ISCTE Business School
Instituto Universitário de Lisboa

ÁRVORES DE DECISÃO E REDES NEURONAIS: APLICAÇÃO A WEB MINING

MARCOS ANDRÉ PAIS HENRIQUES

Projecto de Mestrado
em Prospecção e Análise de Dados

Orientador(a):
Profª. Doutora Diana Mendes, Prof. Associado, Instituto Universitário de Lisboa, ISCTE Business School, Departamento de Métodos Quantitativos

Co-orientador(a):
Profª. Doutora Anabela Costa, Prof. Auxiliar, Instituto Universitário de Lisboa, ISCTE Business School, Departamento de Métodos Quantitativos

Novembro 2010

ÁRVORES DE DECISÃO E REDES NEURONAIS: APLICAÇÃO A WEB MINING

MARCOS ANDRÉ PAIS HENRIQUES

Projecto de Prospecção e Análise de Dados

Orientador(a):
Profª. Doutora Diana Mendes

Co-orientador(a):
Profª. Doutora Anabela Costa

Novembro 2010

ISCTE Business School
Instituto Universitário de Lisboa

ÁRVORES DE DECISÃO E REDES NEURONAIS: APLICAÇÃO A WEB MINING
Marcos André Pais Henriques

- Lombada –

RESUMO

A evolução do conceito de Marketing coloca a relação com o cliente no centro da estratégia da empresa. O fácil acesso a informação torna os clientes mais exigentes e susceptíveis à mudança de marcas com que se relacionam. Assim, as empresas sentem a necessidade de implementar estratégias de CRM – *Customer Relationship Managemen* que permitam obter informação e enviar estímulos aos clientes em todos os pontos de contacto destes com a empresa.

Este trabalho explora o potencial da Internet enquanto ferramenta que permite obter conhecimento sobre os consumidores, centrando-se na análise de dados obtidos através do *site* de um Clube de Fidelização de uma marca de Grande Consumo.

Assim, propõe-se com este trabalho uma metodologia de *Web Mining* de utilização que permita predizer comportamentos futuros, através de dados previamente recolhidos acerca dos comportamentos de utilização por parte dos utilizadores registados.

Para tal, a metodologia proposta assenta em duas etapas: segmentação e modelação. Na segmentação dos utilizadores recorre-se ao algoritmo *Two-Step*, reflectindo os comportamentos ao longo de três anos após a data de registo. Para a modelação, recorre-se a Árvores de Decisão (algoritmo CART) e Redes Neuronais (algoritmo *Backpropagation*), como métodos de classificação. Propõe-se ainda, para além da utilização de cada método de forma individual, a combinação de ambos num Modelo Híbrido.

Espera-se com esta metodologia obter informação que possibilite a incorporação em estratégia de CRM, nomeadamente, possibilitando criar políticas de Marketing geradoras de motivos de interesse e capazes de captar o retorno dos utilizadores ao *Site* de forma continuada.

Palavras-chave: Data Mining, Redes Neuronais, Árvores de Decisão, CRM
Classificação JEL: C44; C45; M31

ABSTRACT

The evolution of the Marketing concept puts the relationship with customer in a central position in the company strategy. Easy access to information makes customers more demanding and likely to change brands to which they relate. As an immediate consequence, the companies feel the need to implement new strategies for CRM - Customer Relationship Management in order to obtain information and send stimuli to customers at all points of contact with the company.

This work explores the potential of the Internet as a tool to obtain knowledge about consumers, focusing on the analysis of data obtained through the site of a Loyalty Club for an FMCG brand.

In order to achieve the main purpose, this work proposes to use a methodology of Web Usage Mining that allows to predict future behaviors, by considering data previously collected about the behavior of Registered Users.

The implemented methodology is based on two steps: segmentation and modeling. In the first step the users are segmented by using the Two-Step algorithm, reflecting the behavior along three years after the date of registration. For modeling, we use the Decision Trees (CART algorithm) and Artificial Neural Networks (Backpropagation algorithm) as methods of classification. It is also proposed, besides the use of each method individually, the combination of both in a Hybrid Model.

With this methodology, we expect to obtain information that facilitates the incorporation into CRM strategy, including the creation of Marketing policies that generate interest and are capable of capturing continuously the return of the Users Site.

Key-words: Data Mining, Neural Networks, Decision Trees, CRM
Classification JEL: C44; C45; M31

AGRADECIMENTOS

Agradeço a Deus as oportunidades que me meu ao longo da vida e pelo privilégio que me concedeu de chegar a este ponto.

Às orientadoras Profª. Doutora Diana Mendes e Profª. Doutora Anabela Costa agradeço toda a ajuda e apoio dado para que o trabalho seguisse o rumo certo.

À Cláudia, minha esposa, quero agradecer todo o apoio, insistência, elogio, pressão e companheirismo que foram os ingredientes necessários para a conclusão desta tese dentro do prazo previsto.

Aos meus Pais e Irmã agradeço todas as condições que me proporcionaram antes para que fosse possível ter chegado a este ponto, bem como todo o incentivo e toda a ajuda dada ao longo deste trajecto.

Agradeço a todos os restantes Familiares, Amigos e Colegas que sempre com uma palavra de interesse, uma palmadinha nas costas ou outros gestos nos incentivaram a continuar.

Ao Bruno, Tiago e à Leonor em especial, agradeço toda a ajuda que me deram na parte escolar do Mestrado no trabalho em grupo e individual. Com os três tudo foi mais fácil de ultrapassar. Agradeço todo o vosso apoio, companheirismo e espírito de equipa. Aprendi muito convosco.

Aos amigos que através do Facebook deixaram palavras de incentivo e constituíram uma boa "pressão social" à conclusão desta tese agradeço a motivação que me deram.

Quero também agradecer à pessoa que me cedeu a Base de Dados, permitindo que eu tivesse boa matéria-prima para trabalhar esta tese.

Agradeço também a Sofia Natal da OgilvyOne a disponibilidade para colaborar, dando uma visão muito pragmática do que se faz no mercado nesta área.

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

ÍNDICE

SUMÁRIO EXECUTIVO

A evolução do Marketing ao longo dos tempos levou à mudança de enfoque da actuação das empresas. A focalização da Gestão e dos recursos passa a centrar-se mais na Gestão da Relação com o Cliente e menos na Gestão do Produto.

Esta evolução releva a importância do conceito de Marketing Relacional que pode ser definido como "atrair, manter e, em organizações multi-serviços, fomentar relações com os clientes" (Berry, 1983).

A focalização das empresas na Gestão da Relação com o Cliente implica ter elevado conhecimento sobre todas as interacções deste com a empresa, independentemente dos pontos de contacto utilizados. Os Clientes mostram-se cada vez mais exigentes com as empresas, sendo menos tolerantes a situações de incoerência entre pontos de contacto ou a situações em que a empresa revele não conhecer o consumidor em algum dos pontos de contacto (ex: quando um cliente solicita informação através do *site* da empresa, espera que num contacto telefónico com a linha de apoio tenham conhecimento desse pedido).

Assim, surge a necessidade das empresas implementarem estratégias de CRM – *Customer Relationship Management.* O CRM é, segundo Lindon *et al.* (2004) "uma estratégia de negócio, uma atitude perante empregados e clientes, apoiada por determinados processos e sistemas em que o objectivo consiste em construir relações duradouras através da compreensão das necessidades e preferências individuais e, desta forma, acrescentar valor à empresa e ao cliente."

A Internet veio dar um enorme contributo para o desenvolvimento do CRM pois permite recolher informação contínua sobre o cliente, ter elevada interactividade com o cliente e estabelecer comunicação bilateral, potenciar o efeito rede entre clientes, maximizado pelas redes sociais melhor estabelecidas (ex: *facebook, hi5, linkedin*), estar perto do cliente em qualquer momento e em qualquer lugar, através do desenvolvimento das tecnologias de acesso móvel e comunicar de forma individualizada com cada destinatário.

A implementação de uma estratégia de CRM integrada e adequada permite ter um relacionamento "360º" do cliente, utilizando as Tecnologias de Informação para que os pontos de contacto funcionem numa perspectiva dupla de comunicação: (1) receber todos os inputs

voluntários e involuntários do cliente (2) enviar estímulos que contribuam para aumentar o lucro e a satisfação do cliente. Estas estratégias deverão contemplar a utilização de metodologias de segmentação adequadas que permitam identificar grupos distintos de clientes com necessidades únicas, com diferentes níveis de satisfação e de lealdade e a quem devem ser oferecidos produtos, serviços e promoções diferenciados. Por sua vez, a aplicação de métodos preditivos permite antecipar comportamentos e com isso desenvolver estratégias de Marketing Relacional com efeito preventivo, nomeadamente, para evitar comportamentos de *churn*.

Nesta tese recorreu-se a uma Base de Dados real que reflecte as interacções dos Utilizadores Registados de um *site* de um Clube de Fidelização de uma marca de Grande Consumo. Foram seleccionados Utilizadores Registados entre 2007 e 2009.

Da análise da informação recolhida, constatou-se que a taxa de utilizadores que não têm qualquer interacção com o *site* ou com *e-mail* enviado pela marca é bastante elevada, aumentando em função do ano de registo (24,2% em 2007 e 54,1% em 10 meses de 2009) e também do tempo decorrido após o registo (entre os utilizadores de 2007 aumenta de 24,2% no ano de registo para 67,7% no ano seguinte). Outros indicadores analisados corroboram também que existe elevada dificuldade neste *site*, tal como noutros similares, em conseguir manter os utilizadores interessados no *site* e manter a sua taxa de visita com níveis elevados.

Assim, pretendeu-se obter um Modelo que permitisse segmentar os comportamentos de interacção dos utilizadores com registo em 2007 nos anos após o registo e definir um Modelo de Classificação que possa ser aplicado aos utilizadores registados em 2008 e 2009 e assim predizer os seus comportamentos futuros com base no segmento a que pertencem.

Assim, recorreu-se ao método de *Two-Step Clustering*, desenvolvido por Chiu *et al.* (2001), para segmentar os utilizadores de 2007 com base nos comportamentos nos Anos 0, 1 e 2 após o momento de registo. Os resultados obtidos permitiram concluir a existência de quatro segmentos de utilizadores:

- Segmento 1 – Visitantes Intensivos Fiéis (5,7%)

- Segmento 2 – Visitantes Moderados Potencialmente Fiéis (44,9%)

- Segmento 3 – Visitantes do *Site Churners* (19,6%)

- Segmento 2 – Não Visitantes (29,8%)

A estrutura de segmentos obtidos diferencia os utilizadores em termos de comportamentos, não existindo diferenças relevantes em termos sócio-demográficos que permitam explicar as diferenças de comportamentos.

Para criar o Modelo de Classificação que permita predizer quais serão os comportamentos dos utilizadores com base nos seus comportamentos no Ano 0 após o registo, aplicaram-se individualmente as Metodologias de Classificação de Árvores de Decisão (com algoritmo CART) e de Redes Neuronais (com algoritmo *Backpropagation*), bem como a combinação de ambas num Modelo Híbrido.

As melhores alternativas obtidas com cada Modelo obtém desempenhos muito similares no que respeita à capacidade global de classificar correctamente a amostra de teste, sendo que foram classificados correctamente 97,57% dos casos com Árvore de Decisão, 97,82% com Redes Neuronais e 97,75% na melhor opção de Modelo Híbrido. Em qualquer das alternativas de Modelo apresentadas constata-se menor capacidade de classificar correctamente os casos do Segmento 1 (Visitantes Intensivos Fiéis) que é também o segmento de menor dimensão mas que tem bastante relevância em termos de fidelização e intensidade de utilização. A capacidade de classificação correcta deste segmento nas melhores alternativas obtidas com cada método varia entre 72,88% com Rede Neuronal e 80% com Modelo Híbrido. Os casos do Segmento 1 que não são classificados correctamente, são na quase totalidade dos casos classificados como Segmento 2, o que pode contribuir para um menor risco na classificação incorrecta, dado que o Segmento 2 é de facto aquele em que os utilizadores apresentam mais semelhanças em termos de comportamento futuro com os do Segmento 1.

1. PROJECTO DE PESQUISA

1.1 Introdução

Para conseguir tirar o máximo partido da relação com o cliente, as empresas necessitam de estabelecer relações sólidas no longo prazo de modo a maximizar a rentabilidade de cada cliente. Para atingir esse desiderato, torna-se fundamental as empresas implementarem estratégias de CRM – *Customer Relationship Management*. O CRM é, segundo Lindon *et al*. (2004) "uma estratégia de negócio, uma atitude perante empregados e clientes, apoiada por determinados processos e sistemas em que o objectivo consiste em construir relações duradouras através da compreensão das necessidades e preferências individuais e, desta forma, acrescentar valor à empresa e ao cliente."

1.2 O Problema

Para uma marca de Grande Consumo o *site* constitui a principal plataforma de contacto com os consumidores no mundo digital. Independentemente do produto ou da marca, existe a clara noção de que quanto mais forte for a relação estabelecida com o consumidor no "mundo digital", mais forte será a vinculação à marca no curto e no longo prazo, com tradução directa na rentabilidade do consumidor. No entanto, manter a relação com os consumidores na *Web* pode se revelar também bastante difícil, pelo que se torna necessário desenvolver mecanismos que possibilitem à marca estabelecer contacto freqüente e bidireccional com o consumidor. A criação de clubes de fidelização visa atingir estes objectivos, através da recolha de informação sobre o consumidor e da obtenção de autorização para enviar comunicação através de vários meios, sendo o e-mail o meio mais frequente.

1.3 Objectivos

Nesta tese pretende-se, através de Árvores de Decisão e de Redes Neuronais desenvolver um Modelo de Classificação que permita antecipar os comportamentos dos Utilizadores

Registados de um *site* de Clube de Fidelização de uma marca de Grande Consumo. Em concreto, pretende-se a partir de segmentação efectuada aos utilizadores registados em 2007 com base nos comportamentos de visita ao *site* nos Anos 0,1 e 2 após o registo, desenvolver Modelo de Classificação que com base na informação do Ano 0 consiga classificar o segmento a que cada utilizador pertence e assim predizer os comportamentos esperados para os Anos 1 e 2.

1.4 Relevância do Estudo

A realização deste estudo permitirá, através da aplicação do Modelo de Classificação aos Utilizadores Registados em 2008 e 2009, predizer após o Ano 0, os comportamentos futuros destes utilizadores nos Anos 1 e 2. Com esta informação a Gestão do *site* poderá desenvolver acções de forma customizada a cada segmento para maximizar as situações de interacção entre o *site* e o utilizador e, consequentemente, evitar comportamentos de abandono e desinteresse. Assim, espera-se com esta informação contribuir para a estratégia integrada da empresa em CRM.

1.5 Estrutura do trabalho

A estrutura do presente trabalho encontra-se definida a partir de um capítulo de Enquadramento (Capítulo 2), na qual é possível compreender o enquadramento teórico deste projecto no que respeita ao Marketing Relacional, ao CRM e ao Data Mining.

No Capítulo 3 são apresentados os Modelos de Classificação (Árvores de Decisão e Redes Neuronais), bem como os respectivos algoritmos (CART e *Backpropagation*, respectivamente). É também apresentado neste capítulo a descrição do Modelo Híbrido que será obtido pela utilização conjunta dos dois anteriores.

Por sua vez, no Capítulo 4 apresenta-se a caracterização da Base de Dados no que respeita ao perfil sócio-demográfico dos utilizadores, ao perfil de registo e utilização efectuada e, por fim, apresenta-se a segmentação obtida.

O Capítulo 5 constitui elemento central deste trabalho e é nele que é apresentado a aplicação do Modelo de Classificação.

No Capítulo 6 procede-se à apresentação das conclusões finais do trabalho.

2. ENQUADRAMENTO

2.1 O Marketing Relacional

A designação de Marketing Relacional, apesar de ser recente, baseia-se em conceitos já utilizados desde sempre no comércio tradicional. A relação estabelecida entre cliente e/ou consumidor e vendedor já neste tipo de comércio se pretendia que fosse próxima, estreita, baseada em elevada confiança e conhecimento mútuo, para que perdurasse no tempo. O retorno do investimento era projectado a longo prazo, o vizinho era o cliente e a garantia da sobrevivência do negócio. Criava-se uma relação de longo prazo com os clientes como se fosse uma extensão da própria família, como referiu McLuhan (1969) na sua obra "Os meios de Comunicação como Extensão do Homem".

Esta é a base da actual designação de Marketing Relacional e que se pretende que seja transposta para o comércio actual das grandes organizações com milhões de clientes. Apesar da elevada dimensão de uma empresa multinacional que opere no mercado de Grande Consumo que possui centenas de milhares de empregados e milhões de consumidores em todo o mundo, a empresa pretende conhecer os clientes de cada uma das suas marcas. As tecnologias de informação vieram dar um contributo muito importante para que as marcas e os consumidores possam interagir em tempo real e as empresas possam conhecer e antecipar comportamentos e atitudes dos seus consumidores melhorando a experiência destes e o respectivo retorno para a marca.

A APAP – Associação Portuguesa das Agências de Publicidade define o Marketing Relacional como "toda a forma de publicidade que visa estabelecer e manter relações entre a marca e o seu consumidor com base em acções personalizadas, interactivas e mensuráveis, criando uma base de conhecimento em constante evolução para a construção da marca." (citado em Lindon *et al.*, 2004).

O conceito de Marketing Relacional surge como evolução e ampliação do conceito de Marketing Directo. O conceito de Marketing Directo foi utilizado a primeira vez em 1967 por Lester Wunderman[1] num discurso no MIT e que levou ao crescimento teórico e prático do

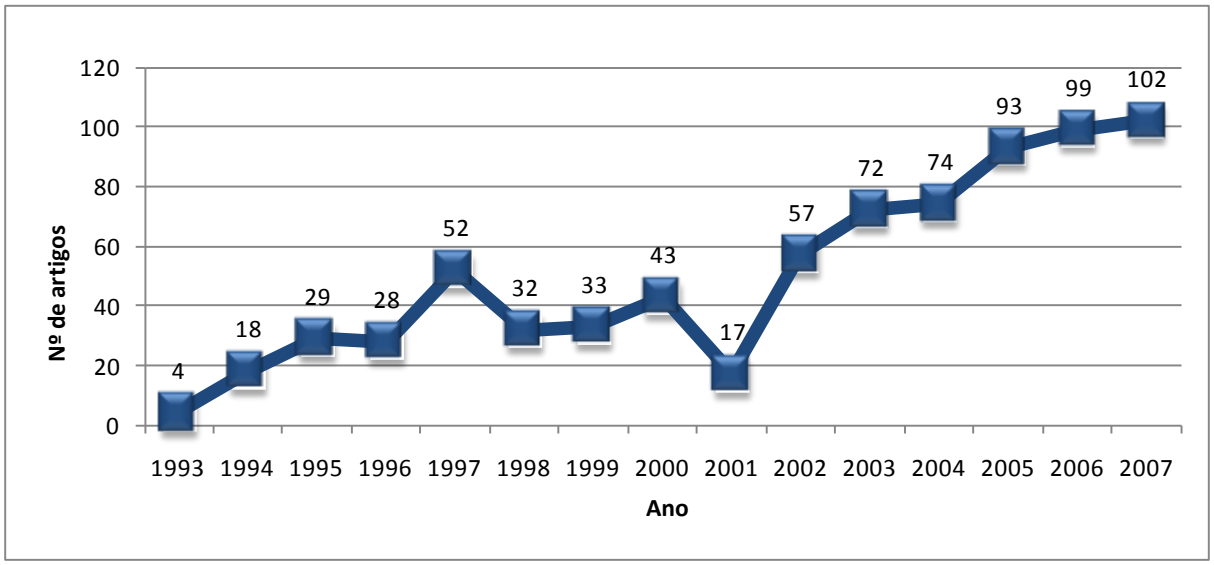
[1] Lester Wunderman (n. 22 de Junho de 1920) é o Fundador da Agência de Comunicação Wunderman, sendo considerado o criador do Marketing Directo actual. As suas inovações incluem, entre outras, uma revista

Marketing Directo. Antes de Wunderman, meios de comunicação publicitária como o Telefone e os Envios Postais eram quase exclusivamente utilizados por empresas de venda directa e/ou à distância.

Por sua vez, segundo Antunes e Rita (2008), no artigo "O marketing relacional como novo paradigma: uma análise conceptual" o termo Marketing Relacional foi utilizado pela primeira vez na literatura de marketing de serviços, por Berry (1983) e a partir desse momento, o Marketing Relacional ganhou cada vez mais interesse junto dos investigadores (Harker e Egan, 2006) e também na prática empresarial.

No entanto, no mesmo artigo os autores analisam a evolução da publicação de artigos nas principais revistas internacionais da especialidade entre 1993 e 2007 e constata-se que a tendência é crescente desde o primeiro ano analisado (ver Figura 1). Em relação aos anos anteriores a 1993, encontram-se esporadicamente artigos relacionados com esta temática no sector dos serviços, nomeadamente no sector bancário.

Figura 1 – Número de artigos de Marketing Relacional publicados nas principais revistas internacionais de marketing, entre 1993 e 2007



Em 1993, Peppers e Rogers lançaram o livro *The One to One Future* e patentearam o conceito "*One-to-One*" contribuindo para uma nova evolução do conceito. Segundo Segadães[2] no prefácio do capítulo sobre Marketing Relacional em Lindon *et al.* (2004), o Marketing *One-to-One* impulsionado pela obra dos autores Peppers e Rogers "parecia ter encontrado a

associada a um cartão de cliente, uma linha de call-center com número grátis, programas de recompensa por fidelização. Foi incluído no Advertising Hall of Fame em 1998.

[2] Luis Segadães desempenhava em 2004 o cargo de Presidente da CP Proximity Portugal (agência de Marketing Relacional) e em 2006 tornou-se Presidente da Touch Me Wunderman, cargo que ocupou até final de 2007.

fórmula certa para explicar o conceito a todos os *marketeers* e sobretudo aos *mass-marketeeres*, cujas mentes estavam ainda muito «presas» aos 4P's, aos mass-media e à publicidade generalista."

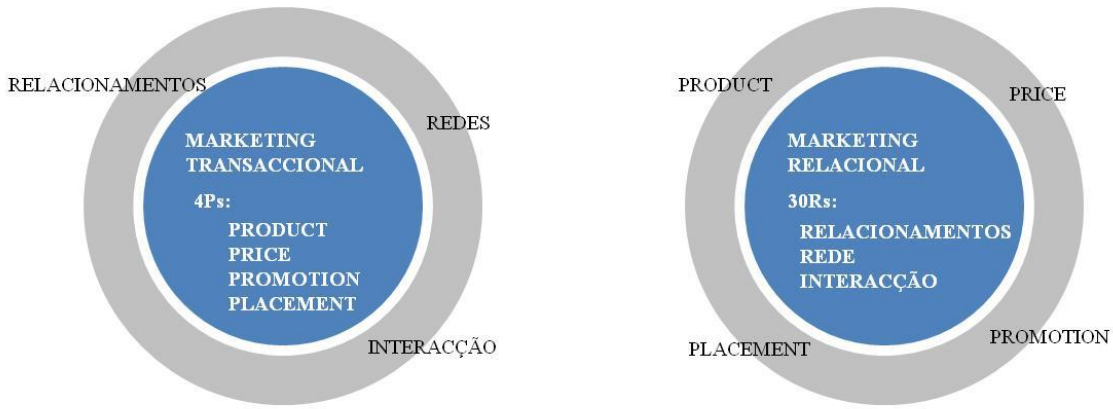
Assim, constata-se que o Marketing Relacional pressupõe uma mudança clara no enfoque quando comparado com o Marketing Transaccional. Na Tabela 1 sistematizam-se as principais diferenças entre o Marketing Transaccional e o Marketing Relacional.

Tabela 1 – Principais diferenças entre Marketing Transaccional e Marketing Relacional (Fonte: Lindon *et al.*, 2004)

	Marketing Transaccional (ou de Massas)	Marketing Relacional
Perspectiva Temporal	Curto prazo	Longo Prazo
Função de Marketing Dominante	Marketing-Mix	Marketing Interactivo
Elasticidade de Preço	Clientes mais sensíveis ao preço	Clientes menos sensíveis ao preço
Dimensão da Qualidade Dominante	Qualidade do produto/output	Qualidade das interacções
Medida da Satisfação do Cliente	Quota de Mercado	Quota de cliente
Sistemas de Informação sobre o Cliente	Estudos de mercado ad-hoc	Feedback em tempo real
Conhecimento do Cliente	Anónimo	Identificação Individual
Frequência dos Contactos	Comunicação esporádica e unilateral	Diálogo constante
Foco da Gestão	Gestão do Produto	Gestão do Cliente

Segundo Gummesson (2005) o Marketing Relacional não abdica dos 4P's (*Product, Price, Promotion* e *Placement*) do Marketing Transaccional mas estes deixam de ter a posição central na estratégia, dando lugar a um modelo de 30R's baseados em: Relacionamentos, Rede e Interacção (ver Figura 2).

