Determinação de Padrões de Desistência em Ginásios

Paulo Pinheiro
Licenciado em Informática, Univ. Aberta, ppinheiro@cedis.pt

Luís Cavique
Univ. Aberta, lcavique@uab.pt

Resumo

O problema da retenção e das causas que levam à desistência da frequência do ginásio é uma questão que os ginásios e academias de *fitness* tentam há muito entender e consequentemente evitar. Tendo em atenção que a indústria do Fitness tem apostado na instalação massificada de sistemas avançados de *Customer Relationship Management* (CRM) existem atualmente bases de dados com dados históricos de grande valia. Este projeto tem por objetivo aplicar técnicas de classificação a estas bases de dados de forma a encontrar um método adequado para determinar padrões que permitam prever quais os utentes que irão abandonar ou cancelar a sua inscrição num período próximo, atendendo ao padrão do perfil de comportamento dos utentes que desistiram nos últimos meses, e criar uma ferramenta que forneça informação que permita aos gestores desses ginásios tomar medidas que permitam prolongar a duração da frequência dos utentes.

Palavras-chave: retenção, data mining, marketing, fitness

Abstract

The retention problem and consequent causes that lead gym and fitness clubs users to cancel their membership is an issue to understand and consequently avoid. Bearing in mind that the fitness industry has focused on mass installation of Customer Relationship Management systems (CRM) there are currently databases with valuable historical data that allow us to study users' behavior. This project aims to apply classification techniques to those databases in order to find a suitable algorithm that will identify patterns allowing us to predict which users will abandon or cancel their membership in a short term, given the profile and behavior patterns of users that have quit in the last months, and create a tool that provides information that allows gym managers to take measures to extend users' membership.

Keywords: churn, data mining, marketing, fitness

1. Introdução

O projeto tem por principal objetivo utilizar técnicas de Classificação para encontrar o algoritmo e o conjunto de atributos mais adequado à determinação de padrões que permitam prever quais os utentes que irão abandonar ou cancelar a sua inscrição num próximo período, tendo por base o perfil-padrão dos utentes que desistiram nos últimos meses, definido a partir de características como o consumo, a forma de utilização e eventuais dados demográficos que se venham a considerar relevantes.

Como segundo objetivo, pretende-se construir uma plataforma que permita automatizar o processo de atualização do padrão, de forma regular, através do modelo escolhido para que seja possível por comparação determinar quais os utentes ativos atuais que se enquadram nesse perfil.

O desenvolvimento deste projeto foi efetuado em duas fases distintas. Numa primeira fase seguiram-se as orientações da metodologia CRISP-DM para determinar o modelo de *data mining* mais adequado ao objetivo que se pretende atingir. Numa segunda fase procedeu-se ao desenvolvimento da aplicação que permite atualizar o modelo com novos dados, apreciar a exatidão do mesmo e produzir uma lista de potenciais desistentes a partir de uma base de dados.

Este artigo está estruturado em 7 secções. Na secção 2 são apresentados os modelos RFM e RM utilizados como ponto de partida para a construção do modelo. Como referido acima, seguiu-se a metodologia CRISP-DM na determinação do modelo. No entanto, para melhor clareza do artigo explica-se na secção 3 as fases de compreensão do negócio e dos dados e o processo ETL aplicado e na secção 4 são apresentados os modelos testados e resultados obtidos. Na secção 5 analisa-se o overfitting dos modelos. Na secção 6 apresenta-se a aplicação desenvolvida e finalmente na secção 7 são elaboradas as conclusões.

2. Bases de Dados de Clientes

Na caracterização de clientes são geralmente utilizados dados relacionados com o indivíduo, que incluem dados sócio-demográficos, geográficos ou de estilo de vida. Para obter este tipo de dados são requeridos dispendiosos inquéritos. Por outro lado, os dados de consumo existem na contabilidade das empresas, na maior parte das vezes basta associar os produtos comprados a um determinado cliente. Neste trabalho o modelo apresentado utiliza exclusivamente dados de consumo, disponíveis em qualquer sistema de informação empresarial, evitando as demoras e os custos dos dados resultantes de inquéritos.

2.1. Base de Dados ou Aplicações para Clientes

Com o apoio das Base de Dados ou Aplicações para Clientes, pretende-se construir relações lucrativas e duradouras, ao comunicar com o cliente certo, utilizando o produto certo, com a mensagem certa (emitida no momento certo e através do canal certo) (Cavique e Themido, 2002). Deste objetivo, distinguem claramente três tipos de conjuntos de dados: os dados do cliente, os dados da compra do cliente e os dados relativos à comunicação com o

cliente (Brito, 2000). Para responder a esta solicitação existe uma estrutura de dados, como se apresenta em seguida:

i) Dados do Cliente: nome, contactos, contato (ou canal) preferencial, datas mágicas, dados demográficos, sociográficos e psicográficos
ii) Compras do Cliente: data, produto, quantidade, preço, forma de pagamento, vendedor, descontos, devoluções, reclamações, ofertas, pontos
iii) Comunicações com o Cliente : data-hora, origem, destino, assunto, conteúdo da mensagem, canal.

As Bases de Dados de Clientes para além de registarem os dados do cliente e das suas compras, como qualquer sistema de contabilidade, integra o registo das comunicações que darão suporte a programas de fidelização.

Para o sucesso das aplicações de cliente, a qualidade dos dados e a agregação dos mesmos são dois fatores críticos. Ao utilizar dados de diversas fontes, a garantia da qualidade dos dados é uma tarefa de extrema importância no que diz respeito à uniformização dos formatos dos dados e à remoção de dados duplicados nos atributos críticos da aplicação. Este tipo de tarefas foram durante muito tempo negligenciadas, estando atualmente a tomar a sua devida importância nas aplicações de clientes. Por outro lado os dados devem estar devidamente consolidados e estáveis por forma a facilitar as pesquisas. Este tipo de conceito desenvolvido para ambientes de Data Warehouse é também necessário nas Bases de Dados de Clientes, permitindo a agregação dos dados e a capacidade de absorver periodicamente novos dado.

2.2. Critérios RFM

Por forma a aumentar as baixas taxas de resposta, em Database Marketing é usual recorrer à técnica de segmentação RFM, onde R (recentidade ou qualidade de ser recente) é dada pela última data da visita à loja, F representa a frequência de compras na loja e M o valor monetário global do cliente.

Existem várias formas de calcular os valores para RFM, neste trabalho vamos utilizar a especificação de Hughes (2000). Se escolhermos para cada atributo RFM, 5 classes, com os números de 1 a 5, obtemos 125 classificações diferentes. Assim o cliente 555 é um cliente muito recente, muito frequente e com um alto volume de compras, enquanto que um cliente 111 é pouco recente, pouco frequente e com baixo volume de compras. Existe uma possibilidade de hierarquizar as classificações, seguindo a valorização que é comum dar aos números inteiros. O critério de segmentação começa por ordenar a tabela de clientes de forma crescente pela data da última compra, num segundo passo classifica os primeiros n/5 clientes com o número 1, os segundos n/5 clientes com o número 2 e assim sucessivamente até ao número 5. O processo repete-se para os atributos da frequência e valor monetário. A concatenação dos três atributos resulta a classificação RFM de cada cliente. Os clientes a selecionar para cada campanha (ou micro-campanha) são os n-ésimos primeiros clientes com maior valor de RFM, permitindo uma única estratégia.

Com a expansão do CRM é introduzida uma nova medida para os clientes, o "customer lifetime value" (CLV). Esta medida é calculada com base no lucro resultante do total de transações dos clientes durante o seu período de vida. Desta forma o CLV devolve valor financeiro de um cliente médio, tornando-se difícil prever a atitude de cada cliente individualmente.

Como forma de ultrapassar as referidas desvantagens, apresentamos um novo critério de segmentação de clientes inspirado nas vantagens dos critérios RFM e CLV. A medida mais importante para o critério RFM é a recentidade, por outro lado a medida do CLV é o valor monetário. O critério proposto RM, cruza as variáveis recentidade e valor monetário, permitindo definir quatro estratégias diferentes de comunicação com o cliente.

Para obter a classificação RM [Cavique 2003] basta ordenar pela data mais recente e afetar R como foi definido anteriormente. O processo repete-se para classificar o valor monetário M. O resultado é uma matriz RM, cujas células têm valores diferentes de clientes. Depois de classificar os clientes com base na recentidade e na frequência, podemos realizar a análise segundo as duas variáveis em simultâneo, dividindo os clientes em quatro quadrantes por forma a adotar estratégias diferentes para cada grupo. Nesta abordagem cada sub-grupo terá associado um verbo que identifica a estratégia a utilizar, tal como: premiar, reter, estimular, esquecer.

3. Compreender e preparar os dados

3.1. Compreender os dados

A base de dados que serviu de suporte a este trabalho foi selecionada de forma a reduzir os constrangimentos com base no volume de dados e de alargar as possibilidades de escolha de atributos para o problema em questão. A base de dados escolhida apresenta um volume de dados adequado, registos de controlo de acessos e de frequência das atividades e conjuga informação de 13 instalações desportivas nos distritos de Braga, Lisboa, Faro, Porto e Viana do Castelo.

Esta base de dados, implementada em Microsoft SQL*Server®, contém toda a informação do negócio das instalações, quer a referente ao *core-business* – atividades de *fitness* -, como também a informação relativa a outras atividades e serviços desportivos que são prestados nas várias instalações. É assim necessário separar os indivíduos relevantes para a aplicação do modelo – aqueles que estão (ou estiveram) inscritos em atividades classificadas como receita principal (*fitness*) - dos que não o são.

Por outro lado, é necessário ter em atenção os atributos que podem ser utilizados pelo modelo, quer por serem referenciados em literatura especializada, quer por estarem disponíveis na base de dados. Neste aspecto, tomou-se como ponto de partida os atributos apresentados na Tabela 1 que podem ser obtidos diretamente ou por aplicação de transformações.

Tabela 1 - Atributos promissores identificados na observação da base de dados

Nº	Atributo a considerar	Justificação
1	Situação atual	Estado da inscrição do Utente. Este é o atributo que se pretende prever.
2	Idade do Utente	Idade do Utente, em anos
3	Género do Utente	Sexo do utente
4	Antiguidade da inscrição	Duração da inscrição, em meses
5	Frequência média de visitas ao clube	Frequência média semanal de visitas ao clube – utilizado como o atributo base para a "Frequência" – o "F" do método RFM
6	Volume de negócios	Volume de negócios total – utilizado como o atributo base para o "Monetário" – o M dos métodos RFM e RM
7	Número de dias sem visitar o clube	Último intervalo de dias em que o utente não frequentou o clube – utilizado como o atributo base para a "Recentidade" – o "R" dos métodos RFM e RM
8	Frequência de Aulas de Grupo	Número total de aulas de grupo frequentadas
9	Número de contatos entre o Clube e o Utente	Contagem do número de registos de contatos
10	Reclamações efetuadas pelo Utente	Contagem do número de registos de reclamações/queixas, por gravidade
11	Familiares ou amigos que frequentam o clube	Contagem do número de familiares associados e de referências dadas
12	Distância a percorrer para chegar ao clube	Conteúdo do campo adicional destinado ao efeito e que contém a distância – em tempo ou quilómetros – a percorrer a partir do local de onde o utente sai para se deslocar ao Ginásio

Para todos os cálculos considera-se como data final a data de aplicação do modelo se a inscrição ainda estiver em vigor, ou a data de cancelamento da inscrição caso contrário.
Constatou-se que os atributos 9, 10, 11 e 12 não se encontravam preenchidos na base de dados pelo que não foram considerados pelos modelos estudados. Os restantes atributos são obtidos a partir de 4 tabelas principais cuja dimensão, em termos de número de registos, é apresentada na Tabela 2.

