Departamento de Ciências e Tecnologias da Informação

# Previsão de tempos de internamento de pacientes via técnicas de *Data Mining*

Nuno Manuel Palhotas Caetano

Dissertação submetida como requisito parcial para obtenção do grau de
Mestre em Gestão de Sistemas de Informação

Orientador:
Doutor Paulo Cortez, Professor Associado com Agregação,
Universidade do Minho

Coorientador:
Doutor Raul M. S. Laureano, Professor Auxiliar,
ISCTE – Instituto Universitário de Lisboa

Setembro, 2013

# Agradecimentos

Concluído o trabalho, importa agradecer a todos que, direta ou indiretamente, permitiram a finalização deste trabalho de investigação.

Ao Professor e Orientador Paulo Cortez pela sua amizade, orientações e total disponibilidade. O Professor Paulo Cortez demonstrou um elevado carácter e profissionalismo, desde as aulas de Sistemas Inteligentes de Apoio à Decisão e principalmente neste último ano pela rápida resposta a todas as minhas solicitações.

Ao Professor Coorientador Raul Laureano pelos conselhos, ensinamentos, esclarecimentos demonstrados durante as aulas de Técnicas Quantitativas de Análise de Dados até à conclusão deste trabalho. Importa referir que ambos foram determinantes no alcançar desta meta, pois em momentos chave desta investigação deram-me total apoio, confiança e incentivo a concluir os objetivos delineados.

A todos os meus Professores do Mestrado em Gestão de Sistemas de Informação, pelos conhecimentos que me transmitiram.

À Direção hospitalar, à chefia do Serviço de Sistemas e Tecnologias de Informação (SSTI), em especial para o Eng.º. Miguel Cordeiro pela autorização da recolha de dados.

Ao meu camarada, Professor Doutor José Augusto Ribeiro da Costa, pelos seus conselhos e incentivo a prosseguir a investigação.

Por último, mas não menos importante, agradeço à minha querida esposa Ana Rita Santos e ao meu querido filho Pedro Caetano, pelo incentivo que sempre me deram durante este processo e pelo entender da minha parcial indisponibilidade em alguns momentos da nossa vida.

A todos aqueles que acreditaram no meu esforço e dedicação, o meu agradecimento por todo o vosso apoio.

# Resumo

Há mais de duas décadas que os hospitais começaram a armazenar a informação clínica electrónica nos seus sistemas de informação hospitalar. Cada vez mais, os hospitais recolhem grandes quantidades de dados através de novos métodos electrónicos de armazenamento de dados, permitindo o aumento do interesse nas áreas da descoberta de conhecimento em bases de dados e *data mining* (DM). Existe então a necessidade de investigar melhores métodos de análise de dados e automatizar esses procedimentos de modo a facilitar a criação de conhecimento.

No passado, objetivos como a necessidade de reduzir o tempo de internamento, aumentar o número de camas disponíveis para novos internamentos, reduzir o tempo de espera na lista de espera cirúrgica e prestar melhores cuidados de saúde têm sido difíceis de cumprir. O DM é então o processo chave neste trabalho através da aplicação de algoritmos de aprendizagem.

Esta dissertação irá focar-se no estudo de caso de uma instituição hospitalar nacional, com base nos dados oriundos do processo de internamento hospitalar entre 2001 e 2013. Obteve-se um modelo preditivo para tempos de internamento através da descoberta de comportamentos e padrões existentes no processo de internamento hospitalar, com base em técnicas de DM.

A concepção de um modelo explicativo permitiu extrair conhecimento útil para a área de negócio hospitalar, possibilitando no futuro, a execução de um processo de internamento mais eficiente, otimizando o número de camas existentes no contexto hospitalar e evitando erros ou desvios no planeamento dos internamentos.

**Palavras-Chave:** *Data Mining*, *Business Intelligence*, Tempos de Internamento, CRISP-DM.

# Abstract

For more than two decades that hospitals began storing information related with electronic clinical information systems. Increasingly, hospitals collect large amounts of data through new methods of electronic data storage, allowing increased interest in the areas of knowledge discovery in databases and data mining (DM). There is thus a need to investigate improved methods of data analysis and automate these procedures to facilitate the creation of knowledge.

In the past, objectives such as the need to reduce the length of stay, increase the number of beds available for new admissions, reduce the wait time on the waiting list and provide the best surgical care has been difficult to meet. DM is then the key process in this work by applying learning algorithms.

This dissertation will focus on the case study of a national hospital, based on data from the process of hospitalization between 2001 and 2013. A predictive model was obtained for the length of stay, through the discovery of behaviors and patterns existing in the hospitalization process, based on DM techniques.

The design of an explanatory model allowed extracting useful knowledge for hospital management, enabling the implementation of a rigorous admission process, optimizing the number of available hospital beds and avoiding errors or deviations in the admission plans.

**Keywords:** Data Mining, Business Intelligence, Length of Stay, CRISP-DM.

# Lista de Abreviaturas

| | |
|---|---|
| AD | Árvore de Decisão |
| ADM | Assistência na Doença aos Militares das Forças Armadas |
| ANN | *Artificial Neural Networks* |
| BI | *Business Intelligence* |
| CART | *Classification and Regression Trees* |
| CRISP-DM | *Cross-Industry Standard Process for Data Mining* |
| DCBD | Descoberta de Conhecimento em Bases de Dados |
| DM | *Data Mining* |
| DRG | *Diagnosis-Related Group* |
| DT | *Decision Trees* |
| DW | *Data Warehouses* |
| EPR | *Electronic Patient Record* |
| ETL | *Extraction, Transformation and Loading* |
| FAP | Força Aérea Portuguesa |
| FN | Falso Negativo |
| FNN | *Feed-forward Neural Network* |
| FP | Falso Positivo |
| GDH | Grupos de Diagnóstico Homogéneos |
| HFA | Hospital da Força Aérea |
| HFAR | Hospital das Forças Armadas |
| HM | Hospital da Marinha |
| HMP | Hospital Militar Principal |
| HMB | Hospital Militar de Belem |
| KDD | *Knowledge Discovery in Databases* |
| LOS | *Hospital Length Of Stay* |
| MAE | *Mean Absolute Error* |
| MLP | *Multi-layer Perceptron* |
| MR | *Multiple Regression* |
| MSE | *Mean Squared Error* |
| MVS | Máquinas de Vetores de Suporte |
| NAREC | *Normalized REC Area* |

| | |
|---|---|
| OLAP | *Online Analytical Processing* |
| PMML | *Predictive Model Markup Language* |
| RAE | *Relative Absolute Error* |
| RATTLE | *R Analytical Tool To Learn Easily* |
| RBF | *Radial Basis Function* |
| REC | *Regression Error Characteristics* |
| RF | *Random Forests* |
| RMSE | *Root Mean Squared Error* |
| RNA | Redes Neuronais Artificiais |
| RNN | *Recurrent Neural Network* |
| ROC | *Receiver Operating Characteristic* |
| RRSE | *Root Relative Squared Error* |
| RSC | *Regression Scatter Characteristics* |
| SEMMA | *Sample, Explore, Modify, Model, Assess* |
| SI | Sistemas de Informação |
| SIAD | Sistemas Inteligentes de Apoio à Decisão |
| SNS | Serviço Nacional de Saúde |
| SQL | *Structured Query Language* |
| SSE | *Sum Squared Error* |
| SSTI | Serviço de Sistemas e Tecnologias de Informação |
| SVM | *Support Vector Machines* |
| TI | Tecnologias de Informação |
| UCI | Unidade Cuidados Intensivos |
| UTI | Unidade de Terapia Intensiva |
| VEC | *Variable Effect Curve* |
| VN | Verdadeiro Negativo |
| VP | Verdadeiro Positivo |

# Índice

# Índice de Figuras

# Índice de Tabelas

# 1. Introdução

## 1.1. Enquadramento e Motivação

Tsumoto e Hirano (2010) referem que, desde há cerca de vinte anos que os hospitais começaram a armazenar a informação clínica electrónica nos seus sistemas de informação hospitalar. Lee et al. (2011) referem que o desenvolvimento mais recente da tecnologia proporcionou o rápido desenvolvimento dos sistemas de informação de apoio à saúde. Atualmente, as instituições de saúde utilizam um sistema de informação para substituir o tradicional papel, apresentando vantagens como a simplificação do processo de comunicação, diminuição da burocracia, aumento da quantidade e qualidade dos dados, melhoria do tempo de resposta e aumento da qualidade do cuidado e segurança dos pacientes. O enorme volume e complexidade da informação gerada e as limitações humanas condicionam a extração de conhecimento, tornando-se difícil a análise e compreensão dos mesmos.

O processo DM surge da necessidade de descoberta de conhecimento, ou seja, desenvolvendo métodos e técnicas de extração de conhecimento a partir de informação guardada em bases de dados (Cruz, 2010). No processo DM são aplicados métodos inteligentes com o objetivo de extrair padrões de dados. Outra definição surge de Han (2006), referindo-o como um processo de identificação de padrões novos e válidos, potencialmente úteis e fundamentalmente inteligíveis. O DM requer conhecimento dos processos por detrás dos dados, de modo a que sejam definidas perguntas úteis para análise, selecionados os dados potencialmente relevantes para resposta às perguntas e interpretados os resultados da análise (Feelders et al., 2000). Consideram ainda que, o conhecimento dos dados dentro e fora da organização também é de extrema importância para a seleção e pré processamento dos mesmos.

Os hospitais apresentam como objetivos a necessidade de reduzir o tempo de internamento, aumentar o número de camas disponíveis para novos internamentos, reduzir o tempo de espera na lista de espera cirúrgica e prestar melhores cuidados de saúde. Torna-se cada vez mais importante efetuar um processo de internamento mais eficiente, otimizando o número de camas existentes no contexto hospitalar.

O uso de técnicas, processos e desenvolvimento de modelos de DM, suportados no sistema de informação de uma instituição hospitalar, com o intuito de criar informação

inteligente sobre o negócio, servirá de apoio à decisão para a gestão e rentabilização dos serviços de internamento, tornando-se um fator crítico de sucesso.

Suthummanon e Omachonu (2004) referem que estudos que têm por base os grupos de diagnóstico homogéneo (GDH) mostram que os hospitais que conseguem controlar os tempos de internamento, diminuem os custos por admissão e os custos diários do doente. O estudo dos tempos de internamento é importante para a gestão hospitalar dada a sua clara relação com os custos hospitalares. Os episódios com tempos de internamento prolongados são responsáveis por uma fatia importante no total de dias de internamentos, apesar que nos últimos anos têm-se verificado que os episódios de longa duração têm diminuído (Freitas, 2006). Este resultado pode indicar uma redução dos custos hospitalares, possivelmente associado a uma melhor gestão hospitalar.

O fato de prestar serviço no SSTI do hospital permitiu obter um conhecimento geral dos processos existentes no contexto hospitalar. Sendo este um caso assente na utilização de dados de uma instituição hospitalar, a questão abordada poderá interessar a responsáveis e gestores de outras unidades hospitalares, pois o estudo poderá abrir novas perspetivas a organizações na área da saúde para a otimização dos seus recursos. Por outro lado, será interessante para académicos com interesses na área do *business intelligence* (BI), ou mesmo conduzir a novas investigações nesta área contribuindo para um incremento do conhecimento.

## 1.2. Objetivos

A importância de melhorar a eficácia (aumento dos resultados) e eficiência (otimização da gestão dos meios) da gestão hospitalar levou à definição dos objetivos deste trabalho. Assim, o objetivo geral é criar um modelo preditivo para tempos de internamento de pacientes numa instituição hospitalar, com base em técnicas de DM. Para o alcance do objetivo geral, é necessário a realização dos seguintes objetivos específicos:

- Prever o tempo de permanência nos serviços de internamento;
- Identificar as melhores técnicas de previsão para obtenção do melhor modelo;
- Identificar os atributos com maior influência na previsão de tempos de internamento.

A previsão de tempos de permanência num serviço de internamento é importante no contexto da disponibilidade de camas livres e de recursos humanos nos serviços de

internamento. Com a previsão do número de dias de internamento pretende-se resolver o problema da lista de espera e reduzir o custo associado ao processo de internamento dos pacientes.

