Universidade do Minho
Escola de Engenharia

Sérgio Manuel da Costa Oliveira

Modelos de Previsão em Gestão Hospitalar recorrendo a técnicas de Data Mining

outubro de 2013

Sérgio Manuel da Costa Oliveira **Modelos de Previsão em Gestão Hospitalar recorrendo a técnicas de Data Mining**

UMinho | 2013

Universidade do Minho
Escola de Engenharia

Sérgio Manuel da Costa Oliveira

Modelos de Previsão em Gestão Hospitalar recorrendo a técnicas de Data Mining

Dissertação de Mestrado
Mestrado em Engenharia e Gestão de Sistemas de Informação

Trabalho efetuado sob a orientação do
Professor Doutor Manuel Filipe Vieira Torres dos Santos

e coorientação do
Mestre Carlos Filipe Portela

outubro de 2013

DECLARAÇÃO

Nome: Sérgio Manuel da Costa Oliveira

Endereço electrónico: pg20671@alunos.uminho.pt

Título dissertação

Modelos de Previsão em Gestão Hospitalar Recorrendo a Técnicas de Data Mining

Orientadores: Professor Doutor Manuel Filipe Vieira Torres dos Santos

Coorientador: Mestre Carlos Filipe Portela

Ano de conclusão: 2013

Designação do Mestrado: Mestrado em Engenharia e Gestão de Sistemas de Informação

Universidade do Minho, ___ /___ /_____

Assinatura: _____

Agradecimentos

Aos meus país, Mário Costa e Maria Oliveira pelo incentivo durante o desenvolvimento desta dissertação e por proporcionar todo o meu percurso académico.

Ao orientador Manuel Filipe Santos e coorientador Filipe Portela pela presença, disponibilidade e orientação ao longo de todo o percurso do projeto. A presença dos orientadores foi fundamental para a realização desta dissertação.

À minha irmã, Daniela Oliveira por toda a preocupação e incentivo.

Por último a todos os meus amigos que estiveram presentes ao longo do meu percurso académico.

Resumo

É notório que as falhas verificadas na gestão hospitalar estão normalmente relacionadas com a falta de informação e a insuficiente gestão de recursos. Estes aspetos são determinantes para a gestão de qualquer entidade organizacional. Foi a partir deste princípio que se abordou o processo de *Data Mining* (DM) neste projeto, com o intuito de identificar dados pertinentes sobre a gestão de doentes e assim proporcionar aos gestores do Centro Hospitalar do Porto (CHP) informações importantes para fundamentar as suas decisões.

Durante a realização desta Dissertação, foram desenvolvidos modelos de DM capazes de realizar previsões em âmbito hospitalar (gestão de altas). O desenvolvimento dos modelos de previsão foi realizado em ambiente real, com dados reais oriundos do CHP. Para isso foi adotada metodologia de investigação *Action Research,* o mesmo foi orientado segundo a metodologia *Cross-Industry Standard Process for Data Mining* (CRISP-DM).

Ao nível do DM foram usadas as técnicas baseadas em Árvores de Decisão, Árvores de Regressão (AR), *Naïve Bayes* e *Support Vector Machine* (SVM) para realizar as tarefas de Classificação e Regressão.

A avaliação e validação dos modelos de Classificação foi efectuada através da utilização da métrica baseada na acuidade. Para os modelos de Regressão foram usadas várias métricas, *Mean Squared Error*, *Mean Absolute Error*, *Relative Absolute Error* e *Regression Error Characteristic.* Para além destas métricas foram ainda usadas as técnicas *Cross Validation* e *Leave-One-Out Cross Validation* para avaliar a capacidade de generalização dos modelos de previsão.

Os modelos de Classificação foram capazes de prever altas de doentes com valores de acuidade compreendidos entre ≈82.69% e ≈94.23%. Alguns dos modelos de Regressão obtiveram um desempenho similar ou inferior ao previsor médio *naïve*, resultados no geral compreendidos entre ≈38.26% e ≈94,89%. Os resultados obtidos permitem suportar decisões ao nível da gestão de altas. Com este trabalho foi também possível concluir que os modelos de Classificação apresentam resultados menos satisfatórios para os serviços de Ortopedia e Obstetrícia e os modelos de Regressão para o serviço de Parto. Porém a Classificação proporcionou bons modelos de previsão para o serviço de Parto e Berçário, e a Regressão para os serviços de Ortopedia, Obstetrícia e Berçário.

Abstract

The hospitals mismanagement is associated with the lack of information and poor management of resources. These aspects are crucial for the management of any organizational entity. It is on this principle that the Data Mining (DM) process was addressed in this project, to identify relevant information about the management of patients and thus provide to the managers of Centro Hospitalar of Porto (CHP) important information to help in their decisions.

While performing this dissertation, several DM models were developed to predict hospital discharge. The development of the predictive models was conducted in a real environment with real data. This project was conducted using the *Action Research* research methodology and the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology.

From the DM techniques, Decision Trees, Naïve Bayes and Support Vector Machine were used to induce Classification and Regression models.

The evaluation and validation of the Classification models was done through the acuity obtained in the results. For Regression models several metrics were used, namely: Mean Squared Error, Mean Absolute Error, Relative Absolute Error and Regression Error Characteristic. In addition to these metrics it was used the Cross Validation and Leave-One-Out Cross Validation techniques to evaluate generalization capacity of the models.

The classification models were able to predict the patient discharges with acuity values ranging from ≈82.69% to ≈94.23%. The regression models achieved a performance similar to or lower than the average naïve prediction, being comprehended between ≈38.26% and ≈94.89%. The results are able to support management decisions, when it comes to patients discharge management, however Classification models for the Orthopedics and Obstetrics services and regression models for Childbirth service presented less satisfactory results. However, Classification provided good predictive models for service Childbirth and Nursery, and Regression to the services of Orthopedics, Obstetrics and Nursery.

Índice de Conteúdos

Índice de Figuras

Índice de Tabelas

Acrónimos

10-folds CV	10-folds Cross Validation
AD	Árvore de Decisão
ACR	Action Research
AR	Árvore de Regressão
BI	Business Intelligence
CHAID	Chi-Square Automatic Interaction Detector
CHP	Centro Hospitalar do Porto
CRISP-DM	Cross-Industry Standard Process of Data Mining
CV	Cross Validation
DM	Data Mining
FN	False Negative
FP	False Positive
GI	Gini Index
IG	Information Gain
LOOCV	Leave-One-Out Cross Validation
MAE	Mean Absolute Error
MC	Matriz de Confusão
MSE	Mean Squared Error
NB	Naïve Bayes
OLAP	On-Line Analytical Processing
PMML	Predictive Model Markup Language
RAE	Relative Absolute Error
REC	Regression Error Characteristic
ROC	Receiver Operating Characteristic
SAD	Sistema de Apoio à Decisão
SEMMA	Sample, Explore, Modify, Model Assessment
SIH	Sistema de Informação Hospitalar
SVM	Support Vector Machine
TN	True Negative
TP	True Positive

1. Introdução

Este capítulo tem como principal propósito identificar a problemática do projeto e determinar a questão de investigação. Em função da questão de investigação foram, propostos vários objetivos que se pretendem alcançar até à conclusão do projeto.

É apresentada a abordagem metodológica que suportará e orientará o processo de conceção do projeto, e por último é apresentada a estrutura do documento.

1.1. Motivação

As organizações acumulam dados de vários processos de negócio. Tendo como base a ideia de que, os grandes volumes de dados podem ser fonte de conhecimento implícita, um novo conhecimento pode ser extraído a partir da utilização de ferramentas adequadas (e.g. *Data Mining*) (Manuel Santos & Azevedo, 2005). Os hospitais operam segundo este princípio, uma vez que as suas bases de dados podem conter conhecimento oculto e importante, por exemplo, no estado clínico de um doente e no seu prognóstico (Manuel Santos, Boa, Portela, Silva, & Rua, 2010).

As técnicas de *Data Mining* (DM) têm sido empregues com sucesso em aplicações destinadas à deteção de fraude (várias áreas), previsão de falência organizacional, na tomada de decisão estratégica, comercialização de base de dados e no desempenho financeiro das organizações (Ozgulbas & Koyuncugil, 2007). O DM também tem auxiliado as organizações na otimização e na alocação de recursos internos, como ainda tem proporcionado uma melhor resposta e entendimento das necessidades dos seus clientes (Nemati & Barko, 2010).

As aplicações de DM no sector da saúde podem ser uma mais-valia. Deste modo, o DM pode ajudar as seguradoras de saúde a detetar fraude ou abuso na tomada de decisão face ao relacionamento com os seu clientes, os médicos podem identificar tratamentos mais eficazes, bem como melhores práticas e os doentes podem receber melhores serviços de saúde. O DM fornece a metodologia e a tecnologia para transformar enormes volumes de dados em informação útil para a tomada de decisão (Koh & Tan, 2005).

Foi a partir deste princípio que se determinou a questão de investigação:

"***É possível construir modelos de DM capazes de suportar o processo de gestão hospitalar, ao nível da gestão de camas, tendo por base o número de altas?***".

Neste projeto serão utilizadas as seguintes técnicas: Árvores de Decisão (AD), Árvores de Regressão (AR), *Naïve Bayes* (NB) e *Support Vector Machine* (SVM), para efetuar previsões seguindo duas abordagens, Classificação e Regressão. As previsões incidem sobre acontecimentos de gestão de doentes, de forma a suportar a tomada de decisão dos gestores em ambiente real.

1.2. Objetivos e Resultados a Atingir

Este trabalho de investigação é importante na medida em que, pretende aferir previsões de acontecimentos de gestão hospitalar com o objetivo de suportar as decisões dos gestores hospitalares. Os objetivos definidos tem como principal propósito responder à questão de investigação definida.

Os objetivos propostos para a realização deste projeto são:

- Aferir previsões capaz de contribuir para a gestão de camas;
- Identificar padrões nas altas hospitalares;
- Obter modelos com boa capacidade de previsão.

Como último objetivo espera-se que, os modelos desenvolvidos possam ser integrados, futuramente, num Sistema de Apoio á Decisão de forma a que este opere em ambiente real. É expectável que o sistema dê respostas em curto espaço de tempo e que seja capaz de reduzir os gastos hospitalares, através de uma melhor gestão de camas.

De forma a atingir os vários objetivos propostos será necessário recorrer a uma fonte de dados. Os dados usados para o desenvolvimento de modelos de DM serão referentes ao número de altas hospitalares, Data, Serviço e Número de Altas, estas serão as variáveis manuseadas.

Os vários modelos a desenvolver serão suportados por técnicas de DM, tais como, AD, AR, NB e SVM.

Quanto aos resultados esperados, pretende-se que no final deste trabalho se tenha adquirido uma perceção apurada da utilidade dos modelos de DM na descoberta de conhecimento em ambiente hospitalar. Espera-se que essa perceção seja resultante do uso, com sucesso e utilidade, das técnicas de DM perante os dados facultados para a realização deste projeto.

1.3. Abordagem Metodológica

Este trabalho de investigação assenta particularmente em aspetos de melhoria na gestão hospitalar. Este fato remete para a necessidade de usar uma metodologia capaz de acompanhar o processo de investigação.

Koshy defende que a metodologia *Action Research* (ACR) é importante em projetos de saúde. Trata-se de um recurso valioso na gestão efetiva de mudança, ou seja, a ACR apoia profissionais e investigadores na procura de práticas capazes de melhorar a qualidade nos serviços de saúde (Koshy, Koshy, & Waterman, 2011).

O uso da metodologia ACR irá possibilitar ao investigador identificar soluções de mudança no contexto em que se insere o projeto, através da observação e da comunicação com pessoas e profissionais do sector da saúde. Durante o desenvolvimento do projeto, o investigador deverá desenvolver e usar um conjunto de ferramentas que possibilite atingir os objetivos, desenvolver um plano, efetuar observações adequadas e realizar avaliações e reflexões críticas. Estes aspetos são importantes para atingir os objetivos propostos (Koshy et al., 2011).

Meyer afirma que a força da metodologia ACR está no foco que ela transmite em gerar soluções de problemas práticos e na sua capacidade em habilitar os investigadores, fazendo com que eles se envolvam na investigação e nas atividades subsequentes de desenvolvimento e de implementação (J. Meyer, 2000). A interação entre Sistemas de Informação e os profissionais do sector da saúde cria um ambiente rico na aplicação do ACR (Kohli & Hoadley, 2007).

A metodologia é composta por cinco fases, trata-se de um processo cíclico, que deve ser implementado dentro de uma entidade/organização em ambiente de pesquisa. Este processo foi definido por Susman e Evered em 1978 (Baskerville, 2007). A Figura 1 ilustra as suas cinco fases.

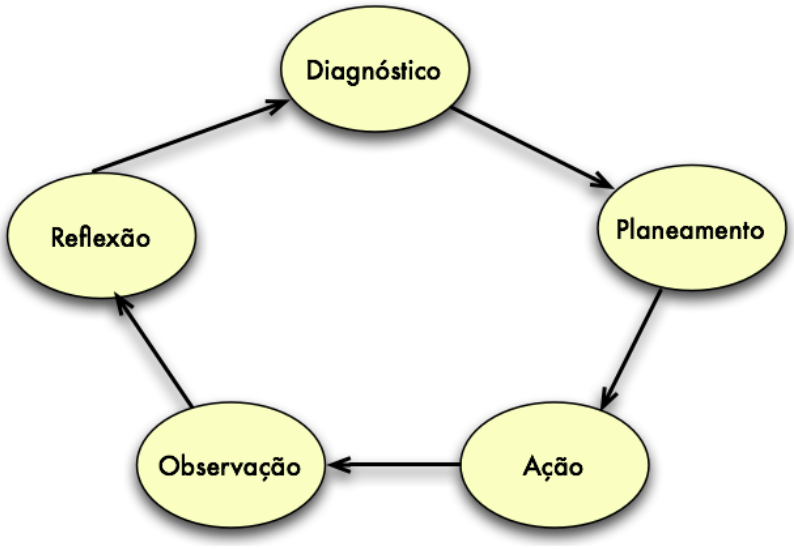

Figura 1 – Ciclo da ACR (adaptado de Baskerville, 2007)

Diagnóstico – Inicialmente será necessário identificar implementações com foco na gestão hospitalar, mais concretamente na gestão de doentes. Está atividade é importante, na medida que, fomentará uma linha orientadora para o desenvolvimento deste projeto. A identificação de implementações reais com foco na gestão hospitalar ou semelhantes serão necessárias para realizar o enquadramento do problema e determinar a situação atual desta temática.

Planeamento – Inicialmente será realizada a revisão de literatura, para adquirir conhecimento dos conceitos que estão diretamente ligados com o tema do projeto de investigação, conceito de *Data Mining* (DM) e técnicas de Previsão. Por último, nesta fase será definida a questão de investigação, bem como os objetivos e resultados que o artefacto deverá atingir.

Esta fase da ACR auxiliará o início do desenvolvimento prático deste projeto, mais concretamente na primeira fase da metodologia *Cross-Industry Standar Process for Data Mining* (CRISP-DM), o Estudo do Negócio, onde será necessário realizar a planificação da respetiva parte prática.

Ação – Numa terceira fase, será necessário compreender os dados fornecidos e posteriormente conceber modelos de previsão através de técnicas de DM, o processo de inserção de dados e de construção de modelos. Esta fase do ACR estará diretamente ligada a três fases do CRISP-DM, Compreensão dos Dados, Preparação dos Dados e Modelação.

Observação – Nesta fase serão analisados os resultados dos modelos desenvolvidos para a gestão de altas, as métricas de avaliação dos modelos o quanto significativos eles podem ser na previsão de altas. A respetiva fase do ACR estará em consonância com a fase de Avaliação do CRISP-DM.

Reflexão – Esta fase ditará a validação do trabalho desenvolvido em cada uma das fases do CRISP-DM. Quer isto dizer, que depois de concluída cada uma das fases do CRISP-DM será da máxima importância determinar se é viável proceder para a fase seguinte. Depois de alcançada e concluída a última fase do CRISP-DM será necessário perceber o quanto importante será para o hospital implementar e seguir o trabalho desenvolvido.

A metodologia de investigação ACR estará presente em cada uma das etapas do trabalho realizado e o seu ciclo de vida foi repetido várias vezes. A metodologia ACR combinada com a metodologia CRISP-DM permite garantir a viabilidade da execução deste trabalho, bem como o seu sucesso a nível científico. No entanto, devido ao elevado número de iterações, não foi possível documentar detalhadamente cada uma delas.

1.4. Estrutura do Documento

A estrutura do documento está dividida em cinco capítulos. O primeiro capítulo, Introdução, descreve a necessidade de responder á questão de investigação, apresenta os objetivos propostos e os resultados que se esperam obter no final do projeto. Por último, no primeiro capítulo é apresentada a metodologia que será seguida para o desenvolvimento da Dissertação (*Action Research*).

No segundo capítulo, Enquadramento Conceptual, são apresentados casos reais de aplicações DM em ambiente hospitalar, neste capítulo são ainda descritos os conceitos que suportam a realização do projeto, tais como, Gestão Hospitalar, *Business Intelligence* (BI), DM e a metodologia CRISP-DM. Ainda no segundo capítulo, são apresentadas as métricas de avaliação dos modelos de Classificação e Regressão. São também apresentadas várias técnicas de DM, as ADs, NBs e os SVMs.

No terceiro capítulo, Trabalho Realizado, é apresentado o desenvolvimento da componente prática, orientado pela metodologia CRISP-DM.

No penúltimo capítulo, Discussão de Resultados, são analisados e discutidos os aspetos mais pertinentes do trabalho realizado.

Por último, no capítulo das Conclusões, é feita uma síntese do trabalho e são também apresentadas as contribuições e recomendações para trabalhos futuros.

2. Enquadramento Conceptual

A revisão de literatura é fundamental para qualquer projeto académico, através da realização de uma revisão eficaz é possível e espectável que o investigador crie uma base sólida na obtenção de conhecimento (Brocke et al., 2009; Webster & Watson, 2002). Seguindo este princípio, foram identificadas algumas das implementações de melhoria na gestão hospitalar, implementações essas que recorreram a técnicas de *Data Mining* (DM). O estudo de implementações de DM em gestão hospitalar prende-se pela necessidade de identificar o que foi feito a nível científico no âmbito do tema da dissertação.

Foram também abordados os conceitos que estão diretamente relacionados com o tema da dissertação, Gestão de Saúde, Gestão Hospitalar, *Business Intelligence* (BI), DM, *Cross-Industry Standar for Data Mining* (CRISP-DM) - metodologia que suportará o processo de desenvolvimento de modelos de previsão, técnicas de DM e respetivas métricas de avaliação. A identificação dos vários conceitos foi importante, pois foi necessário compreender a área de investigação em que este projeto esteve inserido.

2.1. Gestão Hospitalar e os Sistemas de Informação

O hospital é uma empresa onde se devem aplicar os mesmos critérios e princípios de funcionamento e avaliação que se aplicam a uma qualquer empresa de um determinado setor de atividade. Uma empresa é uma organização produtiva de bens ou serviços que visa a sua manutenção e o seu desenvolvimento, isto é, o sucesso. Porém, o hospital é certamente uma empresa que reúne uma série de características específicas que merecem atenção quando analisados os aspetos relacionados com a gestão. Genericamente, o hospital é uma organização extremamente complexa por natureza. É-o por ser uma empresa produtora de serviços altamente diferenciados em que é exigido uma mão-de-obra intensiva e igualmente diferenciada (Proença et al., 2000).

A World Health Organization descreve o hospital como sendo, "... parte integrante de uma organização médica e social, cuja função é fornecer um serviço de saúde completo e de atendimento à população, tanto curativo como preventivo, e cujos serviços de ambulatório

devem chegar às famílias em ambiente doméstico, o hospital também é um centro de formação de trabalhadores de saúde e de investigação biossocial. '' (WHO, 1963).

Os hospitais são organizações burocráticas e como tal seguem o princípio das organizações burocráticas, sendo efetivamente o agrupamento de cargos e o aglomerado de posições individuais organizados de forma hierárquica. Os hospitais são caracterizados como sendo uma organização composta por um sistemas de regras. Essas regras são os limites oficiais para a execução de ações dentro de um hospital (Griffin, 2006).

Os hospitais tal como outras organizações necessitam de uma estrutura organizacional. A mais popular e tradicional estrutura é a pirâmide hierárquica. Segundo esta estrutura todos aqueles que se encontram em cargos perto do topo da pirâmide (e.g. chefes de departamento) possuem um nível específico de autoridade, em que este é transmitido aos colaboradores que se encontram em níveis de poder hierarquicamente inferiores. Deste modo, a autoridade está dispersa pelo organismo hospitalar. Na estrutura em pirâmide os supervisores delegam aos subordinados que, por sua vez, delegam aos seus subordinados (Griffin, 2006).

Para gerir um hospital é necessário que o gestor conheça profundamente a instituição em que trabalha, tenha conhecimento das normas e rotinas, dos serviços que o hospital presta, e que seja capaz de identificar os pontos fortes e aqueles que precisam ser melhorados. Será a partir destes pontos que o gestor deverá traçar um plano claro e organizado de forma a proporcionar uma gestão hospitalar eficiente e eficaz (I. Santos & Arruda, 2012).

A gestão é composta por um conjunto variável de instrumentos técnicos e de tecnológias utilizadas para garantir o sucesso das organizações. Nesse sentido, o que determina uma boa ou má gestão, não é qualidade dos instrumentos e tecnologias usadas mas, o cumprimento da missão da organização em causa e a prossecução dos objetivos institucionais, as contingências das circunstâncias que constituem o ambiente em que se move a organização e os princípios enquadradores e os pressupostos da sua atuação. Desta feita, não existe, em absoluto, nenhum paradigma incontestável de gestão (Proença et al., 2000).

Numa estrutura organizacional como o hospital, em que o "*core business*" é a otimização da passagem de doentes pelo hospital, minimizando o tempo de estadia e maximizando qualidade do tratamento (Neves, 2011). Nesse sentido um dos recursos mais importantes são as camas que se encontram distribuídas pelos vários serviços hospitalares. A sua gestão reflete a eficiência e a qualidade da gestão hospitalar (Yang, Sun, & Lin, 2010).

O Alto Comissariado da Saúde apresentou um estudo, compreendido entre 2004 e 2010 (ACS, 2010), em que este demostra o aumento da procura de serviços hospitalares. Este fenómeno reflete-se na disponibilidade de camas nos hospitais. Por este motivo é imperativo identificar e determinar medidas que permitam atingir maior eficácia na utilização de camas hospitalares (Yang et al., 2010).

Os gestores hospitalares necessitam de informação sobre os doentes e profissionais (médicos, enfermeiros, etc.), para tomarem decisões (Silva & Beuren, 2012). Estes têm que se ir adaptando às mudanças do ambiente de modo a que as organizações alcancem o sucesso. Isto prende-se com o facto de que a economia global tem emergido com o passar do tempo, os recursos humanos, as tecnologias de informação, a rápida tomada de decisão e o conhecimento de equipas multidisciplinares são elementos que podem criar oportunidades competitivas para qualquer organização (Lameirão, 2007). O gestor não se deve esquecer da dimensão financeira no processo de tomada de decisão, se este aspeto for menosprezado, o hospital corre o risco de falência e consequentemente a incapacidade de atender às necessidades de saúde dos doentes (Neto, 2011).

Na gestão hospitalar, bem como em todas as outras áreas da gestão, os processos organizacionais devem ser geridos com intenção de implementar projetos de qualidade com especial foco na produtividade organizacional, quer em instituições públicas da área da saúde, quer em instituições privadas (Roberto & Lira, 2010).

As organizações de saúde necessitam de uma gestão rigorosa e uma estratégia capaz de alocar os recursos disponíveis com base em prioridades (optar por um caminho em detrimento de outro, escolhendo a sua aplicação com base em determinadas características). Tendo como base a complexidade e a multidisciplinaridade dos serviços de um hospital, torna-se evidente o desenvolvimento de um modelo de informação integrado. Este modelo deve estar alinhado com a estratégia de gestão e assim, permitir apoiar as atividades hospitalares (Lameirão, 2007).

Os hospitais iniciaram a aquisição de recursos tecnológicos em meados dos anos 50, desde então, iniciou-se rapidamente a sua inclusão na maioria dos hospitais. Embora tenha havido um grande progresso na aquisição e uso de Sistemas de Informação de Gestão, os sistemas mais sofisticados são usados na área financeira dos hospitais. Porém, além da aplicação de sistemas tecnológicos na área financeira, os hospitais estão a usar cada vez mais os Sistemas de Informação para o planeamento de decisões em sectores destintos, como o marketing, o atendimento, as enfermarias, os laboratórios e a administração (Griffin, 2006).

