimizar o seu *marketing* e as suas campanhas promocionais através das seguinte funcionalidades:

- Obter grupos de clientes pelas suas caraterísticas e comportamentos.
- Saber quais os produtos que mais se vendem conjuntamente.
- Saber quais os produtos mais vendidos para cada grupo.

1.4 Estrutura da dissertação

Esta dissertação encontra-se dividida em 10 principais capítulos, que poderão por sua vez serem subdivididos.

O primeiro capítulo descreve a introdução à dissertação numa perspetiva global da problemática em questão, com o propósito de descrever no que consiste o projeto. É apresentado de forma sumária o contexto e o problema, os principais objetivos a que se propõe e a sua análise de valor.

O segundo capítulo pretende descrever a problemática em detalhe. Aqui é apresentado a MASS Perfumarias, a Softingal e a dimensão dos seus dados, quais os intervenientes deste projeto, importantes conceitos sobre o modelo de negócio do cliente e o seu ERP. É também apresentado uma análise de valor mais pormenorizada e o estado da arte desta dissertação.

O terceiro capítulo tem como objetivo relacionar o problema com as diferentes soluções e tecnologias disponíveis. Apresenta as diferentes ferramentas e tecnologias abordadas para a resolução do problema e quais as que serão utilizadas. Este capítulo descreve ainda a abordagem completa dividida por diferentes fases para elaboração deste projeto.

O quarto capítulo apresenta uma análise detalhada referente aos dados utilizados neste projeto. Refere pormenores sobre a base de dados que contém toda a informação do cliente e quais são os dados necessários a extrair vitais aos objetivos deste projeto. Apresenta ainda a localização e estrutura de informação que armazena estes dados extraídos, tratados e processados. Relata ainda a abordagem de avaliação à qualidade da informação extraída.

O quinto capítulo tem como objetivo descrever a aplicação que é criada. Apresenta detalhes sobre a sua arquitetura e o seu funcionamento e também os requisitos funcionais. Apresenta-se ainda a relação entre o projeto e como este transpõe ao utilizador os objetivos a que se propõe.

O sexto, sétimo e oitavo capítulos descrevem respetivamente em detalhe a abordagem, os algoritmos e os resultados obtidos para a geração de um modelo de classificação de clientes, para a análise de regras de associação e para a recomendação de artigos. Estes três capítulos são responsáveis pelos três grandes objetivos deste projeto.

O nono capítulo é responsável por apresentar as conclusões da dissertação, limitações, trabalho futuro, objetivos realizados e apreciação final deste projeto.

1.5 Resultados Atingidos

Uma das primeiras e mais importantes fases deste projeto é o processo ETL. Processo este que serve de base e reforça a qualidade das análises *Data Mining* realizadas. Após inúmeras iterações a este processo entre implementação e avaliação, foi possível obter um processo que garante a fiabilidade, coerência e padronização de todos os dados necessários.

Um dos objetivos propostos é criar um modelo preditivo de classificação de clientes fidelidade. Este objetivo foi atingido. A aplicação em questão é capaz de criar um modelo de classificação e aplicar este modelo a novos clientes fidelidade, classificando-os num de 5 grupos.

Relativamente a regras de associação é possível gerar regras a diferentes níveis como: marca, tipo e departamento. A classificação de clientes permitiu ir além do idealizado para este objetivo. A aplicação permite também criar regras de associação aplicadas a cada um dos grupos de cliente, permitindo ver o que mais se vende conjuntamente para cada grupo de clientes.

De todos os objetivos propostos destaca-se a recomendação de artigos a clientes. Este objetivo foi também concluído na sua totalidade, satisfazendo todos os requisitos necessários. A aplicação permite recomendar artigos, marcas, tipos e departamentos a cada cliente individualmente. É também possível realizar esta recomendação a cada grupo de clientes.

Para assegurar que todo o desenvolvimento realizado cumpre também a qualidade a que se propôs, avaliou-se devidamente cada um dos objetivos realizados, com base em métricas e métodos de avaliação relativos aos algoritmos realizados. Além de concluídas todas as implementações propostas, e uma análise de resultados muito positivos, é possível concluir que este projeto cumpre com sucesso todos os objetivos a que se propôs, segundo boas práticas e conceitos de desenvolvimento e avaliação.

2 Contexto e Estado da Arte

2.1 Análise e Contexto

A MASS perfumarias é uma cadeia de lojas, 9 físicas e uma loja *online,* que se distribuem pelo grande Porto e Braga. Com mais de 30 anos de existência, a empresa comercializa produtos de perfumaria, cosmética, maquilhagem e higiene corporal. A empresa aposta na fidelização de clientes, recorrendo a vantagens que apresenta com o seu cartão fidelidade e a preços mais baixos relativamente a grandes empresas da concorrência. Esta organização conta com mais de 17,000 clientes fidelidade e tem disponíveis mais de 15,000 artigos para venda. Esta empresa apresenta grandes volumes de dados nos seus sistemas, armazenando mais de 10GB de transações de artigos relativos a um períodos de 4 anos.

Esta empresa para realizar toda a sua gestão (compras, vendas, stocks, tesouraria, salários, recursos humanos) recorre ao ERP *mySoftmais*. O *mySoftmais* é um *ERP* desenvolvido pela Softingal [2] para pequenas e médias empresas. Este ERP existe há 15 anos no mercado e é comercializado por todo o país encontrando-se em funcionamento nos mais diversos tipos de empresas. Todos os dados a serem explorados neste projeto estão armazenados na base de dados do *mySoftmais*. Esta base de dados apresenta uma enorme dimensão fruto da complexidade exigida a este *software*. Dado o volume e a importância dos dados armazenados nos sistemas da empresa, é notório que existe uma grande probabilidade de existir muito conhecimento de negócio que será novo para os gestores da empresa. Este sistema reúne assim excelentes condições para que sejam aplicadas técnicas de *Data Mining* e assim explorado o conhecimento que existe neste.

Para auxiliar a empresa neste processo de exploração é proposto um sistema de recomendação capaz de ler os dados necessários do *mySoftmais*, realizar todo um processo de ETL e aplicá-los nos algoritmos de *Data Mining* necessários de forma totalmente transparente ao utilizador. Este tem apenas que solicitar uma ação indicando o período de análise. A aplicação tratará também de converter a informação gerada pelos algoritmos de *Data Mining* num formato facilmente compreensível pelos gestores da empresa. A informação e conhecimento que este projeto apresenta ao utilizador advém de diferentes naturezas:

- Classificativa: permite obter grupos pré-definidos de clientes com base nas suas caraterísticas e comportamentos.
- Associativa: permite identificar quais os artigos que mais se vendem conjuntamente.
- Recomendativa: com grupos de utilizadores e grupos de artigos que mais se vendem conjuntamente, o sistema é capaz de recomendar artigos personalizadamente a cada cliente.
- Estatística: é apresentado ao utilizador algumas análises estatísticas refentes aos seus clientes, artigos e vendas.

Em suma, este projeto consiste numa aplicação elaborada num ambiente familiar aos gestores desta empresa, capaz de lhes apresentar as informações pretendidas no momento face a um dado período de análise. Esta informação terá como objetivo auxiliar a MASS Perfumarias nas suas decisões estratégicas e operacionais ao classificar clientes em grupos, ao identificar qual a relação entre os artigos que mais se vendem e ao recomendar artigos a grupos de clientes.

A MASS Perfumarias pretende aumentar o número anual de vendas e conquistar novos clientes. Fatores que são possíveis ao conseguir direcionar melhor o seu *marketing*, as suas campanhas e as suas promoções. Neste sentido, este projeto é uma grande mais-valia, porque fornece informação de apoio aos gestores da empresa na hora de tomar decisões que têm influência na conquista destes objetivos.

A empresa não tem até à data uma ferramenta ou analista que consiga fornecer este tipo de informação necessária ao seu maior crescimento. Esta solução além de lhes colmatar esta falta de informação é ainda valorizada pela personalização do projeto exclusivamente à MASS Perfumarias, tanto aos seus dados, como às particularidades do seu negócio. Outra solução existente para este problema está condicionada pela disponibilidade da empresa para a configurar, explorar e aprender a interpretar os resultados obtidos, sem a necessidade de um analista nesta matéria.

O conjunto inicial de dados fornecidos pela empresa tem um grande impacto na construção deste projeto, na medida em que condiciona os mecanismos criados no processo de recolha e tratamento dos dados extraídos do *mySoftmais*. Existem particularidades nos dados que exigem tratamentos/transformações específicas que só são possíveis de detetar com um conjunto considerável de casos de estudo. Condiciona também a execução dos algoritmos de *Data Mining* porque a seleção destes algoritmos deve estar em concordância com o volume e tipo de dados disponíveis. A amostra inicial disponibilizada é um *snapshot* da base de dados do *mySoftmais* com dados de vendas relativos a 4 anos. Esta amostra demonstra ser suficiente para o desenvolvimento deste projeto.

2.2 Análise de Valor

O principal objetivo de uma análise de valor consiste em analisar como aumentar o valor de um produto ou serviço ao menor custo sem sacrificar a qualidade [3]. Esta frase realça a ideia de que uma proposta de valor deve ser bem avaliada devido ao impacto que terá na distinção de um produto ou serviço no mercado. Nesta avaliação deverão ser tidos em conta parâmetros como: características, vantagens/benefícios para todos os envolventes, empresas e produtos concorrentes, custos, preços e retorno.

Este projeto tem como finalidade proporcionar à MASS Perfumarias uma aplicação de recomendação, que lhes permita obter conhecimento sobre a empresa com o fundamento de apoio para futuras melhores decisões. Este conhecimento será recolhido aplicando técnicas de *Data Mining* sobre os dados que são operados diariamente nos seus sistemas de informação. Pretende-se desenvolver uma aplicação que seja capaz de disponibilizar aos gestores conhecimento relativo a clientes, produtos e vendas, sempre que surja a necessidade. A aplicação deverá também ser configurável de forma a ser possível definir diferentes períodos de análise e terá de garantir um funcionamento independente da disponibilidade de acesso à base de dados do ERP *mySoftmais*.

Melhores decisões têm como consequência um maior interesse por parte dos clientes, que por sua vez têm como beneficio o aumento do número anual de vendas e a aquisição de novos clientes. O negócio de perfumaria e produtos de beleza tende a ser sazonal, como tal, em certas alturas do ano (como o dia dos namorados, o dia da mãe, o natal, etc.) o correto direcionamento de determinados artigos ao seu público-alvo é fundamental. Principalmente por serem ocasiões que em que as empresas concorrentes apostam muito na promoção e divulgação. Esta aplicação vem ajudar os gestores na importante questão do direcionamento da publicidade e das campanhas.

Numa perspetiva longitudinal, inicialmente, o cliente ao ser alvo de sugestões com maior fundamento mais atenção desperta sobre os produtos sugeridos, o que proporciona novas visitas às lojas. De seguida, os clientes que já conhecem a MASS Perfumarias e os novos clientes após as primeiras compras, terão uma maior tendência em comprar novamente porque com dados de histórico o sistema conseguirá realizar melhores recomendações. Com melhores recomendações e uma melhor gestão de *marketing* advém a fidelização de clientes. Clientes estes que ao comprarem frequentemente ao longo do ano, acumulam pontos em cartão. Estes pontos em cartão são por si também um incentivo à compra. Como desvantagem os gestores terão de disponibilizar o seu tempo na interação com a aplicação e uma cuidada avaliação/interpretação da informação apresentada.

A empresa pode avaliar o retorno da utilização desta solução comparando diversos critérios (tais como: volume de vendas, número de novos clientes, etc.) com um período homólogo. Esta avaliação pode também ser feita pelo confronto de resultados apresentados pela aplicação com algumas ideias empíricas dos gestores sobre determinados casos.

O modelo de negócio utilizado apresenta uma negociação interativa *win-win*, na medida em que a MASS Perfumarias beneficiará de uma ferramenta de apoio à decisão fornecendo os dados que tornam possíveis a construção deste projeto. É importante que a empresa esteja ciente de que tipo de informações a aplicação é capaz de apresentar e o tipo de possibilidades/ oportunidades que esta proporcionará. É também relevante que a aplicação seja capaz de dar resposta de forma clara e objetiva aos objetivos propostos ao cliente. Caso a solução por motivos técnicos não consiga dar resposta ao que se propôs, é importante dar uma explicação à MASS Perfumarias do porquê numa linguagem acessível a esta.

Para apresentar esta ideia de negócio numa perspetiva estratégica, recorre-se ao modelo Canvas. Este é um mapa visual pré-formatado que contém nove blocos do modelo de negócios, que permitem descrever modelos de negócio novos ou existentes [4]. O modelo Canvas deste projeto encontra-se ilustrado na Tabela 1. O modelo de negócio deste projeto é um sistema de recomendação concebido para esta empresa, que visa auxiliar os seus gestores em melhores tomadas de decisão, tendo em conta um melhor direcionamento dos seus produtos. Como este modelo sugere:

- O segmento de clientes é a Mass Perfumarias cujas suas principais características são a venda de artigos de beleza e perfumaria. Este segmento é muito específico porque este projeto é idealizado exclusivamente para esta empresa, atendendo unicamente aos seus dados e às suas particularidades de negócio.
- A proposta de valor é uma aplicação de recomendação que lhes permita obter conhecimento relativo ao seu negócio e assim apoiar a tomada de melhores decisões no futuro. Melhores decisões terão como consequência um maior volume de vendas conseguindo responder de forma mais assertiva às necessidades dos clientes.
- O canal de comunicação e a relação com o cliente é feita de forma próxima através de reuniões e *email*.
- Os recursos chave para este projeto são: equipamento informático e todo o *software* necessário para cumprir os seus requisitos funcionais como *Visual Studio*, *SQL Server* e R Studio.
- As atividades chave neste projeto são a construção de aplicações, suporte ao cliente na interpretação de resultados e aprofundamento de conteúdos.
- Os parceiros-chave ao sucesso deste projeto são empresas especializadas em *business intelligence* para o auxílio na construção de novas formas de melhor cumprir os objetivos propostos. Também a Softingal, na medida em que detém o conhecimento sobre a estrutura da fonte de dados e como estão contidos os dados necessários a este projeto.
- As estruturas de custo são o local de trabalho/instalações, todo o equipamento informático e *software* necessários à implementação/continuidade do projeto.
- As fontes de receita deste projeto são as renovações anuais de licenças de utilização assim como as avenças mensais de assistência.

Tabela 1 - Modelo Canvas

Parceiros Chave	**Atividades Chave**	**Proposta de Valor**	**Relação com Cliente**	**Segmento de clientes**
Empresas e Consultoras em *business intelligence* Especificamente: Softingal	Construção de aplicações Suporte ao cliente na interpretação de resultados Aprofundamento de conteúdos	Aumentar volume de vendas das empresas Responder adequadamente às necessidades dos clientes Detetar potenciais clientes Reter clientes	Produto dedicado ao cliente Acompanhamento na análise de resultados Assistência especializada e dedicada à empresa	Mass Perfumarias
	Recursos Chave		**Canais de Comunicação**	
	Equipamento informático Software necessário		Reuniões Email	
Estrutura de Custos			**Fontes de Receita**	
Local de Trabalho Equipamento Informático Licenças de Software			Renovações de Licença Avenças de assistência	

A criação de valor pode ser modelada com o Modelo Conceptual para Decomposição do Valor para o Cliente (MCDVC) [5]. Esta decomposição é feita por 4 conceitos: formas de valor e posições de valor temporal; Redes de valor em redes de partilha; o conceito de ativos endógenos e exógenos da empresa; conceito de benefícios e sacrifícios. A decomposição é feita em 3 passos.

1. Construir a rede de valor com o contacto com o cliente identificando como dividir os ativos endógenos e exógenos do produto final. O objetivo é criar componentes simplificados de valor para o cliente.
2. Perceber como o cliente avalia o valor dos componentes criados com objetivo de conhecer a perceção de benefícios e sacrifícios da empresa.
3. Ajustar a posição de valor com base nos benefícios/sacrifícios dos ativos fornecidos, em função da perceção do cliente em relação aos resultados dos benefícios/sacrifícios entendidos por este.

2.3 Estado da Arte

Data Mining é o processo de descoberta de conhecimento a partir de uma fonte com uma quantidade significativa de dados, seja uma base de dados, um *data warehouse* ou qualquer outro tipo de repositório [6].

Este termo nasceu nos anos 90 juntamente com base de dados multidimensionais e novos sistemas com maiores capacidades de armazenamento e processamento. Tem como principal objetivo efetuar análises que permitam contribuir para melhores decisões em diversas áreas tais como: medicina, finanças, marketing, produção, telecomunicações, etc. O *Data Mining* analisa os dados, descobre problemas ou oportunidades escondidas nos relacionamentos, e diagnostica o comportamento dos negócios, requerendo a mínima intervenção do utilizador e assim este poderá dedicar-se apenas na filtragem do conhecimento obtido e na sua interpretação para assim serem produzidas mais vantagens competitivas [7].

A mineração de dados pode gerar novas oportunidades de negócios por duas formas distintas [8]:

- Previsão automatizada de tendências e comportamentos: é possível automatizar o processo de encontrar informações preditivas numa grande base de dados. Perguntas que tradicionalmente exigiam extensas análises manuais, agora podem ser respondidas diretamente a partir dos dados. Um exemplo típico de um problema de previsão é o marketing. A mineração de dados utiliza os dados sobre mailings promocionais do passado para identificar os alvos mais prováveis para maximizar o retorno sobre o investimento em futuras campanhas de *marketing*.

- A descoberta automática de padrões previamente desconhecidos: as ferramentas de mineração de dados através de bases de dados, permitem identificar padrões implícitos nos dados previamente ocultos. Um exemplo de padrão de descoberta é a análise dos dados das vendas para identificar produtos aparentemente não relacionados que são frequentemente comprados juntos. Outros problemas incluem a deteção de padrões de descoberta de diversas naturezas, como por exemplo: transações de cartão de crédito fraudulentas.

O *Data Mining* faz parte de um processo iterativo e interativo de descoberta de conhecimento denominado de KDD – *Knowledge Discovery* from *Data.* A Figura 1 ilustra todos os passos deste processo e onde se enquadra o *Data Mining* [9].

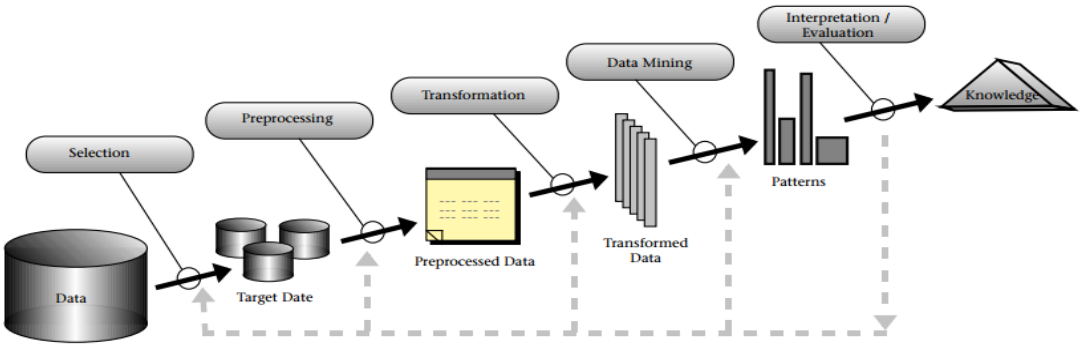


Figura 1 - Processo de descoberta de conhecimento
Fonte: [10]

- Seleção dos Dados: Nesta fase são selecionados apenas os conjuntos ou subconjuntos de dados que terão relevância às análises a serem efetuadas. Esta seleção pode ser feita a qualquer tipo de repositório de informação individual ou a um *data warehouse* que já contenha a integração de vários sistemas de armazenamento.

- Pré-processamento dos dados: Aqui o principal objetivo é assegurar a qualidade dos dados que foram previamente selecionados na fase anterior. Nesta fase realiza-se a limpeza dos dados. A limpeza é composta por operações que envolvem a verificação da consistência dos dados, a correção de possíveis erros e o preenchimento ou a eliminação de valores em falta e/ou redundantes. No pré-processamento são também identificados e removidos dados duplicados e/ou com ruído.

- Transformação dos Dados: Esta fase tem como objetivo converter o conjunto bruto de dados provenientes do pré-processamento e colocá-los num formato padronizado para poder ser posteriormente utilizado pelos algoritmos de *Data Mining*.

- *Data Mining*: é o passo crucial em que os algoritmos de descoberta de conhecimento são aplicados aos dados já devidamente preparados, em que será extraído conhecimento potencialmente útil.

- Interpretação e avaliação do conhecimento: após aplicadas todas as operações de *Data Mining* é necessário apresentar ao utilizador os resultados destas operações. Cabe nesta fase também interpretar o conhecimento que foi descoberto, validando-o através de medidas da qualidade e da perceção de um analista com conhecimentos sobre a área de negócio em questão.

As técnicas analíticas utilizadas na mineração de dados são maioritariamente algoritmos e técnicas estatísticas já bastante conhecidas. O que existe novo é que estas técnicas podem agora ser aplicadas à grande maioria dos negócios para problemas gerais, graças à maior disponibilidade dos sistemas de armazenamento de dados e o maior poder de processamento. Além disso, a evolução dos sistemas de interfaces gráficas permitem disponibilizar os resultados obtidos aos especialistas no negócio, que poderão mais facilmente fazer uso desta informação. Algumas das técnicas usadas para mineração de dados são:

- Redes neurais artificiais: Modelos preditivos não-lineares que aprendem através de treino e assemelham-se a redes neurais biológicas na sua estrutura.

- As árvores de decisão: Estruturas em forma de árvore, que representam conjuntos de decisões. Estas decisões geram regras para a classificação de um conjunto de dados.

- Indução de regras: A extração de regras úteis se-então dos dados que contenham significância estatística.

- Algoritmos genéticos: Técnicas de otimização baseadas nos conceitos de combinação genética, mutação e seleção natural.

- K-Vizinhos mais próximos: Uma técnica de classificação que classifica cada registo com base nos registos mais próximos de acordo com uma função de distância.

O Data Mining como qualquer outra tecnologia possui as suas desvantagens assim como certas condições para a sua utilização:

- Aplicar algoritmos desta natureza sobre muitos *gigabytes* de informação pode levar a tempos de execução algo elevados.

- Elevado custo da sua implementação.

- Necessita de volumes consideráveis de dados. É preciso também prestar grande atenção sobre os dados que são enviados a serem explorados. A qualidade destes refletirá a qualidade do conhecimento extraído.

- Como é uma tecnologia algo recente, ainda não estão disponíveis *softwares* que sejam capazes de expor os resultados obtidos a utilizadores mais inexperientes.

- É uma área nova e ainda bastante complexa e como tal precisa ainda de muita interação com analistas.

Nas empresas tomam-se frequentemente decisões baseadas na intuição do decisor. Estas decisões normalmente não têm o apoio do conhecimento implícito nos sistemas de gestão das empresas. Isto acontece porque nem sempre existe acesso a ferramentas que permitam extrair o devido conhecimento e que o permitam apresentar de forma facilmente interpretável. Esta realidade começa a ser um passado próximo. Existem cada vez mais empresas a prosperar em prol do conhecimento obtido pela utilização do *Data Mining*.

Seguem agora alguns exemplos reais da utilização do *Data Mining* em algumas empresas [11]:

- A rede americana Wall-Mart, pioneira no uso de *Data Mining*, descobriu ao explorar seus números que 60% das mães que compram bonecas *Barbie*, levam também barras de chocolate. Esta empresa descobriu também que as vendas de fraldas estavam fortemente ligadas às vendas de cerveja. Explicação: os pais que saiam à noite para comprar fraldas, compravam também cerveja. A empresa reorganizou as suas prateleiras e conseguiu um aumento de 30% de vendas destes artigos.

- A Sprint, um dos líderes no mercado americano de telefone de longa distância, desenvolveu um método capaz de prever com 61% de segurança, se um consumidor trocaria de companhia telefónica dentro de um período de dois meses. Com um marketing agressivo, conseguiu evitar o abandono de 120,000 clientes e uma perda de 35,000,000 de dólares em faturação.

Atualmente já existem várias empresas que recorrem à utilização do *Data Mining* e aplicam os seus resultados a adaptações e/ou ajustes nos seus negócios. Seguem agora alguns exemplos da sua utilização:

- Identificar padrões de compras inesperadas nos supermercados.

- Otimização de *websites* para maior rentabilidade em ofertas adequadas para cada visitante.

- Otimização de campanhas *e marketing*.

- Prever as taxas de resposta do cliente em campanhas de *marketing*.

- Prever abandono de clientes.

- Distinguir entre os clientes rentáveis e não rentáveis.

- Melhorar a produtividade em processos de produção complexos.

2.4 Tecnologia Relevante

Depois de realizado o estado da arte deste projeto e de ter em conta os objetivos a realizar, destacam-se dois algoritmos que são utilizados e que são vitais para o sucesso deste projeto. Estes algoritmos serão o C5.0 [12] que é responsável pela classificação de clientes e o *Apriori* [13] que é responsável pela geração de regras de associação e pela recomendação de artigos. É apresentada agora uma análise a cada um destes algoritmos onde são ponderadas as suas vantagens e desvantagens de utilização.

2.4.1 C5.0

O algoritmo utilizado neste projeto para gerar os modelos de classificação para clientes da MASS Perfumarias será o C5.0. Este algoritmo permite criar árvores de classificação que tem aplicabilidades como:

- Saber se um empréstimo pode ser aplicado a um dado cliente.
- Prever o abandono de um serviço.
- Criar perfis de utilizadores.
- Entre muitos outros domínios.

Este algoritmo é uma versão melhorada de dois algoritmos que são o ID3 e o C4.5. O algoritmo ID3 foi desenvolvido por Ross Quilan na década de 70. Na década de 80 este algoritmo evoluiu para uma versão melhorada denominada de C4.5. Em 2011 foi lançada uma nova versão evoluída chamada de C5.0. Esta versão permite gerar árvores de decisão mais pequenas, maior performance e melhor cálculo de custo de erros entre outras melhorias face a versão a anterior [14]. O C5.0 para cada nó da árvore, escolhe o atributo que divide de forma mais eficaz o conjunto de dados de acordo com o critério de ganho de informação. O algoritmo, iterativamente, repete este processo nas consequentes das divisões criadas.

Existem outros algoritmos semelhantes ao C5.0 no qual se destacam dois algoritmos

- CART: Este foi implementado por Breiman *et al* em 1984. [15]. Este algoritmo apresenta um comportamento semelhante ao C5.0 mas com algumas desvantagens como [14]:
    - Utiliza a entropia como critério de divisão dos dados.
    - A poda é feita segundo a abordagem pessimista.
    - Não suporta previsores categóricos.

  Outro fator de grande importância sobre o C5.0 é que este permite gerar mais de dois subgrupos para cada divisão oferecendo divisões não binárias (e como tal multi-classificação). Este utiliza o ganho de informação como critério de divisão e permite obter uma árvore de decisão ou uma descrição simples das divisões que foram encontradas pelo algoritmo. Este além de robusto em problemas como: falta de dados e grandes volumes de informação, tem ainda uma funcionalidade de *boosting* para aumentar a precisão da classificação [16]. Este está implementado e disponível em diversas linguagens e ferramentas de *Data Mining*.

- kNN: O k-Nearest Neighbors é um algoritmo também de capacidades preditivas embora tenha um comportamento diferente do CART e do C5.0, uma vez que não gera um modelo. Durante o processo de aprendizagem o que este algoritmo faz é apenas armazenar os dados de treino. Quando surge a necessidade de classificar uma dada entrada, o que este faz é calcular a entrada (k) mais próxima já classificada [17].
  A grande desvantagem deste algoritmo aplicado aos dados deste projeto é que este é bastante eficaz para dados contínuos, o que não acontece para dados discretos.

2.4.2 Medidas de avaliação de modelos de classificação

Quando se constrói um modelo de classificação, independentemente do algoritmo utilizado, é crucial que este seja devidamente avaliado. Na avaliação de modelos existem inúmeras técnicas e métricas passíveis de serem utilizadas.

Será agora apresentada uma lista de algumas das métricas utilizadas para esta avaliação, que são utilizadas também para avaliar os modelos deste projeto. Estas métricas apresentam grande valor quando avaliadas em conjunto. Ao excluir algumas destas como parâmetros de avaliação, aumenta o risco de uma percepção errada da qualidade do modelo.

As métricas apresentadas tem como base uma matriz de confusão ou matriz de acertos/erros, como a ilustrada na Figura 2. Esta matriz permite avaliar a aprendizagem para uma classe objetivo binária, realizada por um classificador com base num conjunto de teste.

		Classe Prevista	
		^Classe +	**^Classe -**
Classe Objectivo	**Classe +**	TP (++)	FN (+-)
	Classe -	FP (- +)	TN (- -)

Figura 2 - Matriz de Confusão
Fonte: [18]

- Tamanho: O tamanho de um modelo é um indicador da complexidade que o modelo tem. É uma medida aplicável quer o modelo seja gerado com base num conjunto de regras ou com recurso a uma árvore de decisão. Quanto maior o tamanho maior o conjunto de regras ou maior a árvore de decisão. Esta medida permite ao analista ter uma perceção da dimensão do modelo, que o auxiliará a decidir se o modelo necessitará de ser simplificado ou se existe margem para aumentar a precisão deste recorrendo a ajustes na dimensão da amostra de treino.

- Taxa de acerto: Esta métrica é uma das mais utilizadas e das que apresenta maior relevância de análise. Esta métrica representa a taxa de acerto de todo o classificador, isto é, a % de exemplos corretamente classificados pelo modelo. Ou seja, o acerto indica quanto melhor um dado modelo consegue prever corretamente um dado atributo objetivo.

$$Taxa\ Acerto = \frac{TP + TN}{TP + FN + FP + TN}$$

- Erros: O número de erros de um modelo indica o número total de vezes em que este falha ao prever um conjunto de registos teste. Esta medida apresenta maior relevância quando analisada com as métricas *Precision* e *Recall*.

$$Taxa\ Erro = 1 - Taxa\ Acerto$$

- Precision: Taxa de registos positivos no conjunto de registos que o classificador previu como positivos.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- Recall: Taxa de registos positivos, corretamente previstos pelo classificador.

$$Recall = \frac{TP}{TP + FN}$$

- TNR: Taxa de acerto na classe Negativa. Percentagem de conjuntos negativos corretamente previstos no modelo.

$$TNR = \frac{TP}{TP + FP}$$

- FPR: Taxa de Falsos Positivos. Percentagem de exemplos negativos previstos como positivos.

$$FPR = \frac{FP}{TN + FP}$$

- F1: Esta métrica é uma média harmónica ponderada entre *Precision* e *Recall*. Esta é importante na medida em que *Precision* e *Recall* não são adequadas para avaliar modelos inferidos a partir de dados com classes desbalanceadas [18].

$$F1 = \frac{2 * TP}{2 *\ TP + FP + FN}$$

- Kappa: É uma métrica que compara o acerto com uma precisão esperada. Este é usado, não só para avaliar um único classificador, mas também para avaliar classificadores entre si [19].
- Curva Roc: Curva *Receiver Operating Characteristic* (ROC) é um importante elemento visual na comparação de modelos. Esta curva mostra o balanço entre a taxa de verdadeiros positivos (TPR) e falsos positivos (FPR) [6]. Em suma esta curva mede os diversos pontos de desempenho de cada classificador.
- Área abaixo da curva: A AUC é uma estimativa da probabilidade de que um classificador irá classificar um registo positivo escolhido aleatoriamente, maior do que um exemplo negativo escolhido aleatoriamente [20]. Esta medida é muito utilizada na *comparação de* modelos.
- Micro-média e Macro-média: são medidas de desempenho avaliadas quando existem modelos multi-classificados (como é o caso dos modelos deste projeto). O mais comum é serem calculados a micro-média e macro-média para a métricas *precision* e o *recall*.

2.4.3 Apriori

O algoritmo utilizado para gerar regras de associação e realizar a recomendação de artigos da MASS Perfumarias neste projeto será o *Apriori*.

As regras geradas por este algoritmo permitem por exemplo:

- Organizar as prateleiras de supermercados segundo vendas conjuntas.
- Recomendar artigo com base na interação do cliente.
- Melhorar a experiência de utilização de páginas *web.*
- Deteção de situações fraudulentas.

O *Apriori* é um algoritmo muito utilizado no *Data Mining* para a geração de regras de associação, por ser simples e rápido nesta tarefa. Identifica os conjuntos frequentes e estende-os a conjuntos de itens maiores se estes conjuntos aparecerem vezes suficientes na sua tabela de transações. Utiliza uma abordagem *bottom up* e decompõe o problema em duas fases [21]:

- Encontrar os conjuntos de itens cujas ocorrências são iguais ou superiores a um dado suporte. Estes conjuntos de itens são chamados de conjuntos frequentes.
- Geração de regras de associação desses conjuntos frequentes cuja confiança seja igual ou superior ao indicado.

Em semelhança ao algoritmo *Apriori* existem outros algoritmos (descartando variâncias do *Apriori*) dos quais se destacam dois:

- AIS: Este foi o primeiro algoritmo desenvolvido que permitia encontrar grupos de itens num conjunto de transações. Os conjuntos candidatos são gerados à medida que as transações são lidas e, para cada transação, é determinado qual dos conjuntos de itens anteriores estão contidos na transação. Novos conjuntos candidatos são gerados por expandir estes conjuntos com outros itens nesta transação [22]. Possui um bom desempenho na sua performance mas gera mais conjuntos candidatos do que o necessário, o que em muitos casos pode resultar num *overflow* de memória [23].
- SETM: Este algoritmo teve como impulso ao seu desenvolvimento o uso de SQL em grandes conjuntos. Este tem um processamento inicial semelhante ao AIS na medida em que gera os conjuntos candidatos à medida que as transações são lidas. No entanto, a geração de novos conjuntos é feita com um *join* aos conjuntos existentes, guardando as transações numa estrutura sequencial organizada. O SETM apresenta as mesmas desvantagens que o AIS, além de que para cada conjunto candidato, existem tantas entradas na estrutura criada quanto o seu valor de suporte [23] [13].

A escolha do algoritmo Apriori baseia-se em alguns dos seus pontos fortes na sua utilização. Permite explorar dados usando diferentes níveis hierárquicos, elimina itens de pequena escala e suporta o uso de várias taxonomias [21]. Assim como o C5.0 este está também implementado e disponível em diversas ferramentas de *Data Mining*. Existem outros algoritmos desta natureza nos quais se destacam o *Eclat* e *FP-growth* [24]. O primeiro foi implementado por Parthasarathy Zaki e Ogihara e é uma abordagem em profundidade, fator que se torna um compromisso aquando da existência de um grande volume de transações a avaliar. O segundo, implementado por J. H. Han, permite encontrar conjuntos frequentes sem a geração de conjuntos candidatos. Este usa um processamento recursivo, que também se torna um compromisso na existência de elevadas transações.

