UNIVERSIDADE ABERTA

MODELOS PARA INCREMENTO DA RETENÇÃO EM SERVIÇOS DESPORTIVOS REGULARES:

ANÁLISE PREDITIVA E AÇÕES DE FIDELIZAÇÃO

PAULO SÉRGIO NUNES PINHEIRO

MESTRADO EM TECNOLOGIAS E SISTEMAS INFORMÁTICOS WEB

2018

UNIVERSIDADE ABERTA

MODELOS PARA INCREMENTO DA RETENÇÃO EM SERVIÇOS DESPORTIVOS REGULARES:

ANÁLISE PREDITIVA E AÇÕES DE FIDELIZAÇÃO

PAULO SÉRGIO NUNES PINHEIRO

MESTRADO EM TECNOLOGIAS E SISTEMAS INFORMÁTICOS WEB

Dissertação orientada pelo

Professor Doutor Luis Manuel Pereira Sales Cavique Santos

2018

A ciência passa por três fases, às quais podemos chamar as fases de Brahe, Kepler e Newton. Na fase Brahe, recolhemos muitos dados, como Tycho Brahe ao registar pacientemente as posições dos planetas noite após noite, ano após ano. Na fase Kepler, aplicamos leis empíricas aos dados, como Kepler fez com os movimentos planetários. Na fase Newton, descobrimos as verdades profundas.

Pedro Domingos, "A Revolução do Algoritmo Mestre"

Resumo

Atrair um novo cliente pode custar até cinco vezes mais do que manter um cliente atual satisfeito e por isso a capacidade de detetar, o mais cedo possível, quais os clientes que irão abandonar ou deixar de adquirir um determinado produto ou serviço, bem como as medidas que podem ser implementadas para evitar esse abandono ou quebra nas vendas são questões que todas as empresas gostariam de ver respondidas.

Neste trabalho é apresentado um sistema inteligente que gera conhecimento acionável ("*actionable knowledge*") baseado em dados reais e orientado para acções de fidelização de clientes de serviços desportivos regulares onde ocorrem elevadas taxas de cancelamento.

Seguindo os passos do *Database Marketing* o sistema evolui em três fases: numa primeira fase, parte de dados obtidos nas bases de dados dos sistemas ERP e CRM existentes nas instalações desportivas, dos quais extrai, transforma e carrega dados num *Data Warehousing*; numa segunda fase são aplicados modelos preditivos para identificar perfis mais suscetíveis de abandono; e por fim, são aplicadas ações de fidelização direcionadas a cada um dos perfis encontrados com o objetivo de aumentar a fidelização e a taxa de retenção.

Palavras-chave: serviços desportivos, *data mining*, análise preditiva, fidelização

Abstract

Attracting a new customer can cost up to five times more than keeping a current customer satisfied and therefore the ability to detect, as early as possible, which customers will abandon or fail to purchase a particular product or service, as well as measures which can be implemented to avoid this abandonment or break in sales are issues that all companies would like to see answered.

This work presents an intelligent system that generates actionable knowledge based on real data and oriented to customer's loyalty actions in regular sports services where high drop out rates occur.

According to Database Marketing's steps the system evolves in three phases: in a first phase, data obtained in the ERP and CRM systems databases existing in the sports facilities, from which it extracts, transforms and loads data in a Data Warehousing; in a second phase, predictive models are applied to identify profiles more susceptible to drop out; and finally, loyalty actions are applied to each one of the found profiles in order to increase loyalty and the retention rate.

Keywords: sport services, data mining, predictive analytics, loyalty

Agradecimentos

Em primeiro lugar fica o agradecimento especial a toda a família e em particular à esposa e aos filhos pela paciência e compreensão nas horas mais difíceis, e também pelas ausências a que por vezes tive de os sujeitar para conseguir levar este projeto até ao fim.

Fica também o maior apreço e os agradecimentos ao Professor Doutor Luis Cavique por ter aceitado a orientação desta dissertação de mestrado, pela enorme disponibilidade e atenção que dedicou a este projeto que foram determinantes para o resultado final, e também pelos permanentes incentivos e participação na redação dos artigos relacionados com este projeto.

Aos colegas desde o primeiro minuto desta aventura, o Bruno Moniz e o Halestino Pimentel, fica também o agradecimento pelas várias noites que passamos "à distância" a desenvolver os vários trabalhos que fizemos em grupo. Seria difícil encontrar melhores colegas para esta aventura e seguramente sem a sua companhia tudo teria sido muito mais complicado e difícil.

Fica também uma palavra de apreço pelos professores das várias unidades curriculares que em muito contribuíram para me tornar num melhor profissional e à Universidade Aberta por me ter proporcionado esta aprendizagem tão útil.

Por fim, fica também o agradecimento ao Fernando Ferreira pela companhia na participação das conferências e seminários que me ajudaram a evoluir nos assuntos relacionados com este projeto.

Índice

Resumo iii
Abstract v
Agradecimentos vii
Índice ix
Índice de Tabelas xi
Índice de Figuras xiii
Lista de Siglas xv
1. Introdução 1
1.1. Contextualização 3
1.2. Objetivos e contribuições 4
1.3. Estrutura do documento 5
2. A problemática dos serviços desportivos regulares contínuos e em épocas 7
2.1. Serviços desportivos regulares em Ginásios, Health-Clubs e Academias 9
2.2. Serviços desportivos regulares em Clubes desportivos, Piscinas e outras instalações 10
3. Trabalho relacionado 13
3.1. Contexto em *Database Marketing* 15
3.2. Pré-processamento de dados: *Extract, Transform, Load* e *Data Warehousing* 15
3.3. *Data Mining*: conceitos, seleção de atributos, algoritmos e métricas 21
3.3.1. Conceitos 21
3.3.2. Seleção de atributos 21
3.3.3. Algoritmos 27
3.3.4. Métricas 30
3.4. Retenção em serviços desportivos, ações de fidelização e experimentação 35
4. Acerca dos dados: DW e ETL 49
4.1. A origem dos dados e os atributos a considerar 51
4.2. A construção do *Data Warehouse* 53
4.3. Estatística descritiva dos dados do DW 62
5. Análise preditiva: classificação 67
5.1. ~~A~~tributos utilizados nos modelos preditivos 69
5.2. Métricas de avaliação da qualidade dos modelos 74

5.3. Obtenção dos perfis acionáveis 80
6. Ações de fidelização 83
6.1. Seleção dos segmentos acionáveis e ações de fidelização 85
6.2. Canais de comunicação e respetivos requisitos.....87
6.3. Avaliação do impacto das ações de fidelização 91
7. Conclusões 95
7.1. Conclusões.....97
7.2. Trabalhos em curso 98
Referências.....99

Índice de Tabelas

Tabela 3-1 - Fórmulas de cálculo das métricas de avaliação de modelos de DM 33
Tabela 3-2 – Síntese dos fatores que influenciam a lealdade dos Utentes 37
Tabela 4-1 - Atributos da tabela "Retencao" 52
Tabela 4-2 - Propriedades das bases de dados de teste a 31/Dez/2017 53
Tabela 4-3 – Sumário de valores estatisticamente significativos por atributo 63
Tabela 4-4 - Nº de Registos, Mínimos e Máximos das classes nos atributos obtidos com a Classificação de Hughes 64
Tabela 4-5 - Número de registos nas tabelas do DW 64
Tabela 5-1 - Atributos utilizados nos modelos preditivos 76
Tabela 5-2 – Profundidade das Árvores e Número de Nós em cada modelo 76
Tabela 5-3 – Métricas *Holdout* dos modelos preditivos (tabela “Retencao”) 77
Tabela 5-4 – Métricas *Holdout* dos modelos preditivos (tabela “Fitness”) 77
Tabela 5-5 – Métricas *Holdout* dos modelos preditivos (tabela “Aquaticos”) 77
Tabela 5-6 – Matriz de confusão e desvios padrão do Modelo Ret71 obtidos com o método *Cross-Validation* 79
Tabela 5-7 – Métricas *Cross-validation* do modelo preditivo Ret71 80
Tabela 5-8 – Perfis de desistência 81
Tabela 6-1 – Perfis de desistência e *workflow* de ações de fidelização a aplicar 87
Tabela 6-2 – Algumas boas práticas para a utilização de emails 88
Tabela 6-3 – Algumas boas práticas para a utilização de SMS 89
Tabela 6-4 - Configuração das experiências 92
Tabela 6-5 - Matriz de suporte ao cálculo do qui-quadrado 93
Tabela 6-6 - Exemplo de preenchimento da matriz de cálculo do qui-quadrado 93
Tabela 6-7 - Cálculos parciais e valor final de $\chi 2$ 93

Índice de Figuras

Figura 3-1 – *Database Marketing* 15
Figura 3-2 - Framework para escolha do método de SA 23
Figura 3-3 - Segmentação RM 39
Figura 3-4 - Modelo *Uplift* 40
Figura 3-5 – Componentes de uma experiência 44
Figura 3-6 – Validações por tipo de experiência 46
Figura 4-1 - Implementação do Data Warehouse 51
Figura 4-2 - Representação dos conjuntos de utentes presentes nas tabelas do DW 54
Figura 4-3 – Diagrama de atividades do processo ETL 60
Figura 5-1 – Histograma dos atributos de entrada após execução dos processos ETL na BD1 em 31/Out/2017 70
Figura 5-2 – Pontuação dos atributos atribuída pelo algoritmo *Microsoft Decision Trees* . 75
Figura 5-3 - Quadro de correlação dos atributos utilizados no Modelo Ret71 79
Figura 5-4 – Árvore de Decisão do Modelo Ret71 81
Figura 6-1 - Pirâmide de ações de fidelização 86
Figura 6-2 - Pirâmide de ações de fidelização 91

Lista de Siglas

AGAP	Associação de Ginásios e Academias de Portugal
BD	Base de Dados
CTA	*Call To Action*
DM	*Data Mining*
DW	*Data Warehouse*
ETL	*Extract, Transform, Load*
GHC	Ginásio ou *Health-Club*
IHRSA	*International Health, Racquet ans SportsClub Association*
MS	Microsoft
RGPD	Regulamento Geral de Proteção de Dados
ROC	*Receiver Operating Characteristic*
SA	Seleção de Atributos
SCI	Sistema de Comunicação Inteligente
SGBD	Sistema de Gestão de Bases de Dados

1. Introdução

1.1. Contextualização

Elevadas taxas de retenção traduzem a satisfação e lealdade dos clientes para com a empresa. Significa que os clientes estão atentos às novidades da empresa para adquirirem mais produtos e serviços ou fazer *upgrade* dos que já tem, falem favoravelmente a outros acerca da empresa, da marca ou dos seus produtos e serviços, tomem menos atenção à concorrência, sejam menos sensíveis ao preço, mantenham um maior nível de envolvimento transmitindo opiniões e ideias. Um cliente atual custa também menos a servir do que um cliente novo, dado que as transações comerciais com ele passam a ser rotina (Kotler & Keller, 2009).

A promoção do exercício físico como um meio para prevenir os crescentes rácios de obesidade e manutenção do bem-estar tem provocado uma proliferação de ginásios e *health-clubs* que competem com instalações desportivas públicas, e consequentemente tem havido uma crescente pressão nos prestadores deste tipo de serviço para manterem vantagens competitivas através de serviços que providenciam maiores níveis de satisfação do cliente (Howat & Assaker, 2016). Num contexto de muita oferta, o setor da prestação de serviços desportivos é caracterizado por uma elevada taxa de desistência (Avourdiadou & Theodorakis, 2014). De acordo com o Barómetro do Fitness da Associação de Ginásios e Academias de Portugal, AGAP, (AGAP, 2016) a taxa de cancelamentos geral em 2016 foi de 69% (o que significa que em 100 utentes ativos, 69 cancelam a sua inscrição) correspondendo portanto a uma taxa de retenção geral de 31%.

O problema das elevadas taxas de desistência / baixas taxas de retenção tem vindo a ser muito estudado recentemente em várias áreas, incluindo na área dos serviços desportivos. Algumas já utilizam técnicas de *data mining* na área da retenção (*churn prediction*), nomeadamente na área das telecomunicações onde também se verificam elevadas taxas de abandono. Contudo, dada a grande dimensão das bases de dados e dos custos envolvidos, a maioria dos estudos existentes nesta área usam pequenas amostras dos registos de clientes, o que pode resultar na baixa confiabilidade e validade dos resultados obtidos (Mahajan, Misra, & Mahajan, 2015).

No caso dos serviços desportivos os estudos efetuados suportam as suas conclusões na aplicação de métodos estatísticos sobre dados obtidos em inquéritos efetuados e apontam

para a relação da lealdade dos clientes com a qualidade das instalações, do pessoal e com a qualidade global dos serviços prestados, resultados que têm vindo a ser utilizados pela gestão das instalações desportivas. Contudo, como se referiu os estudos encontrados são baseados em inquéritos efetuados sobre uma amostra da população o que é apontado pelos próprios autores como uma limitação (Avourdiadou & Theodorakis, 2014; Howat & Assaker, 2016).

Quer motivadas pela legislação atual quer pela necessidade de controlo de acesso às instalações, atualmente as instalações desportivas dispõem de bases de dados de clientes com um grande historial que armazenam dados demográficos, dados de faturação e dados referentes às preferências de frequência dos utentes. E como refere Kotler (Kotler & Keller, 2009) de maneira geral, as empresas podem utilizar as suas bases de dados de clientes de cinco maneiras: para identificar potenciais clientes, para decidir que clientes devem receber uma determinada oferta, para aprofundar a relação com os seus clientes, para reativar compras de clientes e para evitar erros sérios.

Embora as ferramentas disponíveis nos sistemas informáticos atuais disponibilizem formas de segmentar os seus Clientes a capacidade para traçar perfis com base nas compras, utilizações e consumos anteriores é muito limitada e exige recursos técnicos que normalmente não estão à disposição nas instalações desportivas. Esta situação faz com que as instalações desportivas, no sentido de evitarem as desistências, tenham de o fazer através de campanhas dirigidas a todos os clientes, ou através de segmentações que as aplicações que utilizam permitem, independentemente de estarem ou não em situação de pré-abandono, o que leva a um maior consumo de tempo e recursos. É preciso então encontrar meios mais fáceis, rápidos e económicos para responder a esta questão.

1.2. Objetivos e contribuições

Pretende-se demonstrar através deste trabalho que seguindo os passos do *Database Marketing* (Cavique, 2006) – identificação, diferenciação e interação -, e partindo dos dados recolhidos por aplicações utilizadas pelas instalações desportivas é possível conceber um modelo que aplique técnicas de *data mining* / *predictive analytics* / *machine learning* para determinar não só os utentes que se encontram em risco de abandono, como também os perfis de utilização e comportamento que levam à desistência.

Através dos três passos referidos, pretende-se utilizar os dados existentes para identificar os utentes de serviços desportivos regulares em risco de abandono, diferenciando-os dos restantes através de técnicas de *data mining* / *predictive analytics* / *machine learning* que encontrem perfis de utilização e comportamento, e interagir de forma personalizada e com ações de fidelização concretas com os utentes que apresentam esses perfis para evitar o seu abandono. O modelo proposto pode ser apresentado da seguinte forma esquemática:

Dados → Modelos → Fidelização

Propomo-nos efetuar também um planeamento de experiências controladas que permitam validar o sucesso dessas ações de fidelização.

1.3. Estrutura do documento

Este documento tem a seguinte estrutura. No Capítulo 2 é apresentado o problema a tratar e são apresentadas as diferentes formas de prestação dos serviços desportivos regulares e a problemática relacionada. No Capítulo 3 é feita uma abordagem breve a trabalhos relacionados que têm sido efetuados em três áreas que utilizamos no desenvolvimento do trabalho: os processos *Extract / Transform / Load* (ETL), os conceitos, algoritmos e métricas utilizadas em algoritmos de *Data Mining* (DM) na área da classificação e da utilização das árvores de decisão, e na área da retenção dos serviços desportivos, nas ações de fidelização e na área das experiências efetuadas em *Marketing Research*. No Capitulo 4 apresentamos a metodologia e o trabalho desenvolvido na preparação dos dados, no Capitulo 5 introduzimos os modelos preditivos aplicados aos dados tratados no capítulo anterior, medimos os resultados e traçamos alguns perfis baseados nos resultados obtidos com os modelos. No Capítulo 6 apresentamos a metodologia proposta para a implementação das ações de fidelização, bem como o planeamento das experiências e uma proposta para medir os resultados obtidos. Finalmente, no Capítulo 7 apresentamos as conclusões e as contribuições do trabalho.

2. A problemática dos serviços desportivos regulares contínuos e em épocas

2.1. Serviços desportivos regulares em Ginásios, Health-Clubs e Academias

Como indicado na Introdução os serviços desportivos regulares sofrem de elevadas taxas de desistência, muito embora o abandono seja maior em ginásios, health-clubs e academias onde a atividade mais promovida é o *fitness*, do que em instalações desportivas que têm outras valências, nomeadamente piscinas (Frota, 2011).

Por outro lado, a forma de prestação dos serviços varia entre as diversas modalidades praticadas e por vezes também a forma como se consideram ou contabilizam os abandonos não é idêntica para as diversas atividades. O estudo da Retenção nas instalações desportivas está assim diretamente relacionado com a forma como os serviços são comercializados e prestados e por isso interessa compreender as formas de trabalho das entidades que promovem os serviços em causa e a forma como consideram a desistência dos seus utentes.

Em determinadas instalações que se encontram abertas ao público praticamente o ano inteiro (excetuando-se alguns feriados), normalmente de caracter privado tais como Ginásios, Health-Clubs e Academias de Fitness, os serviços são contínuos e os utentes assinam um contrato de fidelização com duração igual ou superior a 6 meses, findos os quais a inscrição pode ser renovada.

A inscrição é paga de forma regular, anualmente, mensalmente, quinzenalmente ou semanalmente sendo que nos intervalos de pagamento mais curtos o pagamento processa-se quase sempre por débito direto na conta bancária do utente enquanto os valores correspondentes a períodos mais longos podem ser pagos de uma só vez através de outras formas de pagamento. Enquanto o utente garante o pagamento da sua inscrição é considerado um utente ativo, é-lhe permitido aceder às instalações através do controlo de acessos e frequentar as atividades que a sua inscrição lhe permite frequentar.

De forma geral é obrigatório que o utente cumpra com os pagamentos correspondentes até ao final do contrato inicial. No entanto, quer por motivos válidos apresentados pelo utente, quer simplesmente por deixar de cumprir com o seu pagamento durante um período de tempo (definido pela gestão da instalação desportiva) o utente é considerado desistente.

A prática da atividade principal destas instalações, o *fitness*, decorre em regimes de livre-trânsito em que o utente pode aceder às instalações todos os dias (ou um determinado número de dias por semana) em horários mais ou menos alargados de acordo com o serviço que contratou sendo que os horários mais alargados, que incluem horas em que há maior afluência de utentes às instalações, são mais caros.

Nas suas visitas às instalações o utente pode frequentar as salas de musculação, de cardiofitness ou frequentar aulas variadas orientadas por um profissional – denominadas aulas de grupo - e de acordo com a sua preferência desde que a lotação não esteja lotada no dia e hora em que deseja frequentá-la.

O serviço regular pode ser complementado por um serviço de treinador pessoal que inclui um conjunto de sessões mensais com acompanhamento exclusivo ou partilhado por um número reduzido de pessoas.

Sobre este tipo de instalação e utilização já foi desenvolvido anteriormente pelo mesmo autor um trabalho de aplicação de modelos preditivos para determinar a probabilidade de desistência dos utentes (Pinheiro & Cavique, 2015).

2.2. Serviços desportivos regulares em Clubes desportivos, Piscinas e outras instalações

Por outro lado, devido à natureza dos seus recintos desportivos, há instalações desportivas que encerram num determinado período do ano. É o caso de alguns complexos desportivos e piscinas que encerram normalmente entre um a dois meses no verão dada a necessidade de efetuar manutenção aos seus equipamentos.

Motivados por esta situação, este tipo de instalações programa os seus serviços desportivos em épocas que decorrem entre Setembro de um ano e Junho ou Julho do ano seguinte de forma semelhante às épocas letivas escolares.

Embora estes tipos de instalação também possam disponibilizar serviços de *fitness* como os referidos anteriormente, na sua grande maioria os utentes inscrevem-se para frequentar aulas de atividades desportivas como a natação que ocorrem em dias da semana e horários fixos, mantendo-se nessa frequência ao longo de toda a época.

Como cada horário tem um número de vagas limitado não são permitidas mais inscrições nos horários que estão completos.

A inscrição é válida para a época em que se inscreveu mantendo-se o utente ativo enquanto não manifestar a vontade de desistir e enquanto mantiver o pagamento das suas mensalidades regularizado.

Ao cancelar a sua inscrição ou ao não regularizar as suas mensalidades num determinado prazo definido em regulamento (normalmente entre 30 a 60 dias), a sua inscrição é cancelada, libertando-se a vaga do horário em que estava inscrito e o utente passa ao estado de desistente.

No final da época, uma vez que terminam as atividades, os utentes passam ao estado de desistente caso não manifestem o seu interesse em manter a sua atividade válida para a época seguinte. Neste caso procedem à renovação da sua inscrição, escolhendo a mesma (ou outra) atividade, frequência semanal e horário.

Nas piscinas municipais é disponibilizado outro tipo de utilização denominado utilização livre. Nesta forma de utilização, os utentes compram senhas ou tempo de utilização que vão utilizando ao longo do tempo, sem qualquer compromisso de utilização regular, salvo o facto de terem de consumir todas as senhas ou valor adquirido num determinado prazo de validade que pode ou não coincidir com a época em que foram adquiridos. Os horários em que o utente pode usufruir deste tipo de utilização depende do espaço deixado vago pelas restantes aulas e a lotação pode variar ao longo do dia, e de dia para dia.

Em alguns casos o valor carregado pode ser utilizado para reservar courts de Ténis, de Squash ou de Padel, sendo descontado o valor correspondente ao tempo de utilização.

Em ambas as formas de prestação do serviço desportivo referidas os utentes podem beneficiar de outros serviços de caracter regular ou pontual, muito embora o foco da retenção da gestão esteja em evitar a desistência dos utentes dos serviços regulares. E por vezes as duas formas de prestação de serviço coexistem na mesma instalação, havendo portanto que ter cuidados acrescidos na promoção da retenção de cada uma das formas.

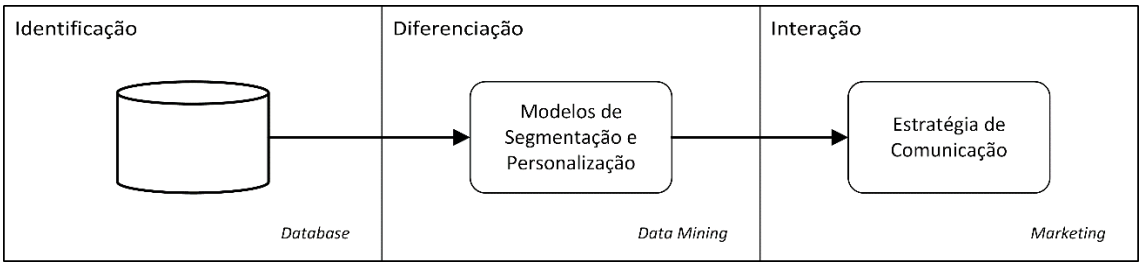
3. Trabalho relacionado

3.1. Contexto em *Database Marketing*

O processo proposto nesta dissertação para traçar os perfis dos clientes em risco de abandono e desencadear mecanismos que permitam evitar esse abandono é um processo que atravessa várias fases.

Em *Database Marketing* (Cavique, 2006) – modelo adaptado do IDIC (*Identify*, *Differentiate*, *Interact*, *Customize*) (Peppers & Rogers, 2004) – procura-se partir dos dados para identificar os utentes de serviços regulares em risco de abandono, diferenciando-os dos restantes através de técnicas de *data mining* / *predictive analytics* / *machine learning* que encontrem perfis de comportamentos, e que permitam interagir de forma personalizada e com ações de fidelização concretas com os utentes que apresentem esses perfis a fim de evitar o seu abandono.

Figura 3-1 – *Database Marketing*



Neste sentido, o trabalho relacionado aborda aspetos relevantes do conhecimento atual nas três fases consideradas: a extração, transformação, carregamento dos dados e a criação do *data warehouse* (DW) de suporte, a seleção de atributos, a aplicação das técnicas de classificação e respetivas métricas do DM, e por fim os aspetos relacionados com as ações de fidelização a tomar e com a os métodos de experimentação a introduzir para avaliar a eficácia das ações propostas.

3.2. Pré-processamento de dados: *Extract, Transform, Load* e *Data Warehousing*

Os processos de extração, transformação e carregamento (ETL) são os responsáveis pela extração de dados a partir das diversas fontes, da sua transformação e do seu carregamento no DW. De uma forma genérica Vassiliadis (Vassiliadis & Simitsis, 2009) refere que o processo “ETL nasceu no primeiro dia em que um programador construiu um programa que utiliza registos de um determinado ficheiro persistente e insere ou enriquece outro ficheiro com essa informação”.

Trujillo (Trujillo & Luján-Mora, 2003) refere que o desenho de um processo ETL compõe-se geralmente por seis tarefas: a) Selecionar a(s) origem(s) dos dados; b) Transformar a(s) origem(s) – filtrar, conversão de códigos, calcular valores derivados, transformar entre diferentes formatos de dados, gerar sequências numéricas, etc.; c) Unir as origens de forma a permitir o seu carregamento no destino; d) Selecionar o destino dos dados; e) Mapear os atributos origem com os de destino e f) Carregar os dados.

Por outro lado, Vassiliadis (Vassiliadis & Simitsis, 2009) refere que ETL é o processo de popular a tabela de factos através de um *workflow* de atividades que efetuam a filtragem, transformações e carregamento dos dados, e identifica nos três passos que dão o nome ao processo os problemas que se levantam em cada um:

a) A Extração como o processo mais simples de todos, mas dificultado devido à necessidade de provocar o mínimo transtorno na origem durante o processo de extração dos dados e pelo facto dos Administradores de Sistema não aceitarem intervenções de vulto nos seus sistemas (normalmente os sistemas origem dos dados);

b) A Transformação, que resolve problemas ao nível do esquema (conflito de nomes, conflitos estruturais, etc.), ao nível dos registos (valores duplicados, inconsistentes, etc.) e ao nível do valor dos campos (diferente representação dos mesmos valores, afetação de valores com NULL, etc.).

Outros problemas ao nível da Transformação referem-se: à possível necessidade de substituição de chaves do sistema de produção por chaves substitutas (*Surrogate Keys*) por questões de performance e homogeneidade e que pode levar à substituição geral de chaves em todo o DW; à necessidade de "desnormalização" para conseguir otimizar determinados tipos de *query* (*Pivoting* e *Unpivoting*), por exemplo, integrando várias tabelas da origem numa só no DW; as operações de atualização do DW a partir de alterações a dados efetuados na origem (*Slowly Changing Dimensions*) e que podem levar a reescrever por cima dos valores antigos, a criar um novo registo ou a utilizar atributos para colocação de valores anteriores; e por fim, a resolução de problemas relacionados com *strings* que pode levar à utilização de ontologias, *thesaurus*, ou expressões regulares, ou outras formas dispendiosas de uniformização de *strings*.

c) Por fim, é necessário proceder ao Carregamento dos dados havendo que optar por carrega-los em massa (*Bulk Insert*) ou por inserção de registos numa sequência de linhas. Questões de *performance* relacionadas com a criação de *logs*, *rollbacks* que podem ocorrer ou sobrecarga proveniente de tratamento individualizado de comandos *insert*, podem levar a optar pela primeira opção, mas questões relacionadas com a descriminação de operações com registos que são inseridos pela primeira vez e de outros que são alterados podem levar a optar pela segunda.

Apesar de algumas questões relacionadas com a limpeza e qualidade dos dados estarem incluídos no passo de Transformação, dado o âmbito alargado e de diversa natureza dos problemas relacionados, estes merecem uma abordagem cuidada ao nível da limpeza dos mesmos. Rahm (Do & Rahm, 2000) refere que a limpeza dos dados é tipicamente efetuada num estágio separado do processo ETL que ocorre antes de carregar os dados no DW. Neste trabalho, o autor identifica os problemas relacionados com a limpeza dos dados, quer os provenientes de uma única origem de dados, quer os que são provenientes de múltiplas origens.