Os Sistemas de Informação Hospitalares (SIH) são amplamente adotados como instrumentos valiosos indispensáveis no domínio do sector de saúde. Eles não são usados apenas para integrar informações dos doentes, mas também são usados para melhorar a eficiência e a qualidade dos cuidados prestados pelo hospital (Müller et al., 2007). A tecnologia tem influenciado fortemente a forma de trabalhar e está a proporcionar novas oportunidade para uma série de diferentes abordagens no sector da saúde (Dwivedi, Bali, & Naguib, 2006).

A necessidade de aplicar novos métodos de gestão hospitalar tem demostrado ser da máxima importância nos últimos anos. As técnicas e as ferramentas de gestão que tem vindo a ser aplicadas têm como propósito a otimização dos processos hospitalares (Lameirão, 2007). Por exemplo, a introdução do BI nos hospitais contribuirá na transformação do conhecimento de negócio em vantagens competitivas e será fundamental para a partilha de conhecimento e para a redução de custos administrativos, assim como permitirá melhorar a qualidade dos cuidados prestados pelo hospital (Bose, 2003).

A inserção de novas tecnologias em hospitais tem proporcionado a oportunidade de melhorias de trabalho, as tarefas são executadas mais rapidamente, com maior consistência e com menor custo (Dwivedi et al., 2006).

2.2. Data Mining no Âmbito Hospitalar

A utilização do DM em especial na área da saúde tem vindo a crescer e destingue-se dos métodos de análise tradicionais porque é caracterizado como uma área que aborda um conjunto de conceitos e técnicas interdisciplinares (Dua & Du, 2011). O DM é descrito como um processo que utiliza técnicas que recorrem à inteligência artificial, estatística e matemática com o objetivo de extrair e identificar informações e conhecimento útil para a tomada de decisão nas organizações (Turban, Sharda, & Delen, 2011).

O DM começou a ser aplicado nas bases de dados hospitalares a partir de 1990 bem como alguns métodos clássicos da estatística (Bose, 2003; S. Tsumoto, Hirano, & Tsumoto, 2011; Shusaku Tsumoto & Hirano, 2009). Desde então, o DM está-se a tornar cada vez mais popular se não, cada vez mais essencial na área da saúde (Koh & Tan, 2005).

A procura da qualidade no sector hospitalar remete para a necessidade de explorar todo o potencial dos dados armazenados de forma eficiente. A análise não dever recair apenas sobre dados clínicos, mas também sobre dados administrativos, de modo a que seja possível melhorar

os diagnósticos e tratamentos, com isto, pretende-se minimizar os custos e melhor o atendimento aos doentes. Neste sentido, o DM pode contribuir positivamente para o sector da saúde, como uma ferramenta capaz de analisar os dados recolhidos do SIH e assim obter modelos e padrões que podem melhorar a assistência aos doentes e ainda proporcionar uma melhor utilização dos recursos farmacêuticos (Alapont et al., 2005).

O DM pode melhorar a tomada de decisão, pois é capaz de descobrir padrões e tendências em grandes e complexas base de dados. O conhecimento adquirido a partir do DM pode influenciar o custo, receitas e eficiência operacional, de forma a facultar o alto nível de prestação de serviços de saúde (Koh & Tan, 2005).

Espera-se que a reutilização dos dados do SIH forneça a compreensão de todas as características do hospital, de forma a que seja possível adquirir conhecimento objetivo sobre a forma como o hospital deve ser gerido, bem como que tipo de cuidados clínicos devem ser fornecidos pelo o hospital (Shusaku Tsumoto & Hirano, 2009).

No Centro Médico de Seton na Califórnia, o DM é usado para diminuir o tempo de espera do doente, evitar complicações clínicas, desenvolver boas práticas na prestação dos serviços e fornecer informações adequadas dos doentes aos médicos. Os profissionais do Centro Médico de Seton vêm o DM como um recurso para manter e melhorar a qualidade do serviço médico (Koh & Tan, 2005).

A Blue Cross, federação Norte Americana de 38 organizações de saúde, fornece seguros de saúde e tem vindo a implementar iniciativas de DM para melhorar os resultados e reduzir gastos através de uma melhor gestão das doenças dos seus clientes. A federação Blue Cross usa os dados dos doentes, tais como, pedidos de internamento, registos farmacêuticos e registos médicos para compreender e identificar o custo que o doente terá e qual o impacto do seu plano de saúde (Koh & Tan, 2005).

A organização Sierra Health Services, tem como principal atividade fornecer e administrar programas de saúde. A necessidade de recorrer ao DM foi inerente para: identificar áreas que devem sofrer ações de melhoria qualitativa, identificar as melhores diretrizes em tratamentos de saúde, controlar as doenças e perceber o custo de gestão para a organização (Koh & Tan, 2005).

2.2.1. Sistema de Gestão do Chiba University Hospital

No hospital Chiba University Hospital, hospital Japonês, foi realizada uma pesquisa com objetivo de perceber que tipo de conhecimento pode ser extraído a partir das base de dados do seu Sistema de Informação (Shusaku Tsumoto & Hirano, 2009).

As receitas dos hospitais japoneses são baseadas no número de pontos por assistência médica, nesse sentido, é importante investigar o facto que determina a quantidade de pontos do hospital Chiba. Para isso, foi necessário efetuar uma análise de estatística descritiva, uma análise exploratória de dados e testes estatísticos. Ambas as análises foram efetuadas sobre os dados das altas hospitalares. As análises relativas aos doentes foram obtidas a partir de, informações básicas (sexo, idade e ocupação), do número de dias que o doente esteve no hospital, das suas doenças e das tendências do seu estado clínico. Durante um período de três anos (1997, 2000) foram realizadas análises dos dados contabilísticos do hospital, bem como a relação entre os pontos de assistência médica e os itens de alta médica, além disso foram ainda efetuadas análises exploratórias, testes estatísticos, análise de regressão e modelos lineares (Shusaku Tsumoto & Hirano, 2009).

No estudo verificou-se que existe uma grande correlação entre os pontos por assistência médica e o tempo de estadia do doente, quanto mais prolongada for a estadia do doente mais caro o doente se torna. No estudo verificou-se que os coeficientes de correlação são de 0,837 e 0,867, no intervalo compreendido entre [-1, 1], ou seja a correlação é muito forte (Shusaku Tsumoto & Hirano, 2009).

Os resultados obtidos no estudo demonstraram que há uma forte possibilidade em combinar os resultados das altas médicas com o sistema contabilístico. Este aspeto resulta numa ferramenta básica na análise de gestão hospitalar. O estudo realizado foi útil para a gestão hospitalar do Chiba Hospital, sendo importante no suporte do processo de decisão (Shusaku Tsumoto & Hirano, 2009).

2.2.2. Reforma Financeira nos Hospitais Públicos na Turquia

Em 2004 cerca de 645 hospitais públicos administrados pelo Ministério da Saúde da Turquia foram submetidos a uma reforma financeira, esse programa de reforma foi intitulado como "Saúde em Transição". Os hospitais que pertenciam à reforma necessitavam de ações de melhoria nas suas estruturas financeira e de gestão. Nesse sentido, foram realizadas análises

minuciosas e confiáveis da situação que estava diretamente ligada com os perfis financeiros e com o desempenho dos hospitais turcos. Com isto, foi possível implementar um projeto eficaz no sector da saúde (Ozgulbas & Koyuncugil, 2007).

O estudo tinha, como propósito determinar os perfis financeiros dos hospitais submetidos à reforma e como objetivo determinar as características financeiras que serviriam como suporte às sugestões de gestão e melhorarias de desempenho (Ozgulbas & Koyuncugil, 2007).

O estudo concluiu que os hospitais públicos podiam ser classificados em 12 perfis distintos, a identificação desses perfis foi conseguida através do *Chi-Square Automatic Interaction Detector* (CHAID). CHAID é um algoritmo pertencente ás Árvores de Decisão e é um dos métodos mais eficientes quando usado para efetuar segmentação de dados (Ozgulbas & Koyuncugil, 2007).

A aplicação do processo de DM neste caso de estudo revelou-se como sendo uma ferramenta útil pois permitiu melhorar o desempenho financeiro dos hospitais públicos da Turquia (Ozgulbas & Koyuncugil, 2007).

2.2.3. Gestão Hospital Através do Processo de Data Mining

Foi desenvolvido um estudo direcionado para a automatização parcial de DM em SIH (Alapont et al., 2005). O estudo realizado demonstrou o desenvolvimento do processo de DM. Este processo seguiu as várias fases da metodologia CRISP-DM, a escolha desta metodologia prendeu-se com o facto de esta suportar a implementação de projetos que se inseriam no mesmo âmbito do estudo apresentado (Alapont et al., 2005). Os hospitais submetidos ao estudo realizado mantêm-se confidenciais, ou seja, estes não são identificados nem localizados (Alapont et al., 2005).

Os objetivos de negócio definidos para o desenvolvimento do projeto focam-se essencialmente na gestão hospitalar: melhorar o uso de recursos hospitalares, evitar a ocupação de camas superior a 100% e planear o cronograma de forma mais adequada aos blocos operatórios (Alapont et al., 2005).

De forma a responder aos objetivos de DM foi necessário determinar quais os dados internos e externos a usar. Os dados internos usados eram referentes aos dados pessoais dos doentes, tais como, sexo, idade e o local onde reside. Também os dados das atividades do utente foram usados, dados como a data de entrada no hospital, duração da estadia, data de

alta, código do serviço médico, código de emergência, diagnóstico inicial e diagnóstico final. Os dados externos são referentes aos dados meteorológicos (temperatura, chuva e vento e sol), fases lunares e características relevantes de cada dia (feriados, festas ou eventos desportivos). Quer os dados internos, quer os externos foram recolhidos de um dos hospitais "piloto". A amostra dos dados eram desde o ano 2000 ao ano 2004. Foram submetidos cerca de 1.459 registos de 17 atributos (Alapont et al., 2005).

Os modelos de previsão realizados foram conseguidos através da aplicação WEKA[1]. Os métodos de aprendizagem usados na aplicação WEKA foram: Regressão Linear, *LestMedSq*, *SMOreg*, *Multilayer Percepton*, *Kstar*, *LWL*, AD *Stump*, árvores *M5P* e *IBK* .

Através do desenvolvimento de modelos de DM foi possível determinar o erro médio absoluto de internamentos por dia, estimado em 13,1 doentes por dia através do modelo de Regressão Linear, a estimativa do número de admissões por dia foi de 12,95 doentes, este valor foi obtido através do modelo de árvore M5P. Além destas verificações, foram ainda desenvolvidos modelos similares, por dia e por turno. Em alguns dos casos verificou-se desvios de resultados, quando comparados com os valores padrão (Alapont et al., 2005).

O sucesso alcançado pelo projeto poderá transformar o DM numa tecnologia disponível para vários hospitais, que não são capazes de pagar aplicações a partir do zero (Alapont et al., 2005).

2.2.4. Análise Inteligente de Ocupação de Camas Hospitalares

As camas hospitalares são um dos recursos mais escassos dos hospitais. Na maioria dos casos encontram-se organizadas por especialidades hospitalares, de forma a proporcionar uma melhor assistência aos doentes.

O presente estudo teve como objetivo demostrar soluções de auxílio aos administradores hospitalares, no planeamento de transferência de camas e em determinar estratégias de alocação das mesmas. O estudo foi realizado no Hospital da Universidade Nacional de Singapura, trata-se de um hospital especializado, fornecedor de serviços médicos de "ponta" (Teow, Darzi, Foo, Jin, & Sim, 2012).

[1] http://www.cs.waikato.ac.nz/ml/weka/

O respetivo estudo segue o princípio de que algoritmos de classificação são mais vantajosos em relação a métodos estatísticos padrão, quer na capacidade de modelar relações não lineares, quer no desenvolvimento de modelos interpretáveis. Várias técnicas de DM foram usadas: Árvore de Decisão (AD), Redes Neuronais Artificiais e Regressão Logística, bem como a abordagem metodológica CRISP-DM (Teow et al., 2012).

O Departamento de Emergência e Acidentes recolheu um conjunto de registos de admissões de doentes no período compreendido entre 1 de Outubro de 2009 a 30 de Março de 2010, no que resulta numa amostra de 6729 registos. Os registos eram compostos por vários campos: idade do doente, gravidade, data e hora de solicitação de cama, tempo de espera, especialidade. Foram ainda definidas variáveis adicionais que derivaram das variáveis existentes, indicador de transbordo, intervalo de horas e dia da semana em que foi solicitada a cama (Teow et al., 2012).

O resultado da aplicação das várias técnicas de DM resultou em acuidades compreendidas entre 74.1% e 76.5%. Os modelos identificam a **especialidade** como a variável de diferenciação chave, isto quer dizer que há incompatibilidade na alocação de camas por departamento. Em particular no serviço de Neurologia, durante as 15:00 e as 7:00 horas apresentou picos de transbordo na ordem dos 70%, enquanto que a Medicina Respiratória apresenta um excesso médio de 10%. Para os restantes departamentos o **tempo de solicitação** de cama foi a variável mais importante a seguir. Verificou-se que havia menos trasbordo entre as 7:00 e as 15:00 horas, em comparação com outros períodos de tempo. Em outros períodos, 00:00-7:00 existiam um maior número de transbordo do que na parte da tarde e da noite, 15:00-00:00. Além disso, o tempo de espera de cama mais elevado encontra-se associado à maior probabilidade de transbordo. A razão que estava por traz deste fenómeno era provavelmente a falta de camas no hospital durante o respetivo período o que levava a que o doente fosse alocado numa outra qualquer cama depois de um grande período de espera no Departamento de Emergência e Acidentes (Teow et al., 2012).

O estudo destacou o desequilíbrio de transbordo nos vários departamentos. O que implicou que algumas especialidades tivessem menos camas à disposição do que outras especialidades, em função da respetiva procura pelas especialidades, desta feita, surge a necessidade de alocar doentes em especialidades com capacidade de alocação.

Após a realização do estudo os administradores possuíam informações importantes para realizar uma melhor gestão de camas, podendo aumentar o número de camas em vários

departamentos, ou redistribuir as camas existentes pelos vários departamentos (Teow et al., 2012).

2.3. Business Intelligence

Os gestores responsáveis pela tomada de decisão nas organizações sabem que a informação atempada e precisa permite melhorar o desempenho do negócio e, por conseguinte, o da organização. O BI é visto como ferramenta essencial para melhorar a qualidade e quantidade de informação disponível para a tomada de decisão (Maribel Santos & Ramos, 2009).

Figura 2 – Componentes de BI (adaptado de Vercellis, 2009)

Para Turban, o termo BI é bastante abrangente, combina arquiteturas, ferramentas, base de dados, ferramentas analíticas, aplicações e metodologias. O principal objetivo do BI é proporcionar o acesso iterativo aos dados de forma a possibilitar a manipulação dos mesmos para que os gestores e analistas realizem análises adequadas (Turban et al., 2011).

Quando são analisados dados históricos, atuais, de situação e de performance, os gestores obtêm informações valiosas que lhes permite tomar as melhores decisões. O processo

de BI baseia-se na transformação de dados em informação, posteriormente em decisões e por fim em ações (Turban et al., 2011).

Os sistemas de BI têm vindo a adquirir funcionalidades de escalabilidade e segurança nos sistemas de gestão de base de dados com o objetivo de construir *Data Warehouses* (repositório de dados) para que posteriormente sejam aplicadas técnicas de *On-Line Analytical Processing (OLAP)* e DM. A aplicação destas técnicas em grandes bases de dados resulta na extração de informação importante para o negócio das organizações (Maribel Santos & Ramos, 2009). O BI é composto por um vasto conjunto de componentes. A Figura 2 ilustra as principais componentes que o constituem.

Se o DM não fosse parte integrante do BI muitas empresas não teriam a capacidade de realizar análises de mercado eficazes, comparar *feedbacks* de clientes, descobrir os pontos fortes e fracos dos concorrentes, reter os clientes de maior valor para as organizações e tomar decisões inteligentes para o negócio. O DM é claramente uma componente importante para o BI. As ferramentas de processamento de análise *online* de BI dependem do armazenamento de dados multidimensionais e do DM. As técnicas de classificação são o núcleo da análise preditiva em BI. Existem muitas aplicações na análise de mercados (exemplo: identificar pontos fortes dos concorrentes), abastecimento (exemplo: na definição e monitorização de rotas) e vendas (exemplo: identificar o volume de vendas dos produtos com desconto e identificar os mais rentáveis) (Han, Kamber, & Pei, 2012) que recorrem a técnicas de DM para suportar o seu negócio.

2.4. Data Mining

O avanço tecnológico tem proporcionado novas formas de criar e de armazenar dados. As organizações acumulam dados relacionados com os seus processos, (faturação, transações comerciais e contabilidade) tendo como base a ideia de que, os grandes volumes de dados podem ser fonte de conhecimento (Manuel Santos & Azevedo, 2005).

O DM é um termo usado para descrever a descoberta de conhecimento a partir de grandes quantidades de dados (Turban et al., 2011).

Do ponto de vista técnico o DM é um processo que recorre a técnicas de inteligência artificial, estatística e matemática para extrair informação e conhecimento útil (ou padrões) a

partir de grandes volumes de dados. Esses padrões podem estar na forma de regras de negócio, afinidades, correlações, termos ou modelos de previsão (Turban et al., 2011).

Vários métodos são usados em DM para diferentes fins e objetivos. As taxonomias são usadas como auxílio na compreensão da variedade de métodos, das suas inter-relações e agrupamentos (Maimon & Rokach, 2010).

Para Berry, Santos e Maimon existem dois tipos de orientações que podem ser usadas para fornecer informações relevantes em DM: orientado à verificação (o sistema verifica as hipóteses do utilizador), e orientado à descoberta (o sistema identifica novas regras e padrões de forma autônoma) (Berry & Linoff, 2000), (Maribel Santos & Ramos, 2009) (Maimon & Rokach, 2010). A Figura 3 representa a taxonomia.

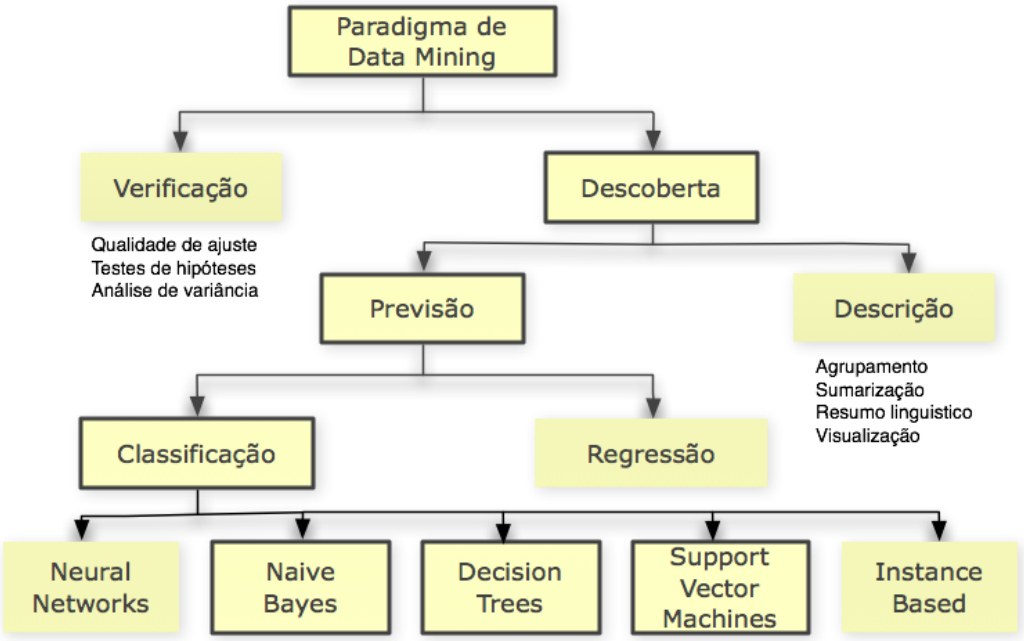

Figura 3 – Taxonomia de DM (adaptado de Maimon & Rokach, 2010)

Os métodos de descoberta conseguem identificar padrões nos dados automaticamente. Os ramos que surgem do nó descoberta são os métodos de **Previsão** e **Descrição**. Os métodos descritivos são orientados para o aumento do conhecimento dos dados (por exemplo, através da visualização), esta é a forma como os dados subjacentes se relacionam com as suas partes. Os métodos de previsão tem como objetivo construir de forma automática um modelo comportamental, obtendo novas amostras e amostras desconhecidas, sendo capaz de prever valores de uma ou mais variáveis que se encontram relacionadas com a amostra. A conceção de padrões que possibilitam a descoberta de conhecimento é de fácil compreensão e podem ser facilmente usados como base de trabalho (Maimon & Rokach, 2010).

A maioria das técnicas orientadas à descoberta é baseada na aprendizagem indutiva, onde o modelo é construído, explicitamente ou implicitamente, pela generalização de um número suficiente de exemplos de treino. O pressuposto subjacente da abordagem indutiva consiste na aplicação do modelo de treino em exemplos futuros imprevistos (Maimon & Rokach, 2010).

O processo de aprendizagem indutivo é dividido em duas fases: fase de aprendizagem e fase de teste. Na fase de aprendizagem o algoritmo analisa os dados e reconhece semelhanças entre objetos de dados. O resultado desta análise será a conceção de uma árvore ou algo equivalente, para um conjunto de regras de produção. A fase de teste é a fase onde, as regras são avaliadas utilizando novos dados e algumas medidas de desempenho são computadas (Cios, Pedrycz, Swiniarski, & Kurgan, 2007).

Os métodos de **Verificação** por outro lado lidam com a avaliação de hipótese sugerida por uma fonte externa, por exemplo, um especialista. Comparando estes métodos com os métodos orientados à descoberta, estes são menos associados ao DM porque a maioria dos problemas do DM passa pela descoberta de uma hipótese ou de um grande conjuntos de hipóteses, em vez de testar o que é conhecido (Maimon & Rokach, 2010).

Outra terminologia usada pela comunidade são as **Máquinas** de **Aprendizagem** que se referem aos métodos de previsão como sendo de, aprendizagem supervisionados e aprendizagem não supervisionados. Aprendizagem não supervisionada refere-se às técnicas com a particularidade de agrupar instâncias de atributos não dependentes. Os métodos de aprendizagem supervisionados por outro lado tentam descobrir a relação entre os atributos de entrada, ou variáveis independentes, e o atributo de saída, ou variável dependente. Os modelos resultantes descrevem e explicam fenómenos que se encontram escondidos num conjunto de dados (Maimon & Rokach, 2010).

É necessário fazer a distinção entre dois dos modelos supervisionados, os modelos de **Classificação** e de **Regressão**. Os modelos de Classificação têm como objetivo identificar uma função que associe um caso a uma classe dentro de diversas classes discretas de classificação, ou seja, os classificadores mapeiam o espaço de entrada em classes predefinidas. Por exemplo, a classe doente possui atributos que descrevem o doente; se determinada pessoa satisfizer as propriedades de classificação do doente, então a pessoa pode ser classificada como doente (Manuel Santos & Azevedo, 2005).

Por outro lado a Regressão mapeia os dados de entrada num domínio de valores reais (Maimon & Rokach, 2010). Por exemplo, um algoritmo de regressão pode prever o período de tempo em que o doente voltará a ser internado, por exemplo, determinar o número de dias.

2.5. Metodologia CRISP-DM

O processo de DM é complexo mas se for enquadrado num contexto metodológico torna-se mais fácil de compreender, implementar e desenvolver. As metodologias mais conhecidas são: CRISP-DM e *Sample, Explore, Modify, Model Assessment* (SEMMA), bem como a especificação *Predictive Model Markup Language* (PMML) (Manuel Santos & Azevedo, 2005).

A metodologia CRISP-DM foi concebida em finais de 1996 por três "veteranos" - consórcio formado pelas empresas NCR (EUA e Dinamarca), DaimlerChrysler AG (Alemanha), SPSS Inc. (EUA) e OHRA (Grupo Bancário Holandês) (Chapman et al., 2000).

Os fundamentos da metodologia CRISP-DM são construídos com base na teoria, nos princípios académicos, na prática e na experiência daqueles que desenvolvem projetos na área do DM (Manuel Santos & Azevedo, 2005).