## 1.3. Abordagem Metodológica

A pesquisa efetuada envolveu levantamento bibliográfico e posterior análise e teve como início a escolha da problemática a abordar.

Tomada a decisão, já previamente justificada, de abordar a questão do problema do agendamento da admissão hospitalar, tornou-se necessário o acesso aos dados residentes num sistema de informação hospitalar. Neste trabalho, explora-se o processo de internamento, pertencente a uma instituição hospitalar, que permitiu a extração da sua informação, mantendo a necessária confidencialidade dos dados.

Para este estudo foram selecionados os dados ocorridos entre Outubro de 2000 e Março de 2013 e durante este período foram efetuados cerca de 26462 episódios de internamento, associados à atividade dos diversos serviços de internamento e especialidades médicas. Foram incluídos, como por exemplo os atributos relacionados com o tipo de internamento, hora de entrada e de alta dos pacientes, apresentando como variável dependente o atributo número de dias de internamento do utente (N_Dias_Intern). Neste trabalho foram ainda testadas cinco técnicas de regressão: método simples da previsão via a média dos valores da saída, regressão múltipla, árvores de decisão (AD), rede neuronal artificial (RNA), *random forest* (RF) e máquinas de vetores de suporte (MVS).

De agora adiante a média dos valores da saída será substituída por *Naive*, a regressão múltipla será substituída por *multiple regression* (MR), a árvore de decisão será substituída por *decision tree* (DT), a rede neuronal artificial será substituída por *artificial neural network* (ANN) e as máquinas de vetores de suporte por *support vector machines* (SVM).

Os resultados obtidos deverão ser interpretados e avaliado o seu impacto em objetivos médicos e de gestão.

## 1.4. Organização

Esta dissertação segue uma estrutura de investigação tradicional encontrando-se organizada noutros capítulos para além desta Introdução que apresenta o tema e os objetivos.

No Capítulo 2, é apresentada uma revisão geral do tema *Business Intelligence* e *Data Mining*. Neste capítulo serão abordados os conceitos de BI, as etapas do processo de descoberta de conhecimento em bases de dados, as fases, os modelos e técnicas associadas ao processo DM, e por fim descritos os principais padrões de DM: PMML (*Predictive Model Markup Language*), SEMMA (*Sample, Explore, Modify, Model, Assess),* e o CRISP-DM (*Cross-Industry Standard Process for Data Mining).*

Uma exposição de vários casos de estudo relacionados com a previsão de tempos de internamento hospitalar, onde foram aplicados métodos de BI na área da saúde, será apresentada no Capítulo 3.

No Capítulo 4, é descrito o problema e a metodologia utilizada para o abordar, concretamente com a aplicação prática da metodologia CRISP-DM, indicando-se para cada fase os procedimentos efetuados.

No Capítulo 5 é efetuada a análise e discussão dos resultados e as conclusões finais serão abordadas no Capítulo 6, realçando-se os contributos desta investigação.

De referir ainda, que alguns tópicos considerados exaustivos e importantes para contextualizar a informação ao longo do trabalho, estão apresentados em Anexos próprios.

# 2. *Business Intelligence* e *Data Mining*

## 2.1. *Business Intelligence*

Pinto (2009) indica que em 1958 Hans Peter Luhn, alemão, cientista informático da IBM, conceptualizou um sistema de BI e que o termo BI foi criado pelo Gartner Group (Consultora na área de Tecnologia de Informação) em 1980.

Associado ao BI, surge a necessidade de descoberta de conhecimento em bases de dados (DCBD), devido à massificação de informação disponível nas organizações e a necessidade de sua organização. De acordo com Fayyad et al. (1996a), as bases de dados eram a infraestrutura necessária para armazenar, aceder e manipular os dados e o *data warehouse* (DW) permitia a aquisição e limpeza dos dados transacionais, tornando-os disponíveis para análise e apoio à decisão. A abordagem mais popular para análise de DW é chamada de *online analytical processing* (OLAP), ou seja, análise multidimensional dos dados. Os processos de *extraction, transformation and loading* (ETL) são utilizados para a seleção, transformação, limpeza, e carregamento dos dados para o DW.

Assim sendo, o conceito de BI surgiu devido à necessidade de sistemas que suportassem a análise de grandes quantidades de dados e integração de dados provenientes de diversos sistemas operacionais. 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 processam seu conhecimento ou como criam novos conhecimentos (Pinto, 2009). O uso do conhecimento é um fator crítico de sucesso para qualquer organização, obrigando ao investimento em mais meios e tornando-as mais eficientes no processo de produzir e disseminar conhecimento (Silva, 2010). Os sistemas de BI apresentam como objetivo auxiliar os decisores na tomada de decisão, transformam os dados existentes em informação útil (Gonçalves et al., 2010) recorrendo a ferramentas de DM. Permitem a exploração e a análise de grandes quantidades de dados com o objetivo de identificar padrões e tendências nos dados (Berry e Linoff, 2004).

Pode-se concluir que o grande benefício de uma plataforma de BI na organização é a transformação da informação em conhecimento. De facto permite a redução de custos, acesso à informação em tempo útil, maior facilidade de análise de dados, eficiência na gestão de recursos e otimização dos investimentos em sistemas de informação (Pinto, 2009).

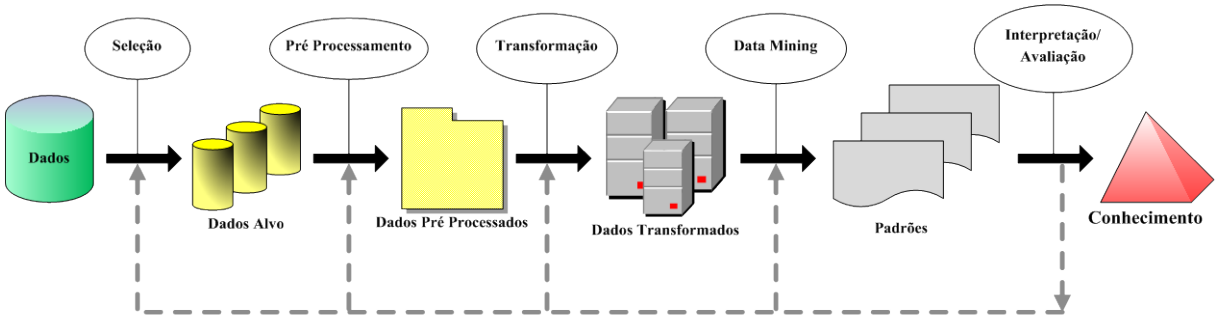
## 2.2. Descoberta de Conhecimento em Base de Dados

A DCBD, ou segundo o termo em inglês *knowledge discovery in databases* (KDD) foi utilizado no primeiro *workshop* KDD em 1989. É um processo não trivial de identificação de novos padrões, extraindo-se conhecimento de alto nível a partir dos dados de baixo nível, no contexto de grandes conjuntos de dados. O KDD inclui todo o processo de descoberta de conhecimento e o DM é somente uma das etapas deste processo (Fayyad et al., 1996b).

Fayyad et al. (1996a) definem o processo de KDD como sendo interativo e iterativo. As etapas são interativas porque o conhecimento sobre o domínio do responsável pela análise dos dados orientará a execução do processo, e iterativa pois não se trata de uma execução sequencial. De modo a refinar os resultados obtidos, são efetuadas sucessivas seleções de parâmetros e aplicadas técnicas de DM (Corrêa e Sferra, 2003). Conforme ilustrado na Figura 1, Fayyad et al. (1996a) dividiu o processo de KDD nas seguintes etapas:

- Definir o domínio da aplicação - Conhecimento prévio e identificação dos objetivos a serem atingidos;
- Criação de um conjunto de dados – Seleção de um conjunto de dados para utilização na descoberta;
- Limpeza de dados e pré-processamento – Remoção de valores discrepantes (*outliers*) e definição de estratégia para lidar com os valores omissos (*missing values*);
- Redução dos dados e projeção – Redução da dimensionalidade ou aplicação de métodos de transformação para reduzir o número de variáveis;
- Escolha da função de DM – Decidir o objetivo do modelo gerado pelo algoritmo de DM, apresentando como exemplo a classificação e a regressão;
- Escolha do algoritmo de DM – Inclui métodos de seleção a ser utilizado para procura de padrões nos dados (seleção apropriada de modelos e parâmetros);
- DM – Procura de padrões, incluindo técnicas aplicadas a problemas de classificação ou de regressão;
- Interpretação – Visualizar e interpretar e os padrões descobertos, traduzindo-os para termos perceptíveis pelo ser humano;
- Uso do conhecimento descoberto – Incorporar o conhecimento para o desempenho do sistema e tomar ações baseadas no conhecimento.

**Figura 1: Etapas do processo KDD**




Fonte: Adaptado de Fayyad et al. (1996a: 41)

Fayyad et al. (1996b) identificam diversos desafios e preocupações a ter em atenção no processo KDD, nos quais se destacam os seguintes:

- Base de dados de grande dimensão - Milhões de registos e grande número de campos (atributos e variáveis). Poderá aplicar-se métodos para reduzir a dimensão do problema e uso do conhecimento prévio para identificar as variáveis irrelevantes;
- Conhecimento prévio - Conhecimento de domínio em todas as etapas do processo, pois um analista poderá não ser um especialista em KDD;
- Sobre ajustamento - O algoritmo procura os melhores parâmetros para o modelo usando um conjunto limitado de dados;
- Alteração dos dados e do conhecimento – A alteração dos dados torna a descoberta de padrões inválida;
- Valores omissos - A falta de informação em atributos importantes pode ser problemática;
- Gestão dos dados e do conhecimento - Alteração dos dados através da modificação ou eliminação dos mesmos, poderá tornar inválidos os padrões anteriormente descobertos.

## 2.3. *Data Mining*

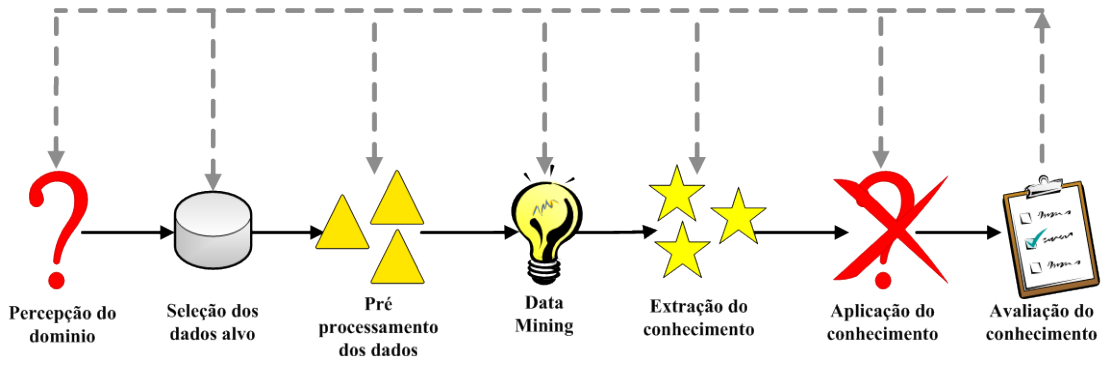
O crescente aumento de dados nas bases de dados organizacionais e a necessidade de técnicas apropriadas para a sua análise facilitou o emergir de novas técnicas de exploração de dados (Ferreira et al., 2006). Várias são as definições existentes na literatura para definir o conceito e objetivos do DM. A primeira definição de DM surge de Fayyad et al. (1996b),

definindo o DM 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. Embora formalmente Fayyad considere o DM como somente uma etapa do KDD, a verdade é se popularizou o termo DM como um sinónimo do KDD. Por exemplo, Bose e Mahapatra (2001) definem o DM como um processo complexo que envolve vários passos iterativos, visando a identificação de padrões, relacionamentos ou modelos implícitos nos dados armazenados em grandes bases de dados (Han e Kamber, 2001). Daí que nesta dissertação o termo DM será utilizado de acordo com esta visão mais ampla, denotando todo o processo de extração de conhecimento útil a partir de dados em bruto.