No que diz respeito aos problemas com origem numa única fonte de dados, são identificados problemas ao nível dos atributos (valores em falta, erros ortográficos, utilização de abreviaturas, valor de um atributo embebido noutro, valores errados nos atributos), do registo (violação de dependências entre atributos), do tipo de registo (duplicação de registos, registos contraditórios) e da própria origem de dados (violação de integridade referencial).

Quando estamos perante a existência de múltiplas origens de dados, para além de poderem ocorrer os mesmos problemas oriundos de uma só origem, podem ainda ocorrer problemas ao nível das diferenças entre os modelos de esquema de dados, como os conflitos estruturais – diferentes representações do mesmo objeto - e de nomeação – o mesmo nome utilizado para objetos diferentes (homónimos) ou diferentes nomes para o mesmo objeto (sinónimos). Podem ocorrer também problemas ao nível da instância – conflitos de dados que podem ocorrer por um mesmo atributo ter codificações diferentes entre as várias origens, ou outros problemas conhecidos como diferentes níveis de

agregação, dados referentes a diferentes períodos no tempo, sobreposição de dados, identidade de objetos (*Object Identity Problem*) ou eliminação de duplicados.

Para além dos problemas de construção do DW, Gama (Gama, Carvalho, Faceli, Lorean, & Oliveira, 2017) refere que a qualidade dos dados pode afetar o desempenho dos algoritmos de DM e identifica, entre outros, um conjunto de aspetos relacionadas com a limpeza de dados e com a transformação dos dados que interessa realçar. No que concerne à limpeza, identifica as seguintes situações:

a) Dados incompletos pela inexistência de valores em alguns atributos indicando três possíveis estratégias para resolver o problema, para além de deixar o atributo por preencher: remover os registos, preencher manualmente ou utilizar um método ou heurística para os preencher de forma automática;

b) Dados inconsistentes (por exemplo, um valor 200 para a idade de uma pessoa) indicando a remoção como uma possível hipótese de resolução do problema;

c) Dados redundantes, em que atributos ou registos possuem valores muito semelhantes, apontando como possível solução a remoção das redundâncias encontradas;

d) Dados com ruído são dados que contêm objetos que, aparentemente, não pertencem à distribuição que gerou os dados analisados ou onde existem diversos *outliers*. São apontadas diversas técnicas estatísticas – baseadas em distribuições conhecidas ou pela utilização do conceito de profundidade na procura do ruído – e também técnicas como as de intervalo, de agrupamento de dados, baseados em distância e baseadas em regressão ou classificação;

e) Deteção de *outliers*, valores anormais ou extremos, sendo apenas apontadas técnicas para deteção dos mesmos;

No que diz respeito à transformação dos dados refere:

a) Conversões simbólico-numéricas, referindo-se à conversão de um valor nominal ou ordinal num valor numérico;

b) Conversões numérico-simbólicas, que convertem valores numéricos em intervalos ou categorias;

c) Transformação de atributos numéricos, necessária quando um valor numérico necessita de ser transformado noutro valor numérico porque os limites inferior e superior são muito

distantes o que causa grande dispersão ou porque os atributos se encontram expressos em escalas diferentes;

A resolução destes problemas e execução das respetivas tarefas é apresentada como um processo em si, que envolve várias fases, nomeadamente análise dos dados, definição do fluxo de transformação e regras de mapeamento, verificação, transformação e fluxo contrário, em direção à(s) origem(s), dos dados já limpos.

Galhardas *et al.* (Galhardas, Florescu, Shasha, & Simon, 2006) introduzem uma *framework* em que uma aplicação de limpeza de dados é modelada como um fluxo direto acíclico de transformações aplicado aos dados de origem. A *framework* oferece três principais serviços: serviço de transformação dos dados para efetuar conversões de esquemas de dados e normalização do valor dos atributos; serviço de correspondência multi-tabela (*Multi-Table Matching* ou *Approximate Join*), que permite produzir uma única tabela a partir de um conjunto de tabelas de *input* depois de resolver problemas de identidade dos objetos; e por fim o serviço de eliminação de duplicados para remoção de duplicados exatos e aproximados através de uma sequência de fases de correspondência, agrupamento e junção.

Ao nível do modelo conceptual e desenho dos processos de ETL, El-Sappagh *et al.* (El-Sappagh, Hendawi, & El Bastawissy, 2011) referem que a modelação dos processos ETL pode ser categorizada em três principais aproximações:

1. Modelação baseada em mapeamento de expressões – corresponde à informação necessária para reconhecer como um atributo de destino é criado a partir dos atributos origem - e diretrizes – conjunto de informação definido pelos programadores no sentido de efetuar o mapeamento entre os atributos dos dois esquemas (origem e destino);

2. Modelação baseada em construções conceptuais;

Neste tipo de modelação, Vassialiaris *et al.* (Vassiliadis, Simitsis, Georgantas, Terrovitis, & Skiadopoulos, 2005) apresenta um modelo lógico formal de um ambiente ETL em camadas (*layers*). O formato completo de um cenário ETL, que envolve atividades, conjuntos de registos e funções pode ser modelado por um grafo, a que chamam Grafo de Arquitetura. É utilizada uma *framework* de modelação de grafos uniforme tanto para modelar a estrutura interna das atividades e o cenário genérico do processo ETL, que permite o

tratamento do ambiente ETL de diferentes formas. Em primeiro lugar, o grafo de arquitetura compreende todas as atividades e armazenamento de dados de um cenário, em conjunto com os seus componentes. Em segundo lugar, captura os fluxos de dados no ambiente ETL. E finalmente compreende também a informação dos tipos das entidades envolvidas e a regulação do cenário de execução.

El-Sappagh *et al.* (El-Sappagh et al., 2011) introduzem o modelo EMD – *Entity Mapping Diagram*. No meta-modelo, são definidas duas camadas, em que a primeira é a camada de abstração em que são claramente definidos cinco objetos (função, *container* de dados, entidade, relações, atributos). Os objetos definidos na camada de abstração são uma vista de alto nível das partes ou objetos que podem ser usados para desenhar um cenário EMD. A segunda é a camada de modelo, que é uma expansão da camada de abstração. Os utilizadores podem também adicionar a sua própria camada para desenho do seu cenário EMD. Os autores incluem ainda neste trabalho uma *framework* para usar este modelo, que consiste em três partes: a de origem dos dados (*Data Source Part*), a do esquema do DW de destino (*Data Warehouse Schema Part*), e parte de mapeamento (*Mapping Part*).

3. Modelação baseada em UML.

Na vertente da modelação baseada em UML, Trujillo (Trujillo & Luján-Mora, 2003) propõe a utilização de UML (*Unified Modeling Language*), uma vez que esta largamente difundida como um *standard* para a análise e desenho de soluções OO (*Object Oriented*) e minimiza o esforço de programadores na aprendizagem de novos diagramas. Além disso, uma vez que a modelação do DW também é feita por UML, a modelação dos processos ETL da mesma forma permitem uma integração total e global. Esta abordagem conceptual permite assim reduzir o tempo de desenvolvimento, facilita a gestão dos repositórios de dados e a administração do DW, e permite fazer análise de dependências (por exemplo, estimar o impacto de uma alteração na origem dos dados).

Os fluxos dos processos ETL são complexos e normalmente são executados debaixo de estritos requisitos de performance, e a sua execução tem de ser completada numa determinada janela de tempo, pelo que a sua otimização em termos de tempo de execução e em termos de custos é crucial para satisfazer os objetivos do trabalho a desenvolver. Os trabalhos analisados, não apresentando uma metodologia geral do processo ETL devido à

diversidade das mesmas, apontam de forma clara os problemas-tipo a resolver nas diversas situações, como múltiplas e heterogéneas origens de dados, problemas de qualidade e limpeza dos dados, e problemas de carregamento de grande volume de dados no destino. Devido à complexidade envolvida, a especificação das atividades dos processos ETL é fundamental e por isso consideramos que uma ferramenta de modelação baseada em UML, como a especificada por Trujillo (Trujillo & Luján-Mora, 2003), é de grande utilidade não só por ser um *standard* muito utilizado em termos de desenvolvimento, como também refere Muñoz (Muñoz, Mazón, & Trujillo, 2010), pelo facto de haver muitas ferramentas de desenvolvimento que permitem a sua utilização.

3.3. *Data Mining*: conceitos, seleção de atributos, algoritmos e métricas

3.3.1. Conceitos

Aprendizagem supervisionada é um paradigma da Extração de Conhecimento de Dados (ECD) que usa conjuntos de dados inicialmente conhecidos, que designamos por conjunto de treino, para criar um modelo que consegue efetuar previsões. O conjunto de treino é constituído por atributos ou características (também designados por variáveis independentes) que descrevem os principais aspetos do registo, considerados os atributos de entrada, e também por um atributo alvo ou de saída. O objetivo de um algoritmo deste tipo consiste em encontrar uma função, a partir dos dados de treino, que possa ser usada para prever um valor que caracterize um novo exemplo, com base nos valores dos seus atributos de entrada (Gama et al., 2017).

Dado que o conjunto de treino é o ponto de partida para a construção deste tipo de algoritmos, para que estes apresentem melhores resultados, quer em termos de precisão, quer em termos de performance, é necessário que se tenha em consideração o volume de dados. Este volume pode estar relacionado quer com o número de exemplos (registos), quer com o número de atributos a considerar, ou mesmo com ambos. Na secção anterior foram referidas algumas técnicas relevantes para o tratamento da quantidade e qualidade dos exemplos.

3.3.2. Seleção de atributos

De seguida, analisaremos as questões relacionadas com a Seleção de Atributos (SA) (*Feature Selection*) cuja importância reside no facto de se tratar de um passo de pré-

processamento de *Machine Learning* que permite reduzir a dimensionalidade, remover dados irrelevantes, melhorar a precisão e a compreensão dos resultados (Yu & Liu, 2003). Outros benefícios da seleção de atributos e variáveis, são por exemplo facilitar a visualização e compreensão dos dados, reduzir os requisitos de armazenamento e de medida, reduzir os tempos de treino e utilização, e reduzir a dimensionalidade para melhorar a performance da previsão (Guyon & Elisseeff, 2003).

Dash (Dash & Liu, 1997) refere que a SA tenta selecionar o menor subconjunto de atributos de acordo com um dos seguintes critérios: não degradar significativamente a precisão da classificação e a classe de distribuição resultante tendo em conta os atributos selecionados, seja tão próxima da distribuição original que conta todos os atributos.

São identificados dois passos importantes na escolha do método de SA:

- O procedimento de geração (*Generation Procedure*), que gera os subconjuntos de atributos a analisar num máximo de 2^n, em que n é o número total de atributos; são consideradas três diferentes abordagens para os procedimentos de geração: a Completa, onde o procedimento faz uma procura completa, não obrigatoriamente exaustiva, do subconjunto de acordo com a função de avaliação, a Heurística, em que são considerados todos os restantes atributos em cada uma das iterações, e a Aleatória, em que a geração de subconjuntos aleatória depende de parâmetros a passar ao procedimento;
- A função de avaliação (*Evaluation Function*), que mede a capacidade do conjunto de atributos ser capaz de distinguir as diferentes classes, que pode pertencer a uma de cinco categorias: medida de distância, medida de ganho de informação, medida de dependência ou correlação, medida de consistência ou medida da taxa de erro de classificação.

O trabalho apresenta uma *framework* para a análise de algoritmos combinando os diferentes tipos de procedimentos de geração com as diferentes categorias de funções de avaliação, identifica um conjunto de métodos e testa-os em conjuntos de dados artificiais que incluem combinações de atributos relevantes, irrelevantes, redundantes e com ruído.

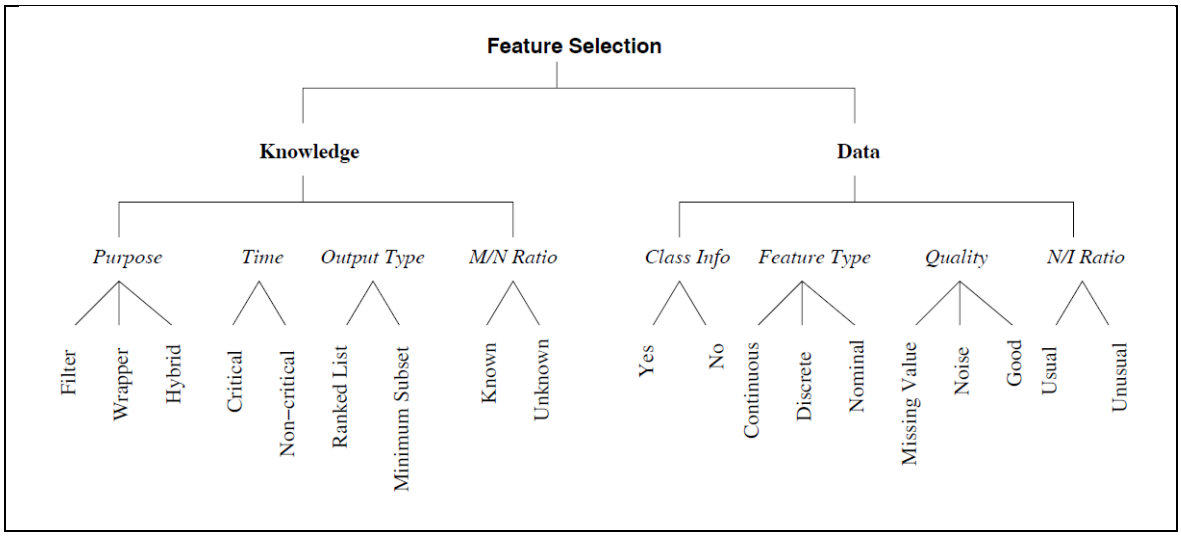
Tendo em consideração que a escolha do método de SA depende de várias características do conjunto de dados (tipo de dados, dimensão e ruído presente) é apresentado um conjunto de linhas orientadoras para identificar as capacidades e características dos métodos avaliados: capacidade para gerir diferentes tipos de dados (contínuos, discretos e

nominais), capacidade de gerir múltiplas classes (mais de duas), capacidade de gerir grandes conjuntos de dados, capacidade de lidar com ruído e capacidade de produzir um conjunto ótimo de atributos não havendo ruído.

Liu (Liu & Yu, 2009) apresenta uma plataforma para permitir uma escolha do método a partir do ponto de vista do utilizador, tendo por base o seu conhecimento e os dados existentes, como ilustra a Figura 3-2.

Por um lado, o fator Conhecimento (*Knowledge*) tem em conta o objetivo da seleção de atributos, o tempo de execução, o tipo de output pretendido, e o rácio entre o número esperado de atributos M e o número total de atributos N. Por outro lado, o fator Dados (*Data*) tem em conta a classe da informação, o tipo de atributo, qualidade dos dados e rácio obtido a partir do número total de atributos N e o número de instâncias I.

Figura 3-2 - Framework para escolha do método de SA



Fonte: (Liu & Yu, 2009)

Esta plataforma permite atingir dois objetivos: agrupar algoritmos existentes com características similares e encontrar as suas forças e fraquezas, e providenciar linhas gerais para a construção de um sistema inteligente de seleção de atributos. O algoritmo encontrado parte do rácio M/N, na medida de consistência e de quatro algoritmos representativos que empregam diferentes estratégias de pesquisa: Focus (*Forward Exhaustive Search*), ABB (*Backward Complete Search*), QBB (*Random Search* e ABB) e *SetCover* (*Sequential Search*). Todos os algoritmos partilham as seguintes características de

conhecimento: usam um critério de avaliação independente (*Filter Model*), procuram um conjunto mínimo de atributos tendo em conta que o tempo de execução é crítico.

Guyon (Guyon & Elisseeff, 2003) foca-se na seleção de um subconjunto de atributos que seja útil para construir um bom sistema de previsão, que pode incluir atributos redundantes, o que nem sempre pode estar em linha com o encontrar ou ordenar as variáveis mais relevantes, que pode excluir atributos redundantes. Descreve a utilização de Métodos de Filtro (*Filter Methods*) que selecionam variáveis pela sua importância tendo em atenção os coeficientes de correlação e indica as vantagens: simplicidade, escalabilidade e os sucessos já obtidos através da implementação prática; e limitações deste processo: eventual redundância dos atributos, pouco ou nenhum ganho de informação com variáveis perfeitamente correlacionadas, ou não consideração de possíveis melhorias de performance através da utilização de variáveis que sozinhas são inúteis mas que em conjunto com outras poderão não o ser.

Este trabalho identifica outras famílias de métodos como os *Wrappers* (*Wrapper Methods*), que definem conjuntos de variáveis de acordo com a sua utilidade para uma dada previsão, utilizando métodos como *Forward Selection* e *Backward Elimination*, métodos embebidos (*Embedded Methods*), que incorporam a seleção de variáveis como parte do processo de treino, métodos de subconjuntos encadeados (*Nested Subset Methods*), que combinam métodos embebidos com estratégias de pesquisas *Greedy* sobre subconjuntos encadeados de variáveis, e por fim referem a otimização direta e objetiva (*Direct Objetive Optimization*) em que otimizam diretamente as funções objetivas de duas partes: um termo de qualidade de ajustamento e uma penalidade pela utilização de uma grande quantidade de atributos.

O artigo refere ainda métodos para a construção de atributos e para reduzir o espaço de dimensões: *Clustering*, substituir um grupo de variáveis similares por uma através de algoritmos como o K-*means*, Factorização Matricial, por exemplo o SVD (*Singular Value Decomposition*) que consiste em formar um conjunto de variáveis que são combinações lineares das variáveis originais, ou SDR (*Sufficient Dimensionality Reduction*) em que as variáveis mais informativas são extraídas através das soluções de um problema de otimização que monitora o compromisso entre a reconstrução e a compressão de dados,

ou a SA Supervisionada, que podem seguir as linhas dos métodos de subconjuntos encadeados, de filtro ou de otimização objetiva direta.

Este trabalho tem ainda o mérito de referir alguns aspetos relevantes na problemática da SA, tais como a variância das variáveis do subconjunto selecionado, a posição das variáveis no contexto de outras, a seleção de variáveis não supervisionada, o *forward* e o *backward selection*, o problema das multi-classes, etc.

Yu (Yu & Liu, 2003) apresenta problemas relacionados com os tipos de algoritmos do modelo de filtro (*Feature Weighting Algorithms* e *Subset Search Algorithms*) para encontrar subconjuntos adequados de atributos em grandes dimensionalidades. Por um lado, os algoritmos puros de *Feature Weighting* não respondem adequadamente aquando da existência de muitos atributos redundantes. Por outro lado, uma vez que os algoritmos de *Subset Search* apresentam complexidade quadrática ou maior em termos de dimensionalidade, não apresentam por isso grande escalabilidade nesta vertente.

Para responder a estas questões, apresenta um modelo baseado no conceito de Correlação Predominante, que parte do princípio que, se a correlação entre duas variáveis é uma boa medida, pode-se considerar um bom atributo se está muito correlacionado com a classe mas não com os outros atributos. Neste sentido, o problema da SA resume-se a encontrar uma medida adequada de correlações entre atributos e um procedimento para selecionar atributos baseado nesta medida.

Cavique (Cavique, Mendes, Funk, 2011) combina a flexibilidade de algoritmos *Rough Sets* e Logical *Analysis of Data* (LAD) para apresentar o algoritmo *Logic Analysis of Inconsistent Data* (LAID).

O objetivo dos *Rough Set* é o de obter regras de decisão a partir de uma tabela de 3 entradas - uma com as observações, outra com os atributos e outra com o atributo de decisão - para encontrar o número mínimo de atributos necessários para explicar a classe. Os atributos podem conter outros tipos de valor para além dos binários.

Por outro lado, o LAD tem por objetivo descobrir um mínimo número de atributos necessários para explicar todas as observações e detetar padrões escondidos num conjunto de dados com duas classes. As observações são classificadas como positivas ou negativas

de acordo com uma função e o objetivo deste método é aproximar esta função com uma união de intervalos.

O objetivo de ambos os métodos (*Rough Set* e LAD) é o de reduzir o número de atributos e de gerar regras para classificar o conjunto de dados disponível. Ambos os procedimentos têm dois passos: o passo de transformação e o passo de redução do número de atributos.

Ambos os métodos apresentam vantagens e desvantagens:

- Os *Rough Set* não excluem ou corrigem as inconsistências dos dados, permitindo obter regras de decisão discordantes, tornando difícil a sua interpretação pelo utilizador. Permite contudo a utilização de dados não binários;
- O LAD apenas trabalha com dados binários, pelo que é necessário utilizar um processo de discretização quando se pretende usa-lo com valores não binários. Por outro lado permite a utilização de custos associados aos atributos, o que permite não só otimizar o número de atributos mas também o custo total; apresenta uma abordagem sistemática, precisa e robusta que evita ambiguidades e é de fácil interpretação pelos utilizadores;

No LAID, cada inconsistência é corrigida pela introdução de uma nova variável "*je ne sais quois*" ("jnsq") que é testada de forma a que os procedimentos do LAD possam ser usados sem necessidade de haver alguma alteração. Desta forma, se duas observações são repetidas mas pertencem a diferentes classes, é adicionada uma nova variável. Se três ou quatro observações são repetidas, então é necessário adicionar duas novas variáveis, e assim sucessivamente.

Para reduzir o conjunto de dados, é introduzido um algoritmo de duas fases (*Two-phase algorithm*). Primeiro o problema é transformado através da geração de uma matriz com uma restrição disjunta. De seguida é escolhido o subconjunto de atributos utilizando o conhecido *Set Covering Problem*. Um dos passos do algoritmo consiste em gerar a matriz disjunta aplicando de seguida uma abordagem heurística para obter uma solução mínima do *Set Covering Problem*.

Esta nova técnica inclui a tolerância a inconsistências e a multiplicidade de classes do *Rough Set*, além da eficiência e da otimização de custos do LAD. O LAID não exclui mas corrige as inconsistências através da nova variável acrescentada "jnsq". A integração das duas

abordagens é tão próxima que se pode considerar o LAID como uma extensão do *Rough Set*.

A SA tem sido uma área de grande investigação e desenvolvimento desde a década de 70 e tem provado ser efetiva para remover atributos irrelevantes e redundantes, aumentando a eficiência das tarefas de aprendizagem, aumentando a precisão das previsões e a compreensão dos resultados obtidos (Yu & Liu, 2003).

De acordo com Chandrashekar (Chandrashekar & Sahin, 2014) a escolha de algoritmos de SA só pode ser feita com o conjunto de dados, uma vez que cada algoritmo se comporta de forma diferente para dados diferentes. Em função da aplicação pretendida, podemos escolher o algoritmo de SA de acordo com a sua simplicidade, estabilidade, número de atributos reduzidos, precisão da classificação obtida, armazenamento e requisitos computacionais. A aplicação de um algoritmo de SA permitirá obter melhores modelos de classificação.

Alguns aspetos na SA não devem ser desconsiderados, tal como refere Guyon (Guyon & Elisseeff, 2003), como por exemplo a análise da variância das variáveis do subconjunto de variáveis selecionado uma vez que pode ser um sintoma de um mau modelo, de que os resultados não são reproduzíveis, que o subconjunto falha na reprodução da visão geral, ou mesmo problemas relacionados com os critérios utilizados (saliência, entropia, suavidade, densidade ou confiabilidade.

Tal como nos processos ETL interessa analisar critérios ou metodologias para escolha do método de SA a utilizar. Liu (Liu & Yu, 2009) apresenta uma metodologia interessante direcionada para o utilizador que tem em conta o seu conhecimento e o conjunto de dados a tratar.

3.3.3. Algoritmos

Partindo dos dados de treino, formado por um subconjunto de registos e atributos selecionados como visto anteriormente, pretende-se formular um modelo ou hipótese capaz de relacionar os valores dos atributos de entrada de um objeto com o valor do seu atributo de saída. Se o domínio do valor a prever for um conjunto de valores nominais, estamos perante um problema de classificação (Gama et al., 2017).

Classificação é pois a tarefa de assignar objetos a uma de várias categorias pré-definidas (Tan, Steinbach, & Kumar, 2006a). Os dados de entrada para uma tarefa de classificação são uma coleção de registos. Cada registo (instância ou exemplo) é caracterizado por um tuplo (x, y), em que x é um conjunto de atributos e y é um atributo especial designado por classe, categoria ou atributo de destino. Nas técnicas de classificação os atributos que fazem parte do conjunto x podem ser discretos ou contínuos, e o atributo de destino tem de ser discreto. Um modelo de classificação pode também ser utilizado como um modelo preditivo para prever uma classe de registos desconhecidos.

De acordo com vários autores (Mahajan et al., 2015), (Pinheiro & Cavique, 2015), (Siegel, 2013), (Tan, Steinbach, & Kumar, 2006b), as árvores de decisão (*decision trees*) são técnicas de classificação adequadas e das mais utilizadas na aplicação de *Data Mining (DM)* a estudos de retenção em várias áreas de negócio.

As árvores de decisão são construídas com base numa parte dos dados disponíveis no DW e a partir da sub-divisão de grupos de características, de acordo com as correlações e padrões encontrados nos atributos relevantes da base de dados que levam à construção de um modelo preditivo que permite a construção de regras do tipo "*if-then-else*".

Como explica Gama (Gama et al., 2017) uma árvore de decisão é um grafo acíclico em que cada nó ou é um nó de divisão, com dois ou mais sucessores, ou um nó folha: um nó folha é rotulado, nos casos mais simples, com uma constante correspondente à moda (especificamente em problemas de classificação); e um nó de divisão que contém um teste condicional baseado nos valores do atributo.

Uma vez estruturadas as regras do tipo "*if-then-else*" numa árvore de decisão, estas tornam-se muito fáceis de interpretar apesar de encontrar uma árvore de decisão ótima ser um problema *NP-complete*.

Para além disso, Tan (Tan et al., 2006a) refere ainda outras características e vantagens das árvores de decisão: não exigem quaisquer pressupostos quanto ao tipo de distribuição de probabilidade da classe ou outros atributos; as técnicas para construir árvores de decisão não são computacionalmente dispendiosas possibilitando a construção rápida de modelos mesmo quando o tamanho do conjunto de treino é muito grande, e mesmo depois de construída a árvore classificar um registo de teste é extremamente rápido e tem

complexidade O(w), em que w é a profundidade máxima da árvore; providenciam uma representação expressiva de funções de valores discretos (embora com alguns problemas de generalização para determinados tipos de problemas com valores booleanos); os algoritmos são bastante resistentes à presença de ruído especialmente quando são utilizados métodos para evitar o *overfitting*; a presença de atributos redundantes (atributos fortemente correlacionados) não afeta a precisão das árvores de decisão, embora possam resultar árvores maiores, o que pode ser evitado com a utilização de algoritmos de SA; Gama (Gama et al., 2017) refere também que o processo de construção de uma árvore seleciona os atributos a usar no modelo de decisão pelo que esta SA produz modelos que tendem a ser bastante robustos em relação à adição de atributos irrelevantes e redundantes.

Como uma desvantagem dos algoritmos de árvores de decisão são apontados os métodos de particionamento recursivos *top-down*, que levam a que os nós inferiores da árvore possam ter um número reduzido de registos – problema conhecido como *data fragmention*. É possível ultrapassá-lo desabilitando a continuação do *splitting* quando o número de registos for inferior a um determinado limite pré-estabelecido.

Outros problemas referidos são o facto de que uma sub-árvore pode estar replicada várias vezes numa mesma árvore o que a torna mais complexa e provavelmente mais difícil de interpretar; e a expressividade da árvore de decisão para modelar relações complexas entre atributos contínuos pode ser limitada;

Gama *et al.* (Gama et al., 2017) acrescenta ainda que uma vez que uma árvore de decisão é uma hierarquia de testes, se o valor de um atributo é desconhecido, surge um problema em decidir que ramo seguir sendo portanto importante que os algoritmos apresentem mecanismos para lidar com valores omissos.

Algumas técnicas utilizam um mesmo algoritmo para gerar classificadores diferentes. A técnica de *Bootstrap Aggregating* (também conhecida por *bagging*) produz réplicas do conjunto de treino por amostragem com reposição. Os classificadores obtidos são usados para classificar cada exemplo no conjunto de teste, sendo que a classificação final é feita através de um esquema de voto uniforme, em que cada classificador contribui de forma igual para a classificação final. Uma vez que as árvores de decisão geram modelos instáveis,

a utilização do voto maioritário dos vários classificadores faz com que se obtenha grandes melhorias nos resultados obtidos uma vez que diminui a variabilidade de cada classificador. Outro algoritmo, conhecido por *boosting*, é iterativo, sendo que em cada iteração associa um peso diferente a cada exemplo do conjunto de treino gerando-se vários e diferentes classificadores. Em cada iteração o peso dos exemplos mal classificados aumenta reduzindo-se o peso dos que foram corretamente classificados. No final das iterações, o classificador final agrega os vários classificadores, em que cada um obtém um peso que é função da sua precisão.

