INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




Exploração de Dados usando Técnicas de Data Mining na Indústria de Retalho de Moda - Caso Estudo Parfois

FREDERICO MIGUEL DE LACERDA BARRIO RIBEIRO DE ALMEIDA
Outubro de 2018

POLITÉCNICO
DO PORTO

Instituto Superior de Engenharia do Porto

Exploração de Dados usando Técnicas de Data Mining na Indústria de Retalho de Moda

Caso de Estudo Parfois

Frederico Miguel de Lacerda Bárrio Ribeiro de Almeida

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
Coorientador: Doutor Carlos Manuel Abreu Gomes Ferreira
Supervisor: Xan Salgado

Porto, Outubro de 2018

"Aqueles que passam por nós, não vão sós, não nos deixam sós.
Deixam um pouco de si, levam um pouco de nós."

Antoine de Saint-Exupéry

Resumo

Com forte crescimento e expansão, a Parfois tem vindo a apostar nos Sistemas de Informação, e no valor acrescentado dos dados, sendo este um forte aliado para a análise e compreensão do seu negócio. Contando já com alguns anos no ativo, o E-Commerce da Parfois tem vindo a evoluir bastante e o seu volume de negócios não tem parado de aumentar acompanhando assim o igual crescimento da Parfois. Como tal, em igual proporção os dados existentes têm vindo a aumentar, tornando assim vasta a quantidade com possibilidade de exploração.

Atualmente a Parfois tem como forte aposta as vendas do E-Commerce, no entanto o desconhecimento do perfil de cliente, invalida a possibilidade de alargar a sua componente de Marketing e Vendas, assim como obter vantagem competitiva.

Este projeto pretende alterar esse paradigma na Parfois E-Commerce, desenvolvendo assim uma plataforma denominada "Parfois Web Client Analytics", que irá auxiliar o negócio na análise de dados e na concretização de melhores decisões, potenciando assim o aumento de vendas e a fidelização de clientes com base em previsões e recomendações

Mais do que definir o correto perfil do cliente Parfois E-Commerce, pretende-se entregar ao negócio uma vasta e complexa ferramenta com o intuito de satisfazer as necessidades do mais exigente utilizador, apoiando-o e capacitando-o na sua vertente de Business & Web Analytics.

Palavras-chave: RFM; CLV; Data Mining; Clustering; Regras de Associação; Classificação; Sistema de Recomendação.

Abstract

Showing strong signs of growth and development, Parfois has invested in Information Systems and the added value of data this being an asset to the evaluation and understanding of the business. Active for a few years, Parfois' E-Commerce has advanced immensely, and its business has grown alongside Parfois' own development. Alongside this development, there has been a growing amount of data which has created more possibilities for exploration and decision support.

Currently, Parfois is investing in E-Commerce sales and so understanding client profiles strengthens the business' capability of increasing the Marketing and Sales sector, as well as creating a competitive advantage.

This project aims to expand on this model, developing a platform named 'Parfois Web Client Analytics' which will support the business through data analysis and more informed decision making. This will also create the potential for growing sales and client loyalty through predictions and recommendations.

More than defining the client's correct E-Commerce profile, this project aims to deliver to the business a vast and complex tool that will satisfy the needs of the most demanding users, supporting and enabling Business & Web Analytics.

Key words: RFM; CLV; Data Mining; Clustering; Association Rules; Classification; Recommendation Engine.

Agradecimentos

À Parfois (Barata & Ramilo, SA), por permitir o desenvolvimento deste projeto, capacitando-o como uma ferramenta para apoiar na tomada de decisão.

À equipa E-Commerce da Parfois, Eliana Silva e Emília Castro, pelo constante acreditar no crescimento das vendas online, na importância do cliente, e apoio nas análises realizadas.

À equipa DSI da Parfois, a todos, mas em particular aos colegas Bruno Lopes, Nelson Vieira, Diogo Costa, João Laureano, Sofia Neves e Leonor Dias, pela disponibilidade, diálogo e fonte de informação.

Aos meus pais, Miguel e Maria José, pela oportunidade de educação e pelo constante acreditar.

Às minhas irmãs, Mafalda, Anastácia e Carolina, as princesas dos meus olhos, que nunca fecharam a porta para mim, e sempre me estenderam a mão.

À minha namorada Maria Pereira por todo o amor incondicional, o universo do meu "Agora".

À Prof.ª Doutora Fátima Rodrigues, orientadora, sempre disponível e ativa, orientando no caminho certo este projeto.

Ao Prof.º Doutor Carlos Ferreira, coorientador, por toda a colaboração, com as críticas e ideias que permitiram a este projeto chegar a bom porto.

Agradeço igualmente a todos os docentes que contribuíram para a minha formação académica, disponíveis sempre para a partilha de conhecimento e esclarecimento de dúvidas.

A Deus e ao Universo, que zelam por mim junto com os meus Anjos.

Por último, um agradecimento, a todos aqueles que mesmo não estando aqui mencionados, contribuíram de alguma maneira para a elaboração desta tese e para o término deste percurso.

Índice

Lista de Figuras

Lista de Tabelas

Acrónimos e Símbolos

Lista de Acrónimos

B2B – *Business to Business*

B2C – *Business to Consumer*

RFM – *Recency, Frequency, Monetary*

ERP – *Enterprise Resource Planning*

ETL – *Extract, Transform, Load*

FPR – *False Positive Rate*

TPR – *True Positives Rate*

AED – Análise exploratória de dados

CLV – *Customer Lifetime Value*

LTV – *Lifetime Value*

IDE – *Integrated Development Environment*

ODBC – *Open Database Connectivity*

JDBC – *Java Database Connectivity*

GDPR – *General Data Protection Regulation*

RGPD – Regulamento Geral Proteção de Dados

SKU – *Stock Keeping Unit*

SaaS – *Software as a Service*

CRM – *Customer Relationship Management*

Notação e Glossário

Matching – Definido no conceito Fashion, como um produto que de alguma forma se relaciona, pode ser utilizado com, ou definir um look.

Look – Visual. Conjunto, composição ou configuração em acessórios e roupas.

Business & Web Analist – Elemento responsável pela realização de análises do negócio na área web/e-commerce.

Reporting – Componente cada vez mais presente em software, que permite ao utilizador extrair informação em diferentes formatos.

Dashboard – Normalmente integrados nas aplicações como painéis, que demonstram métricas e indicadores importantes para o publico alvo.

1 Introdução

O capítulo "Introdução" pretende dar ao leitor uma visão introdutória do projeto desenvolvido no âmbito da unidade curricular Tese de Mestrado de Engenharia Informática (MEI) no ramo de Sistemas de Informação e Conhecimento, lecionada no Instituto Superior de Engenharia do Porto. Neste capítulo é realizada uma contextualização da empresa que se aliou a este projeto, seguido da explicação do problema, e seus objetivos, informando ainda sobre os resultados pretendidos e da abordagem definida. No final do capítulo são apresentados os contributos e a estrutura definida para o documento desenvolvido.

1.1 Contexto

Com forte crescimento e expansão, a Parfois tem vindo a apostar nos Sistemas de Informação, e no valor acrescentado dos Dados, sendo este um forte aliado para a análise e compreensão do seu negócio. Atualmente a Parfois tem como forte aposta as vendas do E-Commerce, no entanto o desconhecimento do perfil de cliente, invalida a possibilidade alargar a sua componente de Marketing e Vendas, assim como obter vantagem competitiva.

1.1.1 Apresentação Empresa

A Parfois é uma empresa de acessórios de moda direcionada ao publico feminino, criada em 1994 pela sua fundadora Manuela Medeiros cujo sonho é garantir produtos únicos e fashion com preços para todos os diferentes públicos.

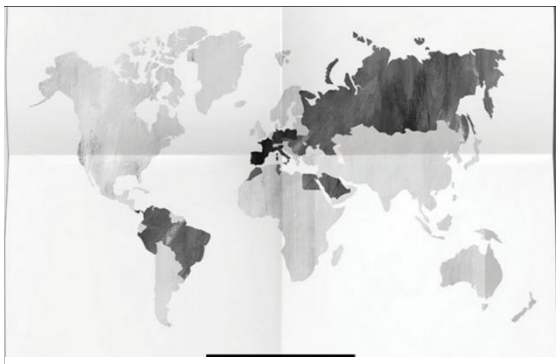


Figura 1 - Parfois World

Dispõe de uma equipa global de designers nos escritórios do Porto e Barcelona, atentos a todas as tendências de moda procura criar produtos inovadores.

A Parfois é hoje uma empresa com mais de 20 anos, contabilizando mais de 700 lojas, em mais de 50 países, continuando a crescer com números surpreendentes. O produto é uma constante em evolução, contando com cerca de 3500 referências por estação/coleção, onde apostam em referências novas em loja todas as semanas.

A satisfação dos seus clientes avalia-se facilmente com números surpreendentes como 70% dos clientes visitam as lojas duas vezes por mês e que cerca de 36% visitam a loja uma vez por semana.

Apostando num conceito de lojas diferenciador com lojas com cerca de 150 m2 e pequenos corners com 50m2, apostando fortemente no seu visual merchandising para conquistar clientes. [1]

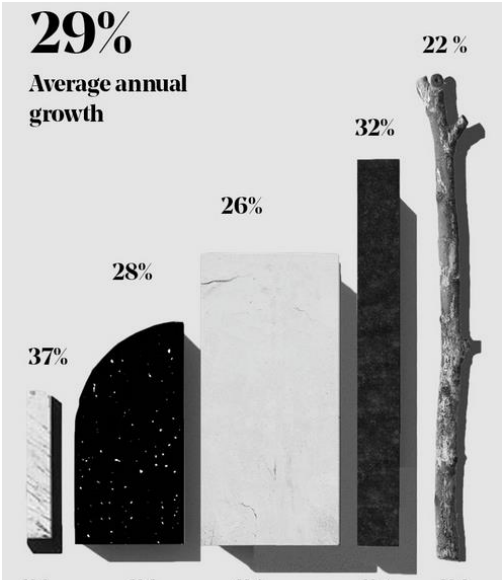


Figura 2 - Crescimento Anual Parfois

Com os seus escritórios sediados em Rio Tinto com os principais departamentos, e contando com mais de 400 colaboradores, apostou recentemente numa plataforma logística em Canelas com mais de 10.000 m2 para garantir forte capacidades de receção e expedição de produto

Contando com lojas espalhadas por mais de 50 países com diferentes modelos comerciais incluindo lojas próprias, franchisados e alguns consignados, os números de crescimento da Parfois não param de surpreender, e têm evoluído bastante ao longo dos anos conforme se pode verificar na Figura 2.

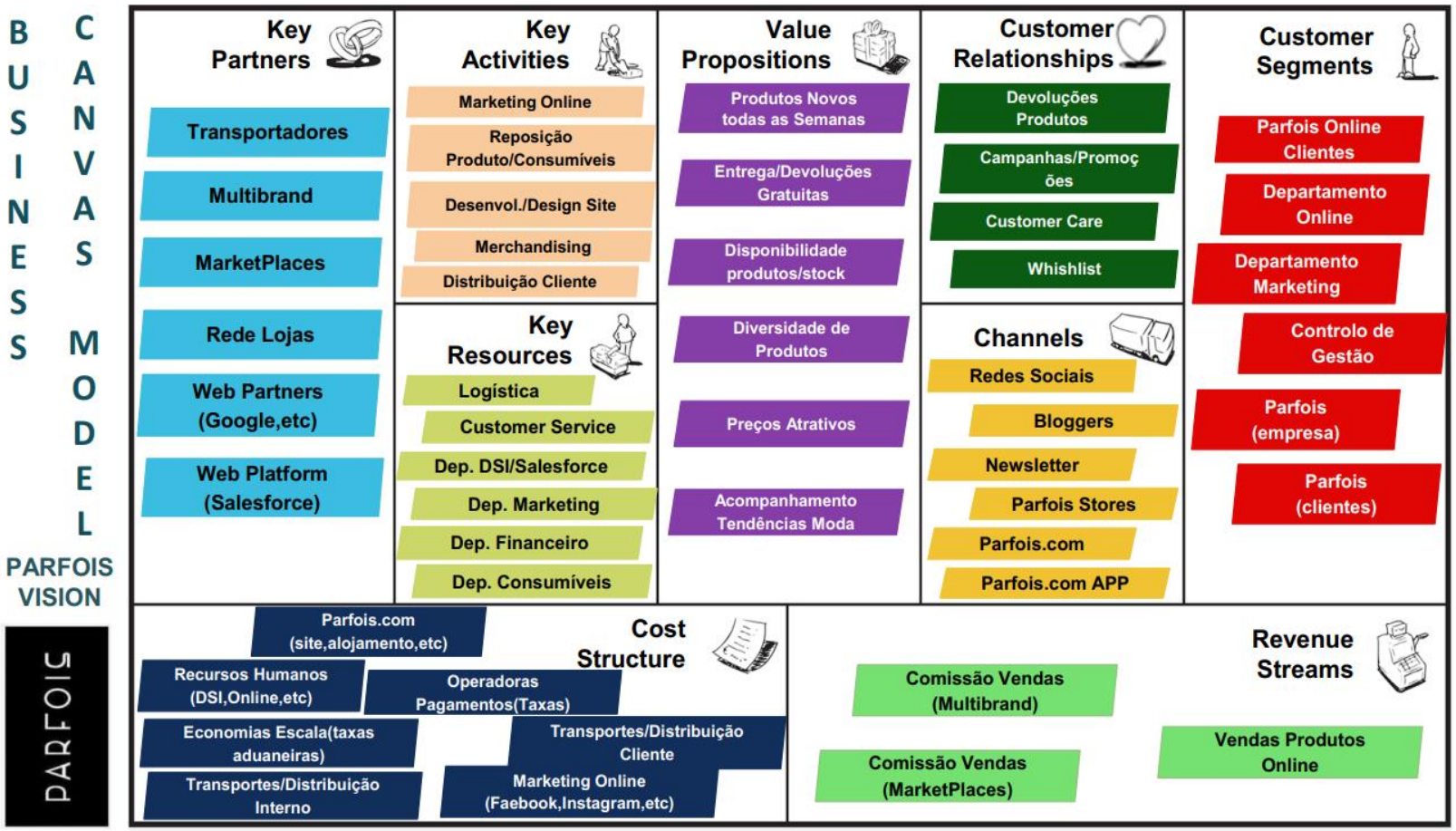
O segmento de mercado E-Commerce tem igualmente vindo a crescer bastante e tem sido uma forte aposta do negócio que acredita fortemente no potencial deste segmento, não somente no B2C, mas também B2B, estando cada vez mais a ser trabalhado nas vertentes de Marketing e de Gestão.

Não somente no seu site oficial (www.parfois.com) mas igualmente disponível em diversos MarketPlaces como Amazon, Debenhams e Tmall, a Parfois quer continuar a crescer no comércio digital.

[1] Fonte de Dados – Site Parfois

1.1.2 Modelo de Canvas

Para perceber o enquadramento do E-Commerce na estratégia da empresa, apresenta-se o modelo de canvas para essa mesma área.



1.2 Problema

Atualmente a inexistência de ferramentas que possibilitem o E-Commerce de aumentar as suas vendas com base em recomendações, e garantir a retenção do cliente através de previsões, são identificadas como falhas e um atraso face à concorrência.

1.3 Objetivos

Recorrendo à análise de dados da empresa Parfois, onde os dados serão extraídos de diversas bases de dados (ERP e Plataforma E-Commerce) de modo a obter o máximo de informação de vendas e devoluções da área de negócio E-Commerce, será realizada a exploração dos dados disponibilizados (2015 e 2016), tentando assim com estes serem definidas algumas análises que apoiem a tomada de decisão, nomeadamente no que diz respeito aos artigos e clientes.

Numa primeira fase com a utilização de um ETL para a extração e tratamento dos dados, estes serão disponibilizados num Cubo, de onde posteriormente se irá recorrer a técnicas de Exploração e Data Mining de modo a que seja possível realizar:

- Classificação de Clientes
- Recomendação de Artigos
- Análise RFM (Recency, Frequency, and Monetary)
- Análise CLV (Customer Lifetime Value)

Serão utilizados os principais algoritmos de Data Mining (a exemplo K-Means, Apriori, Filtragem Colaborativa) para cada um dos tipos de padrões a extrair dos dados, criando modelos de previsão e recomendação, e serão exploradas técnicas de pós-processamento do conhecimento obtido de forma a simplificar o mesmo.

Com o intuito de uma mais fácil análise e utilização dos dados para o Business & Web Analist será desenvolvida uma plataforma denominada "Parfois Web Client Analytics", permitindo a tomada de decisão ao negócio através Reporting e Dashboards.

Devem assim as ferramentas desenvolvidas acrescentar valor ao negócio, permitindo a definição de estratégias e técnicas que aumentem as vendas, e igualmente a aplicabilidade e otimização de técnicas de marketing, garantindo a retenção e fidelização de clientes, e garantindo estar sempre um passo à frente da concorrência neste crescente segmento de mercado do E-Commerce.

1.4 Análise de Valor

Este projeto pretende apoiar a Parfois no crescimento do E-Commerce, desenvolvendo assim uma plataforma denominada "Parfois Web Client Analytics", que irá auxiliar o negócio na análise de dados e na concretização de melhores decisões, potenciando assim o aumento de vendas e a fidelização de clientes.

1.5 Resultados Esperados

A extração e tratamento de toda a informação potenciará a aplicabilidade e otimização de novas técnicas de Marketing junto dos clientes, assim como a definição de novas estratégias deste crescente segmento de mercado.

Mais do que definir o correto perfil do cliente Parfois E-Commerce, pretende-se entregar ao negócio uma vasta e complexa ferramenta com o intuito de satisfazer as necessidades do mais exigente utilizador, apoiando-o e capacitando-o na sua vertente de Business & Web Analytics.

A plataforma é totalmente direcionada ao segmento de clientes “Parfois E-Commerce” garantindo uma nova visão da informação, nomeadamente nas perspetivas:

- Análise de Clientes
    - Classificação de Clientes (RFM)
    - Valor dos Clientes (CLV)

- Gestão de Produtos
    - Recomendação de artigos (baseado em compra)
    - Recomendação de artigos (baseado em matching)

Considerando um enorme volume de dados com uma exploração não direcionada a este âmbito, a informação tem assim elevado valor, embora desconhecido, mas com certezas de potenciar vantagem e alavancar melhores resultados.

A existência de ferramentas idênticas engrandece o nosso caminho definindo-o como o caminho certo, aliado ao nosso know-how na área de retalho e-commerce, e a nossa capacidade de pensamento critico/criativo fazem da plataforma "Parfois Web Client Analytics" uma aposta com potencial.

1.6 Abordagem Preconizada

Contando já com alguns anos no ativo, o E-Commerce da Parfois tem vindo a evoluir bastante, sendo cada vez mais uma aposta da empresa. O seu volume de negócios não tem parado de aumentar acompanhando o igual crescimento da Parfois. Como tal em igual proporção os dados existentes têm vindo a aumentar, sendo vasta a quantidade com possibilidade de exploração.

A implementação deste projeto pretende conforme anteriormente indicado potenciar mais e melhores análises dos dados que possibilitem o E-Commerce de aumentar as suas vendas com base em previsões e recomendações.

O desenvolvimento da plataforma está definida relativamente à sua arquitetura em um modelo com três componentes diferenciadas e cada uma com a sua importância, como se pode verificar na Figura 3 e que em seguida se explica.

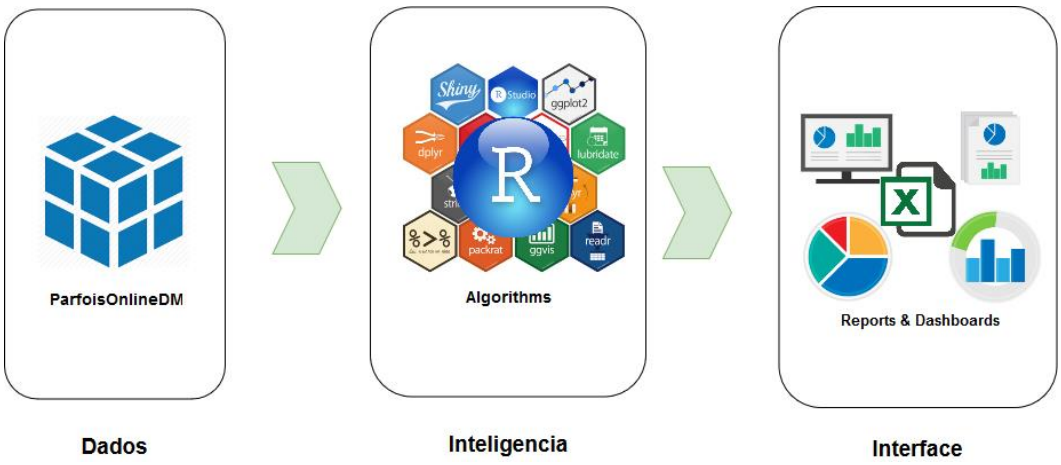


Figura 3 - Modelo Componentes

A denominada componente “Dados” é parte importantíssima neste projeto, sendo eles o ponto de alavancagem e onde se encontra toda a informação inicialmente dispersa, e na sua fase final tratada e segmentada para ser utilizada pelo componente seguinte.

Por sua vez a componente “Inteligência”, será a responsável pela aplicabilidade de diversos algoritmos aos dados resultando numa informação “inteligente” e pronta para ser analisada.

A componente de “Interface” deverá ser uma “user friendly interface” capaz de permitir a um qualquer utilizador explorar os dados através de Reporting e Dashboards.

A plataforma “Parfois Web Client Analytics" irá auxiliar o negócio na análise de dados e na concretização de melhores decisões.

1.7 Contributos da Tese

Em conformidade com os objetivos definidos para a tese, e todo o trabalho de investigação e análise realizado, pode-se constatar que foi identificado duas vertentes onde a tese desenvolvida poderá contribuir valor, identificando-as de seguida.

Contributo para as organizações

A plataforma desenvolvida irá certamente auxiliar a empresa na gestão e controlo da sua área e-commerce, utilizando assim esta ferramenta ágil e user friendly, e com possibilidade de crescimento e expansão, mas também de extensão, uma vez que poderá ser facilmente adaptada para novas empresas que pretendam realizar estas métricas, onde somente teremos de realizar alguns ajustes conforme os requisitos e assumir novas fontes de dados.

Destaca-se que não somente o e-commerce poderá beneficiar com esta solução, mas de igual modo áreas que pretendem definir novas estratégias junto dos clientes, uma vez que é notório que as empresas começam cada vez mais a dar enfâse e importância ao cliente.

Contributo para a sociedade académica

Não sendo um tema totalmente novo no meio académico, a análise de dados terá sempre algo a acrescentar, devido às diversas técnicas aplicadas, no caso em concreto o volume de dados foi um fator diferenciador, aliado à crescente do modelo de negócio de retalho e de e-commerce que tem vindo a evoluir, e que é cada vez mais alvo de caso de estudo, uma vez que é importante percecionar como as grandes empresas se adaptam para a era digital, e como estas se capacitam na análise e interpretação de dados.

Um elemento auxiliar ao indicado anteriormente é a programação na linguagem R destacada neste projeto e com uma enorme capacidade funcional para grandes volumes de dados. De notar a utilização de boas práticas de programação em todo o desenvolvimento, e ferramentas para o seu apoio, o que poderá desde já ser um input válido para alunos e atuais developers.

1.8 Estrutura da Tese

A tese desenvolvida encontra-se estruturada em 8 capítulos, conforme se pode verificar na Figura 4, sendo objetivo que a mesma seja fluida como um todo, permitindo ao leitor uma agradável experiência de leitura e aprendizagem.

Figura 4 - Estrutura Tese

No Capítulo 2, denominado a “Análise de Valor” é transmitido ao leitor o conceito de valor do projeto desenvolvido, permitindo a este ter a perceção de quanto poderá este apoiar o negócio, e quem poderá com este beneficiar, referindo igualmente como poderemos avaliar esse nosso valor.

No Capítulo 3, identificado como “Estado da Arte” o leitor deverá ter uma perspetiva dos principais conceitos abordados ao longo do documento, permitindo um contacto teórico com

os demais temas que foram a base para o desenvolvimento deste projeto identificando alguns dos casos de estudo analisados e considerados de maior relevância.

No Capítulo 4, designado como “Design da Solução” deseja-se dar ao leitor uma visão da arquitetura do projeto proposto, abordando cada um dos diferentes componentes.

No Capítulo 5, assinalado como “Implementação da Solução” incide-se no detalhe sobre o desenvolvimento realizado, onde são descriminadas cada uma das diferentes fases de acordo com o fluxo de desenvolvimento realizado, e interpretando cada uma destas.

No Capítulo 6, intitulado “Demonstração da Solução” o leitor terá a visão da interação com a plataforma, de todos os ecrãs disponíveis, seu funcionamento e informação disponível, sendo cada uma abordada em detalhe.

No Capítulo 7, nomeado como “Avaliação da Solução” o leitor terá a perceção de quais as métricas para a avaliação utilizadas, e que garantem a veracidade do desenvolvimento.

No Capítulo 8, o último, e não menos importante é a denominada “Conclusão” em que se incide sobre uma reflexão de todo o percurso realizado para o fecho deste projeto.

2 Análise de Valor

O capítulo “Análise de Valor” transmite ao leitor o conceito de valor do projeto desenvolvido, permitindo a este ter a perceção de quanto poderá este apoiar o negócio, e quem poderá com este beneficiar, referindo igualmente como poderemos avaliar esse nosso valor.

2.1 Proposta de Valor

"Grow your business knowing your customers" [1]

Não se pode de melhor forma caracterizar a principal proposta de valor, assumindo que se pretende conhecer melhor o Cliente Parfois E-Commerce.

Atualmente esta “loja” representa um significativo volume de vendas, apresentando um grande crescimento nos últimos anos, e identificada como um canal de forte aposta.

Mais do que definir o correto perfil do cliente Parfois E-Commerce, trazer até ao negócio uma vasta e complexa ferramenta com o intuito de satisfazer as necessidades do mais exigente utilizador, apoiando-o e capacitando-o na sua vertente de Business & Web Analytics.

A existência de ferramentas idênticas engrandece o caminho definindo-o como o caminho certo, aliado ao know-how na área de retalho e-commerce, e a capacidade de pensamento critico/criativo fazem da plataforma "Parfois Web Client Analytics" uma aposta vencedora.

2.2 Análise de Valor

O que procura o Cliente? Porque deve este comprar na minha empresa e não na concorrência?

Estas perguntas são a realidade atual de muitas empresas, como criar valor ao cliente, no entanto depende sempre da perspetiva do negócio e da perspetiva do cliente. Isto porque o negócio está focado em garantir qualidade a baixo preço, mantendo as suas margens de lucro, mantendo-se a par da concorrência e numa luta desenfreada pelo conquistar do cliente. Ou seja, garantir o lucro da empresa trabalhando para a satisfação do cliente.

Pelo outro lado o cliente atual já não compra o produto pelas suas características, mas pelo valor/beneficio que este acrescenta, uma vez que no processo de decisão de compra os clientes têm cada vez mais o poder de decidir e comparar os diversos produtos entre os concorrentes. Embora este não tenha a perceção dos custos de produção, tem uma perspetiva global do seu preço, e este preço é importante para o cliente e para o negócio, denominado pelo "perceived value", ou seja, o preço que o cliente tem em mente acerca daquele produto. A Parfois define a sua estratégia de preço junto do cliente, não se colocando no patamar de luxo, mas igualmente não se colocando no patamar do produto barato e baixa qualidade. E esta posição no mercado permite atrair clientes.

Mas o valor para o cliente vai para além do preço que paga pelo produto, mas também pela sua experiência de compra, pelo benefício que este produto lhe garante. A Parfois garante hoje ao cliente uma experiência de compra, privilegiando um atendimento personalizado ao cliente, tendo todo um processo fácil de aquisição e troca do produto. De igual forma o merchandising da exposição de produto e todo o conceito de loja, faz o cliente Parfois se sentir especial, acrescentado valor ao seu processo de compra.

No entanto quando se pensa primeiramente em Valor também têm de se pensar em Custo, e deste processo surgem os benefícios e sacrifícios que vão garantir o denominado "value for the customer" e "perceived value".

De igual modo o projeto proposto pretende ser uma mais valia para a Parfois, tendo como estratégia acrescentar valor para esta. Neste âmbito é necessário realizar uma Análise de Valor onde teremos de analisar cada conceito de valor da plataforma sobre o valor que representará para o cliente caracterizando os seus benefícios e sacrifícios, analisando assim todo o impacto que a plataforma a desenvolver poderá ter no negócio de forma a atingir um valor de melhoria de serviço/ produto em que não aumenta custos nem desqualifica a qualidade do produto / serviço para o cliente final.

De seguida apresenta-se o modelo de análise de valor com cada um dos seus elementos.

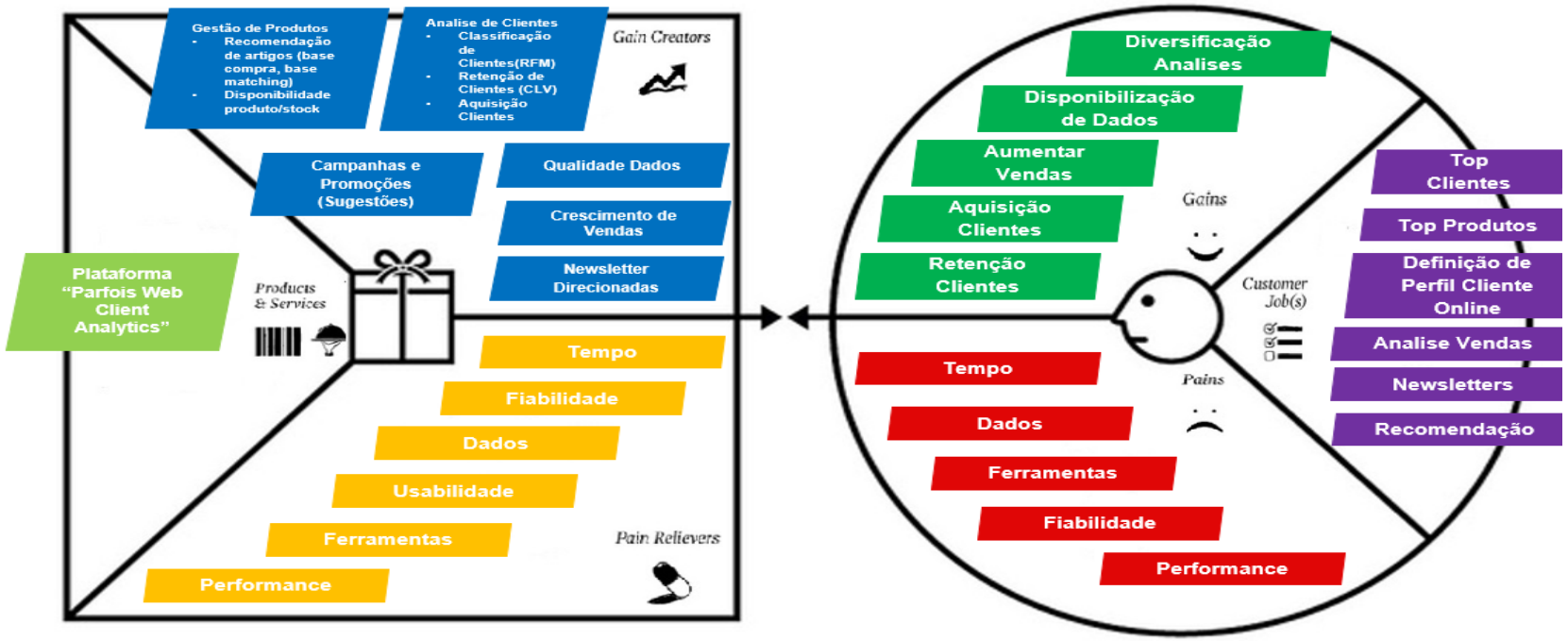
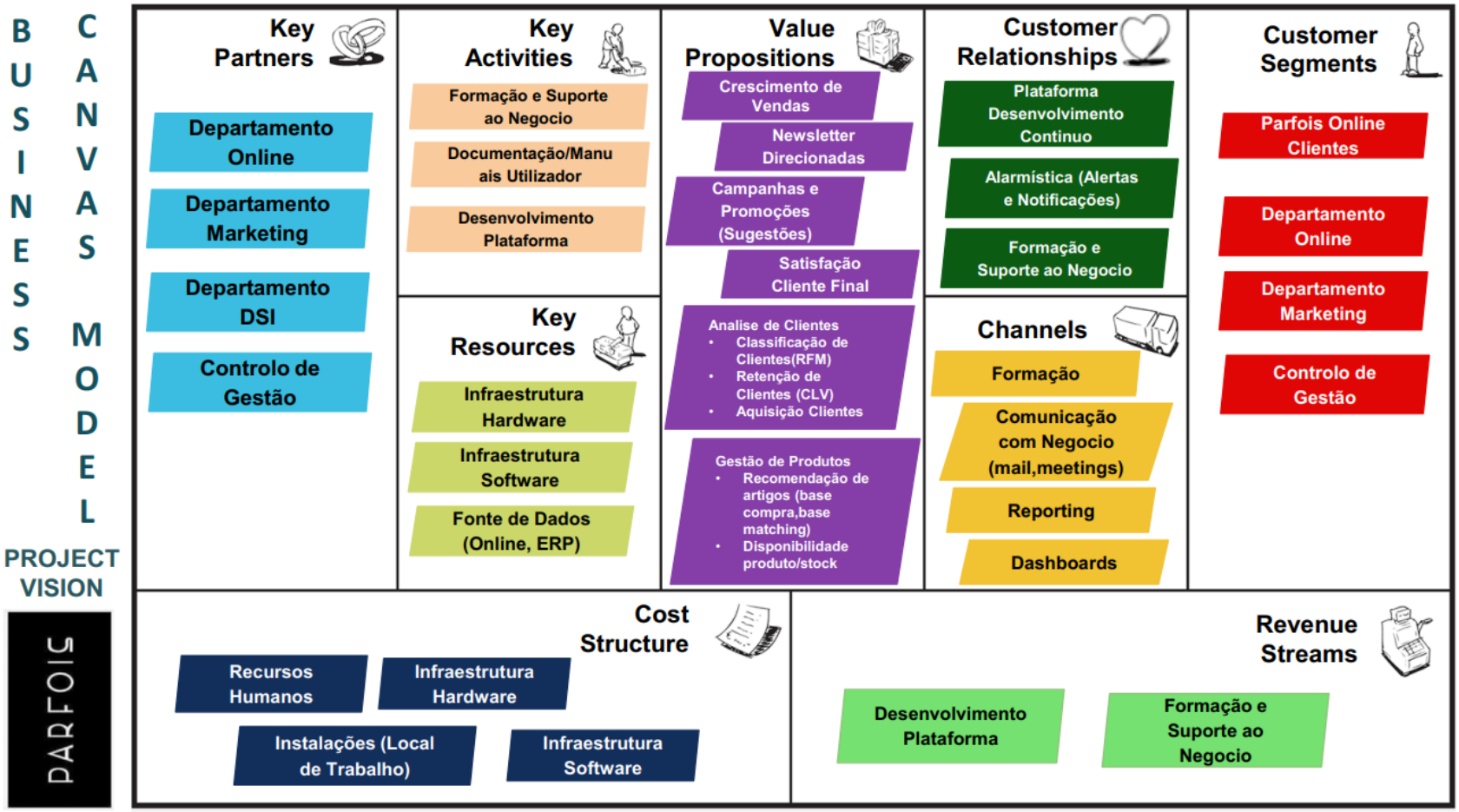


Figura 5 - Modelo Análise Valor

2.3 Modelo de Canvas

Apresenta-se o modelo de canvas sobre a perspetiva do projeto a desenvolver, ou seja, a plataforma "Parfois Web Client Analytics".



2.4 Modelo New Concept Development (NCD)

O modelo NCD (New Concept Development) desenvolvido por Koen em 2001, pretendia fornecer uma visão e terminologia comum ao FFE (Fuzzy Front End), uma vez que identificava como dificuldade a inexistência de uma linguagem e vocabulário comum que permitisse criar conhecimento e diferenciar as diferentes partes do processo de uma forma concreta e percetível para todos. [2]

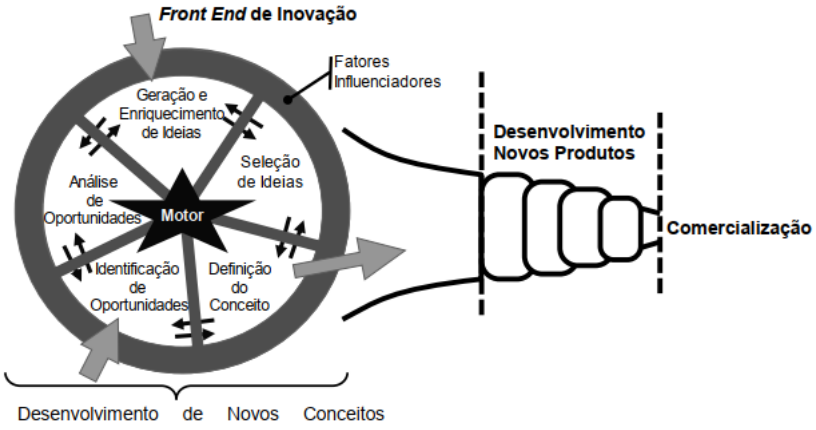


Figura 6 - Modelo NCD [3]

Definido por 3 componentes principais, identificados como o Motor (*Engine*), a Roda (*Wheel*) e o Aro (*Rim*) onde cada um desempenha um papel diferente embora sempre relacionados entre eles.

O motor (*engine*) é denominado como o centro do modelo, a liderança, cultura e visão estratégica do negócio, este é o responsável por controlar o comportamento dos restantes componentes, o impulsionador dos cinco elementos da roda.

A roda (*wheel*) é onde se encontram os principais elementos ativos que se relacionam entre eles, nomeadamente:

- Identificação da Oportunidade (Oportunity Identification)
- Análise da Oportunidade (Opportunity Analysis)
- Definição de Ideias (Idea Genesys)
- Seleção de Ideias (Idea Selection)
- Conceção e Desenvolvimento Tecnológico (Concept & Technology Development)

O aro (*rim*) representa os fatores influenciadores sobre o modelo, que em grande parte não conseguem ser controlados, tratando-se de elementos externos, mas que em muito podem influenciar todo o modelo.

Por último as setas que apontam para o interior do modelo representam onde o processo poderá iniciar, que poderá ser na “Identificação de Oportunidade” ou “Definição de Ideias”. Já

a seta a sair do modelo representa que o processo de "Conceção e Desenvolvimento Tecnológico" deverá evoluir para o processo de "New Product and Process Development" (NPPD).[3]

Valida-se de seguida a denominada "roda" e os seus elementos enquadrando no projeto proposto.

Opportunity Identification - Nesta primeira etapa a empresa tem por objetivo a identificação de oportunidades para potenciar a sua eficiência e eficácia nos diferentes segmentos de mercado. Associado ao grande objetivo da Parfois de aumentar as suas vendas, o segmento Parfois E-Commerce tem vindo a redefinir a sua estratégia de acordo com este objetivo comum, definindo uma estratégia de análise do mercado e concorrência, e verificando as tendências tecnológicas atuais. No seguimento dessa estratégia foi identificada a oportunidade de definir o perfil do cliente Parfois E-Commerce.

Opportunity Analysis - No seguimento da oportunidade de conhecer o "Cliente Parfois E-Commerce", é necessário detalhar de que forma deverá este processo ser realizado, e quais os objetivos a este processo associados.

- Como iremos definir este perfil?
- De que serve a definição desse perfil?
- Que vantagens iremos ter com este processo?

Estas diversas questões fazem parte do processo de brainstorming da análise da oportunidade, onde são avaliados os benefícios assim como os riscos.

Com a oportunidade surge a redefinição da oportunidade e cada vez mais se aproxima de uma ideia final.

Idea Genesis - *"Don't start with the idea, start with the opportunity."* Dizem que toda a ideia nasce de uma oportunidade encontrada, no entanto toda a oportunidade percorre o seu caminho até se tornar uma ideia. Necessário clarificar e discutir sobre a oportunidade encontrada, e como torná-la numa ideia vencedora. [2]

Trazer o negócio e as pessoas até nós com as suas ideias fazendo-as sentir parte da solução. Existe a necessidade de discutir junto com o departamento E-Commerce da Parfois sobre a ideia para definir todos os detalhes da mesma, e ir de encontro com as necessidades do negócio e os objetivos propostos.

Idea Selection - De entre todas as ideias existentes, qual poderá ser a melhor aposta. Um dos grandes dilemas na seleção da ideia acaba por ser o investimento financeiro e o seu retorno, assim como as datas e metas atingir. Poderemos optar por desenvolver externamente, ou talvez

contratar outsourcing, ou então internamente e em conjunto com o negócio realizar o desenvolvimento da ideia, tentando reter dentro de portas o conhecimento.

Das diversas oportunidades transformadas em ideias, e após a sua fase de decisão surge assim a possibilidade de desenvolvimento da plataforma interna "Parfois Web Client Analytics".

Concept & Technology Development - A última etapa do modelo NCD, envolverá o desenvolvimento e concretização da ideia. Como transformar a ideia, num produto válido, funcional e eficiente para o negócio. Para a concretização da denominada plataforma "Parfois Web Client Analytics" existe a necessidade de planear e definir corretamente todo o processo de New Product and Process Development (NPPD), assim como obter o patrocínio e o envolvimento do negócio no atingir dos objetivos propostos. [3]

2.5 Redes de Valor

"People naturally network as they work so why not model itself as network" [4]

Verna Allee define redes de valor "value networks" como a troca de informação complexa entre indivíduos, grupos ou organizações que criam alguma espécie de valor tangível ou intangível. Ou seja, qualquer empresa, ou grupo de empresas comprometidas a gerar benefício tangível ou intangível poderá ser vista como uma rede de valor.[4]

Igualmente Verna Allee desenvolveu a denominada análise de redes de valor "value network analysis", visto como um sistema de análise e mapeamento para desmistificar a criação de valor tangível e intangível entre diversos participantes, acreditando no potencial para a perceção de problemas e mobilização coletiva para a criação de mudança.[5]

Cada vez mais o ser humano não se limita a comunicar dentro da sua rede, mas sim a abrir horizontes ao desconhecido, ou seja, não se limita ao seu departamento de trabalho, seus amigos, mas sim opta por alargar o seu conhecimento conectando-se a outras redes/grupos onde partilha e absorve informação, num contexto de win-to-win. A criação e gestão de uma rede de valor/contactos embora não seja simples permite ao indivíduo fazer chegar a mensagem a mais participantes, assim como obter mais feedback, e obter informação de mais fontes, estando assim a criar valor para si.

Na perspetiva do projeto proposto, a plataforma "Parfois Web Client Analytics" igualmente pretende definir a sua rede de valor, não somente numa ótica interna do departamento E-Commerce, mas também do universo Parfois, e tudo em seu redor. Porque a rede de valor da Parfois é constituída por diversos participantes e todos eles acrescentando de alguma forma valor ao projeto, transcrevendo-se em valor tangível e valor intangível, conforme poderá ser verificado no modelo da rede de valor apresentado de seguida na Figura 7.

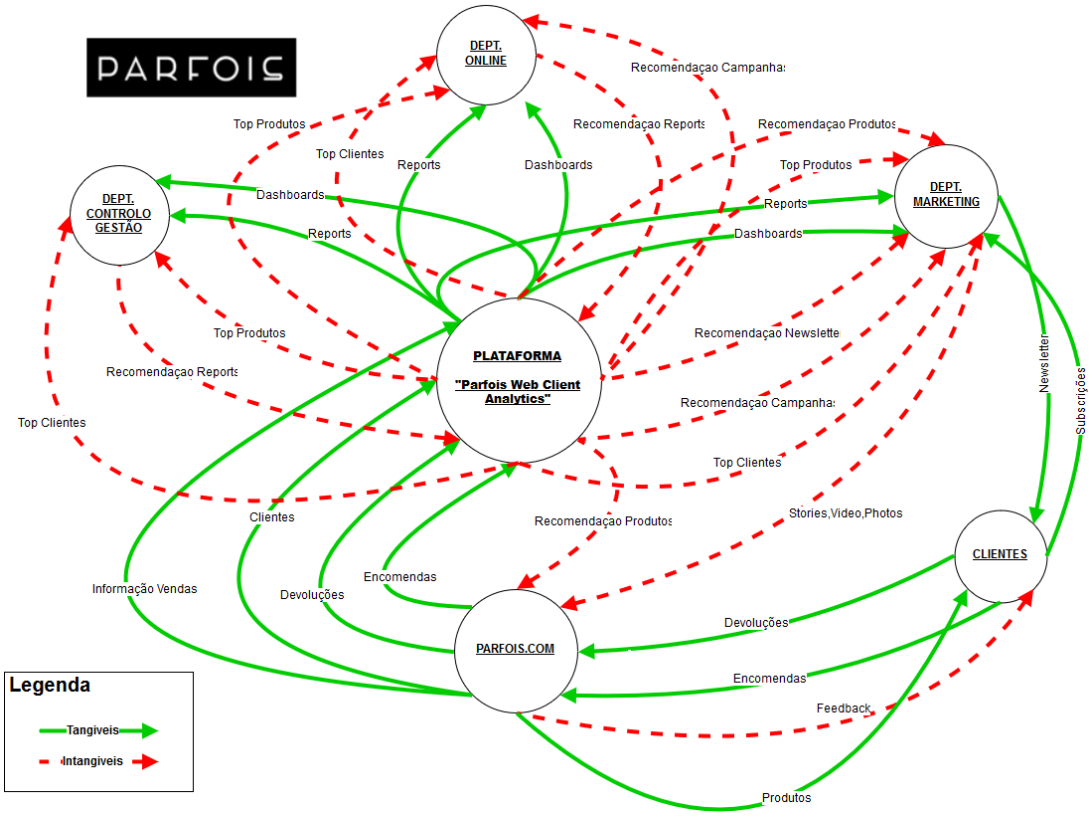


Figura 7 - Rede de Valor

No modelo de rede de valor apresentado, é possível verificar os diferentes e principais participantes que interagem com o projeto proposto “Parfois Web Client Analytics”

Defina-se como valor tangível todos os movimentos onde são envolvidos bens, serviços, ou algum resultado/lucro, não sendo nunca limitado à sua natureza física/palpável, a exemplo no nosso modelo temos as encomendas de clientes.