Deste modo torna-se importante salientar algumas das vantagens associadas ao DM: permite analisar grandes bases de dados; apresenta um método automático de descoberta de padrões nos dados; permite a criação de modelos precisos; construção e atualização dos modelos em tempo útil mesmo tratando-se de grandes quantidades de dados; produção de mais vantagens competitivas pelo fato de fazer um melhor uso dos dados. As desvantagens mais referenciadas são os custos ligados às aplicações de DM, complexidade das ferramentas, o desafio da preparação dos dados, interação muito forte com analistas humanos, uso indevido da informação e informação imprecisa, questões de privacidade e segurança (Cunha, 2009).

Após a apresentação de algumas das definições existentes na literatura, vantagens e desvantagens, é apresentado o resumo do processo e as diversas fases do DM (Figura 2).

**Figura 2: Fases do processo DM**




Fonte: Adaptado de Han e Kamber (2001: 6)

O processo DM apresenta-se como sendo um processo exploratório e iterativo. Através da análise dos dados, novo conhecimento é descoberto e novas hipóteses são formuladas. Durante todo o processo é possível saltar entre as diversas fases (Feelders et al., 2000).

Conforme diagrama demonstrado na Figura 2, o processo de DM engloba várias fases, designadamente a percepção 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 demorada de todo o processo (Freitas, 2006).

Na fase de percepção do domínio, é importante definir o problema, determinar os objetivos, identificar as pessoas chave do processo, definir critérios de sucesso do ponto de vista de DM e preparar um plano com a identificação de todos os passos críticos.

A fase de compreensão dos dados engloba o planeamento sobre o uso dos dados, a autorização para uso dos dados, a descrição da base de dados, a verificação e confidencialidade dos dados (Freitas, 2006).

A fase de preparação dos dados é a fase que consome a maior fatia de tempo do processo. É nesta da fase que são definidos os dados sobre os quais serão aplicados os métodos de DM. Deverá ser justificada a inclusão ou exclusão de dados, realizados testes de significância e correlação, remoção ou correção de valores omissos e produção de novos dados através da transformação de alguns atributos (Pyle, 1999).

A fase de aplicação de algoritmos de DM inclui a seleção de técnicas para modelação dos dados, a definição de procedimentos de treino e de teste, a construção de modelos e a sua avaliação.

Na fase de avaliação do conhecimento descoberto é necessário interpretar os resultados e avaliar o seu impacto nos objetivos inicialmente estabelecidos, sendo aprovados os melhores modelos resultantes dos métodos de DM.

Por fim, na fase de uso do conhecimento descoberto deverá existir um plano para a implementação dos resultados, com a identificação dos problemas associados à sua implementação, podendo ser produzido um relatório a sumarizar os resultados de todo o processo (Freitas, 2006).

Após uma breve apresentação das fases de DM, serão abordados os principais modelos DM. Através da análise da Figura 3, e dependendo do objetivo, a descoberta pode ser classificada em duas categorias.

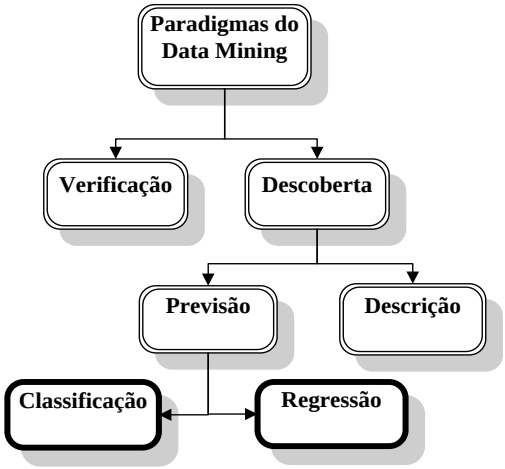
Nos métodos de aprendizagem supervisionada, ou seja, previsão e classificação, pretende-se descobrir a relação entre vários atributos de entrada ou variáveis independentes e o atributo de saída (variável dependente) (Silva, 2010).

A categoria preditiva envolve o uso de variáveis para prever valores desconhecidos ou futuros de outras variáveis, também confirmado por Bose e Mahapatra (2001). A categoria

descritiva foca-se em encontrar padrões interpretáveis que descrevam os dados e possibilitem a interpretação por seres humanos (Cunha, 2009).

Ao nível do objetivo de previsão os problemas são divididos em classificação e regressão. Se o atributo a prever for qualitativo, então trata-se de um problema de classificação. Se o atributo a prever for quantitativo, então trata-se de um problema de regressão que, segundo Moro (2011), o atributo a prever terá um valor numérico que o modelo considere o mais provável face ao que aprendeu através dos dados iniciais.

**Figura 3: Categorias dos objetivos DM**




Fonte: Adaptado de Silva (2010: 7)

Em relação à natureza da tarefa, a aprendizagem não supervisionada refere-se a técnicas que agrupam instâncias sem um atributo dependente, pré-especificado (Maimon e Rokach, 2005), confirmado por Meyfroidt et al. (2009), que indica que os modelos descritivos pertencem à aprendizagem não supervisionada, ocorrendo a modelação de uma variável alvo desconhecida. O objetivo do processo é construir um modelo que descreva regularidades interessantes nos dados, tendo-se como exemplo os problemas de *clustering* e sumarização.

Cruz (2007) indica que num problema de *clustering* pretende-se encontrar um número finito de conjuntos que descrevam os dados. Ou seja, o *clustering* não é nada mais do que o processo de agrupar os dados em classes de modo que os objetos dentro das classes tenham uma alta similaridade em comparação com um outro e, no entanto que sejam diferentes dos objetos dos outros (Han e Kamber, 2001).

Na sumarização pretende-se encontrar uma descrição compacta de um conjunto ou subconjunto de dados (Fayyad et al., 1996b), sendo que a sumarização é principalmente

utilizada no pré-processamento dos dados, quando valores inválidos são determinados por meio do cálculo de medidas estatísticas (Corrêa e Sferra, 2003).

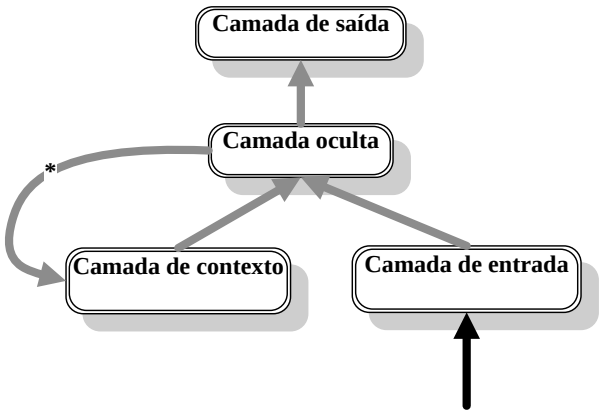
Após breve apresentação dos diversos modelos DM existentes, é importante referenciar as técnicas mais importantes de DM evidenciadas na literatura. Os algoritmos de aprendizagem têm sido aplicados em cenários de DM envolvendo bases de dados de grande dimensão, em que os algoritmos mais utilizados são as árvores de decisão, *random forests*, redes neuronais artificiais, redes bayesianas, máquinas de vetores de suporte e processos gaussianos (Meyfroidt et al., 2009).

O termo rede neuronal é tradicionalmente utilizado para referência a um circuito de neurônios biológicos, no entanto o uso moderno do termo refere-se a redes neurais artificiais (ANN), os quais são compostos de neurônios artificiais (Liao et al., 2012).

As ANN são um conjunto de unidades de processamento simples ou nós, interligados para aumentar o poder computacional de cada unidade. A ANN é robusta a erros na situação de treino dos dados, sendo adequada para a aprendizagem a partir de exemplos ruidosos. Conforme referido por Michalewicz et al. (2006), existem dois tipos diferentes de ANN:

- *Recurrent neural network* (RNN) – Este tipo de rede consiste num conjunto de nós interligados em que a atividade circula em volta da rede (Figura 4):

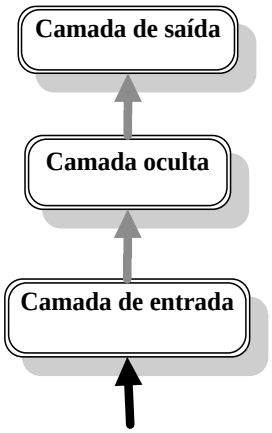
**Figura 4: Exemplo de uma RNN**




Fonte: Adaptado de Michalewicz et al. (2006: 140)

- *Feed-forward neural network* (FNN) - Este tipo de rede não tem ligações recorrentes entre os nós, e a atividade flui num sentido. A atividade é alimentada para a frente passo a passo a partir dos nós de entrada em relação aos nós de saída. (Figura 5).

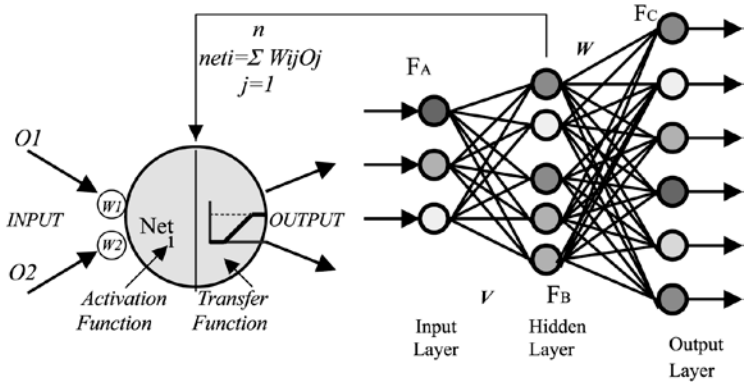
**Figura 5: Exemplo de uma FNN**




Fonte: Adaptado de Michalewicz et al. (2006: 139)

Segundo Olson et al. (2012), as duas arquiteturas ANN mais conhecidas são a *multi-layer perceptron* (MLP) e a *radial basis function* (RBF), em que tipicamente as redes são constituídas por uma camada de neurónios de entrada, uma ou mais camadas de neurónios intermédios e uma camada de saída (Figura 6).

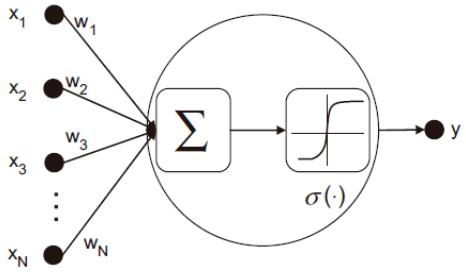
**Figura 6: Exemplo arquitetura MLP**




Fonte: Adaptado de Liu et al. (2001: 993)

A Figura 7 apresenta o esquema gráfico do funcionamento de uma ANN, em que o nó de saída irá construir a sua previsão com base nos nós de entrada com os quais está conectado. Esta decisão é alcançada multiplicando o peso da hiperligação pelo valor de saída do nó e somando estes valores para todos os nós. Se a previsão estiver incorreta, os nós com maior influência na tomada de decisão têm o seu peso alterado para que a previsão errada seja menos provável de acontecer numa próxima iteração (Oliveira, 2009).

**Figura 7: Exemplo de um único nó oculto ou neurónio numa rede**

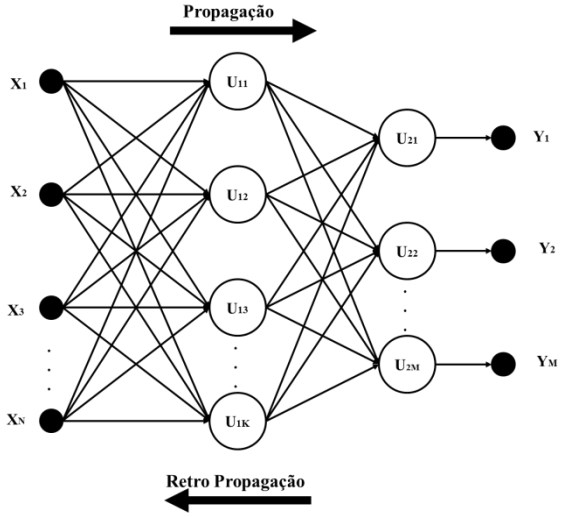



Fonte: adaptado de Meyfroidt et al. (2009: 134)

Um dos algoritmo com maior eficácia é a retro propagação (*backpropagation)* que, segundo Oliveira (2009) é constituído pelos seguintes termos, representados na Figura 8:

- Propagação (*feedfoward*) - Depois de apresentado o padrão de entrada, a resposta de uma unidade é propagada como entrada para as unidades na camada seguinte, até à camada de *output*, onde é obtido o resultado da rede e, consequentemente o erro é calculado;
- Retro propagação - Desde a camada de saída até à camada de entrada são realizadas as alterações nos pesos.