Chapman descreve esta metodologia como sendo um processo hierárquico com um ciclo de vida que se desenvolve em seis fases: Estudo de Negócio, Estudo de Dados, Preparação de Dados, Modelação, Avaliação e Implementação. Estas fases não possuem uma sequência fixa, o desempenho resultante de cada fase é que determina qual a fase ou tarefa que deve ser realizada posteriormente. A sequência lógica das várias fases está ilustrada na Figura 4, através do ciclo externo. A sequência lógica das várias fases nem sempre é seguida, por vezes é necessário recorrer a fases anteriores para reparar ou identificar aspetos importantes. Este cenário é frequente, principalmente quando a primeira abordagem de determinada fase da metodologia é feita prematuramente (Chapman et al., 2000). No Anexo A está ilustrado o ciclo de vida detalhado do CRISP-DM. Esta será a metodologia adotada para desenvolver este projeto.

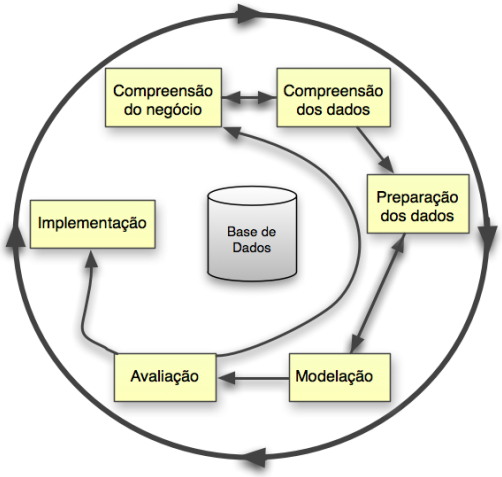

Figura 4 – Ciclo de Vida do CRISP-DM (adaptado de Chapman et al., 2000)

De seguida, são descritas resumidamente as seis fases do CRISP-DM.

2.5.1. Compreensão do Negócio

A primeira fase do CRISP-DM foca-se no entendimento dos objetivos do projeto e nos requisitos do ponto de vista do negócio. O conhecimento adquirido deve ser convertido numa definição do problema de DM e num plano preliminar para atingir os objetivos (Chapman et al., 2000). Esta fase compreende as seguintes tarefas:

- Determinar os objetivos de negócio;
- Avaliação da situação atual;
- Definição dos objetivos do negócio;
- Produzir o plano do projeto.

2.5.2. Compressão dos Dados

Esta fase começa com a junção de dados e prossegue com as atividades que permite: entender os mesmos, identificar problemas relacionados com a qualidade dos dados, identificar as relações entre os dados ou detetar subconjuntos interessantes capazes de formar hipóteses sobre informações que se encontram ocultas (Chapman et al., 2000). Esta fase compreende as seguintes tarefas:

- Recolha inicial dos dados;
- Descrição dos dados;
- Exploração dos dados;
- Verificação da qualidade dos dados.

2.5.3. Preparação dos Dados

A terceira fase abrange todas as atividades necessárias para a construção de um conjunto de dados finais (estes dados serão inseridos na ferramenta de modelação) a partir dos dados iniciais. As tarefas de preparação de dados são suscetíveis de se realizarem várias vezes, a sua repetição não possui ordem prevista. As tarefas incluem a seleção de tabelas, campos e registos, bem como a transformação e limpeza de dados para posteriormente serem usados por ferramentas de modelação (Chapman et al., 2000). Esta fase compreende as seguintes tarefas:

- Seleção dos dados;
- Limpeza dos dados;
- Construção dos dados;
- Integração dos dados;
- Formatação dos dados.

2.5.4. Modelação

Nesta fase, várias técnicas de modelação são selecionadas e aplicadas, também o ajuste dos seus parâmetros deve ser realizado para otimizar os resultados. Normalmente, existem várias técnicas para o mesmo tipo de problema de DM. Algumas técnicas têm requisitos específicos sobre a forma e tipo dos dados. Quer isto dizer que muitas das vezes é necessário regressar à fase de preparação dos dados (Chapman et al., 2000). No âmbito deste trabalho as técnicas aplicadas serão: AD, Árvores de Regressão (AR), *Support Vector Machine* (SVM) e *Naïve Bayes* (NB), para a resolução de problemas de previsão através da Classificação e Regressão. Esta fase compreende as seguintes tarefas:

- Seleção das técnicas de modelação;
- Conceção de modelos de teste;
- Construção do(s) modelos(s);

- Revisão do(s) modelo(s).

2.5.4.1. Árvore de Decisão

Uma AD consiste na forma de representar um conjunto de regras que seguem uma hierarquia de classes e valores. Expressam uma lógica simples e condicional, são facilmente interpretadas pelos utilizadores e graficamente são semelhantes a uma árvore biológica (Manuel Santos & Azevedo, 2005).

A estrutura de uma AD é muito semelhante a uma AR, exceto que nas ARs cada folha prevê um número real e nas ADs prevê decisões (Kantardzic, 2011).

As ADs são compostas por nós e ramos, as ligações existentes entre os diversos nós são estabelecidas através dos ramos que correspondem aos valores dos atributos. Os nós que se encontram na parte inferior da árvore são designados por folhas e indicam as classes em que cada registo pode ser classificado. O nó superior da árvore é designado por nó raiz, este contêm todos os exemplos de treino e podem ser divididos em classes. Todos os nós com a exceção das folhas são chamados de nós de decisão, pois as decisões tomadas são efetuadas nestes nós, tendo como base um único recurso. Cada nó de decisão tem um número finito de nós filhos, esse número é igual ao número de valores que um determinado recurso pode assumir (Cios et al., 2007).

Na Figura 5, está ilustrada uma AD, em que o "Como está o céu?" é o nó raiz da árvore, "Há humidade?" é o nó de decisão, "Sol", "Chuva", "Nublado", "Muita", "Pouca", "Sim" e "Não" são os ramos e "Jogar" e "Não jogar" são os nós folhas.

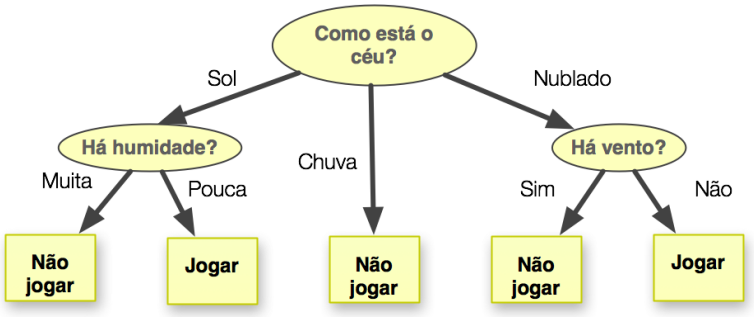

Figura 5 – Árvores de Decisão (M. F. Santos & Azevedo, 2005)

A partir da Figura 5 é possível identificar caminhos que surgem a partir do nó raiz da árvore e do nó de decisão, esses podem ser transformados em regras, através da conjugação de

testes realizados ao longo de cada caminho (Lavrač & Blaž, 2010a). A Figura 5 apresenta uma das regras associadas à ida a praia:

Se "Como está o céu?" = "Sol" e "Há humidade?" = "Pouca" Então "Jogar"

Cada AD possui um grau de complexidade que determina de forma crucial a sua precisão. A complexidade de uma AD é explicitamente controlada pelos critérios de interrupção e pelo método de **poda** empregue. A poda nas AD é um mecanismo usado para eliminar dados ruidosos, tendo como objetivo aumentar a precisão na classificação (Lavrač & Blaž, 2010b).

Os métodos usados na poda das AD tentam superar os problemas de sobre-ajustamento dos modelos, utilizando métricas estatísticas para remover os ramos "menos fiáveis". Este procedimento proporciona normalmente a obtenção de processos de classificação mais rápidos e oferece uma maior capacidade às árvores para classificar corretamente os dados desconhecidos (Maribel Santos & Ramos, 2009).

Podem ser empregues dois tipos de poda:

- Durante o processo de aprendizagem (*forward pruning ou pre-pruning*);
- Após o processo de aprendizagem (*post-pruning ou backward pruning*).

Quando a poda é concebida durante o processo de aprendizagem é ao mesmo tempo construída a árvore, isto é, os atributos são testados durante a construção da árvore. A poda quando realizada após o processo de aprendizagem só é conseguida depois da sua construção estar finalizada, ou seja, depois de uma análise de contexto ser realizada (Manuel Santos & Azevedo, 2005).

Um dos aspetos que dificulta a compreensão de uma AD é também a complexidade. Essa é normalmente medida pelos seguintes resultados: número total de nós, número total de folhas e o número de atributos usados (Lavrač & Blaž, 2010a). Mesmo depois de se utilizar o mecanismo de poda, após aprendizagem, a árvore gerada pode representar uma estrutura complexa, com difícil compreensão (Manuel Santos & Azevedo, 2005).

2.5.4.2. Naïve Bayes

A teoria de Bayes é uma técnica usada fundamentalmente para o reconhecimento de padrões e de classificação. Baseia-se no pressuposto de que a classificação de padrões é expressa em termos probabilísticos. As características estatísticas dos padrões são exprimidas como sendo valores de probabilidades conhecidas que descrevem a natureza aleatória dos padrões e suas características. A teoria de decisão de Bayes fornece uma estrutura para métodos estatísticos na classificação de padrões em classes com base em probabilidades (Cios et al., 2007).

A teoria de Bayes surge através de um trabalho realizado por Tomas Bayes (Bellhouse, 2004) mas efetivamente foi o matemático francês Pierre Simon de Laplace, quem desenvolveu o teorema como ele é conhecido e usado (Manuel Santos & Azevedo, 2005). O teorema de Bayes é definido a partir da seguinte formula matemática:

$$P(H|X) = \frac{P(X|H) * P(H)}{P(X)}$$

Na qual *P(H| X)* é a probabilidade à posterior, onde *H* é a condicionante de *X*, isto é, *P(H| X)* representa a confiança em *X*, a probabilidade condicionada representa a ocorrência do evento *X* condicionando a ocorrência do evento em *H*. *P(H)* corresponde a probabilidade à priori de *H*. A probabilidade à posterior é relativamente diferente da à priori, porque baseia-se em mais informação onde *P(H)* é independente de *X* (Manuel Santos & Azevedo, 2005).

O método de classificação NB foi desenvolvido a partir do teorema de Bayes (Tufféry, 2011).

Uma NB fornece uma abordagem simples, com uma semântica clara, na representação, uso e aprendizagem do conhecimento probabilístico (Witten & Frank, 2005). Há conjuntos de dados em que as NBs não são capazes de obter bons resultados porque os atributos são tratados como se fossem totalmente independentes. A adição de atributos redundantes distorce o processo de aprendizagem (Witten & Frank, 2005).

O funcionamento das NBs está dividido em 4 fases, elas são (Han et al., 2012):

1. Sendo *D* definido como um conjunto de dados (tuplos) de treino associados a uma classe. Cada tuplo é representado por um vector com *n-dimensões* de atributos, *X=(x1, x2,...,xn)*, representam *n-dimensões* realizadas no tuplo de *n-atributos*, respetivamente *A1, A2, ... , An*.

2. Supondo que existem *n classes, C1, C2,..., Cn*. A partir de um tuplo *X*, o classificador irá prever que *X* pertence à classe com maior probabilidade à *posterior* condicionado *X*, isto quer dizer que o classificador das NBs prevê que o tuplo *X* pertence à *classe Ci* se e só se:

$$P(Ci|X) > P(Cj|X) para\ 1 \leq j \leq m, j \neq i$$

Assim consegue-se maximizar *P(Ci| X).* A *classe Ci* para o qual *P(Ci| X)* é maximizada é designada por hipótese máxima a posterior. A relação do teorema de Bayes com as NBs é representado pela seguinte formula:

$$P(Ci|X) = \frac{P(X|Ci) * P(Ci)}{P(X)}$$

3. Como *P(X)* é constante para todas as classes, apenas *P(X| Ci)P(Ci)* necessita de ser maximizado. Se as classes anteriores não forem conhecidas, então é comum que as classes possuam a mesma probabilidade, isto é *P(C1)=P(C2),...,P(Cm)*), desta feita consegue-se maximizar *P(X| Ci)*, caso contrario é maximizado *P(X| Ci)P(Ci)*. As probabilidade das classes anteriores podem ser estimadas por, *P(Ci)= |Ci,D|/|D|*, em que *|Ci,D|* é o numero de tuplos de treino da classe *Ci* em *D*.
4. Um conjunto de dados com muitos atributos, seria extremamente dispendioso do ponto de vista computacional calcular *P(X| Ci)*. Para reduzir essa complexidade é realizada a suposição Naïve da classe condicional. Isto pressupõem que os valores dos atributos sejam condicionalmente independentes uns dos outros, mas é necessário ter em conta o rótulo da classe do tuplo, ou seja, não há relação de dependência entre os atributos, assim

$$P(X|Ci) = \prod_{k=1}^{n} P(xk|Ci) = P(x1|Ci) * P(x2|Ci) * ... * P(xn|Ci)$$

É possível calcular facilmente a probabilidade *P(x1| Ci) ∗ P(x2| Ci) ∗ ... ∗ P(xn| Ci)* a partir dos tuplos de treino. É necessário referir que *xk* representa o valor do atributo *Ak* para cada tuplo de *X*.

Este método é importante porque é muito simples de construir e não necessita de esquemas complicados para estimar parâmetros, isto que dizer que o método pode ser aplicado a grandes conjuntos de dados. As NBs são de fácil interpretação mesmo para quem não é especializado na

área das tecnologias e ainda demonstra ser um modelo eficaz na previsão de acontecimentos (Xindong & Vipin, 2009).

2.5.4.3. Support Vector Machine

O primeiro trabalho realizado em SVM foi apresentado por Vladimir Vapnik, Bernhard Boser e Isabelle Guyon (1992), embora as bases deste método tenham surgido em 1960, com o trabalho de Vapnik e Chervonenkis na teoria de aprendizagem estatística (Han et al., 2012).

O SVM é um método usado na classificação de dados lineares e não lineares. Este método usa o mapeamento não-linear para transformar os dados de treino em dimensões superiores, dentro destas novas dimensões ele procura o hiperplano ótimo separando-o linearmente, ou seja, é criado uma "fronteira de decisão". Para que o mapeamento não linear seja apropriado para uma dimensão suficientemente elevada, os dados das duas classes podem ser separados por um hiperplano. O SVM encontra este hiperplano através de vetores de suporte e de margens, sendo definidos pelos vetores de suporte (Han et al., 2012).

Os SVMs podem ser usados na previsão numérica bem como na classificação. Eles foram aplicados a um série de áreas, incluído reconhecimento de dígitos manuscritos, reconhecimento de objetos, e ainda em testes de *benchmark* em séries de previsões temporais (Han et al., 2012).

O tempo de conceção de um SVM pode ser extremamente lento, porem são altamente precisos devido à sua capacidade em modelar limites complexos em decisões lineares (Han et al., 2012).

O SVM tem uma boa base técnica e necessita apenas de um pequeno número de amostras para a sua formação, várias experiências mostram que é insensível ao número de dimensões de amostras. Um exemplo simples do aspeto descrito está demostrado na Figura 6, onde as amostras pertencem a uma das classes, azul ou amarela (Kantardzic, 2011).

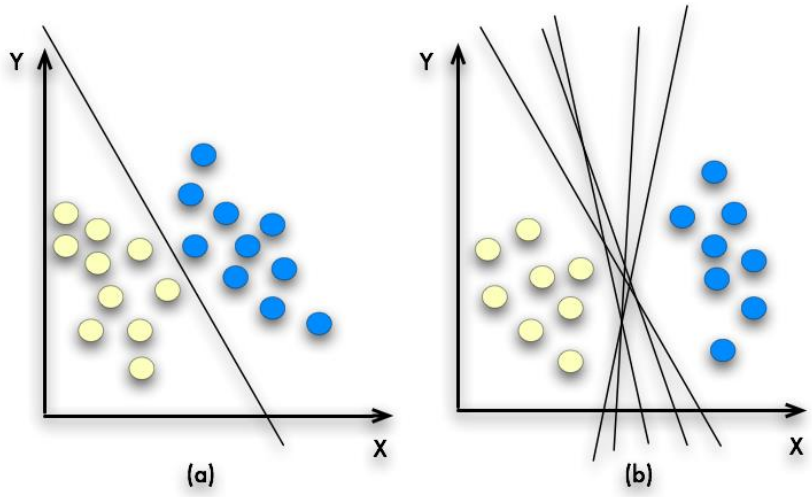

Figura 6 – Separação Linear (Kantardzic, 2011)

Neste problema (Figura 6) o objetivo é separar as duas classes por uma função que é induzida a partir de exemplos disponíveis. Em (b) existem vários classificadores lineares possíveis que podem separa as duas classes das amostras, o mesmo não acontece em (a), pois existe apenas um único classificador linear. A ideia principal é, aumentar o limite de decisão, tão longe quanto possível dos pontos de ambas as classes, há apenas uma que maximiza a margem, isto é, que maximiza a distância entre o classificador e o ponto mais próximo do dado de cada classe (Kantardzic, 2011).

Vapnik (1995) propôs que para selecionar o hiperplano ótimo, este terá de ser aquele que maximiza a largura da margem entre as observações. Então, este hiperplano irá garantir não só o ajuste do modelo, mas também a sua robustez (Tufféry, 2011).

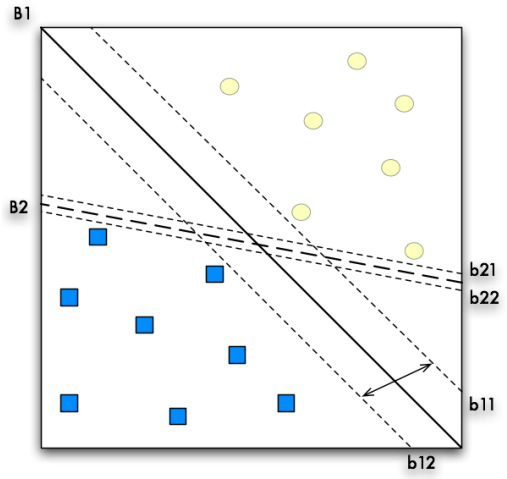

Figura 7 – Exemplo de Hiperplanos (Tufféry, 2011)

Na Figura 7 estão representados dois hiperplanos *B2* e *B1*, ambos podem ser solução, mas *B1* é melhor que *B2*, porque maximiza a margem (Tufféry, 2011).

O SVM pertence à família de modelos lineares generalizados que alcança uma classificação ou decisão de regressão com base no valor da combinação linear de funções de entrada (Turban et al., 2011).

Geralmente muitos hiperplanos lineares são capazes de separar os dados em subsecções múltiplas, cada um pertence a uma classe específica. No entanto, apenas um hiperplano atinge o máximo da separação. Após um processo de aprendizagem o SVM aprende com os casos históricos a construir um modelo matemático que pode perfeitamente atribuir as instâncias de dados às respetivas classes. Se houver apenas duas dimensões, os dados podem ser separados por uma linha (Turban et al., 2011).

2.5.5. Avaliação

Antes de proceder à implementação final do modelo é importante avalia-lo cuidadosamente e rever todos os passos efetuados. Esta tarefa é necessária para ter certeza do melhor modelo e garantir os objetivos de negócio. Um dos principais objetivos é determinar se há algum problema de negócio importante que não tenha sido analisado. No final desta fase a decisão sobre a utilização dos modelos de DM deve ser tomada (Chapman et al., 2000). Esta fase compreende as seguintes tarefas:

- Avaliação dos resultados;
- Revisão dos processos;
- Determinação dos próximos passos.

2.5.5.1. Métricas Associadas à Classificação

Matriz de Confusão (MC): após a geração dos modelos é necessário proceder a avaliação do desempenho. Nos problemas de classificação a técnica mais usada é a MC, na qual se torna possível a definição de várias métricas, taxas de erro e as curvas *Receiver Operating Characteristic* (ROC).

A MC de um classificador indica o número de classificações reais e de previsões efetuadas para cada classe a partir de um modelo de classificação (Manuel Santos & Azevedo,

2005). A MC tem tradicionalmente a dimensão de 2x2, a Tabela 1 representa uma MC para classificadores binários de classe (matriz com duas classes).

Tabela 1 – Matriz de Confusão (Manuel Santos & Azevedo, 2005)

Classe	Previsão C+	Previsão C-
Real C+	Verdadeiros Positivos TP	Falsos Negativos FN
Real C-	Falsos Positivos FP	Verdadeiros Negativos TN

- **Verdadeiros Positivos** (*True Positive*) são designados por TP, correspondem ao número de previsões corretas para saída positiva;
- **Verdadeiros Negativos** (*True Negative*) designados por TN, correspondem ao número de previsões corretas para saídas negativas;
- **Falsos Positivos** (*False Positive*) são designados por FP, corresponde ao número de previsões incorretas para a saída positiva;
- **Falsos Negativos** (*False Negative*) designados por FN, corresponde ao número de previsões incorretas para a saída negativa.

A partir da MC é possível calcular muitas outras medidas: Sensibilidade, Especificidade e Acuidade ou Precisão (Manuel Santos & Azevedo, 2005). A Sensibilidade identifica a quantidade de verdadeiros positivos que estão devidamente classificados como positivos (por exemplo, percentagem de pessoas com frio que estão de facto com frio). A Sensibilidade é calculada através da seguinte equação:

$$sens = \frac{TP}{TP * FN} * 100(\%)$$

A Especificidade identifica a quantidade de verdadeiros negativos, que estão devidamente classificados como negativos (por exemplo: a percentagem de uma pessoa ter realizado um teste para determinar se possui determinada doença, dar resultado negativo e ela realmente não possuir a doença). A Especificidade é calculada através da seguinte equação:

$$espe = \frac{TN}{TN * FP} * 100(\%)$$

Acuidade ou precisão é uma métrica de avaliação de modelos de DM, que corresponde a proporção de acerto do modelo (por exemplo: pode ser usada para identificar o modelo mais preciso, ou seja, o modelo que classificou devidamente os TPs e os TNs). A Acuidade é calculada através da seguinte equação:

$$acui = \frac{TP + TN}{n} * 100(\%)$$

Curva ROC: é uma ferramenta visual útil para comparar modelos de classificação (Han et al., 2012). A partir da curva ROC é possível medir o desempenho de um classificador para duas classes (Sousa, Machado, Rocha, Cortez, & Rio, 2010).

A curva ROC representa a taxa dos verdadeiros positivos (TP) no eixo y e a taxa de verdadeiros negativos (TN) no eixo x. O primeiro é o número de positivos incluídos na amostra, expresso em percentagem, o número total de Positivos (Taxa TP = $\frac{TP}{TP+FN} \times 100$) e o número total de negativos incluídos na amostra, é igualmente expresso em percentagem (Taxa FP= $\frac{FP}{FP+TN} \times 100$) (Witten, Frank, & Hall, 2011). A curva ROC permite visualizar o comprimento entre a sensibilidade (taxa de TP) e a especificidade (1- taxa de FP) do modelo (Manuel Santos & Azevedo, 2005).

Para um modelo de classificação a curva ROC apresenta um *trade-off*, entre a Taxa TP e a Taxa dos FP (Han et al., 2012). Numa situação ideal o modelo deverá possuir indicadores máximos de sensibilidade e de especificidade, ambos com valor igual a um (Manuel Santos & Azevedo, 2005).

2.5.5.2. Métricas Associadas à Regressão

O principal objetivo de um método de regressão é gerar o "melhor" modelo segundo uma estimativa de erro. Em problemas de regressão o erro resíduo e é determinado por: $e = y - \hat{y}$, onde y representa o valor desejado e $\hat{y}$ o valor estimado pelo o modelo (Manuel Santos & Azevedo, 2005).

Os valores previstos para os casos de teste são representados pelo conjunto $\widehat{yi}$ e os valores reais são representados pelo conjunto de dados yi.