Figura 2 – Uma mudança dos 4P's para Relacionamentos, Rede e Interacção



Fonte: Gummesson 2005

Para Gummesson (2005) o Marketing Relacional é "o Marketing baseado em interacção dentro das redes de relacionamentos". Assim, Gummesson defendeu uma nova abordagem do conceito de Marketing assente em três pilares (citado em Antunes, 2008):

- **a relação** – o marketing deve estar orientado para a criação, manutenção e desenvolvimento de relações com os clientes;
- **a interactividade das partes** – as relações entre vendedores e clientes para a criação e entrega mútua de valor exigem um estreito e intenso processo de comunicação entre ambos;
- **o longo prazo** – para criar, manter e desenvolver as relações é necessário um longo espaço temporal.

Para Berry (1983), o Marketing Relacional é "atrair, manter e, em organizações multi-serviços, fomentar relações com os clientes".

Para estabelecer e manter boas relações com os clientes é necessário (Lindon *et al*, 2004):

- Conhecer;
- Ser relevante;
- Comunicar;
- Escutar;
- Recompensar pela sua fidelidade;
- Associar à vida da empresa ou da marca.

Estas mudanças levaram a que a *American Marketing Association* (AMA) sentisse necessidade de redefinir o próprio conceito de Marketing, contemplando agora esta nova realidade de focalização no cliente e nas relações estabelecidas entre a organização e os seus *stakeholders*: *"Marketing is the activity, set of institutions, and processes for creating, communicating, delivering, and exchanging offerings that have value for customers, clients, partners, and society at large."* (Outubro 2007)

2.2 A Estratégia de CRM – *Customer Relationship Management*

CRM ou *Customer Relationship Management* "é uma estratégia de negócio orientada para o cliente concebida para optimizar o lucro e a satisfação do cliente" (Ferrão, 2003). Já em Lindon *et al*. (2004) vão mais longe e definem o conceito como sendo "uma estratégia de negócio, uma atitude perante empregados e clientes, apoiada por determinados processos e sistemas em que o objectivo consiste em construir relações duradouras através da compreensão das necessidades e preferências individuais e, desta forma, acrescentar valor à empresa e ao cliente."

Ambas as definições vão assim de encontro ao conceito base de Marketing Relacional que foi abordado no ponto anterior. Constata-se, que apesar de muitas vezes surgir associado aos *softwares* utilizados pelas consultoras que trabalham nesta área, o CRM é muito mais que isso.

A estratégia de CRM implica ter uma visão única e integrada do cliente, compreendendo "todas as operações, processos e tecnologias que se desenvolvem tendo como objectivo o cliente, e a satisfação das suas necessidades, numa atitude proactiva das empresas." (Ferrão, 2003). A Gestão das Relações com o Cliente, tradução directa do conceito, é então aplicada numa perspectiva multicanal e afecta todas as áreas da empresa. Desta forma conseguir-se-á proporcionar ao cliente uma experiência consistente da marca e proporcionar à empresa uma visão única do cliente, independentemente do meio de contacto escolhido.

A implementação de estratégia de CRM terá como objectivo a maximização do *Life-time-value* dos clientes, através da adequação da oferta e dos *touch-points* entre a marca e os clientes de forma a, com base no conhecimento dos clientes, aumentar a satisfação, fidelizar os clientes, incrementar a venda de produtos mais caros (*up-selling*) e incrementar a venda cruzada de produtos (*cross-selling*).

Segundo Brown (2000) a estratégia de CRM deverá responder às seguintes questões:

- Quais são os segmentos de clientes chave, com base nas necessidades actuais e futuras?
- Existem grupos distintos de clientes com necessidades únicas?
- Existem determinados grupos de clientes a quem devem ser oferecidos produtos e serviços específicos?

- A empresa tem implementadas estratégias específicas para assegurar a lealdade e retenção de clientes?
- A empresa tem estabelecida uma relação *win-win* com o cliente?

Segundo o mesmo autor, a implementação de estratégia de CRM tem vantagens sobre a implementação de estratégias assentes em marketing de massas, nomeadamente:

- Reduz os custos de publicidade;
- Facilita a medição dos resultados de uma dada campanha;
- Permite às empresas competirem pelos clientes com base em serviço e não em preços;
- Previne gastos excessivos com clientes de baixo valor ou gastos insuficientes com clientes de elevado valor;
- Acelera o tempo de desenvolvimento e comercialização de novos produtos;
- Melhora o uso de canais de contacto com o cliente, maximizando cada contacto com o cliente.

Por limitação no acesso a dados sobre consumidores esta tese centrar-se-á na análise de dados obtidos através de um único ponto de contacto: o *site* de Internet. De qualquer modo, espera-se com estes dados obter informação que possibilite a incorporação em estratégia de CRM, nomeadamente, com a definição de acções com efeito não só no canal *Web*, mas também que possam ser transpostas para acções noutros meios de comunicação.

Com o desenvolvimento da economia digital, verificou-se um desenvolvimento exponencial do CRM enquanto estratégia e consequentemente dos *softwares* disponibilizados pelas consultoras nesta área, nomeadamente da *Siebel, Oracle* e *SAP*, havendo soluções especializadas para a gestão de clientes via canal electrónico (ex: *OpenText*, *Soverain* e *Escalate*).

A Internet constitui o meio por excelência para o desenvolvimento de acções de CRM, pois permite como nenhum outro meio:

- Recolher informação contínua sobre o cliente;
- Ter elevada interactividade com o cliente, estabelecendo comunicação bilateral em tempo real e envolvendo-o activamente com a marca;
- Potenciar o efeito rede entre clientes, maximizado pelas redes sociais melhor estabelecidas (ex: *facebook, hi5, linkedin*);
- Estar perto do cliente em qualquer momento e em qualquer lugar, através do desenvolvimento das tecnologias de acesso móvel;

- Comunicar de forma individualizada com cada destinatário.

Assim, a definição de uma correcta estratégia de Internet Marketing é um factor crítico de sucesso para uma boa estratégia de CRM.

Linnof e Berry (2001) descrevem o papel da *Web* no Marketing como tendo mudado a relação existente entre Publicidade e Gestão de Marca e Marketing Directo, por possibilitar saber com algum grau de precisão quem está a ver os anúncios, quem está a aceder a cada anúncio e, também, o que leva os consumidores a fazer compras.

A presente tese explorará o potencial da Internet enquanto ferramenta que permite obter conhecimento sobre os consumidores, não na resposta a anúncios, mas sim através da navegação em *site* corporativo de uma empresa de Grande Consumo.

2.3 A aplicação do *Data Mining*

O *Data Mining* é "o processo de exploração e análise, por meios automáticos ou semi-automáticos, de largas quantidades de dados com o objectivo de descobrir padrões e regras com significado" (Linnof, 1997).

A definição anterior contempla as principais características do *Data Mining* e que constituem eixos fundamentais dos benefícios da sua aplicação em estratégias de CRM. Através da utilização de *Data Mining* as empresas podem transformar quantidades imensas de registos em informação com significado, que permita ter uma imagem clara do comportamento dos clientes e, consequentemente, suportar a estratégia de CRM e o Marketing Relacional.

Assim, através da utilização de *Data Mining*, como uma ferramenta de CRM é possível às empresas aprenderem com o comportamento dos seus clientes e assim acompanhar todo o ciclo de vida do cliente, podendo obter conhecimento sobre comportamentos futuros e tomar acções que, através de soluções *win-win*, contribuam para conduzir a relação Empresa - Cliente na direcção pretendida pela Empresa.

O termo *Data Mining*, segundo Linnof (2000) pode ser utilizado para definir três actividades de *Data Mining* direccionado:

- **Classificação** – consiste em examinar as características de um novo registo na base de dados e atribuir-lhe uma classe predefinida (valor discreto);

- **Estimação** – tal como a classificação, consiste em examinar as características de um novo registo na base de dados, mas neste caso atribuir-lhe um valor contínuo;

- **Predição –** qualquer predição pode ser pensada como Classificação ou Estimação, pelo que o isolamento das actividades anteriores pretende dar mais ênfase à possibilidade que existe de classificar à *priori* comportamentos ou eventos que poderão ocorrer no futuro. Assim, para confirmar a precisão do modelo basta "esperar e ver".

Para qualquer das actividades existem inúmeras situações de aplicação a CRM. Esta tese, tem como objectivo predizer comportamentos futuros, utilizando métodos de classificação que serão explicados com maior detalhe no Capítulo 3.

A Internet constitui-se um canal por excelência para a aplicação de *Data Mining*, dadas as suas características: grande quantidade de dados disponíveis, recolhidos em tempo real e possibilidade de gerar estímulos ao consumidor, também em tempo real com base em padrões e regras obtidos com processos de *Data Mining.*

Na aplicação do *Data Mining* à *Web*, Rud (2001) distingue os conceitos *Web Mining* e *Web Modeling*, considerando que o primeiro se aplica a modelos mais tradicionais como os modelos preditivos, de definição de perfis e de segmentos. Estes modelos são construídos *offline* e os seus *scores* são aplicados aos utilizadores durante as sessões. No segundo conceito (*Web Modeling*), a autora inclui os modelos que são construídos e aplicados durante a sessão do utilizador.

Transversalmente aos dois conceitos definidos por Rud (2001), para a utilização de *Data Mining* aplicado à Internet, Linoff (2001) define três tipos de actividades em que quer os dados quer as razões de utilização diferem:

- ***Mining* de Estrutura** – consiste no processo de extrair informação sobre as ligações entre páginas na *Web*. Permite responder a questões como: Que páginas são as destinatárias de *links* a partir de outras? Que páginas apontam para outras? Que grupos de páginas formam ilhas?

- *Mining* de Utilização – é o processo de extracção de informação da forma como os utilizadores que percorrem os *links* fazem uso dos mesmos. Permite responder a questões como: Que páginas visitam? Quanto tempo ficam em cada página? A que acedem a seguir? Que caminhos pelo *site* levam ao contador de verificação geral e quais levam direito à saída?

- *Mining* de Conteúdos – é o processo de extracção de informação útil a partir do texto, imagens e outras formas de conteúdos que compõem as páginas. Alguns motores de busca ou de recomendação de páginas utilizam *Data Mining* deste tipo para ajudar os utilizadores a encontrar o que procuram na Internet. Permite responder a questões como: Que *site* tem o melhor negócio em "molho picante"? Que páginas estão escritas em alemão? Que páginas estão relacionadas com dança folclórica ou chuva ácida?

Assim, de acordo com a definição de Rud (2001), o âmbito da presente tese insere-se em *Web Mining* visto que os dados foram recolhidos previamente e toda a construção do modelo será efectuada *offline*, pretendendo-se a posterior inclusão no *site* do resultado do modelo definido para classificar novos casos. De acordo com as tipologias de actividades, anteriormente enunciadas e definidas por Linnof (2001), os objectivos desta tese incluem a mesma na extracção de informação acerca da utilização por parte dos utilizadores de um *site*, ou seja, em *Mining* de Utilização.

2.4 Melhores práticas de Marketing Relacional com base na *Web*: aplicação a marcas de Grande Consumo

Em Portugal, a aplicação a marcas de Grande Consumo do Marketing Relacional na *Web* tem vários exemplos. As empresas apostam em *microsites* adaptados à personalidade da marca e aos comportamentos dos consumidores, criando estímulos para o retorno ao *site*. Esta estratégia é bem patente em *site's* como o da *Knorr* (com receitas e promoções) ou como o das marcas *Ben&Jerrys* e *Pringles* (com jogos para crianças), entre outros. Algumas marcas, por sua vez, reconhecem a importância da recolha de informação sobre o cliente e colocam no *site* um local para subscrição de informação que permita enviar aos clientes informação sobre

novidades, promoções ou outras. Conforme se pode observar na Tabela 2, esta alternativa é utilizada pela Unilever em muitas das suas marcas.

Tabela 2 – Exemplos de *sites* de marcas de Grande Consumo com possibilidade de Registo de Utilizadores

Nome	Marca	Empresa	Origem da Empresa
Adagio	Adagio	Lactogal	Nacional
Vaqueiro	Vaqueiro	Unilever	Multinacional
Maizena	Maizena	Unilever	Multinacional
Sun	Comfort	Unilever	Multinacional
Surf	Surf	Unilever	Multinacional
Skip	Skip	Unilever	Multinacional
Mundo Comfort	Comfort	Unilever	Multinacional
Axe	Axe	Unilever	Multinacional

A criação de Clubes de Fidelização, por sua vez, é uma estratégia que implica maior investimento de recursos, mas também o retorno em informação sobre o consumidor e envolvimento do mesmo é superior. Na Tabela 3, sistematizam-se alguns exemplos de Clubes de Fidelização em marcas de Grande Consumo.

Tabela 3 – Exemplos de Clubes de Fidelização de marcas de Grande Consumo

Nome	Marca	Empresa	Origem da Empresa
My Special K	Special K	Kellogg's	Multinacional
Família Mimosa	Mimosa	Lactogal	Nacional
Petnet	Whiskas, Pedigree, Sheba, entre outras	MARS	Multinacional
Aptamil	Milupa Aptamil	Milupa	Multinacional
Clube Bebé Nestlé	Nestlé Nan, Nestlé Cerelac e Nestlé Frutíssima	Nestlé	Multinacional
Clube Nesquik	Nesquik	Nestlé	Multinacional
Saúde de Faca e Garfo	Nestlé, Maggi, Buitoni e outras	Nestlé	Multinacional
Pet Life	Purina, Dog Chow e Tidy Cats	Nestlé	Multinacional
Clube Pescanova	Pescanova	Pescanova	Multinacional
Clube Grumete	Pescanova	Pescanova	Multinacional
Clube Coração Saudável	Becel	Unilever	Multinacional
Clube Olá	Olá	Unilever	Multinacional

Como alternativa aos Clubes de Fidelização, mas com objectivos similares, as marcas apostam também em ter, a partir do seu *site*, fóruns de discussão/partilha de ideias, testemunhos, *blogs*, votações, canais de vídeo e/ou comunidades de consumidores.

No sentido de obter a perspectiva do mercado empresarial sobre o que de melhor se faz em Marketing Relacional para marcas de Grande Consumo, foi efectuada entrevista[3] com Sofia Natal, *Group Account Director Consulting* da *OgilvyOne Worldwide, Lisbon*[4]. Nesta entrevista é realçado o papel da *Web* enquanto meio revolucionário da relação das marcas com os consumidores, pelo seu potencial de interactividade, curto tempo de resposta e menores custos.

As marcas de Grande Consumo apostam em Programas Relacionais, que desempenham um papel fundamental na estratégia das marcas. Estes Programas assentam em segmentações orientadas para tipologias de consumo/consumidores, sendo elaborados planos de contacto próprios, adaptados a cada segmento. Estes Programas visam chegar ao consumidor de forma integrada, multi-plataforma, para maximizar o *Consumer Lifetime Value*.

Actualmente, o *site* é o meio preferido de relação dos consumidores com a marca, em simultâneo com o *e-mail* marketing, sendo o *site* o principal fornecedor de informação para a marca, através do número de visitas, origem da chegada ao s*ite*, *pageviews* e o tempo médio de visita.

O *site* da Nestlé funciona como portal para *sites* específicos centrados nas afinidades com os consumidores das várias marcas (ex: *site* de bebés, crianças, animais, culinária e bem estar). No caso da Nestlé, para além do *site* são utilizadas outras plataformas digitais de comunicação como *e-news* sazonais, *e-mails* marketing e página no Facebook.

Para o sucesso de um Programa Relacional é fundamental a base de dados, isto é, a informação disponível sobre o consumidor. Para que a comunicação seja adequada, estimulante e permita accionar informação que mantenha as pessoas interessadas na marca e envolvidas com a mesma na perspectiva de uma relação duradoura, deve ser segmentada por afinidade, consumo ou por outra forma que seja considerada relevante.

---

[3] Ver entrevista em detalhe no Anexo I

[4] A *OgilvyOne Worldwide, Lisbon* é uma agência que fornece Soluções Integradas de Marketing Relacional.

Actualmente as metodologias de *Data Mining* são utilizadas sobretudo para reconhecimento de grupos de afinidade não naturais, que permitam desenhar acções específicas para cada grupo, no sentido de incentivar as migrações entre grupos de menor valor para maior valor e obter maior fidelização.

Como área de desenvolvimento futuro dos Programas Relacionais está previsto o *reward* dos consumidores mais activos nas plataformas digitais. Assim, é previsível a implementação de técnicas de *Data Mining* e/ou de Segmentação, com base nos comportamentos dos consumidores na utilização das plataformas digitais.

3. CLASSIFICAÇÃO COM ÁRVORES DE DECISÃO E REDES NEURONAIS

Para atingir os objectivos pretendidos para este estudo utilizar-se-ão duas metodologias de classificação: Árvores de Decisão e Redes Neuronais. Estas duas metodologias servirão de base à implementação de um Modelo Híbrido.

Existem vários algoritmos disponíveis para a construção quer de Árvores de Decisão como de Redes Neuronais. Para a construção de Árvores de Decisão podemos utilizar o algoritmo CHAID (Kass, 1980), CART (Breiman et al., 1984) e C5 (Quinlal, 1993). Para a construção de Redes Neuronais, podemos utilizar o algoritmo *Backpropagation*, algoritmos genéticos, entre outros.

Nesta secção apresentar-se-á mais aprofundadamente os algoritmos que serão utilizados nas análises realizadas nas restantes secções do trabalho e que são o algoritmo CART para as Árvores de Decisão e o algoritmo *Backpropagation* para as Redes Neuronais.

3.1. Algoritmo CART

O algoritmo CART (*Classification and Regression Trees*) é um dos algoritmos possíveis de ser utilizado em modelos de classificação (variável alvo nominal) e regressão (variável alvo métrica) com Árvores de Decisão, tal como o nome indica. Foi desenvolvido por Breiman, Friedman, Olshen e Stone (1984) e é um dos algoritmos mais utilizados neste tipo de análises.

Este algoritmo constrói árvores binárias de classificação e regressão, isto é, cada um dos nós de decisão contém apenas duas ramificações, começando sempre no nó raiz da árvore onde se encontra a amostra de treino. Em cada um destes nós de decisão, o crescimento da árvore dá-se através de uma pesquisa exaustiva entre todas as variáveis independentes e todas as separações possíveis de valores, seleccionando a melhor separação de modo a que a diversidade da variável alvo decresça o mais possível nos seus nós descendentes, até aos nós folha, isto é, o processo repete-se até atingir o critério de homogeneidade ou até que sejam satisfeitos outros critérios de paragem impostos à árvore. Assim, este algoritmo é um exemplo

de um algoritmo de partição binária recursiva, pois é aplicado recursivamente a cada um dos subconjuntos gerados, até que não seja possível (ou necessário) efectuar mais nenhuma partição.

Num problema de classificação, e para o cumprimento do critério de minimização da diversidade da árvore, o CART utiliza como medidas de diversidade o índice de Gini ou o índice de Twoing. Segundo os autores deste algoritmo, a construção de árvores utilizando o método CART é pouco sensível à utilização de um ou de outro índice de diversidade, ficando assim ao critério do analista a escolha de qual utilizar. Contudo, a utilização do índice de Twoing inviabiliza a introdução de custos diferenciados de classificação incorrecta no processo de ramificação, ao contrário do índice de Gini em que tal é possível. A fórmula de cálculo do índice de Gini no nó *O* é:

$$G(O) = \sum_{l \neq l^*} p(l;O)\, p(l^*;O) = 1 - \sum_{l=1}^{L} p(l;O)^2 , \quad (1)$$

onde,

- *L*: Número de categorias associado à variável alvo nominal Y;
- *O*: Nó de uma árvore/conjunto de observações;
- *l*: Índice de categorias para a variável alvo nominal Y;
- *p(l;O)*: Probabilidade empírica de ocorrência da categoria *l* de Y em *O*;
- $p(l^*;O)$: Probabilidade efectiva de ocorrência da categoria *l* de Y em *O*.

Por outro lado, num problema de regressão, a medida de diversidade utilizada para uma variável métrica é a variância (S^2) do conjunto de dados.

O decréscimo de diversidade (ou impureza) associada à partição resultante da ramificação Π do nó *O* da árvore *A*, pode ser calculado através da seguinte expressão:

$$\Delta I(\prod; O; A) = \sum_{c=1}^{C} p(O_c) I(O_c) - p(O) I(O) , \quad (2)$$

onde,

- *C*: Número de nós descendentes resultante da ramificação Π do nó *O* (no algoritmo de CART tem-se $C = 2$, uma vez que a árvore é binária);
- *c*: Índice das classes de uma partição que resulta da ramificação Π do nó *O*;
- $p(O_c)$: Probabilidade empírica associada ao nó descendente O_c;

- $I(O_c)$: Medida de diversidade (ou impureza) associada ao nó descendente O_c (a qual, nesta tese, é calculada através do índice de Gini);
- $p(O)$: Probabilidade empírica associada ao nó O;
- $I(O)$: Medida de diversidade (ou impureza) associada ao nó O.

Na aprendizagem de uma Árvore de Decisão é necessário ir avaliando a sua qualidade e o seu ajustamento aos dados da população. Para isso a amostra global é dividida em duas amostras: treino e teste, sendo o primeiro o conjunto de dados a partir dos quais se aprende e a segunda é utilizada para testar a árvore resultante da aprendizagem. O erro de resubstituição é o erro que se associa à amostra do conjunto de treino.

Em regressão, este valor pode ser calculado através do erro quadrático, o qual é definido por:

$$E_n = (y_n - \hat{y}_n)^2, \quad (3)$$

onde,

- y_n : n-ésima observação da variável alvo Y;
- $\hat{y}_n$: Valor estimado para a n-ésima observação da variável alvo Y.

Com o objectivo de obter um ajustamento de qualidade, minimiza-se o erro quadrático total:

$$\sum_{n=1}^{N} E_n, \quad (4)$$

sendo,

- N: Número total de casos observados.

Pode ainda ser calculada a capacidade preditiva do modelo desenvolvido, através da comparação da sua capacidade preditiva com a capacidade preditiva por defeito. Concretamente, a medida de qualidade da regressão pode ser definida por:

$$R^2 = 1 - \frac{\sum_{n=1}^{N} (y_n - \hat{y}_n)^2}{\sum_{n=1}^{N} (y_n - \bar{Y})^2}, \quad (5)$$

onde,

- $\bar{Y}$: Média amostral da variável alvo Y (explicada ou dependente).

Em classificação, o erro é obtido pelo cálculo do número de casos incorrectamente classificados e da sua proporção. O número de casos incorrectamente classificados pode ser determinado por: $\sum_{n=1}^{N} E_n$, (6)

onde,

- N: Número total de casos observados;

- $E_n = \begin{cases} 0, \text{ se } \hat{y}_n = y_n \\ 1, \text{ se } \hat{y}_n \neq y_n \end{cases}$.

Deste modo, a proporção de casos incorrectamente classificados pode ser calculada através da seguinte expressão: $\frac{\sum_{n=1}^{N} E_n}{N}$. (7)

O Índice de Huberty permite analisar a precisão associada ao modelo de classificação desenvolvido, uma vez que, indica a melhoria de precisão de classificação em relação à precisão de classificação por defeito. Em concreto, a expressão de cálculo do Índice de Huberty é dada por:

$$IH = \frac{P - P^{def}}{1 - P^{def}}, \quad (8)$$

onde,

- P : Proporção de casos de uma amostra correctamente classificados por um modelo;
- P^{def} : Proporção de casos de uma amostra que seriam correctamente classificados se fosse atribuída a todos os casos a categoria com maior frequência.

De forma a evitar a memorização dos dados e o sobre-ajustamento do modelo à amostra de teste, é possível impor regras de paragem no crescimento da Árvore de Decisão, isto é, impor alguns parâmetros de forma a parar o crescimento da árvore. Assim, pode impor-se um número máximo de níveis na Árvore de Decisão e um número mínimo de observações nos nós descendentes. Pode ainda assegurar que existe um valor mínimo para a melhoria de uma função de diversidade. A estas regras de paragem, pode chamar-se o momento da pré-poda, *forward prunning*.

O algoritmo CART propõe ainda um processo de poda após a árvore estar completamente construída, ao que se chama *backward prunning*. Esta poda permite a generalização dos resultados sendo que o objectivo deste processo é o de estabelecer um compromisso entre a obtenção de bons resultados e o de obter um modelo simples, de fácil generalização e interpretação. No caso do CART, este processo obedece a um critério de minimização Custo-Complexidade, o qual consiste na minimização da seguinte medida:

$$re_{\delta}(A) = re(A) + \delta |A_F|, \quad (9)$$

onde,

- $re(A)$: Erro de previsão, o qual corresponderá, em classificação, à proporção de casos incorrectamente classificados ($P_{re}(A)$) e em regressão à variância ($S^2_{re}(A)$);
- δ: Constante real que penaliza a complexidade da árvore;
- $|A_F|$: Número de nós folha da árvore *A*.

Nesta medida entra em consideração o erro de previsão sobre a amostra de treino e a complexidade do modelo proposto, traduzido pelo número de nós folha da árvore considerada. Ainda nesta medida, δ é uma constante que penaliza a complexidade da árvore, pois quando esta aumenta, a árvore que minimiza o $re_{\delta}(A)$ terá cada vez menos nós folha. Assim, no método CART, há um aumento sucessivo do valor de δ, dando origem a árvores com o mesmo nó raiz de A, em que a respectiva complexidade vai diminuindo. Uma das árvores geradas durante este processo será seleccionada como modelo final.

