INSTITUTO UNIVERSITÁRIO MILITAR

DEPARTAMENTO DE ESTUDOS PÓS-GRADUADOS

CURSO DE PROMOÇÃO A OFICIAL SUPERIOR DA FORÇA AÉREA

2018/2019

TII

BUSINESS INTELLIGENCE: APLICAÇÃO DE TÉCNICAS DE *DATA MINING* NO HOSPITAL DAS FORÇAS ARMADAS PARA PREVISÃO DE TEMPOS DE INTERNAMENTO

Nuno Manuel Palhotas Caetano

CAP/TINF

INSTITUTO UNIVERSITÁRIO MILITAR

DEPARTAMENTO DE ESTUDOS PÓS-GRADUADOS

BUSINESS INTELLIGENCE: APLICAÇÃO DE TÉCNICAS DE DATA MINING NO HOSPITAL DAS FORÇAS ARMADAS PARA PREVISÃO DE TEMPOS DE INTERNAMENTO

CAP/TINF Nuno Manuel Palhotas Caetano

Trabalho de Investigação Individual do CPOS-FA 2018/19

Pedrouços 2019

INSTITUTO UNIVERSITÁRIO MILITAR
DEPARTAMENTO DE ESTUDOS PÓS-GRADUADOS

BUSINESS INTELLIGENCE: APLICAÇÃO DE TÉCNICAS DE DATA MINING NO HOSPITAL DAS FORÇAS ARMADAS PARA PREVISÃO DE TEMPOS DE INTERNAMENTO

CAP/TINF Nuno Manuel Palhotas Caetano

Trabalho de Investigação Individual do CPOS-FA 2018/19

Orientador: TCOR/TMMA Nuno Alberto Rodrigues Santos Loureiro

Pedrouços 2019

Declaração de compromisso Antiplágio

Eu, **Nuno Manuel Palhotas Caetano**, declaro por minha honra que o documento intitulado ***Business Intelligence*: Aplicação de Técnicas de *Data Mining* no Hospital das Forças Armadas para Previsão de Tempos de Internamento** corresponde ao resultado da investigação por mim desenvolvida enquanto auditor do **Curso de Promoção a Oficial Superior – Força Aérea 2018/2019** no Instituto Universitário Militar e que é um trabalho original, em que todos os contributos estão corretamente identificados em citações e nas respetivas referências bibliográficas.

Tenho consciência que a utilização de elementos alheios não identificados constitui grave falta ética, moral, legal e disciplinar.

Pedrouços**,** 25 de janeiro de 2019

Nuno Manuel Palhotas Caetano
CAP/TINF

Agradecimentos

Aos que fazem parte destas datas

18 de janeiro de 2010

4 de junho de 1977

2 de dezembro de 1976

12 de agosto de 1966

30 de setembro de 1947

6 de setembro de 1947

Índice

Índice de Apêndices

Índice de Figuras

Índice de Tabelas

Business Intelligence: Aplicação de Técnicas de *Data Mining* no Hospital das Forças Armadas para Previsão de Tempos de Internamento

Nuno M. P. Caetano e Nuno A. R. S. Loureiro

Instituto Universitário Militar, Lisboa, Portugal

Notas do autor

Nuno M. P. Caetano, Instituto Universitário Militar, Lisboa, Portugal.

A correspondência relacionada com este artigo deverá ser endereçada para Instituto Universitário Militar, Rua de Pedrouços, 1449-027 Lisboa, Portugal

E-mail: secretaria@ium.pt

Resumo

Com o crescente aumento de dados nos sistemas de informação clínica, tornou-se necessária a exploração de várias tecnologias e metodologias para análise desse valioso conhecimento. Esta investigação teve por objetivo geral obter um modelo preditivo otimizado de tempos de internamento de pacientes no Hospital das Forças Armadas, através da descoberta de comportamentos e padrões existentes no processo de internamento hospitalar, com base em técnicas de *data mining*. Os internamentos realizados no Hospital das Forças Armadas, compreendidos entre 2013 e 2017, foram a população alvo desta investigação e, atendendo aos objetivos e ao problema, a metodologia que se revelou mais adequada foi a metodologia CRISP-DM. Na fase de preparação de dados foram selecionados 19 atributos de entrada, enquanto que na fase de modelação efetuou-se uma abordagem regressiva, aplicando-se cinco técnicas de regressão: *Decision Tree*, *Naive, Multiple Regression*, *Random Forest* e *Support Vector Machines*. Identificou-se como melhor modelo o *Random Forest*, com um coeficiente de determinação de 0,735 e com uma capacidade de prever corretamente 78,5% dos casos. Através de uma análise de sensibilidade, verificou-se que os quatro atributos mais significativos, relacionados com a situação clínica dos pacientes, contribuem em mais de 50% para a capacidade explicativa do modelo gerado: Tipo de Episódio de Internamento (31,9%), Serviço de Internamento (8%), Especialidade Médica (7,5%) e Destino da Alta (7,2%).

Palavras-chave: CRISP-DM, *data mining*, *length of stay*, modelo de previsão.

Abstract

With data collection increasing in clinical information systems, it became necessary to explore various technologies and methodologies to analyze this valuable knowledge. The objective of this investigation was to obtain an optimized predictive model of patient's hospitalization time in the Armed Forces Hospital, through the discovery of behaviors and patterns existing in the hospitalization process, based on data mining techniques. The hospitalizations performed at the Armed Forces Hospital between 2013 and 2017, were the population target of this research and, in view of the objectives and the problem, the methodology that proved to be the most adequate was the CRISP-DM methodology. In the data preparation phase, 19 input attributes were selected, while in the modeling phase a regressive approach was applied, applying five regression techniques: Decision Tree, Naive, Multiple Regression, Random Forest and Support Vector Machines. Random Forest was the best model, with a coefficient of determination 0.735 and predicting correctly 78.5% of the cases. Through a sensitivity analysis, was verified that the four most significant attributes, related to the patient's clinical situation, contribute more than 50% to the explanatory capacity of the new model: Hospital Episode Type (31.9%), Physical Service (8%), Medical Specialty (7.5%) and Discharge Destination (7.2%).

Keywords*: CRISP-DM, data mining, length of stay, prediction model*

Business Intelligence: Aplicação de Técnicas de *Data Mining* no Hospital das Forças Armadas para Previsão de Tempos de Internamento

Introdução

Recentemente, as novas tecnologias proporcionaram o rápido desenvolvimento dos sistemas de informação (SI) de apoio à saúde (Lee et al., 2011). Atualmente, as instituições de saúde utilizam SI que lhes simplificam todo o processo de comunicação, diminuindo a burocracia dos processos, aumentando a quantidade e qualidade dos dados, melhorando o tempo de resposta e aumentando a qualidade no cuidado aos pacientes. O volume e complexidade da informação gerada pelos SI, aliada às limitações humanas, condicionam a extração de conhecimento, tornando-se difícil a análise e compreensão dos mesmos.

Conforme refere Cruz (2007), o processo de *data mining* (DM) surge da necessidade de descoberta de conhecimento, através do desenvolvimento de métodos e técnicas de extração de conhecimento a partir de informação guardada em bases de dados (BD). Han e Kamber (2006) referem-se ao DM como um processo de identificação de novos padrões, potencialmente úteis e fundamentalmente inteligíveis. O DM requer conhecimento dos processos por detrás dos dados, de modo a que sejam definidas perguntas para análise, selecionados os dados relevantes para resposta às perguntas e interpretados os resultados da análise (Feelders, Daniels, & Holsheimer, 2000). Deste modo, torna-se exequível o uso de técnicas, processos e consequente desenvolvimento de modelos de DM, suportados no SI de uma instituição hospitalar, com o intuito de criar informação inteligente sobre o negócio.

Suthummanon e Omachonu (2004) referem ainda que estudos que têm por base os grupos de diagnóstico homogéneo (GDH) mostram que os hospitais que conseguem controlar os tempos de internamento, diminuem significativamente os custos por admissão e a diária do

doente. Segundo Marques (2010) a duração dos internamentos é um fator a ter em conta, uma vez que os internamentos prolongados representam um aumento direto do consumo de recursos. Em linha com o exposto anteriormente, os hospitais apresentam como objetivo a necessidade de reduzir o tempo de internamento, aumentar o número de camas disponíveis para novos internamento e prestar melhores cuidados de saúde aos seus utentes. Estes objetivos também são refletidos por Domingos (2015), referindo que a fusão dos hospitais militares originou a subsequente redução de cerca de 94% das camas de internamento por cada 1000 potenciais utentes. Um modelo de previsão de *length of stay* (LOS) para pacientes hospitalizados permite evitar períodos de internamento prolongados, melhorar os serviços de saúde e gerir de forma mais eficiente os recursos hospitalares.

Noutros estudos nesta área, como o realizado em 1998, Merom, Shohat, Harari, Meir, e Green estimaram a taxa de dias de internamentos inadequados (falha nos critérios estabelecidos para a admissão) e a subsequente identificação das variáveis associadas a essa inadequação. Ao nível do universo estudado, 1369 pacientes de 24 hospitais foram analisados no referido estudo através do uso do modelo de regressão múltipla e atributos como o caso da ocupação; grupo etário; dia inapropriado de internamento; governo; outra entidade hospitalar; outro diagnóstico; sexo; origem da entrada; diagnóstico na admissão e período de estadia, foram referenciados para consequente análise. Deste estudo, obteve-se o valor de 8,6 ± 12,2 dias para internamento inadequado e 6,1 ± 7,3 dias para o internamento adequado.

Em 2007, Abelha, Maia, Landeiro, Neves, e Barros avaliaram o LOS de pacientes adultos, internados numa unidade de cuidados intensivos (UCI) e submetidos a cirurgias não-cardíacas entre outubro de 2004 e julho de 2005. Foram identificados os seguintes atributos para categorizar os pacientes da unidade: idade, género, sexo, índice de massa corporal, estado

físico ASA[1], tipo e magnitude do procedimento cirúrgico, tipo e duração da anestesia, temperatura na admissão, LOS na UCI e no hospital, mortalidade na UCI e no hospital. O modelo adotado neste estudo foi a regressão linear simples, tendo-se obtido um resultado de LOS médio de 4,22 ± 8,76 dias.

Em 2010, Oliveira et al. avaliaram os fatores associados à maior mortalidade e tempo de internamento prolongado numa unidade de terapia intensiva (UTI). Neste estudo participaram 401 pacientes adultos admitidos na UTI, no período de seis meses, tendo-se utilizado os seguintes atributos: sexo, idade, diagnóstico, antecedentes pessoais, APACHE II[2], dias de ventilação mecânica invasiva, reintubação orotraqueal, traqueostomia, dias de internação na unidade de terapia intensiva, alta ou óbito na UTI. Da análise efetuada, concluiu-se o valor médio de 8,2 ± 10,8 dias de internamento na UTI e a identificação dos atributos associados ao objetivo previsto: APACHE II, traqueostomia e a reintubação.

Ainda em 2010, Kalra, Fisher, e Axelrod estudaram as tendências temporais do fluxo de trabalho do serviço de medicina interna de um hospital universitário, analisando os dados obtidos em três diferentes períodos de tempo e abrangendo treze anos. Foram identificados os seguintes atributos mais relevantes: data de admissão, data de saída ou óbito, sexo, idade, código postal de residência, entidade financeira pagadora dos serviços hospitalares e diagnóstico principal. Neste estudo foram utilizados modelos de regressão linear que confirmaram o aumento estatístico mensal do número de admissões do serviço de medicina interna de 117 (1991) para 455 (2004). Por fim confirmou-se que a média de LOS diminuiu de 8,7 para 4,9 dias.

[1] Classificação do estado físico de acordo com a escala da *American Society of Anesthesiologists* (ASA).

[2] O sistema de pontuação APACHE II mede a gravidade da doença em pacientes internados na UTI.

Também em 2010, Pena et al. avaliaram a possibilidade de marcar e antecipar o tempo de permanência numa UTI. Os dados coletados dos 110 pacientes em estudo e das variáveis idade, sexo, género, tipo de cirurgia, prioridade cirúrgica, cirurgia cardíaca prévia e o índice de massa corporal, permitiram obter o resultado de 4,2 dias de permanência em média na UTI.

Mais recentemente, em 2012, Freitas et al. investigaram os LOS discrepantes, estudando os episódios de internamento de hospitais públicos portugueses pertencentes ao sistema nacional de saúde (SNS), compreendidos entre 2000 e 2009. Na análise foram utilizados modelos de regressão logística para examinar a associação de cada variável com os LOS discrepantes, tendo-se obtido os seguintes resultados: nove milhões de episódios de internamento analisados (excluindo os episódios de ambulatório), identificados 3,9% de LOS discrepantes referentes a 19,2% do número total de internamentos. Foram identificados 35,5 dias em média de LOS para o caso de valores discrepantes e 6 dias para os não discrepantes. Concluíram que as variáveis idade, tipo de internamento e tipo de hospital estão significativamente associadas com os altos LOS discrepantes.

Rufino, Gurgel, Pontes, e Freire (2012) estudaram, de agosto de 2010 a março de 2011, os fatores que interferem no LOS numa enfermaria de clínica médica. Participaram no estudo 48 pacientes internados e analisadas variáveis como: idade, sexo, internamento anterior, queixa principal, escolaridade, renda familiar mensal, dor (localização, intensidade, duração), tabagismo, etilismo. Referem ainda que em 2009 a média nacional de LOS era de 6,6 dias, enquanto que em hospitais públicos de média e alta complexidade situava-se em 9,3 dias. Por fim, obtiveram para LOS uma média de 20,9 dias.

No estudo realizado por Caetano (2013) no Hospital da Força Aérea (HFA), concluiu-se que um episódio de internamento em regime de internamento apresentava um LOS de 3,81 dias. O serviço de medicina apresentava um LOS de 3,44 dias, seguindo-se o serviço de

ortopedia (3,14 dias) e por último os serviços de cirurgia, especialidades e pneumologia (2,98 dias). A especialidade médica medicina interna representava 4,42 dias de LOS, seguindo-se a ortopedia (3,39 dias), a especialidade cirurgia geral (3,14 dias) e a urologia com 3,10 dias. No estudo de 2015, Caetano, Cortez e Laureano obtiveram 3,9 dias de LOS para um episódio de internamento em regime de internamento.

No âmbito do processo de reestruturação hospitalar preconizado pela Resolução do Conselho de Ministros n.º 39/2008, de 28 de fevereiro, a Lei Orgânica de Bases da Organização das Forças Armadas, aprovada pela Lei Orgânica n.º 1-A/2009, de 7 de julho, e a Lei Orgânica do Estado-Maior-General das Forças Armadas, aprovada pelo Decreto-Lei n.º 234/2009, de 15 de setembro, consagraram a criação do Hospital das Forças Armadas (HFAR) enquanto hospital militar único, organizado em dois polos hospitalares, um em Lisboa e outro no Porto.

No âmbito do processo de reestruturação do sistema de saúde militar (SSM), foi publicado em Diário da República o Decreto-Lei n.º 187/2012, de 16 de agosto, que veio proceder à criação do Polo de Lisboa (PL) do HFAR, como resultado da fusão dos quatro hospitais militares sediados em Lisboa: O Hospital da Marinha (HM), o Hospital Militar Principal (HMP), o Hospital Militar de Belém (HMB) e o HFA, operada nos termos do Decreto-Lei n.º 200/2006, de 25 de outubro, substituindo estes quatro estabelecimentos hospitalares na prestação de cuidados de saúde aos seus utentes. Posteriormente, o Decreto Regulamentar nº 51/2012 de 10 de dezembro, estabeleceu a estrutura orgânica e a estrutura funcional do PL do HFAR. Decorrente do programa do XIX Governo Constitucional, foi reconhecido como aspeto decisivo para a implementação do SSM, prosseguir com a fusão entre o HM, HMP, HMB e HFA, tendo em vista a operacionalização efetiva do HFAR, o qual constitui um órgão na dependência direta do Chefe do Estado-Maior-General das Forças Armadas (CEMGFA) (Despacho n.º 2943/2014 de 31 de janeiro). Nesse intuito, o Decreto-Lei

nº 84/2014 de 27 de maio veio proceder à criação do HFAR, como um estabelecimento hospitalar militar único, na dependência direta do CEMGFA, constituído pelo PL, sito no designado Campus de Saúde Militar (CSM), em Lisboa, e pelo Polo do Porto (PP), sito nas instalações do antigo Hospital Militar Regional n.º 1 (HMR), no Porto. O HFAR é um estabelecimento hospitalar militar, que se constitui como elemento de retaguarda do SSM em apoio da saúde operacional. Apresenta como missão a prestação de cuidados de saúde diferenciados aos militares das Forças Armadas (FFAA), bem como à família militar e aos deficientes das FFAA.