Tabela 2 – Dimensão, em número de registos, das tabelas que contêm ou a partir das quais se obtêm os atributos promissores, à data da obtenção da base de dados (31/Março/2015)

Tabela	Tipo de conteúdo	Nº de registos
Clientes	Contém dados genéricos e demográficos sobre todos os utentes do Ginásio	165.370
Transição de estados	Contém as transições de estado dos utentes enquanto clientes do Ginásio	1.713.800
Movimentos de acesso	Contém os registos de acesso e de presenças em aulas de grupo	20.083.391
Documentos	Contém o cabeçalho dos documentos de faturação emitidos	1.518.583

A implementação e teste dos modelos foi efetuada com o *Microsoft SQL*Server 2014®*, *Analysis Services Ver. 12.0.2000.8* e *R Ver. 3.1.3* da *The R Foundation for Statistical Computing*. Estas ferramentas foram utilizadas quer na fase de preparação dos dados, quer na fase de implementação dos modelos para comparação de resultados.

3.2. Transformar dos dados

Nesta fase Extract-Transform-Load (ETL) procedeu-se à extração da informação da base de dados origem, transferindo-se para uma outra base de dados separada de forma a que qualquer operação de transformação não altere os dados originais, procedeu-se ao calculo de novos valores necessários, isolou-se e procedeu-se à correção de erros de forma a garantir a integridade e as regras de negócio, tal como se apresenta nos parágrafos seguintes.

Dada a dimensão das tabelas e à necessidade de analisar e eventualmente corrigir alguns dados, optou-se por criar uma estrutura de dados separada para suporte do modelo de *data mining*, apresentada na Figura 1.

São assim levadas a cabo diversas operações de seleção, integração e limpeza, operações de transformação e operações de redução que se refletem numa tabela única ("dmRetencao"), e que pode ser eventualmente alojada noutra base de dados e/ou noutro servidor. Esta abordagem, para além de simplificar a aplicação dos modelos de *Data Mining*, reduz o volume de dados a tratar resultando por isso em menores tempos de execução dos modelos.

É de realçar que se pretende que os padrões de desistência evoluam ao longo do tempo, em resposta às ações implementadas e com as alterações de comportamento dos utentes no que diz respeito aos atributos considerados. Por isso, após a identificação das operações a efetuar sobre os dados e após a deteção de todas as falhas na qualidade dos dados, é implementado um procedimento ("dmPrepareData") que executa todas as operações requeridas e que pode ser agendado para execução periódica, atualizando os dados de suporte ao modelo escolhido.

Nesta fase iniciou-se por implementar procedimentos na base de dados que transferem os dados entre a base de dados "CLIENTE" e a base de dados "DW" e que efetuam os cálculos necessários a partir dos atributos simples disponíveis na base de dados.

3.3. Limpeza dos dados

Após a execução dos referidos procedimentos a tabela "dmRetencao" apresentava um total de 49.875 registos. A diferença entre o número de utentes na base de dados do cliente (165.370) e o número obtido ocorre pelo fato de nas operações de transferência efetuadas terem sido aplicados alguns filtros, nomeadamente referente aos utentes que nunca frequentaram atividades de *fitness* ou que já se tinham tornado desistentes à data inicial considerada para este projeto (1 de Agosto de 2012).

Após a transferência dos dados, iniciou-se a fase de deteção de valores em falta ou duplicados, inconsistências (*noise*) óbvias ou menos óbvias, e á deteção de valores com características consideravelmente diferentes (*outliers*). Através do R, obteve-se sumários

dos atributos carregados no *dataframe* a partir da tabela "dmRetencao" apresentado na Figura 2.

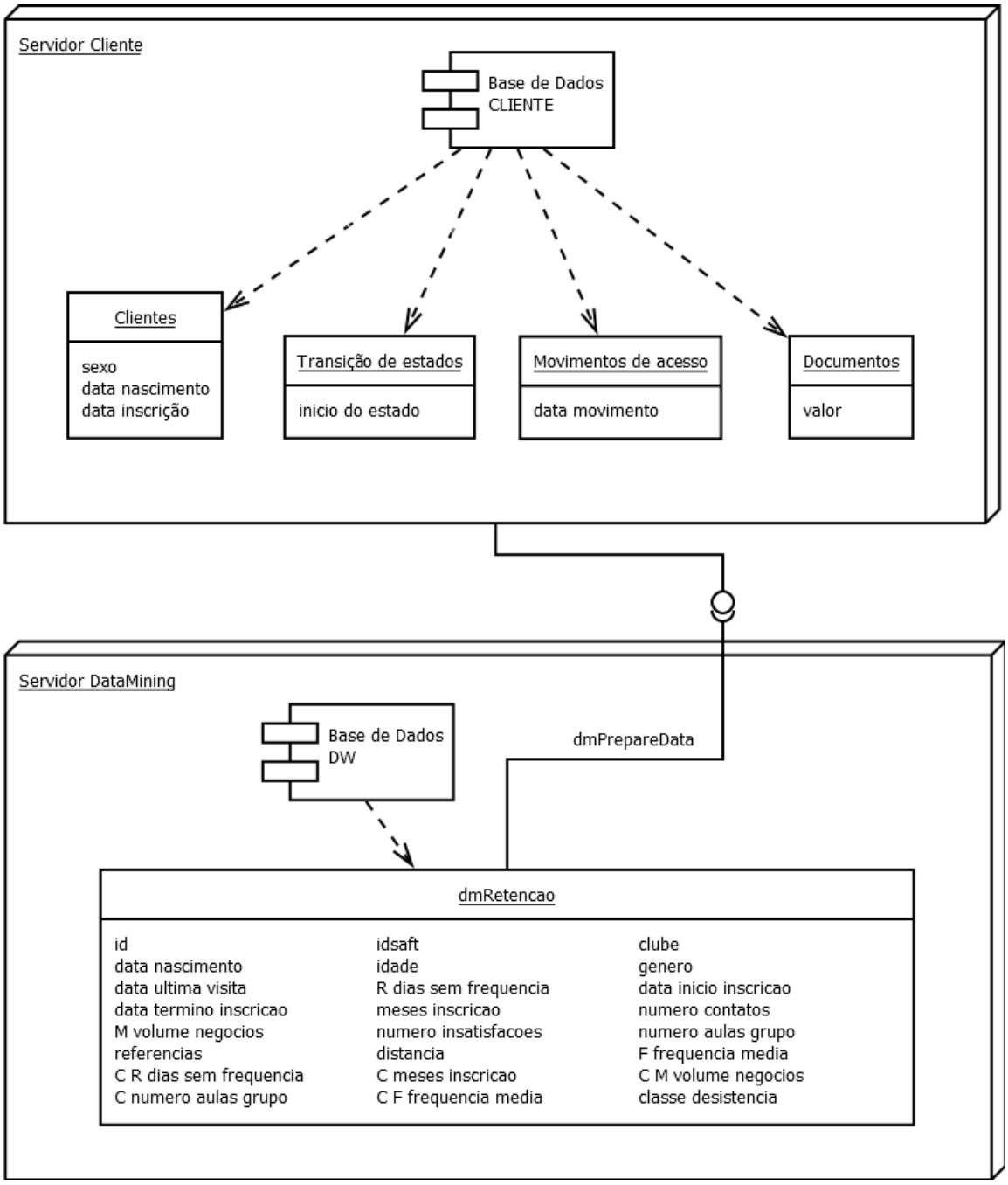


Figura 1 – Diagrama de arquitetura do projeto de Base de Dados

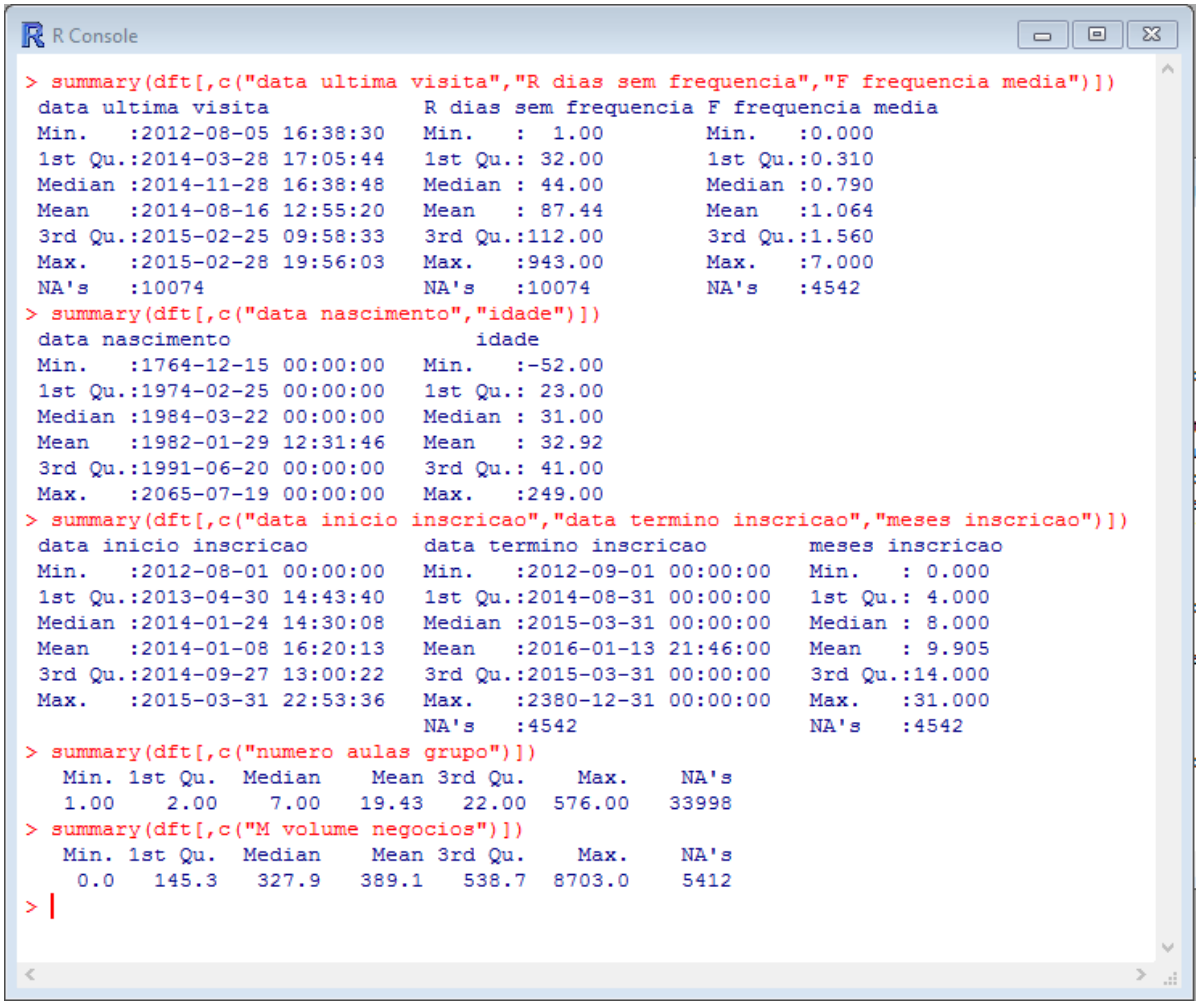


Figura 2 – Sumário das características e quantis dos atributos promissores

A observação dos sumários obtidos permite constatar a existência de vários atributos sem valores disponíveis (*NA's*) pelo que a estratégia seguida, dependendo do atributo, passou pela eliminação direta de registos (registos sem data de nascimento, sem data de termino ou classe de desistência), eliminação por cruzamento de dados (registos sem volume de negócios e em que a data do último período pago é anterior à data de inicio, ou registos sem presenças e sem volume de negócios) ou pela atribuição de valores (atribuição do valor zero ao número de aulas de grupo frequentadas).