Da mesma forma se define valor intangível como tudo o que é movimentado entre a cadeia de valor que nunca será palpável, defina-se como informação, ou partilha de conhecimento, mas que acrescenta valor à relação entre os indivíduos/organizações. A exemplo no nosso modelo consideramos como intangível a informação de “Top Clientes” e “Top Produtos”.

2.6 Método AHP

Um processo de tomada de decisão que permite comparar aspetos quantitativos e qualitativos de cada alternativa é o chamado AHP (Analytic Hierarchy Process).

O método AHP é um dos mais conhecidos métodos de análise multicritério.

Desenvolvido por Thomas Saaty no final da década de 1970, enquanto professor da Pennsylvania's Wharton School. O método aceita variáveis quantitativas e qualitativas, além de permitir que as avaliações sejam feitas com base no conhecimento e em impressões subjetivas que o decisor tem sobre o tema.[6]

O Decision Support Systems Glossary define AHP como "uma aproximação para tomada de decisão que envolve estruturação de multicritérios de escolha numa hierarquia. O método avalia a importância relativa desses critérios, compara alternativas para cada critério, e determina um ranking total das alternativas"[7].

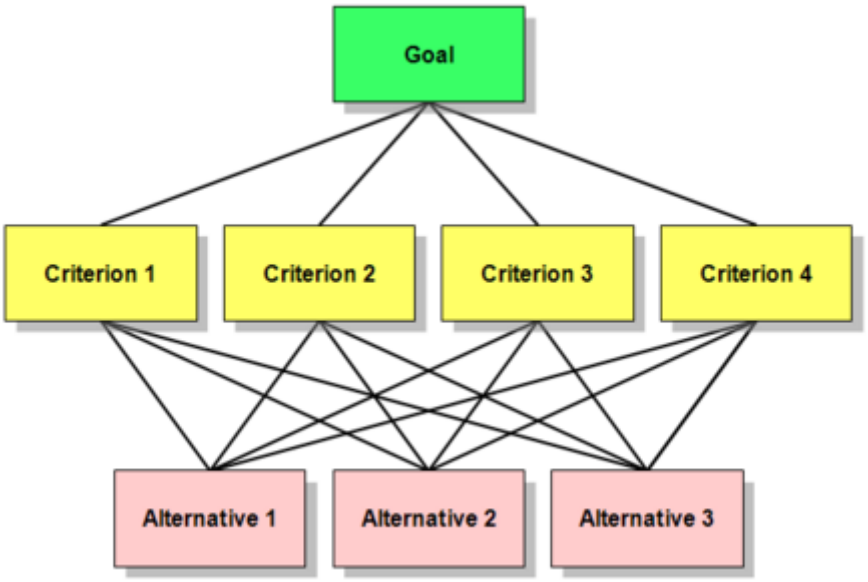


Figura 8 - Representação Abstrata de Hierarquia de Decisão

Forman ilustra de forma clara o modelo AHP na figura "hierarquia de decisão". O primeiro nível significa a "meta final" do decisor, seguidamente são definidos os critérios para a tomada de decisão e no último nível são indicadas as alternativas possíveis para o modelo de decisão [8].

É importante lembrar que o método AHP permite quantos níveis forem necessários. E também inúmeros critérios e subcritérios dentro de cada nível. O mesmo é válido para as alternativas sob avaliação.

No âmbito do projeto proposto pretende-se através do método AHP apoiar na tomada de decisão acerca de qual a melhor opção para a Parfois para responder às suas necessidades acerca da componente E-Commerce.

Para a elaboração do método AHP, será utilizado o software *Super Decisions* que visa apoiar a utilização deste método.[2] [9]

Foi assim definida a seguinte estrutura respeitando o conceito do método AHP, onde se define o objetivo (*goal*), os seus critérios, e as alternativas envolvidas, conforme se pode verificar na Figura 9.

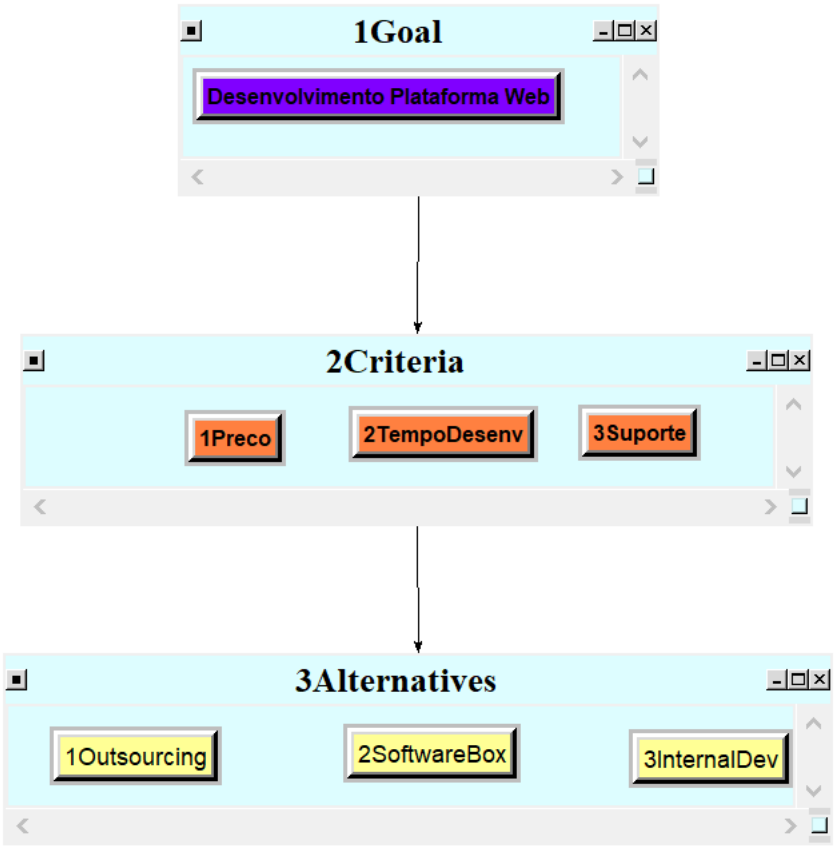


Figura 9 - Hierarquia AHP

Deve-se igualmente definir todos os dados necessários para cada um dos níveis, para ser possível obter um resultado fidedigno, como se verifica com os critérios:

- Preço
- Tempo Desenvolvimento
- Suporte

[2] Software Super Decisions - www.superdecisions.com

Clusters/Nodes	• **1Goal:** *The cluster to contain the goal node* ◦ **Desenvolvimento Plataforma Web:** *Empresa necessita uma plataforma web com informacao de clientes e produtos resultado de vendas do site.* • **2Criteria:** *Criterio para escolha tipo desenvolvimento.* ◦ **1Preco:** *Preco do Produto* ◦ **2TempoDesenv:** *Tempo Desenvolvimento* ◦ **3Suporte:** *Suporte do Produto* • **3Alternatives:** *Alternativas em avaliacao para o desenvolvimento.* ◦ **1Outsourcing:** *Preco- 50Euros/dia 2000Euros Tempo Desenv: -3 semanas analise -2 semanas dev -2 semanas testes -1 semana entrega/formacao Suporte - Gratuito durante desenvolvimento, pago apos entrega.* ◦ **2SoftwareBox:** *Preco- 7000Euros Tempo Desenv: -2 semanas analise -2 semanas dev -1 semanas testes -1 semana entrega/formacao Suporte - Gratuito durante desenvolvimento, pago apos entrega.* ◦ **3InternalDev:** *Preco- 40Euros/dia 1800Euros Tempo Desenv: -2 semanas analise -4 semanas dev -2 semanas testes -1 semana entrega/formacao Suporte - Gratuito durante desenvolvimento, gratuito apos entrega.*

Figura 10 - Clusters Definidos

Saaty sugeriu o uso de uma escala-padrão de valores, que variam de 1 a 9, com o objetivo de avaliar numericamente alternativas e critérios em um processo de decisão. Para cada valor, define-se o seu predicado qualitativo seguido de explicação textual, para diminuir dúvidas no momento da decisão [10].

Intensidade de Importância	Definição	Explicação
1	Mesma importância	As duas atividades contribuem igualmente para o objetivo.
3	Importância pequena de uma sobre a outra	A experiência e o julgamento favorecem levemente uma atividade em relação à outra.
5	Importância grande ou essencial	A experiência e o julgamento favorecem fortemente uma atividade em relação à outra.
7	Importância muito grande ou demonstrada	Uma atividade é muito fortemente favorecida em relação à outra; sua dominação de importância é demonstrada na prática.
9	Importância absoluta	A evidência favorece uma atividade em relação à outra com o mais alto grau de certeza.
2, 4, 6, 8	Valores intermediários entre os valores adjacentes	Quando se procura uma condição de compromisso entre duas definições.
Recíprocos dos valores acima de zero	Se a atividade i recebe uma das designações diferentes acima de zero, quando comparada com a atividade j, então j tem o valor recíproco quando comparada com i.	Uma designação razoável.
Racionais	Razões resultantes da escala	Se a consistência tiver de ser forçada para obter valores numéricos n, somente para completar a matriz.

Tabela 1 - Tabela Padrão de Valores por Saaty

Os valores devem ser estimados tendo como base o conhecimento sobre o negócio do decisor e sua decisão. Aqui podemos ter a decisão de apenas uma pessoa ou de um grupo de pessoas.

A simplicidade das comparações par-a-par permite que o decisor mantenha o foco em cada detalhe do problema. A possibilidade de usar a escala verbal auxilia aqueles que tenham alguma dificuldade com a escala numérica.

Como se pode verificar foram definidos os critérios, com os seguintes valores para as variáveis quantitativas e qualitativas para cada alternativa.

		Critérios		
Alternativas		**Preço €**	**Tempo Desenvolvimento (semanas)**	**Suporte**
Alternativa 1	**Outsourcing**	2000	8	Pago
Alternativa 2	**Software Box**	7000	6	Pago
Alternativa 3	**Internal Dev**	1800	9	Pago

Tabela 2 - Critérios e Alternativas AHP

No entanto para se aplicar corretamente o método AHP, é necessário serem definidas as prioridades dos critérios, mas também as prioridades dos critérios face às alternativas, para isso foi utilizado o software *Super Decisions* que visa apoiar a utilização deste método.[3] [9]

- **Priorização dos Critérios**

Define-se o grau de prioridade entre os critérios definidos.

- Preço tem peso 8x superior ao Suporte.
- Preço tem peso 2x superior ao Tempo.
- Tempo tem peso 4x superior ao Suporte.

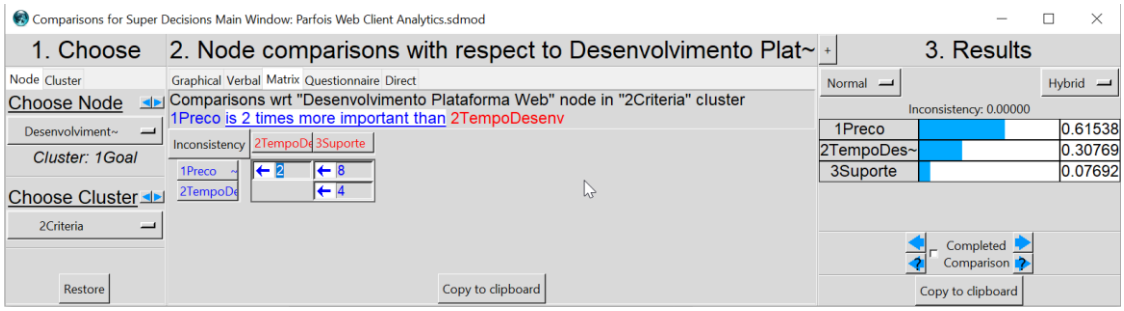


Figura 11 - Priorização dos Critérios

- **Priorização dos Critérios Fase às Alternativas**

Necessário para cada um dos critérios definir as suas importâncias face às alternativas existentes.

Relativamente ao critério Preço, foi realizada a seguinte definição:

- Outsourcing tem peso 2x superior ao SoftwareBox.
- InternalDev tem peso 4x superior ao Outsourcing.
- InternalDev tem peso 8x superior ao SoftwareBox.

[3] Software Super Decisions - www.superdecisions.com

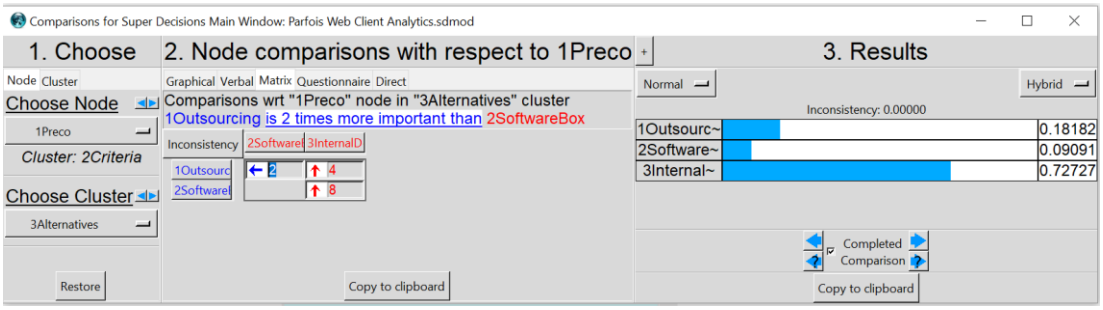


Figura 12 - Priorização Critério Preço

Relativamente ao critério TempoDesenv, foi realizada a seguinte definição:

- SoftwareBox tem peso 2x superior ao Outsourcing.
- Outsourcing tem peso 4x superior ao InternalDev.
- SoftwareBox tem peso 8x superior ao InternalDev.

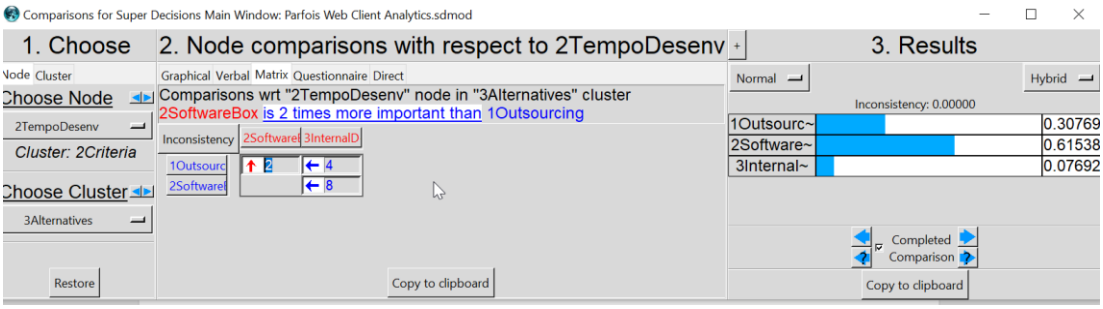


Figura 13 - Priorização Critério Tempo

Relativamente ao critério Suporte, foi realizada a seguinte definição:

- Outsourcing tem peso 2x superior ao SoftwareBox.
- InternalDev tem peso 4x superior ao Outsourcing.
- InternalDev tem peso 8x superior ao SoftwareBox.

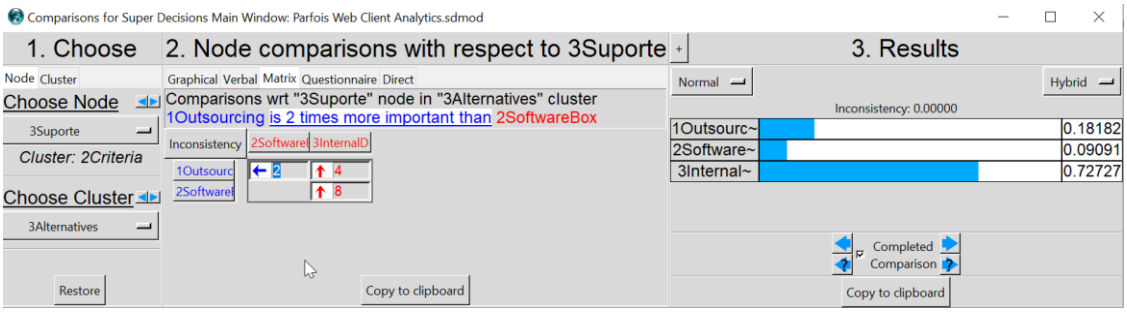


Figura 14 - Priorização Critério Suporte

Uma vez definidos todos os critérios, são realizados os cálculos do método AHP, que irá assim definir conforme todos os dados definidos, um ranking acerca das alternativas, apoiando a decisão.

Alternative Rankings

Graphic	Alternatives	Total	Normal	Ideal	Ranking
	1Outsourcing	0.1103	0.2205	0.4184	3
	2SoftwareBox	0.1261	0.2523	0.4786	2
	3InternalDev	0.2636	0.5272	1.0000	1

Figura 15 - Ranking Alternativas AHP

3 Estado da Arte

O capítulo “Estado da Arte” garante ao leitor uma perspetiva dos principais conceitos abordados ao longo do documento, permitindo um contacto teórico com os demais temas que foram a base para o desenvolvimento deste projeto. No final do capítulo são apresentados alguns dos casos de estudo analisados e considerados de maior relevância.

3.1 Modelo RFM

O modelo RFM existe há mais de quarenta anos tendo sido introduzida por Cullinan, é um dos modelos mais utilizados para efetuar uma segmentação de clientes [11][12].

O valor da análise RFM como um método para identificar os clientes de rápida resposta a promoções de marketing, e para melhorar as taxas de resposta geral é bem conhecido e é amplamente aplicado. Menos compreendido, no entanto, é o valor da aplicação do modelo RFM a uma base de dados com clientes.

O modelo RFM baseia-se nos dados de atividade dos clientes, pode ser algo como as atividades de compras do cliente, como visitas a um website, ou quantas vezes o cliente lança uma aplicação.

A segmentação RFM pode ser aplicada a qualquer tipo de dados relacionados com atividade.

As atividades de dados podem assumir as mais variadas formas:

Figura 16 - RFM Atividades[13]

Desde que os dados possam ser medidos e replicáveis podemos utilizar o modelo RFM.

O modelo RFM possui as seguintes métricas:

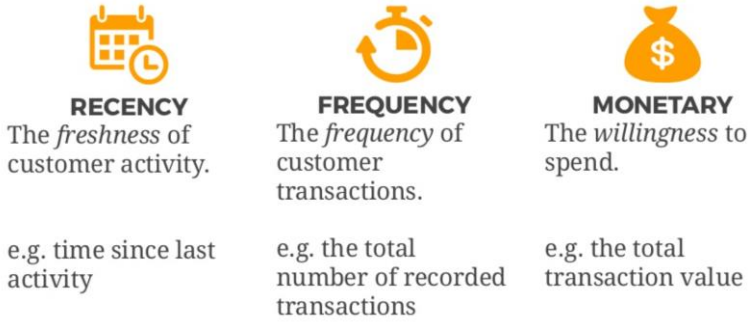


Figura 17 - Métricas RFM[13]

- **Recency**: retrata a assiduidade do cliente.
  Por Exemplo: A última vez que um cliente comprou na Parfois E-Commerce;
  Esta métrica é calculada através da seguinte fórmula $R = NOW - max(Order\ Date)$

- **Frequency**: retrata a frequência do cliente.
  Por exemplo: Número total de transações por cliente na Parfois E-Commerce;
  Esta métrica é calculada através da seguinte fórmula $F = COUNT(TranId)$

- **Monetary**: retrata a vontade que um cliente tem de gastar.
  Por Exemplo: O valor total, em moeda, de quanto o cliente gastou na Parfois E-Commerce.
  Esta métrica é calculada através da seguinte fórmula $M = SUM(TranId\ Value)$

O processo de agrupar um grande número de valores numéricos em um pequeno número de categorias é, às vezes, chamado de categorização. Na análise de RFM, as posições são as categorias classificadas. No entanto existem dois diferentes métodos de categorizar as métricas RFM, e a sua aplicabilidade em muito depende dos objetivos propostos.

Categorização Independente (Independent Binning Method)

Classificações simples são atribuídas a valores de R, F e M. Estas três métricas são designadas de forma independente. A interpretação de cada uma das três métricas do RFM é, portanto, inequívoca; uma pontuação de Frequency de 5 para um cliente significa o mesmo que uma pontuação de Frequency de 5 para outro cliente, independentemente de suas pontuações de Recency. Para amostras menores, isso tem a desvantagem de resultar em uma distribuição menos uniforme da classificação combinada do RFM [14].

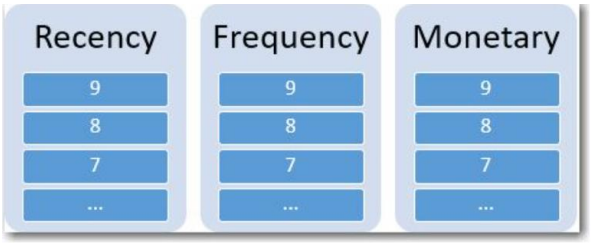


Figura 18 - Independent Binning Method

Categorização Aninhada (Nested Binning Method)

Na categorização aninhada, uma classificação simples é atribuída aos valores de Recency. Dentro de cada classificação de Recency, os clientes recebem uma classificação de Frequency e, dentro de cada classificação de Frequency, é atribuída uma classificação Monetary ao cliente.

Com isso pretende-se fornecer uma distribuição mais uniforme da classificação combinada de RFM, mas tem a desvantagem de tornar a classificação do Frequency e Monetary mais difíceis de interpretar.

Por exemplo, uma classificação de Frequency de 5 para um cliente com classificação de 5 em Recency pode não significar a mesma classificação que uma classificação de 5 para um cliente com classificação de 4, uma vez que a classificação de Frequency depende da classificação de Recency.

Figura 19 - Nested Binning Method

Se pretender uma interpretação rápida e simples, então será melhor utilizar a categorização independente.

Por outro lado, se for pretendido ter em consideração a mudança de compra ao longo do período para planear preços e promoções sazonais, então talvez a categorização aninhada seja o mais indicado [14].

As métricas RFM podem assumir múltiplas definições:

- Totais
    - R – Tempo que passou desde a última transação
    - F – Número total de transações
    - M – Soma do valor das transações feitas por cliente
        - Neste tipo de definição as transações só podem aumentar a importância do cliente na segmentação o que torna o modelo mais fácil de explicar.

- Média
    - R – Tempo que passou desde a última transação
    - F – Média de tempo entre cada transação
    - M- Média, em valor, por transação
        - Neste tipo de modelo as transações podem aumentar ou diminuir a importância do cliente na segmentação. Este modelo torna as campanhas mais complexas.

Atualmente e com o evoluir do estudo do RFM, já tem vindo a ser desenvolvidas novas variantes do RFM, duas das mais utilizadas são as seguintes:

- RFD (Recency, Frequency, Duration), onde a duração é obtida pelo tempo que o cliente passa na loja/site, assumindo que quanto mais tempo frequentar a loja, maior probabilidade de compra existe.
- RFE (Recency, Frequency, Engagment), um pouco similar ao RFD, no entanto é medido o comprometimento do cliente com a marca/produto, onde poderemos efetuar campanhas aos clientes baseado no seu comprometimento com determinados produtos.

3.2 Modelo CLV

Por vezes definidas como CLV (Customer Lifetime Value), ou por LTV (Lifetime Value), significa o “Valor do Tempo de Vida do Cliente”, onde se pretende calcular o valor potencial que pode a vir a ser gerado pelo cliente durante o seu período para com a empresa. No entanto existem diversas definições de diferentes autores.

"CVL é a soma dos fluxos de caixa líquidos descontados, obtidos pela empresa no decorrer do tempo de vida de relacionamento com cliente" [15].

"CLV é o valor que o cliente proporciona à empresa, resume-se ao montante total do desconto líquido da margem de contribuição em termos de tempo de cliente, que consiste na receita obtida a partir do cliente menos o custo associado à manutenção de uma relação com este. De forma simples, consiste no valor presente líquido de todos os ganhos que uma empresa espera receber do cliente ao longo do tempo" [16].

De igual forma se identifica o CLV em duas medidas distintas.

- CLV Histórico, é a soma de todos os lucros das compras do período passado do cliente.
- CLV Preditivo, permite prever quanto lucro um cliente vai gerar ao negócio durante a sua relação com o mesmo. Utiliza dados históricos junto com padrões de comportamento para determinar o valor atual do cliente e prever a sua evolução.

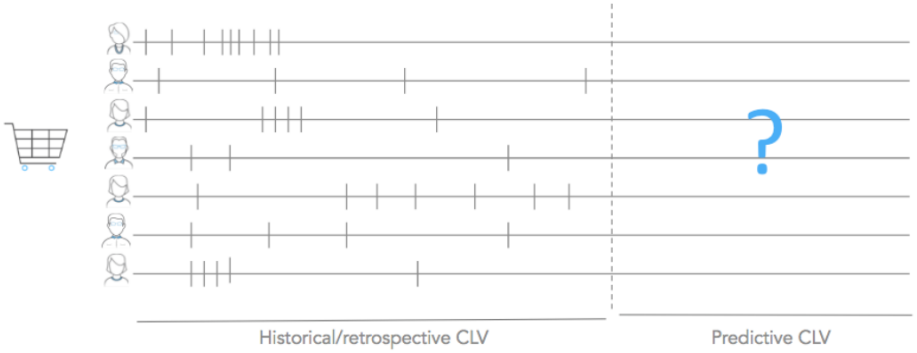


Figura 20 - CLV Histórico e Previsão[17]

Existem diversas e diferentes fórmulas para calcular o CLV, sendo um tema cada vez mais discutido, não existindo uma fórmula ideal ou standard. No entanto uma das fórmulas mais utilizadas é a que apresentamos de seguida, defendida por Gupta [18].

$$CLV = \sum_{t=0}^{T} \frac{(p_t - c_t) r_t}{(1+i)^t} - AC$$

p_t - Preço pago pelo consumidor no tempo t;

c_t - Custo direto de servir o cliente no tempo t;

i - Taxa de desconto ou custo do capital para a empresa;

r_t - Probabilidade do cliente repetir a compra ou estar "ativo" no tempo t;

AC - Custo de aquisição;

T - Horizonte de tempo para a estimação do CLV.

Figura 21 - Formula CLV

Na análise da equação apresentada na Figura 21, é possível verificar que o cálculo do CLV é influenciado por um conjunto de variáveis que determinam de forma efetiva o valor de um cliente, a saber, a margem $(p_t - c_t)$, a taxa de retenção (r_t), a taxa de desconto $(1+i)^t$ e os custos de aquisição de um cliente (AC).

Embora este seja o calculo mais utilizado, no projeto desenvolvido, não iremos optar por esta formula mas algo mais simples e muito utilizado no e-commerce e no retalho, onde se parte do pressuposto de calcular o valor do cliente (CV) multiplicando pela duração media do cliente (demonstrado na Figura 22), ou seja, o tempo em que ele permanece ativo antes de deixar de comprar ou ficar inativo.

Figura 22 - CLV Formula[19]

Todo este processo de calculo é explicado em pormenor na demonstração de resultados do CLV.

Para efeitos práticos o CLV permite a uma empresa perceber se deve ou não apostar num cliente, a exemplo, se o CLV do cliente é de 50€, sugere-se que a empresa gaste até um máximo desse montante para adquirir esse cliente. Se custa à empresa 40€ para adquirir esse cliente, a empresa deve ponderar fazê-lo, o que aumentará o valor da empresa em 10€, no entanto, se este custar à empresa 60€, a empresa deve decidir abdicar do cliente.

Assim com o CLV, o negócio passa a saber exatamente quanto vale para si um cliente em termos monetários e, qual a quantia exata a investir para o adquirir.

"Atrair e manter os mais altos clientes de valor é a base de um programa bem-sucedido de marketing" [20].

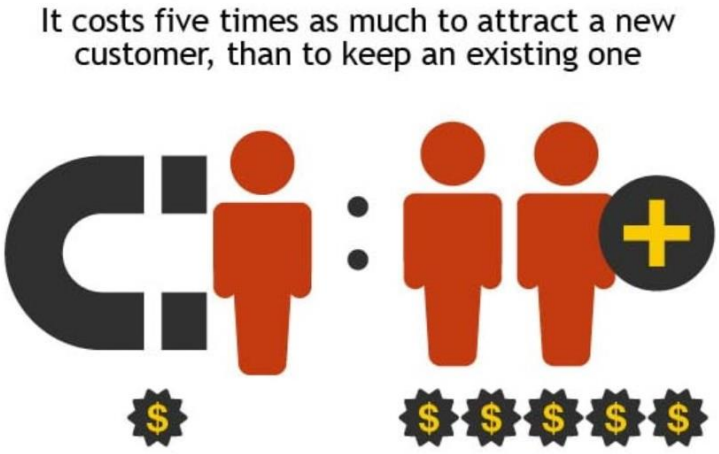


Figura 23 - Custo Cliente

3.3 Sistemas de Recomendação

Para construir um sistema de recomendação, é necessário dois pré-requisitos importantes, primeiro precisamos de obter os dados históricos, e posteriormente pré-processar os mesmos. Por norma são obtidos grandes blocos de dados, onde cada registo informa que o cliente X comprou o item Y.

Um sistema de recomendação consegue assim através de um dado conjunto de clientes, e um conjunto de produtos, e as diversas relações entre esses dois conjuntos, prever uma ou mais novas relações entre clientes e itens.

Como os seus algoritmos subjacentes permitem grande flexibilidade, os sistemas de recomendação podem ser adaptados a diversos cenários, à medida que novas necessidades possam surgir [21].

Os principais sistemas de recomendação dividem-se em duas principais categorias:

- Motores de Filtragem Colaborativa (Collaborative Filtering) - onde as escolhas do cliente X são comparadas a outros clientes; se a maioria das pessoas que compraram o artigo A e B, posteriormente também compraram o artigo C, então se o cliente X comprou o artigo A e B, pela relação anterior poderá fazer sentido o cliente X querer comprar o artigo C.

- Motores Baseados em Conteúdo (Content Based) - onde as similaridades dos itens são definidas de acordo com as propriedades dos próprios itens, ou seja, se o cliente X comprar o artigo A, o motor vai sugerir o item mais similar ao artigo A.

3.3.1 Filtragem Colaborativa

Para se efetuar um sistema de recomendação recorremos a uma técnica denominada de Filtragem Colaborativa (Collaborative Filtering) [22].

Filtragem Colaborativa (CF) é um algoritmo de recomendação popular que baseia as suas previsões e recomendações com base em ratings ou comportamento de outros clientes no sistema. O pressuposto fundamental por trás deste método é que as opiniões de outros clientes podem ser selecionadas e agregadas de modo a proporcionar uma previsão razoável da preferência do cliente. Intuitivamente, eles assumem que, se os clientes concordam sobre a qualidade ou relevância de alguns itens, então provavelmente eles vão concordar sobre outros itens. Se um grupo de clientes gosta das mesmas coisas que a Maria, então é provável que a Maria goste das mesmas coisas que o grupo e até mesmo de coisas que ela não tenha visto.

Para ser possível a aplicabilidade do algoritmo Collaborative Filtering é necessária efetuar os seguintes passos:

1. Transformar os dados numa matriz de frequências para emular uma matriz de ratings
2. Calcular uma matriz de similaridade com base na matriz de ratings
    a. Neste passo é calculada distância através da fórmula do cosseno
3. Neste passo é calculada a pontuação entre os utilizadores
    a. Com base no cluster criado percorremos a matriz de similaridade e procuramos os 3 vizinhos mais próximos. (k-nn)
    b. Obtemos o histórico dos mesmos 3 vizinhos mais próximos
    c. E com base nesse histórico calculamos uma pontuação para essa marca/utilizador
4. Ordenar de forma decrescente as preferências expectáveis para um determinado utilizador.

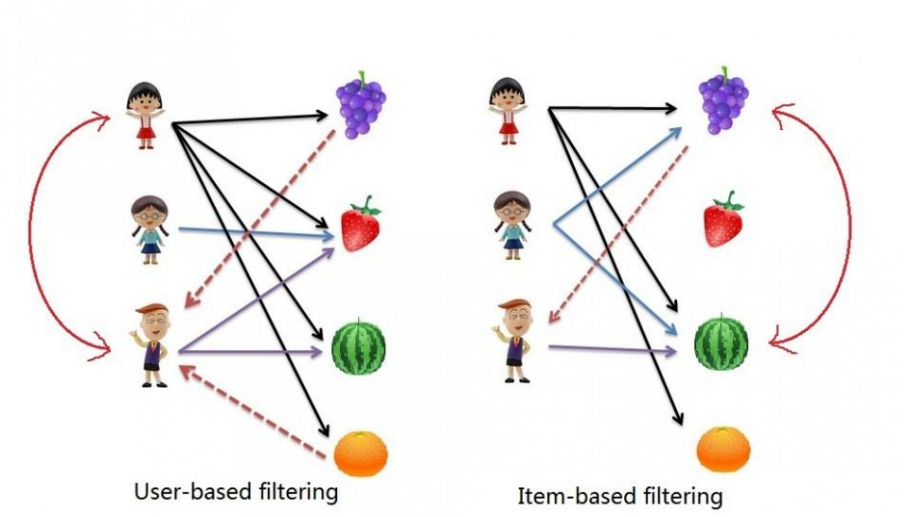


Figura 24 - Filtragem Colaborativa[23]

User Collaborative Filtering

É também conhecido como k-NN collaborative filtering, e foi dos primeiros dos métodos de filtragem colaborativa automáticos. Foi primeiramente introduzido pelo software de recomendação de artigos do GroupLens [24].

É um algoritmo com uma abordagem simples ao problema onde a sua premissa base é encontrar clientes onde o seu passado de comprar é semelhante ao cliente ativo nos dias de hoje e prever o que o cliente ativo poderá eventualmente gostar.

Para prever uma preferência sobre um item que nunca tenha sido comprado pelo cliente o algoritmo olha para outros clientes que tenham comprado o mesmo que o cliente ativo e faz uma previsão sobre itens que os clientes tenham comprado.

Item Collaborative Filtering

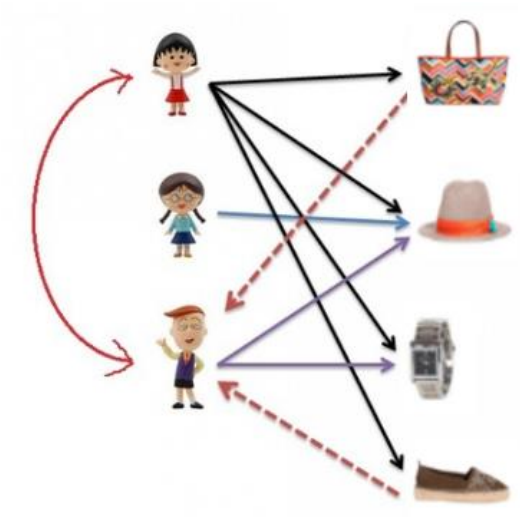
A filtragem colaborativa user-user sofre de problemas de escalabilidade á medida que a base de dados de clientes cresce.

A filtragem colaborativa item-item é um dos algoritmos mais utlizados nos dias de hoje e endereça a maior parte dos problemas da filtragem colaborativa user-user. A filtragem item-item surge pela primeira vez descrita na literatura em 2001 [25].

Em vez de utilizar as semelhanças entre os clientes para prever as preferências dos clientes a filtragem colaborativa item-item utiliza a semelhança entre os itens. Se dois itens podem ou não ser comprados por dois clientes diferentes, então estes clientes são semelhantes, sendo então espectável que tenham as mesmas preferências para itens semelhantes.

3.3.2 UBCF – User Based Collaborative Filtering

A Filtragem Colaborativa Baseada em Clientes analisa as semelhanças entre o histórico de consumo dos clientes.



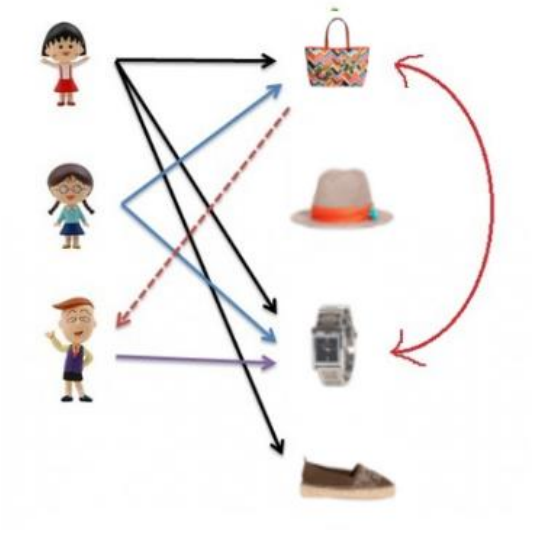
- A Maria comprou uma carteira, um chapéu, um relógio e uns sapatos.
- A Mafalda comprou um chapéu.
- O Frederico comprou um chapéu e um relógio.
- O motor de recomendação vai assim recomendar ao Frederico uma carteira e uns sapatos.

Figura 25 - User Based Collaborative Filtering[23]

A suposição é que clientes com preferências semelhantes avaliarão itens de maneira semelhante. Assim, a falta de classificações para um cliente pode ser prevista primeiro encontrando uma vizinhança de clientes semelhantes e, em seguida, agregar as classificações desses clientes para formar uma previsão [21].

3.3.3 IBCF – Item Based Collaborative Filtering

A Filtragem Colaborativa Baseada em Itens analisa as semelhanças entre o histórico de consumo dos itens.



- A Maria comprou uma carteira, um relógio e uns sapatos.
- A Mafalda comprou uma carteira e um relógio.
- O Frederico comprou um relógio.
- O motor de recomendação vai assim recomendar ao Frederico uma carteira.

Figura 26 - Item Based Collaborative Filtering[23]

É uma abordagem baseada num modelo que produz recomendações baseadas na relação entre itens inferidos da matriz de classificação. A suposição dessa abordagem é que os usuários preferirão itens semelhantes a outros itens de que gostam [21].

3.3.4 Wear It With

Com o objetivo de apoiar o cliente, e potenciar as vendas, a Parfois desenvolveu uma plataforma denominada Wear It With, onde é possível definir alguns “looks” para os clientes de acordo com a análise de tendências realizadas. Assim uma equipa especializada é responsável por definir periodicamente novos “looks” que posteriormente são utilizados para diversas estratégias de marketing junto dos clientes.

Como se verifica na Figura 27, são definidos vários “looks” onde são conjugados diversos produtos, de diversas gamas/categorias de produtos desenvolvidos e comercializados pela Parfois.

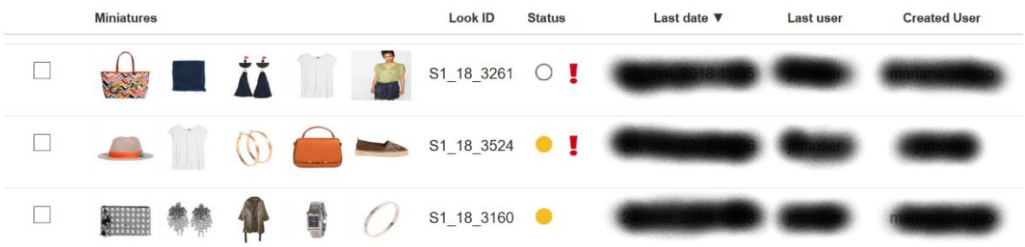


Figura 27 - Wear It With

Quando se analisa em detalhe um dos "looks" definidos conseguimos perceber como se constrói o mesmo, e que produtos o definem. Na Figura 28 é possível verificar que o look definido é constituído pelos seguintes produtos:

- Carteira de Mão
- Brincos
- Casaco
- Relógio
- Pulseira

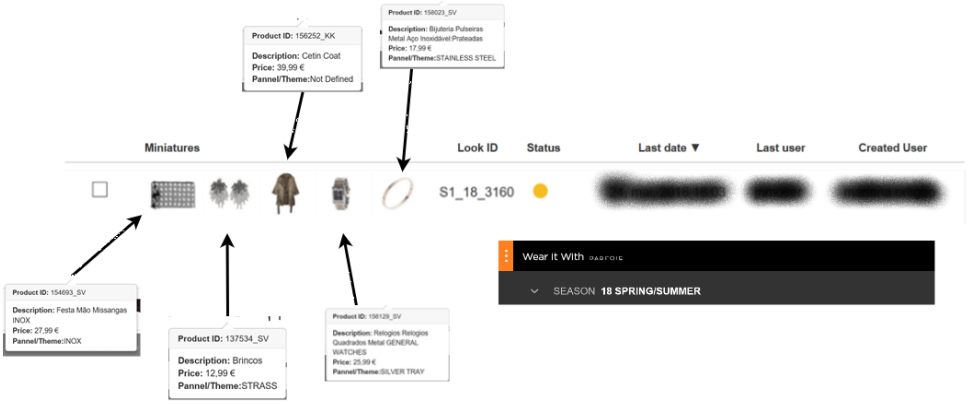


Figura 28 - Wear It With Details

Estes "looks" definidos e os seus produtos constituintes, são assim uma mais valia para o processo de recomendação aos clientes.

Como tal no projeto desenvolvido, esta componente de "looks" foi integrada no processo de recomendação, para sugerir/recomendar ao cliente novos produtos de acordo com o seu histórico de compras.

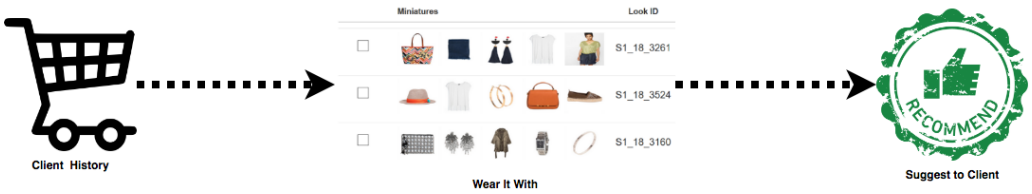


Figura 29 - Wear It With Recommend

3.3.5 Related Items

Com o objetivo de acrescentar valor ao cliente, mas também de potenciar as vendas, a Parfois decidiu evoluir os seus produtos e a relação entre estes. A nova definição de "Related Items", coloca assim esta nova definição "comercial" para os produtos e seus substitutos/equivalentes. De acordo com a análise de tendências realizadas, é possível potenciar a venda de um artigo equivalente, assumindo que o seu artigo original foi um best seller, assumindo que existem padrões de equivalência entre eles.

Existe então uma equipa responsável por definir periodicamente a relação entre produtos, que posteriormente são utilizados para diversas estratégias de marketing junto dos clientes.

Como podemos verificar na Figura 30 são definidos vários "related items" onde são indicados os artigos, e a sua relação.

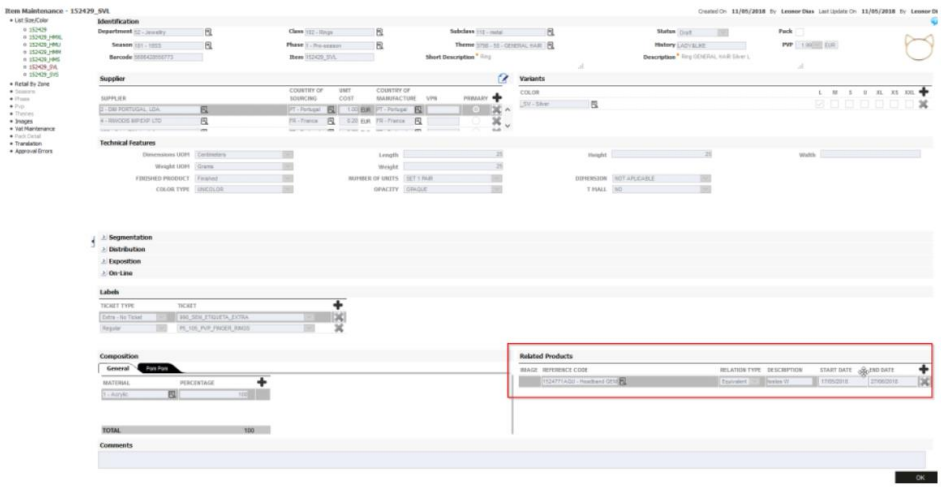


Figura 30 - Related Items

Pressupostos

- Relationship type = EQUI (equivalentes)

- Só são válidas relações equivalentes para as seguintes condições:
  - Têm o mesmo tamanho
  - Têm a mesma cor
  - Têm a mesma estrutura mercadológica (Dept, Class e Subclass) → Pertencentes ao mesmo Grupo

Como tal no projeto desenvolvido, esta componente de "related items" foi integrada no processo de recomendação, para sugerir/recomendar ao cliente novos produtos de acordo com o seu histórico de compras, no entanto o produto recomendado será sempre um produto para o qual existe uma relação de equivalência entre eles, que foi anteriormente definida pela equipa de compras e desenvolvimento do produto.

Figura 31 - Related Items Recommend

3.4 Data Warehouse

Data Warehouse (DW) é uma base de dados onde informação de diferentes fontes são armazenadas. Contém todos os dados históricos de cariz operacional e transacional de uma empresa, permitindo a análise de grandes volumes de dados, e sendo de fácil perceção e usabilidade ao utilizador final.

De salientar duas grandes figuras importantes na história dos Data Warehouse.

Kimball [26] defende uma arquitetura, denominada de Data Warehouse Bus Architecture, demonstrada na Figura 32. Onde o DW consiste no conjunto de diferentes áreas de negócios que cada Data Mart representa, e onde os dados nos Data Mart deverão estar desnormalizados.

Segundo Kimball um DW devera ter as seguintes características principais:

- Facilitar o acesso aos dados da empresa. O conteúdo do DW tem de ser compreendido e intuitivo para todos os utilizadores, e não só para os programadores.
- Adaptável à mudança e garantir que estas mudanças são feitas sem modificar, nem invalidar dados anteriores já existentes.

- Assegurar que os dados estão salvaguardados de possíveis intrusos.
- Apoiar na tomada de decisão.
- Aceitação por toda a comunidade para ser bem-sucedido.