**Figura 8: Representação gráfica de uma ANN**




Fonte: Adaptado de Meyfroidt et al. (2009: 134)

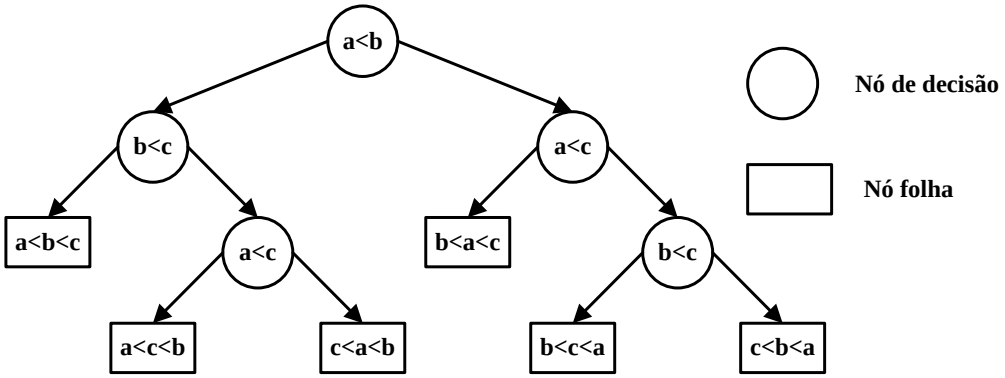
Existe por vezes a necessidade de tomada de decisões importantes no desenho do modelo para uma utilização eficaz das ANN, como por exemplo o processo de paragem para evitar o sobre ajustamento.

No aspeto do sobre ajustamento, as ANN são relativamente boas com dados de treino, mas não o são com novos dados. Para se obter os melhores resultados no processo de aprendizagem, será necessário ter em atenção o tamanho e divisão dos dados em estudo.

Os dados dividem-se em três partes, nomeadamente uma para treino que é utilizada para a atualização dos pesos das sinapses, uma para teste que serve para verificação da resposta da rede a dados não usados para treino e uma parte para validação (Cruz, 2007). Este algoritmo apresenta algumas desvantagens, como por exemplo os longos períodos de treino comparado com outros algoritmos e o desconhecimento do número de neurónios por camada necessários para aprender uma determinada função (Meyfroidt et al., 2009). A principal desvantagem é a dificuldade de interpretação e desconhecimento de como são feitas as previsões, sendo mesmo apelidadas de modelos *black box.*

O algoritmo *decision tree* (DT) têm tido muito sucesso nas áreas de DM pois funcionam bem com grandes conjuntos de dados, muitas variáveis e com diferentes tipos de dados. A DT é um dos métodos mais amplamente utilizados, pois a sua estrutura é relativamente fácil de seguir e compreender (Michalewicz et al., 2006). A Figura 9 representa uma árvore de decisão constituída pela seguinte estrutura: nó raiz (nó com o primeiro teste); nós internos (cada um possui um teste a um atributo dos dados e têm duas ou mais subárvores que correspondem às respostas possíveis); ramos (contendo valores dos atributos); folhas (representam as classes).

**Figura 9: Representação gráfica de uma DT**




Fonte: Adaptado de Preiss (1998: 251)

Para se obter uma previsão para um novo caso, a raiz da árvore é examinada, um teste é realizado e dependendo do resultado do teste, o processo move-se para baixo do ramo apropriado. O processo continua até o nó terminal ser alcançado e o valor do nodo final ser o resultado previsto (Michalewicz et al., 2006).

As DT são utilizadas para todos os tipos de problemas de previsões, no entanto são especialmente populares em problemas de classificação. Se o teste envolver uma variável qualitativa, o número de ramos corresponde ao número de valores possíveis que variável pode ter, ou seja, há um ramo para cada valor possível. Se o teste envolver uma variável quantitativa geralmente há dois ramos, em que o próprio teste determina se o valor é maior ou menor que determinado valor fixo predefinido.

Um dos algoritmos mais populares apresentados na literatura são as c*lassification and regression trees* (CART). As CART possuem a capacidade de construção de árvores binárias, em que cada nó possui duas arestas de saída. A árvore de decisão CART é um processo recursivo de particionamento binário capaz de processar atributos contínuos e nominais, em que a heurística usada para a seleção de atributos é o *gini índex*, apresentando o seguinte funcionamento segundo (Venkatadri e Reddy, 2010):

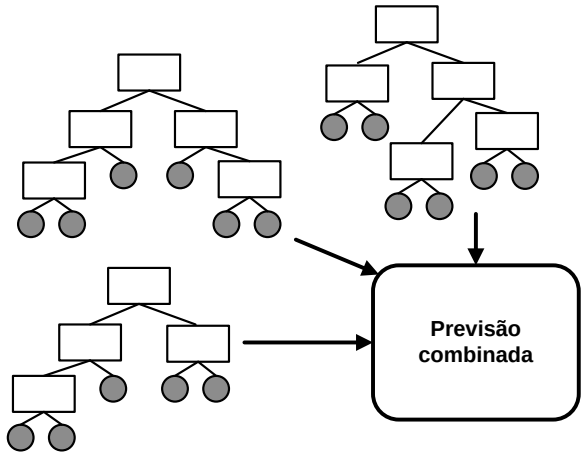
> *"In CART trees are grown, uses gini index for splitting procedure, to a maximum size without the use of a stopping rule and then pruned back (essentially split by split) to the root via cost-complexity pruning. The CART mechanism is intended to produce not one, but a sequence of nested pruned trees, all of which are candidate optimal trees. The CART mechanism includes automatic (optional) class balancing, automatic missing value handling, and allows for cost-sensitive learning, dynamic feature construction, and probability tree estimation."*

O algoritmo CART identifica as variáveis independentes recorrendo aos conceitos de entropia e ganho de informação. A entropia está relacionada com a distribuição dos valores de uma variável, em que uma entropia elevada origina uma distribuição mais uniforme e numa entropia pequena há um valor predominante. O ganho de informação indica a capacidade de uma variável em separar os casos de treino.

O primeiro passo no desenvolvimento de uma DT é o processo de crescimento da árvore. A maioria dos algoritmos de DT efetuam, ou permitem efetuar a poda das árvores de modo a evitar a situação de sobre ajustamento. As árvores crescem até ao seu tamanho máximo e são depois sujeitas à poda para evitar o sobre ajustamento, ou seja, que fique demasiado ajustada aos dados de treino.

Outro conceito existente é o de uma floresta, ou no termo inglês uma *random forest* (RF). Uma RF é composta por um conjunto de DT, construídas com base num conjunto de dados originais. Na Figura 10 verifica-se que a previsão final é obtida pela média das previsões do conjunto das árvores.

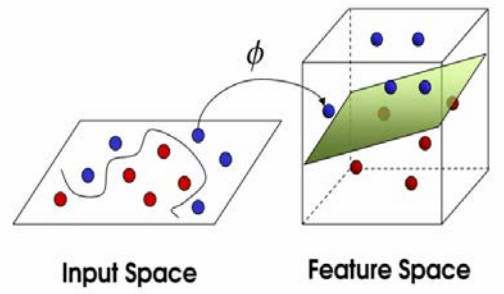
**Figura 10: Representação gráfica de uma RF**




Fonte: Adaptado de Meyfroidt et al. (2009: 133)

Cruz (2007) indica que as máquinas de suporte de vetores (SVM) surgiram pela primeira vez na década de setenta. Criadas inicialmente para problemas de classificação dos dados, mas que recentemente começam a ser aplicadas em problemas de regressão. 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 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). Conforme Figura 11, o algoritmo SVM procura encontrar o melhor híper plano de separação (Meyfroidt et al., 2009).

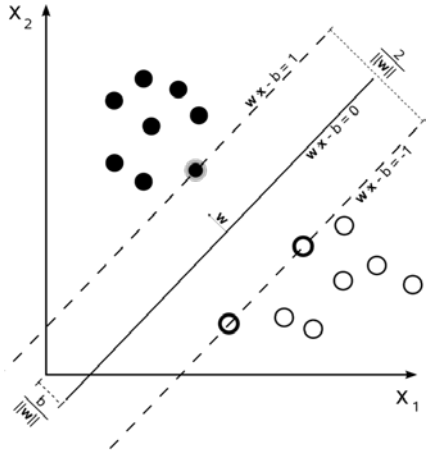
**Figura 11: Princípio de funcionamento de uma SVM**




Fonte: Adaptado de Meyfroidt et al. (2009: 136)

Para aumentar a robustez do classificador e permitir o erro de classificação, margens podem ser definidas em torno do híper plano, determinando o número de exemplos que são autorizados a atravessar o híper plano a uma certa distância (Figura 12).

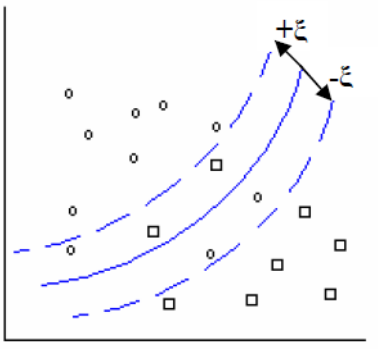
**Figura 12: Representação gráfica de margens para uma SVM**




Fonte: Wikipedia, 2013

A regressão com SVM é conseguida alterando a função do custo para que inclua um parâmetro de distância (Figura 13). Esse parâmetro (ξ) vai permitir a criação de uma margem na qual os dados serão ignorados, pelo que o parâmetro também é chamado de ξ-insensitivo. Utilizar SVM em problemas de regressão torna necessário o controlo de dois parâmetros (C e ξ) (Stitson et al., 1996).

**Figura 13: Margem criada pelo parâmetro ξ-insensitivo**




Fonte: Adaptado de Cruz (2007: 32)

Concluindo, as SVM apresentam diversas vantagens, das quais se referem as mais importantes: possuem uma interpretação geométrica simples; a complexidade computacional da SVM não depende diretamente da dimensionalidade do espaço de entrada; usam minimização do risco estrutural (menos propensas a sobre ajustamento). Apresentam como desvantagens a complexidade algorítmica e elevados requisitos de memória na execução.

Cada uma das técnicas de DM está mais orientada ao tipo de problema que se pretende resolver. A Tabela 1 apresenta algumas tipologias de modelos de DM e o fim a que se destinam.

**Tabela 1: Técnicas de DM e tipos de problemas a que se adequam**

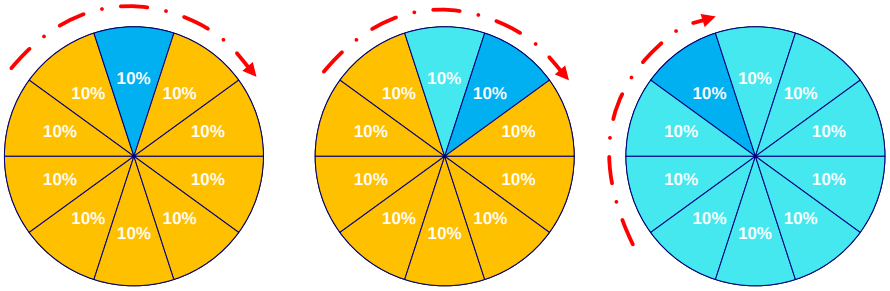
| Técnicas | Tipo de Problema | |
|---|---|---|
| | Classificação | Regressão |
| *Decision Tree* (DT) | X | X |
| *Random Forests* (RF) | X | X |
| *Artificial Neural Network* (ANN) | X | X |
| *Support Vector Machines* (SVM) | X | X |
| Regras de Classificação | X | ---------- |
| *Naive Bayes* (NB) | X | ---------- |
| *Multiple Regression* (MR) | ---------- | X |
| Regressão logística | X | ---------- |
| K-Vizinhos mais próximos (KNN) | X | X |

Fonte: Adaptado de Moro (2011: 16)

Tendo como objetivo a escolha do modelo que melhores resultados obtém, torna-se necessário a aplicação de medidas para avaliar a capacidade de previsão de um modelo obtido. Moro (2011) referencia como técnica de validação o *holdout*. A fim de testar a confiança dos modelos identificados, os dados da amostra são divididos em dois conjuntos. O conjunto de dados de treino é utilizado para construir e identificar o modelo, e o conjunto de dados de teste 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, os quais fazem parte do conjunto de teste. Também Michalewicz et al. (2006) identifica que o processo de construção de um modelo de previsão é constituído pela fase de construção e pela fase de ajustamento dos parâmetros do modelo. Se o conjunto de instâncias inicial for reduzido, o subconjunto definido para treino pode não ter uma dimensão significativa, resultando numa má definição do modelo.