Mean Squared Error (MSE): esta medida utiliza o quadrado das distancias das previsões aos valores reais, o que levará a que grandes distancias sejam relativamente amplificadas quando comparadas com as pequenas distancias. A medida MSE é aplicada quando se está perante situações em que é crucial não cometer erros extremos (Manuel Santos & Azevedo, 2005).

$$MSE = \frac{\sum_{i=1}^{n} |\widehat{yi} - yi|^2}{N}$$

Mean Absolute Error (MAE): esta medida de avaliação de modelos é uma medida alternativa á MSE. Uma vez que a medida MSE tende a exagerar nos cassos de *outliers,* a medida MAE por outro lado trata dos erros de forma uniforme de acordo com a sua magnitude (Witten et al., 2011).

$$MAE = \frac{\sum_{i=1}^{n} |\widehat{yi} - yi|}{N}$$

Relative Absolute Error (RAE): Os erros são normalizados pelo erro da previsão, que faz a previsão dos valores médios (Witten et al., 2011). O $\bar{x}$ representa a média dos valores de treino. Esta medida contrariamente ás outras apresentadas, expressa o valor em percentagem, os valores próximos de 100% correspondem a um modelo com um desempenho equivalente ao previsor médio *naïve*. Quanto menor for valor RAE melhor é o modelo de regressão, um modelo com RAE=0% seria um modelo de regressão perfeito.

$$RAE = \frac{\sum_{i=1}^{n} |yi - \widehat{yi}|}{\sum_{i=1}^{n} |yi - \bar{x}|}$$

Regression Error Characteristic (REC): é uma ferramenta de visualização poderosa com propriedades geométricas (Mittas & Angelis, 2010), tendo como finalidade avaliar e comparar facilmente modelos de previsão (Torgo, 2005). A avaliação dos modelos é feita através da consulta dos gráficos criados (Mittas & Angelis, 2010).

A curva REC determina a tolerância de erro no eixo x (ordenadas) e a precisão da função de regressão no eixo y (abscisas). A precisão é definida em percentagem de pontos contidos no espaço de tolerância. Se a tolerância for zero, apenas são considerados os pontos que a função insere na perfeição. Se a tolerância exceder o erro máximo, então todos os pontos serão considerados precisos. Assim surge a necessidade de *trade-off*, entre a tolerância de erro e

a precisão da função de erro. O conceito de tolerância de erro é pertinente, porque a maioria dos dados de regressão são imprecisos, devido a erros experimentais e erros de medição (Bi & Benett, 2003).

2.5.5.3. Cross Validation

A técnica *Cross-Validation* (CV) é usada para avaliar modelos de previsão, tendo como base um conjunto de partições definidas de forma aleatória de uma amostra de dados (Han et al., 2012).

Em *k-folds* CV, os dados são divididos em k subconjuntos, sendo estes mutuamente exclusivos $D1, D2, \dots Dk$, os vários *folds* possuem tamanhos aproximadamente igual. O treino e o teste é realizado k vezes. Na interação i a partição Di é reservado para teste e as restantes partições são usadas de forma coletiva para treinar o modelo. Isto é, na primeira interação, os subconjuntos $D2, D3, \dots Dk$ são usados coletivamente para treino e assim se obtém o primeiro modelo, o teste é feito com recurso ao conjunto $D1$. Na segunda iteração o modelo é treinado pelos subconjuntos $D1, D3, \dots, Dk$ e é testado em $D2$, este processo é realizado consecutivamente até satisfazer o valor de k (Han et al., 2012).

O método mais comum de prever a taxa de erro de uma técnica de *learning* é através de um conjunto de dados único e fixo, é a estratificação ***10-folds* CV**. É o mais comum porque foram realizados testes em diversas base de dados com recurso a diferentes técnicas de DM. Verificou-se que o 10 é o número certo de *folds* para obter melhores estimativas de erro, este princípio é ainda suportado por evidências teóricas. Embora estas evidências não sejam aceites como totalmente conclusivas, o 10-*folds* CV tornou-se o método padrão em termos práticos (Witten et al., 2011). O ***Leave-One-Out Cross-Validation*** (LOOCV) é um caso especial do *k-folds* CV, em que k é igual ao número de registos do conjunto de dados (Refaeilzadeh, Tang, & Liu, 2009), (Witten et al., 2011), (Han et al., 2012). Por outras palavras, em cada iteração são usados quase todo o conjunto de dados, expecto uma única observação, em que esta observação é usada para teste e as restantes para treino do modelo. Uma estimativa obtida usando o LOOCV é conhecida por ser quase imparcial, possui um elevado desvio, em que este conduz a estimativas pouco fiáveis. Porém, o LOOCV é usado sobretudo em conjuntos de dados de pequena dimensão, onde estes apenas possuem alguma dezenas de amostras de dados (Refaeilzadeh et al., 2009).

2.5.6. Implementação

Esta é a ultima fase da metodologia CRISP-DM, a criação do modelo geralmente não é o fim do projeto. Mesmo que a finalidade do modelo seja aumentar o conhecimento a partir dos dados, o conhecimento adquirido terá de ser organizado e apresentado de forma a que o cliente/gestor possa usá-lo. Muitas das vezes envolve a aplicação em tempo real de modelos dentro da própria organização. A fase de implementação pode ser tão simples como gerar relatórios ou tão complexos como a implementação de um processo de DM escalável em toda a organização, essa complexidade irá sempre depender dos requisitos do projeto (Chapman et al., 2000). Esta fase compreende as seguintes tarefas:

- Planeamento da avaliação dos resultados;
- Planeamento da monitorização e manutenção;
- Produção dos relatório final;
- Revisão do projeto.

A realização deste projeto não contempla a fase de implementação, no entanto, é expectável que os modelos gerados possam ser integrados/implementados em Sistemas de Apoio à Decisão.

3. Trabalho Realizado

Uma vez descritos os conceitos que suportam o desenvolvimento deste trabalho, houve a necessidade de realizar um conjunto de exercícios práticos para, transformar a informação recolhida em conhecimento. Neste Capítulo é apresentado o trabalho prático realizado durante este projeto de dissertação. Também nesta fase do projeto foi abordada a metodologia *Cross–Industry Standard Process for Data Mining* (CRISP-DM), para enquadrar e orientar o trabalho prático realizado.

3.1. Ferramentas Utilizadas

A implementação das técnicas de *Data Mining* (DM) foi conseguida através do ambiente estatístico **R**[2]. O R apresenta-se como sendo uma linguagem de programação e um ambiente de desenvolvimento estatístico (Torgo, 2011). O ambiente de desenvolvimento **R** possui várias versões para diferentes sistemas operativos (p.e. Linux, Windows e Mac OS) (Cortez, 2010), trata-se de um ambiente multiplataforma que foi inicialmente desenvolvido por Ihaka e Gentleman, ambos da Universidade de Auckland, Nova Zelândia. O desenvolvimento atual do ambiente **R** é feito por uma equipa de dezenas de pessoas de diferentes intuições espalhadas por todo o mundo e ainda por uma comunidade *open source* (Torgo, 2011). Em 1 de Agosto de 2012 o repositório de *packages* do CARN fornecia cerca de 4000 *packages* para diferentes fins (Zhao, 2012).

O ambiente **R** não foi desenvolvido especificamente para solucionar problemas de DM, mas possui vários *packages* para esse fim e com uma grande diversidade de algoritmos de DM. Segundo um estudo apresentado pelo KDnuggets[3] realizado por Gregory Piatetsky, foi possível concluir que a segunda ferramenta de DM mais usada nos últimos 12 meses (Junho 2012 até Junho 2013) em projetos profissionais foi o ambiente **R**. O estudo baseia-se em 1880 votos, cerca de 704 votos são relativos ao uso do ambiente **R** em projetos reais, o que representa uma

[2] http://www.r-project.org/

[3] http://www.kdnuggets.com/

taxa de utilização de 37,4%, de seguida está o RapidMiner com uma taxa de cerca de 39,2% (Piatetsky, 2013).

A biblioteca **e1071**[4] (D. Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2012) foi usada para implementar as técnicas de DM, *Support Vector Machine* (SVM), Árvore de Decisão (AD), Árvores de Regressão (AR) e *Naïve Bayes* (NB), e resolver problemas de previsão, tanto de Classificação como de Regressão. Para realizar a avaliação dos modelos de DM foi usada a biblioteca **rminer**[5] (Cortez, 2013).

Para desenvolver as base de dados que alimentaram os modelos de DM, foi usado o sistema de gestão de base de dados **Mysql**[6], que usa a linguagem **Structure Query Language** (SQL) como interface.

3.2. Compreensão do Negócio

O trabalho realizado prende-se pela necessidade de demostrar um estudo empírico, relacionado com a gestão do Centro Hospital do Porto (CHP) esse estudo teve como recursos, a utilização de base de dados e técnicas de DM.

Para realizar o respetivo trabalho foi necessário determinar objetivos que demostrem a necessidade de o desenvolver. Até à data o CHP não possui nenhum mecanismo capaz de prever o fluxo de altas hospitalares. Esta evidência, desencadeia todo um processo de trabalho, na qual resulta o objetivo de negócio, este consiste em prever o número de doentes com alta semanalmente, tendo como principal alvo a melhoria da gestão de camas.

Os dados usados no estudo são proveniente do hospital já mencionado, os modelos usaram como valores de entrada, o número de doentes saídos distribuídos por vários serviços. A amostra fornecida era composta por 62302 registos.

Ao nível de desenvolvimento o estudo realizado tem como objetivos de DM, obter modelos de previsão através de duas abordagens de DM, Classificação e Regressão e obter bons modelos de previsão. Em função dos modelos gerados serão identificados aqueles que

[4] http://cran.r-project.org/web/packages/e1071/index.html

[5] http://cran.r-project.org/web/packages/rminer/

[6] http://www.mysql.com/

apresentam os melhores resultados de modo a que no futuro seja possível a introdução desses modelos num Sistema de Apoio à Decisão (SAD).

Depois de realizada a análise do negócio e de identificar os seus objetivos, tornou-se evidente a necessidade de transpor o conhecimento adquirido para o paradigma de DM. Os modelos de previsão tal como já foi descrito anteriormente, seguirão duas abordagens, Classificação e Regressão, isto porque, não se pretende apenas prever intervalos de valores – classes. A necessidade de usar as duas abordagens de DM também se prende pelo facto de até a presente data não se ter identificado na revisão de literatura qual a abordagem mais adequada para prever a alta de doentes em hospitais.

O plano desenvolvido com as várias fazes do CRISP-DM encontra-se no Apêndice A. Este plano foi desenvolvido com o objetivo do trabalho a realizar termine no período estabelecido, para a entrega do projeto.

3.3. Compreensão dos Dados

Neste trabalho, tal como já foi anteriormente mencionado, foram usados dados relacionados com alta hospitalar de doentes do CHP. O período da a mostra fornecida, está compreendida entre 1/1/2009 e 31/12/2012, sendo referente a 1461 dias. Para o espaço temporal da amostra, os anos 2009, 2010 e 2011 possuem 365 dias e o ano 2012 possui 366 dias, sendo este o único ano bissexto da amostra fornecida.

No respetivo espaço temporal referenciado, foram recolhidas 62302 registos, registos esses que correspondem a altas de noventa e um serviços hospitalares. Cada registo é constituído por três campos, eles são:

- Data: corresponde ao dia, mês e ano em que o(s) doente(s) obtiveram alta hospitalar;
- Serviço: representa o serviço hospital que concedeu a alta hospitalar aos doentes;
- Número de Altas: campo que contém o número de doentes que obtiveram alta hospitalar. Este campo está diretamente relacionado com a data e com o serviço hospitalar, ou seja, os dados estavam agrupados por data e serviço.

Como já fora referido, um dos objetivos é realizar previsões de altas hospitalares por semana, desde logo foi necessário realizar o somatório de doentes com altas por serviço e pela

respetiva semana de cada ano. Em função da respetiva alteração, realizou-se a exploração dos dados de apenas quatro serviços, Ortopedia, Obstetrícia, Parto e Berçário. Neste ponto foram estudados os noventa e um serviços hospitalares, sendo que, apenas estes quatro eram os únicos que apresentavam registos completos desde o início de 2009 até o final de 2012.

A Figura 8 apresenta a variação das altas semanais e respetivos anos no serviço de Ortopedia.

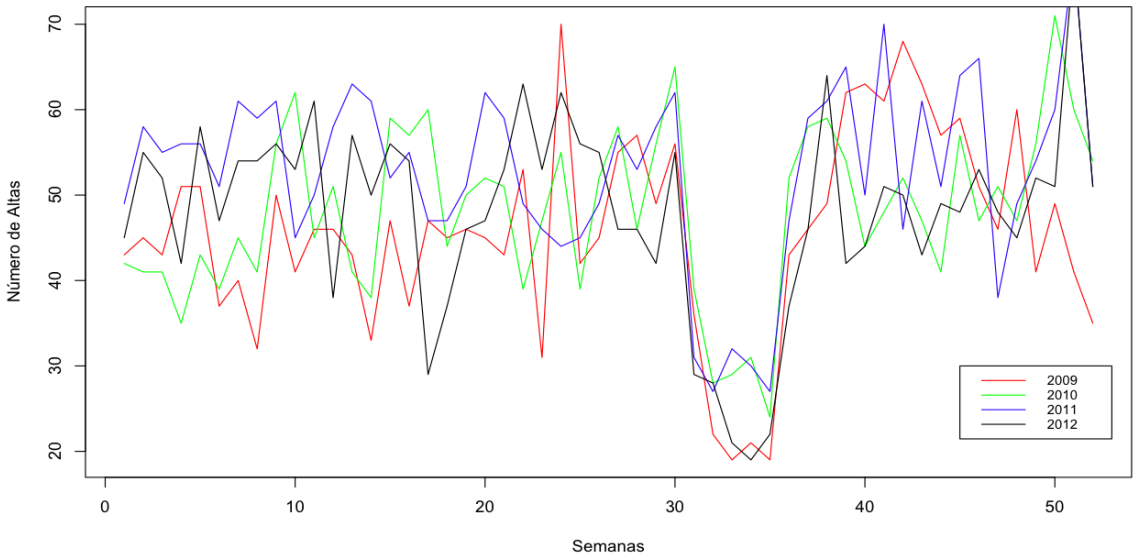

Figura 8 – Variação de altas (Ortopedia)

Perante a Figura 8, observa-se que há uma grande variação de altas hospitalares nos quatro anos para a mesma semana, não existe homogeneidade nas distribuições apresentadas. Por exemplo entre a semana 31 e a 35 verifica-se que há uma queda acentuada em relação às restantes semanas. O respetivo intervalo corresponde ao período de férias.

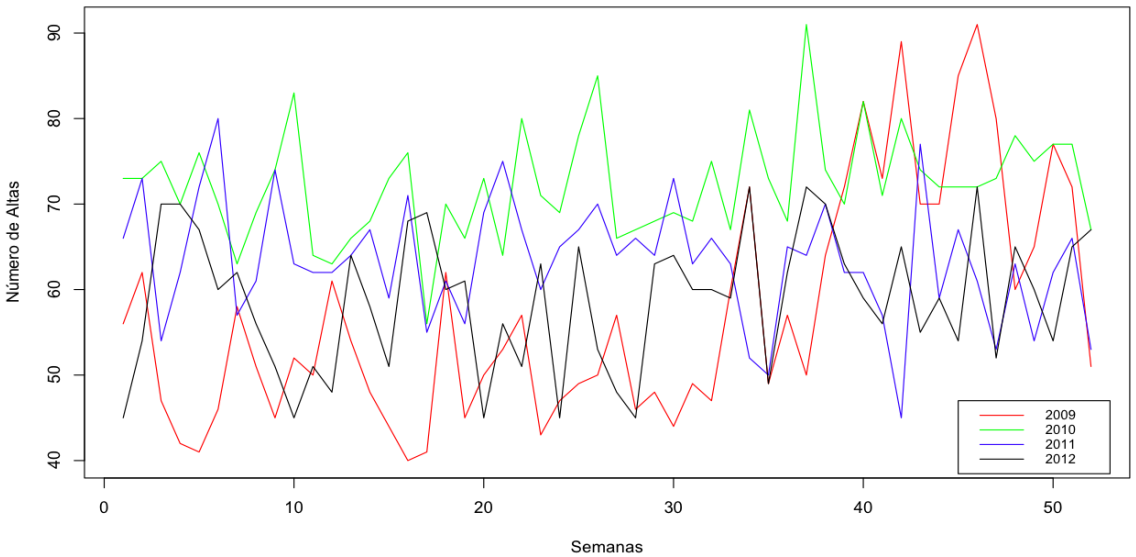

Figura 9 – Variação de altas (Obstetrícia)

A Figura 9 apresenta a variação das altas hospitalares por semana do serviço Obstetrícia. Contrariamente na Figura 8 não se verifica um decréscimo acentuado no número de altas, compreendido entre semanas. Porém, através da Figura 9 é possível observar que no início do ano 2009 (primeira até a décima semana) o número de altas por semana era inferior a dos restantes anos, e no final do ano (semana quarenta e dois até a semana cinquenta e dois), demostra um grande acréscimo em altas hospitalares. Os restantes anos não apresentam variações relevantes. De salientar também o facto de que o número de altas é significativamente diferente entre todos os anos.

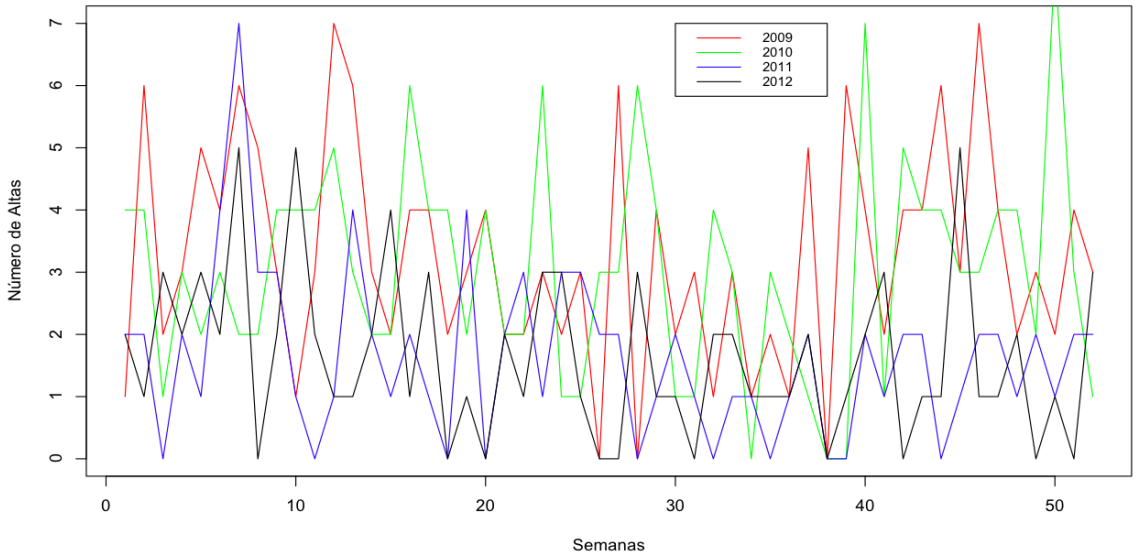

Figura 10 – Variação de Altas (Parto)

A Figura 10, apresenta a variação de altas atribuídas ao longo de quatro anos no serviço de Parto. Através da variação expressa pela Figura 10 torna-se percetível que os anos 2009 e 2010 apresentam maior número de altas atribuídas do que os restantes anos. Ainda na Figura 10 verificou-se um fenómeno que, corresponde ao valor máximo de altas semanais no serviço de Parto, o seu valor é muitíssimo inferior quando comparado com os restantes serviços.

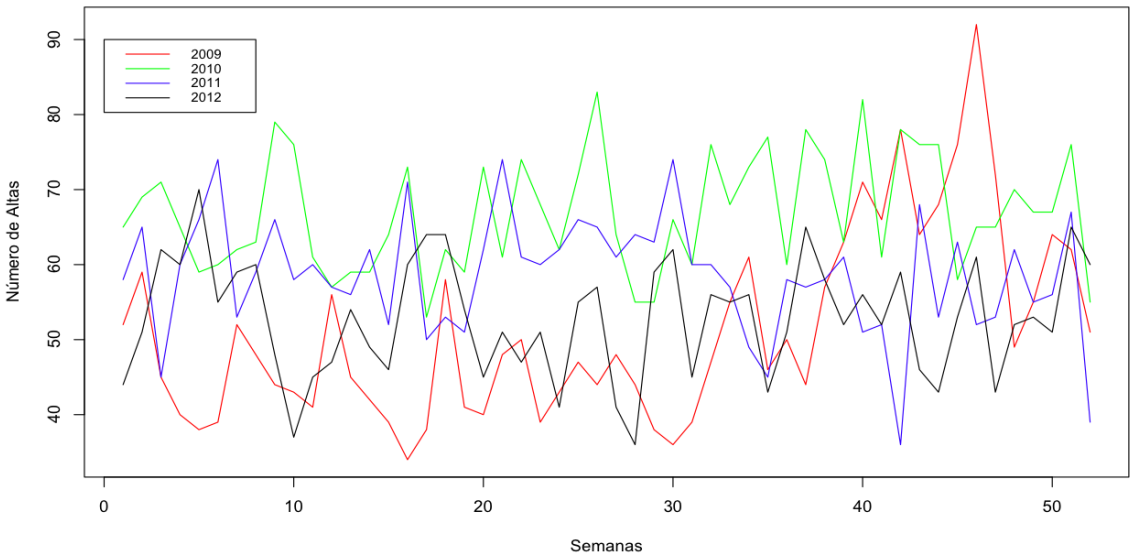

Figura 11 – Variação de altas (Berçário)

Por último é apresentada a Figura 11, que corresponde a variação de altas semanais dos serviço de Berçário. Através da Figura 11 observa-se que no serviço de Berçário, tal como no serviço de Obstetrícia, no início do ano de 2009 (semana um até a décima) o número de altas por semana era inferior à dos restantes anos, e no final do ano (semana quarenta e dois até a semana cinquenta e dois), demostra um grande acréscimo em altas hospitalares. Os anos 2010 e 2011 apresentam variações de valores superiores aos anos 2009 e 2012.

Com objetivo de detalhar mais o estudo dos dados, foi desenvolvida a Tabela 2, que analisa o máximo, o mínimo, a média, coeficiente de variação e o desvio padrão para cada um dos serviços.

Tabela 2 – Estudo dos vários serviços

	Ortopedia	Obstetrícia	Parto	Berçário
Máximo	78	91	8	92
Mínimo	19	40	0	34
Média	≈48.6	≈62.9	≈2.4	≈57.3
Coeficiente de Variação	≈22.634%	≈17.011%	≈70.833%	≈19.197%
Desvio Padrão	11	10.7	1.7	11
Total de Altas	10108	13080	495	11924

A partir da Tabela 2, é possível identificar o serviço com maior número de altas e a sua média, o serviço de Obstetrícia é dos quatro serviços o que mais altas obteve, e a sua média de altas semanais é superior à dos restantes serviços. A dispersão apresentada pelo serviço de Obstetrícia, é inferior à dos serviços Ortopedia, Parto e Berçário. O serviço de Parto apresenta

grande dispersão entre a média e o desvio padrão. Esta conclusão foi obtida através dos valores calculados do coeficiente de variação. Com os valores apresentados pelos coeficientes de variação também é possível identificar as amostras homogéneas e heterogéneas. Os serviços que possuem amostras homogéneas são a Obstetrícia e o Berçário, porque o seu coeficiente de variação é inferior a 20%. Os serviços de Ortopedia e Parto possuem amostras heterógenas uma vez que o seus coeficientes de variação são superiores a 20%.

Após a elaboração da exploração dos dados, foi realizada a verificação da integridade dos mesmos. A amostra fornecida apresenta alguns problemas associados:

- A data não possuía o formato adequado, estava em formato numérico;
- A amostra possuía registos duplicados, registos esses que continham a mesma data e serviço só que um desses registos encontrava-se com o valore zero e o outro com o valor efetivo das altas atribuídas;
- Por último, a amostra é composta por 62302 registos, porém verificou-se que apenas quatro dos noventa e um serviços possuíam 1461 registos, isto quer dizer, que só em quatro serviços foram feitos registos todos os dias durante quatro anos.

3.4. Preparação dos Dados

Partindo do pressuposto, que com este estudo se pretende realizar previsões de altas hospitalares, através de duas abordagens de DM, Classificação e Regressão, é imperativo preparar os dados de modo a que seja possível empregar as duas abordagens de DM. Como tal foram empregues alguma ações de melhoria. O formato da data uma vez não estando no formato correto foi necessário recorrer a transformação do seu valor, colocando-o no formato de Ano/Mês/Dia.

Uma vez que a amostra também possuía registos duplicados foi necessário agrupar os valores de doentes com alta apresentados nos dois registos. Por último, foi analisada a inexistência de registos de vários serviços ao longo dos quatro anos. Como forma de resolver este problema foi necessários descartar todos os serviços que não possuíam 1461 registos. Como medida corretiva foi decidida a não inclusão desses serviços no presente estudo. Neste aspeto foram apenas utilizados os dados de quatro serviços dos 91 serviços existentes no CHP: Ortopedia, Obstetrícia, Parto e Berçário.