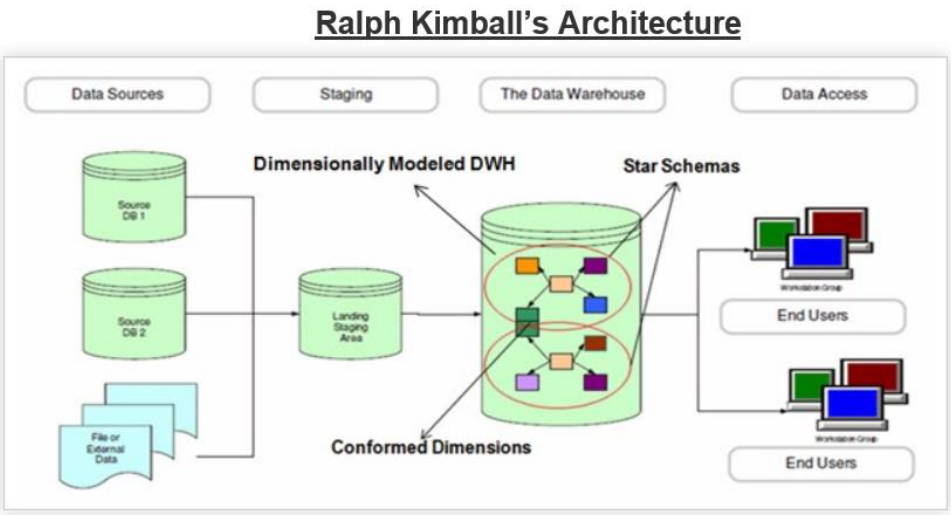


Figura 32 - Arquitetura Kimball

"The Data Warehouse is nothing more than the union of all the constituent data marts." [26]

Imnon [27] defende uma arquitetura onde os dados no DW se encontram normalizados. Já que este é um repositório integrado de todos os dados disponíveis. A desnormalização ocorre nos Data Marts, que representam diferentes processos de negócios. Esta arquitetura é conhecida como Corporate Information Factory, demonstrada na Figura 33.

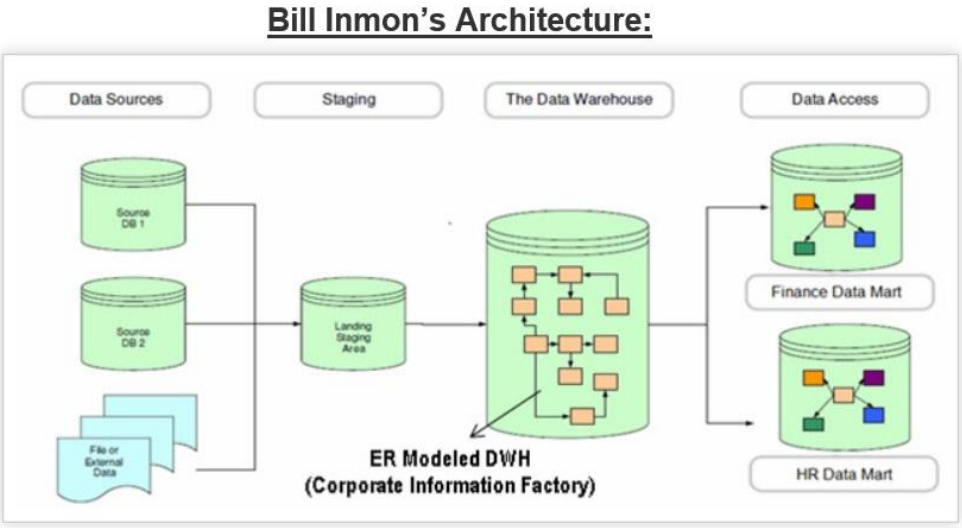


Figura 33 - Arquitetura Inmon

"A data warehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decisions." [27]

Segundo Inmon um DW deverá:

- Ser direcionado ao conteúdo e não à aplicação. Os sistemas operacionais clássicos estão organizados de acordo com as aplicações da empresa. A exemplo, para uma empresa de seguros as aplicações podem ser, seguro de saúde, seguro de vida e seguro automóvel. Num DW a informação estaria dividida pelos diversos tipos de seguros de saúde ou pela política escolhida no seguro de vida.

- Ter os dados associados a uma data, esta associação implica que cada registo deve ter associado um valor que permita enquadrá-lo num registo temporal, seja o momento de modificação ou o momento de carregamento. Desta forma é possível mostrar ao utilizador o valor exato para um determinado momento.

3.5 ETL

O processo ETL (Extract, Transform, Load) é o processo responsável por preencher o DW com os dados oriundos de diferentes bases de dados operacionais. É constituído por três etapas, como o nome indica: Extração, Transformação e Carregamento, visível na Figura 34.

Figura 34 - Etapas ETL

- **E - Extração-** consiste em ler e compreender os dados da fonte copiando os necessários para a staging área, a fim de os mesmos poderem ser manipulados. A principal preocupação a ter durante esta fase é a identificação de registos modificados.

- **T -Transformação-** consiste na limpeza e combinação dos dados, assim como, verificar se existe duplicação dos dados e fornecer novas chaves primárias. A principal preocupação nesta fase é a integridade dos dados.

- **L -Carregamento-** consiste em colocar os dados transformados ao dispor do utilizador, carregando-os para o DW. A principal preocupação nesta fase é o desempenho das consultas.

3.6 Data Mining

Data Mining é visto como um processo de exploração de dados, com o intuito de obter informação de valor acrescentado, nomeadamente padrões e conhecimento.

Na generalidade das vezes, e defendido por muitos autores, o processo de descoberta de conhecimento é definido em 7 etapas, conforme se verifica na Figura 35 , e que se passa a explicar [28].

- Etapa 1 – **Limpeza de Dados** – limpar dados inconsistentes
- Etapa 2 – **Integração de Dados** – combinação de dados de diferentes fontes.
- Etapa 3 – **Seleção de Dados** – seleção e recolha da informação relevante para a análise a realizar
- Etapa 4 – **Transformação de Dados** – consolidação e transformação de dados para os formatos necessários
- Etapa 5 – **Data Mining** – processo onde se aplicam sobre os dados, algoritmos ou métodos inteligentes para extrair padrões de dados
- Etapa 6 – **Avaliação Padrões** – análise e identificação dos padrões de conhecimento mais relevantes
- Etapa 7 – **Disponibilização Conhecimento** – representação do conhecimento obtido através de técnicas de apresentação e visualização aos utilizadores

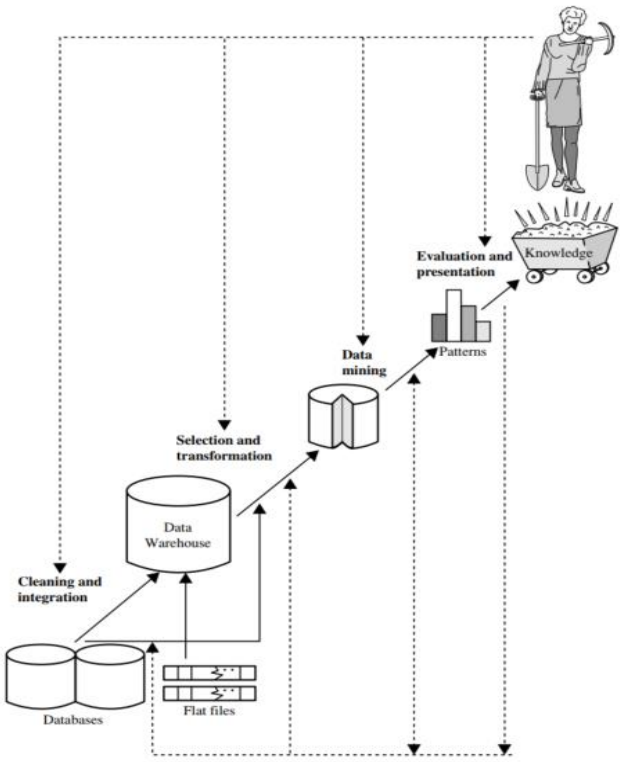


Figura 35 - Data Mining Discovery Steps

3.7 Clustering

Segmentação ou clustering consiste na criação de classes, subconjuntos de registos que representam valores mais próximos em certos atributos, produzindo um esquema de agrupamento que particiona o conjunto de dados em classes. Deve ser usado quando se pretende descobrir nos dados, grupos semelhantes de registos que partilham propriedades comuns sem quaisquer pré-condições acerca do que se possa entender por similaridade – operação não supervisionada [29]. Os resultados da segmentação podem ser usados de duas formas:

- Para resumir o conteúdo de cada segmento da base de dados considerando apenas as características mais relevantes de cada cluster e não de todos os seus registos;

- Como preparação de dados para outros métodos de Data Mining, por exemplo produção de modelos de classificação de cada um dos clusters descobertos.

Os métodos de segmentação ou Clustering são usados para construção de grupos de objetos com base nas semelhanças e diferenças entre os mesmos, de tal maneira que os grupos obtidos sejam os mais homogéneos e mais "distantes" possível [30].

Cálculo do número de clusters

Para calcular o número correto de clusters existem inúmeros índices, neste estudo recorremos a dois índices comumente utilizados:

- Índice Calinski-Harabasz[31]

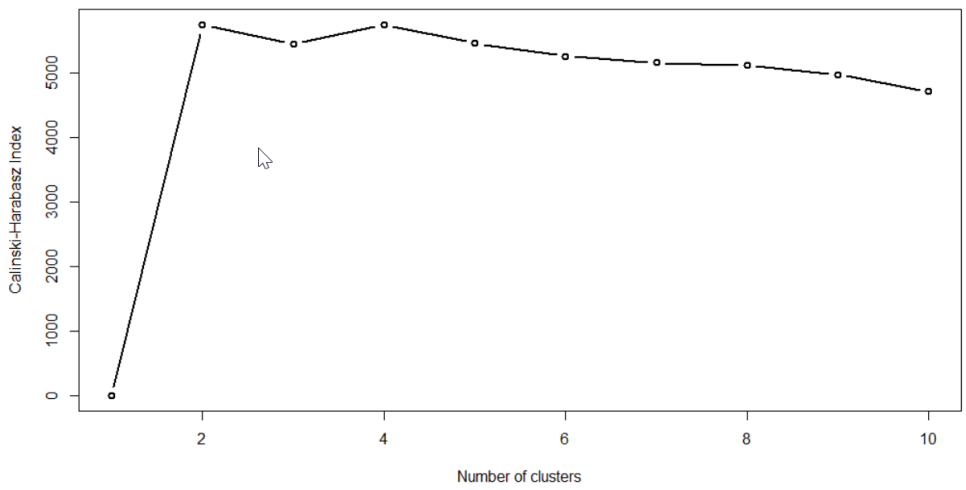


Figura 36 - Calinski-Harabasz

- Coeficiente Silhouette [32]

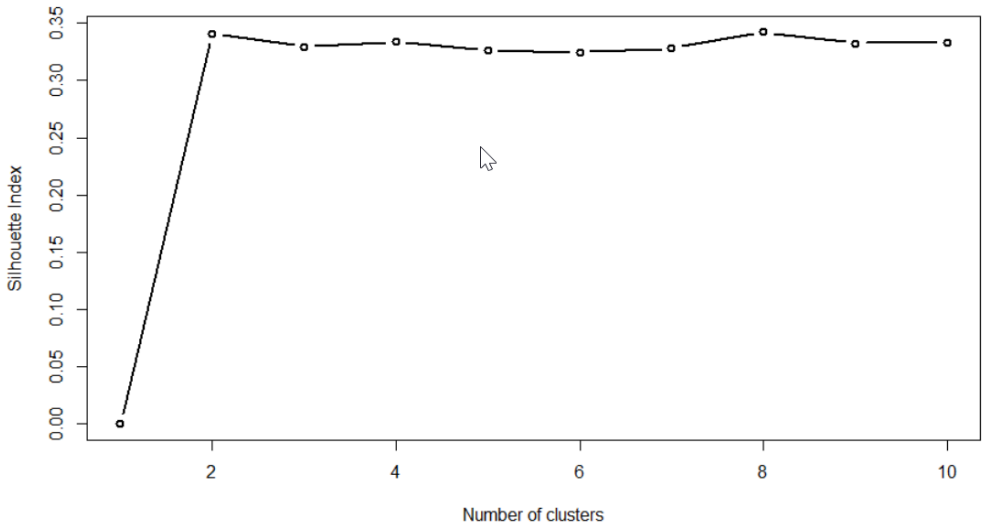


Figura 37 - Silhouette

O número correto de clusters corresponde ao valor de K para o qual exista um "joelho" distinto.

3.7.1 Algoritmo K Means

O algoritmo k-means é o algoritmo mais simples, mais popular e geralmente o mais usado que emprega o critério do erro quadrado [33] [34]. Começa com uma partição inicial aleatória e continua atribuindo aos clusters novas instâncias com base na similaridade entre a instância e o centro do cluster até que um critério de convergência conhecido seja atingido. Por exemplo, se não ocorrer nenhuma reatribuição de uma instância de um cluster a outro, ou o erro quadrado deixar de diminuir significativamente depois de algumas iterações [35].

O algoritmo de k-means é popular porque é fácil de implementar, e a sua complexidade temporal é:

- $O(n \times K \times d)$
  - n - Nº de objetos
  - K - Nº de clusters
  - d - Nº de atributos

O problema principal com este algoritmo é que é sensível à partição inicial e pode convergir para um mínimo local em função do valor do critério da partição inicial escolhida. Este algoritmo tem por função objetivo a soma do quadrado dos erros entre os objetos num cluster e o

respetivo centro. Este tipo de função permite obter bons resultados em clusters isolados e compactos [36].

Explicação Algoritmo K Means (Figura 38**)**

1. Passo - escolha de k centros de clusters coincidentes com k instâncias escolhidas aleatoriamente dentro do conjunto de instâncias;
2. Passo - atribui cada instância ao centro do cluster mais próximo;
3. Passo - recalcula o centro do cluster utilizando os membros correntes;
4. Passo - se o critério de convergência não é atingido, vai para o passo 2 [37].

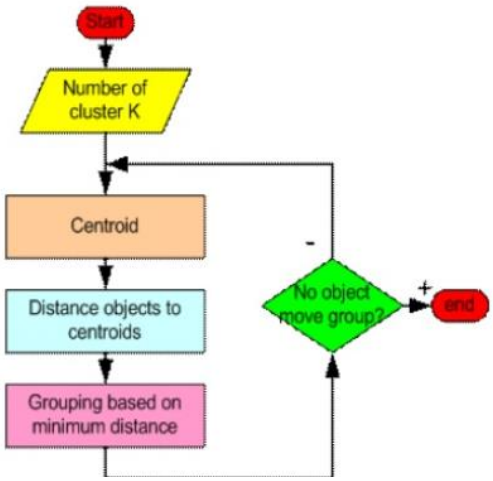


Figura 38 - K Means Passos

Verifica-se igualmente na Figura 39 uma demonstração da aplicabilidade do algoritmo K Means.

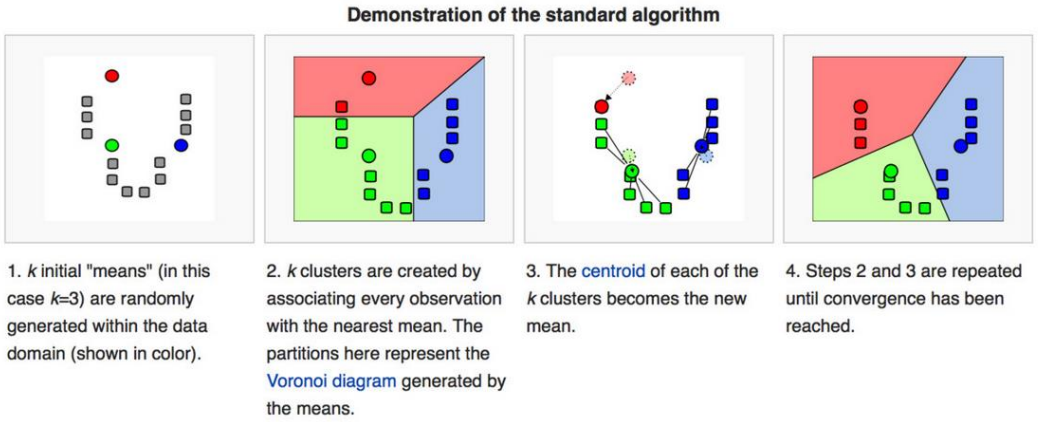


Figura 39 - K Means Demonstração[38]

3.8 Ferramentas Data Mining

Seguidamente se analisa algumas das ferramentas para a aplicabilidade de Data Mining.

3.8.1 Linguagem R

O R é um sistema de computação científica e estatística, programável e que permite o tratamento de vários tipos de dados. Na sua versão base possui um conjunto de ferramentas que permitem o armazenamento, processamento, cálculo, análise e visualização de dados. Possui ainda uma poderosa linguagem de programação que permite a implementação de novas funções com o comportamento definido pelo utilizador. Para além disso, é de acesso livre, existindo uma comunidade bastante ativa de investigadores (designada por CRAN) que desenvolvem funcionalidades que podem ser instaladas para estender as funcionalidades do sistema. [4]

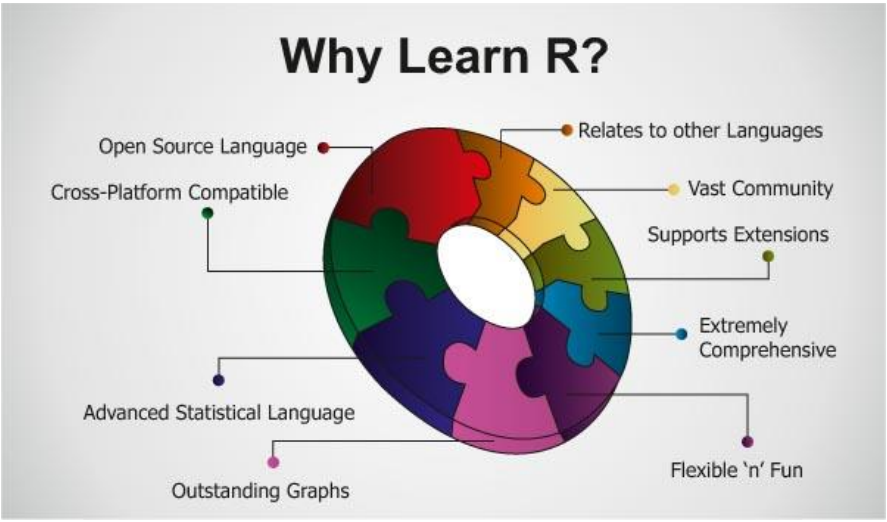


Figura 40 - Linguagem R

O R começou por ser essencialmente desenvolvido como um sistema de tratamento estatístico de dados, mas tem evoluído no sentido de se tornar num ambiente de desenvolvimento coerente, mais genérico e multifacetado. Ainda assim, existem disponíveis na versão base do R muitas ferramentas de tratamento estatístico de dados e muitas outras estão disponíveis para instalação adicional e opcional.

Esta linguagem é abundantemente usada entre estatísticos e analistas de dados para o desenvolvimento de software de estatística e análise de dados.

[4] Fonte - https://cran.r-project.org/doc/manuals/r-release/R-lang.html#Introduction

3.8.2 R Studio

O R Studio é um software IDE (integrated development environment) para a linguagem R, este é composto por uma consola, um editor de syntax capaz de executar comandos, e igualmente utilitários de gráficos, histórico, debug, entre outros.

Disponibilizado em versão open source e versão comercial, para praticamente todos os sistemas operativos, existindo hoje em versão para Desktop e Servidor.

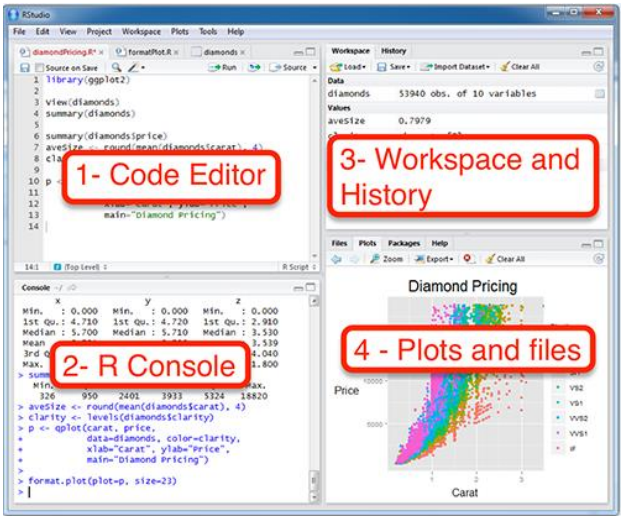


Figura 41 - R Studio Interface

Devido à sua fácil usabilidade tem vindo a crescer muito na comunidade universitária e científica.

Todo o desenvolvimento da plataforma foi realizado utilizando a ferramenta R Studio, que se trata de um sistema de computação científica, estatística e programável que está assente na linguagem de programação R, este sistema é bastante expansível uma vez que existem diversos *packages*/librarias disponíveis, assim como uma grande comunidade em redor desta linguagem de programação.

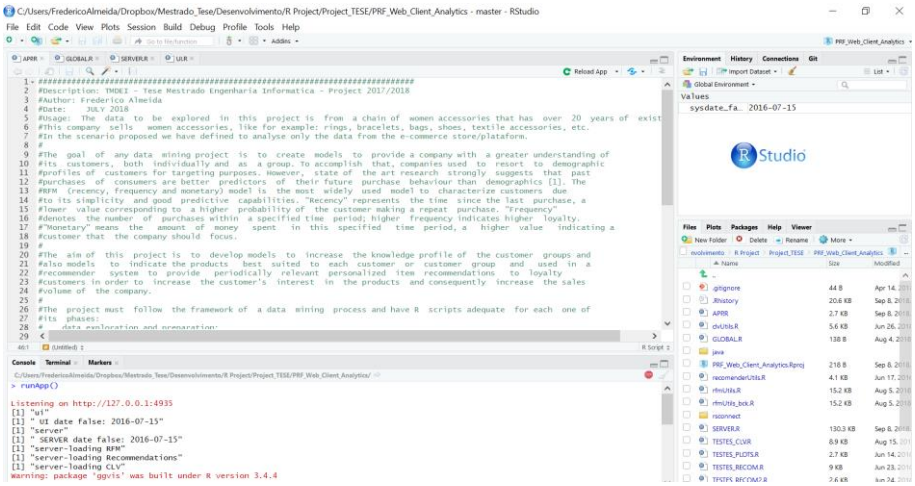


Figura 42 - R Studio

3.8.2.1 R Packages

Existe uma vasta comunidade de developers e utilizadores de R, com isso tem vindo a aumentar o número de pacotes disponíveis para o desenvolvimento de aplicações em R. Estes pacotes podem ser facilmente integrados e utilizados na aplicação aportando valor.

Figura 43 - R Packages

Para a usabilidade do R com maior eficiência, existe uma enorme comunidade de utilizadores que desenvolvem packages/librarias constituídos por diversas funcionalidades, estes podem facilmente ser integrados no desenvolvimento de qualquer aplicação R, permitindo assim ao developer tirar enorme vantagem destes.

No desenvolvimento realizado igualmente efetuamos a utilização de diversos packages/librarias, destacando entre os mais importantes os que de seguida iremos apresentar.

3.8.2.1.1 Shiny

Poderemos indicar que o package Shiny, foi dos packages mais explorados, uma vez que este é responsável pela interface gráfica, permitindo esta ser mais user friendly, possibilitando a criação de dashboards ou aplicações com uma interação bastante acessível, contendo em background todas as potencialidades do R.

Composto por dois componentes principais, o denominado UI (User Interface) e o Server, onde a combinação e comunicação entre estes disponibilizam a aplicação.

3.8.2.1.2 R Markdown

Este package permite através do R a realização de reports mais funcionais e dinâmicos, permitindo exportar para diversos formatos.

3.8.2.1.3 RJDBC

Com o package RJDBC é possível estabelecer uma conexão a base de dados através do padrão JDBC, ou seja, o JDBC é um conjunto de classes e interfaces (API) escritas em Java que permitem envio de instruções SQL para uma determinada base de dados.

3.8.2.1.4 Open XLSX

Para o processo de geração de ficheiros Excel, foi utilizado o package OpenXLSX que permite assim, ler e editar informação de Excel, mas igualmente a possibilidade de criar ficheiros Excel com a informação recolhida.

3.8.2.1.5 RODBC

Com o package RODBC é possível estabelecer uma conexão a base de dados através do padrão ODBC. Devendo igualmente o utilizador configurar no computador a respetiva ligação ODBC que ira ser utilizada por este package.

3.8.2.1.6 GGPLOT2

Indubitavelmente o package ggplot2 é um poderoso aliado para quem pretende disponibilizar a visualização de informação no formato de gráficos.

3.8.2.1.7 RecommenderLab

A utilização do package RecommenderLab permitiu a construção e avaliação de algoritmos de recomendação [21].

3.9 E-Commerce Solutions

Atualmente o E-commerce Parfois utiliza como sua plataforma web, a plataforma Salesforce, de origem americana e fundada em 1999 por Marc Benioff, um ex-executivo da Oracle. Bastante conhecida pelo seu CRM (Sales Cloud) também disponibiliza produtos para atendimento ao cliente, marketing, inteligência artificial, tudo com um funcionamento integrado na denominada "*Customer Success Platform*". Seguidamente vamos conhecer alguns dos seus componentes.

3.9.1 Salesforce Einstein

A plataforma Salesforce Einstein é uma componente de Inteligência Artificial (IA) para o CRM da Salesforce.

O Salesforce Einstein está incorporado em todos os módulos da Salesforce, e torna mais inteligente cada interação com o cliente, este descobre os focos de negócios mais importantes, prevê o que vai acontecer e a seguir, recomenda a melhor ação a ser tomada e, por fim, automatiza certas tarefas para libertar o tempo dos seus utilizadores.

Insights de oportunidade do Einstein

Verifique os principais insights de oportunidade, inclusive feedback do cliente, envolvimento com concorrentes e envolvimento geral com possíveis clientes para entender a probabilidade de o negócio se concretizar.

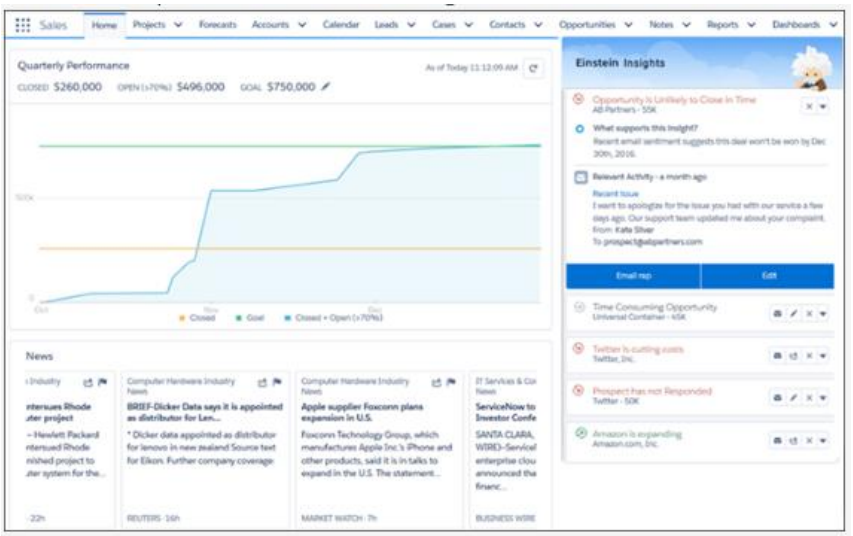


Figura 44 - Salesforce Insights

Recomendações do Einstein (Conteúdo preditivo e recomendações)

Recomende o melhor produto, conteúdo ou oferta para cada indivíduo na web, no email e mesmo em sites e aplicações móveis.

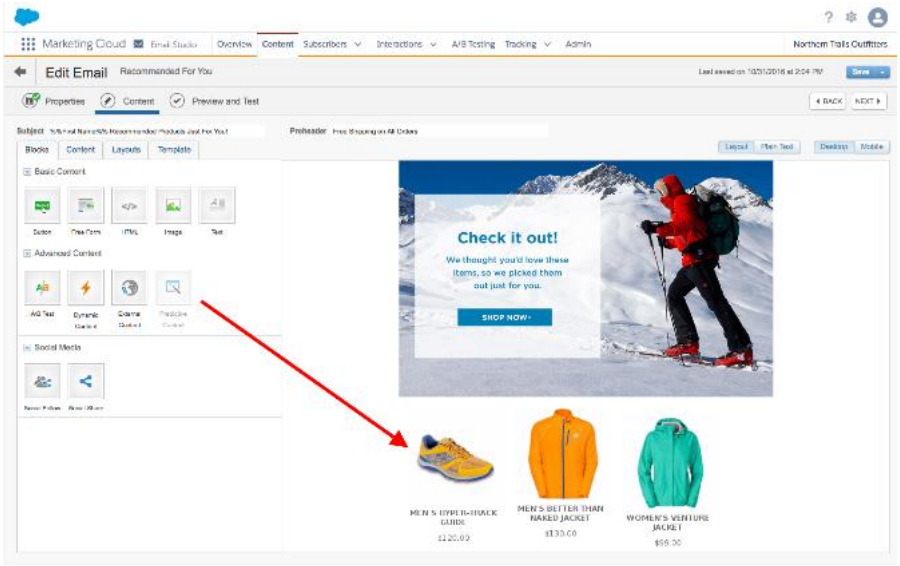


Figura 45 - Salesforce Recomendações

Classificação preditiva do Einstein

A classificação preditiva do Einstein usa dados do cliente para oferecer classificações de produtos mais personalizadas, diminuindo o tempo que um cliente gasta procurando pelo que ele quer e aumentando a conversão. A classificação preditiva personaliza os resultados da pesquisa explícita (pesquisa pela caixa de pesquisa) e também da pesquisa implícita (navegação pelo catálogo da loja).

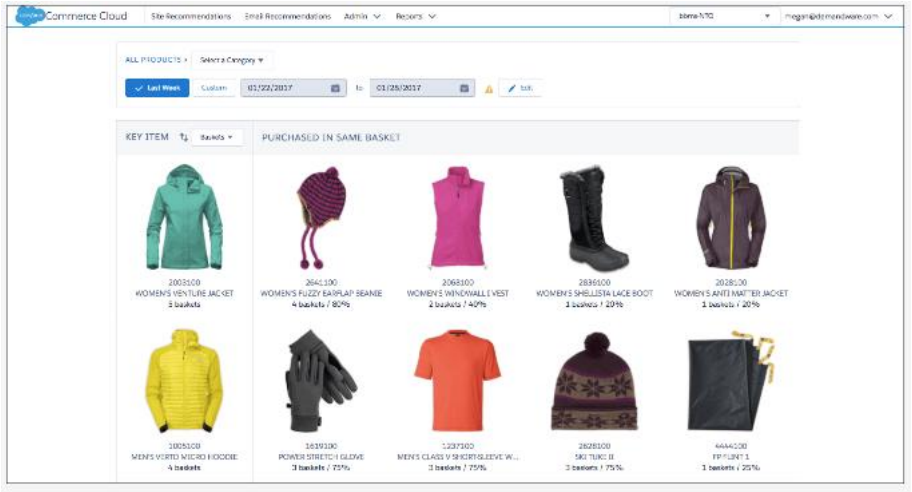


Figura 46 - Salesforce Preditiva

Alguns dos módulos do Einstein Salesforce.

Einstein do Sales Cloud	Orientar os vendedores para os melhores leads e oportunidades para que eles possam se concentrar em fechar os negócios certos
Einstein do Service Cloud	Prestar um serviço proativo, ajudando os clientes a encontrar suas próprias respostas e recomendar o conteúdo certo para que o Customer Care possa resolver os casos mais rapidamente
Einstein do Marketing Cloud	Ajudar os profissionais de marketing a criar campanhas de marketing mais personalizadas para prever os prováveis próximos passos dos clientes e recomendar conteúdo e produtos com base no canal e nas preferências do público
Einstein do Analytics Cloud	Automatizar e priorizar insights que você precisa saber no futuro
Einstein do Commerce Cloud	Personalizar as experiências dos clientes ao recomendar as ofertas e produtos certos no momento certo para impulsionar o compromisso, maximizar as conversões e aumentar o valor do pedido
Einstein do Salesforce IoT	Automatizar e prever eventos, além de permitir que os utilizadores finais atuem de acordo com seus insights mais importantes

De referir que os módulos apresentados, se encontram disponíveis para comercialização, no entanto não foram adquiridos ainda para a Parfois, daí se ter apostado no desenvolvimento interno de uma solução.

3.10 Casos de Estudo Semelhantes

3.10.1 Estudos sobre RFM

No estudo realizado na Turquia [44], foi utilizado um conjunto de dados da plataforma e-commerce de uma loja de desporto na Turquia, para demonstrar os benefícios da utilização do modelo desenvolvido.

Este modelo propõem uma nova abordagem utilizando o RFM em três etapas, incluindo o clustering, regras de classificação, e regras de associação para providenciar alguma inteligência ao processo de análise de mercado, e potenciar melhores estratégias de marketing.

Na primeira etapa, os segmentos de clientes com valores semelhantes de RFM são identificados para poder adotar diferentes estratégias de marketing para diferentes segmentos de clientes.

Já na segunda etapa, as regras de classificação são exploradas usando variáveis demográficas (idade, sexo, nível de escolaridade, etc.) e valores de RFM de segmentos de clientes para prever comportamentos futuros de clientes e segmentar perfis de clientes de forma mais clara.

Na terceira e última etapa, são exploradas regras de associação para identificar as associações entre segmentos de clientes, perfis de clientes e itens de produtos adquiridos e, consequentemente, recomendar produtos com classificações associadas, o que resulta em melhor satisfação do cliente e vendas cruzadas [44].

No estudo realizado por Jo-Ting Wei, Shih-Yen Lin, e Hsin-Hung Wu_[12], era pretendido efetuar uma análise das diversas aplicabilidades do modelo RFM, tentando responder as questões:

- Quais são as definições e esquemas de pontuação do modelo de RFM?
- Como o modelo de RFM é aplicado?
- Quais são as vantagens e desvantagens do modelo RFM?
- Quais são as vantagens e desvantagens relativas do modelo de RFM e outros modelos?
- Como o RFM é combinado com outras variáveis ou outros modelos?

Esta revisão do modelo de RFM é essencial e pode fornecer informações proveitosas para investigadores e decisores. Na verdade, o modelo de RFM provou ser muito bem-sucedido numa variedade de áreas práticas.

Como tal, o RFM pode ajudar a identificar clientes valiosos e desenvolver uma estratégia de marketing eficaz não apenas para organizações de lucro, mas também organizações sem fins lucrativos e governamentais.

Para os investigadores, eles podem obter uma explicação completa sobre a visão geral do modelo de RFM, para que possam ter mais ideias sobre a aplicação refinada do RFM [12].

3.10.2 Estudos sobre CLV

No estudo realizado por Mahboubeh e seus colegas [45], foi utilizado um conjunto de dados de uma empresa de produtos de beleza e saúde, para realizar a segmentação de clientes através do CLV.

Neste são utilizadas duas abordagens diferentes. Na primeira utiliza-se a análise RFM para a segmentação de clientes, já na segunda abordagem é utilizado o RFM considerando um parâmetro adicional denominado Item Count.

Foi realizado o agrupamento de clientes em segmentos de acordo com os parâmetros RFM e Extended RFM usando o algoritmo K-means. Agrupar os clientes em diferentes grupos ajuda os tomadores de decisão a identificar os segmentos de mercado com mais clareza e, assim, desenvolver estratégias de marketing e venda mais eficazes para a retenção de clientes.

A comparação dos resultados dessas abordagens mostra que a adição do Item Count como um novo parâmetro ao RFM não faz diferença no resultado do agrupamento, portanto, o CLV é calculado com base no método de RFM ponderado para cada segmento.

Como os pesos do RFM variam de acordo com as características do setor, o método AHP foi aplicado para determinar a importância relativa das variáveis do RFM com base no ponto de vista do especialista

De acordo com parâmetros ponderados de RFM, o valor de CLV foi calculado para cada segmento de cliente. Em seguida, a classificação CLV foi atribuída a cada segmento com base em seu valor de CLV. O valor atual fornece um ponto de vista financeiro e o valor potencial indica oportunidades de vendas cruzadas.

Os resultados do CLV calculado para diferentes segmentos podem ser usados para explicar estratégias de marketing e vendas pela empresa [45].

Um outro estudo e como resultado de intensas discussões acerca do CLV no decorrer da "Thought Leadership Conference" organizado pela Universidade do Connecticut, entre académicos, especialistas e implementadores, foi produzido um estudo [18], que analisa vários modelos de CLV implementados que são úteis para a segmentação de mercado e a alocação de recursos de marketing para aquisição, retenção e venda cruzada. Os autores revisam vários insights empíricos que foram obtidos a partir desses modelos, concluindo que ainda existem diversas áreas que necessitam de mais pesquisa e estudo.

Primeiro, é apresentada uma estrutura conceitual que mostra como o CLV se adapta na cadeia de valor e quais são seus principais impulsionadores. Em seguida, são apresentadas várias abordagens de modelagem que foram adotadas para abordar o CLV. Essas abordagens variam de modelos econométricos a técnicas de ciência da computação. Realiza-se igualmente uma discussão detalhada de áreas para pesquisas futuras, terminando o estudo com algumas observações finais [18].

4 Design da Solução

O capítulo “Design da Solução” pretende dar ao leitor uma visão da arquitetura do projeto proposto, abordando cada um dos diferentes componentes.

4.1 Arquitetura

No seguimento da abordagem ao modelo realizada anteriormente iremos agora detalhar um pouco mais cada um dos três componentes representados neste modelo.

A componente “Dados” é a responsável pelos dados disponíveis para a aplicabilidade dos algoritmos na componente de “Inteligência”, que por sua vez após esse mesmo processo ficam disponíveis para realização de Reports e Dashboards na componente “Interface”.

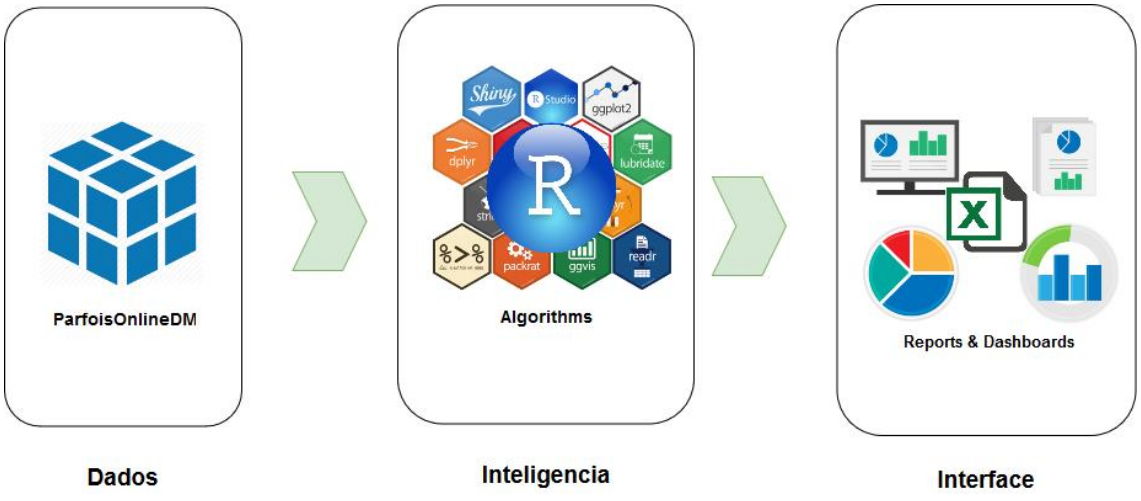


Figura 47 - Modelo Componentes

Estes três componentes estão diretamente encadeados e dependentes, uma vez que cada um dos componentes necessita que os processos desenvolvidos no seu componente antecedente tenham sucesso. Ou seja, se aquando a interação do utilizador na componente de "Interface" esta efetua um pedido à componente "Inteligência" e esta por sua vez tenta obter os conteúdos da componente "Dados", mas se os dados não estão disponíveis, então o utilizador vai ver o seu pedido rejeitado/negado devido à indisponibilidade/inexistência de dados. Assim como igualmente pode acontecer se a componente de "Inteligência" não conseguir aplicar os algoritmos, da mesma forma o utilizador vai ver o seu pedido rejeitado/negado. Assim de certo modo e analisando numa perspetiva de processo contínuo podemos indicar que eles são dependentes entre si.

Iremos seguidamente dentro deste capítulo abordar individualmente cada um destes componentes.

4.1.1 Dados

A componente "Dados" é o ponto de partida, uma vez que a reduzida exploração de dados é um dos fatores de lançamento deste projeto. Será este componente o responsável pela nossa fonte de dados, que após diversas análises e tratamentos deverá ser disponibilizada num Cubo denominado "ParfoisOnlineDM".

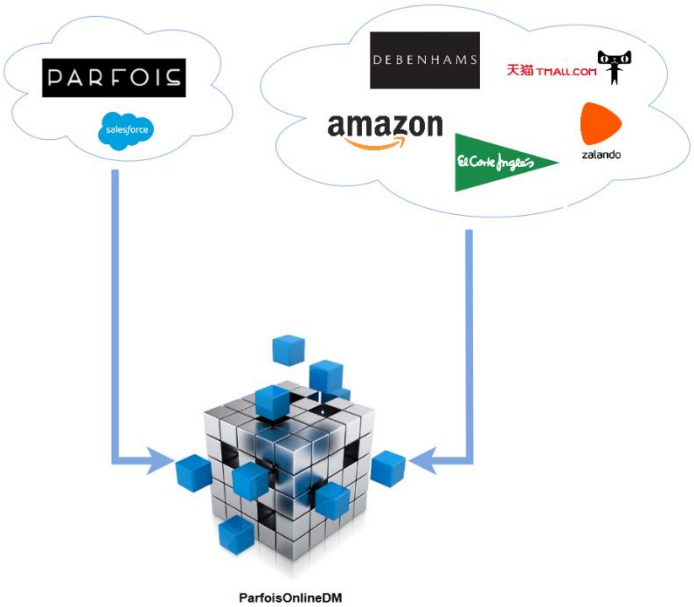


Figura 48 - Data Warehouse Parfois

Não obstante este tema será abordado com mais detalhe para se perceber corretamente os dados, as suas fontes e a sua evolução.

A componente “Dados” é conforme se pode constatar na Figura 49 a construção de um data warehouse, seguindo a metodologia de Kimball é realizado o processo ETL constituído por três etapas, como o nome indica: Extração, Transformação e Carregamento.

Iniciando na etapa de Extração dos dados de distintas fontes, nomeadamente no ERP da Parfois, onde se encontram as faturas e guias realizadas para os clientes finais e MarketPlaces, e igualmente na plataforma Verdinho, onde se verificam os dados de clientes e das suas encomendas. Igualmente existem outras fontes menos relevantes que são visíveis, mas que não serão abordadas. Posteriormente é realizada a etapa de Transformação responsável pela limpeza, tratamento e transformação de dados, com o intuito de garantir a integridade dos dados. Por último realiza-se a etapa de Carregamento onde é realizada a passagem dos dados para o DW.

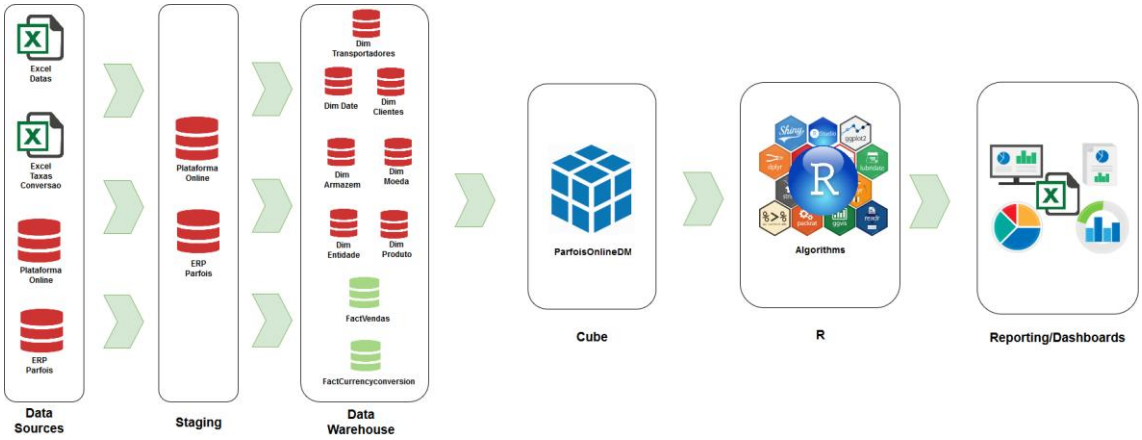


Figura 49 - Arquitetura "Dados"

Embora seja visível na Figura 49 toda a definição da construção do DW, nomeadamente a sua arquitetura, não é previsível que a mesma seja desenvolvida ao longo deste projeto, mas sim utilizado o seu cubo final como fonte de dados.

Uma vez definido o DW este fornecerá os dados ao cubo OLAP denominado “ParfoisOnlineDM” que será a fonte de dados do módulo Inteligência.

4.1.2 Inteligência

A componente “Inteligência” é indubitavelmente a mais complexa de todas, uma vez que é nesta etapa que são aplicados aos dados os diferentes algoritmos com o intuito de responder aos objetivos propostos.

Para o desenvolvimento e aplicabilidade dos algoritmos iremos utilizar um sistema de computação científica, estatística e programável denominada por linguagem de programação R, anteriormente apresentada.

Conforme se verifica na Figura 50 pretende-se aplicar sobre os dados uma diversidade de diferentes processos/técnicas, usando estes alguns algoritmos para assim obter os resultados esperados. Embora seja expectável receber dados com qualidade e devidamente tratados, é igualmente necessário realizar um pré processamento dos mesmos para garantir o sucesso dos próximos processos.