Dada a utilização de pesos a técnica de *boosting* tem como principal vantagem a redução da variância e da variabilidade aleatória das hipóteses individuais.

O algoritmo de florestas aleatórias (*random forests*) é baseado em injeções de aleatoriedade, em que os exemplos são obtidos por amostragem com reposição do *bagging* e a seleção de atributos é feita de forma aleatória, o que gera várias árvores de decisão cujas previsões são combinadas por votação uniforme. Resultados experimentais têm revelado que este algoritmo é dos mais competitivos (Gama et al., 2017).

Em comparação direta com as árvores de decisão, estas últimas técnicas (*bagging*, *boosting* e *random forests*) apresentam melhorias significativas de precisão. Em contrapartida, dado gerarem múltiplos classificadores, há um aumento da complexidade e da dificuldade de interpretação das regras obtidas.

3.3.4. Métricas

Existem vários métodos de avaliação da performance de um modelo. Tan (Tan et al., 2006a) refere quatro métodos e aponta as limitações de cada um.

No método da retenção (*Holdout method*) os dados originais são particionados em dois conjuntos disjuntos designados por conjunto de treino e conjunto de teste. O modelo é induzido a partir do conjunto de treino e depois testado contra o conjunto de teste. Normalmente são utilizados 70% dos dados para treino e 30% para teste, apesar de poderem ser utilizadas outras proporções. A precisão do modelo é calculada a partir da aplicação do modelo aos dados de teste.

O método das amostras aleatórias (*Random Sampling*) aplica várias vezes o método da retenção com diferentes conjuntos de treino e de teste, calculando a precisão do modelo com a média das precisões obtidas nas várias aplicações do método da retenção.

No método da validação cruzada (*Cross-Validation*) cada registo é usado o mesmo número de vezes para treinar e apenas uma vez para testar. Na validação cruzada *k-fold* os registos são segmentados em k partições de igual tamanho. Em cada execução, uma das partições é usada para testar enquanto as outras são usadas para treinar o modelo. Este procedimento é repetido k vezes para que cada partição seja usada apenas uma vez para testar o modelo.

Finalmente no método *Bootstrap* a amostragem dos registos de treino são escolhidos com substituição, isto é, um registo colocado para treinar o modelo pode de novo ser selecionado para o mesmo efeito. Os registos que não fazem parte da amostra de treino fazem parte dos registos de teste onde é aplicado o modelo e avaliada a sua precisão. O processo de amostragem é repetido *b* vezes para gerar *b* amostragens *Bootstrap*.

As limitações apontadas ao método da retenção referem-se ao facto de ficarem menos registos disponíveis para treino dada a reserva de registos para teste. Por outro lado, é apontado que este método apresenta um grande intervalo de confiança uma vez que quanto mais pequeno for o conjunto de treino maior será a variância do modelo ou, se pelo contrário o conjunto de treino for muito grande, a precisão calculada com a aplicação do modelo sobre os poucos registos de testes que sobram será menos fiável. Corre-se ainda o risco de uma determinada classe poder estar muito representada num conjunto e pouco no outro, o que pode fazer incorrer o modelo em menor precisão.

O método das amostras aleatórias sofre de alguns dos mesmos problemas do método anterior, nomeadamente pelo facto de continuarem a estar disponíveis menos registos para treino do modelo. Para além disso não tem controlo sobre o número de vezes que um registo é usado para treino e para teste o que pode fazer com que um registo seja mais usado para treino que outros desvirtuando assim o modelo.

No caso da validação cruzada mais registos estão disponíveis para treino e os conjuntos de teste são mutuamente exclusivos, havendo no entanto um custo computacional superior para apurar a precisão do modelo.

Escolhido o método de avaliação do modelo, a avaliação processa-se através da contagem de registos de teste, que envolvem a fração restante dos dados do DW de suporte à construção do modelo, e que o modelo classifica corretamente, os verdadeiros positivos e verdadeiros negativos, e incorretamente, os falsos positivos e falsos negativos. Estas quatro contagens permitem construir uma matriz que se designa por Matriz de Confusão sobre a qual são calculadas as métricas de avaliação do modelo.

As variações na matriz de confusão correspondem a características específicas dos dados e podem ou não provocar alterações nos valores das métricas. Sokolova (Sokolova & Lapalme, 2009) denomina estas sensibilidades das métricas como Invariância. Assim, uma medida é invariante se o seu valor não muda quando há alterações na matriz de confusão. Estas propriedades podem ou não ser benéficas em função do objetivo com que se considere a métrica.

Ainda Sokolova (Sokolova & Lapalme, 2009) apontam as seguintes métricas para os modelos de classificação binária:

Accuracy rate (vs Error rate): indica a eficácia geral do modelo uma vez que diz respeito a todos os resultados corretamente classificados (ou mal classificados, no caso da *Error rate*). Esta medida é invariante no que diz respeito aos resultados corretamente classificados (positivos e negativos) uma vez que não distingue os verdadeiros positivos dos verdadeiros negativos (ou o caso contrário, para a *Error rate*).

Precision: indica a concordância do modelo com os valores positivos corretamente classificados. Esta medida é invariante no que diz respeito à correta classificação dos resultados positivos, negligenciando a classificação correta dos negativos.

Recall/Sensitivity/True Positive Rate: indica a efetividade do modelo a identificar casos positivos. Tal como a métrica anterior, esta medida é invariante no que diz respeito à correta classificação dos resultados positivos, negligenciando a correta classificação dos casos negativos.

Specificity/True Negative Rate: aponta para a efetividade do modelo na classificação de casos negativos. Esta métrica é invariante no que diz respeito a alterações na contagem dos verdadeiros positivos e falsos negativos, bem como a alterações uniformes de positivos e negativos.

F-score: indica a relação entre os casos positivos realmente existentes e os casos positivos identificados pelo classificador. Uma vez que se trata de uma métrica de síntese das métricas *Precision* e *Recall*, também se aplica a propriedade de invariância no que diz respeito à classificação dos casos negativos.

As métricas referidas anteriormente apresentam valores entre 0 e 1 (ou entre 0 e 100%).

Outro indicador de qualidade do modelo, o ***Kappa***, indica a extensão em que a concordância observada excede a concordância hipoteticamente esperada (Landis & Koch, 1977). Os valores de *Kappa* constam no intervalo [-1,1]. Quando *Kappa*=1 está-se perante uma concordância perfeita, quando *Kappa*=0 a concordância é aleatória e quando *Kappa*=-1 está-se perante uma discordância "perfeita" (Eugenio & Glass, 2004).

Por diversos fatores, quer pelo facto de se poder chegar a resultados discordantes, quer pela arbitrariedade das escalas de concordância, esta métrica tem vindo a cair em desuso (Eugenio & Glass, 2004) (Pontius et al., 2011).

A tabela 3-1 apresenta as fórmulas de cálculo das métricas indicadas anteriormente.

Tabela 3-1 - Fórmulas de cálculo das métricas de avaliação de modelos de DM

Métrica		Formula
Accuracy		$(tp + tn) \,/\, (tp + fp + tn + fn)$
Precision		$(tp) \,/\, (tp + fp)$
Sensitivity / True Positive Rate / Recall		$(tp) \,/\, (tp + fn)$
F-score		$\left((\beta^2 + 1)\, tp\right) / \left((\beta^2 + 1)\, tp + \beta^2 fn + fp\right)$
		$(2tp) \,/\, (2tp + fp + fn)$
Specificity / True Negative Rate		$(tn) \,/\, (fp + tn)$
Kappa	K_0	$(tp + tn) \,/\, (tp + fp + tn + fn)$
	K_e	$\left((tn + fn)(tn + fp) + (fp + tp)(fn + tp)\right) / (tp + fn + fp + tn)^2$
		$(K_0 - K_e) \,/\, (1 - Ke)$

Uma outra forma de avaliar os modelos preditivos de classificação é através da análise ROC (*Receiver Operating Characteristic*).

De acordo com Ferri-Ramírez *et al.* (Ferri-Ramírez, C; Flach, P; Hernández-Orallo, 2002), a análise ROC permite avaliar a performance de um modelo de uma forma mais independente e completa do que usando outras métricas, como a *Accuracy*. É normalmente utilizada em modelos preditivos binários e é fácil de definir e de interpretar. Uma métrica utilizada em conjunto com a análise ROC é a *AUC (Area Under the Curve)* que aponta para a habilidade do modelo em evitar classificações falsas.

O gráfico ROC é desenhado colocando o rácio *True Positive Rate* (*Sensitivity*) no eixo y, e o *False Positive Rate* (sendo que *False Positive Rate* = 1 – *True Negative Rate* ou *False Positive Rate* = 1 - *Specificity*) no eixo x, sendo que cada ponto marcado representa o *tradeoff* entre os falsos positivos e os falsos negativos.

O ponto (0,0) representa o ponto em que o modelo classifica tudo como negativo, e o ponto (1,1) onde o modelo classifica tudo como positivo. No caso de uma classificação perfeita do modelo, a curva ROC atingirá o ponto teórico de concordância perfeita do modelo, em que *Sensitivity* = 100% e a *Specificity* = 100%, e a métrica AUC terá o valor mais alto, isto é, a maior probabilidade (Parodi et al., 2012).

Normalmente neste gráfico, para termos comparativos da performance do modelo, traça-se a linha em que x = y que corresponde à obtenção, de forma aleatória, do valor da classe que se pretende prever.

Uma vez que os modelos binários são representados por pontos individuais no espaço ROC, o ponto obtido é facilmente representável no gráfico.

Este método tem ainda a particular vantagem de permitir comparar de forma simples dois modelos comparando as suas curvas: se as duas curvas não se intersectam, um dos modelos predomina sobre o outro; se houver interseção, um dos métodos é melhor para alguns rácios de custo, e o outro é melhor para outros rácios de custo.

Para valores de AUC iguais ou inferiores a 0.5 considera-se estar perante um modelo aleatório e para valores próximos ou iguais a 1 perante um modelo muito bom ou perfeito.

A significância das métricas apresentadas e traduzidas nas fórmulas da tabela 3-1 permite aferir a pertinência destas métricas no que diz respeito à avaliação dos modelos preditivos a criar.

No caso concreto dos modelos passiveis de serem utilizados para a segmentação base das ações de fidelização, salienta-se a relvância das métricas *accuracy*, dado avaliar os casos corretamente classificados (positivos ou negativos), a *precision*, a *sensitivity* e *f-score* uma vez que relacionam a concordância do modelo com os casos positivos (aqueles que realmente se prevê virem a ser desistentes), e por fim a *specificity* dado apontar a conformidade do modelo com os casos negativos.

3.4. Retenção em serviços desportivos, ações de fidelização e experimentação

Segundo Surujlal & Dhurup (Surujlal & Dhurup, 2012) existem vários benefícios associados com a retenção dos clientes: o custo de aquisição de um cliente só tem lugar no inicio da relação e portanto quanto maior for a duração da relação, menor será o custo; a probabilidade de os clientes de longa-duração mudarem é baixa, tendem a ser menos sensíveis aos preços e são mais propensos a dar referências; estão também mais propensos a adquirirem produtos auxiliares e suplementos que têm margens maiores; e devido ao seu conhecimento da organização é menor o custo de os servir; Marisa (Marisa & Pedragosa, 2006) refere também que um dos fatores estratégicos de desenvolvimento dos centros especializados para a atividade desportiva é a sua orientação para a fidelização dos consumidores aos produtos/serviços oferecidos. Independente do estado de evolução de um *health-club* no mercado, o seu objetivo principal é fidelizar sócios, cujos comportamentos de compra se tornam repetitivos ao longo do tempo ignorando a agressiva concorrência. Hoffman *et al. apud* Marisa (Marisa & Pedragosa, 2006) indicam que 95% dos lucros são originados por clientes de longo-prazo, e são derivados das vendas, custos operacionais reduzidos e recomendações.

Devido às grandes taxas de abandono neste sector há muito que a retenção nos serviços desportivos é uma preocupação, pelo que muitos trabalhos têm sido desenvolvidos nesta área, quer sectorialmente nomeadamente na área dos Ginásios, Health-Clubs e Academias, quer de forma geral para todo o tipo de instalações desportivas.

Avourdiadou & Theodorakis (Avourdiadou & Theodorakis, 2014) procuram determinar o impacto da qualidade dos serviços e da satisfação na lealdade entre novos utentes e utentes experientes no contexto de centros de desporto e *fitness*, e concluem pelos resultados obtidos que a qualidade do serviço afeta de forma consistente a satisfação geral quer dos novos, quer dos utentes experientes, embora sendo a principal impulsionadora da lealdade juntos dos novos utentes. No entanto, adiantam que a longo prazo a satisfação geral é o maior orientador do comportamento futuro dos utentes.

Surujlal & Dhurup (Surujlal & Dhurup, 2012) identificam várias estratégias de relações com os clientes e de retenção tendo concluído que a maior parte dos ginásios e *health-clubs* implementam estratégias direcionadas para gestão de reclamações, para lidar com a

concorrência, para monitorizar os serviços e a rendibilidade, e de incentivo à renovação das inscrições por parte dos utentes. Referem ainda como estratégias comuns a melhoria da qualidade dos serviços, utilização de equipamento atualizado e tecnologias emergentes.
Howat (Howat & Assaker, 2016) procura sistematizar dimensões que influenciem a lealdade dos utentes. Através das dimensões "apresentação das instalações", "serviços principais disponibilizados", "pessoal" e "existência de parqueamento" agrupadas em "qualidade do processo"; "sucesso na competição", "sociabilidade", "saúde e fitness", "relaxação" e "libertação do stress" agrupadas em "qualidade dos resultados"; e finalmente o "valor" e "satisfação global"; concluem que as dimensões agrupadas na "qualidade do processo" e a "satisfação global" são as que têm mais impacto na lealdade, seguidas da "qualidade dos resultados" e do "valor".
Procurando conhecer as razões que podem influenciar a atitude de permanência num serviço de *fitness* ou GHC, seguindo a linha de vários autores Gonçalves (Gonçalves, 2012) considera dois tipos de condicionantes: as variáveis externas e as internas ao individuo.
Como variáveis externas identifica as variáveis situacionais e ambientais, as sociais e culturais e os grupos de referência e classe social; e como variáveis internas o género, a idade e a etnicidade, a satisfação, a motivação e as expetativas, a perceção e atitudes, a personalidade e a imagem de si próprio e os estilos de vida e bem-estar.
Como resultado, a autora conclui que há uma distinção de comportamento relativamente à retenção entre os géneros (o género feminino apresenta mais retenção), pela idade (os menores de 20 anos e os maiores de 65 anos mostram maior retenção) e de comportamento perante a prática (os sócios que têm uma duração de treino maior revelam maior retenção).
Num trabalho sobre Gestão da Retenção Frota (Frota, 2011) compila diversos estudos na área da psicologia do exercício e estudos efetuados por grandes organizações de suporte à gestão de *Health-Clubs* para procurar compreender os fatores que podem levar ao abandono e aponta algumas estratégias utilizadas para melhorar as adesões e a retenção, entre outras a prescrição de exercício, a regulação da frequência e da intensidade do exercício, bem como a duração do mesmo, para além de outras no âmbito da qualidade das instalações, pessoal e qualidade do atendimento.

A Tab. 3-2 sintetisa de forma genérica os fatores que influenciam a lealdade dos utentes encontrados nos trabalhos relacionados analisados.

Tabela 3-2 – Síntese dos fatores que influenciam a lealdade dos Utentes

Género, Idade, Etnicidade, Classe social, Grupos de referência, Estilos de vida e bem-estar, Personalidade, Sociabilidade, Motivação	Demográficos
Serviços disponibilizados, Parqueamento, Sucesso na competição	Contratação do serviço
Frequência, Intensidade do exercício, Duração do exercício	Frequência
Pessoal, Prescrição do exercício, Apresentação das instalações, Equipamento atualizado, Utilização de tecnologias emergentes, Gestão das reclamações, Lidar com a concorrência, Satisfação geral, Expetativas, Perceção e atitudes	Qualidade do serviço

Da leitura dos trabalhos relacionados com a retenção em serviços desportivos referidos anteriormente, interessa reter que na sua totalidade os trabalhos assentam na realização de inquéritos aos utentes ou aos próprios gestores das instalações e nunca baseados nos dados constantes nos sistemas de faturação e controlo deste tipo de instalações.

Interessa ainda referir de novo Frota (Frota, 2011) que aponta como principais indicadores do controlo da retenção os conceitos de Permanência Média (somatório dos meses de permanência dos membros que cancelaram a dividir pelo número de membros que cancelaram), Cancelamento Precoce (menos de 6 meses de permanência ou subdividindo em menos de 3 meses e entre 4 a 6 meses de permanência), Taxa de cancelamento – *Attrition* (soma do número de cancelamentos dos últimos 12 meses a dividir pela média do número de clientes ativos), Taxa de Retenção (1 – Taxa de cancelamento), Turnover (número total de cancelamentos de um mês a dividir pelo número de clientes ativos), Nº de acessos mensais e NPS – *Net Promotor Score*.

Frota (Frota, 2011) refere a IHRSA num estudo realizado sobre a evasão dos membros dos Ginásios e Health-Clubs tendo concluído que 25% das desistências estão relacionadas com o clube, sendo que 45% destas são recuperáveis, 24% abandona por razões pessoais, 22% dos abandonos relacionam-se com o dinheiro, sendo 31% destas recuperáveis, e finalmente 29% referem-se a problemas situacionais, dos quais se podem recuperar 44%, resumindo que de uma forma geral, nos clubes exclusivamente de *fitness* a evasão dos membros ronda os 40 a 50% por ano, enquanto que nos clubes com outras atividades, que dispõem de piscina e de pavilhões polidesportivos, a evasão é um pouco menor encontrando-se entre os 35 a 45%.

Desta forma, a capacidade de detetar o mais cedo possível quais os clientes que irão abandonar ou deixar de adquirir um determinado produto ou serviço é uma questão que todas as empresas e entidades vendedoras ou prestadoras de um serviço gostariam de ver respondida, porque atrair um novo cliente pode custar até cinco vezes mais do que manter um cliente atual satisfeito.

Por outro lado, em vez de dispersarem os seus esforços, as empresas procuram cada vez mais concentrar-se nos clientes em que terão mais hipóteses de os satisfazer e/ou nos que serão mais rentáveis. Por isso, procuram identificar padrões de semelhanças e necessidades em grupos de clientes, processo que se denomina por Segmentação.

De acordo com Kotler (Kotler & Keller, 2009) os consumidores podem ser segmentados de acordo com as suas características geográficas, demográficas, psicográficas e comportamentais. Para serem uteis, os segmentos devem obedecer a cinco critérios: devem ser **mensuráveis** no que diz respeito ao tamanho, ao poder de compra e às características que o definem -, **substanciais** – devem ser suficientemente grandes e homogéneos para que justifiquem o desenvolvimento de um programa de marketing -, **acessíveis** – no sentido de que é possível chegar aos clientes que constituem o segmento - , **diferenciáveis** – no sentido de que a diferenciação de características corresponde a necessidades diferentes - e **acionáveis** – no sentido de que são realizáveis ações direcionadas para o(s) segmento(s).

No modelo de *Database Marketing* (Cavique, 2006) inspirado no modelo IDIC (*identify, differentiate, interact, customize*) de Peppers & Rogers (Peppers & Rogers, 2004), pretende-se satisfazer a necessidade dos clientes e construir uma relação lucrativa e duradoura ao comunicar regularmente com o cliente certo, utilizando a oferta de produtos certa, e com a mensagem certa (emitida no momento certo, através do canal certo). Para construir um modelo deste tipo identificam-se três passos: a preparação dos dados suportados num DW, a segmentação e personalização, e por fim a estratégia de comunicação (Marketing). Para a segmentação e personalização, Cavique (Cavique, 2003) parte do *Customer Lifetime Value* (CLV) – medida calculada com base no lucro resultante do total de transações dos clientes durante o seu "período de vida" – e da técnica de segmentação RFM (Recenticidade da última visita, Frequência de compras e valor

Monetário), e propõe o RM que cruza as variáveis Recenticidade e Valor Monetário, permitindo definir quatro estratégias diferentes de comunicação com o cliente.

Para obter a classificação RM de cada cliente, ordena-se os mesmos pela data mais recente classificando os primeiros n/5 clientes com o valor R=1, os segundos n/5 com o valor R=2 e assim sucessivamente até ao número 5. Procede-se da mesma forma para o atributo valor monetário. O resultado é uma matriz RM 2x2 com quatro quadrantes, ilustrada na Figura 3-2.

Para os grupos de clientes em cada quadrante são propostas estratégias diferentes: para os Clientes do Grupo R+M+, que compram mais e com mais frequência, é proposta uma estratégia de premiação da fidelização com caracter permanente; nos Clientes R+M- propõe-se a estimulação da compra, de forma a migrá-los para o quadrante R+M+; para os clientes R-M+ devem ser implementadas estratégias de retenção, também com o intuito de os migrar para o quadrante R+M+; para os clientes R-M- não se propõe qualquer estratégia, dado que são pouco rentáveis e podem inclusivamente ser prejudiciais à empresa.

Figura 3-3 - Segmentação RM

	+ Recentes	–
+ V. Monetário	Premiar	Reter
–	Estimular	Esquecer

Recorrendo a modelos preditivos Siegel (Siegel, 2013) indica o modelo *lift* como uma forma de segmentação, uma vez que o modelo preditivo identifica os clientes mais suscetíveis a uma determinada comunicação ou ação de marketing. No caso, o modelo preditivo pode identificar quais os clientes que se encontram num determinado patamar de pré-desistência e depois fazer-se incidir as ações de fidelização sobre estes.

Por outro lado, ainda Siegel (Siegel, 2013) refere outra forma de segmentar para obter os grupos-alvo a contatar através do modelo *uplift*[1], modelo que prevê a influência sobre o

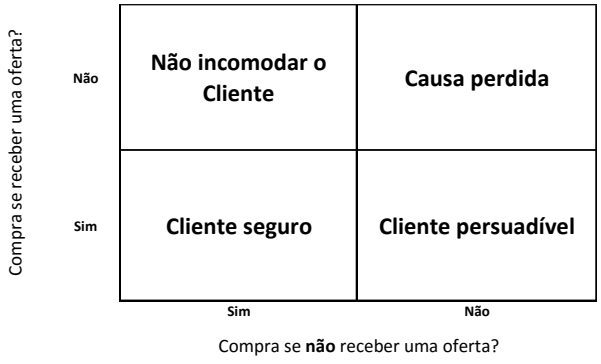
[1] O modelo *Uplift* é também conhecido por *diferencial response*, *impact*, *incremental impact*, *incremental response*, *net lift*, *net response*, *persuasion, true lift*, ou *true response modeling.*

comportamento de um individuo e que resulta da aplicação de uma forma de tratamento sobre outra – por exemplo, de clientes contatados sobre clientes não contatados. Não se pretende obter os clientes com mais probabilidade de comprar ou com menos probabilidade de comprar, mas sim os clientes que são mais influenciáveis pelos contatos de marketing que possam ser efetuados.

Para aprender a distinguir os clientes influenciáveis – aqueles a quem faz diferença efetuar algum tratamento – o modelo *Uplift* aprende de ambos os tipos de clientes, os que foram contatados e os que não foram, pelo que é necessário utilizar dois conjuntos de dados para treino do modelo, constituindo-se um grupo de clientes que "é tratado" – grupo de tratamento - e um outro grupo de clientes que não o é – grupo de controlo.

Como resultado, cada cliente é segmentado num dos quatro quadrantes indicados na Fig. 3-4, propondo-se a supressão dos contactos com os clientes nos quadrantes dos "clientes certos" (*Sure things*), das "causas perdidas" (*Lost causes*) e dos clientes a "não incomodar (*Do-not-disturb*).

Figura 3-4 - Modelo *Uplift*



O método *Uplift* utiliza também técnicas de DM para efetuar a segmentação através de árvores *uplift* que, da mesma forma que as árvores de decisão, utilizam atributos para automaticamente identificarem sub-grupos, mas de forma distinta destas, tentam identificar segmentos extremos pela diferença de tratamentos, identificando segmentos que são particularmente influenciáveis.

Os métodos de segmentação abordados – RM, *Lift* e *Uplift* – providenciam uma oportunidade de redução de custos evitando ações desnecessárias no intuito de aumentar a retenção. O método RM oferece a possibilidade de segmentar os clientes numa forma

simples e prévia à aplicação de técnicas de DM, combinando dois atributos fáceis de determinar para cada cliente. Dispondo de um modelo preditivo, o modelo *Lift* é simples de aplicar. Por outro lado, o modelo *Uplift* é mais complexo e combina os paradigmas da modelação preditiva com a comparação de resultados entre grupos de tratamento e grupos de controlo.

A maximização dos benefícios de utilização dos modelos preditivos de retenção é obtida através da integração com modelos que permitam a realização de ações de marketing personalizadas. O desafio que se coloca é o de como associar as ações corretas com os clientes corretos. Gorgoglione & Panniello (Gorgoglione, 2011) identificam cinco possíveis abordagens para a criação de ações personalizadas:

- A abordagem Computacional permite automatizar completamente a geração de ações baseando-se nos perfis de clientes, não sendo necessária nenhuma intervenção humana. Uma vez que esta abordagem faz uso de comportamentos, ações e reações de clientes registadas na base de dados, só pode ser usada após terem ocorrido previamente outras ações que permitam a construção do modelo.
- A abordagem baseada em Similaridades são as usadas pelos sistemas de recomendação e pelos métodos de personalização de conteúdos web. Este tipo de abordagem assume que as ações estão relacionadas com as preferências dos clientes, podem ser inferidas através dos perfis de clientes, e parte do princípio que clientes similares se comportam de forma similar e ações similares causam reações similares.
- A abordagem *Bottom-Up* inclui os métodos de *knowledge discovery* e implementa-se em dois passos separados: 1) criação dos perfis de clientes e 2) decidir que ações são adequadas. Neste caso, apenas o primeiro passo é efetuado por um algoritmo, havendo intervenção humana no segundo.
- A abordagem *Top-Down* consiste nos mesmos dois passos da abordagem *Bottom-Up*. Contudo a decisão sobre que ações se pretendem implementar são tomadas antes de definir os perfis de clientes.
- A abordagem Personalizada oferece aos clientes diversas opções diferentes, ficando ao seu critério escolher a que preferem.

Kotler (Kotler & Keller, 2009) refere que é cada vez mais difícil selecionar um meio eficaz para levar as mensagens aos clientes dada a fragmentação e desordem dos atuais canais de comunicação, e classifica-os em dois grandes grupos: canais pessoais e não pessoais. Os canais pessoais permitem que duas ou mais pessoas comuniquem pessoalmente ou através de correio, telefone ou e-mail e a sua efetividade deriva da sua capacidade de personalização e de permitirem obter feedback sobre as ações efetuadas. Os canais não pessoais saem fora do âmbito deste trabalho, pelo que não iremos abordá-los.

Merisavo (Merisavo & Raulas, 2004), num estudo efetuado sobre dados recolhidos de 890 consumidores de uma multinacional de cosméticos, observou que o envio regular de e-mails de contacto tem efeitos positivos na lealdade, levando os clientes a criar relações fortes com a marca, a efetuar compras regularmente, a recomendarem a marca aos seus amigos e a estimular a visita de lojas.

Os consumidores que receberam os e-mails consideraram interessantes, e pela ordem indicada, os seguintes conteúdos: ofertas de vendas, informação sobre novos produtos, concursos, notícias genéricas acerca de beleza, informação sobre eventos, *links* para páginas relevantes e informação sobre tendências internacionais.

Kotler (Kotler & Keller, 2009) acrescenta que para os e-mails terem os efeitos pretendidos têm de ser relevantes, direcionados a um *target* e enviados em *timing* adequado.

Pousttchi (Pousttchi & Wiedemann, 2006) após análise de 30 casos de estudo, identifica cinco objetivos para o marketing móvel: divulgar a marca, alterar a imagem da marca, melhorar a lealdade à marca, construir bases de dados de clientes e como veículo para o "boca-a-boca". De acordo com as características, identificam quatro tipos característicos nas mensagens enviadas: informação, entretenimento, sorteio e cupões / vouchers. Kotler (Kotler & Keller, 2009) adianta ainda que os *Smartphones* também permitem programas de promoção da retenção em que os clientes podem receber recompensas ao visitarem as lojas.

