risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Revista Ibérica de Sistemas y Tecnologías de Información

Recebido / Recibido: 10/04/2014
Aceitação / Aceptación: 03/06/2014

Previsão de tempos de internamento num hospital português: aplicação da metodologia CRISP-DM

Raul M. S. Laureano [1], Nuno Caetano [2], Paulo Cortez [3]

raul.laureano@iscte.pt, nuno.caetano@defesa.pt, pcortez@dsi.uminho.pt

[1] Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (UNIDE-IUL), Escola de Gestão (IBS), Av. das Forças Armadas, 1649-026 Lisboa, Lisboa, Portugal

[2] HFAR - Hospital das Forças Armadas, Azinhaga Ulmeiros, 1620-060 Lisboa, Lisboa, Portugal

[3] ALGORITMI Research Centre, Department of Information Systems, University of Minho, Campus de Azurém, 4800-058 Guimarães, Guimarães, Portugal

DOI: 10.4304/risti.13.83-98

Resumo: Com base nos dados disponíveis num hospital português relativos aos processos de internamento, ocorridos no período de 2000 a 2013, e seguindo a metodologia de *data mining* CRISP-DM, obteve-se um modelo de previsão dos tempos de internamento baseado no algoritmo *random forest* que apresentou uma elevada qualidade, e superior à obtida com outras técnicas de *data mining*, e que permitiu identificar os atributos clínicos do paciente como os mais importantes para a explicação dos tempos de internamento.

Palavras-chave: *Data Mining*; Tempo de internamento; Modelo de previsão; Regressão; CRISP-DM.

Predicting inpatient length of stay in a Portuguese hospital using the CRISP-DM methodology

Abstract: *Using data collected from a Portuguese hospital, within the period 2000 to 2013, we adopted the CRISP-DM methodology to predict inpatient length of stay. The best method (random forest algorithm) achieved a high quality prediction. Such model allowed the identification of the most relevant input features, which are related with the patients' clinical attributes.*

Keywords*:* Data Mining; Length of Stay; Prediction model; Regression; CRISP-DM.

1. Introdução

Os hospitais têm vindo a beneficiar dos avanços da tecnologia e, em particular, dos sistemas de informação de apoio à saúde (Lee *et al.*, 2011). No entanto, dado o enorme volume e complexidade da informação armazenada, torna-se difícil a análise e

compreensão dos dados e, consequentemente, a obtenção de informação útil à tomada de decisão.

No que diz respeito à gestão hospitalar, os hospitais apresentam como objetivos reduzir o tempo de internamento, aumentar o número de camas disponíveis para novos internamentos e reduzir as listas de espera. Neste contexto, os hospitais têm necessidade de prever os tempos de permanência num serviço de internamento.

Dispondo os hospitais, desde há cerca de vinte anos, de bases de dados com informação clínica dos seus milhares de pacientes (Tsumoto & Hirano, 2010), porque não usarem de forma regular esse manancial de dados para a criação de modelos de apoio à decisão. De facto, estudos mostram que os hospitais que conseguem controlar os tempos de internamento diminuem os custos por admissão e os custos diários do doente (Suthummanon & Omachonu, 2004).

Por sua vez, o crescente aumento de dados nas bases de dados organizacionais e a necessidade de técnicas apropriadas para a sua análise facilitou o emergir de novas técnicas de exploração de dados (Ferreira *et al.*, 2006). A área do *data mining* visa a extração de conhecimento útil a partir de dados em bruto (Fayyad *et al.*, 1996). O uso de processos e técnicas de *data mining* para e desenvolvimento de modelos de apoio à decisão para a gestão e rentabilização dos serviços de internamento, suportados no sistema de informação de uma instituição hospitalar, torna-se um fator crítico de sucesso.

Assim, os objetivos do estudo são: criar um modelo preditivo dos tempos de internamento de pacientes numa instituição hospitalar (HFAR – Hospital das Forças Armadas – Polo de Lisboa); e identificar os atributos, clínicos e sociodemográficos dos pacientes, com maior influência no tempo de internamento. Um modelo de previsão de tempo de internamento (LOS, do inglês *length of stay*), que é uma medida empregue em todo o mundo para medir o consumo dos recursos hospitalares e a monitorização de desempenho (Castillo, 2012), permite evitar períodos de internamento prolongados, melhorar os serviços de saúde e gerir de forma mais eficiente os recursos hospitalares.

De facto, neste estudo foram selecionados os dados referentes a 26.462 episódios de internamento, associados à atividade dos diversos serviços de internamento e especialidades médicas, ocorridos entre outubro de 2000 e março de 2013. Foram selecionados, com a colaboração de especialistas de saúde, 29 atributos (incluído o atributo a prever) explicativos do número de dias de internamento do utente, entre outros, os relacionados com o tipo de internamento, hora de entrada do paciente, e os de caracterização sociodemográfica dos pacientes (e.g., sexo). Foram também testadas diversas técnicas de regressão: método simples da previsão baseado na média (AP), regressão múltipla (MR), árvores de decisão (DT), redes neuronais artificiais (ANN), *random forest* (RF) e máquinas de vetores de suporte (SVM). Os resultados obtidos foram interpretados tendo em consideração os impactos em objetivos médicos e de gestão.

O resto do artigo está organizado da seguinte forma: na secção 2 apresenta-se uma exposição de vários casos de estudo relacionados; na secção 3 é descrito o problema e a metodologia CRISP-DM (*Cross-Industry Standard Process for Data Mining*) utilizada, indicando-se para cada fase os procedimentos efetuados; na secção 4 é efetuada a

análise e discussão dos resultados; e na última secção são apresentadas as conclusões finais, realçando-se os contributos da investigação.