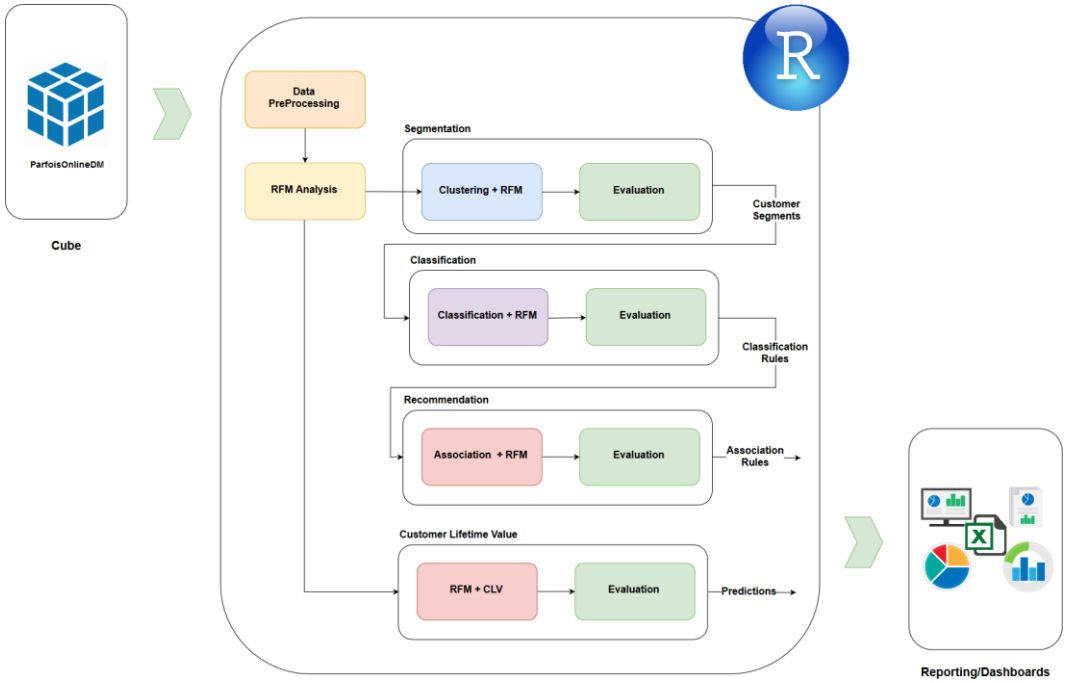


Figura 50 - Arquitetura "Inteligência"

4.1.3 Interface

A componente “Interface” é onde o valor, a criatividade e o “user friendly” se devem conjugar, resultando em Reports e Dashboards que disponibilizem a informação necessária no momento certo aos utilizadores. Deve-se garantir que esta plataforma seja intuitiva quer para o utilizador inexperiente assim como cumpra os requisitos do utilizador avançado, nomeadamente o Business & Web Analist.

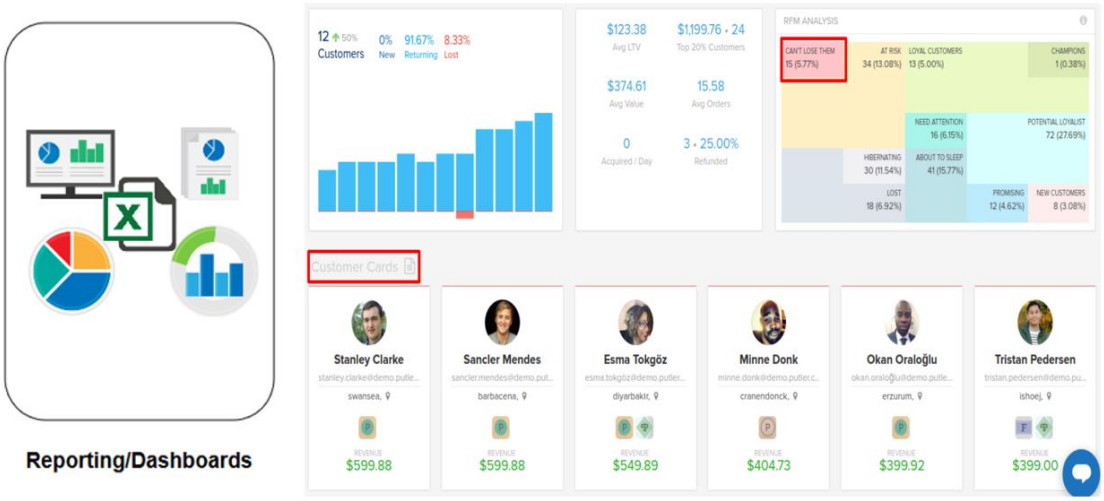


Figura 51 - Exemplo Dashboards para Utilizador

É pretendido realizar o desenvolvimento desta plataforma utilizando a interface gráfica Shiny que consiste num package que está integrado no R, e que permite a criação de uma interface web bastante user friendly e com interação direta com o R.

5 Implementação da Solução

O capítulo “Implementação da Solução” incide no detalhe sobre o desenvolvimento realizado, focando para o leitor os pontos de maior importância, garantindo que a sua compreensão é pactuante com o projeto efetivamente implementado. São descriminadas cada uma das diferentes fases de acordo com o fluxo de desenvolvimento realizado, e interpretando cada uma destas de forma explicativa para o leitor.

5.1 Dados

A componente “Dados” é de extrema importância para a plataforma, uma vez que garantir a qualidade dos mesmos, é fulcral para orientar o Negócio nas corretas decisões. No entanto este processo obedece a várias etapas conforme já explicamos anteriormente. Pretende-se nos próximos parágrafos explicar ao leitor como foi desenvolvido este processo.

5.1.1 Origem dos Dados

Para a avaliação dos dados fundamentais para a plataforma foi necessário efetuar diversos estudos com base nos inputs do Negócio para conseguir concretizar os seus objetivos. Para tal ser possível é necessário perceber corretamente as necessidades deste, quais as suas diversas análises mais frequentes, quais as métricas mais importantes, quais os campos/atributos essenciais.

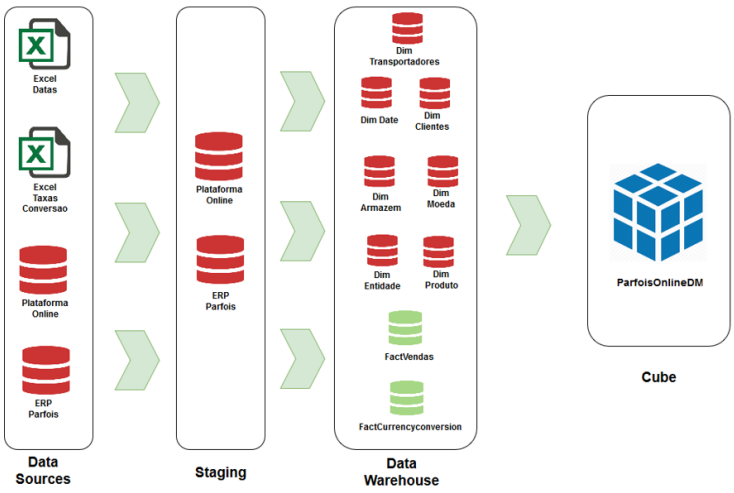


Figura 52 - Arquitetura Dados Desejável

Numa fase inicial foi avaliada a possibilidade de utilizar o cubo ParfoisOnlineDM cujo desenvolvimento do mesmo está a ser realizado internamente pela empresa, conforme transcrito neste documento, e com a arquitetura apresentada na Figura 52. No entanto ao longo do desenvolvimento verificamos que parte das necessidades eram inexistentes neste cubo, e a sua adaptação era morosa e não prioritária face aos desenvolvimentos daquela equipa.

Não existindo toda a informação no denominado cubo ParfoisOnlineDM, com esta limitação, urge a necessidade de construir alternativas para o cumprimento dos objetivos propostos para este projeto.

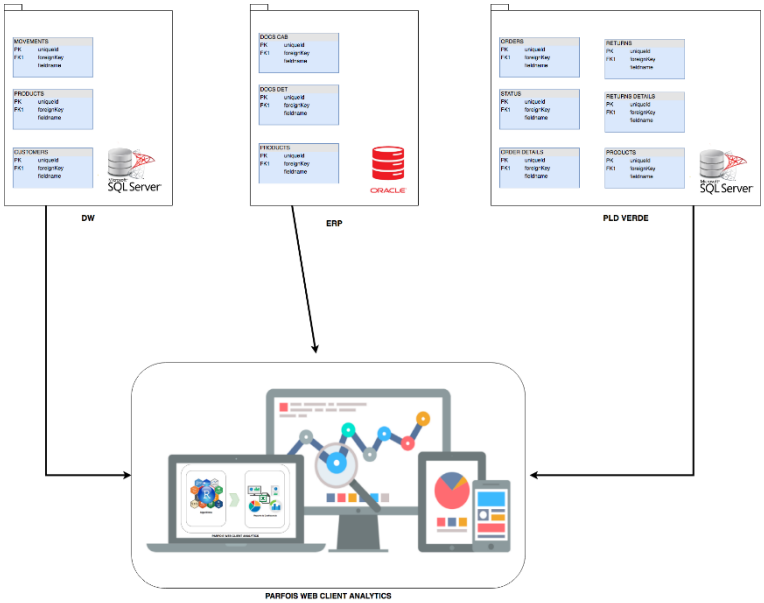


Figura 53 - Arquitetura Dados Implementada

Foi então avaliada a possibilidade de a plataforma se conectar diretamente às fontes de dados tal como acontece com o cubo, numa perspetiva de somente consultar informação, com regras e perfis bem delineados.

Após analisada e validada essa alternativa, realizamos diretamente na plataforma a definição de conexões para as diferentes fontes de dados, conforme se demonstra na Figura 53.

5.1.2 Conjunto de Dados (dataset)

No âmbito do desenvolvimento deste projeto foram utilizados dados de uma empresa de retalho de acessórios femininos, nomeadamente a empresa Parfois, referentes à plataforma de e-commerce da empresa para um período de dois anos, particularmente entre 2015 e 2016.

Este conjunto de dados disponível para análise totalizava números bastante interessantes, nomeadamente:

- Período entre 01-01-2015 a 31-12-2016 (24 meses)
- 172.461 Encomendas
- 24,3€ Valor Médio Encomenda
- 2 Un. Quantidade Média por Encomenda
- 162.681 Faturas
- 128.032 Clientes Diferentes
- 18.123 Artigos Diferentes
- 4.093 Devoluções

Das encomendas existentes as mesmas somente se referem à plataforma PARFOIS.COM descartando os restantes MarketPlaces (exemplo da Amazon, El Corte Inglês, etc.).

Dos vários campos existentes para as Encomendas, Faturas e Devoluções foram utilizados o código encomenda, produto, código cliente, data encomenda, valor encomenda, quantidade encomenda, tipo encomenda, etc.

Referente aos artigos o seu principal campo foi o denominado SKU, assim como a descrição do artigo, e gamas/departamento.

Por sua vez os dados dos clientes não tiveram grande usabilidade uma vez que estão protegidos ao abrigo do GDPR/RGPD, e de momento a sua utilização encontra-se limitada.

5.1.3 Transformação e Problemas com os Dados

Tal como em qualquer projeto de exploração de dados, a etapa de transformação deve ser cuidadosamente realizada, e carece de bastante dedicação. A limpeza e manipulação de dados deve ser uma decisão bem ponderada, uma vez que poderá ter impacto para o futuro.

Assumindo as três conexões existentes para as diferentes fontes de dados, cada uma destas mereceu a sua atenção e resoluções diferenciadas.

O grande dilema dos dados, é como transformar os mesmos em informação relevante e de qualidade para com estes gerar conhecimento.

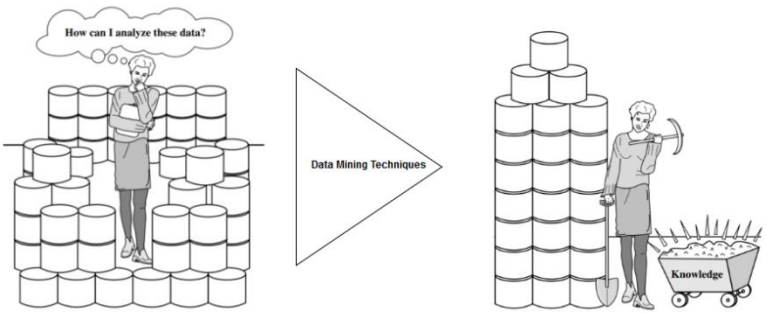


Figura 54 - Data Mining, Data Problem

Avaliamos de seguida cada uma das conexões revelando as técnicas utilizadas nos dados em cada um destes. Nas três conexões, conexão DW, conexão PLD e conexão ERP, essencialmente transformaram-se formatos de datas e removeram-se colunas menos relevantes.

Problemas de Dados

- Diferentes Motores de Base de Dados, usabilidade de Oracle e Microsoft SQL.
- Uniformização de Clientes, verificado que o mesmo email poderá existir em diversos códigos de cliente.
- Segmentação por Idades, não é obrigatório em site a colocação de datas de nascimento, como tal a amostra de dados preenchidos é reduzida não podendo ser usado para segmentar.
- Validação de Países e Códigos Postais, o site não realiza essas validações, como tal surgem diversos erros, que impedem a segmentação por países/zonas, etc.

- Fidelização de Cliente, não obrigatoriedade de ficha do cliente para a realização de compras, existindo o denominado “Convidado”.

5.2 Arquitetura Projeto R

O projeto R desenvolvido e que resulta na plataforma “Parfois Web Client Analytics”, obedece à arquitetura comum dos projetos R.

5.2.1 R Files

Durante todo o desenvolvimento da plataforma, foram sendo desenvolvidos cada um dos principais processos que iriam fazer parte da plataforma, como tal os mesmos foram desenvolvidos individualmente e de forma integrada como um todo. De seguida será demonstrada a estrutura adotada na Figura 55.

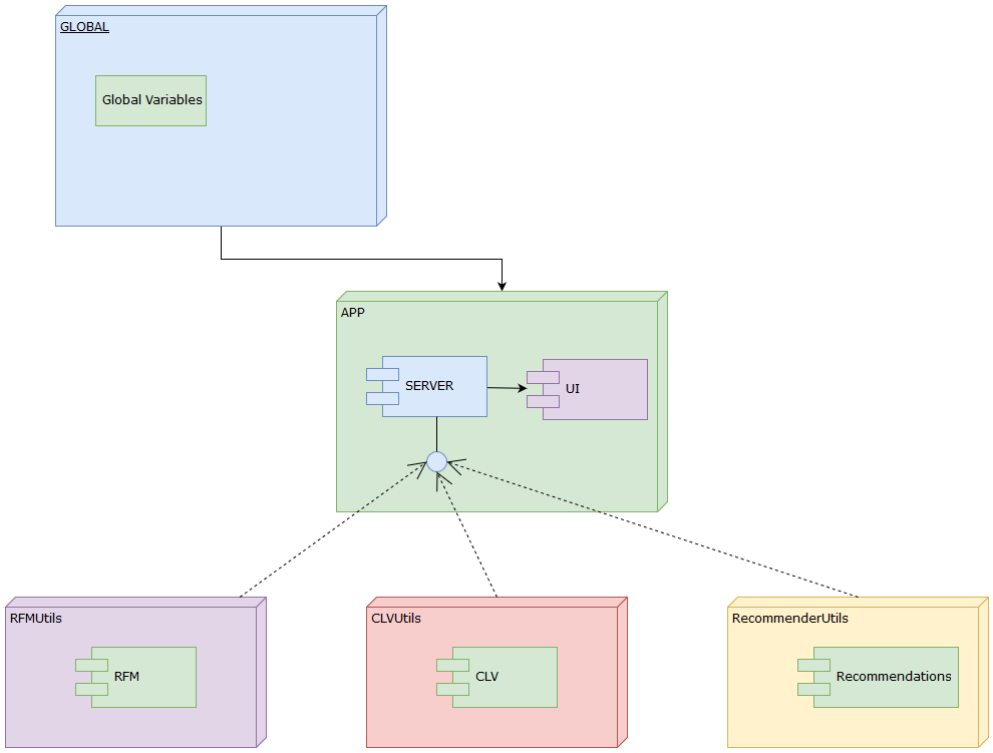


Figura 55 - R Files

O bloco GLOBAL é onde são definidas as variáveis globais a serem utilizadas pela plataforma.

O bloco APP é o principal componente, onde este invoca o modulo SERVER e o modulo UI, através do package Shiny.

No módulo UI, é onde é definida toda a interface gráfica do utilizador, responsável pela construção da estrutura dos diversos menus laterais, do painel principal e respetivos tabs existentes nestes. Este invoca as diversas funcionalidades desenvolvidas no modulo SERVER, assim como a utilização de alguns packages.

No módulo SERVER é onde está construída toda a inteligência da plataforma, com as diversas funcionalidades invocadas pelo modulo UI. Para a concretização dos distintos processos este utiliza diversos packages, assim como faz uso dos diferentes blocos RFMUtils, CLVUtils e RecommenderUtils.

O bloco RFMUtils integra o modulo RFM, onde está construída toda a estrutura de funções necessárias para a implementação do RFM de clientes.

O bloco RecommenderUtils integra o modulo Recommendations, onde existe toda a estrutura de funções necessárias para a utilização dos diferentes algoritmos de recomendação de clientes e artigos.

O bloco CLVUtils integra o modulo CLV, onde está construída toda a estrutura de funções necessárias para a implementação do CLV de clientes.

5.2.2 R Publish to Web

Com o objetivo de proporcionar ao utilizador uma maior mobilidade, assim como obter capacidades de performance melhoradas, foi explorada a possibilidade de publicar para a Web/Cloud o projeto, uma vez que se necessitava de colocar um protótipo para uma fase de avaliação dos utilizadores.

Para essa mesma fase de avaliação junto dos utilizadores, efetuamos assim a avaliação das plataformas disponíveis, escolhendo utilizar a *SHINYAPPS.IO* que detém de uma integração com o RStudio, permitindo ao programador publicar facilmente as suas aplicações Shiny para essa mesma plataforma conforme demonstrado na Figura 56 - Publicação para Web.

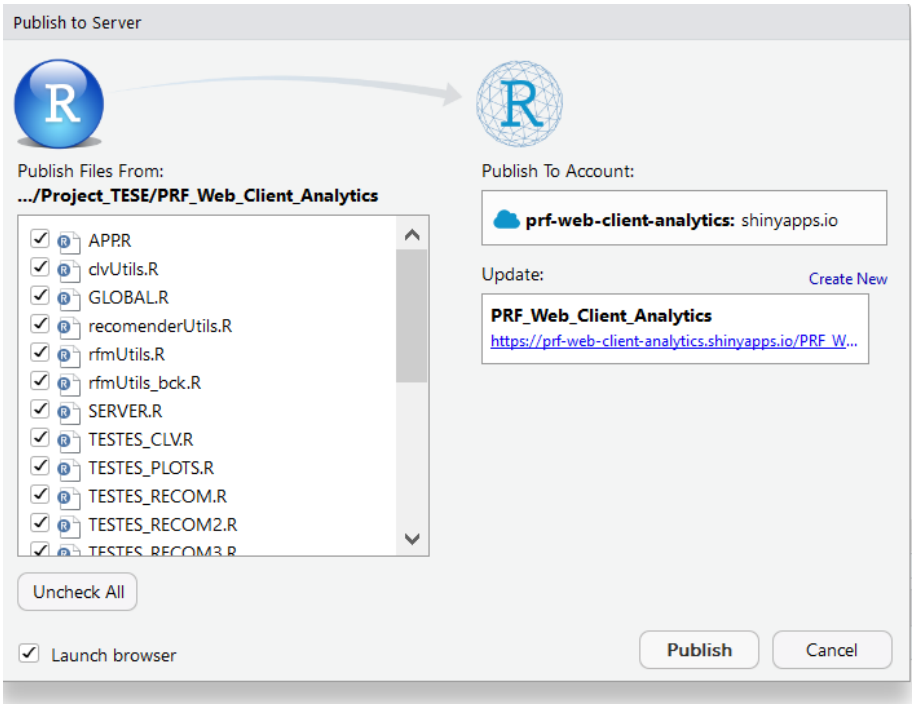


Figura 56 - Publicação para Web

SHINYAPPS.IO – identificada como uma plataforma self-service permitindo partilhar as aplicações desenvolvidas em Shiny para a Web. Utilizada muito para prototipagem de aplicações, e igualmente nas suas versões comerciais como uma verdadeira plataforma de apoio ao negócio. Tecnicamente é um SaaS (Software as a Service), ou seja, um serviço em execução na Cloud distribuída por diversos servidores, onde cada aplicação é independente utilizando dados carregados com a aplicação, ou através de ligações a terceiros, como web-services ou bases de dados.

O processo de publicação do projeto para a Cloud é de extrema simplicidade, indicando ao utilizador todas as etapas, e permitindo somente publicar as alterações da aplicação. No entanto revelou-se um autêntico desafio a tentativa de funcionamento da aplicação na plataforma ShinyApps.io, uma vez que projeto se utilizam 3 diferentes fontes de dados nomeadamente bases de dados em máquinas diferentes e protegidas por firewall. Com isso se verifica na plataforma Shinyapps.io a impossibilidade de conexão a essas mesmas fontes de dados, indicada na Figura 57.

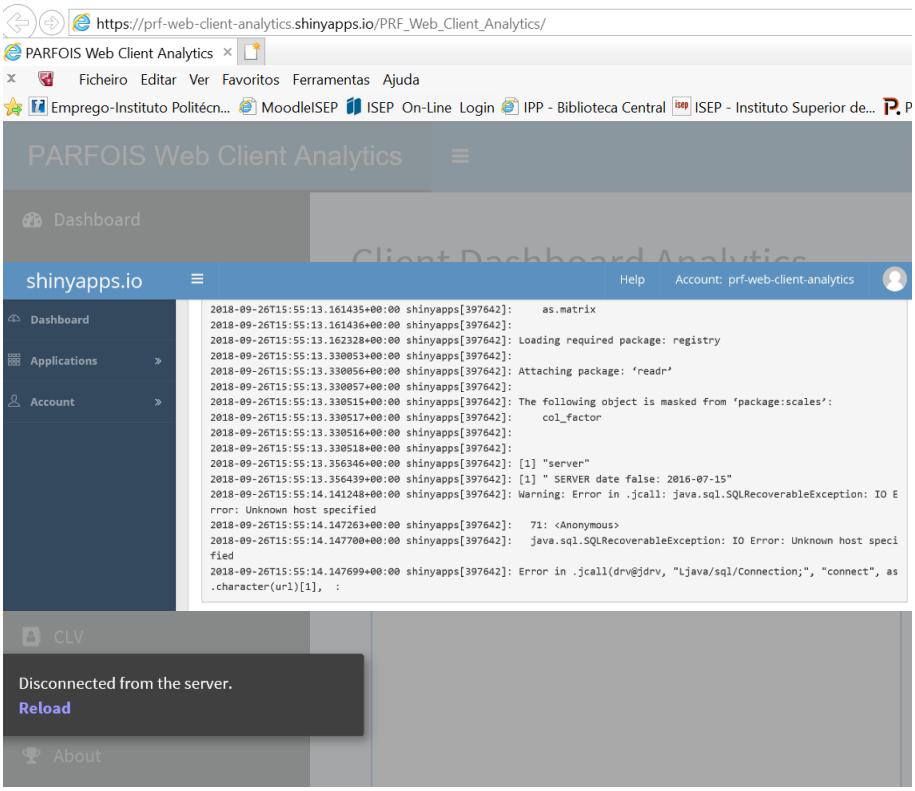


Figura 57 - ShinyApp Connection Erro

Validado junto com os fóruns de suporte da Shinyapps.io, foi indicado que para o funcionamento com bases de dados externas, é necessário junto dos administradores de rede da empresa requisitar a alteração de regras de firewall para permitir o acesso aos IP públicos dos servidores Shinyapps.io, pedido este prontamente recusado pela equipa IT uma vez que estaríamos a abrir um "buraco" na rede para acesso a bases de dados, com informação confidencial.

Tema este que retorna para um tema cada vez mais discutido nas organizações, a denominada Cloud. Não obstante das vantagens existentes para o uso da Cloud, nem todas as organizações estão totalmente abertas a esse mesmo uso, uma vez que o valor dos dados para a mesma pode ser bastante grande, e igualmente por questões de confidencialidade.

Atualmente a alternativa mais adotada pelas equipas de desenvolvimento em aplicações R Shiny, tem vindo a ser a ferramenta Shiny Server/Shiny Server Pro, onde num comparativo com o Shinyapps.io, o autor Ian Pylvainen ressalva que para a escolha entre estas duas opções deveremos saber responder as seguintes questões:

- A empresa permite aplicações fora da sua firewall?
- A empresa está consciente de que os dados vão ser colocados na Cloud para uso da aplicação?

- A empresa aceita uma plataforma de computação compartilhada para suas análises? (a exemplo, não existindo nenhum SLA definido)

A resposta negativa a qualquer uma destas questões, só por si justifica a usabilidade software dentro da empresa e não na Cloud.

5.3 Arquitetura do Modelo

No desenvolvimento deste projeto uma das componentes de maior importância foi a construção de uma aplicação, que utiliza no seu background vários modelos de técnicas de data mining em R que reunisse os diversos algoritmos capazes de realizar as diferentes etapas definidas, tal como é apresentado na Figura 58.

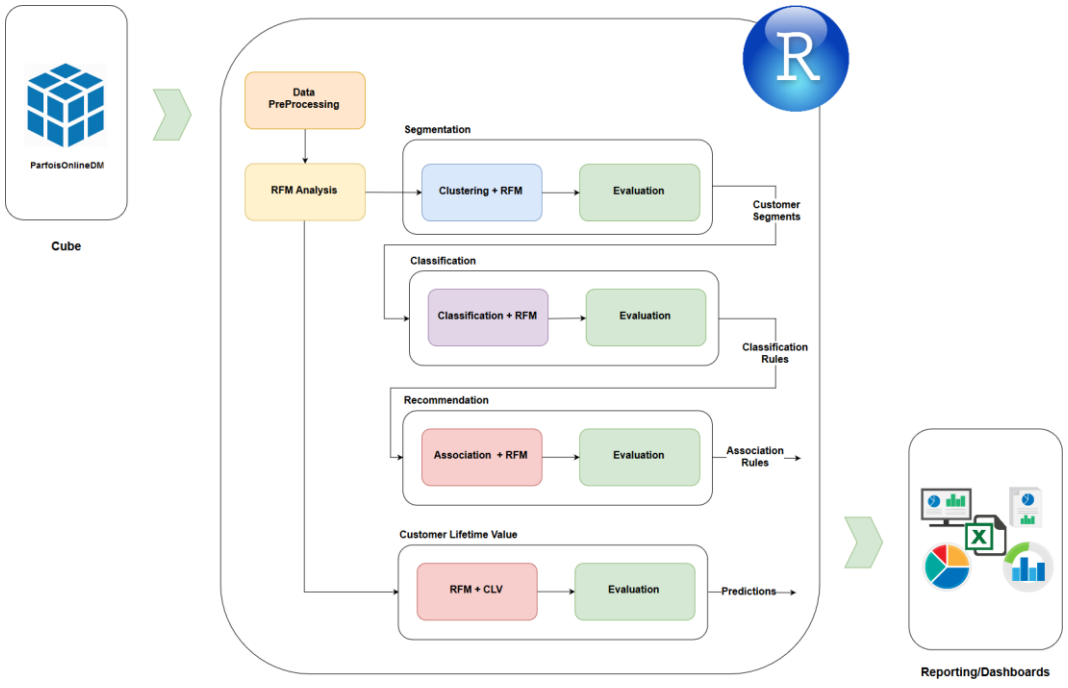


Figura 58 - Modelo Técnicas Data Mining em R

Esta aplicação está definida em diversas etapas, qua vamos explicar de seguida.

- Segmentação
    - O processo de Segmentação deverá através do algoritmo k-means, realizar a segmentação dos clientes, no caso em concreto optou-se por fazer a segmentação utilizando diversos dados, tais como valor total de encomendas,

número de clientes, número de encomendas, resultando em clusters agregados ao país.

- Classificação
    - O processo de Classificação deverá permitir classificar os clientes, mais concretamente utilizando o modelo RFM identificar os clientes nas diferentes métricas, *Recency* (assiduidade), *Frequency* (frequência) e *Monetary* (valor), este modelo será fundamental, uma vez que será a base para alguns dos diferentes processos.

- Recomendação
    - No processo de Recomendação é pretendido desenvolver a componente de recomendação de produtos para os clientes utilizando algoritmos de associação/recomendação tais como o *apriori* e a filtragem colaborativa.

- CLV
    - Por último, o processo de "CLV", ou seja, o "Valor do Tempo de Vida do Cliente", onde se pretende calcular o valor potencial que pode vir a ser gerado pelo cliente durante o seu período para com a empresa.

Estas diferentes etapas, vão ser exploradas nos próximos capítulos, efetuando a interpretação dos seus resultados e como os obter.

5.4 Segmentação

Na análise de dados, muitas vezes temos volumes de dados muito grandes, que são muitas vezes semelhantes entre si, portanto, podemos organizá-los em alguns clusters com observações semelhantes dentro de cada cluster. Por exemplo, no caso de dados de clientes, embora possamos ter milhões de clientes, esses clientes podem pertencer apenas a alguns segmentos: os clientes são semelhantes dentro de cada segmento, mas diferentes entre os segmentos. Muitas vezes, podemos querer analisar cada segmento separadamente, pois eles podem se comportar de maneira diferente (por exemplo, diferentes segmentos de mercado podem ter diferentes preferências de produto e padrões comportamentais) [46].

A segmentação de clientes pretende apoiar a empresa na divisão de grupos de clientes validando as semelhanças entre estes, o que poderá ser bastante útil para definição de campanhas direcionadas.

5.4.1 Clustering

Uma definição recorrente de clustering é “o processo de organizar dados em grupos cujos membros são similares de alguma forma”. Um cluster é um grupo de dados que partilham atributos semelhantes.

A análise de cluster pode ter muitas aplicabilidades, por exemplo, pode ser usado para identificar segmentos de clientes ou conjuntos competitivos de produtos, ou para segmentação geo-demográfica, etc. Geralmente é necessário dividir os dados em segmentos e executar qualquer análise subsequente dentro de cada segmento, a fim de desenvolver insights específicos do segmento [46].

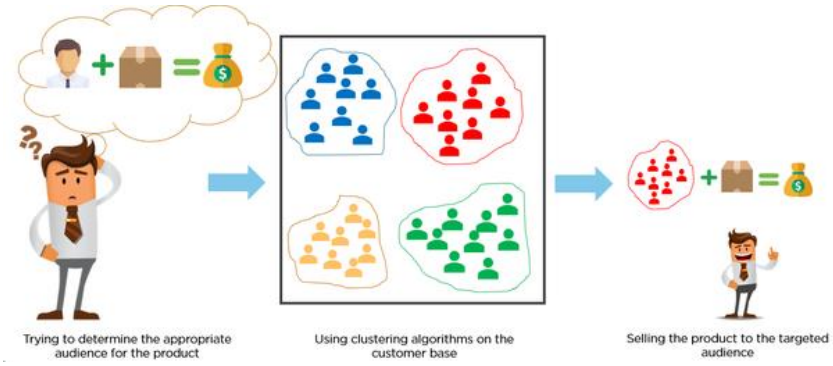


Figura 59 - Segmentação Clientes por Atributos

Através do clustering poderemos assim conhecer melhor os nossos clientes e nos direcionarmos para os corretos, conforme se exemplifica na Figura 59, de igual modo o clustering é útil como preparação de dados para outros métodos de data mining, por exemplo produção de modelos de classificação de cada um dos clusters descobertos, onde no cenário desenvolvido, os clusters são viáveis para uma melhor classificação do RFM e do CLV.

5.4.1.1 Interpretação dos Resultados do Clustering

Para a interpretação e avaliação do Clustering, efetuou-se a definição de um período de análise, neste cenário apresentou-se um período de doze meses.

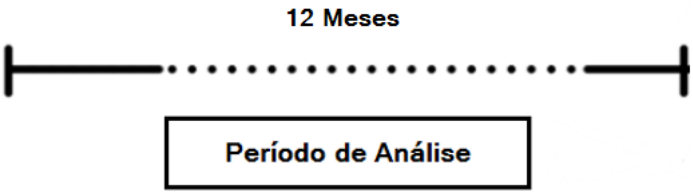


Figura 60 - Período Análise Clustering

No cenário que de seguida apresentamos para segmentar a base de dados de clientes foi assim definido o seguinte período, tendo em consideração o atributo demográfico, neste caso o país do cliente.

- Período de Análise – 12 Meses (01-01-2016 a 31-12-2016)

Iniciamos o processo recolhendo alguns dados para realizar esta análise para o período de 2016.

ENTITY	TOTAL	YEAR	COUNTRY
[redacted]	8.08	2016	
[redacted]L.COM	39.98	2016	
MA[redacted]@HOTMAIL.COM	22.49	2016	
HE[redacted].COM	36.99	2016	
CARI[redacted].COM	12.49	2016	
BEG[redacted].COM	27.97	2016	AL
SA[redacted]O.DE	31.93	2016	AT
ALE[redacted]OM	57.96	2016	AT
PE[redacted].NET	93.96	2016	AT
R[redacted]M	14.49	2016	AT
RI[redacted]M	36.98	2016	AT
LI[redacted]M	42.98	2016	AT
SA[redacted].AT	47.98	2016	AT
VI[redacted]G.AT	77.97	2016	AT
K[redacted].DE	54.98	2016	AT

Resultando num conjunto de dados de 101130 registos com quatro variáveis, no entanto estes apresentavam alguns problemas, nomeadamente a inexistência dos países, onde tivemos de realizar tratamento.

Para ser possível trabalhar com estes dados foram transformados os registos nulos para “NotDefined”.

Posterior a esta etapa, foi necessário efetuar diversas agregações de informação para obter os dados necessários, que passamos a indicar:

- Total em Valor por País
- Total de Clientes por País
- Total de Encomendas por País

Após todo este tratamento os resultados foram os seguintes, demonstrado no exemplo da Figura 61.

COUNTRY	TOTAL_AMOUNT	TOTAL_CUSTOMERS	TOTAL_ORDERS	YEAR
AL	27.97	1	1	2016
AT	12465.00	208	257	2016
BE	25427.78	444	526	2016
CZ	7045.75	143	168	2016
DE	99120.81	1910	2159	2016
DK	3041.12	53	69	2016
EE	1161.17	24	31	2016
ES	1080981.02	32965	39277	2016
FI	1911.77	38	45	2016
FR	380227.55	9002	10219	2016
GR	1955.41	41	43	2016
HR	8020.22	163	208	2016
HU	31056.18	681	860	2016
IE	84210.89	1956	2251	2016

Figura 61 - Totais por Países

Para cada país foi atribuído um número para utilização posterior na definição de clusters, onde com os dados obtidos e após calculados resulta no apresentado na Figura 62.

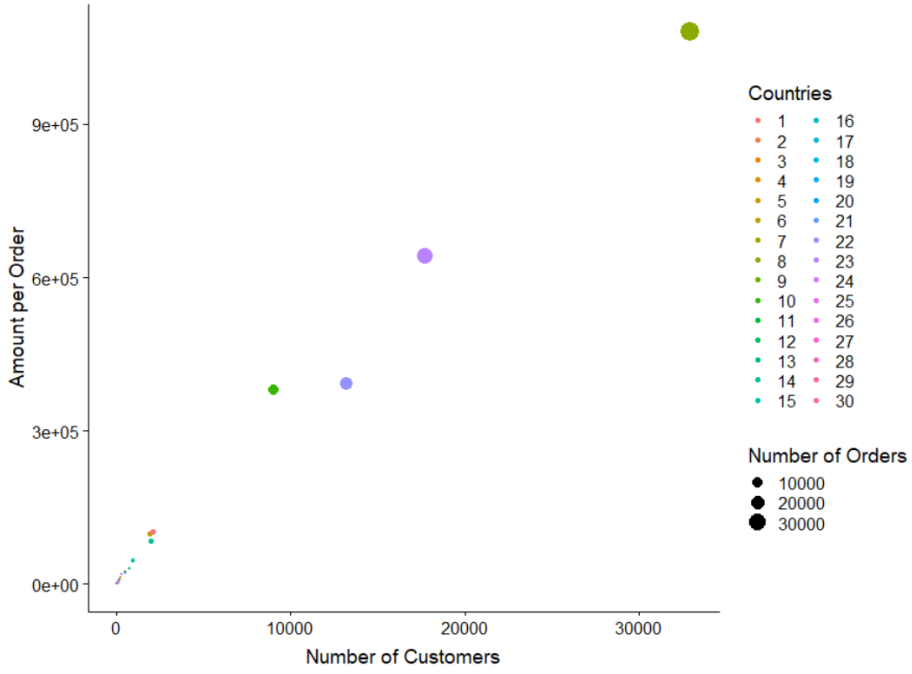
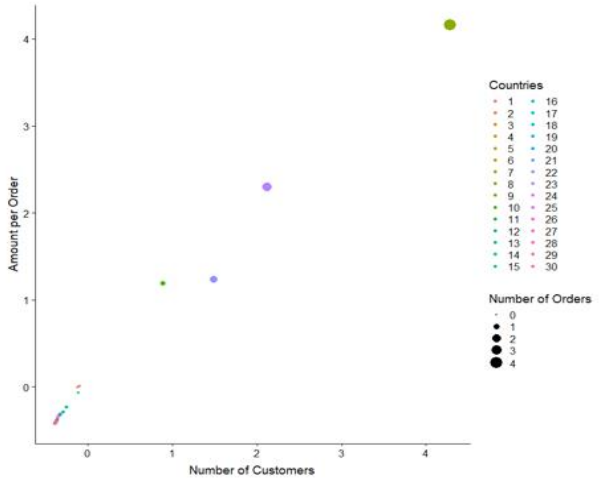


Figura 62 - Dados Países

Para se conseguir trabalhar corretamente com os dados foi necessário normalizar os mesmos nos atributos utilizados.

```
> summary(df_country_k)
  TOTAL_AMOUNT      TOTAL_CUSTOMERS     TOTAL_ORDERS          CODE
Min.   :      20   Min.   :    1.00   Min.   :    1.0   Min.   : 1.00
1st Qu.:    1923   1st Qu.:   38.75   1st Qu.:   43.5   1st Qu.: 8.25
Median :    6190   Median :  127.50   Median :  150.0   Median :15.50
Mean   :   99406   Mean   : 2758.20   Mean   : 3371.0   Mean   :15.50
3rd Qu.:   41882   3rd Qu.:  880.50   3rd Qu.: 1037.8   3rd Qu.:22.75
Max.   :1080981   Max.   :32965.00   Max.   :39277.0   Max.   :30.00
```




```
> summary(rescale_df_country_customers)
      CODE          cust_scal.V1        amou_scal.V1        ord_scal.V1
Min.   : 1.00   Min.   :-0.391143   Min.   :-0.421215   Min.   :-0.388925
1st Qu.: 8.25   1st Qu.:-0.385787   1st Qu.:-0.413151   1st Qu.:-0.384020
Median :15.50   Median :-0.373197   Median :-0.395066   Median :-0.371729
Mean   :15.50   Mean   : 0.000000   Mean   : 0.000000   Mean   : 0.000000
3rd Qu.:22.75   3rd Qu.:-0.266375   3rd Qu.:-0.243799   3rd Qu.:-0.269275
Max.   :30.00   Max.   : 4.285206   Max.   : 4.160079   Max.   : 4.143835
```

Figura 63 - Resultado da Normalização

Neste processo foi utilizado o algoritmo K-Means, que como já explicado anteriormente, o algoritmo começa por atribuir aleatoriamente uma observação a cada uma das *k* categorias (centros de cada categoria) e em seguida atribui cada observação à categoria cujo centro seja mais próximo, calcula a média de cada categoria, sendo estas os novos centros de cada categoria. Em seguida, ele reatribui cada observação à categoria com centro mais próximo antes de recalcular novos centros. Esta etapa é repetida várias vezes até que não sejam necessárias mais reatribuições [47].

Executando assim o algoritmo k-means para este conjunto de dados, onde o k assume o valor de 2, 3 e 4, foi possível obter o seguinte resultado na Figura 64.

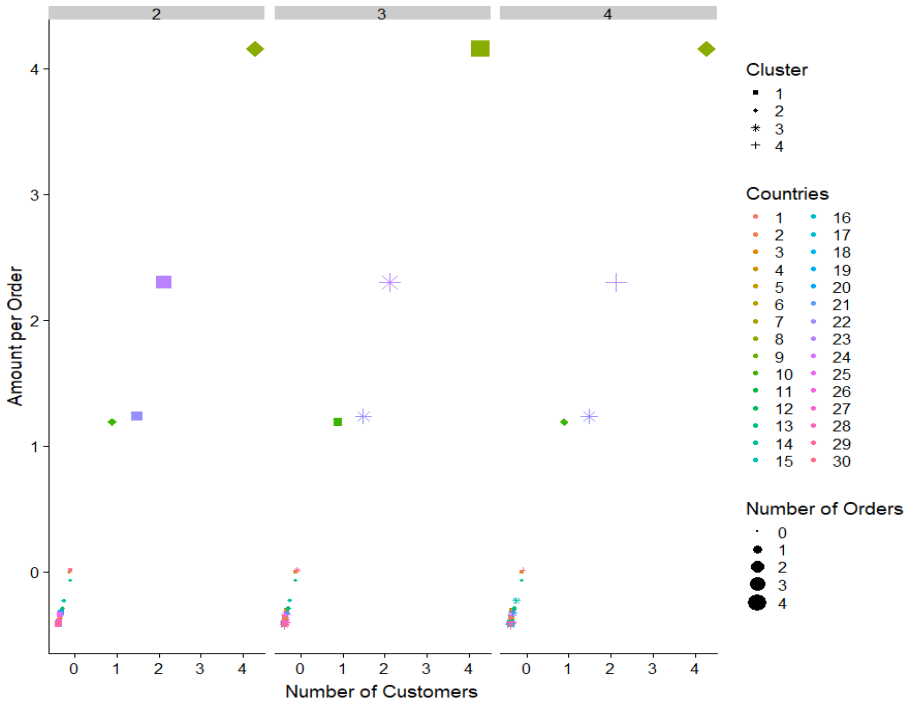


Figura 64 - Resultado K-Means

É possível perceber que quando o k é igual a 3 ou 4, os clientes são divididos em segmentos mais pequenos.

Como determinar o número certo de k para escolher? Não há grandes abordagens algorítmicas para responder a essa pergunta, mas normalmente o realizado é executar o algoritmo k-means em diferentes valores de k, e medir a quantidade de erro que é reduzida pela adição de mais clusters, sendo que, ao adicionar mais clusters, reduza o erro.

Por outro lado, é necessário ter em consideração que à medida que se adiciona mais clusters, existe o risco de sobrecarregar os dados, e em caso extremo, acabar tendo cada ponto de dados como o seu próprio cluster [48].

Para avaliarmos o número certo de clusters, utilizou-se três diferentes abordagens, utilizando o método do cotovelo (Elbow), o método Silhouette, e o pacote NbClust que testa diferentes índices.

O método cotovelo (Elbow) que se trata de um método de interpretação e validação de cluster, projetado para ajudar a encontrar o número apropriado de clusters num determinado conjunto de dados. O método do cotovelo é usado com erro quadrado (sse) ou pela soma de erros dentro do cluster(wcss), baseado na observação de que o aumento do número de clusters pode ajudar a reduzir a soma das variâncias “dentro do cluster”.

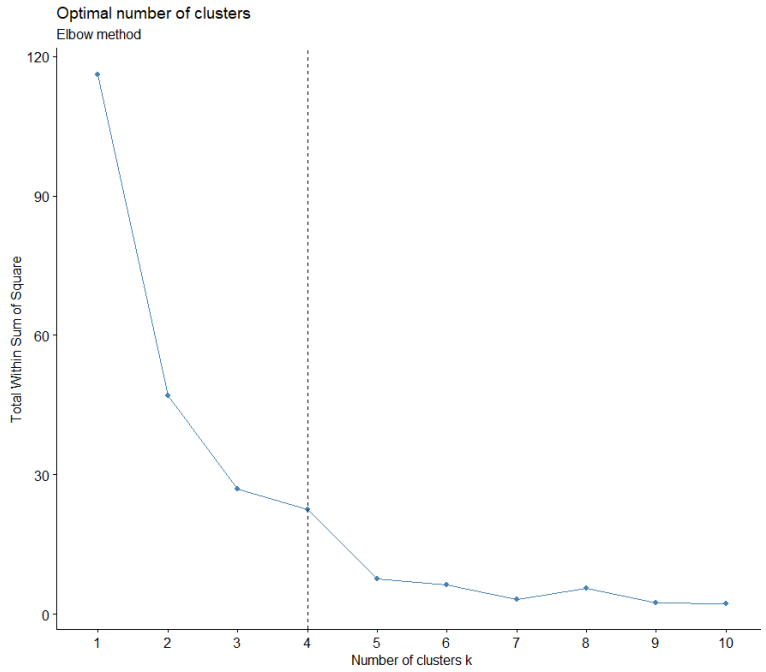


Figura 65 - Método Elbow

Com o método do cotovelo (Elbow) obteve-se a indicação como melhor cluster, para o número de 4 clusters, demonstrado na Figura 65.

Para o segundo método utilizado, o Silhouette que também se trata de um método de interpretação e validação de cluster, onde o valor da silhueta é uma medida de como um objeto é semelhante ao seu próprio cluster (coesão) comparado a outros clusters (separação), ou seja, determina quão bem cada objeto está dentro de seu cluster. Um valor médio elevado da silhueta indica um bom agrupamento.

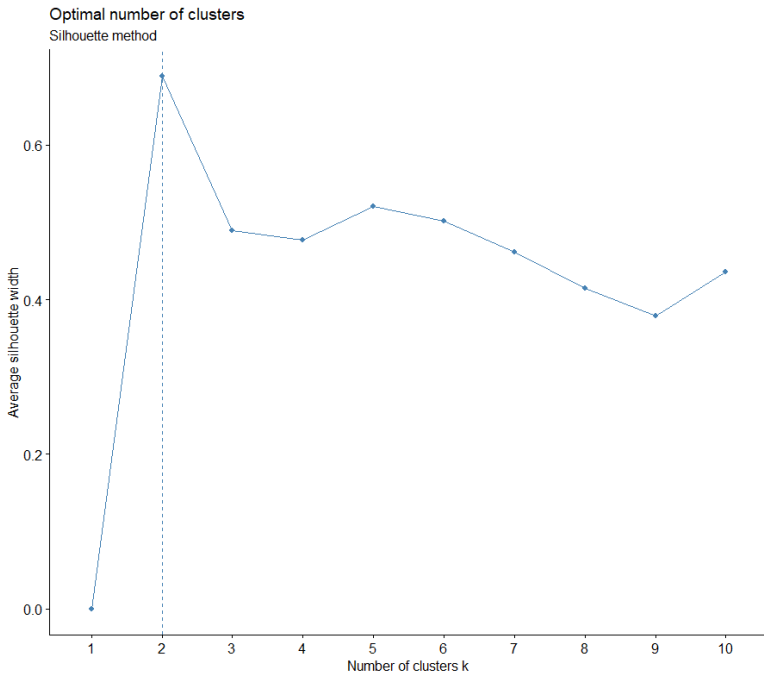


Figura 66 - Método Silhouette

Com o método do Silhouette obteve-se a indicação como melhor cluster, para o número de 2 clusters, demonstrado na Figura 66.