Deste modo, em cada passo do processo de poda, o CART procura identificar sucessivos elos mais fracos (os quais, correspondem aos nós cujos descendentes irão ser podados). Em concreto, para cada nó *O* não terminal, calcula-se o valor de δ a partir do qual faz sentido podar a subárvore descendente de *O*, sendo este dado por:

$$\delta(O) = \frac{re(O) - re(A_O)}{|A_O| - 1}, \quad (10)$$

onde,

- $re(O)$: Erro de previsão no nó *O*;
- $re(A_O)$: Erro de previsão na subárvore descendente do nó *O*;
- $|A_O|$: Número de nós folha na subárvore descendente do nó *O*.

A poda será efectuada no nó O com subárvore descendente A_O, ao qual está associado o valor mínimo de $\delta(O)$, o que originará uma nova árvore.

Por fim, a selecção da árvore podada incide na interpretação das estimativas do erro apropriadas, isto é, pode basear-se no erro mínimo avaliado sobre a amostra de teste. Breiman *et al.* (1984) propõem ainda, para melhorar este processo de escolha, que se seleccione a árvore podada mais simples (i.e., com menos nós folha) com o erro associado não superior ao erro mínimo obtido acrescido de um erro padrão. Em concreto, definindo o erro mínimo por:

$$re_{\min} = \min_k \left\{ re(A^{(k)}) \right\}, \quad (11)$$

a árvore seleccionada, que se designa por $A^{(k*)}$, deverá respeitar a seguinte desigualdade:

$$re(A^{(k^*)}) \leq re_{min} + SE(re_{min}), \quad (12)$$

onde,

- $A^{(k)}$: k-ésima árvore obtida pelo processo de poda;
- $SE(re_{min})$: erro padrão associado à medida de erro.

As previsões são realizadas através do modelo de previsão resultante da árvore podada seleccionada.

Relativamente aos procedimentos para se lidar com os dados omissos, o algoritmo CART propõe que se um caso tiver uma observação omissa da variável seleccionada para a ramificação de um nó, essa variável é substituída pela candidata seguinte, ou seja, por aquela que tiver um maior grau de semelhança com a partição adoptada (na qual não existem casos omissos). A este método adoptado pelo CART denomina-se de *Surrogate Spliting* (ramificação substituta). A ramificação substituta é sempre aquela que, associado à sua semelhança com a ramificação seleccionada, possibilita um maior decréscimo de diversidade na árvore.

Os autores do algoritmo CART propõem ainda, como apoio à interpretação de uma Árvore de Decisão, uma medida de importância (M) das variáveis explicativas X_j´s usadas na construção da árvore.

Esta medida de importância baseia-se na redução da diversidade proporcionada pelo uso da variável X_j em cada ramificação ou pelo seu potencial uso traduzido no conceito de ramificação substituta. A medida de importância relativa à variável X_j é definida por:

$$M(X_j) = \sum_{O \in A} z^{jO} \, \Delta I(\Pi^j, O; A), \quad (13)$$

onde,

- Π^j : Ramificação no nó O que se associa à variável X_j;
- Π^r : Ramificação seleccionada no nó O;
- Π^s : Ramificação substituta no nó O;
- $z^{jO} = \begin{cases} 1, \Pi^j \in \Pi^s \cup \Pi^r \\ 0, \Pi^j \notin \Pi^s \cup \Pi^r \end{cases}$.

Quanto maior o valor da medida de importância da variável X_j, maior será o seu poder explicativo da variável alvo.

Com esta medida é possível conhecer a contribuição potencial dessa variável para a previsão pois, embora uma variável possa não aparecer na árvore final como responsável pela ramificação, a sua medida de importância pode ser elevada.

3.2. Algoritmo *Backpropagation*

O Algoritmo *Backpropagation* é um algoritmo de treino/ aprendizagem de modelos de Redes Neuronais Artificiais e foi pela primeira vez proposto por Paul Werbos em 1970, sendo no entanto só a partir de 1986 com Rumelhart e McClelland que se tornou mais popular. Os modelos de Redes Neuronais, à semelhança de outros utilizados em *Data Mining*, funcionam através de métodos de treino/ aprendizagem para a resolução de problemas de previsão e de classificação.

Assim, o Algoritmo *Backpropagation* é um conjunto de instruções que permitem levar a cabo a tarefa de optimização do desempenho na resolução destes tipos de problemas (classificação e regressão), fazendo-o através da experiência. Partindo da definição de uma tarefa inicial, através de um processo de tentativa e erro com mudanças em alguns parâmetros a definir pelo investigador, o algoritmo permitirá chegar a uma solução para o problema em estudo, a qual

garante a minimização do erro de previsão ou classificação, através do ajustamento dos pesos associados às ligações entre os neurónios artificiais.

O algoritmo *Backpropagation* é utilizado em aprendizagem supervisionada, isto é, nos casos em que os valores da variável alvo na amostra de treino são conhecidos, sendo aplicável em redes com múltiplas camadas de tipo *feed-forward* (camada(s) que apenas recebe(m) como input o output de neurónio(s) da camada anterior).

Para optimizar o desempenho do algoritmo *Backpropagation* existem vários métodos para minimizar o erro das estimativas obtidas, sendo o Método do Gradiente Descendente um dos mais conhecidos. Este método permite a identificação da direcção das mudanças a efectuar nos pesos para que a aprendizagem caminhe no sentido da minimização do erro, sendo por isso mais indicado para redes de grande dimensão. Pelo contrário, o algoritmo de Levenberg-Marquardt, que tem como objectivo a minimização do erro quadrático e é bastante eficiente, deve ser aplicado em redes de pequena dimensão, uma vez que não indica a direcção do ajustamento a efectuar nos pesos.

Existem outros algoritmos que utilizam o Método do Gradiente Descendente como o Algoritmo QuickProp e RPROP que são considerados melhores para aproximar relações funcionais, mas são menos utilizados que o *Backpropagation*.

Dois dos obstáculos à utilização do *Backpropagation* são o tempo e a capacidade computacional necessários para a resolução de problemas complexos, o que levou à introdução de um parâmetro adicional no algoritmo: o *Momentum*. Esta nova versão do algoritmo passou então a chamar-se *Backpropagation with Momentum* e permite optimizar o processo de aprendizagem tornando-o mais rápido, pois evita que o algoritmo fique estanque em mínimos locais tomando-os como sendo os mínimos globais da função.

O processo de aprendizagem com o algoritmo *Backpropagation* divide-se em duas etapas:

- *Forward*: a partir de pesos gerados aleatoriamente pelo investigador (parâmetros *w*'s que devem variar entre 0 e 1), para as ligações desde as variáveis preditivas X_j´s no sentido entrada-saída da rede, o algoritmo considera-os fixos e calcula, para cada observação, o valor de $\hat{y}$ (output estimado) de acordo com a expressão do neurónio de saída. Este valor estimado é comparado com o resultado verdadeiro (*y* - valor conhecido da variável alvo) e obtém-se o erro $(\hat{y} - y)$.

- *Backward*: Com base no erro calculado, o algoritmo aprende e regressa ao início da rede tentando ajustar os pesos *w*'s (tendo em conta as suas variações Δw's) em cada ligação de forma a iniciar novamente o processo. O ajustamento é efectuado recorrendo ao método, já referido anteriormente, do Gradiente Descendente, isto é, utilizando a informação de qual deve ser a direcção do ajustamento dos valores dos pesos *w*'s de forma a minimizar o erro em cada neurónio.

Em concreto, no método do Gradiente Descendente, na fase *Backward*, o ajustamento é efectuado através da Regra Delta, na qual a variação associada ao peso w_{mp} é definida por:

$$\Delta w_{mp} = -\eta_r \frac{\partial E_n}{\partial w_{mp}}, \quad (14)$$

onde,

- m: Índice de número para as variáveis auxiliares (associadas aos neurónios das camadas intermédias);
- p: Índice de número para as variáveis explicativas;
- w_{mp}: Peso associado à ligação entre a p-ésima variável explicativa e a m-ésima variável auxiliar;
- η_r: Taxa de aprendizagem associada ao r-ésimo ciclo do algoritmo.

Apenas o parâmetro η_r - taxa de aprendizagem é definido pelo investigador na versão original do algoritmo (sem *Momentum*). A taxa de aprendizagem regula a velocidade de aprendizagem, isto é, o valor dos ajustamentos em cada ciclo *r*. Normalmente utiliza-se o valor 0,1. O sinal negativo que surge na fórmula indica a direcção do ajustamento a efectuar em cada ligação no sentido de minimizar o erro.

Em redes com camadas intermédias, dada a possibilidade de serem encontrados vários mínimos locais que não correspondam a mínimos globais, pode-se definir o parâmetro *Momentum* (α é um valor entre [0,1[) de modo a que em cada ciclo a variação em *w* tenha em consideração o ajustamento efectuado na iteração anterior. A inclusão deste parâmetro no ajustamento designa-se por Regra Delta Generalizada, na qual a variação associada ao peso w_{mp} na iteração *n* (do ciclo *r*) é definida por:

$$\Delta w_{mp}^{(n)} = -\eta_r \frac{\partial E_n}{\partial w_{mp}^{(n)}} + \alpha \Delta w_{mp}^{(n-1)} \ , \quad (15)$$

onde,

- $\Delta w_{mp}^{(n-1)}$: ajustamento em w_{mp} efectuado na iteração anterior (i.e., na (n–1) ésima iteração).

Os pesos são actualizados em cada observação até completar um ciclo. Considera-se que um ciclo está terminado quando as duas etapas *Forward* e *Backward* são efectuadas à totalidade das observações da amostra de treino. O número de ciclos a efectuar pode ser definido pelo investigador no sentido de evitar sobreajustamento dos dados e também limitar o tempo de computação.

O algoritmo *Backpropagation* pode ser aplicado a problemas de predição (regressão) e de classificação. Os problemas de classificação têm como alvo uma variável nominal, correspondendo os neurónios de saída a cada uma das categorias da variável alvo. O algoritmo nestes casos terá em conta na fase *backward* a função de activação (logística ou sigmóide) que permite obter as estimativas da probabilidade de pertença de cada observação a cada uma das categorias alvo, assim como uma função do erro adequada à variável nominal.

4. CARACTERIZAÇÃO DA BASE DE DADOS UTILIZADA

4.1 Identificação da Base de Dados utilizada

Para a aplicação prática da teoria atrás enunciada recorreu-se a uma Base de Dados real proveniente de um *site* de um Clube de Fidelização de uma marca de Grande Consumo. Esta base de dados é constituída por duas tabelas principais:

- Tabela de Utilizadores Registados;
- Tabela de *logs*, isto é, interações dos utilizadores com o *site*, e-mails ou newsletters gerados pela empresa.

Para a realização deste estudo foram seleccionados os utilizadores registados com data de registo entre 1/1/2007 e 2/11/2009.

4.2 Caracterização da Base de Dados

Originalmente, as variáveis que constituíam a Base de Dados eram as que se encontram sistematizadas na Com base nestas variáveis foram criadas novas variáveis (ver Anexo II), cuja análise é efectuada no ponto seguinte.

Tendo por base a selecção dos utilizadores com data de registo entre 1/1/2007 a 2/11/2009, conforme enunciado anteriormente, a dimensão inicial da Base de Dados era de 13.399 utilizadores registados (número de registos na Tabela de utilizadores registados) a que correspondiam 171.175 *logs* (número de registos na Tabela de *Logs*).

Todos os utilizadores registados foram classificados por Distrito e Região. Em 392 utilizadores não havia informação para esta classificação e dado que os mesmos apresentavam outras características pouco credíveis, como respostas iguais, indiciando terem sido utilizados para testes, esses utilizadores e os respectivos logs foram eliminados.

Assim, o universo do estudo é constituído por 13.007 Utilizadores Registados e 169.376 *logs*. Tabela 4.

Com base nestas variáveis foram criadas novas variáveis (ver Anexo II), cuja análise é efectuada no ponto seguinte.

Tendo por base a selecção dos utilizadores com data de registo entre 1/1/2007 a 2/11/2009, conforme enunciado anteriormente, a dimensão inicial da Base de Dados era de 13.399 utilizadores registados (número de registos na Tabela de utilizadores registados) a que correspondiam 171.175 *logs* (número de registos na Tabela de *Logs*).

Todos os utilizadores registados foram classificados por Distrito e Região. Em 392 utilizadores não havia informação para esta classificação e dado que os mesmos apresentavam outras características pouco credíveis, como respostas iguais, indiciando terem sido utilizados para testes, esses utilizadores e os respectivos logs foram eliminados.

Assim, o universo do estudo é constituído por 13.007 Utilizadores Registados e 169.376 *logs*.

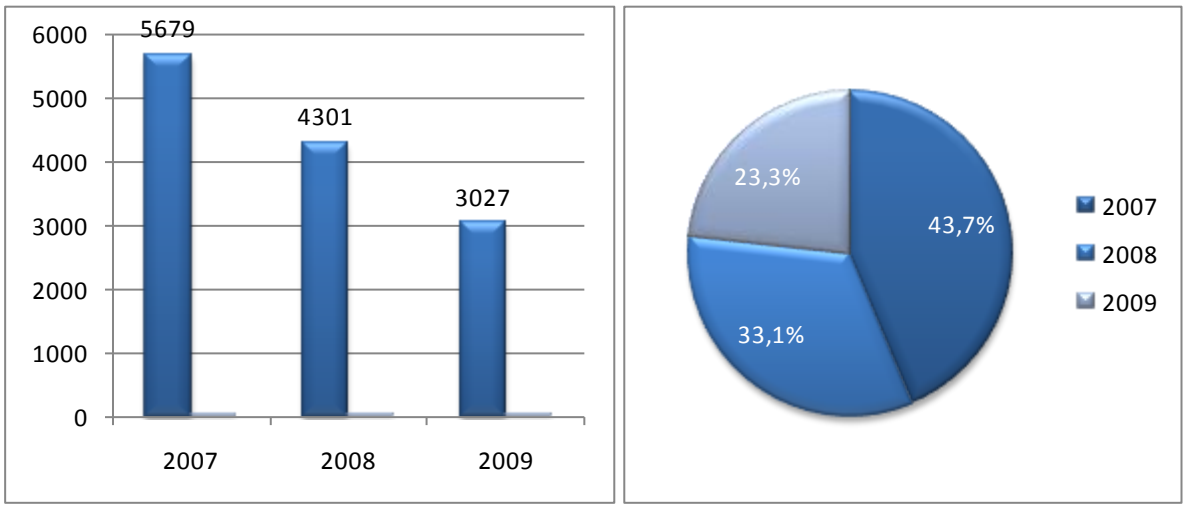
Tabela 4 – Variáveis originais da Base de Dados

Tabela de Utilizadores Registados		Tabela de *Logs* (interacções)	
Nome da variável	**Descrição**	**Nome da variável**	**Descrição**
id_User	Código único que permite identificar cada utilizador registado e estabelece a relação entre as duas Tabelas	**Data**	Data e hora da interacção
cp4	Código Postal de 4 Dígitos	**id_user**	Código único que permite identificar cada utilizador registado e estabelece a relação entre as duas Tabelas
cp3	Código Postal de 3 Dígitos	**pag**	Nome da página, link ou área do site activada em cada interacção
data_nasc	Data de Nascimento do utilizador		
foto_home	Se tem ou não fotografia		
localidade	Localidade de residência		
sexo	Sexo		
id_orig	Código identificador da origem de entrada na Base de Dados		
data_reg	Data de Registo na Base de Dados		
data_act	Data de Actualização dos Dados		
aut_tel	Autorização de comunicação via telefone		
aut_email	Autorização de comunicação via e-mail		
aut_postal	Autorização de comunicação via postal		
aut_sms	Autorização de comunicação via sms		
dt_activacao	Data de Activação após Registo		

4.2.1. Caracterização do registo

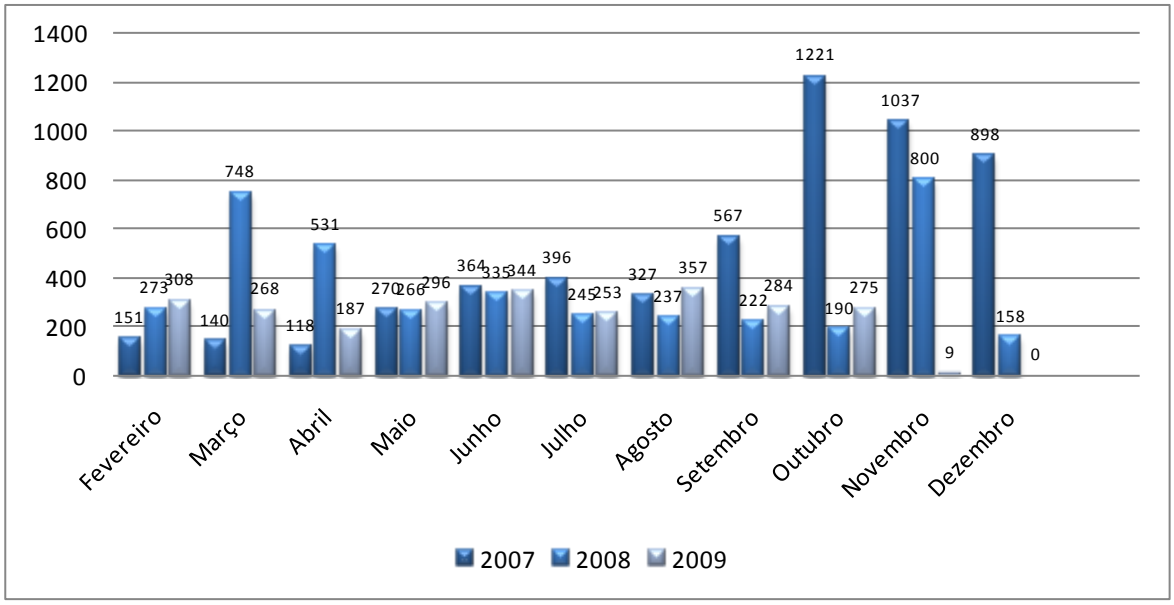
Os utilizadores registados que constam na base de dados distribuem-se de forma diferenciada por ano de registo e revelam tendência decrescente entre 2007 e 2009, variando entre 5679 em 2007 e 3027 em 2009, conforme pode ser observado na Figura 3.

Figura 3 – Número de utilizadores registados



Dado que o número de utilizadores registados na BD no ano de 2009 apenas contém os registos efectuados até 2 de Novembro e de Maio a Outubro o número de utilizadores registados em 2009 foi sempre superior ao período homólogo do ano anterior (ver Figura 4), é expectável que no total do ano 2009 se tenha conseguido atingir valores muito próximos de 2008, mas claramente existe tendência decrescente face a 2007.

Figura 4 – Número de registos



Na Figura 5 é possível constatar que apenas cerca de 70% dos utilizadores registados completaram este processo realizando a activação no *site*, atingindo o valor máximo entre os registados em 2008 (76,1%). A etapa correspondente à activação não constitui requisito para a visita ao *site* mas permite confirmar validade do endereço de e-mail e é também indiciador de maior envolvimento com o *site*, pois os indicadores de visita são superiores entre os utilizadores activos, conforme é analisado na secção 4.2.3. (na qual se efectua a Caracterização do perfil de utilização). Conforme se pode observar na Tabela 5, dada a importância da activação, foi feito um esforço da gestão do *site* no sentido de diminuir o tempo entre o registo e a activação, tendo o número de dias entre estas duas etapas, reduzido, em média, de 51,94 dias em 2007 para 10,15 em 2009. O prazo que ocorre com maior frequência é zero dias, isto é, a activação ocorre no dia do registo. Informa-se ainda que em 2007, 51,8% dos utilizadores activou o registo no próprio dia, enquanto em 2009, esta percentagem foi de 65,8%.

Figura 5 – Activação no *site* após registo

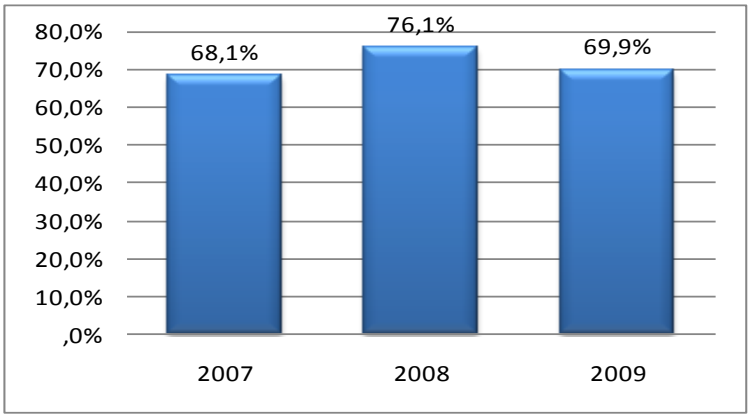


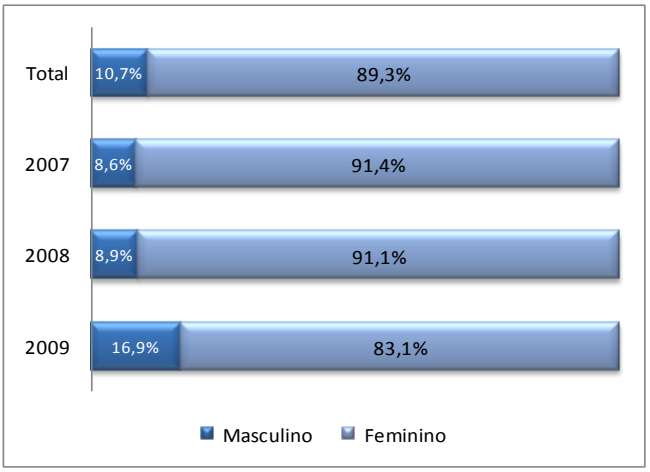
Tabela 5 – Nº de dias entre registo e activação

	Ano de Registo		
	2007	2008	2009
Média	51,94	35,35	10,15
Mínimo	0,00	0,00	0,00
Mediana	0,00	1,00	0,00
Máximo	844,00	636,00	293,00
Moda	0,00	0,00	0,00
Desvio Padrão	123,37	90,02	32,45

4.2.2. Caracterização dos utilizadores registados

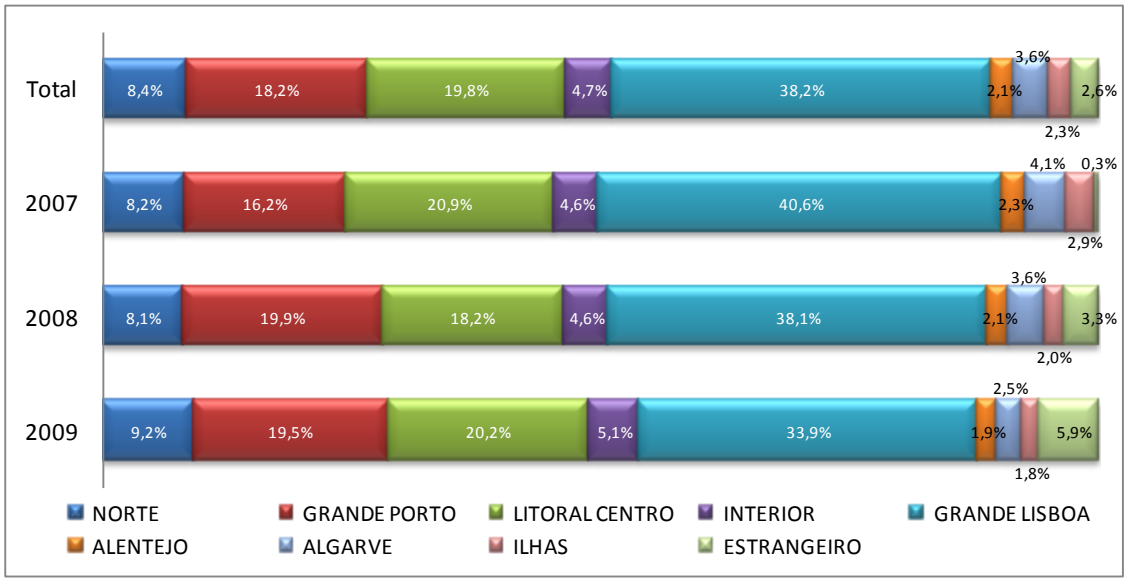
No que respeita à caracterização demográfica dos utilizadores registados neste *site*, constata-se, a partir da Figura 6, que a grande maioria dos utilizadores são do sexo feminino (89,3%) e que em 2009 a percentagem de utilizadores do sexo Masculino é praticamente duplicada (16,9% em 2009 face a 8,9% em 2008).

Figura 6 – Género sexual dos utilizadores



De acordo com a Figura 7, no total de utilizadores registados, Grande Lisboa (38,2%), Litoral Centro (19,8%) e Grande Porto (18,2%) são as regiões com maior importância no que respeita à proveniência dos utilizadores. Analisando a respectiva evolução por ano, constata-se diminuição do peso da Grande Lisboa (40,6% em 2007 e 33,9% em 2009), por contraponto ao aumento dos utilizadores do estrangeiro (0,3% em 2007 e 5,9% em 2009). De realçar que a maioria dos utilizadores do estrangeiro são provenientes do Brasil.