2.4.4 Medidas de avaliação de regras de associação

Para avaliar um conjunto de regras de associação existem medidas adequadas a este efeito. As medidas mais utilizadas e maior relevância são:

- Suporte: Frequência com que os itens ocorrem relativamente à sua totalidade. Esta métrica é usada para filtrar regras não interessantes

$$sup(X \Rightarrow Y) = \frac{(X \cup Y)}{N}$$

- Confiança: Percentagem de casos em que o *LHS* prevê corretamente a ocorrência de *RHS*.

$$conf(X \Rightarrow Y) = \frac{\sup(X \cup Y)}{\sup(X)}$$

Outra métrica que é utilizada na avaliação da importância de uma regra é o *Lift*. Esta tem o mesmo significado que o interesse da regra e é a razão entre a confiança e a confiança esperada.

$$lift(X \Rightarrow Y) = \frac{\sup(X \cup Y)}{\sup(X) * \sup(Y)}$$

Este conjunto de métricas é transversal a qualquer algoritmo utilizado para a geração das regras. Em semelhança com a avaliação de modelos, etas métricas apresentam maior valor quando avaliadas em conjunto.

2.5 Exemplos de Aplicações

Neste projeto o uso de *Data Mining* terá como foco a recomendação de artigos, com o objetivo de permitir à MASS Perfumarias melhorar os alvos das suas campanhas e otimizar as suas estratégias de Marketing. *ERPs* e *Data Mining* são conceitos utilizados conjuntamente há alguns anos, o que leva a que já existam várias aplicações de *Data Mining* em ERPs em circunstâncias e com objetivos semelhantes aos deste projeto. Apresenta-se agora um conjunto destes contributos.

1. Customers Behavior Modeling by Semi-Supervised Learning in Customer Relationship Management [25]
    - Dados/Objetivos: Neste projeto são utilizados 2 conjuntos de dados. O 1º relaciona-se com informação bancária e contém 1,000 registos com 32 variáveis. O 2º está relacionado com informação de uma seguradora, contém 8,820 registos e 22 variáveis. O objetivo da aplicação de técnicas *Data Mining* no 1º conjunto é classificar cada registo como Bom ou Mau e no 2º conjunto é rotular cada registo com Sim ou Não relativamente à aprovação de crédito.
    - Métodos: É usada uma técnica de aprendizagem semi-supervisionada. O método semi-supervisionado proposto é uma rede neural *feed-forward* treinada por um algoritmo de propagação retroativa (*perceptron* multicamada), com o objetivo de prever a categoria de um cliente desconhecido ou de potenciais clientes. Este documento apresenta também a combinação entre uma rede neural com conjuntos de dados com apenas uma parte já classificada, de modo a aumentar a performance do classificador. É também apresentada uma alteração ao algoritmo original de redes neurais com o objetivo de aumentar a precisão do classificador para estes casos de estudo.
    - Implementação: *Rapid Miner* 5.1.
    - Resultados: Para o 1º conjunto de dados foram obtidos os seguintes valores: 74.67% acerto; 80.18% de TP; 60.24% de TN; 39.76% de FP; 18.82% de FN. Para o 2º conjunto de dados foram obtidos os seguintes valores: 80.19% acerto; 84.44% de TP; 48.80% de TN; 51.20% de FP; 15.56% de FN.

2. An Efficient CRM- Data Mining Framework for the Prediction of Customer Behavior [26]
    - Dados/Objetivos: O conjunto de dados é relativo a uma base de dados de marketing direto de uma aplicação de depósito de longo prazo. Estes dados estão relacionados com um banco português e fazem parte de uma amostra de um período entre maio de 2008 e novembro de 2010. O conjunto de dados contém 4,521 registos com 16 variáveis e o objetivo é obter um modelo de classificação dos clientes segundo: aderiu ou não aderiu à aplicação de depósito de longo prazo.
    - Métodos: Este documento propõe provar que um modelo de classificação de clientes utilizando redes neurais, leva a melhores percentagens de acerto do que um classificador Naive Bayes.
    - Implementação: WEKA.
    - Resultados: Modelo classificador de redes neurais mostra uma melhor precisão (88.63%) versos (87.97%) entre os dois modelos experimentados. Naive Bayes apresenta melhores valores de TVP (0.47%), TFP (0.067%) e área de ROC (0.858). O tempo necessário para construir o modelo é consideravelmente alto recorrendo a redes neurais (1767.75s) relativamente ao de Naive Bayes (0.08s).

3. Prediction of Churn Behavior of Bank Customers Using Data Mining Tools [27]
    - Dados/Objetivos: O conjunto de dados consiste em 1,474 registos, dos quais 1,163 são clientes ativos e 311 são clientes que abandonaram os serviços do banco. Este conjunto é também composto por 8 variáveis relacionadas com os clientes. O objetivo é criar um modelo de classificação capaz de prever o abandono de clientes ao banco.
    - Métodos: Este documento apresenta a previsão de abandono com base em ferramentas de *Data Mining* no setor bancário. Neste documento é feito um estudo sobre o comportamento de clientes de banco da India. As experiências realizadas baseiam-se em 2 algoritmos de classificação: CART e C5.0. As experiências demonstram que a taxa de sucesso de previsão de abandono recorrendo ao CART é bastante elevada e que o algoritmo C5.0 mostra resultados pobres nesta previsão, mas que no entanto, a taxa de sucesso da previsão da classe ativo pelo C5.0 é mais eficaz do que a outra técnica.
    - Implementação: Não especificada.
    - Resultados: A utilização do método CART resulta numa percentagem de acerto para ativo de 86.30% e para abandono uma percentagem de 91.22%. Este método gerou 12 regras válidas e verdadeiramente aplicáveis ao negócio. O método C5.0 apresenta uma taxa de acerto de 96.26% para ativos e 68.4% de abandono.

4. Research of the Bank's CRM Based on Data Mining Technology [28]
    - Dados/Objetivos: O conjunto de dados é composto por 53,872 registos de clientes, com 6 variáveis. Cada conjunto será classificado em 3 grupos tais como: alto, médio e baixo. Este conjunto é referente a uma agência Chongqing Branch Agricultural Bank of China. Esta implementação visa ajudar o banco a entender melhor o comportamento dos clientes que cumprem os requisitos de adesão a um grupo de elite a partir de um grande conjunto de dados.
    - Métodos: Este documento apresenta a construção de um modelo de classificação para analisar o tipo de clientes de um banco através da aplicação do algoritmo ID3.
    - Implementação: Microsoft Visual Studio tools of SQL Sever2005
    - Resultados: A percentagem de acerto desta árvore de decisão atinge 88%. Através desta árvore de decisão foram geradas algumas conclusões: A qualidade de crédito dos clientes desempenha um papel vital no desenvolvimento saudável de um banco. A primeira filtragem de cliente advém da renda mensal dos clientes. O rendimento mensal não é um fator absoluto, embora seja importante para um cliente tornar-se num cliente de alto valor. O fundo de educação e registo de crédito através de POS pode ser considerada como um fator também importante de referência.

5. Targeting customers via discovery knowledge for the insurance industry [29]
    - Dados: A fonte de dados neste documento é composta por 188,464 registos que representam clientes alvo de aplicações de seguros. Cada registo é composto por 7 variáveis. O objetivo é gerar um conjunto de regras que sejam usadas para encontrar potenciais clientes para um dado produto de seguro. Para gerar estas regras os autores recorreram ao uso do método ID3 para a construção de modelo classificador.
    - Objetivos/Métodos: O caso de estudo que é apresentado neste documento pretende explorar as regras de decisão para uma determinada empresa de seguros no mercado.
    - Implementação: Não especificada.
    - Resultados: Com um *threshold* de 61 foram obtidas 1,264 regras de decisão. Estas após um processo de seleção por parte da empresa alvo, resultaram num conjunto total de 12 regras com verdadeiro significado segundo a lógica de funcionamento da empresa.

6. Data Mining and ERP: An Application in Retail Sector [30]
    - Dados/Objetivos: A fonte de dados deste documento é um conjunto de transações feitas por clientes do maior retalhista de computadores e eletrónica da Turquia. Estes dados provêm do ERP MRPII e exploram o período de Maio de 2014. Esta amostra é composta por 4,875 registos com 85 campos cujos valores são 0 ou 1. Cerca de 65.23% da amostra são do sexo masculino e os restantes 23.86% do sexo feminino. Pretende-se que estas regras sejam úteis para fins tais como o *layout* da loja, ajuste da disposição de prateleira na loja,

etc., com o objetivo de aumentar a outras estratégias promocionais e de vendas.

- Métodos: Para executar esta tarefa o método utilizado para gerar as regras de associação foi o algoritmo *Apriori*.
- Implementação: *SPSS Clementine*.
- Resultados: Este documento apresenta ainda de forma resumida as conclusões determinadas pelo autor após uma análise às regras geradas: 71,094% dos clientes que compraram uma câmara, também compraram um cartão de memória (suporte de 5,251%); 23,577% dos clientes que compraram peças de computador, também compraram cabos de alimentação (suporte de 5,046%); 21,545% dos clientes que compraram peças de computadores, também compraram uma CPU. (suporte de 5,046%); 18,689% dos clientes que compraram um telemóvel, também compraram acessórios para este (suporte de 8,451%).

7. Discovering Interesting Association Rules in the Web Log Usage Data [31]
    - Dados/Objetivos: A amostra utilizada neste artigo é composta pelo log de informações relativas às solicitações ao *website* oficial de uma determinada instituição de educação. Esta é composta pelos *logs* de um dia selecionado arbitrariamente, sem consciência de quaisquer atividades especiais na instituição naquele dia. O objetivo deste trabalho passa por fornecer informação sobre a utilidade de regras de associação quando são aplicadas ao conjunto de dados desta instituição. Este trabalho pretende provar também que a descoberta de regras interessantes e potencialmente úteis de associação em dados de uso referentes a *websites*, é uma tarefa demasiado demorada com consequências negativas na performance/desempenho do *website*.
    - Métodos: Para construção das regras de associação foi utilizado o algoritmo *Apriori* baseado no trabalho descrito em Agrawal and Srikant (1994).
    - Implementação: WEKA.
    - Resultados: Este estudo resultou num total de 300 regras, apenas 19 das quais apresentam confiança igual a 1.0, enquanto 69 apresentam confiança maior do que 0,85. O resultado final apresentado é composto por 12 regras. Estas foram o resultado de uma filtragem com diversos critérios como por exemplo: páginas sem interesse elucidativo, clusters de páginas, páginas que fazem parte da *Home Page*, etc.

8. Mining Rare Association Rules from e-Learning Data [32]
    - Dados/Objetivos: Os dados utilizados fazem parte de uma amostra de 230 alunos em 5 cursos distintos no Moodle da Universidade de ciência da computação de Córdoba. Esta amostra é composta por 10 variáveis relativas ao aluno, cadeiras, *posts*, leituras e tempos de utilização. Todas estas variáveis foram descaracterizadas para 4 valores discretos. O objetivo é obter um conjunto de regras descobertas por um conjunto de algoritmos para ajudar os docentes a detetar discrepâncias de comportamento em relação aos estudantes numa plataforma como o Moodle.
    - Métodos: Neste artigo é relatada a extração de regras de associação sobre o uso por parte dos estudantes do sistema *Moodle*. Os autores mencionam que esse tipo de regra é difícil de encontrar com a aplicação de algoritmos de *Data Mining* convencionais. Neste projeto foram utilizados 4 algoritmos: *Apriori-Frequent* com *threshold* de 0.05; *Apriori-Infrequent* com suporte de 0.1; *Apriori-Inverse* com *threshold* de 0.1; *Apriori-Rare* com *threshold* de 0.1; A confiança definida para todos os algoritmos foi de 0.7.
    - Implementação: Não especificada.
    - Resultados: *Apriori-Frequent* gerou 788 regras com suporte e confiança médios de 0.162 e 0.717; *Apriori-Infrequent* gerou 388 regras com suporte e confiança médios de 0.058 e 0.863; *Apriori-Inverse* gerou 46 regras com suporte e confiança médios de 0.056 e 0.883; *Apriori-Rare* gerou 44 regras com suporte e confiança médios de 0.056 e 0.883. Concluiu-se ainda que alguns algoritmos específicos, como *Apriori-Inverse* e *Apriori-Rare*, são melhores a descobrir regras de associação raras do que outros algoritmos standart, como *Apriori-Frequent* e *Apriori-Infrequent*.

9. Association Rules in Data Mining: An Application on a Clothing and Accessory Specialty Store [33]
    - Dados/Objetivos: Na implementação do processo descrito, a fonte de dados utilizada pertence a uma loja de roupa e acessórios de moda que operam na província de Osmaniye, Turquia. Estes dados representam uma amostra referente a 2012. Este conjunto é composto por 42,390 transações de vendas, com 9,000 tipos diferentes de produtos subdivididos em 35 categorias. O estudo presente neste documento tem como objetivo gerar regras de associação, de modo a descobrir o que mais se vende conjuntamente nesta loja e assim possibilitar a esta melhorar as suas promoções e disposição dos seus conteúdos aos clientes.
    - Métodos: O método utilizado para a geração destas regras foi o *Apriori* segundo R. Agrawal e R. Srikant em 1994. Este método foi ajustado com um suporte mínimo de 0.05% e confiança de 50%.
    - Implementação: *SPSS Clementine*.

- Resultados: Este processo conclui 25,470 regras de associação, nas quais foram selecionadas apenas 8. Esta seleção baseia-se na escolha daquelas cujos valores de confiança e *lift* são significativamente maiores. As regras mais importantes são: um cliente que compra um cinto, fato e gravata compra também uma camisa, com 100% de probabilidade. Este grupo tem 4,24% de probabilidade de ser encontrado em conjunto, no âmbito das operações no conjunto total de dados.

10. Retail Market analysis in targeting sales based on Consumer Behaviour using Fuzzy Clustering – A Rule Based Model [34]
    - Dados/Objetivos: Os dados referidos neste documento são de uma cadeia de retalho composta por 10 lojas em diferentes zonas do país. Neste documento não são especificados detalhes sobre o real número de clientes e transações, é apenas indicado que para efeitos de ilustração serão somente utilizados 15 registos de transação com 9 variáveis. O objetivo deste estudo é auxiliar a empresa na identificação e compreensão das necessidades dos clientes, de acordo com as características demográficas e assim criar novos conjuntos de especificações e ofertas para os seus clientes.
    - Métodos: Esta tarefa será levada a cabo utilizando o método *Apriori*. Não foram também apresentados detalhes sobre os ajustes de confiança e suporte dados a este método.
    - Implementação: Não especificada.
    - Resultados: Com base nos dados apresentados resultaram 35 regras de associação, como por exemplo: Se Idade = 35 e Sexo = Masculino sugerido cerveja e cigarros com suporte de 0.4 e confiança de 0.6.

3 Avaliação de Soluções e Abordagem

3.1 Tecnologias

Após apresentados diversos detalhes relativos a este projeto, é notório que são necessárias diversas tecnologias que utilizadas em conjunto, proporcionam uma solução que permite dar resposta não só aos objetivos definidos como também aos requisitos impostos pela MASS Perfumarias. Estas tecnologias são agora descritas individualmente para que seja possível conhecê-las melhor, e são também contextualizadas no âmbito deste projeto para que se conheça o seu papel no desenrolar do funcionamento do sistema.

3.1.1 SQL Server

É utilizado o *SQL Server* na construção de toda a estrutura de armazenamento deste projeto. É também utilizado no processo de recolha, de alguns fatores de limpeza e transformação de dados. Esta tecnologia de base de dados é utilizada neste projeto por diversos motivos:

- É uma tecnologia com qual a MASS Perfumarias possui licenças de utilização e de servidores dedicados à sua utilização.
- Experiência de utilização.
- Maturidade e potencialidades desta ferramenta Microsoft.
- Boa performance em bases de dados não relacionais.

3.1.2 C#

A aplicação deste projeto é construída sobre a *framework Windows Form Aplication* (WFA) recorrendo-se à linguagem de programação C#. Com esta linguagem são elaboradas todas as interações com o utilizador e parcialmente a camada de gestão de dados. Os formulários são contruídos com a utilização de uma *framework* de apresentação denominada de *MetroFramework*. O uso da linguagem C# e da *framework* WFA devem-se aos seguintes fatores:

- Experiência de utilização.
- O foco deste projeto são análises *Data Mining* a serem elaboradas e a exploração dos dados da empresa, como tal, é importante que a ferramenta de desenvolvimento da interface da aplicação seja de simples e prática programação.
- Os utilizadores deste projeto trabalham diariamente com outras interfaces WFA.
- Existem algumas bibliotecas que permitem uma fácil interação entre o C# e a linguagem R.

A comunicação com a camada de inteligência é feita com o recurso a R.NET [35]. Esta tecnologia desenvolvida por investigadores da universidade de Carnegie Mellon, consiste numa ponte operativa entre tecnologias .NET e as bibliotecas de R. Existem três formas de comunicar entre C# e a linguagem R: chamadas ao sistema operativo, R COM ou R.NET. As chamadas ao sistema operativo para comunicar com o motor R apresentam algumas desvantagens tais como: implica permissões especiais, confirmações de caminhos e variáveis de ambiente, entre outras configurações que devem ser totalmente transparentes ao utilizador. A utilização de R COM já não requer configurações como as chamadas ao sistema operativo, o que apresenta ser uma melhor solução face à anterior. No entanto, quando comparado com R.NET esta apresenta algumas desvantagens: requer uma instalação extra no sistema operativo; a ligação ao motor R é mais simples de implementar; R.Net além de utilizar a sintaxe C# é de mais simples utilização, como demonstram a Figura 3 e Figura 4.

```
//R COM
StatConnector Sc1 = new StatConnectorSRVLib().StatConnectorClass();
Sc1.Init("&ldquo;R&rdquo;");
```

Figura 3 - Exemplo R COM

```
//R.NET
REngine.SetEnvironmentVariables();
engine = REngine.GetInstance();
engine.Initialize();
```

Figura 4 - Exemplo R.NET

3.1.3 R

Toda a camada de inteligência foi desenvolvida recorrendo à linguagem R. Optou-se pelo uso do R em relação a outras alternativas, maioritariamente por ser uma linguagem *open source* com uma comunidade muito dedicada e pelas potencialidades que oferece nos diferentes tipos de análises realizados neste projeto. Outra vantagem que se demonstrou decisiva é ser a linguagem que mais facilmente permite uma integração com outras. As perspetivas de utilização num futuro pós este projeto são maiores, assim como permite ter uma enorme flexibilidade na aplicação dos algoritmos necessários. O desenvolvimento em R elaborou-se em *RStudio* que consiste num *IDE* de desenvolvimento muito utilizado dedicado apenas a esta linguagem.

3.2 Abordagem

Esta secção tem como objetivo documentar de forma sucinta a abordagem ao problema, relatando as diferentes fases da construção deste projeto, de forma cronológica, para que seja possível compreender por que etapas e com que sequência este projeto se realiza desde o início até ao seu término. Este ponto reflete também o tipo de documentação de apoio que foi necessário ao seu desenvolvimento e em que circunstâncias foi aplicado.

A primeira fase deste projeto realiza-se com o primeiro contacto com o cliente, neste caso, com a MASS Perfumarias. Nesta fase são abordados diversos aspetos relativos ao tipo de informação a ser extraída, quais os resultados pretendidos com este projeto e qual a finalidade do conhecimento a ser extraído.

A segunda fase deste projeto consiste no contacto com a base de dados que contém toda a informação disponibilizada pela MASS Perfumarias. Esta base de dados armazena todos os artigos vendidos e já não disponíveis para venda, todo o histórico de clientes e dados relativos a vendas de 2011 a 2014. Devido à sua dimensão, foi necessário bastante tempo de análise à estrutura desta fonte de dados e apoio prestado pela Softingal tanto no esclarecimento de dúvidas como em apoio no primeiro contacto com esta.

Após a compreensão da estrutura de dados fornecida e numa quarta fase, procede-se a uma exploração de dados relativas a vendas, artigos e clientes. Esta fase tem uma grande importância na medida em que permite conhecer a informação com a qual este projeto se vai desenvolver, visto que estes dados terão um impacto máximo no conhecimento a ser extraído.

Numa quinta fase, após a análise realizada no ponto anterior, é possível uma definição mais concreta dos objetivos, isto é, numa definição exata do tipo de análises a serem realizadas. Esta definição mais precisa dos objetivos só é possível elaborar nesta fase, pelo facto de o tipo, a dimensão e a qualidade da informação presente na fonte de dados ser uma grande condicionante no possível conhecimento a ser extraído. Em suma, os dados disponíveis ditam se é possível ou não realizar as análises planeadas inicialmente.

A sexta fase deste projeto consiste na aprendizagem da linguagem de programação R. R é um conjunto integrado de *software* que permite a manipulação de dados, cálculos e representações gráficas. Esta linguagem além de *open source*, faculta uma grande coleção de ferramentas dedicadas a análises de dados [36]. Após uma análise de algumas alternativas existentes e face a estas vantagens apresentadas, R foi um das linguagens escolhidas para este projeto. Sendo R uma matéria desconhecida tornou-se fundamental um estudo intensivo sobre esta linguagem. A documentação de apoio neste processo de aprendizagem de R: documentação oficial da linguagem; tutorias; *papers*; cursos *online* das conhecidas escolas *Udacity* [37] e *Coursera* [38].

Após consolidadas aptidões de R e com os objetivos já devidamente definidos, numa sétima fase, é necessário aprofundar e rever alguns conceitos relativos a *Data Mining*. Para este efeito recorre-se ao material letivo disponibilizado durante as aulas do mestrado, alguns *papers* e

assim como a alguns livros da área. Toda esta documentação consultada estará disponível em detalhe na secção de referências deste documento.

Com todas as condições reunidas que possibilitam o início do desenvolvimento deste projeto, inicia-se uma oitava fase que consiste na estruturação deste. Nesta fase é concebida uma solução capaz de responder aos requisitos necessários impostos pelo cliente, assim como capaz de responder aos objetivos estipulados. Nesta fase é desenhada uma solução que permite ao cliente ter acesso às análises objetivadas neste projeto, sempre que surja essa necessidade e cujos resultados sejam apresentados ao utilizador para que este consiga facilmente interpretá-los. A solução permite também que todo o processo de recolha e transformação da informação seja configurável e transparente ao utilizador.

Numa nona fase, inicia-se a implementação da solução e das análises em questão. Aqui é construída toda a solução idealizada na fase anterior, desde a recolha, tratamento, leitura, e processamento da informação; a construção da aplicação; de toda a estrutura R de apoio às análises; de toda a interface de exposição de resultados.

Na décima e última fase, após concluída toda a solução encontram-se reunidas as condições ideais para uma análise geral de resultados. Nesta fase será feita uma apreciação relativa a cada uma das análises efetuadas, assim como uma apreciação global de resultados obtidos com os dados disponibilizados pela MASS Perfumarias.

4 Projeto

4.1 Design

Com base nestes requisitos, idealizou-se uma aplicação com uma interface de utilização simples num ambiente usual aos utilizadores. Esta aplicação contará com um conjunto de ecrãs disponíveis, em que permitirá ao utilizador selecionar qual a informação a que deseja ter acesso, ao qual a aplicação deverá apresentar de forma organizada as grandes análises efetuadas, simplificando todo o processo e tornando-o mais compreensível ao utilizador.

Esta aplicação encontra-se dividida em três camadas:

- Interface: Todas interações que os utilizadores façam com o sistema serão realizadas nesta camada. Esta será responsável por apresentar aos utilizadores todos os ecrãs que estes solicitem, assim como levar a cabo a tarefa de expor as informações pretendidas de forma intuitiva e compreensível. Esta terá também a tarefa de comunicar com o sistema de ETL. Terá também que comunicar com o sistema R e solicitar que este processe as análises solicitadas pelo utilizador.

- Gestão de dados: Uma das três partes será a responsável por todo o processo de ETL, e como tal será denominada de gestão de dados. Esta será responsável pela ligação e recolha ao sistema *mySoftmais*, por todo o processo de limpeza e transformação e por fim terá também de persistir a informação tratada na *Staging Area* deste projeto. Em suma, a gestão de dados será responsável por todo o fluxo descrito nos capítulos 5.3 Staging Area e 5.4 Processo ETL.

- Inteligência: Haverá uma terceira e mais importante camada que será responsável por todo o processamento de *Data Mining*. Esta parte será um conjunto de ficheiros escritos em R que neles contêm todas as implementações necessárias para que sejam aplicados os algoritmos construídos que irão produzir os resultados pretendidos, sobre os dados presentes na *Staging Area*. Em suma será nesta camada que terão lugar as tarefas de classificação de clientes, geração de regras e recomendação de artigos.

4.2 Arquitetura

A Figura 5 ilustra o diagrama de componentes sobre o sistema e as suas partes constituíntes que foram agora descritas. Esta figura permite compreender também como as três grandes partes deste projeto se relacionam.

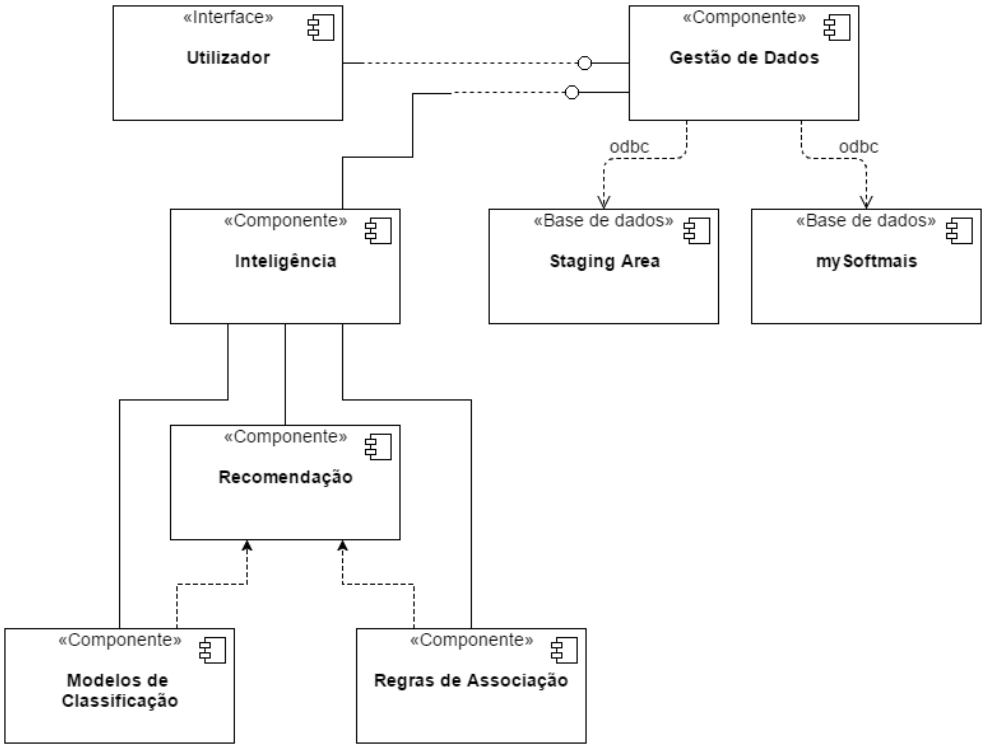


Figura 5 – Diagrama de Componentes

As ações são despoletadas pelo utilizador diretamente na camada de *interface*. Esta camada ao recolher o pedido feito pelo utilizador, redireciona-a para a camada de gestão de dados e aguarda pela resposta, a qual a apresentará ao utilizador.

Por sua vez a camada de gestão de dados transpõe esse pedido para a camada de inteligência. No entanto, precisa primeiramente de se assegurar que já existem na *Staging Area* todos os dados que a camada de inteligência necessita para realizar as operações solicitadas. Esta irá consultar a *Staging Area* e verificar se as informações solicitadas se encontram disponíveis. Caso estas não se encontrem disponíveis, esta camada irá ligar-se à base de dados do *mySoftmais* extrair, transformar e carregar toda a informação na *Staging Area.*

A camada de *interface* após se assegurar de que já se encontram reunidas as condições necessárias para responder ao pedido do utilizador, irá solicitar à camada de inteligência que efetue os processamentos necessários ao pedido do utilizador.

A camada de inteligência por sua vez, irá recolher da *Staging Area* todos os dados que entende convenientes e aplicá-los nos algoritmos responsáveis pelo processamento dos resultados solicitados. Após o processamento destes algoritmos, os resultados serão enviados para a camada de *interface*, que se encarregará de apresentar ao utilizador os resultados obtidos de uma forma amigável e de fácil compreensão.

É possível observar que a camada de inteligência é composta por 3 subpartes que correspondem diretamente aos principais objetivos deste projeto. Como o nome indica, a subparte modelos de classificação é responsável por gerar os modelos preditivos de classificação de clientes fidelidade. Geração de regras de associação será a componente, que com base nas transações inseridas na *Staging Area* irá criar as regras de associação necessárias.

Os resultados gerados por estas duas subpartes serão condicionantes para a terceira subparte que tem como objetivo a recomendação de artigos a clientes. Para que o consiga elaborar, necessita dos grupos de utilizadores gerados pela classificação de clientes e das regras obtidas pela geração de regras de associação.

5 Dados

Neste capítulo serão apresentados detalhes relativos aos dados que servirão de suporte para todas as análises que serão efetuadas neste projeto. Dá-se especial relevância a este capítulo pelo facto de que a qualidade da informação dos dados extraídos irão ditar a qualidade das análises feitas. O adágio "*garbage in, garbage out*" não é dado ao acaso. Os dados devem ser tratados como um ativo chave e estratégico para garantir que a sua qualidade seja imperativa. Se a informação a ser analisada estiver desatualizada, inconsistente, incompleta ou desintegrada, é natural que os resultados obtidos nas análises não terão grande valor [39]. É importante referir também que as observações feitas neste capítulo, estão condicionadas pela dimensão da amostra fornecida pela Mass Perfumarias e terão como foco apenas artigos, clientes e vendas. Esta amostra representa os dados da empresa no período de 2011 a 2014.

5.1 Análise

As análises a serem efetuadas neste projeto baseiam-se em três grandes partes: clientes, produtos e vendas. Sobre cada uma existem alguns conceitos importantes que serão agora descritos.

Existem dois tipos de entidades: consumidores individuais ou empresas. Dentro dos consumidores individuais existe ainda uma separação lógica destes em três tipos distintos:

- Aqueles que aderem ao cartão fidelidade e que são denominados de clientes fidelidade. Neste caso o cliente é identificado pelo número do seu cartão fidelidade. Este cartão existe com o objetivo estratégico de fidelização de clientes. Este permite a acumulação de pontos de compras que resultam em vales quando atingido determinado número de pontos. O cartão também beneficiará o cliente em épocas especiais de desconto e benefícios em algumas ofertas, como ofertas de aniversário, entre outros.
- Aqueles que não se identificam no ato da venda e são denominados de clientes finais ou de consumidores finais.
- Aqueles que não pretendem aderir ao cartão fidelidade mas que se identificam no ato da venda, com o seu nome e número de contribuinte. Esta identificação é fundamental para obter um histórico de compras deste tipo de clientes.

Relativamente a produtos, a empresa disponibiliza para venda sensivelmente 15,000 artigos. A natureza dos bens em questão são a cosmética, perfumaria e produtos de beleza. Cada um destes artigos está relacionado com diferentes tabelas de preço a aplicar em diferentes ocasiões e cada um caraterizado em diversas categorias tais como: Família, Marca, Linha, Secção, etc. Além dos atributos mais comuns de caracterização como: nome, código, etc. cada artigo é ainda caracterizado por quantidades por caixa e por tipo de embalagem.

As vendas de artigos a clientes são representadas no sistema como documentos. É importante para este projeto entender o que é um documento e como está estruturado. Um documento é uma estrutura de dados que representa os movimentos, vendas, compras, transferências e as diversas ocasiões em que artigos ou dinheiro estão envolvidos. É com base em documentos que o programa se baseia para grande parte da gestão do negócio. É também sobre o documento que o *software* assegura que todas as exigências legais impostas pela legislação portuguesa em vigor sejam cumpridas. Quando neste trabalho existe uma referência vendas, o que é pretendido referenciar são documentos relacionados com as vendas.

Um documento está dividido em três partes principais:

- Cabeçalho - No cabeçalho são armazenadas informações essenciais para identificar o documento, as suas entidades, datas, entre muitas outras configurações e informações tanto relacionadas com legislação, como relacionadas com gestão interna da aplicação.
- Linhas - São aqui armazenados os diferentes produtos tratados em quaisquer documento. É armazenado um artigo para cada linha, junto com várias informações, tais como: quantidades, embalagem, taxas, preços de venda, descontos, etc.
- Totais - Aqui são guardados os valores totais do documento, onde são armazenados os diferentes valores para diferentes taxas de IVA em vigor, descontos financeiros, descontos globais, etc.

É também importante fazer referência a como um documento é identificado e como garantir que não há documentos repetidos. Seguindo um procedimento de identificação legal de documentos, esta é composta por uma série, um código e um número. Estes são os campos que nos permitem conhecer a sua natureza (vendas, compras, movimentação de *stock*, etc.) e onde são originadas (neste caso, em que lojas foram produzidos).

5.2 Fonte de dados

A fonte de informação de todo o processo de gestão da empresa é feita por um sistema ERP que a empresa utiliza para gerir todas as suas atividades, processos e fluxo de trabalho. Este ERP apresenta uma complexidade bastante significativa em diversos aspetos. Este é um programa com mais de 15 anos de existência, robusto e adaptável a diversas áreas de negócio.

Para que o programa seja capaz de responder a toda esta complexidade, requer uma estrutura de dados de igual complexidade. Esta estrutura contém todos os dados de origem para todas as análises que serão feitas.

Esta fonte de dados em questão, é uma base de dados com mais de 300 tabelas e mais de 10GB de informação (face ao período compreendido entre 2011 e 2014). Esta base de dados é tão complexa quanto a variância de fluxos e as características de vários sectores de atividade existentes. Algumas das tabelas mais complexas podem ter mais de 100 campos com numerosas relações com outras tabelas.

Devido à já descrita complexidade da base de dados deste sistema ERP, é necessário investir tempo a conhecer a forma como esta se encontra estruturada, de modo a possibilitar o correto enquadramento da informação que se pretende extrair, com o menor número de erros na interpretação da estrutura das tabelas.