Vários trabalhos (Kotler & Keller, 2009), (Merisavo & Raulas, 2004), (Pousttchi & Wiedemann, 2006) apresentam os e-mails e a comunicação móvel (essencialmente SMS – *Short Message Service* - e MMS – *Multimedia Messaging Service*) como potenciais canais de comunicação para promover ações cujo objetivo seja o de aumentar a lealdade dos

clientes e consequentemente também as taxas de retenção. Contudo, todos eles referem também a necessidade das comunicações serem integradas num contexto global das ações de marketing comunicacional da organização (IMC – *Integrated Marketing Communications*) e enviadas com base em permissões dadas pelos clientes (*opt-in*), em respeito pela sua privacidade e em conformidade com as leis vigentes, para que não tenham o efeito contrário ao pretendido.

Como refere Kotler (Kotler & Keller, 2009), em *Marketing Research* existem cinco possíveis abordagens de investigação para recolher informação sobre o impacto de ações de marketing: a observacional, através de *focus groups*, através de inquéritos, a comportamental e a experimental.

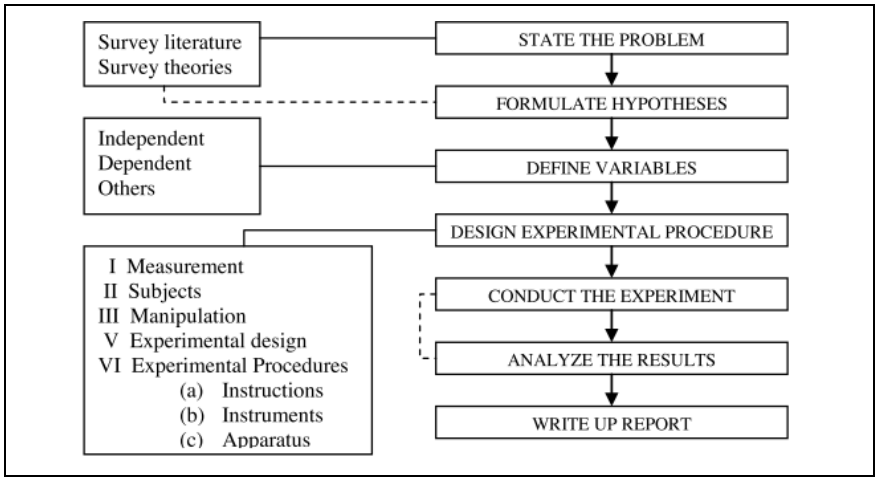
De forma genérica, na abordagem observacional os sujeitos são observados à medida que compram ou consomem os produtos ou serviços; nos *focus groups* são reunidas entre 6 a 10 pessoas que obedecem a determinadas características e que discutem os tópicos de interesse moderados por um profissional que questiona e tenta compreender as suas motivações; os inquéritos são utilizados para questionar e percecionar conhecimento, preferências e satisfação obtendo-se uma escala de magnitude para estes fatores; a abordagem comportamental procura encontrar as preferências dos consumidores através dos dados existentes nas bases de dados, das compras que efetuam e de outras preferências que manifestam; a experimental destina-se a captar as relações causa-efeito através do controlo de fatores que afetam a variável em estudo, procurando a eliminação de influências externas e estranhas. É considerado como o método de investigação mais válido em Marketing Research (Kotler & Keller, 2009) (Zikmund, William; Ward, Steven; Lowe, Ben; Winzar, Hume; Babin, 2011).

As experiências seguem um conjunto de passos relacionados como indicado na Fig. 3-5.

Uma experiência procura assim relacionar uma causa com um efeito e pode ser de dois tipos: natural, em que o investigador apenas mede os efeitos causados; ou controlada, em que ocorrem dois tipo de intervenção: a manipulação de pelo menos um aspeto da experiência (ou variável) e a afetação aleatória de sujeitos a dois grupos, o grupo experimental, o que sofre o efeito da experiência, e o grupo de controlo, o que não sofre

qualquer intervenção, sendo efetuadas medições sobre ambos de forma a se poder comparar os resultados.

Figura 3-5 – Componentes de uma experiência



Fonte: (Smith & Albaum, 2010)

As experiências envolvem três tipos de variáveis: a variável independente e que afeta outra variável na experiência. O efeito da variável independente na outra variável é o que a experiência procura medir; a variável dependente, que é a variável que se altera quando a variável independente varia; e outras variáveis que não são manipuladas mas que podem afetar a experiência.

Smith (Smith & Albaum, 2012) aponta aspetos que têm de ser tidos em conta na preparação de uma experiência e devidamente validados. Como prioritários indica as validações internas, que permitem verificar se na realidade existe causalidade entre as variáveis ou se a causalidade se deve a fatores estranhos, e as externas, que verificam a possível aplicação a outras situações e generalizações.

A nível da validade interna os principais fatores a ter em conta são:

- A duração (*"history"*), porque quanto mais tempo durar a experiência, mais provável será que fatores externos à experiência a influenciem;
- A maturação, dado que os participantes evoluem e a resposta dada pode alterar-se ao longo da experiência;
- Os testes, uma vez que a realização dos mesmos pode intervir nas respostas de testes subsequentes;

- A instrumentação, caso existam alterações nos instrumentos de medida ou nos processos no decurso da experiência;
- A forma como a seleção dos participantes é feita no que concerne à sua colocação nos grupos experimental e de controlo: sendo aleatória o efeito de seleção será uma medida de variação aleatória; não sendo, os grupos podem diferir em características importantes que podem influenciar a variável dependente;
- A mortalidade, uma vez que os participantes nos grupos podem abandoná-los no decurso da experiência, o que pode criar variações estranhas à experiência entre cada medida;
- A regressão estatística, uma vez que há a tendência das pontuações altas se tornarem mais baixas nas medidas seguintes e vice-versa, o que leva a que as pontuações sucessivas se aproximem das médias.

Ao nível da validação externa são apontados:

- Os efeitos reativos aos testes, uma vez que, por exemplo, a aplicação de pré-testes pode alterar o comportamento de sujeitos que não sabem que estão a ser alvo de uma experimentação;
- Os efeitos reativos à própria situação experimental que pode fazer com os sujeitos reajam à mesma;
- A interação entre o apuramento dos resultados da experiência e a sua duração, dado que a medida extemporânea dos resultados da experiência pode invalidar os resultados da mesma;
- A interação entre o tratamento e a seleção, dado que o método de seleção pode condicionar a generalização dos resultados;

Consideram-se normalmente dois grupos de modelos de experiências: o clássico e o estatístico. Os modelos clássicos focam-se no impacto de uma variável dependente de cada vez. Os modelos estatísticos examinam o impacto de duas ou mais variáveis independentes. Em geral, este tipo de desenho lida com formatos que afetam objetos de teste a diferentes níveis de tratamento e as medidas são apenas obtidas após os tratamentos serem efetuados. Não estando no âmbito deste trabalho, mantemos a discussão apenas a nível dos modelos clássicos.

Existem vários tipos de modelos clássicos: os pré-experimentais não têm qualquer controlo e têm pouca utilidade para estabelecer causalidade; os ditos *quasi-experimental* têm controlo mas não afetam de forma aleatória os sujeitos aos grupos.

A Fig. 3-6 apresenta as classes dos vários modelos, a forma como a experiência é montada em termos da construção dos grupos e sequência de eventos, e as validações que ficam resolvidas em cada classe pela forma como a experiência é montada.

Figura 3-6 – Validações por tipo de experiência

	Internal								*External*		
	History	Maturation	Testing	Instrumentation	Regression	Selection	Mortality	Interaction of Selection and Others	Interaction of Testing and *X*	Interaction of Selection and *X*	Reactive Arrangements
One-Shot Case Study *X O*	-	-				-	-			-	
One-Group Pretest-Posttest Design *O X O*	-	-	-	-	?	+	+	-	-	-	?
Time Series *O O O X O O O*	-	+	+	?	+	+	+	+	-	?	?
Multiple Time-Series *O O O X O O O* *O O O O O O*	+	+	+	+	+	+	+	+	-	-	?
Static-Group Comparison *X O* *O*	+	?	+	+	+	-	-	-		-	
Nonequivalent Control Group Design *O X O* *O O*	+	+	+	+	?	+	+	-	-	?	?
Posttest-Only control Group Design *R X O* *R O*	+	+	+	+	+	+	+	+	+	?	?
Separate-Sample Pretest-Posttest Design *R O (X)* *R X O*	-	-	+	?	+	+	-	-	+	+	+
Pretest-Posttest Control Group Design *R O X O* *R O O*	+	+	+	+	+	+	+	+	-	?	?
Solomon Four-Group Design *R O X O* *R O O* *R X O* *R O*	+	+	+	+	+	+	+	+	+	?	?

Note: In the tables, a minus (-) indicates that the factor is not controlled, a plus (+) indicates that the factor is controlled, a question mark (?) indicates a possible source of concern, and a blank indicates that the factor is not relevant.

Fonte: (Smith & Albaum, 2010)

A classificação dos modelos experimentais clássicos apresentados na Fig. 3-6 utiliza a seguinte notação:

- Um *X* para representar a exposição de grupos de teste a uma experiência cujo efeito se pretende observar e/ou medir;
- Um *O* para representar uma observação ou uma medida;
- Um *R* para indicar que os sujeitos foram aleatoriamente designados a grupos para diferentes tratamentos;

A leitura da esquerda para a direita indica a sequência de eventos. Quando na mesma linha, estamos perante eventos que ocorreram sobre o mesmo grupo (experimental ou de controlo); a leitura de cima para baixo indica a ocorrência de eventos em simultâneo sobre diferentes grupos;

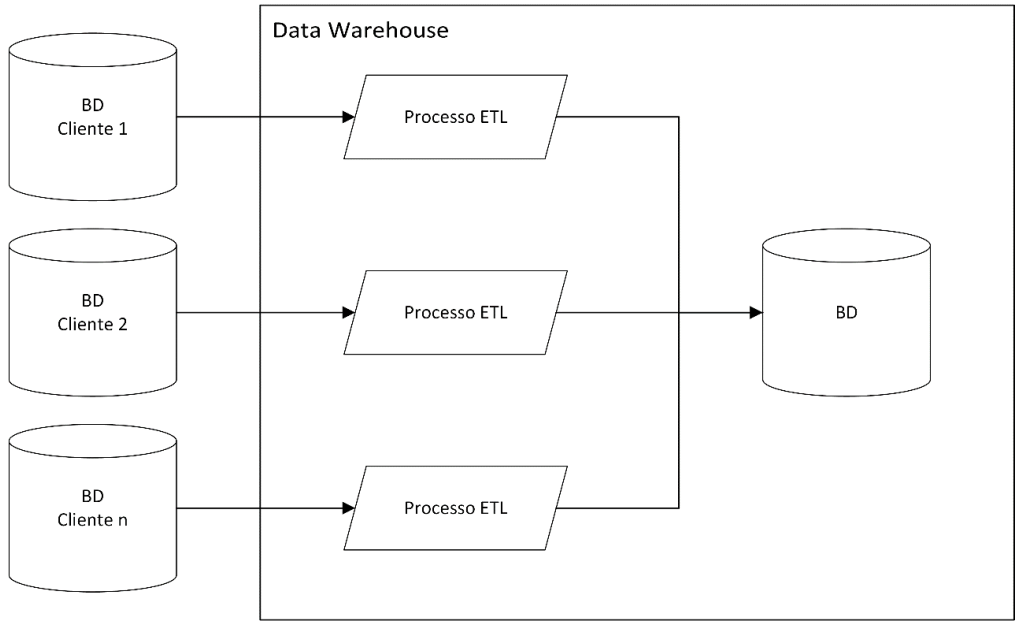
4. Acerca dos dados: DW e ETL

4.1. A origem dos dados e os atributos a considerar

No trabalho que nos propusemos desenvolver, e de acordo com o modelo proposto em *Database Marketing* (Cavique, 2006), o primeiro passo - identificar – requere um conjunto de dados de origem - atributos e registos - a partir dos quais se possa popular uma tabela de factos no DW onde se irão concentrar os atributos relevantes que permita partir para a segmentação (diferenciação) através dos algoritmos de DM. Neste sentido, pode haver a necessidade destes dados serem obtidos em diferentes conjuntos pelo que devem ser integrados de forma a constituir uma única tabela (Gama et al., 2017).

A obtenção dos dados, o seu tratamento e síntese devem causar o mínimo impacto nas bases de dados de origem – às quais se poderá ter acesso apenas de leitura (Vassiliadis & Simitsis, 2009) – pelo que se projeta um DW com a arquitetura apresentada na Fig. 4-1.

Figura 4-1 - Implementação do Data Warehouse



Com base no trabalho relacionado apresentado na seção Retenção em Serviços Desportivos, ações de fidelização e experimentação - (Howat & Assaker, 2016) (Avourdiadou & Theodorakis, 2014) (Surujlal & Dhurup, 2012) (Marisa & Pedragosa, 2006) (Gonçalves, 2012) (Frota, 2011) - há que ter em atenção a existência de fatores com maior ou menor impacto na retenção das instalações desportivas (ver Tab. 3-2) pelo que o DW deve conter um conjunto de diversos atributos relevantes que sejam passíveis de serem extraídos dos dados registados nas bases de dados dos sistemas ERP utilizados pelas

instalações desportivas. Pelas suas características, esses atributos podem ser agrupados em:

- Atributos demográficos, como a idade e o género;
- Atributos relacionados com a contratação do serviço como a frequência contratada, o número de meses da inscrição e o volume de negócios;
- Atributos relacionados com a frequência nomeadamente frequência real, frequência média e número de dias sem visitar as instalações;
- Outros atributos relacionados com a qualidade do serviço como o número de reclamações, outras manifestações de insatisfação, abordagens e contatos pessoais ou por outras formas, avaliações da condição física ou outro tipo de avaliação comunicada regularmente ao utente;

Considerando estas bases de dados e o elevado número de atributos que disponibiliza, procedeu-se à seleção dos atributos que podem corresponder a aspetos considerados relevantes para a retenção mencionados anteriormente. Em resultado, identificaram-se os atributos que foram inumerados na tabela de factos do DW, que designamos por "*Retencao*" e que se apresenta na Tab. 4-1.

Tabela 4-1 - Atributos da tabela "Retencao"

Atributo	Tipo	Atributo	Tipo
Id	Uniqueidentifier	Distancia	Int
Atividade_aquaticas	Bit	Dtultvisita *	Date
Atividade_atletismo	Bit	Duracaotreino	Int
Atividade_coletivos	Bit	Freqcontratada *	Int
Atividade_combate	Bit	Freqcontratadasemanal	Int
Atividade_danca	Bit	Freqmedia	Decimal
Atividade_especiais	Bit	Freqmediaaulas	Decimal
Atividade_fitness	Bit	Freqreal *	Int
Atividade_natureza	Bit	Genero	Bit
Atividade_outra	Bit	Idade	Int
Atividade_raquete	Bit	Inicio *	Date
Classe_idade	Int	Mesesinscricao	Int
Classe_desistencia	Bit	Natividades	Int
Classe_diassemfrequencia	Int	Naulas	Int
Classe_duracaotreino	Int	Ncontatos	Int
Classe_freqmedia	Int	Nfrequencias	Int
Classe_freqmediaaulas	Int	Ninsatisfacao1	Int
Classe_mesesinscricao	Int	Ninsatisfacao2	Int
Classe_naulas	Int	Ninsatisfacao3	Int
Classe_nfrequencias	Int	Ninsatisfacoes	Int
Classe_ratiofreqcontratadareal	String	Nreferencias	Int
Classe_volnegocios	Int	Nrenovacoes	Int
Classe1_diassemfrequencia	String	Ratiofreqcontratadareal	Decimal
Classe1_freqmedia	String	Referencias	Bit
Classe1_mesesinscricao	String	Termino *	Date
Datanasc *	Date	Utilizacao_livre	Bit
Diassemfrequencia	Int	Volnegocios	Decimal
* Atributos auxiliares			

Tendo em atenção a arquitetura apresentada na Fig. 4-1 e necessidade de carregar o DW com uma quantidade adequada de registos com os atributos necessários considerou-se como BD Cliente 1 (BD1) uma base de dados operacional utilizada numa instalação desportiva da região da grande Lisboa que suporta uma aplicação de mercado (e@sport) à qual foram aplicados os processos ETL conforme descritos por Trujillo (Trujillo & Luján-Mora, 2003).

Desta base de dados consideraram-se apenas os utentes que se inscreveram entre 1 de Junho de 2014 e 31 de Dezembro de 2017. A Tab. 4-2 apresenta o número de registos relevantes nesta base de dados a 31 de Dezembro de 2017.

Tabela 4-2 - Propriedades das bases de dados de teste a 31/Dez/2017

	BD1 01-Jun-2014 a 31-Dez-2017
Nº de Utentes admitidos	21 755
Nº de movimentos no Controlo de Acessos	3 344 947
Nº de inscrições em atividades	122 806

4.2. A construção do *Data Warehouse*

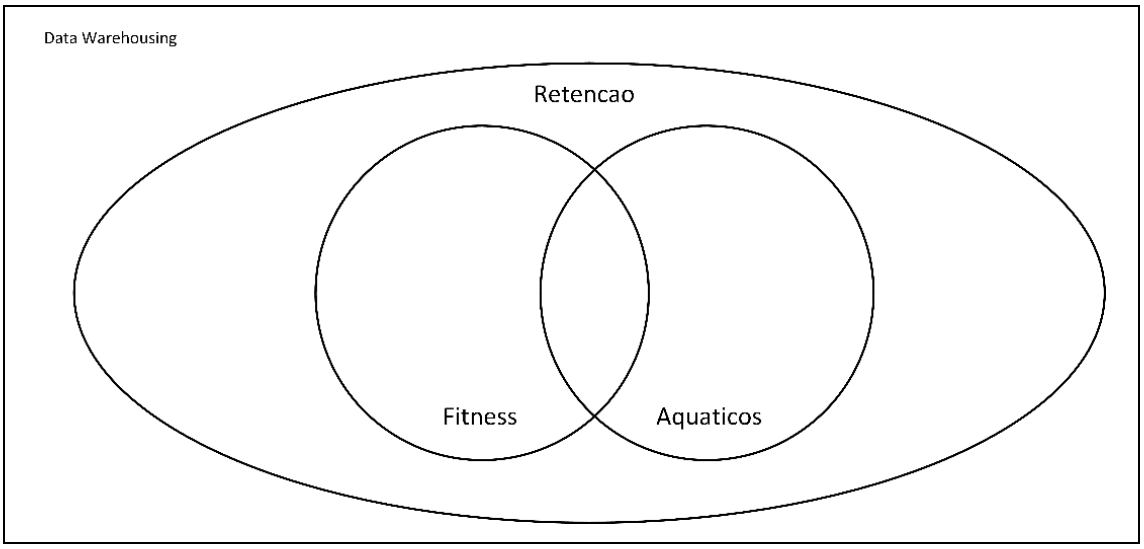
A base de dados do cliente (BD1) está suportada em MS SQL*Server®. O DW foi também implementado em MS SQL*Server e os procedimentos ETL são *Storage Procedures* implementados em *Transact*-SQL, dialeto SQL próprio deste SGBD.

Uma vez que será necessário na fase seguinte observar o comportamento dos modelos preditivos aplicados aos praticantes das diferentes atividades, nomeadamente das atividades aquáticas e de *fitness* que se apresentam com um maior número de utentes, optou-se nesta fase por criar duas tabelas adicionais no DW, com estrutura idêntica à tabela de factos que concentra todos os utentes ("*Retencao*"), uma com o nome "*Aquaticos*" que contém os utentes que frequentaram pelo menos uma vez atividades aquáticas durante a sua inscrição, e outra com o nome "*Fitness*" que contém os utentes que frequentaram atividades de *fitness* pelo menos uma vez durante o período da sua inscrição. Na prática, as tabelas "*Fitness*" e "*Aquaticos*" contêm subconjuntos dos utentes presentes na tabela "*Retencao*" como ilustrado no diagrama de Venn da Fig. 4-2. Apesar da redundância das situações referidas, evitam-se complicações derivadas de filtragem na fase seguinte.

Tendo em vista que a construção do DW se destina à aplicação de técnicas de DM e dado que, como refere Gama (Gama et al., 2017), o desempenho de algumas técnicas estão

limitadas à manipulação de valores de determinado tipo ou o próprio desempenho é influenciado pelo intervalo de variação de valores, alguns atributos foram transformados e foram criados atributos adicionais que derivam de classificações e transformações efetuadas nos dados ou nos atributos originais.

Figura 4-2 - Representação dos conjuntos de utentes presentes nas tabelas do DW



A transformação dos dados envolve limpeza, correção ou remoção de dados inconsistentes, deteção de ruído e *outliers*, bem como a realização de conversões simbólico-numéricas, conversões numérico-simbólicas e transformação de atributos numéricos como explicado anteriormente. Uma vez que a transformação dos dados ocorre na importação ou diretamente no DW a fase de carregamento de dados ocorre em simultâneo com as fases referidas.

O processo ETL apresentado na Fig. 4-1, e implementado no protótipo, é executado em diversos passos executados de forma sequencial. Os dados são importados da base de dados do cliente para a base de dados do DW, realizando-se ao mesmo tempo a escolha dos atributos relevantes, reduzindo o conjunto de atributos que irá posteriormente servir de base à construção dos modelos preditivos.

Desta forma, os atributos relevantes indicados nas tabelas referidas são preenchidos de quatro formas diferentes: a) por importação direta do valor da base de dados origem para o DW; b) por transformação do valor na base de dados origem para um novo valor no DW;

c) por *queries* ou cálculos efetuados na base de dados origem; d) por cálculo direto do valor no DW a partir de atributos obtidos por uma das outras formas;

a) Atributos obtidos por importação direta

O atributo "*Id*" corresponde a uma pseudonimização do Utente a que dizem respeito os dados contidos nos restantes atributos de cada registo;

O atributo "*datanasc*", do tipo *date*, é um atributo de importação direta destinado a suportar o cálculo da idade do utente;

b) Atributos obtidos por transformação do valor na base de dados origem para um novo valor no DW

O atributo "*idade*", do tipo *int*, não existe na base de dados origem pelo que é calculado com base na data de nascimento. A idade é calculada à data em que o utente se tornou desistente ou no último dia do mês anterior aquele em que se efetua a integração dos dados no DW;

O atributo "*genero*", do tipo *bit*, corresponde a uma transformação simbólica-numérica a partir da base de dados origem: o valor 0 (zero) corresponde ao valor F na origem; o valor 1 (um) corresponde ao valor M na origem;

c) Atributos obtidos por *queries* ou cálculos efetuados na base de dados origem

O atributo "*dtulvisita*", do tipo *date*, corresponde à data em que o utente entrou nas instalações pela última vez;

Os atributos "*inicio*" e "*termino*", do tipo *date*, são auxiliares (destinam-se a suportar cálculos relacionados com a duração da inscrição e com as frequências médias) e correspondem à data de início e término da relação do utente com a instalação desportiva;

O atributo "*volnegocios*", do tipo *decimal*, corresponde ao somatório de todas as faturas emitidas para a utente ao longo da sua inscrição;

O atributo "*utilizacao_livre*", do tipo *bit*, é assinalado com 1 caso o utente esteja ou tenha estado inscrito em utilização livre[2], ou com 0 caso contrário;

Os atributos "*atividade_atletismo*", "*atividade_aquaticas*", "*atividade_fitness*", "*atividade_danca*", "*atividade_coletivos*", "*atividade_raquete*", "*atividade_combate*",

[2] Consultar a seção **2. Serviços desportivos regulares contínuos e em épocas** para a definição do serviço de utilização livre

"*atividade_especiais*", "*atividade_natureza*", "*atividade_outra*" são do tipo *bit* e identificam a(s) atividade(s) que o utente frequenta (ou frequentou) – caso em que o atributo tem o valor 1 – ou não frequentou – caso em que o atributo tem o valor 0. Dada a disparidade de termos utilizados para nomear as atividades em cada instalação desportiva, e uma vez que se pretende uma generalização que permita a adaptação a todas as instalações desportivas verificou-se a necessidade de criar uma tabela de sinónimos (atividades) para "traduzir" os termos próprios de cada instalação nos termos gerais considerados neste trabalho.

O atributo "*nfrequencias*", do tipo *int*, corresponde ao número total de visitas às instalações durante o período de vigência da inscrição;

O atributo "*naulas*", do tipo *int*, corresponde ao número de aulas que o utente frequentou durante o período da sua inscrição. Este contador distingue-se do anterior dado que corresponde apenas a frequência de aulas, e não a visitas em modo de utilização livre;

O atributo "*nrenovacoes*", do tipo *int*, conta o número de renovações entre épocas efetuadas pelo utente;

O atributo "*nreferencias*", do tipo *int*, conta o número de familiares ou qualquer outro utente que esteja associado ao utente. Na prática, pode ser considerado como o grau de centralidade do utente;

O atributo "*classe_desistencia"* é do tipo *bit*, em que o valor 1 corresponde a uma desistência, e o valor 0 (zero) corresponde a um utente ativo.

Dado que o comportamento do utente no período que precede a sua desistência pode ser diferente do que era o seu comportamento anterior, consideraram-se alguns atributos calculados com base nos últimos dois meses da inscrição. No caso dos desistentes, o valor determinado refere-se aos dois meses anteriores à desistência; no caso dos ativos, os valores referem-se aos últimos dois meses anteriores que terminam no último dia do mês anterior à data da integração. São assim calculados o atributo "*freqcontratadasemanal*", do tipo *int*, que apresenta o número de visitas semanais que o utente pode efetuar à instalação de acordo com o que lhe permite o serviço que contratou; e os atributos auxiliares "*freqcontratada*" e "*freqreal*" que dizem respeito, respetivamente, ao número de visitas que o utente poderia ter efetuado nos últimos dois meses da sua inscrição, e ao

número de visitas que efetivamente realizou. Estes dois últimos atributos são utilizados para cálculo direto no DW de dois novos atributos ("*ratiofreqcontratadareal*" e "*classe_ ratiofreqcontratadareal*");

d) Atributos obtidos por cálculo direto do valor no DW a partir de atributos obtidos por uma das outras formas

O atributo "*diassemfrequencia*" é do tipo *int* e contém o número de dias, contados desde a última visita (atributo "*dtultvisita*") até à data da desistência, se é um desistente, ou até ao último dia do mês anterior aquele em que se efetua a integração dos dados no DW, caso ainda seja um utente ativo;

O atributo "*freqmedia*", do tipo *decimal*, indica a frequência média do utente em todo o período da sua inscrição. O cálculo é efetuado pela divisão do número de frequências total (atributo "*nfrequencias*") pelo número de semanas da inscrição, excetuando-se o período de encerramento entre épocas;

O atributo "*freqmediaaulas*", do tipo *decimal*, é calculado da mesma forma que o atributo "*freqmedia*", embora seja utilizado o atributo "*naulas*" em vez do atributo "*nfrequencias*";

O atributo "*natividades*", do tipo *int*, resulta da contagem dos bits com valor 1 referentes aos campos "*atividade_atletismo*", "*atividade_aquaticas*", "*atividade_fitness*", "*atividade_danca*", "*atividade_coletivos*", "*atividade_raquete*", "*atividade_combate*", "*atividade_especiais*", "*atividade_natureza*", "*atividade_outra*";

O atributo "*mesesinscricao*", do tipo *int*, parte das datas de início e término da inscrição para calcular o tempo de vida, em meses, da relação entre o utente e a instalação desportiva. Se a relação se mantém, o número de meses da inscrição é calculado à data do último dia do mês anterior aquele em que se está a efetuar a integração de dados no DW;

O atributo "*ratiofreqcontratadareal*", do tipo decimal, apresenta a relação entre o número de visitas que o utente pode realizar nos últimos dois meses, e o número de visitas que efetivamente realizou. O cálculo é obtido pela divisão entre o valor do atributo "*freqreal*" e o atributo "*freqcontratada*";

O atributo "*classe_ratiofreqcontratadareal"* é um atributo nominal e corresponde a uma classificação do valor do atributo anterior. Os valores possíveis são "Nunca", "<=25", "<=50", "<=75" e ">75";

O atributo "*referencias*", do tipo *bit*, contém o valor 0 (zero) caso o atributo "*nreferencias*" tenha o valor 0 (zero), ou o valor 1 (um), caso o atributo "*nreferencias*" apresente um valor maior ou igual a 1;

Alguns atributos calculados diretamente no DW apresentam-se ainda numa forma especial dado que se referem a algum tipo de classificação que permite agrupar valores em classes de forma a reduzir a amplitude e/ou a grande variedade dos valores que apresentam. Nestes casos optou-se por efetuar uma conversão numérico-simbólica em duas formas possíveis: a) classificação de acordo com intervalos conhecidos e apresentados como relevantes em trabalhos relacionados com fidelização e retenção de utentes em serviços desportivos; b) a classificação de Hughes, já efetuada em trabalhos anteriores (Pinheiro & Cavique, 2015), onde se ordenam os atributos individualmente por ordem crescente e agrupam-se os registos em categorias de 1 a 5 (o primeiro 1/5 dos registos é atribuída a categoria 1, aos seguintes 1/5 registos a categoria 2 e assim sucessivamente) em que cada categoria fica com o mesmo número de registos (+/- 1 dependendo do resto da divisão do número de registos por 5).