Após a fusão dos principais hospitais militares, torna-se essencial que esta investigação estude os LOS hospitalares, devido à estrita relação com os custos hospitalares. Esta preocupação também é identificada por Freitas (2006), afirmando que os episódios com LOS prolongados são responsáveis por uma percentagem importante no total de dias de internamentos. Torna-se assim importante para a comunidade, continuar o estudo desenvolvido de 2000 a 2012 e poder comparar com os dados de 2013 a 2017, devido a todas as alterações organizacionais identificadas anteriormente.

Com a elaboração desta investigação, pretende-se apoiar adequadamente a tomada de decisão do HFAR, e que as suas conclusões potenciem um processo de internamento mais eficiente, otimizando a gestão das camas disponíveis, proporcionando uma ocupação média mais elevada e menor desperdício de recursos hospitalares (Azari, Janeja, & Mohseni, 2012). Assim sendo, apresenta-se como objetivo geral (OG), obter um modelo preditivo otimizado de LOS de pacientes no HFAR, através da descoberta de comportamentos e padrões existentes no processo de internamento hospitalar, com base em técnicas de DM. Para o alcance do objetivo geral, é necessário a realização dos seguintes objetivos específicos (OE): OE1 – Construir um

modelo para previsão do LOS; OE2 - Avaliar os resultados obtidos com estudos idênticos anteriores.

Ao nível da problemática da investigação, pretende-se responder a duas perguntas de investigação (PI):

PI1 - "Qual o melhor modelo de previsão de LOS de pacientes no HFAR, com base em técnicas de DM, por forma a otimizar o LOS hospitalar?"

PI2 – "De que forma, após a fusão dos hospitais militares, o HFAR conseguiu otimizar o LOS hospitalar?"

Decorrente das PI, foram formuladas as seguintes hipóteses (H): H1 - Obteve-se um modelo que permitiu efetuar previsões com uma margem de erro inferior a 30%; H2 - Após a fusão dos hospitais militares, o HFAR otimizou o LOS hospitalar, implementando uma melhor gestão e planeamento das altas hospitalares, o que permitiu aumentar o número de camas disponíveis para novas admissões.

Metodologicamente, esta investigação será desenvolvida de acordo com o raciocínio hipotético-dedutivo, com uma estratégia de investigação quantitativa, num horizonte temporal longitudinal e um desenho de pesquisa do tipo estudo de caso e comparativo. Os internamentos realizados no HFAR compreendidos entre 2013 e 2017 serão a população alvo deste estudo e, atendendo aos objetivos e ao problema, a metodologia que se revela mais adequada é a metodologia CRISP-DM, obtendo uma preferência de 43% entre os profissionais para resolução de problemas que envolvam o DM (Piatetsky, 2014).

Método

No decurso da fase exploratória e com o fim da pesquisa bibliográfica inicial, identificou-se um conjunto de conceitos importantes para a presente investigação. Segundo Pinto (2009),

o termo *business intelligence* (BI) surge em 1958, quando Hans Peter Luhn, alemão, cientista informático da IBM, conceptualizou um primeiro sistema de BI. Em 1980, o *Gartner Group*, consultora na área de tecnologia de informação, torna pública a criação do termo BI.

O conceito BI surgiu da necessidade de análise de grandes quantidades de dados e integração de dados provenientes de diversos sistemas operacionais, sendo que, a aquisição e tratamento de informação seriam irrelevantes se os mesmos não gerassem conhecimento. A crescente importância do conhecimento exige inovação técnica, levantando questões sobre como as organizações criam e processam novos conhecimentos (Pinto, 2009).

O uso do conhecimento é um fator crítico de sucesso de uma organização, obrigando ao investimento em mais meios e tornando-a mais eficiente no processo de produzir e disseminar conhecimento (Silva, 2010). Os sistemas de BI, de uma forma resumida, permitem a exploração e a análise de grandes quantidades de dados com o objetivo de identificar padrões e tendências nos dados (Berry & Linoff, 2004).

O termo *data mining* (DM) foi também um dos conceitos identificados e a primeira definição de DM surge de Fayyad, Piatetsky-Shapiro, e Smyth (1996), como uma das fases do processo de descoberta de conhecimento em bases de dados. Consiste na produção de modelos, através da aplicação de análise dos dados e de algoritmos de descoberta, dentro de limitações computacionais aceitáveis. Definição semelhante de DM surge de Han e Kamber (2001), com a identificação de padrões, relacionamentos ou modelos implícitos nos dados armazenados em grandes bases de dados.

O processo DM apresenta-se como sendo um processo exploratório e iterativo, pois através da análise dos dados, novo conhecimento é descoberto e novas hipóteses são formuladas. Engloba várias fases, designadamente, a perceção do domínio, a compreensão dos dados, a preparação dos dados, a aplicação de algoritmos de DM, a avaliação do conhecimento

descoberto, o uso do conhecimento descoberto, sendo a fase de preparação dos dados a fase mais longa de todo o processo (Freitas, 2006).

O *hospital length of stay*, já anteriormente identificado como LOS, foi outro conceito a ser identificado, termo definido como a duração de um episódio de internamento, calculado a partir do dia da admissão até o dia da alta, com base no número de noites passadas no hospital. Os pacientes admitidos no mesmo dia da alta têm um tempo de internamento menor que um dia.

O último dos conceitos identificados foi o *cross-industry standard process for data mining* (CRISP-DM). A metodologia CRISP-DM, descreve algumas aproximações utilizadas por especialistas para resolver problemas e construir modelos de análise preditiva. A metodologia em questão surgiu da necessidade de definir um modelo processual padrão, não-proprietário e gratuito, para sistematizar a descoberta de conhecimento.

Figura 1: Representação dos quatro níveis metodologia CRISP-DM

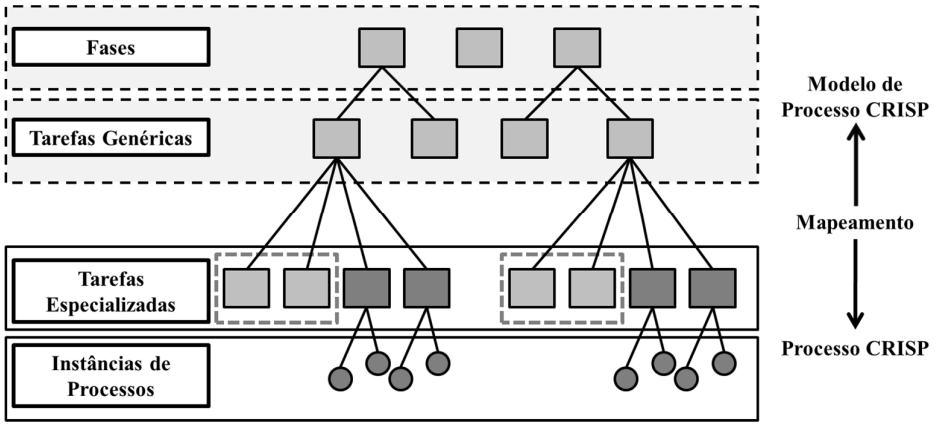

Figura 1. A metodologia CRISP-DM é representada por um modelo hierárquico de processos e constituída por um conjunto de tarefas com quatro níveis de abstração: Fases, tarefas genéricas, tarefas especializadas e instâncias de processos. Adaptado de "CRISP-DM 1.0 - Step-by-step data mining guide" por P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth, 2000, *CRISP-DM Consortium*, p. 6.

Figura 2: Representação das seis fases da metodologia CRISP-DM

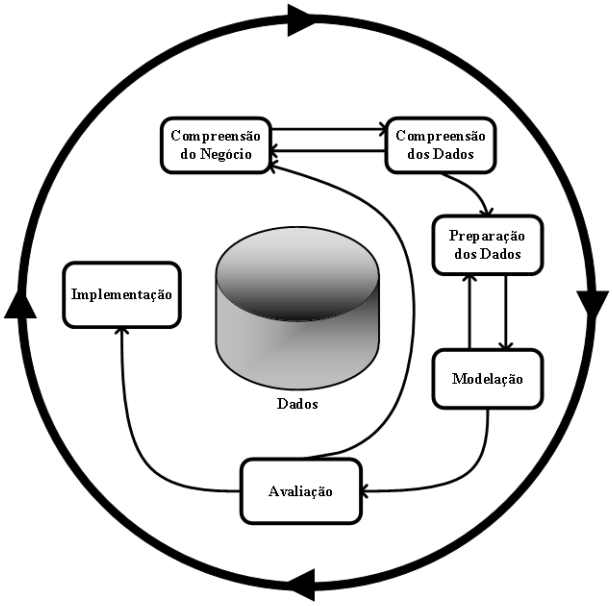

Figura 2. O CRISP-DM divide-se nas seguintes fases: *Business understanding, data understanding, data preparation, modeling, evaluation, deployment*. As fases do ciclo de vida são flexíveis, pois em qualquer ponto do projeto poderá ser necessário recuar para fases anteriores, sendo que o resultado de cada fase determina qual a próxima fase a executar. Adaptado de "CRISP-DM 1.0 - Step-by-step data mining guide" por P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth, 2000, *CRISP-DM Consortium*, p. 10. Copyright 2000 pela SPSS Inc.

Compreensão do Negócio

Nesta fase pretendeu-se compreender os objetivos e requisitos da perspetiva do negócio, convertendo esse conhecimento para a definição de um problema de DM, definição de um plano e critérios de sucesso para alcançar os objetivos (Chapman et al., 2000).

Ao nível dos objetivos de negócio foi apresentado como objetivo principal obter um modelo preditivo de LOS de pacientes no HFAR, através da descoberta de comportamentos e padrões existentes no processo de internamento hospitalar, com base em técnicas de DM. Para apoiar adequadamente a tomada de decisão do HFAR e potenciar um processo de internamento mais eficiente, ao nível dos critérios de sucesso do negócio, a escolha recairá no modelo de regressão que produza os valores mais próximos dos dados de origem.

Na subfase relativa à avaliação da situação atual, e de modo a definir o plano de projeto, realizou-se uma entrevista exploratória à BGEN/MED Regina Ramos, Diretora do HFAR. Questionada sobre a fusão dos hospitais militares e consequente redução de camas de internamento disponíveis, a Diretora do HFAR afirmou que a fusão dos hospitais era uma inevitabilidade, lógica e muito vantajosa do ponto de vista da gestão de recursos. A redução de camas não teve qualquer consequência negativa porque a taxa de ocupação não é a desejada para uma melhor produtividade. Na questão seguinte, solicitou-se a opinião sobre sistemas de BI que permitam prever e otimizar os LOS, à qual respondeu que uma gestão adequada de qualquer empresa, incluindo um hospital, só pode retirar vantagens da utilização das ferramentas disponíveis atualmente, nomeadamente para a previsão de LOS.

No âmbito da recolha de dados, o HFAR dispõe de um sistema de informação para registo hospitalar, suportado numa BD relacional Oracle 11G. Através de manipulação de dados via *structured query language* (SQL), os dados para futura análise serão extraídos para o ficheiro internamento em formato **.xls, *.csv*, mantendo a confidencialidade da informação e o anonimato dos utentes. Após a recolha de todos os dados, proceder-se-á à análise e tratamento dos mesmos, por meio das ferramentas *open source (R, Rattle),* como também o caso da biblioteca *rminer*.

O ambiente de programação R foi desenvolvido por Ross Ihaka e Robert Gentleman do Departamento de Estatística da Universidade de Auckland, Nova Zelândia. Conforme referenciado por Costa (2011), este sistema permite uma programação direcionada para a estatística e análise de dados, sendo que a adoção da biblioteca *rminer* para a ferramenta R facilita o uso de algoritmos de DM nas tarefas de regressão, através de um conjunto reduzido e coerente de funções (Cortez, 2010). Esta biblioteca permite ainda o cálculo de diversas

métricas e gráficos, incluindo procedimentos de análise de sensibilidade para extrair informação a partir de modelos treinados (Costa, 2011).

No passo seguinte pretendeu-se definir as metas de negócio e transpor os objetivos de negócio para a análise de dados. Deste modo, em coordenação com a direção do HFAR, identificou-se que o modelo deverá prever o LOS com uma margem de erro inferior a 30%. Conforme referido por R. R. Mateus (entrevista por *email*, 20 de setembro de 2018):

> No caso da saúde, esse cálculo de previsibilidade deverá ficar exclusivamente no âmbito dos gestores e não poderá refletir-se ou ter qualquer impacto na abordagem clínica dos doentes. Isto é, os médicos que têm os doentes à sua responsabilidade não podem estar, de forma alguma, condicionados por taxas ou estatísticas, não podem gerir a sua atividade no sentido de cumprir qualquer taxa. O interesse do doente é mandatório.

Para alcançar a meta estabelecida anteriormente, pretende-se escolher o modelo de regressão que produza os valores mais próximos dos dados, em que o modelo de regressão ideal apresenta um valor próximo de 0%. Por fim, para a fase de modelação dos dados foram identificadas as seguintes cinco técnicas de regressão: *decision tree* (DT), *naive, multiple regression* (MR), *random forest* (RF) e *support vector machines* (SVM).

Compreensão dos Dados

Esta fase iniciou-se com a aquisição dos dados iniciais, descrição, exploração e verificação da qualidade dos mesmos. Todo o processo de internamento fora identificado em investigação anterior, sendo que todos os atributos extraídos já se resumiam a um registo por utente e por número de processo de internamento.

A extração de dados foi efetuada via ferramenta *SQL Navigator*, executado o *script*[3] para aquisição dos dados iniciais, o que permitiu a transposição para o ficheiro de manipulação das técnicas de DM (*internamento.xls*). Foram inicialmente selecionados 44 atributos e os dados respeitam ao período de 1 de janeiro de 2013 a 31 de maio de 2017, registando 20291 episódios de internamento. A Tabela 1 do Apêndice A apresenta uma breve descrição dos 28 atributos que servirão de suporte à continuação da investigação, anteriormente referenciados por um painel de especialistas (Caetano, 2013). Nas tarefas seguintes de exploração, análise estatística e distribuição dos atributos, verificação da qualidade dos dados, recorreu-se às ferramentas *microsoft excel* para eliminação de carateres portugueses (incompatíveis com a biblioteca *open source rminer*) e *open source* R.

A análise aos dados[4] identificou que os atributos qualitativos representam a maioria dos atributos selecionados (68%) e os quantitativos representam a minoria (32%). Verificou-se a existência de valores omissos para alguns dos atributos (por exemplo o estado civil com 8434 valores em falta), e que alguns dos atributos qualitativos apresentam um elevado número de valores possíveis (a titulo de exemplo as 12703 horas possíveis de entrada no internamento), representando uma dispersão elevada, podendo dificultar a utilização destes mesmos atributos pelas técnicas de DM.

Procedeu-se de seguida à visualização gráfica das distribuições dos atributos, utilizando o diagrama de frequências e o diagrama de extremos e quartis (*boxplot*), que permitiram identificar os valores discrepantes (*outliers*) e analisar o grau de assimetria das distribuições. Tendo em vista a validação dos diversos atributos e sendo que o seu processamento e análise

[3] Documento suplementar *script sql* para aquisição do *dataset* inicial.

[4] Documento suplementar frequências dos atributos selecionados para LOS.

são complexos, apresentam-se como exemplo a Figura 3 abaixo, com o diagrama de frequências, extremos e quartis, e a Tabela 2 do Apêndice A com o código R gerado.

Figura 3: Diagrama de frequência, de extremos e quartis do atributo Idade_Intern

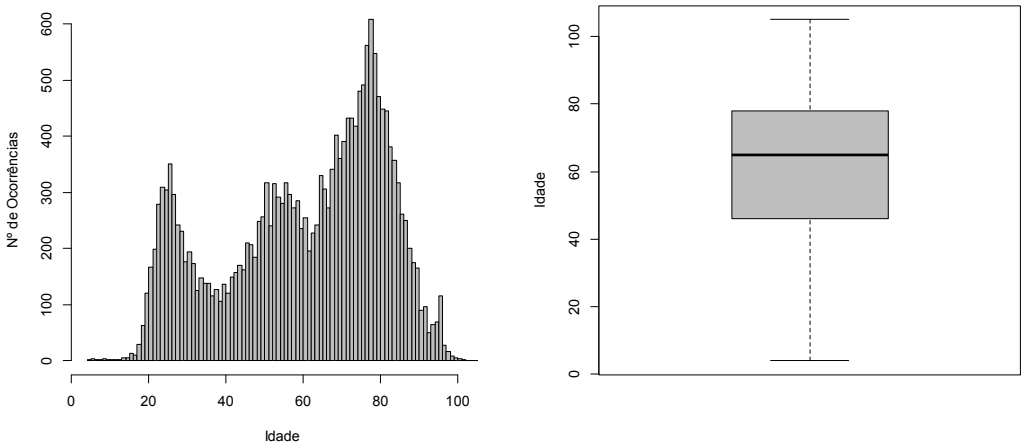

Figura 3. Na figura verifica-se que o atributo associado à idade do utente apresenta valor máximo de 100 anos de idade, em que a mediana se situa nos 65 anos de idade.