Para último método utilizou-se o pacote NbClust (), que pode ser usado para calcular simultaneamente muitos outros índices e métodos para determinar o número de clusters, onde através deste comparou-se 26 índices diferentes, sugerindo 0, 2, 3, 5, 8, 9 e 10 clusters.

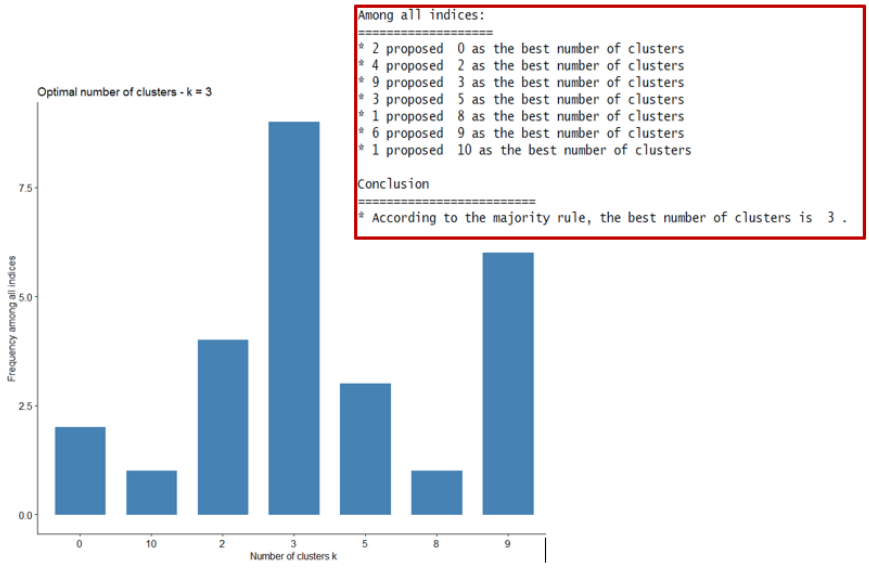


Figura 67 - Nbclust Clusters

No entanto número de clusters indicado pela maioria dos índices, foi o número de 3 clusters, indicado por 9 índices, visível na Figura 67.

Com estes resultados obtidos é possível arriscar que a melhor escolha poderá ser considerar entre 3 clusters, ou seja segmentos de clientes, como se é possível percecionar na Figura 68.

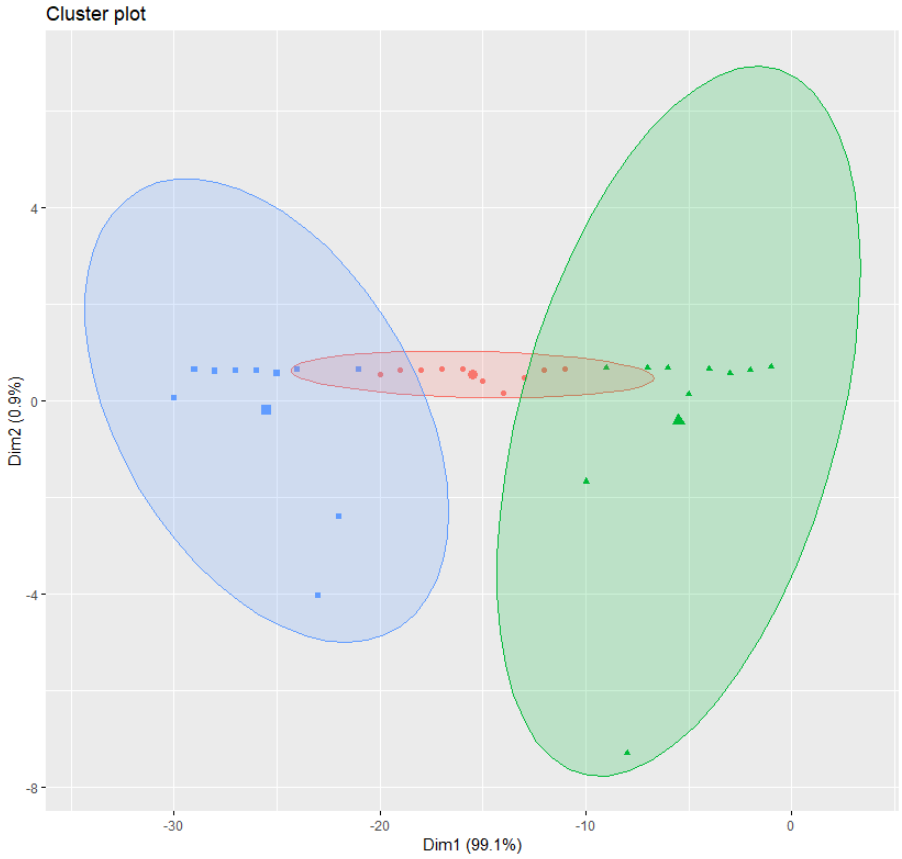


Figura 68 - Cluster=3

No entanto o número ideal de clusters é muitas vezes definido como subjetivo e depende do método usado para medir semelhanças e os parâmetros usados para particionamento.

5.5 Classificação

A classificação de clientes assume atualmente um papel fundamental para uma empresa, realizada em distintos departamentos e com diferentes métricas, tem vindo a ter cada vez mais ênfase na área de Marketing.

"identificação de indivíduos ou organizações com características semelhantes que têm implicação significativa para a determinação de uma estratégia de marketing"[49]

5.5.1 RFM

RFM significa Recency, Frequency e Monetary. A análise RFM é uma técnica de marketing usada para analisar o comportamento do cliente, por exemplo, quando um cliente comprou

recentemente (Recency), com que frequência o cliente compra (Frequency) e qual o valor que o cliente adquire (Monetary).

É uma técnica útil para melhorar a segmentação de clientes, dividindo os clientes em vários grupos para futuros processos de marketing e para identificar os clientes com maior probabilidade de responder a campanhas e promoções [44].

5.5.1.1 Interpretação dos Resultados do RFM

Para a interpretação e avaliação da análise RFM, efetuou-se a definição de um período de análise, neste cenário apresentou-se um período de três meses.

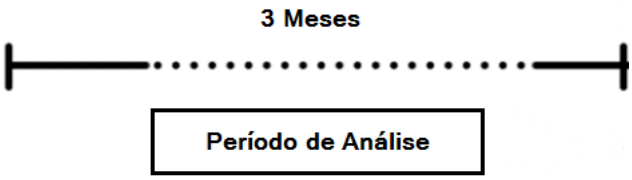


Figura 69 - Período Análise RFM

No cenário que de seguida se apresenta para avaliar os resultados obtidos do RFM foi assim definido o seguinte período:

- Período de Análise – 3 Meses (22-03-2016 a 20-06-2016)

O processo de definição do RFM é realizado em quatro fases distintas, iniciando-se pela Recency, posterior a Frequency e o último a ser calculado a Monetary, sendo a última fase um processo simples de concatenação destes três anteriores resultados.

Verifique-se seguidamente uma amostra de cinco clientes obtidos de todo o conjunto de dados do período de análise anteriormente definido, onde este conjunto apresenta os dados essenciais para a definição do RFM, nomeadamente:

- CustomerId - código único de cliente
- Date - data da encomenda
- Amount - valor da encomenda

Customer Id	Date	Amount
273	02/04/2016	22,48

310	25/05/2016	31,98
310	24/05/2016	37,98
966	06/05/2016	9,07
3320	02/06/2016	29,98
17565	06/06/2016	59,91
17565	04/05/2016	62,9
17565	21/04/2016	57,93
17565	20/04/2016	61,96
17565	07/04/2016	54,94
17565	03/04/2016	58,89

Tabela 3 - Dados Cliente Base

É necessário com estes dados tratar os mesmos de forma a conseguir obter três novas variáveis, nomeadamente:

- CustomerId- código único de cliente
- Date Most Recent - data encomenda mais recente
- Frequency - número de encomendas realizadas
- Monetary - valor total gasto em encomendas

Para obter estes dados efetua se os seguintes cálculos:

- Primeira etapa – obter o total por cliente, somatório da coluna Amount por CustomerId (Monetary)
- Segunda etapa - obter o número de encomendas por clientes, para cada CustomerId somar o número de registos da coluna Date. (Frequency)
- Terceira etapa - obter a data mais recente de cada encomenda de cliente, através das colunas CustomerId e Date. (Date Most Recent)

Customer Id	Date Most Recent	Frequency	Monetary
273	02/04/2016	1	22,48
310	25/05/2016	2	69,96
966	06/05/2016	1	9,07
3320	02/06/2016	1	29,98
17565	06/06/2016	6	356,53

Tabela 4 - Dados Cliente Tratados

Fica assim em falta obter a variável Recency, necessária para dar início ao processo de classificação de clientes por RFM. É necessário ter em consideração que para efeitos de definição da métrica Recency, é necessário indicar qual a data em que se realiza esta análise, sendo que no cenário apresentado foi definida a data de 20-07-2018, que será útil para o cálculo do Recency.

- Recency = Data Analise – Data Encomenda mais recente (Date Most Recent)

Posteriormente ao tratamento destes dados, identificados dados em "bruto", fica-se com os dados necessários para avançarmos com a análise RFM, nomeadamente os seguintes dados:

- CustomerId - código único de cliente
- Recency - período desde a última data de encomenda (comparável à data de analise definida)
- Frequency - número de encomendas realizadas
- Monetary - valor total gasto em encomendas

Customer Id	Recency	Frequency	Monetary
273	109	1	22,48
310	56	2	69,96
966	75	1	9,07
3320	48	1	29,98
17565	44	6	356,53

Tabela 5 - Dados Base para RFM

De realçar que para a classificação RFM realizada, foi definido uma segmentação de 5 grupos, distribuindo-se entre 1 e 5, sendo o 1 o valor mais baixo do grupo, e o 5 o valor mais alto do grupo, onde para cada uma das métricas, poderemos efetuar a seguinte leitura:

- Recency
  - Classificação 1 – Cliente mais antigo.
  - Classificação 5 – Cliente mais recente.
- Frequency
  - Classificação 1 – Cliente menos frequente.
  - Classificação 5 – Cliente mais frequente.

- Monetary
  - Classificação 1 – Cliente menor valor.
  - Classificação 5 – Cliente maior valor.

Com os novos dados obtidos, inicia-se a segmentação dos clientes começando por classificar os clientes pela sua Recency, ou seja, pelo tempo desde a sua última compra, dividindo em 5 grupos iguais (quintil), considerando para os 20% clientes mais recentes a classificação de 5,

para os próximos 20% uma classificação de 4, e assim sucessivamente até à classificação de 1 para os clientes mais antigos [50].

Customer Id	Recency	Frequency	Monetary	R
273	109	1	22,48	1
310	56	2	69,96	4
966	75	1	9,07	3
3320	48	1	29,98	4
17565	44	6	356,53	5

Tabela 6 - Classificação Cliente Recency

Para a classificação do Frequency, ou seja os cliente que mais encomendas realizam, igualmente se realizou a abordagem da divisão por 5 grupos iguais (quintil), no entanto verifica-se que esta divisão por quintil na métrica de Frequency não ficava equilibrada, resultado somente na classificação de 1 e de 5, ou seja, os 20% melhores, e os 20% piores. Como tal realiza-se outra abordagem, onde se encontra o valor máximo da Frequency e se divide por 5, o resultado deste, é utilizado como a métrica de "dimensão" para cada grupo, a exemplo, se o valor máximo para a Frequency, então teremos 12/5=2,4 , assume-se assim que cada grupo será dividido por intervalos de 2, nomeadamente:

- Grupo 1 = 0 a 2 (inclusive)
- Grupo 2 = 2 a 4 (inclusive)
- Grupo 3 = 4 a 6 (inclusive)
- Grupo 4 = 6 a 8 (inclusive)
- Grupo 5 = maior que 8

Somente aplicando esta nova abordagem conseguimos ter classificações entre 1 e 5.

Customer Id	Recency	Frequency	Monetary	R	F
273	109	1	22,48	1	1
310	56	2	69,96	4	1
966	75	1	9,07	3	1
3320	48	1	29,98	4	1
17565	44	6	356,53	5	3

Tabela 7 - Classificação Cliente Frequency

No último processo de calculo de classificação é para o Monetary, isto é, classificar os clientes que maior ou menor valor despende nas suas encomendas. Também foi realizado a abordagem da divisão por 5 grupos iguais (quintil), considerando para os 20% clientes com maior valor adquirido a classificação de 5, para os próximos 20% uma classificação de 4, e assim sucessivamente até à classificação de 1 para os clientes com menor valor adquirido.

Customer Id	Recency	Frequency	Monetary	R	F	M
273	109	1	22,48	1	1	2
310	56	2	69,96	4	1	5
966	75	1	9,07	3	1	1
3320	48	1	29,98	4	1	3
17565	44	6	356,53	5	3	5

Tabela 8 - Classificação Cliente Monetary

A última fase do processo é como já anteriormente indicado, um simples processo de concatenação das três métricas anteriormente obtidas, nomeadamente o Recency, Frequency e Monetary, sendo assim possível de obter 125 diferentes classificações (5x5x5) para os diversos clientes. Onde têm o valor mais elevado o RFM de 555, e para o valor mais baixo o RFM de 111.

Customer Id	Recency	Frequency	Monetary	R	F	M	RFM
273	109	1	22,48	1	1	2	112
310	56	2	69,96	4	1	5	415
966	75	1	9,07	3	1	1	311
3320	48	1	29,98	4	1	3	413
17565	44	6	356,53	5	3	5	535

Tabela 9 - Classificação Cliente RFM

No que diz respeito à interpretação de cada cliente com base na classificação RFM obtida, verifica-se respostas, como por exemplo:

Cliente 17565 obteve uma classificação de 535, ou seja, é um cliente que embora com um valor monetário elevado, e encomendas recentes, não tem um número grande de encomendas. Aqui poderemos tentar persuadir o cliente a realizar mais encomendas.

Cliente 273 obteve a classificação de 112, aqui é importante perceber o que funcionou errado com este cliente, uma vez que este não tem encomendas recentes, nem várias encomendas. Poderá ter tido algum problema com a encomenda, ou outra justificação que não o faça voltar a comprar o produto, eventualmente também poderá ter encontrado uma alternativa/concorrência o que fez com que não regressasse. Neste cenário deve-se apelar à criatividade para realizar campanhas para reconquistar os clientes [51].

5.6 Recomendação de Produtos

A recomendação de produtos é atualmente uma ferramenta muito poderosa e que pode com alguma facilidade ser implementado sobre uma base dados com histórico.

Um motor de recomendação consegue assim através de um dado conjunto de clientes, e um conjunto de produtos, e as diversas relações entre esses dois conjuntos, prever uma ou mais novas relações entre clientes e itens.

Essas recomendações podem ajudar a melhorar a percentagem de vendas, ajudando o cliente a encontrar produtos que ele deseja, fazendo-o comprar mais rapidamente, promover a venda cruzada sugerindo produtos adicionais, e melhorar a fidelidade do cliente por meio da criação de um relacionamento de valor acrescentado para ambas as partes [21].

5.6.1 Interpretação dos Resultados da Recomendação

Para a interpretação e avaliação das recomendações propostas, efetuou-se a definição de um período de análise, neste cenário apresentou-se um período de um mês.

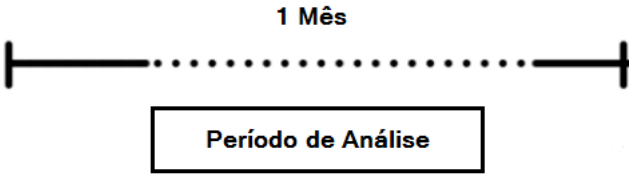


Figura 70 - Período Análise Recomendação

No cenário que de seguida se apresenta para avaliar os resultados obtidos foi assim definido o seguinte período:

- Período de Análise – 1 Mês (01-01-2016 a 30-01-2016)

Primeiro foi necessário obter um conjunto de dados de encomendas para esse período.

```
day_recom_min <- as.Date('2016-01-01' ,format='%Y-%m-%d')

day_recom_max <- as.Date('2016-01-30' ,format='%Y-%m-%d')


purchases_cust_detail = sqlQuery(
  Conn_DW,
  paste(
    "SELECT  UPPER(M.cust_email_sfk) ENTITY,
    P.PROD_COD ARTICLE,
    M.ORDER_ID,
    M.ONL_MOV_QTY
    FROM VCUB_FCT_ONL_MOVEMENTS_D M WITH (nolock)
    INNER JOIN nq_vcub_DIM_PRODUCT P WITH (nolock) ON P.prod_spk = M.prod_sfk
    where M.cod_type='PUR' and M.COD_PLATFORM='PARFOIS' and M.ONL_MOV_QTY>0 and M.inv_cod is not null
    AND convert(date,CONVERT(varchar(10),M.dat_mov_sfk,101)) between '",
    day_recom_min,
    "' and '",
    day_recom_max,
    "'",
    sep = ""
  )
)
```

Resultando num conjunto de dados de 26373 registos com quatro variáveis.

ENTITY	ARTICLE	ORDER_ID	ONL_MOV_QTY
31	138219PR_M	PUR/247754	1
189	139549MU_	PUR/238838	1
189	122426MU2	PUR/238838	1
189	136781MU_	PUR/238838	1
189	86634VE_	PUR/238838	1
368	123437PA_L	PUR/250601	1
368	136370DO_L	PUR/250601	1
368	136802PA_M	PUR/250601	1
368	138239PA_L	PUR/250601	1
368	136696PA_L	PUR/250601	1
368	138885PA_L	PUR/250601	1
488	127400PA_	PUR/242207	1

```
Data
purchases_cust_ 26373 obs. of 4 variables
  ENTITY : int 13998 13998 13998 13998 13998 13998 14036 1
  ARTICLE : Factor w/ 3439 levels "119004PR_","119004TR_",
  ORDER_ID : Factor w/ 11144 levels "PUR/236339","PUR/2363
  ONL_MOV_QTY: num 1 1 1 1 1 1 1 1 1 1 ...
```

```
> summary(purchases_cust_detail)
     ENTITY              ARTICLE            ORDER_ID       ONL_MOV_QTY
 Min.   :    31   134689PR_L:  284   PUR/243458:   44   Min.   : 1.000
 1st Qu.: 98680   134689CA_L:  200   PUR/247171:   44   1st Qu.: 1.000
 Median :198240   134689MZ_L:  142   PUR/250018:   29   Median : 1.000
 Mean   :199646   137707PR_L:  131   PUR/256770:   29   Mean   : 1.022
 3rd Qu.:300365   134673BE_M:  130   PUR/237480:   21   3rd Qu.: 1.000
 Max.   :401374   134673MZ_M:  128   PUR/245004:   20   Max.   :20.000
                  (Other)   :25358   (Other)   :26186
```

Para ser possível trabalhar com estes dados foram transformados numa matriz através da função **acast**, e posteriormente para uma matriz classificada com “realRatingMatrix”.

```
product_item <-
  acast(purchases_cust_detail[, c("ENTITY", "ARTICLE", "ONL_MOV_QTY")],
        ENTITY ~ ARTICLE,
        fun.aggregate=sum
        ,value.var = "ONL_MOV_QTY")


# turn that matrix into a rating matrix
product_item_rmatrix <- as(product_item, "realRatingMatrix")
```

Seguido da normalização dos registos da matriz classificada, ou seja, garantir a centralização removendo o desvio, subtraindo o registo pela média de todos os registos, com a função **normalize**.

```
61  #normalize the rating matrix
62  product_item_rmatrix_norm <- normalize(product_item_rmatrix)
63  head(as(product_item_rmatrix_norm, "data.frame"))
64
65:1  (Top Level)
Console  Terminal  Markers
C:/Users/FredericoAlmeida/Dropbox/Mestrado_Tese/Desenvolvimento/R Project/Project_TESE/PRF_Web_Clie
> head(as(product_item_rmatrix_norm, "data.frame"))
      user      item        rating
1       31 119004PR_ -0.0003565062
6486    31 119004TR_ -0.0003565062
12971   31 119137DO_ -0.0003565062
19456   31 119139PA1 -0.0003565062
25941   31 119321RO_ -0.0003565062
32426   31 119329MU_ -0.0003565062
>
```

Para possibilitar o tratamento mais rápido e eficiente da matriz, deve-se converter a mesma em registos de 0 e 1 através da função **binarize**.

```
65  #binarize the rating matrix
66  product_item_b_rmatrix  <-
67    binarize(product_item_rmatrix_norm, minRating = 1)
68
69  head(as(product_item_b_rmatrix, "data.frame"))
70
72:1  (Top Level)
Console  Terminal  Markers
C:/Users/FredericoAlmeida/Dropbox/Mestrado_Tese/Desenvolvimento/R Project/Project_TESE/PI
> head(as(product_item_b_rmatrix, "data.frame"))
     user       item rating
5  101263 134115MZ_M      1
10 104763 134689PR_L      1
24  11059 138383PR_M      1
30 112686 140173PA_M      1
4  122782 133959PR_L      1
17  12864 137836WI_S      1
>
```

Para realizar um modelo de recomendação foi usado a função **recommender** do package recommenderlab [21], onde especificamos um algoritmo dos diversos disponíveis (POPULAR, RANDOM, IBCF, UBCF).

Seguidamente para gerar as recomendações utiliza-se a função **predict** que utiliza o modelo de recomendação anteriormente definido e os dados de novos clientes.

```
#tipos algoritmos - POPULAR - RANDOM - UBCF - IBCF
rec_ALGOR <- Recommender(product_item_b_rmatrix[1:nrow(product_item_b_rmatrix)],method="UBCF")


#ALL - 1 sugestion per ENTITY
predi_all <-predict(rec_ALGOR, product_item_b_rmatrix,n=1)
```

Foi possível então gerar distintas recomendações para diferentes algoritmos.

- UBCF – 10355 recomendações
- IBCF – 115 recomendações
- RANDOM – 10355 recomendações
- POPULAR – 400 recomendações

	ARTICLE	ENTITY
1	134685MZ_L	68149
4	134685MZ_L	151372
5	134685MZ_L	194860
6	134685MZ_L	259587
2	134689CA_L	104763
3	134697SK_M	135352
7	134350LA_39	292357
8	134689MZ_L	298325
9	132414MZ1M	342660

Figura 71 - UBCF

ARTICLE	ENTITY
134689PR_L	3086
134689PR_L	7020
134689PR_L	7308
134689PR_L	9471
134689PR_L	11059
134689PR_L	12156
134689PR_L	12864
134689PR_L	16295
134689PR_L	18288

Figura 72 - POPULAR

ARTICLE	ENTITY
138486CU_	3086
136936CU_	9471
137949PR_M	18288
134730BR_M	19540
134730BR_M	25486
134689CA_L	29605
128591DO_	41203
138655ORNM	42680
134881AZ_M	52068

Figura 73 - IBCF

ARTICLE	ENTITY
139005BE_37	31
140141TP_S	189
127655DO_M	368
133947BE_M	488
140079PR_	584
138987PA_	1505
137005BX_	1603
78085VR_	1829
137085AZ_	1830

Figura 74 - RANDOM

No entanto as recomendações geradas devem ser avaliadas para se definir qual será a que obtém menor erro e que poderá ser a melhor a utilizar[52].

```
##VALIDATION

# create evaluation scheme splitting taking 90% of the date for training and leaving 10% for validation or test
#k   - number of folds/times to run the evaluation (defaults to 10 for cross-validation and bootstrap and 1 for split)
#given   - single number of items given for evaluation or a vector of length of data giving the number of items given

eval_scheme <- evaluationScheme(product_item_rmatrix, method="split", train=0.9, k=1, given=3, goodRating=1)



# creation of recommender model based on UBCF
Rec.ubcf <- Recommender(getData(eval_scheme, "train"), "UBCF")
# creation of recommender model based on IBCF for comparison
Rec.ibcf <- Recommender(getData(eval_scheme, "train"), "IBCF")


# making predictions on the test data set
p.ubcf <- predict(Rec.ubcf, getData(eval_scheme, "known"), type="ratings")
# making predictions on the test data set
p.ibcf <- predict(Rec.ibcf, getData(eval_scheme, "known"), type="ratings")


# obtaining the error metrics for both approaches and comparing them
error_metric.ubcf<-calcPredictionAccuracy(p.ubcf, getData(eval_scheme, "unknown"))
error_metric.ibcf<-calcPredictionAccuracy(p.ibcf, getData(eval_scheme, "unknown"))

error_metric <- rbind(error_metric.ubcf,error_metric.ibcf)
rownames(error_metric) <- c("UBCF","IBCF")
error_metric



## predict missing ratings
## (results in RMSE, MSE and MAE)

##UBCF
ev_ubcf <- evaluate(eval_scheme, "UBCF", type="ratings")
avg(ev_ubcf)


##IBCF

ev_ibcf <- evaluate(eval_scheme, "IBCF", type="ratings")
avg(ev_ibcf)
```

Para essa mesma avaliação utiliza-se as fórmulas mais recorrentes na avaliação de rácios preditivos, nomeadamente através do MAE (Mean Average Error) e do RMSE (Root Mean Square Error) que vamos explicar.

MAE (Mean Average Error) - Calcula a magnitude média dos erros num determinado conjunto de previsões, sem considerar sua direção. É a média sobre a amostra de teste das diferenças absolutas entre a previsão e a observação real, em que todas as diferenças individuais têm peso igual [53].

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

Figura 75 - Formula MAE

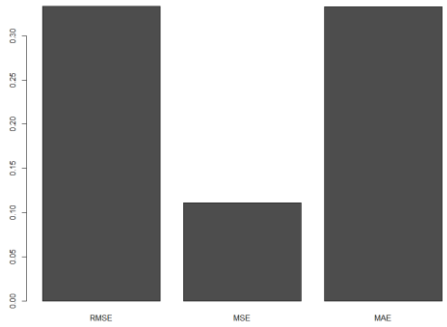
RMSE (Root Mean Squared Error) - É uma regra de pontuação quadrática que também mede a magnitude média do erro, ou seja, a medida do desempenho do seu modelo. É a raiz quadrada da média das diferenças quadradas entre os valores previstos e os valores reais [53].

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

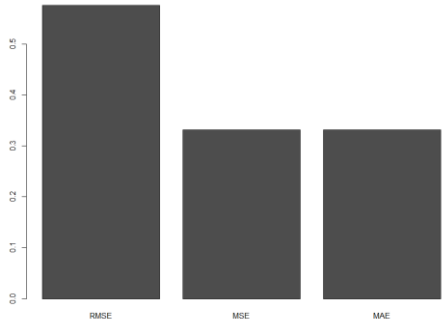
Figura 76 - Formula RMSE

Essa diferença entre sua previsão e a observação real é o termo de erro. O termo de erro é importante porque geralmente o pretendido é minimizar o erro.

Efetuados os cálculos verifica-se para cada um dos algoritmos apresenta os seguintes valores.




```
> ev_ubcf <- evaluate(eval_scheme, "UBCF", type="ratings")
UBCF run fold/sample [model time/prediction time]
         1  [6.64sec/365.26sec]
> avg(ev_ubcf)
        RMSE       MSE       MAE
res 0.333378 0.1111409 0.3331615
```




```
> ev_ibcf <- evaluate(eval_scheme, "IBCF", type="ratings")
IBCF run fold/sample [model time/prediction time]
         1  [2257.54sec/0.51sec]
> avg(ev_ibcf)
        RMSE       MSE       MAE
res 0.5758756 0.3316327 0.3316327
```

Com os valores indicados pode-se concluir que o algoritmo UBCF apresenta um menor termo de erro de previsão comparativamente com o IBCF.

O processo de avaliação apresentado foi obtido tendo como base as diretrizes de avaliação de modelos de recomendação indicados pelo package RecommenderLab [21].

5.7 CLV

Com o Customer Lifetime Value, ou CLV, ou seja, o "Valor do Tempo de Vida do Cliente", é uma métrica frequentemente negligenciada que pode prever com precisão o valor dos clientes, onde se pretende calcular o valor potencial que pode vir a ser gerado pelo cliente durante o seu período para com a empresa.

O CLV fornece informações cruciais sobre o quanto deveremos gastar na aquisição dos clientes, mas também o valor de retorno que estes trarão para a empresa no longo prazo.

Em vez de apenas tentarmos arranjar clientes, pode-se ter a perceção em quais clientes se devem concentrar e, mais importante ainda, as razões porque nos devem concentrar neles.[19]

5.7.1 Interpretação dos Resultados do CLV

Para a interpretação e avaliação dos modelos definidos, efetuou-se a definição dos respetivos períodos. Para o período de treino foi definido 12 meses, por sua vez para o período de teste, efetua-se a definição de um Curto Período (2 meses) e um Longo Período (4 meses) para assim melhor avaliar os modelos.

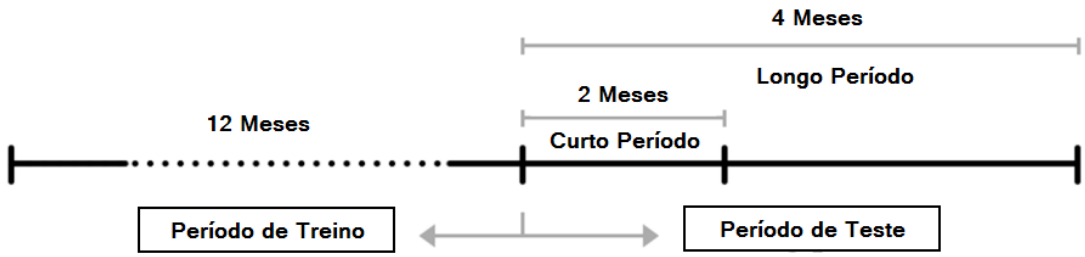


Figura 77 - Períodos Análise CLV

Os dados de teste e de treino foram submetidos a uma limpeza, isto é, unificar os clientes, retirando os novos clientes recém-chegados do período de teste, garantindo que o comparativo de clientes somente é realizado com clientes existentes em ambos os períodos.

No cenário que de seguida apresentamos para avaliar os resultados obtidos foram definidos os seguintes períodos:

- Período de Treino – 12 Meses (14-03-2015 a 13-03-2016)
- Período de Teste – 4 Meses (16-03-2016 a 14-07-2016)

Nestes mesmos períodos e sem aplicar a limpeza de clientes, e sem a definição de RFM, foi possível obter para o período de treino 74333 registos, e para o período de testes 25223 registos.

ENTITY	DATE	TOTAL
0	2015-10-09	5.99
31	2016-01-14	18.68
80	2015-12-02	62.41
189	2016-01-04	38.96
189	2015-12-03	39.96
213	2015-12-29	32.50
213	2015-07-05	36.95
249	2015-09-07	33.10
292	2015-10-09	5.99
292	2015-08-19	10.49
302	2016-02-24	9.99
368	2016-01-19	30.94
370	2015-09-14	19.99

Figura 78 - Dados Treino

ENTITY	DATE	TOTAL
273	2016-04-02	22.48
310	2016-05-25	31.98
310	2016-05-24	37.98
714	2016-05-25	17.95
716	2016-04-12	17.99
718	2016-03-17	34.99
828	2016-05-21	24.99
842	2016-06-30	19.59
882	2016-07-04	17.95
882	2016-05-17	21.23
907	2016-06-22	20.99
966	2016-05-06	9.07
1004	2016-06-03	32.96

Figura 79 - Dados Teste

Uma vez tratados os dados de ambos os períodos e classificados de acordo com o RFM, verifica-se a seguinte amostra para o período de treino foram encontrados 61463 registos, e para o período de teste 22857 registos.

ENTITY	recency_days	transaction_count	amount	Recency	Frequency	Monetary	rfm_score	Buy
512	236	1	18.39	3	1	1	311	0
550	323	1	7.99	2	1	1	211	0
584	192	1	47.98	4	1	5	415	0
591	373	1	37.98	2	1	4	214	0
632	374	1	35.98	2	1	4	214	0
694	326	1	2.99	2	1	1	211	0
741	216	1	33.98	4	1	3	413	0
849	383	1	59.40	2	1	5	215	0
882	377	1	30.68	2	1	3	213	1
891	237	1	35.98	3	1	4	314	0
910	233	1	8.78	3	1	1	311	0

Figura 80 - RFM Período Treino

ENTITY	recency_days	transaction_count	amount	Recency	Frequency	Monetary	rfm_score
273	109	1	22.48	1	1	2	112
310	56	2	69.96	3	5	5	355
714	56	1	17.95	3	1	1	311
716	99	1	17.99	1	1	2	112
718	125	1	34.99	1	1	4	114
828	60	1	24.99	3	1	2	312
842	20	1	19.59	5	1	2	512
882	16	2	39.18	5	5	4	554
907	28	1	20.99	4	1	2	412
966	75	1	9.07	2	1	1	211

Figura 81 - RFM Período Teste

Sobre estes mesmos dados é importante identificar quais os clientes que efetuaram compras no período de treino e igualmente no período de teste, detetando-se assim 4409 clientes correspondendo a 7% de clientes que voltaram a comprar.

ENTITY	recency_days	transaction_count	amount	Recency	Frequency	Monetary	rfm_score	Buy
882	377	1	30.68	2	1	3	213	1
1004	210	1	29.99	4	1	3	413	1
1137	233	1	12.77	3	1	1	311	1
3364	151	1	9.07	5	1	1	511	1
5200	226	1	51.99	4	1	5	415	1
5226	371	1	15.47	2	1	1	211	1
5282	175	2	111.33	5	5	5	555	1
5447	142	1	54.50	5	1	5	515	1
5487	166	2	72.12	5	5	5	555	1
5667	136	1	24.76	5	1	2	512	1
5693	189	1	45.98	4	1	4	414	1
5735	177	1	30.97	5	1	3	513	1
5777	193	1	27.96	4	1	3	413	1

Figura 82 - Clientes em Ambos Períodos

Com estes dados calculados, e identificados os clientes com compras em ambos os períodos pode-se assim segmentar os mesmos conforme cada métrica do RFM e avaliar a mesma [54].

Avaliando os clientes de acordo com a Recency (tempo desde a última compra) os mesmos ficam segmentados em 5 grupos, onde o grupo 1 são os clientes mais antigos, e o grupo 5 os clientes mais recentes.

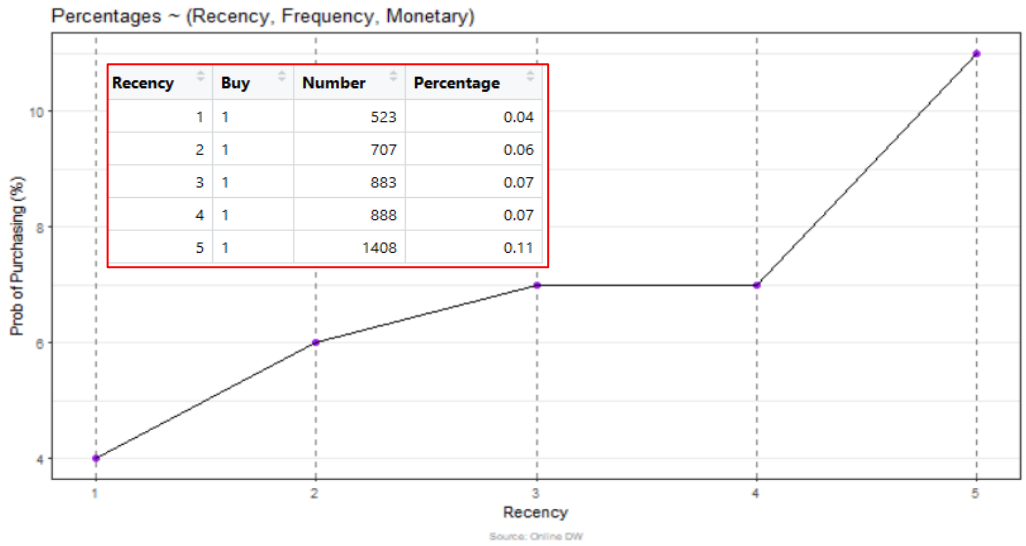
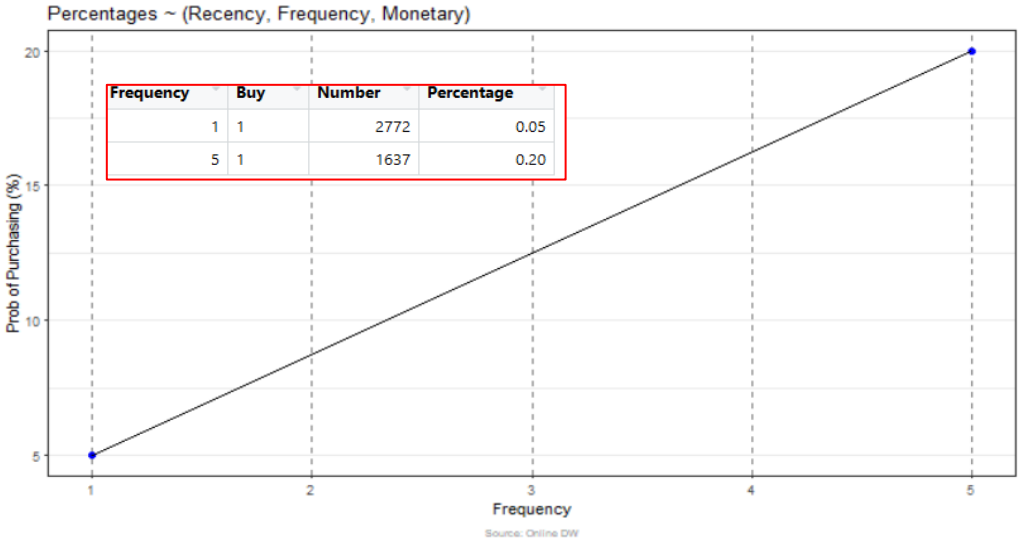


Figura 83 - Probabilidade Compra - Recency

Pode-se então indicar que a percentagem de clientes do grupo 5 (clientes mais recentes) que voltaram a comprar foi de 11%.

Avaliando os clientes de acordo com a Frequency (número de compras efetuadas) os mesmos ficam segmentados em 2 grupos, onde o grupo 1 são os clientes menos, e o grupo 5 os clientes mais frequentes.

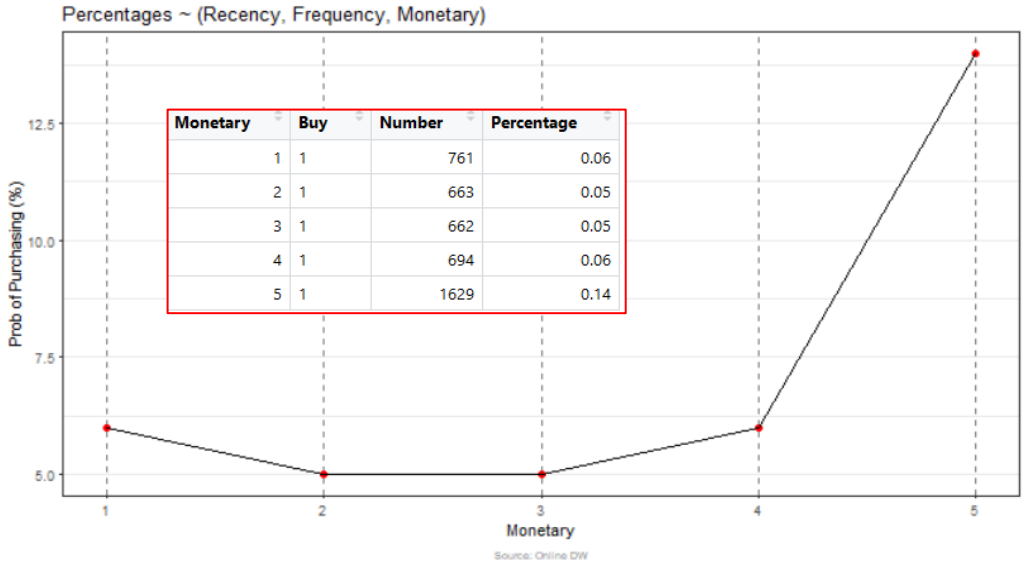


Frequency	Buy	Number	Percentage
1	1	2772	0.05
5	1	1637	0.20

Figura 84 - Probabilidade Compra - Frequency

Verifica-se então que no grupo de clientes mais frequentes (grupo 5) estes foram 20% que voltaram a comprar no período de teste face ao período de treino.

Por último e avaliando os clientes de acordo com a Monetary (valor gasto em compras) os mesmos ficam segmentados em 5 grupos, onde o grupo 1 são os clientes com menor valor de compra, e o grupo 5 os clientes com maior valor de compra.



Monetary	Buy	Number	Percentage
1	1	761	0.06
2	1	663	0.05
3	1	662	0.05
4	1	694	0.06
5	1	1629	0.14

Figura 85 - Probabilidade Compra - Monetary

Neste cenário existe uma constatação importante, que o grupo de clientes de maior valor realizado em compras no período de treino, corresponde a 14% de compras realizadas no período de teste.

Análise ao Cliente

Selecionando um cliente em específico demonstra-se como proceder para calcular o seu CLV.

ENTITY	DATE	TOTAL
741	2015-12-17	33.98
849	2015-07-03	59.40
882	2015-07-09	30.68
891	2015-11-26	35.98

Figura 86 - Compras P. Treino

ENTITY	DATE	TOTAL
828	2016-05-21	24.99
842	2016-06-30	19.59
882	2016-07-04	17.95
882	2016-05-17	21.23
907	2016-06-22	20.99

Figura 87 - Compras P. Teste

Este cliente como se pode constatar tem compras no período de treino assim como no período de teste, até podemos verificar que no período de teste já realizou mais compras do que no seu período de treino.

Seguidamente é necessário calcular o RFM para o cliente no seu período de treino, assim como identificar se este fez ou não compras no período de teste (coluna Buy=1).

ENTITY	recency_days	transaction_count	amount	Recency	Frequency	Monetary	rfm_score	Buy
741	216	1	33.98	4	1	3	413	0
849	383	1	59.40	2	1	5	215	0
882	377	1	30.68	2	1	3	213	1
891	237	1	35.98	3	1	4	314	0
910	233	1	8.78	3	1	1	311	0
915	291	1	45.87	2	1	4	214	0

Figura 88 - Cliente P. Treino

De igual forma também se calcula o RFM do cliente para o seu período de teste, para posterior comparação.

ENTITY	recency_days	transaction_count	amount	Recency	Frequency	Monetary	rfm_score
828	60	1	24.99	3	1	2	312
842	20	1	19.59	5	1	2	512
882	16	2	39.18	5	5	4	554
907	28	1	20.99	4	1	2	412
966	75	1	9.07	2	1	1	211

Figura 89 - Cliente P. Teste

É possível constatar, que o cliente para além de ter feito um maior número de encomendas, também teve um maior valor de aquisição, resultando isto na sua mudança de RFM nos

diferentes períodos, de um cliente RFM 213, para um cliente RFM 554, demonstrando uma evolução positiva do cliente.

Para a realização do cálculo do CLV do cliente procede-se a 4 etapas de calculo diferentes, respetivamente:

- AOV (Average Order Value) - Valor Médio da Encomenda
- PF (Purchase Frequency) - Frequência de Compra
- CV (Customer Value) - Valor do Cliente
- CAL (Customer Average Lifespan) - Duração Média do Cliente
- CLV (Customer Lifetime Value) - Valor do Tempo de Vida do Cliente

Inicia-se então calculando o AOV para o cliente, que representa o valor médio monetário que um cliente efetua em cada encomenda. Para obter esse valor divide-se o valor total monetário nas diversas encomendas, pelo número de encomendas realizadas.

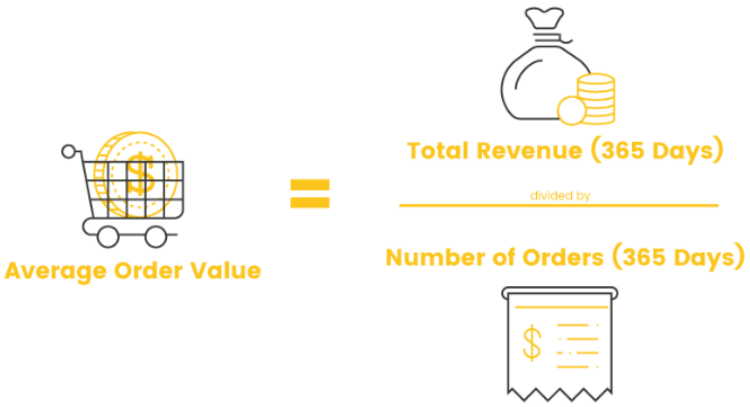


Figura 90 - AOV Formula[55]

- Valor Médio Encomenda = Total Monetário Compras / Número de Compras
    - Average Order Value = Total Sales / Order Count

Para o cliente selecionado temos então o cálculo de:

- Valor Médio Encomenda = 30,68€ / 1 Encomenda = 30,68€

O cálculo do PF representando a quantidade média de encomendas por cliente naquele período, é calculado pelo número total de encomendas dividido pelo número total de clientes na amostra.

Figura 91 - PF Formula[55]

- Frequência de Compra = Total Encomendas / Número de Clientes
    - Purchase Frequency = Total Orders / Total Customers

Para o período selecionado temos então o cálculo de:

- Frequência de Compra = 7433 Encomendas / 61463 Clientes = 1,20939 Encomendas

Para o cálculo do CV que representa o valor médio do cliente para o período indicado, calculado através Valor Médio Encomenda multiplicado pela Frequência de Compra.

Figura 92 - CV Formula[55]

- Valor Cliente = Valor Médio Encomenda x Frequência de Compra
    - Customer Value = Average Order Value X Purchase Frequency

Para o período selecionado temos então o cálculo de:

- Valor Cliente = 30,68€ x 1,20939 Encomendas = 37,104€

Por último e para que seja possível calcular o CLV de cada cliente tem que se conseguir calcular o CAL, ou seja a duração média do cliente, isto é, o tempo médio em que o cliente está ativo antes de deixar de comprar, ficando como inativo. No entanto os dados existentes não permitem com exatidão calcular este valor, e como tal tem de se utilizar o valor de 3 anos, uma vez que é o defendido por especialistas [55].