No caso das classificações efetuadas de acordo com intervalos conhecidos (a) temos os atributos:

- "*classe_idade*" em que o utente é colocado numa das seguintes classes "<20", "<35", "<49", "<65" e "≥65" de acordo com a sua idade (atributo "idade") (Gonçalves, 2012);
- "*classe1_diassemfrequencia*" tendo os utentes sido classificados de acordo com as classes nominais "[00-07]", "]07-15]", "]15-30]", "]30-60]" e "]60-inf[" correspondendo respetivamente a 0 a 7 dias sem frequentar, de 8 a 15 dias sem frequentar, de 16 a 30 dias sem frequentar, de 31 a 60 dias sem frequentar, e mais de 60 dias sem frequentar;
- "classe1_freqmedia" com as classes nominais "<=0.1", "<=0.2", "<=0.5", "<=1", "<=2", "<=3" e ">3", correspondendo a uma frequência média igual ou inferior a 0.1 dias por semana, a uma frequência média igual ou inferior a 0.2 dias por semana, a uma frequência média igual ou inferior a 0.5 dias por semana, a uma frequência média igual ou inferior a 1 dia por semana, a uma frequência média igual ou inferior a 2 dias por semana, a uma frequência média igual ou inferior a 3 dias por semana e a uma frequência média superior a 3 dias por semana, respetivamente;

- “*classe1_mesesinscricao*” em que o utente é colocado numa das seguintes classes “[00-01]”, “]01-02]”, “]02-04]”, “]04-06]”,”]06-09]”, “]09-12]” e “]12-inf[” em função do número de meses que tem (ou teve) a sua inscrição;

Os atributos obtidos com a classificação de Hughes (b) são “*classe_diassemfrequencia”*, “*classe_mesesinscricao”*, “*classe_volnegocios”*, “*classe_freqmedia”*, “*classe_naulas”*, “*classe_nfrequencias”* e “*classe_freqmediaaulas”*, correspondendo cada um ao atributo com o mesmo nome, sem o prefixo “*classe_*”.

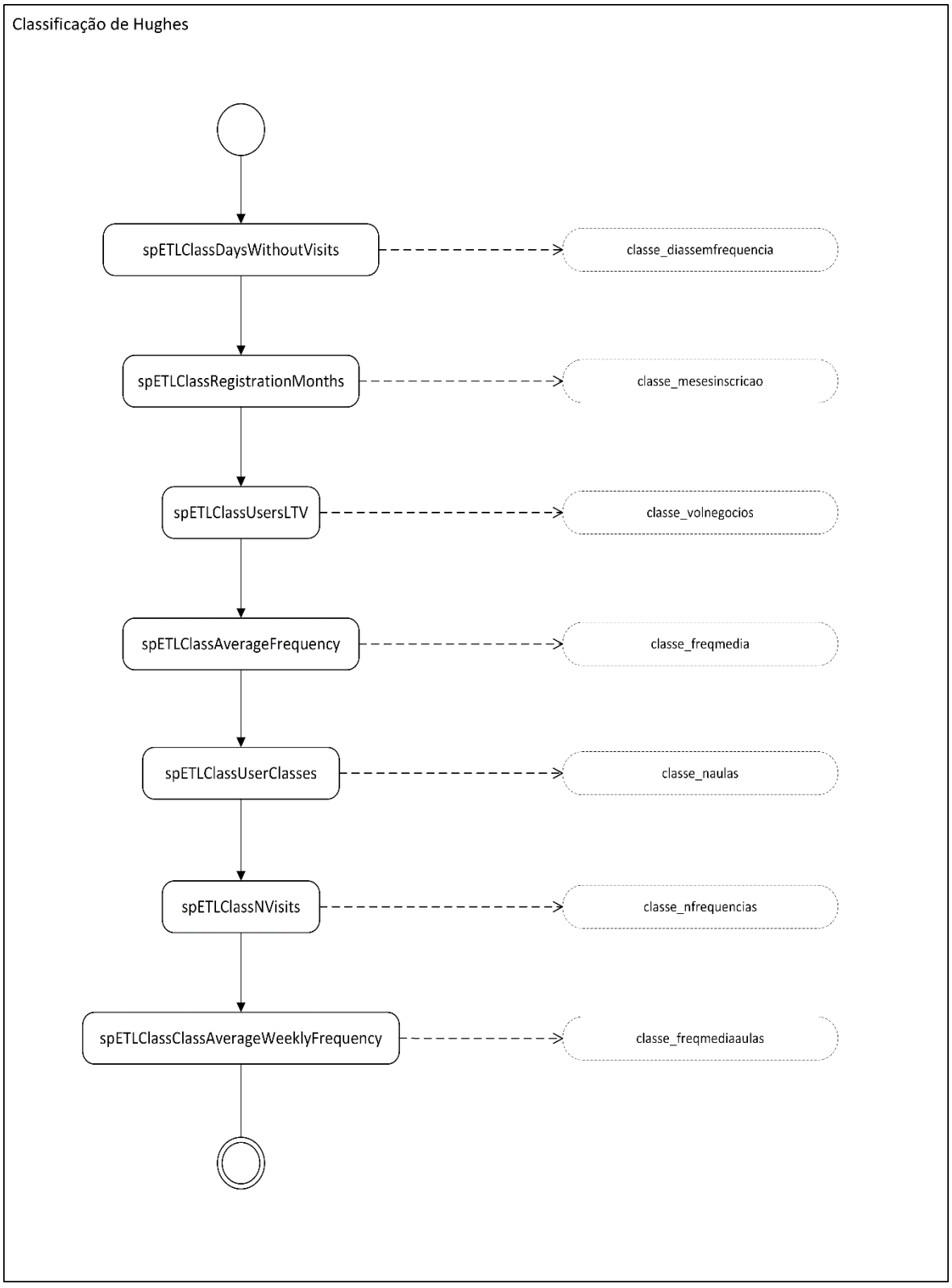
A observação dos dados existentes nas bases de dados origem permitiu concluir pela inexistência de dados relativos a alguns dos atributos que poderiam ser relevantes, nomeadamente os atributos relacionados com a qualidade dos serviços (“*ncontatos*”, “*ninsatisfacoes*”). Por outro lado, optou-se por não se considerar outros atributos pela inexatidão que poderia resultar de uma tentativa de cálculo dos mesmos nomeadamente nos casos dos atributos “*distancia*” ou “*duracaotreino*”.

No primeiro caso (“*distancia*”), seria necessário que a morada indicada na base de dados correspondesse à morada de onde o utente se desloca para ir à instalação desportiva, o que grande parte das vezes não corresponde a realidade porque a morada na base de dados é a de casa e em grande parte dos casos o utente desloca-se para a instalação desportiva partindo do local de trabalho; noutros casos utiliza o transporte público e noutros o carro. Para que este atributo se tornasse útil seria de considerar um atributo na origem (a nível da base de dados do cliente) com a indicação do tempo necessário para a deslocação, uniformizando assim a unidade de medida da deslocação.

No caso da duração do treino e respetiva classificação (“*classe_duracaotreino*”), em muitas instalações desportivas, o controlo de acessos só é feito à entrada o que torna impossível calcular com segurança o tempo que o utente permaneceu nas instalações.

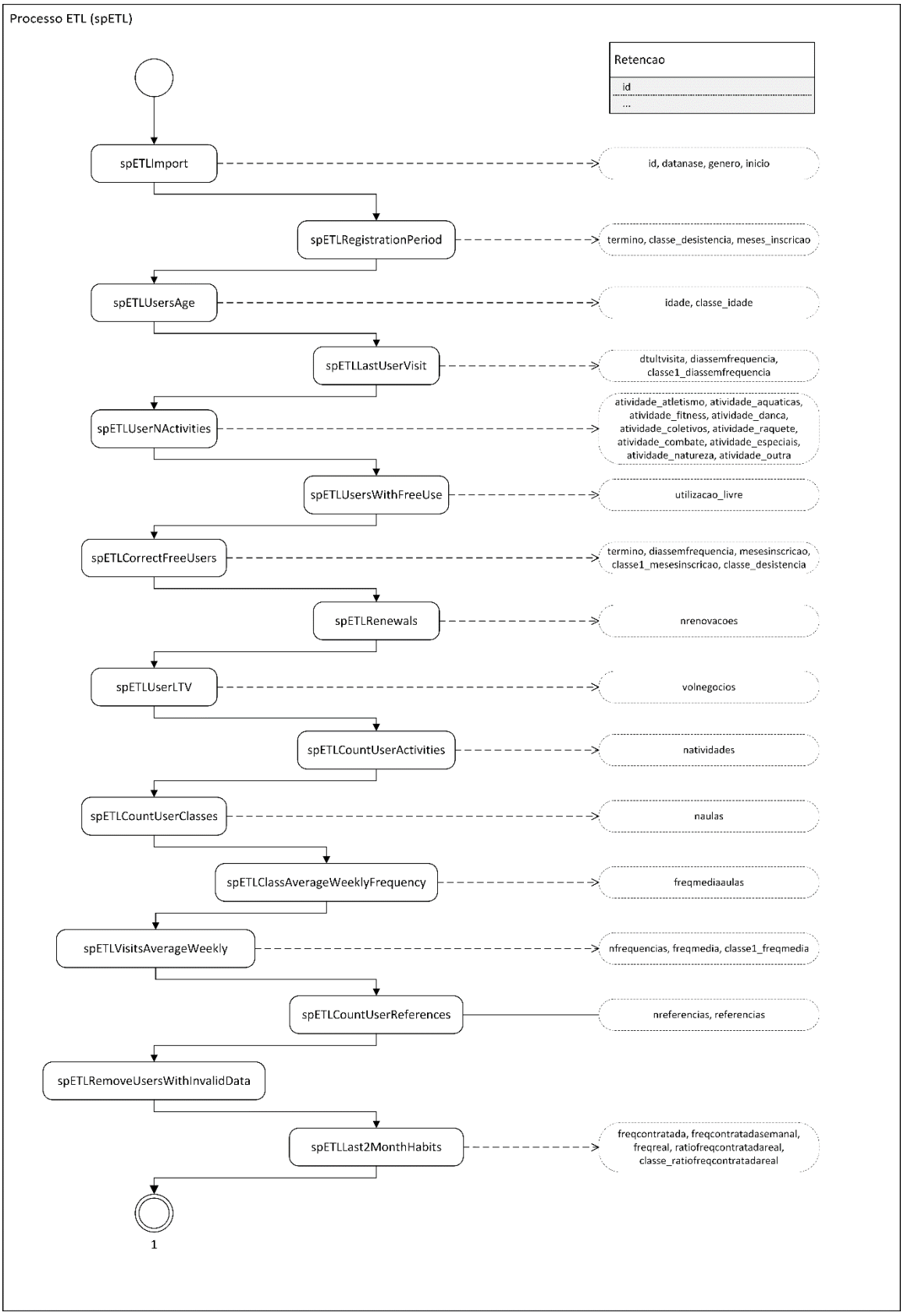
A Fig. 4-3 apresenta os diagramas de atividades do processo ETL geral e dos procedimentos de classificação de Hughes.

Figura 4-3 – Diagrama de atividades do processo ETL



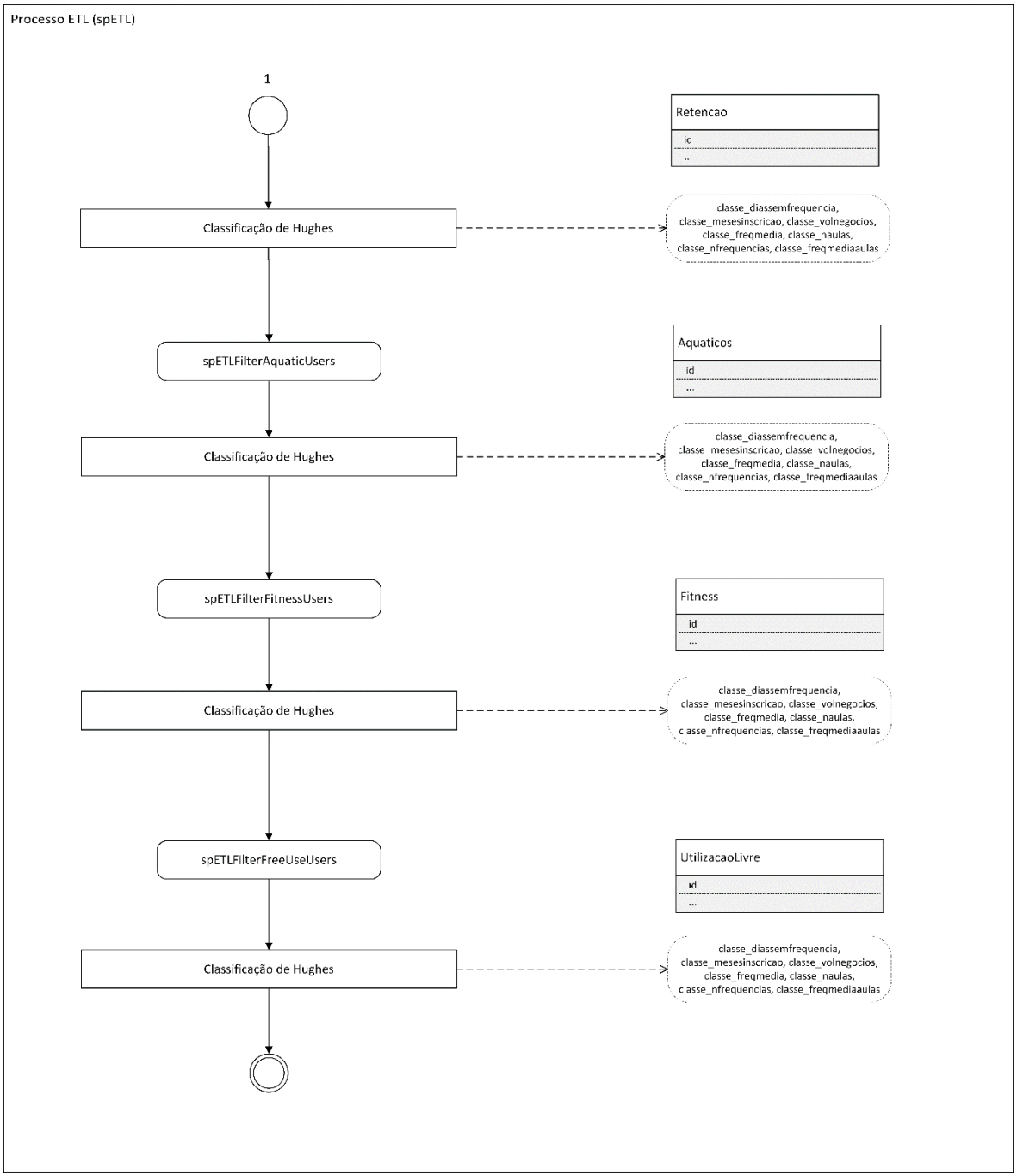
a) Subrotina para determinação da classificação de Hughes

Figura 4-3 – Diagrama de atividades do processo ETL



b) Parte I do Processo ETL

Figura 4-3 – Diagrama de atividades do processo ETL



c) Parte II do Processo ETL

4.3. Estatística descritiva dos dados do DW

Nas Tab. 4-3 e Tab. 4-4 apresenta-se, respetivamente, o sumário de valores estatisticamente significativos de cada atributo, e o número de registos, valor mínimo e valor máximo das classes obtidas para cada atributo que utiliza a classificação de Hughes após a execução do processo ETL sobre a BD1 em Outubro de 2017.

Tabela 4-3 – Sumário de valores estatisticamente significativos por atributo

a) Atributos com duas classes (do tipo Bit)							
Classe	Nº Utentes	Classe	Nº Utentes	Classe	Nº Utentes	Classe	Nº Utentes
Atividade_aquaticas		Atividade_atletismo		Atividade_coletivos		Atividade_combate	
Frequentou	2924	Frequentou	13	Frequentou	83	Frequentou	159
Não Frequentou	5451	Não Frequentou	8362	Não Frequentou	8292	Não Frequentou	8216
Atividade_danca		Atividade_especiais		Atividade_fitness		Atividade_natureza	
Frequentou	0	Frequentou	138	Frequentou	5722	Frequentou	0
Não Frequentou	8375	Não Frequentou	8237	Não Frequentou	2653	Não Frequentou	8375
Atividade_outra		Atividade_raquete		Classe_desistencia		Genero	
Frequentou	2	Frequentou	47	Desistente	6445	Feminino	5211
Não Frequentou	8373	Não Frequentou	8328	Não Desistente	1930	Masculino	3164
Referencias		Utilizacao_livre					
Não apresentou	8142	Frequentou	388				
Apresentou	233	Não Frequentou	7987				

b) Atributos não discretizados (tipos Date, Int e Decimal)							
Datanasc		Dtultvisita		Diassemfrequencia		Freqcontratada	
Mínimo	1929-12-16	Mínimo	2014-07-03	Mínimo	0.00	Mínimo	0.14
1º Quartil	1985-12-22	1º Quartil	2015-11-30	1º Quartil	6.00	1º Quartil	28.98
Mediana	1993-10-19	Mediana	2016-11-04	Mediana	38.00	Mediana	30.03
Média	1991-01-29	Média	2016-09-17	Média	69.94	Média	39.27
3º Quartil	1996-12-14	3º Quartil	2017-09-29	3º Quartil	88.00	3º Quartil	60.97
Máximo	2017-04-17	Máximo	2017-10-31	Máximo	1084.00	Máximo	62.02
Freqcontratadasemanal		Freqmedia		Freqmediaaulas		Freqreal	
Mínimo	1.000	Mínimo	0.0100	Mínimo	0.0000	Mínimo	0.000
1º Quartil	7.000	1º Quartil	0.3400	1º Quartil	0.0000	1º Quartil	2.000
Mediana	7.000	Mediana	0.6500	Mediana	0.0000	Mediana	4.000
Média	5.853	Média	0.8512	Média	0.1381	Média	6.182
3º Quartil	7.000	3º Quartil	1.1300	3º Quartil	0.0500	3º Quartil	8.000
Máximo	7.000	Máximo	10.3300	Máximo	4.6700	Máximo	89.000
Idade		Inicio		Mesesinscricao		Natividades	
Mínimo	0.00	Mínimo	2014-06-02	Mínimo	0.00	Mínimo	1.000
1º Quartil	20.00	1º Quartil	2015-03-02	1º Quartil	4.00	1º Quartil	1.000
Mediana	23.00	Mediana	2015-11-05	Mediana	9.00	Mediana	1.000
Média	25.69	Média	2016-01-01	Média	11.24	Média	1.085
3º Quartil	31.00	3º Quartil	2016-10-04	3º Quartil	14.00	3º Quartil	1.000
Máximo	87.00	Máximo	2017-10-30	Máximo	50.00	Máximo	5.000
Naulas		Nfrequencias		Nreferencias		Nrenovacoes	
Mínimo	0.000	Mínimo	1.00	Mínimo	0.0000	Mínimo	0.0000
1º Quartil	0.000	1º Quartil	7.00	1º Quartil	0.0000	1º Quartil	0.0000
Mediana	0.000	Mediana	17.00	Mediana	0.0000	Mediana	1.0000
Média	7.538	Média	34.12	Média	0.0314	Média	0.9848
3º Quartil	2.000	3º Quartil	41.00	3º Quartil	0.0000	3º Quartil	2.0000
Máximo	410.000	Máximo	592.00	Máximo	3.0000	Máximo	5.0000
Ratiofreqcontratadareal		Termino		Volnegocios			
Mínimo	0.00	Mínimo	2014-07-11	Mínimo	0.0		
1º Quartil	3.87	1º Quartil	2016-02-04	1º Quartil	77.6		
Mediana	13.12	Mediana	2016-12-14	Mediana	148.6		
Média	19.00	Média	2016-12-02	Média	255.3		
3º Quartil	27.61	3º Quartil	2017-10-31	3º Quartil	317.3		
Máximo	100.00	Máximo	2018-07-31	Máximo	3 747.2		

c) Atributos discretizados (tipo Int e String)							
ClasseIdade		Classe_ratiofreqcontratadareal		Classe1_diassemfrequencia		Classe1_freqmedia	
<20	1993	Nunca	1168	[00, 07]	2327	<= 0.1	445
<35	4727	<=25	4874	]07, 15]	637	<= 0.2	656
<49	1002	<=50	1677	]15, 30]	674	<= 0.5	2215
<65	485	<=75	457	]30, 60]	1977	<= 1	2534
>=65	168	>75	199	]60, +∞[	2760	<= 2	1901
						<= 3	457
						> 3	167

Tabela 4-3 – Sumário de valores estatisticamente significativos por atributo

Classe1_mesesinscricao	
[00-01]	427
]01-02]	789
]02-04]	1164
]04-06]	1027
]06-09]	975
]09-12]	1438
]12-+∞[	2555

Tabela 4-4 - Nº de Registos, Mínimos e Máximos das classes nos atributos obtidos com a Classificação de Hughes

Classe	Nº Registos	Mínimo	Máximo	Classe	Nº Registos	Mínimo	Máximo
Classe_diassemfrequencia				Classe_mesesinscricao			
1	1675	0	4	1	1675	0	3
2	1675	4	24	2	1675	3	6
3	1675	24	49	3	1675	6	11
4	1675	49	112	4	1675	11	18
5	1675	112	1084	5	1675	18	50
Classe_volnegocios				Classe_freqmedia			
1	1675	0.00	69.60	1	1675	0.01	0.29
2	1675	69.60	113.20	2	1675	0.29	0.51
3	1675	113.20	194.60	3	1675	0.51	0.80
4	1675	194.60	376.50	4	1675	0.80	1.29
5	1675	376.55	3 747.20	5	1675	1.29	10.33
Classe_nfrequencias				Classe_naulas			
1	1675	1	6	1	1675	0	0
2	1675	6	13	2	1675	0	0
3	1675	13	24	3	1675	0	0
4	1675	24	50	4	1675	0	6
5	1675	50	592	5	1675	6	410
Classe_freqmediaaulas							
1	1675	0.00	0.00				
2	1675	0.00	0.00				
3	1675	0.00	0.00				
4	1675	0.00	0.21				
5	1675	0.21	4.67				

A Tab. 4-5 apresenta o número de registos após a execução do processo ETL sobre a base de dados origem nos meses de Outubro, Novembro e Dezembro de 2017.

Tabela 4-5 - Número de registos nas tabelas do DW após execução do processo ETL sobre a BD1

Tabela do DW	Out/2017	Nov/2017	Dez/2017
Retencao	8375	8546	8601
Aquaticos	2924	2988	3006
Fitness	5722	5831	5869
UtilizadoresLivres	388	389	389

Tendo em atenção o número de registos final no DW, o número de registos removidos corresponde a 60,46% dos Utentes em Dezembro de 2017. Os motivos que justificam a remoção deste número de registos referem-se a: a) utentes que apesar de registados, nunca concretizaram nenhuma inscrição ou nunca efetuaram o pagamento da mesma; b)

registos com algum dos atributos inconsistentes ou em falta; c) registos correspondentes a utentes que frequentam atividades que não dispõem de mecanismos de controlo de acessos não sendo portanto possível determinar dados relativos a frequências. Devido à ausência destes dados, optou-se por não considerar estes registos no DW procedendo também à sua eliminação.

De qualquer forma, resulta um número significativo de utentes que podem servir de base à construção dos modelos preditivos pretendidos.

5. Análise preditiva: classificação

5.1. Atributos utilizados nos modelos preditivos

Neste trabalho, pretende-se utilizar a análise preditiva para encontrar perfis que caracterizem utentes em fase de pré-desistência de forma a que se atue sobre eles antes que venham a concretizar a sua desistência.

No caso dos ginásios e de outras instalações públicas e privadas que prestam serviços desportivos regulares, a situação de um cliente é indicada pelo seu estado, que pode ser Ativo, se é cliente, ou Desistente se deixou de o ser. No DW criado, o atributo "*classe_desistencia*" é o que caracteriza o estado do utente à data e portanto o atributo que se pretende prever (o atributo alvo). Por questões de ordem prática optou-se por definir o atributo como um valor binário, correspondendo o valor 1 a um Desistente, e o valor 0 a um Ativo como apresentado na seção anterior do trabalho. Estamos assim perante uma tarefa de classificação dado que, conforme refere Gama (Gama et al., 2017), pretendemos formular um modelo ou hipótese capaz de relacionar os valores dos restantes atributos constantes no DW com o valor do atributo alvo, que é um atributo nominal. Na prática pretendemos construir um modelo preditivo que seja capaz de identificar conjuntos de características que permitam classificar um utente no que diz respeito ao patamar do seu estado de pré-desistente.

Sendo um algoritmo de classificação, e de acordo com vários autores que apontam as árvores de decisão ("*decision trees*") como adequado e dos mais utilizados em estudos relacionados com a retenção, perfila-se como uma alternativa viável para o problema colocado.

Este algoritmo, através de uma hierarquia de testes, permite traçar uma árvore em que cada nó pode ser um nó de divisão ou um nó folha, formando diversos ramos. Cada ramo forma uma regra com uma parte condicional, os nós de divisão, em que cada nó indica uma condição sobre um dos atributos utilizado pelo algoritmo, e uma conclusão no nó-folha à qual se chega através das várias condições existentes nos nós de divisão entre o nó folha e a raiz da árvore.

Para construir e validar modelos preditivos uteis partindo do DW construído utilizou-se o *Microsoft SQL Server Analysis Services Designer* Ver. 13.0.1701.8 e o algoritmo *Microsoft Decision Trees* (Microsoft, 2017b).

Este algoritmo requere três tipos de atributos: o atributo-chave, que identifica cada registo (ou exemplo) de forma única (atributo "*Id*"); o atributo a prever (atributo "*classe_desistencia*"); e os atributos de entrada (qualquer um dos restantes atributos), que podem ser discretos ou contínuos.

Uma vez que no caso o atributo a prever é discreto (pode assumir os valores 1 ou 0, desistente ou não desistente respetivamente), o algoritmo faz a previsão baseada nas relações entre os atributos de entrada do conjunto de dados. Utiliza os valores desses atributos, conhecidos como estados ou classes, para prever os estados do atributo a prever. O modelo DM é criado pelo algoritmo que adiciona nós à árvore sempre que determina que um atributo de entrada está significativamente correlacionado com o atributo a prever.

Os histogramas que cruzam cada atributo de entrada com as possíveis classificações do atributo a prever, apresentados na Fig. 5-1, permitem obter uma visualização gráfica da importância que cada atributo pode assumir no(s) modelo(s) que se pretende(m) construir. Conforme referido (Microsoft, 2017c), quando o algoritmo cria um conjunto possível de valores de entrada, aplica a SA para identificar os atributos e valores que providenciam mais informação, e deixa de considerar os valores que são muito raros.