Sumarizando a análise dos 28 atributos selecionados, no âmbito do processo de internamento, concluiu-se que o utente do HFAR é na sua maioria do sexo masculino, predominantemente no estado civil casado(a), possuindo ao nível da escolaridade a formação básica (3º ciclo). Foi durante o mês de janeiro que se registaram mais episódios de internamento, principalmente nos serviços Departamento Cirúrgico - Ala B e Departamento Cirúrgico - Ala A. Por fim, encontra-se disponível na Tabela 3 do Apêndice A uma análise estatística dos atributos quantitativos para consulta.

Preparação dos Dados

Nesta fase pretendeu-se analisar os atributos anteriormente selecionados, transformá-los, limpá-los, abrangendo todas as atividades necessárias para construir o conjunto de dados final (*dataset*) (Chapman et al., 2000).

Decorrente da tarefa de limpeza dos dados, diversas ações foram efetuadas, entre as quais: remoção de três episódios de internamento, pois ocorreram em situação de testes aplicacionais; eliminação do número de níveis dos diversos atributos, de modo a facilitar a tarefa de modelação e duplicação dos dados. No caso do atributo Escolaridade, 1767 instâncias foram substituídas pelo valor "NA" (*Not Available*), pois os códigos não apresentavam descritivo associado. Processo semelhante ocorreu com as 215 instâncias do atributo Est_Civil, pois apresentavam o descritivo "Desconhecido". Ainda no decorrer desta tarefa, foram excluídos dez atributos devido ao seu elevado número de níveis, existência de atributo redundante, elevado número de valores omissos e baixa relevância teórica (Tabela 1 do Apêndice B). Após a realização de uma primeira limpeza dos dados, totalizou-se 18 atributos significativos e 20287 episódios.

Para resolver a questão dos valores omissos em diversos atributos, pois a sua existência contribui para a diminuição da qualidade dos modelos, pode-se, por exemplo, ignorar os registos ou substituí-los aplicando diversas técnicas: c*ase substitution* (valor atribuído por um perito), *cold deck* (valor retirado de uma base de dados) e *hot deck*. Nesta investigação aplicou-se a técnica *hot deck* (exemplo código R na Tabela 2 do Apêndice B), que consiste em procurar o exemplo mais próximo ou semelhante e consequente substituição dos valores omissos pelo valor encontrado (Brown & Kros, 2003).

Tabela 1: Tratamento dos valores omissos (*hot deck*)

Atributo	Nº valores omissos
Escolaridade	17986
Est_Civil	8646
Proc_Principal	3661
Diag_Principal	3342
Diag_Inicial	7684
GCD	3350

Nota. Dimensão da amostra (20287).

Figura 4: Diagrama de frequência, de extremos e quartis do atributo N_Dias_Intern

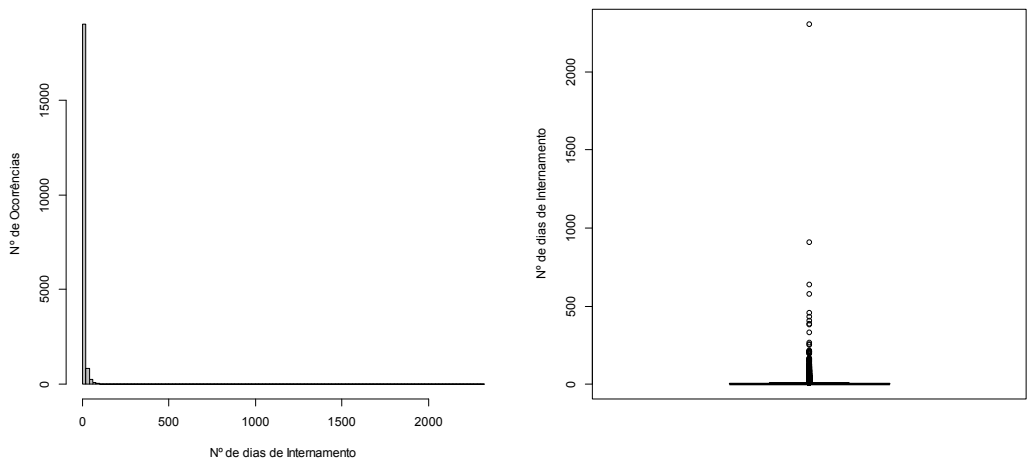

Figura 4. Na figura verifica-se o caso de uma instância discrepante, que foi considerada valor errado (episódio de testes) e posteriormente eliminada. O caso discrepante apresentava 2306 dias de internamento hospitalar. O maior agrupamento de valores observados situam-se próximos da média (6,19) e menores frequências aquando do seu afastamento.

Figura 5: Diagrama de frequência, de extremos e quartis do atributo N_Intern_Anterior

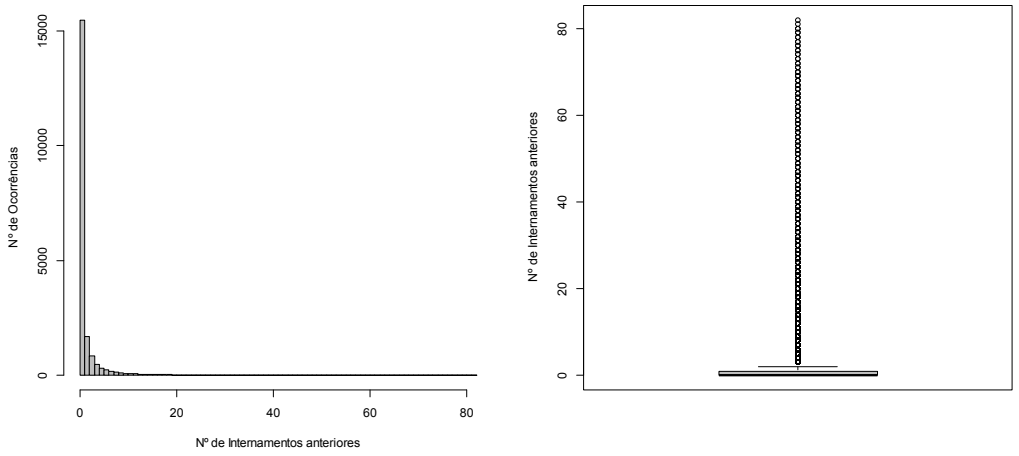

Figura 5. Na figura de extremos e quartis verifica-se que o maior agrupamento de valores situam-se próximos da média (1,82), existindo diversos valores discrepantes que se afastam muito do padrão, no entanto não foram considerados valores errados.

As variáveis N_Dias_Intern e N_Intern_Anterior (Figuras 4 e 5) encontram-se enviesadas para o estremo esquerdo do seu domínio de valores, sendo que foi aplicada uma técnica da transformação (função *log1p(x)* disponível na ferramenta R), que permitiu calcular o $\ln(x+1)$ com precisão e facilitar o processo de aprendizagem na fase seguinte, tendo-se obtido novos atributos transformados (LG_N_Dias_Intern e LG_N_Intern_Anterior).

Tabela 2: Sumário estatístico: atributos LG_N_Dias_Intern e LG_N_Intern_Anterior

Atributo	Média	Min	P25	Mediana	P75	Máximo
LG_N_Intern_Anterior	0,54	0,00	0,00	0,00	0,69	4,42
LG_N_Dias_Intern	1,27	0,00	0,00	1,39	1,10	6,82

Nota. Min – Mínimo, P25 – Percentil 25, P75 – Percentil 75.

Figura 6: Diagrama de frequência e *bloxpot* do atributo LG_N_Dias_Intern

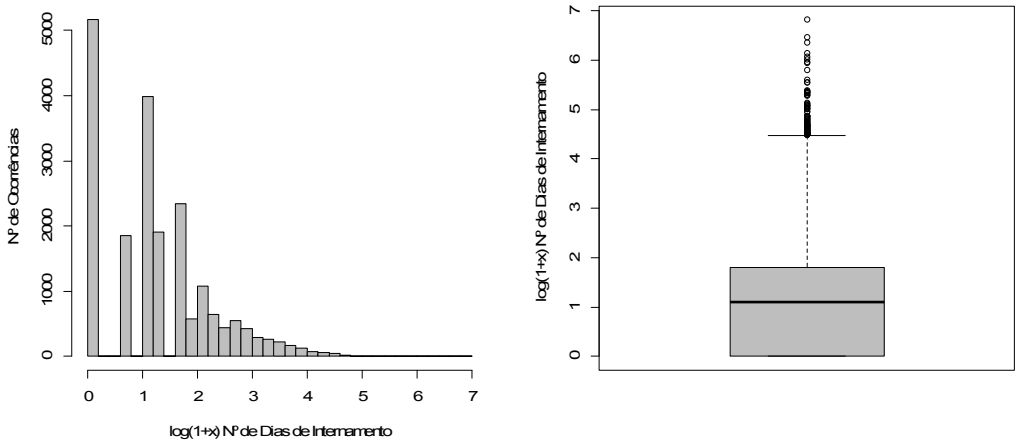

Figura 6. Na figura constata-se que o diagrama de frequência apresenta uma distribuição assimétrica à esquerda. O diagrama de extremos e quartis apresenta um maior enviesamento no extremo superior e um enviesamento inferior na zona central dos dados.

Figura 7: Diagrama de frequência e *bloxpot* do atributo LG_N_Intern_Anterior

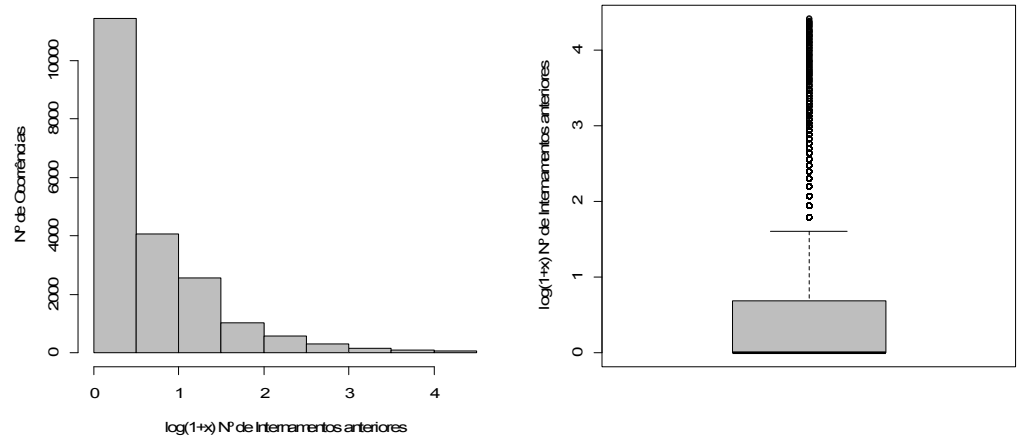

Figura 7. Na figura constata-se que o diagrama de frequência apresenta uma distribuição assimétrica à esquerda. O diagrama de extremos e quartis apresenta um relativo enviesamento no extremo superior (dados estão menos concentrados na parte superior que na inferior) e valores não considerados *outliers*, apesar de residirem acima do intervalo superior.

Concluída a tarefa de limpeza dos dados, avançou-se para a tarefa seguinte: construção dos dados com a inclusão de novos atributos (derivados) que apresentaram uma boa relevância para a restante investigação, podendo-se analisar os mesmos na Tabela 3 e Figuras 1 e 2 do Apêndice B.

Tabela 3: Atributos derivados incluídos no *dataset*

Nome do Atributo	Atributo Origem	Descrição	Valores Possíveis	Classificação Original	Tipo de Tratamento
DiaSemana_Intern	Dt_Internamento	Dia de semana do internamento.	[1 (Segunda) a 7 (Domingo)]	Ordinal	Numérica
Trimestre_Intern	Mes_Intern	Trimestre do internamento.	[1 – 4]	Ordinal	Numérica

Nota. Atributos derivados – criados com base noutros atributos.

As Figuras 3 e 4 do Apêndice B representam a transformação (formato alterado para "HH") dos atributos Hora_Intern e Hora_Alta_Intern, pois possuíam 12703 e 8336 níveis respetivamente.

Tabela 4: Recodificação do atributo Escolaridade

Código Antigo	Código Novo	Descrição	Frequências
200	1	Sem habilitações	151
310	2	Básico (1. Ciclo)	4053
320	3	Básico (2. Ciclo)	3378
330	4	Básico (3. Ciclo)	5549
400; 620; 630; 700	5	Secundário	1289
810; 820; 830; 840; 850	6	Superior	5867

Nota. O atributo foi recodificado de 13 para 6 níveis.

Os atributos Escolaridade (Tabela 4 e Figura 5 do Apêndice B), Proc_Principal (Tabela 3 do Apêndice B e Figura 6 do Apêndice B), Diag_Principal[5] (Figura 7 do Apêndice B), Diag_Inicial[6] (Figura 8 do Apêndice B), Idade_Intern (Tabela 5 e Figura 9 do Apêndice B) apresentavam ainda muitos valores possíveis, tendo sido simplificados através da função *delevels*, reduzindo o seu número de níveis para análise.

Tabela 5: Recodificação do atributo Idade_Intern

Código Antigo	Novo Código	Descrição	Frequências
[0 – 14]	1	< 15 Anos	30
[15 – 44]	2	15 – 44 Anos	4802
[45 – 64]	3	45 – 64 Anos	5045
[65 – 84]	4	65 – 84 Anos	8485
[85 – 150]	5	≥ 85 Anos	1925

Nota. O atributo foi recodificado de 102 para 5 níveis.

Tabela 6: Recodificação do atributo Mes_Intern

Código Antigo	Novo Código	Frequências
Jan	1	2221
Feb	2	1971
Mar	3	2014
Apr	4	1712
May	5	1880
Jun	6	1534
Jul	7	1374
Aug	8	1138
Sep	9	1555
Oct	10	1612
Nov	11	1751
Dec	12	1525

Nota. O atributo foi recodificado de nominal para ordinal.

[5] Documento suplementar recodificação do atributo Diag_Principal.

[6] Documento suplementar recodificação do atributo Diag_Inicial.

Por fim, e para melhor modelação dos dados, o atributo Mes_Intern também foi transformado conforme informação da Tabela 6 e Figura 10 do Apêndice B.

Após a realização dos procedimentos necessários para se obter uma boa modelação dos dados, avançou-se para a fase seguinte com um total de 20 atributos e 20287 registos, tendo-se analisado novamente as frequências selecionadas[7] e disponibilizado na Tabela 4 do Apêndice B diversos exemplos das funções pertencentes à biblioteca *rminer* executadas nesta fase.

Modelação

Conforme Chapman et al., (2000), nesta fase selecionam-se e aplicam-se diversas técnicas de modelação, tendo sido os seus parâmetros ajustados de forma a otimizar os resultados.

Torna-se importante referir que, sendo quantitativo o atributo a prever, a abordagem escolhida e explorada foi a regressão, tendo-se testado diferentes técnicas e métricas, como por exemplo a seleção das cinco técnicas de regressão mencionadas anteriormente na fase de compreensão de negócio (DT, *naive,* MR, RF, SVM).

A técnica *naive* representa a previsão da média dos valores da variável de saída, a DT funciona bem com grandes conjuntos de dados, muitas variáveis, com diferentes tipos de dados, sendo a sua estrutura relativamente fácil de seguir e compreender (Z. Michalewicz, Schmidt, Michalewicz & Chiriac, 2006). A MR é um modelo que exige pouco processamento e de fácil interpretação pelo ser humano, enquanto que a RF é composta por um conjunto de DT, construídas com base num conjunto de dados originais. Por fim, as SVM baseiam-se na definição e utilização de vetores de suporte que contenham apenas os exemplos mais representativos do universo de treino, aplicando posteriormente uma transformação não linear

[7] Documento suplementar frequências finais dos atributos.

aos atributos de entrada através de uma função de *kernel,* que permite definir o híper plano ótimo de separação entre as possíveis classes de saída (Moro, 2011).

Tendo como objetivo a escolha do modelo que melhores resultados obtém, tornou-se necessário a aplicação de procedimentos de validação para avaliar a capacidade de previsão dos modelos obtidos.