Figura 7 – Perfil geográfico dos utilizadores

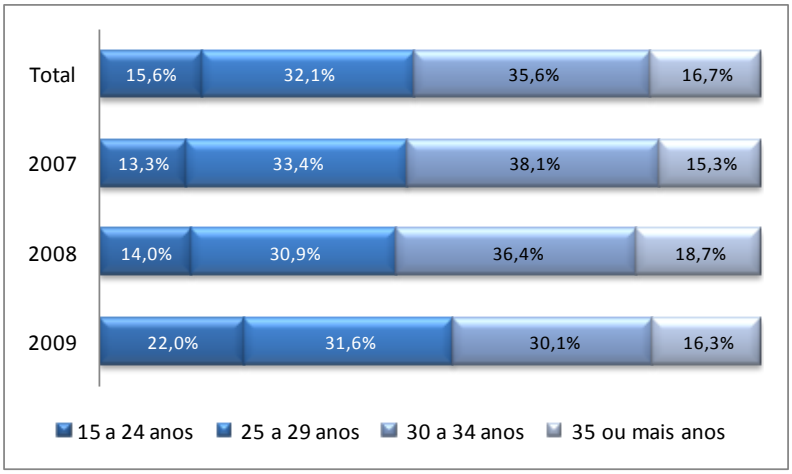


No que respeita ao perfil etário dos utilizadores, constata-se que a idade média é de 30 anos em 2007 e em 2008, sendo ligeiramente mais baixa em 2009 (29 anos). Com base na Figura 8, conclui-se que os utilizadores do *site* são bastante jovens, dado que 15,6% têm menos de 25

anos e 16,7% têm 35 ou mais anos. Conclui-se assim que dois terços dos utilizadores têm idade compreendida entre 25 e 34 anos.

Considerando a mesma figura, refere-se ainda que os utilizadores registados em 2009 apresentam estrutura etária ainda mais jovem, assinalando-se 22,0% dos utilizadores com menos de 25 anos.

Figura 8 – Estrutura etária dos utilizadores



Esta evolução da estrutura etária é transversal a todas as regiões, não sendo por isso só justificada pelo aumento do número de registados do estrangeiro.

4.2.3. Caracterização do perfil de utilização

Para caracterizar o perfil de utilização dos utilizadores registados deste *site*, foi necessário criar novas variáveis a partir da tabela de registo de *logs*. Assim, o primeiro passo foi classificar os *logs* em E-mail e em *Site*. Os primeiros dizem respeito a *logs* em que o utilizador leu um E-mail que lhe foi enviado pelo *site* (ex: E-mail de Aniversário). Os *logs Site* correspondem a todos os outros e significam que ocorreu uma interacção do utilizador com o *site*. Com esta divisão inicial conclui-se, com base na Figura 9, que 5.103 utilizadores (39,2%) nunca visitaram o *site*, sendo assim classificados de Não Visitantes. De realçar que 32% dos utilizadores não tiveram mesmo qualquer interacção com o *site*, isto é, não leram o E-mail e não acederam ao *site*.

Figura 9 – Status de Visita ao *Site*

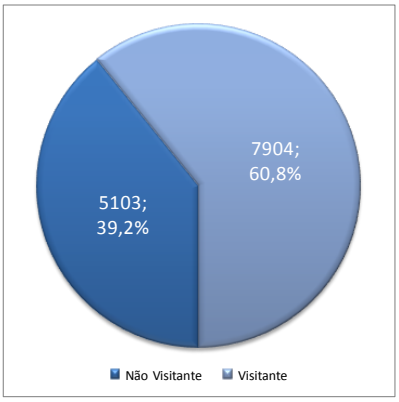
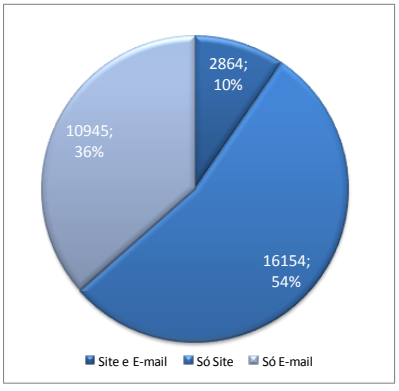


Figura 10 – Tipo de sessão



Após esta classificação inicial, todos os *logs* foram agrupados em sessões tendo-se considerado como critério

para definição de uma sessão um ou mais *logs* referentes ao mesmo utilizador com intervalo de uma ou mais horas entre os próximos *logs* do mesmo utilizador, dado que a partir deste tempo sem realizar alguma acção no *site* era necessário voltar a introduzir as credenciais de acesso (*login*). Assim, foram obtidas 29.963 sessões que foram classificadas de acordo com o tipo de *log* (nomeadamente Site e E-mail, Só Site e Só E-mail), e cuja distribuição pode ser observada na Figura 10. Nesta figura constata-se que a grande maioria das sessões (90%) são de um só tipo e apenas 10% conjugam acesso ao *Site* e ao E-mail. O Tipo de Sessão mais comum é a visita exclusiva ao *Site*, mas 36% das sessões são exclusivamente de leitura de E-mail.

Com esta classificação das sessões obteve-se para cada utilizador em cada um dos anos o número total de sessões (Site e/ou E-mail) e também o número de sessões Só Site. Para que esta análise fosse mais comparável entre anos de registo, transferiu-se a informação de cada ano relativamente ao ano de registo, considerando-se para os primeiros 365 dias após o registo o ano em que ocorreu o registo (Ano 0) e, sempre que possível, o 1º e o 2º Ano após o registo (Ano 1 e Ano 2, respectivamente).

Em função da data de registo alguns utilizadores não completaram um dos anos e não têm actividade no ano seguinte. Este facto ocorre com maior impacto nos utilizadores de 2009, havendo assim maior desconhecimento sobre os comportamentos futuros destes utilizadores o que constitui por si só um aspecto interessante a analisar. Assim, como se pode observar na Tabela 6 não existe informação acerca do comportamento no 2º ano após o registo para 100% dos utilizadores de 2009 e de 2008 e ainda para 33,2% dos utilizadores registados em 2007 (registados entre 3/11/2007 e 31/12/2007). A informação acerca do primeiro ano após o registo não é possível de caracterizar para 100% dos registados em 2009 e 22,2% dos registados em 2008. Existe também elevada percentagem de casos em que existem ciclos anuais incompletos, nomeadamente para a totalidade dos utilizadores registados em 2009, o que exerce alguma influência nas comparações entre os vários anos.

Tabela 6 – Status de informação disponível

		Ano de Registo							
		2007		2008		2009		Total	
Ano	Status Informação	Nº de Utilizadores	(%)	Nº de Utilizadores	(%)	Nº de Utilizadores	(%)	Nº de Utilizadores	(%)
Ano 0	Ano Completo	5679	100,0%	3351	77,9%	0	,0%	9030	69,4%
	Ano Incompleto	0	,0%	950	22,1%	3027	100,0%	3977	30,6%
	Sem Informação	0	,0%	0	,0%	0	,0%	0	,0%
Ano 1	Ano Completo	3793	66,8%	0	,0%	0	,0%	3793	29,2%
	Ano Incompleto	1886	33,2%	3346	77,8%	0	,0%	5232	40,2%
	Sem Informação	0	,0%	955	22,2%	3027	100,0%	3982	30,6%
Ano 2	Ano Completo	0	,0%	0	,0%	0	,0%	0	,0%
	Ano Incompleto	3793	66,8%	0	,0%	0	,0%	3793	29,2%
	Sem Informação	1886	33,2%	4301	100,0%	3027	100,0%	9214	70,8%

A Tabela 7 é um exemplo onde a existência de Anos Incompletos tem influência na análise, pois esta diz respeito ao número total de sessões desde o momento do registo até à data de extracção da base. Verifica-se que a percentagem de utilizadores que não tinham efectuado qualquer interacção com *site* ou e-mail é superior nos utilizadores registados em 2009 (54,1% face a 21,7% dos utilizadores registados em 2007).

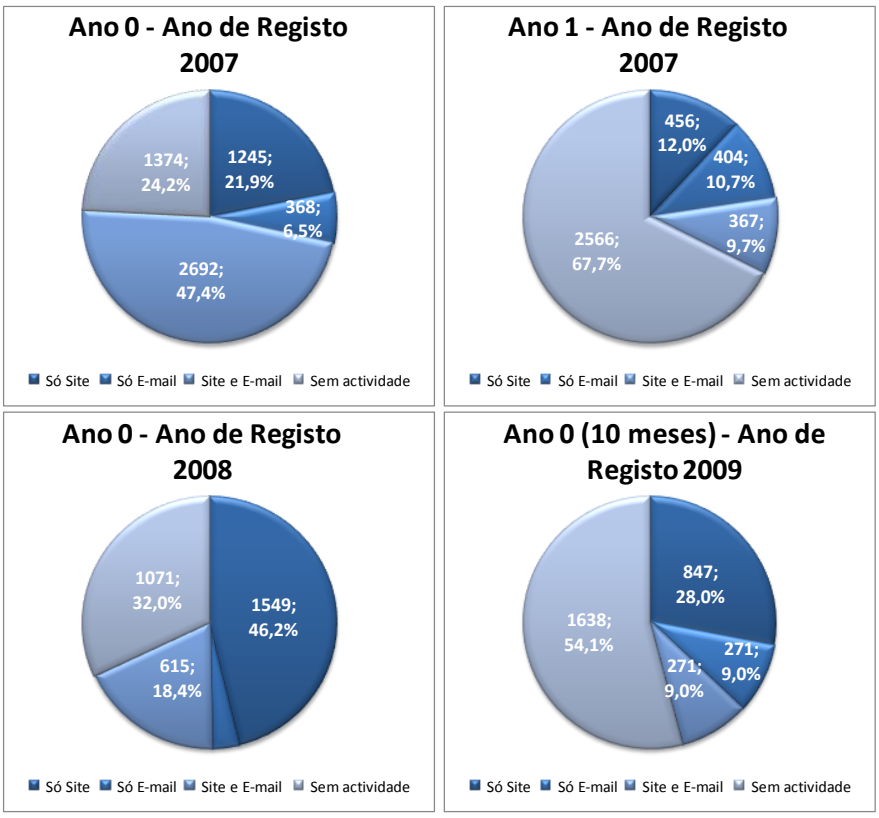
Ainda na Tabela 7 é possível observar que a realização de Sessões Só Site é bastante inferior ao total de sessões. Realça-se que a elevada percentagem de utilizadores registados que nunca visitaram o *site* em exclusivo, isto é, sem serem estimulados através de E-mail. Mesmo entre os utilizadores registados em 2007 mais de 60% efectuou uma ou menos sessões de visita exclusiva ao *site*.

Tabela 7 – Nº de sessões entre 2007 a 2009

		Ano de Registo					
		2007		2008		2009	
		Nº de Utilizadores	(%)	Nº de Utilizadores	(%)	Nº de Utilizadores	(%)
Nº de Sessões Site e/ou Email 2007 a 2009	0	1230	21,7%	1295	30,1%	1638	54,1%
	1	969	17,1%	1169	27,2%	910	30,1%
	2	714	12,6%	767	17,8%	306	10,1%
	3	594	10,5%	439	10,2%	110	3,6%
	4	477	8,4%	273	6,3%	32	1,1%
	5	413	7,3%	145	3,4%	20	,7%
	6	346	6,1%	65	1,5%	6	,2%
	7	244	4,3%	51	1,2%	3	,1%
	Mais de 8	692	12,2%	97	2,3%	2	,1%
	TOTAL	**5679**	**100%**	**4301**	**100%**	**3027**	**100%**
Nº de Sessões SóSite 2007 a 2009	0	1858	32,7%	1614	37,5%	1912	63,2%
	1	1642	28,9%	1404	32,6%	876	28,9%
	2	981	17,3%	694	16,1%	177	5,8%
	3	549	9,7%	286	6,6%	45	1,5%
	4	287	5,1%	139	3,2%	9	,3%
	5	139	2,4%	70	1,6%	5	,2%
	6	74	1,3%	20	,5%	3	,1%
	7	48	,8%	25	,6%	0	,0%
	Mais de 8	101	1,8%	49	1,1%	0	,0%
	TOTAL	**5679**	**100%**	**4301**	**100%**	**3027**	**100%**

Considerando apenas os anos completos pode-se observar na Figura 11 que no Ano 0 em 2008, 46,2% dos utilizadores registados só têm sessões do tipo Só Site, enquanto que nos utilizadores de 2007 este valor foi de apenas 21,9%. Entre os utilizadores de 2007 tem mais peso o número de utilizadores com os dois tipos de sessão, isto é, de Site e de E-mail (47,4% face a 18,4% em 2008). No Ano 1, dos utilizadores registados em 2007 constata-se maior peso de sessões exclusivamente de leitura de E-mail (10,7%) e um aumento muito significativo da percentagem de utilizadores que não tiveram qualquer tipo de actividade (67,7% face a 24,2% no Ano 0). Em 2008, constata-se a diminuição da taxa de actividade, existindo 32% dos utilizadores registados sem qualquer sessão no ano em que ocorreu o registo. Para o ano de 2009, não sendo directamente comparável com os anteriores, constata-se que a taxa de actividade é inferior à verificada em 2008 (45,9% face a 68,0%). Também a taxa de visita exclusiva ao *site* tem menor importância quando comparada com as sessões Só E-mail ou Site e E-mail em 2009 (28% para 18%), sendo que em 2007 este rácio é de 21,9% para 53,9%. Os resultados anteriores permitem levantar a hipótese de a tendência de menor Actividade verificada de 2007 para 2008 seja agravada em 2009 mesmo após o ano completo.

Figura 11 – Tipo de sessão por ano de registo



Atendendo à Tabela 8, é possível afirmar que o Tipo de Sessão no Ano 1 aparenta estar relacionado com o Tipo de Sessão no Ano 0, nomeadamente no que respeita à realização ou não de algum tipo de sessão. Dos 92,1% de utilizadores que não tiveram qualquer sessão no Ano 0, também não realizaram qualquer sessão no Ano 1 e o mesmo acontece com 83,5% dos Utilizadores que só visitaram o Site no Ano 0. Entre os utilizadores que tiveram sessões Só E-

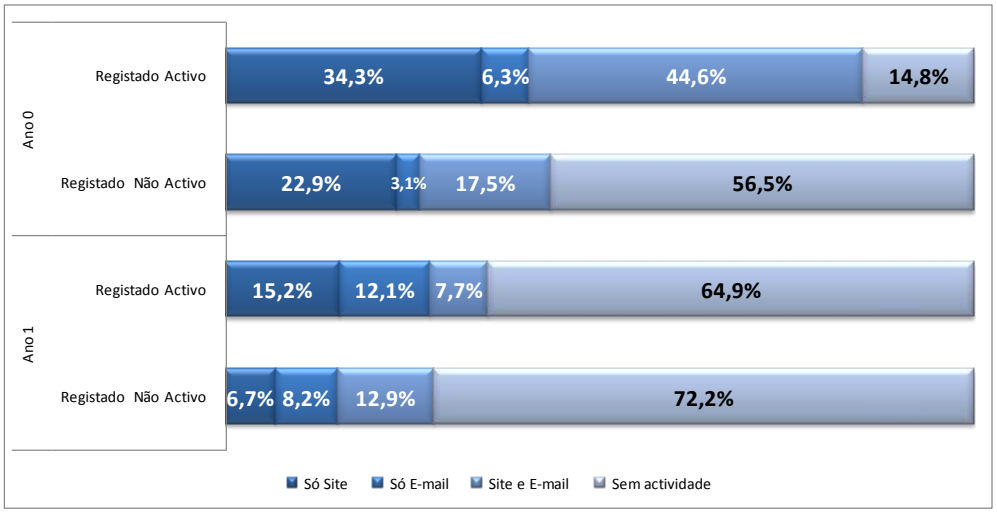
mail ou Site e E-mail ocorre maior actividade no Ano 1, sendo que aproximadamente 50% dos utilizadores que acedem a Site e E-mail no Ano 0, têm algum tipo de sessão no Ano 1.

Tabela 8 – Tipo de sessão Ano 0 vs Ano 1

		Tipo de Sessão no Ano 1			
		Só Site	Só E-mail	Site e E-mail	Sem actividade
Tipo de Sessão no Ano 0	Só Site	11,9%	2,7%	1,9%	83,5%
	Só E-mail	14,4%	12,1%	6,6%	66,9%
	Site e E-mail	15,5%	17,2%	16,9%	50,4%
	Sem actividade	3,8%	2,7%	1,4%	92,1%
	TOTAL	12,0%	10,7%	9,7%	67,7%

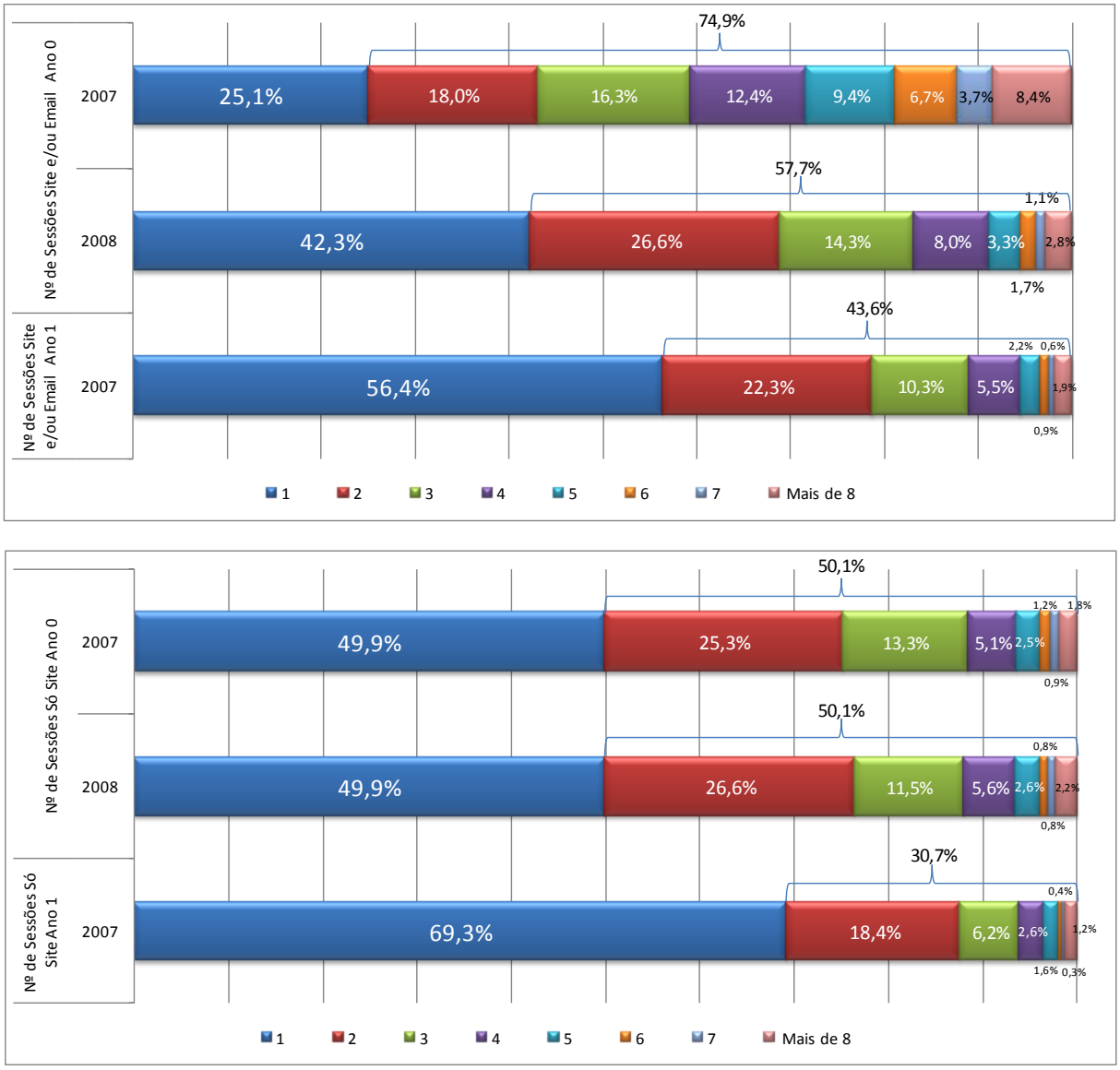
A análise do Tipo de Sessão no Ano 0 e no Ano 1 para os utilizadores registados em 2007 em função do Status de Activação permite concluir que a activação tem elevada importância no acesso a Site e/ou E-mail no ano após o registo (ver Figura 12). De facto, apenas 14,8% dos Registados Activos não tiveram qualquer tipo de actividade no Ano 0, enquanto que entre os Registados Não Activos esta percentagem é de 56,5%. No Ano 1, esta diferença desvanece-se e é menos relevante, dado que, também, entre os Registados Activos a percentagem de utilizadores sem actividade é muito elevada (64,9%).

Figura 12 – Tipo de sessão por status de activação



Na sequência das análises anteriores, conclui-se através dos gráficos apresentados na Figura 13 que em 2008 também ocorreu diminuição da percentagem de utilizadores que teve mais do que uma sessão (de 74,9% para 57,7%). Para as sessões Só Site, as percentagens são bastante similares entre 2007 e 2008, o que confirma as conclusões anteriormente estabelecidas e que permite levantar a hipótese de que terá havido maior envio de E-mails de estímulo em 2007.

Figura 13 – Distribuição do número de sessões por tipo e ano



Na Tabela 9 pode observar-se a evolução do número médio de sessões por trimestre completo. A análise detalhada por trimestre do ano 0 permite constatar que o número médio de sessões Site e/ou E-mail, por trimestre completo decresce sucessivamente de 2007 para 2008 e para 2009. No 1º Trimestre após o registo cada utilizador teve em média 1,12 sessões enquanto que em 2008 a média foi 0,99 e em 2009 apenas 0,52. Nos trimestres seguintes esta tendência é ainda mais acentuada, nomeadamente no 2º trimestre em que passou de 0,66 em 2007 para 0,26 em 2008 e 0,12 em 2009. Em função do que já foi referido anteriormente sobre o facto de em 2008 terem tido maior importância as sessões de Site e menos de E-mail, o número médio de sessões Só Site no 1º Trimestre de 2008 é superior ao verificado em 2007. Comparando os mesmos dados em função do Status de Activação constata-se que o número

médio de sessões é sempre superior entre os utilizadores activos, mas a diminuição verificada no 2º Trimestre ocorre de forma similar em ambos os *status*.

Tabela 9 – Número médio de sessões por trimestre completo

	Ano de Registo			Status de Activação	
	2007	2008	2009	Registado Activo	Registado Não Activo
Site e/ou Email no Ano0 Trim1	1,12	0,99	0,52	1,09	0,51
Site e/ou Email no Ano0 Trim2	0,66	0,26	0,12	0,51	0,26
Site e/ou Email no Ano0 Trim3	0,64	0,12	0,16	0,47	0,19
Site e/ou Email no Ano0 Trim4	0,38	0,10		0,30	0,21
Só Site no Ano0 Trim1	0,74	0,92	0,49	0,86	0,41
Só Site no Ano0 Trim2	0,21	0,15	0,06	0,20	0,05
Só Site no Ano0 Trim3	0,23	0,10	0,03	0,20	0,05
Só Site no Ano0 Trim4	0,18	0,07		0,15	0,06

Analisando o número médio de sessões em função do perfil sócio-demográfico dos utilizadores (ver Tabela 10), verifica-se que a diminuição deste indicador do Ano 0 para o Ano 1 é transversal a todos os estratos. O número médio de sessões mais elevado é verificado entre os utilizadores do sexo Feminino, com idade entre 25 e 39 anos ou das Regiões Grande Porto, Grande Lisboa, Alentejo ou Algarve. Este padrão é similar entre os utilizadores registados de 2007 e de 2008. Os utilizadores cujo número médio de sessões mais diminuiu de 2007 para 2008 são os do sexo Feminino, com idade entre 30 e 39 anos ou residentes no Alentejo, Litoral, Grande Porto ou Algarve. Relativamente aos utilizadores de 2009, ainda que o número médio de sessões por utilizador não possa ser comparável com os anteriores, por corresponder a apenas 10 meses, constata-se que os estratos que apresentam maior intensidade de utilização são similares aos que têm o mesmo comportamento em 2007 e 2008.