Figura 5-1 – Histograma dos atributos de entrada após execução dos processos ETL na BD1 em 31/Out/2017

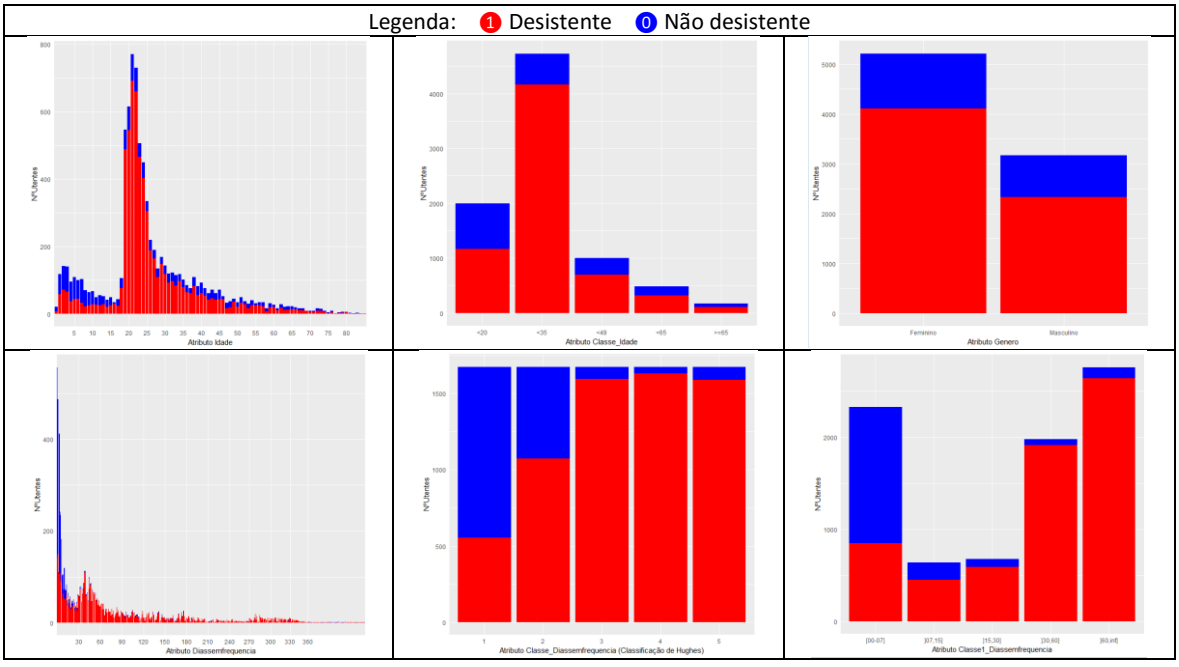


Figura 5-1 – Histograma dos atributos de entrada após execução dos processos ETL na BD1 em 31/Out/2017

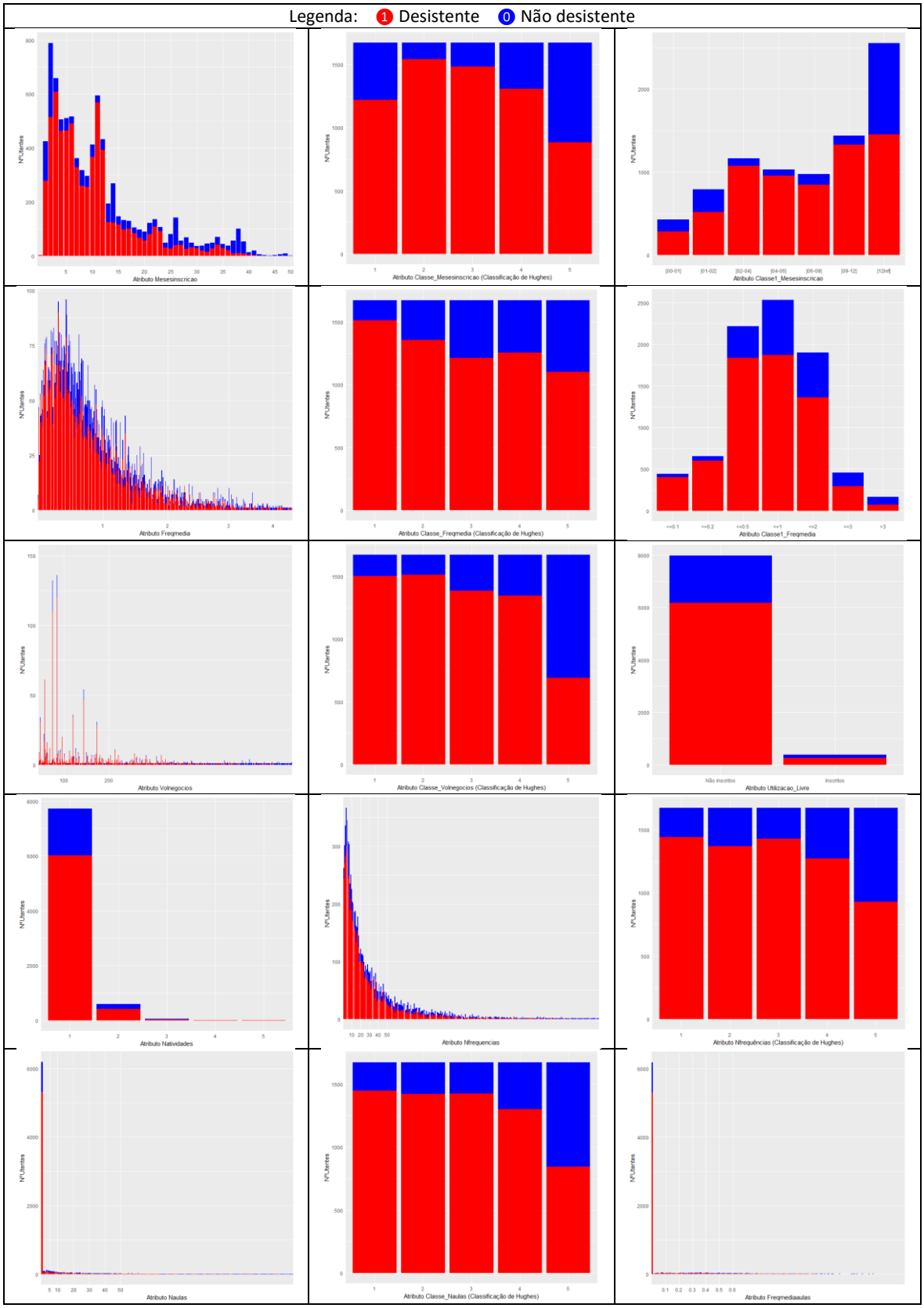


Figura 5-1 – Histograma dos atributos de entrada após execução dos processos ETL na BD1 em 31/Out/2017

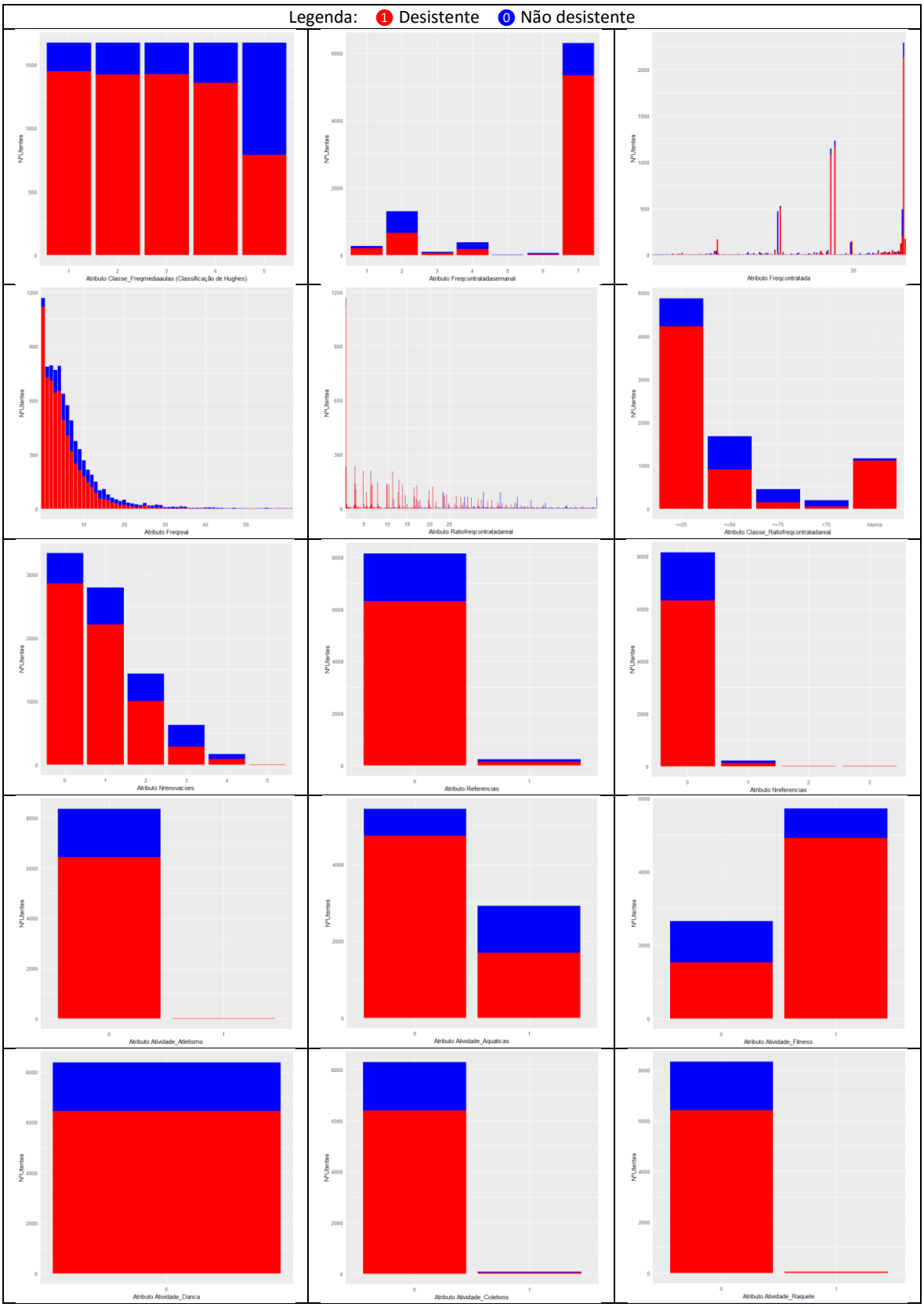
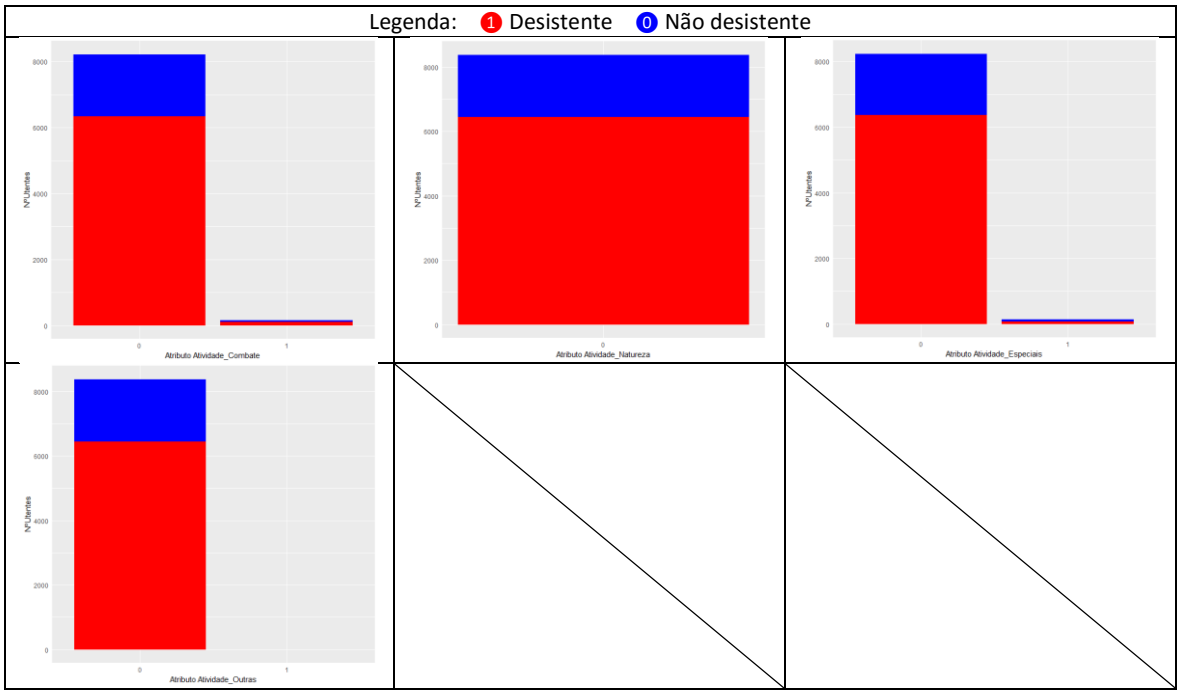


Figura 5-1 – Histograma dos atributos de entrada após execução dos processos ETL na BD1 em 31/Out/2017



No sentido de otimizar a performance, o algoritmo agrupa também os valores em caixas ("*bins*"), criando agrupamentos de valores que podem assim ser processados como uma unidade.

Após ter correlacionado os atributos, é utilizada uma equação que calcula o ganho de informação para cada atributo, e o que tiver melhor pontuação é o atributo utilizado para dividir os exemplos em subconjuntos, que por sua vez são analisados pelo mesmo processo, recursivamente, até que não se possa mais efetuar sub-divisões.

A equação usada para analisar o ganho de informação depende dos parâmetros configurados, do tipo do atributo a prever, e do tipo dos atributos de entrada.

O algoritmo "*decision trees*" utilizado pelo produto em questão aplica automaticamente um processo de SA providenciando vários métodos cuja aplicação depende do tipo dos atributos (Microsoft, 2017a).

O método "*interestingness score*" é o método utilizado por defeito para classificar e ordenar atributos que contêm dados numéricos contínuos não binários, utilizando uma medida baseada em entropia. A entropia de um determinado atributo é comparada com a entropia de todos os outros atributos através da fórmula:

$$interestingness(atrib) = -\big(m - entropia(atrib)\big) \times \big(m\ - entropia(atrib)\big)$$

em que **m** é a medida central de entropia.

Os métodos "Entropia de *Shannon*" e os dois métodos "*Bayesianos*" podem ser utilizados em atributos discretos ou discretizados de forma automática pelo próprio algoritmo.

A "Entropia de *Shannon*" mede a incerteza de uma variável aleatória para um determinado resultado através da seguinte fórmula:

$$H(x) = -\sum P(x_i) log\big(P(x_i)\big)$$

Os restantes dois métodos são baseados em redes "*Bayesianas*" que são grafos acíclicos direcionados de estados e respetivas transições entre eles. A primeira variante, o K2, é escalável e pode analisar múltiplas variáveis, mas requere ordenação nas variáveis de entrada. A segunda variante, o "*Bayesian Dirichlet with Uniform Prior*" assume um caso especial da distribuição de Dirichlet, em que uma constante é usada para criar uma distribuição fixa ou uniforme de estados anteriores.

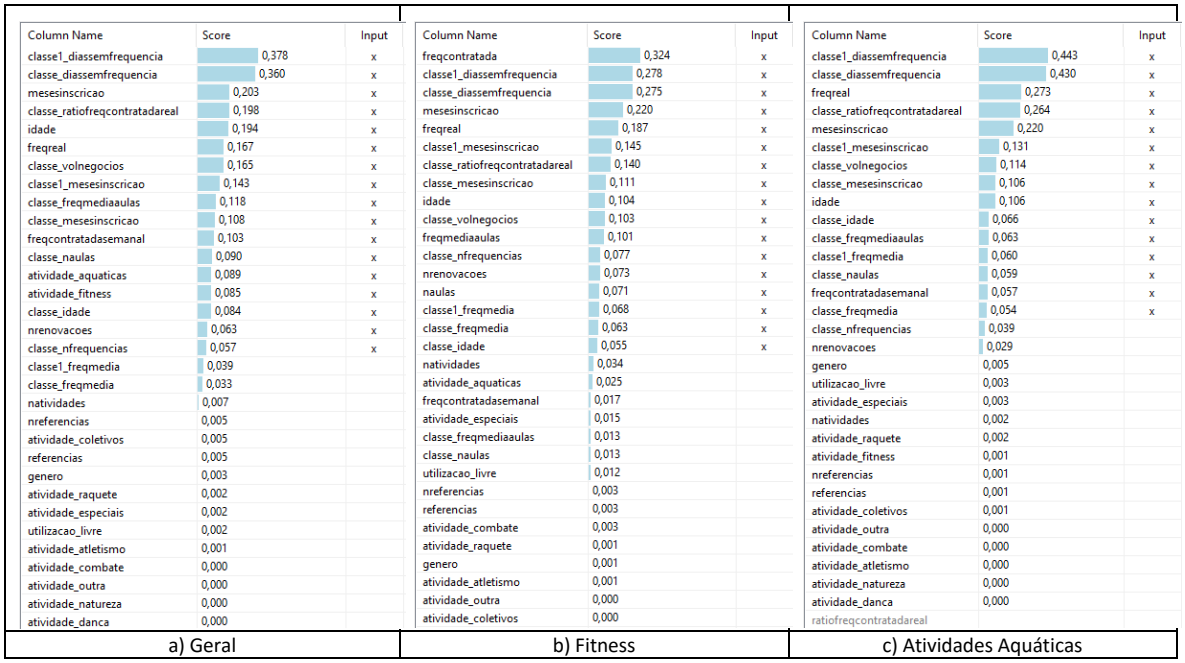
A sugestão de atributos resultante da pontuação obtida pelo algoritmo, utilizando o método *Bayesian Dirichlet with Uniform Prior* com o qual se obteve os melhores resultados, é apresentada na Fig. 5-2, sendo que o quadro à esquerda é a pontuação obtida sobre a tabela do DW que contém todos os utentes (tabela "*Retencao*"), o do meio apresenta a pontuação obtida sobre a tabela que contém apenas utentes que frequentaram a atividade *fitness* (tabela "Fitness") e o da direita apresenta a pontuação obtida sobre a tabela que contém utentes que frequentaram atividades aquáticas (tabela "*Aquaticos*").

5.2. Métricas de avaliação da qualidade dos modelos

A Fig. 5-2 permite observar que existem atributos relevantes com boa pontuação, como por exemplo os atributos "*classe_diassemfrequencia*" e "*classe1_diassemfrequencia*" nas sugestões dadas, que são, na prática, diferentes classificações de um mesmo valor. Apesar de, como indica Gama (Gama et al., 2017), o processo de construção de uma árvore selecionar os atributos a usar resultando modelos que tendem a ser bastante robustos em relação à adição de atributos irrelevantes e redundantes não podemos perder de vista que o objetivo deste trabalho é o de obter modelos preditivos que conduzam a perfis acionáveis. Torna-se pois necessário considerar combinações de atributos de entrada que

permitam atingir esse objetivo sem que com isso se degrade de forma significativa a precisão dos modelos. Neste sentido partiu-se para a construção dos modelos, utilizando num primeiro modelo todos os atributos sugeridos, construindo-se depois outros modelos com variações no conjunto dos atributos de entrada através de remoção de atributos redundantes observando-se a complexidade da árvore resultante no que diz respeito à profundidade e número de nós, e obtendo-se a precisão de cada modelo.

Figura 5-2 – Pontuação dos atributos atribuída pelo algoritmo *Microsoft Decision Trees*



Column Name	Score	Input
classe1_diassemfrequencia	0,378	x
classe_diassemfrequencia	0,360	x
mesesinscricao	0,203	x
classe_ratiofreqcontratadareal	0,198	x
idade	0,194	x
freqreal	0,167	x
classe_volnegocios	0,165	x
classe1_mesesinscricao	0,143	x
classe_freqmediaaulas	0,118	x
classe_mesesinscricao	0,108	x
freqcontratadasemanal	0,103	x
classe_naulas	0,090	x
atividade_aquaticas	0,089	x
atividade_fitness	0,085	x
classe_idade	0,084	x
nrenovacoes	0,063	x
classe_nfrequencias	0,057	x
classe1_freqmedia	0,039	
classe_freqmedia	0,033	
natividades	0,007	
nreferencias	0,005	
atividade_coletivos	0,005	
referencias	0,005	
genero	0,003	
atividade_raquete	0,002	
atividade_especiais	0,002	
utilizacao_livre	0,002	
atividade_atletismo	0,001	
atividade_combate	0,000	
atividade_outra	0,000	
atividade_natureza	0,000	
atividade_danca	0,000	

a) Geral

Column Name	Score	Input
freqcontratada	0,324	x
classe1_diassemfrequencia	0,278	x
classe_diassemfrequencia	0,275	x
mesesinscricao	0,220	x
freqreal	0,187	x
classe1_mesesinscricao	0,145	x
classe_ratiofreqcontratadareal	0,140	x
classe_mesesinscricao	0,111	x
idade	0,104	x
classe_volnegocios	0,103	x
freqmediaaulas	0,101	x
classe_nfrequencias	0,077	x
nrenovacoes	0,073	x
naulas	0,071	x
classe1_freqmedia	0,068	x
classe_freqmedia	0,063	x
classe_idade	0,055	x
natividades	0,034	
atividade_aquaticas	0,025	
freqcontratadasemanal	0,017	
atividade_especiais	0,015	
classe_freqmediaaulas	0,013	
classe_naulas	0,013	
utilizacao_livre	0,012	
nreferencias	0,003	
referencias	0,003	
atividade_combate	0,003	
atividade_raquete	0,001	
genero	0,001	
atividade_atletismo	0,001	
atividade_outra	0,000	
atividade_coletivos	0,000	

b) Fitness

Column Name	Score	Input
classe1_diassemfrequencia	0,443	x
classe_diassemfrequencia	0,430	x
freqreal	0,273	x
classe_ratiofreqcontratadareal	0,264	x
mesesinscricao	0,220	x
classe1_mesesinscricao	0,131	x
classe_volnegocios	0,114	x
classe_mesesinscricao	0,106	x
idade	0,106	x
classe_idade	0,066	x
classe_freqmediaaulas	0,063	x
classe1_freqmedia	0,060	x
classe_naulas	0,059	x
freqcontratadasemanal	0,057	x
classe_freqmedia	0,054	x
classe_nfrequencias	0,039	
nrenovacoes	0,029	
genero	0,005	
utilizacao_livre	0,003	
atividade_especiais	0,003	
natividades	0,002	
atividade_raquete	0,002	
atividade_fitness	0,001	
nreferencias	0,001	
referencias	0,001	
atividade_coletivos	0,001	
atividade_outra	0,000	
atividade_combate	0,000	
atividade_atletismo	0,000	
atividade_natureza	0,000	
atividade_danca	0,000	
ratiofreqcontratadareal		

c) Atividades Aquáticas

A Tab. 5-1 apresenta os modelos criados tendo por base os dados da tabela "*Retencao*" do DW - pelo que dizem respeito aos Utentes que frequentaram qualquer uma das atividades de *fitness* ou aquáticas - e os atributos utilizados em cada um.

A Tab. 5-2 apresenta os modelos criados, a profundidade da árvore criada pelo modelo e o número de nós da árvore. Nesta tabela, são apresentados os modelos adicionais com o nome Ret11, Ret21, Ret31, Ret71, Ret81 e Ret91 que foram obtidos com os mesmos atributos dos modelos com o mesmo prefixo, Ret1, Ret2, Ret3, Ret7, Ret8 e Ret9 respetivamente, mas com o parâmetro MINIMUM_SUPPORT = 50 para que os nós da árvore nunca tenham menos de 50 registos, número que consideramos suficientemente significativo para justificar uma ação. A observação desta tabela permite constatar a redução da profundidade da árvore e do número de nós relativamente ao modelo inicial, em que o parâmetro apresentava o valor por defeito (MINIMUM_SUPPORT = 10).

Tabela 5-1 - Atributos utilizados nos modelos preditivos

Modelo	Atributos utilizados
Ret1	Atividade_aquaticas, Atividade_atletismo, Atividade_especiais, Atividade_fitness, Atividade_raquete, Classe_diassemfrequencia, Classe_freqmedia, Classe_freqmediaaulas, Classe_idade, Classe_mesesinscricao, Classe_Naulas, Classe_nfrequencias, Classe_Rationfreqcontratadareal, Classe_volnegocios, Classe1_diassemfrequencia, Classe1_freqmedia, Classe1_mesesinscricao, Freqcontratadasemanal, Freqreal, Genero, Idade, Mesesinscricao, Natividades, Nreferencias, Nrenovacoes, Utilizacao_livre
Ret2	Atividade_aquaticas, Atividade_atletismo, Atividade_especiais, Atividade_fitness, Atividade_raquete, Classe_diassemfrequencia, Classe_freqmedia, Classe_freqmediaaulas, Classe_idade, Classe_mesesinscricao, Classe_Naulas, Classe_nfrequencias, Classe_Rationfreqcontratadareal, Classe_volnegocios, Freqcontratadasemanal, Genero, Idade, Mesesinscricao, Natividades, Nreferencias, Nrenovacoes, Utilizacao_livre
Ret3	Atividade_aquaticas, Atividade_atletismo, Atividade_especiais, Atividade_fitness, Atividade_raquete, Classe_idade, Classe_Naulas, Classe_nfrequencias, Classe_Rationfreqcontratadareal, Classe_volnegocios, Classe1_diassemfrequencia, Classe1_freqmedia, Classe1_mesesinscricao, Freqcontratadasemanal, Genero, Natividades, Nreferencias, Nrenovacoes, Utilizacao_livre
Ret7	Atividade_aquaticas, Atividade_atletismo, Atividade_especiais, Atividade_fitness, Atividade_raquete, Classe_idade, Classe_Rationfreqcontratadareal, Classe_volnegocios, Classe1_diassemfrequencia, Classe1_freqmedia, Classe1_mesesinscricao, Freqcontratadasemanal, Natividades, Nreferencias, Nrenovacoes
Ret8	Atividade_aquaticas, Atividade_fitness, Classe_volnegocios, Classe1_diassemfrequencia, Classe1_freqmedia, Classe1_mesesinscricao, Freqcontratadasemanal, Natividades, Nrenovacoes
Ret9	Classe_idade, Classe_volnegocios, Classe1_diassemfrequencia, Classe1_mesesinscricao, Freqcontratadasemanal, Nrenovacoes

Tabela 5-2 – Profundidade das Árvores e Número de Nós em cada modelo

Dados obtidos após o processo ETL em 31/Out/2017 na BD1

Modelo	Profundidade da Árvore	Número de Nós
Ret1	6	34
Ret11	5	26
Ret2	6	41
Ret21	5	33
Ret3	8	44
Ret31	6	28
Ret7	7	38
Ret71	6	30
Ret8	7	57
Ret81	6	37
Ret9	7	53
Ret91	7	45

A Tab. 5-3 apresenta a matriz de confusão e as métricas obtidas em cada um dos modelos. Como explicado anteriormente, uma vez que as formas de utilização nas atividades aquáticas e de *fitness* são diferentes, foram também criados modelos apenas para os Utentes que frequentaram *fitness*, apresentados na Tab. 5-4, e para os Utentes que frequentaram atividades Aquáticas, apresentados na Tab. 5-5. Nestes modelos, não foi considerado o atributo de frequência da atividade, nomeadamente o atributo "*atividade_fitness*" e "*atividade_aquaticas*", respetivamente.

As métricas de avaliação dos modelos preditivos foram obtidas através do método *Holdout*, com os dados particionados em dois conjuntos disjuntos: o conjunto de treino com 70% dos dados, e o conjunto de teste com os restantes 30%.