No método de validação *holdout*, os dados da amostra são divididos em dois conjuntos: O conjunto de dados de treino (2/3) é utilizado para construir, identificar e para estimar os parâmetros do modelo, enquanto que o conjunto de dados de teste (1/3) é utilizado para avaliar a precisão e o desempenho do modelo. O modelo de eleição deverá ser o que melhor generalize os dados treinados e o que melhor se identifique na aprendizagem de novos casos.

Outro método de validação utilizado é o *k-fold*, exigindo um procedimento mais elaborado. Os dados são divididos em k partições de igual tamanho e em cada execução é testado um determinado subconjunto, sendo que os restantes são utilizados para treino do modelo. No caso k = 5, em cada rotação é treinado um modelo e a estimativa global do modelo é dada pelo erro médio do teste das k rotações.

Realça-se mais uma vez, que o ambiente de programação escolhido foi o ambiente R e a biblioteca *open source rminer*. O problema de regressão iniciou-se com a exploração de modelos de previsão mais simples (*naive*, MR e DT), passando posteriormente para modelos mais complexos (SVM e RF).

Conforme informação disponível na Tabela 7, da biblioteca *open source rminer* destacam-se as funções *mining* e *savemining*. *Mining* é uma função que treina e testa um modelo ajustado às diversas execuções e métodos de validação: O parâmetro *runs* indica o número de execuções do processo, o parâmetro *model* indica a técnica escolhida e o parâmetro *method* indica a forma de validação. O método de validação *holdout* foi o selecionado, com 2/3

para treino e 1/3 para testes, e com a finalização deste processo, a função *savemining* arquiva a informação e resultados obtidos do modelo para posterior análise e tratamento.

Tabela 7: Código para obtenção e teste do modelo RF – *runs* = 1

```
library(rminer)
library(randomForest)
d<-read.table("internamentoR.csv",header=TRUE,sep=",")
M=mining(LG_N_Dias_Intern~.,data=d,Runs=1,method=c("holdout",2/3),model="randomforest",
search="heuristic10")
savemining(M,"internamento_randomforest.model")
```

Nota. A tabela descreve o código gerado para obtenção do modelo RF através da função *mining,* com uma execução do método.

Finalizada esta experiência inicial e por forma a obter-se maior robustez dos resultados alcançados, realizaram-se 20 execuções de cada técnica dos métodos de validação cruzada *holdout* e *5-fold*, salientando que foram aplicadas 20 execuções a cada procedimento *5-fold*, perfazendo o total de 100 experiências para cada teste.

Tabela 8: Código para obtenção e teste do modelo RF – *runs* = 20

```
d<-read.table("internamentoR.csv",header=TRUE,sep=",")
M=mining(LG_N_Dias_Intern~.,data=d,Runs=20,method=c("holdout",2/3),model="randomforest",
search="heuristic10")
savemining(M,"internamento_randomforest20.model")
M=mining(LG_N_Dias_Intern~.,data=d,Runs=20,method=c("kfold",5),model="randomforest",
search="heuristic10")
savemining(M,"internamento_randomforest20_5.model")
```

Nota. A tabela descreve o código gerado para obtenção do modelo RF através da função *mining*, com 20 execuções do método.

Terminadas todas as tarefas para obtenção dos diversos modelos, avançou-se para a fase de avaliação dos resultados, encontrando-se descrito em material suplementar[8] os exemplos dos métodos em código *rminer* executados nesta fase.

Avaliação

O objetivo desta fase foi avaliar os modelos concebidos, verificando o seu comportamento em ambiente de teste e assegurando que cumprem os objetivos de negócio (Chapman et al., 2000), sendo que os modelos foram avaliados seguindo os objetivos de negócio anteriormente definidos.

Nos modelos de regressão pretende-se escolher aquele que produz valores mais próximos dos dados, tendo-se utilizado métricas de regressão para avaliar os diversos modelos (por exemplo a métrica R^2 permite medir o grau de correlação entre as previsões e valores observados). Com base na ferramenta *open source* R, concretamente a função *mmetric* pertencente à biblioteca *rminer,* tornou-se possível calcular as métricas de regressão e avaliar a qualidade dos modelos obtidos.

Tabela 9: Descrição das métricas de regressão

Métrica	Descrição	Expressão	Avaliação
R^2	Coeficiente de determinação	$R^2 = \frac{\sum_{i=1}^{n}(Y_{ai} - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$	]-Inf,1] - Melhores modelos situam-se com valores próximo do 1.
MAE	Erro médio absoluto	$MAE = \frac{\sum_{i=1}^{n}\lvert Y_i - Y_{ai}\rvert}{n}$	[0,Inf[- Melhores modelos situam-se com valores próximos do 0.
RMSE	Raiz do erro quadrático médio	$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - Y_{ai})^2}{n}}$	

Nota. n: dimensão da amostra de teste; y_i: valor observado para cada indivíduo; y_{ai}: valor estimado para cada indivíduo; $\overline{Y}$: valor médio do atributo a prever.

[8] Documento suplementar código R utilizado durante a fase de modelação.

Outro método utilizado para avaliar e comparar os modelos de regressão foi a *regression error characteristics* (REC). A curva REC permite comparar vários modelos de regressão num único gráfico, expondo a taxa de acerto global (eixo das ordenadas) para diversos valores de tolerância (T) de erro absoluto (eixo das abcissas). Conclui-se que a taxa de acertos é definida como a percentagem de pontos que se encaixam dentro da tolerância, sendo que se a tolerância fosse zero, apenas os pontos de previsão perfeita seriam considerados (Silva, 2010).

O último método utilizado para avaliar a qualidade dos resultados obtidos pelo melhor modelo foi o gráfico de dispersão *regression scatter characteristics* (RSC), em que nos eixos das abcissas estão representados os valores observados e no eixo das ordenadas os valores das previsões.

Por fim, e de modo a obter uma informação mais precisa e expressiva do melhor modelo obtido, fez-se uso do *relative input importance barplot* (IMP) e da *variable effect curve* (VEC), proposta em Cortez e Embrechts (2013). A função *mgraph* da biblioteca *rminer* permite obter a curva VEC para os atributos mais importantes na construção do modelo.

Excerto do código executado em R para calculo das métricas e obtenção dos gráficos encontram-se descriminados com maior detalhe no Apêndice C.

Implementação

Apesar de os modelos obtidos apresentarem boa qualidade, o tempo disponível para esta investigação, não permite o desenvolvimento de software de base para implementação destes modelos. Um dos cenários possíveis de ser seguido pela direção do HFAR, passa pela incorporação dos modelos obtidos no seu sistema atual, ou seja, no *electronic patient record system* (EPR) que faz parte do *hospital information system* (HIS).

Resultados e Discussão

Com a aplicação das métricas de regressão, após seleção nas fases anteriores das técnicas de regressão e aplicação do método de validação *holdout* (2/3 para treino e 1/3 para testes), foi possível a primeira avaliação da qualidade dos modelos obtidos. Posteriormente, efetuaram-se 20 execuções de cada técnica, obtendo-se os resultados dos valores de erro médio de cada modelo através da função *meanint* (calculo da média e intervalo de confiança).

Tabela 10: Métricas obtidas dos testes de validação *holdout*

Nº Execuções	Modelo	Métricas		
		R^2	MAE	RMSE
Runs = 1	Naive	0	0,8317796	1,057427
	MR	0,6532844	0,414906	0,6225638
	DT	0,60098	0,4324518	0,6680122
	RF	**0,7262002**	0,3248634	0,5545339
	SVM	0,6846953	0,3642389	0,5987021
Runs = 20	Naive	0	0,822 ± 0,003	1,043 ± 0,004
	MR	0,656 ± 0,003	0,407 ± 0,002	0,611 ± 0,004
	DT	0,604 ± 0,005	0,436 ± 0,008	0,655 ± 0,004
	RF	**0,73 ± 0,003**	0,314 ± 0,002	0,541 ± 0,003
	SVM	0,692 ± 0,004	0,355 ± 0,003	0,583 ± 0,005

Nota. Runs = nº de execuções de cada técnica de regressão.

Sendo estes os primeiros modelos, é de realçar o bom desempenho do modelo RF que obteve na métrica R^2 o resultado de 0,726 (*run* = 1) e 0,73 (*run* = 20). Salienta-se também o bom resultado nas restantes métricas comparativamente ao modelo SVM.

De forma a obter-se maior robustez nos resultados, efetuaram-se 20 execuções a cada método de validação *5-fold*.

Tabela 11: Métricas obtidas dos testes de validação *k-fold* (k = 5)

Modelo	Métricas		
	R^2	MAE	RMSE
Naive	0,000 ± 0,000	0,822 ± 0,000	1,042 ± 0,000
MR	0,657 ± 0,000	0,405 ± 0,000	0,61 ± 0,000
DT	0,606 ± 0,001	0,431 ± 0,002	0,655 ± 0,001
RF	**0,735 ± 0,000**	0,31 ± 0,000	0,537 ± 0,000
SVM	**0,695 ± 0,001**	0,352 ± 0,000	0,578 ± 0,001

Nota. Runs = 20.

Com base nos valores descriminados na tabela anterior, confirmou-se que as avaliações efetuadas pela métrica R^2, nomeadamente o valor do modelo RF (0,735), resultado do método de validação *5-fold*, é superior ao valor do modelo RF (0,73), resultado do método de validação *holdout* (*run* = 20), e que o mesmo se enquadra nos objetivos elencados pela direção do HFAR, identificando-se um modelo com capacidade de previsão do LOS com uma margem de erro inferior a 30%. Confirmou-se ainda que os melhores resultados foram obtidos pelo modelo de RF, que supera outros modelos de DM para todas as três métricas de erro.

Para verificar que os resultados dos valores de erro médio dos dois melhores modelos (RF e SVM) do método de validação *5-fold* são efetivamente diferentes entre si, foi efetuado o *t-test* (*student test*). O *t-test* é um teste de hipótese que permite rejeitar ou não uma hipótese nula, sendo que a probabilidade de erro se denomina de *p-value* e o nível de confiança por 1-*p-value* (probabilidade de o intervalo de confiança conter o valor do parâmetro). O *open source* R dispõe da função *t-test* para efetuar o teste paramétrico descrito na Tabela 1 do Apêndice D, tendo-se definido as seguintes hipóteses:

H0: A diferença do resultado das médias dos modelos RF e SVM é igual a 0.

Ha: A diferença do resultado das médias dos modelos RF e SVM é diferente de 0.

Com um nível de confiança de 95%, foi rejeitada a H0 (hipótese nula) pelo facto de existirem evidências estatísticas para se afirmar que a diferença do resultado das médias dos modelos RF e SVM são significativas, ou seja, o valor de $p < 0{,}05$. Com a rejeição da H0, concluiu-se que o resultado das médias do modelo RF é significativamente superior à média do modelo SVM.

$t_{(32{,}808)} = 133{,}3374$ e $p = 1{,}909882e^{-46}$

Para avaliar os resultados dos modelos da Tabela 11, usufruiu-se da função *mgraph* da biblioteca *rminer* (Tabela 2 do Apêndice C) que permitiu obter a análise REC expressa na Figura 8.

Figura 8: Gráfico da curva REC para os modelos RF e SVM

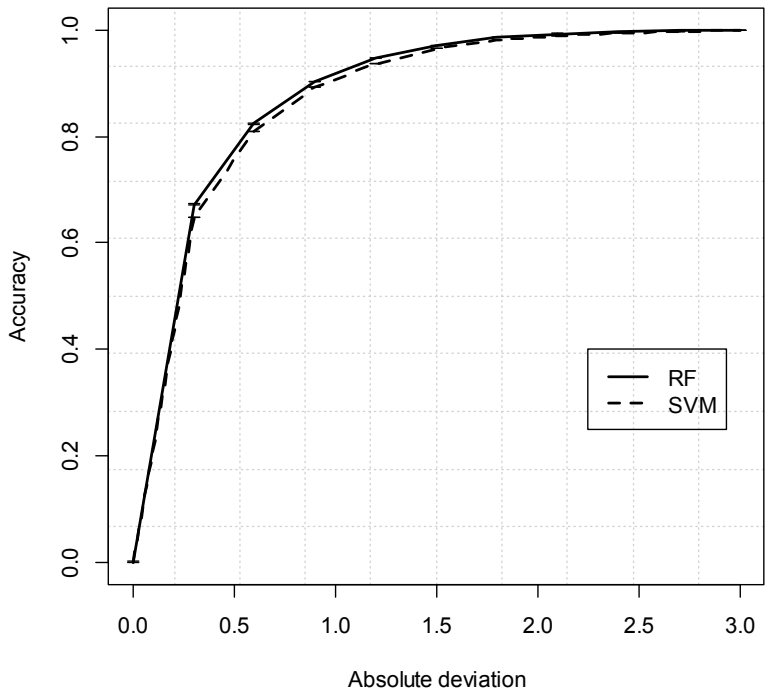

Figura 8. Na figura constata-se que os modelos RF e SVM apresentam uma boa capacidade de previsão. O modelo RF apresenta uma curva superior e bastante regular comparativamente com o modelo SVM.

A análise do gráfico REC confirmou o modelo RF como o melhor modelo preditivo, apresentando sempre uma maior precisão (eixo y) para qualquer valor de tolerância (eixo x). Por exemplo, para uma tolerância de 0,5 (na escala de transformação logarítmica), o modelo RF prevê corretamente 78,5% dos exemplos do conjunto de testes. Após uma primeira análise do gráfico REC, no próximo passo, foi possível através das métricas *normalized rec area* e *tolerance* pertencentes à função *mmetric* do *open source* R (código disponível na Tabela 2 do Apêndice D), obter os valores precisos dos dois melhores modelos (RF e SVM) para onze valores de desvio absoluto no intervalo [0,1].

Tabela 12: Precisão dos modelos RF e SVM para valores de desvio absoluto

Desvio Absoluto	Precisão RF	Precisão SVM	Valor (RF – SVM)
0,0	0,0%	0,0%	0,0%
0,1	44,8%	36,5%	8,3%
0,2	58,2%	55,2%	3,0%
0,3	67,3%	64,8%	2,5%
0,4	**73,6%**	71,8%	1,8%
0,5	**78,5%**	77,0%	1,5%
0,6	82,4%	81,0%	1,4%
0,7	85,6%	84,2%	1,4%
0,8	88,2%	86,9%	1,3%
0,9	90,3%	89,1%	1,2%
1,0	92%	90,9%	1,1%

Nota. Precisão adquirida com base na média das 20 execuções.

Se o valor do desvio for de 0,5 a taxa de acerto para o modelo RF será de 0,785 e de 0,77 para o modelo SVM (0,015 inferior ao modelo RF). Verificou-se também que para todos os valores de desvio absoluto, o modelo RF apresentou uma melhor precisão em comparação com o modelo SVM, com uma diferença que varia de 1,1% (para uma tolerância de 1,0) a 8,3% (para uma tolerância de 0,1).

Concluindo, garantindo o requisito de obtenção de um modelo que permitisse efetuar previsões com uma margem de erro inferior a 30%, o modelo RF com um desvio absoluto de 0,4 (na escala de transformação logarítmica) conseguiu prever acertadamente 73,6% dos casos e com um desvio absoluto de 0,5 (na escala de transformação logarítmica) obteve-se uma taxa de acertos de 78,5%.

Por último, aplicou-se novamente a função *mgraph* com o parâmetro *graph="RSC"* (Tabela 3 do Apêndice C), que permitiu obter o gráfico de dispersão RSC dos melhores modelos disponíveis na Figura 9 e os respetivos cálculos para previsão do erro máximo nos extremos inferior e superior, expressos na Tabela 13.

Tabela 13: Previsão erro máximo nos extremos dos modelos RF e SVM

Escala logarítmica: y = log(x+1)
Função inversa: x = exp(y)-1
A – Cálculo para o extremo inferior
0,5 = log(x1+1) - log(x2+1)
Erro máximo => 0,5
log(x2+1) = 0 => aplicando exp(0) - 1 => x2 = 0
log(x1+1) = 0,5 => aplicando exp(0,5) - 1 => x1 = 0,648721
significa que x1 - x2 = **0,65** => desvio em dias normais para **A**
B – Cálculo para o extremo superior
Valor mais alto é 4,4 e a tolerância é 0,5
4,4 = log(x1+1)
3,9 = log(x2+1)
log(x1+1) = 4,4 => aplicando exp(4,4) - 1 => x1 = 80,450
log(x2+1) = 3,9 => aplicando exp(3,9) - 1 => x2 = 48,402
significa que x1 - x2 = **32** => desvio em dias normais para **B**

Nota. Tolerância = 0,5.