Tabela 10 – Nº médio de sessões por tipo e por características demográficas

		Ano de Registo									
		2007				Var. Site e/ou Email (Ano 1 - Ano 0)	Var. Só Site (Ano 1 - Ano 0)	2008		2009	
		Site e/ou Email Ano 0	Só Site Ano 0	Site e/ou Email Ano 1	Só Site Ano 1			Site e/ou Email Ano 0	Só Site Ano 0	Site e/ou Email Ano 0	Só Site Ano 0
Sexo	Masculino	2,16	1,04	0,52	0,28	-1,64	-0,77	1,34	1,12	0,39	0,25
	Feminino	2,88	1,39	0,65	0,33	-2,23	-1,06	1,64	1,36	0,78	0,52
Idade	15 a 19 anos	2,24	1,18	0,27	0,13	-1,97	-1,05	0,90	0,74	0,49	0,41
	20 a 24 anos	2,60	1,39	0,55	0,34	-2,05	-1,05	1,31	1,10	0,60	0,42
	25 a 29 anos	2,75	1,34	0,65	0,33	-2,10	-1,01	1,71	1,43	0,75	0,51
	30 a 34 anos	2,92	1,39	0,69	0,35	-2,23	-1,05	1,74	1,40	0,81	0,53
	35 a 39 anos	2,92	1,34	0,61	0,30	-2,31	-1,05	1,67	1,43	0,69	0,44
	40 ou mais anos	2,66	1,08	0,54	0,18	-2,12	-0,90	1,22	1,05	0,50	0,30
Região	Norte	2,74	1,28	0,66	0,36	-2,08	-0,92	1,68	1,42	0,59	0,40
	Grande Porto	2,89	1,40	0,66	0,37	-2,22	-1,03	1,67	1,38	0,78	0,49
	Litoral	2,78	1,39	0,53	0,28	-2,25	-1,12	1,58	1,30	0,71	0,47
	Interior	2,70	1,25	0,58	0,26	-2,12	-0,99	1,66	1,37	0,57	0,39
	Grande Lisboa	2,81	1,34	0,69	0,34	-2,12	-1,00	1,73	1,43	0,81	0,55
	Alentejo	2,85	1,46	0,56	0,32	-2,29	-1,14	1,43	1,20	0,59	0,42
	Algarve	2,88	1,40	0,70	0,35	-2,18	-1,05	1,64	1,37	0,61	0,42
	Ilhas	2,45	1,14	0,44	0,22	-2,01	-0,91	1,11	0,95	0,89	0,64
	Estrangeiro	1,12	0,65	0,63	0,25	-0,49	-0,40	0,43	0,42	0,25	0,21

Atendendo à Tabela 11, a evolução do número médio de *logs* por sessão apresenta um comportamento similar com o verificado anteriormente, constatando-se o aumento do número de *logs* de 2007 a 2009. Este aumento decorre da maior percentagem de sessões com visita ao Site em detrimento das sessões de E-mail (contam apenas como um *log*). Nos anos em que é possível comparação entre o Ano 0 e os seguintes constata-se diminuição sucessiva do número médio de *logs* por sessão. À semelhança do verificado no número médio de sessões, também no número médio de *logs* por sessão se constata enorme diferença entre os utilizadores Activos e Não Activos no Ano 0 (6,06 face a 2,48, respectivamente), mantendo-se nos anos seguintes mas já de forma menos expressiva.

Tabela 11 – Número médio de *logs* por sessão

	Ano de Registo			Status de Activação	
	2007	2008	2009	Activo	Não Activo
Ano0	4,34	5,45	7,53	6,06	2,48
Ano1	1,51	1,92	.	1,74	1,16
Ano2	0,35	.	.	0,39	0,29

Em suma, todos os resultados apresentados permitem concluir que o principal desafio deste *site*, à semelhança do que ocorre neste tipo de *sites*, é conseguir gerar motivos de interesse que consigam captar o retorno dos utilizadores ao *site* após o registo inicial e ao longo do

tempo, pelo que se torna por demais evidente a necessidade de desenvolver estratégias de comunicação que vão de encontro às necessidades dos consumidores/utilizadores.

4.2.4. Segmentação dos consumidores/utilizadores

Para servir de suporte às estratégias de comunicação *online* e *offline* junto dos utilizadores registados do *site* pretende-se desenvolver um Modelo de Classificação que permita, a partir de uma segmentação inicial em *Clusters* com base nos comportamentos de visita dos utilizadores de 2007 nos Anos 0, 1 e 2, classificar os utilizadores de 2008 e 2009 e adaptar as estratégias de comunicação *online* e *offline* dirigidas a estes utilizadores.

Assim, foram seleccionados para serem segmentados em *Clusters* 3808 dos 5679 utilizadores registados em 2007, que correspondem aos utilizadores sobre os quais, em função da data de registo, existe informação sobre comportamento de utilização no Ano 1 e no Ano 2.

Como metodologia de segmentação utilizou-se o método de *Clustering Two-Step*, desenvolvido por Chiu, Fang, Chen, Wang e Jeris (2001). A metodologia *Two-Step* é um algoritmo que actua em duas fases ou passos, como o seu nome indica. No 1º passo os casos são sumarizados em muitos pré-*clusters* e no 2º passo os pré-*clusters* são reagrupados no número de clusters desejado ou definido pelo algoritmo automaticamente.

Analisando em detalhe o 1º passo de sumarização, constata-se que é utilizada uma abordagem sequencial, em que cada caso é processado individualmente sendo decidido pelo algoritmo se cada novo caso deve ser junto com algum pré-*cluster* existente ou se, por outro lado, deve ser criado um novo pré-*cluster*.

O algoritmo funciona assim pela construção de uma árvore de *Cluster Features entries* (CF) que consiste numa árvore com vários níveis de nós e em que os elementos nos nós são subconjuntos de observações com características semelhantes. Cada *Cluster Feature Entry* sumariza informação sobre o subgrupo em que está inserida:

- Número de elementos do subgrupo;

- Média e variância para cada variável contínua;

- Frequência associada a cada categoria de cada variável discreta.

Assim, no processamento de cada nova observação esta é dirigida por cada nó até ao nó folha correcto, tendo sempre como regra os parâmetros definidos pelo utilizador no que respeita a:

- *B* – Número de nós descendentes numa ramificação;

- *D* – Profundidade máxima da árvore (número de níveis abaixo do nó raíz);

- *T* – distância de fusão de dois subgrupos, isto é, a distância a partir da qual um novo caso não integra um subgrupo já existente.

Quando chega a um nó folha, a observação procura a observação com informação mais similar. Se a observação estiver dentro da distância de fusão definida, a observação é absorvida para o nó folha e actualiza a informação sumarizada na correspondente *CF entry*. Se não houver possibilidade desta observação ser integrada nesse nó folha então é criado um novo nó folha e o nó folha original divide-se em dois. Caso a árvore cresça para além do limite máximo definido a árvore é gerada novamente com base na anterior aumentando a distância de fusão.

Para efectuar este 1º passo do algoritmo a ordem dos dados deverá ser aleatória, pois a ordem influencia os resultados neste passo.

O 2º passo do algoritmo é efectuado após terminado o 1º passo e agrupa os sub-*clusters* resultantes da etapa anterior em grupos até ao número máximo de clusters pedido pelo utilizador (ou estipulado pelo utilizador). Na fase de agrupamento, o algoritmo utiliza um método hierárquico e, dado que os resultados foram pré-agrupados durante o 1º passo, obtém desempenhos bastante satisfatórios e de fácil computação.

O processo hierárquico de agrupamento consiste na junção recursiva dos pré-*clusters* até ao ponto máximo em que um *cluster* contém a informação de todas as observações. Inicialmente, esta fase considera um *cluster* para cada sub-*cluster* produzido no passo anterior. De seguida, os grupos são comparados e os pares de *clusters* com a menor distância calculada pela medida de logaritmo da verosimilhança são seleccionados e agrupados num só *cluster*. O processo repete-se assim até se obter um só *cluster*.

Desta forma, são facilmente comparáveis soluções com diferentes números de *clusters* utilizando-se, para o efeito, os critérios BIC (*Bayesian Information Criterion*), AIC (*Akaike's*

Information Criterion), bem como a comparação entre os valores de rácio das medidas de distância entre as soluções com menor BIC e/ou AIC.

O algoritmo *Two-Step* também permite a aplicação do modelo a novos casos. Nessa situação. a cada novo caso é atribuído um *score* e esse *score* é comparado com os *scores* do *cluster*. Cada caso é assim agregado ao *cluster* que se encontra mais próximo. Se a distância a qualquer um dos *clusters* for superior ao limite estabelecido esse novo caso é considerado um *outlier* e não é atribuído a nenhum *cluster*.

Do exposto anteriormente conclui-se que o algoritmo *Two-step* tem como principal vantagem a capacidade de lidar com grandes quantidades de informação essencialmente pela forma como no primeiro passo a informação de várias observações são sumarizadas obtendo-se assim diminuição relevante dos recursos computacionais necessários. Pelo contrário, uma grande desvantagem deste método reside no impacto que a ordenação dos dados tem na segmentação final obtida, pois esta grandes alterações em função de ordenações diferentes (Jesus e Cardoso, 2007).

Dado que nesta tese o objectivo da segmentação é agrupar os utilizadores com base nos seus comportamentos de visita nos Anos 0, 1 e 2 de modo a aplicar um Modelo de Classificação que utilize apenas variáveis cuja informação seja conhecida no Ano 0, foram seleccionadas sobretudo variáveis do Ano 0, bem como as variáveis mais estruturantes do comportamento nos Anos 1 e 2, as quais são apresentadas na Tabela 12.

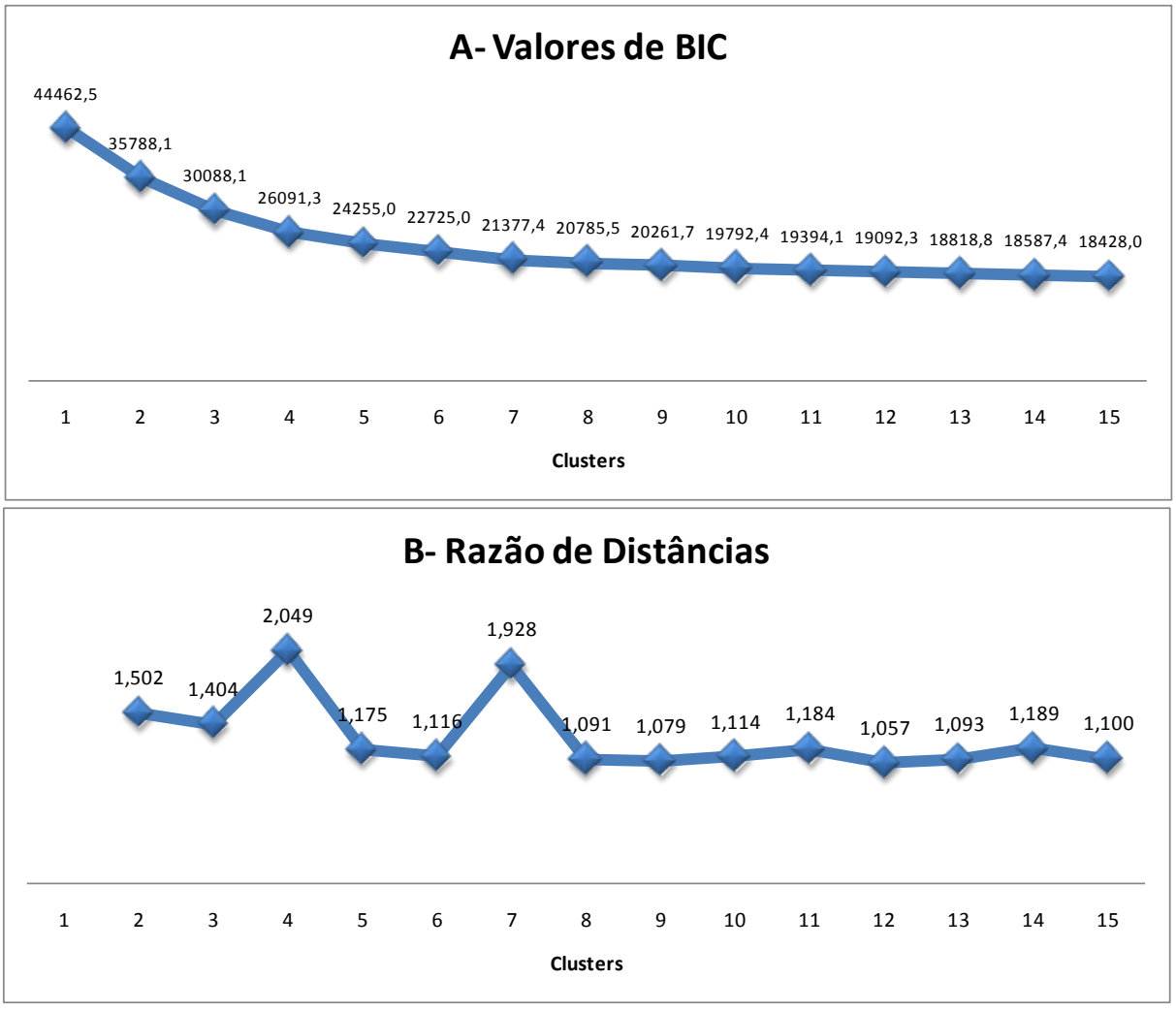
Tabela 12 – Variáveis incluídas na segmentação de utilizadores

Ano a que se refere	Nome da variável	Descrição
Ano 0	Activ_Ano0	Tipo de Páginas Visitadas no Ano 0
	paginas_por_sessao_Ano0	Número médio de páginas por sessão no Ano0
	N_de_sessoes_Ano0_Trim1	Número médio de sessões Site e/ou Email no Ano0 Trim1
	N_de_sessoes_Ano0_Trim2	Número médio de sessões Site e/ou Email no Ano0 Trim2
	N_de_sessoes_Ano0_Trim3	Número médio de sessões Site e/ou Email no Ano0 Trim3
	N_de_sessoes_Ano0_Trim4	Número médio de sessões Site e/ou Email no Ano0 Trim4
	N_de_sessoes_SóSite_Ano0_Trim1	Número médio de sessões Só Site no Ano0 Trim1
	N_de_sessoes_SóSite_Ano0_Trim2	Número médio de sessões Só Site no Ano0 Trim2
	N_de_sessoes_SóSite_Ano0_Trim3	Número médio de sessões Só Site no Ano0 Trim3
	N_de_sessoes_SóSite_Ano0_Trim4	Número médio de sessões Só Site no Ano0 Trim4
Ano 1	Activ_Ano1	Tipo de Páginas Visitadas no Ano 1
Ano 2	Activ_Ano2_2	Tipo de Páginas Visitadas no Ano 2

Segundo Hair *et al*. (2006) "a existência de multicolinearidade actua como um processo de atribuição de pesos" às dimensões em estudo. Neste caso, a utilização de mais variáveis do Ano 0 teve como objectivo atribuir maior peso na definição dos grupos às variáveis desse ano, pois são estas que serão utilizadas como variáveis independentes no Modelo de Classificação.

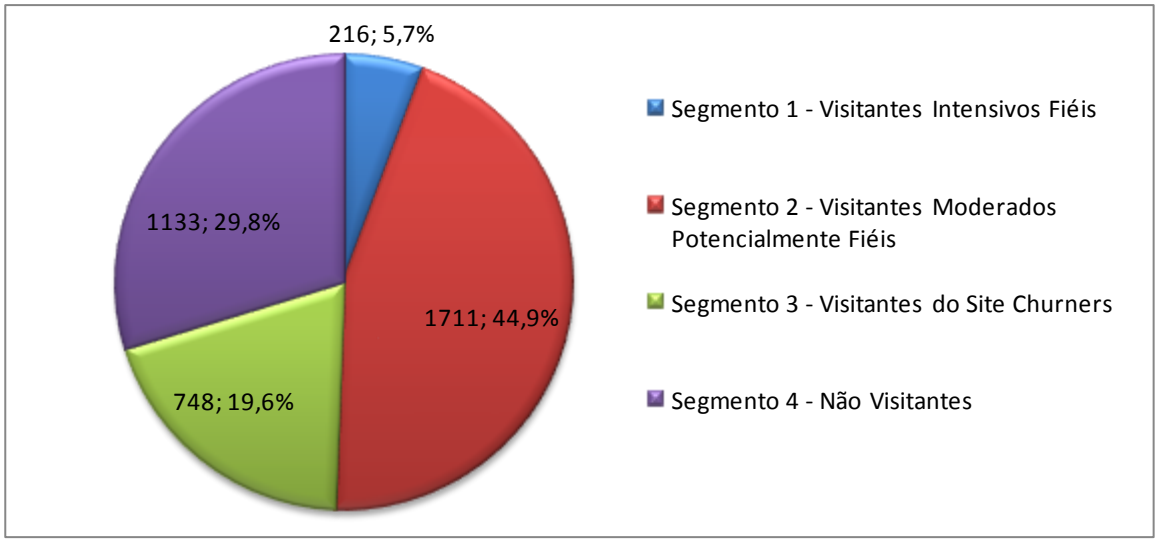
Na aplicação do método não foi restringido o número máximo de *clusters* a obter. A escolha final do número de *clusters* foi efectuada com base no critério *Bayesian Information Criteria* e da Razão de Distâncias. No 1º critério constata-se que o decréscimo do BIC atenua logo a partir dos quatro grupos, enquanto no 2º critério verifica-se que a Razão de Distâncias é superior para quatro *clusters*, ainda que para sete *clusters* este valor também seja bastante elevado. Ambos os critérios são coincidentes e, apesar de ambas as alternativas de quatro e sete *clusters* poderem ser escolhidas, optou-se pela solução de quatro *clusters* (ver Figura 14).

Figura 14 – Resultados dos critérios de avaliação do número de clusters



Como resultado da opção por quatro *clusters*, obtiveram-se os segmentos apresentados na Figura 15.

Figura 15 – Segmentação dos utilizadores



Genericamente, o Segmento 1 caracteriza-se pela reduzida dimensão e por incluir visitantes que possuem maior intensidade de utilização e que revelam maior potencial de continuar a utilizar o Site e/ou o E-mail nos Anos 1 e 2.

O Segmento 2 é o de maior dimensão e é constituído por visitantes que revelam menor intensidade de utilização que os do Segmento 1, mas têm potencial de fidelização superior aos Segmentos 3 e 4.

O Segmento 3 é constituído por utilizadores que no Ano 0 visitam o *site*, mas nos anos seguintes, maioritariamente, não têm actividade com o Site ou o E-mail.

O Segmento 4 é constituído na sua maioria por Não Visitantes, isto é, utilizadores sem actividade ou apenas com visualização do E-mail.

A análise em detalhe de cada segmento de acordo com a Tabela 13 permite constatar que no Ano 0 os utilizadores dos Segmentos 1 e 2 têm sessões de ambos os tipos: Site e/ou Email, enquanto que 98,3% dos utilizadores do Segmento 3 são apenas visitantes do Site. Em relação ao Segmento 4, a percentagem de utilizadores que não efectuaram qualquer visita ao *site* atinge 77,7% e a percentagem dos que só visualizaram o E-mail é de 22,3%.

A análise do comportamento nos Anos 1 e 2 permite constatar que o Segmento 1 é de facto o mais fiel uma vez que , no Ano 1, 78,7% dos utilizadores mantém a actividade e no Ano 2

essa percentagem assume o valor 46,3%. No Segmento 2, a percentagem de utilizadores que mantém actividade nos Anos 1 e 2 é de 46,2% e 26,3%, respectivamente. Pelo contrário, nos Segmentos 3 e 4 a percentagem de Utilizadores Sem Actividade é muito elevada, verificando-se no Ano 1 para os Segmentos 3 e 4 percentagens de ausência de actividade de 84% e 86,7%, respectivamente. Para além disso, no Ano 2 essas percentagens sofrem novo aumento passando a assumir os valores 92,2% e 93,7% nos Segmentos 3 e 4, respectivamente.

Tabela 13 – Tipo de páginas visitadas por segmento

		Segmento 1 - Visitantes Intensivos Fiéis	Segmento 2 - Visitantes Moderados Potencialmente Fiéis	Segmento 3 - Visitantes do Site Churners	Segmento 4 - Não Visitantes
Tipo de Páginas Visitadas no Ano 0	Só Site	5,1%	-	98,3%	-
	Só E-mail	1,9%	-	-	22,3%
	Site e/ou E-mail	93,1%	100,0%	1,7%	-
	Sem actividade	-	-	-	77,7%
Tipo de Páginas Visitadas no Ano 1	Só Site	24,5%	14,4%	11,8%	6,2%
	Só E-mail	19,0%	17,1%	2,5%	4,9%
	Site e/ou E-mail	35,2%	14,8%	1,7%	2,3%
	Sem actividade	21,3%	53,8%	84,0%	86,7%
Tipo de Páginas Visitadas no Ano 2	Só Site	2,8%	-	1,5%	0,1%
	Só E-mail	32,4%	26,1%	6,3%	6,0%
	Site e/ou E-mail	11,1%	0,2%	-	0,2%
	Sem actividade	53,7%	73,7%	92,2%	93,7%

A Tabela 14 apresenta o número de sessões efectuadas no Ano 0, o qual permite aprofundar a distinção entre os 3 primeiros Segmentos. O Segmento 1 distingue-se claramente pelo facto de ter elevado Número de Sessões, sendo que 79% tem mais de 8 sessões (Site e/ou E-mail) e 22,2% tem mais de 8 sessões exclusivas de visita ao *site*. O Segmento 2 tem uma utilização mais moderada, com 56% dos seus elementos com 4 ou menos sessões Site e/ou E-mail e com 81,2% dos seus elementos com 1 a 3 sessões exclusivas de visita ao Site. No Segmento 3 constata-se que a distribuição do Número de Sessões Site e/ou E-mail é similar à do Número de Sessões Só Site, uma vez que cerca de 60% dos utilizadores tem apenas 1 sessão.

Tabela 14 – Número de sessões por segmento

		Segmento 1 - Visitantes Intensivos Fiéis	Segmento 2 - Visitantes Moderados Potencialmente Fiéis	Segmento 3 - Visitantes do Site Churners	Segmento 4 - Não Visitantes
Nº de Sessões Site e/ou Email_ Ano 0	,00	-	-	-	77,7%
	1,00	-	3,7%	59,5%	11,6%
	2,00	-	14,1%	22,5%	5,2%
	3,00	-	56,1% 20,3%	10,0%	2,9%
	4,00	1,4%	18,0%	5,3%	1,7%
	5,00	6,9%	16,5%	1,7%	0,3%
	6,00	6,5%	12,3%	0,7%	0,2%
	7,00	6,0%	6,8%	0,3%	0,2%
	Mais de 8	79,2%	8,2%	-	0,4%
Nº de Sessões_Só Site _ Ano 0	,00	2,3%	13,0%	0,3%	10-
	1,00	11,1%	41,8%	60,2%	-
	2,00	7,4%	81,2% 26,2%	22,6%	-
	3,00	9,7%	13,2%	9,6%	-
	4,00	11,6%	4,4%	4,8%	-
	5,00	14,8%	1,2%	1,7%	-
	6,00	11,1%	0,2%	0,5%	-
	7,00	9,7%	-	0,3%	-
	Mais de 8	22,2%	-	-	-

Em termos médios, de acordo com a Tabela 15 verifica-se que em todos os Segmentos existe diminuição do número médio de sessões de ano para ano. Através deste indicador é bastante visível a diferença existente entre os Segmentos 1 e 2 em todos os anos. Os Utilizadores do Segmento 1 têm, em média, 11,87 sessões no Ano 0, face a 4,44 do Segmento 2. No entanto, no Ano 1 esta média diminui de forma relevante em ambos os Segmentos, em concreto, no Segmento 1 o número médio de sessões é 2,69, enquanto que no Segmento 2 é 0,84.

Tabela 15 – Número médio de sessões por segmento

		Segmento 1 - Visitantes Intensivos Fiéis	Segmento 2 - Visitantes Moderados Potencialmente Fiéis	Segmento 3 - Visitantes do Site Churners	Segmento 4 - Não Visitantes
Número médio de sessões	Nº de Sessões Site e/ou Email_ Ano 0	11,87	4,44	1,70	,44
	Nº de Sessões _ Só Site Ano 0	6,18	1,58	1,67	,00
	Nº de Sessões Site e/ou Email_ Ano 1	2,69	,84	,24	,19
	Nº de Sessões _ Só Site Ano 1	1,57	,38	,19	,09
	Nº de Sessões Site e/ou Email_ Ano 2	,79	,33	,09	,08
	Nº de Sessões _ Só Site Ano 2	,17	,00	,01	,00

Comparando o Número médio de páginas por sessão representado na Tabela 16, para cada um dos anos, constata-se que no Ano 0 o Segmento 3 apresenta maior Número médio de páginas por sessão (9,16), enquanto que no Segmento 1 este valor é de 6,29 e no Segmento 2 é 4,63.

Neste indicador o Segmento 3 consegue obter um valor elevado devido ao tipo de páginas visitadas, isto é, devido a 98,3% dos utilizadores só terem sessões Só Site e a este tipo de sessão estar associado um Número médio de páginas superior ao afecto a uma sessão de leitura de E-mail. No Ano 1, o comportamento *Churner* dos utilizadores do Segmento 3 faz diminuir, de forma abrupta, o Número médio de páginas por sessão (concretamente, o número médio de páginas passa de 9,16 no Ano 0 para 0,99 no Ano), enquanto que os comportamentos de maior fidelização dos utilizadores dos Segmentos 1 e 2 mantêm este indicador em 3,52 e 1,67, respectivamente.