Neste estudo, tal como já foi mencionado, pretende-se realizar previsões de altas semanalmente, para que isso fosse possível, foi necessário agrupar os registos diários em registos semanais. Por convenção uma semana tem início no Domingo e termina no Sábado seguinte. Seguindo este princípio, as altas foram agrupadas em 52 semanas, pelos respetivos anos 2009, 2010, 2011 e 2012.

Depois de agrupados os registos diários em semanas foi necessário recorrer a métodos capazes de determinar intervalos de altas hospitalares, para deste modo realizar previsões através da abordagem de Classificação.

A seleção do número de classes ou de intervalos não constitui nenhum método rigoroso e científico, nem existe nenhum método de seleção que possa ser considerado o mais correto (Reis, 2008). Assim sendo, foram implementados alguns métodos para criar classes: Média, Quartis, Média-Desvio Padrão e Regra de Sturges.

Média: A média é uma medida de **localização** que, geralmente, indica um valor central da distribuição (Murteira, Ribeiro, Silva, & Pimenta, 2010). Foi essencialmente através da média que foram criadas duas classes, como tal foi necessário identificar o seu valor, que se traduz na média de altas pelos quatro serviços, os limites das classes foram definidos em função do máximo e o mínimo de altas. A média foi calculada através da expressão matemática $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} xi$, onde n representa o número de semanas e xi o número de altas hospitalares da respetiva semana. Os intervalos são expressos por, $[min, \overline{x}]$ e $]\overline{x}, max]$.

A Figura 12 representa as frequências das classes (Média) criadas para os vários serviços.

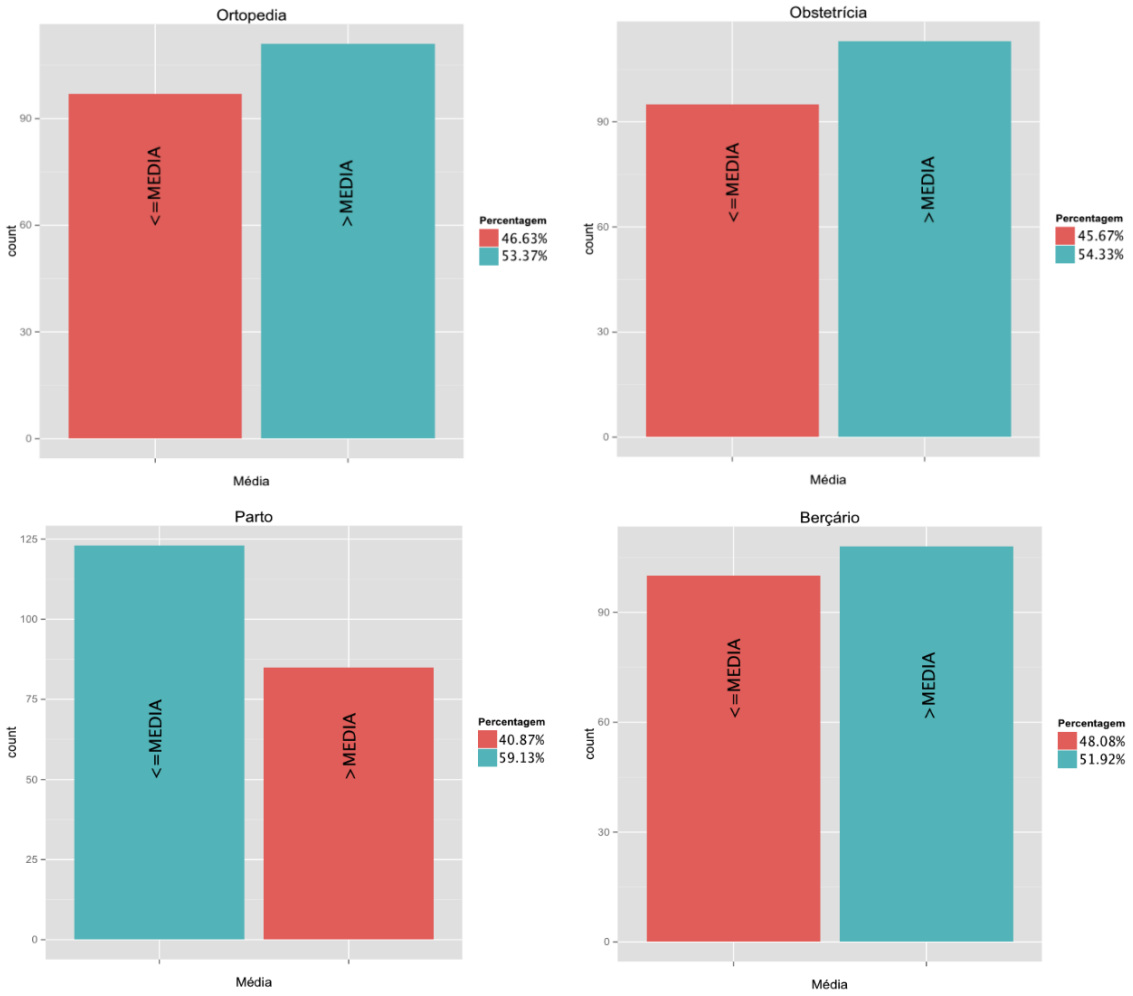

Figura 12 – Frequência de classes (Média)

Quartis: Este método recorre a uma medida de tendência não central. Os quartis identificados como valores da variável que dividem a distribuição de frequências em quatro partes iguais (Reis, 2008). Então, perante uma amostra $x = \{x1, x2, x3, \dots, xn\}$ pretende-se determinar um conjunto de quatro classes destintas, a partir dos quartis. Em primeira instância foi necessário ordenar o conjunto de dados por ordem crescente e posteriormente identificar o valor máximo e mínimo.

Para determinar a primeira classe é necessário identificar o número mínimo de doentes com alta da amostra e o número de doentes com alta que se encontram na posição que corresponde ao primeiro quartil, para determinar esse valor foi necessário recorrer ao seguinte cálculo $xn = (25\% * n)/100\%$. A primeira classe definida apresentas-se como um intervalo entre dois valores, [min, $xn = (25\% * n)/100\%$], a classe seguinte é determinada através do intervalo] $xn = (25\% * n)/100\%, xn = (50\% * n)/100\%$], a terceira classe é definida através do intervalo]$xn = (50\% * n)/100\%, xn = (75\% * n)/100\%$], e por fim a quarta classe] $xn = (75\% * n)/100\%, max$].

A Figura 13 representa as frequências das classes (Quartis) criadas para os vários serviços.

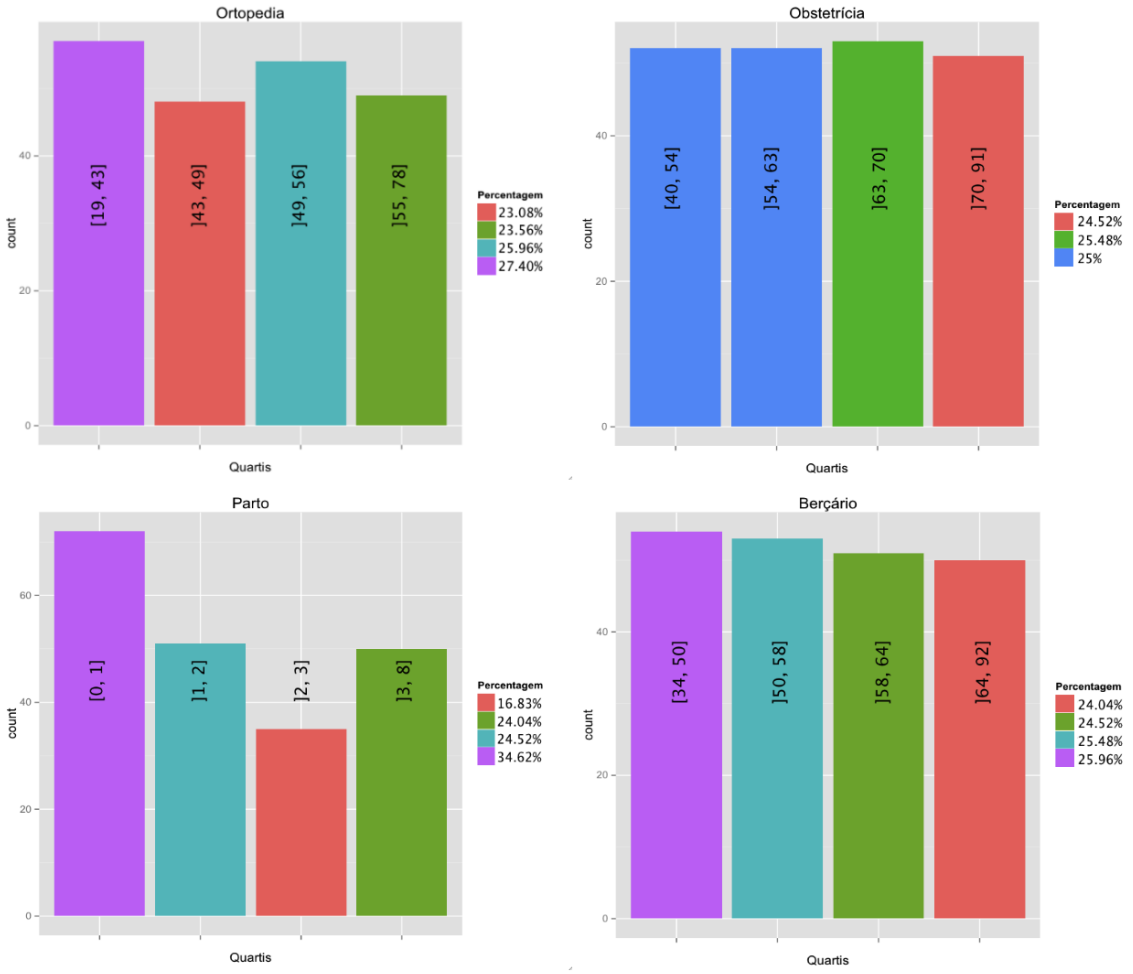

Figura 13 – Frequência de classes (Quartis)

Média e Desvio Padrão: este método requer o mesmo cálculo efetuado no primeiro método descrito, só que necessita do valor do Desvio Padrão. Uma vez identificado o valor da média, expresso por $\overline{x}$, foi calculado o Desvio Padrão, o seu valor foi determinado por:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(xi - \overline{x})^2}{n-1}}$$

A variável xi representa o número de altas hospitalares por semana e n, o número de semanas.

A parir dos valores determinados, s e $\overline{x}$ foram definidas classes de valores, em que estas representam intervalos de valores.

Um intervalo pode ser definido da seguinte forma, $]\overline{x},\ \overline{x} + s]$, neste caso é necessário ter em atenção o intervalo definido, pois este não pode possuir valores superiores ao max ou

inferiores ao min, o que em determinados casos se pode verificar, mas terão de ser reajustados. Por isso é que esta método não possui um processo trivial para determinar classes.

A Figura 14 representa as frequências das classes (Média-Desvio Padrão) criadas para os vários serviços.

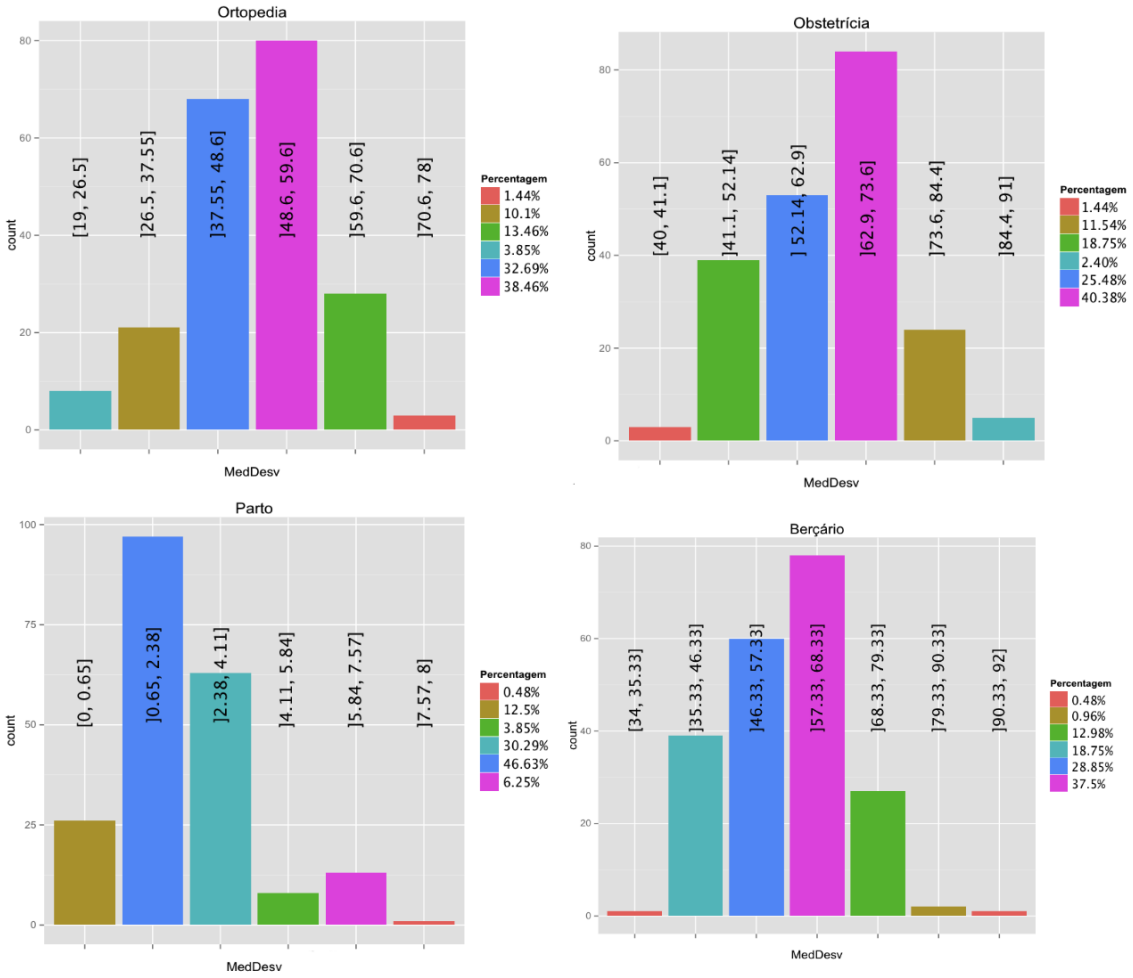

Figura 14 – Frequência de classes (Média - Desvio Padrão)

Regra de Sturges: por último, foram determinadas as classes com recurso à Regra de Sturges. O número de classes foi determinado a partir da expressão, $k = 1 + 3{,}322 * log_{10}208$ e a amplitude de cada classe foi determinada em função de, $amplitude = (\max - min)/k$ (Pestana & Velosa, 2008).

A primeira classe corresponde ao intervalo que é determinado por meio do valor min e da aplicação do seguinte cálculo, $h1 = min + amplitude$, o primeiro intervalo é apresentado como $k1 = [min, h1]$. O segundo intervalo será apresentado como $k2 =]h1, h1 + amplitude]$, os restantes intervalos são calculados segundo este processo, consecutivamente até determinar o conjunto de classes k.

A Figura 15 representa as frequências das classes (Regra de Sturges) criadas para os vários serviços.

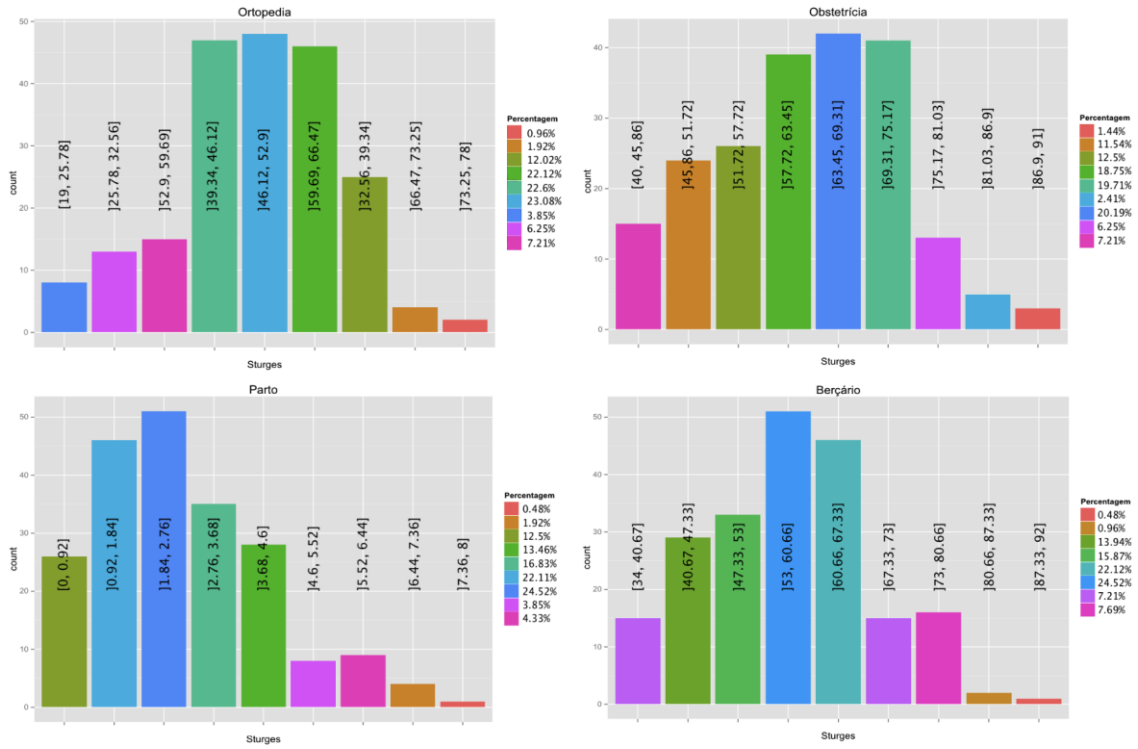

Figura 15 – Frequência de classes (Regra de Sturges)

Através do conjunto de métodos descritos, as tabelas de dados foram criadas em função do número de semanas e dos respetivos anos, ou seja, as linhas são o número de semanas (52 na totalidade), as colunas correspondem aos anos, 2009 a 2012. Esta representação de dados é definida como sendo **convencional**.

A seguinte Tabela 3 representa uma amostra da tabela de dados com recurso ao método de representação de dados convencional.

Tabela 3 – Representação de dados convencional (Obstetrícia)

NUM_SEMANA	ANO_2009	ANO_2010	ANO_2011	ANO_2012
1	56	73	66	45
2	62	73	73	54
3	47	75	54	70
4	42	70	62	70
5	41	76	72	67

Porém, foi concebido outro metódo de representação de dados, tendo como base as tabelas de dados descritas. Este método é designado por **janela deslizante**. Nesta

representação os registos estão distribuídos em função do tempo (linhas) e os atributos são precedentes da semana n (colunas).

Os valores ou as classes de altas hospitalares foram introduzidas desde a primeira semana de 2009 até a última de 2012 no campo "SEMANA_N", o que resulta em duzentos e oito registos, o campo que precede à esquerda da "SEMANA_N" designado por "SEMANA_N_1" possui o primeiro registo nulo, pois o valor da primeira semana encontra-se no segundo registo, isto repete-se mais duas vezes, dessa forma o primeiro registo possui quatro valores de altas hospitalares para diferentes semanas.

A partir deste método de representação as tabelas iriam possuir duzentos e onze registos, porém a partir deste método verificou-se que a tabela iria possuir seis registos com ausência de valores, porque o primeiro registo (linha 1 coluna 4) apenas iria possuir um único valor (não incluído o campo de contagem, NUM_SEMANA), total de doentes que saíram na primeira semana de 2009, o segundo registo iria possuir apenas dois valores (linha 2 coluna 3 e 4), total de doentes que saíram na segunda e primeira semana de 2009. Dessa feita foram eliminadas todas as linhas com ausência de valores, o que resultou em tabelas compostas por duzentos e cinco registos. Esta tabela possui cinco atributos (NUM_SEMANA, SEMANA_N_3, SEMANA_N_2, SEMANA_N_1, SEMANA_N). Seguindo a ordem apresentada das semanas, o primeiro registo possui no campo SEMANA_N o valor de altas hospitalares da quarta semana de 2009, para o campo SEMANA_N_1 o valor de altas hospitalares da terceira semana de 2009, para o campo SEMANA_N_2 o valor de altas hospitalares da segunda semana de 2009 e no campo SEMANA_N_3 encontra-se o valor de altas hospitalares da primeira semana de 2009. No segundo registo é introduzido o valor da quinta semana de 2009 e por conseguinte é descartado o valor da primeira semana, passando a estar o valor da segunda semana, assim sendo o segundo registo possui os valores da quinta semana até a segunda. Este processo repete-se duzentas e cinco vezes.

A seguinte Tabela 4 representa uma amostra da tabela de dados com recurso ao método de representação de dados janela deslizante.

Tabela 4 – Representação em janela deslizante (Obstetrícia)

NUM_SEMANA	SEMANA_N_3	SEMANA_N_2	SEMANA_N_1	SEMANA_N
1	56	62	47	**42**
2	62	47	**42**	41
3	47	**42**	41	46
4	**42**	41	46	58
5	41	46	58	51

No presente estudo foram concebidos trinta e seis tabelas de dados destintas para as duas abordagem de DM. Para a Regressão apenas foram desenvolvidas duas tabelas destintas para cada um dos serviços. Os modelos de DM que seguem a abordagem de Regressão terão como recurso na sua totalidade, oito tabelas de dados.

As tabelas de dados usadas nos modelos de Classificação foram todas aquelas que nesta fase se apresentam com valores em formato de classe. Os quatro métodos usados para criar classes foram implementadas em tabelas de dados convencionais, apenas o método da média é que não foi usada em tabelas do tipo janela deslizante. Isto resulta em sete tabelas de dados por serviço, resultando assim em vinte e oito tabelas de dados que serão usados pelos modelos de Classificação.

As tarefas de manipulação, organização e acesso aos dados foram conseguidas através do sistema de gestão de base de dados *MySQL*.

3.5. Modelação

Depois de terminadas as três primeiras fases da abordagem CRISP-DM, deu-se inicío à presente fase, a Modelação.

Tal como já foi descrito anteriormente, foram usadas duas abordagem de DM, Classificação e a Regressão, com o objetivo de resolver problemas de Previsão. Na Classificação foram usadas variáveis qualitativas ordinais, as classes criadas seguem a ordenação ascendente, na Regressão as variáveis são quantitativas discretas, tratando-se de valores inteiros.

As técnicas usadas nos modelos de Classificação foram, SVM, AD e NB, já para os modelos de Regressão foram usados as técnicas SVM e AR.

De forma a implementar um mecanismo de teste para os modelos, foram selecionados dois métodos de amostragem, *10-folds Cross Validation* (10-*folds* CV) e o *Leave-One-Out Cross*

Validation (LOOCV). O *10-folds* CV foi implementado, devido aos bons resultados que tem demostrado em base de dados multidisciplinares (Witten et al., 2011). Quanto ao LOOCV, a sua implementação prende-se pelo facto de ser um método adequado para base de dados com apenas algumas dezenas de registos (Refaeilzadeh et al., 2009), o que neste estudo é algo que esta evidente, as tabelas de dados apenas possuem 52 ou 205 registos.

Todas as técnicas foram submetidas à função ***tune*,** função que provem da biblioteca *e1071*, tendo como objetivo realizar a pesquisa em grelha dos intervalos dos hiperparâmetros previamente fornecidos e identificará consecutivamente o melhor modelo e respetivos hiperparâmetros.

Para a técnica SVM, foram usados dois *kernels* destintos, *Radial-Basic Function* (RBF) e Linear. Perante os diferentes *kernels* usados, surgiu a necessidade de realizar diferentes parametrizações, pois os seus hiperparâmetros diferem de *kernel* para *kernel*. Em função dos *kernels* usados pelos SVMs foi determinado um intervalo de valores que o parâmetro custo C poderia assumir, o seu intervalo foi definido em função dos valores obtidos pela potência $2^{(1,..,4)} = [2, \dots, 16]$, em que $C > 0$.

O parâmetro custo C, introduz alguma flexibilidade na separação das categorias de forma a controla o compromisso (*trade-off)*, entre erros no treino e a rigidez das margens (Kantardzic, 2011).