2. Trabalhos Relacionados com Previsão de Tempos

Os sistemas inteligentes na medicina têm sido aplicados em diversas vertentes. Entre outras, na fase de diagnóstico, na identificação de melhores terapias para uma patologia e na investigação de novas formas de tratamento (por exemplo, Santos e Portela, 2011). No que diz respeito ao tempo de internamento, existem diversos estudos. Em geral, os modelos consideram as características sociodemográficas e clínicas do paciente e, também, as características relacionadas com o próprio internamento como explicativas do tempo de internamento. Estes modelos são estimados, quer recorrendo a técnicas estatísticas mais tradicionais, em geral, regressão logística e regressão linear (por exemplo, Merom *et al.*, 1998; Abelha *et al.*, 2007; Liu *et al.*, 2010; Pena *et al.*, 2010; Freitas *et al.*, 2012), quer recorrendo a técnicas mais associadas ao *data mining*, em geral, redes neuronais (por exemplo, Walczak *et al.*, 1998), quer ainda a diversas técnicas e comparando-as (por exemplo, Liu *et al.*, 2006; Tanuja *et al.*, 2011; Sheikh-Nia, 2012).

Alguns autores optam por considerar o número de dias de internamento como um atributo categorial, correspondendo a internamentos de curta, média e longa duração (Azari *et al.*, 2012). Noutra perspetiva, identificam-se estudos que têm por base uma amostra de pacientes de um único hospital com uma única especialidade ou de apenas uma especialidade dum hospital, como o de Tanuja *et al.* (2011), em que são analisados 401 episódios de um hospital de geriatria. No campo oposto, também se identificam estudos em que a amostra é composta por episódios de diferentes hospitais de um mesmo país ou de diferentes países. Por exemplo, Zhan & Miller (2003), que analisam 7,45 milhões de internamentos de 994 hospitais norte-americanos no ano de 2000, ou Liu *et al.* (2010) que recolhem dados de 155.474 hospitalizações ocorridas entre 2002 e 2005 em 17 hospitais da Califórnia (Estados Unidos da América).

Em relação aos atributos considerados como potencialmente explicativos do tempo de internamento estes variam de acordo com os objetivos do estudo. Em qualquer dos casos identificam-se em muitos estudos características sociodemográficas dos pacientes, itens do seu estado fisiológico à admissão e outros atributos relacionados com o diagnóstico clínico e sua gravidade, atributos relacionados com o historial clínico do paciente e, também, com o próprio internamento e hospital.

3. Problema e Metodologia

3.1. Problema

A previsão de tempos de internamento é uma tarefa complexa, sendo que deverá ser efectuada pouco tempo após a chegada do doente ao hospital. No contexto deste estudo, o problema é modelado via uma tarefa de regressão, pretendendo-se analisar o desempenho de diversas técnicas de *data mining* para prever o tempo de internamento de um hospital português. Associado a este problema de regressão surgem três questões: preparação dos dados, modelação dos dados e avaliação dos modelos.

3.2. Metodologia CRISP-DM

Atendendo aos objetivos e ao problema a metodologia que se revela mais adequada é a metodologia CRISP-DM, que é mais usual em problemas que envolvam *data mining*, tendo uma taxa de preferência entre os profissionais de 42% (KDnuggets, 2007). A seleção desta metodologia, em detrimento de outras como o SEMMA (*Sample*, *Explore*, *Modify*, *Model*, *Assess*) ou da PMML (*Predictive Model Markup Language*), reside no facto de esta ser mais completa e iniciar-se com o estudo do negócio, ou seja, o CRISP-DM é uma metodologia que se centra nas necessidades dos gestores e na resolução dos seus problemas de gestão. Esta metodologia contempla seis fases que são flexíveis (Clifton & Thuraisingham, 2001), sendo que todas foram abordadas neste estudo exceto a última (implementação).

Compreensão do negócio

Em 2012 foi criado o Hospital das Forças Armadas (HFAR) enquanto hospital militar único, tendo como missão a prestação de cuidados de saúde aos beneficiários da ADM (Assistência à Doença dos Militares), em cooperação e articulação com o SNS (Serviço Nacional de Saúde). Uma das dificuldades com que este hospital se depara reside em garantir camas suficientes para pacientes oriundos da consulta e do serviço de urgência, devido à fusão dos hospitais militares de Lisboa.

Em relação aos recursos tecnológicos disponíveis, a instituição hospitalar dispõe de um sistema de informação para registo de informação clínica relevante, suportada numa base de dados relacional Oracle 10G, e privilegia as ferramentas *open source* (R, Rattle) para o tratamento dos dados. Por outro lado, o estudo teve um prazo de um ano para a sua concretização (final de setembro de 2013) e teve que garantir a segurança e proteção dos dados dos pacientes. Definiu-se também que a solução para prever os tempos de permanência nos serviços de internamento deveria permitir efetuar previsões com uma margem de erro inferior a 20% e teria por base as técnicas de *data mining* para problemas de regressão, tendo-se adotado a biblioteca rminer para a ferramenta R que facilita o uso de algoritmos nas tarefas de regressão (Cortez, 2010).