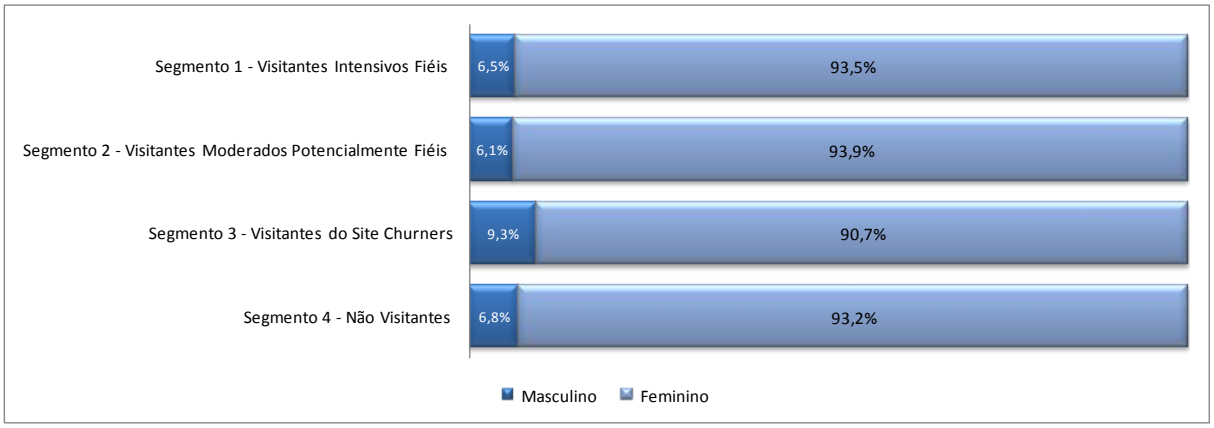
Tabela 16 – Número médio de páginas por sessão por segmento

		Segmento 1 - Visitantes Intensivos Fiéis	Segmento 2 - Visitantes Moderados Potencialmente Fiéis	Segmento 3 - Visitantes do Site Churners	Segmento 4 - Não Visitantes
Número médio de páginas por sessão	Ano 0	6,29	4,63	9,16	0,25
	Ano 1	3,52	1,67	0,99	0,54
	Ano 2	0,94	0,34	0,08	0,22

A caracterização sócio-demográfica dos quatro segmentos permite concluir que não existem diferenças muito relevantes no perfil dos 4 segmentos, mas pontualmente ocorrem diferenças ligeiras.

A comparação do Género Sexual dos utilizadores de cada segmento, de acordo com a Figura 16, permite identificar uma grande semelhança na composição dos 4 grupos, sendo a maior parte dos grupos constituídos por cerca de 93% de elementos do género Feminino. A excepção é o Segmento 3 que tem 9,3% de elementos do género Masculino.

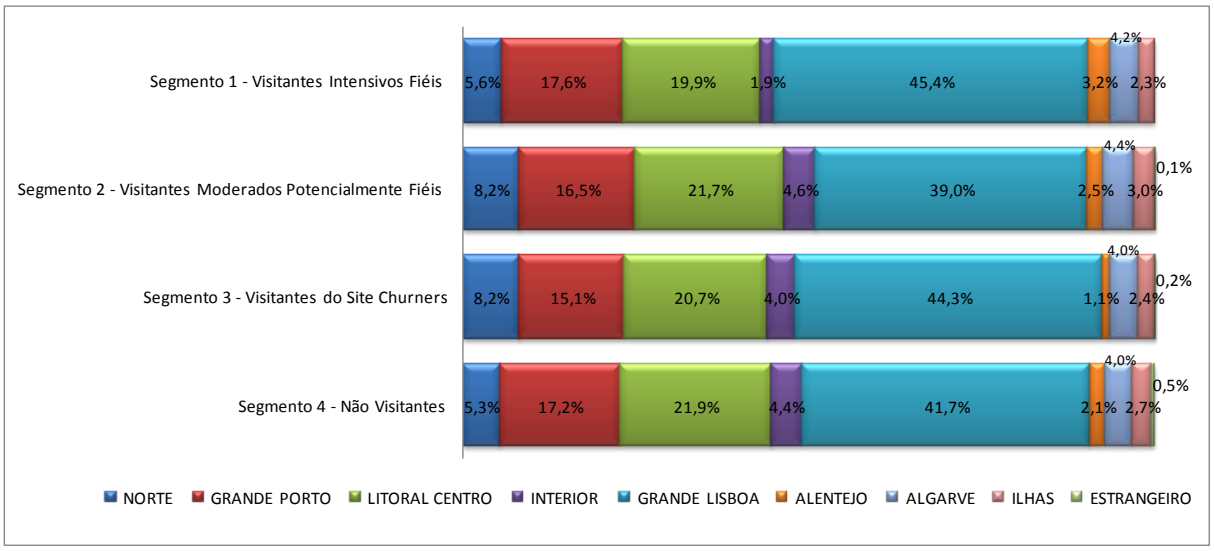
Figura 16 – Género sexual dos utilizadores por segmento



Atendendo à Figura 17, pode afirmar-se que geograficamente também não existem diferenças relevantes na distribuição dos utilizadores de cada segmento. Verifica-se, no entanto, que nos

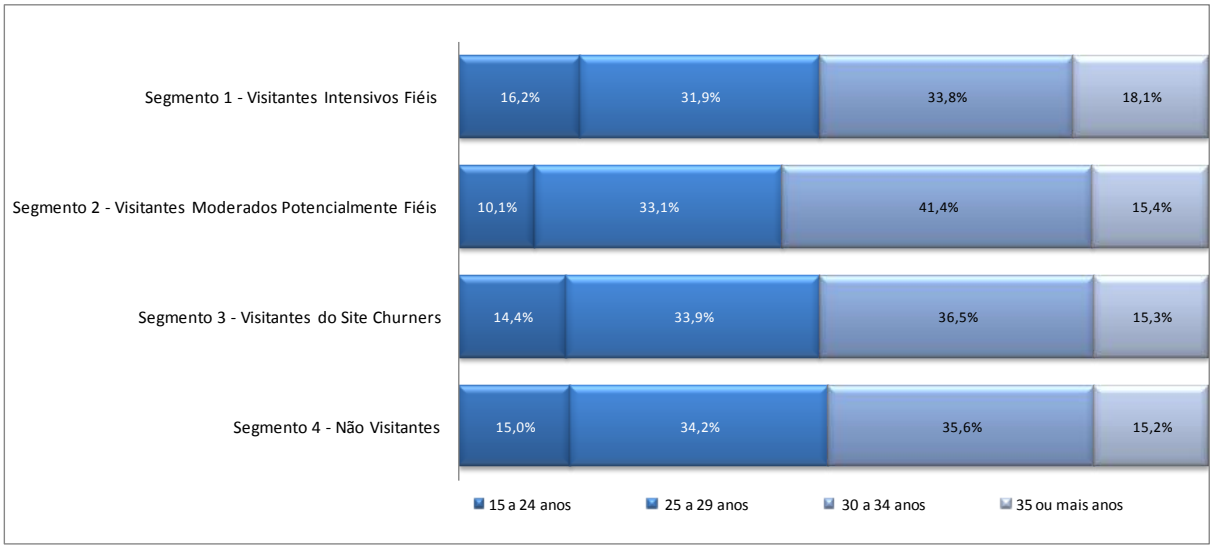
Segmentos 1 e 3 existem mais elementos da Grande Lisboa do que nos Segmentos 2 e 4 (concretizando, 45,4%, 44,3%, 39% e 41,7%, respectivamente para os Segmentos 1,3,2 e 4).

Figura 17 – Perfil geográfico dos utilizadores por segmento



Em relação ao Perfil Etário por Segmento, a análise da Figura 18 permite concluir que no Segmento 1 existe maior percentagem de utilizadores com 35 ou mais anos (18,1% face a cerca de 15% nos restantes segmentos) e também de utilizadores com menos de 25 anos (16,2% face a cerca de 15% dos Segmentos 3 e 4 e 10,1% do Segmento 2). Pelo contrário, o Segmento 2 é o que apresenta maior concentração de elementos nas faixas etárias intermédias, isto é, entre os 25 e os 34 anos. Neste segmento, 74,5% dos utilizadores tem entre 25 e 34 anos, enquanto que nos restantes segmentos esta percentagem não ultrapassa os 70%.

Figura 18 – Perfil etário dos utilizadores por segmento



Em síntese, conclui-se que a segmentação obtida permite diferenciar de forma relevante os comportamentos dos Utilizadores Registados em 2007, com base na utilização nos três períodos em análise (Anos 0, 1 e 2) e que a caracterização Sócio-demográfica não explica a diferença de comportamentos entre os 4 segmentos. Deste modo, as estratégias de CRM a adoptar a cada um dos Segmentos terão que ter em conta a heterogeneidade demográfica dos mesmos.

.

5. APLICAÇÃO DO MODELO DE CLASSIFICAÇÃO

Conforme referido anteriormente, pretende-se aplicar um Modelo de Classificação que permita predizer qual o segmento a que pertencem os utilizadores registados em 2008 e em 2009, com base no seu comportamento no Ano 0. Para efectuar a classificação recorrer-se-á a um Modelo Híbrido com base nas metodologias de classificação: Árvores de Decisão e Redes Neuronais.

A aplicação do Modelo Híbrido terá por base a abordagem efectuada por Lin and McClean (2001) e também aplicada a um caso de *Web Mining* por E. Suh et al. (2004). Para este caso em concreto, a adaptação desta abordagem será efectuada tendo em conta que só estão a ser utilizados duas metodologias e que a variável alvo tem quatro categorias e não duas como no caso apresentado por E. Suh et al (2004).

Para aplicar o Modelo de Classificação foram seleccionadas as variáveis do Ano 0 que foram utilizadas no processo de segmentação, às quais foram adicionadas outras variáveis, também referentes ao mesmo Ano. Estas variáveis foram seleccionadas após a realização de vários testes com combinações de variáveis, tendo sido obtidos os melhores resultados com o conjunto de variáveis enunciadas na Tabela 17.

Tabela 17 – Variáveis incluídas no Modelo de Classificação

Ano a que se refere	Nome da variável	Descrição	Variável Utilizada no Processo de Segmentação
Ano 0	Status_Ano0	Status Activação e Visita Ano 0	✗
	Status_de_Activação_Ano0	Status de Activação no Ano 0	✗
	Status_de_Visita_Ano0	Status de Visita no Ano 0	✗
	dias_ate_activacao_cod	Nº de Dias entre Registo e Activação – Categorizada	✗
	Activ_Ano0	Tipo de Páginas Visitadas no Ano 0	✓
	N_de_sessoes_Ano0	Nº de Sessões Site e/ou Email_ Ano 0	✗
	N_de_sessoes_SóSite_Ano0	Nº de Sessões _ Só Site Ano 0	✗
	paginas_por_sessao_Ano0	Número médio de páginas por sessão no Ano0	✓
	N_de_sessoes_Ano0_Trim1	Número médio de sessões Site e/ou Email no Ano0 Trim1	✓
	N_de_sessoes_Ano0_Trim2	Número médio de sessões Site e/ou Email no Ano0 Trim2	✓
	N_de_sessoes_Ano0_Trim3	Número médio de sessões Site e/ou Email no Ano0 Trim3	✓
	N_de_sessoes_Ano0_Trim4	Número médio de sessões Site e/ou Email no Ano0 Trim4	✓
	N_de_sessoes_SóSite_Ano0_Trim1	Número médio de sessões Só Site no Ano0 Trim1	✓
	N_de_sessoes_SóSite_Ano0_Trim2	Número médio de sessões Só Site no Ano0 Trim2	✓
	N_de_sessoes_SóSite_Ano0_Trim3	Número médio de sessões Só Site no Ano0 Trim3	✓
	N_de_sessoes_SóSite_Ano0_Trim4	Número médio de sessões Só Site no Ano0 Trim4	✓

Legenda: ✓ **- Sim** ✗ **- Não**

A aprendizagem e validação de ambos os modelos foi realizada através da divisão dos 3808 utilizadores alvo do modelo em Amostra de Treino e Amostra de Teste, respectivamente. Assim, foram utilizadas duas partições:

- 70% de Amostra de Treino e 30% de Amostra Teste

- 50% de Amostra de Treino e 50% de Amostra Teste.

Ambas as partições foram definidas previamente pelo SPSS e aplicadas de igual forma a todos os modelos de modo a que os resultados obtidos não fossem influenciados por alterações nas partições. Tal significa que para cada partição os casos considerados como Amostra de Teste foram assim considerados em todas as alternativas de modelo que utilizassem a mesma partição, quer nas Árvores de Decisão quer nas Redes Neuronais.

Assim, nos pontos seguintes apresentar-se-ão os resultados obtidos através de cada uma das metodologias isoladamente e de seguida a combinação resultante das melhores soluções obtidas anteriormente.

5.1. Modelo com Árvores de Decisão

Para aplicação da metodologia de Árvores de Decisão foi utilizado o algoritmo CART já descrito anteriormente.

Com base nas variáveis e partições da amostra atrás enunciadas e tendo o objectivo de desenvolver um modelo que possua boa capacidade classificativa, foram testados vários modelos com diferentes parâmetros, dos quais se revelam algumas das alternativas testadas.

Inicialmente procurou-se definir um modelo que se adaptasse o mais possível à amostra de treino, pelo que foram definidos os seguintes parâmetros:

- Profundidade da Árvore (Número máximo de níveis): 20;
- Nº Mínimo de observações por nível Nó Pai: 20;
- Nº Mínimo de observações por nível Nó Filho: 5;
- Valor mínimo para função de diversidade: 0,0001;
- Sem Poda.

Apesar dos parâmetros anteriores permitirem a obtenção de uma árvore bastante complexa, a primeira versão mostrou-se relativamente simples na sua estrutura, tendo apenas 7 níveis e 16 nós folha. Este modelo obteve excelentes resultados na classificação dos utilizadores em segmentos, conseguindo classificar correctamente 98% dos casos na Amostra de Treino e 97,4% dos casos na Amostra de Teste. Estes resultados revelam não só excelente capacidade de classificação, mas também excelente consistência entre Amostra de Treino e Amostra de Teste, não havendo sinais significativos de sobreajustamento.

A partir desta solução, foram também experimentadas outras alternativas de valores para os parâmetros cuja definição tinha como objectivos incrementar o desempenho de classificação e/ou diminuir a complexidade. Assim, na Tabela 18 sistematizam-se as modificações efectuadas nos valores dos parâmetros e os respectivos resultados obtidos.

Tabela 18 – Resultados obtidos nas Árvores de Decisão

		Árvore 1	Árvore 2	Árvore 3	Árvore 4	Árvore 5	Árvore 6	Árvore 7	Árvore 8
Amostra de treino/ teste		70/30	70/30	50/50	70/30	70/30	70/30	70/30	50/50
Pré-Poda	Nº Máximo de Níveis	20	20	20	20	20	20	20	20
	Nº Mínimo de observações por nível Nó Pai	20	20	20	30	30	100	100	100
	Nº Mínimo de observações por nível Nó Filho	5	5	5	15	15	50	50	50
	Valor mínimo para função de diversidade	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
Poda	Valor Máximo de diferença no risco (nos erros padrão)	Sem poda	Com poda (1)	Com poda (1)	Sem poda	Com poda (1)	Sem poda	Com poda (1)	Com poda (1)
Estrutura	Nº de nós folha	16	9	9	6	6	5	5	4
	Profundidade	7	5	5	4	4	3	3	2
Ajustamento Amostra de treino	Casos Bem Classificados (%)	98,0	97,8	97,9	97,0	97,0	96,3	96,3	96,2
	Casos bem classificados por defeito (regra maioria) (%)	44,9	44,9	44,6	44,9	44,9	44,9	44,9	44,6
	Índice de Huberty	96,4	96,0	96,3	94,5	94,5	93,2	93,2	93,1
Ajustamento Amostra de teste	Casos Bem Classificados (%)	97,4	97,6	97,2	96,9	96,9	96,8	96,8	95,6
	Casos bem classificados por defeito (regra maioria) (%)	45,1	45,1	45,3	45,1	45,1	45,1	45,1	45,3
	Índice de Huberty	95,3	95,6	94,9	94,4	94,4	94,1	94,1	91,9
	Casos Bem Classificados Segmento 1 (%)	76,9	76,9	67,3	67,7	67,7	56,9	56,9	68,2
	Casos Bem Classificados Segmento 2 (%)	97,6	98,0	98,2	98,4	98,4	99,4	99,4	94,7
	Casos Bem Classificados Segmento 3 (%)	99,1	99,1	99,5	97,8	97,8	97,8	97,8	99,0
	Casos Bem Classificados Segmento 4 (%)	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Na Árvore 2 mantiveram-se os mesmos parâmetros enunciados anteriormente e que foram aplicados à Árvore 1 mas realizou-se a Poda, tendo-se utilizado o critério de Poda baseado nos erros Padrão. Segundo Fonseca (1994), apresentado em Quinlan (1993), o algoritmo de Poda baseada no erro visa contornar o problema da utilização de um conjunto independente de exemplos para esse efeito, sendo que utilizar o próprio conjunto de treino para efectuar a poda da árvore tem a vantagem de não obrigar à redução do mesmo por separação, em conjunto de treino e conjunto de teste.

Quer na Árvore 2, quer nas seguintes em que foi efectuada Poda utilizou-se sempre o valor um como parâmetro do Valor Máximo de diferença no risco. Os resultados obtidos na Árvore 2 foram ainda ligeiramente melhores do que os obtidos na alternativa anterior, conseguindo-se classificar de forma correcta 97,6% dos casos da Amostra de Teste.

Na Árvore 3, utilizaram-se os mesmos parâmetros da Árvore 2, mas utilizou-se a segunda partição da amostra, isto é, o algoritmo treinou sobre 50% da amostra e foi testado sobre os restantes 50%. Com menos dados para treinar, os valores das percentagens de casos bem classificados sofreram um ligeiro decréscimo face às alternativas anteriores, continuando a demonstrar elevada consistência.

Nas alternativas seguintes testou-se o impacto da definição de valores para os parâmetros relativos ao número de casos em cada nó, de forma a impor regras de paragem mais restritivas ao crescimento das árvores, sem e com poda.

Com estas alternativas obtêm-se apenas ligeiros decréscimos na percentagem de Casos Bem Classificados com redução quer no número de Nós Folha, quer na Profundidade. No entanto, dado que a Árvore que apresenta o melhor resultado (Árvore 2) apresenta reduzido grau de complexidade será a Árvore 2 que será utilizada para aplicação no Modelo Híbrido.

Analisando em detalhe os resultados obtidos com esta alternativa de Árvore, considerando-a um Modelo de Classificação, conclui-se que existe elevada capacidade classificativa (97,6% dos casos em Amostra de Teste) e elevada consistência com o desempenho obtido na Amostra de Treino (97,8% dos casos bem classificados). O Índice de Huberty de 95,6% confirma também a boa capacidade preditiva do modelo, significando este valor que o modelo contribuiu para aumentar em 95,6% a capacidade de classificação correcta face à classificação que se obteria por defeito.

Por segmento, constata-se que a capacidade preditiva é bastante elevada no Segmento 2 (98,0%), Segmento 3 (99,1%) e Segmento 4 (100%). O Segmento 1 que representa apenas 5,7% do total da amostra é aquele em que se verifica menor capacidade preditiva do modelo (com 76,9% dos casos bem classificados). Para este segmento, atendendo à Tabela 19 constata-se que existe alguma dificuldade do modelo em distinguir se os elementos pertencem ao Segmento 1 ou a outros, sendo que a maior confusão ocorre com o Segmento 2 (na Amostra de Teste 15,4% dos utilizadores do Segmento 1 são classificados como pertencentes ao Segmento 2). Excluindo os casos de má classificação entre os Segmentos 1 e 2, existem apenas 0,8% de casos mal classificados na Amostra de Teste.

Tabela 19 – Matriz de classificação associada à Árvore de Decisão 2

Segmentos Observados		Segmentos Preditos				
		Segmento 1- Visitantes Intensivos Fiéis	Segmento 2 - Visitantes Moderados Potencialmente Fiéis	Segmento 3 - Visitantes do Site Churners	Segmento 4 - Não Visitantes	Casos Bem Classificados (%)
Amostra de Treino	Segmento 1	115	25	9	2	76,2%
	Segmento 2	20	1187	2	0	98,2%
	Segmento 3	0	2	522	0	99,6%
	Segmento 4	0	0	0	811	100,0%
	Total	5,0%	45,0%	19,8%	30,2%	97,8%
Amostra de Teste	Segmento 1	50	10	3	2	76,9%
	Segmento 2	8	492	2	0	98,0%
	Segmento 3	0	2	222	0	99,1%
	Segmento 4	0	0	0	322	100,0%
	Total	5,2%	45,3%	20,4%	29,1%	97,6%

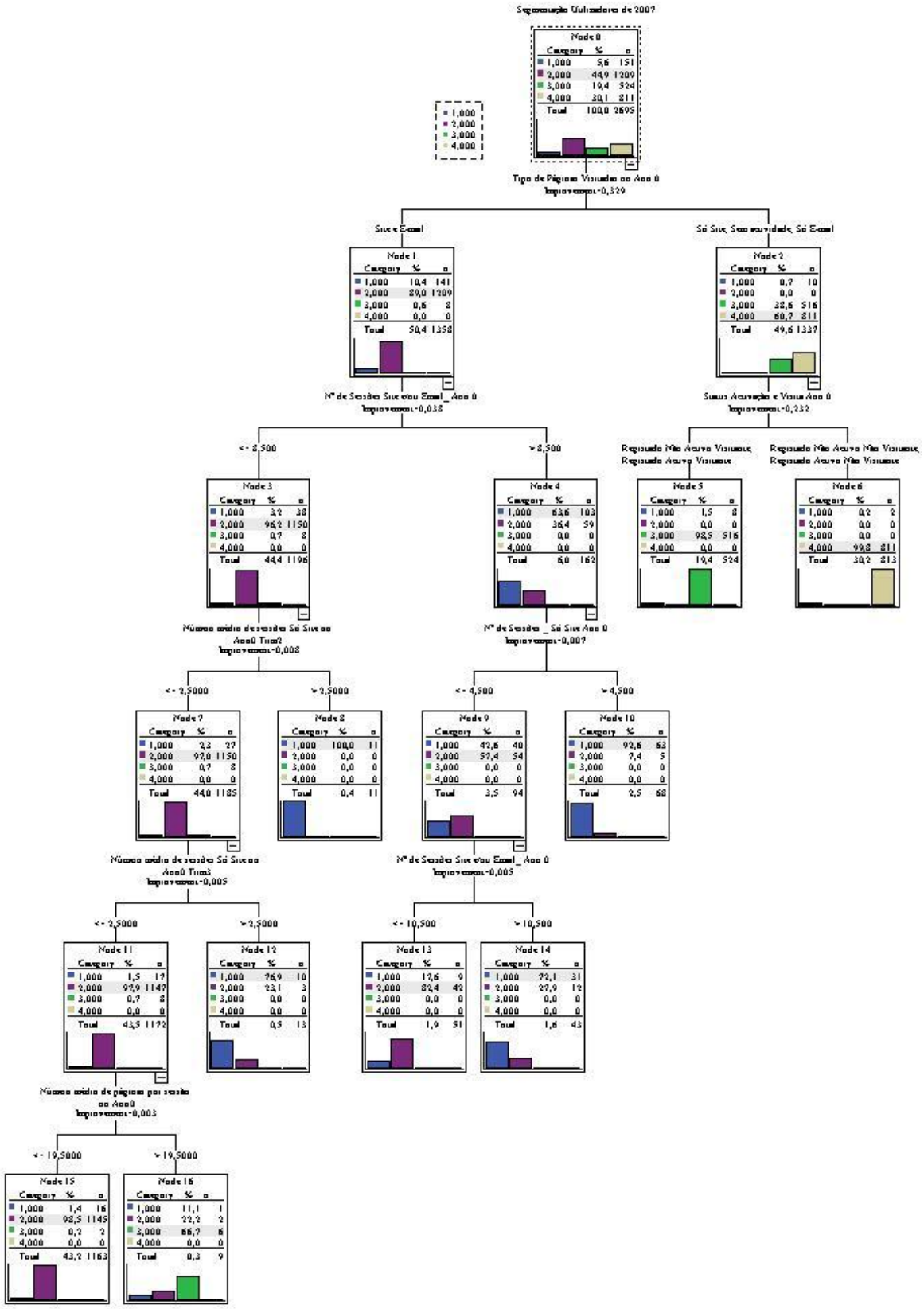
A árvore seleccionada é apresentada na Figura 19 e é constituída por dezassete nós, dos quais nove são nós folha. O número mínimo de casos nos Nós Pai é 24 (no Nó 9) e nos Nós Filho é 2 (no Nó 12).

Uma análise em detalhe da Árvore de Decisão seleccionada leva a afirmar que logo no nível 1 é efectuada a separação total entre os Segmentos 1 e 2 e os Segmentos 3 e 4 através da variável Tipo de Páginas Visitadas no Ano 0. A redução da diversidade proporcionada através da inclusão desta variável, neste nível, é a de maior valor (0,329), conforme se pode verificar na Figura 19.

Passando para o segundo nível da Árvore de Decisão, a variável Status de Activação e Visita no Ano 0 permite por si só separar os Segmentos 3 e 4 e classificar a quase totalidade dos casos correctamente. Por sua vez, a partir do Nó 1 a variável Nº de Sessões Site e/ou E-mail_Ano0 divide-se entre quem tem valor inferior ou igual a 8,5 Sessões e quem tem valor superior, sendo que no primeiro caso encontram-se sobretudo utilizadores classificados como Segmento 2 (96,2%) e no segundo caso são classificados 71,9% como elementos do Segmento 1 e 28,1% como do Segmento 2.