Após as operações referidas, os sumários em R deixaram de apresentar atributos com *NA's* pelo que não se aplicou mais nenhuma outra técnica de resolução como a de imputação do valor médio, determinação por interpolação ou regressão linear, "vizinho mais próximo" ou outras.

Importa no entanto observar o sumário dos atributos, agora sem NA's, no intuito de procurar atributos inconsistentes ou que contenham valores que por natureza da atividade não devam ser considerados (*outliers*).

A simples observação do atributo relativo à data de nascimento no sumário R permite concluir de imediato que não é provável haver pessoas a frequentar a atividade de *fitness* com 249 anos de idade ou com -52 anos. No entanto, não será tão claro observar *outliers* nos outros atributos pelo que se optou pela produção dos gráficos *box-and-whiskers* e de histogramas com base nos valores dos diferentes atributos, para realçar as possíveis inconsistências e dispersão de valores.

No que concerne à idade, e ao atributo base correspondente (data de nascimento) foram consideradas idades corretas entre os 16 e os 93 anos de idade, tendo sido atribuído NULL aos registos com valores fora deste intervalo.

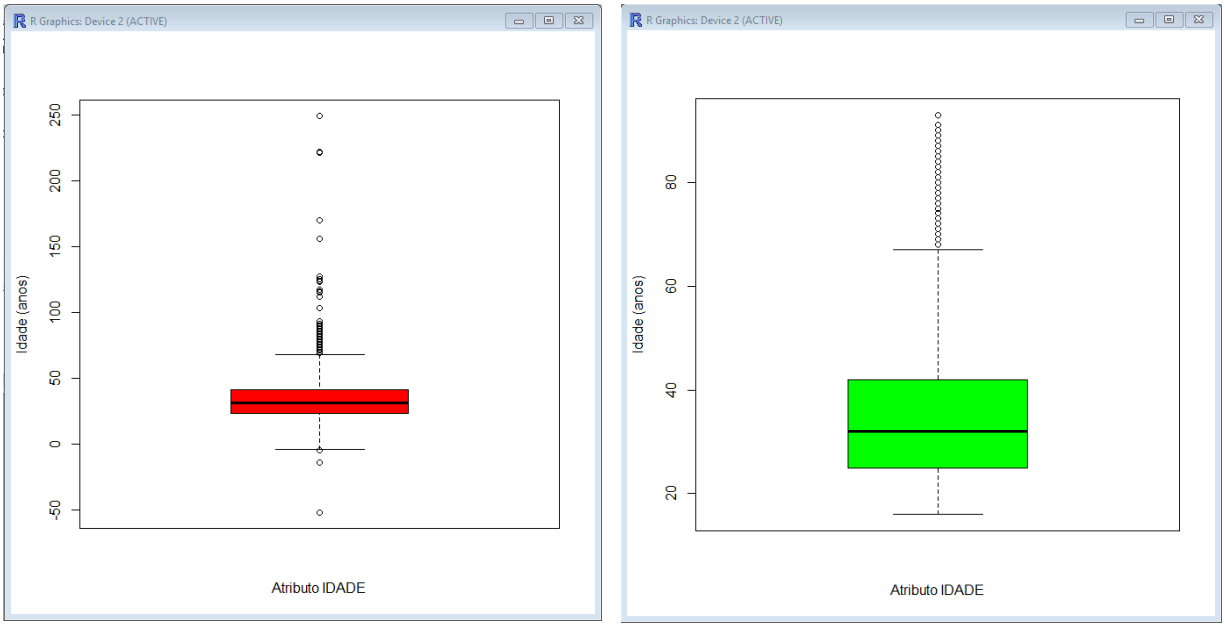


Figura 3 – Gráfico box-and-whiskers do atributo idade
a) Antes da limpeza **b)** Após a limpeza

A análise do sumário R e dos gráficos *box-and-whiskers* dos restantes atributos levou às seguintes conclusões:
- o atributo "R dias sem frequencia" apresenta 11.107 registos acima do 3º quartil (mais de 112 dias sem frequência), sendo que 10.069 são inscrições com menos de 365 dias sem frequência, e 986 são inscrições com menos de 730 dias sem frequência; nenhuma operação foi efetuada sobre este atributo dado se considerar não haver valores irregulares uma vez que em inúmeras situações, os utentes assinam contratos de fidelização de 12 ou 24 meses assumindo durante este período uma prática muito irregular acabando mesmo por deixar de frequentar o ginásio;
- dado o período considerado para este projeto (01 de agosto de 2012 a 31 de março de 2015), o atributo "meses inscricao" não apresenta nenhuma duração anómala uma vez que não foram encontrados registos com duração superior a 32 meses;
- apesar do atributo "M volume negocios" apresentar alguns registos acima do 3º quartil (538,20€), apenas 103 utentes apresentam um valor médio pago por mês acima dos

100,00€, o que não é um valor exagerado tendo em atenção alguns serviços adicionados, nomeadamente o serviço de treinador pessoal;
- o atributo "F frequencia media" apresenta um número razoável de valores acima do 3º quartil (média de frequência superior a 1,56/semana) pelo que se procurou determinar qual o número de utentes que apresentavam uma frequência média igual ou superior a 2/semana tendo-se obtido os valores indicados na Tabela 3. Também aqui não se depreende haver valores fora do normal;

Tabela 3 – Nº de utentes com frequência média semanal superior a 2/semana

Média de frequência	**Nº de Utentes**
>=2	4909
>=3	1766
>=4	572
>=5	130
>=6	31
>=7	1

- finalmente, o gráfico correspondente ao atributo "numero aulas grupo" apresenta um aspeto atípico (Figura 4), mas a execução de um *query* sobre a base de dados relacionando o número de aulas de grupo frequentadas com o número de meses da inscrição resultou em que apenas 27 utentes frequentam 7 aulas de grupo por semana. A observação manual dos extratos de presenças destes utentes confirma os números calculados, pelo que também aqui nada há a corrigir.

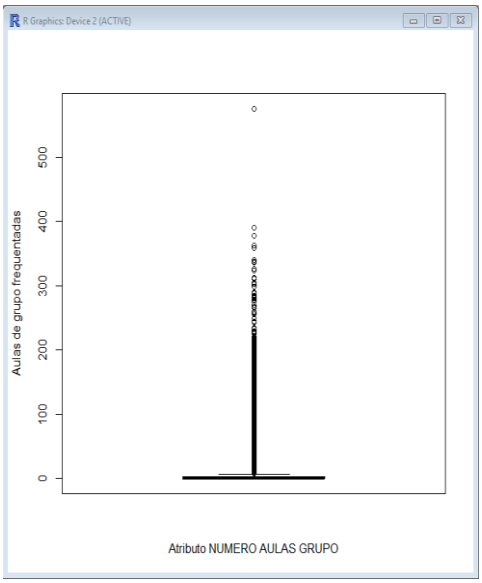


Figura 4 – Gráfico *box-and-whiskers* do atributo "numero aulas grupo"

Atendendo à dispersão observada em alguns atributos e à concentração de valores noutros atributos, o que pode complicar não só a performance como também a observação do resultado dos modelos, optou-se por criar classificações, criando novos atributos que resultam da classificação dos anteriores. Para cada um dos atributos em questão utilizou-se uma técnica semelhante à de Hughes (2015), ordenando cada um dos atributos e agrupando os registos em grupos de 1 a 5. Este processo transforma os atributos contínuos em atributos discretos. Para facilitar a compreensão, os atributos classificados têm praticamente o mesmo nome dos atributos não classificados, com a diferença de apresentarem um "C" no inicio do nome do atributo.

Tabela 4 – Atributos promissores

Atributo	Tipo	Observações
idsaft		Atributo-chave de identificação única do registo
idade	Continuo	
genero	Discreto	
R dias sem frequencia	Continuo	
C R dias sem frequencia	Discreto	Classificação de 1 a 5 do atributo "dias sem frequencia"
meses inscricao	Continuo	
C meses inscricao	Discreto	Classificação de 1 a 5 do atributo "R meses inscricao"
M volume negocios	Continuo	
C M volume negocios	Discreto	Classificação de 1 a 5 do atributo "M volume negocios"
numero aulas grupo	Continuo	
C numero aulas grupo	Discreto	Classificação de 1 a 5 do atributo "numero aulas grupo"
F frequencia media	Continuo	
C F frequencia media	Discreto	Classificação de 1 a 5 do atributo "F frequencia media"
classe desistencia	Discreto	Atributo que se pretende prever

A discretização dos atributos criou as classes indicadas na Tabela 5:

Tabela 5 – Classes criadas e intervalo de valores correspondentes (Mar/2015)

Atributo > Classe		C R dias sem frequencia	C meses inscricao	C M volume negocios	C numero aulas grupo	C F frequencia media
1	De	1 dia	0 meses	0 €	0 aulas	0,00 x semana
	A	32 dias	3 meses	105,60 €	0 aulas	0,24 x semana
2	De	32 dias	3 meses	105,60 €	0 aulas	0,24 x semana
	A	35 dias	6 meses	244,80 €	0 aulas	0,59 x semana
3	De	35 dias	6 meses	244,80 €	0 aulas	0,59 x semana
	A	62 dias	12 meses	438,90 €	0 aulas	1,04 x semana
4	De	62 dias	12 meses	438,90 €	0 aulas	1,04 x semana
	A	131 dias	16 meses	588,60 €	5 aulas	1,79 x semana
5	De	131 dias	16 meses	588,60 €	5 aulas	1,79 x semana
	A	971 dias	31 meses	8703,00 €	576 aulas	7,00 x semana

Entender a correlação dos atributos é um fator que pode determinar quer a criação inicial do modelo, quer posteriormente a simplificação dos mesmos. A Figura 6 apresenta a relação de correlação dos atributos simples (contínuos, à esquerda) e classificados (discretos, à direita).