Compreensão dos dados

Os episódios de pedido de internamento têm a sua origem em episódios de consulta, de urgência ou de plano operatório anteriormente registado, gerando um episódio de pré-internamento. Com a entrada do paciente é gerado um episódio associado a um serviço físico de internamento, médico e valência hospitalar, sendo que o internamento em causa pode ser considerado em regime de internamento ou de ambulatório. O paciente considera-se internado até obtenção de alta médica e após saída física do serviço de internamento. O fluxo de trabalho existente no hospital encontra-se representado na Figura 1.

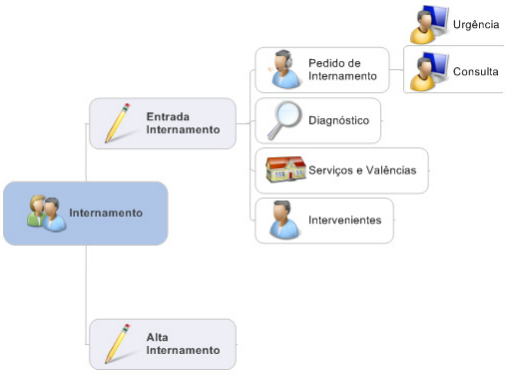

Figura 1 - Diagrama do processo de internamento hospitalar

A análise da estrutura da base de dados permitiu identificar o relacionamento entre as diversas tabelas associadas ao processo de internamento e constatar que os atributos disponibilizados já se resumiam a um registo por paciente e por número de processo de internamento, permitindo a sua transposição direta para o ficheiro de entrada às técnicas de *data mining*. A manipulação de dados efetuou-se via ferramenta SQL Navigator 6.4 e foi utilizada a ferramenta Microsoft Excel para visualização e tratamento do conjunto inicial de dados. Constatou-se que os dados respeitavam ao período de outubro de 2000 a março de 2013, contemplando 26.462 episódios de internamento associados às diversas especialidades médicas.

Atendendo aos atributos disponíveis na base de dados e aos identificados na literatura foram selecionados 28 atributos (Tabela 1), que foram confirmados e validados por um painel de nove especialistas de diversas especialidades médicas do hospital, nomeadamente, cirurgia geral (2), cirurgia plástica, gastrenterologia, ginecologia, medicina interna (2), neurocirurgia, e pneumologia.

A análise aos dados identificou que os atributos qualitativos representam a maioria dos atributos selecionados. Por outro lado, verificou-se a existência de valores omissos em alguns dos atributos (por exemplo, o Diagnóstico Principal que apresenta 19.268 valores em falta) e que alguns atributos qualitativos apresentavam um elevado número de categorias possíveis, representando uma dispersão muito elevada, podendo dificultar a utilização destes mesmos atributos pelas técnicas de *data mining* escolhidas aquando da modelação (a título de exemplo, 11.195 datas de nascimento e 2.436 localidades diferentes). Também a visualização gráfica das distribuições dos atributos permitiu identificar *outliers*, que nos casos em que correspondiam a erros foram eliminados (por exemplo, uma idade de 270 anos) e analisar o grau de assimetria das distribuições. Foram igualmente analisadas as relações entre atributos tendo-se identificado apenas uma relação muito forte entre os atributos GDH e GCD, podendo evidenciar redundância nos dados.

Tabela 1 – Atributos selecionados e sua validação por outros estudos e painel de especialistas

Atributo	Nº validações (painel)	Suporte
Sexo **(Sexo)**	8	Merom *et al.* (1998); Abelha *et al.* (2007); Kalra *et al.* (2010); Oliveira *et al.* (2010); Pena *et al.* (2010); Castillo (2012); Sheikh-Nia (2012)
Idade **(Idade_Intern)**	8	Abelha *et al.* (2007); Kalra *et al.* (2010); Oliveira *et al.* (2010); Pena *et al.* (2010); Castillo (2012); Freitas *et al.* (2012); Sheikh-Nia (2012)
Escolaridade **(Escolaridade)**	3	Castillo (2012)
Estado civil **(Est_Civil)**	4	Painel especialistas
Data de nascimento **(Dt_nascimento)**	1	Painel especialistas
País **(Pais)**	3	Painel especialistas
Localidade **(Localidade)**	5	Painel especialistas
Procedimento cirúrgico principal **(Proc_Principal)**	9	Abelha *et al.* (2007); Castillo (2012)
Tipo de episódio de origem **(T_Episod_Origem)**	5	Merom *et al.* (1998)
Diagnóstico principal **(Diag_Principal)**	7	Merom *et al.* (1998); Kalra *et al.* (2010); Oliveira *et al.* (2010); Azari *et al.* (2012);
Especialidade médica **(Espec_Medica)**	8	Castillo (2012); Sheikh-Nia (2012) Azari *et al.* (2012); Sheikh-Nia (2012)
Tipo de episódio de internamento **(T_Episod_Intern)**	5	Castillo (2012); Freitas *et al.* (2012)
Nº de internamentos anteriores **(N_Intern_Anterior)**	9	Castillo (2012) Sheikh-Nia (2012)
Tratamento **(Tratamento)**	8	Painel especialistas
Diagnóstico inicial **(Diag_Inicial)**	7	Painel especialistas
GDH **(GDH)**	5	Painel especialistas
GCD **(GCD)**	2	Painel especialistas
Ano do internamento **(Ano_Intern)**	1	Freitas *et al.* (2012); Sheikh-Nia (2012)
Data de admissão internamento **(Dt_Internamento)**	4	Kalra *et al.* (2010)
Data de saída internamento **(Dt_Alta_Intern)**	4	Kalra *et al.* (2010); Oliveira *et al.* (2010)
Nº de dias de internamento **(N_Dias_Intern)**	9	Merom *et al.* (1998); Abelha *et al.* (2007); Oliveira *et al.* (2010); Azari *et al.* (2012); Castillo (2012)
Serviço de internamento **(Serv_Intern)**	7	Castillo (2012)
Médico alta **(N_Med_Alta)**	5	Sheikh-Nia (2012)
Destino após alta **(Dest_Alta)**	7	Painel especialistas
Data pedido internamento **(Dt_Ped_Intern)**	2	Painel especialistas
Mês internamento **(Mes_Intern)**	2	Painel especialistas
Hora internamento **(Hora_Interm)**	4	Painel especialistas
Hora alta internamento (**Hora_Alta_Inter)**	2	Painel especialistas