Figura 9: Gráficos RSC dos modelos RF e SVM

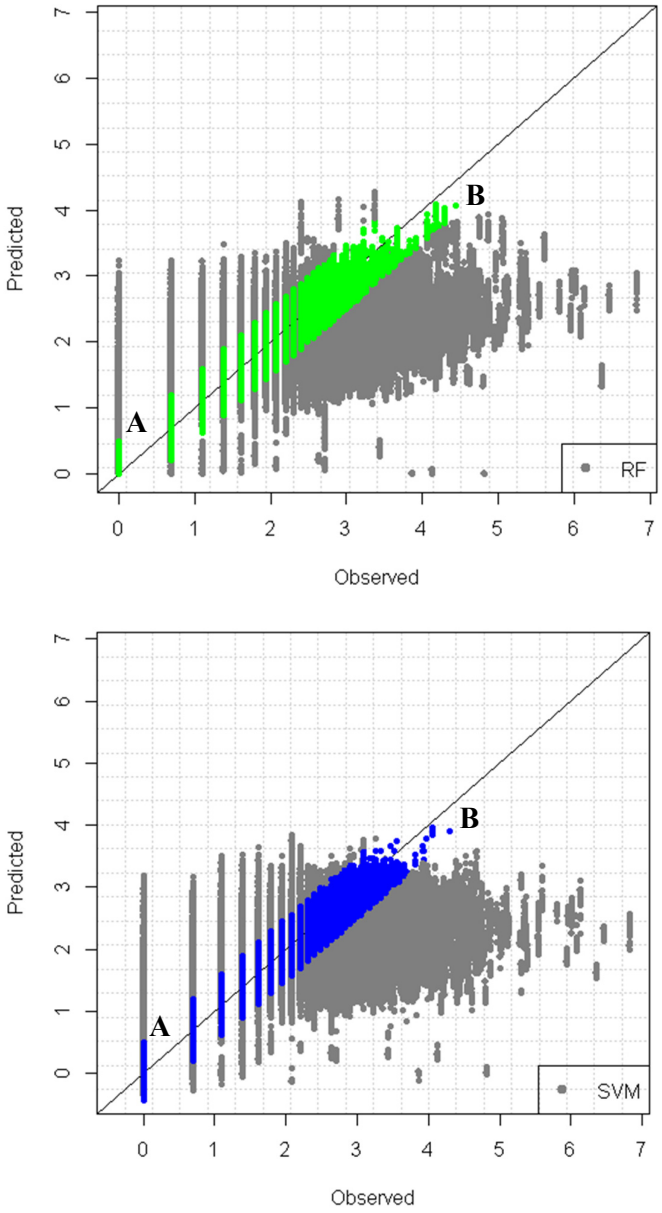

Figura 9. Nas figuras verifica-se que a maioria dos pontos se situa próximo da diagonal, demonstrando a qualidade dos modelos. Os pontos com desvio absoluto de 0,5 encontram-se representados nos gráficos com a cor verde e azul respetivamente.

Se para desvio absoluto de 0,5 o modelo RF obteve uma taxa de acertos de 78,5% de LOS, tal significa que dá um erro máximo de 0,65 dias para o extremo inferior da escala (0) e um erro máximo aproximado de 32 dias no extremo superior da escala (4,4).

Através da análise de sensibilidade de Cortez e Embrechts (2013), foi possível obter graficamente a importância relativa dos diversos atributos, demonstrado na Figura 10. Na Tabela 3 do Apêndice D verifica-se a aplicação da função *fit* e *importance* com o parâmetro *method="DSA"* da biblioteca *rminer*, permitindo extrair o conhecimento do modelo gerado, em termos de importância e média dos resultados dos atributos analisados.

Figura 10: Gráfico IMP do modelo RF

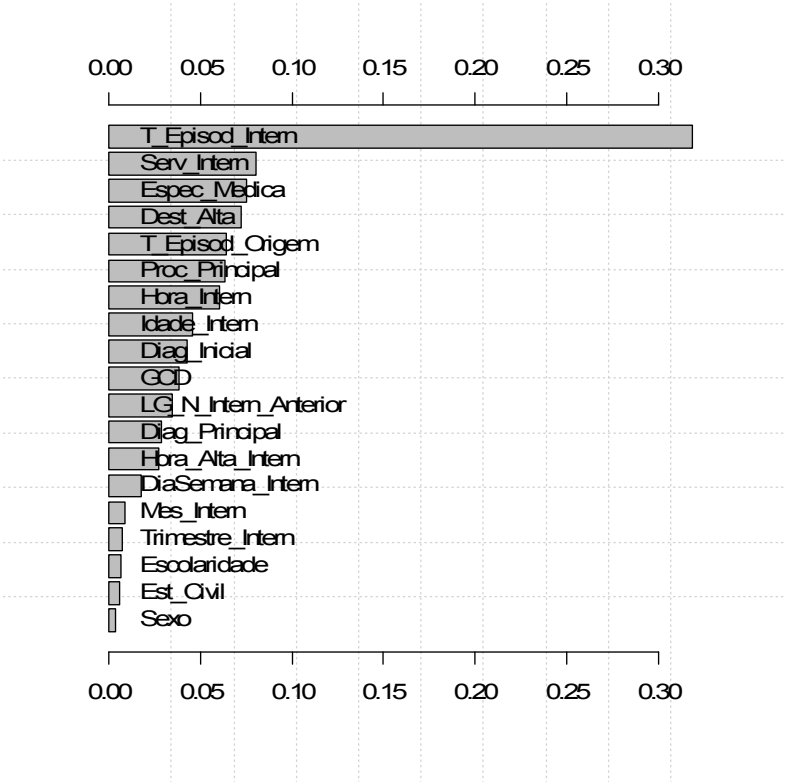

Figura 10. Na figura verificaram-se os seguintes valores de importância para a definição do modelo: T_Episod_Intern (31,9%); Serv_Intern (8%); Espec_Medica (7,5%); Dest_Alta (7,2%); T_Episod_Origem (6,4%); Proc_Principal (6,4%); Hora_Intern (6%); Idade_Intern (4,5%); Diag_Inicial (4,3%); GCD (3,8%); LG_N_Intern_Anterior (3,4%); Diag_Principal (2,8%); Hora_Alta_Intern (2,7%); DiaSemana_Intern (1,8%); Mes_Intern (0,9%); Trimestre_Intern (0,7%); Escolaridade (0,7%); Est_Civil (0,6%); Sexo (0,3%).

Conforme análise do gráfico anterior, confirmou-se que o atributo tipo de episódio de internamento (T_Episod_Intern) apresenta uma importância de 31,9% na definição do modelo gerado, seguindo-se como atributos mais significativos: serviço de internamento (Serv_Intern)

com 8%, especialidade médica (Espec_Medica) com 7,5% e por último o destino da alta (Dest_Alta) com 7,2%. Por outro lado, verificou-se ainda que estes quatro atributos mais significativos relacionados com a situação clínica dos pacientes, contribuem em mais de 50% para a capacidade explicativa do modelo gerado.

Na última fase de análise e de modo a detalhar a influência dos atributos mais significativos na construção do modelo RF, aplicou-se o gráfico VEC em conjunto com a função *importance* conforme código explicativo na Tabela 4 do Apêndice D.

Figura 11: Gráfico de influência do atributo Tipo de Episódio de Internamento

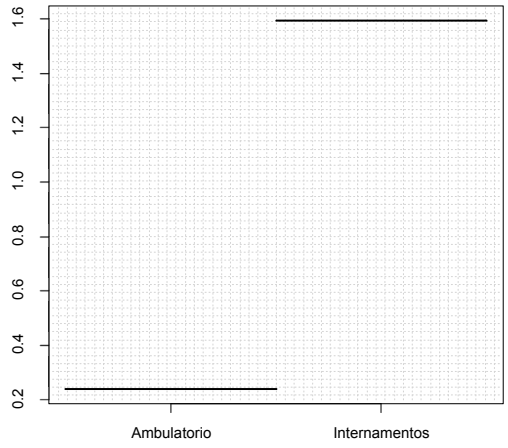

Figura 11. Na figura os segmentos de linha VEC mostram a influência média do tipo de episódio de internamento (eixo x) na saída do modelo RF (eixo y).

A Figura 11 mostra a influência global do atributo de entrada mais relevante (Tipo de Episódio de Internamento), que é um atributo nominal com duas classes. Os segmentos da linha VEC claramente confirmaram que um episódio de internamento em regime de Ambulatório está relacionado com um LOS médio mais baixo (0,24 na escala de transformação logarítmica,

0,27 dias na escala normal) quando comparado com o regime em Internamento (1,59 na escala de transformação logarítmica, 3,9 dias na escala normal).

Figura 12: Gráfico de influência do atributo Serviço de Internamento

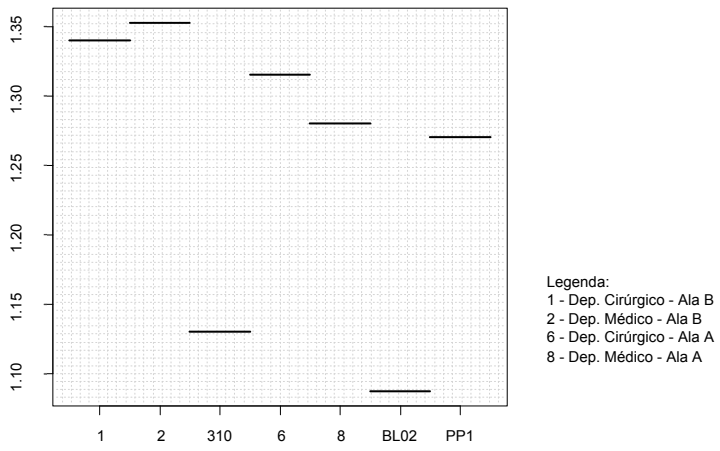

Figura 12. Na figura os segmentos de linha VEC mostram a influência média do serviço de internamento (eixo x) na saída do modelo RF (eixo y).

Relativamente à media de influência do atributo Serviço de Internamento representado na Figura 12, observou-se um maior LOS no serviço Departamento Médico - Ala B com um valor de 2,86 dias (1,35 na escala de transformação logarítmica), seguindo-se o serviço Departamento Cirúrgico - Ala B com um valor de 2,82 dias (1,34 na escala de transformação logarítmica), o serviço Departamento Médico - Ala A com um valor de 2,74 dias (1,32 na escala de transformação logarítmica), e por fim o serviço Departamento Cirúrgico - Ala A com o valor de 2,6 dias (1,28 na escala de transformação logarítmica).

Figura 13: Gráfico de influência do atributo Especialidade Médica

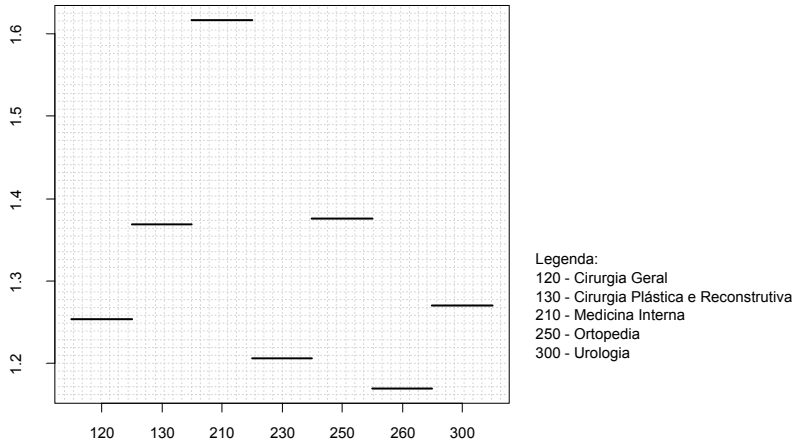

Figura 13. Na figura os segmentos de linha VEC mostram a influência média da especialidade médica (eixo x) na saída do modelo RF (eixo y).

Da análise do terceiro atributo mais relevante representado na Figura 13, a Especialidade Médica, verificou-se que a Medicina Interna está associada ao maior LOS (1,61 correspondendo a 4 dias), seguida da Ortopedia (1,38 correspondendo a 2,97 dias), da Cirurgia Plástica e Reconstrutiva (1,37 correspondendo a 2,94 dias) e por fim a Urologia (1,27 correspondendo a 2,56 dias).

De modo a cumprir o OE2, confirmando a H2 e respondendo à PI2, foram comparados os resultados dos melhores modelos obtidos em estudos anteriores (Caetano (2013); Caetano et al. (2015)). Importante realçar que no estudo de 2013 foram utilizados 18 atributos, 14 atributos em 2015 e 19 atributos nesta investigação para previsão do LOS (2018).

Tabela 14: Análise do atributo Tipo de Episódio de Internamento

Estudo	Importância em %	Caraterística	Nº Dias
2013	26,1 %	Internamentos	**3,81**
		Ambulatório	0,27
2015	30,1 %	Internamentos	3,9
		Ambulatório	**0,1**
2018	**31,9 %**	Internamentos	3,9
		Ambulatório	0,27

Nota. Episódio em regime de ambulatório até 24 horas de LOS.

Da Tabela 14, concluiu-se que o atributo Tipo de Episódio de Internamento em regime de Ambulatório foi mais baixo no estudo de 2015 (0,1 dias). Em regime de Internamento apresentou o valor mais baixo no estudo de 2013 (3,81 dias), não representando uma diferença muito significativa para os restantes estudos.

Tabela 15: Análise do atributo Serviço de Internamento

Estudo	Importância em %	Caraterística	Nº Dias
2013	**12,3 %**	Serviço Medicina	3,44
		Serviço Ortopedia	3,14
		Serviços Cirurgia e Pneumologia	2,98
2015	**12,3 %**	Serviço Medicina	3,3
		Serviço Ortopedia	3
		Serviços Cirurgia e Especialidades	2,9
2018	8 %	Dep. Médico - Ala B	2,86
		Dep. Cirúrgico - Ala B	2,82
		Dep. Médico - Ala A	2,74
		Dep. Cirúrgico - Ala A	**2,6**

Nota. 2013 e 2015 – Serviços de internamento do HFA.

Relativamente ao atributo Serviço de Internamento apresentado na Tabela 15, observou-se uma diminuição generalizada do LOS no estudo de 2018, comparativamente com os restantes estudos, confirmando totalmente a H2 apresentada.

Tabela 16: Análise do atributo Especialidade Médica

Estudo	Importância em %	Caraterística	Nº Dias
2013	**10,9 %**	Medicina Interna	4,42
		Ortopedia	3,39
		Cirurgia Geral	3,14
		Urologia	3,10
2015	10,1 %	Medicina Interna	4,2
		Ortopedia	3,5
		Cirurgia Geral	3,1
		Urologia	3,1
2018	7,5 %	Medicina Interna	4
		Ortopedia	2,97
		Cirurgia Plástica e Reconstrutiva	2,94
		Urologia	**2,56**

Nota. Especialidades médicas organizadas por ordem decrescente de LOS.

Da análise do terceiro atributo mais relevante representado na Tabela 16, verificou-se que as especialidades médicas do último estudo apresentam valores significativamente inferiores às suas homólogas dos estudos anteriores, concluindo-se que os valores médios do LOS associados aos atributos Serviço de Internamento e Especialidade Médica diminuíram relativamente aos estudos de 2013 e 2015.

Por fim, de modo a confrontar o valor médio de 3,9 dias de internamento obtido nesta investigação, no contexto internacional recorreu-se a dados obtidos no relatório estatístico do Eurostat (2018), confirmando-se que em Portugal (2015) o LOS médio situou-se em 7,9 dias,

valor este significativamente mais alto que o extraído neste estudo (+ 4 dias de internamento). Nos estudos referentes a 2016, a Holanda (4,5 dias) e a Turquia (4,2 dias) apresentaram os valores mais baixos, mas ligeiramente superiores ao observado nesta investigação para o HFAR.

Conclusão

Com o crescente aumento de dados nos sistemas de informação clínica, tornou-se necessária a exploração de várias tecnologias e metodologias para análise desse valioso conhecimento. As técnicas de DM têm vindo a ser utilizadas com sucesso em diversas áreas de negócio, no entanto a sua utilização no sector da saúde para previsão de LOS em hospitais, poderá tornar-se um dos maiores desafios para os gestores hospitalares.

A presente investigação teve por OG obter um modelo preditivo otimizado de LOS de pacientes no HFAR, através da descoberta de comportamentos e padrões existentes no processo de internamento hospitalar, com base em técnicas de DM.

Metodologicamente, esta investigação foi desenvolvida de acordo com o raciocínio hipotético-dedutivo, assente numa estratégia de investigação quantitativa, num horizonte temporal longitudinal e num desenho de pesquisa do tipo estudo de caso e comparativo. O percurso metodológico seguido nesta investigação passou por três fases: exploratória, analítica e conclusiva (Santos & Lima, 2016).