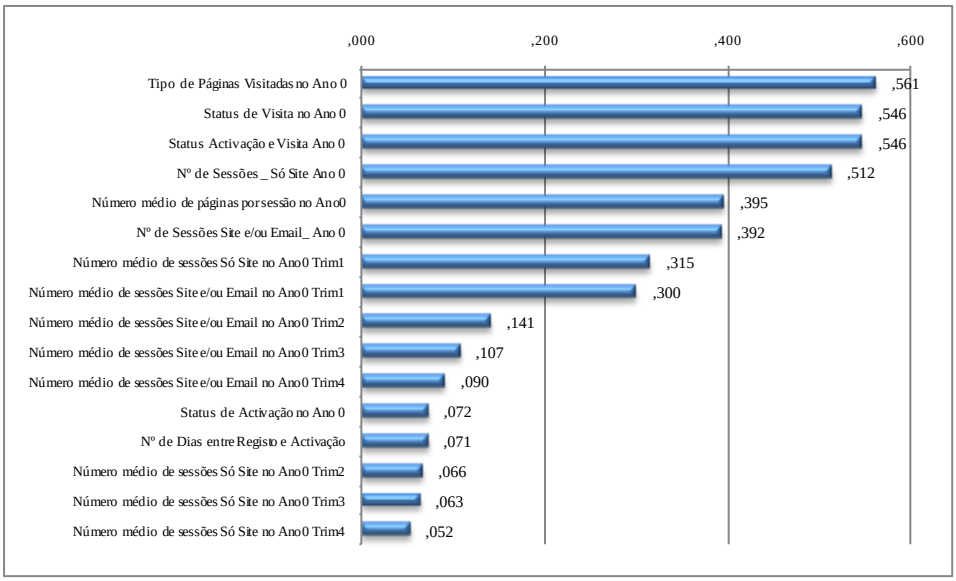
Nos níveis 3 e 4 as variáveis que têm poder explicativo para classificar os Segmentos 1 e 2 são relacionadas com o número médio de sessões num trimestre ou no total do ano. Por sua vez, a ramificação no nível 5 é efectuada com a variável Número Médio de Páginas por Sessão e quando esta assume valor inferior ou igual a 19,5, 98% dos casos são classificados como pertencentes ao Segmento 2.

Figura 19 – Árvore de Decisão 2



Na Figura 20 é observável a importância das várias variáveis para o modelo, sendo que quatro delas apresentam importância acima de 0,5: Tipo de Páginas Visitadas no Ano 0 (0,561), Status de Visita no Ano 0 (0,546), Status Activação e Visita no Ano 0 (0,546) e Nº de Sessões_Só Site Ano 0 (0,512). As variáveis que apresentam menor importância para a definição da Árvore de Decisão são as variáveis que reflectem o comportamento de Activação e as variáveis relativas aos comportamentos dos Utilizadores a partir do 2º Trimestre.

Figura 20 – Importância das variáveis no modelo da Árvore de Decisão 2



Em conclusão, o algoritmo CART revela uma excelente capacidade para classificar correctamente novos casos, tendo-se obtido desempenhos classificativos bastante elevados, com elevada consistência e facilidade de interpretação. Deste modo, conclui-se que o algoritmo de CART por si só seria um método adequado para a exploração deste problema. Ainda assim, caso o menor desempenho obtido na classificação dos utilizadores do Segmento 1 fosse considerado crítico poderiam ser associados custos de classificação incorrecta de modo a minimizar esses impactos.

5.2. Modelo com Redes Neuronais

A modelação utilizando Redes Neuronais considerou as variáveis e as partições de Amostra descritas na Secção 5 para o modelo com Árvores de Decisão.

No entanto, segundo Santos *et al.* (2005) as variáveis nominais devem ser modeladas por meio de indicadores binários associados a cada uma das suas categorias e as variáveis quantitativas devem ser estandardizadas de modo a terem média nula e desvio padrão único, para que o protagonismo destas no modelo não esteja dependente da sua escala de medição. De realçar que, apesar destes procedimentos de transformação das variáveis terem sido executados previamente, utilizando o SPSS, não se verificam diferença de resultados face a modelação em que se utilizam as variáveis originais. Assim, é de admitir que o SPSS executa estes procedimentos de forma automática.

Tal como nas Árvores de Decisão, também nas Redes se testaram várias alternativas das quais se sintetizam os resultados obtidos nas mais relevantes. Em todas as alternativas apresentadas o Número de épocas definido para treino foi 1000, embora tenham sido testadas outras opções, as quais não produziram efeitos relevantes no ajustamento, quer à amostra de treino quer à amostra de teste. Como Método de Optimização utilizou-se o Gradiante Descendente por ser o que produz melhores resultados e ser o mais eficiente em redes de grande dimensão, como é o caso desta.

A função de activação, também chamada de função de transferência, é uma função matemática que aplicada à combinação linear entre as variáveis de entrada e pesos que chegam a determinado neurónio, retorna o seu valor de saída. Existem diversas funções matemáticas que são utilizadas como função de activação. Nas alternativas que produziram melhores resultados recorreu-se à função Tangente Hiperbólica como função de Activação na camada de Output, enquanto que na(s) camada(s) escondidas a utilização desta função ou da função Sigmóide produziu também bons resultados. A utilização da função Tangente Hiperbólica é, segundo Pereira (1999), bastante utilizada em problemas de classificação devido ao facto de, em algumas situações práticas, a função Tangente Hiperbólica acelerar a convergência do algoritmo de treino da rede neuronal. Entretanto, não é claro se diferentes funções de activação têm maiores efeitos no desempenho da rede.

Da definição do número de camadas escondidas, ou intermédias, vai depender também o número de neurónios na camada escondida, sendo que, quanto menor o número de camadas maior o número de neurónios. Assim, foram testadas soluções com uma e duas camadas escondidas, não se verificando diferenças relevantes na amostra de teste, no que diz respeito ao número de casos bem classificados. No entanto, nas situações com duas camadas existe

maior ajustamento aos dados de treino, ainda que não se possa dizer que se trata de sobreajustamento, pois as diferenças entre as duas amostras são sempre pequenas.

Na literatura não se encontra um critério geral que permita definir o número de neurónios na camada intermédia. Em geral, redes neuronais com poucos neurónios escondidos são preferíveis, visto que elas tendem a possuir um melhor poder de generalização, reduzindo o problema de sobreajustamento (*overfitting*). Entretanto, redes com poucos neurónios escondidos podem não possuir a habilidade suficiente para modelar e aprender os dados em problemas complexos, podendo ocorrer *underfitting*, ou seja, a rede não converge durante o período de treino (Pereira, 1999).

Com base nestas considerações optou-se por não limitar o número máximo de unidades em cada camada, verificando-se em todas as alternativas testadas que os resultados obtidos não revelam sinais de sobreajustamento.

A Taxa de Aprendizagem foi fixa em 0,1 pois é o valor mais recomendado na literatura, pois permite incrementos mais pequenos em cada ciclo e assim dá maior garantia de que mais soluções são experimentadas. Ainda assim, foram testados outros valores para este parâmetro e constata-se que a melhor alternativa foi encontrada com o valor fixo em 0,2. Utilizando valores do parâmetro acima de 0,2, os resultados obtidos são menos consistentes, quer quando se compara a Amostra de Treino com a Amostra de Teste, quer entre repetições do processo.

Em todas as opções apresentadas atribui-se o valor 0,9 ao parâmetro *Momentum* pois valores elevados do *Momentum* tendem a ajudar mais o algoritmo a evitar os mínimos locais. Em zonas planas da superfície do erro o *Momentum* ajuda a aprendizagem a prosseguir e em zonas em que o gradiente se mantém, o *Momentum* ajuda a acelerar a aprendizagem (Cardoso). Esse valor é também o definido por defeito pelo SPSS.

Tal como as Árvores de Decisão, também as Redes Neuronais se revelam bastante adequadas para a resolução deste problema, pois o desempenho obtido na classificação é muito bom (ver Tabela 20).

O melhor resultado com Redes Neuronais está associado às Redes 6 e 7, as quais são apresentadas na Tabela 20. Estas Redes foram obtidas com uma Camada Intermédia, Taxa de Aprendizagem de 0,1 e 0,2, respectivamente, *Momentum* de 0,9, Tangente Hiperbólica como função de activação na Camada de Output e Sigmóide como função de activação na Camada Intermédia. Ambas as Redes possuem 11 Neurónios na camada intermédia.

Estas redes produzem os melhores resultados globais na percentagem de Casos Bem Classificados na Amostra de Teste com 97,8%, sendo que, na Amostra de Treino, a Rede 6 obtém o resultado de 98,4% enquanto a Rede 7 obtém 97,8%. Tais valores revelam elevada consistência dos resultados obtidos e, consequentemente, reduzido sobreajustamento à amostra de treino, em ambas as redes.

Tabela 20 – Resultados obtidos nas Redes Neuronais

		Rede 1	Rede 2	Rede 3	Rede 4	Rede 5	Rede 6	Rede 7	Rede 8
Amostra de treino/ teste		70/30	70/30	50/50	70/30	70/30	70/30	70/30	70/30
Opções de Computação	Nº de épocas	1000	1000	1000	1000	1000	1000	1000	1000
	Método de Optimização	Gradiante Descendente	Gradiante Descendente	Gradiante Descendente	Gradiante Descendente	Gradiante Descendente	Gradiante Descendente	Gradiante Descendente	Gradiante Descendente
	Função Activação Camada Output	Tangente Hiperbólica	Tangente Hiperbólica	Tangente Hiperbólica	Tangente Hiperbólica	Tangente Hiperbólica	Tangente Hiperbólica	Tangente Hiperbólica	Tangente Hiperbólica
	Função Activação Camada(s) Escondida(s)	Tangente Hiperbólica	Tangente Hiperbólica	Tangente Hiperbólica	Tangente Hiperbólica	Tangente Hiperbólica	Sigmoide	Sigmoide	Sigmoide
Parâmetros	Nº de Camadas Intermédias	2	1	2	2	2	1	1	2
	Máximo Unidades Camada 1	Auto	Auto	Auto	Auto	Auto	Auto	Auto	Auto
	Máximo Unidades Camada 2	Auto	-	Auto	Auto	Auto	-	-	Auto
	Taxa de Aprendizagem	0,1	0,1	0,1	0,2	0,5	0,1	0,2	0,1
	Momentum	0,9	0,9	0,9	0,9	0,9	0,9	0,9	0,9
Ajustamento na Amostra de Treino	Casos Bem Classificados (%)	96,6	98,0	98,6	96,2	98,3	98,4	97,8	98,1
Ajustamento na Amostra de Teste	Casos Bem Classificados (%)	96,0	97,2	96,9	96,9	97,6	97,8	97,8	97,7
	Casos Bem Classificados Segmento 1 (%)	49,2	66,1	51,1	59,3	71,2	74,6	72,8	72,9
	Casos Bem Classificados Segmento 2 (%)	98,2	98,6	98,6	98,8	99,0	99,0	99,2	99,0
	Casos Bem Classificados Segmento 3 (%)	97,7	98,1	99,2	98,1	98,1	98,1	98,1	98,1
	Casos Bem Classificados Segmento 4 (%)	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Conforme se pode observar na Tabela 20, a Rede 7 produz os melhores resultados de classificação nos segmentos 2,3 e 4, enquanto que a Rede 6 é a rede que melhor classifica os elementos do Segmento 1 (74,6%). Assim, para efeitos de integração no Modelo Híbrido, utilizar-se-á a Rede 7 pois apresenta a melhor capacidade de classificação no Segmento 2.

Analisando em detalhe os resultados obtidos na Rede 7 constata-se a elevada capacidade preditiva no Segmento 2 (99,2%), Segmento 3 (98,1%) e Segmento 4 (100%). O Segmento 1 que representa apenas 5,7% do total da amostra é, tal como nas Árvores de Decisão o único em que se verifica menor capacidade preditiva do modelo com 72,8% dos casos bem classificados. Tal como nas Árvores de Decisão e atendendo à Tabela 21, constata-se que existe alguma dificuldade do modelo em distinguir se os elementos pertencem ao Segmento 1

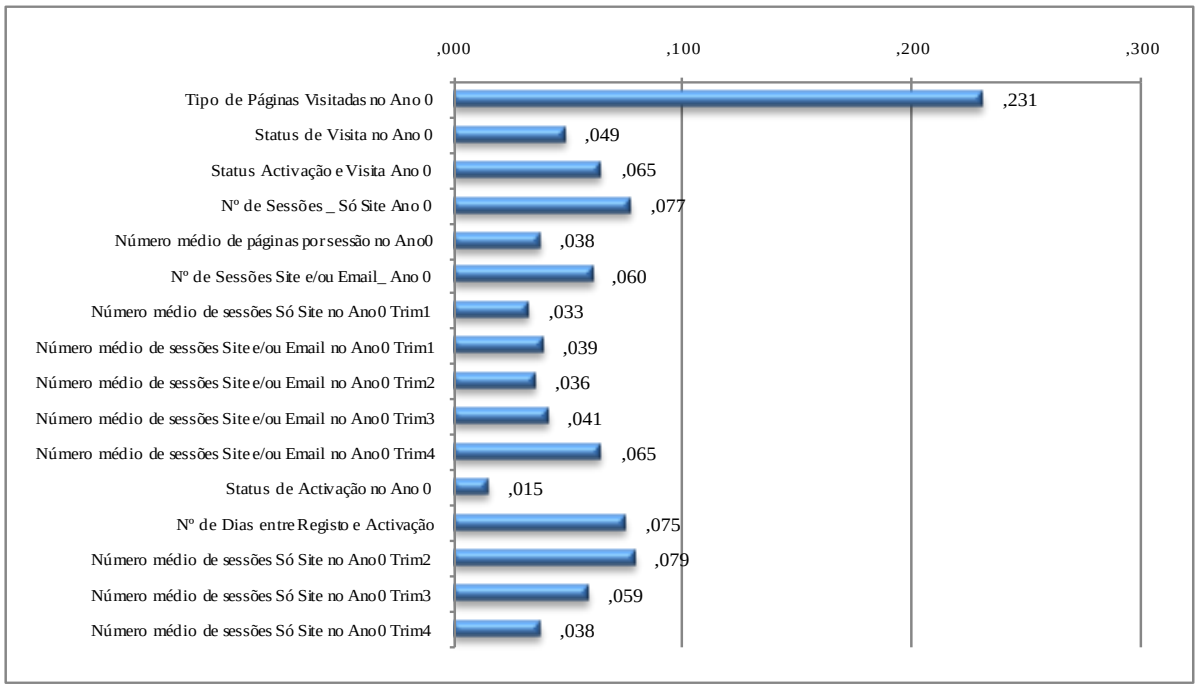
ou a outros, sendo que a maior confusão ocorre com o Segmento 2 na amostra de teste (20,3% dos utilizadores do Segmento 1 são classificados como pertencentes ao Segmento 2). Excluindo os casos de má classificação entre o Segmento 1 e o Segmento 2, na amostra de teste existem apenas 0,7% de casos mal classificados.

Tabela 21 – Matriz de classificação para a Rede Neuronal 7

		Segmentos Preditos				
Segmentos Observados		Segmento 1 - Visitantes Intensivos Fiéis	Segmento 2 - Visitantes Moderados Potencialmente Fiéis	Segmento 3 - Visitantes do Site Churners	Segmento 4 - Não Visitantes	Casos Bem Classificados (%)
Amostra de Treino	Segmento 1	108	33	8	2	71,5%
	Segmento 2	7	1202	0	0	99,4%
	Segmento 3	0	8	516	0	98,5%
	Segmento 4	0	0	0	811	100,0%
	Total	4,3%	46,1%	19,4%	30,2%	97,8%
Amostra de Teste	Segmento 1	43	12	2	2	72,9%
	Segmento 2	4	498	0	0	99,2%
	Segmento 3	0	4	212	0	98,1%
	Segmento 4	0	0	0	322	100,0%
	Total	4,3%	46,8%	19,5%	29,5%	97,8%

A análise da importância das variáveis, ainda que os valores não possam ser directamente comparados com os obtidos nas Árvores de Decisão, permite comparar a hierarquia da importância das variáveis. Assim, na Figura 21 apresenta-se, pela ordem de hierarquia de importância que as variáveis têm na Árvore de Decisão escolhida (ver Figura 19), o somatório da importância de cada categoria das variáveis nominais e da importância das variáveis contínuas independentes.

Figura 21 – Importância das variáveis no modelo da Rede Neuronal 7



Assim, pode-se concluir que o Tipo de Páginas Visitadas no Ano 0 é em ambos os modelos a variável considerada mais importante para a classificação dos casos.

No entanto, as restantes variáveis têm um nível de importância bastante inferior. Além disso, a variável que aqui surge como a segunda mais importante do Ranking é o Número médio de sessões Só Site no Ano 0 Trim3, que nas Árvores de Decisão é considerada a décima quarta variável mais importante.

Em conclusão, as Redes Neuronais com o algoritmo *Backpropagation* revelam excelente capacidade para classificar correctamente novos casos, tendo-se obtido desempenhos classificativos ainda melhores do que os alcançados nas Árvores de Decisão. É ainda de assinalar a elevada consistência dos resultados considerando as Amostras de Treino e de Teste. Face às Árvores de Decisão, as Redes Neuronais têm como principal dificuldade a incapacidade de interpretação sobre como as variáveis e os valores das mesmas permitem classificar os grupos e a impossibilidade de atribuição de custos de classificação incorrecta.

5.3. Modelo Híbrido

Apesar dos excelentes resultados obtidos por ambos os modelos pretende-se aferir se a combinação das duas metodologias num Modelo Híbrido permite obter melhores resultados, sobretudo na classificação dos indivíduos do Segmento 1.

A abordagem utilizada, como foi referido anteriormente, segue a metodologia utilizada por E. Suh et al. (2004), isto é, tem como ponto de partida as probabilidades de cada caso pertencer a um ou a outro segmento, calculadas em cada uma das metodologias individualmente utilizadas.

Assim, para cada utilizador foram calculadas novas probabilidades de pertencer a cada segmento obtidas a partir da média das respectivas probabilidades em:

- Modelo Árvore 2 e Modelo Rede 6;

- Modelo Árvore 2 e Modelo Rede 7.

De realçar que nas Redes Neuronais os valores obtidos não são probabilidades mas sim pseudo-probabilidades pelo que em cada caso a soma das pseudo-probabilidades de pertencer

a um segmento não totaliza o valor 1. Este facto tem impacto nos casos em que se obtém a probabilidade de pertencer a mais de um segmento como sendo superior a 0,5. Assim, a Classificação do Segmento foi efectuada de acordo com o Segmento para o qual existe maior probabilidade de pertença. No Anexo III, sistematiza-se o exemplo dos resultados obtidos para um conjunto casos.

Dado que a partição da Amostra na proporção de 70% para a Amostra de Treino e 30% para a Amostra de Teste nos vários modelos foi efectuada recorrendo-se à mesma variável, calculou-se, também para os Modelos Híbridos (ver

Tabela 22), a percentagem de casos bem classificados em cada segmento nas duas Amostras.

Comparando os resultados obtidos pelas metodologias de forma individualizada com os Modelos Híbridos, constata-se que em ambos os Modelos Híbridos ocorre diminuição da capacidade de classificar correctamente:

- o Segmento 2 quando comparados com ambos os Modelos de Redes Neuronais;

- o Segmento 3 quando comparados com o Modelo de Árvore de Decisão.

Tabela 22 – Comparação das matrizes de classificação

		Casos Bem Classificados (%)				
Segmentos Observados		Árvore de Decisão 2	Rede Neuronal 6	Rede Neuronal 7	Modelo Híbrido (Árvore 2 e Rede 6)	Modelo Híbrido (Árvore 2 e Rede 7)
Amostra de Treino	Segmento 1- Visitantes Intensivos Fiéis	76,16%	81,46%	71,52%	79,47%	78,81%
	Segmento 2 - Visitantes Moderados Potencialmente Fiéis	98,18%	99,42%	99,42%	99,09%	99,01%
	Segmento 3 - Visitantes do Site Churners	99,62%	98,47%	98,47%	98,47%	98,47%
	Segmento 4 - Não Visitantes	100,00%	100,00%	100,00%	100,00%	100,00%
	Total	97,77%	98,40%	97,85%	98,14%	98,07%
Amostra de Teste	Segmento 1- Visitantes Intensivos Fiéis	76,92%	74,58%	72,88%	78,46%	80,00%
	Segmento 2 - Visitantes Moderados Potencialmente Fiéis	98,01%	99,00%	99,20%	98,21%	98,41%
	Segmento 3 - Visitantes do Site Churners	99,11%	98,15%	98,15%	98,21%	98,21%
	Segmento 4 - Não Visitantes	100,00%	100,00%	100,00%	100,00%	100,00%
	Total	97,57%	97,82%	97,82%	97,57%	97,75%

Globalmente, o desempenho dos Modelos Híbridos é ligeiramente inferior ao das Redes Neuronais mas igual ou superior ao da Árvore de decisão. Ainda assim, o objectivo de melhorar a classificação do Segmento 1 é de facto atingido através do Modelo Híbrido, alcançando-se o valor máximo de 80% de Casos Bem Classificados na Amostra de Teste no Modelo Híbrido com a Rede Neuronal 7.

6. CONCLUSÕES

Com esta tese pretende-se, utilizando metodologias de segmentação e de predição de comportamentos aplicadas ao *site* de um Clube de Fidelização, dar um contributo para uma componente do que será a estratégia de CRM de uma empresa de Grande Consumo.

O âmbito da tese enquadra-se na definição de *Web Mining*, segundo Rud (2001), visto que os dados foram recolhidos previamente e toda a construção do modelo será efectuada *offline*, pretendendo-se a posterior inclusão no *site* do resultado do modelo definido para classificar novos casos. Por sua vez, os objectivos desta tese prendem-se com a extracção de informação acerca da utilização por parte dos utilizadores de um *site* o que, segundo as tipologias de actividades definidas por Linnof (2001), inclui-se em *Mining* de Utilização.

Nesta tese recorreu-se a uma Base de Dados real que reflecte as interacções dos Utilizadores Registados de um *site* de um Clube de Fidelização de uma marca de Grande Consumo. Foram seleccionados Utilizadores Registados entre 2007 e 2009.

Da análise da informação recolhida, constatou-se que a taxa de utilizadores que não têm qualquer interacção com o *site* ou com *e-mail* enviado pela marca é bastante elevada, aumentando em função do ano de registo (24,2% em 2007 e 54,1% em 10 meses de 2009) e também do tempo decorrido após o registo (entre os utilizadores de 2007 aumenta de 24,2% no ano de registo para 67,7% no ano seguinte). Outros indicadores analisados corroboram também que existe elevada dificuldade neste *site*, tal como noutros similares, em conseguir manter os utilizadores interessados no *site* e manter a sua taxa de visita com níveis elevados.

Assim, pretendeu-se obter um Modelo que permitisse segmentar os comportamentos de interacção dos utilizadores com registo em 2007 nos anos após o registo e definir um Modelo de Classificação que possa ser aplicado aos utilizadores registados em 2008 e 2009 e assim predizer os seus comportamentos futuros com base no segmento a que pertencem.

Assim, recorreu-se ao método de *Two-Step Clustering*, desenvolvido por Chiu *et al.* (2001), para segmentar os utilizadores de 2007 com base nos comportamentos nos Anos 0, 1 e 2 após o momento de registo. Os resultados obtidos permitiram concluir a existência de quatro segmentos de utilizadores:

- Segmento 1 – Visitantes Intensivos Fiéis (5,7%)

- Segmento 2 – Visitantes Moderados Potencialmente Fiéis (44,9%)

- Segmento 3 – Visitantes do *Site Churners* (19,6%)

- Segmento 2 – Não Visitantes (29,8%)

A estrutura de segmentos obtidos diferencia os utilizadores em termos de comportamentos, não existindo diferenças relevantes em termos sócio-demográficos que permitam explicar as diferenças de comportamentos.

Para criar o Modelo de Classificação que permita predizer quais serão os comportamentos dos utilizadores com base nos seus comportamentos no Ano 0 após o registo, aplicaram-se individualmente as Metodologias de Classificação de Árvores de Decisão (com algoritmo CART) e de Redes Neuronais (com algoritmo *Backpropagation*), bem como a combinação de ambas num Modelo Híbrido.

As melhores alternativas obtidas com cada Modelo obtém desempenhos muito similares no que respeita à capacidade global de classificar correctamente a amostra de teste, sendo que foram classificados correctamente 97,57% dos casos com Árvore de Decisão, 97,82% com Redes Neuronais e 97,75% na melhor opção de Modelo Híbrido. Em qualquer das alternativas de Modelo apresentadas constata-se menor capacidade de classificar correctamente os casos do Segmento 1 (Visitantes Intensivos Fiéis) que é também o segmento de menor dimensão mas que tem bastante relevância em termos de fidelização e intensidade de utilização. A capacidade de classificação correcta deste segmento nas melhores alternativas obtidas com cada método varia entre 72,88% com Rede Neuronal e 80% com Modelo Híbrido. Os casos do Segmento 1 que não são classificados correctamente, são na quase totalidade dos casos classificados como Segmento 2, o que pode contribuir para um menor risco na classificação incorrecta, dado que o Segmento 2 é de facto aquele em que os utilizadores apresentam mais semelhanças em termos de comportamento futuro com os do Segmento 1.

A escolha por qualquer um dos modelos deve ser feita tendo em conta a importância de cada Segmento, o custo de não classificar correctamente um dos segmentos e a complexidade da classificação de novos casos, sendo que neste último critério os Modelos Híbridos estão em desvantagem. Para ajudar no processo de decisão de qual o modelo a escolher recomenda-se que em trabalhos futuros seja efectuada a aplicação dos Modelos obtidos para classificar os

Utilizadores Registados em 2008 e em 2009 e comparar as estruturas dos segmentos preditos com a segmentação observada nos Utilizadores Registados em 2007 e avaliar comparativamente a sua consistência interna. Ainda assim, antecipa-se que qualquer um dos modelos desenvolvidos permite obter excelentes resultados na classificação do Segmento a que pertencem os novos casos.