Figura 93 - CAL Formula[55]

- Duração Média Cliente = 3 anos
    - Customer Average Lifespan = 3 Years

Reunidos todos os cálculos necessários, e anteriormente explicados como foram obtidos, pode então calcular o CLV para o nosso cliente.

Figura 94 - CLV Formula[55]

- Valor do Tempo de Vida do Cliente = Valor Cliente x Duração Média Cliente
    - Customer Lifetime Value = Customer Value x Customer Average Lifespan

Para o período selecionado temos então o cálculo de:

- Valor do Tempo de Vida do Cliente = 37,104€ x 3 anos= 111,27€

Foi obtido o valor de 111,27€ que é o potencial valor que o cliente esteja disposto a gastar no período de 3 anos, em que se espera que este esteja ativo.

6 Demonstração da Solução

O capítulo “Demonstração da Solução” pretende dar ao leitor uma interação da plataforma, de todos os ecrãs disponíveis, seu funcionamento e informação disponível, abordando em detalhe todos os seus diferentes componentes.

6.1 Painel de Indicadores

O Painel de Indicadores é o primeiro contacto do utilizador, ao iniciar a plataforma este é carregado com as informações mais relevantes para o dia a dia da equipa Online.

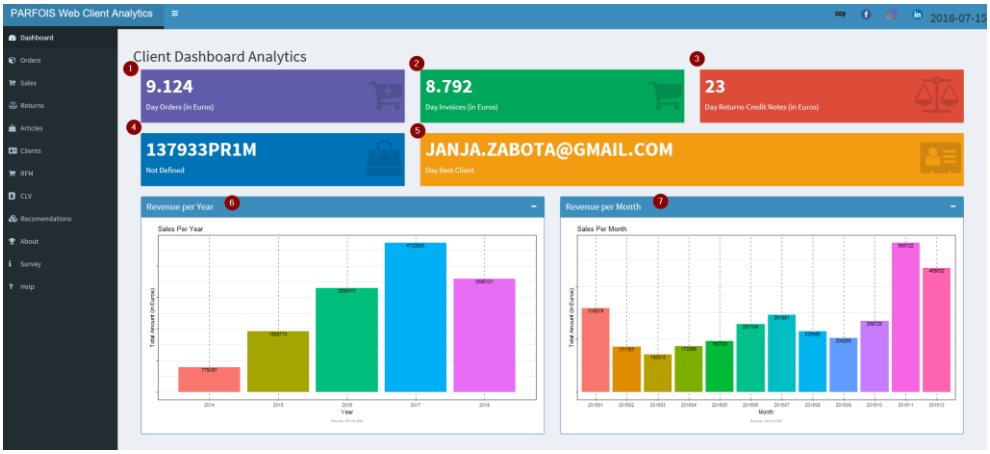


Figura 95 - Dashboard

Segue uma breve descrição dos pontos acima identificados:

1. Valor diário de encomendas
2. Valor diário de faturas
3. Valor diário de devoluções (notas de crédito)
4. Artigo mais vendido do dia
5. Melhor cliente do dia
6. Crescimento anual de vendas
7. Crescimento mensal de vendas

Esta informação permite ter uma perspetiva, anual, mensal e diária dos principais indicadores.

6.2 Encomendas

Na componente Encomendas, o utilizador tem a possibilidade de consultar para determinada data o volume de encomendas realizadas, e ter a perceção da sua grandeza comparativamente com os cinco dias anteriores e cinco dias posteriores.

O estado das diversas encomendas do dia selecionado, igualmente é verificado no gráfico disponibilizado.

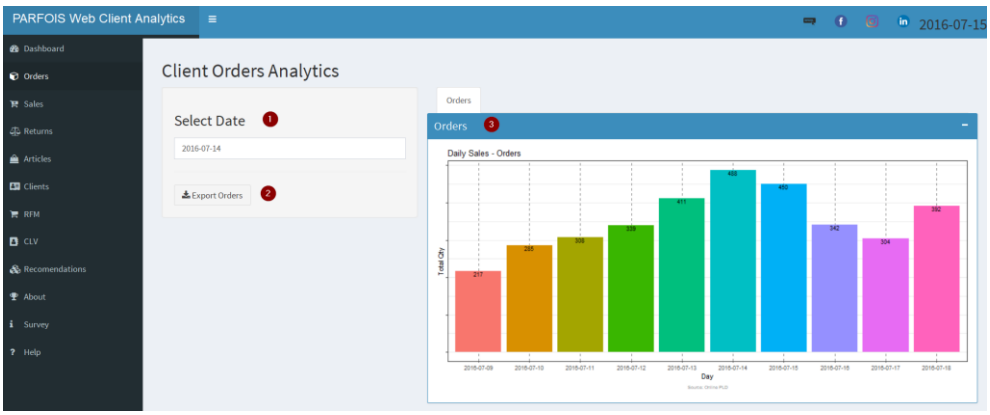


Figura 96 - Orders

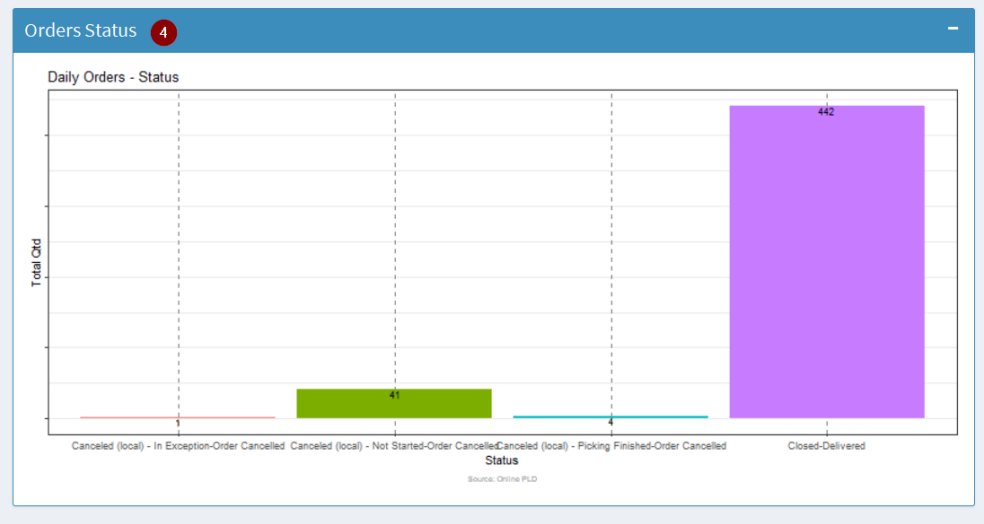


Figura 97 - Order Status

Segue uma breve descrição dos pontos acima identificados:

1. Seleção de data
2. Exportação para PDF dos gráficos definidos
3. Progressão de encomendas
4. Estado das encomendas do dia

Igualmente é facultada a possibilidade de exportar para report (pdf) os gráficos disponibilizados.

6.3 Vendas

Na componente Vendas, o utilizador poderá consultar para determinada data o volume de vendas realizadas, e ter a perceção da sua grandeza comparativamente com os cinco dias anteriores e cinco dias posteriores. Também pode ser validada a quantidade de produtos vendidos para cada departamento na data selecionada.

Nesta componente existem duas secções, apresentando informação disponibilizada pelo ERP, e informação da plataforma online (PLD).

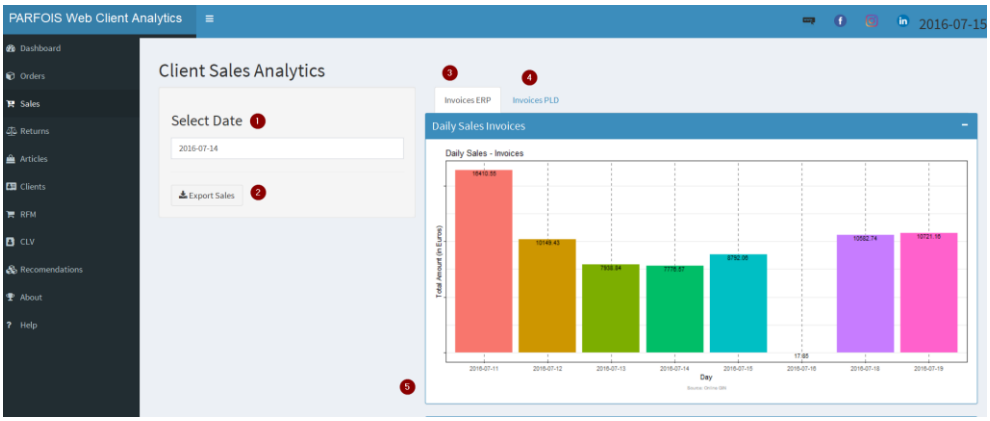


Figura 98 - Sales

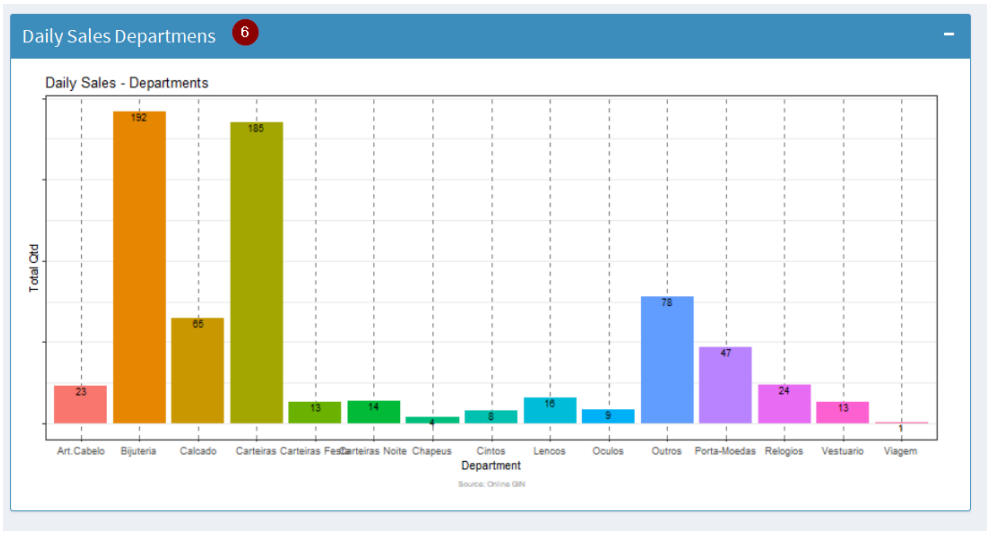


Figura 99 - Sales Departments

Figura 100 - Sales PLD

Segue uma breve descrição dos pontos acima identificados:

1. Seleção de data
2. Exportação para PDF dos gráficos definidos
3. Tab com informação de faturas do ERP
4. Tab com informação de faturas da plataforma online
5. Progressão de faturas ERP
6. Informação de faturas ERP por departamento/gamas
7. Progressão de faturas da plataforma online

Igualmente é facultada a possibilidade de exportar para report (pdf) os gráficos disponibilizados.

6.4 Devoluções

Na componente Devoluções, o utilizador poderá consultar para determinada data o volume de devoluções realizadas pelo Cliente, e ter a perceção da sua grandeza comparativamente com os cinco dias anteriores e cinco dias posteriores. Estas poderão dar ou não origem a Notas de Crédito conforme a aceitação da devolução.

Também pode ser validada a quantidade de produtos devolvidos para cada departamento na data selecionada.

Nesta componente existem duas secções, apresentando informação relativamente a Notas de Crédito, e informação de Devoluções.

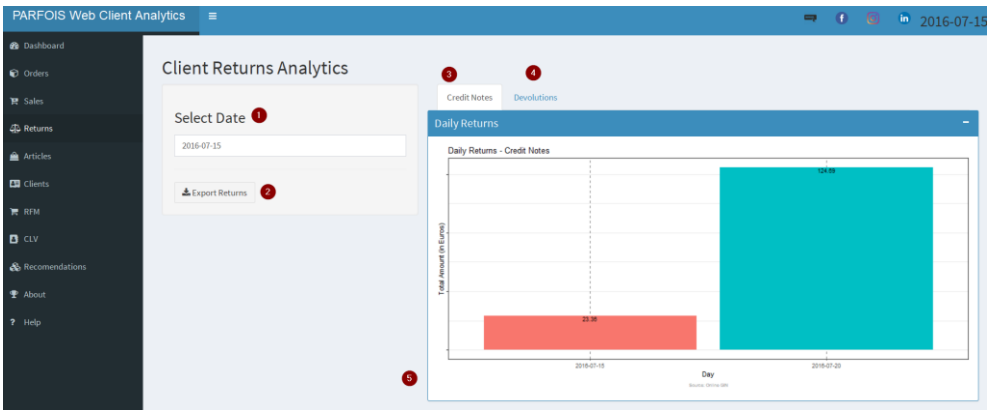


Figura 101 - Returns Credit Notes

Figura 102 - Returns Credit Notes Departments

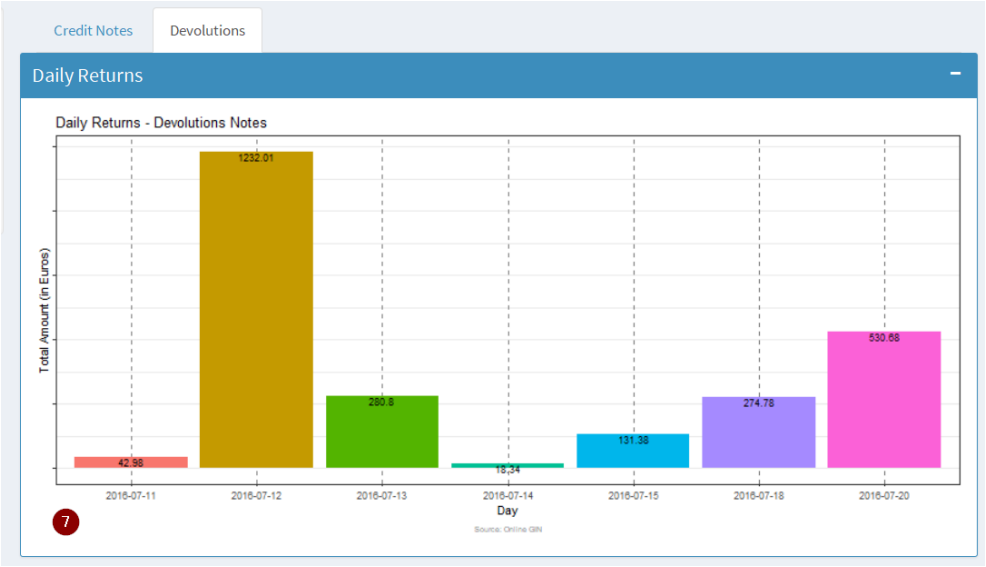


Figura 103 - Returns Devolutions

Figura 104 - Returns Devolutions Departments

Segue uma breve descrição dos pontos acima identificados:

1. Seleção de data
2. Exportação para PDF dos gráficos definidos
3. Tab com informação de notas de crédito
4. Tab com informação de devoluções
5. Progressão de notas de crédito
6. Informação de notas de crédito por departamento/gamas
7. Progressão de devoluções
8. Informação de devoluções por departamento/gamas

Igualmente é facultada a possibilidade de exportar para report (pdf) os gráficos disponibilizados.

6.5 Artigos

Uma das componentes mais completas é a componente Artigos, o utilizador poderá realizar diversas consultas sobre os artigos para o Ano selecionado.

Nesta componente existem quatro secções, apresentando informação de artigos que incide essencialmente sobre as Vendas e Devoluções de Clientes, onde pode ser detalhada para aquele Ano, Ano anterior, e Ano posterior.

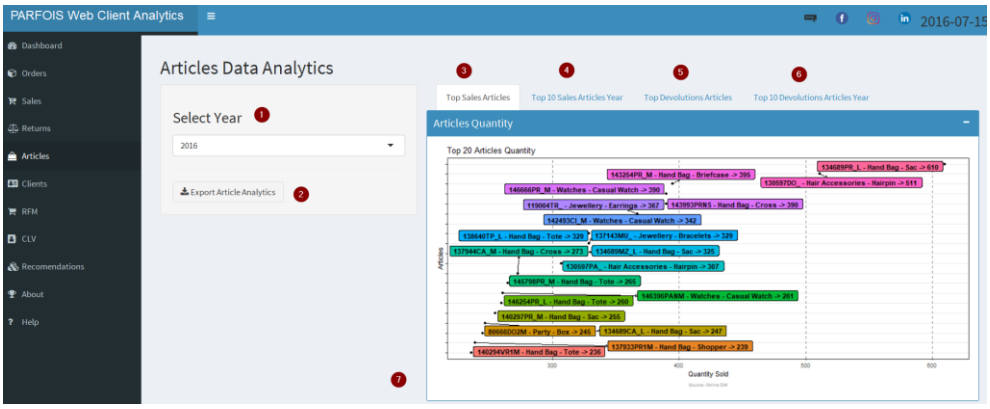


Figura 105 - Articles Tops

Segue uma breve descrição dos pontos acima identificados:

1. Seleção de data
2. Exportação para PDF dos gráficos definidos
3. Tab com informação de Vendas - Top Artigos
4. Tab com informação de Vendas - Top 10 Artigos/Ano
5. Tab com informação de Devoluções - Top Artigos
6. Tab com informação de Devoluções - Top 10 Artigos/Ano
7. Vendas - Top 20 Artigos Vendidos (em quantidade) (Ano Selecionado)
8. Vendas - Top 20 Artigos Vendidos (em valor) (Ano Selecionado)
9. Vendas - Top 10 Artigos Vendidos (em quantidade) (Ano Selecionado, Posterior e Anterior)
10. Vendas - Top 10 Artigos Vendidos (em valor) (Ano Selecionado, Posterior e Anterior)
11. Vendas - Top 20 Artigos Devolvidos (em quantidade) (Ano Selecionado)
12. Vendas - Top 20 Artigos Devolvidos (em valor) (Ano Selecionado)
13. Vendas - Top 10 Artigos Devolvidos (em quantidade) (Ano Selecionado, Posterior e Anterior)
14. Vendas - Top 10 Artigos Devolvidos (em valor) (Ano Selecionado, Posterior e Anterior)

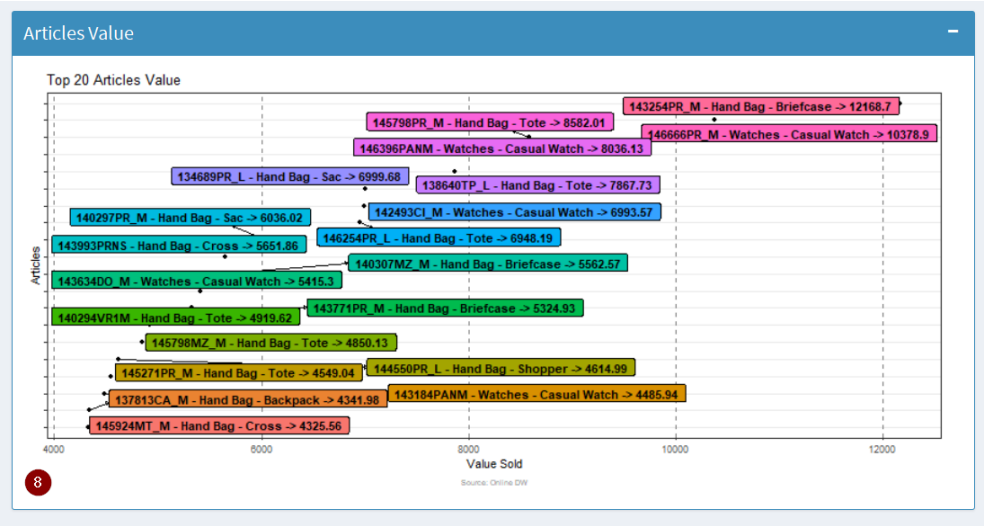


Figura 106 - Top 20 Articles Sales Value

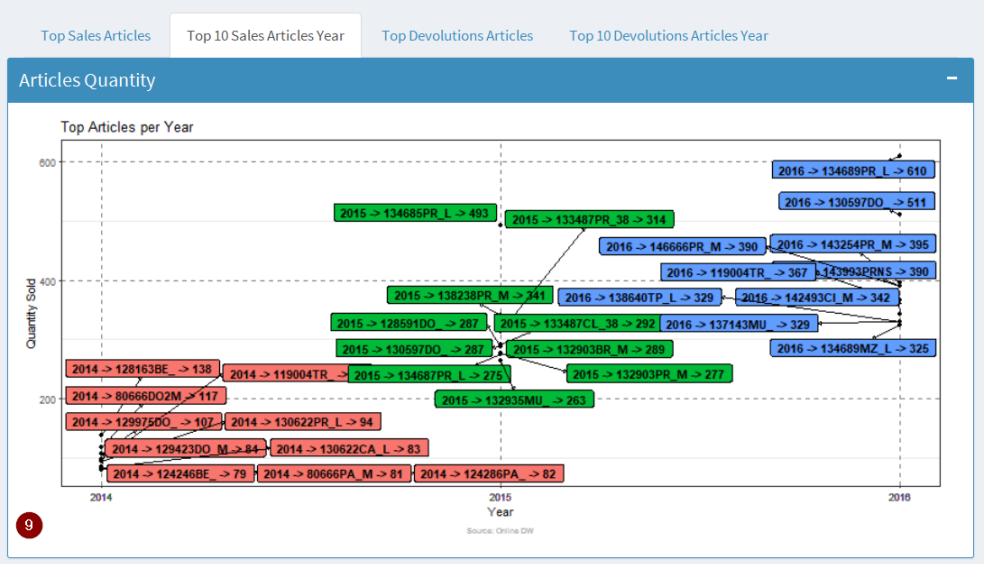


Figura 107 - Top 10 Articles Sales Year Quantity

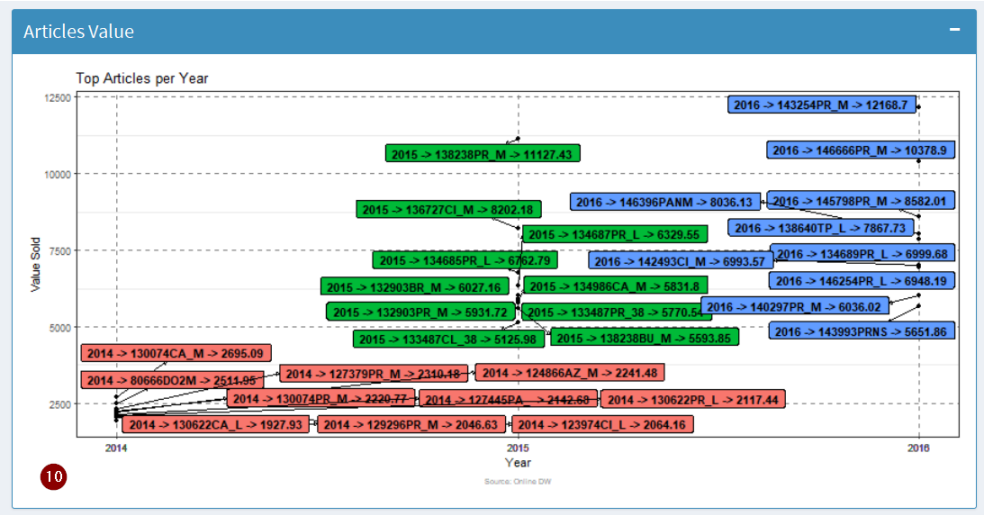


Figura 108 - Top 10 Articles Sales Year Value

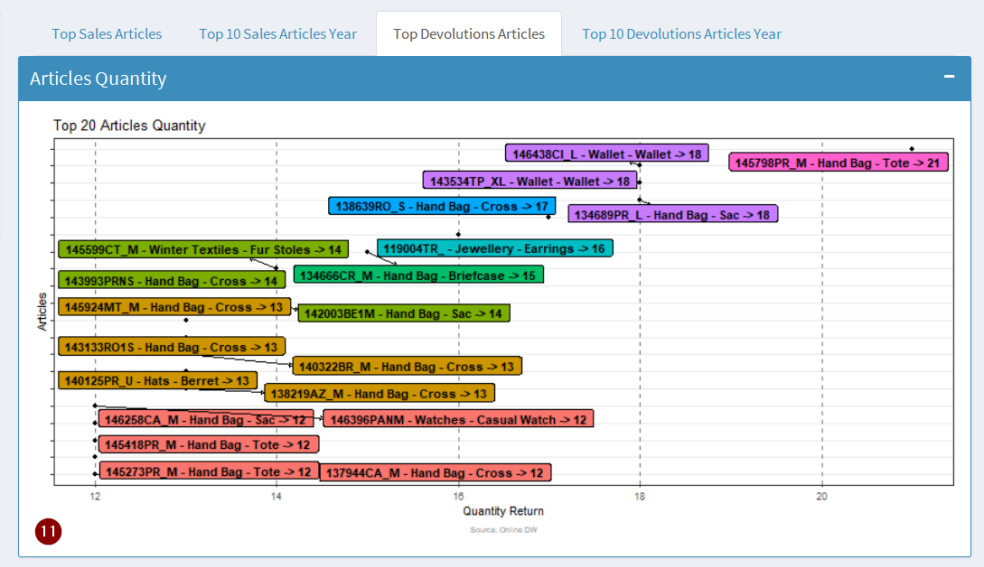


Figura 109 - Top 20 Articles Returns Year Quantity

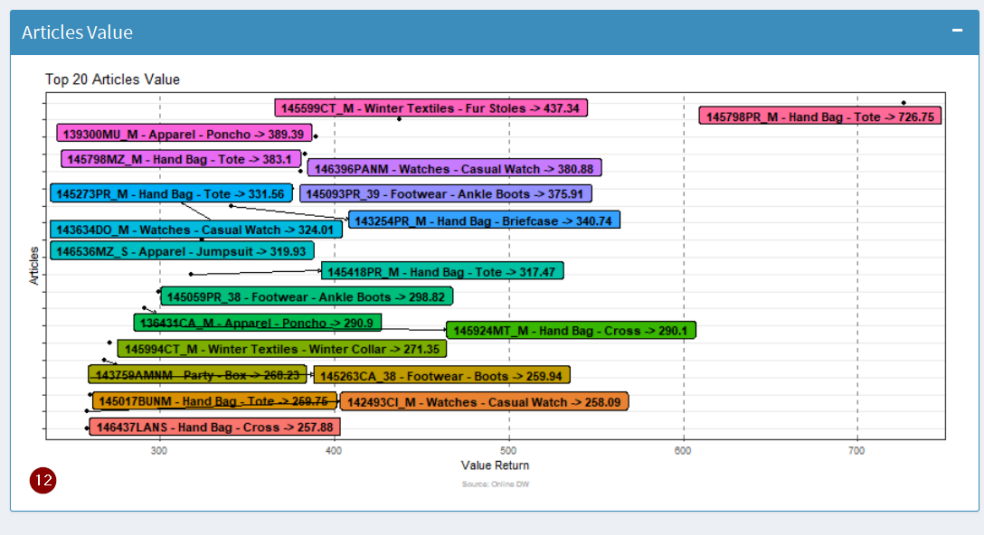


Figura 110 - Top 20 Articles Returns Year Value

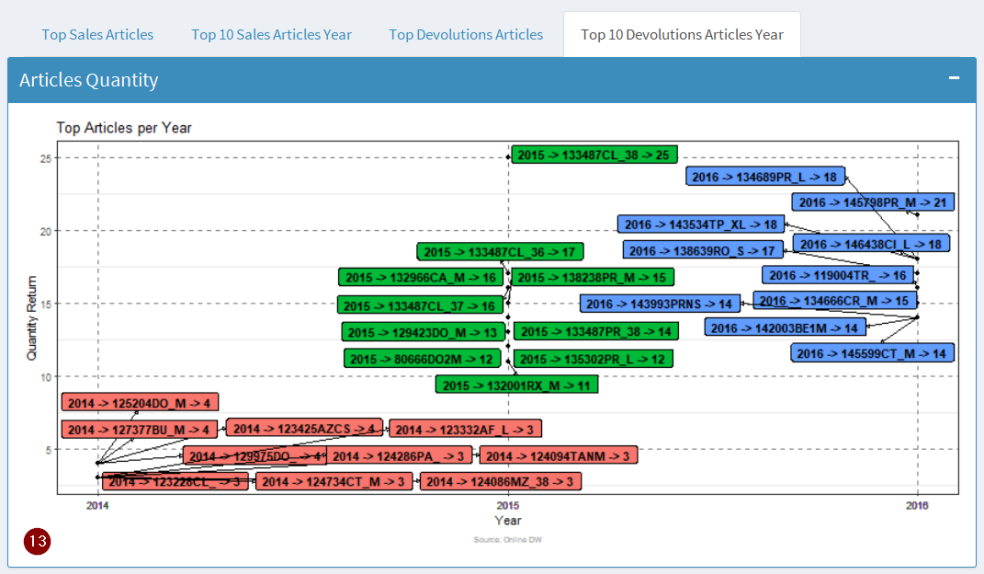


Figura 111 - Top 10 Articles Returns Year Quantity

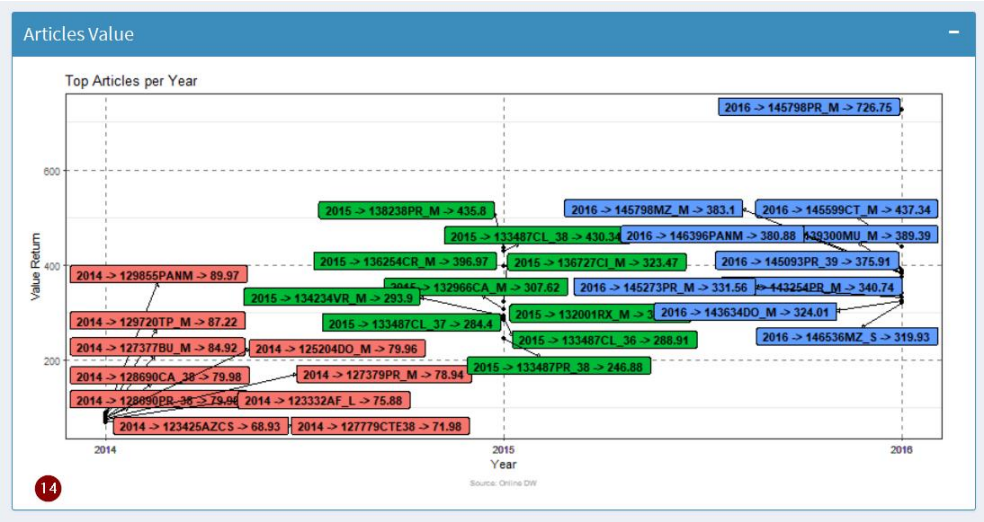


Figura 112 - Top 10 Articles Returns Year Value

Igualmente é facultada a possibilidade de exportar para report (pdf) os gráficos disponibilizados.

6.6 Clientes

Na componente Clientes, o utilizador poderá realizar diversas consultas sobre os principais clientes para o Ano selecionado.

Nesta componente existem duas secções, uma apresentando informação de Top Clientes do ano selecionado, e a outra secção com os Top 5 Clientes para os três últimos a anos.

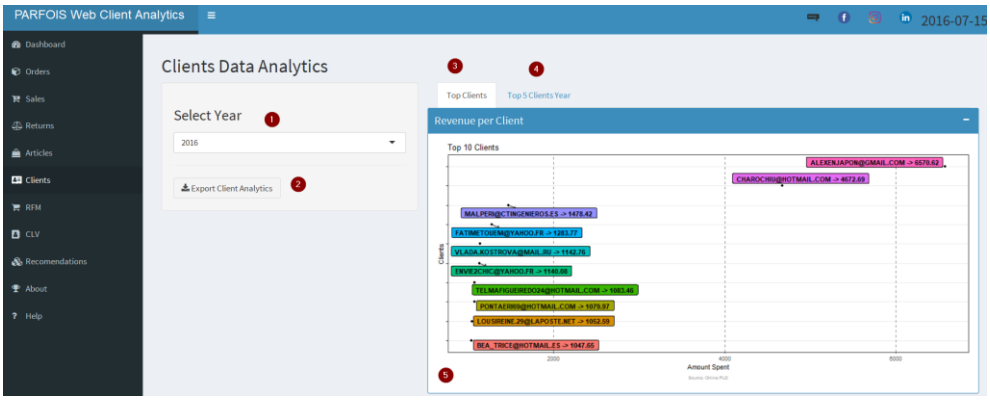


Figura 113 - Top 10 Clientes

Segue uma breve descrição dos pontos acima identificados:

1. Seleção de ano
2. Exportação para PDF dos gráficos definidos
3. Tab com informação de Top Clientes
4. Tab com informação de Top Clientes 5 Anos
5. Clientes - Top 10 Clientes (em valor) (Ano Selecionado)
6. Clientes - Top 5 Clientes/Ano (em valor) (Ano Selecionado e Anteriores)

Figura 114 - Top 5 Clientes/Ano

Igualmente é facultada a possibilidade de exportar para report (pdf) os gráficos disponibilizados.

6.7 RFM

Uma das componentes com mais valor acrescentado é a componente RFM, o utilizador poderá realizar diversas consultas para o período indicado.

Nesta componente existem três secções, apresentando informação de RFM que incide essencialmente sobre as Recency (cliente mais recentes), Frequency (clientes mais frequentes) e Monetary (clientes com mais gastos).

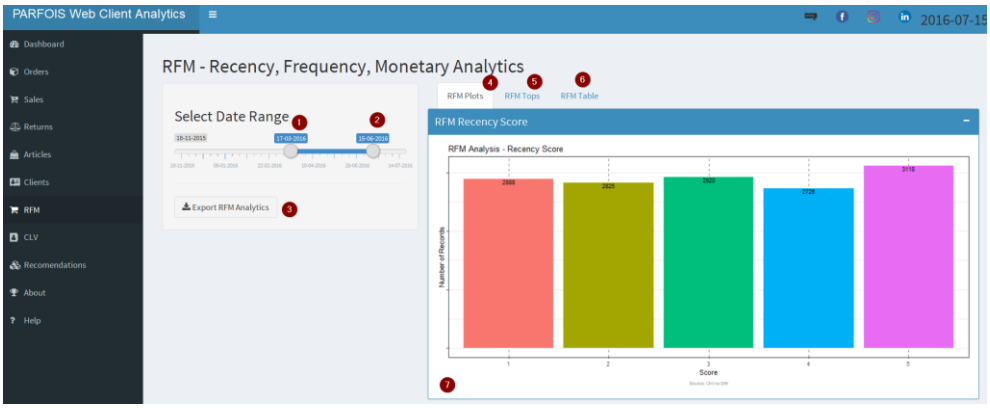


Figura 115 - RFM Recency

Segue uma breve descrição dos pontos acima identificados:

1. Seleção de data início analise
2. Seleção de data fim analise
3. Exportação para PDF dos gráficos definidos
4. Tab com informação de RFM Plots
5. Tab com informação de RFM Tops
6. Tab com informação de RFM Table
7. RFM – Recency Score

Figura 116 - RFM Monetary

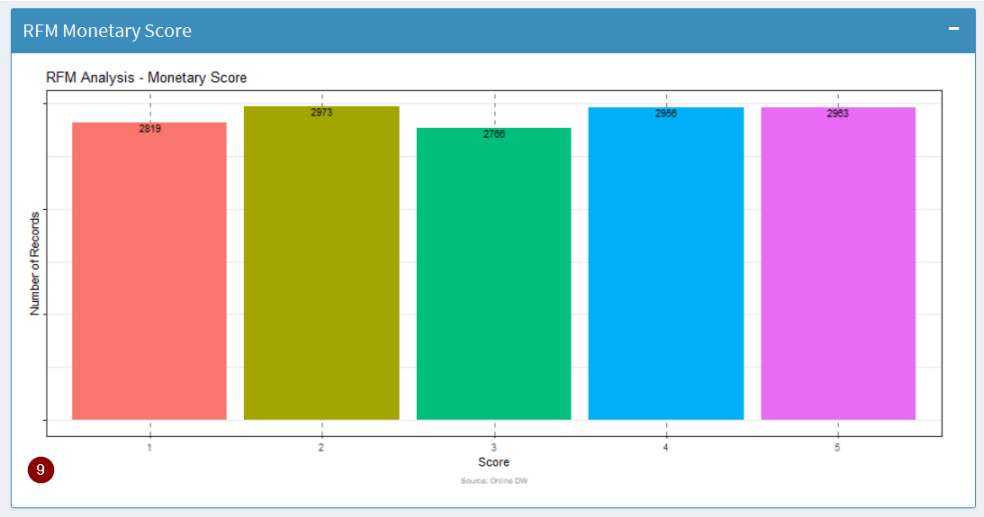


Figura 117 - RFM Monetary

Também é disponibilizado os clientes Tops RFM onde é percetível a variação dos diferentes grupos de clientes.

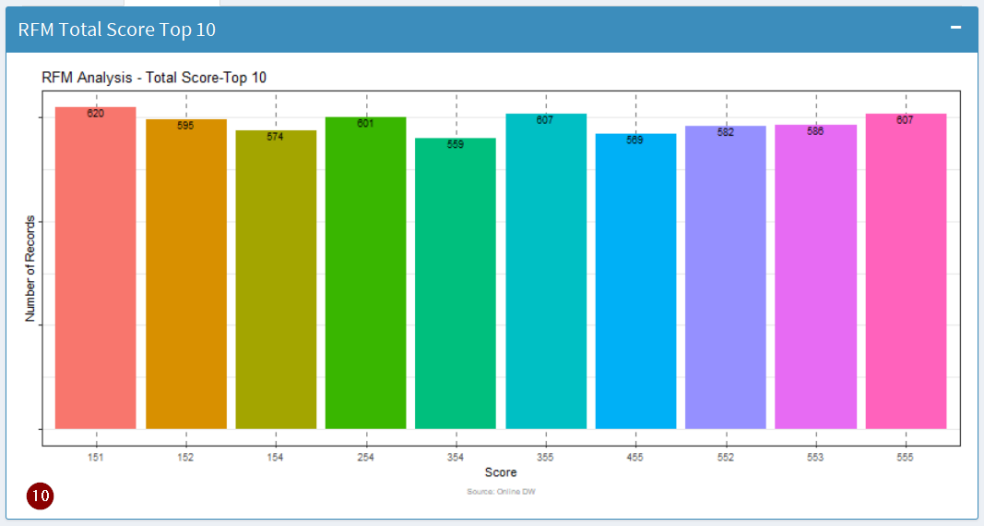


Figura 118 - RFM Top 10

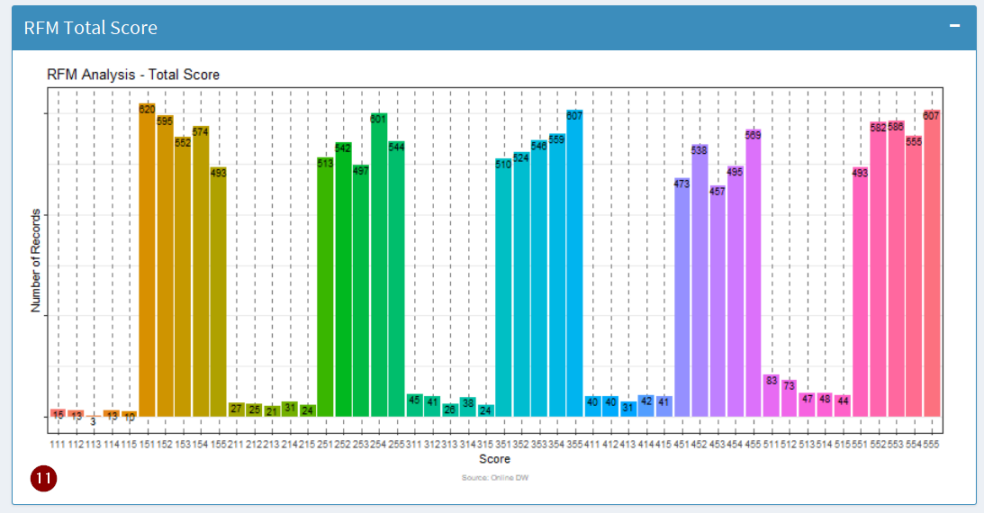


Figura 119 - RFM Tops

E a secção RFM Table onde podemos consultar os dados em formato tabela e exportar os mesmos para Excel.

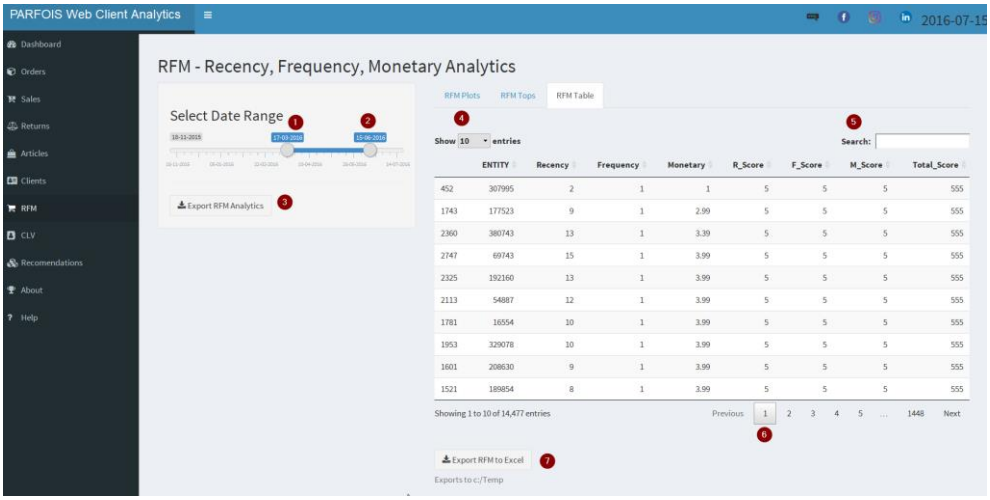


Figura 120 - RFM Table

Igualmente é facultada a possibilidade de exportar para report (pdf) os gráficos disponibilizados em toda a componente do RFM.

6.8 CLV

A componente CLV é bastante complexa, uma vez que o utilizador para obter resultados mais fiáveis deverá respeitar as regras indicadas, nomeadamente o uso de um período histórico de 12 meses, e um período de comparação de 4 meses.

Nesta componente existem duas secções, apresentando a informação de CLV que incide essencialmente sobre a Probabilidade de Compra sobre as diferentes métricas (Recency, Frequency, Monetary).

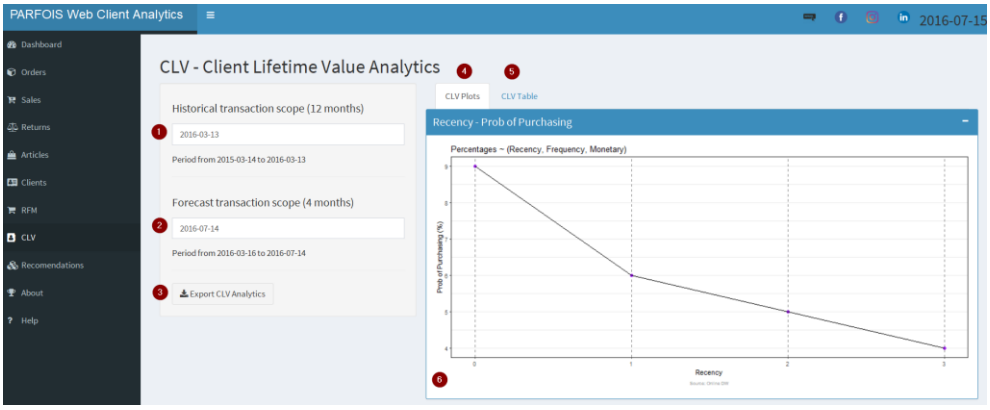


Figura 121 - CLV Recency

Segue uma breve descrição dos pontos acima identificados:

1. Seleção de data (para efeito do histórico 12 meses)
2. Seleção de data (para efeito do forecast 4 meses)
3. Exportação para PDF dos gráficos definidos
4. Tab com informação de CLV Plots
5. Tab com informação de CLV Table
6. CLV – Recency Probabilidade de Compra

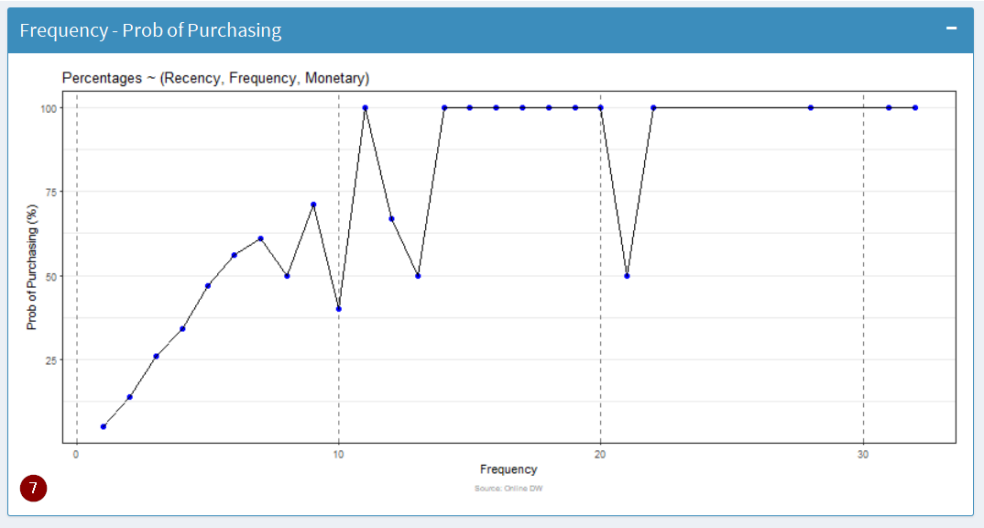


Figura 122 - CLV Frequency

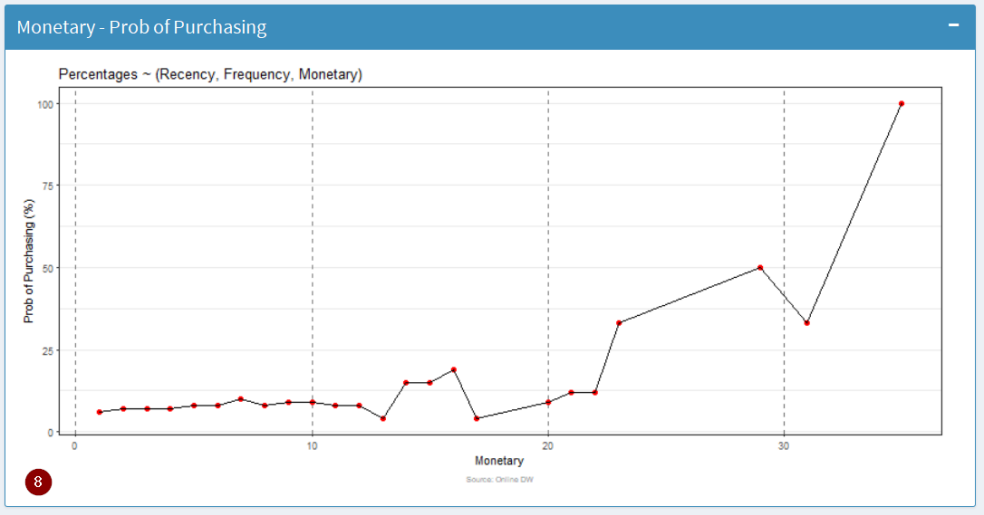


Figura 123 - CLV Monetary

Na secção CLV Table podemos consultar os dados em formato tabela e exportar os mesmos para Excel.