Por conseguinte, existe uma necessidade de conhecer bem a forma como a informação é dividida pelas tabelas da base de dados, como elas se relacionam e quais aquelas que são responsáveis pelo armazenamento dos dados necessários para as análises.

Com a finalidade de explorar a base de dados e interiorizar o seu esquema, foi feito um levantamento de todas as tabelas que serão necessárias para explorar. Estas tabelas representam todas as informações que efetivamente contêm os dados necessários para permitir realizar as análises pretendidas.

De seguida e em mais detalhe foram levantados todos os campos de cada umas das tabelas que continham informação relevante aos casos de estudo. Para realizar esta tarefa, foram feitas diversas folhas de *Excel*. Cada folha contém todos os detalhes sobre a tabela, por exemplo: tipos de dados, o que cada campo representa, o seu objetivo, se vai ser utilizado ou não, quantos registos possuem dados válidos, entre outros fatores.

As principais tabelas de *Excel* seguem em anexo (Anexos II, III e IV) a este documento, para apresentar em detalhe todo o levantamento à base de dados do *mySoftmais*. Neste ponto, a ajuda da *Softingal* demonstrou-se um fator importante tanto no ritmo de avanço nesta fase de análise, assim como na qualidade do levantamento feito. Foi de facto necessária a ajuda da *Softingal* nos seguintes casos:

- Interpretação da nomenclatura de campos e tabelas, que devido às suas abreviações ocultavam o seu significado.
- Os campos apresentavam nomes perfeitamente claros, mas o seu conteúdo estava sujeito a interpretações erradas do seu verdadeiro significado.
- Existiam também casos de campos cuja utilização por parte dos operadores era incorreta e/ou incoerente.

Para cada campo que apresentasse um significado que correspondesse à informação necessária às análises em questão, foi feita uma verificação que consistia em obter o número de registos preenchidos e o número de registos válidos. Foram apenas levantados os campos com elevado número de registos preenchidos e baixo número de registos inválidos face ao número total de registos da tabela. Esta análise efetuada à base de dados do *mySoftmais* poderá ser analisada em maior detalhe com as tabelas dos Anexos II, III e IV.

5.3 Staging Area

Com os detalhes relatados nas subsecções anteriores sobre a fonte de dados deste projeto, podemos sumariar que esta é uma base de dados do ERP *mySoftmais*, que se caracteriza por ser extensa e complexa. Com a necessidade de ler a informação pretendida neste repositório de dados, é imperativo que se analise primeiramente soluções para efetuar esta leitura. É necessário que se analisem alternativas ao modo de leitura a esta base de dados pelos seguintes motivos:

- A base de dados do *mySoftmais* pode nem sempre estar acessível aquando da necessidade da utilização da aplicação.
- Pretende-se que seja possível utilizar esta aplicação sem recurso a uma ligação à internet ou uma ligação em rede ao servidor que contém a base de dados.
- As análises poderão ter caracter interativo e a base de dados de trabalho em questão pode não conter toda a informação. Este ponto é frequente acontecer nesta empresa, visto que as lojas trabalham apenas com uma base de dados semelhante á fornecida neste projeto, ou seja, uma base de dados que represente apenas os últimos 4 anos.
- A Informação encontra-se muito normalizada, o que dificulta a rapidez no acesso aos dados necessários em diferentes ocasiões.

Face a estes problemas optou-se pela utilização de uma *Staging Area.*

Uma S*taging Area* é uma área de armazenamento intermédia e temporária onde são guardados dados provenientes de um ou mais repositórios de dados [40]. Geralmente situa-se diretamente entre as fontes de dados de um sistema e os seus sistemas alvo. A sua utilização é muito comum como ferramenta de apoio a processos de extração, tratamento e carregamento (ETL), que serão abordados posteriormente neste documento [41].

O uso de uma *Staging Area* é facilmente justificado pelas diversas vantagens que oferece, face aos problemas acima mencionados e aos custos que implicam a construção de uma estrutura deste tipo. Por conseguinte, podemos afirmar que neste caso, o uso de uma *Staging Area* traduz-se nas seguintes vantagens [42]:

- Permite guardar apenas a informação necessária face ao vasto volume das fontes originais de informação.
- Permite criar "*snapshots*" de determinada informação armazenada.
- É um grande auxílio no processo das diferentes fases do processamento ETL, porque permite persistir a informação transformada entre cada fase.
- Permite descentralizar o processamento das análises e a fonte de informação original.
- Permite adicionar informação temporária e outros tipos de informação auxiliar ao processamento das análises.

Compreender bem a informação que será necessária para abordar todas as análises propostas e conhecer bem a estrutura da base de dados que contém todos os dados disponíveis, são fatores chave para que se possa dar início à construção da *Staging Area*. Esta construção consiste em desenvolver uma estrutura de dados que consegue armazenar toda a informação necessária, devidamente formatada e padronizada.

A estrutura da *Staging Area* foi desenhada de modo a ser possível armazenar toda a informação necessária para apoiar as análises a serem executadas pela aplicação, e como tal, precisa de albergar os dados recolhidos e tratados relativos a clientes, produtos e vendas. Esta será uma base de dados construída em *SQL Server* não normalizada, de localização configurável, composta por três tabelas. Existe uma tabela que será responsável pelo armazenamento de dados de identificação, caracterização, localização e comportamento relativos a cliente fidelidade.

Existe também uma tabela que guardará toda a informação relativa aos artigos comercializados pela empresa. Esta terá campos relativos a identificação, descrição, caracterização e valores estatísticos de preços acumulados até ao fim do período da amostra fornecida. Nesta base de dados poderá também ser encontrada uma tabela de maior dimensão que será encarregue de armazenar informações relativas a vendas de clientes fidelidade. Cada linha desta tabela corresponde a uma única linha de uma determinada venda realizada em qualquer uma das lojas da empresa (seja física ou *online*).

Esta identifica cada venda como um documento, com código, série e ano e número, qual o cliente que a originou e qual o artigo que foi comprado. São também armazenados dados sobre descontos, valores de tabela, valores de compra, quantidades, totais, índices, entre outros valores.

A estrutura criada para armazenamento na *Staging Area* é composta pelas seguintes tabelas.

Tabela 2 – *Staging Area*: Clientes Fidelidade

Clientes Fidelidade	
Atributo	**Valor**
Número Contribuinte	Texto
Data Registo	Data
Ok	Texto
Código Postal	Texto
Abreviatura Código Postal	Texto
Cidade	Texto
Idade	Inteiro
Fidelidade	Texto
Valor de Pontos	Decimal
Pontos	Decimal
Total de Pontos	Decimal
Total de itens Comprados	Inteiro
Número de Compras	Inteiro
Gasto Total	Decimal
Valor Médio de Compra	Decimal
Itens Médios por Compra	Decimal
Frequência à loja	Decimal
Classe	Texto
Data Última Compra	Data
Número Dias sem Comprar	Inteiro
Compras nas Lojas	Texto
Loja Habitual	Texto
Número de Vouchers	Inteiro
Vouchers por Ano	Decimal
Anos de Registo	Inteiro
Compras de Acessórios	Inteiro
Compras de Cosmética	Inteiro
Compras de Beleza	Inteiro
Compras de Perfumaria	Inteiro
Compras de Toucador	Inteiro
Zona Geográfica	Texto

Tabela 3 – *Staging Area*: Produtos

Produtos	
Atributo	**Valor**
Código	Texto
Ok	Texto
Data Registo	Data
Quantidade por Caixa	Inteiro
PMC	Decimal
PMD	Decimal
UPC	Decimal
INDPMC	Decimal
INDPMD	Decimal

Tabela 4 – *Staging Area*: Valores Auxiliares

Valores Auxiliares	
Atributo	**Valor**
Ano Início	Inteiro
Ano Fim	Inteiro
Data Início	Data
Data Fim	Data
Meses de inatividade	Inteiro
Data de Referencia	Data

Tabela 5 – *Staging Area*: Vendas de Clientes Fidelidade

Vendas de Clientes Fidelidade	
Atributo	**Valor**
Entidade	Texto
Data	Data
Código de Documento	Texto
Série	Texto
Ano	Inteiro
Número	Inteiro
Linha	Inteiro
Código de Produto	Texto
Descrição	Texto
Marca	Texto
Tipo	Texto
Departamento	Texto
Fornecedor	Texto
IVA	Texto
Unidade	Texto
Fator de Unidade	Inteiro
Quantidade	Decimal
Tabela de Preço	Texto
Preço Unitário de Tabela	Decimal
Desconto de Tabela	Decimal
Preço Unitário de Venda	Decimal
Desconto 1	Decimal
Desconto 2	Decimal
PMC	Decimal
PMD	Decimal
UPC	Decimal
Embalagem	Inteiro
Total	Decimal

5.4 Processo ETL

Como já mencionado neste documento, os dados de origem deste projeto provêm de uma complexa base de dados de um sistema *ERP* chamado *mySoftmais*. Foi também mencionado no capítulo anterior que será utilizada uma *Staging Area* para armazenar os dados extraídos da base de dados do mySoftmais devidamente selecionados, limpos e padronizados. Assim, o processo de extração referido neste capítulo, será feito a partir da base de dados do *mySoftmais* para a área de armazenamento da *Staging Area* deste projeto.

Contudo, a movimentação de informação entre a base de dados do *mySoftmais* e a *Staging Area* apresenta diversos problemas que impedem que esta transição seja feita de forma linear:

- As estruturas são bastante distintas.
- Existem valores em falta.
- Existem valores duplicados.
- Os campos recolhidos são em grande parte tratados por utilizadores, o que leva a que os valores inseridos sejam muito díspares entre si.
- Existem diversas inconsistências de informação.

Face a estes problemas é imperativo que se recorra a um processo de *ETL* devidamente estruturado e adaptado às questões e estruturas deste projeto.

Um processo de *ETL – Extraction Transform Load*, como o nome indica é um processo composto por 3 fases que consistem na extração de informação de uma ou diversas fontes, no seu tratamento (de inconsistências, faltas de integridade, etc.) e no seu carregamento cujo destino será repositório de informação (regra geral, este comunica diretamente com os sistemas alvo):

- Extração: serão feitas recolhas às diversas tabelas do *mySoftmais*, onde será selecionada apenas a informação necessária.
- Transformação: após a extração, serão feitas as seguintes transformações aos dados recolhidos:
    - Padronização
    - Formatação
    - Codificação
    - Derivações e cálculos
- Carregamento: serão carregados para a *Staging Area* os dados resultantes das transformações aplicadas no passo anterior. Após esta fase, a informação estará pronta a ser utilizada pelo sistema alvo.

Durante o processo de ETL será necessário realizar complexas recolhas de dados, transformações e padronizações aos dados fonte. Para tal, existem várias abordagens a este processo e, não é possível afirmar que existe uma melhor abordagem para realizar os diferentes procedimentos neste processo. Para este projeto foram tidas em conta duas possíveis alternativas.

- Recorrer à ferramenta SSIS - *SQL Server Integration Services* da *Microsoft* [43].
- Recorrer ao *SQL* para a recolha, armazenamento e tratamento de informação e a uma linguagem de programação genérica para realizar toda a gestão envolvente ao processo.

A primeira opção é apresentada como uma ótima alternativa. Esta ferramenta (não é só mas também destinada para tais problemas) apresenta soluções de extração, carregamento e de mapeamento de campos, com ótimos níveis de desempenho. No entanto, devido à necessidade de dotar a empresa de uma solução que terá de ser tão autónoma quanto possível em todos os níveis, requer que se utilize o menor número de configurações possíveis ao seu funcionamento, facto que condiciona a utilização desta ferramenta. A segunda alternativa seria uma solução que permitisse a utilização de SQL para a extração, armazenamento e transformação dos dados necessários e usar uma linguagem de programação genérica de operações tais como: saber que informação parametrizar e recolher, gerir as localizações das bases de dados, agendar processos, etc.

Neste caso, poderia ser uma parte integrante de toda aplicação e assim facilitar todo o âmbito, desde recolher a dados na base dados do *mySoftmais* até à apresentação das análises ao utilizador. Esta alternativa apresenta ainda o inconveniente de ser bastante trabalhosa, na medida em que todas as implementações terão de ser realizadas na íntegra.

Face às vantagens e desvantagens de cada alternativa será utilizada neste projeto a segunda opção. Em suma, os mecanismos e rotinas de ETL e a gestão do armazenamento da *Staging Area* serão então da responsabilidade da aplicação. Esta irá consultar uma instância SQL (o que contém os dados originais) para que lhe sejam fornecidos todos os dados devidamente tratados, de forma adequada. Depois desta consulta irá inserir esses dados nas respetivas tabelas da *Staging Area*.

Todo este fluxo encontra-se ilustrado na Figura 6, onde é possível observar os subprocessos do processo ETL, a ligação entre a extração e base de dados do *mySoftmais*, o processo de transformação da informação extraída (recorrendo-se a SQL e uma linguagem de programação genérica) e ao processo de carregamento da informação transformada para a *Staging Area*.

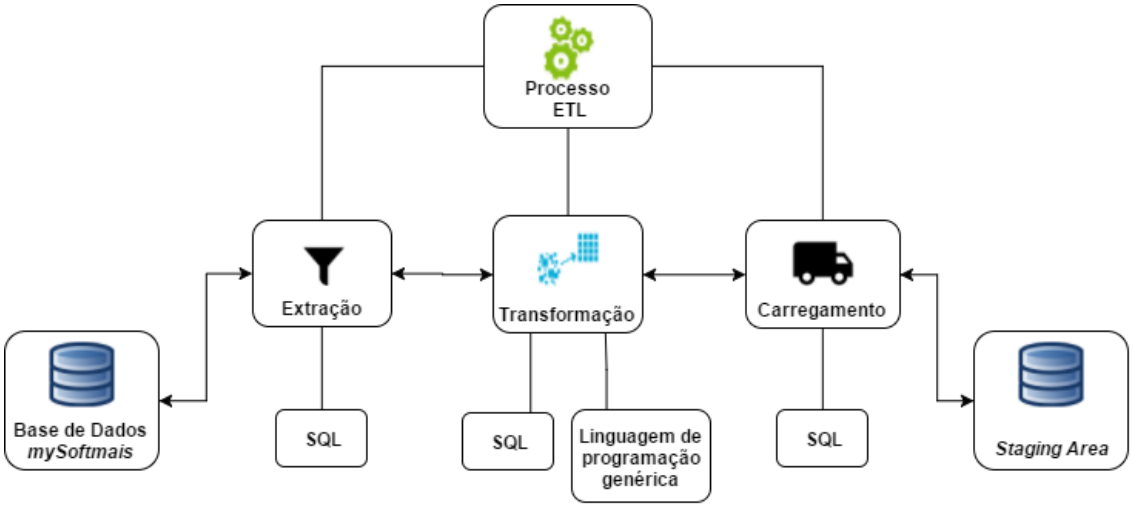


Figura 6 - Esquema do Processo ETL

O processo de ETL será agora descrito de forma pormenorizada. Este processo é dividido em três partes principais: clientes, produtos e vendas.

5.4.1 Clientes

Como referido anteriormente, existem três tipos de clientes aos quais apenas um tem a relevância necessária para ser analisado: os clientes que aderem ao cartão de fidelidade (de agora em diante nomeados clientes fidelidade). Estes clientes têm o valor necessário pelas seguintes razões:

- Estão devidamente identificados, o que permite distingui-los individualmente.
- Realizam compras nas lojas com relativa regularidade.
- Aderem a promoções e acumulam pontos de fidelidade que se refletem em vales.

Todas as informações necessárias relativas a clientes fidelidade estão numa única tabela no *mySoftmais*, denominada de T49ENTVD_GLB. A partir desta tabela, foram recolhidas as seguintes informações:

- Número de Contribuinte: os clientes fidelidade são identificados exclusivamente pelo seu número de contribuinte. Este campo permite assegurar que cada cliente é único no sistema. Aos clientes que não pretendem fornecer o seu número de contribuinte, é atribuído a este campo o número do cartão fidelidade, precedido da sigla SNC (sem número de contribuinte).
- Código postal e localidade: estes dois campos permitem formar grupos de entidades pela sua localização geográfica. Estes apresentam uma elevada importância na preparação dos dados para as análises em questão.

- Zona: Este atributo será deduzido pelo código postal com a finalidade de atribuir a entidade a uma de quatro zonas. Estas zonas são divisões geográficas criadas com base na presença das lojas físicas da MASS Perfumarias.
- Primeira data de compra: esta data permite saber quando foi a primeira vez que o cliente fez uma compra na loja e se identificou. Coincide também com a data em que o cliente foi registado no sistema.
- Idade: este campo apresenta relevância porque permite associar clientes a faixas etárias.
- Valor de pontos e número de pontos: estes campos indicam quantos pontos e qual valor destes que o cliente tem no seu cartão. O número de pontos e o seu valor terão grande influência na medida em que são convertidos em vales, que por si só são um fator de fidelização de clientes.
- Número de compras, número de itens comprados e respetivos valores médios: este campo indica o número de compras efetuadas em qualquer uma das lojas (que são também consideradas idas à loja). O número de itens comprados foi também recolhido por ser um valor importante na caracterização de um cliente, assim como na avaliação do seu comportamento. Este número aliado ao número de compras efetuadas permite saber, por exemplo, se é um cliente que vai poucas vezes à loja mas em cada ida compra muitos produtos ou se visita a loja com frequência mas adquire poucos produtos.
- Valor total gasto e valor médio gasto por compra: estes valores, em semelhança ao número de compras e de itens comprados, possuem um grande valor de caracterização e avaliação de clientes. Como consequência estes terão uma grande importância para as análises a serem executadas.
- Rácio de frequência à loja: este rácio é o produto entre o número de compras efetuadas (que corresponde a visitas à loja) e o número de anos do período referente à amostra.
- Caracterização: com base no modelo de pirâmide do *Application Template for Customer Relationship Management* escrito pela SPSS [44], foram definidas 5 caracterizações possíveis de atribuir aos clientes (inativo, pequeno, médio, grande, topo). Este critério teve como base de distribuição o valor total despendido nas lojas.
- Data da última compra, número de vouchers e respetiva média anual, assim como o valor total gasto e o valor gasto por compra, apresentam uma grande importância nos casos de estudo pela sua influência na caracterização e avaliação de clientes.
- Loja habitual: Este campo indica se um dado cliente visita sempre a mesma loja ou se distribui a suas compras por mais do que uma loja. Este campo será recolhido pelos mesmos motivos que os valores no ponto anterior.
- Número de cada tipo de artigo adquirido: serão recolhidos o número de itens adquiridos separados pelos seguintes tipos: acessórios, cosmética, beleza, perfumaria, toucador e sem classificação. Estes atributos terão um papel fulcral na caracterização de clientes na medida que permite, por exemplo, saber se é um cliente que compra apenas artigos de determinada categoria. Este exemplo quando relacionado com outros atributos como: idade, localização, última visita à loja, etc. proporciona a este atributo um papel fundamental nas análises a serem estudadas.

Existem 17,756 clientes fidelidade registados no sistema, o que demonstra ser uma generosa amostra de clientes. No entanto, nem todos esses clientes possuem preenchidos todos os dados relativos aos seus detalhes (informação geográfica, idade, primeira data de compra, etc.). Dado o volume inicial da amostra de clientes, foi decidido recolher apenas aqueles cujos campos de registos estivessem maioritariamente preenchidos. As restrições feitas foram as seguintes:

- Número de Contribuinte: Não poderia ser vazio ou nulo e não igual ao número de consumidor final (999999990). Existem clientes que não pretendem apresentar o seu número de contribuinte aquando do seu registo, o que implica o uso de outro número, que também é único. A solução que a *Softingal* encontrou para este problema foi recorrer ao número do cartão de fidelidade precedido da sigla "SNC" (sem número de contribuinte).
- Código postal: não poderia ser vazio ou nulo e teria de corresponder ao formato de código postal Português. Esta restrição foi aplicada pelo facto de na amostra fornecida, não existirem quaisquer registos de moradas estrangeiras em fichas de clientes fidelidade. Por conseguinte, todos os registos cujo código postal não correspondessem ao formato português, invalidavam o cliente.
- Localidade: não poderia ser vazio ou nulo e o seu comprimento deve ser superior a 4 caracteres.
- Primeira data de compra: não poderia ser vazio ou nulo e o seu valor teria de ser uma data compreendida entre o ano atual e 1995.
- Idade - não poderia ser vazio ou nulo e o seu valor teria de estar compreendido entre 15 e 100.

Durante este processo foram detetadas algumas inconsistências de informação relativas a clientes. Um exemplo das inconsistências encontradas foi a detecção da existência dos mesmos clientes em ambas as tabelas referentes a clientes (fidelidade e empresas). Foram encontrados clientes registados com o mesmo número de contribuinte considerados empresas e considerados também clientes fidelidade. Não serão aproveitadas estas entidades porque os mesmos números de contribuinte existem em ambas as tabelas. Como este número é o único elemento que permite identificar unicamente um cliente fidelidade, estes casos não serão considerados.

Após filtrados todos os casos que apresentavam inconsistências ou que não respeitavam as restrições impostas, a amostra passou para 17,262 clientes que transitarão para a *Staging Area*.

5.4.2 Produtos

Neste processo de ETL também foram extraídos a partir da base de dados *mySoftmais* os artigos que a empresa comercializa para ambos os tipos de clientes. A tipologia dos bens em questão são a cosmética, perfumaria e produtos de beleza. Para gerir produtos, o *mySoftmais* recorre a diversas tabelas das quais 4 foram utilizadas para extração:

- Tabela de produto: nesta tabela estão armazenadas informações de identificação e descrição, especificações para cada tipo de artigo e também variáveis de controlo interno sobre os produtos.
- Tabela de unidades: aqui estão contidas todas as unidades de cada artigo armazenado. Por exemplo: quilos, caixas, embalagens, etc.
- Tabela de embalagem: esta tabela identifica e lista cada artigo para seus tipos de embalagens. Por exemplo: grande, pequeno, médio, etc.
- Tabela de acumulados: A função desta tabela é manter registos de valores mensais relativos às quantidades de *stock*, preços e custos. Esta tabela existe com o propósito de acelerar o processo de geração de diversos relatórios relativos a inventários, preços médios de compra e venda, margens, entre outros.

Tendo em consideração as necessidades das análises a serem efetuadas e as informações disponíveis relativas a produtos, foram recolhidos os seguintes dados:

- Código de artigo: este campo representa o código do produto. Este código identifica exclusivamente o produto no sistema.
- Artigo estado: este campo indica se o produto está disponível para venda.
- Data da inserção no sistema: esta data indica o dia em que o produto foi registado no sistema pela primeira vez o que permite saber qual a primeira época deste artigo.
- Data da última venda e data da última compra: estas datas são fundamentais conjuntamente, para calcular qual a taxa de rotação de cada artigo.
- Unidade, embalagem e respetivos fatores de conversão: como já mencionado neste documento os produtos são de uma natureza que são frequentemente vendidos em embalagens de diferentes unidades. Estes campos indicam que o artigo é vendido a determinadas unidades numa embalagem. O que demonstra ser uma característica importante na caracterização do artigo.
- Primeiro preço do artigo, preço médio de compra, preço médio ponderado, último preço de compra, custos indiretos aplicados ao preço médio ponderado, custos indiretos aplicados ao último preço de compra: todos estes campos apresentam os seus respetivos valores acumulados à data o último registo utilizado na amostra. Estas são também características importantes dos artigos, porque retratam preços e o preço tem como regra geral uma quota grande na influência de uma venda.
- Vendido: este campo indica se o produto foi vendido pelo menos uma vez (no espaço temporal da amostra).
- Fornecedor, marca e departamento: estes são campos que permitem agrupar artigos em famílias e categorias, que têm uma especial importância em análises do tipo recomendação de artigos. Como conseguinte, estes campos foram levantados para serem posteriormente utilizados em análises deste género.

Existem 38,619 produtos introduzidos no sistema, no entanto, nem todos ainda são vendidos atualmente. Ao unir a tabela de produtos com a tabela acumulados, é possível saber quais os que ainda são vendidos atualmente. Isto é devido ao processo de limpeza que a Softingal realiza

na base de dados para reduzir consideravelmente o seu tamanho, excluindo muitos dados que não são de todo utilizados. Aplicando a união nas duas tabelas, o resultado apresenta um total de 14,992 produtos disponíveis para venda. Os artigos requerem que sejam inseridos no sistema, por operadores especializados, com toda a informação correta ao seu funcionamento. Portanto estes não estão sujeitos às subjetividades dos dados inseridos manualmente como se constata no caso de clientes. Como tal, todos os registos inseridos encontram-se devidamente preenchidos, não requerendo especial atenção no filtro e limpeza das informações relativas a artigos. Estes requerem apenas a padronização de alguns valores.

Os critérios utilizados para determinar se um registo está devidamente preenchido são:

- Quantidade por embalagem e unidade deve ser maior que zero
- Campos relacionados com os preços têm de ser superiores a zero, de forma a não contemplar ofertas.
- Os campos referentes a valores acumulados não podem ser nulos.

Os critérios de filtragem que se aplicaram nesta fase não restringiram qualquer dos artigos lidos no processo de extração, mantendo-se assim uma amostra de excelente qualidade com 14,992 artigos comercializados.

Apesar das mencionadas boas condições dos dados em torno de artigos, existem porém algumas situações que inviabilizam uma melhor caracterização do artigo. Os campos marca e departamento (dos dados recolhidos) fazem parte de uma estrutura que tem mySoftmais, que permite criar famílias e subfamílias estruturadas para os produtos. Esta estrutura é muito útil na caracterização de produtos ou criar grupos de artigos aglomerados por família. Embora seja possível combinar várias famílias ou subfamílias, a maioria dos artigos não está caracterizado pelas mesmas famílias. A tabela a seguir mostra as famílias utilizadas pela empresa para catalogar, o número de produtos que não podem ser catalogados e o número de vendas que são afetados.

Tabela 6 - Produtos e caraterizações por família

Família	Produtos não catalogados	Vendas afetas
Marca	37	3513
Linha	9079	280237
Secção	2408	216670
Subfamília	3141	224608
Tipo de produto	9040	86177
Departamento	3	25

Seria ótimo se fosse possível caracterizar os produtos para mais famílias, mas foram usados apenas aqueles que implicam menos perda de elementos caracterizados (marca e departamento).

Uma outra particularidade relevante encontrada surgiu ao tentar recolher todos os preços de tabela para cada produto, onde foram encontrados 22 artigos cujas vendas não tinham preços tabelados. Após esta questão examinada, a conclusão foi que estes produtos são utilizados como uma solução alternativa para representar algumas vendas em campanhas promocionais ou representar a oferta de alguns produtos. Foi decidido que os artigos de venda que contenham este tipo de particularidade seriam excluídos. A empresa utiliza este esquema para representar ocasiões promocionais, porque não recorrem aos recursos de gestão promocional oferecidos pelo *mySoftmais*. Existiam também outros artigos de oferta que possuíam tabela de preço que também foram descartados.

5.4.3 Vendas

O processo de ETL sobre as vendas mostrou-se bastante complexo e algo moroso. Nesta fase está envolvida essencialmente a gestão de documentos, que é a parte mais complexa da estrutura da base de dados *mySoftmais*. Como referido anteriormente é com base na gestão documental que o *software* processa todo o fluxo de trabalho da empresa, desde as encomendas a fornecedores, passando pela distribuição de artigos e gestão de stock até ao processo da venda ao cliente.

Nesta fase serão feitas recolhas de dados às maiores e mais complexas tabelas e as que contêm maior volume de dados. Dois exemplos de tal complexidade podem ser dados pela tabela do cabeçalho do documento e a tabela de linhas de documento. O primeiro apresenta uma enorme estrutura com cerca de 130 campos, com múltiplas relações com várias tabelas (totais, linhas, regimes, taxas, clientes, moradas, etc.). A segunda apresenta enormes volumes de dados (com um relação de N para 1 com a tabela de documentos) com mais de 3.500.000 registos na amostra fornecida.

Com a separação lógica efetuada entre clientes fidelidade e empresas, decidiu-se criar também uma separação entre os documentos para cada um dos tipos de clientes. Todo o processo que será agora descrito será comum para a extração de vendas, quer para clientes fidelidade como para empresas. A principal diferença reside no tipo de entidade alvo do documento e em algumas pequenas particularidades derivadas de mecanismos de controlo interno do *mySoftmais*.

Depois de avaliadas as necessidades das análises no que toca a vendas de clientes, foram ponderados e recolhidos os seguintes campos provenientes da união de diversas tabelas como cabeçalhos de documento, linhas, clientes, artigos entre outras tabelas:

- Número de contribuinte de cliente: este campo será o elemento que permite saber para quem se destina o documento ou seja, para que cliente foi feita a venda.

- Data: a data em que o documento foi feito é um fator que permite localizar temporalmente um documento e como tal, é de grande importância.
- Código, série, ano e número: a união destes 4 campos formam o identificador único deste documento, que é um fator necessário distinguir em qual das lojas foi feito, qual é o seu tipo de documento e qual a sua sequência.
- Código do produto: o identificador único do produto vendido é um elemento crucial no registo de uma linha de venda. Este fator torna essencial a recolha deste campo.
- Marca, departamento, fornecedor, embalagem, unidade e fator conversão unitário e taxa de IVA: todos estes campos referentes aos artigos foram necessários recolher, porque podem sofrer alterações no momento da venda em relação aos dados registados na sua ficha. A recolha destes campos sujeitos a alterações poderão ser relevantes aquando da avaliação de determinados comportamentos e características das vendas.
- Quantidade vendida: a quantidade unitária do artigo a ser vendido é uma característica importante de uma linha de venda para quantificar quantidades vendidas nas mais diversas circunstâncias de análise.
- Preço de venda, primeiro e segundo descontos utilizados e preço total: estes valores monetários são os principais influenciadores de vendas, dado que são condicionantes importantes à concretização ou não da venda por parte do cliente.
- Preço médio de compra, preço médio ponderado, último preço de compra, custos indiretos imputados ao produto para o preço médio ponderado, custos do produto indiretos imputados ao último preço de compra: estes valores são recolhidos no momento da venda e utilizados neste projeto, porque permitem percecionar a influência do preço de venda nos valores médios acumulados dos artigos vendidos.

Como estes dados estão devidamente controlados pelo sistema, não foram necessários controlos adicionais, filtragem de dados incorretos ou em falta. No entanto, existem algumas particularidades que necessitam de ser controladas na recolha de vendas:

- É necessário descartar todas as linhas que sejam faturadas como ofertas ao cliente. Estas serão descartadas por representarem artigos de amostras, ofertas promocionais, brindes, entre outros.
- Garantir a correta separação das vendas para ambos os tipos de clientes, visto que as vendas para cada um dos tipos são registadas internamente no *mySoftmais* com diversas particularidades que impedem que o mecanismo de recolha seja igual para ambos os casos.
- Aplicar cálculos e concluir valores que não estão presentes na base de dados, mas que são calculados pelo *mySoftmais.* Este programa possui diversas camadas de lógica de negócio e tratamento da informação que (de forma correta) não está explícita na base de dados (como *triggers*, *procedures*, *views*, etc.). Estudar o código do programa de forma a compreender como este processa a gestão documental é uma hipótese fora de questão, pelo imenso tempo que seria necessário despender para o fazer. Nesta fase a ajuda prestada pela Softingal nas orientações dadas à compreensão do processamento

existente no seu programa, demonstrou ser um fator de grande importância, porque possibilitou ter uma grande certeza na interpretação da informação existente na amostra fornecida.

Todos os campos que foram recolhidos e mencionados neste capítulo não são resultantes de uma única análise à estrutura de dados do *mySoftmais*. Estes são uma consequência de um ciclo iterativo entre o processo de ETL e o Análise Exploratória de Dados (AED). Além do levantamento das necessidades de dados, este ciclo permitiu também descobrir necessidades de padronização, de limpeza e de valores desajustados.

5.5 Exploração

Esta secção tem como objetivo descrever a metodologia utilizada na exploração da informação obtida até ao momento, numa fase pós elaboração do processo de ETL.

Na secção anterior foi detalhado todo o processo de extração, transformação e carregamento feito à amostra disponibilizada pela MASS Perfumaria. Este teve como objetivo preparar toda a informação necessária à realização dos grandes objetivos deste projeto, para o repositório de dados com que a aplicação irá funcionar. Como descrito anteriormente este repositório é uma *Staging Area* construída justamente para este propósito.

Depois de obtida a informação necessária, devidamente limpa e padronizada, é imperativo que o analista esteja bem familiarizado com a informação presente. Como esta área de estudo ainda depende muito do analista humano, é importante que este conheça bem a informação com a qual está a lidar, para que seja possível maximizar a qualidade das interpretações feitas aos resultados obtidos pelos algoritmos de *Data Mining*. Conhecer bem a informação trará também vantagens na medida em que permite uma melhor seleção dos dados a explorar pelos algoritmos de *Data Mining*, reduzir espaços de procura, diminuir o número de experiências necessárias, entre outros.

Para conhecer melhor os dados recolhidos após o processo de ET, e também validar a sua qualidade, recorreu-se ao R como ferramenta de aplicação de técnicas de AED. AED são um conjunto de técnicas gráficas e quantitativas criadas por John Wilder Tukey em 1977 com o objetivo de auxiliar a análise de amostras de dados [45]. Estas técnicas permitem [46]:

- A deteção de erros
- Verificações de hipóteses assumidas
- Seleções primárias para construção de modelos
- Extrair variáveis importantes
- Detetar anomalias e valores extremos

Estas podem ser aplicadas aos mais diferentes tipos de valores (nominais, contínuos, discretos, etc.) através da aplicação de:

- Gráficos de caixa
- Histogramas
- Gráficos de múltiplas séries
- Gráfico de dispersão
- Diagrama de Pareto
- Trimean
- Entre muitas outras técnicas de análise de valores

O processo de exploração resulta também na verificação da necessidade de aplicar fatores adicionais de limpeza de dados ou na necessidade de mais dados adicionais no processo de ETL. A AED juntamente com o processo de ETL formam um ciclo que permitem melhorar iterativamente a informação a que está sujeita a análise.

Em seguimento da lógica do processo de ETL esta fase de exploração foi também subdividida em três partes nas quais se explorou a informação relativa a clientes, produtos e vendas. Numa primeira fase foram explorados individualmente todos os atributos de cada uma destas partes em questão. Nesta exploração individual o interesse consistia em saber qual o alcance dos valores, médias, desvios padrões, valores mais e menos frequentes, e entre outras medidas quantitativas. Numa segunda fase, para cada parte em questão, com base em assunções e taxas de correlação, foram explorados conjuntamente dois ou mais atributos. O objetivo desta segunda fase consiste em detetar comportamentos ou tendências entre atributos significativamente importantes para a área de negócio em questão.