No decorrer da fase exploratória e de modo a compreender a temática, foram realizadas diversas atividades, das quais se destacam a recolha de informação bibliográfica relevante e a entrevista exploratória à BGEN/MED Regina Ramos – Diretora do HFAR. Os internamentos realizados no HFAR compreendidos entre 2013 e 2017 foram a população alvo desta

investigação e, atendendo aos objetivos e ao problema, a metodologia que se revelou mais adequada foi a metodologia CRISP-DM.

Na fase analítica, a informação foi analisada e tratada nas quatro das seis fases relativas a esta metodologia: compreensão do negócio, compreensão dos dados, preparação dos dados e modelação.

Por fim, na fase conclusiva, através das restantes fases da metodologia CRISP-DM (avaliação e implementação), foi efetuada a avaliação e discussão dos resultados obtidos nas fases anteriores. Far-se-á ainda a apresentação das conclusões finais, validação das H, resposta às PI, contributos para o conhecimento, as limitações que condicionaram o desenrolar da investigação e as recomendações de possíveis abordagens em futuros trabalhos.

Como já referido anteriormente, na fase de compreensão de negócio, em coordenação com a direção do HFAR, identificou-se que o modelo deveria prever o LOS com uma margem de erro inferior a 30%.

Na fase inicial da compreensão dos dados foram selecionados os dados respeitantes ao período de 1 de janeiro de 2013 a 31 de maio de 2017, registando 20291 episódios de internamento, tendo sido posteriormente selecionados 28 dos 44 atributos iniciais devido à relevância para o estudo em causa. Neste processo foram utilizadas ferramentas computacionais *open source*, nomeadamente a biblioteca *rminer* do ambiente R. Da análise aos dados, identificou-se para alguns dos atributos a existência de valores omissos, valores discrepantes e elevado número de valores possíveis (níveis), dificultando a utilização destes mesmos atributos pelas técnicas de DM.

Durante a preparação dos dados foram realizadas as atividades necessárias para construir o conjunto de dados final. Diversas ações foram efetuadas para limpeza dos dados, como por exemplo a exclusão de dez atributos devido ao seu elevado número de níveis, existência de

atributo redundante, elevado número de valores omissos ou baixa relevância teórica. Para resolver a questão dos valores omissos, aplicou-se a técnica *hot deck* e para facilitar o processo de aprendizagem na fase seguinte aplicou-se a técnica de transformação (*log1p(x)*) em alguns atributos. Diversas atividades foram efetuadas ao nível da construção dos dados, nomeadamente a inclusão de novos atributos derivados dos originais ou simplesmente a redução do número de níveis com a aplicação da função *delevels.* Após a realização dos procedimentos necessários para se obter uma boa modelação dos dados, avançou-se para a fase seguinte com um total de 19 atributos significativos, atributo a prever e 20287 episódios.

Sendo quantitativo o atributo a prever, efetuou-se uma abordagem regressiva na fase de modelação, aplicando-se cinco técnicas de regressão: DT, *naive,* MR, RF e SVM. Tendo como objetivo a escolha do modelo que melhores resultados obtém, tornou-se necessário a aplicação de procedimentos de validação para avaliar a capacidade de previsão dos modelos obtidos (*holdout* e o *k-fold*). Com base na função *mining* pertencente à biblioteca *open source rminer* obtiveram-se diversos modelos, possibilitando a avaliação dos seus resultados na fase seguinte.

Durante a avaliação dos resultados, pretendeu-se assegurar que os modelos concebidos correspondiam aos objetivos de negócio anteriormente definidos, fazendo-se uso de diversas métricas de regressão (R^2, MAE, RMSE), da curva REC, dos gráficos RSC, IMP e VEC para avaliação dos modelos.

Os resultados das avaliações efetuadas pela métrica R^2, identificaram o modelo RF (0,735), resultado do método de validação *5-fold*, como sendo superior ao modelo RF (0,73), resultado do método de validação *holdout* (*run* = 20). Sendo que, ambos os modelos anteriormente identificados, apresentaram uma métrica R^2 superior aos objetivos estabelecidos (0,7), pode-se afirmar que se conseguiu gerar bons modelos. Nesta fase o modelo SVM (0,695),

resultado do método de validação *5-fold*, apresentou um resultado no limite do inicialmente estabelecido.

A análise do gráfico REC confirmou o modelo RF como o melhor modelo preditivo, em comparação com o modelo SVM, apresentando sempre uma maior precisão para qualquer valor de tolerância. O modelo RF com um desvio absoluto de 0,4 (na escala de transformação logarítmica) previu acertadamente 73,6% dos exemplos do conjunto de testes e com um desvio absoluto de 0,5 obteve-se uma taxa de acertos de 78,5%.

Posteriormente, analisou-se o gráfico de dispersão RSC para prever o erro máximo nos extremos inferior e superior dos modelos RF e SVM. O modelo RF com um desvio absoluto de 0,5 obteve uma previsão acertada de 78,5% dos exemplos do conjunto de testes, significando que dá um erro máximo de 0,65 dias para o extremo inferior da escala (A = 0) e um erro máximo aproximado de 32 dias no extremo superior da escala (B = 4,4).

De modo a extrair a importância dos atributos que contribuíram para a criação do melhor modelo (RF), procedeu-se à análise de sensibilidade através do gráfico IMP e da função *fit* e *importance* da biblioteca *rminer*. O atributo Tipo de Episódio de Internamento apresentou uma importância de 31,9% na definição do modelo gerado, seguindo-se os seguintes atributos: Serviço de Internamento com 8%, Especialidade Médica com 7,5% e por último o Destino da Alta com 7,2%. Por outro lado, verificou-se ainda que estes quatro atributos mais significativos relacionados com a situação clínica dos pacientes, contribuem em mais de 50% para a capacidade explicativa do modelo gerado.

Finalizando a fase de avaliação, com o auxilio do gráfico VEC, procedeu-se a uma análise mais detalhada da influência dos valores de entrada mais importante na construção do modelo RF, aplicando-se a função e^{y}- 1, função inversa de $\ln(x+1)$, para obter o numero de dias na escala normal para os atributos selecionados na explicação do melhor modelo: Tipo de

Episódio de Internamento, Serviço de Internamento e Especialidade Médica. Um episódio de internamento em regime de Ambulatório está relacionado com um LOS médio mais baixo (0,27 dias) quando comparado com o regime em Internamento (3,9 dias). Ao nível do serviço de internamento, observou-se um maior LOS no serviço Departamento Médico - Ala B com um valor de 2,86 dias, seguindo-se o Departamento Cirúrgico - Ala B com um valor de 2,82 dias, o Departamento Médico - Ala A com um valor de 2,74 dias, e por fim o Departamento Cirúrgico - Ala A com o valor de 2,6 dias. Por fim, com base na analise da especialidade médica, verificou-se que a Medicina Interna possui o maior LOS com 4 dias, seguida da especialidade Ortopedia com 2,97 dias, da Cirurgia Plástica e Reconstrutiva com 2,94 e por fim a Urologia com 2,56 dias.

Respeitante ao OE1, *construir um modelo para previsão do LOS*, a H1 foi confirmada, uma vez que se obteve um modelo que permite efetuar previsões com uma margem de erro inferior a 30%. Em resposta à PI1, concluiu-se que o melhor modelo de previsão de LOS de pacientes no HFAR, com base em técnicas de DM, por forma a otimizar o LOS hospitalar, é o modelo RF com um coeficiente de determinação de 0,735 e com capacidade de prever corretamente 78,5% dos casos.

Relativamente ao OE2, *avaliar os resultados obtidos com estudos idênticos anteriores*, a H2 foi confirmada, sendo que após a fusão dos hospitais militares, o HFAR otimizou o LOS hospitalar, implementando uma melhor gestão e planeamento das altas hospitalares, o que permitiu aumentar o número de camas disponíveis para novas admissões. Em resposta à PI2, concluiu-se que os processos implementados após a fusão dos hospitais militares permitiram a diminuição do valor do LOS hospitalar, conforme os resultados apresentados nesta investigação, quando comparado com os estudos anteriores.

Uma das limitações que condicionaram de alguma forma o desenrolar da investigação foi a concretização da fase de implementação do CRISP-DM, propondo-se a incorporação dos modelos obtidos no atual sistema clínico do HFAR.

O modelo de previsão obtido trouxe contributos para o conhecimento, sendo um incentivo para as instituições hospitalares apostarem numa melhoria da eficiência dos seus processos internos e na extração de informação útil para apoiar a sua tomada de decisão. Com uma estimativa precisa do tempo de internamento, o hospital pode planear uma melhor gestão das camas disponíveis, uma utilização eficiente dos recursos, proporcionando uma ocupação média mais elevada e menor desperdício de recursos hospitalares (Azari et al., 2012).

Esta investigação proporciona diversas perspetivas de trabalho futuro, sendo que, estudo semelhante poderá ser aplicado na previsão de tempos de ocupação do bloco operatório hospitalar (BLO). Na aeronáutica, o estudo de modelos para previsão de tempos de manutenção de aeronaves poderá ser deveras pertinente, permitindo otimizar os recursos operacionais, humanos e materiais existentes.

Referências

Abelha, F., Maia, P., Landeiro, N., Neves, A., & Barros, H. (2007). Determinants of Outcome in Patients Admitted to a Surgical Intensive Care Unit. *Arquivos de Medicina, 21*(5/6), 135-143.

Azari, A., Janeja, V. P., & Mohseni, A. (2012). Predicting Hospital Length of Stay (PHLOS): A Multi-Tiered Data Mining Approach. *2012 IEEE 12th International Conference on Data Mining Workshops* (pp. 17-24).

Berry, M. J., & Linoff, G. S. (2004). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management.* Hoboken: John Wiley & Sons.

Brown, M. L., & Kros, J. F. (2003). Data mining and the impact of missing data. *Industrial Management & Data Systems, 103*(8), 611-621.

Caetano, N. M. P. (2013). *Previsão de tempos de internamento de pacientes via técnicas de Data Mining*. (Dissertação de Mestrado). ISCTE – Instituto Universitário de Lisboa, Lisboa, Portugal.

Caetano, N., Cortez, P., & Laureano R. M. S. (2015). Using Data Mining for Prediction of Hospital Length of Stay: An Application of the CRISP-DM Methodology. In: Cordeiro J., Hammoudi S., Maciaszek L., Camp O., Filipe J. (Ed.), *Enterprise Information Systems. ICEIS 2014. Lecture Notes in Business Information Processing*, *227*, Springer.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., et al. (2000). *CRISP-DM 1.0 - Step-by-step data mining guide.* Consultado em 30 set. 2018. Disponível em https://www.the-modeling-agency.com/crisp-dm.pdf

Cortez, P. (2010). Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool. In: Perner P. (Ed.), *Advances in Data Mining - Applications and*

Theoretical Aspects, Proceedings of the 10th Industrial Conference on Data Mining (ICDM 2010), LNAI 6171 (pp. 572–583). Berlin: Springer.

Cortez, P., & Embrechts, M. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences, 225*, 1-17.

Costa, J. F. (2009). *Um Ambiente Gráfico para Facilitar Tarefas de Data Mining via Ferramenta R*. (Dissertação de Mestrado). Universidade do Minho - Escola de Engenharia, Guimarães.

Cruz, A. J. (2007). *Data Mining via Redes Neuronais Artificiais e Máquinas de Vectores de Suporte*. (Dissertação de Mestrado). Universidade do Minho – Escola de Engenharia, Guimarães, Portugal.

Decreto Regulamentar n.º 51/2012, de 10 de dezembro (2012). *Estabelece a estrutura orgânica e a estrutura funcional do Polo de Lisboa do Hospital das Forças Armadas, bem como os princípios de gestão que lhe são aplicáveis*. Diário da República, 1.ª Série, 238, 6926-6930. Lisboa: Ministério da Defesa Nacional.

Decreto-Lei n.º 84/2014, de 27 de maio (2014). *Cria o Hospital das Forças Armadas*. Diário da República, 1.ª Série, 101, 2960-2963. Lisboa: Ministério da Defesa Nacional.

Decreto-Lei n.º 187/2012, de 16 de agosto (2012). *Cria o Polo de Lisboa do Hospital das Forças Armadas*. Diário da República, 1.ª Série, 158, 4490-4492. Lisboa: Ministério da Defesa Nacional.

Decreto-Lei n.º 200/2006, de 25 de outubro (2006). *Estabelece o regime geral de extinção, fusão e reestruturação de serviços públicos e de racionalização de efectivos*. Diário da República, 1.ª Série, 206, 7389-7393. Lisboa: Ministério das Finanças e da Administração Pública.

Decreto-Lei n.º 234/2009, de 15 de setembro (2009). *Aprova a Lei Orgânica do Estado-Maior-General das Forças Armadas*. Diário da República, 1.ª Série, 179, 6444-6455. Lisboa: Ministério da Defesa Nacional.

Despacho n.º 2943/2014, de 31 de janeiro (2014). *Reforma do Sistema de Saúde Militar*. Diário da República, 2.ª Série, 37, 5386-5388. Lisboa: Ministério da Defesa Nacional.

Domingos, M. S. (2015). *Como transformar o Hospital das Forças Armadas num Hospital de excelência*. (Trabalho de Investigação Individual do CPOG). IESM - Instituto de Estudos Superiores Militares, Lisboa, Portugal.

Eurostat (2018). *Hospital discharges and length of stay statistics*. Consultado em 10 de jan. 2019. Disponível em https://ec.europa.eu/eurostat/statistics-explained/index.php/Hospital_discharges_and_length_of_stay_statistics

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM, 39*(11), 27-34.

Feelders, A., Daniels, H., & Holsheimer, M. (2000). Methodological and practical aspects of data mining. *Information & Management, 37*(5), 271-281.

Freitas, A., Silva-Costa, T., Lopes, F., Garcia-Lema, I., Teixeira-Pinto, A., Bradzil, P., et al. (2012). Factors influencing hospital high length of stay outliers. (B. Central, Ed.) *BMC Health Services Research, 265*(12), 1-10.

Freitas, J. A. (2006). *Uso de Técnicas de Data Mining para Análise de Bases de Dados Hospitalares com Finalidades de Gestão*. (Dissertação de doutoramento). Faculdade de Economia da Universidade do Porto, Porto, Portugal.

Han, J., & Kamber, M. (2001). *Data Mining: Concepts and Techniques.* Burlington: Morgan Kaufmann Publishers.

Han, J., & Kamber, M. (2006). *Data Mining – Concepts and Techniques* (2 ed.). Amsterdam: Elsevier.

Kalra, A. D., Fisher, R. S., & Axelrod, P. (2010). Decreased Length of Stay and Cumulative Hospitalized Days Despite Increased Patient Admissions and Readmissions in an Area of Urban Poverty. (S. o. Medicine, Ed.) *J Gen Intern Med, 25*(9), 930-935.

Lee, T.-T., Liu, C.-Y., Kuo, Y.-H., Mills, M. E., Fong, J.-G., & Hung, C. (2011). Application of data mining to the identification of critical factors in patient falls using a web-based reporting system. *International Journal of Medical Informatics, 80*(2), 141-150.

Lei Orgânica n.º 1-A/2009, de 7 de julho (2009). *Aprova a Lei Orgânica de Bases da Organização das Forças Armadas*. Diário da República, 1º Suplemento, 1.ª Série, 129, 4344-(2) a 4344-(9). Lisboa: Assembleia da República.

Marques, J. C. (2010). *Causas do Prolongamento do Internamento: O caso de um serviço de Medicina Interna*. (Dissertação de Mestrado). Universidade da Beira Interior, Covilhã, Portugal.

Merom, D., Shohat, T., Harari, O., Meir, G., & Green, M. S. (1998). Factors associated with inappropriate hospitalization days in internal medicine wards in Israel: a cross-national survey. (Oxford, Ed.) *International Journal for Quality in Health Care, 10*(2), 155-162.

Michalewicz, Z., Schmidt, M., Michalewicz, M., & Chiriac, C. (2006). *Adaptive Business Intelligence* (1 ed.). New York: Springer.

Moro, S. M. (2011). *Optimização da Gestão de Contactos via Técnicas de Business Intelligence: aplicação na banca.* (Dissertação de Mestrado). ISCTE – Instituto Universitário de Lisboa, Lisboa, Portugal.

Oliveira, A. B., Dias, O. M., Mello, M. M., Araújo, S., Dragosavac, D., Nucci, A., et al. (2010). Fatores associados à maior mortalidade e tempo de internação prolongado em uma

unidade de terapia intensiva de adultos. *Revista Brasileira de Terapia Intensiva, 22*(3), 250-256.

Pena, F. M., Soares, J. d., Peixoto, R. S., Júnior, H. R., Paiva, B. T., Moraes, F. V., et al. (2010). Análise de um modelo de risco pré-operatório específico para cirurgia valvar e a relação com o tempo de internação em unidade de terapia intensiva. *Rev. bras. ter. intensiva, 22*(4), pp. 339-345.