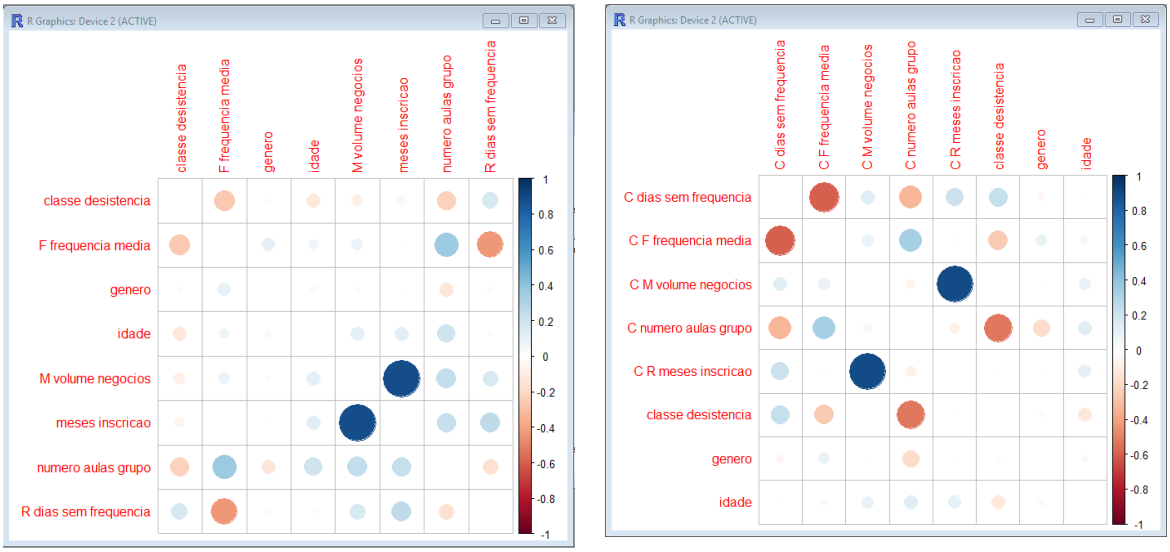


Figura 6 - Correlação de atributos (ordem alfabética) em R

a) Atributos contínuos **b)** Atributos discretos

A análise do resultado da correlação quantitativa e gráfica (Figura 6) leva-nos a concluir que:

- algumas correlações sofrem alterações quando se transforma os atributos contínuos em discretos;
- há uma correlação elevada entre o número de meses de inscrição e o volume de negócios;
- há uma correlação elevada entre o número de aulas de grupo e a frequência média;
- há uma correlação negativa relativamente forte entre o número de aulas de grupo e a desistência;
- há uma correlação negativa relativamente forte entre os dias sem frequência e a frequência média;
- há uma correlação forte entre o atributo referente ao número de dias sem frequência e a desistência.

4. Modelação

Estando os dados devidamente analisados e tratados, passamos a fase de implementação dos modelos, combinando os diferentes atributos com os vários algoritmos para obter os melhores resultados possíveis.

Relativamente aos algoritmos a utilizar, optou-se pela utilização de algoritmos de Classificação uma vez que de acordo com Tan (2005), as técnicas de classificação são adequadas a previsão ou descrição de um conjunto de dados com categorias binárias ou nominais, o que se enquadra com o objetivo proposto em que se pretende determinar se o utente virá ou não a ser desistente (categoria binária).

De acordo com as ferramentas disponíveis, avançou-se para a implementação dos modelos em *Microsoft SQL*Server® Analysis Services* utilizando o algoritmo *Microsoft Decision Trees*, e em R utilizando o algoritmo *Decision Tree* disponível nos packages "tree", "party" e "rpart" do R. Foram também testados os algoritmos *Microsoft Logistic Regression* e *Microsoft Neural Network* disponiveis em *Microsoft SQL*Server® Analysis Services*, no entanto, dado que os resultados obtidos são claramente inferiores aos obtidos com os algoritmos baseados em *Decision Tree* optou-se pela sua omissão neste artigo.

Os algoritmos *Decision Tree* procedem à construção da árvore de uma forma recursiva, analisando os atributos em cada situação e criando novas ramificações em função dos valores do atributo. O algoritmo pára quando todos os valores de um atributo na folha da árvore são idênticos ou quando a sua sub-divisão não apresenta mais-valias na determinação da classe que se pretende prever. Ficamos assim perante uma árvore em que cada nó interior corresponde a um atributo utilizado, as ligações entre os nós apresentam os valores que esse atributo pode assumir, e as folhas da árvore contêm a estatística sobre as ocorrências em que os atributos têm os valores apresentados no caminho a partir da raiz da árvore.

Os algoritmos podem criar apenas sub-divisões binárias dos atributos - criando eventualmente árvores mais verticais - ou, por outro lado, criando mais sub-divisões em cada atributo em função dos seus valores possíveis, criando árvores mais horizontais. Utiliza-se o termo "split" para nomear esta operação.

Por outro lado, cada algoritmo utiliza diferentes metodologias e/ou métricas para, em cada passo, escolher o atributo mais adequado, o que normalmente significa o atributo que apresenta um maior nivel de homogeneização. São do conhecimento público métricas como Gini Index, Information Gain, Condition Inference, etc.

Os algoritmos utilizados apresentam os métodos de "split" e de escolha do atributo indicados na Tabela 5.

Tabela 5 – Métodos de *split* e de escolha do atributo dos algoritmos *Decision Tree* utilizados

	Microsoft SQL*Server Analysis Services	**Sistema R**		
Nome do algoritmo	Microsoft Decision Tree	Package "tree"	Package "rpart"	Package "party ctree"
Método de "split"	Completo (Multiway)	Binario	Binario	Binario
Método de escolha do atributo	Bayesian Dirichlet Equivalent	Deviance	Gini index	Condition Inference (p-value)

Tendo em atenção estes conceitos, métodos e métricas, cada algoritmo treina o modelo baseado num subconjunto de dados provenientes da base de dados e, numa fase posterior, aplica o modelo a outro subconjunto de dados onde é possível verificar e medir a sua validade.

Neste trabalho, atendendo ao número de registos na tabela "dmRetencao" (44.820), determinou-se utilizar 70% dos registos para treino e os restantes 30% para teste do modelo. Todos os indicadores apresentados seguidamente referem-se às métricas aplicadas sobre os dados de teste.

Após a aplicação do modelo aos dados de teste, e de forma a aplicar as métricas conhecidas e especificadas na Tabela 7, é construída uma matriz, conhecida por Matriz de Confusão ou Matriz de Classificação, onde são apresentados os valores determinados pelo modelo e os valores efetivamente reais. A matriz apresenta a configuração indicada na Tabela 6.

Tabela 6 – Matriz de Confusão

	Valor real	
Valor determinado pelo modelo	Falso	Verdadeiro
Falso	TN	FN
Verdadeiro	FP	TP

Legenda:
TP – True positive – valor corretamente assinalado como positivo (desistente)
TN – True negative – valor corretamente assinalado como negativo (não desistente)
FP – False positive – valor incorretamente assinalado como positivo (falso desistente)
FN – False negative – valor incorretamente assinalado como negativo (falso não desistente)

Para comparar a performance dos algoritmos escolheram-se os indicadores apresentados na Tabela 7, calculados a partir da matriz de confusão obtida em cada modelo.

Tabela 7 – Fórmulas dos indicadores utilizados para avaliação dos modelos

Indicador	Fórmula
Taxa de Precisão (Accuracy)	(TP + TN) / n
Taxa de Erro	(FP + FN) / n
Taxa de Verdadeiros Positivos (Recall, Sensitivity)	TP / (TP + FN)
Taxa de Falsos positivos (Fall-out)	FP / (FP + TN)
Taxa de Verdadeiros Negativos (Specificity)	
Taxa de Precisão (Precision)	TP / (TP + FP)
Kappa	$K_e = [(TN + FN) * (TN + FP) + (FP + TP) * (FN + TP)] / n^2$ $K_0 = (TN + TP)/n$ $K = [K_0 – K_e] / [1 – K_e]$

A Tabela 8 é preenchida a partir dos resultados obtidos na implementação dos modelos que utilizam os diferentes algoritmos e conjugam os atributos disponíveis.

A observação dos dados da Tabela 8 permite realçar que:

- o modelo que utiliza todos os atributos classificados (modelo 4) apresenta um incremento substancial na taxa de precisão (*Accuracy*=84,20%) relativamente ao modelo que utiliza os atributos não classificados (modelo 1 com *Accuracy*=77,70%); O mesmo acontece com os modelos equivalentes que não utilizam os atributos "genero" e "idade";
- a não inclusão dos atributos demográficos ("genero" e "idade") individualmente ou em conjunto, quer nos modelos que utilizam os atributos não classificados (modelos 2 e 3) ou nos que utilizam os atributos classificados (modelos 5 e 6) não altera significativamente a taxa de precisão (*Accuracy*) relativamente ao modelo que inclui estes dois atributos (modelos 1 e 4, respetivamente);
- os modelos FM, RFM e RM (modelos 7, 9 e 11 respetivamente) que utilizam apenas os atributos não classificados apresentam melhores taxas de precisão (*Accuracy*) que os respetivos modelos que utilizam atributos classificados.

No que diz respeito à implementação dos modelos em R, a Tabela 9 apresenta os resultados obtidos.

Tabela 8 – Modelos mais significativos, atributos e algoritmos utilizados, matrizes de confusão e indicadores em *Microsoft SQL*Server® Analysis Services*

Nº	Model Name	Used Attributes	Confusion Matrix			Accuracy	True Positive	Precision	True Negative
				Actual		-------	-------	-------	
			Predicted	TRUE	FALSE	Error	False Positive	Kappa	
1	**ALL**	R dias sem frequencia, F frequencia media, meses inscricao, numero aulas grupo, M volume negocios, genero, idade	**TRUE**	4556	1616	77,70%	76,71%	73,82%	78,47%
			FALSE	1383	5891	22,30%	21,53%	54,96%	
2	**ALL-GENDER**	R dias sem frequencia, F frequencia media, meses inscricao, numero aulas grupo, M volume negocios, idade	**TRUE**	4493	1681	76,97%	76,05%	72,77%	77,70%
			FALSE	1415	5857	23,03%	22,30%	53,49%	
3	**ALL-GENDER-AGE**	R dias sem frequencia, F frequencia media, meses inscricao, numero aulas grupo, M volume negocios	**TRUE**	4509	1479	77,53%	74,52%	75,30%	80,00%
			FALSE	1542	5916	22,47%	20,00%	54,57%	
4	**ALL-CLASS**	C R dias sem frequencia, C F frequencia media, C meses inscricao, C numero aulas grupo, C M volume negocios, genero, idade	**TRUE**	4754	911	84,20%	79,66%	83,92%	87,82%
			FALSE	1214	6567	15,80%	12,18%	67,82%	
5	**ALL-GENDER-CLASS**	C R dias sem frequencia, C F frequencia media, C meses inscricao, C numero aulas grupo, C M volume negocios, idade	**TRUE**	4732	936	83,93%	79,44%	83,49%	87,50%
			FALSE	1225	6553	16,07%	12,50%	67,27%	
6	**ALL-GENDER-AGE-CLASS**	C R dias sem frequencia, C F frequencia media, C meses inscricao, C numero aulas grupo, C M volume negocios	**TRUE**	4813	959	84,69%	81,40%	83,39%	87,27%
			FALSE	1100	6574	15,31%	12,73%	68,84%	
7	**FM**	F frequencia media, M volume negocios	**TRUE**	3718	1711	71,10%	63,09%	68,48%	77,35%
			FALSE	2175	5842	28,90%	22,65%	40,79%	
8	**FM-CLASS**	C F frequencia media, C M volume negocios	**TRUE**	3698	2228	66,65%	62,11%	62,40%	70,26%
			FALSE	2256	5264	33,35%	29,74%	32,39%	
9	**RFM**	R dias sem frequencia, F frequencia media, M volume negocios	**TRUE**	4802	1649	79,38%	81,05%	74,44%	78,07%
			FALSE	1123	5872	20,62%	21,93%	58,57%	
10	**RFM-CLASS**	C R dias sem frequencia, C F frequencia media, C M volume negocios	**TRUE**	4248	1882	73,11%	71,01%	69,30%	74,79%
			FALSE	1734	5582	26,89%	25,21%	45,69%	
11	**RM**	R dias sem frequencia, M volume negocios	**TRUE**	4770	1723	79,00%	81,26%	73,46%	77,26%
			FALSE	1100	5853	21,00%	22,74%	57,83%	
12	**RM-CLASS**	C R dias sem frequencia, C M volume negocios	**TRUE**	4060	1775	72,59%	68,00%	69,58%	76,25%
			FALSE	1911	5700	27,41%	23,75%	44,35%	