Ambas as fases foram implementadas utilizando os recursos e potencialidades da linguagem R. Todos os gráficos realizados na exploração foram desenvolvidos recorrendo a uma biblioteca desenvolvida para este efeito, denominada de GGplot2 [47]. Esta biblioteca foi criada por Hadley Wickham e oferece uma poderosa linguagem gráfica para a criação de gráficos visualmente apelativos e no entanto complexos. Esta tem sido cada vez mais utilizada dentro da comunidade R nos últimos anos. Permite criar gráficos que representam atributos numéricos, categóricos e até de múltiplas variáveis [48]. A sintaxe muito específica de GGplot2 requer algum tempo de habituação, mas que se mostra compensadora pelas grandes potencialidades da biblioteca e pelo *design* apelativo dos gráficos criados.

5.6 Avaliação de Resultados

A avaliação de resultados tem uma vertente de grande importância na verificação do trabalho elaborado, independentemente da sua natureza. Por conseguinte é importante que definam corretas medidas de avaliação mesmo antes da implementação. Esta avaliação deverá ser qualitativa e se possível seguir medidas/boas práticas de engenharia.

Como referido anteriormente, é crucial obter uma boa qualidade dos dados que são tratados neste projeto. Como tal, é importante avaliar a qualidade da informação que é extraída da amostra de dados durante o processo de ETL. Então, são feitas diversas iterações entre o processo ETL e as Análises Exploratórias de Dados (AED) [49] até que seja assegurada uma amostra de dados a ser utilizada, devidamente limpa, padronizada, sem valores de ruído ou em falta. Ao aplicar o AED é utilizada uma avaliação de métricas para detetar valores nulos, negativos, desajustados, etc. É utilizada a apreciação do analista para filtrar determinados situações como: produtos nunca vendidos, ofertas, clientes com idades falsas, empresas registadas como clientes, etc.

6 Classificação

A Mass Perfumarias é uma empresa que ao longo dos seus 30 anos de existência realizou uma forte aposta na fidelização de clientes e em preços mais baixos relativos à generalidade da concorrência, adquiriu um conjunto bastante significativo de clientes que se deslocam às suas lojas com relativa regularidade. Estes clientes embora todos consumidores de produtos de perfumaria e beleza, serem maioritariamente senhoras e da região norte do país, apresentam diversas diferenças entre si.

Estas diferenças encontram-se no vasto leque etário, na antiguidade, no tipo de itens adquiridos, preferências de marcas, rácio de frequência à loja, nas diferentes épocas do ano em que visitam a loja, entre outros diversos aspetos pessoais. Estes clientes além das suas diferentes características apresentam também comportamentos distintos. Existem clientes que frequentam a loja 1 vez de 2 em 2 anos, existem clientes que a frequentam 3 vezes por ano, clientes que compram poucos itens mas de elevado valor, clientes que compram vários itens mas de pequeno valor. Esta diversidade de clientes e das suas características apresentam um problema para empresa.

A MASS Perfumarias encontra grandes dificuldades quando pretende classificar os seus clientes, de modo a que seja possível distinguir as particularidades de um bom cliente, de um cliente médio ou ainda de um cliente de topo de um cliente fraco.

Um dos objetivos deste projeto passa por criar um modelo preditivo de classificação para clientes, que permita à empresa ter um mecanismo capaz de prever a classificação de novos clientes e assim direcionar melhor o seu marketing e apresentar melhores sugestões de produtos.

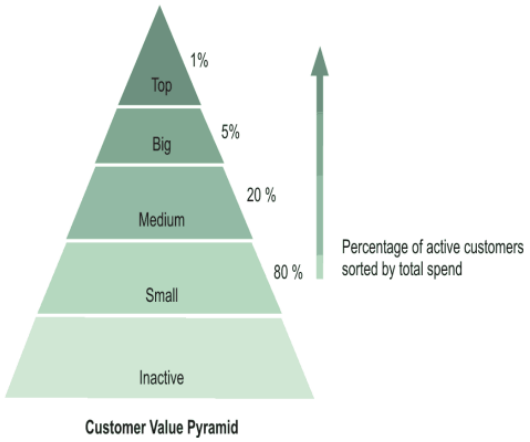


Figura 7 – Pirâmide de valor de cliente
Fonte: [44]

O atributo objetivo desta classificação será o atributo classe. Como já referido esta classe é gerada com base na pirâmide do *Application Template for Customer Relationship Management* [44] recorrendo ao valor total gasto. Esta caracterização passa pela atribuição de uma das seguintes classificações: inativo; pequeno; médio; grande; topo. Uma representação desta pirâmide pode ser observada na Figura 7.

É importante realçar que o modelo de classificação a ser gerado, não deverá ser aplicado imediatamente a novos clientes fidelidade. Os atributos que caracterizam o cliente não são suficientes para criar um modelo de classificação. É necessário a adição de outros atributos característicos do seu comportamento (número de compras, número de itens comprados, número de pontos, etc.) para a construção deste modelo preditivo. Como tal, a caracterização de novos clientes deverá ser feita, por exemplo, no final de um ano. É recomendável que exista compasso de espera entre a adesão do cliente e a sua classificação, para que hajam registos suficientes para avaliar o comportamento do cliente e assim este poder ser corretamente classificado.

6.1 Definição

Criar um modelo de classificação consiste em criar modelos que descrevem o grupo ao qual o item pertence, através da análise dos itens já classificados e pela inferência de um conjunto de regras [50]. Consiste em examinar determinadas características nos dados e atribuir uma classe previamente definida. Desta forma os dados podem ser associados a classes ou a conceitos através de um processo de discriminação ou de caracterização [51].

A construção de um modelo de classificação é um processo de dois passos:

- O primeiro passo consiste em construir um classificador que descreve um conjunto pré-determinado de classes ou conceitos, na qual se denomina fase de treino ou função $y = f(X)$ [6]. Neste passo, são utilizados algoritmos de classificação que descobrem relações entre atributos que tornam possível a previsão de uma classificação. Os algoritmos são aplicados a uma parte da amostra, a qual se denomina conjunto de treino e o seu resultado será um classificador capaz de prever uma classe para um dado elemento de conjunto.
- No segundo passo aplica-se o classificador criado numa outra parte da amostra denominada de conjunto de teste, com a finalidade de obter a precisão deste modelo. A precisão é dada pelo número de casos que foram corretamente previstos pelo classificador ou seja, representa a taxa de acerto do classificador.

6.2 Implementação

Este capítulo pretende descrever todo o processo de implementação que está inerente à implementação de modelos de classificação de clientes. Serão abordados quais os cuidados com os dados antes da implementação dos modelos, quais os conjuntos candidatos, resultados das experiências efetuadas, otimizações aos modelos criados e as comparações efetuadas entre estes.

6.2.1 Preparação de Dados para Classificação

Com a diversidade de atributos recolhidos e calculados sobre estas entidades é necessário que primeiro sejam realizadas mais algumas operações sobre os dados a utilizar. Esta fase denomina-se de preparação de dados para classificação e previsão [6]:

- Detetar dados que provoquem ruído.
- Detetar campos redundantes.
- Aplicar generalização ou discretização de valores contínuos ou discretos.
- Verificar a existência de atributos redundantes e/ou irrelevantes através de análises de correlação.

Esta fase permitiu identificar que existem atributos que necessariamente teriam de ser generalizados e outros que poderiam ser excluídos, por terem uma alta taxa de correlação quando relacionados com outros atributos ou por serem causadores de ruído. Verificou-se após esta fase, que poderiam ser feitos os seguintes ajustes:

- Descartados os campos: pontos atuais e valores de pontos, por apresentarem uma elevada taxa de correlação com o número total de pontos.
- Os campos: cidade e código postal por serem causa de ruído foram descartados.
- Os campos: idade, número total de pontos e dias sem compras foram discretizados.

Após reunidos os diversos atributos de clientes fidelidade, concluídas algumas iterações entre ETL e AED, elaborada a arquitetura deste projeto e seleção da técnica de classificação mais adequada ao presente cenário e preparados os dados para a classificação, encontram-se então reunidas as condições necessárias à implementação de modelos de classificação sobre clientes.

6.2.2 Conjuntos Candidatos

Com base nos diferentes atributos recolhidos e tratados foi possível obter três conjuntos de atributos candidatos à construção dos modelos. Esta separação de atributos foi feita dado que, existem atributos cujo significado seja semelhante a outros e até correlacionados mas com uma perspetiva de utilização diferente, por exemplo: número médio itens comprados versos número de idas à loja e número de itens comprados.

Serão agora apresentados individualmente cada conjunto candidato. Para cada conjunto serão apresentados os respetivos atributos que o constituem, assim como o seu respetivo índice de atributo no conjunto. Serão apresentados quais os atributos que foram utilizados e descartados pelo C5.0, o tamanho da árvore resultante, a percentagem de acerto, o número de erros de previsão e a respetiva matriz de confusão relativos ao conjunto de teste.

6.2.2.1 Conjunto 1

Este conjunto é composto por características de cliente comportamentais e valores médios relativos a valores de compras.

Tabela 7 - Resumo do conjunto candidato 1

<table>
<tr><td colspan="4">Conjunto 1</td></tr>
<tr><td colspan="2">Dados em Análise</td><td colspan="2">Árvore de decisão</td></tr>
<tr><td>Índice</td><td>Atributo</td><td></td><td></td></tr>
<tr><td>15</td><td>Valor médio compra</td><td>Tamanho</td><td>51</td></tr>
<tr><td>16</td><td>N. médio itens comp.</td><td>Erros</td><td>403(3,3%)</td></tr>
<tr><td>17</td><td>Freq. visita à loja</td><td rowspan="14">Atributos utilizados C5.0</td><td rowspan="14">100.00% N. médio itens comp.
60.16% Cosmética
58.90% Vouchers p. ano
55.74% Perfumaria
48.08% Toucador
17.21% Anos de registo
3.19% Valor médio compra
1.39% Freq. visita à loja
1.27% Loja habitual
0.57% Zona</td></tr>
<tr><td>22</td><td>Compras em</td></tr>
<tr><td>23</td><td>Loja habitual</td></tr>
<tr><td>26</td><td>Vouchers p. ano</td></tr>
<tr><td>27</td><td>Anos de registo</td></tr>
<tr><td>28</td><td>Acessórios</td></tr>
<tr><td>29</td><td>Cosmética</td></tr>
<tr><td>30</td><td>Beleza</td></tr>
<tr><td>31</td><td>Produtos s. classificação</td></tr>
<tr><td>32</td><td>Perfumaria</td></tr>
<tr><td>33</td><td>Toucador</td></tr>
<tr><td>34</td><td>Zona</td></tr>
<tr><td>35</td><td>Grupo etário</td></tr>
<tr><td colspan="2">Observações</td><td>Acerto</td><td>0.9784621</td></tr>
<tr><td colspan="4"></td></tr>
</table>

Tabela 8- Matriz de confusão do modelo gerado com o conjunto candidato 1

Matriz de Confusão						
Previsão	**Inativo**	**Pequeno**	**Médio**	**Grande**	**Top**	Total
Grande	0	0	22	42	8	72
Inativo	2063	0	0	0	0	2063
Médio	0	74	399	31	1	503
Pequeno	0	2439	59	0	0	2498
Top	0	0	0	10	27	37
Total	2063	2513	480	83	36	5175

6.2.2.2 Conjunto 2

O conjunto 2 é composto por atributos de comportamento e valores médios relativos a valores de compras. Inclui também o número de pontos, compras e o número de dias sem compras.

Tabela 9 - Tabela resumo do conjunto candidato 2

Conjunto 2			
Dados em Análise		**Árvore de decisão**	
Índice	**Atributo**		
13	Compras	Tamanho	55
15	Valor médio compra	Erros	380(3,1%)
16	N. médio itens comp.	Atributos utilizados C5.0	100.00% N. médio itens comp. 60.16% Cosmética 57.22% Perfumaria 55.31% Grupo de pontos 48.42% Toucador 16.92% Anos de registo 10.75% Vouchers p. ano 2.66% Valor médio compra 1.20% Loja habitual 1.13% Freq. visita à loja 0.98% Zona 0.50% N. dias sem compras 0.26% Grupo etário
17	Freq. visita à loja		
37	N. dias sem compras		
22	Compras em		
23	Loja habitual		
26	Vouchers p. ano		
27	Anos de registo		
28	Acessórios		
29	Cosmética		
30	Beleza		
31	Produtos s. classificação		
32	Perfumaria		
33	Toucador		
34	Zona		
35	Grupo etário		
36	Grupo de pontos		
Observações		Acerto	0,9608145

Tabela 10 - Matriz de confusão do modelo gerado com o conjunto candidato 2

Matriz de Confusão						
Previsão	Grande	Inativo	Médio	Pequeno	Top	Total
Grande	0	0	26	45	7	78
Inativo	2063	0	0	0	0	2063
Médio	0	75	392	26	2	495
Pequeno	0	2438	62	0	0	2500
Top	0	0	0	12	27	39
Total	2063	2513	480	83	36	5175

6.2.2.3 Conjunto 3

Este conjunto é composto por características de cliente comportamentais e valores totais relativos a valores de compras.

Tabela 11 - Resumo do conjunto candidato 3

Conjunto 3			
Dados em Análise		**Árvore de decisão**	
Índice	**Atributo**		
12	Total Itens comprados	Tamanho	41
13	Compras	Erros	411(3,4%)
37	N. dias sem compras	Atributos utilizados C5.0	100.00% Compras 60.16% Cosmética 57.22% Perfumaria 52.44% Grupo de pontos 46.92% Total Itens comprados 16.44% Anos de registo 10.86% Vouchers 1.85% Loja habitual 0.98% Zona 0.50% N. dias sem compras 0.34% Toucador 0.26% Grupo etário
22	Compras em		
23	Loja habitual		
25	Vouchers		
27	Anos de registo		
28	Acessórios		
29	Cosmética		
30	Beleza		
31	Produtos s. classificação		
32	Perfumaria		
33	Toucador		
34	Zona		
35	Grupo etário		
36	Grupo de pontos		
Observações		Acerto	0.9656213

Tabela 12 - Matriz de confusão do modelo gerado com o conjunto candidato 3

Matriz de Confusão						
Previsão	Grande	Inativo	Médio	Pequeno	Top	Total
Grande	0	0	24	37	8	69
Inativo	2063	0	0	0	0	2063
Médio	0	95	382	30	1	508
Pequeno	0	2418	73	1	0	2492
Top	0	0	1	15	27	43
Total	2063	2513	480	83	36	5175

Com os conjuntos apresentados levantam-se duas importantes questões:

- Os modelos apresentam um número muito elevado de folhas. Este fator indica que são modelos complexos que não produzem grande relevância analítica. Será necessário que estes sejam sujeitos a um processo que os simplifique, de modo a que estes produzam árvores de decisão mais pequenas e assim mais compreensíveis.
- A existência de vários modelos candidatos levanta a necessidade de descobrir qual o que representará um melhor modelo de classificação. É necessário recorrer a um processo de comparação recorrendo a medidas de avaliação/comparação de modelos de classificação.

6.2.3 Simplificação de Modelos

Como mencionado no final do subcapítulo anterior, os modelos candidatos a modelos de classificação de clientes apresentam uma complexidade elevada. Esta complexidade é dada pelo grande número de folhas que estes apresentam (média de 50 folhas) face ao número de atributos que os constituem. Será necessário que o modelo eleito possua um número de folhas significativamente menor, com a finalidade de obter um modelo simplificado para uma melhor compreensão.

O processo de simplificação utilizado neste projeto consiste em ajustar o volume de dados utilizado como amostra para treino e teste. Todos os modelos candidatos são passíveis de serem simplificados devido à sua elevada taxa de acerto, cuja média ronda os 95%. Esta taxa de acerto é suficientemente alta para permitir que se sacrifique alguma precisão por um modelo mais simples. Outro fator que possibilita este processo é a dimensão da amostra inicial, cujos 17 mil clientes devidamente selecionados e padronizados permite que se criem amostras aleatórias mais pequenas sem consequência na qualidade.

Com a necessidade de simplificar vários modelos e dada a natureza cíclica deste processo, optou-se pelo desenvolvimento de um algoritmo escrito em R cujo objetivo será simplificar um dado modelo.

Este algoritmo será composto por uma função, cujo retorno será a percentagem de dados da amostra que minimiza o número de folhas face a um dado limite de percentagem de acerto. A Figura 8 apresenta a assinatura da função criada (denominada de *ModelAdjustment*) que tem como parâmetros de entrada um qualquer modelo, o seu atributo objetivo e qual a percentagem de acerto mínima.

```
ModelAdjustment <- function(Model, goalAttribute, accuracy) {
```

Figura 8 - Assinatura da função de simplificação de modelos

O funcionamento de *ModelAdjustment* encontra-se ilustrado no fluxograma apresentado na Figura 9 que se segue.

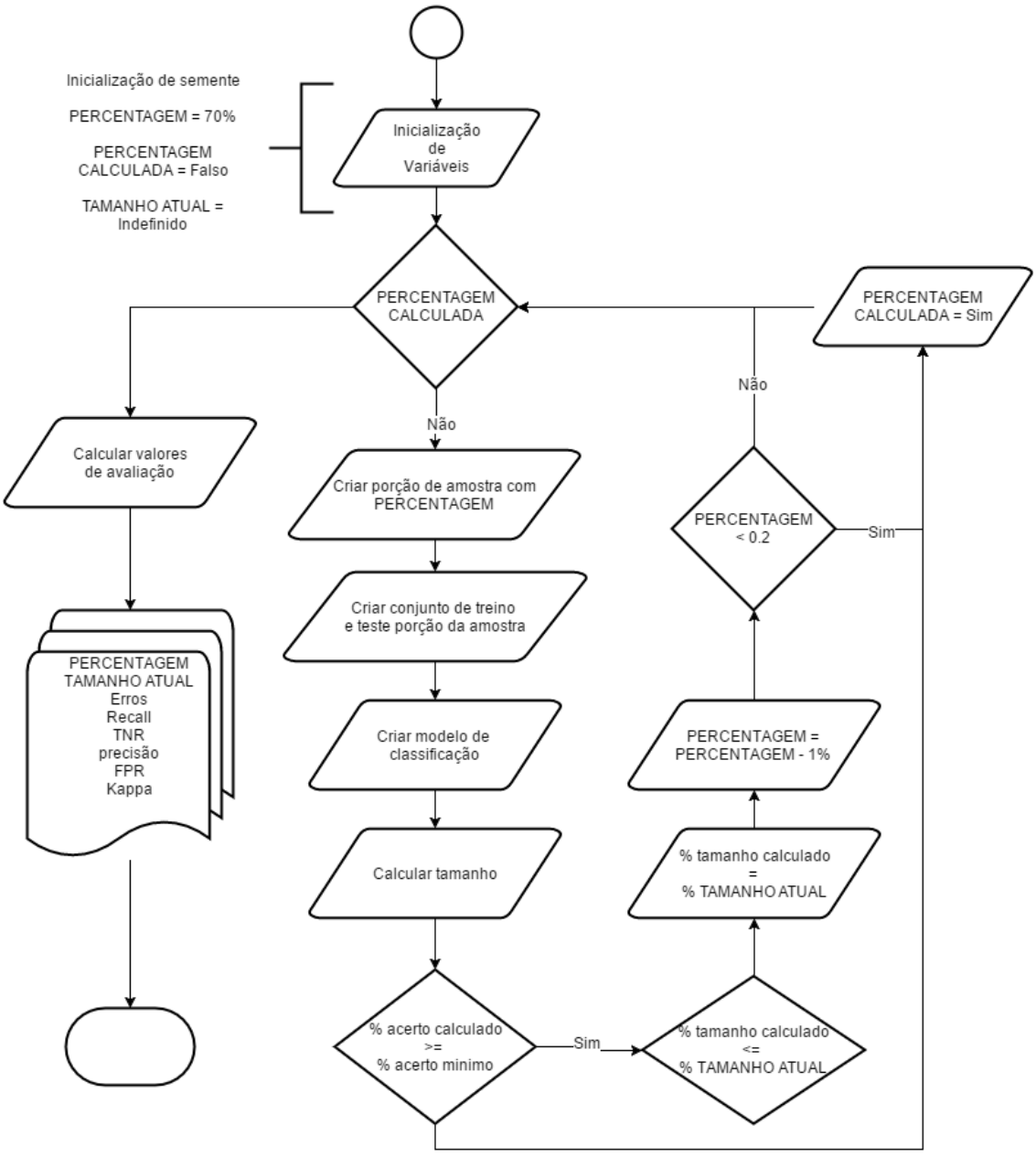


Figura 9 - Fluxograma de ModelAdjustment

A função de simplificação de modelos apresenta como *output*, além da percentagem da amostra utilizada, outras informações relevantes à avaliação de modelos (avaliação esta que será abordada de seguida neste documento), o número de iterações feitas, de folhas e o de erros. Um exemplo do output da função *ModelAdjustment* pode ser observado na Figura 10 e na Figura 11.

```
[1] "Data Percentage:  0.58"
[1] "Iterations:  13"
[1] "Size:  41"
[1] "Errors:  277( 3.3%) "
[1] "accuracy:  0.955474656958835"
[1] "recall:  0.542372881355932"
[1] "tnr:  1"
[1] "precision:  0.551724137931034"
[1] "fpr:  0"
[1] "f1:  0.547008547008547"
[1] "kappa:  0.947052833492949"
```

Figura 10 - Output da função *ModelAdjustment*

```
[1] "0.2 of data percentage reached"
[1] "Data Percentage:  0.2"
[1] "Iterations:  51"
[1] "Size:  23"
[1] "Errors:  130( 3.4%) "
[1] "accuracy:  0.942632850241546"
[1] "recall:  0.423076923076923"
[1] "tnr:  1"
[1] "precision:  0.407407407407407"
[1] "fpr:  0"
[1] "f1:  0.415094339622641"
[1] "kappa:  0.931778976372421"
```

Figura 11 - Output de percentagem mínima de 20%

Existe ainda um indicador de percentagem mínima da amostra de 20%. O output apresentado em casos onde a função itere até atingir os 20% de amostra é apresentado o *output* ilustrado na Figura 11. O indicador "0.2 of data percentage reached" indica que a função tentou utilizar até à percentagem mínima de otimização de 20% da amostra inicial e não conseguiu otimizar o conjunto até à percentagem de acerto especificada. Significa que utilizar menos de 20% da amostra não irá reduzir substancialmente o conjunto, de modo a que se justifique a perda da qualidade da amostra.

ModelAdjustment apresenta ainda um gráfico representativo da curva de aprendizagem que o modelo apresenta, comparando o acerto e a percentagem de dados utilizada ao longo do processo. A Figura 12 mostra um exemplo de uma curva de aprendizagem calculada por *ModelAdjustment* para um dado conjunto.

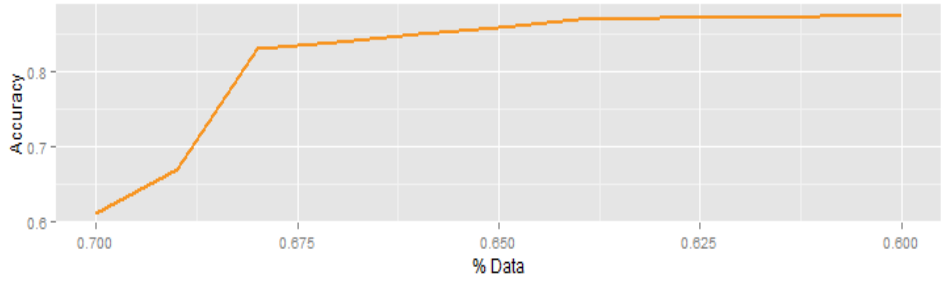


Figura 12 - Exemplo de Curva de Aprendizagem

Após a elaboração desta função é possível simplificar os três modelos candidatos deste projeto de forma autónoma e padronizada. Os três modelos serão agora processados por esta função, de modo a obter a percentagem de informação que otimiza a dimensão do modelo.

Para cada modelo será agora apresentada a percentagem de informação que a função recomenda dado o limite de 94% de acerto e os restantes valores do *output* apresentado. Com base nesta percentagem e avaliando os restantes parâmetros do output serão gerados três conjuntos candidatos simplificados.

6.2.3.1 Simplificação Conjunto 1

Ao aplicar a função *ModelAdjustment* ao conjunto 1 com a uma taxa de acerto mínima de 94% são apresentados os seguintes valores.

Tabela 13 - Valores resultantes da simplificação do conjunto 1

Atributo	**Valor**
Percentagem de dados	0.3%
Iterações	41
Tamanho	21
Erros	3.4%
Taxa de acerto	0.9473335
Recall	0.433333
TNR	1
Precision	0.52
FPR	0
F1	0.4727273
Kappa	0.937203

Com base neste *output* será utilizada a percentagem de 30% da amostra inicial sobre o conjunto 1 na geração do modelo de classificação. Como indicado pela função, esta percentagem permitirá reduzir o modelo de 51 folhas para 23 com o aumento 0.1% de erros de teste.

É necessário ter em conta que a dimensão e qualidade da amostra permitem que se utilize 30% desta sobre o conjunto 1 e os restantes conjuntos, caso contrário a percentagem de acerto teria de ser ajustada para que a função retornasse uma percentagem de amostra mais elevada.

6.2.3.2 Simplificação Conjunto 2

Relativamente ao conjunto candidato 2, em semelhança com o conjunto 1 apresenta um número elevado de 55 folhas face à dimensão da amostra. Ao ser aplicada a função *ModelAdjustment* ao conjunto 2 com a uma taxa de acerto mínima de 94% é apresentado o seu *output* na Tabela 14.

Tabela 14 - Valores resultantes da simplificação do conjunto 2

Atributo	**Valor**
Percentagem de dados	0.37%
Iterações	34
Tamanho	22
Erros	4%
Taxa de acerto	0.945223
Recall	0.458333
TNR	1
Precision	0.594594
FPR	0
F1	0.517647
Kappa	0.946744

Com base neste *output* apresentado na tabela anterior, será utilizada a percentagem de 37% da amostra inicial relativamente ao conjunto candidato 2 na geração do modelo de clientes. Como indicado pela função, esta percentagem permitirá reduzir o modelo de 55 folhas para apenas 22 com o aumento 0.9% de erros de teste.

6.2.3.3 Simplificação Conjunto 3

O conjunto candidato 3 por sua vez também apresenta um número elevado de 41 folhas face à dimensão da amostra. Ao aplicar a função de simplificação *ModelAdjustment* ao conjunto 3 com uma igual taxa de acerto mínima de 94% é apresentado o seguinte *output* ilustrado na Tabela 15.

Tabela 15 - Valores resultantes da simplificação do conjunto 3

Atributo	**Valor**
Percentagem de dados	0.23%
Iterações	48
Tamanho	14
Erros	4%
Taxa de acerto	0.938044
Recall	0.285714
TNR	1
Precision	0.125
FPR	0
F1	0.173913
Kappa	0.926315

Com base nos valores apresentados na tabela anterior, será utilizada a percentagem de 23% da amostra inicial relativamente ao conjunto candidato 3 na geração do modelo de clientes. Como indicado pela função, esta percentagem permitirá reduzir o modelo de 41 folhas para apenas 14 com o aumento 0.6% de erros de teste.

A curva de aprendizagem apresentada por *ModelAdjusment* para os três conjuntos deste projeto não é abordada, devido ao facto de a variância do acerto para estes modelos ser muito pequena. Com uma variância na casa dos 0.3%, o gráfico não apresenta uma curva relevante ao auxilio do analista na escolha de uma percentagem de dados adequada ao seu objetivo e/ou limites.

O processo de simplificação agora descrito, permitiu a simplificação dos 3 conjuntos candidatos com uma redução do tamanho dos modelos de classificação em mais de 50% com um aumento médio da taxa de erros de 5%.

Com três conjuntos candidatos convertidos em modelos simplificados de classificação é necessário descobrir qual o melhor destes modelos ou seja, qual o modelo que melhor representa clientes fidelidade da MASS Perfumarias. Para realizar esta tarefa, deu-se início a um processo de comparação de modelos.

6.2.4 Comparação de Modelos

Como mencionado anteriormente neste capítulo existem três conjuntos candidatos a modelos de classificação. Com a existência de mais do que um modelo candidato, advém a necessidade de descortinar qual dos modelos será uma melhor representação de clientes fidelidade da empresa em questão. Para que se possa avaliar um determinado modelo, existem enumeras métricas e técnicas passíveis de serem utilizadas. Existem medidas que são transversais a quaisquer modelos e outras que dependem da natureza deste. A decisão de quais as métricas a utilizar e avaliação desses mesmos resultados ficam ao encargo do analista. Este facto deve-se à natureza dos dados, da diversidade dos cenários de avaliação e dos objetivos pretendidos.

Para desempenhar a tarefa de comparar os diferentes modelos, elaborou-se um algoritmo projetado numa função representada na linguagem R, que expõe de forma comparativa diversas medidas de avaliação de modelos. O objetivo desta função é auxiliar um analista na comparação de mais do que um modelo, independente da sua natureza. Esta é denominada de *ModelComparison*.

ModelComparison apresenta 3 argumentos de entrada, como sugere a Figura 13.

```
ModelComparison <- function(modelVec, goalAttribute, modelIDs) {
```

Figura 13 - Assinatura da função *ModelComparison*

O primeiro argumento (*modelvec*) representa um vetor de conjuntos de modelos a serem avaliados. O argumento *goalAttribute* é a especificação de qual o atributo objetivo. *ModelIDs* representa um vetor com uma identificação de cada modelo no vetor *modelvec* (por ex. [modelo quantitativo, modelo de valores médios,…]).

O funcionamento desta função é apresentado na Figura 14. Como esta sugere, a função percorre uma iteração cíclica para cada conjunto a ser comparado. Para cada conjunto são calculadas diversas medidas de avaliação de modelos de classificação. Todos os cálculos efetuados são apresentados ao analista para que este facilmente consiga comparar os modelos enviados para *ModelComparison*.

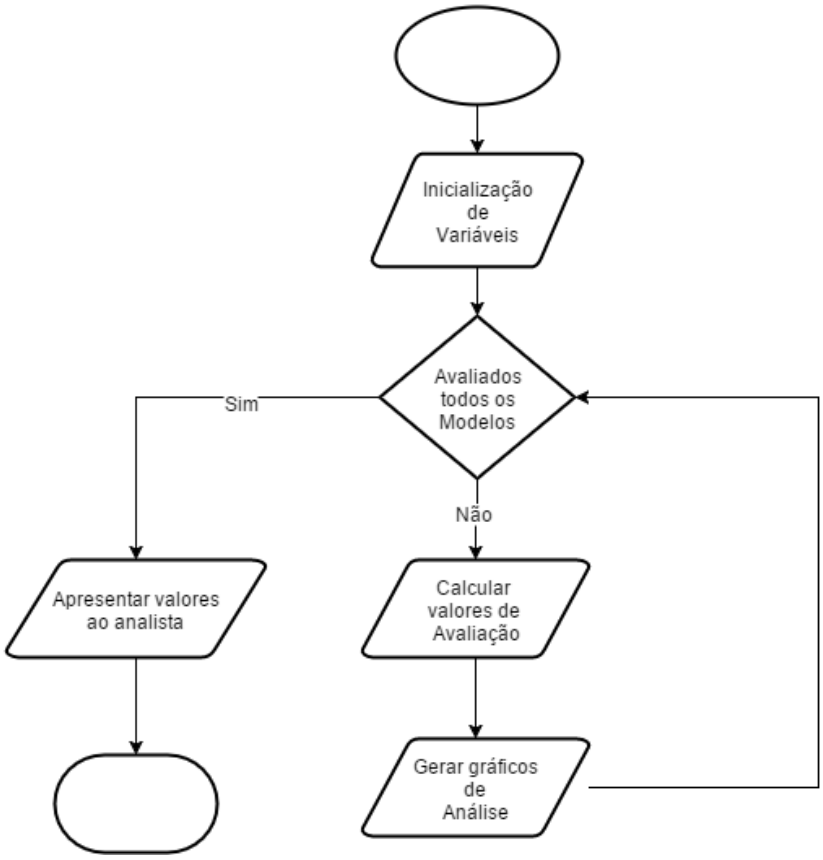


Figura 14 - Fluxograma da função *ModelComparison*

Neste algoritmo não foram verificadas todas as métricas existentes de avaliação de modelos. Foram abordadas aquelas que permitem avaliar modelos, independentemente da sua natureza assim como outras aplicáveis aos modelos aqui elaborados. Estas métricas são as descritas na subsecção Medidas de avaliação de modelos 2.4.2 Medidas de avaliação de modelos.

Um exemplo de parte do output gerado por esta função pode ser visto na Figura 15. Todas as medidas de avaliação e gráficos apresentados por esta função, serão detalhados na seguinte subsecção.

```
Model Comparison

  X.Model1. size      errors accuracy  kappa recall precision     f1 tnr fpr
1    Model1   57 207( 2.4%)    0.9539 0.9452 0.4405    0.6379 0.5211   1   0
  X.Model2. size      errors accuracy kappa recall precision     f1 tnr fpr
1    Model2   51 236( 2.8%)    0.9638 0.957  0.725       0.5 0.5918   1   0
  X.Model3. size      errors accuracy  kappa recall precision     f1 tnr fpr
1    Model3   50 193( 2.3%)    0.9578 0.9498 0.4722    0.5862 0.5231   1   0

Areas under curve
Model1 :  88.5512
Model2 :  91.6651
Model3 :  90.2591
```

Figura 15 – Parte do *Output* de *ModelComparison*

```
Macro Precisions
Model1 :  0.2124
Model2 :  0.2095
Model3 :  0.2444

Micro Precisions
Model1 :  0.0296
Model2 :  0.0288
Model3 :  0.0375

Macro Recall
Model1 :  0.2056
Model2 :  0.2172
Model3 :  0.1843

Micro Recall
Model1 :  0.03
Model2 :  0.0293
Model3 :  0.0381
```

Figura 16 – Parte do Output de *ModelComparison*

Esta função apresenta ainda ao analista um importante elemento de comparação de modelos que é a curva ROC. Um exemplo desta curva gerada por *ModelComparison* é sugerido pela Figura 17.