Piatetsky, G. (2014). *CRISP-DM, still the top methodology for analytics, data mining, or data science projects*. Consultado em 26 de set. 2018. Disponível em https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Pinto, D. S. (2009). *Business Intelligence – O Poder Do Conhecimento*. (Dissertação de Mestrado). Business School - ISCTE – Instituto Superior de Ciências do Trabalho e da Empresa, Lisboa, Portugal.

Resolução do Conselho de Ministros n.º 39/2008, de 28 de fevereiro (2008). *Aprova as orientações para a execução da reorganização da estrutura superior da defesa nacional e das Forças Armadas*. Diário da República, 1.ª Série, 42, 1328-1332. Lisboa: Presidência do Conselho de Ministros.

Rufino, G. P., Gurgel, M. G., Pontes, T. d., & Freire, E. (2012). Avaliação de fatores determinantes do tempo de internação em clínica médica. *Rev Bras Clin Med., 10*(4), 291-297.

Santos, L., & Lima, J. (Coords.). 2016. *Orientações Metodológicas para a Elaboração de Trabalhos de Investigação*. Cadernos do IESM, 8. Lisboa: Instituto de Estudos Superiores Militares.

Silva, F. J. (2010). *Aplicação de técnicas de Data Mining na avaliação da qualidade da carne de cordeiro*. (Tese de Mestrado). Universidade do Minho – Escola de Engenharia, Guimarães, Portugal.

Suthummanon, S., & Omachonu, V. K. (2004). DRG-Based Cost Minimization Models: Applications in a Hospital Environment. *Health Care Management Science* (Vol. 3, pp. 197-205). Dordrecht: Kluwer Academic Publishers.

Apêndice A - Compreensão dos Dados

Tabela 1: Atributos selecionados para a previsão do LOS

Atributo	Descrição	Nº validações (painel)	Nº seleções (estudos)	Classificação original	Tipo de tratamento
Caraterísticas utente:					
Sexo	Sexo	8	7	Nominal	Nominal
Dt_Nascimento	Data de nascimento	1	-	Data	Nominal
Idade	Idade	8	7	Numérica	Numérica
Pais	País	3	-	Nominal	Nominal
Localidade	Morada	5	-	Nominal	Nominal
Escolaridade	Escolaridade	3	1	Nominal	Nominal
Est_Civil	Estado civil	4	-	Numérica	Numérica
Dados processo internamento:					
T_Episod_Origem	Episódio de origem	5	1	Nominal	Nominal
T_Episod_Intern	Episódio de internamento	5	2	Nominal	Nominal
Dt_Ped_Intern	Data pedido de internamento	2	-	Data	Nominal
Serv_Intern	Serviço de internamento	7	1	Nominal	Nominal
Espec_Medica	Especialidade médica	8	2	Nominal	Nominal
Dest_Alta	Destino após a alta clinica	7	-	Nominal	Nominal
N_Med_Alta	Médico alta clinica	5	1	Nominal	Nominal
Tratamento	Tratamento clinico	8	-	Ordinal	Numérica
Proc_Principal	Procedimento principal	9	2	Ordinal	Numérica
Diag_Principal	Diagnóstico principal	7	6	Nominal	Nominal
Diag_Inicial	Diagnóstico inicial	7	-	Nominal	Nominal
GDH	GDH	5	-	Ordinal	Numérica
GCD	GCD	2	-	Ordinal	Numérica
Dt_Internamento	Data de internamento	4	1	Data	Nominal
Mês_Intern	Mês de internamento	2	-	Nominal	Nominal
Ano_Intern	Ano de internamento	1	2	Ordinal	Numérica
Hora_Intern	Hora de internamento	4	-	Data	Nominal
Dt_Alta_Intern	Data de alta	4	2	Data	Nominal
Hora_Alta_Intern	Hora de alta	2	-	Data	Nominal
N_Intern_Anterior	Nº de internamentos anteriores	9	2	Numérica	Numérica
Atributo a prever:					
N_Dias_Intern	Nª de dias de internamento	9	5	Numérica	Numérica

Nota. GDH - Código Grupo Diagnóstico Homogéneo, GCD - Código Grande Categoria Diagnóstico.

Tabela 2: Código R dos gráficos representados do atributo Idade_Intern

hist(d$Idade_Intern, plot = TRUE, main="", xlab="Idade", ylab="Nº de Ocorrências", breaks=100, col="gray")
boxplot(d$Idade_Intern, main="", xlab="", ylab="Idade", col="gray")

Nota. hist – Diagrama de frequência, *bloxpot* - Diagrama de extremos e quartis.

Tabela 3: Informação estatística dos atributos quantitativos

Atributo	Nº Válidos	Média	Desvio Padrão	Min	P25	Mediana	P75	Máximo
Idade_Intern	20291	60,64	20,74	4,00	46,00	65,00	78,00	105,00
N_Intern_Anterior	20291	1,82	5,73	0,00	0,00	0,00	1,00	82,00
N_Dias_Intern	20291	6,19	23,36	0,00	0,00	2,00	5,00	2306,00
Ano_Intern	20291	-	-	2007	2014	2015	2016	2017

Nota. Min – Mínimo, P25 – Percentil 25, P75 – Percentil 75.

Apêndice B - Preparação dos Dados

Figura 1: Diagrama de frequência e *bloxpot* do atributo DiaSemana_Intern

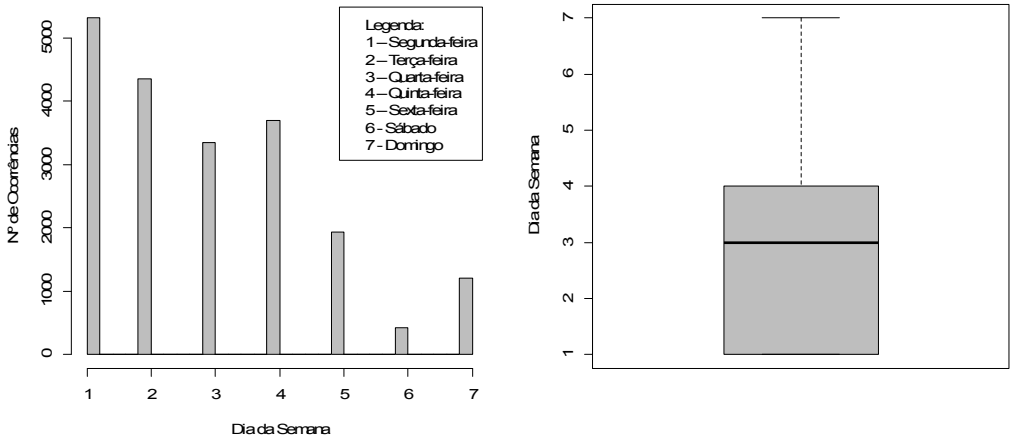

Figura 1. Na figura constata-se que o diagrama de frequência apresenta uma maior distribuição às segunda e terça-feira.

Figura 2: Diagrama de frequência e *bloxpot* do atributo Trimestre_Intern

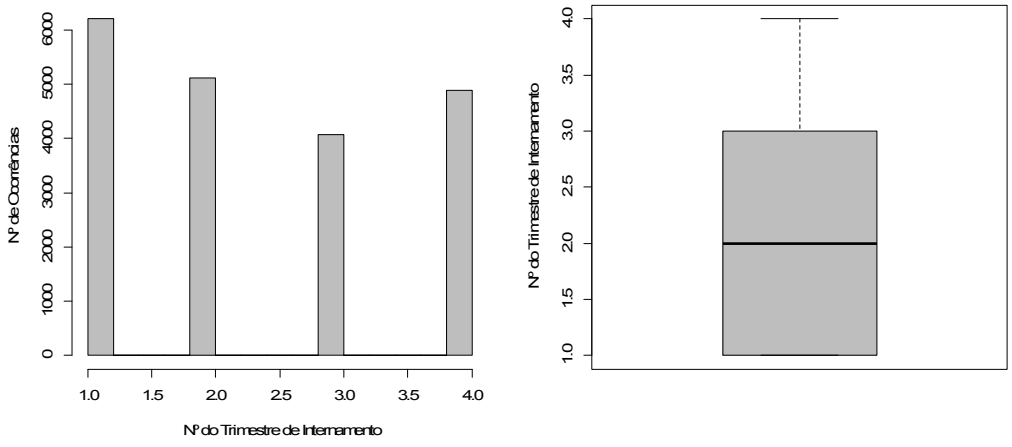

Figura 2. Na figura constata-se que o diagrama de frequência apresenta uma maior distribuição no primeiro trimestre de cada ano.

Figura 3: Diagrama de frequência e *bloxpot* do atributo transformado Hora_Intern

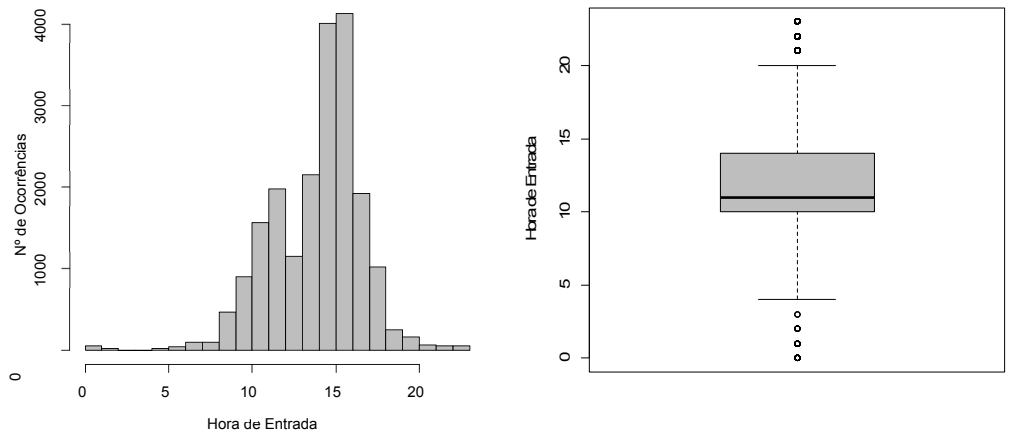

Figura 3. Na figura constata-se que o diagrama de frequência apresenta uma maior distribuição entre as 15:00 e as 16:59.

Figura 4: Diagrama de frequência e *bloxpot* do atributo transformado Hora_Alta_Intern

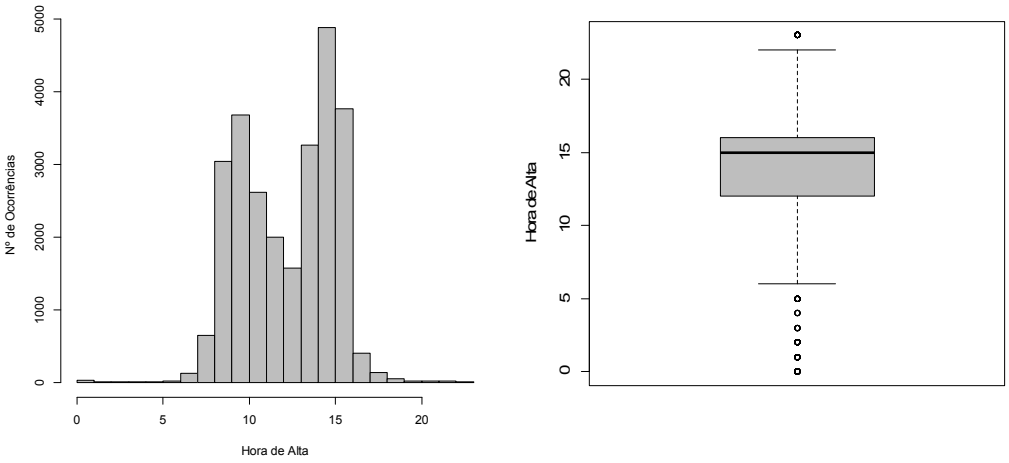

Figura 4. Na figura constata-se que o diagrama de frequência apresenta uma maior distribuição entre as 15:00 e as 15:59.

Figura 5: Diagrama de frequência e *bloxpot* do atributo transformado Escolaridade

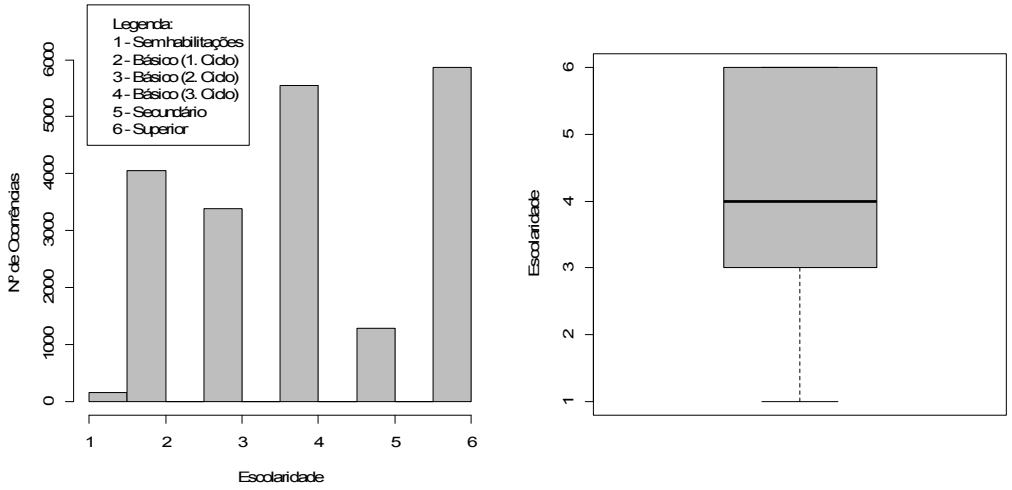

Figura 5. Na figura constata-se que o diagrama de frequência apresenta uma maior distribuição no nível de escolaridade superior.

Figura 6: Diagrama de frequência do atributo transformado Proc_Principal

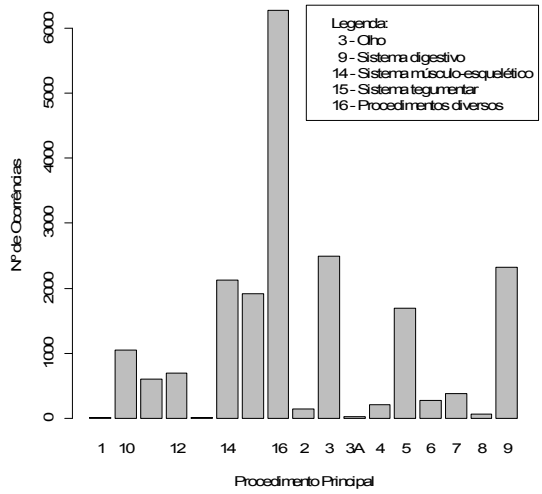

Figura 6. Na figura constata-se que o diagrama de frequência apresenta uma maior distribuição ao nível de procedimentos diversos.

Figura 7: Diagrama de frequência e *bloxpot* do atributo transformado Diag_Principal

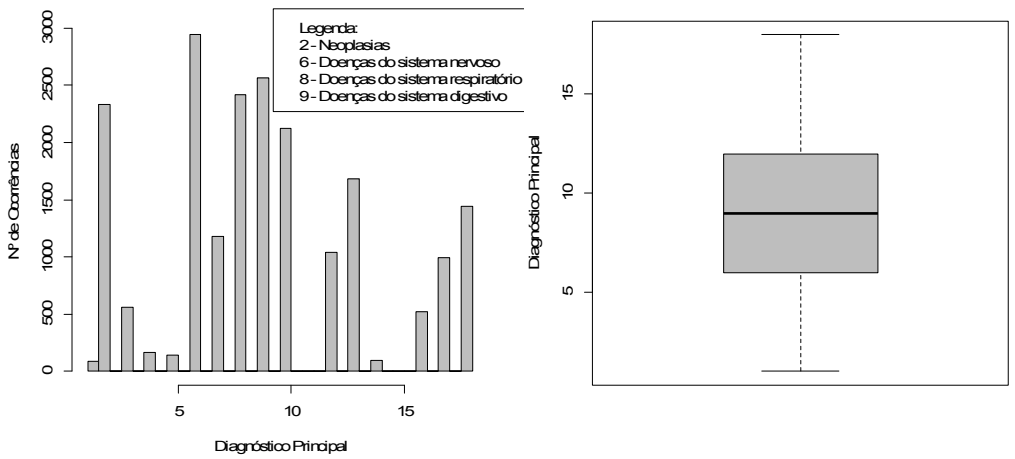

Figura 7. Na figura constata-se que o diagrama de frequência apresenta uma maior distribuição ao nível das doenças do sistema nervoso.