Tabela 5-3 – Métricas *Holdout* dos modelos preditivos (tabela "Retencao")

Dados obtidos após o processo ETL em 31/Out/2017 na BD1

Modelo	Matriz de Confusão Previsto / Real				*Accuracy*	*Error Rate*	*Sensitivity*	*Specificity*	*False Positive Rate*	*Precision*	*F-Score*
	F/F	F/V	V/F	V/V							
Ret1	409	80	179	1844	89.59%	10.31%	95.87%	69.56%	30.44%	91.15%	93.44%
Ret11	409	95	179	1829	89.09%	10.91%	95.06%	69.56%	30.44%	91.09%	93.03%
Ret2	423	96	165	1828	89.61%	10.39%	95.01%	71.94%	28.06%	91.72%	93.34%
Ret21	409	98	179	1826	88.97%	11.03%	94.91%	69.56%	30.44%	91.07%	92.95%
Ret3	442	119	146	1805	89.45%	10.55%	93.81%	75.17%	24.83%	92.52%	93.16%
Ret31	406	102	182	1822	88.69%	11.31%	94.70%	69.05%	30.95%	90.92%	92.77%
Ret7	429	120	168	1795	88.54%	11.46%	93.73%	71.86%	28.14%	91.44%	92.57%
Ret71	433	140	164	1775	87.90%	12.10%	92.69%	72.53%	27.47%	91.54%	92.11%
Ret8	437	133	160	1782	88.34%	11.66%	93.05%	73.20%	26.80%	91.76%	92.40%
Ret81	433	140	164	1775	87.90%	12.10%	92.69%	72.53%	27.47%	91.54%	92.11%
Ret9	423	117	174	1798	88.42%	11.58%	93.89%	70.85%	29.15%	91.18%	92.51%
Ret91	417	132	180	1783	87.58%	12.42%	93.11%	69.85%	30.15%	90.83%	91.95%

Tabela 5-4 – Métricas *Holdout* dos modelos preditivos (tabela "Fitness")

Dados obtidos após o processo ETL em 31/Out/2017 na BD1

Modelo	Matriz de Confusão Previsto / Real				*Accuracy*	*Error Rate*	*Sensitivity*	*Specificity*	*False Positive Rate*	*Precision*	*F-Score*
	F/F	F/V	V/F	V/V							
Fit1	159	27	94	1436	92.95%	7.05%	98.15%	62.85%	37.15%	93.86%	95.96%
Fit11	128	23	125	1440	91.38%	8.62%	98.43%	50.59%	49.41%	92.01%	95.11%
Fit2	86	14	167	1449	89.45%	10.55%	99.04%	33.99%	66.01%	89.67%	94.12%
Fit21	131	70	122	1393	88.81%	11.19%	95.22%	51.78%	48.22%	91.95%	93.55%
Fit3	193	141	60	1322	88.29%	11.71%	90.36%	76.28%	23.72%	95.66%	92.93%
Fit31	161	115	92	1348	87.94%	12.06%	92.14%	63.64%	36.36%	93.61%	92.87%
Fit7	138	83	115	1380	88.46%	11.54%	94.33%	54.55%	45.45%	92.31%	93.31%
Fit71	161	115	92	1348	87.94%	12.06%	92.14%	63.64%	36.36%	93.61%	92.87%
Fit8	154	98	99	1365	88.52%	11.48%	93.30%	60.87%	39.13%	93.24%	93.27%
Fit81	152	98	101	1365	88.40%	11.60%	93.30%	60.08%	39.92%	93.11%	93.21%
Fit9	155	113	98	1350	87.70%	12.30%	92.28%	61.26%	38.74%	93.23%	92.75%
Fit91	161	115	92	1348	87.94%	12.06%	92.14%	63.64%	36.36%	93.61%	92.87%

Em termos de precisão dos modelos preditivos, a observação das Tab. 5-3, 5-4 e 5-5 permite constatar que não há ganhos significativos nas métricas dos modelos aplicados aos utentes das atividades aquáticas ou do *fitness* em separado relativamente aos modelos

Tabela 5-5 – Métricas *Holdout* dos modelos preditivos (tabela "Aquaticos")

Dados obtidos após o processo ETL em 31/Out/2017 na BD1

Modelo	Matriz de Confusão Previsto / Real				*Accuracy*	*Error Rate*	*Sensitivity*	*Specificity*	*False Positive Rate*	*Precision*	*F-Score*
	F/F	F/V	V/F	V/V							
Aq1	333	51	38	455	89.85%	10.15%	89.92%	89.76%	10.24%	92.29%	91.09%
Aq11	334	69	37	437	87.91%	12.09%	86.36%	90.03%	9.97%	92.19%	89.18%
Aq2	332	78	39	428	86.66%	13.34%	84.58%	89.49%	10.51%	91.65%	87.98%
Aq21	304	60	67	446	85.52%	14.48%	88.14%	81.94%	18.06%	86.94%	87.54%
Aq3	333	51	38	455	89.85%	10.15%	89.92%	89.76%	10.24%	92.29%	91.09%
Aq31	334	69	37	437	87.91%	12.09%	86.36%	90.03%	9.97%	92.19%	89.18%
Aq7	333	51	38	455	89.85%	10.15%	89.92%	89.76%	10.24%	92.29%	91.09%
Aq71	334	69	37	437	87.91%	12.09%	86.36%	90.03%	9.97%	92.19%	89.18%
Aq8	316	33	55	473	89.97%	10.03%	93.48%	85.18%	14.82%	89.58%	91.49%
Aq81	309	42	62	464	88.14%	11.86%	91.70%	83.29%	16.71%	88.21%	89.92%
Aq9	316	33	55	473	89.97%	10.03%	93.48%	85.18%	14.82%	89.58%	91.49%
Aq91	309	42	62	464	88.14%	11.86%	91.70%	83.29%	16.71%	88.21%	89.92%

criados com base na globalidade dos utentes. Por outro lado, apesar das formas de utilização serem distintas, os atributos prioritários encontrados pelo algoritmo nestas situações são praticamente os mesmos que os encontrados pelo algoritmo quando aplicado sobre a tabela do DW que contém todos os utentes, embora com pontuações diferentes (Fig. 5-2) e dando origem a árvores de decisão também diferentes.

Considerando os modelos em que o número mínimo de exemplos por nó é igual ou superior a 50, que apresentam menos nós e uma menor profundidade da árvore do modelo, consequentemente menos dispersos, mais significativos e minimizando a ocorrência de *overfitting* através da redução da fragmentação, constata-se pela observação da Tab. 5-3 que não há diferenças significativas, em termos das avaliações obtidas com as métricas entre os vários modelos que têm em consideração todos os utentes – maior diferença é de 1.51% na métrica *Accuracy* entre o modelo Ret11 e o modelo Ret91.

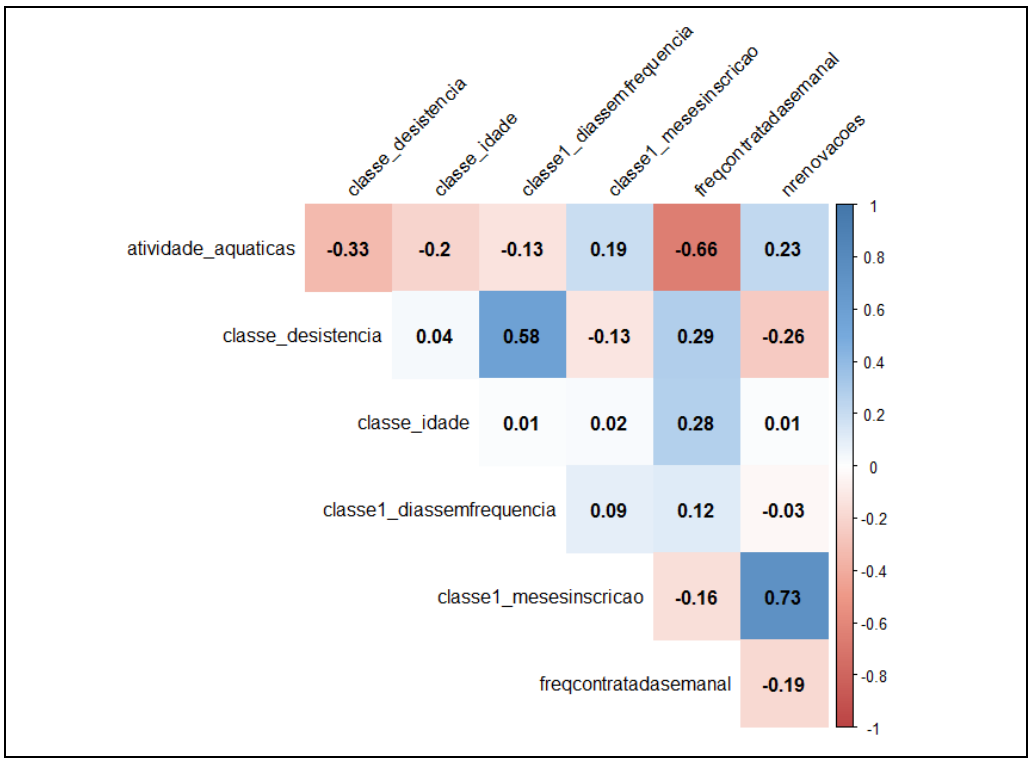
Embora o modelo Ret71 considere mais atributos a entrada que o Ret81 (Tab. 5-1) a matriz de confusão obtida é idêntica e consequentemente também as respetivas métricas. Contudo o conjunto dos atributos utilizados na prática pelo algoritmo diferem e são diferentes do conjunto de entrada indicado ao algoritmo para cada modelo. Do conjunto de entrada indicado na Tab. 5-1 o modelo Ret71 só utiliza os atributos "*classe1_diassemfrequencia*", "*atividade_aquaticas*", "*classe1_mesesinscricao*", "*nrenovacoes*", "*classe_idade*" e "*freqcontratadasemanal*"; e o modelo Ret81 adiciona o atributo "*atividade_fitness*" aos indicados para o modelo Ret71.

Há no entanto uma diferença no número de nós da árvore de decisão de um e outro modelo. O Ret71 apresenta uma árvore com 30 nós enquanto que o Ret81 apresenta uma árvore com 37 nós.

Assim sendo, não se podendo considerar haver um modelo substancialmente melhor entre os modelos criados, a escolha sobre o modelo a utilizar recai sobre um modelo criado com base em todos os utentes e que apresenta uma árvore em que os atributos utilizados são menos redundantes e com menos nós (Princípio da Parcimónia / "*Occam's Razor*"), uma vez que reduz a complexidade e facilita a criação de ações na fase seguinte. Assim, com base nos modelos criados sobre os dados obtidos pelo processo ETL sobre a BD1 em

Out/2017 optou-se pelo modelo preditivo Ret71 apresentando-se o quadro de correlação dos seus atributos na Fig. 5-3.

Figura 5-3 - Quadro de correlação dos atributos utilizados no Modelo Ret71



Para complementar a avaliação deste modelo, utilizou-se também o método *Cross-Validation*, tendo-se segmentado os registos em 10 partições de igual tamanho. Em cada execução, uma das partições foi usada para testar enquanto as outras foram usadas para treino do modelo, repetindo o processo 10 vezes sendo que cada partição é usada apenas uma vez para testar o modelo. Com este método de avaliação obtém-se a matriz de confusão e os desvios padrão apresentadas na Tab. 5-6 que permite concluir, em função do coeficiente de variação, haver uma pequena variação nos estados corretamente classificados e muito pequena nos casos incorretamente classificados.

Tabela 5-6 – Matriz de confusão e desvios padrão do Modelo Ret71 obtidos com o método *Cross-Validation*

	Classificações médias		Desvios Padrão		Coeficiente de Variação	
Previsto / Real	Não Desistente	Desistente	Não Desistente	Desistente		
Não Desistente	100.300	36.702	6.784	5.622	0.068	0.153
Desistente	32.997	416.298	6.528	5.622	0.198	0.014

Os resultados das métricas obtidas com o método *Cross-Validation* indicadas na Tab. 5-7 são ligeiramente melhores do que as obtidas com o método *Holdout* nas métricas de avaliação geral (*Accuracy* com mais 0.21% e consequentemente a *Error Rate* com menos 0.21% e *Precision* com mais 1.12% e *F-Score* com mais 0.17%). No que diz respeito às métricas relacionadas com a classificação dos não desistentes, ocorre uma oscilação simétrica, melhorando a classificação dos verdadeiros negativos (*Specificity* com mais 2.71%) e piorando a dos falsos positivos (*False Positive Rate* com -2.71%). A métrica relacionada com a classificação dos verdadeiros positivos degrada-se em 0.79%.

Tabela 5-7 – Métricas *Cross-validation* do modelo preditivo Ret71

Dados obtidos após o processo ETL em 31/Out/2017 na BD1

Modelo	*Accuracy*	*Error Rate*	*Sensitivity*	*Specificity*	*False Positive Rate*	*Precision*	*F-Score*
Ret71	88.11	11.89	91.90	75.24	24.76	92.66	92.28

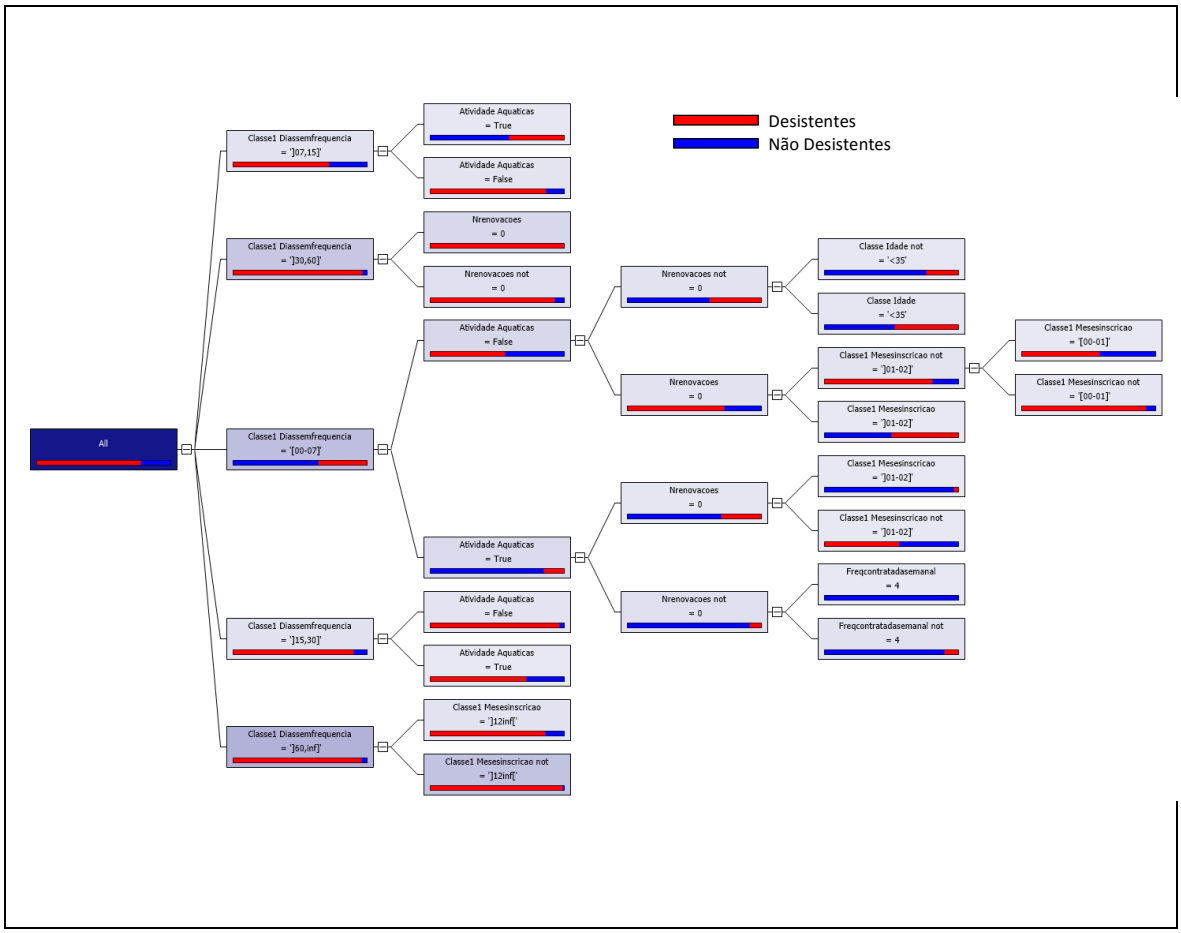
A árvore resultante do modelo, onde estão expressas as regras que caracterizam os nós, é apresentada na Fig. 5-4.

5.3. Obtenção dos perfis acionáveis

Na árvore construída pelo algoritmo, cada folha é definida por um conjunto de regras que a caracterizam, definidas pelo caminho da folha até à raiz. Uma vez que o parâmetro MINIMUM_SUPPORT = 50 sabe-se que cada folha tem pelo menos um número significativo de exemplos, leia-se utentes, e que podem existir em cada uma dessas folhas utentes que são desistentes (porção assinalada a vermelho na Fig. 5-4) e utentes que não o são (porção assinalada a azul na Fig. 5-4). A relação entre essas quantidades define, nessa folha, um limiar de probabilidade de desistência para o conjunto de regras que a definem. Uma vez que podemos obter o conjunto de regras que definem essas folhas, podemos, a partir delas, traçar o perfil de possíveis desistentes.

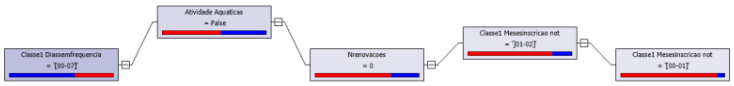
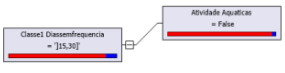
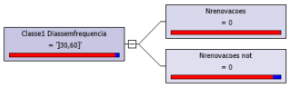
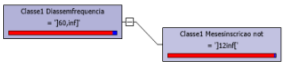
Se, por exemplo, considerarmos o limiar de probabilidade de desistência nos 90% encontramos na árvore de decisão da Fig. 5-4 algumas folhas que apresentam limiares superiores.

Figura 5-4 – Árvore de Decisão do Modelo Ret71



A Tab. 5-8 apresenta as condições que definem essas folhas e que traçam diferentes perfis.

Tabela 5-8 – Perfis de desistência

	**Perfil A (92.66%)** Utente que não visita as instalações há menos de 7 dias, não frequenta atividades aquáticas, nunca renovou a sua inscrição e a sua inscrição tem mais de 2 mês
	**Perfil B (96.22%)** Utente que não visita as instalações entre 16 e 30 dias e não frequenta atividades aquáticas
	**Perfil C (96.38%)** Utente que não visita as instalações entre 31 a 60 dias
	**Perfil D (98.42%)** Utente que não visita as instalações há mais de 60 dias, mas cuja inscrição tem menos de 12 meses

Para encontrar os modelos preditivos que resultou na seleção do modelo Ret71 com a árvore de decisão apresentada na Fig. 5-4 seguiram-se as orientações da metodologia CRISP-DM (IBM, 2011). De acordo com esta metodologia o processo decorre num ciclo de seis fases iniciando-se com o conhecimento do negócio, o estudo e conhecimento dos dados, seguindo-se a sua preparação e a criação dos modelos preditivos. A satisfação ou insatisfação com as métricas de avaliação obtidas pode levar à implementação dos modelos ou a rever de novo o processo a partir das fases iniciais esperando obter-se no final de cada ciclo melhores modelos, não só no que diz respeito aos valores das métricas, mas também no que diz respeito a obter-se árvores de decisão que apresentem segmentos acionáveis.

Por outro lado, com a implementação do sistema proposto neste trabalho, é possível que surjam novos atributos relevantes e alterações nos padrões obtidos a partir dos dados existentes. Em resultado, espera-se que as árvores de decisão resultantes da aplicação do modelo ao longo do tempo apresentem diferentes estruturas e possam inclusivamente fazer uso de outros atributos.

6. Ações de fidelização

6.1. Seleção dos segmentos acionáveis e ações de fidelização

Para concentrar os esforços de aumento da taxa de retenção em grupos mais suscetíveis de abandono dos serviços desportivos regulares seguiu-se uma abordagem *bottom-up* conforme refere Gorgoglione (Gorgoglione, 2011) propondo-se a aplicação de ações de fidelização sobre a segmentação obtida com o modelo preditivo referido no Capitulo 5. Os segmentos pretendidos correspondem às folhas da árvore que apresentam um determinado limiar de probabilidade de desistência. Por sua vez, as regras de construção da árvore de decisão que levam da raiz às folhas que definem esses segmentos indicam o perfil dos utentes sobre os quais pretendemos fazer incidir as ações de fidelização.

Este processo de segmentação obedece aos critérios de utilidade de Kotler (Kotler & Keller, 2009) que indica que uma segmentação só é útil se os segmentos obedecerem a cinco critérios: sejam mensuráveis, substanciais, acessíveis, diferenciáveis e acionáveis. São mensuráveis uma vez que é possível determinar o número de utentes enquadrado em cada segmento. São substanciais porque é possível determinar um número mínimo de utentes nesses segmentos que justifique um tratamento diferenciado e estão acessíveis porque basta comparar as características de cada utente com o perfil definido para o enquadrar, ou não, no segmento. São diferenciáveis porque cada segmento definido por cada folha da árvore de decisão depende do valor ou classe de cada atributo utilizado para definir as condições que levam da raiz da árvore de decisão até à folha, havendo sempre diferenças de ramo para ramo. Finalmente, e também pelo mesmo motivo, as condições que definem o segmento apresentam as características dos utentes em situação de desistência, pelo que a sua análise permite o planeamento e a implementação de ações que tentem evitar essas desistências.

O processo de escolha dos segmentos acionáveis começa pela análise da árvore e pela escolha do limiar de probabilidade de desistência. Esta indicação seleciona de forma automática as folhas a considerar e consequentemente os perfis acionáveis. Como se viu no exemplo do Capitulo 5, a definição do limiar de probabilidade de desistência nos 90% levou à seleção dos perfis indicados na Tab. 5-9.

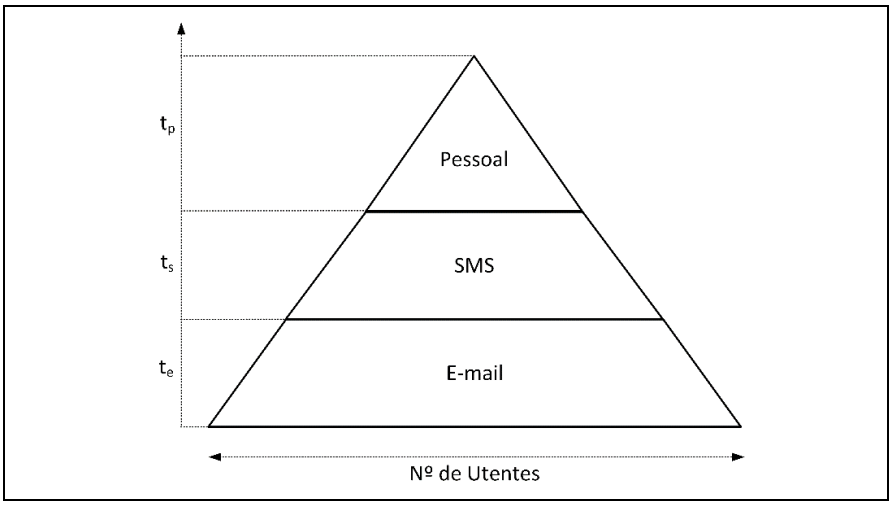
Para cada um dos perfis obtido deve ser definido um fluxo de ações condicionalmente sequenciais em que, num primeiro momento, todos os utentes que apresentam esse perfil

são alvo dessa ação. Num segundo momento, apenas os utentes que não mudaram o seu comportamento são alvo da segunda ação e finalmente, num terceiro momento, só são alvo da terceira ação os que não alteraram o seu comportamento após a primeira e a segunda ação.

Caso o comportamento do utente se altere, leia-se "a alteração das características e comportamento do utente deixam de o colocar na folha da árvore onde inicialmente foi colocado", o utente deixa de ser alvo dessa sequência ou fluxo de ações. Se essa alteração o enquadra num segmento alvo de outro *workflow*, então o utente passa a ser enquadrado no fluxo de ações previsto para o novo perfil.

No intuito de reduzir o esforço envolvido nas ações de fidelização propõe-se a utilização de canais de comunicação cujo custo seja inversamente proporcional ao número de utentes que envolve. Uma vez que é esperado que em cada estágio pelo menos alguns dos utentes alvo de cada um dos tipos de ação alterem o seu comportamento no sentido de não se tornarem desistentes, propõe-se a implementação do *workflow* em três estágios iniciando-se com o e-mail, onde o custo é praticamente zero, passando-se depois à utilização do SMS e por fim ao contato pessoal, construindo-se assim um encadeamento em pirâmide das ações a desenvolver, conforme ilustrado na Fig. 6-1.

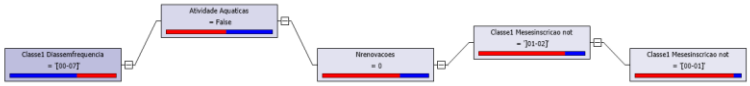
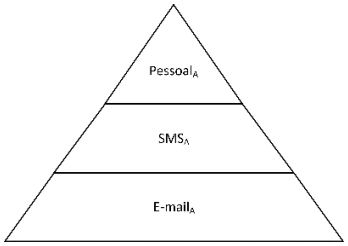
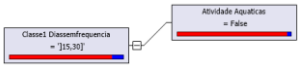
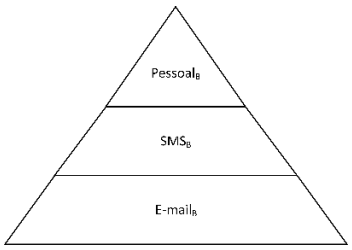
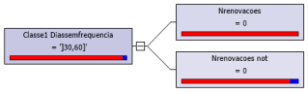
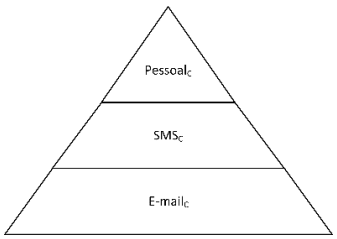
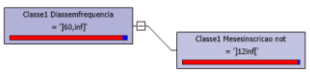
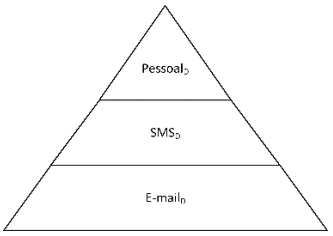
Figura 6-1 - Pirâmide de ações de fidelização



Uma vez que a árvore de decisão criada pelo algoritmo pode apresentar diversas folhas em que o limiar de desistência é superior ao indicado, podem ser desenhados e implementados

diversos *workflows* com diferentes mensagens adequadas a cada um dos perfis considerados. A Tab. 6-1, construída a partir da Tab. 5-9, ilustra o que referimos.

Tabela 6-1 – Perfis de desistência e *workflow* de ações de fidelização a aplicar

	**Perfil A (92.66%)** 
	**Perfil B (96.22%)** 
	**Perfil C (96.38%)** 
	**Perfil D (98.42%)** 

6.2. Canais de comunicação e respetivos requisitos

A escolha dos canais de comunicação não é fácil dada a sua atual fragmentação e desordem (Kotler & Keller, 2009). Na escolha, para além do custo é importante ter em conta a sua capacidade de personalização e de obtenção de *feedback* e por isso vários autores (Kotler & Keller, 2009), (Pousttchi & Wiedemann, 2006), (Merisavo & Raulas, 2004) atribuem ao e-mail e ao telefone (quer no que diz respeito às chamadas de voz como ao envio de SMS –

Short Message Service) grande potencial na promoção de ações cujo objetivo é o de aumentar a lealdade e as taxas de retenção.

No que diz respeito à utilização do e-mail existem essencialmente dois riscos que devem ser tidos em conta: o risco de, por algum motivo, o email não ser entregue; e o risco do destinatário não o ler.

Para minimizar a ocorrência destes riscos devem ser consideradas algumas boas práticas (ResultadosDigitais, 2017) na implementação dos e-mails. A Tab. 6-2 sintetiza algumas dessas boas práticas.

Tabela 6-2 – Algumas boas práticas para a utilização de emails

Boa prática	Descrição
HTML bem construído	O HTML utilizado na construção do email deve utilizar as boas práticas de codificação desta linguagem de marcação
Incluir uma versão em texto pleno	Os emails enviados em HTML devem conter uma versão em texto pleno para que possam ser lidos em sistemas que não permitem a leitura em HTML. A consistência entre o texto pleno e o HTML é importante para a validação efetuada pelos filtros antispam
Indicar um *Preheader*	O *Preheader* é uma informação complementar ao assunto e ajuda o destinatário a determinar a relevância do email sem ter de o abrir.
Utilizar uma proporção adequada entre imagens e texto	A proporção ideal é de 40% de imagens e 60% de texto. Não deve definir o corpo da mensagem apenas com uma imagem, sem texto adicional. Uma vez que muitos ISP interpretam o conteúdo do email, se o conteúdo for apenas uma imagem, o email pode ser automaticamente direcionado para SPAM
Não utilizar imagens "pesadas"	Cada imagem deve ter no máximo 100Kb, e o somatório do tamanho das imagens não deve exceder os 470Kb
Personalizar o email	O email deve indicar dados específicos do destinatário, como o nome, motivo pelo qual o email lhe é dirigido, temas que se sabe à priori que o destinatário prefere, etc.
Indicar um assunto sucinto e explícito	O assunto deve estar limitado a 50 caracteres e não deve ser indicado utilizando apenas letras maiúsculas. Deve ser explicito no que diz respeito ao conteúdo do email
Indicar um *Call-To-Action*	Um *Call-To-Action* é um estimulo para levar o destinatário do email a fazer qualquer coisa. Deve ser incluído um *Call-To-Action* e título nos primeiros 300px de altura do email.
Evitar a utilização de termos banais e de pontuação abusiva	Evitar a utilização de termos como "promoção", "grátis", "crédito", etc. Evitar a utilização de pontuação exagerada do tipo "!!!!" ou "!?!?!?".
Escolher um horário de envio adequado	Os emails devem ser enviados em horas em que o destinatário tenha a caixa de correio aberta para que possa reagir de imediato à sua receção
Incluir assinatura	Os emails devem apresentar uma assinatura que identifique sem margem para dúvidas o emissor da mensagem
Solicitar o consentimento prévio	A solicitação de consentimento para o envio de emails publicitários é, para além de uma boa prática, obrigatória por imposição do Regulamento Geral de Proteção de Dados.
Disponibilizar uma opção para remover a subscrição	Deve incluir uma opção para que o destinatário possa optar por não receber mais emails (opt-out). A inexistência desta opção pode levar o destinatário a indicar o seu email como SPAM dificultando o envio de mensagens seguintes. Por outro lado, de acordo com o RGPD, é uma indicação que deve ser seguida

A utilização dos SMS, sendo um canal de comunicação pessoal mais controlado porque tem um custo associado, não apresenta tantos riscos, quer no que diz respeito à entrega, quer no que diz respeito à própria leitura. No entanto, há também boas práticas que devem ser seguidas como indica a Tab. 6-3 (The Mobile Experience Company Inc., 2018).