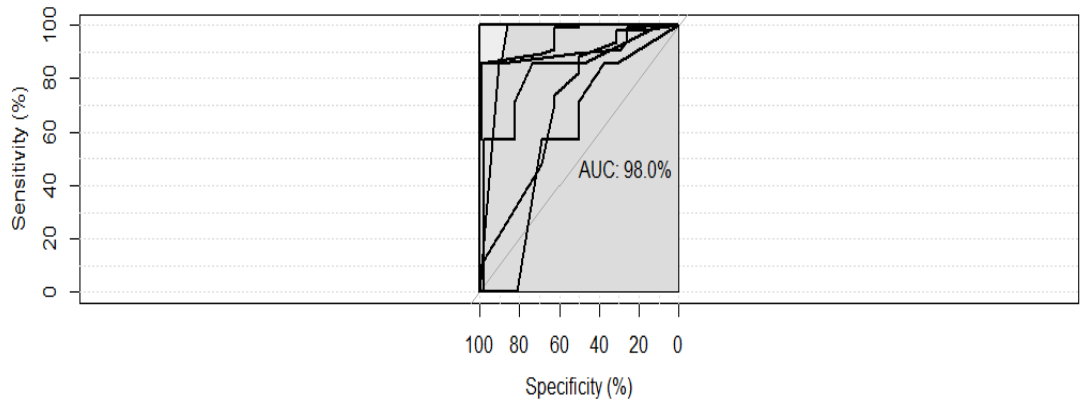
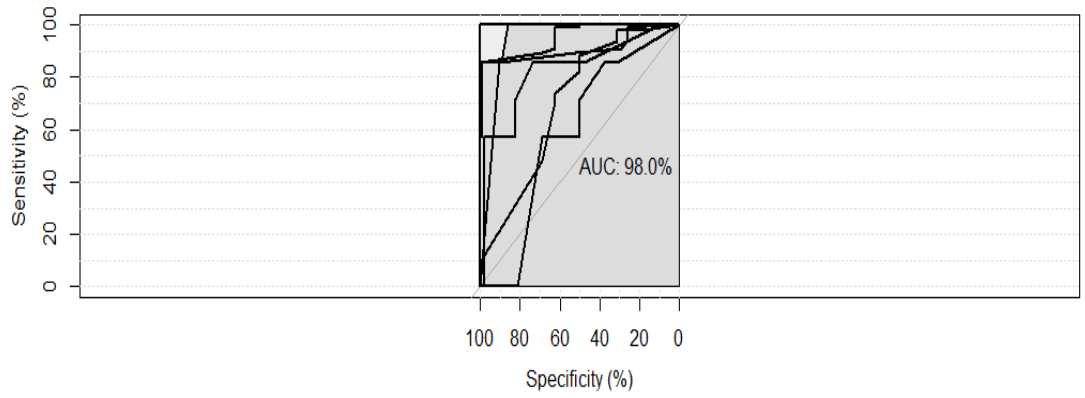


Figura 17 - Exemplo Curva ROC

Aplicando *ModelComparison* com os três modelos deste projeto já simplificados com a função, obtém-se os seguintes *outputs*:

```
  X.Model1. size      errors accuracy  kappa recall precision     f1 tnr fpr
1    Model1   23  122( 3.4%)   0.9473 0.9372 0.4333      0.52 0.4727   1   0
  X.Model2. size      errors accuracy  kappa recall precision     f1 tnr fpr
1    Model2   22  212( 4.0%)   0.9452 0.9467 0.4583    0.5945 0.5176   1   0
  X.Model3. size      errors accuracy  kappa recall precision     f1 tnr fpr
1    Model3   14  97( 4.0%)    0.9380 0.9263 0.2857      0.125 0.1739   1   0

Areas under curve
Model1 :  89.2538
Model2 :  84.2698
Model3 :  93.3736

Macro Precisions
Model1 :  0.1597
Model2 :  0.1719
Model3 :  0.0729

Micro Precisions
Model1 :  0.021
Model2 :  0.0229
Model3 :  0.0115

Macro Recall
Model1 :  0.1757
Model2 :  0.1819
Model3 :  0.1512

Micro Recall
Model1 :  0.0207
Model2 :  0.0226
Model3 :  0.0111
```



Analisando o *output* anterior, conclui-se que o modelo que melhor representa clientes fidelidade da MASS perfumarias é o modelo 3. Este modelo é o que apresenta menos folhas, logo é o mais simples, é o modelo que apresenta uma taxa de erro menor, embora seja um pouco menos preciso que os restantes. Isto é, compensado pela simplicidade do modelo e assim ser mais fácil de interpretar.

6.2.5 Resultados

Um dos grandes objetivos deste projeto é a criação de um modelo de classificação para clientes fidelidade da MASS Perfumarias. Para concretizar este objetivo foi realizado um estudo das possíveis técnicas/objetivos, que permitiam criar estes modelos dentro das necessidades deste projeto. De seguida foram reunidos três conjuntos candidatos a modelos de classificação que foram consequência de um processo de ETL construído para este efeito. De seguida estes três conjuntos candidatos foram transformados em três respetivos modelos de classificação recorrendo-se ao algoritmo C5.0 e à linguagem R.

Estes modelos tiveram a necessidade de passar por duas fazes: simplificação e comparação. A fase de simplificação foi necessária devido à dimensão e complexidade que os modelos apresentavam. As árvores de decisão consequentes destes apresentavam uma dimensão demasiado grande para serem facilmente compreendidas. A fase de comparação foi igualmente importante, porque naturalmente dos três modelos apenas seria utilizado um como o modelo representante de clientes.

Após todas estas fases, elegeu-se o conjunto candidato/modelo 3 como o melhor conjunto representativo de clientes fidelidade. O resultado final deste grande objetivo pode ser concluído com a árvore de decisão ilustrada na Figura 18.

```
Purchases <= 0: INACTIVE (964)
Purchases > 0:
:...Cosmetic <= 11:
    :...Perfumery > 6:
    :   :...TotalPointsGroup = Sem pontos: MEDIUM (0)
    :   :   TotalPointsGroup = G. Pontos 1: SMALL (32/1)
    :   :   TotalPointsGroup in {G. Pontos 2,G. Pontos 3,G. Pontos 4}:
    :   :   :...Purchases <= 9: SMALL (24/9)
    :   :       Purchases > 9: MEDIUM (99/25)
    :   Perfumery <= 6:
    :   :...Purchases <= 33: SMALL (1059/15)
    :       Purchases > 33:
    :       :...Purchases > 70: MEDIUM (7/1)
    :           Purchases <= 70:
    :           :...TotalPointsGroup in {G. Pontos 1,G. Pontos 2,G. Pontos 4,
    :               :                    Sem pontos}: SMALL (39/6)
    :               TotalPointsGroup = G. Pontos 3: MEDIUM (12/4)
    Cosmetic > 11:
    :...TotalPointsGroup = G. Pontos 1: SMALL (16)
        TotalPointsGroup in {G. Pontos 2,G. Pontos 3,
        :                    Sem pontos}: MEDIUM (101/16)
        TotalPointsGroup = G. Pontos 4:
        :...Cosmetic <= 32:
            :...Vouchers <= 7: MEDIUM (30/13)
            :   Vouchers > 7: BIG (6/2)
            Cosmetic > 32:
            :...Zone = BRAGA: TOP (0)
                Zone = RIO TINTO - GONDOMAR: MEDIUM (3/1)
                Zone in {OUTROS,PORTO}:
                :...Vouchers <= 5: BIG (8)
                    Vouchers > 5: TOP (21/6)
```

Figura 18 - Árvore de decisão do modelo de classificação

6.3 Avaliação de Resultados

Na classificação de clientes, uma vez que o volume de dados disponível é bastante elevado, os modelos são criados e avaliados recorrendo ao método *holdout*. Com este método são criados dois conjuntos disjuntos de dados, cerca 70% dos dados serão dados de treino e os restantes 30% dados de teste.

Ao usarem-se dois conjuntos de dados disjuntos para criação e teste dos modelos criados faz-se uma avaliação mais rigorosa e imparcial. Existem métricas que permitem avaliar a qualidade dos modelos criados como: taxa de acerto, *precision*, *recall*, *kappa* e F1 obtidas a partir da matriz de acertos/erros (matriz de confusão) do modelo. Estas métricas são abrangentes à maioria dos modelos de classificação, no entanto os modelos gerados neste projeto são multi-classificados. Por conseguinte, para uma correta avaliação, necessitam de outras métricas adicionais Micro-Média, Macro-Média, *AUC* e curva *ROC*.

Ao longo deste capítulo foram analisados três conjuntos de dados candidatos a modelos de classificação de clientes fidelidade. Apesar destes três modelos apresentarem boas métricas de avaliação, o modelo que se distinguiu entre os três foi o conjunto 3. Este modelo quando comparado com os dois restantes apresenta melhores valores de avaliação. Estes valores são apresentados seguinte tabela.

Tabela 16 – Avaliação modelo de classificação

Atributo	Valor
Tamanho	14
Erros	91 (4.0%)
Taxa de acerto	0.93
Recall	0.28
Kappa	0.92
Precision	0.125
F1	0.174
TNR	1
FPR	0
AUC	93.37
Macro Precision	0.07
Micro Precision	0.0115
Macro Recall	0.1512
Micro Recall	0.0111

Das métricas apresentadas, salientam-se os valores mais importantes como:

- Tamanho: 14 – Produz uma árvore de decisão simples. O que permite entender melhor a classificação que é feita assim como acelerar o processo de classificação de novos clientes.
- Erros: 4.0% – Dado o volume de clientes fidelidade (17,262) esta percentagem de erros é bastante aceitável, na medida em que o acerto é de 93%.

- Acerto: 93% – Esta é a métrica mais utilizada na avaliação de modelos. Significa que o classificador prevê corretamente a classificação feita 93% das vezes. Como regra, um classificador possui uma taxa de acerto aceitável na casa dos 80%.

6.4 Funcionamento

Relativamente ao modus operando da aplicação sobre os modelos de classificação, salienta-se que esta é capaz de gerar um modelo classificador de clientes fidelidade e de os classificar de uma forma autónoma com base nas suas características e comportamentos. A classificação de clientes é agora documentada no âmbito da aplicação e da interação com o utilizador.

Para tal é disponibilizado ao utilizador uma interface que permite a seleção de um período de análise, como demostra a Figura 19. Este período consiste num alcance de datas de inscrições de clientes. Como o intuito é classificar novos clientes periodicamente, estas datas fornecem ao utilizador a flexibilidade de selecionar o período de inscrições de clientes.

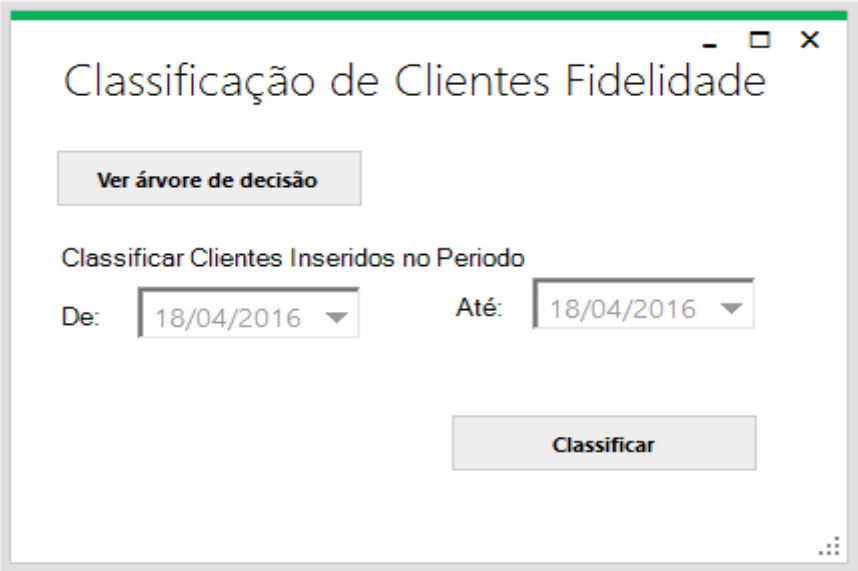


Figura 19 - Interface de classifcação de clientes

O clicar em “classificar”, faz com que a *interface* transmita um pedido á camada de gestão de dados que trate da resposta ao pedido do utilizador. Esta por sua vez, após verificar que existe na *Staging Area* todos os dados necessários, irá comunicar com a camada de inteligência, à qual solicitará que efetue a classificação de todos os clientes fidelidade inscritos no período indicado.

A camada de inteligência utilizará um conjunto de bibliotecas referentes á classificação de modelos, que por sua vez aplicam o algoritmo C5.0. Após a execução deste algoritmo a resposta é devidamente tratada e apresentada ao utilizador pela camada de *interface.*

A Figura 19 demostra ainda que é possível visualizar qual a árvore de decisão a qual o sistema terá como base para gerar a classificação de clientes fidelidade. Esta árvore foi gerada durante o processo de implementação descrito na subsecção anterior. A árvore de decisão é apresentada ao utilizador com o intuito de se apresentar ao utilizador como um fundamento à classificação que é gerada.

Após indicado o período a aplicação irá proceder à classificação dos clientes cuja data de inscrição se encontra entre o alcance especificado. É então apresentado ao utilizador uma página que começa por indicar de forma sumária quantos clientes foram classificados como pequeno, médio, grande, topo. É também apresentada de forma individual a classificação obtida para cada um dos clientes fidelidade. Este output é ilustrado pela Figura 20.

Classificação de Clientes Fidelidade

Classe	Quantidade
SMALL	520
MEDIUM	43
BIG	2
TOP	0

Cliente	Classificacao
100369790	SMALL
100370616	SMALL
102372562	MEDIUM
100483682	INACTIVE
100484980	SMALL
100486304	SMALL
1008687252	SMALL

Figura 20 - Apresentação da classificação de clientes

7 Regras de Associação

A Mass Perfumarias é uma empresa que iniciou o seu negócio há 25 anos. Com tantos anos de mercado e com boas estratégias de crescimento a empresa tem vindo a crescer todos anos. Atualmente contam já com 9 lojas físicas e uma loja *online*. Este crescimento aqui indicado é entendido como o crescimento do volume anual de vendas. Segundo a amostra fornecida para estudo, podemos constatar na Figura 21 que o volume de vendas da empresa tem aumentado. Salienta-se que a amostra de 2014 só contempla o primeiro semestre.

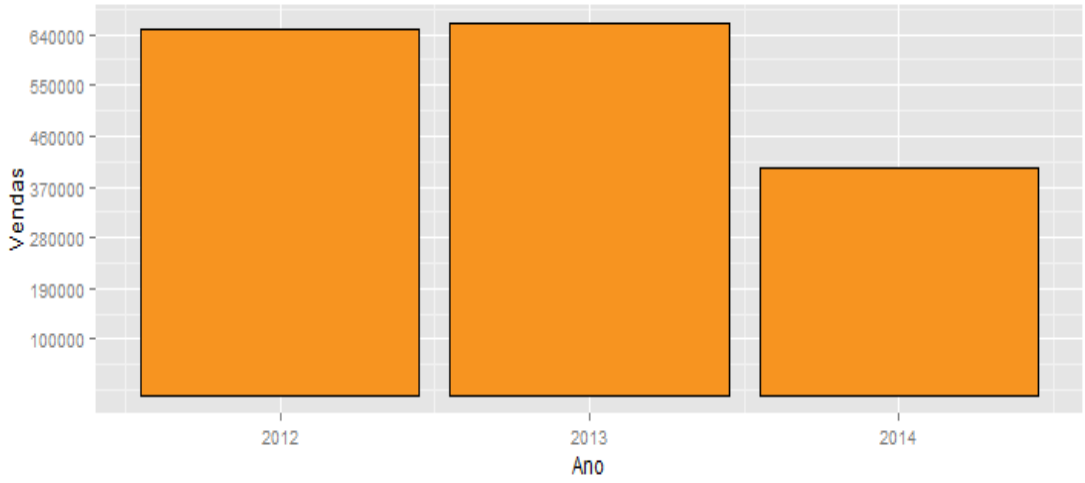


Figura 21 - Volume anual de vendas da amostra

É possível constatar também que o volume de vendas é bastante elevado. Este fator aliado ao número de lojas e ao elevado número de artigos para venda, contribuem para que seja difícil para os gestores da empresa saberem o que mais se vende conjuntamente.

É um fator estratégico muito importante para os gestores de empresas de venda de artigos ao público terem conhecimento do que mais se vende em conjunto nas suas lojas. Para tal é necessário recorrer a técnicas de Data Mining, mais especificamente a regras de associação.

As regras de associação são muito úteis na análise preventiva de comportamento. Estas permitem descobrir quais são os cestos de compras mais comuns dos seus clientes. Ao saber como são constituídos estes cestos, é possível agrupar melhor os produtos nas prateleiras, nos catálogos e até definir *layout* das lojas.

7.1 Definição

As regras de associação têm como objetivo encontrar elementos que implicam a presença de outros elementos numa transação ou seja, encontrar relações ou padrões frequentes entre conjuntos de dados. O termo transação indica quais os itens que foram consultados numa determinada operação [52].

Os autores deste conceito foram Imielinski Agrawal e Swami em 1993. A sua aplicabilidade permite extrair os mais importantes tipos de conhecimento sobre os dados, nos mais diversos tipos de mercado. Este problema foi definido originalmente como: encontrar todas as regras de associação que possuam suporte e confiança maiores ou iguais a um suporte e confiança mínimo indicados.

As regras de associação são elementos compostos por duas partes (*left hand side* (***LHS****)* e *right hand side (**RHS**)*). O ***LHS*** rep\resenta a expressão antecedente da regra e o ***RHS*** o lado consequente. Ambos lados de uma regra de associação podem ser formados por conjuntos de um ou mais itens.

A regra de associação é uma implicação na forma $LHS \Rightarrow RHS$ em que $LHS \subseteq A$ e $RHS \subseteq A$ e $LHS \cap RHS = \emptyset$. A regra $LHS \Rightarrow RHS$ ocorre no conjunto de transações ***T*** com um suporte ***sup*** se em $100 \times sup\%$ das transações em ***T*** ocorre $LHS \cup RHS$. A regra $LHS \Rightarrow RHS$ ocorre no conjunto de transações ***T*** com uma confiança ***conf*** se em $100 \times conf\%$ das transações de ***T*** em que ocorre ***LHS*** também ocorre ***RHS*** [53].

O funcionamento típico deste conceito, começa pela especificação de qual a fonte de dados onde estão todas as transações a serem analisadas. De seguida são especificados o suporte e confiança mínimos. O próximo passo consiste em criar um conjunto de itens frequentes. Estes conjuntos são compostos por itens que aparecem pelo menos tão frequentemente juntos quando o especificado no suporte. O último passo é gerar regras com base nestes conjuntos, cuja confiança seja igual ou superior ao indicado na confiança previamente definida.

O algoritmo utilizado é o *Apriori*, como detalhado na subsecção 2.4 Tecnologia Relevante.

7.2 Implementação

O número de transações realizadas por clientes da MASS Perfumarias é bastante elevado. Este facto além das estratégias comerciais da empresa, deve-se essencialmente a dois fatores: o crescimento anual do volume de vendas e o número de anos em atividade. Um número de transações muito elevado tem como consequência os tempos de resposta e a viabilidade da mineração de regras de associação. Isto deve-se ao elevado poder computacional que os algoritmos de regras de associação requerem. Esta necessidade deve-se à natureza do seu funcionamento.

Tipicamente percorrem todas as transações várias vezes. O primeiro ciclo é feito para determinar o suporte individual de cada item e verificar se cada um cumpre com o suporte mínimo indicado. Os ciclos seguintes têm um comportamento semelhante ao primeiro, mas recorrem apenas aos itens recolhidos pelo ciclo antecedente criando conjuntos maiores a cada iteração. Este processo termina quando não forem encontrados mais itens cujo suporte seja pelo menos igual ao especificado. Em suma, o número elevado de transações aplicado em algoritmos computacionalmente existentes, como neste caso, inviabiliza este processo de descoberta de regras.

Este problema aplica-se neste projeto, face à dimensão dos dados fornecidos pela empresa. A amostra disponibilizada pela empresa é composta por 1,691,251 transações distribuídas por 4 anos. Existem 8,151 transações registadas na última quinzena de 2011, 629,535 em 2012, 650,276 em 2015 e 403,289 no primeiro semestre de 2014. Este número de transações tem uma influência muito negativa nos tempos de resposta em tempo útil que a aplicação terá de ser capaz de responder. Este número dificulta também o estudo, implementação e testes que são feitos na construção desta parte do sistema.

Tendo em consideração estes factos é necessário selecionar apenas um grupo de transações, com uma dimensão mais adequada, que serão as utilizadas neste projeto. Como este projeto consiste numa aplicação que será utilizada ao longo de diversos anos e não apenas um estudo pontual, é necessário que este critério seja adequado não só à amostra, mas também a qualquer período de análise selecionado pelo utilizador. Para que seja possível satisfazer este requisito, para esta análise, serão selecionadas as transações feitas por clientes fidelidade. Este critério de seleção permite além de recolher um número de transações adequado, apresenta ainda a vantagem de selecionar transações de clientes que compram com alguma regularidade.

Tendo em conta a amostra inicial com 1,691,251 transações, após selecionadas apenas as que pertencem a clientes fidelidade, este número reduz-se para 307,969. Estas encontram-se distribuídas também por 4 anos, das quais 2028 pertencem à última quinzena de 2011, 116665 a 2012, 115660 a 2015 e 73616 ao primeiro semestre de 2014.

O mySoftmais permite criar múltiplas classificações e subclassificações de artigos. Esta funcionalidade permite descobrir relações entre artigos a diferentes níveis. No caso da MASS Perfumarias os artigos estão classificados por marca, tipo e departamento. Então, além de ser possível descobrir as principais relações entre artigos, é possível descobrir também que marcas, tipos de artigo ou departamentos se vendem mais conjuntamente.

Dado que é possível obter uma classificação de clientes, este facto é aproveitado em conjunto com as regras de associação para que seja possível apresentar ao utilizar o que determinado grupo de clientes mais compra conjuntamente.

Dadas estas possibilidades é necessário adaptar a aplicação do algoritmo *Apriori* às transações em questão. Esta adaptação é implementada segundo um algoritmo que recebe as instruções do utilizador, recolhe as transações adequadas ao pedido, aplica o *Apriori* e trata o seu *output* para obter o resultado desejado. Este algoritmo recebe dois parâmetros:

- Tipo: Indica qual o nível da transação que será aplicado (produto, marcas, tipos ou departamentos).
- Grupo: Parâmetro opcional que indica se o utilizador pretende restringir a procura a algum grupo de clientes (pequeno, médio, grande, topo).

Este algoritmo é apresentado na Figura 22. Como é possível observar este algoritmo começa por recolher as transações adequadas. Transações estas que são filtradas segundo o descrito acima nesta subsecção, com o filtro das transações de um determinado grupo caso este parâmetro seja especificado. De seguida é aplicado o algoritmo *Apriori*. Tendo todas as regras geradas por este, que são de um volume considerável, é necessário ainda recolher as melhores regras para um resultado mais fidedigno. Então, para cada artigo transacionado, seleciona-se as 3 melhores regras onde estes constam. Após esta fase ainda se obtém um número muito elevado de regras, cujos valores de confiança podem ser explorados. É aplicada uma união a todas as regras e recolhidas as N regras com maior confiança. Tendo as melhores N regras, são recolhidos todos os artigos presentes no *RHS* destas regras e este é o *output* final deste algoritmo.

Este valor de N representa um número compreendido entre 2 e 22. Os valores 2 a 22 são os extremos do número de artigos adquiridos nas transações. N toma um valor neste intervalo que maximiza o valor da medida F1.

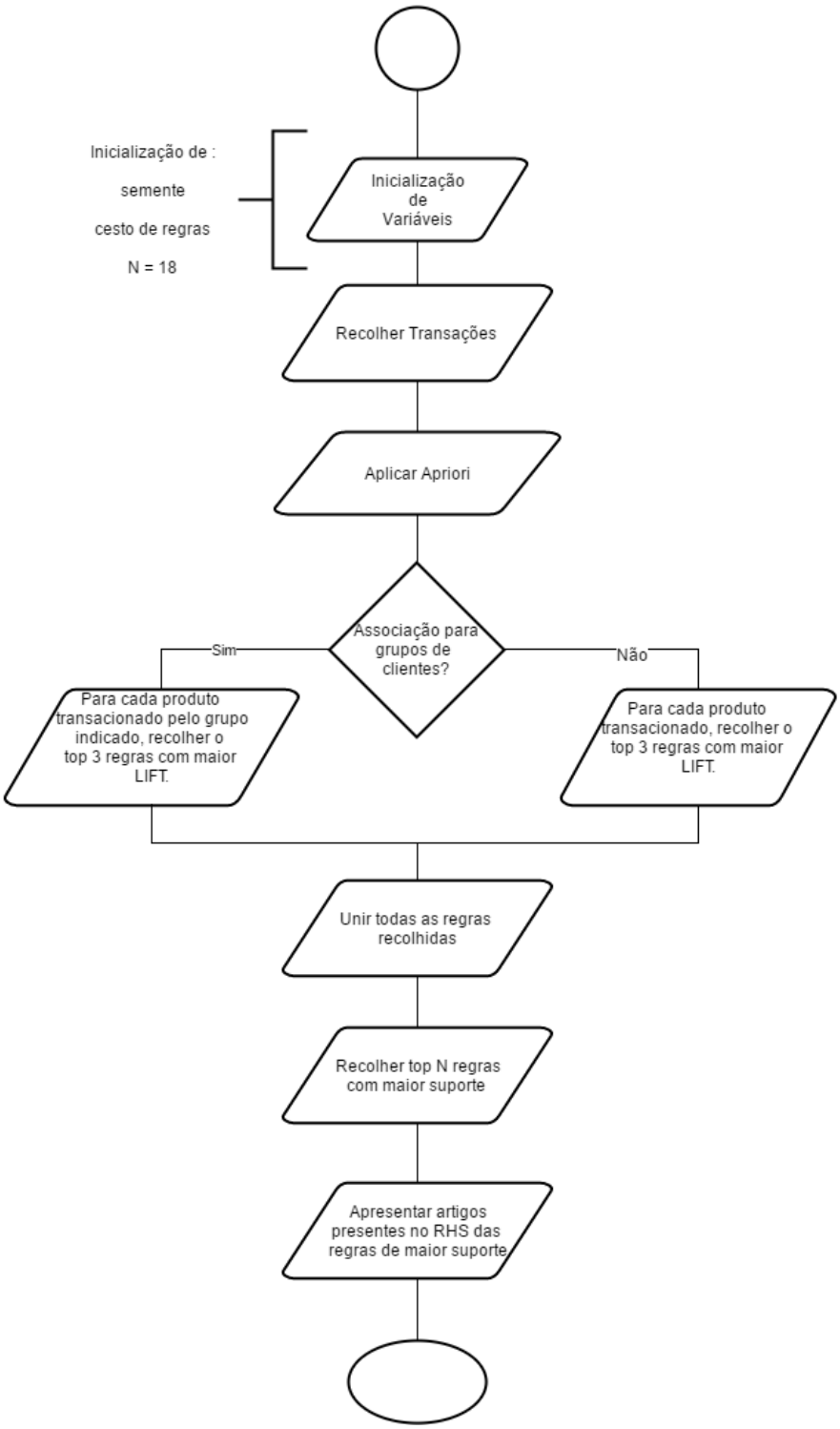


Figura 22 – Algoritmo de geração de associações entre artigos

7.3 Avaliação de Resultados

Relativamente a regras de associação existem medidas objetivas que permitem ajustar e filtrar a relevância das regras: suporte, confiança, *Lift*. Estas medidas carecem de serem reajustadas mediante o número de transações envolvidas e o número de artigos existentes. É necessário também selecionar, das regras mais relevantes, aquelas que realmente são mais objetivas tendo em conta o contexto e funcionamento da MASS Perfumarias. Em suma estas métricas permitem a definição da relevância das regras, mas a avaliação da sua qualidade ficará à responsabilidade do analista.

A amostra fornece um volume enorme de transações. Após a aplicação dos critérios de filtragem de transações descrito na subsecção 7.2 Implementação, é possível diminuir o número de transações de 1,691,251 para 307,969.

Com base nestas transações a aplicação é capaz de descortinar o que mais se vende conjuntamente (produtos, marcas, tipos ou departamento) de forma global ou para um dado grupo de clientes (pequeno, médio, grande ou topo). Para avaliar corretamente este ponto, foram geradas regras de associação entre todas as combinações entre produtos, marcas, tipos ou departamento e os grupos pequeno, médio, grande ou topo. Para cada uma das combinações foram registados os respetivos valores de suporte, confiança, *Lift* e calculado o valor médio para cada métrica. Um importante fator a ter em conta aquando da avaliação das regras além do elevado número de transações é o número de artigos transacionados. Este número de artigos é também muito elevado dado que a MASS Perfumarias disponibiliza para venda cerca de 15000 artigos.

Este volume de transações e artigos realçam a importância da métrica *Lift* nesta avaliação. O *Lift* de uma regra A ⇒ B indica o quanto mais frequente se torna B, quando A ocorre. Esta medida é calculada com a seguinte expressão: *Lift*(***LHS*** ⇒ **RHS**) = *Conf*(***LHS*** ⇒ **RHS**) ÷ *Sup*(**RHS**).

Contudo, os valores obtidos para suporte e confiança demostram ser bastante satisfatórios. Estes valores são agora apresentados na Tabela 17.

Tabela 17 – Avaliação regras de associação

Atributo	Valor
Suporte	0.36
Confiança	0.87
LIFT	1.86

7.4 Funcionamento

Relativamente a regras de associação, a aplicação é capaz de apresentar ao utilizador o que mais se vende conjuntamente ao nível de artigo, marca, tipo e departamento. Esta é também capaz de influenciar a resposta caso o utilizador pretenda ver o que mais se vende conjuntamente apenas para um determinado grupo de clientes (pequenos, médios, grandes, topo).

Para esta tarefa é apresentado ao utilizador a seguinte interface que é ilustrada na Figura 23.

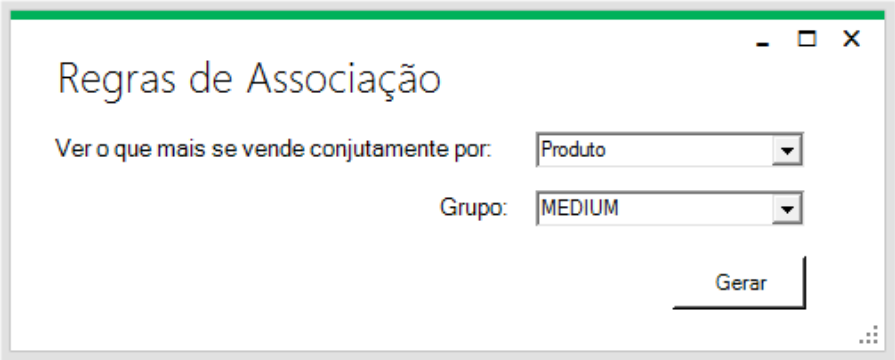


Figura 23 – Interface de regras de associação

Após indicado o que o utilizador pretende ver e com a influência ou não de algum grupo é apresentado ao utilizador o *output* da Figura 24. Este output apresenta ao utilizador os artigos com relações mais fortes ou seja, os artigos que mais frequentemente se vendem em conjunto. Estas relações mais fortes são estipuladas pela confiança, suporte e lift que cada uma das regras apresenta.

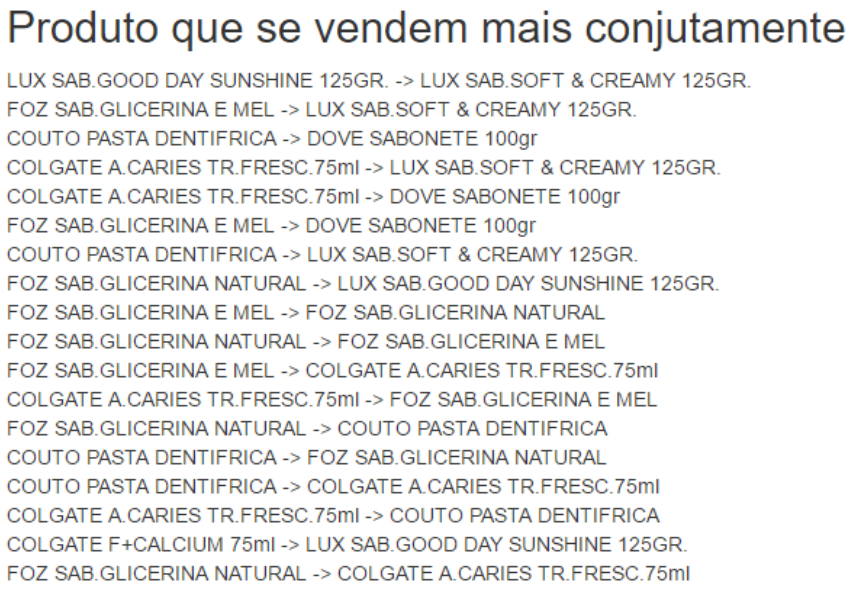


Figura 24 – *Output* das regras de associação

8 Recomendação de Produtos

Em qualquer negócio de venda ao público, principalmente na venda a retalho, existe uma grande variedade de artigos para uma grande variedade de clientes. Como tal, a tarefa de gerir o marketing da empresa ou gerir campanhas e/ou promoções é uma tarefa complexa. Campanhas e marketing bem definidas e estruturadas terão um impacto direto nas vendas da empresa. Isto é, ao conseguir direcionar determinados grupos de artigos a determinados grupos de clientes, fará com que os clientes aumentem o seu interesse em pelo menos visitar a loja.

Todas a empresas têm a necessidade de investir na divulgação. As que não o fazem correm o risco de sofrerem as consequências de uma forte concorrência ou gastar recursos em tentativas e erros de marketing. A MASS Perfumarias é uma empresa que não foge a esta regra. É uma empresa que vende a retalho, que tem para venda uma enorme variedade de artigos e uma grande carteira de clientes.

Uma divulgação eficaz requer conhecimentos específicos sobre os clientes e as suas compras. O auxílio de uma ferramenta desta natureza é essencial neste processo, devido ao elevado número de clientes, vendas e produtos disponíveis.

Com a aplicação que se projeta neste estudo é possível minimizar o esforço e o que investimos em campanhas e marketing, sendo capaz de direcionar corretamente produtos a clientes. Uma aplicação desta natureza é capaz de auxiliar os gestores a segmentar as suas campanhas da melhor forma possível. Proporciona também a vantagem de possibilitar uma análise mais concreta de resultados das suas campanhas.

Como referido anteriormente a aplicação irá permitir à Mass Perfumarias uma recomendação de artigos aos seus clientes. Esta recomendação poderá ser feita a determinados clientes especificamente, ou a grupos de clientes. Estes grupos de clientes são clientes com a mesma classificação obtida e descrita no capítulo 6 Classificação.