Deste modo, para evitar a situação referida anteriormente, existe a possibilidade de efetuar a validação cruzada *10-fold*, representado na Figura 14. Os dados são divididos em N blocos de dimensão semelhante. A aprendizagem faz-se com recurso a N iterações, em que a cada iteração são utilizados N-1 blocos para aprendizagem e o outro para teste, sendo este diferente a cada iteração (Cruz, 2007).

**Figura 14: Representação gráfica de uma validação cruzada (*10-fold*)**




Fonte: Adaptado de Delen et al. (2012: 549)

Kohavi e Provost (1998) referem que em modelos de classificação, a matriz de confusão permite uma visualização inequívoca dos resultados de um modelo, sendo os resultados apresentados sob a forma de tabela de duas entradas: uma das entradas é constituída pelas classes desejadas, a outra pelas classes previstas pelo modelo. Na Figura 15 encontra-se representada uma matriz de confusão, sendo que: VP = Verdadeiro Positivo, VN = Verdadeiro Negativo, FP = Falso Positivo e FN = Falso Negativo. Os valores da diagonal principal, ou seja, VP e VN são os casos corretamente classificados, FP e FN que se encontram na outra diagonal são os casos incorretamente classificados (Freitas, 2006).

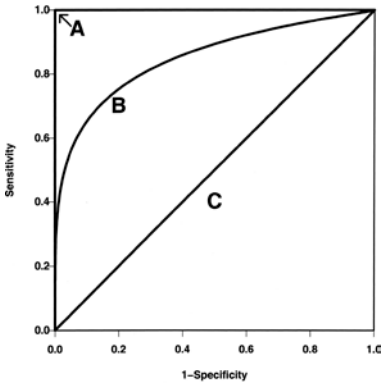
**Figura 15: Representação da matriz de confusão**

| | | Classe Observada | | |
|---|---|---|---|---|
| | | **Positivo** | **Negativo** | |
| **Classe Prevista** | **Sim** | Verdadeiro Positivos | Falso Positivos | ***s*** |
| | **Não** | Falso Negativos | Verdadeiro Negativos | ***n*** |
| | | **P** | **N** | |

Fonte: Adaptado de Freitas (2006: 27)

Outro dos métodos utilizados em modelos de classificação é o gráfico *receiver operating characteristic* (ROC), representado na Figura 16. O gráfico ROC representa-se em duas dimensões, com o valor de VP no eixo dos Y e o valor de FP no eixo dos X, permitindo visualizar a relação entre os verdadeiros positivos e os falsos positivos (Freitas, 2006). De notar que quanto mais perto estiver do canto superior esquerdo, ponto (0,1), melhor será o classificador, pois terá maior taxa de VP e menor taxa de FP.

**Figura 16: Representação da curva ROC**




Fonte: Adaptado de Zou et al. (2007: 656)

Nos modelos de regressão pretende-se escolher aquele que produz valores mais próximos dos dados. A diferença entre o valor real ($y$) e o previsto ($\hat{y}$) é designada por erro ou resíduo ($e_i$), e pode-se calcular um erro global, ou seja, de todos os valores previstos, usando as seguintes métricas para modelos de regressão:

SSE - *Sum squared error* (regressão, "<", [0, Inf[) $\text{SSE} = \sum_i^N e_i^2$ (1)

MSE - *Mean squared error* (regressão, "<", [0, Inf[) $\text{MSE} = \frac{SSE}{N}$ (2)

RMSE - *Root mean squared error* (regressão, "<", [0, Inf[)

$$\text{RMSE} = \sqrt{\sum_{i=1}^{N}(\gamma i - \hat{\gamma} i)^2 / N} \quad (3)$$

RRSE – *Root relative squared error* (regressão, "<", [0%, Inf[)

$$\text{RRSE} = \sqrt{\frac{\sum_{i=1}^{N}(\gamma i - \hat{\gamma} i)^2}{\sum_{i=1}^{N}(\gamma i - \overline{\gamma} i)^2}} \times 100\ (\%) \quad (4)$$

MAE - *Mean absolute error* (regressão, "<", [0, Inf[) $\text{MAE} = \frac{1}{N} \times \sum_{i=1}^{N} |\gamma i - \hat{\gamma}\, i|$ (5)

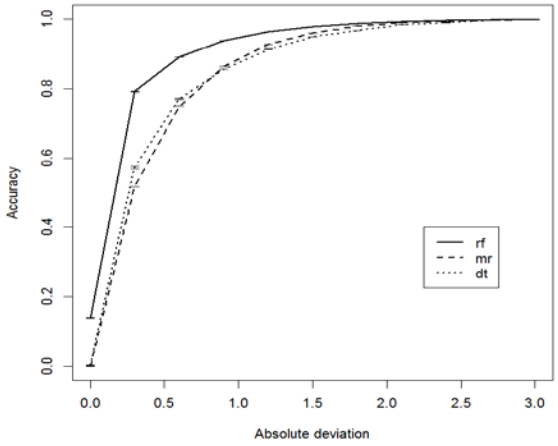
Outra métrica para modelos de regressão é a *relative absolute error* (RAE). O RAE é independente da escala dos valores da variável de saída, e valores próximos de 100% correspondem a um modelo que tem um desempenho similar ao do previsor médio naïve. Quanto menor o RAE, melhor é o modelo de regressão, pelo que o modelo de regressão ideal apresenta um valor próximo de 0 (Witten e Frank, 2005a).

RAE - *relative absolute error* (regressão, "<", [0%, Inf[) $\text{RAE} = \frac{1}{N} \times \sum_{i=1}^{N} \frac{|\gamma_i - \hat{\gamma} i|}{|\gamma_i - \overline{\gamma} i|}$ (6)

Outro método para avaliação e comparação de modelos de regressão é o gráfico *regression error characteristics* (REC), representado na Figura 17 e elaborado pelo autor. A

curva REC mostra a taxa de acerto global (eixo das ordenadas) para diversos valores de tolerância (T) de erro absoluto (eixo das abcissas). A precisão, ou taxa de acertos, é definida como a percentagem de pontos que se encaixam dentro da tolerância. Se a tolerância fosse zero, apenas os pontos de previsão perfeita seriam considerados (Silva, 2010).

**Figura 17: Representação da curva REC**




A qualidade dos dados é um ponto de grande preocupação em qualquer sistema de informação. A existência de muitos dados em falta origina a diminuição da qualidade da informação e dos modelos. Pode-se tratar a problemática dos valores omissos ignorando os registos com valores omissos ou substituindo um valor omisso, mais concretamente da seguinte forma: c*ase substitution* (valor atribuído por um perito); valor médio, mediana ou moda do atributo; *cold deck* (valor retirado de uma base de dados); *hot deck* (valor do exemplo mais semelhante ou próximo); valor estimado por regressão linear; m*ultiple imputation* (combinação dos métodos anteriores) (Brown e Kros, 2003).

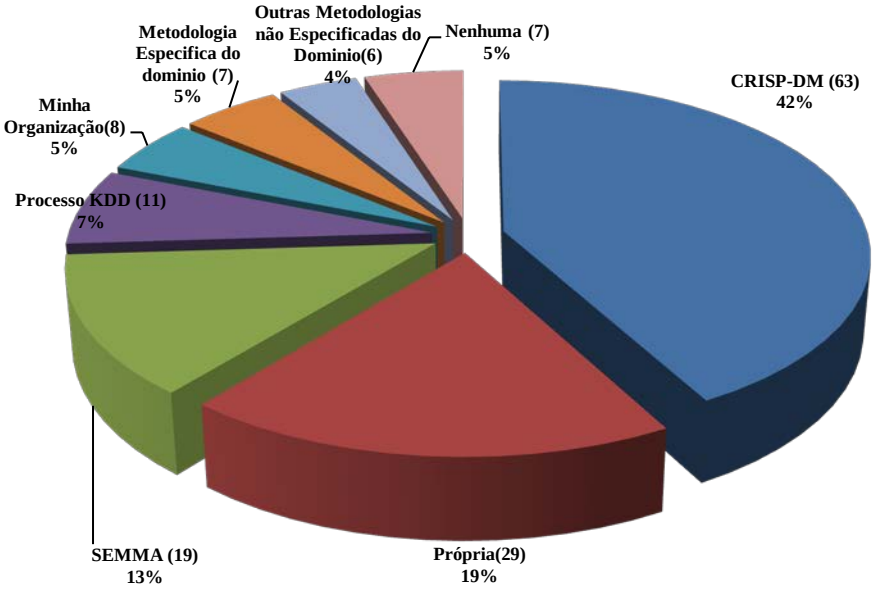
## 2.4. Padrões DM

Nos últimos anos tem havido um enorme crescimento e consolidação do campo do DM, incluindo a implementação de padrões (*standards*) como o caso do PMML, SEMMA e do CRISP-DM.

O PMML define uma linguagem baseada em XML para modelos de DM que incluem por exemplo árvores de decisão e regressão logística. O SEMMA foca-se na aplicação e visualização exploratória de dados estatísticos baseados em técnicas de DM. O CRISP-DM

divide-se nas fases de compreensão do negócio (*business understanding*), compreensão dos dados (*data understanding*), preparação dos dados (*data preparation*), modelação (*modeling*), avaliação (*evaluation*) e implementação (*deployment*). A Figura 18 reflete um estudo publicado na internet em 2007, com base em 150 respostas, identificando a metodologia CRISP-DM como principal escolha dos profissionais (Cunha, 2009).

**Figura 18: Estudo sobre a preferência da metodologia usada para DM**




Fonte: Adaptado de Cunha (2009: 42)

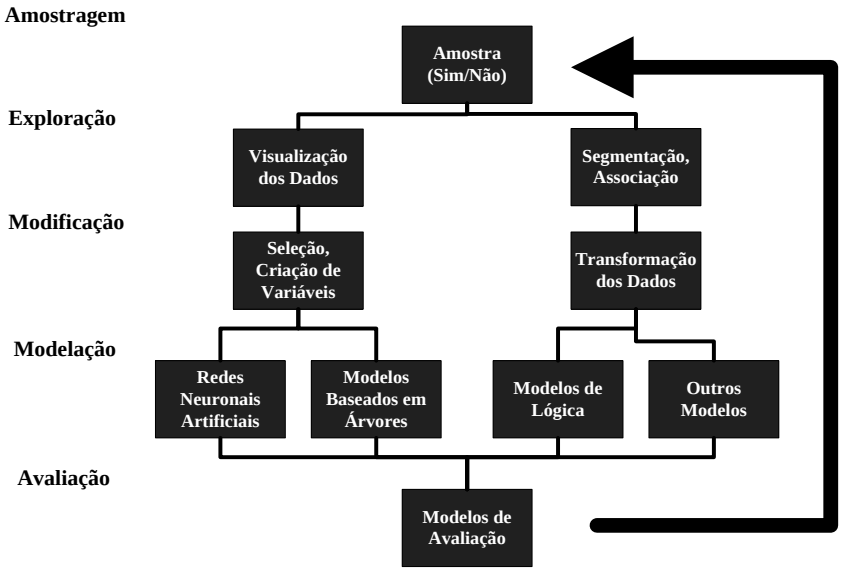
Foi o Grupo DM que desenvolveu o PMML, encontrando-se dividido na seguinte estrutura (Clifton e Thuraisingham, 2001):

- Dicionário de Dados - Nomes e definição dos tipos de campos de entrada e de saída do modelo;
- Esquema de DM – Definição das entradas particulares no dicionário de dados utilizados como entrada e saída através de um modelo particular. Pode também especificar um intervalo de valores aceitáveis e como os valores fora desta faixa são para ser tratados;
- Estatística - Contém estatísticas sobre um campo. Exemplos para atributos numéricos seriam o mínimo, máximo, média, desvio padrão e mediana;
- Normalização - Algumas ferramentas podem esperar entradas compreendidas num determinado intervalo, como tal o modelo descreve como poderá ser efetuado para cada campo;

- O modelo atual - Vários tipos de modelos, como por exemplo, as redes neuronais e as árvores de decisão.