Tabela 9 – Modelos mais significativos, atributos e algoritmos utilizados, matrizes de confusão e indicadores em R

Nº	Model & Package Name	Used Attributes	Confusion Matrix			Accuracy	True Positive	Precision	True Negative
				Actual		-------	-------	-------	
			Predicted	TRUE	FALSE	Error	False Positive	Kappa	
21	ALL_tree	R dias sem frequencia, F frequencia media, meses inscricao, numero aulasgrupo, M volume negocios, genero, idade	**TRUE**	4063	1930	78,24%	80,36%	67,80%	76,96%
			FALSE	993	6447	21,76%	23,04%	55,29%	
22	ALL_rpart		**TRUE**	4996	1027	82,27%	78,60%	82,95%	85,55%
			FALSE	1360	6080	17,73%	14,45%	64,33%	
23	ALL_ctree		**TRUE**	4527	1466	76,43%	72,70%	75,54%	79,66%
			FALSE	1700	5740	23,57%	20,34%	52,49%	
24	ALL_GENDER_tree	R dias sem frequencia, F frequencia media, meses inscricao, numero aulas grupo, M volume negocios, idade	**TRUE**	4063	1930	78,24%	80,36%	67,80%	76,96%
			FALSE	993	6447	21,76%	23,04%	55,29%	
25	ALL_GENDER_rpart		**TRUE**	4996	1027	82,27%	78,60%	82,95%	85,55%
			FALSE	1360	6080	17,73%	14,45%	64,33%	
26	ALL_GENDER_ctree		**TRUE**	4471	1522	78,69%	76,94%	74,60%	80,03%
			FALSE	1340	6100	21,31%	19,97%	56,76%	
27	ALL_GENDER_AGE_tree	R dias sem frequencia, F frequencia media, meses inscricao, numero aulas grupo, M volume negocios	**TRUE**	4627	1366	79,62%	77,14%	77,21%	81,63%
			FALSE	1371	6069	20,38%	18,37%	58,77%	
28	ALL_GENDER_AGE_rpart		**TRUE**	4996	1027	82,27%	78,60%	82,95%	85,55%
			FALSE	1360	6080	17,73%	14,45%	64,33%	
29	ALL_GENDER_AGE_ctree		**TRUE**	4817	1176	80,00%	76,13%	80,38%	83,45%
			FALSE	1510	5930	20,00%	16,55%	59,76%	
30	ALL_CLASS_tree	C R dias sem frequencia, C F frequencia media, C meses inscricao, C numero aulas grupo, C M volume negocios, genero, idade	**TRUE**	3762	2231	80,32%	90,11%	62,77%	75,90%
			FALSE	413	7027	19,68%	24,10%	58,96%	
31	ALL_CLASS_rpart		**TRUE**	3762	2231	80,32%	90,11%	62,77%	75,90%
			FALSE	413	7027	19,68%	24,10%	58,96%	
32	ALL_CLASS_ctree		**TRUE**	4582	1411	83,85%	85,81%	76,46%	82,57%
			FALSE	758	6682	16,15%	17,43%	66,98%	
33	ALL_GENDER_CLASS_tree	C R dias sem frequencia, C F frequencia media, C meses inscricao, C numero aulas grupo, C M volume negocios, idade	**TRUE**	3762	2231	80,32%	90,11%	62,77%	75,90%
			FALSE	413	7027	19,68%	24,10%	58,96%	
34	ALL_GENDER_CLASS_rpart		**TRUE**	3762	2231	80,32%	90,11%	62,77%	75,90%
			FALSE	413	7027	19,68%	24,10%	58,96%	
35	ALL_GENDER_CLASS_ctree		**TRUE**	4551	1442	83,79%	86,10%	75,94%	82,30%
			FALSE	735	6705	16,21%	17,70%	66,83%	
36	ALL_GENDER_AGE_CLASS_tree	C R dias sem frequencia, C F frequencia media, C meses inscricao, C numero aulas grupo, C M volume negocios	**TRUE**	3762	2231	80,32%	90,11%	62,77%	75,90%
			FALSE	413	7027	19,68%	24,10%	58,96%	
37	ALL_GENDER_AGE_CLASS_rpart		**TRUE**	3762	2231	80,32%	90,11%	62,77%	75,90%
			FALSE	413	7027	19,68%	24,10%	58,96%	
38	ALL_GENDER_AGE_CLASS_ctree		**TRUE**	4370	1623	83,66%	88,43%	72,92%	80,89%
			FALSE	572	6868	16,34%	19,11%	66,36%	
39	RFM_tree	R dias sem frequencia, F frequencia media, M volume negocios	**TRUE**	3894	2099	77,59%	81,02%	64,98%	75,67%
			FALSE	912	6528	22,41%	24,33%	53,75%	
40	RFM_rpart		**TRUE**	3915	2078	77,70%	81,02%	65,33%	75,84%
			FALSE	917	6523	22,30%	24,16%	54,02%	
41	RFM_ctree		**TRUE**	4939	1054	80,70%	76,25%	82,41%	84,85%
			FALSE	1538	5902	19,30%	15,15%	61,26%	
42	RM_tree	R dias sem frequencia, M volume negocios	**TRUE**	3894	2099	77,59%	81,02%	64,98%	75,67%
			FALSE	912	6528	22,41%	24,33%	53,75%	
43	RM_rpart		**TRUE**	3915	2078	77,70%	81,02%	65,33%	75,84%
			FALSE	917	6523	22,30%	24,16%	54,02%	
44	RM_ctree		**TRUE**	4847	1146	80,79%	77,16%	80,88%	83,97%
			FALSE	1435	6005	19,21%	16,03%	61,30%	

A observação dos dados da Tabela 9 permitiu concluir que:

- não há qualquer alteração de valores na matriz de confusão da aplicação do algoritmo do package "rpart" quer se inclua os dois atributos demográficos ("genero" e "idade") (modelos 22 e 31), apenas um (modelos 25 e 34) ou nenhum (modelos 28 e 37);
- no caso do algoritmo do package "tree", e apenas nos modelos que utilizam os campos não classificados, o modelo (modelo 27) ganha precisão quando se retiram ambos

os campos demográficos, relativamente aos dois modelos que incluem os 2 campos demográficos ou apenas o atributo "idade" (modelos 21 e 24);

- ao contrário do que acontece nos modelos em *Microsoft SQL*Server Analysis Services*, os modelos "ctree" (modelos 23, 26 e 29) melhoram ligeiramente a taxa de precisão (*Accuracy*) quando se retiram os campos demográficos;
- nos modelos que utilizam atributos classificados e que utilizam o algortimo "ctree", o indicador Taxa de Verdadeiros Positivos melhora quando o modelo não inclui os 2 atributos (modelos 32, 35 e 38);
- ao contrário dos outros algoritmos em R, o algoritmo "ctree" faz uso de todos os atributos indicados na construção da árvore;
- embora sem variações muito significativas, o indicador Kappa apresenta o seu melhor valor no modelo "ctree" que inclui os dois atributos demográficos "genero" e "idade" (modelo 32 com 66,98%) e pior no modelo que não inclui os dois atributos demográficos (modelo 38 com 66,36%).

Constatando a melhor precisão dos modelos "ctree" que utilizam atributos classificados (modelos 32, 35 e 38) procedeu-se à obtenção dos gráficos ROC, e constata-se que o que apresenta melhor configuração é o correspondente ao modelo 33 que não utiliza os atributos demográficos.

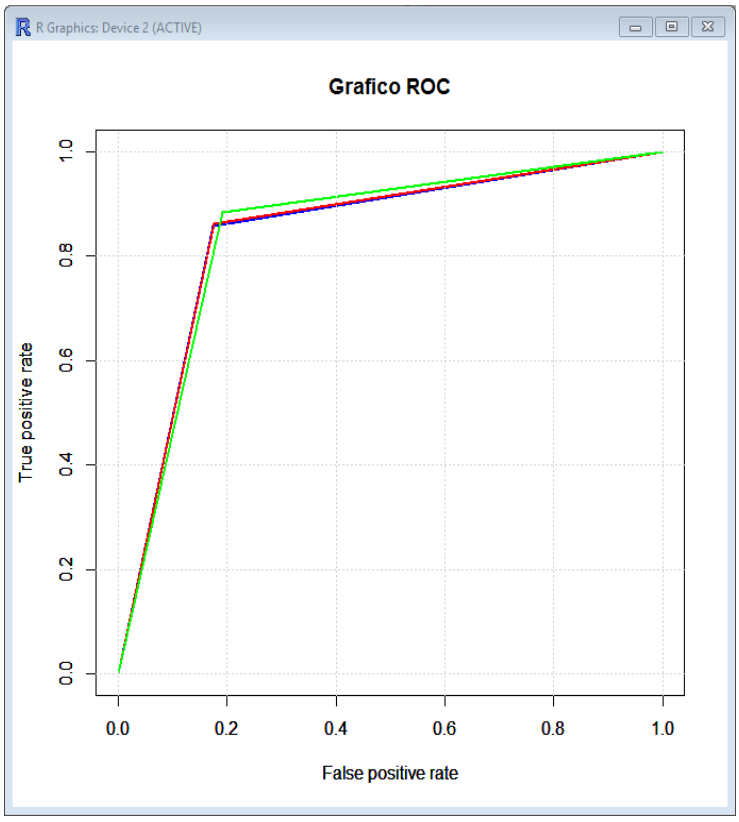


Figura 7 – Gráfico ROC dos modelos "ctree" 32 (linha azul), 35 (linha vermelha) e 38 (linha amarela)

Comparando agora os dados das tabelas 8 e 9, observamos ainda que:

- os algoritmos que utilizam os atributos demográficos "genero" e "idade" combinados com os atributos classificados (modelos 4, 5, 32 e 35) apresentam taxas de verdadeiros positivos inferiores aos modelos sem os referidos atributos (modelo 6 com 81,40% e modelo 35 com 86,10%);
- a ligeira alteração das taxas de precisão (*Accuracy*) dos modelos 4 (84,20%), 5 (83,93%) e 6 (84,69%) em *Microsoft SQL*Server Analysis Services*, e dos modelos 32 (83,85%), 35 (83,79%) e 38 (83,66%) revelam a presença de *overfitting* e sugerem a não utilização dos atributos demográficos;
- constata-se ainda haver uma ligeira diferença entre a taxa de precisão (*Accuracy*) obtida no modelo 6 (84,69%) e a obtida no modelo 38 (83,66%).