Preparação dos dados

Nesta análise foram eliminados dados incorretos (por exemplo, um número de dias de internamento de 2.294 dias referente a um episódio de ambulatório) e registos relativos a 29 episódios de internamento associados ao código de serviço 9 (serviço virtual para testes aplicacionais). Verificou-se também a correspondência entre os

códigos dos vários atributos e seus respetivos atributos descritivos, de modo a eliminar um grande número de níveis existentes (por exemplo, na Escolaridade os códigos 10, 31, 99, 999 foram substituídos por "NA" (*Not Available*), pois não apresentavam descritivo associado, no Estado Civil o valor 9 e "A" foram substituídos por "NA", pois apresentavam o descritivo "Desconhecido"). Nesta fase foram ainda excluídos 14 atributos (Tabela 2) devido, essencialmente, à existência de redundância nos dados, de um grande número de valores omissos, de um elevado número de valores possíveis ou de falta de relevância teórica e/ou prática.

Os valores omissos foram substituídos nas 26.431 observações recorrendo à técnica *hot deck* (Brown & Kros, 2003), que consiste em procurar o exemplo mais semelhante (*1-neareast neighbor*) e substituir o valor omisso pelo valor encontrado no exemplo mais próximo. No caso do Sexo optou-se pela exclusão de 12 registos com o valor indefinido.

Tabela 2 – Atributos eliminados

Atributo	**Motivo de exclusão**
Dt_Nascimento	Existência de atributo para análise da idade (Idade_Intern).
País	99,96% dos casos correspondem a residência em Portugal, enquanto os restantes 0,02% são países PALOP e 0,02% valores omissos.
Localidade	28,6% de valores omissos e muitos valores possíveis (2.436).
Dt_Ped_Intern	Sem relevância para o estudo. Elevado número (47,9%) de valores omissos.
Dt_Internamento	Existência de atributo para análise do mês e hora (Mes_Intern e Hora_Intern).
Ano_Intern	Sem relevância para o estudo.
N_Med_Alta	Baixo número (19,1%) de valores omissos. Elevado número de níveis (156 códigos únicos de identificação de cada médico). O atributo Espec_Medica irá permitir efetuar agrupamentos por especialidades médicas associadas ao episódio de internamento.
Diag_Inicial	63% de valores omissos.
Tratamento	Atributo não conhecido no processo de admissão hospitalar do utente.
GDH	Atributo não conhecido no processo de admissão hospitalar do utente.
GCD	Atributo não conhecido no processo de admissão hospitalar do utente.
Dest_Alta	Atributo não conhecido no processo de admissão hospitalar do utente.
Hora_Alta_Intern	Atributo não conhecido no processo de admissão hospitalar do utente.
Dt_Alta_Intern	Atributo não conhecido no processo de admissão hospitalar do utente.

Foram ainda transformados alguns atributos. Para o número de internamentos anteriores e para o número de dias de internamento considerou-se o *LN (x+1),* sendo esta transformação muito comum quando se tem uma distribuição fortemente assimétrica positiva. Foram ainda criados valores transformados para o atributo Hora Internamento pois possuía 746 níveis. O seu formato foi alterado para "HH", obtendo no máximo 24 níveis possíveis. Também o atributo Escolaridade foi recodificado para corresponder às habilitações académicas usuais (Sem habilitações, Básico (1. Ciclo), Básico (2. Ciclo), Básico (3. Ciclo), Secundário e Superior) e foi criado o atributo Escalão Etário (<15 anos, 15 – 44 anos, 45 – 64 anos, 65 – 84 anos, e ≥85 anos). Os atributos Procedimento Principal e Diagnóstico Principal apresentavam demasiadas categorias que foram agrupadas para apresentarem menos níveis (em ambos os atributos 15 categorias que correspondem a grupos naturais de procedimentos ou diagnósticos). Por fim, um maior conhecimento da atividade hospitalar levou à criação de um novo atributo com relevância para o objetivo proposto, o Dia da Semana do

Internamento (Dia_Semana_Intern), que assume os valores de 1 (Segunda) a 7 (Domingo).

Modelação

Foram testadas diferentes técnicas de regressão, nomeadamente, AP, MR, DT, ANN, RF e SVM, que se encontram descritas pormenorizadamente em Hastie *et al.* (2008). Para analisar a validade do modelo a opção recaiu inicialmente no método de validação *holdout* que divide aleatoriamente os dados em dois conjuntos: conjunto de treino (para estimar os parâmetros do modelo, 2/3 da amostra), e o conjunto de teste (para avaliar a precisão do modelo, 1/3 da amostra). Complementou-se a análise com o método de validação cruzada *k-fold* com funcionamento semelhante ao anterior, mas os dados são divididos aleatoriamente 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. Definiu-
-se k=5, em que em cada rotação é treinado um modelo, sendo que a estimativa global do modelo é dada pelo erro médio do teste das k rotações. Por forma a obter-
-se maior robustez dos resultados realizaram-se 20 execuções de cada técnica nos dois métodos de validação (*holdout* e 5-*fold*) e calcularam-se os intervalos de confiança para cada métrica de qualidade, usada na fase da avaliação dos modelos.