Como indicado na subseção 7.2 Implementação, o mySoftmais permite criar múltiplas classificações e subclassificações a artigos e estão criadas 3 classificações: marca, tipo e departamento. Esta funcionalidade permite que a recomendação possa ser realizada por cada classificação ou subclassificações atribuída aos artigos, além da recomendação individual de cada artigo. Estas classificações foram elaboradas hierarquicamente, fator que facilita a recomendação a diferentes números de artigos. Existem disponíveis aproximadamente 15,000 artigos para venda, no entanto existem apenas 276 marcas, 95 tipos e 6 departamentos. Será então possível recomendar não só produtos mas marcas, tipos de produto ou produtos de determinados departamentos.

8.1 Definição

Sistemas de recomendação é uma aplicação de *Data Mining* aplicada ao problema de relacionar clientes a artigos que estes gostariam de adquirir. A recomendação pode ser vista como uma função ***u : C x I → R*** em que ***C ∈ Clientes***, ***I ∈ Itens***, ***R ∈ Ratigns***, cujo objetivo seria responder a C1 gostará de I2? [54]

Tabela 18 - Matriz de demonstração da função: C x I → R

	I1	I2	…		In
C1	4	?	1	5	6
C2		5	7	6	
…					
Cn		3			3

Esta recomendação é feita pela produção de um conjunto de top-N artigos recomendáveis para um dado cliente. Este conjunto pode ser obtido através de três abordagens [55]:

- Filtragem colaborativa - Consiste na análise do comportamento passado de um cliente ou dos itens adquiridos anteriormente por este ou ambas as circunstâncias mas realizadas por outros clientes.
- Baseada em conteúdo - Utiliza um conjunto de características distintas de um item, com a finalidade de recomendar itens adicionais com propriedades semelhantes.
- Hibrida: Utiliza filtragem colaborativa em simultâneo com a recomendação baseada em conteúdo, dado que ambas abordagens se podem complementar.

A abordagem que será utilizada neste projeto será uma abordagem hibrida. São utilizadas as compras de outros clientes igualmente classificados como base de recomendação, assim como produtos não adquiridos de características semelhantes aos artigos adquiridos. É possível recorrer a regras de associação para desenvolver esta lista de top-N artigos recomendáveis. Para tal é preciso criar um processo com os seguintes passos [56]:

1. Gerar um conjunto de regras de associação com base em todas as transações que respeitem um dado suporte e confiança mínimos.
2. Para cada cliente a ser alvo da recomendação, deverão ser selecionados todos os artigos que estes adquiriram num período de teste.
3. Para cada um destes artigos selecionar todas as regras cujos artigos pertencem ao lado esquerdo das regras.
4. Selecionar todos os itens únicos das regras e remover os que o cliente já adquiriu anteriormente.
5. Ordenar os artigos de forma decrescente pela confiança das regras onde estes foram previstos.
6. Selecionar os N primeiros artigos de maior confiança. Serão estes N artigos recomendáveis a cada cliente.

8.2 Implementação

Os dados com os quais será feita a recomendação necessitam ainda de uma maior atenção relativamente à sua qualidade. Este maior cuidado deve-se ao facto de ser fácil gerar recomendações em demasia tendo em conta o grande número de transações, clientes e produtos da MASS Perfumarias. É necessária atenção também à geração de falsos positivos (produtos recomendados sem interesse) e de falsos negativos (produtos não recomendados mas com interesse).

Para evitar estes problemas foi necessário realizar uma filtragem de diversos critérios às transações relativamente à sua entidade e aos seus artigos. Estes critérios são:

- Transações cujas entidades sejam clientes fidelidade. Salienta-se a importância deste critério, porque garante que as transações selecionadas pertencem a clientes que compram com alguma regularidade, o que permite reduzir bastante quais os artigos com maior potencial de recomendação.
- Os clientes fidelidade incluídos nas transações têm de ter pelo menos uma compra para cada ano do período ou seja, nesta amostra só serão recolhidas as transações dos clientes fidelidade com pelo menos uma compra em 2012 outra em 2013 e outra em 2014. Este critério é um complemento ao critério anterior e segue o mesmo propósito e importância.
- Transações cujos artigos tenham sido adquiridos pelo menos 10 vezes no período de avaliação.
- Não recolher transações cujos artigos não se encontrem devidamente identificados na sua marca, tipo e departamento.
- Os artigos presentes nas transações devem ser transacionados pelo menos 1 vez por cada ano do período de avaliação.

Nestas condições é possível que a recomendação seja feita ao nível do artigo, marca ou tipo. Esta é feita a um dado cliente individual ou para um dado grupo de clientes (topo, grandes, médios ou pequenos). Estas diversas formas de recomendação proporcionam à MASS Perfumarias uma grande flexibilidade de recomendação, aproveitando assim todos os seus recursos disponíveis, ao invés de se restringir a recomendação de somente artigos a apenas um dado cliente.

Para levar a cabo a tarefa de recomendação com as exigências e possibilidades mencionadas, foi elaborado um algoritmo capaz de levar a cabo esta tarefa. Este é agora apresentado na Figura 25.

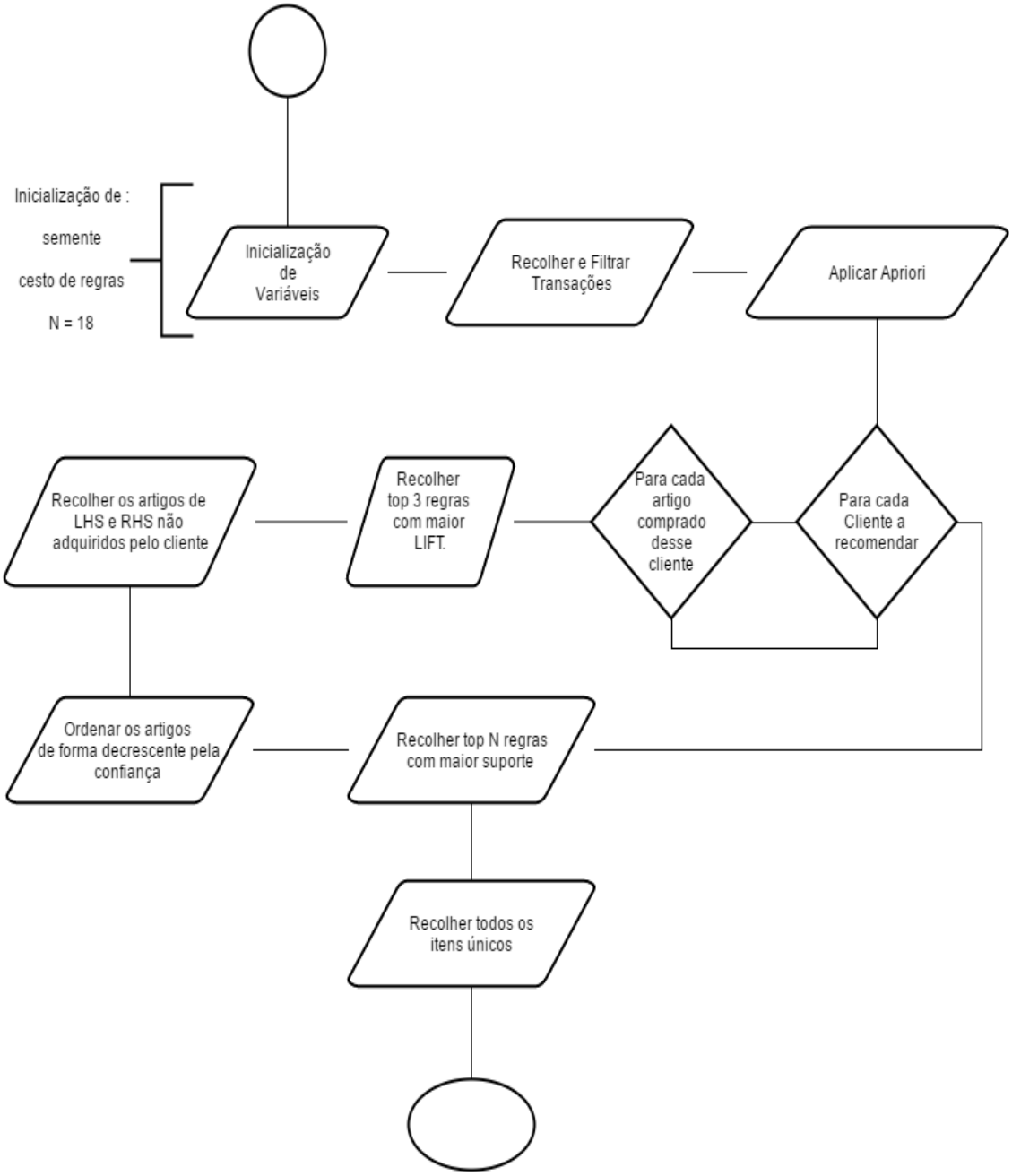


Figura 25 – Algoritmo de recomendação

Este algoritmo foi codificado em R e como ilustrado na figura anterior, este começa por recolher e filtrar as transações necessárias à recomendação. Esta filtragem segue os critérios descritos anteriormente que asseguram a qualidade da recomendação. Uma vez recolhidas as transações estas são aplicadas ao algoritmo *Apriori* cujo *output* é um conjunto enorme de regras de associação.

De seguida é aplicada uma primeira iteração para cada cliente que será alvo da recomendação, em que este poderá ser apenas um cliente como um grupo de clientes. Nesta iteração é realizada outra iteração tendo em conta cada produto adquirido pelo cliente a ser iterado. Para cada artigo iterado de cada cliente, são recolhidas as 3 regras com maior *lift* onde esse artigo aparece, quer seja no *LHS* como no *RHS*.

Depois de iterado todos os artigos adquiridos de um cliente, são recolhidos todos os artigos presentes nas regras de maior *lift* que não foram adquiridos por este cliente. Tendo um conjunto de artigos de maior relevância e não comprados anteriormente pelo cliente, são recolhidas as N regras com maior suporte onde estes artigos se encontram. Este conjunto de regras contêm os artigos a serem recomendados, em que precisam apenas de serem apresentados ao utilizador sem repetições.

Esta implementação deu origem a um artigo [57] sobre recomendação de artigos baseado no comportamento dos clientes. Este artigo pode ser consultado no Anexo I deste documento.

8.3 Avaliação de Resultados

A abordagem à avaliação de resultados da recomendação de artigos neste projeto é feita pelo confronto entre um período de testes e treino. Segundo a amostra disponível o período de treino será entre os anos 2012 e 2013 e o período de teste será o primeiro semestre de 2014. Este confronto de valores permite obter uma percentagem de acerto deste sistema de recomendação. Analisando as vendas entre 2012 e 2013 o sistema realiza um conjunto de recomendações de artigos que são comparados com as transações do primeiro semestre de 2014. Os valores que a serem confrontados são: *Precision*, *Recall e F1*.

Esta avaliação utiliza o sistema de recomendação implementado em que se recorre a uma recomendação ao nível do artigo para cada um dos grupos de clientes. Para cada cliente é calculado a *Precision* o *Recall* e *F1.*

Após iterados todos os clientes de todos os grupos são calculados os valores médios destas 3 métricas.

A avaliação feita desta forma não apresenta resultados satisfatórios.

Tabela 19 – Avaliação 1 sistema de recomendação

Atributo	Valor
Precision	0.09
Recall	0.06
F1	0.07

Estes valores devem-se a dois motivos relacionados com o modo de funcionamento da empresa, ao qual foi necessário uma profunda análise às transações em questão. Estes motivos são:

- Constatou-se que os clientes tendem a repetir 50% das suas compras. Como este sistema produz recomendações de artigos ainda não adquiridos pelos clientes, consegue acertar forçosamente em apenas metade das compras realizadas em 2014.
- A Mass Perfumarias utiliza sobre o *mySoftmais* um controlo sobre os artigos em que o mesmo artigo, com combinação de embalagens ou unidades diferentes, são considerados artigos diferentes. Por exemplo, o perfume X numa embalagem de 50ml é um artigo, o mesmo com uma embalagem de 75ml ou 100ml já é considerando um artigo distinto. Este facto causa um problema na medida em que estes artigos sendo exatamente o mesmo são considerados diferentes para o *mySoftmais* e, como consequência, para este sistema também.

  Para resolver este problema optou-se por uma metodologia de duas avaliações. A primeira consiste em avaliar como até então usando a *Precision*, *Recall* e *F1* sobre o acerto dos artigos em 2014. Na segunda, é avaliado o acerto não sobre o artigo, mas sim sobre o tipo de produto, caso não haja acerto no artigo. Esta abordagem é feita dado que 2 artigos iguais mas com caixas diferentes têm em comum o seu tipo.

Com esta nova avaliação de dois passos, as métricas de avaliação subiram para os valores da Tabela 20. Ao observar estes valores é preciso ter em consideração o facto das compras realizadas se repetem 50% das vezes.

Tabela 20 – Avaliação 2 sistema de recomendação

Atributo	Valor
Precision	0.23
Recall	0.17
F1	0.19

8.4 Funcionamento

Como descrito até agora o processo de recomendação de artigos é complexo e composto por diferentes fases. É necessário cuidado processo de ETL e AED, entender bem a complexidade e proporção dos dados disponíveis, estudar e implementar um algoritmo de recomendação adequado.

Apesar desta complexidade é necessário que a interação deste processo com o utilizador seja simples, quer ao nível dos *inputs* a especificar quer na apresentação dos resultados obtidos. Para permitir ao utilizador indicar os *inputs* pretendidos para a recomendação, é apresentado o painel ilustrado na Figura 26. Este permite ao utilizador indicar os diferentes parâmetros que pretende especificar como: o que recomendar (produto, marca, modelo ou departamento) e se o alvo será um dado cliente ou um grupo de clientes (topo, grande, médio ou baixo).

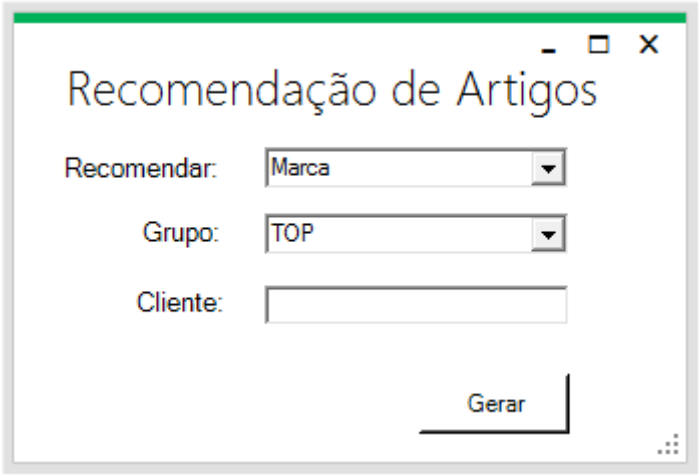


Figura 26 - Interface de recomendação

Após indicados quais os parâmetros de entrada da recomendação, estes são processados pela camada de inteligência cujo retorno são os produtos, marcas, tipos ou departamentos recomendados para um dado cliente ou um grupo de clientes. Um exemplo de um *output* apresentado é ilustrado na Figura 27.

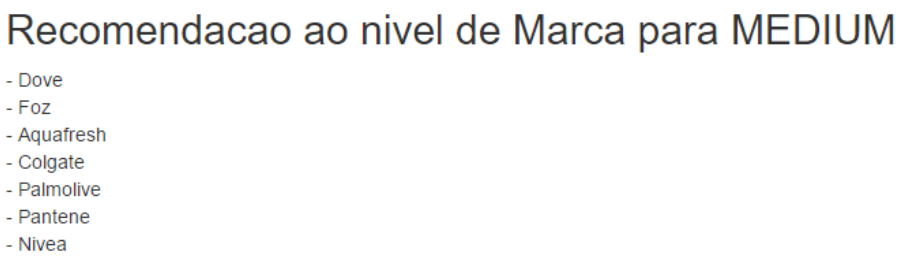


Figura 27 - *Output* de recomendação

9 Conclusões

Tal como apresentado neste documento, a solução pretendida está focada em criar um sistema de recomendação. Este documento relata o estudo e a implementação de uma aplicação que alberga funcionalidades, que permitem à MASS Perfumarias tomar melhores decisões operacionais e estratégicas, fornecendo-lhes *outputs* devidamente tratados e facilmente percetíveis, face aos *inputs* indicados. Dada a dimensão e complexidade do projeto, é de notar que existe ainda espaço para melhorias e trabalho futuro. Todavia, o projeto no seu estado atual, corresponde na totalidade aos objetivos a que se propôs.

9.1 Objetivos e Trabalho Futuro

Realça-se que o trabalho de ETL levado a cabo exigiu diversas iterações até chegar ao estado atual. Este é capaz de selecionar, transformar e carregar os dados provenientes da vasta base de dados do *mySoftmais* e assegurar a qualidade dos dados que são tratados pelos algoritmos de *Data Mining*.

Relativamente à classificação de clientes, foi proposto criar um modelo preditivo de classificação de clientes fidelidade. Este objetivo foi atingido a 100%. Este é capaz, autonomamente, de ler os clientes presentes na *Staging Area*, criar um modelo de classificação e aplicar este modelo a novos clientes fidelidade. O modelo classifica cada cliente fidelidade em topo, grande, médio ou pequeno.

Neste módulo de classificação, foram desenvolvidas duas bibliotecas capazes de simplificar e comparar modelos de classificação. Estas foram desenvolvidas tendo em conta as necessidades deste projeto, tendo ainda espaço para trabalho futuro em ambas:

- Simplificação de modelos: a simplificação de modelos é feita pela redução da quantidade de dados utilizados para treino, dado um limite mínimo de taxa de acerto. Neste módulo a geração dos classificadores é feita utilizando apenas o algoritmo *C5.0*, dado que foi o utilizado neste projeto. Pretende-se num futuro evoluir esta biblioteca para que suporte outros algoritmos de classificação como o *CART* e o *kNN*. Pretende-se também gerar os dados de testes recorrendo a outras formas além do método *holdout* como: validação cruzada, l*eave one out* ou *Bootstrap*.
- Comparação de modelos: de momento a biblioteca é capaz de receber *N* modelos de classificação e apresentar ao analista, diversas métricas de avaliação que o auxiliem a selecionar o melhor dos *N* modelos. Uma das melhorias a implementar é além de mostrar estas métricas, ser capaz de as avaliar e indicar de imediato qual dos modelos é o mais favorável. Outra melhoria será o suporte de diversos modelos de classificação, dado que atualmente suporta apenas modelos de classificação gerados pelo algoritmo C5.*0*.

Sobre as regras de associação, podemos afirmar que este objetivo também se encontra num estado que responde ao que se propôs. O proposto inicialmente era gerar as associações mais fortes entre os artigos vendidos. No entanto foi possível ir mais além. É possível gerar estas associações a diferentes níveis que não apenas ao nível do artigo como: marca, tipo e departamento. É também possível criar estas associações restringindo a pesquisa a transações feitas a grupos de utilizadores (topo, grande, médio ou pequeno).

Dado o volume enorme de transações, foi necessário estudar um conjunto de filtros a aplicar que permitem obter um grupo de transações que minimizasse o número de falsos positivos (produtos recomendados sem interesse) e de falsos negativos (produtos não recomendados mas com interesse). Um destes filtros é utilizar apenas transações de clientes fidelidade. Pretende-se melhorar este filtro de transações para que seja também possível utilizar transações de clientes não identificados. Este passo será muito desafiante na medida em que em apenas 1 ano, o número de transações destas entidades é superior a 1,000,000.

O módulo de recomendação de artigos encontra-se na totalidade concluído dado ao objetivo a que se propôs. Este sofre do mesmo problema do número de transações que as regras de associação, sendo ainda necessário aplicar outros filtros ao nível dos artigos transacionados. Em semelhança foi também possível ir mais além e aplicar a recomendação a um grupo específico de clientes (pequeno, médio, grande ou topo). É possível recomendar não só produtos, mas também marcas, modelos ou departamentos.

Este módulo poderá ser melhorado no futuro. Atualmente a recomendação é feita apenas para clientes fidelidade, mas poderá ser estendida a clientes não identificados. A recomendação teria forçosamente de ser direcionada a pequenos grupos dado o enorme número de vendas e clientes não registados. Estes grupos poderiam ser, por exemplo, divididos por zona geográfica, grupos etários, antiguidade, etc.

As análises gráficas efetuadas permitem dar a conhecer alguns dados combinados ao utilizador, para que este tenha uma melhor perceção da informação presente nos seus dados. Foram feitas diversas análises reativas a clientes, produtos e vendas. No entanto, este tipo de gráficos tem imensas aplicabilidades e perspetivas de mostrar informação. Como tal, existe sempre espaço para novos gráficos e novas perspetivas mediante o que o utilizador pretenda.

Após alguma utilização da aplicação por parte da MASS Perfumarias existirão, certamente, alguns ajustes a realizar quer ao processo de ETL quer às implementações efetuadas. Num futuro próximo e após aplicadas estratégias na empresa com base nas recomendações feitas, será importante comparar algumas métricas com o período homólogo e medir o impacto que este projeto tem no crescimento da empresa.

9.2 Apreciação Final

Este projeto teve também como finalidade colocar os intervenientes em contacto com o mundo do Data Mining e do *Business Intelligence*. Este foi desenvolvido desde o seu início, o que exigiu a aplicação de matérias lecionadas durante o Mestrado de Engenharia Informática, assim como o estudo de outras técnicas e vertentes necessárias à realização deste projeto.

Devido ao gosto de explorar novas técnicas e para tornar este projeto também numa forma de crescimento técnico, existiram inúmeras situações de aprendizagem transversais a todas as matérias envolventes a este projeto.

Com o período de tempo disponível foi conseguido um bom planeamento de uma estrutura capaz de albergar todas as funcionalidades idealizadas. Foi também concluída a aplicação com as funcionalidades planeadas, que permitem que esta aplicação seja utilizada pela empresa. Explorar os dados da MASS Perfumarias mostrou-se uma ótima experiência a nível profissional pelos conhecimentos adquiridos. Estes fatores trazem uma grande satisfação profissional e pessoal.

Referências

[1] "mysoftmais," [Online]. Available: http://www.softingal.pt/pt/?m=services&id=48. (05/07/2016)

[2] "softingal," [Online]. Available: www.softingal.pt. (05/07/2016)

[3] S. Nicola, Análise de Valor de Negócio, ISEP Instituto Superior de Engenharia do Porto: GECAD Knowledge Engineering and Decision Support Research Center.

[4] "wikipedia," [Online]. Available: https://pt.wikipedia.org/wiki/Business_Model_Canvas. (05/07/2016)

[5] S. Nicola, E. P. Ferreira e . J. J. P. Ferreira, A Quantitative Model for Decomposing & Assessing the Value for the Customer, Journal of Innovation Management, 2014.

[6] J. Han e M. Kamber, Data Mining: Concepts and Techniques, Elsevier Inc, 2006.

[7] J. R. Carvalho, "conceitos-e-tecnicas-sobre-Data Mining," [Online]. Available: http://www.devmedia.com.br/conceitos-e-tecnicas-sobre-Data Mining/19342. (05/07/2016)

[8] D. Alexander. [Online]. Available: http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/.(05/07/2016)

[9] A. d. J. D. Pacheco, L. Machado, C. F. Jung e C. S. Ten Caten, 2013 06 02. [Online]. Available: http://www.revistaespacios.com/a13v34n06/13340615.html. (05/07/2016)

[10] U. Fayyad, G. Piatetsky-Shapiro e P. Smyth, From Data Mining to Knowledge Discovery in Databases, American Association for Artificial Intelligence.

[11] F. Vessoni. [Online]. Available: https://portogente.com.br/portopedia/datamining-73758. (05/07/2016)

[12] "qizeresearch," [Online]. Available: https://qizeresearch.wordpress.com/2014/05/25/decision-tree-c5-0-example/.(05/07/2016)

[13] R. Agrawal e R. Srikant, Fast Algorithms for Mining Association Rules, IBM Almaden Research Center.

[14] M. Kuhn, Classification Using C5.0, Pfizer Global R&D, 2013.

[15] "CART," [Online]. Available: https://en.wikipedia.org/wiki/Predictive_analytics#Classification_and_regression_trees. (05/07/2016)

[16] IBM, "01.ibm," [Online]. Available: https://www-01.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.modeler.help/c50node_general.htm. (05/07/2016)

[17] R. Li, "rayli," [Online]. Available: http://rayli.net/blog/data/top-10-Data Mining-algorithms-in-plain-english/#8_kNN. (05/07/2016)

[18] F. Rodrigues, Descoberta de Conhecimento - Avaliação de Modelos, Porto: ISEP, 2014.

[19] B. D. Eugenio, On the usage of Kappa to evaluate agreement on coding tasks, Chicago: Electrical Engineering and Computer Science - University of Illinois at Chicago.

[20] D. M. Rice, "kdnuggets," Rice Analytics, St. Loui, [Online]. Available: http://www.kdnuggets.com/2010/09/pub-is-auc-the-best-measure.html. (05/07/2016)

[21] F. Rodrigues, Descoberta de Conhecimento - Regras de Associação, Departamento de Engenharia Informática (DEI/ISEP).

[22] D. S. Sayad, Association Rules, University of Toronto, 2010.

[23] G. C. Silva, Mineração de regras de associação aplicada a dados da Secretaria Minunipal de Saúde de Londrina, Universidade Federal do Rio Grande do Sul.

[24] "wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/Association_rule_learning. (05/07/2016)

[25] S. Emtiyaz e M. Keyvanpour, Customers Behavior Modeling by Semi-Supervised Learning in Customer Relationship Management.

[26] F. T. Bahari e S. M. Elayidom , An Efficient CRM-Data Mining Framework for the Prediction of Customer Behaviour.

[27] D. U. Prasad e S. Madhavi, Prediction of Churn Behavior of Bank Customers Using Data Mining Tools, Business Intelligence Journal.

[28] Y. Wang e D. S. Wu, Research of the Bank's CRM Based on Data Mining Technology, School of Economics and Business Administration, Chongqing University.

[29] C.-H. Wua, S.-C. Kaob, Y.-Y. Suc e C.-C. Wud, Targeting customers via discovery knowledge for the insurance industry, Expert Systems with Applications, 2005.

[30] O. A. KASAPOGLU e U. T. GURSOY, Data Mining and ERP: An Application in Retail Sector, Vienna: International Academic Conference, 2015.

[31] M. Dimitrijević e Z. Bošnjak, Discovering Interesting Association Rules in the Web Log Usage Data, Interdisciplinary Journal of Information, Knowledge, and Management, 2010.

[32] C. Romero, J. R. Romero, J. M. Luna e S. Ventura, Mining Rare Association Rules from e-Learning Data, Spain: Dept. of Computer Science, University of Córdoba.

[33] M. Y. Avcilar e E. Yakut, Association Rules in Data Mining: An Application on a Clothing and Accessory Specialty Store, Canadian Social Science, 2014.

[34] D. Bhanu e S. P. Madeshwari, 10. Retail Market analysis in targeting sales based on Consumer Behaviour using Fuzzy Clustering – A Rule Based Model, JOURNAL OF COMPUTING, 2009.

[35] "rdotnet," [Online]. Available: https://rdotnet.codeplex.com/.(05/07/2016)

[36] W. N. Venables e D. M. Smith, An Introduction to R, 2015.

[37] D. Eckles, M. Burke, C. Saden e S. Messing, "Udacity," [Online]. Available: https://www.udacity.com/course/data-analysis-with-r--ud651. (05/07/2016)

[38] R. D. Peng, J. Leek e B. Caffo, "Coursera," [Online]. Available: https://www.coursera.org/course/rprog. (05/07/2016)

[39] K. Wright, "Make Your Data a Strategic Asset," SAS company, [Online]. Available: http://www2.sas.com/proceedings/sugi29/103-29.pdf. (05/07/2016)

[40] "data-warehouses.net," Data Warehouses and Business Intelligence, [Online]. Available: http://data-warehouses.net/architecture/staging.html. (05/07/2016)

[41] M. Allen e D. Cervo , Master Data Management in Practice: Achieving True Customer MDM, New Jersey: John Wiley & Sons, Inc., 2011.

[42] D. Team, "dwbi.org," 31 12 2014. [Online]. Available: http://dwbi.org/etl/etl/52-why-do-we-need-staging-area-during-etl-load. (05/07/2016)

[43] "Microsoft SSIS," Microsoft, [Online]. Available: https://msdn.microsoft.com/en-us/library/ms141026.aspx. (05/07/2016)

[44] S. Inc, Clementine® Application Template for Customer Relationship Management, United States of America, 2001.

[45] “wikipedia,” [Online]. Available: https://en.wikipedia.org/wiki/Exploratory_data_analysis. (05/07/2016)

[46] S. J. Howard , Experimental Design and Analysis, 2015.

[47] “ggplot2,” Hadley Wickham, [Online]. Available: http://ggplot2.org/.

[48] R. I. Kabacoff, “statmethods,” [Online]. Available: http://www.statmethods.net/advgraphs/ggplot2.html. (05/07/2016)

[49] “AED,” [Online]. Available: https://en.wikipedia.org/wiki/Exploratory_data_analysis.

[50] “wikipedia,” Julho 2015. [Online]. Available: https://pt.wikipedia.org/wiki/Minera%C3%A7%C3%A3o_de_dados. (05/07/2016)

[51] S. C. Côrtes, R. M. Porcaro e S. Lifschitz, Mineração de Dados – Funcionalidades, Técnicas e Abordagens, PUC-RioInf.MCC, 2002.

[52] “wikipedia,” [Online]. Available: https://pt.wikipedia.org/wiki/Regras_de_associa%C3%A7%C3%A3o. (05/07/2016)

[53] Sousa R., Rodrigues F. Mining association rules with rare and frequent items, International Journal of Knowledge Engineering and Data Mining, pp. 237-247, Vol. 2, N. 4, Inderscience Publishers, 2013

[54] D. Rodriguez, Recommender Systems, University of Alcala, 2011.

[55] “wikipedia,” [Online]. Available: https://en.wikipedia.org/wiki/Recommender_system. (05/07/2016)

[56] B. Sarwar, G. Karypis,, J. Konstan e J. Riedl, Analysis of Recommendation Algorithms for E-Commerce, University of Minnesota: Department of Computer Science and Engineering.

[57] F. Rodrigues e B. Ferreira, Product recommendation based on shared customer’s, in Proceedings 11th Conference on Enterprise Information Systems, Porto, 2016.

[58] “centeris,” [Online]. Available: http://centeris.eiswatch.org/.(05/07/2016)

[59] F. S. G. d. G. Abreu, DESMISTIFICANDO O CONCEITO DE ETL, Revista de Sistemas de Informação, 2008.

[60] D. A. Keim, Visual Techniques for Exploring Databases, Institute for Computer Science, University of Halle-Wittenberg.

[61] F. Rodrigues, Descoberta de Conhecimento - Classificação, Departamento de Engenharia Informática (DEI/ISEP).

[62] J. STEFANOWSKI, Data Mining - Evaluation of Classifiers, Poland: Institute of Computing Sciences Poznan University of Technology, 2010.

[63] A. Wasilewska, APRIORI Algorithm.

[64] E. C. Gonçalves, Instituto Brasileiro de Geografia e Estatística – IBGE, [Online]. Available: http://www.devmedia.com.br/Data Mining-de-regras-de-associacao-parte-1/6533. (05/07/2016)

[65] A. Bellogín, I. Cantador, F. Díez, E. Chavarriaga e P. Castells, An Empirical Comparison of Social, Collaborative Filtering, and Hybrid Recommenders, ACM Transactions on Intelligent Systems and Technology (TIST), 2013.

[66] R. Agrawal e R. Srikant, Fast Algorithms for Mining Association Rules, 650 Harry Road, San Jose, CA 95120: IBM Almaden Research Center.

Anexos

Anexo I – Product recommendation based on shared customer's Behaviour

Conference on ENTERprise Information Systems / International Conference on Project MANagement / Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist 2016, October 5-7, 2016

Product recommendation based on shared customer's behaviour

Fátima Rodrigues[a,b], Bruno Ferreira[b,*]

[a]*GECAD, Research Group on Intelligent Engineering and Computing for Advanced Innovation and Development*
[b]*Institute of Engineering – Polytechnic of Porto (ISEP/IPP)*

Abstract

Today consumers are exposed to an increasing variety of products and information never seen before. This leads to an increasing diversity of consumers' demand, turning into a challenge for a retail store to provide the right products accordingly to customer preferences. Recommender systems are a tool to cope with this challenge, through product recommendation it is possible to fulfill customers' needs and expectations, helping maintaining loyal customers while attracting new customers. However the huge size of transactional databases typical of retail business reduces the efficiency and quality of recommendations. In this paper a hybrid recommendation system that combines content-based, collaborative filtering and data mining techniques is proposed to surpass these difficulties. The recommendation algorithm starts to obtain similar groups of customers using customer lifetime value. Next an association rule mining approach based on similar shopping baskets of customers of the same cluster, in a specific time period is implemented in order to provide more assertive and personalized customer product recommendations. The algorithm was tested with data from a chain of perfumeries. The experimental results show that the proposed algorithm when compared with a base recommendation (made solely on past purchases of customers) can increase the value of the sales without losing recommendation accuracy.

Peer-review under responsibility of SciKA - Association for Promotion and Dissemination of Scientific Knowledge.

Keywords: clustering; market basket; association rules; product recommendation

1. Introduction

Consumers are permanently involved in multi-category decision-making. In a retail context, such multi-category decision processes result in the shopping-baskets that comprise the set of items that the consumers purchase on one visit to the store. Both on-line and off-line retailers are traditionally interested in understanding the composition of their customers' market baskets, since valuable insights for designing micro-marketing and/or targeted cross-selling programs can be derived[1]. Recommender systems are technologies that assist businesses to implement such strategies. Schafer[2] presented a detailed taxonomy of recommender systems in e-commerce, and determined how they can increase the probability of cross-selling; establish customer loyalty; and fulfill customer needs by discovering products in which they may be interested. The need to reduce information overload by retrieving the most relevant information and services from a huge amount of data, and also, the development of recommendation approaches and techniques, has determined a rapid proliferation of recommender systems grouped into eight main domains[3]: e-government, e-business, e-commerce/e-shopping, e-library, e-learning, e-tourism, e-resource services and e-group activities. Recommender systems are usually classified based on how recommendations are made[4]. A content-based recommender system is based on similar items to those a

given user has liked in the past[5,6]. A collaborative filtering makes recommendations based on items owned by users whose taste is similar to those of the given user[7,8]. Combining content-based and collaborative recommendations originate hybrid approaches[9], which are commonly used, considering that both types of recommendations may complement each other.

In the recommender system here described two sources of information are used. First, it is used clustering to obtain groups of customers with similar interests based on prior purchase patterns. Second, rule association mining is performed on baskets of the same cluster, in order to derive relationships between products. Since these relationships are based on purchases of similar customers that have also purchased, in the same time period, at least one same product, it is expected to identify additional product relationships that are not captured using only the past baskets of the customer. In summary, this recommender system uses collaborative filtering combined with the ideas from content-based filtering.