O hiperparâmetro *Gamma* γ, foi definido tal como C, o seu intervalo de variação foi determinado em função dos valores obtidos pela potencia, $2^{(-1,0,1)} = [0.5, 1, 2]$, a sua parametrização foi usada para o *kernel* RBF. O valor de γ determina a curvatura da fronteira de decisão (Ben-Hur & Weston, 2010). Os intervalos do *Custo* e do *Gamma*, foram definidos após realizar 10 execuções dos modelos em que os parâmetros se encontravam em *default,* depois de observados os valores dos respetivos parâmetros *Custo* e *Gamma* é que posteriormente foram criados os intervalos definidos pela potencia de dois.

A função de perda $\varepsilon - insensível$ foi introduzida nos SVMs. A sua utilização proporcionará aos modelos ignorar erros, ou seja, será criado um tubo em volta dos resíduos e os erros são ignorados. Se o valor do ε for muito elevado pode levar a ocorrência de *overfitting* (Witten et al., 2011). O seu valor será determinado por *default*, sendo o valor de ε igual a 0,1.

As ADs e ARs foram usadas para realizar Previsão através das abordagens de Classificação e de Regressão, a implantação desta técnica foi conseguida através do algoritmo CART.

Foram usados dois métodos de seleção de atributos ou regras de *splitting*, *Information Gain* (IG) e o *Gini Index* (GI), na qual o IG não foi usado nos modelos de Regressão.

A medida de seleção de atributos IG determina qual o atributo com maior ganho de informação e usa-o para fazer a divisão de um nó (Han et al., 2012). O IG é determinado pela diferença entre o requisito de informação original (isto é com base em apenas a proporção de classes), e o novo requisito (ou seja, obtidos após o particionamento de A), está diferença é expressa por: $Gain(A) = Info(D) - Info_A(D)$. O atributo A que tiver maior ganho de informação, $Gain(A)$ é o atributo de divisão no nó n (Han et al., 2012).

O objetivo do GI é calcular o seu valor para cada atributo, utilizando para o nó o atributo com menor índice de impureza (Manuel Santos & Azevedo, 2005). O índice de GI mede a impureza de D, a partir de uma partição de dados ou de um conjunto de atributos de formação $Gini(D) = 1 - \sum_{i=1}^{m} pi^2$, onde pi corresponde á probabilidade de um atributo de D pertencer à classe Ci, esse valor é estimado por $|Ci \cap D|/|D|$. O somatório é calculado em função de m classes (Han et al., 2012).

Por último o algoritmo NB apenas foi usado para realizar Previsão através da abordagem de Classificação, onde não foi realizada qualquer parametrização manual para este algoritmo, este apenas recorreu a função **tune** para identificar o método de amostragem que deveria seguir, estando este previamente determinado.

Nos modelos de Classificação como nos de Regressão as variáveis de entrada variam em função do método de representação de dados, podendo ser do tipo **convencional** ou **janela deslizante**. Na representação de dados do tipo convencional, as variáveis de entrada foram ANO_2009, ANO_2010 e ANO_2011, o *target* usado foi o atributo ANO_2012, com base neste cenário foram realizadas previsões para o ano 2012 em função dos anos anteriores.

Para os modelos que usaram a representação de dados do tipo janela deslizante, os mesmos possuíam como valores de entrada a SEMANA_N_3, SEMANA_N_2 e SEMANA_N_1, o *target* usado foi o atributo designado por SEMANA_N, com base neste cenário foram realizadas previsões para a SEMANA_N em função das semanas anteriores. Ou seja, a previsão realizada para a quarta semana do ano 2009 é conseguida com os valores das três semanas antecedentes.

Para a abordagem de Classificação foram desenvolvidos vários modelos:

- Quatro formas de criar classes para cada um dos quatro serviços:
  - Média;
  - Quartis;
  - Média e Desvio Padrão;
  - Regra de Sturges.
- Dois métodos destintos de representação de dados:
  - Convencional;
  - Janela deslizante.
- Três técnicas de DM que resulta em cinco aplicações:
  - Nos SVMs foram usados dois *kernels* (Linear e RBF);
  - Nas ADs foi aplicado o IG e o GI no que resultou em duas aplicações;
  - Para as NBs apenas foi realizado um tipo de aplicação, *default*.
- Dois métodos de amostragem:
  - 10-*folds* CV;
  - LOOCV.

Estas componentes conjugadas resultam em 320 modelos para realizar previsões através da Classificação.

Na abordagem de Regressão foram usados:

- Dois métodos de representação de dados para quatro serviços hospitalares:
  - Convencional;
  - Janela deslizante.
- Duas técnicas de DM:
  - Nos SVMs foram usados dois *kernels* (Linear e RBF);
  - Nas ARs foi aplicado o método de seleção GI.
- Dois métodos de amostragem:
  - 10-*folds* CV;
  - LOOCV.

Estas componentes conjugadas resultam em 48 modelos para realizar previsões através da Regressão.

Os modelos desenvolvidos podem ser representados pela seguinte expressão, $M_n = A_f + S_i + C_x + MRD_z + TDM_y + MA_k$. O modelo M_n deve pertencer a uma abordagem (A) Classificação ou Regressão, é composto por um serviço (S), um tipo de classe (C), um método de representação de dados (MRD), uma técnica de DM (TDM) e por um método de amostragem (MD).

$$A_f = \{Classificação_1, Regressão_2\}$$
$$S_i = \{Ortopedia_1, Obstetrícia_2, Parto_3, Berçário_4\}$$
$$C_x = \{Média_1, Quartis_2, Média\ e\ Desvio\ Padrão_3, Sturges_4, Número_5\}$$
$$MRD_z = \{convencional_1, janela\ deslizante_2\}$$
$$TDM_y = \{SVML_1, SVMRBF_2, ADGI_3, ADIG_4, ARGI_5, NB_6\}$$
$$MA_k = \{10 - folds\ CV_1, LOOCV_2\}$$

Através da notação de representação de modelos de DM é possível apresentar um exemplo de um modelo implementado. O modelo de DM que segue a abordagem de Regressão, com recurso aos dados do serviço de Obstetrícia, com o método de representação de dados do tipo convencional, usa a técnica SVM com *kernel* RBF e o método de amostragem 10-*folds* CV, é expresso por: $M_1 = A_2 + S_2 + C_5 + MRD_1 + TDM_2 + MA_1$. Os modelos criados seguem a ordem do seguinte **Algoritmo 1** representado em **pseudo-código**. Através do algoritmo procedeu-se a codificação em linguagem R.

Algoritmo 1 Conceção de modelo de DM

Requer: ligação à base de dados

1: Seleção da técnica de modelação
2: Determinar método de amostragem (número de partições)
3: Determinar o número de vezes que o treino deve ser repetido
4: Sintonizar valores de controlo
5: Selecionar *target*
6: Selecionar *inputs*
7: Aplicar a técnica de DM
8: Identificar o resultado

3.6. Avaliação

Depois de aplicados os modelos de DM foi necessário realizar avaliações dos respetivos modelos apresentados. Para avaliar os resultados apresentados pelos modelos de DM que recorrem a abordagem de Classificação foi usada a medida de Precisão ou Acuidade, já para os modelos que recorreram a abordagem de Regressão foram usadas as métricas *Mean Squared Error* (MSE), *Mean Absolute Error* (MAE) e *Relative Absolute Error* (RAE).

Uma vez que os resultados apresentados pelos modelos de DM dependem da divisão dos subconjuntos mutuamente exclusivos, foram implementados dois procedimentos 10-*folds* CV e LOOCV. Na aplicação dos respetivos procedimentos de divisão foram realizadas dez execuções para cada um deles, quer para os modelos de Classificação quer para os de Regressão.

Foram realizadas 100 experiências para cada configuração de teste em modelos que usam o procedimento 10-*folds* CV. Já o número de experiencias de configuração de teste para modelos que usam o procedimento LOOCV depende do tipo representação de dados, na representação de dados convencional foram realizadas 520 experiências de teste e na representação de dados do tipo janela deslizante, foram realizadas 2050 experiências.

Na seguinte Tabela 5 estão representados os valores das Acuidades obtidas pelos melhores modelos de Classificação. Os restantes resultados obtidos pelos modelos de Classificação podem ser consultados no Apêndice B.

Tabela 5 – Avaliação dos modelos de Classificação

Modelo	Serviço	Técnica	Método de Amostragem	Classe	Representação de dados	Acuidade
M_1	S_1	TDM_2	$MA_{1,2}$	C_4	MRD_1	≈82.69%
M_2	S_2	$TDM_{1,2}$	$MA_{1,2}$	C_4	MRD_1	≈82.69%
M_3	S_3	TDM_2	$\mathrm{MA}_{1,2}$	C_4	MRD_1	≈94.23%
M_4	S_4	TDM_2	$MA_{1,2}$	C_4	MRD_1	≈90.38%

A técnica usada pelos modelos de DM que melhore resultados originou foi o SVM, a diferença de resultados obtidos pelos modelos de DM que usaram a técnica SVM é significativa quando comparados com as restantes. Os resultados apresentados na Tabela 5 demostram que a aplicação dos SVMs foi incontestavelmente superior às restantes técnicas. Em 28 aplicações

dos SVMs os modelos gerados foram superiores em 21 vezes o que resulta numa taxa de superioridade de 75%.

Os métodos de amostragem usados não demostraram ser determinantes para os resultados apresentados, ambos foram equivalentes.

Dos dois métodos usados para realizar a representação de dados o método convencional foi o que sempre proporcionou a obtenção dos melhores modelos.

Através da Tabela 5 verifica-se que a Regra de Sturges usada na criação de classes, foi a que proporcionou melhores resultados. Os resultados obtidos (acuidades) pelos modelos de Classificação que usaram classes definidas pela Regra de Sturges, foram superiores em cerca de ≈31,25%. Ou seja das 80 utilizações da Regra de Sturges nos modelos de Classificação, 25 obtiveram melhores resultados que os restantes métodos de criação de classes.

Os resultados obtidos pelos modelos de DM, demostram valores de precisão bastante aceitáveis, embora apenas os resultados dos modelos M_3 e M_4 serão capazes de suportar decisões. Para os gestores os resultados obtidos através da avaliação dos modelos de previsão (valores das acuidades), poderão ser usados para tomarem decisões. Porém, os valores da Sensibilidade e Especificidade de cada classe obtidos pelos melhores modelos encontra-se no Apêndice C, o valor informacional que estas métricas expressão são capazes de determinar os intervalos de doentes que realmente tiveram alta e os que não obtiveram alta.

De seguida, na Figura 16 serão apresentados as classes resultantes dos modelos de previsão e respetiva frequência. As classes são representadas pelo seu valor máximo (ponto a verde) e pelo seu mínimo (ponto a azul), se os valores reais (pontos a vermelho) estiverem inseridos entre o máximo e o mínimo significa que a previsão da classe foi devidamente realizada, ou seja, os valores reais estavam enquadrados nos limites da classe prevista. As classes apresentadas foram obtidas através da Regra de Sturges.

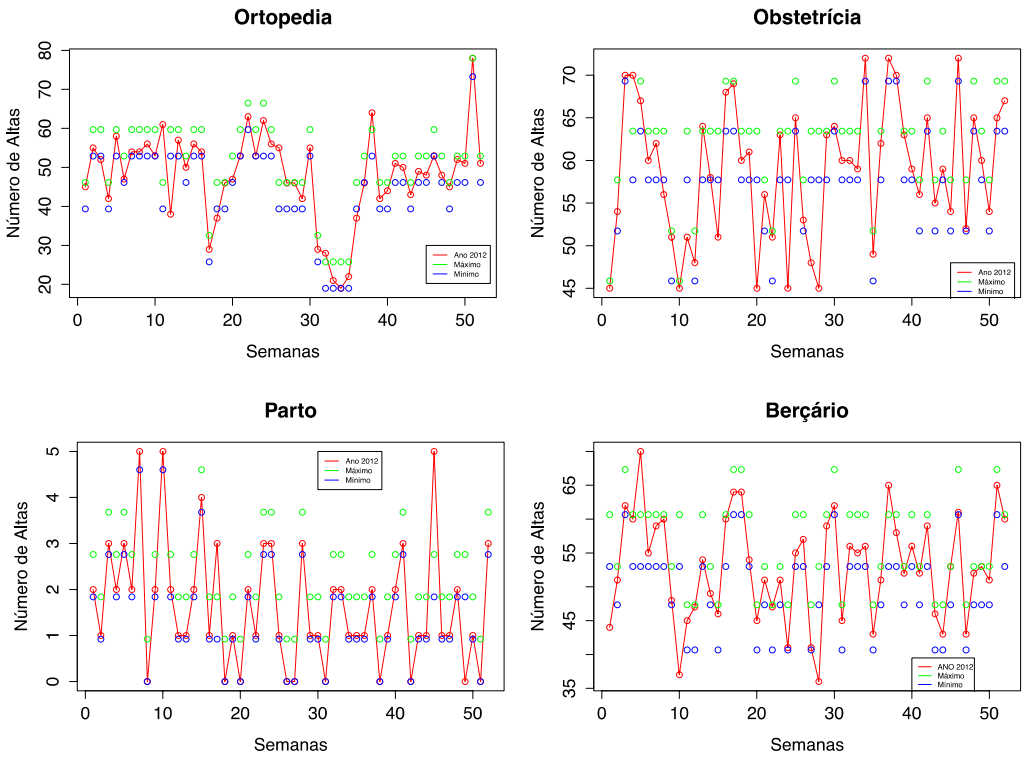

Figura 16 – Previsões (Classificação)

A representação gráfica da curva *Receiver Operating Characteristic* (ROC) para os melhores modelos de Classificação não foi realizada, devido ao elevado número de diferentes classes previstas, onde são sempre superior a 2 classes. Pois esta representação é feita normalmente para 2 classes (Ferri, Orallo, & Salido, 2003).

Uma vez apresentados os resultados dos modelos de Classificação, é imperativo apresentar os resultados obtidos pelos modelos de Regressão, a seguinte Tabela 6 expressa esses mesmos resultados. Os restantes resultados podem ser observados no Apêndice D.

Tabela 6 – Avaliação dos modelos de Regressão

Modelo	Serviço	Técnica	Método de Amostragem	Representação de dados	MAE	MSE	RAE
M_5	S_1	TDM_2	$MA_{1,2}$	MRD_1	≈4,030	≈34,432	≈74,075%
M_6	S_2	TDM_2	$MA_{1,2}$	MRD_1	≈1,863	≈10,925	≈38,260%
M_7	S_3	TDM_2	$MA_{1,2}$	MRD_1	≈0,619	≈0,989	≈96,894%
M_8	S_4	TDM_2	$MA_{1,2}$	MRD_1	≈2,433	≈20,150	≈53,818%

A técnica que proporcionou a obtenção dos melhores resultados apresentados pelos modelos de Regressão foi o SVM, com o *kernel* RBF. Através da representação gráfica das curvas *Regression Error Characteristic* (REC) é possível identificar a superioridade da aplicação

dos SVMs em relação as ARs para os melhores modelos desenvolvidos, tal observação pode ser constatada na Figura 17.

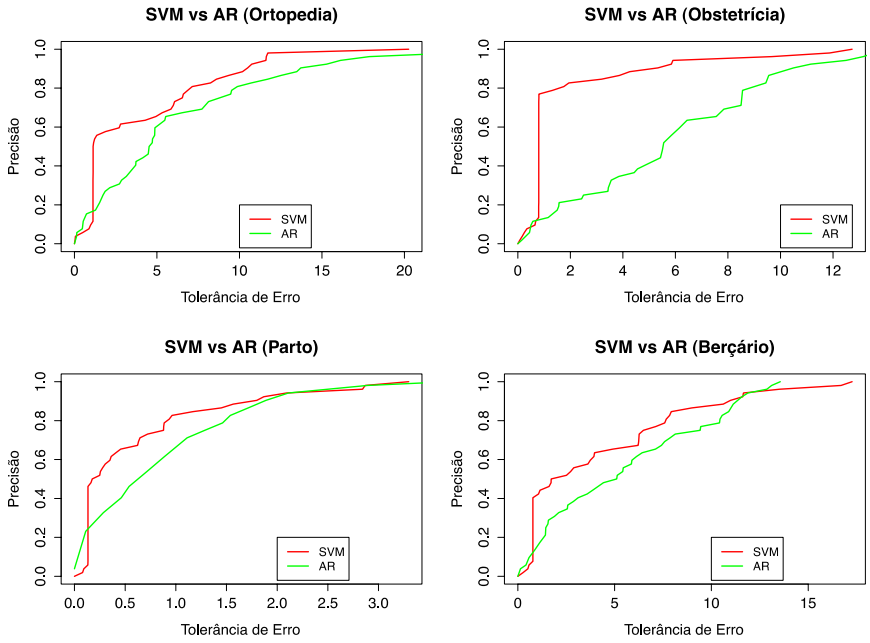

Figura 17 – Curvas REC para modelos de previsão

Os melhores resultados apresentados foram igualmente obtidos pelos dois métodos de amostragem, 10-*folds* CV e LOOCV.

Quanto à representação dos dados mais uma vez, verificou-se que o método convencional proporcionou a obtenção dos melhores resultados.

Através dos valores expressos pela métrica MAE é possível observar que os erros médios absolutos são significativamente baixos para os modelos M_6 e M_8.

De forma a identificar as previsões realizadas com valores muitos discrepantes dos valores reais, ou seja, identificar a existência de *outliers*, foi usada a métrica MSE. Os valores apresentados pela métrica demostram que nos vários serviços ocorrerem previsões desajustadas para algumas semanas, esta observação é pertinente uma vez que se pretende identificar valores o menos discrepantes possíveis.

Ainda de forma a realizar a avaliação e validação dos modelos gerados recorreu-se à métrica RAE, esta auxiliou na identificação dos melhores modelos. Para os modelos M_5 e M_7 verifica-se que o desempenho de cada modelo foi similar ao previsor médio *naïve*, em particular para o serviço de Parto (M_7), pois o seu valore de RAE é muito próximo de 100%. Os modelos M_6 e M_8, representam bons modelos de Regressão, é ainda de salientar o resultado obtido pelo

o serviço de Berçário, que se encontra significativamente a baixo dos resultados apresentados pelos restantes serviços, com o valor de RAE ≈38,260%.

De forma a demostrar a qualidade dos resultados obtidos, foram realizadas representações gráficas para os vários serviços estudados, os resultados podem ser observados na Figura 18.

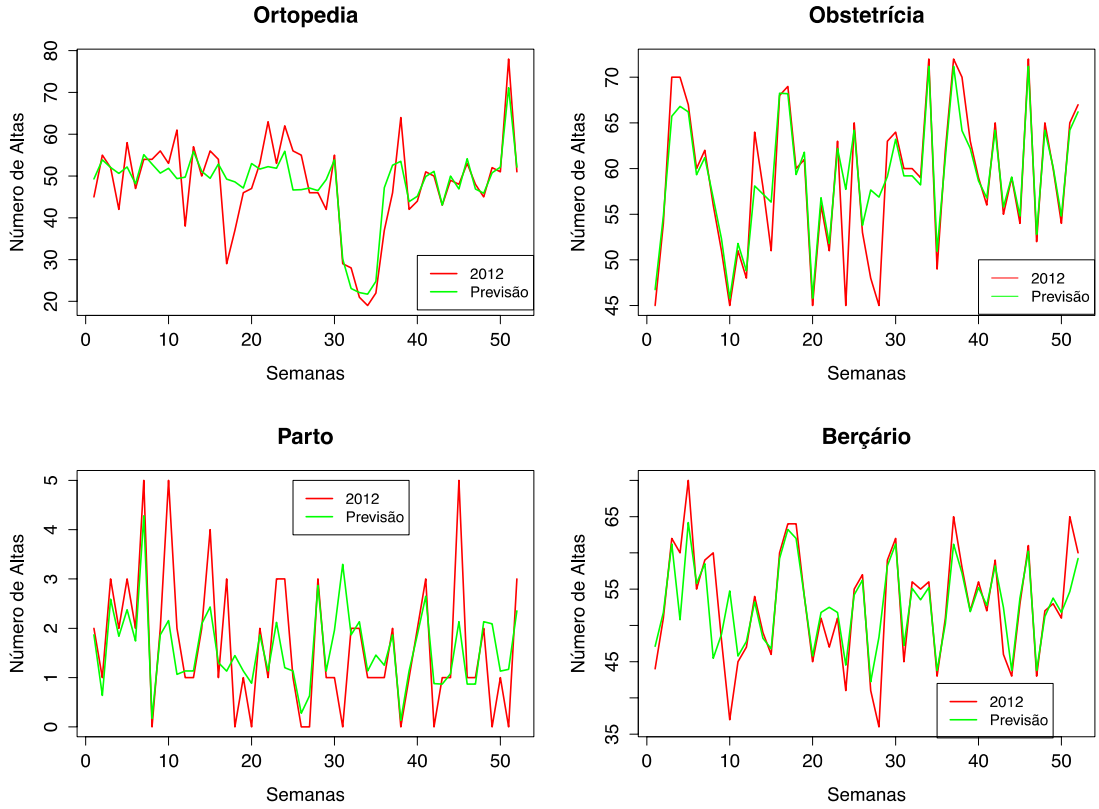

Figura 18 – Previsões (Regressão)

Depois de realizada a avaliação dos modelos de DM pelas abordagens de Classificação e Regressão é da maior importância identificar o modelo que melhor se ajusta na resolução do presente problema. De entre os modelos gerados que seguem a abordagem de Classificação, destaca-se pela positiva, o modelo M_3 apresentado na Tabela 5, referente ao serviço de Parto, obteve uma excelente acuidade ≈94,23%, porém na Regressão foi o serviço que obteve um resultado menos positivo, comparativamente aos restantes.

Tornou-se pertinente determinar qual das abordagens de DM seguidas apresenta melhor precisão para determinado serviço. Através da revisão de literatura não foi identificada qualquer métrica que proporcionasse tal comparação. Então de modo experimental foi necessário recorrer a uma métrica que se pudesse aplicar aos modelos gerados por Classificação e Regressão.

Ficou desde logo determinado que para realizar a respetiva comparação seria necessário recorrer aos valores que delimitam as classes criadas que originaram os melhores modelos de previsão por Classificação. Assim sendo, foi necessário converter classes previstas em dois valores reais. Os valores correspondem aos máximos e mínimos que se encontram representados na Figura 16.

Posteriormente, foi desenvolvido um algoritmo capaz de determinar a diferença mínima entre o valor real e o valor mais próximo, dos dois valores que determina a amplitude das classes previstas. O seguinte pseudo-código representa o algoritmo 2 usado para determinar o valor de MAE através dos valores das classes previstas por Classificação.

Algoritmo 2 Calcular o erro médio absoluto

```
Requer: valores máximos e mínimos das classes previstas
Função determinar erro mínimo
    Para todas as semanas Fazer
        Se distância máxima for superior à distância mínima Então
            Inserir valor máximo no vetor MAX
        Se não
            Inserir valor mínimo no vetor MAX
        Fim se
        Se distância máxima for inferior à distância mínima Então
            Inserir valor máximo no vetor MIN
        Se não
            Inserir valor mínimo no vetor MIN
        Fim se
    Fim para
    Somar valores mínimos do vetor MIN e dividir por 52
    Retornar MAE
Fim função
```

Para o modelo M_1 foi determinado o valor mínimo possível de MAE ≈2,468, o valor desta métrica traduz o número médio de altas previstas erradamente, este valor é inferior ao valor determinado pelo modelo M_5, com MAE ≈4,030. O MAE calculado em função dos valores (que determinam um classe de valores) mais próximos dos valores reais pelo modelo M_1 é inferior ao MAE obtido pelo modelo M_5. O erro determinado pelo algoritmo representa ser um erro relativamente pequeno quando comparado com o valor médio de altas semanais do serviço de Ortopedia, $(2{,}468 * 100)/48{,}6$ ≈5,078%. Com uma acuidade de 82,69%, o modelo de Classificação M_1, apresenta melhores resultados.

Para o modelo M_2 foi determinado o valor mínimo possível de MAE ≈2,642, este valor é superior ao valor determinado pelo modeloM_6, com MAE ≈1,862. O valor de MAE obtido pelo modelo de Regressão M_6, quando comparado com o valor médio de altas semanais do serviço de Obstetrícia, representa um erro pequeno de $(1{,}863 * 100)/62{,}9$ ≈2,962%. Comparando as duas abordagens, torna-se evidente que o modelo de Regressão M_6, é mais preciso na realização de previsões.