Avaliação

Para avaliar os diversos modelos consideraram-se três métricas de regressão: coeficiente de determinação (R^2), erro médio absoluto (MAE) e raiz do erro quadrático médio (RMSE). De facto, em problemas de regressão pretende-se escolher o modelo que estima valores mais próximos dos dados, isto é, aquele que minimiza os erros (diferença entre o valor real observado e o valor previsto pelo modelo). A Tabela 3 apresenta as expressões que permitem calcular as três métricas selecionadas para os diferentes modelos de regressão e suas características.

Tabela 3 – Métricas selecionadas para avaliação dos modelos de regressão

Métrica	Expressão	Características da métrica
R^2	$R^2 = \frac{\sum_{i=1}^{n}\left(Ya_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}$	Métrica popular por não ser expressa na unidade de medida do atributo a prever. Valores próximos de 1 traduzem melhores modelos.
MAE	$MAE = \frac{\sum_{i=1}^{n}\lvert Y_i - Ya_i\rvert}{n}$	Expressa na unidade de medida do atributo a prever. Valores próximos de 0 traduzem melhores modelos.
RMSE	$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - Ya_i\right)^2}{n}}$	

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

Adicionalmente compararam-se os resultados dos diversos modelos através da curva *regression error characteristics* (REC). A curva REC mostra a taxa de acerto global no eixo das ordenadas, para diversos valores de tolerância de erro absoluto no eixo das

abcissas, sendo a precisão definida como a percentagem de pontos que se encaixam dentro da tolerância.

Desta forma, analisou-se o comportamento de cada modelo em ambiente de teste de dados e verificou-se se cumpriam os objetivos de negócio.

Implementação

Apesar de os modelos obtidos terem apresentado boa qualidade, a gestão hospitalar optou por não proceder já à implementação dos mesmos por pretender desenvolver mais investigação nesta área. Por exemplo, pretende investigar uma modelação via classes e uma modelação especializada para alguns tipos de serviços (como a ortopedia).

Alguns autores consideram que se obtêm modelos com maior capacidade preditiva se o problema de previsão for um problema de classificação. Liu *et al.* (2006) consideram internamentos com duas durações, até 10 dias e mais de 10 dias, Rufino *et al.* (2012) consideram três escalões de duração, 0 a 7 dias, 7 a 14 dias e 15 a 30 dias, e outros autores definem as categorias da duração de acordo com a especialidade do hospital (Lowell & Davis, 1994). Assim, há a necessidade de avaliar a existência de ganhos com a aplicação de modelações alternativas e comparar os resultados a obter com os agora obtidos (com um modelo global para o hospital considerando o tempo de internamento numérico).

Um outro caminho passível de ser seguido pela Direção do Hospital pode incluir o desenvolvimento de um sistema CRM (*Customer Relationship Management*) em que os modelos obtidos seriam incorporados no sistema, passando, neste caso, o utente a ser o centro dos processos do negócio.

4. Análise e Discussão dos Resultados

A análise aos dados permite identificar o perfil do paciente deste hospital. Este é do sexo masculino (57,6%), com idade acima de 50 anos (metade tem 58 ou mais anos), é casado (63,4%) e possui escolaridade ao nível do 3º ciclo do ensino básico (17,9%). Os internamentos ocorreram em maior número em janeiro (10,4%) e respeitam principalmente aos serviços físicos de cirurgia (27,5%) e de especialidades (26,7%). Em média estão internados 7,14 dias (desvio-padrão=23,8 dias) e metade dos pacientes está, no máximo, internado 3 dias.

4.1. Modelo preditivo dos tempos de internamento

A avaliação das diferentes técnicas com o método *k-fold* (k=5) leva a concluir que o melhor modelo é o *random forest* (RF) com um coeficiente de determinação (R^2) médio de 0,813. De facto, os valores apresentados na Tabela 4 evidenciam que os melhores resultados referem-se aos três últimos modelos (ANN, RF e SVM), que apresentam maiores coeficientes de determinação e menores valores nas outras duas métricas (MAE e RMSE). Tendo-se obtido um coeficiente de determinação superior a 0,8 conclui-se que se gerou um bom modelo, com uma qualidade ao nível do muito bom.

Tabela 4 – Métricas obtidas dos testes de validação *k-fold* (k=5) com 20 execuções

Parametrização	Modelo	Métricas		
		R^2	MAE	RMSE
K-fold (5), Execuções=20	AP	0,000± 0,000	0,861 ± 0,000	1,085 ± 0,000
	MR	0,641 ± 0,000	0,446 ± 0,000	0,650 ± 0,000
	DT	0,622 ± 0,001	0,415 ± 0,001	0,667 ± 0,001
K-fold (5) Procura *heuristic*, Execuções=20	ANN	0,736 ± 0,001	0,340 ± 0,001	0,558 ± 0,001
	RF	**0,813± 0,000**	**0,224±0,000**	**0,469± 0,000**
	SVM	0,745 ± 0,001	0,296 ± 0,001	0,547 ± 0,002