Apesar dos dois modelos utilizarem o mesmo conjunto de atributos, a forma de criação da árvore de decisão difere entre os dois modelos, uma vez que o algoritmo *Microsoft Decision Tree* utiliza por defeito um *split* múltiplo, e o "ctree" utiliza um *split* binário (Tabela 5).

Constata-se assim haver dois modelos que se destacam dos outros (modelo 6 em *Microsoft SQL*Server Analysis Services* e modelo 38 em R), mas que entre si não apresentam diferenças substanciais que levem a optar por um ou outro.

5. O underfitting e o overfitting dos modelos

Os problemas de "underfitting" e de "overfitting" dos modelos provocam erros de previsão que podem ser evitados ou corrigidos, desde que se compreenda a natureza e a origem do problema.

Um modelo demasiado simples provoca a ocorrência de problemas, conhecidos como problemas de "underfitting" tanto no treino do modelo, como posteriormente na aplicação do modelo aos dados de teste.

Em alguns modelos implementados o problema de "underfitting" é observável através dos resultados mais fracos das métricas utilizadas (por exemplo, nos modelos 10 e 12 e ainda mais noutros modelos implementados que por questões de simplicidade optamos por não incluir neste artigo).

Um modelo demasiado complexo e/ou que contenha ruído pode provocar a criação de árvores de decisão mais complexas do que o necessário. Este problema é conhecido por "overfitting".

Apesar dos problemas de "overfitting" não serem de fácil deteção existem técnicas próprias que podem ser utilizadas. Neste trabalho, aplicamos os conceitos do Principio da Parcimónia que refere que a explicação mais simples é normalmente a que deve ser

assumida como a correta, pois evita a probabilidade do erro associado às explicações complexas.

A Figura 8 apresenta a árvore de decisão criada pelo modelo 38, que apresenta 65 nós terminais e 10 niveis de profundidade, com Accuracy=83,66%, Sensibilidade=88,43%, Especificidade=80,89% e Kappa=66,36%.

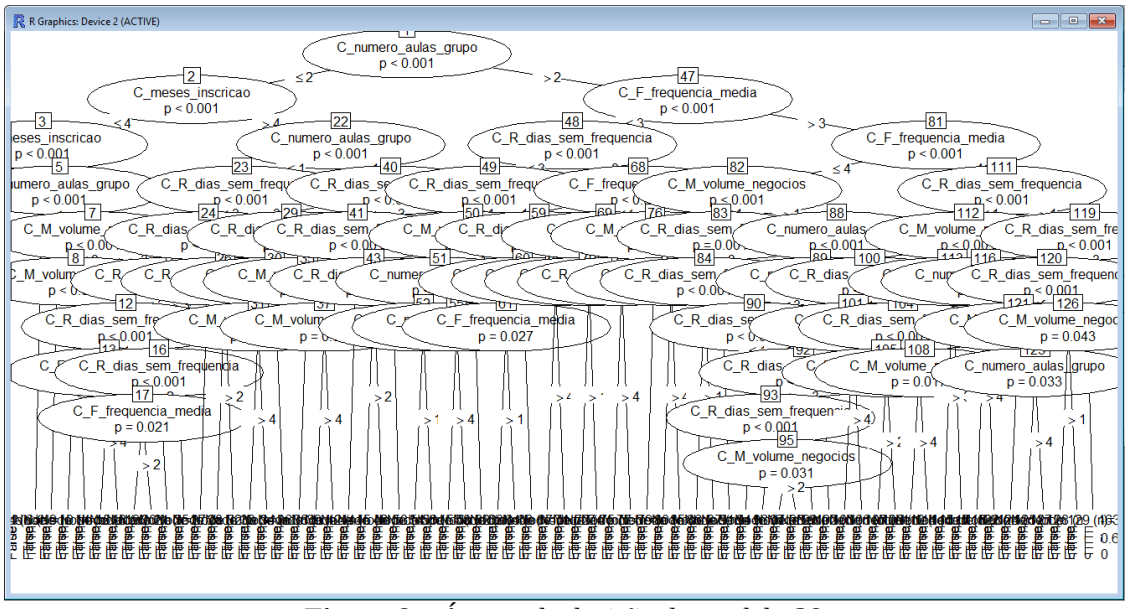


Figura 8 – Árvore de decisão do modelo 38

Se "cortarmos" sucessivamente na profundidade da árvore obtida com os dados de treino - método conhecido como "prunning" - obtemos os resultados apresentados no gráfico da Figura 9, ao aplicarmos o modelo sobre os dados de teste.

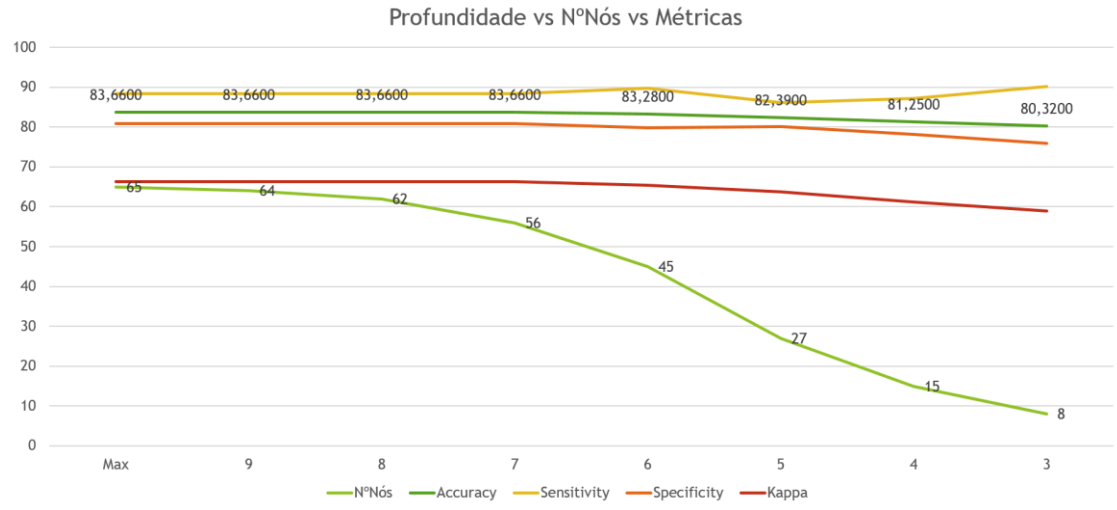


Figura 9 – Resultados obtidos pelo "corte" na profundidade da árvore no modelo 38

Como se pode observar, os indicadores mantêm-se inalteráveis quando o corte é efetuado até ao nível de profundidade 7, embora o número de nós terminais seja reduzido para 56.

Se o corte continuar a ser aplicado até ao nível de profundidade 3, obtém-se uma árvore de decisão com apenas 8 nós terminais (Figura 10) embora com uma degradação geral dos indicadores, mais acentuada no indicador Kappa.

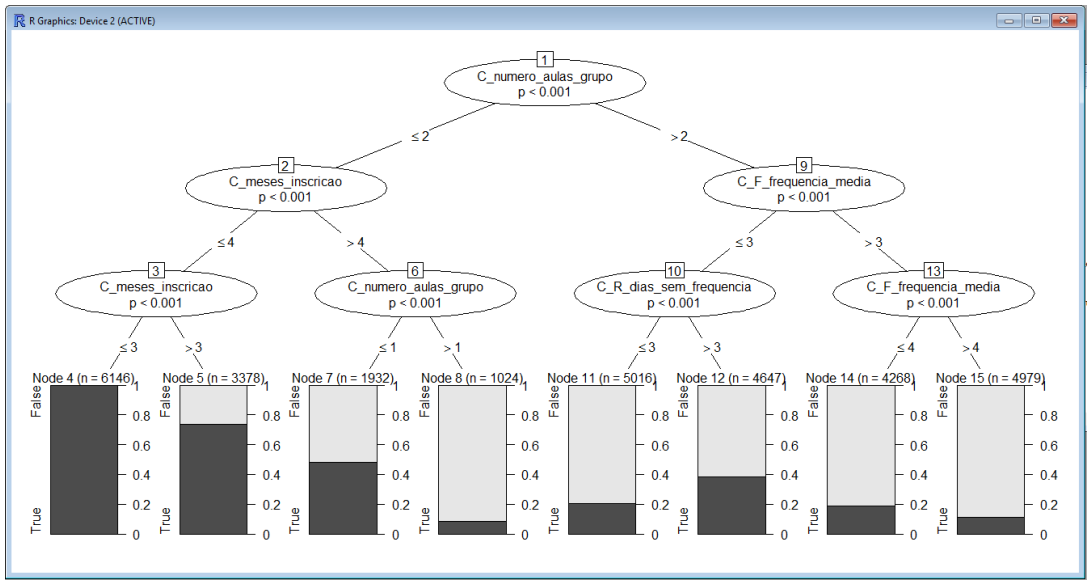


Figura 10 – Resultados obtidos pelo "corte" na profundidade da árvore no modelo 38 (maxdepth=3)

A árvore de decisão obtida é apresentada apenas a título de exemplo de forma a permitir a sua visualização gráfica. É consideravelmente mais simples e, apesar de perder precisão, permite a focagem em estudos de casos que incluam um grande grupo de ocorrências.

Por exemplo, a observação da árvore da Figura 10 permite observar que no caso do modelo nas condições indicadas, todos os utentes encontrados nas classes 1, 2 e 3 do atributo "C meses inscricao" e simultaneamente nas classes 1 e 2 do atributo "C numero aulas grupo" se tornaram desistentes. Pelo lado positivo, podemos observar que uma percentagem muito baixa dos utentes encontrados nas classes 4 e 5 do atributo "C meses inscricao" e simultaneamente nas classes 2, 3, 4 e 5 do atributo "C numero aulas grupo" é que se torna um desistente.

6. A aplicação desenvolvida

O desenvolvimento tinha por objetivo consumir o modelo de *data mining* apurado na fase anterior implementando uma aplicação com quatro principais funcionalidades ilustradas no diagrama de casos da Figura 11.