Para o modelo M_3 foi determinado o valor mínimo possível de MAE ≈0,219, este valor é inferior ao valor determinado pelo modelo M_7, com MAE ≈0,619. O valor de MAE obtido pelo modelo M_3 quando comparado com o valor de médio de altas semanais do serviço de Parto, representa um erro não muito significativo de $(0{,}219 * 100)/2{,}4$ ≈9,125%. Comparando as duas abordagens, o modelo de Classificação M_3, é mais preciso na realização de previsões.

Por último, para o modelo M_4 foi determinado o valor mínimo possível, de MAE ≈2,428. O valor de MAE obtido pelo modelo de M_4 é inferior ao valor obtido pelo modelo de M_8, MAE ≈2,433. O valor de MAE determinado pelo modelo M_4 quando comparado com o valor médio de altas semanais do serviço de Berçário, representa um erro pequeno de $(2{,}428 * 100)/57{,}3$ ≈4,237%. Para o serviço de Berçário o modelo M_4, demostra ser o mais preciso na realização de previsões.

A Tabela 7 representa o melhor modelo de DM para cada serviço e respetiva abordagem.

Tabela 7 - Melhores modelos de DM

Modelo	Serviço	Técnica	Método de Amostragem	Representação de dados	Abordagem
M_1	S_1	TDM_2	$MA_{1,2}$	MRD_1	A_1
M_6	S_2	TDM_2	$MA_{1,2}$	MRD_1	A_2
M_3	S_3	TDM_2	$MA_{1,2}$	MRD_1	A_1
M_4	S_4	TDM_2	$MA_{1,2}$	MRD_1	A_1

3.7. Implementação

O trabalho de investigação realizado demostra ser relevante e pode ser entendido como uma oportunidade para o desenvolvimento de sistemas capazes de realizar previsões de altas hospitalares, através de modelos de DM.

Os resultados obtidos pelos modelos de previsão demostram ser precisos, e segundo esta evidência o presente estudo poderá servir como base para o desenvolvimento de um SAD, capaz de operar em ambiente e tempo real. No entanto, será de extrema importância alargar o presente estudo aos restantes serviços hospitalares, serviços esses que deverão englobar internamentos e identificar se as precisões obtidas nesses serviços são significativas. A monitorização dos modelos de DM deve ser realizada. Para que o SAD se mantenha atualizado e optimizado, será necessário a introdução de novos dados e controlar a parametrização dos modelos de previsões.

O desenvolvimento do respetivo SAD irá possibilitar aos gestores obterem informações precisas referentes às altas hospitalares e deste modo será possível realizar uma melhor gestão das camas hospitalares pelos respetivos serviços que o hospital dispõem. Porém o desenvolvimento do SAD não se encontra inserido no âmbito deste projeto de dissertação.

3.8. Sumário

Antes de ter início a componente prática, foi necessário determinar qual seria a linha orientadora a seguir, para tal foi estudada a metodologia CRISP-DM. Desta forma o desenvolvimento prático seguiu a metodologia escolhida, visto este ser um projeto de DM.

De seguida foi realizado o levantamento de ferramentas de DM. Do conjunto de ferramentas identificadas, optou-se pelo uso do ambiente de programação **R.** A escolha do ambiente **R** para solucionar o problema deve-se ao facto de este se apresentar como uma das ferramentas mais usadas para solucionar problemas de DM e de possuir um conjunto de recursos capazes de satisfazer as necessidades desta dissertação.

O propósito do estudo foi identificar se era possível a partir das abordagens de Classificação e Regressão, gerar modelos capazes de prever com elevada precisão as altas hospitalares nos vários serviços do CHP. Neste aspeto, o estudo abrangeu apenas quatro serviços, Ortopedia, Obstetrícia, Parto e Berçário, uma vez que apenas estes apresentavam a totalidade dos dados para o período especificado (2009 – 2012).

Através dos dados fornecidos e dos objetivos propostos surgiu a necessidade de criar classes para realizar previsões através de Classificação, como tal foram aplicado 4 métodos de criação de classes: Média, Quartis, Média-Desvio Padrão e Regra de Sturges. Depois de

determinados os métodos de criação de classes foram aplicados em 2 métodos de representação de dados: Convencional e Janela Deslizante.

No processo de modelação foram usadas 4 técnicas de DM: AD, AR, SVM e NB, com diferentes parametrizações

O trabalho prático desenvolvido foi bastante aprofundado, levando ao teste e exploração de diversos cenários/técnicas. Foram desenvolvidos 320 modelos para realizar previsões com a Classificação e 48 modelos com Regressão, o que no total foram desenvolvidos 368 modelos destintos. De todos os modelos foram selecionados os 4 melhores modelos (um para cada serviço) para a abordagem de Classificação e Regressão.

4. Discussão de Resultados

É importante referir que o presente documento não representa todo o trabalho realizado, foram desenvolvidos modelos de Classificação para realizar previsões por Semestre, Trimestre, Estações do Ano e ainda Diariamente, só que, em reunião com os orientadores, chegou-se à conclusão que a agregação dos dados deveria ser feita semanalmente. Pois a previsão semanal possui o detalhe adequado para os gestores realizarem a planificação e gestão de doentes e camas.

Os resultados apresentados demostram ser bastante aceitáveis, devido às avaliações das previsões realizadas. Nos modelos de Classificação as melhores previsões resultaram em acuidades superiores a ≈82%. Os serviços de Ortopedia e Obstetrícia são os que apresentam os piores resultados dos quatro serviços, com ≈82,69% de acuidade, já os resultados obtidos nos serviços de Parto e Berçário são de ≈94,23% e ≈90,38%. As previsões realizadas para os serviços Parto e Berçário apresentam resultados suficientemente satisfatórios para suportar a decisão.

Nos modelos de Regressão as previsões realizadas foram submetidas a três métrica de avaliação *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE) e *Relative Absolute Error* (RAE), para compara os modelos recorreu-se a representação gráfica da curva REC. Os resultados obtidos com a métrica MAE apresentaram um baixo nível de erro para os quatro serviços em especial para os serviços de Obstetrícia e Berçário, com resultados de ≈1,863 e ≈2,433 doentes. Estes erros correspondem a um valor muito pequeno quando comparado com o valor médio das altas de doentes destes dois serviços. Para os serviços de Ortopedia e Parto os resultados também são bastante aceitáveis.

A aplicação da métrica MSE demostra que as previsões realizadas apresentam valores desajustados com os valores reais, isto demostra que existem *outliers* nas previsões realizadas, porém são pouco frequentes, mesmo assim este aspeto é importante pois em futuras previsões é necessário ter em atenção este tipo de ocorrências pois pode induzir os gestores a cometerem erros.

Os resultados obtidos com a métrica RAE demostram que foram obtidos bons modelos de Regressão, o serviço que apresenta o resultado menos positivo foi o de Parto com um valor ≈96,894%, este valor representa que o modelo teve um desempenho próximo do previsor médio

naïve. Para os restantes serviços os valores de RAE variam entre ≈38,26% e ≈53,818%, o que são bastante bons. A representação gráfica da curva *Regression Error Characteristic* (REC) demostra que os *Support Vector Machine* (SVM) foram superiores às Árvores de Regressão (AR).

As previsões realizadas através da Classificação demostram ser mais precisas para os serviços Ortopedia, Parto e Berçário, porém a Regressão demostra ser mais precisa na previsão de altas hospitalares para o serviço de Obstetrícia. Os resultados apresentados pelas previsões por Regressão não devem ser descartados para trabalhos futuros, pois as previsões realizadas são significativas.

Dos dois métodos de amostragem usados 10-*folds Cross Validation* (10-*folds* CV) e *Leave-One-Out Cross Validation* (LOOCV), o 10-*folds* CV é o mais adequado para o estudo realizado, não pelos resultados obtidos, pois as diferenças observadas são muito pequenas e muitas das vezes são iguais, mas sim pelos tempos de execução dos modelos. Os modelos que usam método de amostragem LOOCV necessitam de bastante tempo para terminar a sua execução.

É necessário também referir que o método de representação de dados convencional originou sempre melhores modelos quer na Classificação ou Regressão, assim sendo o método de janela deslizante não se verificou relevante.

Uma vez que os dados usados são reais, a inclusão destes modelos num Sistema de Apoio a Decisão (SAD) torna-se espectável. Assim o conhecimento gerado a partir da utilização de DM deve ser utilizado de modo a que possa influenciar a eficiência operacional, facilitando as decisões de alto nível e a prestação de serviços nos serviços estudados.

Por último, é importante referir que a ferramenta R revelou-se adequada para este trabalho. Embora a curva de aprendizagem seja mais acentuado do que algumas aplicações de DM com representação gráfica, demostrou que com algum desenvolvimento de código em R fosse possível carregar dados, aplicar as técnicas de DM e realizar a avaliação dos modelos. Isto também se verificou porque as bibliotecas usadas, *e1071* e *rminer* continham os recursos (p.e. técnicas de DM e métricas) necessários para levar o trabalho a bom porto.

5. Conclusões

Este capítulo é composto por três subcapítulos, o primeiro apresenta uma síntese do trabalho prático realizado e a relação deste com a problemática abordada durante o projeto. O segundo analisa o cumprimento dos objetivos propostos, a questão de investigação e o contributo científico que este trabalho apresenta. Por último são descritos os aspetos que devem ser tidos em contra para trabalhos futuros.

5.1. Síntese

Para que a gestão de um hospital seja devidamente conseguida é importante que os seus gestores conheçam devidamente a instituição em que laboram. Devem ser capazes de traçar um plano claro e organizado de forma a proporcionar uma gestão hospitalar eficiente e eficaz. O "*core business*" de um hospital é a otimização do tempo de internamento de doentes no hospital. Como tal a gestão hospitalar deve ter como foco a gestão de doentes e os recursos necessários, impreterivelmente as camas.

O desenvolvimento deste trabalho aborda a problemática referida, a gestão hospitalar, mais concretamente a gestão de altas hospitalares que por sua vez está diretamente ligada a gestão de camas. O estudo recaiu sobre dados reais de doentes internados (requerem cama) em quatro serviços distintos (Ortopedia, Obstetrícia, Parto e Berçário), serviços esses fornecidos pelo Centro Hospitalar do Porto (CHP).

Uma vez que este trabalho visa realizar previsões eficazes de altas hospitalares para os quatro serviços já referidos o estudo baseou-se no processo de *Data Mining* (DM). A extração do conhecimento foi concretizada através da aplicação de técnicas de DM, a sua aplicação sobre as fontes de dados resultou no procedimento chave. Na descoberta de conhecimento contido na base de dados fornecida pelo CHP.

Com o objetivo de tornar o processo de DM menos complexo recorreu-se à metodologia *Cross-Industry Standard Process for Data Mining* (CRISP-DM), e com o desenvolvimento do processo de DM verificou-se a boa capacidade de criar linhas orientadoras para o desenvolvimento de trabalhos deste cariz.

Devido à falta de conhecimento sobre qual a melhor abordagem de DM a seguir, optou-se por determinar o objetivo como sendo de Classificação e Regressão. O carácter das variáveis teve de ser alterado para que a aplicação de modelos de Classificação fosse possível. O que desta feita determinou-se que as variáveis a usar pelos modelos de Classificação deveriam de ser do tipo qualitativas ordinais e para os modelos de Regressão quantitativas discretas.

Para que os modelos de DM fossem possíveis de criar, foram selecionadas quatro técnicas: Árvores de Decisão (AD), Árvores de Regressão (AR), *Support Vector Machine* (SVM) e *Naïve Bayes* (NB); das quais as AD, SVM e NB foram usadas para modelos de Classificação e as AR e SVM para modelos de Regressão.

Como por si só a aplicação das técnicas de DM não é suficiente para determinar a precisão das previsões realizadas, foi selecionado um conjunto de métricas para avaliar a relevância dos modelos de previsão. Desta feita, verificou-se a significância dos modelos através do cálculo dos valores das acuidades para os modelos de Classificação. De forma geral os modelos podem ser considerados como sendo bons modelos de previsão, tendo particular atenção os valores obtidos pelos serviços de Parto e Berçário que obtiveram acuidades superiores ou iguais a 90%. A avaliação dos modelos de Regressão gerados debruça-se sobre a aplicação de três métricas: *Mean Squared Error* (MSE*), Mean Absolute Error* (MAE) e *Relative Absolute Error* (RAE). Para identificar o desempenho das duas técnicas AR e SVM foi usada a curva *Regression Error Characteristic* (REC). Os resultados obtidos pelas várias métricas foram satisfatórios, apresentando um elevado nível de qualidade na previsão, em especial os serviços de Obstetrícia e Berçário, com valores de MAE de 1,863 e 2,433 doentes.

5.2. Contribuições

Depois de concluído um projeto, é importante confrontar os objetivos delineados no início do projeto com os resultados obtidos. Neste sentido, é essencial seguir uma postura crítica nas observações efetuadas entre os dois pontos referidos.

Em primeira instância é necessário referir a complexidade que o projeto apresentou. A complexidade é resultante dos objetivos propostos para o desenvolvimento do respetivo projeto e da pouca realização de estudos que incidem nesta problemática, previsões de altas hospitalares com recurso ao DM. Um aspeto bastante importante a ter em consideração foi o facto de o problema ter sido abordado utilizando uma diminuta quantidade de atributos (Data, Serviço e

Número de altas). O facto de usar só este tipo de dados tinha como principal objetivo perceber até que ponto era possível prever o valor futuro de uma variável, utilizando como dados de entrada essa variável recolhida em vários anos. Assim foi necessário recorrer a diferentes abordagens de DM, Classificação e Regressão pois na revisão de literatura não foi identificada qual a abordagem indicada a seguir.

Mediante os resultados apresentados, resultantes do processo de DM, torna-se evidente o comprimento dos objetivos propostos para o desenvolvimento do projeto. O que por sua vez, reflete e responde á questão de investigação, sendo esta positiva. Porém, não responde única e exclusivamente a viabilidade da gestão de camas, mas reponde ao fator que possibilita essa mesma gestão que é, a gestão de altas para serviços de internamente.

O contributo deste trabalho é expresso pelos modelos gerados por Classificação e Regressão, pois estes apresentam-se como sendo adequados para solucionar problemas que se enquadram no ambiente estudado. Também, como referido anteriormente, podemos concluir que o método de representação de dados convencional, a criação de classes pela Regra de Sturges, o método de amostragem *10-folds CV* e a técnica SVM, demostraram ser os mais adequados para estudos que abrangem esta temática.

Uma vez que o desenvolvimento do projeto não contempla uma única abordadem de DM foi desenvolvido um algoritmo capaz de calcular o valor de MAE, atravez dos valores reais (máximo e mínino) contidos nas classes previstas atravez da abordagem de Classificação. Assim foi possivel usar uma métrica que permitisse comparar de forma abrangente os resultados apresentados por Classificação e Regressão. Segundo isto, o algoritmo desenvolvido representa uma contribuição deste trabalho.

5.3. Trabalho Futuro

Depois de concluído o presente trabalho é de grande importância indicar um conjunto de direções que podem ser tomadas. Os trabalhos futuros que seguirem a mesma linha de investigação desta Dissertação, gestão hospitalar com especial atenção à gestão de camas, devem ter em consideração os seguintes aspetos:

- Incorporar novas variáveis nos modelos de previsão, tais como o sexo e faixas etárias dos doentes. E determinar se as previsões poderão ser mais detalhadas. A introdução do número de entrada de doentes, tempo de ocupação e rácio de

ocupação das camas devem ser tidas em conta, para criar um sistema de gestão de camas mais completo;

- Repetir as experiências para mais serviços hospitalares com novos dados;
- Implementar um protótipo de um SAD em ambiente real que faculte previsões aos gestores e identifique se o sistema está em conformidade com a realidade.

Referências

ACS. (2010). *Evolução dos Indicadores do PNS 2004-2010*. Alto Comissariado da Saúde.

Alapont, J., Bella, A., Ferri, C., Hernández, J., Llopis, J., & Quintana, R. (2005). Specialised Tools for Automating Data Mining for Hospital Management. *Proc. First East European Conference on Health Care Modelling and Computation*.

Baskerville, R. (2007). Educing Theory from Practice. In *Information Systems Action Research - An Applied View of Emerging Concepts and Methods*. Springer.

Bellhouse, D. (2004). The Reverend Thomas Bayes, FRS: A Biography to Celebrate the Tercentenary of His Birth. *Statistical Science*, *19*, 3–43. doi:10.1214/088342304000000189

Ben-Hur, A., & Weston, J. (2010). A User's Guide to Support Vector Machines. In O. Carugo & F. Eisenhaber (Eds.), *Data Mining Techniques for the Life Sciences*. Humana Press.

Berry, M., & Linoff, G. (2000). *Mastering Data Mining: The Art and Science of Customer Relationship Management*. John Wiley & Sons, Inc.

Bi, J., & Benett, K. (2003). Regression Error Characteristic Curves. Presented at the Proceedings of the Twentieth International Conference on Machine Learning, Washington DC.

Bose, R. (2003). Knowledge management-enabled health care management systems: capabilities, infrastructure, and decision-support. *Expert Systems with Applications*, *24*(1), 59–71. doi:10.1016/S0957-4174(02)00083-0

Boser, B., Guyon, I., & Vapnik, V. (1992). A training Algorithm for Optimal Margin Classifiers. Presented at the COLT '92 Proceedings of the fifth annual workshop on Computational learning theory, New York, NY, USA: ACM.

Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., & Cleven, A. (2009). Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process. Presented at the Proceeding of the 17th European Conference on Information Systems.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide. SPSS inc.

Cios, K., Pedrycz, W., Swiniarski, R., & Kurgan, L. (2007). *Data Mining A Knowledge Discovery Approach*. Springer.

Cortez, P. (2010). Data mining with neural networks and support vector machines using the R/rminer tool. Presented at the Advances in Data Mining. Applications and Theoretical Aspects, Springer.

Cortez, P. (2013). Simpler use of data mining methods (e.g. NN and SVM) in classification and regression. Retrieved from http://cran.r-project.org/web/packages/rminer/rminer.pdf

Dua, S., & Du, X. (2011). *Data Mining and Machine Learning in Cybersecurity*. CRC Press - Taylor & Francis Group.

Dwivedi, A., Bali, R., & Naguib, R. (2006). Building New Healthcare Management Paradigms: A Case for Healthcare Knowledge Management. Presented at the Healthcare Knowledge Management Issues, Advances, and Successes, Health Informatics Series.

Ferri, C., Orallo, J., & Salido, M. A. (2003). Volume under the ROC Surface for Multi-class Problems. In *Machine Learning: ECML 2003*. Springer Berlin Heidelberg.

Griffin, D. (2006). *Hospitals What They Are and How They Work* (3ª Edição.). Jones And Bartlett Publishers.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques* (3ª Edição.). Morgan Kaufmann.

Kantardzic, M. (2011). *Data Mining Concepts, Models, Methods, and Algorithms* (2ª Edição.). Wiley - IEEE Press.

Koh, H., & Tan, G. (2005). Data Mining Applications in Healthcare. *Journal of Healthcare Information Management — Vol. 19, No. 2*.

Kohli, R., & Hoadley, E. (2007). Healthcare Fertile Ground for Action Research. In *Information Systems Action Research - An Applied View of Emerging Concepts and Methods*. Springer.

Koshy, E., Koshy, V., & Waterman, H. (2011). *Action Research in Healthcare*. SAGE Publications Ltd.

Lameirão, S. (2007). *Gestão Hospitalar e o uso dos Sistemas de Informação: Aplicação ao CHVR-PR*. Universidade de Trás-os-Montes e Alto Douro.

Lavrač, N., & Blaž, Z. (2010a). Classification Trees. In *Data Mining and Knowledge Discovery Handbook* (2ª Edição.). Springer.

Lavrač, N., & Blaž, Z. (2010b). Data Mining in Medicine. In *Data Mining and Knowledge Discovery Handbook* (2ª Edição.). Springer.

Maimon, O., & Rokach, L. (2010). Introduction to Knowledge Discovery and Data Mining. In *Data Mining and Knowledge Discovery Handbook* (2ª Edição.). Springer.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2012). Misc Functions of the Department of Statistics (e1071). Retrieved from http://cran.r-project.org/web/packages/e1071/e1071.pdf

Meyer, J. (2000). Qualitative research in health care: Using qualitative methods in health related action research. *British Medical Journal*, *320*(7228), 178–181. doi:10.1136/bmj.320.7228.178

Mittas, N., & Angelis, L. (2010). Visual comparison of software cost estimation models by regression error characteristic analysis. *The Journal of Systems and Software*.

Müller, M., Markó, K., Daumke, P., Paetzold, J., Roesner, A., & Klar, R. (2007). Biomedical Data Mining in Clinical Routine: Expanding the Impact of Hospital Information Systems. Presented at the MEDINFO 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics, IOS Press.

Murteira, B., Ribeiro, C., Silva, J., & Pimenta, C. (2010). *Introdução à Estatística*. Escolar Editora.

Nemati, H., & Barko, C. (2010). Organizational Data Mining. In *Data Mining and Knowledge Discovery Handbook* (2ª Edição.). Springer.

Neto, L. (2011). Análise da situação económico-financeira de hospitais. *O Mundo Da Saúde*, *35*(3).

Neves, M. (2011). Os Médicos vão ter de ser os motores da reforma do sistema. *Revista Portuguesa De Gestão & Saúde*, (5).

Ozgulbas, N., & Koyuncugil, A. (2007). Financial profiling of public hospitals: an application by data mining. Presented at the international journal of health planning and management, Wiley - A John Wiley and Sons, Ltd.

Pestana, D., & Velosa, S. (2008). *Introdução à Probabilidade e à Estatística* (4ª Edição.). Fundação Calouste Gulbenkian.

Piatetsky, G. (2013). KDnuggets Annual Software Poll: RapidMiner and R vie for first place. Retrieved from http://www.kdnuggets.com/2013/06/kdnuggets-annual-software-poll-rapidminer-r-vie-for-first-place.html

Proença, J., Vaz, A., Escoval, A., Candoso, F., Ferro, D., Carapeto, C., ... Roeslin, V. (2000). *O Hospital Português*. Vida Económica-Conferforum.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-Validation. In *Encyclopedia of Database Systems*. Springer.

Reis, E. (2008). *Estatística Descritiva* (7ª ed.). Edição Sílabo.

Roberto, W., & Lira, R. (2010). O Gestor Hospitalar e a sua atuação frente ao suprimento de materiais, *Volume 4*(13).

Santos, I., & Arruda, J. (2012). Análise do Perfil Profissional dos Gestores dos Hospitais Particulares da Cidade de Aracaju- SE. *Revista Eletronica Da Faculdade José Augusto Vieira*, *Nº -7*.

Santos, Manuel, & Azevedo, C. (2005). *Data Mining Descoberta de conhecimento em base de dados*. FCA - Editora de Informática, Lda.

Santos, Manuel, Boa, M., Portela, F., Silva, Á., & Rua, F. (2010). Real-time prediction of organ failure and outcome in intensive medicine. In *2010 5th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1 –6).

Santos, Maribel, & Ramos, I. (2009). *Business Intelligence Tecnologia da Informação na Gestão de Conhecimento* (2ª Edição.). FCA - Editora de Informática, Lda.

Silva, M., & Beuren, I. (2012). Características Bibliométricas dos Artigos Sobre Gestão Hospitalar Publicos em Periódicos de Alto Impacto. Presented at the XV SemeAd - Seminários em Administração.

Sousa, P., Machado, A., Rocha, M., Cortez, P., & Rio, M. (2010). A Collaborative Approach for Spam Detection. In *2010 Second International Conference on Evolving Internet* (pp. 92–97).

Teow, K., Darzi, E., Foo, E., Jin, X., & Sim, J. (2012). Intelligent Analysis of Acute Bed Overflow in a Tertiary Hospital in Singapore. *Springer US*.

Torgo, L. (2005). Regression error characteristic surfaces. Presented at the KDD '05 Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, ACM. doi:10.1145/1081870.1081959

Torgo, L. (2011). *Data Mining with R: Learning with Case Studies*. CRC Press - Taylor & Francis Group.

Tsumoto, S., Hirano, S., & Tsumoto, Y. (2011). Towards Data-Oriented Hospital Services: Data Mining in Hospital Management. In *SRII Global Conference (SRII), 2011 Annual* (pp. 349 –356). doi:10.1109/SRII.2011.47

Tsumoto, Shusaku, & Hirano, S. (2009). Data mining in hospital information system for hospital management (pp. 1 –5). Presented at the ICME International Conference on Complex Medical Engineering, 2009. CME. doi:10.1109/ICCME.2009.4906685

Tufféry, S. (2011). *Data Mining and Statistics for Decision Making*. John Wiley & Sons, Ltd.