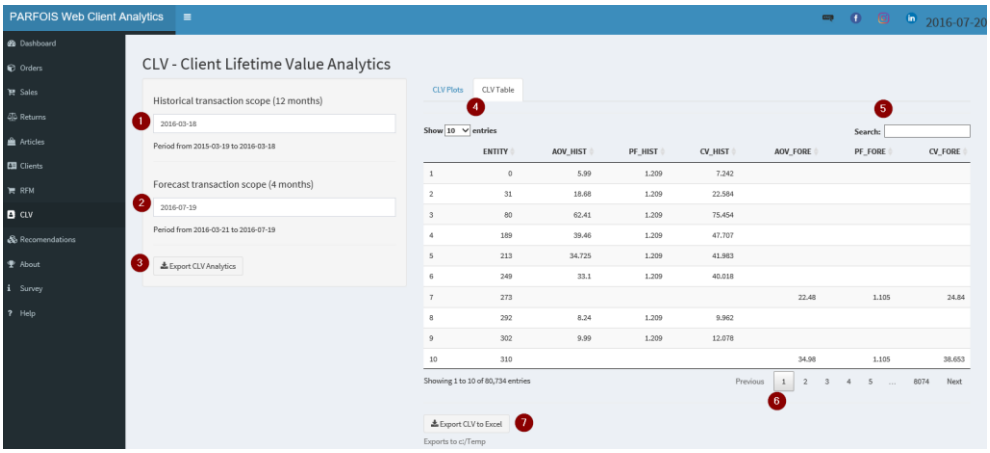


Figura 124 - CLV Table

Igualmente é facultada a possibilidade de exportar para report (pdf) os gráficos disponibilizados.

6.9 Recomendações a Clientes

Na componente Recomendações a Clientes, o utilizador poderá gerar recomendações para os clientes considerando um determinado período selecionado.

Igualmente é disponibilizada a opção de selecionar diferentes algoritmos para gerar as recomendações, nomeadamente os algoritmos:

- *POPULAR* - sugestão dos artigos mais populares
- *UBCF - User Based Collaborative Filtering*
- *IBCF - Item Based Collaborative Filtering*
- *RANDOM* - sugestão aleatória de artigos

O utilizador poderá ainda indicar quantas recomendações pretende gerar para cada cliente.

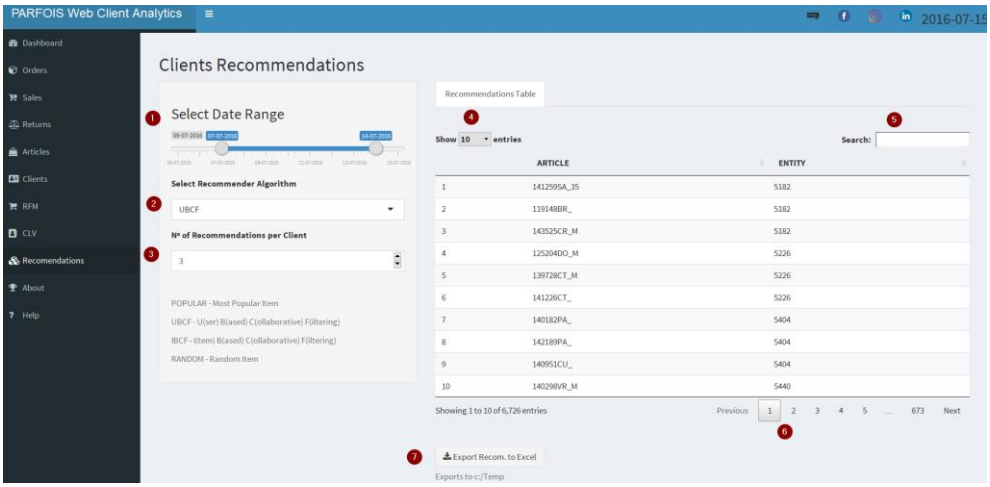


Figura 125 - Clients Recommendations

Segue uma breve descrição dos pontos acima identificados:

1. Seleção de período de data
2. Seleção de algoritmo (disponíveis 4 algoritmos diferentes)
3. Seleção do número de recomendações por cliente
4. Indicação do número de linhas a serem visíveis na tabela
5. Pesquisa de informação na tabela
6. Navegação nos próximos registos da tabela
7. Exportação para Excel da tabela definida

Nesta secção podemos consultar os dados em formato tabela e exportar os mesmos para Excel.

6.10 Acerca do Autor

A componente Acerca do Autor, é somente de cariz informativo, resultando em informação acerca do âmbito do projeto desenvolvido.

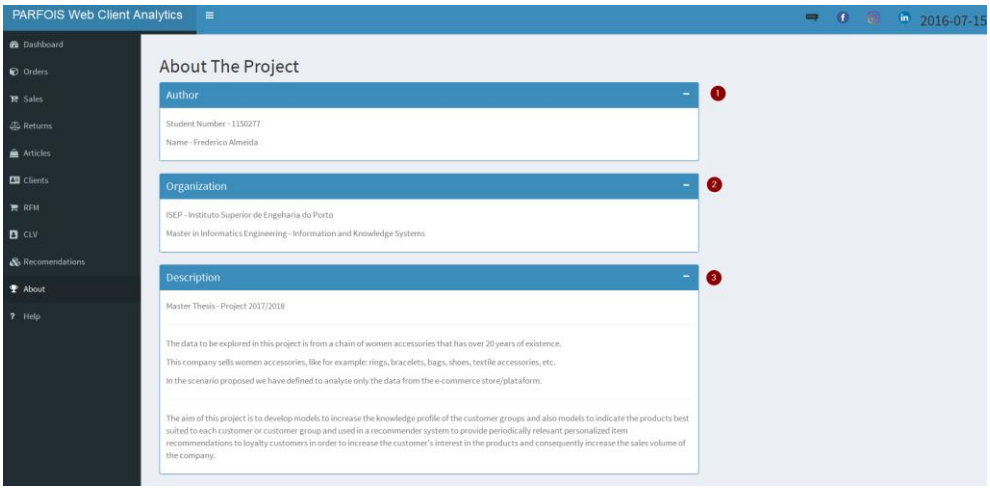


Figura 126 - About

6.11 Inquérito

A componente Inquérito, permitirá ao utilizador realizar um pequeno inquérito de satisfação da plataforma desenvolvida, possibilitando assim dar o seu feedback acerca da mesma.

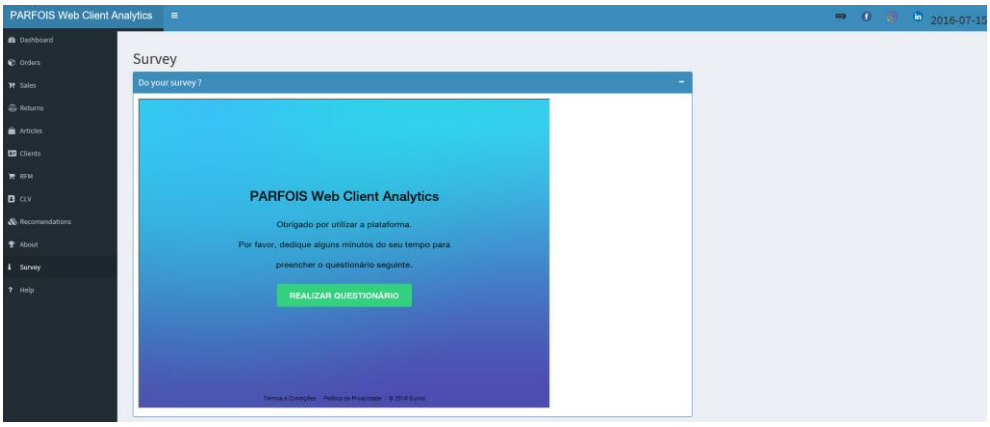


Figura 127 - Survey

6.12 Ajuda

A componente de Ajuda, permitirá ao utilizador obter mais informação acerca dos tópicos mais focados pela plataforma, nomeadamente o tema RFM e o tema CLV, acedendo assim diretamente para a informação existente na Wikipédia sobre estes temas.

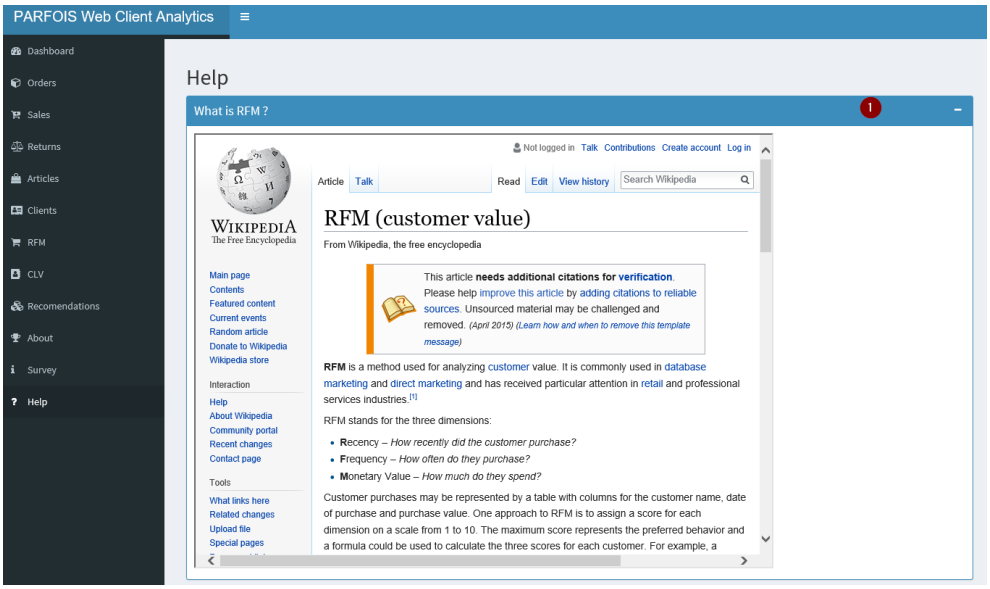


Figura 128 - Help RFM

Figura 129 - Help CLV

6.13 Relatórios

Nas diferentes componentes desenvolvidas, é possibilitado ao utilizador exportar para PDF os diferentes gráficos gerados.

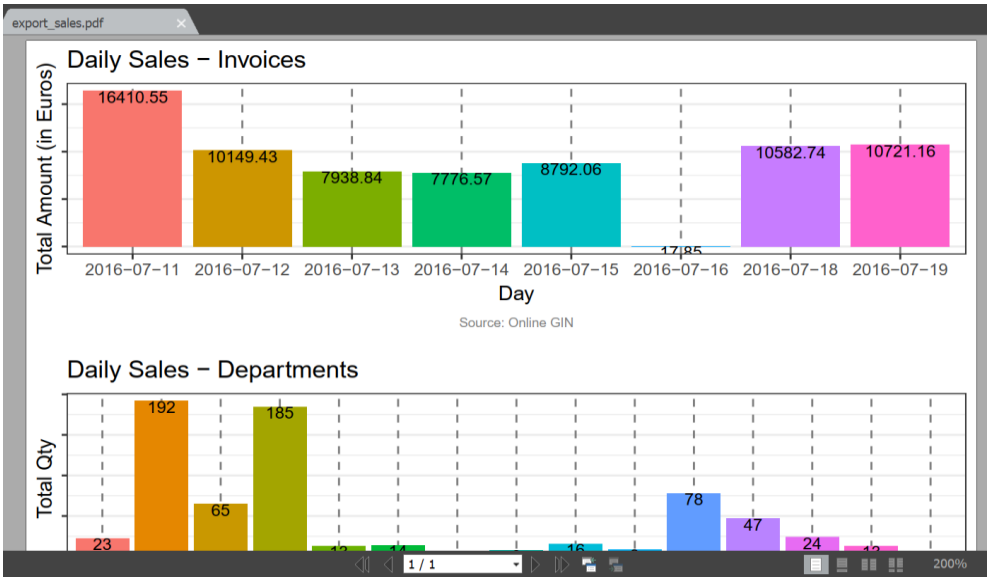
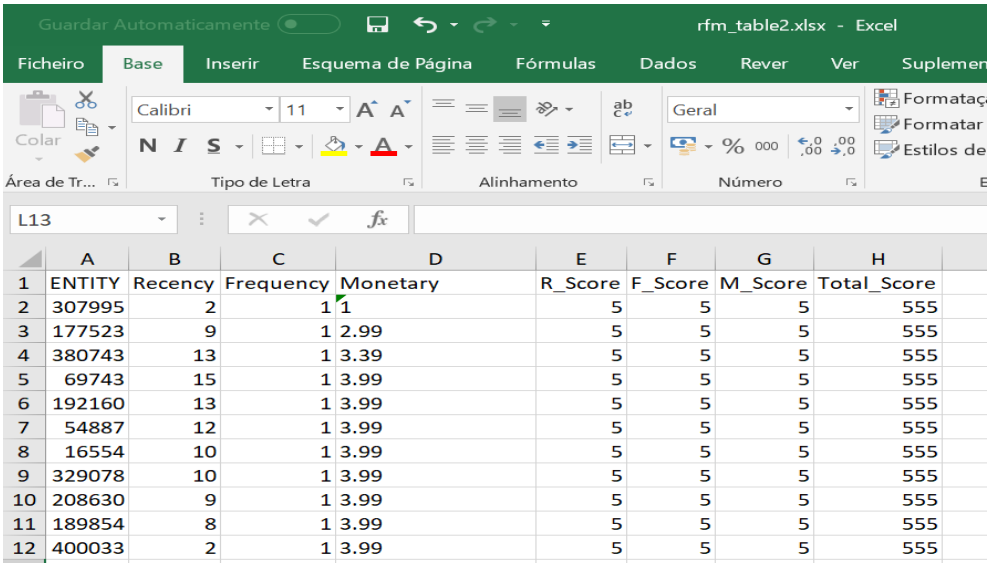


Figura 130 - Exportação Reports

6.14Exportação Excel

Em algumas das componentes desenvolvidas, e nomeadamente aquando a apresentação dos dados no formato tabela, é possibilitado ao utilizador exportar para Excel estes mesmos dados.



	A	B	C	D	E	F	G	H
1	ENTITY	Recency	Frequency	Monetary	R_Score	F_Score	M_Score	Total_Score
2	307995	2	1	1	5	5	5	555
3	177523	9	1	2.99	5	5	5	555
4	380743	13	1	3.39	5	5	5	555
5	69743	15	1	3.99	5	5	5	555
6	192160	13	1	3.99	5	5	5	555
7	54887	12	1	3.99	5	5	5	555
8	16554	10	1	3.99	5	5	5	555
9	329078	10	1	3.99	5	5	5	555
10	208630	9	1	3.99	5	5	5	555
11	189854	8	1	3.99	5	5	5	555
12	400033	2	1	3.99	5	5	5	555

Figura 131 - Exportação para Excel

7 Avaliação da Solução

O capítulo "Avaliação da Solução" é para o leitor a constatação da veracidade do resultado obtido e permite ao leitor ter a perceção de quais as métricas para a avaliação da mesma.

7.1 Avaliação Global

Para se avaliar a solução desenvolvida teremos de dividir em 4 fases distintas de avaliação, uma vez que se diferenciam entre elas, embora relacionadas como um todo. Seguidamente vamos detalhar cada uma destas.

7.2 Primeira Fase de Avaliação

Primeira Fase de Avaliação – Recolha e exploração de dados.

- Recolha e avaliação da qualidade dos dados
    - Verificações de dados de amostra e dados representativos junto do cliente.
- Análises Exploratórias de Dados (AED)
    - Conhecer os dados, variáveis e seu comportamento.
    - Detetar assunções.
    - Identificar padrões e tendências.
    - Detetar erros e amostras.
    - Descobrir relações e interdependências.

7.2.1 Avaliação da Qualidade dos Dados

Para se avaliar a qualidade de dados, foi necessário recorrer à utilização do ambiente de pré-produção e contaram igualmente com a ajuda de algumas equipas do negócio, nomeadamente:

- E-commerce Development Team
- Business Intelligence Development Team

Foi necessário verificar junto destes que os dados recolhidos correspondiam efetivamente aos dados que estão atualmente a ser disponibilizados ao negócio. Também uma das componentes validadas junto com estes foi garantir que todas as transformações que fossem realizadas sobre os dados iam de encontro aos resultados comparáveis.

Já anteriormente foram indicados alguns dos problemas detetados nas diversas fontes de dados, que em diversas fases colocaram em causa alguns dos desenvolvimentos. É importante referir que a plataforma esta diretamente ligada a estas fontes de dados, e que caso surjam alterações ao modelo de dados, este poderá colocar a plataforma com incongruências, apresentando dados inválidos, e ao limite de indicar erros ao utilizador ou indisponibilizar por completo a plataforma.

7.3 Segunda Fase de Avaliação

Segunda Fase de Avaliação – Modelos/Algoritmos Desenvolvidos

- Modelos de Recomendação
    - Medidas especificas para validar os modelos de recomendação
        - MAE (Mean Average Error)
        - RMSE (Root Mean Square Error)

7.4 Terceira Fase de Avaliação

Terceira Fase de Avaliação – Plataforma "Parfois Web Client Analytics"

- Qualidade do código
    - Avaliação de através de ferramentas específicas.
- Tempos de desenvolvimento
    - Avaliação de através de ferramentas específicas.
- Documentação existente
    - Avaliação junto de coordenador do projeto e key users.
- Definição e realização de Testes
    - Testes unitários
    - Testes funcionais
    - Testes automatização
    - Testes segurança

7.4.1 Qualidade do Código

Como qualquer outra linguagem, a linguagem R com a sua evolução começou a sentir necessidade de utilizar boas praticas, embora muitos dos programadores já as tenham implícitas de outras linguagens, o *Google's R Style Guide* veio assumir-se como um guia para os programadores de R, tornando assim o código mais fácil de ler e validar entre toda a comunidade. Alguns autores como Hadley Wickham em "*Advanced R Book*", Graham Williams no livro "*Sharing R Code - With Style*" e também Laurent Gatto no tutorial "*Writing better R code"*, defendem boas praticas para a programação em R [56].

Para a usabilidade dessas boas praticas surgiram ferramentas para apoiar os programadores, surgiu o "*Tidyverse style guide*" [57] derivado do anteriormente apresentado "*Google's R style guide*", mas atualizado e evoluído para os dias de hoje. Igualmente Yihui Xie desenvolveu em 2017 um pacote de R denominado formatR, que permite formatar o código de R automaticamente, tendo também em base essas boas práticas de desenvolvimento de código em R [58].

No entanto em R existem dois pacotes/librarias dentro da conhecida coleção de pacotes Tidyverse, que permitem assim garantir uma maior qualidade do código realizado, indo de encontro às boas condutas anteriormente discutidas. Neste desenvolvimento utiliza-se esses

mesmos pacotes/librarias denominados LINTR e STYLER que será apresentado numa visão prática do projeto desenvolvido.

7.4.1.1 LINTR

O pacote/libraria LINTR executa uma validação ao código desenvolvido, avaliando se o mesmo vai de encontro ao estilo de código definido, verificando a sintaxe de erros e questões de semântica, e alertando o utilizador das irregularidades [59].

Neste caso pratico utiliza-se a função LINT do pacote LINTR sobre diversos ficheiros de código desenvolvido, conforme visível na Figura 132.

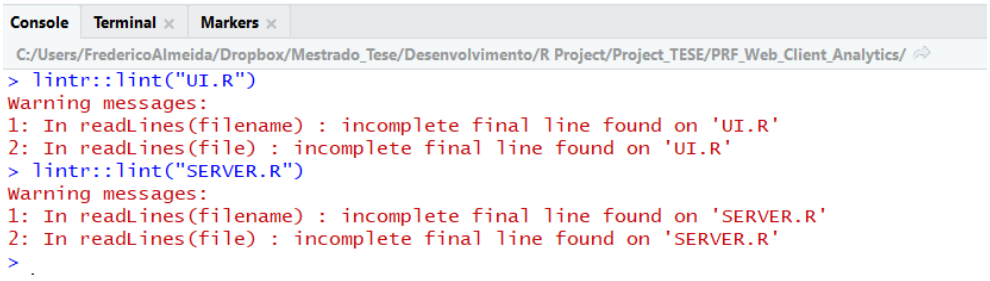


Figura 132 - Pacote LINTR

Este processo efetua assim a validação do código existente nos ficheiros indicados, retornando ao utilizador as respetivas irregularidades e gravidade, tal como demonstrado na Figura 133.

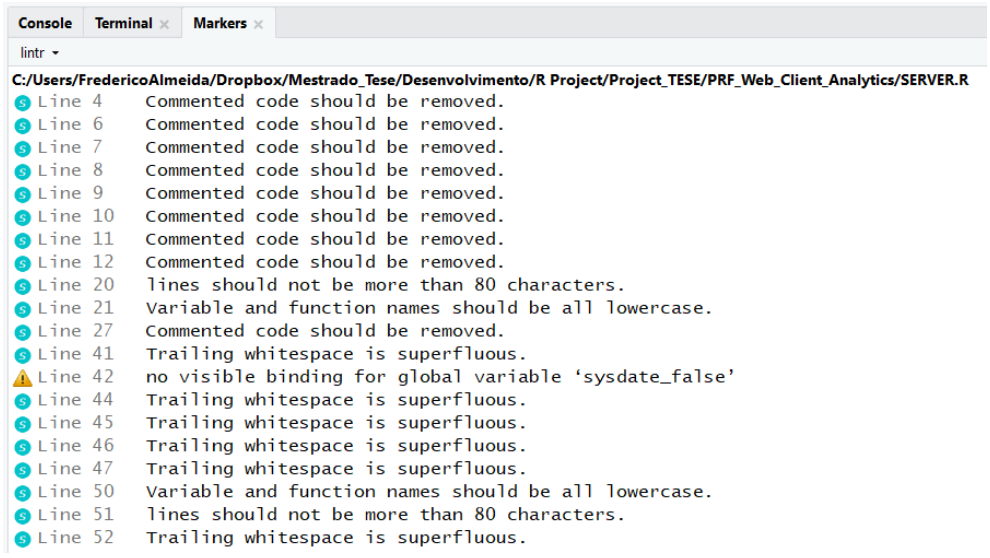


Figura 133 - Resultados LINTR

7.4.1.2 STYLER

O pacote/libraria STYLER incorporado no Rstudio como Addin, efetua de forma automática a formatação do código de acordo com as regras do "*Tidyverse style guide*" permitindo alguma customização pelo programador [60].

A grande vantagem do STYLER é uniformizar todo o código para permitir sempre a mesma sintaxe entre todo o desenvolvimento, e corrigindo o mesmo caso necessário, como visível na Figura 134 e Figura 135.

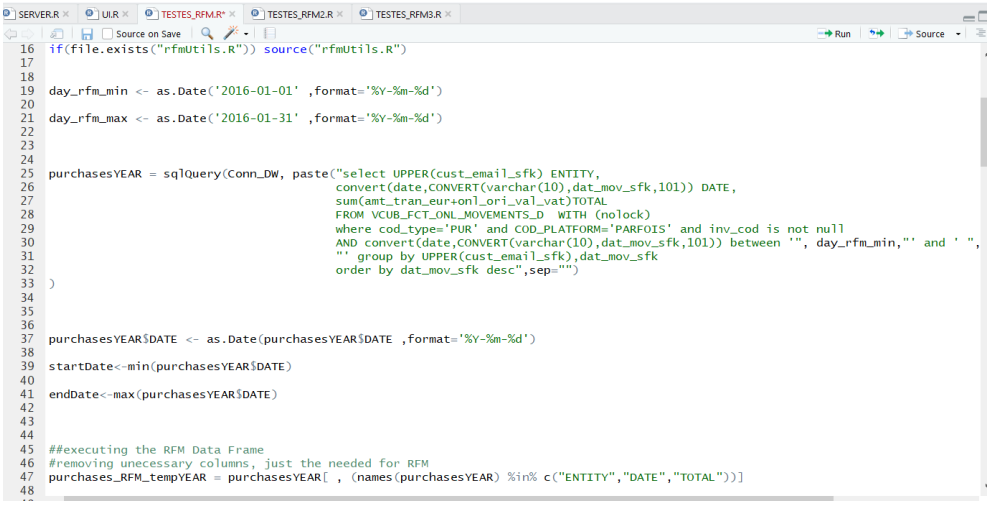


Figura 134 - Antes do STYLER

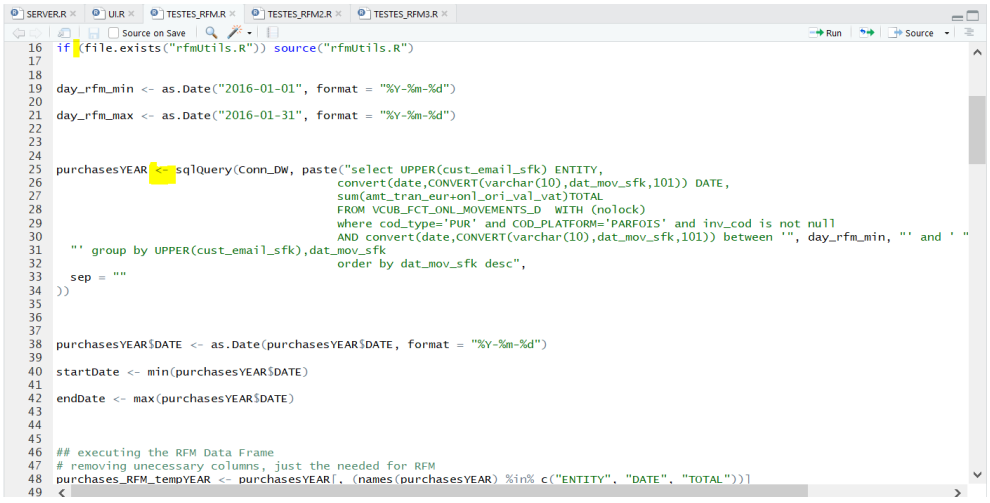


Figura 135 - Após o STYLER

7.4.2 Tempos de Desenvolvimento

Ao longo de todo o processo de desenvolvimento foi utilizado a ferramenta GIT, permitindo assim controlar o código desenvolvido e todas as alterações realizadas, esta ferramenta igualmente poderia permitir uma melhor gestão de desenvolvimento do código em equipa.

Assim que efetuada a primeira versão estável, foi realizado o primeiro "*commit*" denominado a V1 na data de 2018-04-14 conforme a Figura 136 representa.

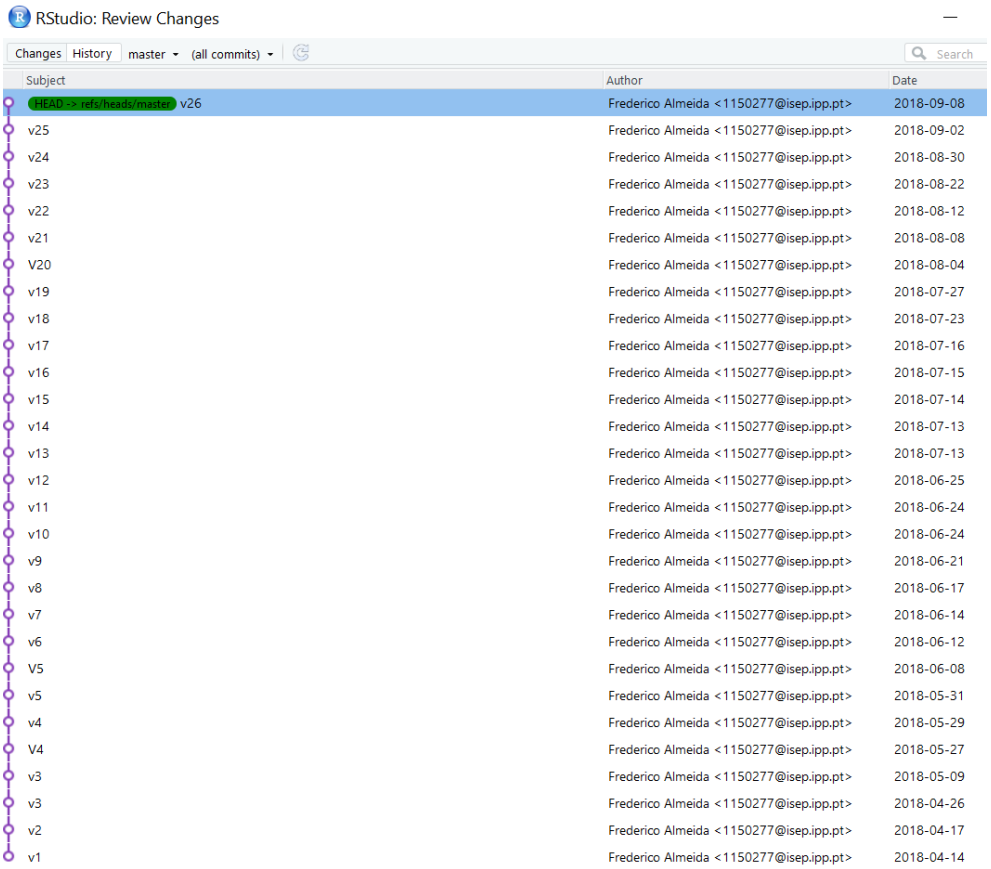


Figura 136 - Histórico Desenvolvimento

Na data de 2018-09-08 foi definida a última versão denominada V26 e sobre a qual é apresentado este projeto.

7.4.3 Testes ao Desenvolvimento

Cada vez mais o software é complexo e interligado com diversas plataformas e dispositivos, com isso a importância de testes ganha cada vez mais relevo. Utilizar a correta metodologia de testes é um dos pontos fundamentais. Garantir o cumprimento dos requisitos e obter o sucesso de projeto somente é possível com uma boa definição e execução de testes.

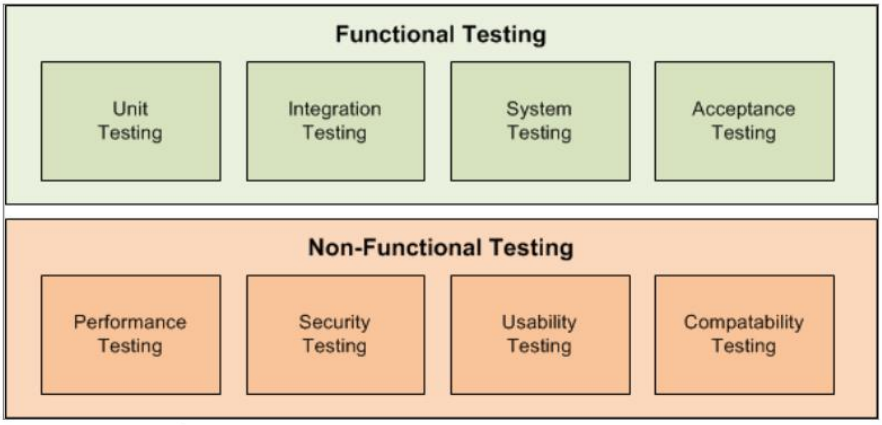


Figura 137 - Metodologia Testes [61]

Embora seja desejável a realização de testes funcionais e de testes não funcionais conforme a Figura 137, não foi de todo possível a realização de testes não funcionais, sendo que ao longo das várias etapas de desenvolvimento foram realizados somente testes funcionais que iremos descrever de seguida:

- Testes unitários
- Testes de integração
- Testes de sistema/aplicativo
- Testes de aceitação

Para um melhor desenvolvimento de todo o projeto foram utilizados distintos ambientes, nas diversas fases de desenvolvimento. Considerando os seguintes:

- Ambiente de Testes
    - Utilizado ao longo do processo de desenvolvimento e na realização de testes unitários e dos testes de integração.
- Ambiente Pré-Produção

- Utilizado para realizar os testes de sistema/aplicação, estes com dados próximos aos reais.
- Utilizado para a realização dos testes de aceitação com o negócio.

- Ambiente de Produção
  - Implementação da plataforma desenvolvida.

7.4.3.1 Testes Unitários

Os testes unitários são definidos como testes individuais a componentes ou módulos que compõem a plataforma. estes testes são normalmente realizados antes de os componentes/módulos serem integrados na plataforma.

Um dos exemplos de testes unitários realizados, foram a validação das conexões com sucesso a cada uma das fontes de dados, conforme a Figura 138.

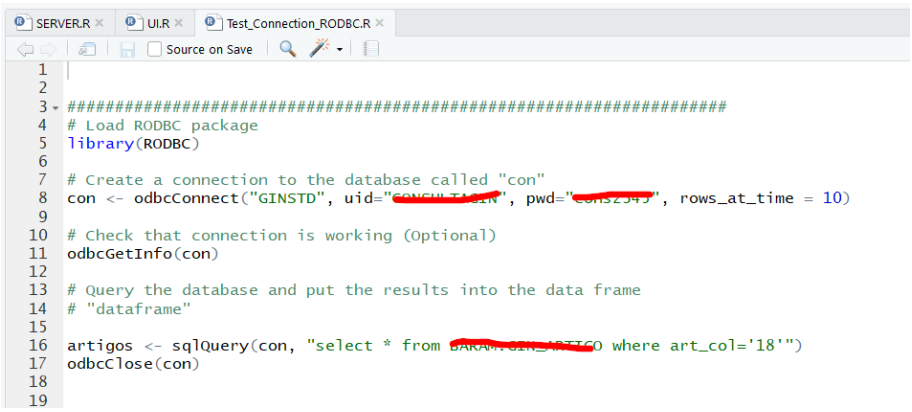


Figura 138 - Teste Conexão

7.4.3.2 Testes de Integração

A realização de testes de integração é o teste dos diferentes módulos / componentes que foram testados com sucesso (nos testes unitários) mas agora quando integrados para executar tarefas e atividades específicas. Esse teste geralmente é feito com uma combinação de testes unitários automatizados e testes manuais, dependendo de como os mesmos conseguem ser integrados para obter resultados específicos.

A realização de testes de integração é bastante extensa uma vez que é necessário validar corretamente o resultado de cada um dos diferentes componentes.

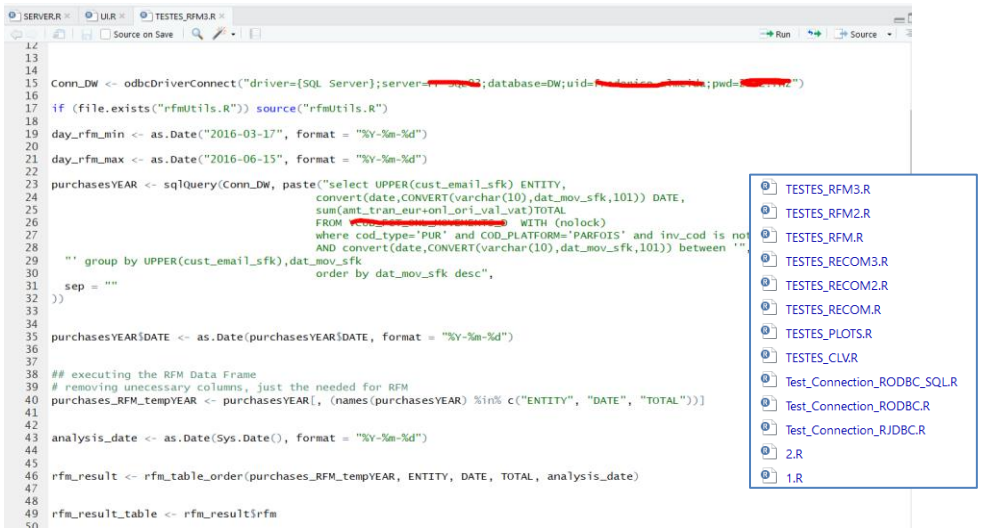


Figura 139 - Teste de Integração

Embora se apresente na Figura 139, o exemplo de testes para a componente do RFM, também é visível que igualmente foi necessário testar o sucesso dos diferentes componentes.

7.4.3.3 Testes de Sistema/Aplicação

Os testes do sistema/aplicação envolvem testar todo a aplicação em busca de bugs e erros. Este teste é realizado através da interface dos componentes de hardware e software de todo o sistema (que foram testados anteriormente e testados quanto à integração), e em seguida, testados como um todo. Estes testes também conhecidos como teste de caixa preta, onde a aplicação é avaliada num ambiente distinto, e o mais próximo das condições de trabalho esperadas pelo utilizador, podendo ser afetada pelos diversos problemas existentes no ambiente e não diretamente relacionados com a aplicação.

Estes testes foram realizados no ambiente de pré-produção e contaram igualmente com a ajuda de algumas equipas do negócio, nomeadamente:

- E-commerce Development Team
- Quality Assurance Team
- Business Intelligence Development Team

7.4.3.4 Testes de Aceitação

A realização dos testes de aceitação é a fase final do teste de software funcional e envolve garantir que todos os requisitos de aplicação / projeto foram atingidos e que os utilizadores finais e clientes testaram o sistema para garantir que ele funcione conforme esperado, respondendo a todos os requisitos definidos.

Estes testes foram concretizados no ambiente de pré-produção, tendo sido realizadas duas sessões e com o apoio de quatro elementos do negócio, nomeadamente:

- E-commerce Development Team
- E-commerce Analytics Manager
- E-commerce Operations Controller

Destas mesmas sessões para além da validação dos principais objetivos propostos, surgiram algumas indicações importantes para melhorias futuras. Igualmente estes elementos foram convidados a realizaram o preenchimento do inquérito acerca da plataforma desenvolvida.

7.5 Quarta Fase de Avaliação

Quarta Fase de Avaliação – Negócio

- Inquérito de satisfação ao cliente/utilizadores.

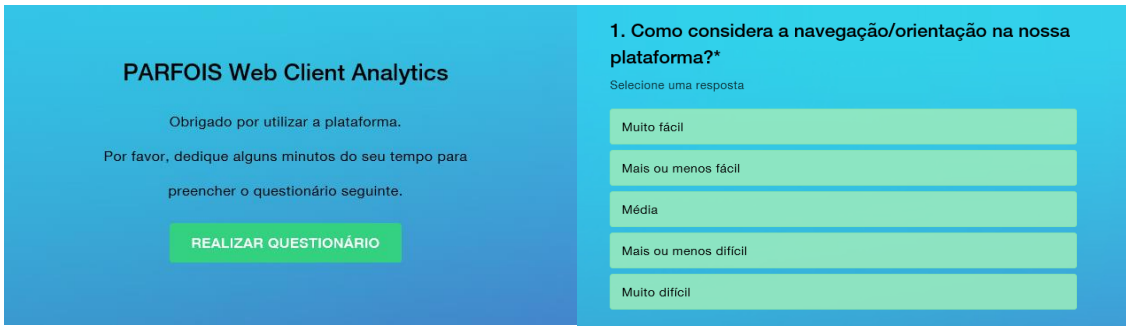
Futuramente e para melhor avaliação da plataforma desenvolvida, poderão ser realizadas duas análises:

- ROI (return of investment) – O negócio poderá avaliar o retorno de investimento através de comparação dos principais objetivos definidos para o E-Commerce, tais como:
    - Volume de vendas
    - Aquisição/retenção de clientes
    - Redução devoluções
    - Adesão a recomendações

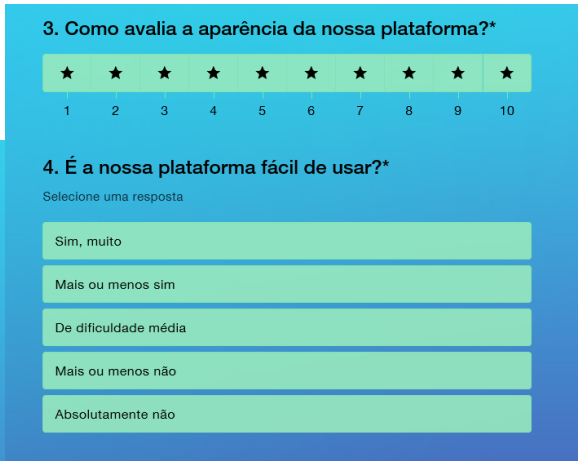
- Comparativo de Resultados, podem ser comparados o volume de dados e resultados em diferentes períodos (Like For Like), assim como a perceção dos gestores sobre os problemas ou benefícios adquiridos.

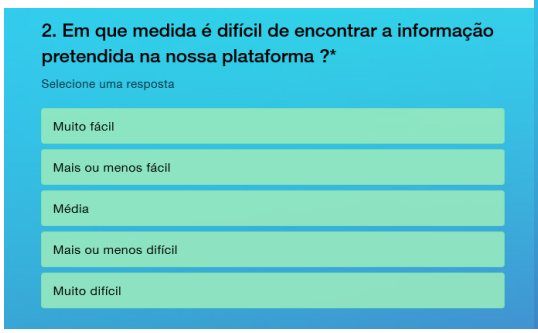
7.5.1 Inquérito de Satisfação

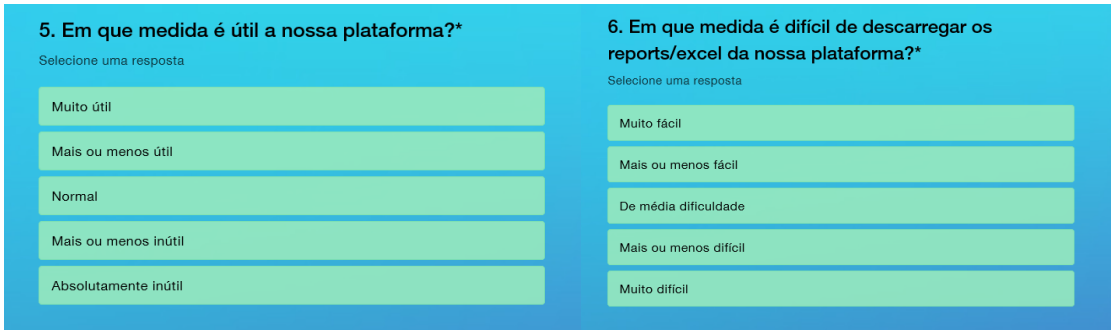
Era de extrema importância avaliar a nossa plataforma junto dos utilizadores da plataforma desenvolvida, para realizarmos essa mesma avaliação realizamos um inquérito através da plataforma SURVIO, e disponibilizamos o mesmo diretamente na plataforma.