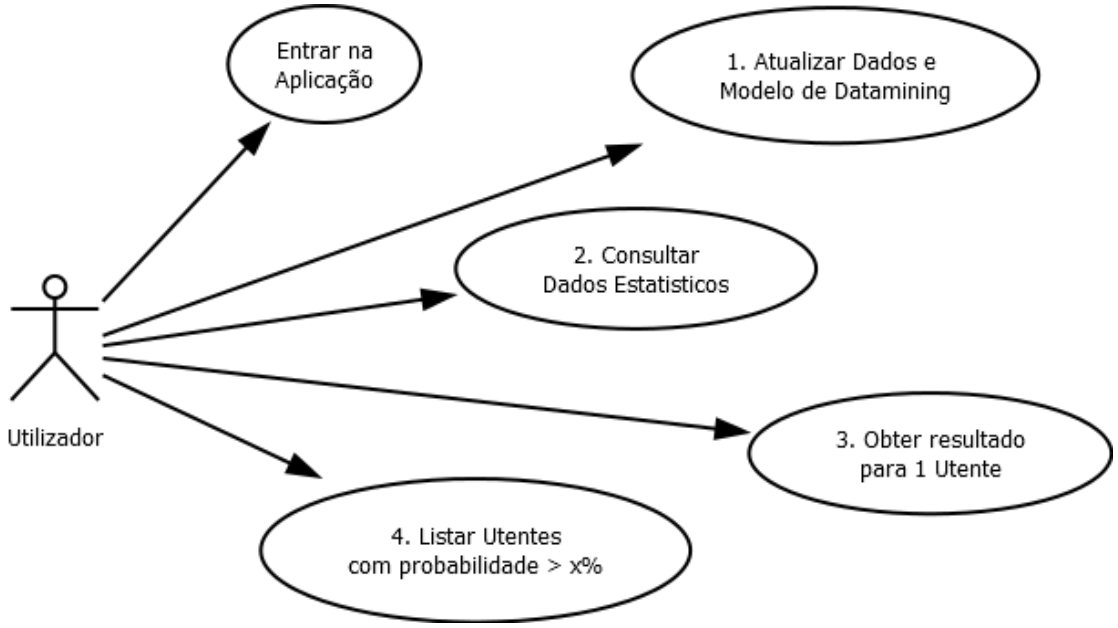


Figura 11 – Diagrama de casos de utilização

No primeiro caso de utilização (*Atualizar Dados e Modelo de Data Mining*), a aplicação importa os dados da base de dados do cliente para a base de dados de Data Wharehouse e procede à limpeza dos mesmos de acordo com o descrito na Secção 3. Por fim, aplica-se o modelo de Data Mining (modelo 6) atualizando a árvore de decisão, e procede ao cálculo dos indicadores (Tabela 7) do modelo relativos ao novo período.

No caso *Consultar Dados Estatísticos* (caso de utilização 2) obtém-se a performance do modelo em cada período (mês e ano) através da apresentação dos indicadores calculados na atualização do modelo.

No caso de utilização *Obter resultado para 1 Utente* (ou Simular) solicita-se ao utilizador a indicação dos dados do utente, necessários ao cálculo dos atributos utilizados no modelo (data da inscrição para calcular o numero de meses da inscrição, número de entradas no ginásio para calcular a frequência média semanal, número de aulas de grupo que frequentou, valor total pago e há quantos dias não vem ao ginásio); são calculados os valores correspondentes e classificados de acordo com os valores mínimos e máximos das tabelas de classificação (Tabela 5 do período em questão); e por fim a árvore de decisão é percorrida com essas características para obter o nó adequado. Uma vez que cada nó contém a informação necessária, é apresentado ao utilizador a percentagem dos utentes que desistem e que não desistem, e que têm as essas características.

Finalmente, o último caso de utilização permite, partindo da lista de todos os utentes ativos á data, listar todos aqueles que, face às suas características atuais, se enquadram em nós que apresentam uma percentagem de desistentes superior a um valor indicado pelo utilizador. Em cada utente, a aplicação procede como no indicado no caso de utilização anterior para determinar a probabilidade de desistência.

Para o desenvolvimento da aplicação utilizou-se uma metodologia de desenvolvimento em três níveis, tal como representado no Diagrama de Arquitetura da Figura 12.

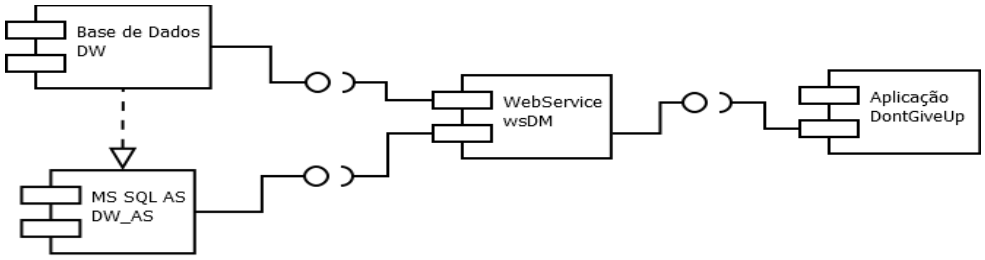


Figura 12 – Diagrama de arquitetura do projeto

Ao nível da base de dados DW (onde reside a tabela "dmRetencao" que contém a informação de suporte ao modelo) foram integrados os procedimentos desenvolvidos na fase de preparação dos dados, foram implementadas algumas funções escalares para simplificação e generalização dos cálculos e desenvolvidos novos procedimentos para implementar as funcionalidades requeridas pela aplicação.

As funções escalares e procedimentos foram implementados no dialeto SQL do *Microsoft SQL*Server*: *Transact SQL*.

Na segunda camada foi implementado um serviço web ("wsDM"). A criação do serviço permitiu atingir dois objetivos: o de servir esta aplicação e posteriormente permitir a sua utilização por outras aplicações em redes internas e através da internet. O consumo dos serviços disponibilizados é efetuado através de mensagens SOAP em formato XML.

Foram implementados quatro métodos principais:
- o método "wsDataAndModelUpdate" para atualizar o modelo, correspondente ao primeiro caso de utilização;
- o método "wsSelectModelStat" para selecionar as informações sobre o estado do modelo (segundo caso de utilização);
- o método "wsSelectUserProbability" que, mediante as características de um utente retorna a probabilidade do mesmo se tornar num desistente (terceiro caso de utilização);
- e finalmente, o "wsSelectGiveUpUsers" que retorna todos os utentes, atualmente ativos, que apresentam uma probabilidade de se virem a tornar desistentes acima de um valor indicado (quarto e ultimo caso de utilização).

Por fim, a camada de apresentação limita-se a consumir os serviços web implementados e a apresentar ao utilizador a informação retornada em formulários próprios, tendo sido utilizados componentes de um produtor conhecido de mercado (UI for ASP.NET AJAX da Telerik) para melhorar o aspeto visual final.

O serviço web e a aplicação foram desenvolvidos em *Visual Studio 2010* em linguagem C#. O acesso aos dados e procedimentos na base de dados DW é feito através do modelo

Entities. Contudo, uma vez que o *Microsoft Analysis Services* armazena a informação sobre os modelos de *data mining* numa base de dados separada (na Figura 12 indicada como DW_AS), para estabelecer ligação e obter resultados é necessário utilizar uma *framework* própria, o ADOMD.NET.

Por outro lado, para executar o comando de atualização do modelo após a execução dos procedimentos de importação e limpeza dos dados, é necessário utilizar XMLA (*XML for Analysis*). Após a atualização do modelo, o método executa os procedimentos "SystemGetClassificationMatrix" e "SystemGetLiftTable" para registar os resultados do modelo para futura comparação com resultados obtidos noutros períodos.

Os métodos "wsSelectUserProbability" e "wsSelectGiveUpUsers" utilizam ainda DMX (*Data Mining eXtensions*) para consultarem diretamente a árvore de decisão construída pelo modelo.

A Figura 13 apresenta o formulário principal da aplicação desenvolvida, onde se realça no topo as quatro opções disponíveis correspondentes aos quatro casos de utilização apresentados anteriormente: *Informação Estatística*, *Simular*, *Relatório de Desistentes* e *Atualizar Modelo*.

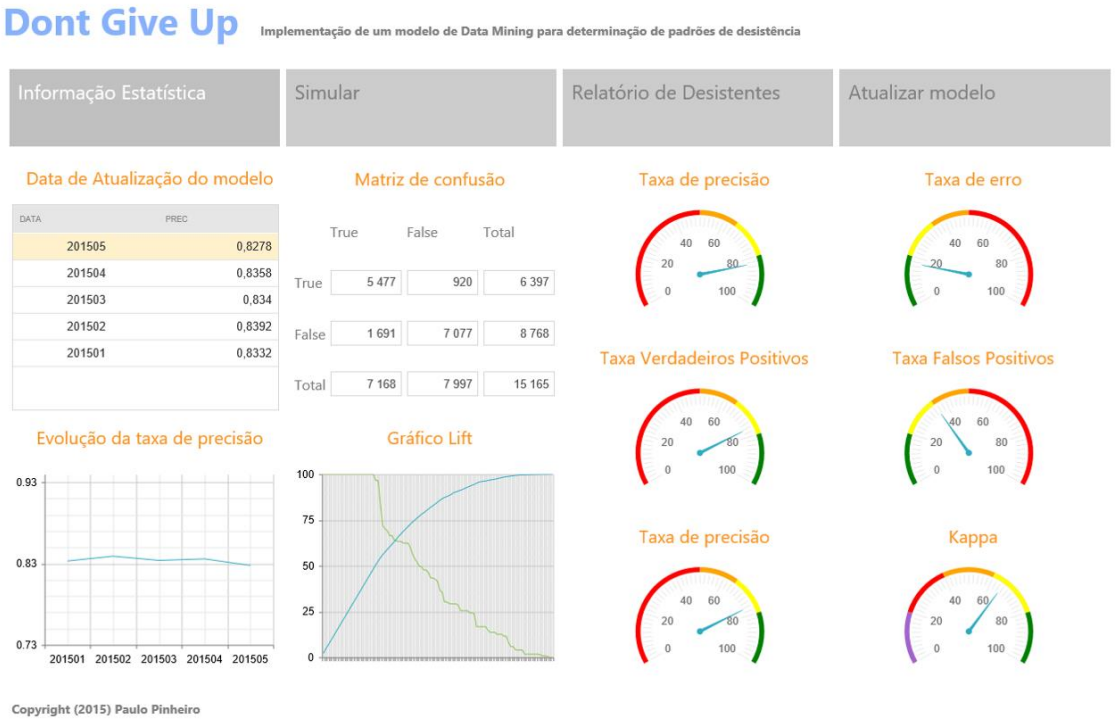


Figura 13 – Formulário de entrada da aplicação ***DontGiveUp***
Resultado da seleção do Ano/Mês - apresentação da informação estatística sobre o modelo do período escolhido

Os indicadores apresentados na aplicação são calculados através das formulas indicadas na Tabela 6.

7. Conclusões

7.1. Sobre o algoritmo de classificação escolhido e os atributos utilizados

Tendo em atenção que o projeto tem por objetivo encontrar padrões de desistência - o que em termos dos modelos criados se traduz na determinação de verdadeiros positivos -, optou-se por escolher um modelo que apresente melhores taxas de precisão e que tendencialmente privilegie os positivos, ainda que sejam falsos positivos, em detrimento de uma maior precisão a nível dos negativos verdadeiros (não desistentes). Será sempre preferível considerar que um utente poderá tornar-se num desistente, ainda que tal não venha a acontecer, do que não o considerar, e acabar por se tornar num.

Nestas considerações, o modelo 6 baseado no algoritmo *Microsoft Decision Tree* e o modelo 38 baseado no algoritmo "ctree" do Package "party" em Sistema R foram os que apresentaram melhores resultados nos indicadores observados, como se pode ver nos gráficos da Figura 14.