There are usually quite a lot of products to be considered in a recommender system. It would be very inefficient if every product needs to be considered before making recommendations. Dimensionality reduction techniques have been incorporated to produce quickly quality recommendations for large-scale problems[10,11]. However these systems have some disadvantages, for example, require extra attributes about users or products to group the users into clusters and require the number of clusters be given in advance, which is a big burden on the user. With the proposed approach, clustering is done totally based only on derived attributes about products purchased by costumers, without the necessity of collecting extra attributes about customers and products. Besides, when selecting the baskets for recommending products, we consider only baskets of clients of the same cluster, bought in a specific time period, resulting in a much greater reduction of the number of products to consider. Due to dimensionality reduction on the number of products, the processing time for making recommendations by our approach is much reduced. Experimental results show that the proposed recommender system can enhance the recommendations with a good performance without compromising the recommendation quality.

The remainder of this paper is organized as follows: in section 2 a brief explanation of concepts and algorithms used to implement the recommender system is made. In the following section the hybrid recommendation algorithm is explained. In the next section details of the recommender system are provided. Section 6 presents the experimental results and in last section conclusions and suggestions for future work are disclosed.

2. Background

2.1. Customer Lifetime Value Analysis and RFM Evaluation

Customer lifetime value is typically used to identify profitable customers and to develop strategies to target customers. The RFM (recency, frequency and monetary) model is the most widely used model to characterize customers due to its simplicity and good predictive capabilities. “Recency” represents the time since the last purchase, a lower value corresponding to a higher probability of the customer making a repeat purchase. “Frequency” denotes the number of purchases within a specified time period; higher frequency indicates higher loyalty.

"Monetary" means the amount of money spent in this specified time period, a higher value indicating a customer that the company should focus[9]. In fact, these three variables characterize the customer in terms of his behavior and can be used as the segmenting variables by observing customers' attitudes toward the product, brand, benefit, or even loyalty.

2.2. Cluster Analysis

Cluster analysis groups data objects based only on information in data that describes the objects and their relationships. The goal is that the objects within a group be similar to one another and different from the objects in the others groups. The grouping of objects is based on a distance or similarity function, so that clusters can be formed from objects with a high similarity to each other. Several clustering algorithms have been developed, here we will use a partitional algorithm to group customers with similar lifetime value. A partition algorithm initially defines K seed points x_k, one for each cluster, and iteratively update these points to optimize some objective-function. At each iteration, each object x_i is assigned to the most similar seed point. When the attributes are real values, the seed points are referred as centroids if they represent the arithmetic mean of each cluster, or as medoids if the seed points must be objects of the data set (the closest to the clusters' centers). When the features are categorical, the seed points are designated as modes. The most known partitional clustering algorithm is the K-means[10] that (locally) minimizes the sum of squared errors between the cluster centroids and the objects in the corresponding clusters:

$$J_{KM} = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \bar{x}_k\|^2 \quad (1)$$

where $\|.\|$ represents the l2-norm. K-means takes as input the number of clusters K, and starts by randomly defining K centroids,$\{\bar{x}_1, \ldots, \bar{x}_k\}$. Then, it iterates between two steps: assigning each object to the cluster represented by the closest centroid; and updating the clusters' centroids as the means of each cluster:

$$\bar{x}_k = \frac{\sum_{x_i \in C_k} x_i}{|C_k|} \quad \forall_{k \in \{1,..,K\}} \quad (2)$$

The process repeats until no point changes clusters, or equivalently, until the centroids remain the same.

2.3. Association Rule Mining

Association rule mining is used to identify relationships among a set of items in a database. These relationships are based on the co-occurrence of the data items. Therefore, the main purpose of extracting association rules is to find out synchronous relationships by analyzing the random data and to use this data as reference during decision-making. The association rule is defined as follows[11]:

Let D be a database composed by a collection of items $A=\{a_1, a_2, ..., a_m\}$ and a group of transactions $T=\{t_1, t_2, ..., t_m\}$, where each transaction $t_i \in T$ is composed by a set of items such that each item set is a non-empty sub-item set of A, $t_i \subseteq A$. Each item set X holds in T with support Sup, if Sup% of the transactions in T contain X. The support is a measure that evaluates the statistical importance of X in the database D.

The association rule is an implication in the form $X \rightarrow Y$ in that $X,Y \subseteq A$ and $X \cap Y = \varnothing$. The rule means that if X is purchased, Y can be bought at the same time. The rule $X \rightarrow Y$ holds in T with confidence Conf, if Conf% of transactions in T that support X also support Y; i.e. Conf $(X \rightarrow Y)$ = Sup $(X \cup Y, D)$ / Sup(X,D). A high confidence ensures the predictability of the rule. The confidence measure doesn't detect independence among the items in a rule, and rules with no correlated items can have high confidence value. This happens because the confidence measures ignores the support of the item set in the rule consequent. One way to address this problem is by applying the lift metric, Lift$(X \rightarrow Y)$ = Conf$(X \rightarrow Y)$ / Sup(Y,D), which measures how many times X and Y occur together more than expected, if they were statistically independent.

In this work we will use the Apriori algorithm[11]. Even though Apriori was the first algorithm developed to extract association rules, it is still one of the most widely used algorithms. Ease of implementation, simplicity, efficiency, and empirical success are the main reasons for its popularity.

3. Hybrid Recommendation Algorithm

The cluster K to which a customer C belongs is first identified. Then, the set of all products previously purchased by customer C in a specific time period is selected - PS_{TC}. In order to avoid an explosive rule generation and get customized associations, previous to rule generation it is selected all the customers' transactions from cluster K, in the same time period forming the set of cluster transactions for that period - TS_{TK}. From TS_{TK} are eliminated the baskets with only one item, as it is not possible to generate rules from them.

With this specific TS_{TK} it is calculated the range of support values of its items, and it is selected the minimum value of the support range to be used as minimum support (Sup_{min}) in the Apriori algorithm. The goal of this Sup_{min} is to obtain all possible rules from this specific set of cluster transactions. As confidence is a measure of the rule's strength, in order to avoid items purchased together occasionally, it is defined as minimum confidence, $Conf_{min}$=100%. The Apriori association rule mining algorithm is applied to find the recommendation rules RS_{CK} relate to this subset of transactions TS_{TK}. For each product P from PS_{TC} are selected all the rules from RS_{CK} that contains in his left hand side the product P. To get the items more relate with those purchased by the customer, the RS_{CK} is sorted by the lift measure, and the top-N rules with highest lift are added to the recommendation rule set for customer C, RRS_{TC}. This specific selection, based on items a costumer has bought in the past, and on items other similar costumers of the same cluster have also bought, allows making a hybrid recommendation.

The set of candidate products for recommendation to customer C, RPS_{TC} is the set of all products of the RRS_{TC} minus previously bought products, PS_{TC} (RRS_{TC} - PS_{TC}). Previously bought products are excluded from the recommend list since the recommender is meant to broaden each customer's purchase products.

All candidate products from RPS_{TC} are sorted and ranked according to its support. The N highest ranked candidate products (top sellers) are selected as the top-N recommended products.

4. Recommender System

For the implementation of the proposed methodology, a recommender system was developed for a chain of perfumeries. The goal of the recommender system is to provide periodically relevant personalized item recommendations to loyalty customers, in order to increase the customer's interest in the stores products and consequently increase the sales volume. This company has over 30 years of existence; sells perfumery, cosmetics, make-up and body care products and early bet on customer retention through a loyalty card. The company has already 25000 customers and has more than 11000 items available for sale. The company's management is made through an Enterprise Resource Planning (ERP) system that the company uses to manage all its activities, processes and workflow. The ERP presents a very significant complexity with a database of over 300 tables, corresponding to 10GB of data for the 2012-14 period.

4.1. Data Staging Area

In order to centralize all information relating to customers into a single data source and provide quickly and accurately customer information to the recommender system, a specific data staging area was created. Due to the wide variety and heterogeneity of data in the ERP, collecting and cleaning transactional data are done through an Extraction Transformation Loading (ETL) process. As the name implies this is a three phase process involving: the extraction of data from the ERP system; data transformation such as, the treatment of inconsistencies, mapping data to a single naming convention, handling missing data and errors, integrity faults, etc.; and finally, the loading of data into the staging area. The staging area is a database built on SQL Server not standardized with configurable location, consisting of three tables: customers, products and sales. The mechanisms and routines of the ETL process and the storage management of the staging area are conducted using the SQL language and a generic programming language for performing the management of all surrounding processes.

From the data collected with the ETL process, this study will be focused on the 2012-2014 period. The recommendations will be only to loyalty customers that is, customers who made purchases on the three years 2012-2014. The first two years, 2012 and 2013, will be used as training set, and the first semester of 2014 year as test set, that is, the sales of the two first years will be used to recommend products, that will be checked with the sales of the first semester of 2014 year. The reduced dataset contains 3245 loyalty customers. There aren't any known preferences of customers, only what they bought in the last two years.

4.2. Costumer Segmentation

The system starts to perform customer segmentation based only on customer lifetime value, built based on the R–F–M customer attributes concerning the 2012-2013 purchases period. This segmentation will give groups of customers with identical behaviour shopping in terms of when they buy, how often they buy and how much they buy.

Prior to any analysis that uses distance calculations it is necessary to normalize or to standardize the features to the same scale. Some features can dominate solely because they tend to have larger values than others. Doing normalization this problem is avoided. Data scaling depends on the data distribution. As can be seen in the boxplots of RFM attributes in Fig. 1, these attributes have too many outliers, denoted as circles or dots beyond the whiskers.

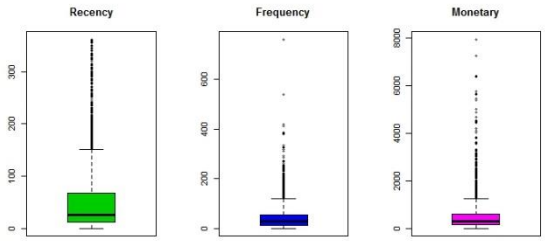


Fig. 1. Boxplot of RFM attributes.

The sigmoidal normalization is a nonlinear transformation that transforms the input data into the range -1 to 1, using a sigmoid function. It is an appropriate approach to capture the very large outlier values while mapping the input data into a range of values.

For the majority of clustering algorithms, including the k-means, it is necessary to specify the number of clusters k in advance. Various cluster evaluation measures can be used to approximately determine the correct or natural number of clusters. A given evaluation measure will work better on some datasets than others, because of that we compute two distinct indices: the total Within Sum of Squares (WSS) and the Calinski-Harabasz index[12] for values of k between 2:10. Fig. 2 shows a plot of the Calinski-Harabasz versus the number of clusters and also a plot of WSS versus the number of clusters. There is a distinct peak in the Calinski-Harabasz and a distinct knee in the WSS when the number of clusters is 3. So, for both indices the best value of k is 3, which means this dataset has 3 clusters of customers that have similar RFM behaviour.

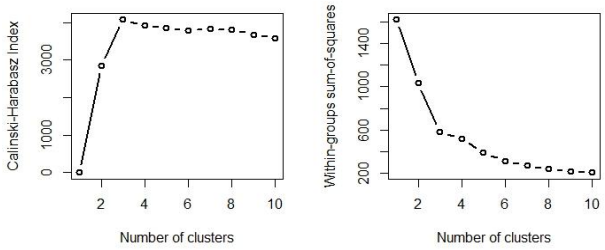


Fig. 2. WSS and Calinski-Harabasz versus number of clusters.

The k-means algorithm is not guaranteed to have a unique stopping point. K-means can be fairly unstable, in that the final clusters depend on the initial cluster centers. It is good practice to run k-means several times with different random starts, and then select the segmentation with the lowest total WSS. For that we run K-means with k=3, 25 random starts, and 100 maximum iterations per run. Table 1 describes quantitatively the three clusters obtained in terms of their average RFM values.

Table 1. Average RFM parameters for each cluster.

Cluster	Number of Customers	Recency	Frequency	Monetary
1	1663	33.51	29.42	251.63
2	842	24.38	96.51	970.99
3	740	181.99	20.33	199.45

To better analyse the clusters, we categorize the R, F, and M parameters in five categories (Very Low, Low, Medium, High, and Very High) accordingly to the quantile values of these parameters (table 2).

Table 2. RFM Categories.

	Very Low	Low	Medium	High	Very High
Recency	0-14	14-33	33-65	65-87	87-362
Frequency	4-17	17-31	31-45	45-56	56-699
Monetary	7-124	124-256	256-426	426-510	510-6165

Comparing the result of RFM parameters' values of each cluster (table 1) with categorical values (table 2), the category of the three parameters is identified to qualitatively characterize each cluster as shown in table 3.

Table 3. Clusters Characterization.

Cluster	Recency	Frequency	Monetary
1	Medium	Low	Low
2	Low	Very High	Very High
3	Very High	Low	Low

The segmentation performed using only the RFM features has discovered well-separated groups of customers with significantly different purchase attraction. Customers in cluster 1 with the pattern R (M) F (L) M (L) are customers who buy few and low value items but, because of their medium recency, they are more likely to make a repeat purchase. Cluster 2 includes high valuable and loyalty customers. They should be treated especially not to lose them. And finally, cluster 3 includes customers who very rarely visit the shops and made very few and cheap purchases. They are valueless customers that the retailer should consider whether or not to make product recommendations.

5. Personalized Product Recommendation

Recommendations should be carefully used, because its misuse, recommendations of no interest to the customer or excessive number of recommendations, could have an opposite desired effect - customer loss. Poor recommendations can cause two types of characteristic errors: false negatives, which are products that are not recommended, though the customer would like them, and false positives, which are products that are recommended, though the customer does not like them. In a recommender system, the most important errors to avoid are false positives, because these errors will lead to unhappy customers and thus they will be unlike to return to the store. Making recommendation only for customers who are likely to buy recommended products could be a solution to avoid the false positives of the poor recommendation. From the segmentation obtained it is clear distinct cluster consumption habits. So, recommendations must be specific to each one of the clusters, and due to the high number of customers in each cluster, specific target customers must be selected.

The selection of target customers for all clusters is made according to the regularity of customer's visits to the stores in the two previous years. The customers that are selected for recommendations in March month for instance are those that have made purchases in March in the last two years. Doing this it is respected the incursion pattern of the customer to the store. The recommendations will be made for periods of two months; this prevents recommending seasonal items at inappropriate times of the year. As the test set includes data for the first half of 2014, it will be evaluated three periods of recommendation for each cluster. Table 4 shows the number of target customers by cluster for each one of the three periods.

Table 4. Number of target customers by period.

Cluster\Period	Jan-Feb	Mar-Apr	May-Jun
1	460	504	555
2	589	607	631
3	68	68	88

The data considered in this study relative to the purchases of the 2012-2014 period consists in 76208 baskets, with 1373 distinct items. The company uses a three-level hierarchical item taxonomy that divides the 1373 items across 71 types and 4 categories: accessories, cosmetics, perfumery and toiletries.

Each cluster has associated a very sparse dataset with few items per basket. Table 5 presents a characterization of the clusters in terms of their baskets.

Table 5. Cluster baskets characterization.

Cluster	Number Baskets	Number Items	Type	Items/basket		
				Min	Mean	Max
1	32982	1365	70	1	1.77	15
2	36546	1371	71	1	2.46	20
3	6680	1246	70	1	1.86	16

The average item included in a single market basket is only 2.11 and the average item purchased by single customer is only 2.04. Baskets with only one item cannot be used to generate association rules therefore they will be removed.

In order to avoid an explosive rule generation and also get customized associations for each cluster, previous to rule generation it is selected the cluster transactions specific to the time period of recommendation. The Apriori association rule mining algorithm is applied to this specific-basket, with a Sup_{min} equal to the minimum support of the items of the basket, and with a $Conf_{min}$=100%. The goal is to find the maximum recommendation rules relate to this subset of transactions - RS_{CK}. Table 6 shows the characterization of the rule set for each cluster, by period of recommendation.

Table 6. Cluster Rule Set characterization by Period.

Cluster	Period	Items	Items/basket		Sup_{min}	Number Rules
			Mean	Max		
1	Jan-Feb	872	2.50	8	0.0007429	1374
1	Mar-Apr	921	2.56	8	0.000645	1358
1	May-Jun	992	2.71	13	0.0006108	2572
2	Jan-Feb	1016	3.13	13	0.000429	2468
2	Mar-Apr	1060	3.16	16	0.0003805	3211
2	May-Jun	1111	3.26	13	0.0003455	2436
3	Jan-Feb	518	2.71	11	0.00268	2072
3	Mar-Apr	567	2.67	11	0.002392	2083
3	May-Jun	590	2.76	8	0.002277	1309

For each target customer it is only selected, from the RS_{CK}, the rules that include in its left hand side at least one item from the set of all products previously purchased (PS_{TC}) by the customer. This rule set is ordered by the lift measure, and the top-N rules with highest lift are added to the recommendation rule set for customer C - RRS_{TC}. The set of candidate products for recommendation to customer C - RPS_{TC} is the set of all products of the $RRS_{TC} - PS_{TC}$, to recommend items, which the customer has not purchased before. All candidate products are sorted and ranked according to their support and then the top-N products are recommended. Due to the fact that the number of recommended products (N) has influence in the recommendation accuracies, and the baskets size have different ranges and characteristics, experiments for each one of the datasets were made with N ranging from 2 to the maximum number of items/basket in that cluster/period, to get the best number of items to recommend. Table 7 presents the best number of items to recommend by cluster/period.

Table 7. Number of items to recommend by cluster/Period.

Cluster\Period	Jan-Feb	Mar-Apr	May-Jun
1	5	6	7
2	7	4	7
3	7	2	2

6. Experimental Results

The proposed system will be used to make individual customer recommendations for specific periods of 2014 year. To measure the potential interest of customers in the recommendation performed, there are two measures extensively adopted in the related literature[13], recall and precision defined in Eqs. 3, 4, respectively.

$$recall = \frac{n(BI \cap RI)}{n(BI)} \quad (3)$$

$$precision = \frac{n(BI \cap RI)}{n(RI)} \quad (4)$$

Where BI represents all items contained in the basket bought by the customer in the specific period of 2014 year, and RI stands for the items recommended to this basket. Recall is defined as the ratio of the number of correctly recommended items (i.e., the number of items recommended really purchased by customers) to the total number of purchased items. Precision is defined as the ratio of the number of correctly recommended items to the total number of recommended items. Increasing the number of recommended items tends to increase the recall and reduce the precision. However, the F_1 metric can be used to balance the trade-off between precision and recall. F_1 metric assigns equal weight to precision and recall.

$$F_1 = \frac{2 \times recall \times precision}{recall + precision} \quad (5)$$

It is important to remember that the recommendation list, by design, will contain no products previously purchased by the customer. Once the assessment will be made with new products not previously purchased by the customer, the algorithm is evaluated by the number of new recommended items purchased by the customer, and for the recommended items not purchased it is checked whether their category match with the category of the items bought by the customer.

In order to quantify the impact of the proposed recommender algorithm, it will be used a control recommendation that works as a "placebo" such as, the list of the past items bought by the customer that we call base recommendation. Base recommendation is a recommendation made only with the past items bought by the customer, that is, for a customer C and a period p, it is recommended the products purchased by C in the same period p, of the two previous years. Another goal of our recommender system is to provide a quantification of the increase in the value of sales achieved by the suggested recommendation. Tables 8 and 9 present the evaluation metrics and the average value of sales obtained with the base recommendation and with proposed hybrid algorithm, for the three first bi-months of 2014 year.

Table 8. Evaluation metrics of Base Recommendation.

Cluster	Period	recall	precision	F1	Avg Value of Sales
1	Jan-Feb	0,204	0,138	0,165	2,323
1	Mar-Apr	0,269	0,14	0,184	2,992
1	May-Jun	0,224	0,133	0,167	2,537
2	Jan-Feb	0,232	0,189	0,208	4,827
2	Mar-Apr	0,146	0,188	0,164	3,676
2	May-Jun	0,193	0,161	0,176	4,911
3	Jan-Feb	0,236	0,12	0,159	1,935
3	Mar-Apr	0,113	0,153	0,13	1,097
3	May-Jun	0,086	0,099	0,092	0,999
Overall				0,161	2,811

As table 9 shows that in all cases the recommendation accuracy of the proposed algorithm is higher than that of base recommendation. The maximum F_1 (0,208) of the base recommendation, Cluster = 2, period Jan-Feb, is worse than that of the hybrid recommendation (0,323). The average value of the sales is also significantly larger than that of the base recommendation. The hybrid algorithm increases 96% the average value of the sales when compared with base recommendation. Meanwhile, the overall average F_1 of the hybrid recommendation (0.227) is also slightly better than that of base recommendation (0.161). Therefore, the hybrid algorithm proposed remarkably improves the average value of the sales without decreasing the recommendation accuracy.

Table 9. Evaluation metrics of Hybrid Recommendation.

Cluster	Period	recall	precision	F1	Avg Value of Sales
1	Jan-Feb	0,264	0,188	0,22	3,915
1	Mar-Apr	0,303	0,174	0,221	3,934
1	May-Jun	0,258	0,161	0,198	3,731
2	Jan-Feb	0,342	0,305	0,323	8,967
2	Mar-Apr	0,231	0,289	0,257	9,976
2	May-Jun	0,326	0,294	0,309	9,754
3	Jan-Feb	0,298	0,158	0,207	3,285
3	Mar-Apr	0,158	0,214	0,182	2,434
3	May-Jun	0,116	0,129	0,122	3,539
Overall				0,227	5,504

7. Conclusions

In this study a recommendation algorithm based on customer segmentation, followed by association rule generation to extract the best products to recommend to a target group of customers was described. The segmentation of customers based on costumer consumption behaviour through RFM attributes was adequate, since it has separated distinct groups of customers with different buying habits. Moreover, clustering customers into different groups not only improves the quality of recommendation but also allows selecting baskets of customers with similar buying habits. Also performing recommendations on specific periods of time is advantageous, because it permits to make recommendations specific to the period, which is important in a seasonal business like the perfumeries business. The experimental results show that the proposed algorithm indeed can yield recommendations of higher quality. However, evaluating the performance of a recommender system essentially requires feedback from the user. The success of the deployed system in influencing the customers can be really measured through the change in customer behaviour, such as the number of recommendations that are followed, or the change in revenue.

As future work we intend to monitor future sales and check them with our system recommendations.

Acknowledgements

The authors would like to thank *Mass Perfurmarias* for making available for this study their data sources and information about their customers, products and sales.

References

1. Mild, A., Reutterer, T. *An improved collaborative filtering approach for predicting cross-category purchases based on binary market basket data. Journal of Retailing and Consumer Services*, 10(3), pp. 123-133, 2003.
2. Schafer, J. B., Konstan, J. A., Riedl, J. *E-commerce recommendation applications*. In Applications of Data Mining to Electronic Commerce. Springer US, 115-153, 2001.
3. *Lu J., Wu D., Mao M., Wang W., Zhang G. Recommender system application developments: a survey. Decision Support Systems, 74, 12-32, 2015.*
4. *Park D. H., Kim H. K., Choi I. Y., Kim J. K. A literature review and classification of recommender systems research, Expert Systems with Applications, 39(11), 10059-10072, 2012.*
5. *Liu L., Mehandjiev N., Xu D.-L. Context similarity metric for multidimensional service recommendation, Int. J. Electron. Comm.,* 18 (1), 73–104, 2013.

6. *Huang S.-L. Designing utility-based recommender systems for e-commerce: evaluation of preference-elicitation methods Electron. Comm. Res. Appl.*, 10 398–407, 2011.
7. *Guo G., Zhang J., Thalmann D., Yorke-Smith N. Leveraging prior ratings for recommender systems in e-commerce Electron. Comm. Res. Appl., 13, 440–455, 2014.*
8. *Cui H., Zhu M. Collaboration filtering recommendation optimization with user implicit feedback J. Comput. Inf. Syst., 10 (14), 5855–5862, 2014.*
9. *Bellogin A., Cantador I., Diez F., Castells P., Chavarriaga E. An empirical comparison of social, collaborative filtering, and hybrid recommenders, ACM Transactions on Intelligent Systems and Technology (TIST) 4, 1–29, 2013.*
10. *Ba Q., Li X., Bai Z. Clustering collaborative filtering recommendation system based on SVD algorithm, IEEE Int. Conf. Software Eng. Service Sci., 963–967, 2013*
11. *Cai Y. fung Leung H. Li Q. Min H. Tang J. Li J. Typicality-based collaborative filtering recommendation, IEEE Trans. Knowl. Data Eng., 26 (3), 766–779, 2014*
12. *Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., Sriram, S. Modeling customer lifetime value. Journal of Service Research, 9(2), 139-155, 2006.*
13. *Miglautsch J. Thoughts on RFM scoring, Journal of Database Marketing 8 (1), 67–72, 2000.*
14. *Jain A. K. Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666, 2010.*
15. *Agrawal R., Imielinski T., Swami A. N. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216, 1993.*
16. *Caliński, T., & Harabasz, J. A dendrite method for cluster analysis. Communications in Statistics-theory and Methods, 3(1), 1-27, 1974.*
17. *Gunawardana, A., & Shani, G. A survey of accuracy evaluation metrics of recommendation tasks. The Journal of Machine Learning Research, 10, 2935-2962, 2009*

Anexo II – Ficheiro Excel Processo ETL: Clientes

CAMPO	ANALISAR	DESCRIÇÃO	MOTIVO	N. Reg. Válidos	N. Reg. Inválidos	% Dados Corretos
C78TERCEIRO	Sim	Campo discriminador. C - clientes; O - outros credores e devedores; F - fornecedores;	Campo fundamental na caracterização de uma entidade. Permitirá distinguir clientes de fornecedores e outros.	3154	#REF!	#REF!
C78COD	Sim	Código da entidade.	Campo que permitirá identificar unicamente um cliente. Será importante para a análise.	3154	#REF!	#REF!
C78NOME	Não	Nome da entidade.	Este campo não se apresenta como uma característica relevante de um cliente e não apresenta importância para as análises a efetuar.	----	----	----
C78OBS	Não	Observações relacionadas com a entidade.	Campo que não apresenta qualquer relevância. Apresenta-se vago, de livre utilização e de pouca utilização.	----	----	----
C78NUMCTB	Sim	Número de contribuinte.	Campo também permite identificar unicamente um cliente. Será então importante para a análise, e também pelo significado do 1º dígito deste que permitirá validar a natureza da entidade.	3151	#REF!	#REF!

C78INTRACOM	Não	Indica se a entidade é intracomunitária.	Este campo apresenta-se como um ótimo fator que permite a distinção entre clientes nacionais ou estrageiros. Mas como nunca é utilizado, e dado o domínio da empresa, todos os clientes serão considerados portugueses.	----	----	----
C78MOE	Não	Moeda de transação com a entidade.	Este campo não é relevante no sentido de que não permite distinguir a nacionalidade do cliente, e visto que todas as entidades desta tabela apresentam todas a mesma moeda.	----	----	----
C78RGLIVA	Não	Regime de IVA aplicado à entidade.	Este campo será relevante numa análise a fornecedores. Neste caso não existe a necessidade de o ter em conta, até porque penas uma das entidade se encontra taxada com taxa simplificada,	----	----	----
C78CNTBANCO	Não	Número de conta bancária.	O número da conta bancária de um cliente não se apresenta de todo relevante para as análises a efetuar.	----	----	----
C78NIB	Não	NIB da entidade.	O NIB de uma entidade não se apresenta de todo relevante	----	----	----

			para as análises a efetuar. Como tal, será descartado.			
C78DESC1	Não	Desconto aplicado à entidade aquando cliente.	O desconto será um fator importante de análise. No entanto não será tido em conta por não existirem registos com a utilização deste campo referente a clientes.	----	----	----
C78DESC2	Não	Desconto aplicado à entidade aquando cliente.	O desconto será um fator importante de análise. No entanto não será tido em conta por não existirem registos com a utilização deste campo referente a clientes.	----	----	----
C78DESCF1	Não	Desconto aplicado à entidade aquando fornecedor.	Não será analisado por ser destinado a fornecedores.	----	----	----
C78DESCF2	Não	Desconto aplicado à entidade aquando fornecedor.	Não será analisado por ser destinado a fornecedores.	----	----	----
C78DESCF3	Não	Desconto aplicado à entidade aquando fornecedor.	Não será analisado por ser destinado a fornecedores.	----	----	----
C78CNDPAG	Sim	Condições de pagamento da entidade.	A condição de pagamento apresenta-se com um fator importante numa venda, como tal, este campo será tomado em consideração.	3151	#REF!	#REF!
C78IDI	Não	Idioma da entidade.	O idioma de um cliente poderá ser determinado por campos já	----	----	----

			tidos em consideração. Todos os registos apontam para o mesmo valor "Português".			
C78IDIEXP	Não	Idioma da expedição.	Campo aplicado a fornecedores, não será tido em conta.	----	----	----
C78EXP	Não	Modo de expedição da entidade.	Campo aplicado a fornecedores, não será tido em conta.	----	----	----
C78VND	Sim	Vendedor que se relaciona com a entidade .	Quem normalmente está em contacto com o cliente, é um fator bastante relevante, como tal será tido em conta.	2726	#REF!	#REF!
C78PLFCC	Não	Plafond da conta corrente da entidade.	O campo é relevante porque pela influência em vendas. No entanto, só dois registos apresentam valores para este campo. Por tal, este não será tido em conta.	----	----	----
C78PLFLETRA	Não	Plafond de letras.	Não será analisado por ser destinado a fornecedores.	----	----	----
C78PLFENC	Não	Plafond de encomendas.	Não será analisado por ser destinado a fornecedores.	----	----	----
C78PLFTOTAL	Não	Valor total do plafond.	Este campo neste caso será igual ao C78PLFCC, visto que apenas fornecedores utilizam os restantes campos relacionados com plafond. Como o campo C78PLFCC não será tido em conta, este como	----	----	----

			consequência também não.			
C78PLFSEGURO	Não	Valor seguro do plafond.	Não será analisado por ser destinado a fornecedores.	----	----	----
C78PLFAUTO	Não	Plafond automático.	Não será analisado por ser destinado a fornecedores.	----	----	----
C78ANTIGUIDADE	Sim	Se a entidade é devedora à mais de 50 dias.	Este campo é relevante na medida que permite saber quais os clientes que são devedores.	----	----	----
C78CORTACRD	Não	Campo de controlo relacionado com a antiguidade da entidade. Simboliza se será cortado o crédito à entidade.	Campo que não apresenta relevância de análise, como tal, será descartado.	----	----	----
C78CNTCC	Não	Número da conta corrente da entidade.	Campo de controlo interno da aplicação. Por esse motivo, será descartado.	----	----	----
C78CNTTITREC	Não	Número da conta corrente da entidade para títulos recebidos.	Campo destinado a fornecedores. Por esse motivo, será descartado.	----	----	----
C78TITSACADO	Não	Número da conta corrente da entidade para títulos sacados.	Campo destinado a fornecedores. Por esse motivo, será descartado.	----	----	----
C78CNTTITDESC	Não	Número da conta corrente da entidade para títulos descontados.	Campo destinado a fornecedores. Por esse motivo, será descartado.	----	----	----
C78CNTTITREFORM	Não	Número da conta corrente da entidade para títulos reformados	Campo destinado a fornecedores. Por esse motivo, será descartado.	----	----	----

C78CNTTITEND	Não	Número da conta corrente da entidade para títulos endossados.	Campo destinado a fornecedores. Por esse motivo, será descartado.	----	----	----
C78CARTA	Não	Número da carta inicio de aviso de liquidação.	Campo destinado a fornecedores. Por esse motivo, será descartado.	----	----	----
C78CARTAFIM	Não	Número da carta fim de aviso de liquidação.	Campo destinado a fornecedores. Por esse motivo, será descartado.	----	----	----
C78ENTGESTAO	Não	Flag para saber se trata ou não filiais.	Campo destinado a fornecedores. Por esse motivo, será descartado.	----	----	----
C78CODDESPP	Não	Código da tabela de descontos de pronto pagamento.	Campo destinado a fornecedores. Por esse motivo, será descartado.	----	----	----
C78TABPRC	Não	Tabela de preços da entidade.	Este campo aparenta ser relevante na medida em que aborda preços. No entanto, os dados mostram que todos os registos apresentam o uso da mesma tabela de preços ou de nenhuma.	----	----	----
C78BANCO	Não	Campo relação com a tabela de bancos.	Campo de controlo interno da aplicação. Por esse motivo, será descartado.	----	----	----
C78BALCAO	Não	Balcão do banco com que a entidade trabalha.	Campo de controlo interno da aplicação. Por esse motivo, será descartado.	----	----	----
C78MODPAG	Sim	Modo habitual de pagamento da entidade.	Este campo será tido em consideração por se considerar que poderá ser útil em diversas análises.	2002	#REF!	#REF!