PARFOIS Web Client Analytics
Obrigado por utilizar a plataforma.
Por favor, dedique alguns minutos do seu tempo para
preencher o questionário seguinte.
REALIZAR QUESTIONÁRIO

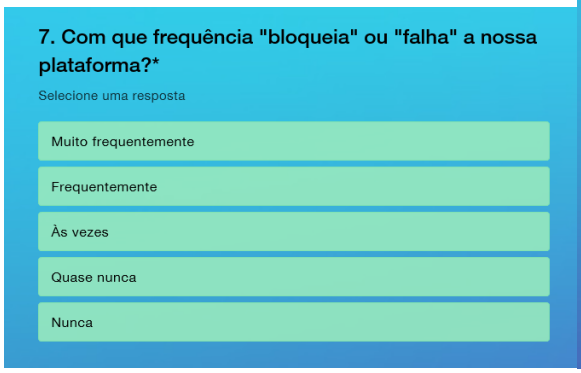
1. Como considera a navegação/orientação na nossa plataforma?*
Selecione uma resposta
Muito fácil
Mais ou menos fácil
Média
Mais ou menos difícil
Muito difícil

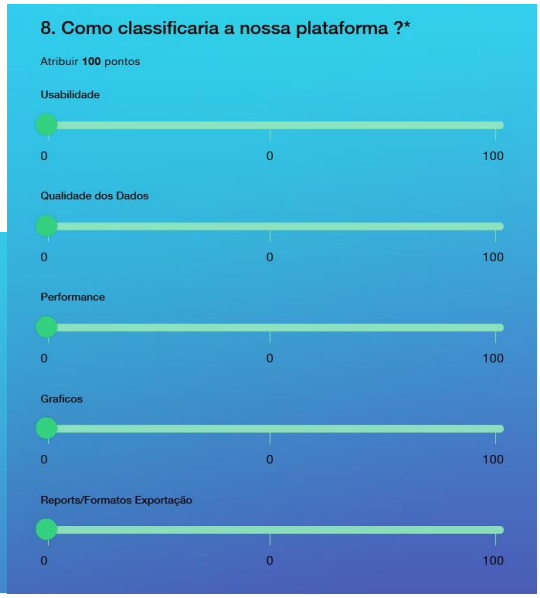

3. Como avalia a aparência da nossa plataforma?*
1 2 3 4 5 6 7 8 9 10
4. É a nossa plataforma fácil de usar?*
Selecione uma resposta
Sim, muito
Mais ou menos sim
De dificuldade média
Mais ou menos não
Absolutamente não


2. Em que medida é difícil de encontrar a informação pretendida na nossa plataforma ?*
Selecione uma resposta
Muito fácil
Mais ou menos fácil
Média
Mais ou menos difícil
Muito difícil


5. Em que medida é útil a nossa plataforma?*
Selecione uma resposta
Muito útil
Mais ou menos útil
Normal
Mais ou menos inútil
Absolutamente inútil

6. Em que medida é difícil de descarregar os reports/excel da nossa plataforma?*
Selecione uma resposta
Muito fácil
Mais ou menos fácil
De média dificuldade
Mais ou menos difícil
Muito difícil


7. Com que frequência "bloqueia" ou "falha" a nossa plataforma?*
Selecione uma resposta
Muito frequentemente
Frequentemente
Às vezes
Quase nunca
Nunca


8. Como classificaria a nossa plataforma ?*
Atribuir 100 pontos
Usabilidade
0 0 100
Qualidade dos Dados
0 0 100
Performance
0 0 100
Graficos
0 0 100
Reports/Formatos Exportação
0 0 100

9. Em que medida tem confiança nos dados da nossa plataforma?*

Selecione uma resposta

- Tenho confiança absoluta
- Tenho muita confiança
- Tenho confiança normal
- Não o acho digno de confiança
- Não tenho confiança nenhuma

10. Recomendaria a plataforma a outros elementos da equipa ?*

Selecione uma resposta

- Definitivamente sim
- Provavelmente sim
- Não sei
- Provavelmente não
- Definitivamente não

11. Como está satisfeito/a com o funcionamento da nossa plataforma?*

Selecione uma resposta

- Muito satisfeito/a
- Satisfeit/a
- Médio satisfeito/a
- Insatisfeito/a
- Muito insatisfeito/a

12. Apostaria na continuidade da plataforma e seu crescimento ?*

Selecione uma resposta

- Definitivamente sim
- Provavelmente sim
- Não sei
- Provavelmente não
- Definitivamente não

13. No global como classificaria a nossa plataforma?*

★ ★ ★ ★ ★ ★ ★ ★ ★ ★

1 2 3 4 5 6 7 8 9 10

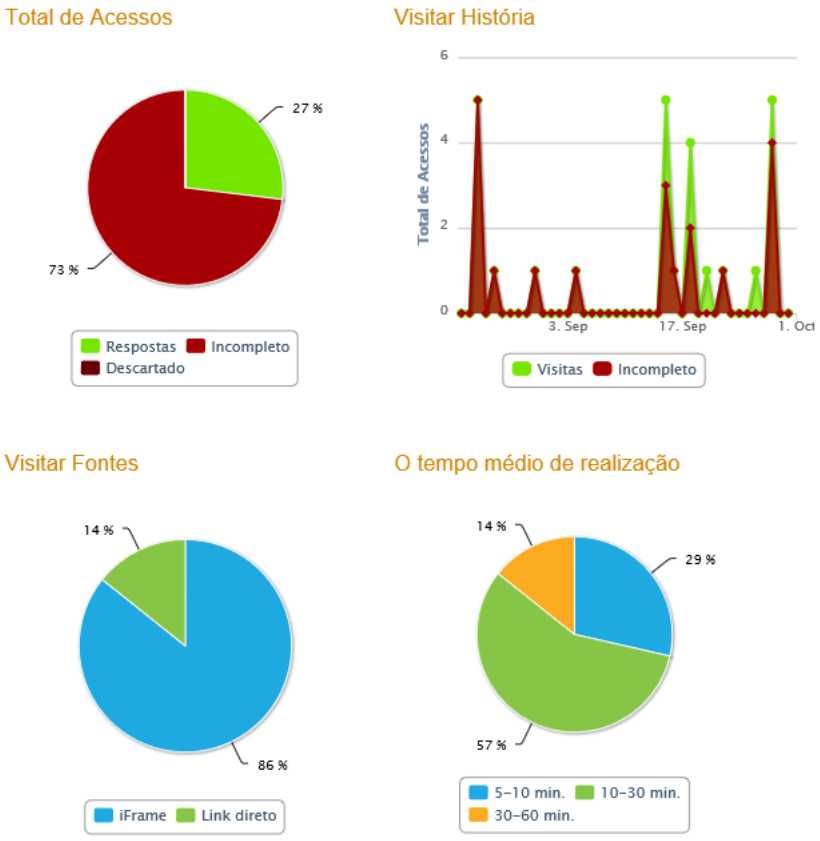
14. Como podemos melhorar a nossa plataforma ?*

Escreva uma ou algumas palavras...

500

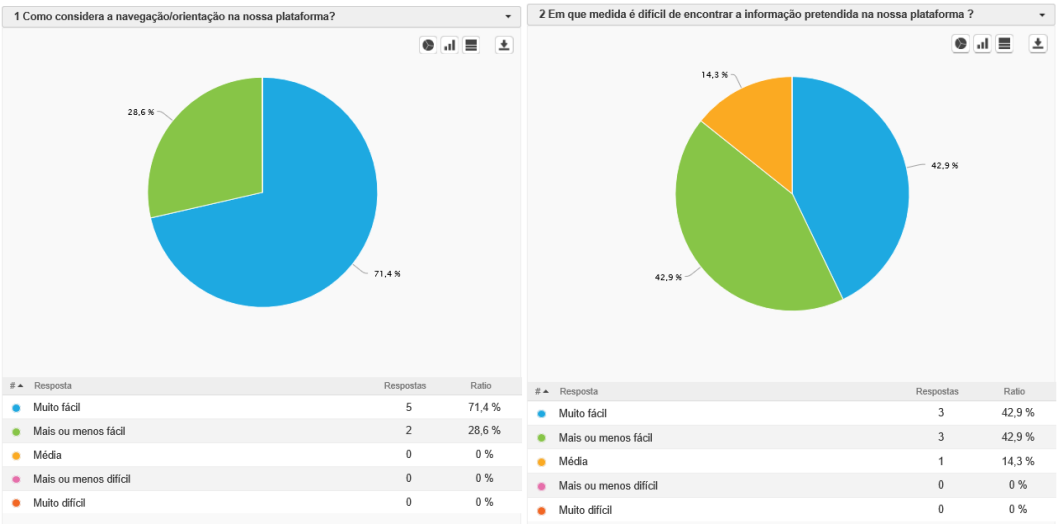
7.5.2 Resultados do Inquérito de Satisfação

Embora não tenha sido atingido o objetivo proposto de quinze inquéritos, o que também não significaria a utilização da plataforma na plenitude por este mesmo número de utilizadores, foram registados sete inquéritos sendo com estes possíveis tirar algumas lições para o futuro.

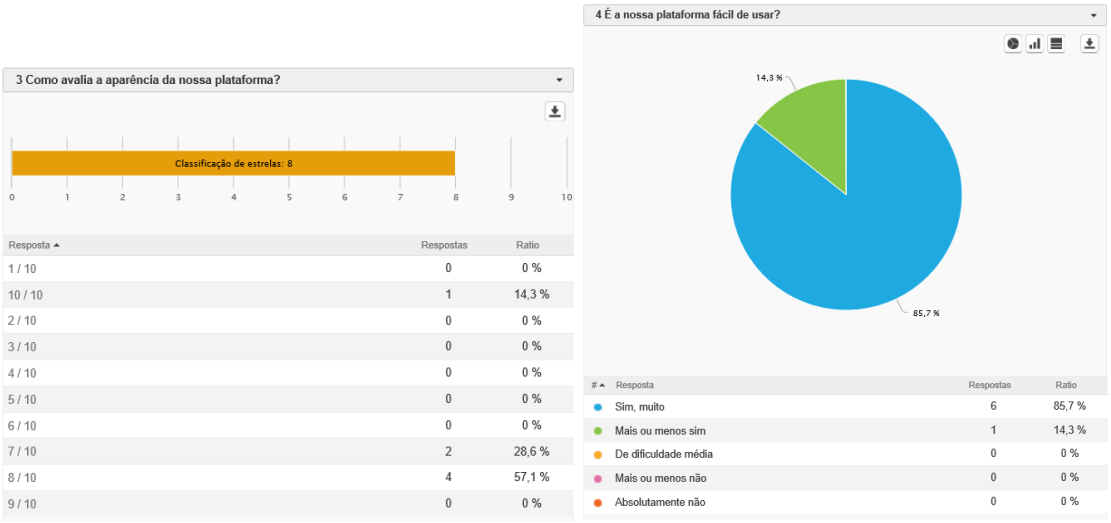


Alguns dos dados estatísticos obtidos apontam que não foi possível a todos os utilizadores realizarem corretamente o inquérito, apoiando o baixo número de inquéritos para avaliação.

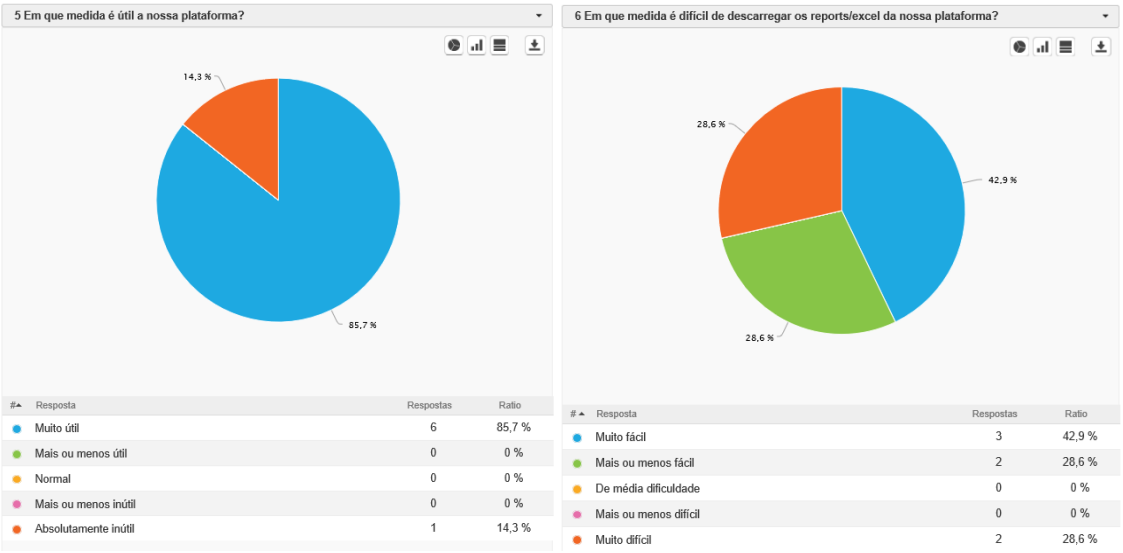
Relativamente aos resultados dos inquéritos realizados, vamos avaliar cada um destes.



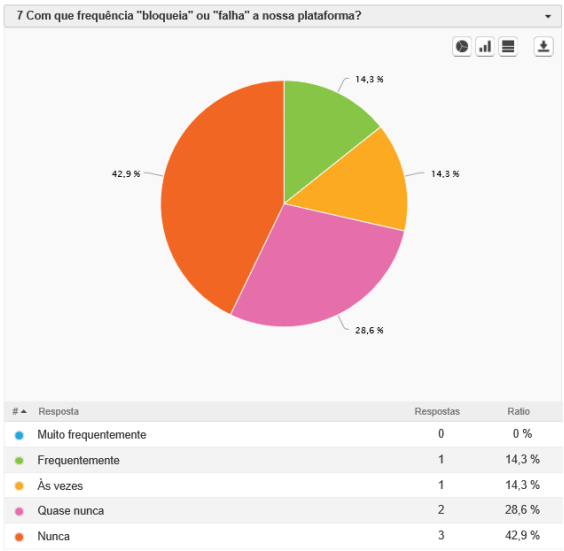
Para a questão 1 é de referir que é praticamente consenso entre todos que a plataforma é de fácil navegação, no entanto para a questão 2 já existe uma maior diferença nas respostas, onde embora seja um saldo positivo, nem todos sentem a mesma facilidade em obter os dados.



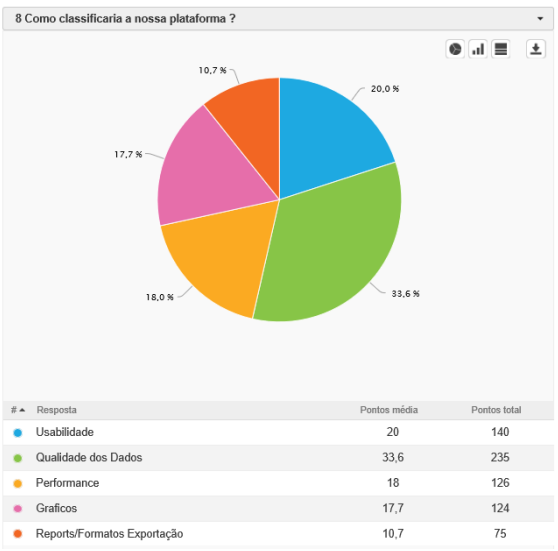
Relativamente à questão 3 é claramente percetível que a aparência da plataforma é positiva, também mediante as respostas da questão 4 poderemos obter que a mesma também apresenta uma fácil usabilidade.



Com as respostas à questão 5 garante-se que plataforma é bastante útil para o negócio, já por outro lado com os resultados da questão 6, percebe-se que é importante trabalhar no tema da exportação de relatórios e excel, uma vez que não são muito positivos estes valores.

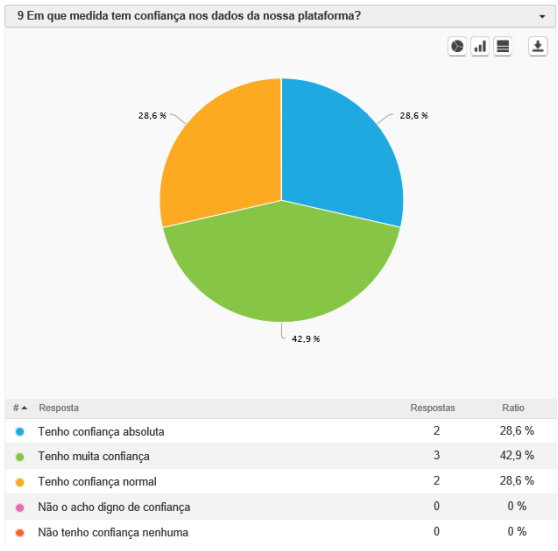


#	Resposta	Respostas	Ratio
1	Muito frequentemente	0	0 %
2	Frequentemente	1	14,3 %
3	Às vezes	1	14,3 %
4	Quase nunca	2	28,6 %
5	Nunca	3	42,9 %

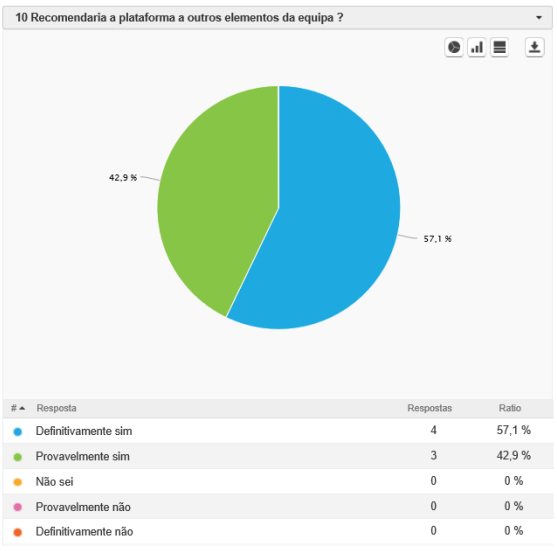


#	Resposta	Pontos média	Pontos total
1	Usabilidade	20	140
2	Qualidade dos Dados	33,6	235
3	Performance	18	126
4	Graficos	17,7	124
5	Reports/Formatos Exportação	10,7	75

As respostas obtidas para a questão 7 foram algo que deve ser considerado alarmante, uma vez que a performance da plataforma pode colocar em causa o uso desta pelos utilizadores. Na questão 8 consegue-se concluir que a usabilidade e a qualidade dos dados são os mais positivos para os utilizadores, mas por lado oposto, a performance e relatórios são os pontos mais negativos.

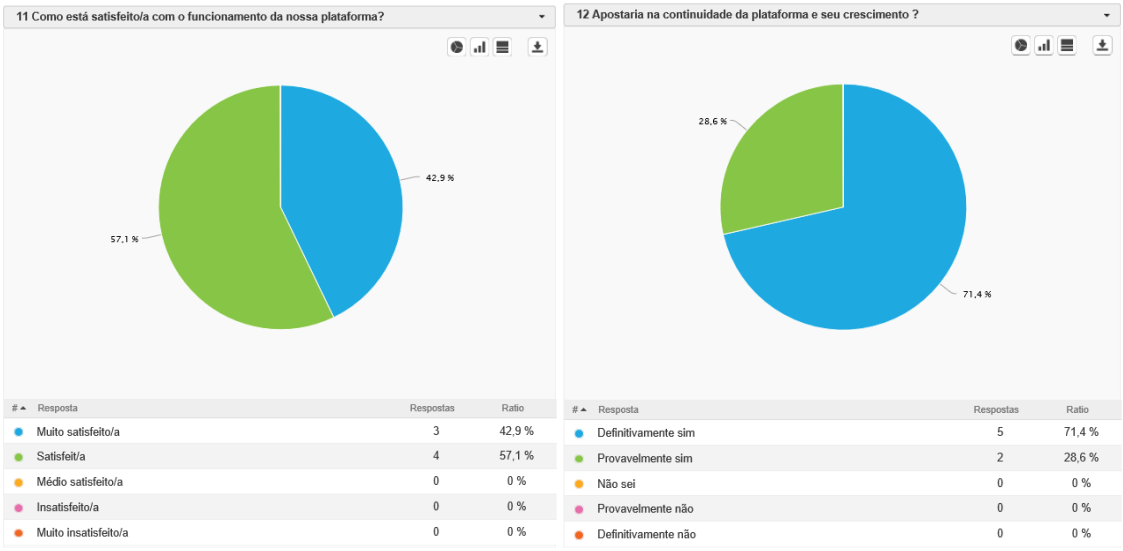


#	Resposta	Respostas	Ratio
1	Tenho confiança absoluta	2	28,6 %
2	Tenho muita confiança	3	42,9 %
3	Tenho confiança normal	2	28,6 %
4	Não o acho digno de confiança	0	0 %
5	Não tenho confiança nenhuma	0	0 %

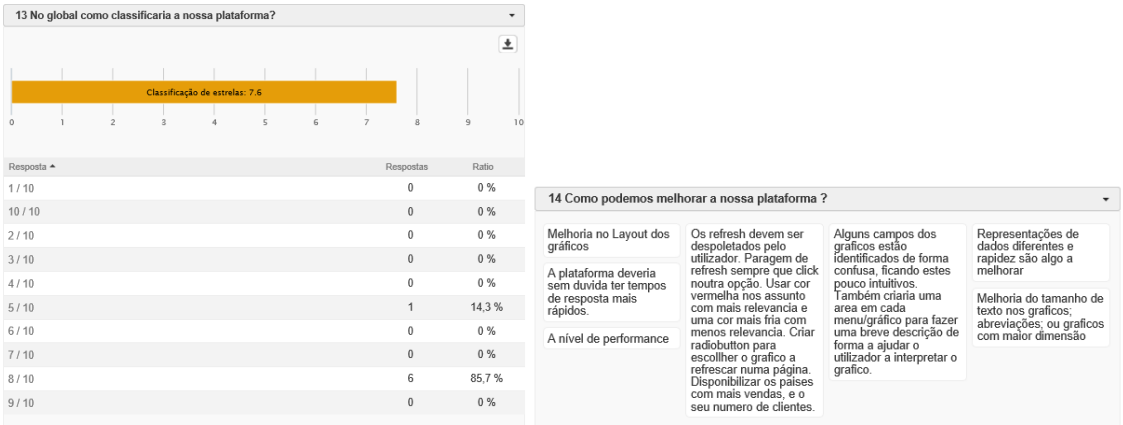


#	Resposta	Respostas	Ratio
1	Definitivamente sim	4	57,1 %
2	Provavelmente sim	3	42,9 %
3	Não sei	0	0 %
4	Provavelmente não	0	0 %
5	Definitivamente não	0	0 %

Com os resultados da questão 9 é claramente verificado que a confiança dos dados da plataforma é positiva, e também mediante as respostas da questão 10 poderemos obter que a mesma poderá ser também utilizada por outros elementos.



Com resultados positivos para a questão 11 e para a questão 12, obtém-se um feedback positivo que permite identificar não só uma satisfação nos utilizadores, mas também que estes apoiam o seu crescimento.



Em conclusão verificamos que os utilizadores que responderam a estes inquéritos têm em comum dois pontos que destacamos como duas grandes prioridades futuras deste projeto, nomeadamente:

- Tempo de Resposta dos Dados/Performance.
    - Para ultrapassar este problema, deverão ser criadas algumas condições a nível da base de dados, tornando esta funcional na agregação de dados, principalmente dados históricos, nomeadamente a criação de vistas agregadas ao ano. De igual modo com a implementação de plataforma num servidor dedicado, são esperadas algumas vantagens.

- Extração de Dados e Relatórios.

- A ferramenta utilizada para a exportação de Excel, não é a mais adequada, sendo que teremos de explorar outros pacotes disponíveis para realizar a mesma. No que respeita à exportação de relatórios, os mesmo devem ser uniformizados no que diz respeito à sua apresentação de acordo com o template da empresa, e também de como a informação é organizada dentro do documento gerado.

8 Conclusão

O capítulo “Conclusão” incide sobre uma reflexão de todo o percurso realizado para a conclusão deste projeto. O leitor ficará consciente de que o caminho teve as suas corretas e erradas decisões, mas que se tornaram crescimento e aprendizagem para o futuro.

8.1 Objetivos Concluídos

A Parfois encontra-se atualmente numa fase de crescimento exponencial na sua área de e-commerce, onde cada vez mais o cliente é exigente, e garantir níveis de qualidade e satisfação elevados são mais do que uma necessidade, são uma obrigatoriedade.

Com isso em mente os desafios eram claros, e os objetivos ambiciosos, garantir à empresa uma ferramenta que auxiliasse na análise de dados e na concretização de melhores decisões, potenciando assim o aumento de vendas e a fidelização de clientes.

Numa era onde os Dashboards começam a fazer cada vez mais parte das nossas vidas, e se destacam entre as ferramentas de apoio à tomada de decisão, pretendia-se que a plataforma de igual modo tivesse como ecrã inicial, algo que transmitisse de imediato uma visão ampla sobre este modelo de negócio.

De igual forma a análise de requisitos realizada junto com o negócio na fase inicial, obrigou que fossem respeitados os principais KPI´s do E-Commerce, e que estes estivessem presentes na solução, uma vez que são estes que apoiam na tomada de algumas decisões. O ciclo de vida do processo de compra online foi obrigatoriamente contemplado neste desenvolvimento, sendo abordado os dois principais processos, a encomenda e respetiva faturação, assim como a devolução e respetiva nota de crédito, sendo que estes correspondem a métricas importantes para a avaliação deste modelo de negócio.

O papel do cliente, e a importância que este deve assumir foram os pontos onde se pretendia maior destaque, pois era aqui que existia a necessidade de descobrir/redescobrir o nosso cliente, conseguindo definir de que forma poderemos avaliar este, algo que com o modelo RFM foi possível de concretizar, permitindo assim a possibilidade de criar com este uma relação win-win, o que com o modelo CLV veio acrescentar ainda mais informação na obtenção, retenção e fidelização de clientes em que este se apoia. Novamente o papel do cliente surge em destaque no processo de recomendação, onde junto com este se pretende obter uma maior taxa de vendas, mas de igual forma das preferências do cliente.

Estas três componentes referenciadas anteriormente, nomeadamente o RFM, o CLV e a Recomendação, vão permitir aos diversos beneficiários da plataforma a definição de novas e mais direcionadas estratégias junto do cliente.

Para além das diversas vantagens já referidas, ainda poderemos referenciar algumas outras fruto deste desenvolvimento, e que já referimos ao longo deste documento, o tempo de preparação e tratamento dos dados é agora inexistente para o utilizador, a fiabilidade é aumentada garantindo o uso das fontes corretas e centralizadas, a usabilidade permite uma maior abrangência de utilizadores, não estando limitado à sua experiência e capacidade de tratamento e análise.

Este ponto anterior, de certa forma refere um dos requisitos neste projeto, que foi a possibilidade de extração de informação. A definição da ferramenta Excel como utilizada em larga escala pelos analistas da empresa, foi o nosso objetivo e respeitado na análise de RFM e análise de CLV, facultando os diversos dados obtidos para extração pelo utilizador.

No entanto convém realçar que não somente o E-Commerce é uma equipa que adquire vantagem com este projeto, mas que igualmente diversas equipas como o Marketing, Controlo de Gestão e também o cliente final são elementos da nossa rede de valor beneficiados com esta nova ferramenta.

Da necessidade de tomada de decisão mais facilitada para altos quadros organizacionais advém o objetivo de possibilitar a plataforma com reports de fácil legibilidade, e somente com a informação estritamente necessária para decisão, de igual forma os mesmos deveriam ser seguros e não editáveis, exportados assim da plataforma no formato PDF para posterior distribuição pelos demais destinatários.

Numa perspetiva mais técnica do projeto, poderemos assumir que a elevada capacidade de computação de uma ferramenta como o R permitiu que o tratamento de dados fosse muito mais célere, flexível e fiável, não somente para os principais utilizadores, mas para os responsáveis por garantirem o suporte e continuidade do produto, onde é importante referir que de igual modo o uso das normas e regras de programação em R, junto com as ferramentas de validação inseridas no R, permitem uma maior qualidade do código, mas também maior facilidade de adaptação para qualquer programador que no futuro pretenda analisar ou dar continuidade ao projeto.

A usabilidade de pacotes, funções e librarias do R, assim como os diversos fóruns existentes abordando todo o crescimento existente em volta da linguagem R, acrescentaram imenso valor

ao projeto, permitindo usar técnicas mais recentes, e validação dos resultados obtidos. Mas mais importante é a consciência generalizada da importância dos dados, e das imensas possibilidades de tratamento sobre estes, quer clustering, algoritmos, etc, e que cada vez mais com o crescimento exponencial dos dados, abrem portas para um futuro onde não somente os dados, mas o que se obtém destes vai ser o foco.

Deve-se de igual modo destacar a flexibilidade do Shiny, embutido no R, permitiu a criação desta plataforma com uma usabilidade bastante simplificada e de fácil compreensão ao utilizador, mas também sendo de fácil aprendizagem para programadores para darem continuidade a desenvolvimentos futuros.

Consegue-se também que com o apoio das diversas equipas da DSI, quer na fase de obtenção de dados, quer na fase de testes, tivéssemos taxas de fiabilidade e qualidade dos dados mais elevadas, no entanto para comprovar o funcionamento da plataforma nada melhor do que a equipa e-commerce realizá-lo, sendo um fator fundamental na sua validação.

Embora existissem alternativas para corresponder aos requisitos da empresa, como a aquisição de novos módulos para o website existente, a empresa apostou no desenvolvimento no âmbito desta tese, uma vez que os custos seriam bastante mais reduzidos, o conhecimento ficava retido, o suporte estaria garantido, e as possibilidades de crescimento futuro eram mais flexíveis e imediatas.

O comprometimento das diversas equipas e elementos envolvidos neste projeto, foram indubitavelmente o fator de sucesso.

8.2 Limitações e Trabalho Futuro

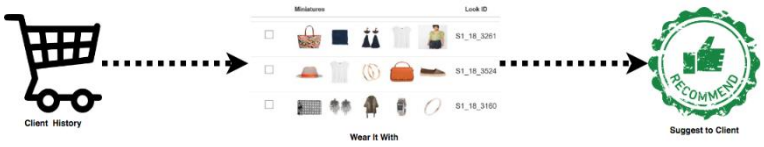
Não é a solução perfeita, não existe uma solução final. E esse foi um dos requisitos deste projeto, porque hoje tudo é feito de mudança e adaptabilidade, e isso era algo pretendido nesta plataforma. A sua capacidade de crescimento, de flexibilidade e de adaptação às diversas sinergias criadas pelo modelo de negócio e-commerce, que cada vez mais evolui e se torna fator de aposta das médias e grandes empresas, uma vez que permite a globalização destas.

Deste capítulo deveremos destacar duas componentes distintas de análise, as limitações do projeto desenvolvido, e o que poderemos perspetivar para o seu futuro.

Iniciando esta análise nas limitações que ao longo do desenvolvimento deste projeto fomos encontrando são destacadas entre as demais, as que consideramos mais críticas, nomeadamente:

- Recomendação utilizando o Wear it With
    - Nas diversas abordagens realizadas com o Wear It With verificamos que a ausência de registos para o período disponibilizado como amostra, não permitia

identificar os artigos para looks existentes, sendo que o status do look também deve ser considerado, estando estes já atualmente inativos.



- Registo de cliente em site/execução de encomenda
    - A não obrigatoriedade de registo de cliente no site, permitindo a possibilidade de cliente "anónimo" invalida a existência de uma maior base de dados de clientes.
- Impossibilidade de testes em ambiente real
    - Com a execução de alguns testes no ambiente de produção, conseguir-se-ia ter métricas de avaliação mais recentes, assim como a realização de testes de performance totalmente reais ao ambiente do dia a dia dos utilizadores.
- Qualidade da informação no Cubo
    - O Cubo atual apresenta algumas limitações no que concerne a dados, e aos tratamentos realizados sobre estes, tais como tratamento de nulos, de igualdade de tipos para colunas idênticas, entre outros.
- Performance da plataforma
    - A inviabilidade de melhores recursos na fase de desenvolvimento da plataforma, impossibilita que algumas das métricas de tempo, sejam em grande parte incomparáveis com o ambiente real utilizado pelos utilizadores.
- Base dados de clientes
    - A inexistência de uma base de dados consolidada, e automaticamente atualizável, inviabiliza algumas análises, nomeadamente no que podem ser clientes duplicados ou semelhantes.
- Performance do Cubo
    - De igual modo ao problema de performance da plataforma, também a performance do Cubo, é impactada com existência de recursos mínimos no definido ambiente de desenvolvimento.
- Recomendação utilizando os Related Items
    - Não possível adicionar a componente de Related Items na plataforma, uma vez que foi verificado que o volume de dados existente não era suficiente para gerar recomendações, apresenta-se este baixo volume sendo um tema bastante recente no negócio, de igual modo para os artigos do período disponibilizado não se verificam equivalências realizadas.

No que diz respeito ao que se considera que possam ser melhorias futuras, e novos desenvolvimentos que acrescentem valor na plataforma desenvolvida destaca-se os seguintes pontos:

- Usabilidade de Variáveis Globais
    - Permitindo assim a reutilização de alguns dados previamente computados.
- Implementação da Plataforma em arquitetura Cliente-Servidor
    - Existe a possibilidade de aquisição de modulo cliente-servidor do R Studio, garantindo assim maior disponibilidade, manutenção e performance do produto desenvolvido.
- Implementação da Plataforma em arquitetura Web/Cloud
    - Fazendo uso das novas tecnologias, existe a possibilidade de publicar para Web/Cloud os projetos desenvolvidos em R, permitindo uma maior mobilidade aos utilizadores e melhores capacidade de performance. Estruturas como SHINYAPPS.IO permitem hoje uma integração direta com o RStudio.
- Segmentação de Clientes por Atributos
    - Existindo diversos atributos associados à ficha de cliente, é possível a realização de diversas segmentações, como por exemplo por faixa etária, zona geográfica, etc.
- Classificação de Clientes
    - Através das suas classificações individuais obtidas com o RFM, é possível igualmente segmentarmos e classificarmos os clientes conforme se verifica na Figura 140 .

Segment	Description	R	F	M
Champions	Bought recently, buy often and spend the most	4 - 5	4 - 5	4 - 5
Loyal Customers	Spend good money. Responsive to promotions	2 - 5	3 - 5	3 - 5
Potential Loyalist	Recent customers, spent good amount, bought more than once	3 - 5	1 - 3	1 - 3
New Customers	Bought more recently, but not often	4 - 5	<= 1	<= 1
Promising	Recent shoppers, but haven't spent much	3 - 4	<= 1	<= 1
Need Attention	Above average recency, frequency & monetary values	2 - 3	2 - 3	2 - 3
About To Sleep	Below average recency, frequency & monetary values	2 - 3	<= 2	<= 2
At Risk	Spent big money, purchased often but long time ago	<= 2	2 - 5	2 - 5
Can't Lose Them	Made big purchases and often, but long time ago	<= 1	4 - 5	4 - 5
Hibernating	Low spenders, low frequency, purchased long time ago	1 - 2	1 - 2	1 - 2
Lost	Lowest recency, frequency & monetary scores	<= 2	<= 2	<= 2

Figura 140 - Classificação de Clientes RFM

- Análise Continua de Clientes
    - Cada vez mais é importante o cliente para as organizações, como tal o mesmo deve ser avaliado continuamente e não isoladamente, como se demonstra na Tabela 10. Perceber as oscilações dos clientes ao longo do tempo vai permitir a realização de melhores campanhas junto deste, e abordar os clientes certos no momento certo.

	Customer Id	RFM
Análise A - 01/01/2015	273	154
	310	412
	966	355
Análise B - 01/01/2016	273	515
	310	112
	966	555

Tabela 10 - Análise Continua Clientes RFM

- Segurança e controlo de acessos integrado com AD (Active Directory)
    - Esta componente poderia permitir acrescentar na plataforma a possibilidade de definição de perfis e utilizadores, garantindo um maior controle sobre quem e a que informação acede.
- Melhoria de Reporting
    - Definição de reports totalmente direcionados a utilizadores, e com layout global e percetível para todos.
- Base de Dados de Resultados
    - Toda a informação gerada pela plataforma poderia ser de igual forma guardada numa base de dados, para permitir outras ferramentas utilizarem estes dados, para posterior transformação e análise.

Pretende-se assim que os pontos acima representados possam ser elementos válidos para a continuidade da plataforma, e a consequente aposta no seu desenvolvimento contínuo dentro da empresa.

Referências

[1] M. Treacy and F. Wiersema, *The discipline of market leaders: choose your customers, narrow your focus, dominate your market,*. 1997.

[2] P. Koen *et al.*, "Providing Clarity and a Common Language To the 'Fuzzy Front End.,'" *Res. Technol. Manag.*, vol. 44, no. 2, pp. 46–55, 2001.

[3] P. A. Koen *et al.*, "Fuzzy Front End : and Techniques," *Ind. Res.*, 1996.

[4] V. Allee, "Value network analysis and value conversion of tangible and intangible assets," *J. Intellect. Cap.*, vol. 9, no. 1, pp. 5–24, 2008.

[5] V. Allee, "Value Network Analysis What we will cover • Importance of network models," pp. 1–21, 2012.

[6] T. L. Saaty, "Decision-making with the AHP: Why is the principal eigenvector necessary," *Eur. J. Oper. Res.*, vol. 145, no. 1, pp. 85–91, 2003.

[7] A. H. U. M. Tchemra, "Tabela De Decisão Adaptativa Na Tomada De Decisão Multicritério," *Tese Diss.*, p. 172, 2009.

[8] E. H. Forman and S. I. Gass, "The Analytic Hierarchy Process—An Exposition," *Oper. Res.*, vol. 49, no. 4, pp. 469–486, 2001.

[9] W. Adams and R. Saaty, "Super Decisions Software Guide," *Super Decis.*, 2003.

[10] T. L. Saaty, "Deriving the AHP 1-9 scale from first principles," *Sixth Int. Symp. Anal. Hierarchy Process*, no. August, p. pg. 245, 2001.

[11] S. Gupta *et al.*, "Modeling customer lifetime value," *J. Serv. Res.*, vol. 9, no. 2, pp. 139–155, 2006.

[12] J. Wei, S. Lin, and H. Wu, "A review of the application of RFM model," *African J. Bus. Manag.*, vol. 4, no. 19, pp. 4199–4206, 2010.

[13] Kamil Bartocha, "RFM Segmentation," 2015. [Online]. Available: https://www.slideshare.net/WhiteRavenPL/rfm-segmentation. [Accessed: 10-Oct-2018].

[14] IBM, “IBM Knowledge Center - RFM Binning.” [Online]. Available: https://www.ibm.com/support/knowledgecenter/en/SSLVMB_24.0.0/spss/rfm/idh_rfm_binning_transactions.html. [Accessed: 10-Oct-2018].

[15] R. C. Blattberg, E. C. Malthouse, and S. A. Neslin, “Customer Lifetime Value: Empirical Generalizations and Some Conceptual Questions,” *J. Interact. Mark.*, vol. 23, no. 2, pp. 157–168, 2009.

[16] P. D. Berger and N. I. Nasr, “Customer lifetime value: Marketing models and applications,” *J. Interact. Mark.*, vol. 12, no. 1, pp. 17–30, 1998.

[17] Jean-Rene Gauthier, “An Introduction to Predictive Customer Lifetime Value Modeling,” 2017. [Online]. Available: https://www.datascience.com/blog/intro-to-predictive-modeling-for-customer-lifetime-value. [Accessed: 10-Oct-2018].

[18] S. Gupta *et al.*, “Modeling Customer Lifetime Value,” *J. Serv. Res.*, vol. 9, no. 2, pp. 139–155, 2006.

[19] Kevin Donnely, “Why Customer Lifetime Value Matters,” 2017. [Online]. Available: https://www.shopify.com/blog/customer-lifetime-value. [Accessed: 10-Oct-2018].

[20] R. C. Blattberg and J. Deighton, “Manage marketing by the customer equity test.,” *Harv. Bus. Rev.*, vol. 74, no. 4, pp. 136–144, 1996.

[21] M. Hahsler, “recommenderlab: A Framework for Developing and Testing Recommendation Algorithms,” *Nov*, pp. 1–37, 2011.

[22] J. S. Breese, D. Heckerman, and C. Kadie, “Empirical analysis of predictive algorithms for collaborative filtering,” *Proc. 14th Annu. Conf. Uncertain. Artif. Intell.*, pp. 43--52, 1998.

[23] Salem Marafi, “Collaborative Filtering with R : Salem Marafi,” 2014. [Online]. Available: http://www.salemmarafi.com/code/collaborative-filtering-r/. [Accessed: 28-Sep-2018].

[24] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, “GroupLens : An Open Architecture for Collaborative Filtering of Netnews,” *Proc. 1994 ACM Conf. Comput. Support. Coop. Work*, pp. 175–186, 1994.

[25] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, “Item-based collaborative filtering recommendation algorithms,” in *Proceedings of the tenth international conference on World Wide Web - WWW ’01*, 2001, pp. 285–295.

[26] R. Kimball, “The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses,” *Architecture*, p. 771, 1998.

[27] W. H. Inmon, “The data warehouse and data mining,” *Commun. ACM*, vol. 39, no. 11, pp. 49–50, 1996.

[28] H. Jiawei and M. Kamber, *Data mining: concepts and techniques*, vol. 5. 2001.

[29] F. Rodrigues, “Clustering.” Apontamentos DESCO, unidade curricular Mestrado Engenharia Informática DEI/ISEP, Porto, p. 84, 2015.

[30] A. K. Jain, “Data clustering: 50 years beyond K-means,” *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.

[31] T. Calinski and J. Harabasz, “A dendrite method for cluster analysis,” *Commun. Stat. -*

Theory Methods, vol. 3, no. 1, pp. 1–27, 1974.

[32] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math.*, vol. 20, no. C, pp. 53–65, 1987.

[33] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A K-Means Clustering Algorithm,” *Appl. Stat.*, vol. 28, no. 1, p. 100, 1979.

[34] J. A. Hartigan, “Clustering Algorithms,” *Inf. Retr. Data Struct. Algorithms*, vol. 2, pp. 419–442, 1975.

[35] Z. Huang, “Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values,” *Data Min. Knowl. Discov.*, vol. 2, no. 3, pp. 283–304, 1998.

[36] A. Likas, N. Vlassis, and J. J. Verbeek, “The global k-means clustering algorithm,” *Pattern Recognit.*, vol. 36, no. 2, pp. 451–461, 2003.

[37] parry prabhu, “K mean-clustering algorithm,” 2015. [Online]. Available: https://www.slideshare.net/parryprabhu/k-meanclustering-algorithm. [Accessed: 28-Sep-2018].

[38] Weston Pace, “Wikipedia K Means Example,” 2007. [Online]. Available: https://en.wikipedia.org/wiki/K-means_clustering. [Accessed: 28-Sep-2018].

[39] M. Avcilar, M. Y. Avcilar, and E. Yakut, “Association Rules in Data Mining: An Application on a Clothing and Accessory Specialty Store,” *Can. Soc. Sci.*, vol. 10, no. 3, pp. 75–83, 2014.

[40] C. Mihaescu, “Mining Frequent Itemsets – Apriori Algorithm,” *Lab. Modul. 8*, 2012.

[41] A. Bhandari, A. Gupta, and D. Das, “Improvised apriori algorithm using frequent pattern tree for real time applications in data mining,” in *Procedia Computer Science*, 2015, vol. 46, pp. 644–651.

[42] M. Hegland, “The Apriori Algorithm – a Tutorial,” *Inst. Math. Sci. Prepr. Ser.*, 2005.

[43] M. Mittal, S. Pareek, and R. Agarwal, “Efficient Ordering Policy for Imperfect Quality Items Using Association Rule Mining,” in *Encyclopedia of Information Science and Technology, Third Edition*, IGI Global, pp. 773–786.

[44] D. Birant, “Data Mining Using RFM Analysis,” *Knowledge-Oriented Appl. Data Min.*, no. iii, pp. 91–108, 2011.

[45] M. Khajvand, K. Zolfaghar, S. Ashoori, and S. Alizadeh, “Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study,” *Procedia Comput. Sci.*, vol. 3, pp. 57–63, 2011.

[46] T. Evgeniou, “Cluster Analysis and Segmentation.” [Online]. Available: http://inseaddataanalytics.github.io/INSEADAnalytics/CourseSessions/Sessions45/ClusterAnalysisReading.html. [Accessed: 13-Oct-2018].

[47] R. Ait, A. Amine, B. Bouikhalene, and R. Lbibb, “Customer Segmentation Model in E-commerce Using Clustering Techniques and LRFM Model : The Case of Online Stores in Morocco,” vol. 9, no. 8, pp. 1976–1986, 2015.

[48] I. Outlier AI, “Clustering: K-means Clustering, in Practice.” [Online]. Available:

https://outlier.ai/data-driven-daily/clustering-k-means-clustering-in-practice/. [Accessed: 14-Oct-2018].

[49] D. Jobber and F. Ellis- Chadwick, *Principles and practice of marketing*. McGraw-Hill Higher Education, 2013.

[50] Anish Nair, “RFM Analysis For Successful Customer Segmentation - Putler.” [Online]. Available: https://www.putler.com/rfm-analysis/. [Accessed: 10-Oct-2018].

[51] Pushpa Makhija, “RFM Analysis for Customer Segmentation | CleverTap,” 2018. [Online]. Available: https://clevertap.com/blog/rfm-analysis/. [Accessed: 10-Oct-2018].

[52] Bigdata Doc, “Recommender Systems 101 - a step by step practical example in R,” 2014. [Online]. Available: http://bigdata-doctor.com/recommender-systems-101-practical-example-in-r/. [Accessed: 10-Oct-2018].

[53] JJ, “MAE and RMSE — Which Metric is Better? – Human in a Machine World – Medium,” *2016*. [Online]. Available: https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d. [Accessed: 10-Oct-2018].

[54] Hoa K. Quach, “RPubs - Customer Lifetime Value (CLV),” 2017. [Online]. Available: https://rpubs.com/hoakevinquach/Customer-Lifetime-Value-CLV. [Accessed: 10-Oct-2018].

[55] Alex McEachern, “The Easy Way to Calculate Customer Lifetime Value,” 2017. [Online]. Available: https://blog.smile.io/easy-way-to-calculate-and-increase-customer-lifetime-value. [Accessed: 06-Oct-2018].

[56] Joseph Rickert, “Writing Good R Code and Writing Well · R Views,” 2016. [Online]. Available: https://rviews.rstudio.com/2016/12/02/writing-good-r-code-and-writing-well/. [Accessed: 10-Oct-2018].

[57] Hadley Wickham, “The tidyverse style guide.” [Online]. Available: http://style.tidyverse.org/. [Accessed: 10-Oct-2018].

[58] Yihui Xie, “formatR - Format R code automatically - Yihui Xie | 谢益辉,” 2017. [Online]. Available: http://yihui.name/formatr/. [Accessed: 10-Oct-2018].

[59] Jim Hester, “LINTR Static Code Analysis for R,” 2014. [Online]. Available: https://github.com/jimhester/lintr. [Accessed: 10-Oct-2018].

[60] M. Kirill and W. Lorenz, “Non-Invasive Pretty Printing of R Code • styler,” 2017. [Online]. Available: http://styler.r-lib.org/. [Accessed: 10-Oct-2018].

[61] Inflectra, “Software Testing Methodologies - Learn The Methods & Tools,” 2018. [Online]. Available: https://www.inflectra.com/ideas/topic/testing-methodologies.aspx. [Accessed: 29-Sep-2018].

Anexos

PARFOIS

ACORDO DE CONFIDENCIALIDADE

Entre:

Barhold, S.A., sociedade anónima com o capital social de € 10.000.000, sede na Rua de Sistelo, 755 – Lugar de Santegãos, 4435 – 429 Rio Tinto, titular do cartão de pessoa coletiva número 506 810 860, registada na Conservatória do registo Comercial de Gondomar sob o mesmo número, de ora em diante apenas designada por «**2ª Outorgante**».

E

Frederico Miguel Lacerda Barrio Ribeiro de Almeida, residente na Rua da Casa Nova, n.º 247, Ilha dos Amores – Costa, 4810 – 087 Guimarães, portador do Cartão de Cidadão número 12574558, Contribuinte Fiscal número 194639134, inscrito na Segurança Social com o número 12019363791, doravante identificado como «**2º Outorgante**».

É celebrado um ACORDO DE CONFIDENCIALIDADE, nos termos e nas condições seguintes:

Primeira

O **2º Outorgante** está vinculado à **1ª Outorgante** por via de um contrato de trabalho celebrado em 27/09/2010.

Segunda

1. O **2º Outorgante** está a frequentar o curso de Mestrado no Instituto Superior de Engenharia do Porto.
2. A Tese de Mestrado a efetuar pelo **2º Outorgante** irá versar sobre a **1ª Outorgante**.

Terceira

1. O **2º Outorgante** compromete-se a manter total confidencialidade sobre as informações que utilizar no âmbito do referido Mestrado. Incluem-se neste dever, quer as informações específicas da **1ª Outorgante,** quer informações dos seus fornecedores ou clientes.
2. A publicação ou divulgação da Tese de Mestrado, estudo, artigo ou trabalho que verse sobre a empresa carece sempre de autorização escrita.

Feito em Rio Tinto, a 15 de dezembro de 2017, em duplicado ficando um exemplar em poder de cada um dos outorgantes.

P´la **1ª OUTORGANTE**

BARATA & RAMILO, S.A.
NIF: 500 590 753
Rua do Sisteio, 755
Lugar de Santegãos
4435-429 Rio Tinto

2º OUTORGANTE

FREDERICO RIBEIRO ALMEIDA