Relativamente à curva REC, à exceção do modelo AP, todos os modelos apresentam uma capacidade de previsão bastante boa, destacando-se o modelo RF, com uma curva bastante regular e superior à dos restantes modelos, sem propriamente pontos acentuados de mudança de comportamento (Figura 2). Por exemplo, se o valor de tolerância for de 0,5, a taxa de acerto para o modelo RF é de 0,854 (prevê-
-se acertadamente 85% dos casos). A qualidade do modelo RF é evidenciada no gráfico de dispersão *regression scatter characteristics* (RSC), que, para um dado valor da tolerância, representa no eixo das abcissas os valores observados e no eixo das ordenadas os valores previstos (Figura 3). Constata-se que a maioria dos pontos se situa próximo da diagonal, pelo que se tem um bom modelo de previsão (os pontos dentro duma tolerância de 0,5 estão representados no gráfico pela cor preta). Assim, para uma tolerância de 0,5, o erro máximo é de 0,7 dias, para o extremo inferior da escala (0), e de aproximadamente 26 dias, no extremo superior da escala (4,2).

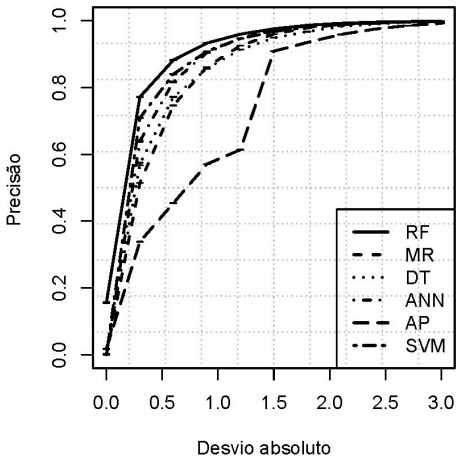

Figura 2 - Curva REC dos modelos gerados

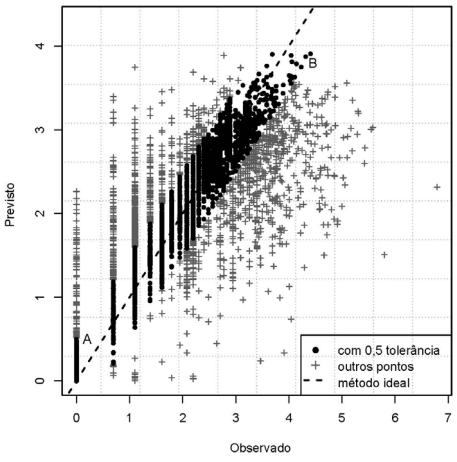

Figura 3 - RSC do modelo RF

4.2. Atributos explicativos do tempo de internamento

Para avaliar a importância relativa de cada atributo explicativo no melhor modelo obtido e também para caracterizar a sua relação com o número de dias de internamento recorreu-se a uma análise de sensibilidade (Cortez e Embrechts, 2013), em que graficamente se representam os valores observados para o atributo explicativo

(eixo das abcissas) e os valores previstos pelo modelo para o tempo de internamento (eixo das ordenadas). Os três atributos que contribuem em mais de 10% para a capacidade explicativa do modelo gerado estão relacionados com a situação clínica do paciente (Tabela 5). O Tipo de Episódio de Internamento é o atributo que mais explica (30,6%), sendo seguido pelos atributos Serviço de Internamento (12,3%) e Especialidade Médica (10,1%).

Tabela 5 – Importância relativa dos atributos explicativos no modelo RF

Atributo	Importância	Atributo	Importância
T_Episod_Intern	0,306	Hora_Intern	0,033
Serv_Intern	0,123	T_Episod_Origem	0,025
Espec_Medica	0,101	Est_Civil	0,037
Proc_Principal	0,086	LG_N_Intern_Anterior	0,061
Idade_Intern	0,031	DiaSemana_Intern	0,029
Diag_Principal	0,067	Mes_Intern	0,017
Escolaridade	0,058	Sexo	0,024

Tal como esperado um episódio de internamento em regime de ambulatório corresponde a um número médio de dias de internamento baixo (0,1 dias, correspondendo a 0,1 na escala de transformação logarítmica), quando comparado com um episódio em regime de internamento, em que o tempo médio de estadia no hospital é mais elevado (3,9 dias, correspondendo a 1,58 na escala transformada).

Relativamente ao serviço de internamento (Figura 4) verifica-se um maior número de dias de estadia no internamento do serviço de medicina com um valor estimado de 3,3 dias (1,45 na escala de transformação logarítmica), seguindo-se o serviço de ortopedia com um valor estimado de 3 dias (1,39 na escala de transformação logarítmica). Os serviços com menos dias de internamento estão relacionados com cirurgias plásticas (menos de 2 dias).

Em relação à especialidade médica (Figura 5) a medicina interna destaca-se como sendo a especialidade com bastante mais tempo de internamento (média de 4,2 dias, correspondendo a 1,64 na escala de transformação logarítmica), seguida pela especialidade de ortopedia (média de 3,5 dias, correspondendo a 1,50 na escala transformada). Com bastante menos tempo de internamento identificam-se as especialidades de oftalmologia e otorrinolaringologia (médias inferiores a 2,7 dias).

Os resultados encontrados para os atributos mais importantes na explicação do tempo de internamento confirmam os resultados encontrados em outros estudos. Por exemplo, os atributos Tipo de Episódio de Internamento (Castillo, 2012; Freitas *et al.*, 2012), Serviço de Internamento (Castillo, 2012), Especialidade Médica (Azari *et al.*, 2012; Sheikh-Nia, 2012) e Procedimento Principal (Abelha *et al.*, 2007; Castillo, 2012).