C78MODENVIO	Não	Modo de envio de mercadoria ao cliente.	Não será tido em conta este campo, porque após uma análise aos registos, todos eles apresentam o mesmo valor: "Correios"	----	----	----
C78DIASPAG	Não	Referência à altura mensal na qual a entidade habitualmente efetua pagamentos.	Campo destinado a fornecedores. Por esse motivo, será descartado.	----	----	----
C78CAE	Não	Campo refente ao CAE da entidade.	Este campo não apresenta relevância às análises em questão.	----	----	----
C78DATAINS	Sim	Data de inserção da ficha de entidade.	Este campo apresenta relevância porque permite saber quando é que o cliente foi registado no sistema.	1736	#REF!	#REF!
C78ENCSATISFEITA	Não	Campo de controlo relacionado com encomendas.	Campo de controlo interno da aplicação. Por esse motivo, será descartado.	----	----	----
C78DATAALT	Não	Data da última alteração dos dados da entidade.	Este campo não apresenta relevância às análises em questão, na medida em que saber a data da última alteração dos seus dados não é relevante.	----	----	----
C78CLICONTRATO	Não	Campo de controlo relacionado com contratos.	Campo de controlo interno da aplicação. Por esse motivo, será descartado.	----	----	----
C78TABCOMISSCLI	Não	Referência à tabela de comissões.	Este campo embora se considere importante, não apresenta registos	----	----	----

			utilizados. Como tal, não será utilizado.			
C78OK	Sim	Indica se a entidade se encontra ativa.	Este campo será fundamental para determinar quais os clientes que ainda se encontram ativos.	2878	#REF!	#REF!
C78ALERTA	Não	Definir alerta da entidade (0=branco, 1=verde, 2=amarelo, etc.).	Campo que se destina a alertas. Além dos dados mostrarem que raramente é utilizado, não produz valor analítico.	----	----	----
C78CONTENCIOSO	Não	Define se é uma entidade que se encontra em contencioso.	Este campo além de não produzir valor às análises em questão, nunca é utilizando (segundo os registos).	----	----	----
C78TIPOENT	Não	Campo descritivo de caracterização da entidade.	Campo descritivo. Como tal, será descartado.	----	----	----
C78ESTADO	Não	Estado da entidade é uma descrição livre.	Por ser uma descrição livre, apresenta-se como um campo com uma descrição que não produz valor em análise.	----	----	----
C78GUIAFACTURA	Não	Permite definir se cada guia obriga a uma fatura.	Campo destinado a fornecedores. Por esse motivo, será descartado.	----	----	----
C78REFERENCIA	Não	Define se o campo V/referência é obrigatório ou não.	Campo destinado a fornecedores. Por esse motivo, será descartado.	----	----	----
C78ENTDINHEIRO	Não	Permite definir se é uma entidade a dinheiro ou não.	Este campo nunca é utilizado, como tal, será descartado.	----	----	----

C78NATUREZA	Não	No caso de outros devedores e credores definir se é credora ou devedora a entidade.	Campo de controlo interno da aplicação. Por esse motivo, será descartado.	----	----	----
C78COBRADOR	Não	Campo referente ao cobrador à entidade.	Quem normalmente está em contacto com o cliente, além do vendedor. Como neste caso não é utilizado, não será tido em conta.	----	----	----
C78ENTAUTORIZADO	Não	S - Autorizar entregar mercadoria por nome. B - Autorizar entregar mercadoria por B.I.	Campo destinado a fornecedores. Por esse motivo, será descartado.	----	----	----
C78DESCMENOS	Não	Campo referente a desconto percentual de compromisso.	Este campo não é utilizado segundo os registos apresentados, como tal, não será utilizado.	----	----	----
C78PLFCC2	Não	Campo referente ao plafond de conta corrente.	Campo de controlo interno da aplicação. Por esse motivo, será descartado.	----	----	----
C78PLFTOTAL2	Não	Campo referente ao plafond total da entidade.	Campo de controlo interno da aplicação. Por esse motivo, será descartado.	----	----	----
C78ANTIGUIDADE2	Não	Indica se o cliente possui dividas à mais de 50 dias.	Campo de controlo interno da aplicação. Por esse motivo, será descartado.	----	----	----
C78DTLIMPLF	Não	Data limite do Plafond.	Campo de controlo interno da aplicação. Por esse motivo, será descartado.	----	----	----

C78TABGERCLI	Não	Tabela de preços geral.	Este campo não é utilizado segundo os registos apresentados, como tal, não será utilizado.	----	----	----
C78SUJENCCNT	Não	Sujeito a Encontro de Contas.	Este campo não é utilizado segundo os registos apresentados, como tal, não será utilizado.	----	----	----
C78ESTCIVIL	Não	Campo que indica o estado civil da entidade.	Este campo não é utilizado segundo os registos apresentados, como tal, não será utilizado.	----	----	----
C78CONJUGE	Não	Indica o cônjuge da entidade em questão.	Este campo não é utilizado segundo os registos apresentados, como tal, não será utilizado.	----	----	----
C78ENTEMPRE	Não	Descrição de entidade empregadora da entidade.	Este campo não é utilizado segundo os registos apresentados, como tal, não será utilizado.	----	----	----
C78FACTORING	Não	Campo para tratamento de pagamento Factoring. Este Indica a conta a ser faturado.	Este campo não possui grande valor para as pesquisas em questão. Também nunca é utilizada, como tal, será descartado.	----	----	----
C78NOVOCODIGO	Não	Descrição de outro código.	Campo de controlo interno da aplicação. Por esse motivo, será descartado.	----	----	----
C58CCONTENCIOSO	Não	Número da conta corrente da entidade para contencioso.	Campo de controlo interno da aplicação. Por esse motivo, será descartado.	----	----	----
C78CCONTENCIOSO	Não	Número da conta corrente da entidade para contencioso.	Campo de controlo interno da aplicação. Por esse motivo, será descartado.	----	----	----

C78ARMAZEM	Não	Campo que referencia o armazém por defeito da entidade.	Campo de controlo interno da aplicação. Por esse motivo, será descartado.	----	----	----
C78TRATAQUOTA	Não	Campo referente a quotas.	Este campo serve para controlo de quotas, para empresas que o façam. O que não é o caso. Como tal, este será descartado.	----	----	----
C78QUOTAPAGA	Não	Campo referente a quotas.	Este campo serve para controlo de quotas, para empresas que o façam. O que não é o caso. Como tal, este será descartado.	----	----	----
C78SNC	Não	Campo de controlo de conversão de POC para SNC	Campo de controlo interno da aplicação. Por esse motivo, será descartado.	----	----	----
C78CNTSEGURADORA	Não	Campo referente a seguradoras.	Campo de controlo interno da aplicação. Por esse motivo, será descartado.	----	----	----
C78SERIE_SEGURADORA	Não	Campo referente a seguradoras.	Campo de controlo interno da aplicação. Por esse motivo, será descartado.	----	----	----
C78CONSUMIDORFINAL	Não	Indica se a entidade é faturada como consumidor final.	Como todos os registos apresentam o mesmo valor "Não", este campo será descartado	----	----	----
C78ENTIDADEGLOBAL	Não	Campo de controlo interno.	Campo de controlo interno da aplicação. Por esse motivo, será descartado.	----	----	----
C78RIC	Não	Define se a entidade se rege pelo regime de IVA de caixa.	Este campo refere-se a cálculos contabilísticos, por esse motivo, será descartado.	----	----	----

C78IBAN	Não	Indica o IBAN do cliente.	O IBAN de um cliente não se apresenta de todo relevante para as análises a efetuar. Como tal, será descartado.	----	----	----
C78BIC	Não	Indica o BIC do cliente.	O BIC de um cliente não se apresenta de todo relevante para as análises a efetuar. Como tal, será descartado.	----	----	----

Anexo III – Ficheiro Excel Processo ETL: Artigos

CAMPO	ANALISAR	DESCRIÇÃO	MOTIVO	N. Reg. Válidos	N. Reg. Inválidos	% Dados Corretos
C58CODARTIGO	Sim	Código do artigo.	Este campo será importante para identificar unicamente cada artigo.	38619	0	100,00%
C58CODALTERNATIVO	Não	Código alternativo ao código de artigo (campo imediatamente acima).	Este campo não produz qualquer relevância pela existência do campo C58CODARTIGO, como tal, não será utilizado.	----	----	----
C58CODBARRAS	Não	Código de barras do artigo.	Este campo não produz qualquer relevância pela existência do campo C58CODARTIGO, como tal, não será utilizado.	----	----	----
C58UNI	Sim	Unidade do artigo.	Este campo serve como referência à tabela T135UNIDADE. Tabela esta que é fundamental na caracterização do produto.	----	----	----
C58LOTE	Não	Lote do artigo.	Este campo permite agrupar artigos por lote, fator que poderá ser importante apenas em análises focadas na qualidade do artigo. Como tal, não será utilizado.	----	----	----

C58STKNEGATIVO	Não	Indica se é permitido contabilizar stock negativo sobre o artigo.	Este campo relaciona-se apenas com gestão de stocks. Como tal, não será utilizado.	----	----	----
C58CODIVACOMPRA	Não	Taxa de IVA da compra do artigo.	Este campo como se relaciona com o iva taxado nas compras, não será utilizado.	----	----	----
C58CODIVAVENDA	Não	Taxa de IVA da venda do artigo.	Este campo apresenta-se relevante porque está relacionado com vendas. No entanto, será vantajoso recolher a taxa do artigo posteriormente no momento da sua venda. Como tal, não será utilizado.	----	----	----
C58PESO	Não	Peso do artigo.	Devido à tipologia de artigos em questão, este campo não é utilizado. Como tal, não será utilizado.	----	----	----
C58VOLUME	Não	Volume do artigo.	Devido à tipologia de artigos em questão, este campo não é utilizado. Como tal, não será utilizado.	----	----	----
C58OBS	Não	Observações relativas ao artigo.	Este campo não apresenta qualquer relevância. Apresenta-se	----	----	----

			vago, de livre utilização e de pouca utilização.			
C58CODPAUTAL	Não	É um código específico de alguns artigos que tem o nome de código pautal.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----
C58ESTADO	Sim	Estado do artigo	Este campo será importante para determinar se o artigo ainda é vendido.	38619	0	100,00%
C58DATAERROPMC	Não	Data de erro de ultimo preço de compra.	Campo de controlo da aplicação. Como tal, não será utilizado.	----	----	----
C58STKMINIMO	Não	Quantidade mínima de stock para o artigo.	Este campo relaciona-se apenas com gestão de stocks. Como tal, não será utilizado.	----	----	----
C58STKMAXIMO	Não	Quantidade máxima de stock para o artigo.	Este campo relaciona-se apenas com gestão de stocks. Como tal, não será utilizado.	----	----	----
C58STKREPOSICAO	Não	Quantidade de reposição de stock para o artigo.	Este campo relaciona-se apenas com gestão de stocks. Como tal, não será utilizado.	----	----	----
C58PONTOENC	Não	Ponto de stock onde é necessário realizar uma	Este campo relaciona-se apenas com gestão de stocks. Como	----	----	----

		encomenda a fornecedor.	tal, não será utilizado.			
C58DATAULTCOMPRA	Sim	Data da última compra do artigo.	Este campo apesar de estar relacionado com a gestão de stocks, será utilizado porque poderá ser útil em diversos cenários nas análises a efetuar.	8353	30266	21,63%
C58DATAULTVENDA	Sim	Data da última venda do artigo.	Este campo apesar de estar relacionado com a gestão de stocks, será utilizado porque poderá ser útil em diversos cenários nas análises a efetuar.	20871	17748	54,04%
C58QNTVALORIZAR	Não	Define quantidade a valorizar para preço médio de custo.	Este campo apresenta-se importante na medida em que permite saber a valorização do artigo mediante quantidade. No entanto nunca é utilizado. Como tal, não será utilizado.	----	----	----
C58CMPADISTKVEN	Não	Campo adicional de stock de venda.	Este campo relaciona-se apenas com gestão de stocks. Como tal, não será utilizado.	----	----	----

C58CMPADIPRCVEN	Não	Campo adicional de preço de venda.	Este campo relaciona-se apenas com gestão de stocks. Como tal, não será utilizado.	----	----	----
C58FLGABSVEN	Não	Preço x Valor (Campo Adicional).	Este campo não apresenta sempre o mesmo valor: "1". Como tal, não será utilizado.	----	----	----
C58CMPADISTKCOM	Não	Campo adicional de stock de compra.	Este campo relaciona-se apenas com gestão de stocks. Como tal, não será utilizado.	----	----	----
C58CMPADIPRCCOM	Não	Campo adicional de preço de compra.	Este campo relaciona-se com compras. Fator que não será analisado. Como tal, não será utilizado.	----	----	----
C58FLGABSCOM	Não	Preço x Valor (Campo Adicional) x Quantidade.	Este campo apresenta-se importante na medida em que permite saber a valorização do artigo mediante quantidade e o seu preço. No entanto nunca é utilizado. Como tal, não será utilizado.	----	----	----
C58MARGEMCOMERC	Não	Margem comercial do artigo.	Este campo apresenta-se importante na medida em que permite saber a margem comercial aplicada no	----	----	----

			preço, no entanto nunca é utilizado. Como tal, não será utilizado.			
C58MARGEM	Não	Margem bruta sobre o artigo.	Este campo apresenta-se importante na medida em que permite saber a margem comercial aplicada no preço, no entanto nunca é utilizado. Como tal, não será utilizado.	----	----	----
C58PRCQNT	Não	Indica se trata ou não preços por quantidade.	Este campo não apresenta sempre o mesmo valor: "Não". Como tal, não será utilizado.	----	----	----
C58UNICOMPRA	Não	Unidade de compra do artigo.	Este campo não será utilizado porque os seus registos não se encontram coerentes com os campos C58UNIVENDA e C58UNI.	----	----	----
C58UNIVENDA	Não	Unidade de venda do artigo.	Este campo não será utilizado porque os seus registos não se encontram coerentes com os campos C58COMPRA e C58UNI.	----	----	----
C58FACTORCOMPRA	Não	Facto de compra (Quantidades). 1 por defeito	Este campo relaciona-se com compras. Fator que não será analisado.	----	----	----

			Como tal, não será utilizado.			
C58FACTORVENDA	Não	Facto de venda (Quantidades). 1 por defeito	Este campo embora importante, apresenta 99% da vezes o mesmo valor: "1". Como tal, não será utilizado.	----	----	----
C58IMG	Não	Imagem do artigo.	Este campo não apresenta qualquer relevância. Como tal, não será utilizado.	----	----	----
C58PRCVOLUME	Não	Se é aplicável preços diferentes por volume de venda.	Este campo não será utilizado porque todos os seus registos apresentam o mesmo valor: "N". Como tal, não será utilizado.	----	----	----
C58DATAINS	Sim	Data de inserção do artigo no sistema.	Este campo apresenta relevância na medida que permitirá saber a data de inserção de um artigo.	36186	2433	93,70%
C58DATAALT	Não	Data de última alteração.	A data da última alteração do artigo, não é relevante para as análises a efetuar. Como tal, não será utilizado.	----	----	----
C58ARTCONTRATO	Não	Indica se o artigo está sujeito a contrato.	Este campo não é utilizado pela empresa apresentando sempre o valor	----	----	----

			de "Não". Como tal, não será utilizado.			
C58ARTCOMP	Não	Indica se é artigo composto.	Este campo não é utilizado pela empresa apresentando sempre o valor de "Não". Como tal, não será utilizado.	----	----	----
C58QNTMINFORN	Não	Quantidade mínima a encomendar ao fornecedor.	Este campo relaciona-se apenas com gestão de stocks. Como tal, não será utilizado.	----	----	----
C58QNTEMBALAGEM	Não	Quantidade do artigo por embalagem.	Este campo relaciona-se apenas com gestão de stocks. Como tal, não será utilizado.	----	----	----
C58ARTQUADRO	Não	Flag se é um artigo quadro (tratamento de artigos para GAS).	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----
C58EMBALAGEM	Não	Qual é a tabela das embalagens?	Este campo apesar de ser muito importante apresenta sempre o mesmo valor: "G". Como tal, não será utilizado.	----	----	----

C58TIPOEMB	Não	Tipo de embalagem.	Este campo apesar de ser muito importante apresenta sempre o mesmo valor: "G". Como tal, não será utilizado.	----	----	----
C58RESERVA	Não	Indica se é permitido efetuar reservas do artigo.	Este campo relaciona-se apenas com gestão de stocks. Como tal, não será utilizado.	----	----	----
C58UNIDVAR	Não	Indica se a unidade é variável.	Este campo não é utilizado pela empresa. Como tal, não será utilizado.	----	----	----
C58PORENC	Não	Indica se o artigo não existir em stock, se estará disponível por encomenda.	Este campo relaciona-se apenas com gestão de stocks. Como tal, não será utilizado.	----	----	----
C58CMPADICIONAL	Não	Campo adicional.	Este campo relaciona-se apenas com gestão de stocks. Como tal, não será utilizado.	----	----	----
C58ARTFOR	Não	Indica se existe controlo por fornecedor.	Este campo relaciona-se apenas com gestão de stocks. Como tal, não será utilizado.	----	----	----
C58DIASVALIDADE	Não	Dias de validade do artigo.	Este campo relaciona-se apenas com gestão de stocks. Como	----	----	----

			tal, não será utilizado.			
C58SECTOR	Não	Setor do artigo	Este campo relaciona-se apenas com gestão de stocks. Como tal, não será utilizado.	----	----	----
C58MULPESO	Não	Indica se o artigo está disponível em diversos pesos	Este campo relaciona-se apenas com gestão de stocks. Como tal, não será utilizado.	----	----	----
C58QNTCAIXA	Sim	Quantidade do artigo por caixa.	Este campo será muito importante, porque permitirá determinar a influencia que terá o número de artigos por caixa nas vendas.	36186	2433	93,70%
C58SAZONAL	Não	Sazonalidade do artigo.	Este campo apesar de importante, não existem registos com a sua utilização. Como tal, não será utilizado.	----	----	----
C58OPEINS	Não	Operador que inseriu o artigo.	Este campo não apresenta qualquer relevância. Como tal, não será utilizado.	----	----	----
C58OPEALT	Não	Operador que alterou o artigo pela última vez.	Este campo não apresenta qualquer relevância.		----	----

			Como tal, não será utilizado.			
C58PRCINICIAL	Sim	Preço inicial do artigo.	Este campo será utilizado porque permite saber qual o seu preço inicial e o impacto que poderá ter em diversos cenários.	26532	12087	68,70%
C58UNID2	Não	Unidade alternativa do artigo.	Não existem registos com a sua utilização deste campo. Como tal, não será utilizado.	----	----	----
C58UNIETIQ	Não	Unidade de etiqueta.	Não existem registos com a sua utilização deste campo. Como tal, não será utilizado.	----	----	----
C58PRCLOTE	Não	Indica se o preço será distinto mediante o lote.	Este campo seria relevante se fosse utilizado pela empresa, o que não acontece. Como tal, não será utilizado.	----	----	----
C58LOJA	Não	Indica se é restrito a determinadas lojas.	Este campo seria relevante se fosse utilizado pela empresa, o que não acontece. Como tal, não será utilizado.	----	----	----
C58ECOVALOR	Não	Campo de valor do Ecovalor. (taxa para alguns artigos)	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como		----	----

			tal, não será utilizado.			
C58DATAMERGE	Não	Data do último merge de cliente.	Campo de controlo da aplicação. Como tal, não será utilizado.	----	----	----
C58COR	Não	Cor do artigo.	Este campo apresentar de importante, não existem registos com a sua utilização. Como tal, não será utilizado.	----	----	----
C58TAMANHOS	Não	Referência à tabela de tamanhos	Este campo apresentar de importante, não existem registos com a sua utilização. Como tal, não será utilizado.	----	----	----
C58CODGERAL	Não	Código geral do artigo.	Campo de controlo da aplicação. Como tal, não será utilizado.	----	----	----
C58AUTORES	Não	Referência à tabela de autores	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----
C58SNC	Não	Campo de controlo de conversão de POC para SNC	Campo de controlo da aplicação. Como tal, não será utilizado.	----	----	----
C58LARGURA	Não	Largura do artigo.	Este campo não é utilizado pela empresa nem aplicável à	----	----	----

			tipologia de artigos em questão. Como tal, não será utilizado.			
C58ESPESSURA	Não	Espessura do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----
C58TIPOMATERIAL	Não	Tipo de material do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----
C58REFERENCIA	Não	Referência do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----
C58PESOLINEAR	Não	Valor da medida de preço linear do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----
C58FERRAMENTA	Não	Ferramenta de manuseio do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como	----	----	----

			tal, não será utilizado.			
C58NORMAS	Não	Indica as normas do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----
C58CINTAS	Não	Indica as cintas do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----
C58TUBOS_ATADO	Não	Indica tubos atado do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----
C58MT_TUBO	Não	Indica os metros de tubo do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----
C58KG_METRO	Não	Indica o preço por quilograma ou metro do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como	----	----	----

			tal, não será utilizado.			
C58SECCAO	Não	Indica a secção do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----
C58COMPRIMENTO	Não	Indica o comprimento do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.		----	----
C58DENSIDADE	Não	Indica a densidade do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----
C58UNI_ATADO	Não	Indica a unidade atado do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----
C58FORMA	Não	Indica a forma do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----

C58LADO_A	Não	Indica preço para determinadas particularidades do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----
C58LADO_B	Não	Indica preço para determinadas particularidades do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----
C58VELOC_PROD	Não	Indica velocidade de produção do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----
C58CONDFORN	Não	Indica as condições de fornecimento do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----
C58MATERIAL	Não	Indica o material do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----

C58TIRA	Não	Indica a tira do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----
C58TOLERANCIA	Não	Indica tolerância do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----
C58RUGOSIDADE	Não	Indica rugosidade do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----
C58RAIOARESTA	Não	Indica o raio da aresta do artigo.	Este campo não é utilizado pela empresa nem aplicável à tipologia de artigos em questão. Como tal, não será utilizado.	----	----	----

Anexo IV – Ficheiro Excel Processo ETL: Vendas

CAMPO	ANALISAR	DESCRIÇÃO	MOTIVO	N. Reg. Válidos	N. Reg. Inválidos	% Dados Corretos
C57CODDOC	Sim	Código do documento. Pertence à indentação de um documento.	Este campo é bastante importante porque faz parte da composição da identificação de um documento, como tal, será utilizado.	3565904	0	100,00%
C57SERIE	Sim	Série do documento. Pertence à indentação de um documento.	Este campo é bastante importante porque faz parte da composição da identificação de um documento, como tal, será utilizado.	3565904	0	100,00%
C57ANO	Sim	Ano do documento. Pertence à indentação de um documento.	Este campo é bastante importante porque faz parte da composição da identificação de um documento, como tal, será utilizado.	3565904	0	100,00%
C57NUMDOC	Sim	Número do documento. Pertence à indentação de um documento.	Este campo é bastante importante porque faz parte da composição da identificação de um documento, como tal, será utilizado.	3565904	0	100,00%

C57LINHA	Não	Linha do documento.	Este campo não apresenta relevância por servir apenas para enumerar as linhas de um documento.	----	----	----
C57CODDOCORI	Não	Código do documento origem.	Este campo não apresenta relevância porque serve de mecanismo para a aplicação relacionar documentos. Como tal, não será utilizado.	----	----	----
C57SERIEORI	Não	Série do documento origem.	Este campo não apresenta relevância porque serve de mecanismo para a aplicação relacionar documentos. Como tal, não será utilizado.	----	----	----
C57ANOORI	Não	Ano do documento origem.	Este campo não apresenta relevância porque serve de mecanismo para a aplicação relacionar documentos. Como tal, não será utilizado.	----	----	----
C57NUMDOCORI	Não	Número do documento origem.	Este campo não apresenta relevância porque serve de mecanismo para a aplicação relacionar	----	----	----

			documentos. Como tal, não será utilizado.			
C57LINHAORI	Não	Linha origem do documento origem.	Este campo não apresenta relevância porque serve de mecanismo para a aplicação relacionar documentos. Como tal, não será utilizado.	----	----	----
C57FLGDEBCRD	Não	Indica se a linha é a Débito ou a Crédito	Este campo não apresenta relevância porque serve de mecanismo para a aplicação controlar entradas ou saídas de dinheiro.	----	----	----
C57ARMAZEM	Não	Qual o armazém a que se destina o artigo.	Este campo não apresenta relevância porque não se relaciona com vendas, mas sim com outros tipos de documentos (ex.: Transferências, Guias). Como tal, não será utilizado.	----	----	----
C57MODOFACT	Não	Modo de faturação.	Este campo não apresenta relevância por estar relacionado com faturação. Como tal, não será utilizado.	----	----	----

C57INTRODUCAO	Não	Método de introdução (Artigo, Descrição, Rubrica ou Texto).	Este campo não apresenta relevância por ser apenas de controlo interno da aplicação. Como tal, não será utilizado.	----	----	----
C57ARTIGO	Sim	Indica qual o artigo em questão a ser processado na linha do documento.	Este campo é bastante importante porque indica qual o artigo a ser processado. Este fator é essencial para conhecer quais os artigos que são vendidos, como tal, será utilizado.	3565904	0	100,00%
C57RUBRICA	Não	Indica qual a rúbrica a ser processada na linha.	Este campo não apresenta relevância, por nunca ser aplicado nos casos em análise. Como tal, não será utilizado.	----	----	----
C57DSC	Não	Descritivo do artigo.	Este campo não apresenta relevância por ser representar o descritivo de um artigo. Como tal, não será utilizado.	3565904	0	100,00%
C57CODIVA	Sim	Indica qual a taxa de IVA aplicada à linha do documento.	Este campo poderá apresentar relevância, por ter influência direta nos preços dos artigos. Como	3565904	0	100,00%

			tal, será utilizado.			
C57UNI	Sim	Unidade do artigo. Ex: 1 Unidade, 1 Caixa, etc.	Este campo poderá apresentar relevância, por ter influência direta nos preços dos artigos. Como tal, será utilizado.	3565904	0	100,00%
C57FACTORCONV	Sim	Fator de conversão da unidade.	Este campo poderá apresentar relevância, por ter influência direta nos preços dos artigos, por refletir o fator da unidade. Como tal, será utilizado.	3565904	0	100,00%
C57QNT	Sim	Quantidade do artigo a ser processado.	Este campo poderá apresentar relevância, por ter influência direta nos preços dos artigos. Como tal, será utilizado.	3565904	0	100,00%
C57PRCUNT	Sim	Preço do artigo a ser processado.	Este campo poderá apresentar grande relevância, por indicar qual o preço dos artigos. Como tal, será utilizado.	3565904	0	100,00%

C57PERDESC1	Sim	Percentagem de desconto na linha do documento.	Este campo poderá apresentar relevância, por ter influência direta nos preços dos artigos. Como tal, será utilizado.	3565904	0	100,00%
C57PERDESC2	Sim	Percentagem de desconto na linha do documento.	Este campo poderá apresentar relevância, por ter influência direta nos preços dos artigos. Como tal, será utilizado.	3565904	0	100,00%
C57VALDESC1	Sim	Valor de desconto na linha do documento.	Este campo poderá apresentar relevância, por ter influência direta nos preços dos artigos. Como tal, será utilizado.	3565904	0	100,00%
C57VALDESC2	Sim	Valor de desconto na linha do documento.	Este campo poderá apresentar relevância, por ter influência direta nos preços dos artigos. Como tal, será utilizado.	3565904	0	100,00%
C57LOTE	Não	Indica qual o lote do artigo a ser processado na linha.	Este campo não apresenta relevância por apenas indicar qual o lote do artigo. Este facto não afeta diretamente uma venda.	----	----	----

			Como tal, não será utilizado.			
C57PERCOMVND	Não	Percentagem de comissão de vendedor por linha.	Este campo apresenta relevância por influenciar um vendedor numa venda. No entanto os registos apresentam sempre o mesmo valor, 2%. Como tal, não será utilizado.	----	----	----
C57PMC	Sim	Preço médio de custo.	Este campo apresenta relevância aquando da necessidade de analisar artigos, por indicar qual o seu preço médio de custo. Como tal, será utilizado.	3565904	0	100,00%
C57PMD	Sim	Preço médio ponderado	Este campo apresenta relevância aquando da necessidade de analisar artigos, por indicar qual o seu preço médio ponderado. Como tal, será utilizado.	3565904	0	100,00%
C57UPC	Sim	Ultimo preço de compra	Este campo apresenta relevância aquando da	3565904	0	100,00%

			necessidade de analisar artigos, por indicar qual o seu ultimo preço de compra. Como tal, será utilizado.			
C57CUSTOSINDPMC	Sim	Custo indireto imputado ao preço médio de custo.	Este campo apresenta relevância aquando da necessidade de analisar artigos, por indicar qual o seu custo indireto imputado ao preço médio de custo. Como tal, será utilizado.	3565904	0	100,00%
C57CUSTOSINDPMD	Sim	Custo indireto imputado ao preço médio ponderado.	Este campo apresenta relevância aquando da necessidade de analisar artigos, por indicar qual o seu custo indireto imputado ao preço médio ponderado. Como tal, será utilizado.	3565904	0	100,00%
C57CUSTOSINDUPC	Sim	Custo indireto imputado ao último preço de compra.	Este campo apresentar relevância aquando da necessidade de analisar artigos, por indicar qual o seu custo indireto	3565904	0	100,00%

			imputado ao ultimo preço de compra. Como tal, será utilizado.			
C57QNTEXISTENTE	Não	Quantidade existente em stock do artigo no momento do processamento da linha.	Este campo não apresenta relevância por estar relacionado com a gestão de stocks. Como tal, não será utilizado.	----	----	----
C57DATAULTCMPVND	Não	Data da última compra ou venda dependendo do documento.	Este campo não apresenta relevância por estar relacionado com a gestão de stocks. Como tal, não será utilizado.	----	----	----
C57DATAENTREGA	Não	Data da entrega do artigo.	Este campo não apresenta relevância porque não se relaciona com vendas, mas sim com outros tipos de documentos (ex.: Encomendas). Como tal, não será utilizado.	----	----	----
C57QNTANULADA	Não	Quantidade anulada da linha em processamento.	Este campo não apresenta relevância, porque serve para anular uma quantidade da linha. Como tal, não será utilizado.	----	----	----

C57ARMAZEMDES	Não	Qual o armazém a que se destina o artigo.	Este campo não apresenta relevância porque não se relaciona com vendas, mas sim com outros tipos de documentos (ex.: Transferências, Guias). Como tal, não será utilizado.	----	----	----
C57LOCALDES	Não	Indica qual o destino para o artigo.	Este campo não apresenta relevância porque não se relaciona com vendas, mas sim com outros tipos de documentos (ex.: Transferências, Guias). Como tal, não será utilizado.	----	----	----
C57FLGENTSAI	Não	Indica o tipo de movimentação de stock (entrada ou saída).	Este campo não apresenta relevância por estar relacionado com a gestão de stocks. Como tal, não será utilizado.	----	----	----
C57LINHAREAL	Não	Campo de controlo interno da aplicação	Este campo não apresenta relevância por ser controlo interno da aplicação. Como tal, não será utilizado.	----	----	----
C57DOCEMUSO	Não	Indica se o documento está em uso	Este campo não apresenta relevância por ser controlo	----	----	----

		por outro utilizador.	interno da aplicação. Como tal, não será utilizado.			
C57TABORIGEM	Não	Separador origem da linha no documento.	Este campo não apresenta relevância por ser controlo interno da aplicação. Como tal, não será utilizado.	----	----	----
C57LINHASATISFEITA	Não	Indica se a linha já se encontra satisfeita.	Este campo não apresenta relevância por ser controlo interno da aplicação. Como tal, não será utilizado.	----	----	----
C57NIVELLINHA	Não	Nível da linha.	Este campo não apresenta relevância por utilizar-se por exemplo em artigos compostos para definir o nível na arvore do produto. Como tal, não será utilizado.	----	----	----
C57ABATESTOCK	Não	Indica se abate stock ou não.	Este campo não apresenta relevância por estar relacionado com a gestão de stocks. Como tal, não será utilizado.	----	----	----
C57PRECOGRUPO	Não	Indica se trata preço de grupo ou não.	Este campo não apresenta relevância por ser controlo interno da aplicação.	----	----	----

			Como tal, não será utilizado.			
C57CMPAD1	Não	Campo adicional de texto.	Este campo não apresenta relevância por ser controlo interno da aplicação. Como tal, não será utilizado.	----	----	----
C57PENDENTE	Não	Indica se a linha se encontra pendente.	Este campo não apresenta relevância por ser controlo interno da aplicação. Como tal, não será utilizado.	----	----	----
C57EMBALAGEM	Sim	Embalagem do artigo. Ex: Grande, Pequena, etc.	Este campo poderá apresentar relevância, por ter influência direta nos preços dos artigos. Como tal, será utilizado.	3565904	0	100,00%
C57FACTOREMB	Sim	Fator de conversão da embalagem.	Este campo poderá apresentar relevância, por ter influência direta no preços dos artigos, por refletir o fator da embalagem. Como tal, será utilizado.	3565904	0	100,00%
C57REFDOCORI	Não	Referencia a um documento de origem.	Este campo não apresenta relevância por ser controlo interno da aplicação.	----	----	----

			Como tal, não será utilizado.			
C57OBSLINHA	Não	Observações da linha.	Este campo não apresenta relevância por representar o descritivo de um artigo. Como tal, não será utilizado.	----	----	----
C57PKILO	Não	Preço por quilo.	Este campo não apresenta relevância por não se aplicar ao tipo de produtos comercializados pela empresa. Como tal, não será utilizado.	----	----	----
C57PVOLUME	Não	Preço por volume.	Este campo não apresenta relevância por não se aplicar ao tipo de produtos comercializados pela empresa. Como tal, não será utilizado.	----	----	----
C57CUSTO	Não	Campo utilizado para reajuste de custo do artigo nas compras.	Este campo não apresenta relevância porque não se relaciona com vendas, mas sim com outros tipos de documentos. Como tal, não será utilizado.	----	----	----

C57COMPRIMENTO	Não	Comprimento.	Este campo não apresenta relevância por não se aplicar ao tipo de produtos comercializados pela empresa. Como tal, não será utilizado.	----	----	----
C57LARGURA	Não	Largura.	Este campo não apresenta relevância por não se aplicar ao tipo de produtos comercializados pela empresa. Como tal, não será utilizado.	----	----	----
C57VALOROFERTA	Não	Valor de oferta do artigo	Este campo apresenta relevância mas não será extraído. Será apenas utilizado para filtrar as ofertas de todas as vendas efetuadas.	----	----	----
C57VALORIVAOFERTA	Não	Valor de oferta	Este campo poderá apresentar relevância, por ter influência direta nos preços dos artigos. No entanto, este campo nunca é utilizado. Como tal, não será utilizado.	----	----	----

C57VALORANULADO	Não	Valor anulado.	Este campo não apresenta relevância, porque não se relaciona com vendas, mas sim com outros tipos de documentos. Como tal, não será utilizado.	----	----	----
C57ECOVALOR	Não	Eco valor.	Este campo não apresenta relevância por não se aplicar ao tipo de produtos comercializados pela empresa. Como tal, não será utilizado.	----	----	----
C57VALDESC3	Não	Valor de desconto na linha do documento.	Este campo poderá apresentar relevância, por ter influência direta nos preços dos artigos. No entanto, este campo nunca é utilizado. Como tal, não será utilizado.	----	----	----
C57PERDESC3	Não	Percentagem de desconto na linha do documento.	Este campo poderá apresentar relevância, por ter influência direta nos preços dos artigos. No entanto, este campo nunca é utilizado. Como tal, não será utilizado.	----	----	----

C57QNTENVIADA	Não	Quantidade enviada.	Este campo não apresenta relevância, porque não se relaciona com vendas, mas sim com outros tipos de documentos (ex.: Transferências, Guias). Como tal, não será utilizado.	----	----	----
C57SNC	Não	Campo de controlo de conversão de POC para SNC	Campo de controlo interno da aplicação. Por esse motivo, será descartado.	----	----	----
C57SEQ	Não	Sequência.	Este campo não apresenta relevância por ser controlo interno da aplicação. Como tal, não será utilizado.	----	----	----
C57CODMOTIVOIVA	Não	Código do motivo do IVA	Este campo não apresenta relevância por ser aplicado apenas a artigos isentos de IVA. Este tipo de artigos não são comercializados pela empresa. Como tal, não será utilizado.	----	----	----