Conforme referido por Santos e Azevedo (2005) e Cunha (2009), o processo SEMMA está compreendido em cinco fases representadas na Figura 19.

**Figura 19: Representação das cinco fases da metodologia SEMMA**




Fonte: Adaptado de Cunha (2009: 46)

A fase de amostragem (*sample*) consiste na amostragem dos dados, extraindo parte do *dataset* que contenha informação significativa para manipulação rápida. Uso de dados de treino para redução do tempo necessário para processamento através de escolha aleatória, de todas as n observações, de observações estratificadas e/ou as 1ªs n observações. A fase exploração (*explore*), consiste na exploração dos dados através da pesquisa de tendências e anomalias não previstas. Na fase modificação (*modify*) será efetuada a alteração dos dados através da criação, seleção e transformação das variáveis. Alguns passos passam pela transformação das variáveis para melhorar o ajuste do modelo, eliminação de valores omissos e a substituição de valores omissos pela média. A fase modelação (*model*) consiste na modelação dos dados, permitindo que o *software* efetue automaticamente a combinação dos dados para atingir o resultado desejado. Escolha do melhor modelo é efetuada através de métodos de otimização e de testes estatísticos significativos. Por último, a fase avaliação (*assessment*) consiste na avaliação dos dados e avaliação dos resultados do processo DM, com base num *framework* para comparação de modelos e previsões.

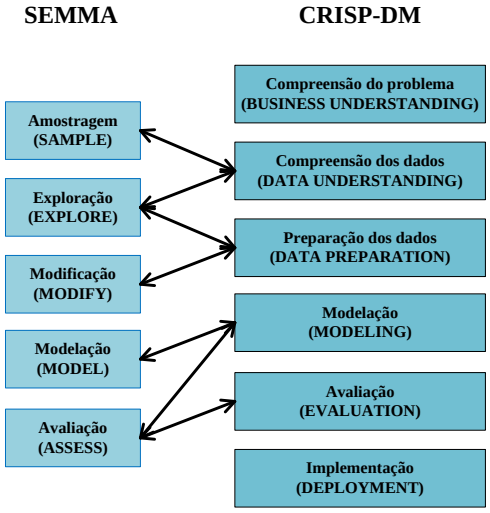
Conforme pesquisa efetuada por Azevedo e Santos (2008), as cinco etapas do processo SEMMA podem ser vistas como a aplicação prática das cinco fases do processo KDD. Na Tabela 2 encontra-se um sumário da correspondência, sendo que o processo CRISP-DM será abordado no ponto 4.2. Na Figura 20 constata-se que muitas das fases do processo SEMMA estão diretamente relacionadas com as fases da metodologia CRISP-DM.

**Tabela 2: Correspondências entre KDD, SEMMA e CRISP-DM**

| KDD | SEMMA | CRISP–DM |
|---|---|---|
| Pre KDD | ---------- | Compreensão do negócio |
| Seleção | Amostragem | Compreensão dos dados |
| Pre processamento | Exploração | |
| Transformação | Modificação | Preparação dos dados |
| DM | Modelação | Modelação |
| Interpretação/Avaliação | Avaliação | Avaliação |
| Post KDD | ---------- | Implementação |

Fonte: Adaptado de Azevedo e Santos (2008: 185)

**Figura 20: Comparação das fases das metodologias SEMMA e CRISP-DM**




Fonte: Adaptado de Cunha (2009: 48)

# 3. Previsão de Tempos de Internamento Hospitalar

## 3.1. Enquadramento

Com o crescente aumento de dados na área da saúde, tornou-se necessária a exploração de várias tecnologias e metodologias de análise por parte dos clínicos. O DM na medicina é aplicado de variadas formas, como por exemplo, na fase de diagnóstico, na identificação de melhores terapias e na pesquisa de novas formas de tratamento. No artigo de Meyfroidt et al. (2009), são descritas algumas barreiras na utilização de base dados hospitalares para fins de pesquisa, como por exemplo, o problema da confidencialidade, a grande quantidade e a qualidade dos dados.

Os sistemas de BI podem ser classificados em dois grandes tipos: baseados em modelos e baseados em dados. Os sistemas baseados em modelos utilizam construções analíticas como a previsão, algoritmos de otimização e árvores de decisão. Os sistemas baseados em dados lidam com DW, bases de dados e tecnologia OLAP (Barrento et al., 2010). Os sistemas de BI, aplicados aos hospitais apresentam como finalidade a faturação e a gestão de cuidados de saúde dos pacientes, contêm informação demográfica do paciente, tempo de internamento, taxas e detalhe de faturação por serviço. Poderá ainda apresentar a finalidade de gestão de camas em hospitais com o objetivo de otimizar ao máximo a utilização das camas (Barrento et al., 2010).

Outra opinião apresenta Marshall et al. (2005), defendendo que a melhor forma de avaliar a atividade do sistema é efetuar a medição do fluxo de pacientes nos hospitais. Um modelo preciso e confiável de fluxo de pacientes permite aos gestores hospitalares prever a atividade futura nas enfermarias dos serviços de internamento. Tais previsões seriam extremamente úteis para avaliar o uso futuro da cama e exigências futuras de recursos hospitalares, tais como o número de camas necessárias e o período de tempo durante o qual as camas são necessárias.

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. Com uma estimativa precisa do tempo de internamento dos pacientes, 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). A mesma opinião é

partilhada por Castillo (2012), alegando que o LOS é uma das medidas empregues em todo o mundo para medir o consume dos recursos hospitalares e a monitorização de desempenho.

## 3.2. Estudos

Merom et al. (1998) apresentam no seu estudo o objetivo de obter uma estimativa da taxa nacional de dias de internamento inadequados (falha dos critérios estabelecidos para internamento) e identificam as variáveis associadas a essa inadequação. Durante as fases do estudo foram analisados 1369 pacientes, efetuados 1003 internamentos em 33 enfermarias num total de 24 hospitais. Foram utilizadas para análise as seguintes variáveis: ocupação; grupo etário; dia inapropriado de internamento; governo; outra entidade hospitalar; outro diagnóstico; sexo; origem da entrada; diagnóstico na admissão e período de estadia. Relativamente à análise estatística, o teste qui-quadrado foi utilizado para avaliar as diferenças nos dias de internamento (adequado/inadequado). O modelo de regressão logística múltipla foi utilizado para avaliar a associação independente entre a variável dependente (dia inapropriado de internamento) e as variáveis associadas com os dias de internamento inapropriado. Neste estudo, concluiu-se que 182 (18,1%) dos dias foram considerados inadequados. Da análise da regressão logística múltipla verificou-se que as variáveis significativamente associadas com os dias inadequados foram: «*government vs other hospital ownership*; *diagnosis on admission of acute cardiac event vs other diagnosis*; *diagnosis on admission of acute cardiac event vs period of the stay*». Concluiu-se ainda que, dos 62,6% dos dias inadequados, 72% foram justificados pela espera dos pacientes por testes ou consultas. O internamento inadequado fixou-se nos 8,6 ± 12,2 dias enquanto o internamento adequado ficou em 6,1 ± 7,3 dias. Por fim, em 6,7% do total de dias analisados, nenhuma justificação foi encontrada para a continuação do internamento.

Gago et al. (2005) apresentam na sua publicação o sistema INTCare, que é um sistema de informação inteligente baseado no processo automático de descoberta de conhecimento e no paradigma de agentes. O sistema foi projetado para funcionar em ambiente UCI, apoiando as decisões dos profissionais de saúde através de modelos DM. Este sistema é composto pelos seguintes agentes: *clinical data entry; pre-processing; data mining; performance; model initialization; data retrieval; prediction; scenario evaluation; interface; ensemble; connection agent.* As técnicas utilizadas permitem prever a falha de órgãos para, por exemplo, o sistema

renal, cardiovascular e nervoso central. Conclui que, os testes não são significativos devido ao baixo número de pacientes considerados.

Abelha et al. (2007) apresentam como objetivo avaliar o tempo de permanência de pacientes internados numa unidade de cuidados intensivos (UCI). Foram analisados todos os pacientes adultos, admitidos entre Outubro de 2004 e Julho de 2005, submetidos a cirurgias não-cardíacas. Os atributos utilizados para categorizar os pacientes foram: idade, género, sexo, índice de massa corporal, estado físico ASA, tipo e magnitude do procedimento cirúrgico, tipo e duração da anestesia, temperatura na admissão, tempo de permanência (LOS) na UCI e no hospital, mortalidade na UCI e no hospital. Para melhor análise dos dados foram criados os seguintes grupos: pacientes com permanência prolongada na UCI; pacientes sem UCI prolongada; pacientes que faleceram na sua estadia no hospital; pacientes que sobreviveram na sua estadia no hospital. Posteriormente, os grupos foram comparados, para avaliar cada variável clínica com a permanência prolongada na UCI ou mortalidade hospitalar, através da análise uni variada por meio da regressão logística binária simples com *odds ratio* e intervalo de confiança de 95%. Foi ainda realizado o teste t para amostras independentes e ainda o teste $\chi 2$ e o teste de *Fisher*. Dos resultados obtidos, verificou-se que o tempo médio de internamento na UCI é de 4,22 ± 8,76 dias.

Aranha et al. (2009) expõem o objetivo de identificar um método estatístico para cálculo do tempo da presença do doente na sala de operação. Foram utilizados os métodos de análise de sobrevivência e o estimador *Kaplan-Meier* para análise de uma amostra aleatória simples de 71 indivíduos sujeitos a cirurgias cardíacas realizadas no ano de 2008. Foram utilizadas como variáveis o procedimento cirúrgico, tempo real da operação, tempo exato da operação e tempo da presença do doente na sala de operação. Conclui que a matriz de relação permite a otimização do tempo do doente na sala de operação, apresentando como média diária a realização de uma operação até 520 minutos, para período disponível de 720 minutos.

Kalra et al. (2010) propõem estudar as tendências temporais do hospital *Temple University*, mais concretamente o fluxo de trabalho no serviço de medicina interna. Os dados analisados foram obtidos no referido serviço para três diferentes períodos de tempo que abrangem 13 anos: 01 de Maio a 30 de Junho de 1991; 01 de maio a 30 de junho de 1998 e 01 de maio até 30 de junho de 2004. Os dados mais relevantes utilizados neste estudo foram: data de admissão, data de saída ou óbito, sexo, idade, código postal de residência, entidade financeira pagadora dos serviços hospitalares e diagnóstico principal. Utilizou a ferramenta *microsoft excel* para se efetuar a análise, comparado o número de admissões para cada período de estudo de sessenta e um dias através do cálculo da probabilidade dos eventos de *Poisson.*

Para variáveis dicotómicas foram efetuados testes para tendências lineares através do teste do qui-quadrado *Mantel-Haenszel.* Para variáveis contínuas foram efetuados testes para as tendências lineares recorrendo aos quadrados de regressão linear para a variável de tempo versus interesse. Neste estudo concluiu-se o seguinte: o número de admissões no serviço de medicina interna aumentou de 1991 (117/mês) para 2004 (455/mês); o tempo médio de internamento no serviço diminuiu de 8,7 para 4,9 dias; a percentagem de pacientes readmitidos num prazo de 12 meses desde a última alta aumentou de 42,3% para 49,5%; a média acumulada do tempo de permanência de 12 meses (incluindo readmissões), diminuiu de 15,8 para 12,5 dias.