Turban, E., Sharda, R., & Delen, D. (2011). *Decision Support and Business Intelligence Systems* (9ª Edição.). Prentice Hall.

Vapnik, V., & Cortes, C. (1995). Support Vector Networks - Machine Learning. Kluwer Academic Publishers.

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: writing a literature review. *MIS Q.*, *26*(2), xiii–xxiii.

WHO. (1963). *Expert Committee on Health Statistics* (No. 261).

Witten, I., & Frank, E. (2005). *Data Mining Pratical Machine Learning Tools and Techniques* (2ª Edição.). Morgan Kaufmann.

Witten, I., Frank, E., & Hall, M. (2011). *Data Mining Pratical Machine Learning Tools and Techniques* (3ª Edição.). Morgan Kaufmann.

Xindong, W., & Vipin, K. (2009). *The Top Ten Algorithms in Data Mining*. CRC Press - Taylor & Francis Group.

Yang, G., Sun, L., & Lin, X. (2010). Six-stage Hospital Beds Arrangement Management System. Presented at the Management and Service Science, IEEE.

Zhao, Y. (2012). *R and data mining examples and case studies*. Academic Press.

Apêndice A – Plano de Atividades Detalhado

Task	Duration	Start	End
1) CRISP-DM	30w 2d	13/03/01...	13/09/30 17:00
1.1) Estudo do Negócio	8w 3d	13/03/01...	13/04/30 17:00
1.1.1) Determinar objetivos	4w 3d	13/03/01 08:00	13/04/02 17:00
1.1.2) Avaliar situação atual	2w 2d	13/03/25 08:00	13/04/09 17:00
1.1.3) Desenvolver plano de trabalho	3w 2d	13/03/28 08:00	13/04/19 17:00
1.1.4) Apresentar plano	1w 2d	13/04/22 08:00	13/04/30 17:00
1.2) Estudo de Dados	8w 4d	13/04/01...	13/05/31 16:00
1.2.1) Recolha inicial dos dados	3w 1d	13/04/01 16:00	13/04/23 16:00
1.2.2) Descrição dos dados	2w 3d	13/04/16 16:00	13/05/03 16:00
1.2.3) Exploração dos dados	3w 1d	13/05/01 16:00	13/05/23 16:00
1.2.4) Formatação dos dados	1w 1d	13/05/23 16:00	13/05/31 16:00
1.3) Preparação de Dados	4w 1d 1h	13/06/01...	13/07/02 09:00
1.3.1) Seleção dos dados	1w 1h	13/06/03 08:00	13/06/10 09:00
1.3.2) Limpeza dos dados	1w 2d	13/06/06 09:00	13/06/17 09:00
1.3.3) Construção de dados	1w	13/06/16 08:00	13/06/21 17:00
1.3.4) Integração dos dados	3d 7h	13/06/19 09:00	13/06/24 17:00
1.3.5) Formatação dos dados	1w 1h	13/06/25 08:00	13/07/02 09:00
1.4) Modelação	8w 4d 6h	13/07/02...	13/09/02 16:00
1.4.1) Seleção de técnicas de modelação	2w 4d	13/07/02 09:00	13/07/22 09:00
1.4.2) Formulação de testes	3w 1d 7h	13/07/18 09:00	13/08/09 17:00
1.4.3) Construção de modelos	3w	13/08/08 08:00	13/08/28 17:00
1.4.4) Revisão de modelos	3d 7h	13/08/28 08:00	13/09/02 16:00
1.5) Avaliação	8w 3d	13/08/01...	13/09/30 17:00
1.5.1) Avaliação de resultados	4w 2d	13/08/01 08:00	13/08/30 17:00
1.5.2) Revisão do processo	3w 3d	13/08/28 08:00	13/09/20 17:00
1.5.3) Determinar os próximos passos	1w 1d	13/09/21 08:00	13/09/30 17:00
2) Revisão	3w	13/09/30 08:00	13/10/18 17:00

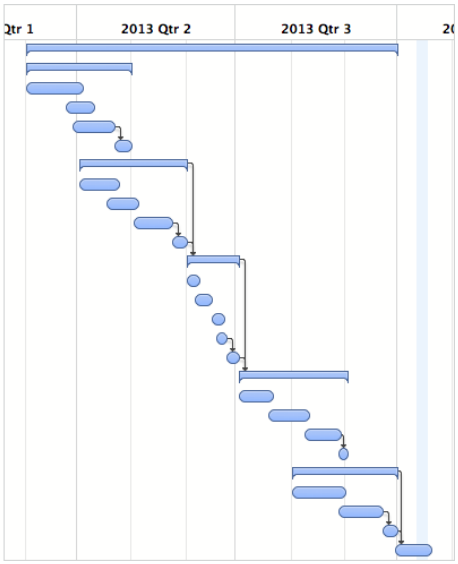

Figura 19 – Plano de Atividades (CRISP-DM)

Apêndice B – Avaliação dos Modelos de Classificação

Tabela 8 – Avaliação por Classificação (Ortopedia)

Abordagem: Classificação
Serviço: Ortopedia

Representação de Dados: Convencional			
Classes: Média			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	61,54%
		LOOCV	61,54%
	RBF	10-folds CV	65,38%
		LOOCV	65,38%
DT	Gini	10-folds CV	61,54%
		LOOCV	61,54%
	Information	10-folds CV	61,54%
		LOOCV	61,54%
NB		10-folds CV	61,54%
		LOOCV	61,54%

Representação de Dados: Convencional			
Classes: Quartis			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	57,69%
		LOOCV	57,69%
	RBF	10-folds CV	80,77%
		LOOCV	80,77%
DT	Gini	10-folds CV	51,92%
		LOOCV	51,92%
	Information	10-folds CV	53,85%
		LOOCV	53,85%
NB		10-folds CV	55,77%
		LOOCV	55,77%

Representação de Dados: Convencional			
Classes: Média – Desvio Padrão			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	59,62%
		LOOCV	59,62%
	RBF	10-folds CV	69,23%
		LOOCV	69,23%
DT	Gini	10-folds CV	50%
		LOOCV	50%
	Information	10-folds CV	50%
		LOOCV	50%
NB		10-folds CV	61,54%
		LOOCV	61,54%

Representação de Dados: Convencional			
Classes: Sturges			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	75%
		LOOCV	75%
	RBF	10-folds CV	82,69%
		LOOCV	82,69%
DT	Gini	10-folds CV	46,15%
		LOOCV	46,15%
	Information	10-folds CV	44,23%
		LOOCV	44,23%
NB		10-folds CV	73,08%
		LOOCV	73,08%

Representação de Dados: Janela Deslizante			
Classes: Quartis			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	39,62%
		LOOCV	39,62%
	RBF	10-folds CV	50,95%
		LOOCV	52,83%
DT	Gini	10-folds CV	39,62%
		LOOCV	39,62%
	Information	10-folds CV	41,51%
		LOOCV	41,51%
NB		10-folds CV	33,96%
		LOOCV	33,96%

Representação de Dados: Janela Deslizante			
Classes: Média – Desvio Padrão			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	43,39%
		LOOCV	43,39%
	RBF	10-folds CV	54,72%
		LOOCV	60,38%
DT	Gini	10-folds CV	45,28%
		LOOCV	45,28%
	Information	10-folds CV	45,28%
		LOOCV	45,28%
NB		10-folds CV	43,39%
		LOOCV	43,39%

Representação de Dados: Janela Deslizante			
Classes: Sturges			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	49,06%
		LOOCV	49,06%
	RBF	10-folds CV	71,70%
		LOOCV	71,70%
DT	Gini	10-folds CV	58,60%
		LOOCV	58,60%
	Information	10-folds CV	56,60%
		LOOCV	56,60%
NB		10-folds CV	49,06%
		LOOCV	49,06%

Tabela 9 – Avaliação por Classificação (Obstetrícia)

Abordagem: Classificação
Serviço: Obstetrícia

Representação de Dados: Convencional			
Classes: Média			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	65,38%
		LOOCV	65,38%
	RBF	10-folds CV	65,38%
		LOOCV	65,38%
DT	Gini	10-folds CV	63,46%
		LOOCV	63,46%
	Information	10-folds CV	63,46%
		LOOCV	63,46%
NB		10-folds CV	65,38%
		LOOCV	65,38%

Representação de Dados: Convencional			
Classes: Quartis			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	55,77%
		LOOCV	55,77%
	RBF	10-folds CV	60,58%
		LOOCV	63,46%
DT	Gini	10-folds CV	50%
		LOOCV	50%
	Information	10-folds CV	50%
		LOOCV	50%
NB		10-folds CV	51,92%
		LOOCV	51,92%

Representação de Dados: Convencional			
Classes: Média – Desvio Padrão			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	57,69%
		LOOCV	57,69%
	RBF	10-folds CV	67,31%
		LOOCV	67,31%
DT	Gini	10-folds CV	51,92%
		LOOCV	51,92%
	Information	10-folds CV	51,92%
		LOOCV	51,92%
NB		10-folds CV	53,85%
		LOOCV	53,85%

Representação de Dados: Convencional			
Classes: Sturges			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	82,69%
		LOOCV	82,69%
	RBF	10-folds CV	82,69%
		LOOCV	82,69%
DT	Gini	10-folds CV	50%
		LOOCV	50%
	Information	10-folds CV	42,31%
		LOOCV	42,31%
NB		10-folds CV	61,54%
		LOOCV	61,54%

Representação de Dados: Janela Deslizante			
Classes: Quartis			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	16,13%
		LOOCV	16,13%
	RBF	10-folds CV	24,2%
		LOOCV	22,58%
DT	Gini	10-folds CV	29,03%
		LOOCV	29,03%
	Information	10-folds CV	29,03%
		LOOCV	29,03%
NB		10-folds CV	24,19%
		LOOCV	24,19%

Representação de Dados: Janela Deslizante			
Classes: Sturges			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	27,42%
		LOOCV	27,42%
	RBF	10-folds CV	19,35%
		LOOCV	19,35%
DT	Gini	10-folds CV	14,52%
		LOOCV	14,52%
	Information	10-folds CV	9,68%
		LOOCV	9,68%
NB		10-folds CV	20,97%
		LOOCV	20,97%

Representação de Dados: Janela Deslizante			
Classes: Média – Desvio Padrão			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	41,94%
		LOOCV	43,55%
	RBF	10-folds CV	30,65%
		LOOCV	30,65%
DT	Gini	10-folds CV	24,19%
		LOOCV	24,19%
	Information	10-folds CV	24,19%
		LOOCV	24,19%
NB		10-folds CV	37,1%
		LOOCV	37,1%

Tabela 10 – Avaliação por Classificação (Parto)

Abordagem: Classificação
Serviço: Parto

Representação de Dados: Convencional			
Classes: Média			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	76,92%
		LOOCV	76,92%
	RBF	10-folds CV	76,92%
		LOOCV	76,92%
DT	Gini	10-folds CV	76,92%
		LOOCV	76,92%
	Informati on	10-folds CV	76,92%
		LOOCV	76,92%
NB		10-folds CV	76,92%
		LOOCV	76,92%

Representação de Dados: Convencional			
Classes: Média – Desvio Padrão			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	65,38%
		LOOCV	65,38%
	RBF	10-folds CV	73,08%
		LOOCV	73,08%
DT	Gini	10-folds CV	59,62%
		LOOCV	59,62%
	Informati on	10-folds CV	59,62%
		LOOCV	59,62%
NB		10-folds CV	67,31%
		LOOCV	67,31%

Representação de Dados: Janela Deslizante			
Classes: Quartis			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	41,94%
		LOOCV	40,81%
	RBF	10-folds CV	25,81%
		LOOCV	25,81%
DT	Gini	10-folds CV	24,19%
		LOOCV	24,19%
	Information	10-folds CV	24,19%
		LOOCV	24,19%
NB		10-folds CV	33,87%
		LOOCV	33,87%

Representação de Dados: Convencional			
Classes: Quartis			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	63,46%
		LOOCV	63,46%
	RBF	10-folds CV	82,69%
		LOOCV	82,69%
DT	Gini	10-folds CV	57,69%
		LOOCV	57,69%
	Information	10-folds CV	53,85%
		LOOCV	53,85%
NB		10-folds CV	51,92%
		LOOCV	51,92%

Representação de Dados: Convencional			
Classes: Sturges			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	75%
		LOOCV	76,92%
	RBF	10-folds CV	94,23%
		LOOCV	94,23%
DT	Gini	10-folds CV	57,69%
		LOOCV	57,69%
	Information	10-folds CV	57,69%
		LOOCV	57,69%
NB		10-folds CV	65,38%
		LOOCV	65,38%

Representação de Dados: Janela Deslizante			
Classes: Média – Desvio Padrão			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	40,32%
		LOOCV	58,06%
	RBF	10-folds CV	38,71%
		LOOCV	38,71%
DT	Gini	10-folds CV	32,26%
		LOOCV	32,26%
	Information	10-folds CV	54,84%
		LOOCV	54,84%
NB		10-folds CV	54,84%
		LOOCV	54,84%

Representação de Dados: Janela Deslizante			
Classes: Sturges			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	12,1%
		LOOCV	11,29%
	RBF	10-folds CV	20%
		LOOCV	22,58%
DT	Gini	10-folds CV	25,81%
		LOOCV	25,81%
	Information	10-folds CV	32,26%
		LOOCV	32,26%
NB		10-folds CV	8,06%
		LOOCV	8,06%

Tabela 11 – Avaliação por Classificação (Berçário)

Abordagem: Classificação
Serviço: Berçário

Representação de Dados: Convencional			
Classes: Média			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	71,15%
		LOOCV	71,15%
	RBF	10-folds CV	71,15%
		LOOCV	71,15%
DT	Gini	10-folds CV	69,32%
		LOOCV	69,32%
	Information	10-folds CV	69,32%
		LOOCV	69,32%
NB		10-folds CV	71,15%
		LOOCV	71,15%

Representação de Dados: Convencional			
Classes: Quartis			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	53,85%
		LOOCV	53,85%
	RBF	10-folds CV	67,31%
		LOOCV	67,31%
DT	Gini	10-folds CV	50%
		LOOCV	50%
	Information	10-folds CV	48,08%
		LOOCV	48,08%
NB		10-folds CV	53,85%
		LOOCV	53,85%

Representação de Dados: Convencional			
Classes: Média – Desvio Padrão			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	58,27%
		LOOCV	59,62%
	RBF	10-folds CV	67,31%
		LOOCV	67,31%
DT	Gini	10-folds CV	51,92%
		LOOCV	51,92%
	Information	10-folds CV	53,85%
		LOOCV	53,85%
NB		10-folds CV	59,62%
		LOOCV	59,62%

Representação de Dados: Convencional			
Classes: Sturges			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	75%
		LOOCV	63,46%
	RBF	10-folds CV	90,38%
		LOOCV	90,38%
DT	Gini	10-folds CV	48,08%
		LOOCV	48,08%
	Information	10-folds CV	48,08%
		LOOCV	48,08%
NB		10-folds CV	59,62%
		LOOCV	59,62%

Representação de Dados: Janela Deslizante			
Classes: Quartis			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	35,48%
		LOOCV	35,48%
	RBF	10-folds CV	37,42%
		LOOCV	38,71%
DT	Gini	10-folds CV	29,03%
		LOOCV	29,03%
	Information	10-folds CV	32,26%
		LOOCV	32,26%
NB		10-folds CV	37,1%
		LOOCV	37,1%

Representação de Dados: Janela Deslizante			
Classes: Sturges			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	24,52%
		LOOCV	24,52%
	RBF	10-folds CV	23,71%
		LOOCV	24,19%
DT	Gini	10-folds CV	25,81%
		LOOCV	25,81%
	Information	10-folds CV	19,35%
		LOOCV	19,35%
NB		10-folds CV	20,97%
		LOOCV	20,97%

Representação de Dados: Janela Deslizante			
Classes: Média – Desvio Padrão			
Técnica		Método de Amostragem	Acuidade
SVM	Linear	10-folds CV	27,74%
		LOOCV	28,06%
	RBF	10-folds CV	35,48%
		LOOCV	35,48%
DT	Gini	10-folds CV	33,87%
		LOOCV	33,87%
	Information	10-folds CV	25,81%
		LOOCV	25,81%
NB		10-folds CV	37,1%
		LOOCV	37,1%

Apêndice C – Análise de Sensibilidade e Especificidade

Tabela 12 – Sensibilidade e Especificidade (Ortopedia)

Serviço: Ortopedia

Classes	Sensibilidade	Especificidade
[19, 25.78]	75%	100%
]25.78, 32.56]	100%	98%
]32.56, 39.34]	0%	94,23%
]39.34, 46.12]	71,43%	97,36%
]46.12, 52.9]	91,67%	97,5%
]52.9, 59.69]	82,35%	97,14%
]59.69, 66.47]	100%	96%
]73.25, 78]	100%	100%

Tabela 13 – Sensibilidade e Especificidade (Obstetrícia)

Serviço: Obstetrícia

Classes	Sensibilidade	Especificidade
[40, 45.86]	100%	94%
]45.86, 51.72]	100%	93,75%
]51.72, 57.72]	100%	97,73%
]57.72, 63.45]	62,5%	100%
]63.45, 69.31]	100%	97,67%
]69.31, 75.17]	100%	97,87%

Tabela 14 – Sensibilidade e Especificidade (Parto)

Serviço: Parto

Classes	Sensibilidade	Especificidade
[0, 0.92]	100%	97,67%
]0.92, 1.84]	94,74%	100%
]1.84, 2.76]	85,71%	100%
]2.76, 3.68]	100%	97,78%
]3.68, 4.6]	100%	100%
]4.6, 5.52]	100%	98%

Tabela 15 – Sensibilidade e Especificidade (Berçário)

Serviço: Berçário

Classes	Sensibilidade	Especificidade
[34, 40.67]	0%	96,15%
]40.67, 47.33]	100%	97,5%
]47.33, 53]	92,31%	100%
]53, 60.66]	80,95%	100%
]60.66, 67.33]	100%	97,83%
]67.33, 73]	0%	98,08%

Apêndice D – Avaliação dos Modelos de Regressão

Tabela 16 – Avaliação por Regressão (Ortopedia)

Abordagem: Regressão
Serviço: Ortopedia

Representação de Dados: Convencional					
Técnica		Método de Amostragem	MAE	MSE	RAE
SVM	Linear	10-folds CV	6,67	70,82	117,43%
		LOOCV	6,67	70,82	117,43%
	RBF	10-folds CV	4,03	34,43	73,08%
		LOOCV	4,03	34,43	73,08%
AR	Gini	10-folds CV	6,35	76,62	116,49%
		LOOCV	6,35	76,62	116,49%

Representação de Dados: Janela Deslizante					
Técnica		Método de Amostragem	MAE	MSE	RAE
SVM	Linear	10-folds CV	7,97	102,83	147,33%
		LOOCV	7,97	102,83	147,33%
	RBF	10-folds CV	8,68	128,97	188,76%
		LOOCV	8,51	121,92	199,02%
AR	Gini	10-folds CV	8,61	125,17	153,73%
		LOOCV	8,61	125,17	153,73%

Tabela 17 – Avaliação por Regressão (Obstetrícia)

Abordagem: Regressão
Serviço: Obstetrícia

Representação de Dados: Convencional					
Técnica		Método de Amostragem	MAE	MSE	RAE
SVM	Linear	10-folds CV	6,23	62,08	341,77%
		LOOCV	6,23	62,08	341,77%
	RBF	10-folds CV	1,86	10,93	38,26%
		LOOCV	1,86	10,93	38,26%
AR	Gini	10-folds CV	5,91	49,06	195,82%
		LOOCV	5,91	49,06	195,92%

Representação de Dados: Janela Deslizante					
Técnica		Método de Amostragem	MAE	MSE	RAE
SVM	Linear	10-folds CV	7,07	80,06	125,32%
		LOOCV	6,98	79,04	120,68%
	RBF	10-folds CV	7,80	95,64	132,92%
		LOOCV	7,75	94,6	131%
AR	Gini	10-folds CV	7,84	96,43	123,41%
		LOOCV	7,52	91,08	117.8%

Tabela 18 – Avaliação por Regressão (Parto)

Abordagem: Regressão
Serviço: Parto

Representação de Dados: Convencional					
Técnica		Método de Amostragem	MAE	MSE	RAE
SVM	Linear	10-folds CV	1	1,72	473,32%
		LOOCV	1	1,72	473,32%
	RBF	10-folds CV	0,62	0,99	96,89%
		LOOCV	0,62	0,99	96,89%
AR	Gini	10-folds CV	0,99	1,60	300,01%
		LOOCV	0,99	1,60	300,01%

Representação de Dados: Janela Deslizante					
Técnica		Método de Amostragem	MAE	MSE	RAE
SVM	Linear	10-folds CV	1,31	2,8	323,6%
		LOOCV	1,33	2,87	278,2%
	RBF	10-folds CV	1,49	3,5	220,16%
		LOOCV	1,46	3,41	214,59%
AR	Gini	10-folds CV	1,39	3,07	224,1%
		LOOCV	1,38	3,04	313,32%

Tabela 19 – Avaliação por Regressão (Berçário)

Abordagem: Regressão
Serviço: Berçário

Representação de Dados: Convencional					
Técnica		Método de Amostragem	MAE	MSE	RAE
SVM	Linear	10-folds CV	6,18	59,82	553,83%
		LOOCV	6,18	59,82	553,83%
	RBF	10-folds CV	2,43	20,15	53,82%
		LOOCV	2,43	20,15	53,82%
AR	Gini	10-folds CV	5,55	47,87	209,67%
		LOOCV	5,55	47,87	209,67%

Representação de Dados: Janela Deslizante					
Técnica		Método de Amostragem	MAE	MSE	RAE
SVM	Linear	10-folds CV	5,86	77,32	129,45%
		LOOCV	7,09	81,2	127,27%
	RBF	10-folds CV	7,35	86,41	129,85%
		LOOCV	7,74	93,55	134,6%
AR	Gini	10-folds CV	7,82	95,47	123,99%
		LOOCV	7,97	101,42	122,35%

Anexo A – Ciclo de Vida do CRISP-DM Detalhado

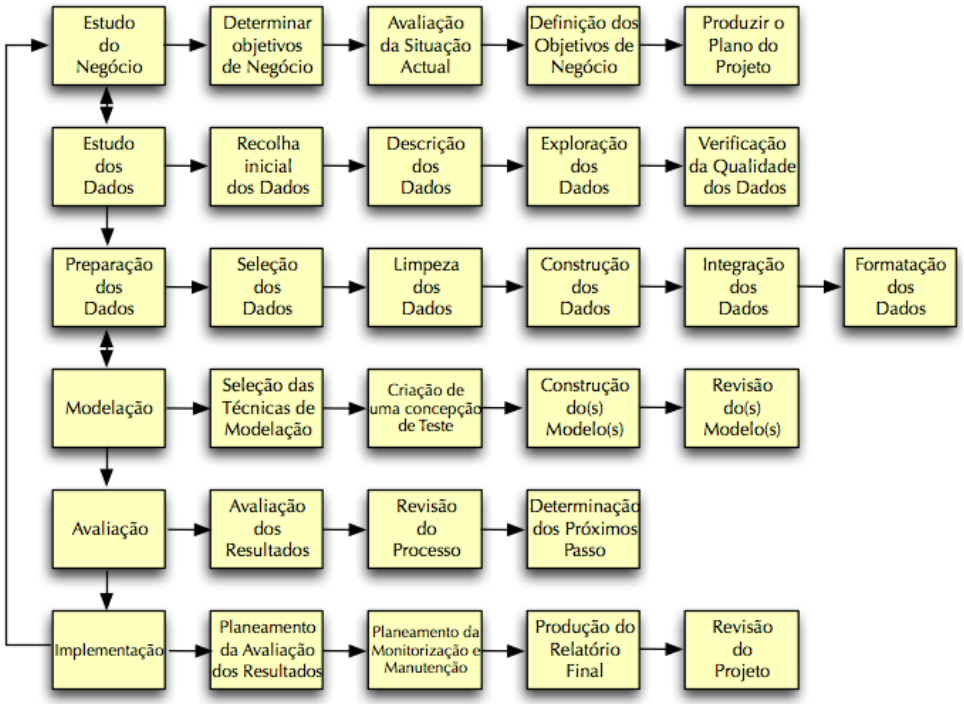

Figura 20 – CRISP-DM Detalhado