Para trabalhos futuros, não obstante os bons resultados obtidos com esta metodologia, é desejável que sempre que possível os dados de utilização sejam cruzados com informação sobre os conteúdos visualizados e as áreas do *site* por onde os utilizadores navegam e nas quais demonstram mais interesse. Tal não foi possível de efectuar nesta tese, dado que o acesso muito limitado à estrutura de construção do *site* e aos respectivos metadados não tornou possível a extracção desse tipo de informação.

A conjugação desta informação com informação sobre comportamentos de compra e de consumo *offline* dos clientes, ainda que mais difícil de obter, poderá também ter um contributo muito positivo para melhorar ainda mais os resultados. A possibilidade de estudar a evolução dos comportamentos *online* e *offline* dos consumidores e tentar estabelecer relações causais entre ambos os comportamentos poderá dar *insights* muito importantes para a gestão do *site*.

Por outro lado, poderá também ser interessante aprofundar esta metodologia no sentido de conseguir antecipar o momento a partir do qual é possível predizer os comportamentos futuros, ou seja, efectuar nova segmentação e respectivos Modelos de Classificação que permitam conseguir predizer os comportamentos nos Anos 1 e Anos 2, apenas com informação dos primeiros 6 meses, por exemplo.

No que respeita à aplicação dos métodos, quer de segmentação quer de classificação, pensa-se que será possível melhorar o desempenho desta metodologia de forma a melhorar a precisão e a accionabilidade dos Modelos de Classificação explorando uma nova estrutura de segmentos que permita diferenciar melhor o Segmento 1 dos restantes e assim obter um Segmento 1 que contenha menos elementos com características passíveis de serem classificados como Segmento 2.

Apesar das melhorias que são possíveis de aplicar em trabalhos futuros conclui-se que toda a metodologia aplicada nesta tese, desde o tratamento da informação contido originalmente na BD aos Modelos de Classificação, passando pela segmentação efectuada, constitui uma

ferramenta que pode ser aplicada noutros *sites*, mesmo de outros sectores de actividade, possuindo assim elevada relevância prática. Esta metodologia poderá assim contribuir para a estratégia de CRM, possibilitando criar políticas de Marketing que sejam geradoras de motivos de interesse e que consigam captar o retorno dos Utilizadores ao *Site* de forma continuada.

7. REFERÊNCIAS BIBLIOGRÁFICAS

Livros

Brown, S. A. (2000), *Customer Relationship Management - A Strategic Imperative in the world of e-business*. Ontario, Canadá: John Wiley & Sons Canada, Ltd.

Cardoso, M. M. (2006), *Textos não publicados sobre CART e Redes Neuronais*.

Ferrão, F. (2003), *CRM - Marketing e Tecnologia*. Lisboa, Portugal: Escolar Editora.

Fonseca, J. M. M. R. da (1994), Indução de Árvores de Decisão - HistClass - Proposta de um *algoritmo não paramétrico*. Departamento de Informática, Universidade Nova de Lisboa.

Gummesson, E. (2005), *Marketing de Relacionamento Total*. Porto Alegre, Brasil: Bookman.

Hair, J.F., Black, B., Babin, B., Anderson, R., Tatham, R. (2006), *Multivariate Data Analysis 6th edition*. New Jersey, EUA: Pearson.

Haykin, S. (1999), *Neural networks : a comprehensive foundation. 2ª edition*: Upper Saddle River : Prentice Hall.

Hughes, A. (2000), *Strategic Database Marketing*. : McGraw-Hill.

Lindon, D., Lendrevie, J.,Lévy, J.,Dionísio,P.,Rodrigues, J. (2004), *Mercator XXI*. Lisboa, Portugal: Dom Quixote.

Linnof, G. S. e Berry, M. J.A. (1997), *Data Mining Techniques for Marketing, Sales, and Customer Support*. Nova Iorque: Wiley Computer Publishing.

Linnof, G. S. e Berry, M. J. A. (2000), *Mastering Data Mining: The Art and Science of Customer Relationship Management*. Nova Iorque: John Wiley & Sons, Inc.

Linnof, G. S. e Berry, M. J. A. (2001), *Mining the Web - Transforming Customer Data into Customer Value*. Nova Iorque, EUA: John Wiley & Sons, Inc.

Peppers, D. e Rogers, M. (1999), *Le one to one en pratique*. : Éditions d'Organisation.

Pereira, BB (1999), *Introduction to Neural Networks in Statistics: Center of Multivariate Analysis*, Technical Report; Penn. State Univerisity.

Rud, O. P. (2001), *Data Mining Cookbook Modeling Data for Marketing, Risk, and Customer Relationship Management*. : John Wiley & Sons Canada, Ltd.

Yovits, M. C. (1993), *Advances in Computers Volume 37*. San Diego: Academic Press, Inc..

Periódicos científicos

Almeida, F. C. de; Siqueira, J. de O. e Onusic, L. M. (2005), Data Mining no Contexto de Customer Relationship Management. *Caderno de Pesquisas em Administração*, São Paulo, v. 12, n. 2, (pp. 85-97).

Antunes, J. e Rita, P. (), O marketing relacional como novo paradigma: uma análise conceptual. *Rev. Portuguesa e Brasileira de Gestão* 2008, Abril, (pp. 36-46).

Cabete, N.P. (2006), Algoritmo CART: Previsão do Desempenho na Matemática do Secundário. *Revista de Ciências da Computação* Volume I, Ano I, Nº1, (pp.).

Chiu, T.; Fang, D.; Chen, J.;Wang, Y.; Jeris, C., (2001), A robust and scalable clustering algorithm for mixed type attributes in large database environment. 7th *ACM SIGKDD Conference Proceedings*, (pp. 263-268).

Hosseini S.M.S. et al., (2010), Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications* 37, (pp. 5259–5264).

Jesus, N.B.; Cardoso, M.G.M.S., (2007), Análise de Agrupamento Incremental - Segmentação de Pontos de Retalho. *Revista de Ciências da Computação* Volume II, Ano II, Nº2, (pp.).

Ngai, E.W.T. et al., (2009), Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications* 36, (pp. 2592–2602).

Rygielski, C. et al., (2002), Data mining techniques for customer relationship management. *Technology in Society* 24, (pp. 483-502).

Santos et al., (2005), Usando Redes Neurais Artificiais e Regressão Logística na Predição da Hepatite A. *Revista Brasileira de Empidemiologia 8* (2), (pp. 117-126). "

Suh E. et al., (2004), A prediction model for the purchase probability of anonymous customers to support real time web marketing: a case study. *Expert Systems with Applications* 24, (pp. 245-255).

Referências não publicadas retiradas da internet

Wunderman, L., 3-6-2010, <http://www.wunderman.com>.

AMERICAN MARKETING ASSOCIATION, . Definition of Marketing , 3-6-2010, <http://www.marketingpower.com>.

8. APÊNDICE

Anexo I – Entrevista sobre a Importância dos *Sites* e dos Clubes de Fidelização para as marcas de Grande Consumo

Quando questionada acerca da importância dos *Sites* e dos Clubes de Fidelização para as marcas de Grande Consumo, Sofia Natal começa por diferenciar os conceitos, separando-os em duas componentes: Plataformas digitais (nos quais se incluem os *sites*) e Programas Relacionais (nos quais se incluem os Clubes de Fidelização). **"As Plataformas digitais e os Programas Relacionais são fundamentais para as marcas, nomeadamente, no sector do Grande Consumo. O digital veio, de facto, revolucionar as relações com os consumidores no seu potencial de interactividade e tempo de resposta. Os *sites* são uma vertente já bastante redutora daquilo que é hoje o potencial de presença e interacção de uma marca no online, na via digital. Actualmente, ter um *site* é uma situação «higiénica», um «*must have*» para ter significado no mundo das opções de compra. Por sua vez, os Programas Relacionais desempenham um papel verdadeiramente estratégico, com activos de *Intelligence* a partir de informação em base de dados, accionável. Estes programas, com planos de contactos próprios e segmentações orientadas para tipologias de consumo/consumidores procuram agir sobre todas as dimensões do ciclo de relacionamento com o consumidor, obviamente trabalhando na prossecução do já muito falado *Consumer Lifetime Value*, mas fazendo uma análise e acção holística."**

O desenvolvimento do *online* veio trazer **"uma nova dimensão na relação com os consumidores. Se por um lado o meio em si permite o que jamais no *offline* poderia traduzir em termos de interactividade, também a «actualização» do meio em si permite uma dinâmica arrasadoramente diferente das práticas tradicionais. Por outro lado, são notórias as economias de custos no envio de comunicação quando comparamos com correio físico (imagine-se a frequência de comunicação possível para 200.000 lares qualificados da Base de Dados Nestlé, a cada novidade de um portfolio de negócio como o desta Companhia).**

Neste contexto, como referido antes, **"os *sites* são a identificação da marca actualmente. Na Ogilvy, não há cliente que não tenha um *site* e novas propostas para *sites* são sempre propostas apresentadas estrategicamente de presença no online. Mais do que ter um *site*, hoje importa saber como estar ou não estar nas redes sociais, como ouvir o que as pessoas dizem na *web* sobre a marca e como reagir e influenciar. Ainda mais tacticamente, até como poder influenciar a visibilidade do *site* naquela grande «montra» que é actualmente o Google. Mais do que ter um *site*, que é «higiénico», importa saber como estar de uma maneira diferenciada, visível, e até, como criar uma experiência - mas claro, sempre alinhado com os objectivos da marca (e do marketing)."**

Apesar do crescente papel de outras plataformas digitais, **"a relação dos *sites* com os Programas Relacionais estabelece-se com frequência porque o *site* é actualmente o canal preferido de relação dos consumidores com a marca, em simultâneo com o e-mail marketing. É a plataforma mais cómoda e com a qual é mais fácil de interagir. Embora não seja verdade para todas as marcas, na grande maioria é. Do ponto de vista de *KPIs*[5] de uma marca, o *site* continua a ser o fornecedor *main*:**

- as visitas mostram a visibilidade, assim como as origens (posicionamento em motores de busca, que é um dos serviços também em crescimento para os *sites*);

- as *pageviews* e o tempo médio de visita mostram a interactividade e o interesse (de grosso modo)."

Acerca dos Programas Relacionais**,** Sofia Natal realça o papel da base de dados e do respectivo tratamento para o sucesso de um Programa Relacional: "**Nestes Programas o aspecto mais importante é a informação: obter o conhecimento do consumidor e encontrar um canal de comunicação frequente e autorizado. É fundamental que a comunicação seja segmentada por afinidade, consumo ou outra qualquer tipologia identificada, que nos vai permitir comunicar com maior relevância, e logo, com maior sucesso. A base de dados é um activo de valor incalculável para permitir programas de fidelização de sucesso. Mais do que dados, a qualidade e *insight* desses dados permitem accionar informação que mantenha as pessoas interessadas na marca e envolvidas com a mesma, na perspectiva de uma relação duradoura."**

[5] KPIs – Key Performance Indicators. São indicadores que traduzem o desempenho da empresa.

A título de exemplo, Sofia Natal refere o Programa de Fidelização da Nestlé[6] um dos Programas Relacionais com maior sucesso na área do Grande Consumo e que é desenvolvido pela *OgilvyOne*: **"Há mais de 7 anos que a Nestlé constrói uma verdadeira base de dados, contando hoje com mais de 1 milhão de Lares, embora em contactos de email tenha cerca de 250.000 disponibilidades. Facilmente consegue contactar com 250.000 pessoas em Portugal, da noite para o dia. O *site* é o principal canal de aquisição e contamos com dados muito elaborados sobre os nossos consumidores, principalmente, informação estratégica, tal como: se têm ou não filhos, idades dos residentes no lar, se têm ou não animais de estimação e chegamos até ao nível dos consumos. O Programa Relacional da Nestlé centra-se em dois aspectos fundamentais:**

- Base de dados: a base de dados é uma activo fundamental para qualquer marca/gestor da Nestlé, que dado o esforço da Companhia, pode usufruir desta visibilidade; é um activo dificilmente copiado pela concorrência!

- Relação e conteúdo que oferece online: o *site* nestle.pt faz parte do Programa, enquanto plataforma e tem uma visibilidade fantástica. O *site* da Nestlé funciona ainda assim como "portal", para aquilo que definimos como *affinities* naturais, cada uma com *site* específicos (bebés, crianças, animais, culinária e bem estar); e tem tido um sucesso enorme, é uma verdadeira plataforma de visibilidade em massa.

Para além do *site* nestle.pt, utilizamos outras plataformas digitais de comunicação com os consumidores, tais como, e-News sazonais que são enviadas para a BD, *sites* de afinidade mencionados anteriormente (em que o mais importante é o de culinária), enviamos e-mails marketing com regularidade e temos uma página no Facebook com cerca de 13.000 amigos (criada este ano) e que tem gerado um tráfego fantástico."

Inovando na utilização das plataformas digitais, a *OgilvyOne* desenvolveu a primeira rede social de animais. A PETNET (do cliente Mars[7]) constitui um autêntico sucesso e é um programa relacional que diferencia a marca da concorrência e **"quem lá chegasse primeiro, ganhava o espaço, o conceito"** refere Sofia Natal**.** Outro exemplo de inovação na utilização das plataformas digitais foi a criação do primeiro reality show online através de redes sociais,

[6] Não existe qualquer relação entre este exemplo e a BD utilizada neste estudo.
[7] A MARS Inc. actua na área da alimentação para animais de estimação através das marcas Whiskas, Pedigree, Sheba, entre outras.

o ICONS, para o lançamento do Ford Fiesta que visava atingir o público jovem. Esta acção ganhou o primeiro prémio do Sapo 2009 na categoria de Eficácia.

Ainda que não possuam estudos que permitam ter dados concretos sobre a relação entre menor ou maior utilização dos *sites* por parte do consumidor e a relação deste com a marca (quantidade, repetição ou frequência de compra e satisfação ou fidelização à marca), Sofia Natal refere que **"se falarmos em bens alimentares, começamos a assistir que há cada vez mais pessoas a procurar, por exemplo, informação alimentar/nutricional, para tomar as suas decisões de compra. Acreditamos sobretudo que, pelo potencial de visibilidade e interacção, um *site* pode conseguir estabelecer uma boa relação com um consumidor. Trabalhando-se muito bem o posicionamento e os valores da marca, tendo por trás uma estratégia de relacionamento, conseguir-se-á, sobretudo, a fidelização do consumidor. Temos casos de clientes que quiseram apenas lançar um *site* informativo, e que estão lá, mas poucos os vêem ...e poucos voltam. O segredo está em explorar o potencial da plataforma *site* enquanto veículo de fácil actualização e de relação com o consumidor. E, claro, nunca esquecer que ele está numa «prateleira» chamada Google."**

Os principais desafios que se colocam aos gestores destas plataformas digitais para que consigam maximizar o seu potencial, segundo Sofia Natal, são três:

- **"Visão mais holística das plataformas digitais, não centrada apenas no *site* e com partilha de informação.**
- **Estratégias Relacionais, online ou 360º, que implicam base de dados, *Intelligence* e planos de contacto.**
- **Sensibilidade para o investimento em *Search Engine Optimization*[8]."**

Para desenvolvimento dos Programas Relacionais, as metodologias de *Data Mining* são utilizadas sobretudo para **"reconhecimento de Grupos de Afinidade não naturais. Os naturais seriam por exemplo, os bebés, a culinária, os animais, as crianças. Já fizemos isso para a Nestlé, por exemplo, a partir de um questionário muito completo que sempre tivemos. Conseguimos analisar comportamentos de consumo e consumos efectivos de produtos em todo o portfólio e conseguimos extrair oito grupos de afinidade para os quais faria sentido comunicar de maneira diferenciada, pois tinham necessidades e**

[8] Search Engine Optimization (SEO) é o processo de incrementar a visibilidade de um site ou página de internet nos motores de busca através da via natural ou não paga ("orgânica" ou "algorítmica").

motivações diferenciadas. Por outro lado, também temos matrizes de valor – fidelização, situando os grupos em quadrantes, desenhando acções especificas para cada grupo, no sentido de incentivar as migrações de quadrante, para maior valor e claro, maior fidelização."

Em resumo, "**para uma parte importante dos clientes da *OgilvyOne*, e dado o perfil da agência, os *sites* são uma situação higiénica, onde a imagem impera. Existe muita procura obviamente para *sites* de campanha, com duração limitada. Do ponto de vista de médio longo prazo, estrategicamente, cada vez se procuram mais *sites* que assentam em Programas Relacionais."**

Como área de desenvolvimento futuro dos Programas Relacionais está previsto o *reward* dos consumidores mais activos nas plataformas digitais. Assim, é previsível a implementação de técnicas de *data mining* e/ou de clusterização, com base nos comportamentos dos consumidores na utilização das plataformas digitais.

Anexo II – Novas variáveis criadas a partir das variáveis originais

Variável	Descrição
Status_Ano0	Status Activação e Visita Ano 0
Status_de_Activação_Ano0	Status de Activação no Ano 0
Status_de_Visita_Ano0	Status de Visita no Ano 0
Status	Status Activação e Visita
Status_de_Activação	Status Activação
Status_de_Visita	Status Visita
dias_ate_activacao	Nº de Dias entre Registo e Activação
Idade_cod	Idade
Distrito_COD	Distrito
Região_Cod	Região
Ano_de_Registo	Ano de Registo
Mês_de_Registo	Mês de Registo
dias_ate_activacao_cod	Nº de Dias entre Registo e Activação
N_de_sessoes_2007	Nº de Sessões Site e/ou Email_ 2007
N_de_sessoes_2008	Nº de Sessões Site e/ou Email_ 2008
N_de_sessoes_2009	Nº de Sessões Site e/ou Email_ 2009
N_de_sessoes_SóSite_2007	Nº de Sessões Só Site_ 2007
N_de_sessoes_SóSite_2008	Nº de Sessões Só Site_ 2008
N_de_sessoes_SóSite_2009	Nº de Sessões Só Site_ 2009
Activ_2007	Tipo de Sessão 2007
Activ_2008	Tipo de Sessão 2008
Activ_2009	Tipo de Sessão 2009
Activ_Ano0	Tipo de Sessão Ano 0
Activ_Ano1	Tipo de Sessão Ano 1
Activ_Ano2	Tipo de Sessão Ano 2
N_de_sessoes_Ano0	Nº de Sessões Site e/ou Email_ Ano 0
N_de_sessoes_ano0_cod	Nº de Sessões Site e/ou Email_ Ano 0
N_de_sessoes_Ano1	Nº de Sessões Site e/ou Email _ Ano 1
N_de_sessoes_ano1_cod	Nº de Sessões Site e/ou Email _ Ano 1
N_de_sessoes_Ano2	Nº de Sessões Site e/ou Email _ Ano 2
N_de_sessoes_ano2_cod	Nº de Sessões Site e/ou Email _ Ano 2
N_de_sessoes_SóSite_Ano0	Nº de Sessões _ Só Site Ano 0
N_de_sessoes_SóSite_ano0_cod	Nº de Sessões_Só Site _ Ano 0
N_de_sessoes_SóSite_Ano1	Nº de Sessões_Só Site _ Ano 1
N_de_sessoes_SóSite_ano1_cod	Nº de Sessões_Só Site _ Ano 1
N_de_sessoes_SóSite_Ano2	Nº de Sessões_Só Site _ Ano 2
N_de_sessoes_SóSite_ano2_cod	Nº de Sessões_Só Site _ Ano 2
Ano0	Status Informação Ano0
Ano1	Status Informação Ano1
Ano2	Status Informação Ano2
N_de_sessoes_total	Nº de Sessões Site e Email 2007 a 2009
N_de_sessoes_total_cod	Nº de Sessões Site e Email 2007 a 2009
N_de_sessoes_SóSite_total	Nº de Sessões _SóSite 2007 a 2009
N_de_sessoes_SóSite_total_cod	Nº de Sessões SóSite 2007 a 2009
paginas_por_sessao_Ano0	Número médio de logs por sessão no Ano0
paginas_por_sessao_Ano1	Número médio de logs por sessão no Ano1
paginas_por_sessao_Ano2	Número médio de logs por sessão no Ano2
paginas_por_sessao_Ano2_completo	Número médio de logs por sessão no Ano2 excepto Ano Incompleto
N_de_sessoes_Ano0_Trim1	Número médio de sessões Site e/ou Email no Ano0 Trim1
N_de_sessoes_Ano0_Trim2	Número médio de sessões Site e/ou Email no Ano0 Trim2
N_de_sessoes_Ano0_Trim3	Número médio de sessões Site e/ou Email no Ano0 Trim3
N_de_sessoes_Ano0_Trim4	Número médio de sessões Site e/ou Email no Ano0 Trim4
N_de_sessoes_SóSite_Ano0_Trim1	Número médio de sessões Só Site no Ano0 Trim1
N_de_sessoes_SóSite_Ano0_Trim2	Número médio de sessões Só Site no Ano0 Trim2
N_de_sessoes_SóSite_Ano0_Trim3	Número médio de sessões Só Site no Ano0 Trim3
N_de_sessoes_SóSite_Ano0_Trim4	Número médio de sessões Só Site no Ano0 Trim4

Anexo III – Exemplo de aplicação do Modelo Híbrido

	Árvore de Decisão 2					Rede Neuronal 6					Modelo Híbrido					Segmento Observado
	Probabilidade de Pertencer ao…				Segmento Predito	Probabilidade de Pertencer ao…				Segmento Predito	Probabilidade de Pertencer ao…				Segmento Predito	
	Segmento 1	Segmento 2	Segmento 3	Segmento 4		Segmento 1	Segmento 2	Segmento 3	Segmento 4		Segmento 1	Segmento 2	Segmento 3	Segmento 4		
Caso 1	0,35	0,65	0,00	0,00	2	0,75	0,17	-0,02	0,00	1	0,55	0,41	-0,01	0,00	1	1
Caso 2	0,01	0,99	0,00	0,00	2	0,00	0,95	-0,01	0,01	2	0,01	0,97	0,00	0,01	2	2
Caso 3	0,01	0,99	0,00	0,00	2	0,04	0,94	0,00	0,01	2	0,03	0,96	0,00	0,00	2	2
Caso 4	0,00	0,00	1,00	0,00	3	0,00	0,00	0,99	0,00	3	0,00	0,00	1,00	0,00	3	3
Caso 5	0,00	0,00	1,00	0,00	3	0,00	0,00	0,99	0,00	3	0,00	0,00	1,00	0,00	3	3
Caso 6	0,00	0,00	1,00	0,00	3	0,00	0,00	0,99	0,00	3	0,00	0,00	1,00	0,00	3	3
Caso 7	0,02	0,00	0,00	0,98	4	0,08	-0,04	0,00	0,97	4	0,05	-0,02	0,00	0,98	4	4
Caso 8	0,11	0,22	0,00	0,67	4	0,03	0,95	0,00	0,01	2	0,07	0,59	0,00	0,34	2	4
Caso 9	0,02	0,00	0,00	0,98	4	0,13	-0,04	0,01	0,96	4	0,07	-0,02	0,01	0,97	4	4
Caso 10	0,02	0,00	0,00	0,98	4	-0,05	0,04	0,01	0,98	4	-0,01	0,02	0,00	0,98	4	4

	Árvore de Decisão 2					Rede Neuronal 7					Modelo Híbrido					Segmento Observado
	Probabilidade de Pertencer ao…				Segmento Predito	Probabilidade de Pertencer ao…				Segmento Predito	Probabilidade de Pertencer ao…				Segmento Predito	
	Segmento 1	Segmento 2	Segmento 3	Segmento 4		Segmento 1	Segmento 2	Segmento 3	Segmento 4		Segmento 1	Segmento 2	Segmento 3	Segmento 4		
Caso 1	0,35	0,65	0,00	0,00	2	0,76	0,11	-0,01	0,00	1	0,55	0,38	-0,01	0,00	1	1
Caso 2	0,01	0,99	0,00	0,00	2	-0,03	0,93	0,02	0,03	2	-0,01	0,96	0,01	0,02	2	2
Caso 3	0,01	0,99	0,00	0,00	2	0,01	0,92	0,00	0,02	2	0,01	0,96	0,00	0,01	2	2
Caso 4	0,00	0,00	1,00	0,00	3	0,00	0,00	0,99	0,00	3	0,00	0,00	0,99	0,00	3	3
Caso 5	0,00	0,00	1,00	0,00	3	0,00	0,00	0,99	0,00	3	0,00	0,00	0,99	0,00	3	3
Caso 6	0,00	0,00	1,00	0,00	3	0,00	0,00	0,99	0,00	3	0,00	0,00	0,99	0,00	3	3
Caso 7	0,02	0,00	0,00	0,98	4	0,01	-0,04	0,02	0,98	4	0,01	-0,02	0,01	0,98	4	4
Caso 8	0,11	0,22	0,00	0,67	4	-0,03	0,93	0,01	0,03	2	0,04	0,58	0,00	0,35	2	4
Caso 9	0,02	0,00	0,00	0,98	4	-0,04	0,11	0,00	0,98	4	-0,01	0,06	0,00	0,98	4	4
Caso 10	0,02	0,00	0,00	0,98	4	0,01	-0,03	-0,01	0,99	4	0,01	-0,02	0,00	0,99	4	4