Mazier et al. (2010) abordam o problema do agendamento da admissão de internamento num hospital com estadia incerta nos serviços de internamento e com parte significativa de pacientes oriundos do serviço de urgência. Neste estudo propõem-se uma técnica de amostragem para estimar o número de camas necessárias para pacientes oriundos da urgência e internados, com base em três abordagens: uma determinista, uma que considera taxas de serviços e por último a otimização de Monte Carlo. Estabelece como variáveis de estudo as seguintes: número de novos pacientes da urgência; número de novos pacientes para agendamento; número de saídas. Conclui-se que o método de otimização de Monte Carlo proporciona melhores resultados.

Oliveira et al. (2010) apresentam como objetivo avaliar os fatores associados à maior mortalidade e tempo de internamento prolongado numa unidade de terapia intensiva (UTI). Refere que o tempo médio de permanência do paciente nas UTIs brasileiras é de um a seis dias, enquanto Williams et al. (2005) apontam uma média de 5,3 ± 2,6 dias de internamento em UTIs internacionais. Participaram deste estudo 401 pacientes admitidos consecutivamente na UTI de adultos, clínica/cirúrgica do Hospital das Clínicas da Universidade Estadual de Campinas, no período de seis meses. Foram recolhidos dados como: sexo, idade, diagnóstico, antecedentes pessoais, APACHE II, dias de ventilação mecânica invasiva, reintubação orotraqueal, traqueostomia, dias de internação na unidade de terapia intensiva, alta ou óbito na UTI. Ao nível dos resultados, a média de internamento na UTI foi de 8,2 ± 10,8 dias, concluindo o estudo ao afirmar que o APACHE II, traqueostomia e a reintubação estiveram associados à maior taxa de mortalidade e tempo de permanência prolongado em UTI.

Pena et al. (2010) tiveram como objetivo no seu estudo, avaliar a capacidade de previsão do escore de *Ambler* em marcar e antecipar o tempo de permanência numa UTI. Os dados foram obtidos através da cirurgia a 110 pacientes e foram analisadas variáveis resultantes de exames pré-operatórios e de fatores de risco, como por exemplo o tipo de

cirurgia, a idade, o sexo, o género, tipo de cirurgia, prioridade cirúrgica, cirurgia cardíaca prévia e o índice de massa corporal. Ao nível da análise estatística, o desempenho preditivo do escore de *Ambler* foi obtido através da curva ROC e a área sobre as curvas dos modelos aditivo e logístico foram comparadas por meio do teste de *Hanley-MacNeil.* Do estudo obteve-se o resultado de 4,2 dias de permanência em média na UTI e concluiu-se que o escore de *Ambler* obteve uma boa capacidade de previsão.

Santos e Portela (2011) referem que os sistemas inteligentes de apoio à decisão (SIAD) têm vindo a ganhar interesse para previsão da falha de órgãos e resultados dos pacientes a fim de apoiar e orientar a decisão clinica baseada na noção de eventos críticos e nos dados provenientes dos monitores em tempo real. Evidencia a necessidade de redesenhar o sistema de aquisição de dados e a sua arquitetura. O novo modelo permite que os dados sejam adquiridos e armazenados de modo electrónico e a extinção de toda a informação em papel na UCI. Este trabalho apresenta também a necessidade de recolha de mais dados do que os outros previamente recolhidos, principalmente os dados de resultados laboratoriais e de sistemas de medicamentos. O módulo *Data Acquisition* apresenta os novos/reformulados agentes: *Vital Signs Acquisition* (monitorização dos sinais vitais); *ENR Agent* (captura informação clínica médica e de enfermagem)*; LR* (captura informação clínica proveniente dos dados de laboratório).

Azari et al. (2012) propõem uma nova abordagem para previsão de tempos de internamento hospitalar. Apresenta uma metodologia que emprega *clustering* para definição de conjuntos de treino para diferentes algoritmos de classificação. Apresenta ainda como principais atributos os seguintes: *Specialty* (especialidade do serviço), *lenghtOfStay* (n.º de dias que permanece internado), *DSFS* (dias passados desde o primeiro ato desse ano), *Primary condition group* (código generalizado do diagnóstico principal) e *CharlsonIndex* (categorias de códigos de diagnóstico). Dividiu o tempo de internamento em três diferentes grupos: o primeiro grupo corresponde ao intervalo de um a dois dias, em que a intervenção é mínima para reduzir o tempo de internamento. O segundo grupo corresponde ao intervalo de maior que dois e menor que sete, representando uma grande proporção dos pacientes. Por fim, o terceiro grupo corresponde ao intervalo maior que sete dias inclusive, em que são esses os utentes que representam os episódios prolongados e dispendiosos. Conclui que o desempenho das diversas técnicas de classificação, através do uso do *clustering* para formar os conjuntos de treino, permite melhores resultados de previsão em comparação com conjuntos de treino *non-clustering.*

Bachouch et al. (2012) apontam como objetivo o desenvolvimento de um modelo matemático para planeamento de camas hospitalares com base na programação linear, considerando duas classes de pacientes: elegíveis e agudos. Define como variáveis significativas para cada paciente os seguintes: categoria (elegível ou agudo), primeiro e último período de internamento, sexo (masculino ou feminino), tempo de permanência, patologia, estado de contágio, taxas de atividade, custo de um dia de atraso de internamento, disponibilidade de camas, localização da cama, data inicial do internamento e data fim do internamento.

Castillo (2012) pretende desenvolver um modelo estatístico para prever a duração da estadia do paciente em hospitais públicos mexicanos. Utiliza os seguintes atributos*: Age* (idade do paciente), *First Diagnosis* (problema diagnosticado na primeira avaliação médica no hospital), *Diagnosis* (problema de saúde que causa a hospitalização), *Gender* (género), *Occupation* (atividade profissional), *Education level* (grau de escolaridade), *Lenght of stay* (n.º de noites passadas no hospital), *Origin* (origem do internamento), *Previous visits* (n.º de internamentos anteriores), *Surgical procedure* (procedimento cirúrgico principal), *Number of surgical procedures* (n.º total de procedimentos cirúrgicos efetuados durante o internamento), *Ward* (enfermaria). Na generalidade sugere um *two-component Lognormal mixture model* para descrever o tempo de internamento e a criação de uma nova variável intitulada *LOS category* dividida em duas categorias: curta (pacientes com tempos de internamentos até dois dias) e média/longa (pacientes com tempos de internamento superior a três dias). Para o hospital ISSEMyM a variável *LOS category* é redefinida para as seguintes categorias: curta/média (pacientes com tempos de internamento até onze dias) e longa (pacientes com tempos de internamento superiores a doze dias). Por fim, para o hospital MRC a variável foi redefinida para a categoria curta (pacientes com tempos de internamento até três dias) e média/longa (pacientes com tempos de internamento superior a quatro dias). Conclui que o modelo probabilístico para tempos de internamento foi realizado com sucesso, utilizando um modelo baseado análise de *cluster* com modelos de mistura finita.

Freitas et al. (2012) abordam a problemática dos tempos de internamentos discrepantes *(length of stay outliers).*O seu estudo é baseado em episódios de internamento de hospitais públicos pertencentes ao sistema nacional de saúde (SNS), entre o período de 2000 e 2009. As variáveis utilizadas para análise foram as seguintes: *year of discharge, comorbidities, age, A-DRG complexity, readmission, admission and DRG type, discharge status, distance from residence to hospital, hospital type*. Na análise foram utilizados modelos de regressão logística para examinar a associação de cada variável com os tempos de internamentos

discrepantes, e o modelo de regressão linear multivariada com todas as variáveis para calcular o *odds ratio* ajustado e respetivos intervalos de confiança de 95%. Ao nível dos resultados, em nove milhões de episódios de internamento analisados (excluindo os episódios de ambulatório), foram verificados 3,9% de tempos de internamentos discrepantes referentes a 19,2% do número total de internamentos. A mediana/média do tempo de internamento para o caso de valores discrepantes foi de 25/35,5 dias e de 4/6,0 dias para os valores não discrepantes. Concluindo, as variáveis *age, type of admission* e *hospital type* estão significativamente associadas com os altos tempos de internamentos discrepantes.

Rufino et al. (2012) pretendem avaliar os fatores que interferem no tempo de internamento numa enfermaria de clínica médica. O estudo foi realizado num hospital universitário, no período de agosto de 2010 a março de 2011 a uma amostra composta por 48 pacientes internados. Neste estudo foram utilizadas as seguintes variáveis: idade, sexo, internamento anterior, queixa principal, escolaridade, renda familiar mensal, dor (localização, intensidade, duração), tabagismo, etilismo, comorbidades (hipertensão arterial sistémica, diabetes *mellitus*), diagnóstico sindrômico, diagnóstico etiológico, medicação de que faz uso (regularmente ou continuamente) interrupção da medicação após alta hospitalar, intercorrências da internação (infecção, perda de peso, demora para realização de exames). Durante 2009 a média do tempo de internamento de pacientes em hospitais públicos de média e alta complexidade era de 9,3 dias e a média nacional era de 6,6 dias. Através da análise dos dados, os pacientes foram divididos em G1 com um tempo de internamento inferior a 10 dias e em G2 com um tempo de internamento superior a 10 dias. Realizou-se uma análise estatística descritiva através do teste qui-quadrado e do teste de *Mann-Whitney* a um nível de significância de 5%. Obteve-se como resultado deste trabalho a média de 20,9 dias de tempo de internamento.

Sheikh-Nia (2012) pretende demonstrar que os algoritmos genéricos de classificação podem ser empregues em conjunto para prever o tempo de internamento hospitalar de um paciente no próximo ano, com base no seu histórico clínico. Apresenta como objetivo a previsão da variável *Days in Hospital* de cada paciente no ano 2 (Y2), através da informação registada no ano anterior (Y1), referente a esse mesmo paciente. Apresenta como principais atributos os seguintes: *MemberId* (n.º único de utente), *AgeAtFirstClaim* (idade no momento do registo), *Sex* (género), *ProviderId* (médico responsável), *Year* (ano), *Specialty* (nome da especialidade médica), *LenghtOfStay* (n.º de dias desde o dia da admissão), *DSFS* (n.º de dias desde o primeiro registo), *Primary ConditionGroup* (categoria do diagnostico), *CharlsonIndex*, *Procedure group* (categoria do procedimento), *DaysInHospital* (n.º de dias no

hospital no ano anterior). Os resultados mostraram que todos os classificadores independentes superaram a linha de base, por um fator de 1,78 pela ANN, 1,20 por KNN, 1,17 por DT e 1,12 por NB. Na Tabela 3 apresenta-se um breve resumo dos atributos e resultados dos diversos estudos existentes na literatura.