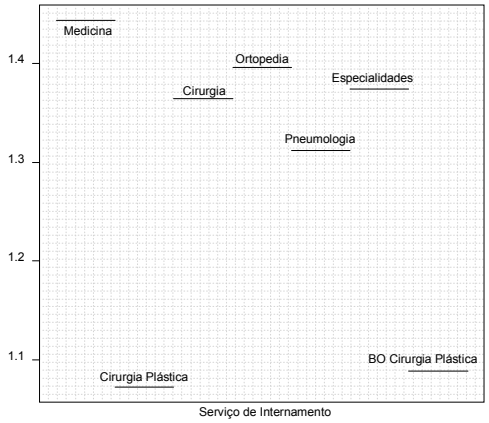

Figura 4 – Influência do serviço de internamento no tempo de internamento

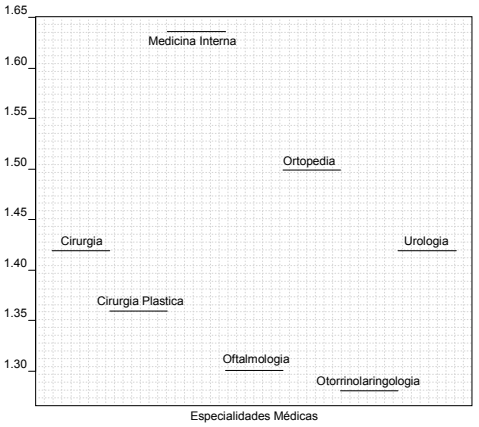

Figura 5 – Influência da especialidade médica no tempo de internamento

Por outro lado, tendo o Hospital das Forças Armadas a função de dar apoio médico aos militares no ativo, na reserva, ou na reforma, e aos seus familiares, bem como a doentes de outros subsistemas de saúde (como, por exemplo, a ADSE) com os quais o hospital tem acordos, e sendo um hospital com valências iguais ou similares às de outros hospitais (bloco operatório, urgência básica, serviços de internamento médico e cirúrgico, várias especialidades médicas, entre outras, gastrenterologia, cirurgia vascular, neurocirurgia, cirurgia plástica, oftalmologia e urologia), o perfil de doente não difere muito do de outros hospitais com características semelhantes. Este hospital apenas tem um acesso a consultas mais restrito do que em outros hospitais públicos. Assim, os resultados obtidos podem, de certa forma, ser utilizados por outros hospitais para uma melhor gestão dos seus recursos.

5. Conclusões

As técnicas de *data mining* têm vindo a ser utilizadas com sucesso em diversas áreas de negócio. A sua utilização no sector da saúde para previsão de tempos de internamento em hospitais, uma das atividades mais desafiantes para os gestores hospitalares, revelou, neste estudo, ser um sucesso. De facto, tendo por base uma grande amostra de episódios de internamentos, relativos a um hospital português, foi possível obter um modelo de apoio à decisão, baseado no algoritmo *random forest*, com boa qualidade, ou seja, que garante erros médios de previsão relativamente baixos. Este processo de descoberta de conhecimento foi guiado pela metodologia CRISP-DM e foram utilizadas ferramentas computacionais *open source*, nomeadamente a biblioteca rminer do ambiente R.

O modelo selecionado apresenta um coeficiente de determinação de 0,81, valor bastante superior aos 0,5 ou 0,6 tidos como valores mínimos aceitáveis em modelos de previsão, e que pode prever corretamente em 85% dos casos. O modelo permitiu igualmente identificar os atributos clínicos, em detrimento dos atributos sociodemográficos dos pacientes, como os mais importantes na previsão dos tempos de

internamento. Destacaram-se o Tipo de Episódio de Internamento, o Serviço de Internamento, e a Especialidade Médica.

O modelo de previsão obtido é um incentivo para as instituições hospitalares apostarem numa melhoria da eficiência dos seus processos internos e na extração de informação útil para apoiar a tomada de decisão. Só com bons modelos, que permitam reduzir os tempos de internamento (evitando, por exemplo, tempos de internamento prolongados) e melhor gerir o plano operatório do Bloco Operatório Principal e, consequentemente, aumentar as camas disponíveis para novos internamentos, se consegue reduzir as listas de espera (e melhor informar o doente quanto ao tempo de espera para uma cirurgia) e, assim, prestar melhores cuidados de saúde aos cidadãos. Paralelamente consegue-se uma gestão dos recursos mais eficaz e eficiente que se traduz numa redução dos custos hospitalares, ao permitir, por exemplo, uma melhor gestão de recursos humanos dos serviços de internamento (algumas especialidades médicas evidenciam valores superiores de ocupação nos seus serviços), quer através de um melhor planeamento das férias, quer na planificação da contratação temporária de pessoal.

Por outro lado, a metodologia CRISP-DM permitiu aos responsáveis do hospital escalonar as diversas atividades para obtenção do resultado final e obrigou-os a pensar o negócio, o que por si só é uma mais-valia da metodologia adotada. O decorrer do estudo, com o envolvimento de um painel de especialistas e de gestores do Hospital das Forças Armadas, revelou que o quadro geral da gestão hospitalar, em especial o relacionado com o processo de internamento, mostra a necessidade de um novo e atual processo de planeamento nesta área.

O modelo gerado pode vir a ser integrado num sistema de apoio à decisão e constituir-se como um elemento de precioso auxílio à área de negócio hospitalar, permitindo a otimização, gestão e rentabilização dos serviços de internamento. No futuro, outras técnicas de *data mining* poderão ser utilizadas, explorando mais atributos explicativos, nomeadamente os ligados à caracterização sociodemográfica dos pacientes, incluindo, entre outros, a classe social e o esquema de pagamento (existência de seguro de saúde ou a obtenção de crédito à saúde) e o número de elementos do seu agregado familiar.