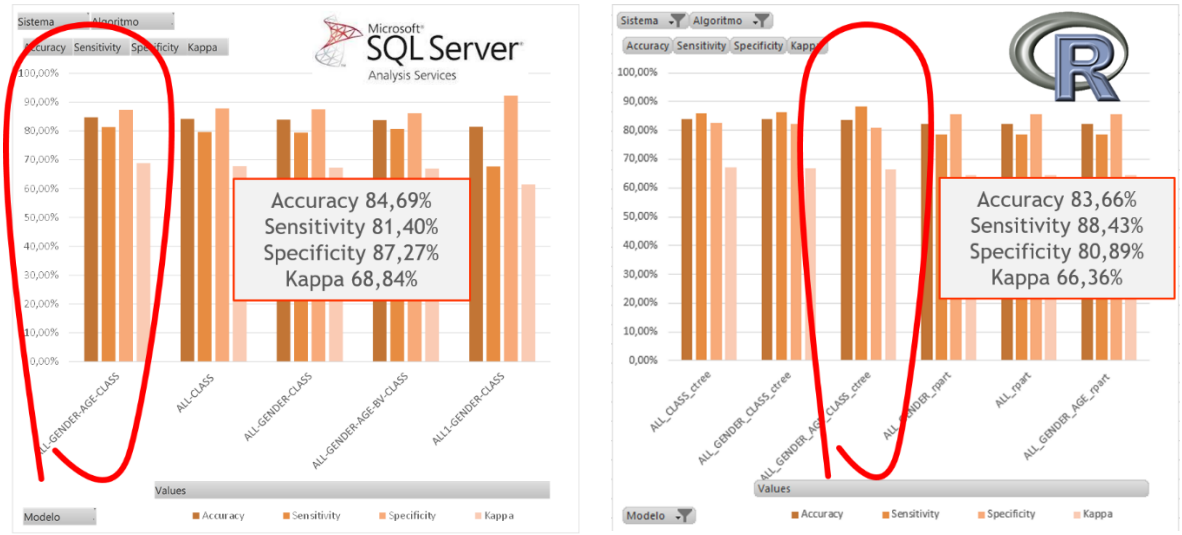
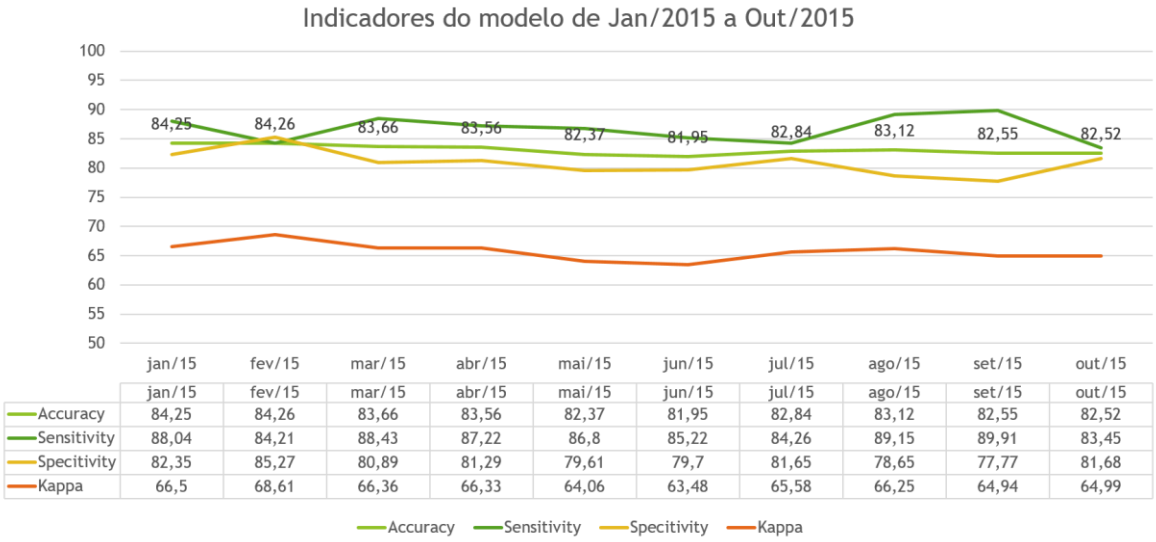


Figura 14 – Gráficos de indicadores dos melhores modelos em Microsoft SQL*Server Analysis Services (à esquerda) e em Sistema R (à direita)

Uma vez que se pretende aplicar o modelo periodicamente, numa análise regular mês-a-mês, avaliou-se também o comportamento do modelo 38 nos primeiros 10 meses de 2015, de acordo com os dados disponíveis. O resultado é apresentado na Figura 15.



	jan/15	fev/15	mar/15	abr/15	mai/15	jun/15	jul/15	ago/15	set/15	out/15
Accuracy	84,25	84,26	83,66	83,56	82,37	81,95	82,84	83,12	82,55	82,52
Sensitivity	88,04	84,21	88,43	87,22	86,8	85,22	84,26	89,15	89,91	83,45
Specitivity	82,35	85,27	80,89	81,29	79,61	79,7	81,65	78,65	77,77	81,68
Kappa	66,5	68,61	66,36	66,33	64,06	63,48	65,58	66,25	64,94	64,99

Figura 15 – Gráficos de indicadores do modelo 38 em Sistema R ao longo de 10 meses

Da análise do gráfico pode-se concluir que o modelo não sofre grandes variações ao longo do ano no que diz respeito à precisão (Accuracy ±2,3%), embora os indicadores de sensibilidade e especificidade possam sofrer oscilações maiores, eventualmente relacionados com períodos mais atípicos, como é o caso dos meses de ferias e pós-natal.

Dos modelos efetuados e testados tanto em Microsoft SQL*Server Analysis Services como no Sistema R, e no que diz respeito aos atributos a considerar, os que apresentam melhores resultados são os que resultam da conjugação dos atributos classificados como o número de aulas de grupo frequentadas, o número de meses da inscrição, o número de dias sem frequência, o volume de negócios e a frequência média, exatamente por esta ordem de relevância do atributo para o modelo. Conclui-se também que os atributos correspondentes aos dados demográficos idade e género não apresentam mais-valias significativas no modelo determinado, revelando a presença de *overfitting* nalguns modelos que utilizavam estes atributos.

Ainda no que diz respeito aos atributos, é importante referir que por ausência de dados, alguns atributos candidatos não foram utilizados nos modelos, e que a sua eventual utilização poderá trazer melhorias aos resultados obtidos.

Dos dois modelos que melhores resultados apresentaram, por razões de ordem prática relacionadas com a passagem a produção, a opção recaiu pela utilização do modelo 6 no restante desenvolvimento deste projeto. Refere-se no entanto que ambos os produtos apresentam opções de afinação dos algoritmos que não foram testadas neste âmbito, e como tal poderão melhorar a performance dos modelos obtidos.

7.2. A solução implementada

A aplicação final apresenta-se com utilização muito simples e prática, e permite fornecer dados úteis para a tomada de opções e ações com vista à melhoria da taxa de retenção, bem como apreciar a evolução do modelo ao longo do tempo.

A implementação da aplicação em camadas permitirá alimentar o modelo com dados provenientes de outras aplicações disponíveis no mercado, para além da que foi utilizada neste projeto. Por sua vez, o serviço web, dados os métodos disponibilizados, permitirá fornecer informação a outras aplicações.

7.3. O resultado final

A taxa de precisão final obtida (*Accuracy* de 84,69% no modelo 6 e 83,66% no modelo 38) bem como a relação da Sensibilidade com a Especificidade apresentam bons resultados. A constatação deste fato pode ser obtida pela observação dos gráficos ROC dos modelos em que as linhas dos modelos se aproximam bastante do topo superior esquerdo, apresentando uma grande área abaixo da linha do modelo.

Por outro lado, de acordo com Landis e Koch (1977) e a escala de interpretação do indicador Kappa, o valor obtido nos modelos 6 (68,84%) e 38 (66,83%) significam uma concordância substancial com o modelo perfeito.

Na ótica do Applied Business Analythics, a análise detalhada dos nós terminais do modelo, dos valores dos atributos que levam a essa classificação e dos padrões que formam, permite a tomada de ações que levem a alteração do comportamento dos utentes, prolongando a duração da sua inscrição e consequentemente no aumento da taxa de retenção.

Referências

Brito C.M. (2000), O Marketing Relacional in Os Horizontes do Marketing, ed. Brito C.M. e Lencastre P., Editorial Verbo, Lisboa, pp. 61-84.

Cavique L. (2003), "Micro-Segmentação de Clientes com Base em Dados de Consumo: Modelo RM-Similis", Revista Portuguesa e Brasileira de Gestão, volume 2, nº3, pp. 72-77.

Cavique L., I. Themido (2002), Estratégias de Comunicação em CRM in E- Portugal, L. Valadares Tavares e M.J. Pereira, eds, pp. 13-20.

Correia A., Sacavém A., Colaço C. (2006) "A Indústria do Welness" in Manual de Fitness & Marketing, Visão e Contextos, pp. 55-66

Coulouris, G., Dollimore, J., Kindberg T., Blair G., (2012) "Distributed Systems Concepts and Design", Addison-Wesley

Han J., Kamber M., (2006) "Data Mining Concepts and Techniques", Elsevier

Hughes A.M. (2000), Database Marketing, McGraw-Hill Companies.

Hughes A. M., (2015) "Why RFM works in Predicting Response", http://www.dbmarketing.com/articles/Art245.htm [6 de Maio de 2015]

Hothorn T., Hornik K., Strobl C., Zeileis A. (2015), http://party.R-forge.R-project.org [8 de Junho de 2015]

IBM SPSS Modeler CRISP-DM Guide (2011), IBM Corporation

Landis R., Koch G. (1977) "The Measurement of Observer Agreement for Categorical Data" pp. 159-173

Ledolter J. (2013) "Data Mining and Business Analytics with R", Wiley

MICROSOFT, "Data Mining (SSAS)", https://msdn.microsoft.com/en-us/library/bb510516.aspx [30 de Abril de 2015]

Murteira B., Ribeiro C. S., Silva J. A., Pimenta C. (2008) "Introdução à Estatística", McGraw Hill

Tan P., Steinbach M., Kumar V. (2006) "Introduction to Data Mining", Addison-Wesley, pp. 145-198,

Ripley B. (2015) "Package tree", http://cran.r-project.org/web/packages/tree/tree.pdf [6 de Junho de 2015]

Paulo Pinheiro é licenciado em Informática pela Universidade Aberta. Atualmente é Diretor Executivo da CEDIS, da qual é também sócio fundador, tendo como principal responsabilidade a direção do Departamento de Desenvolvimento e responsável pela orientação de projetos nas áreas de Gestão de Instalações Desportivas, Educação, Bilhética e Parques de Campismo. Foi programador freelancer para o IEFP e formador na RSP / LISNAVE. Foi o autor de diversos artigos para as revistas AMSTRAD e SPOOLER.

Luís Cavique, Professor Auxiliar no Departamento de Ciências e Tecnologia (DCeT), Secção de Informática, Física e Tecnologia (SIFT). Vice-coordenador da Licenciatura em Informática e Coordenador do Mestrado Tecnologias e Sistemas Informáticos Web no biénio 2014–2016. Licenciado em Engenharia Informática em 1988 pela FCT-UNL. Obteve o grau Mestre em Investigação Operacional e Eng. Sistemas pelo IST-UTL em 1994. Obteve o grau de Doutor em Eng. Sistemas pelo IST-UTL em 2002. Tem como áreas de interesse, a intersecção da Informática (Computer Science) com a Engenharia de Sistemas (Management Science) designadamente a área de "Data Mining".