Figura 8: Diagrama de frequência e *bloxpot* do atributo transformado Diag_Inicial

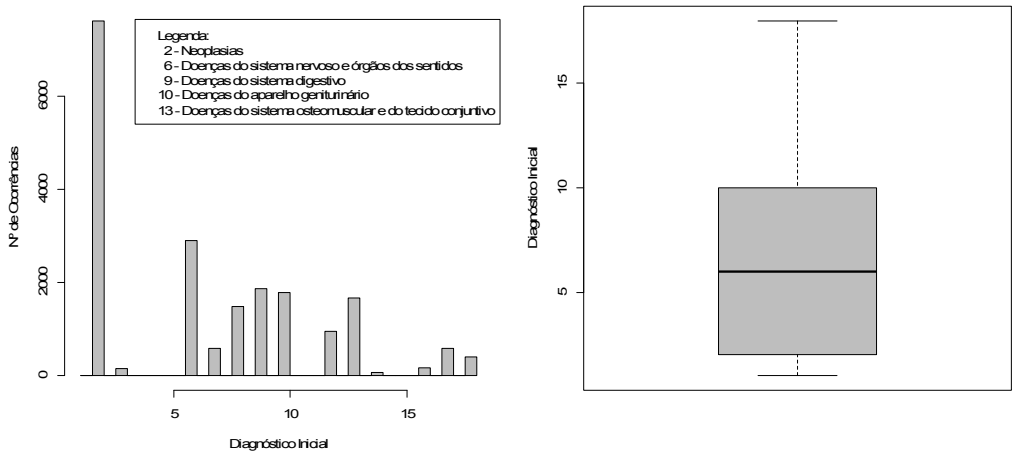

Figura 8. Na figura constata-se que o diagrama de frequência apresenta uma maior distribuição ao nível das neoplastias.

Figura 9: Diagrama de frequência e *bloxpot* do atributo transformado Idade_Intern

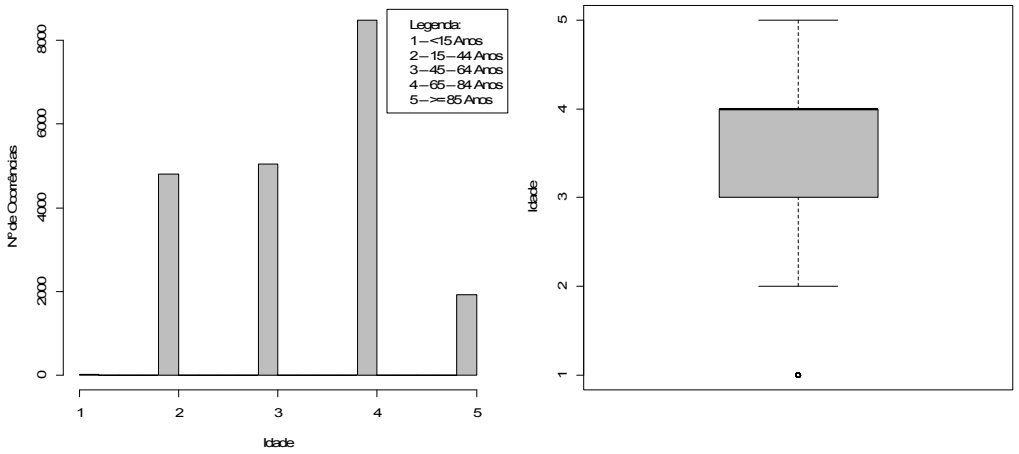

Figura 9. Na figura constata-se que o diagrama de frequência apresenta uma maior distribuição no intervalo de idades 65 a 84 anos.

Figura 10: Diagrama de frequência e *bloxpot* do atributo transformado Mes_Intern

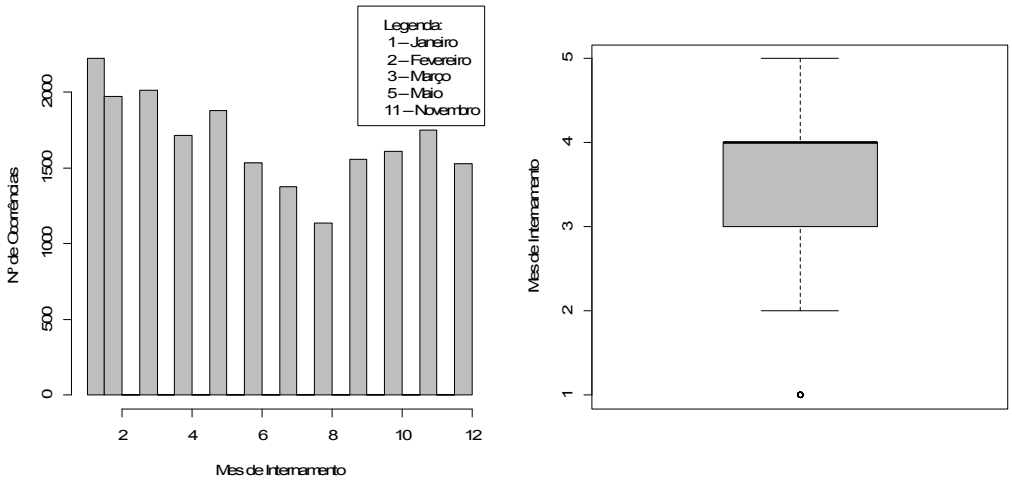

Figura 10. Na figura verifica-se que o diagrama de frequência apresenta uma maior distribuição no mês de janeiro.

Tabela 1: Informação dos atributos excluídos

Atributo	Motivo de Exclusão
Dt_Nascimento	Atributo redundante - (Idade_Intern).
País	Baixa relevância teórica - (97,23% dos casos apontam Portugal como país).
Localidade	Elevado número de níveis (3002).
Dt_Ped_Intern	Baixa relevância teórica e elevado número (34,4%) de valores omissos.
Dt_Internamento	Atributos redundantes do mês e hora – (Mes_Intern, Hora_Intern).
Ano_Intern	Baixa relevância teórica.
Dt_Alta_Intern	Atributo redundante da hora e dias de internamento - (Hora_Alta_Intern, N_Dias_Intern).
N_Med_Alta	Atributo redundante da especialidade médica – (Espec_Medica).
Tratamento	Atributo redundante do procedimento médico - (Proc_Principal).
GDH	Elevado número de níveis (539), atributo redundante do diagnóstico – (GCD).

Nota. O valor dos omissos obteve-se dividindo o valor do campo *missing* dos atributos pela dimensão da amostra (20291).

Tabela 2: Exemplo código R para tratamento dos valores omissos

Fase de pré-processamento da BD:
library(rminer)
d<-read.table("Internamentonewatribut_comNA.csv", header=TRUE,sep=";")
summary(d)
Substituição de valores omissos pelo valor encontrado:
A<-imputation(imethod = "hotdeck", d, Attribute = "Escolaridade", Missing = NA, Value = 1)
A<-imputation(imethod = "hotdeck", d, Attribute = "Est_Civil", Missing = NA, Value = 1)
A<-imputation(imethod = "hotdeck", d, Attribute = "Proc_Principal", Missing = NA, Value = 1)
A<-imputation(imethod = "hotdeck", d, Attribute = "Diag_Principal", Missing = NA, Value = 1)
A<-imputation(imethod = "hotdeck", d, Attribute = "Diag_Inicial", Missing = NA, Value = 1)
A<-imputation(imethod = "hotdeck", d, Attribute = "GCD", Missing = NA, Value = 1)
Gravar a base de dados com as alterações realizadas:
write.table(A,file="Internamentonewatribut_semNA.csv",row.names=FALSE,col.names=TRUE,sep=";")

Nota. Library - Executar a biblioteca *rminer.*

Tabela 3: Recodificação do atributo Proc_Principal

Código Antigo	Novo Cód.	Descrição	Freq.
[01–05] [010–059]	1	Operações no sistema nervoso	11
[06–07] [060–079]	2	Operações sobre o sistema endócrino	147
[08–16] [080–169] [800–1699]	3	Operações no olho	2489
17 [170–179] [1700–1799]	3A	Outros procedimentos terapêuticos e de diagnósticos	28
[18–20] [180–209] [1800–2099]	4	Operações na orelha	211
[21–29] [210–299] [2100–2999]	5	Operações no nariz, boca e faringe	1695
[30–34][300–349][3000–3499]	6	Operações sobre o sistema respiratório	275
[35–39][350–399][3500–3999]	7	Operações sobre o sistema cardiovascular	384
[40–41][400–419][4000–4199]	8	Operações no sistema sanguíneo e linfático	57
[42–54][420–549][4200–5499]	9	Operações sobre o sistema digestivo	2323
[55–59][550–599][5500–5999]	10	Operações no sistema urinário	1046
[60–64][600–649][6000–6499]	11	Operações nos órgãos genitais masculinos	602
[65–71][650–719][6500–7199]	12	Operações nos órgãos genitais femininos	692
[72–75][720–759][7200–7599]	13	Procedimentos obstétricos	8
[76–84][760–849][7600–8499]	14	Operações no sistema músculo-esquelético	2127
[85–86][850–869][8500–8699]	15	Operações no sistema tegumentar	1917
[87–99][870–999][8700–9999]	16	Procedimentos diagnósticos e terapêuticos diversos	6275

Nota. O atributo foi recodificado de 1066 para 17 níveis.

Tabela 4: Código R executado na fase preparação dos dados

Fase de pré-processamento da BD

```
library(rminer) #executar a biblioteca RMINER
d<-read.table("Internamentonewatribut_grup.csv", header=TRUE,sep=";")
summary(d$Mes_Intern)
table(d$Mes_Intern)
C=unique(d$Mes_Intern) #retorna os valores únicos
```

Exemplo de substituição do valor da classe pelo novo valor

```
g=factor(d$Mes_Intern)
G=delevels(f,c("Jan"),"1")
print(table(G))
```

Gravar a base de dados com as alterações realizadas

```
write.table(G,file="Internamentonewatribut_grup.csv",row.names=FALSE,col.names=TRUE,sep=";")
```

Diagrama de frequências do atributo transformado Idade_Intern

```
hist(d$Idade_Intern, plot = TRUE , main = "",  xlab = "Idade",  ylab = "Nº de Ocorrências", breaks=16, col = "gray")
legend(locator(1), xpd=TRUE, legend=c("Legenda:", "1 – <15 Anos", "2 – 15 – 44 Anos", "3 – 45 – 64 Anos", "4 – 65 – 84 Anos", "5 – >= 85 Anos"))
```

Diagrama de extremos e quartis do atributo transformado Idade_Intern

```
boxplot(d$Idade_Intern, main="", xlab="", ylab="Idade", col="gray")
```

Aplicação da técnica da transformação nas variáveis N_Intern_Anterior e N_Dias_Intern

```
d<-read.table("ficheiro.csv", header=TRUE,sep=";")
A=log1p(d$N_Intern_Anterior)
A=log1p(d$N_Dias_Intern)
write.table(A,file=" Internamentonewatribut_LOG.csv ",row.names=FALSE,col.names=TRUE,sep=";")
```

Nota. Exemplos de programação realizada nesta fase de preparação de dados.

Apêndice C - Avaliação

Tabela 1: Código R gerado para avaliar a qualidade dos modelos

Métricas de Regressão

```
print(mmetric(M,metric=c("R2","MAE","RMSE"),aggregate="no"))
miR=meanint(mmetric(M,metric="R2",aggregate = "no"))
cat("R2=",round(miR$mean,digits=3),"+-",round(miR$int,digits=3),"\n")
miR=meanint(mmetric(M,metric="MAE",aggregate = "no"))
cat("MAE=",round(miR$mean,digits=3),"+-",round(miR$int,digits=3),"\n")
miR=meanint(mmetric(M,metric="RMSE",aggregate = "no"))
cat("RMSE=",round(miR$mean,digits=3),"+-",round(miR$int,digits=3),"\n")
```

Nota. Exemplos de programação realizada nesta fase de modelação de dados.

Tabela 2: Código R para obtenção da curva REC e comparação dos modelos

Regression Error Characteristic curve

```
RF=loadmining("internamento_randomforest20_5.model")
SVM=loadmining("internamento_svm20_5.model")
L=vector("list",2); L[[1]]=RF;  L[[2]]=SVM;
mgraph(L,graph="REC",xval=3,leg=list(pos=c(2.25,0.4),leg=c("RF","SVM")),Grid=10)
```

Nota. Exemplos de programação realizada nesta fase de modelação de dados.

Tabela 3: Código R para obtenção do gráfico RSC

Regression Scatter Characteristic curve

```
M=loadmining("internamento_randomforest20_5.model")
for(r in 1:M$runs){
X=M$test[[r]]
Y=M$pred[[r]]
if(r==1)
        mgraph(X,Y,graph="RSC",leg=c("rf"), col="gray50",cex=0.7, Grid=20,  main="")
else points(X,Y,col="gray50", pch=19,cex=0.7)
    I=which(abs(Y-X)<=T)
if(length(I)>0){
        points(X[I],Y[I],col="green",pch=19,cex=0.7)}}
```

Nota. T = 0.5 # valor para a tolerância a admitir.

Apêndice D - Resultados

Tabela 1: Código R para obtenção do *t.test*

Student Test	
RF=loadmining("internamento_randomforest20_5.model")	
SVM=loadmining("internamento_svm20_5.model")	
mRF=mmetric(RF,metric="R2",aggregate = "no")	
mSVM=mmetric(SVM,metric="R2",aggregate = "no")	
t.test(mRF,mSVM, alternative =c("two.side"),conf.level=0.95)	
Welch Two Sample t-test	95 percent confidence interval:
data: mRF and mSVM	0.03877089
t = 133.3374, df = 32.808, p-value < 2.2e-16	0.03997266
alternative hypothesis: true difference in means is not	
equal to 0	sample estimates:
cat("p-value:",t.test(mRF,mSVM)$p.value)	mean of x: 0.7346390
p-value: 1.909882e-46	mean of y: 0.6952672

Nota. Nível de confiança *(conf.level* = 0.95*).*

Tabela 2: Código R para definir valor de tolerância de erro absoluto

Tolerância de 0,5
M=loadmining("internamento_randomforest20_5.model")
M=loadmining("internamento_svm20_5.model")
Tol=meanint(mmetric(M,metric="TOLERANCE", val=0.5,aggregate = "no"))
Nar=meanint(mmetric(M,metric="NAREC", val=0.5,aggregate = "no"))
cat("TOLERANCE=",round(Tol$mean,digits=3),"+-",round(Tol$int,digits=3),"\n")
cat("NAREC=",round(Nar$mean,digits=3),"+-",round(Nar$int,digits=3),"\n")
TOLERANCE= 0,785 ± 0,000
NAREC= 0,593 ±0,001

Nota. Exemplos de programação realizada nesta fase de avaliação de resultados.

Tabela 3: Código R para aquisição gráfico IMP

Relative Input Importance Barplot

```
library(rminer)
library(randomForest)
d<-read.table("internamentoR.csv", header=TRUE,sep=",")
M=fit(LG_N_Dias_Intern~.,data=d,model="randomforest", task="reg", search="heuristic10")
savemining(M,"internamento_randomforest_importance.model")

M=loadmining("internamento_randomforest_importance.model")
d<-read.table("internamentoR.csv", header=TRUE,sep=",") # ler a bd
Imp=Importance (M, data=d, method="DSA")
print(round(Imp$imp,digits=3))
L=list(runs=1,sen=t(Imp$imp),sresponses=Imp$sresponses)
mgraph(L,graph="IMP",leg=names(d),col="gray",Grid=10)
```

Nota. Extração da importância dos diversos atributos.

Tabela 4: Código R para aquisição dos gráficos VEC

Variable Effect Curve

```
M=loadmining("internamento_randomforest_importance.model")
d<-read.table("internamentoR.csv", header=TRUE,sep=",")
Imp=Importance (M, data=d, method="DSA")
vecplot(Imp,graph="VEC",xval=6,Grid=50,main="", xlab="", ylab="")
vecplot(Imp,graph="VEC",xval=7,Grid=50,main="", xlab="", ylab="")
vecplot(Imp,graph="VEC",xval=8,Grid=50,main="", xlab="", ylab="")
```

Responses

```
print(sresp$y[2,])
sresp=Imp$sresponses[[6]]
0.2399989 1.5926556
sresp=Imp$sresponses[[7]]
1.340239 1.352631 1.129945 1.315778 1.280244 1.087606 1.270459
sresp=Imp$sresponses[[8]]
1.253772 1.369267 1.616451 1.206656 1.375973 1.170200 1.270499
```

Nota. Código gerado para os três atributos mais significativos.