Por outro lado, face à grande implementação de *Smartphones*, os SMS podem incluir ligações que permitam ao utente aceder a uma página da Internet, tal como se tratasse de

uma *Call-To-Action* (CTA) num email. Contudo, as limitações relacionadas com o tamanho da mensagem e o respetivo custo associado pode representar um problema, pelo que o utilizador deve ter cuidados adicionais quer com a utilização de caracteres portugueses, já que em muitos casos a utilização destes caracteres duplica o tamanho da mensagem; quer com a utilização de variáveis na pré-definição da mensagem, que são posteriormente substituídas pelo valor do atributo a que correspondem o que em muitos casos pode fazer variar o tamanho previsto da mensagem.

Tabela 6-3 – Algumas boas práticas para a utilização de SMS

Boa prática	Descrição
Solicitar o consentimento prévio	A solicitação de consentimento para o envio de SMS publicitários é, para além de uma boa prática, obrigatória por imposição do Regulamento Geral de Proteção de Dados.
Enviar a mensagem em horários de expediente	O sinal sonoro emitido pelo telemóvel aquando da receção da mensagem pode ser incomodativo principalmente se a mensagem for recebida em horas de descanso
Utilizar conteúdos e linguagem adequada	Dado a pequena dimensão da mensagem, deve ser evitada a tendência de utilização de abreviaturas que tornam a mensagem difícil de ler e de entender.
Disponibilizar uma opção para remover a subscrição	Deve incluir uma opção para que o destinatário possa optar por não receber mais SMS (opt-out). Por outro lado, de acordo com o RGPD, é uma indicação que deve ser seguida

Por defeito, as mensagens SMS utilizam um conjunto de caracteres GSM 03.38 com codificação de 7 bits. Quando são usados apenas caracteres deste conjunto, o tamanho máximo para uma mensagem é de 160 caracteres. No entanto, a utilização de caracteres fora deste conjunto altera a forma como é calculado o tamanho da mensagem. Se forem usados caracteres da tabela de extensão GSM 03.38, cada um destes caracteres conta como dois. Se forem usados caracteres fora destes conjuntos, passa a ser usado, de forma automática, a codificação Unicode (UCS-2) pelo que o tamanho da mensagem fica reduzido a 70 caracteres.

Se forem usados mais de 160 caracteres (ou de 70, no caso de se ter usado UCS-2) a mensagem é sub-dividida e enviada em partes. Para que o sistema não perca a ordem pela qual a mensagem deve ser apresentada ao destinatário, é ainda utilizado um cabeçalho especial (UDH) que reduz o tamanho em 7 caracteres (ou 3 se estiver a ser utilizado o UCS-2).

Por fim, no último estágio do *workflow,* também os contatos a efetuar por telefone apresentam os seus riscos: o utente pode não atender, o momento pode não ser o adequado ou pode mesmo não estar interessado em apresentar justificações ou argumentar com quem lhe telefona.

Existem técnicas conhecidas aplicadas nos questionários telefónicos que podem ser utilizadas para minimizar estas situações (Smith & Albaum, 2012). No entanto, conforme propomos, se o contato telefónico for efetuado pelo Profissional de Desporto que ministra as aulas que o utente frequenta haverá uma intimidade relativa que levará a que o utente não encare esse contato com desconforto. Por outro lado, a apresentação de incentivos pode também ajudar a ultrapassar esta barreira, não só no contato telefónico pessoal, mas em qualquer um dos canais indicados nos restantes estágios do *workflow*.

As ações possíveis de aplicar em cada estágio do *workflow* podem ser agrupadas de acordo com a sua finalidade em quatro grandes grupos:

- Ações informativas, com informação personalizada sobre os horários que pode frequentar e sobre benefícios incluídos no seu serviço de que não está a fazer uso;
- Ações de perceção da satisfação, nomeadamente de inquéritos Net Promotor Score (NPS), do motivo das ausências e de outros inquéritos de qualidade;
– Oferta de benefícios, nomeadamente ao nível da frequência gratuita de aulas de outras atividades não incluídas no serviço contratado, aulas de substituição, participação em master-classes ou outros eventos, utilização gratuita de recintos, brindes ou vales;
- Remoção da subscrição e/ou atualização de consentimentos, ao abrigo do Regulamento Geral de Proteção de Dados (RGPD);

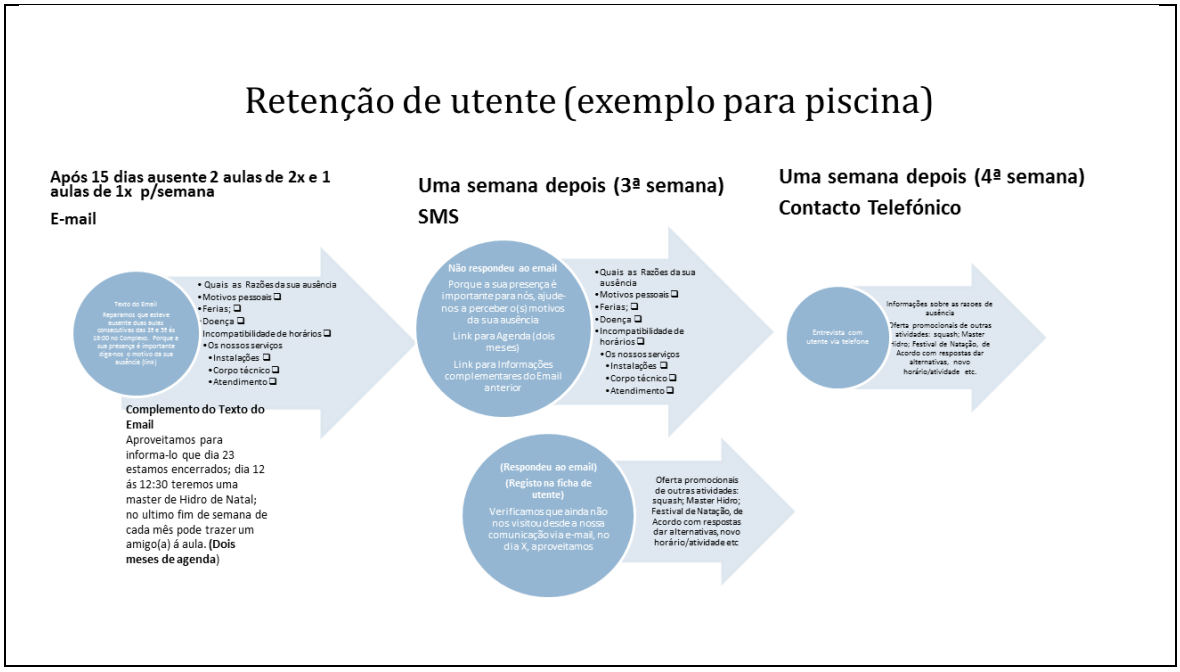
As CTA, quer nos emails, quer nos SMS - simuladas através da utilização de ligações ("*links*") - assumem especial importância para automatizar o processo de "levar" as ações referidas aos utentes. A disponibilização das CTA nos emails e nos SMS permite redirecionar os utentes para uma "*landing page*" onde obtêm a informação ou benefício a que diz respeito a ação. Esta situação permitirá, por um lado, contabilizar o número de utentes que leram o email ou SMS e que o "acionaram" e por outro lado saber quem foram os utentes que o fizeram. Por último permitirá uma automatização no processo de oferta de informação ou de incentivo.

Importa referir no entanto que o modelo apresentado neste trabalho não propõe ações de fidelização concretas, mas apenas a forma como as mesmas devem ser implementadas e que canais de comunicação devem utilizar. As ações de fidelização devem ser construídas e preparadas para serem dirigidas aos perfis identificados pelo modelo preditivo, de acordo

com o entendimento da gestão das instalações e de acordo com os recursos que podem ser disponibilizados para o efeito.

A título de exemplo a Fig. 6-2 apresenta um conjunto de ações proposto pela gestão de uma instalação desportiva na implementação do protótipo que acompanha este trabalho.

Figura 6-2 - Pirâmide de ações de fidelização



6.3. Avaliação do impacto das ações de fidelização

Após a definição do modelo descrito anteriormente, torna-se necessário avaliar a sua eficácia de acordo com as seguintes hipóteses:

***H_0**: Após executar as ações de fidelização, o número de desistentes é o mesmo do que se não tivessem sido efetuadas ações de fidelização;*

***H_1**: Após executar as ações de fidelização, o número de desistentes é menor do que se não tivessem sido efetuadas as ações de fidelização;*

A determinação de qual das hipóteses se verifica, num determinado intervalo de confidência, permitirá concluir se há ou não uma relação causal entre a aplicação das ações aos perfis construídos e a redução do número de desistências.

De acordo com a terminologia indicada no trabalho relacionado propõe-se a validação das hipóteses através do planeamento de experiências na configuração indicada na Tab. 6-4.

Tabela 6-4 - Configuração das experiências

Grupos		e-Mail		SMS		Pessoal	
R_t	O_{1t}	X_1	O_{2t}	X_2	O_{3t}	X_3	O_{4t}
R_c	O_{1c}		O_{2c}		O_{3c}		O_{4t}

As experiências são construídas através da implementação de testes A/B e avaliadas através do método qui-quadrado ("*chi-square*") que permitirá aferir uma conclusão estatística para o problema em questão.

Os testes A/B são usados manipulando uma variável causal e onde se procura determinar o impacto dessa manipulação em dois grupos diferentes de indivíduos, um onde incide a experiência e outro de controlo. Neste caso, os grupos são criados dividindo os utentes alvo das ações destinadas a um perfil em dois grupos de utentes: aqueles sobre quem se faz incidir as ações de fidelização – que constituem o grupo de teste - e outro grupo sobre o qual não serão aplicadas as ações – que constituem o grupo de controlo.

Para a criação dos grupos propõe-se que a sua constituição se faça no início da experiência pela divisão em partes iguais dos utentes que apresentam o perfil definido pela folha da árvore sobre o qual se pretende aplicar o conjunto de ações. Uma vez que os utentes que apresentam o perfil definido pela folha da árvore têm probabilidades de abandono diferentes, propõe-se a ordenação dos mesmos por ordem decrescente dessa probabilidade, alternando-se a sua colocação em cada um dos grupos de teste e controlo até se esgotarem os utentes. Este processo, apesar de não ser puramente aleatório, cria grupos homogéneos e equivalentes em termos de probabilidade de desistência o que permite evitar problemas com a validação no que diz respeito a seleção para a constituição dos grupos e generalização dos resultados da experiência (Smith & Albaum, 2012).

A aplicação do método do qui-quadrado a estes grupos deve ter dois objetivos pelo que deve ser efetuada em dois passos. Num primeiro passo, a aplicação do método deve ser efetuada após cada ação, o que permitirá avaliar o desempenho da ação em concreto. Num segundo passo, a aplicação do método deve ser efetuada após o decorrer de todas as ações para avaliar o modelo como um todo.

A aplicação do método pode ser efetuada pela construção de uma matriz onde se registam os valores observados e esperados do número de utentes que desistiram e não desistiram

no início e após a aplicação de cada ação, ou, no caso da avaliação global, após terem decorrido todas as ações.

A Tab. 6-5 define as formulas a utilizar para a aplicação do método.

Tabela 6-5 - Matriz de suporte ao cálculo do qui-quadrado

	Observados		Total	Esperados	
	O_t	O_c	$T = O_t + O_c$	$E_t = En_t + Ed_t$	$E_c = En_c + Ed_c$
Não desistentes	On_t	On_c	$N = On_t + On_c$	$En_t = (N / T) * O_t$	$En_c = (N / T) * O_c$
Desistentes	Od_t	Od_c	$D = Od_t + Od_c$	$Ed_t = (D / T) * O_t$	$Ed_c = (D / T) * O_c$

Após a construção da tabela, a determinação do qui-quadrado faz-se pela aplicação da seguinte fórmula aos valores On, Od, En e Ed:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Em termos de exemplificação, suponha-se que pretendemos validar a experiência efetuada pela aplicação da primeira ação de fidelização (X_1) a um perfil em que se encontram 300 utentes, pelo que teríamos inicialmente 150 utentes nos grupos de teste e controlo (R_t = 150 utentes e R_c = 150 Utentes).

Supondo que após a ação de fidelização teríamos 45 desistentes no grupo de teste e 70 desistentes no grupo de controlo, a Tab. 6-5 seria preenchida como indicado na Tab. 6-6:

Tabela 6-6 - Exemplo de preenchimento da matriz de cálculo do qui-quadrado

	Observados		Total	Esperados	
	$O_t = 150$	$O_c = 150$	$T = 300$	$E_t = 150$	$E_c = 150$
Não desistentes	$On_t = 105$	$On_c = 80$	$N = 185$	$En_t = 92.5$	$En_c = 92.5$
Desistentes	$Od_t = 45$	$Od_c = 70$	$D = 115$	$Ed_t = 57.5$	$Ed_c = 57.5$

A Tab. 6-7 apresenta o resultado do cálculo das iterações para cada valor de i e o valor final de χ^2.

Com o resultado obtido para $\chi^2 = 7.871$ podemos concluir, através da consulta da distribuição do qui-quadrado e com um nível confiança de 99%, que a ação de fidelização efetuada através do envio de e-mails (X_1) teve o efeito desejado.

Tabela 6-7 - Cálculos parciais e valor final de χ^2

i	Substituição de valores na fórmula	Valor
1	$(105 - 92.5)^2 / 92.5$	1.689
2	$(80 - 92.5)^2 / 92.5$	0.748
3	$(45 - 57.5)^2 / 57.5$	2.717
4	$(70 - 57.5)^2 / 57.5$	2.717
	$\chi^2 =$	7.871

Esta avaliação pode ser repetida para a segunda (X_2) e terceira ação de fidelização (X_3), com o intuito de avaliar individualmente cada ação de fidelização, e também de uma forma global para avaliar o modelo proposto, preenchendo a Tab. 6-5 em função do total de desistentes dos grupos de teste e controlo no final da experiência, ao efetuar a observação O_4 e após ter decorrido o período das experiências ($t_e + t_s + t_p$).

Se o resultado obtido para a avaliação global do modelo apresentar um nível de confiança superior a 95%, H_0 pode ser rejeitada podendo-se concluir com relevância estatística, que após executar as ações de fidelização, o número de desistentes é menor do que se não tivessem sido efetuadas as ações de fidelização.

7. Conclusões

7.1. Conclusões

Com este trabalho procuramos apresentar um contributo para aumentar a fidelização e as taxas de retenção nos serviços desportivos regulares através da construção de um sistema que gera conhecimento acionável ("*actionable knowledge*") baseado em dados reais, que utiliza informação relevante e é acionável.

Em concordância com o modelo de *Database Marketing* (Cavique, 2006) este sistema de comunicação inteligente (SCI) é baseado em três passos:

1) A partir da utilização de dados que podem ser encontrados em bases de dados de suporte a sistemas ERP e CRM de instalações desportivas, foram identificados registos e conjuntos de atributos relevantes que caracterizam o comportamento dos utentes das instalações desportivas;
2) A partir dos atributos relevantes encontrados, construíram-se modelos preditivos baseados em árvores de decisão que permitem num determinado limiar de certeza diferenciar os utentes desistentes dos restantes;
3) As diferenças encontradas permitem segmentar e interagir com os utentes através da implementação de ações de fidelização especificamente dirigidas as características de cada segmento encontrado;

A aplicação dos modelos preditivos aos dados e atributos obtidos assume-se como a principal característica diferenciadora e que introduz "inteligência" no sistema, permitindo traçar o perfil dos utentes desistentes. Os perfis obtidos permitem obter segmentos com as características indicadas por Kotler (Kotler & Keller, 2009) o que permite por um lado, quantificar as ações a levar a cabo, e por outro diferenciá-las de acordo com as características indicadas pelo perfil. Além disto, do método preditivo utilizado, as árvores de decisão, podem resultar vários perfis de pré-desistência acionáveis por diferentes ações de fidelização.

É também de realçar que os modelos preditivos que apresentam melhores resultados atribuem maior relevância aos atributos relacionados com a frequência em detrimento de outros atributos, como o *género*, considerado como um fator determinante noutros trabalhos (Gonçalves, 2012).

Comparativamente a outros trabalhos realizados na área da fidelização, o sistema construído difere na forma de tratar a retenção, uma vez que parte de dados reais que dizem respeito ao comportamento e preferências de todos os utentes da instalação desportiva, ao contrário de outros estudos na área que utilizam inquéritos realizados sobre amostras dos utentes.

As ações de fidelização são levadas aos utentes através de três canais de comunicação (e-mail, SMS e telefonicamente/pessoalmente) dispostos num fluxo piramidal que permitem minimizar o esforço envolvido no aumento da fidelização e consequentemente da taxa de retenção. É proposto a realização de experiências controladas (Testes AB) como forma de avaliação da relação entre as ações de fidelização individuais e no seu conjunto (as causas) e o aumento da taxa de retenção (o efeito).

7.2. Trabalhos em curso

Resta referir que é necessário não esquecer a íntima relação deste trabalho com o Regulamento Geral de Proteção de Dados que entrou em vigor em 25 de Maio de 2018 e que diz respeito a todo o espaço europeu. No enquadramento deste trabalho, este novo regulamento impõe restrições nomeadamente ao tipo de dados que são recolhidos, ao período durante o qual os dados são armazenados, à utilização dos dados para a construção de perfis e à comunicação com o próprio utente. Contudo, as questões colocadas são facilmente ultrapassadas desde que se utilizem processos como a "anomização" e a obtenção de consentimentos por parte dos titulares dos dados.

Para agilizar o funcionamento do sistema foi desenvolvido um protótipo que abrange os três passos do sistema e permite determinar o sucesso das ações de fidelização através do método referido, estando em curso a preparação das ações de fidelização concretas e a sua implementação em duas instalações desportivas.

Referências

AGAP. (2016). Barómetro 2016.

Avourdiadou, S., & Theodorakis, N. D. (2014). The development of loyalty among novice and experienced customers of sport and fitness centres. *Sport Management Review*, *17*(4), 419–431. http://doi.org/10.1016/j.smr.2014.02.001

Cavique, L. (2002). *Meta-heurísticas na Resolução do Problema da Clique Máxima e Aplicação na Determinação do Cabaz de Compras*. Instituto Superior Técnico.

Cavique, L. (2003). Micro-Segmentação de Clientes com Base em Dados de Consumo : Modelo RM-Similis. *Revista Portuguesa e Brasileira de Gestão*, *2*(nº 3), 72–77.

Cavique, L. (2006). Relatório da Unidade Curricular de Database Marketing, 2005-2006. *Escola Superior de Comunicação Social, Instituto Politécnico de Lisboa*, (unpublished).

Cavique, L., Mendes, A. B., & Funk, M. (2011). Logical Analysis of Inconsistent Data (LAID) for a Paremiologic Study.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering*, *40*(1), 16–28. http://doi.org/10.1016/j.compeleceng.2013.11.024

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intl. J. of Intelligent Data Analysis*, *1*(97), 131–156. Retrieved from http://www.machinelearning.net/feature-selection/DashLiu1997.pdf

Do, H. H., & Rahm, E. (2000). Data Cleaning: Problems and Current Approaches. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, *23*(4), 3–13.

El-Sappagh, S. H. A., Hendawi, A. M. A., & El Bastawissy, A. H. (2011). A proposed model for data warehouse ETL processes. *Journal of King Saud University - Computer and Information Sciences*, *23*(2), 91–104. http://doi.org/10.1016/j.jksuci.2011.05.005

Eugenio, B. Di, & Glass, M. (2004). The Kappa Statistic: A Second Look. *Computational Linguistics*, *30*(1), 95–101. http://doi.org/10.1162/089120104773633402

Ferri-Ramírez, C; Flach, P; Hernández-Orallo, J. (2002). Multi-dimensional ROC Analysis with Decision Trees, 1–36.

Frota, M. (2011). Gestão da Retenção. In *Manual de Gestão de Ginásios e Health Clubs - Excelência no sector do Health & Fitness* (pp. 103–148).

Galhardas, H., Florescu, D., Shasha, D., & Simon, E. (2006). An Extensible Framework for Data Cleaning. *Proceedings of the 16th International Conference on Data Engineering (ICDE 2000)*.

Gama, J., Carvalho, A. P. de L., Faceli, K., Lorean, A. C., & Oliveira, M. (2017). *Extração de Conhecimento de Dados*. (E. Silabo, Ed.) (3ª edição).

Gonçalves, C. (2012). Variáveis Internas e Externas ao Indivíduo que influenciam o Comportamento de Retenção de Sócios no Fitness. *PODIUM Sport, Leisure and Tourism Review*, *1*(2), 28–58.

Gorgoglione, M. (2011). Beyond Customer Churn: Generating Personalized Actions to Retain Customers in a Retail Bank by a Recommender System Approach. *Journal of Intelligent Learning Systems and Applications*, *03*(02), 90–102. http://doi.org/10.4236/jilsa.2011.32011

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research (JMLR)*, *3*(3), 1157–1182. http://doi.org/10.1016/j.aca.2011.07.027

Howat, G., & Assaker, G. (2016). Outcome quality in participant sport and recreation service quality models: Empirical results from public aquatic centres in Australia. *Sport Management Review*, *19*(5), 520–535. http://doi.org/10.1016/j.smr.2016.04.002

IBM. (2011). IBM SPSS Modeler CRISP-DM Guide, 53.

Kotler, P., & Keller, K. L. (2009). *Marketing Management*. *Organization* (Vol. 22). http://doi.org/10.1080/08911760903022556

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. http://doi.org/10.2307/2529310

Liu, H., & Yu, L. (2009). Toward Integrating Feature Selection Algorithms for Classification and Clustering.

Mahajan, V., Misra, R., & Mahajan, R. (2015). Review of Data Mining Techniques for Churn Prediction in Telecom. *Journal of Information and Organizational Research*, *39*(2), 183–197.

Marisa, V., & Pedragosa, D. (2006). O Processo de Fidelização no Health Club Bioritmo.

Merisavo, M., & Raulas, M. (2004). The impact of e-mail marketing on brand loyalty. *Journal*

of Product & Brand Management, *13*(7), 498–505. http://doi.org/10.1108/10610420410568435

Microsoft. (2017a). Feature Selection (Data Mining). Retrieved June 20, 2018, from https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/feature-selection-data-mining?view=sql-server-2014

Microsoft. (2017b). Microsoft Decision Trees Algorithm. Retrieved June 20, 2018, from https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/microsoft-decision-trees-algorithm?view=sql-server-2014

Microsoft. (2017c). Microsoft Decision Trees Algorithm Technical Reference. Retrieved June 20, 2018, from https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/microsoft-decision-trees-algorithm-technical-reference?view=sql-server-2014

Muñoz, L., Mazón, J.-N., & Trujillo, J. (2010). Systematic review and comparison of modeling ETL processes in Data Warehouse. *5th Iberian Conference on Information Systems and Technologies, CISTI 2010*. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-77957854091&partnerID=40&md5=0284c9578a4b21847767b1083fc8aa7d

Parodi, S., Muselli, M., Carlini, B., Fontana, V., Haupt, R., Pistoia, V., & Corrias, M. V. (2012). Restricted ROC curves are useful tools to evaluate the performance of tumour markers. *Statistical Methods in Medical Research*. http://doi.org/10.1177/0962280212452199

Peppers, D., & Rogers, M. (2004). *Managing Customer Relationships*. *RESEARCH ANALYSIS AND EVALUATION International Indexed & Referred Research Journal* (Vol. 30097). Wiley.

Pinheiro, P., & Cavique, L. (2015). Determinação de Padrões de Desistência em Ginásios. *Revista de Ciências Da Computação*, *10*, 33–60. Retrieved from http://lead.uab.pt/OJS/index.php/RCC/article/view/97/72

Pontius, R. G., Millones, M., Pontius, Robert, Gilmore, J., Millones, M., Pontius, R. G., & Millones, M. (2011). Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*,

32(15), 4407–4429. http://doi.org/10.1080/01431161.2011.552923

Pousttchi, K., & Wiedemann, D. (2006). A contribution to theory building for mobile marketing: Categorizing mobile marketing campaigns through case study research. *Mobile Business, 2006. ICMB'06. …*, (2925). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4124096

ResultadosDigitais. (2017). 21 dicas de Email Marketing para alavancar suas campanhas. Retrieved June 20, 2018, from https://materiais.resultadosdigitais.com.br/dicas-email-marketing

Siegel, E. (2013). *PREDICTIVE ANALYTICS The power to predict who will click, buy, lie or die*. Wiley.

Smith, S. M., & Albaum, G. S. (2010). *An Introduction to Marketing Research*. http://doi.org/10.1016/S1529-1839(04)70026-3

Smith, S. M., & Albaum, G. S. (2012). *Basic Marketing Research: Volume 1 Handbook for Research Professionals* (Vol. 1).

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427–437. http://doi.org/10.1016/j.ipm.2009.03.002

Surujlal, J., & Dhurup, M. (2012). Establishing and maintaining customer relationships in commercial health and fitness centres in South Africa. *International Journal of Trade, Economics and Finance*, *3*(1), 14–18. http://doi.org/10.7763/IJTEF.2012.V3.165

Tan, P.-N., Steinbach, M., & Kumar, V. (2006a). *Classification : Basic Concepts , Decision Trees , and*. *Introduction to Data Mining* (Vol. 67). http://doi.org/10.1016/0022-4405(81)90007-8

Tan, P.-N., Steinbach, M., & Kumar, V. (2006b). *Introduction to Data Mining*. Pearson.

The Mobile Experience Company Inc. (2018). 7 Best Practices for SMS Marketing Every Agency Needs To Know. Retrieved from https://www.mobilexco.com/blog/7-best-practices-for-sms-marketing-every-agency-needs-to-know

Trujillo, J., & Luján-Mora, S. (2003). A UML based approach for modeling ETL processes in data warehouses. *Conceptual Modeling-ER 2003*, *2813*, 307–320. http://doi.org/ETL processes, Data warehouses, conceptual modeling, UML

Vassiliadis, P., & Simitsis, A. (2009). *EXTRACTION, TRANSFORMATION, AND LOADING*. *Encyclopedia of Database Systems*. Springer US.

Vassiliadis, P., Simitsis, A., Georgantas, P., Terrovitis, M., & Skiadopoulos, S. (2005). A generic and customizable framework for the design of ETL scenarios. *Information Systems*, *30*(7), 492–525. http://doi.org/10.1016/j.is.2004.11.002

Yu, L., & Liu, H. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. *International Conference on Machine Learning (ICML)*, 1–8. http://doi.org/citeulike-article-id:3398512

Zikmund, William; Ward, Steven; Lowe, Ben; Winzar, Hume; Babin, B. (2011). *Marketing Research*. *Business* (Second Asi). Retrieved from http://books.google.com/books?hl=en&lr=&id=tCspQP0CYgcC&oi=fnd&pg=PR7&dq=Strategy+process,+cintent,+context+an+international+perspective&ots=-gMZvHbLMt&sig=vGe72LHzg4hofsDcvBGeWHr88RI