Ao nível da investigação importa testar o modelo desenvolvido com dados de outro hospital português (por exemplo, com dados do antigo Hospital Militar Principal) para verificar se este continua válido. De referir que existe alguma evidência de que o tempo médio de internamento varia consideravelmente de país para país e até dentro do mesmo país de hospital para hospital. Por exemplo, Tiessen *et al.* (2013) comparam os tempos de internamento de dois hospitais japoneses, país com mais tempo de internamento na OCDE (Organização para a Cooperação e Desenvolvimento Económico), com os tempos de dois hospitais canadianos, país com tempo médio de internamento próximo da média da OCDE, e apontam que as diferenças se devem não a fatores clínicos dos pacientes, mas sim a fatores relacionados com normas profissionais ou culturais, esquemas de pagamento diferenciados e acesso a cuidados de saúde de longo prazo. Dentro do mesmo país a grande variabilidade nos tempos de internamento entre hospitais deve-se às características demográficas e/ou clínicas dos seus pacientes e/ou ao próprio ambiente hospitalar onde o paciente é tratado.

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, 17-24.

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

Castillo, M. G. (2012). Modelling Patient Length of Stay in Public Hospitals in Mexico. Doctoral Thesis, School of Management Southampton: University of Southampton.

Clifton, C., & Thuraisingham, B. (2001). Emerging standards for data mining. Computer Standards & Interfaces, 23(3), 187-193.

Cortez, P. (2010). Data mining with neural networks and support vector machines using the R/rminer tool. In P. Perner (Ed.), Advances in Data Mining - Applications and Theoretical Aspects, 10th Industrial Conference on Data Mining (ICDM 2010), LNAI 6171, 572–583.

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

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.

Ferreira, C., Fernandes, H., Alves, V., & Santos, M. Y. (2006). O data mining na compreensão do fenómeno da dor: uma proposta de aplicação. Conferência Ibérica de Sistemas e Tecnologias de Informação, 1-15.

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

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

Hastie, T., Tibshirani, R. & Friedman, J. (2008). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition, NY: Springer-Verlag.

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. Journal of General Internal Medicine, 25(9), 930-935

KDnuggets (2007). KDnuggets Polls: Data Mining Methodology (Aug 2007). www.kdnuggets.com [acedido em 28 de março de 2013].

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.

Liu, P., Lei, L., Yin, J., Zhang, W., Naijun, W. & El-Darzi, E. (2006). Healthcare data mining: Prediction inpatient length of stay. 3rd International IEEE Conference Intelligent Systems, 832-837.

Liu, V., Kipnis, P., Gould, M. K. & Escobar, G. (2010). Length of stay predictions: improvements through the use of automated laboratory and comorbidity variables. Medical Care, 48(8), 739-744.

Lowell, W. E. & Davis G. E. (1994). Predicting length of stay for psychiatric diagnosis-related groups using neural networks. Journal of the American Medical Informatics Association, 1(6), 459-466.

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. International Journal for Quality in Health Care, 10(2), 155-162.

Oliveira, A. B., Dias, O. M., Mello, M. M., Araújo, S., Dragosavac, D., Nucci, A. & Falcão, A. L. (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. S., Peixoto, R. S., Júnior, H. R., Paiva, B. T., Moraes, F. V., Engel, P. C., Gomes, N. C. & Pena, G. S. (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. Revista Brasileira de Terapia Intensiva, 22(4), 339-345.

Rufino, G. P., Gurgel, M. G., Pontes, T. d. & Freire, E. (2012). Avaliação de fatores determinantes do tempo de internação em clínica médica. Revista Brasileira Clínica Médica, 10(4), 291-297.

Santos, M., & Portela, F. (2011). Enabling ubiquitous data mining in intensive care: features selection and data pre-processing. ICEIS 2011 - 13th International Conference on Enterprise Information Systems, 261-266.

Sheikh-Nia, S. (2012). An Investigation of Standard and Ensemble Based Classification Techniques for the Prediction of Hospitalization Duration. Tese de Mestrado. University of Guelph, Guelph.

Suthummanon, S., & Omachonu, V. K. (2004). DRG-based cost minimization models: Applications in a hospital environment. In R. I. Field (Ed.), Health Care Regulation in America: Complexity, Confrontation, and Compromise, 3, 197-205.

Tanuja. S., Acharya, U. D. & Shailesh, K. R. (2011). Comparison of different data mining techniques to predict hospital length of stay. Journal of Pharmaceutical and Biomedical Sciences, 7(7).

Tiessen, J., Kambara, H., Sakai, T., Kato, K., Yamauchi, K. & McMillan, C. (2013). What causes international variations in length of stay: A comparative analysis for two impatient conditions in Japanese and Canadian hospitals. Health Services Management Research, 26(2-3), 86-94.

Tsumoto, S., & Hirano, S. (2010). Risk mining in medicine: Application of data mining to medical risk management. Fundamenta Informaticae, 98(1), 107-121.

Walczak, S., Scorpio, R. J. & Pofahl, W. E. (1998) Predicting hospital length of stay with neural networks. Eleventh International FLAIRS Conference, 333-337.

Zan, C. & Miller, M. R. (2003). Excess length of stay, charges, and mortality attributable to medical injuries during hospitalization. JAMA, 290(14), 1868-1874.