**Tabela 3: Resumo dos atributos e resultados na literatura.**

| Bibliografia | Atributos | Resultados |
|---|---|---|
| Merom et al. (1998) | Ocupação; Grupo etário; Dia inapropriado de internamento; Governo; Outra entidade hospitalar; Outro diagnóstico; Sexo; Origem da entrada; Diagnóstico na admissão; Período de estadia. | • 18,1% dias foram considerados inadequados;<br>• Dos 62,6% dos dias inadequados, 72% foram justificados pela espera dos pacientes por testes ou consultas;<br>• O LOS dias de internamento inadequado foi de 8,6 ± 12,2 dias;<br>• O LOS dias de internamento adequado foi de 6,1 ± 7,3 dias;<br>• 6,7% dias pesquisados sem nenhuma justificação para continuação do internamento; |
| Abelha et al. (2007) | Índice de massa corporal; Estado físico ASA; Magnitude do procedimento cirúrgico; Tipo da anestesia; Duração da anestesia; Temperatura na admissão; Tempo de permanência (LOS) na UCI; Mortalidade na UCI; Mortalidade no hospital; Idade; Sexo; Tipo do procedimento cirúrgico; Tempo de permanência (LOS) no hospital. | • O tempo médio de internamento na UCI é de 4.22 ± 8.76 dias; |
| (Aranha, et al., 2009) | Tempo real da operação (T1);Tempo exato da operação (T2); Tempo da presença do doente na sala de operação (T3); Procedimento cirúrgico. | • Média diária a realização de uma operação até 520 minutos, para período disponível de 720 minutos. |
| (Kalra et al., 2010) | Código postal de residência; Entidade financeira pagadora dos serviços hospitalares; Data de admissão; Data de saída ou óbito; Sexo; Idade; Diagnóstico principal. | • O número de admissões no serviço de medicina interna aumentou de 1991 (117/mês) para 2004 (455/mês);<br>• O tempo médio de internamento no serviço diminuiu de 8,7 para 4,9 dias;<br>• A percentagem de pacientes readmitidos num prazo de 12 meses desde a última alta aumentou de 42,3% para 49,5%;<br>• A média acumulada do tempo de permanência de 12 meses (incluindo readmissões), diminuiu de 15,8 para 12,5 dias. |
| (Mazier, et al., 2010) | Número de novos pacientes da urgência; Número de novos pacientes para agendamento; Número de saídas. | • O método de otimização de Monte Carlo proporciona melhores resultados. |
| (Oliveira, et al., 2010) | Antecedentes pessoais; APACHE II; Dias de ventilação mecânica invasiva; Reintubação oro traqueal; Traqueostomia; Sexo; Idade; Diagnóstico; Dias de internamento na unidade de terapia intensiva; Alta ou óbito na unidade de terapia intensiva. | • A média de internamento na unidade de terapia intensiva foi de 8,2 ± 10,8 dias, concluindo o estudo ao afirmar que o APACHE II, traqueostomia e a reintubação estiveram associados à maior taxa de mortalidade e tempo de permanência prolongado em unidade de terapia intensiva. |
| (Pena, et al., 2010) | Prioridade cirúrgica; Cirurgia cardíaca prévia; Tipo de cirurgia; Índice de massa corporal; Idade; Sexo. | • Obteve-se o resultado de 4,2 dias de permanência em média na unidade de terapia intensiva e concluiu-se que o escore de Ambler |

| Bibliografia | Atributos | Resultados |
|---|---|---|
| | | obteve uma boa capacidade de previsão. |
| (Azari et al., 2012) | DSFS; Indice charlson; Especialidade; Tempo de internamento; Condição primária do grupo. | • O desempenho das diversas técnicas de classificação, através do uso do *clustering* para formar os conjuntos de treino, permite melhores resultados de previsão em comparação com conjuntos de treino *non-clustering*. |
| (Bachouch et al., 2012) | Categoria; Primeiro e último período de internamento; Patologia; Estado de contágio; Taxas de atividade; Custo de um dia de atraso de internamento; Disponibilidade de camas; Localização da cama; Sexo; Tempo de permanência; Data inicial do internamento; Data fim do internamento. | • Prioridade no agendamento dos pacientes agudos, de modo a libertar os recursos do serviço de urgência. |
| (Castillo et al., 2012) | Nº de procedimentos cirúrgicos; Primeiro diagnóstico; Ocupação; Idade; Diagnóstico; Género; Nível de educação; Tempo de internamento; Origem; Internamentos anteriores; Procedimento cirúrgico; Enfermaria | • Um modelo de mistura de dois componentes, lognormal parecia ser mais adequado para descrever o tempo de permanência, obtendo-se a criação de uma nova variável chamada *LOS* com duas categorias: Curto (pacientes com permanência até 2 dias) e Médio/Longo (pacientes com permanência com mais de 3 dias). |
| (Freitas, et al., 2012) | Comorbidades, Complexidade A-DRG; Readmissão; Tipo DRG; Estado de descarga; Distância da residência ao hospital; Tipo de hospital; Ano de descarga; Idade; Tipo de admissão. | • Verificou-se que 3,9% de tempos de internamentos discrepantes referentes a 19,2% do número total de internamentos;<br>• A mediana/média do tempo de internamento para o caso de valores discrepantes foi de 25/35,5 dias e de 4/6,0 dias para os valores não discrepantes;<br>• As variáveis idade, tipo de admissão, tipo de hospital estão significativamente associadas com os altos tempos de internamentos discrepantes. |
| Rufino et al., 2012) | Idade; Sexo, internamento anterior; queixa principal; escolaridade; renda familiar mensal; dor; tabagismo; etilismo; comorbidades; diagnóstico sindrômico; diagnóstico etiológico; interrupção da medicação após alta hospitalar; intercorrências da internação. | • No ano de 2009 a média do tempo de internamento de pacientes em hospitais públicos de média e alta complexidade era de 9,3 dias e a média nacional era de 6,6 dias.<br>• Neste trabalho obteve-se a média de 20,9 dias de tempo de internamento. |
| (Sheikh-Nia, 2012) | DSFS; Índice charlson; Grupo de procedimento; Dias no hospital; Nº de identificação; Idade na primeira queixa; Sexo; Médico; Ano; Especialidade; Tempo de internamento; Condição de grupo primária. | • Os resultados mostraram que, em termos de medida-F (medida combinada de precisão e recuperação), todos os classificadores independentes superaram a linha de base.<br>• Fator de 1,78 pela ANN, 1,20 por KNN, 1,17 por DT e 1,12 por NB. |

# 4. Trabalho Realizado

## 4.1. Contextualização

O Hospital da Força Aérea (HFA) teve a sua génese no então denominado Hospital Militar da Terra Chã, na Ilha Terceira, Açores (1943-1975). Inicialmente, esta instituição hospitalar pertenceu ao Exército Português para apoiar os militares do Corpo Expedicionário Português, estacionados na Ilha Terceira e os militares ingleses que ali desembarcaram.

Mais tarde, com a chegada das forças norte-americanas, esta unidade hospitalar passou para a dependência da Força Aérea Portuguesa (FAP), tendo sido o único hospital deste ramo, até 1975, ano em que foi extinto pelo Decreto-lei nº 525/75, de 25 de setembro. Este hospital teve um papel singular, quer no tratamento de militares evacuados dos teatros de operações da guerra colonial (1961-1975), quer na formação e experiência do corpo clínico da FAP.

Em 1972, foi criado o Núcleo Hospitalar Especializado da Força Aérea n.º 1 (NHEFA1), com base no Centro de Diagnóstico e Tratamento da Formação de Adidos da Força Aérea, no Paço do Lumiar, em Lisboa, compreendendo uma enfermaria geral e o Centro Médico Psicológico. No ano de 1979, o NHEFA1 passou a denominar-se Hospital da Força Aérea, vocacionado essencialmente para a prevenção, tratamento e reabilitação dos militares da FAP e suas famílias. Em 2000, o HFA iniciou o processo de modernização do seu sistema de saúde ao nível dos sistemas de informação (SI) e tecnologias de informação (TI), dividido nas seguintes fases: instalação de base de dados para suporte às futuras plataformas electrónicas; implementação da plataforma de gestão clínica e hospitalar e do *electronic patient record* (EPR), permitindo maior flexibilidade no registo clínico.

Ao HFA competia-lhe: tratar e reabilitar os militares da Força Aérea, os seus familiares e quando superiormente autorizado, outros doentes; colaborar com outros estabelecimentos hospitalares na prestação de serviços, em ações de formação e investigação científica.

No âmbito do processo de reestruturação hospitalar, foi publicada a criação do Hospital das Forças Armadas (HFAR) enquanto hospital militar único. O polo de Lisboa do HFAR será o resultado da fusão do Hospital da Marinha (HM), o Hospital Militar Principal (HMP), o Hospital Militar de Belém (HMB) e o HFA.

No contexto nacional, Carriço (2012) escreve no jornal de negócios a vontade do ministério da saúde em reduzir o tempo de internamento nos hospitais e libertar camas. Ao abrigo do acordo com a *troika*, Portugal comprometeu-se a cortar 15% nas despesas com os

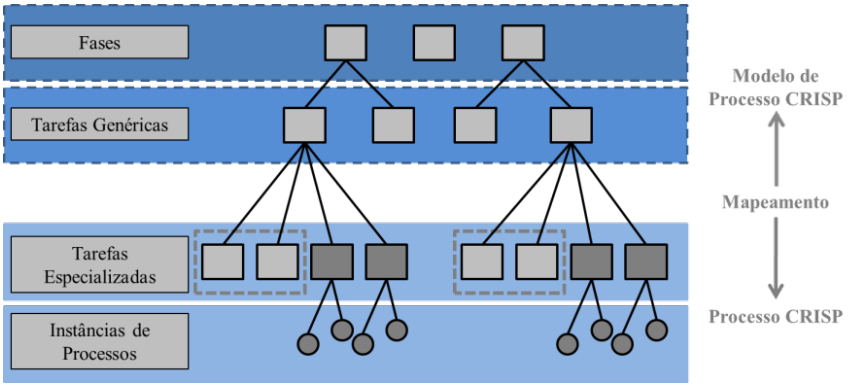
hospitais, o que será feito pelo Ministério da Saúde através da redução do tempo de internamento e do aumento das cirurgias em ambulatório (mais baratas). Estando o HFAR debaixo da alçada do Ministério da Defesa, prevê-se uma situação idêntica devido aos cortes orçamentais previstos.

## 4.2. Metodologia

Pesquisar significa, de forma bem simples, procurar respostas para indagações propostas, uma atividade básica das ciências na sua descoberta da realidade. Do ponto de vista da forma de abordagem do problema, trata-se de uma pesquisa qualitativa. Do ponto de vista de seus objetivos, enquadra-se como uma pesquisa exploratória, visando proporcionar familiaridade com o problema com vista a torná-lo explícito.

Este tipo de pesquisa envolve levantamento bibliográfico e teve como início a escolha da problemática a abordar. Conforme referido no ponto 2.4, a metodologia escolhida para esta investigação foi a metodologia CRISP-DM, que é mais completa quando comparada com a metodologia SEMMA. Esta metodologia 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. Segundo a Figura 21, esta metodologia encontra-se representada por um modelo hierárquico de processos, representados por um conjunto de tarefas com quatro níveis de abstração: fases, tarefas genéricas, tarefas especializadas e instâncias de processos (Chapman et al., 2000).

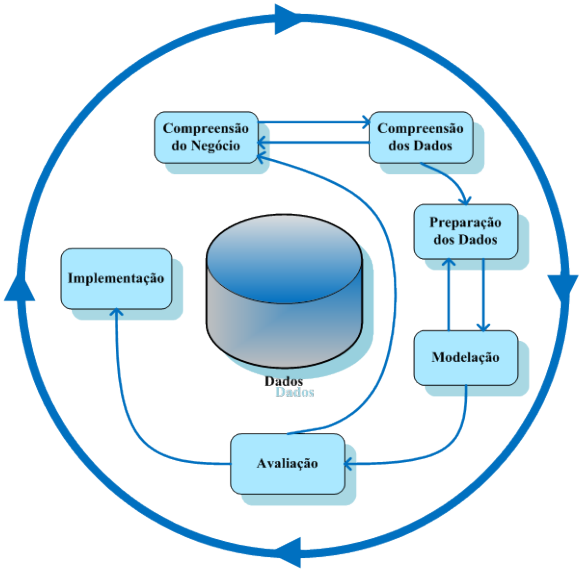
**Figura 21: Representação dos quatro níveis metodologia CRISP-DM**




Fonte: Adaptado de Chapman et al. (2000: 6)

De seguida, é apresentada uma breve descrição das seis fases relativas a esta metodologia, representadas na Figura 22.

**Figura 22: Representação das seis fases da metodologia CRISP-DM**




Fonte: Adaptado de Chapman et al. (2000: 10)

Na fase compreensão do negócio (*business understanding*), é definido o objetivo de negócio, avaliada a situação, determinadas as metas de DM e produzido um plano de projeto.

A fase compreensão dos dados (*data understanding*) é iniciada com a aquisição dos dados iniciais, descrição, exploração e verificação da qualidade dos dados.

Na fase preparação dos dados (*data preparation*) pretende-se selecionar os dados, posteriormente limpar, construir, integrar e formatar os dados.

Modelação (*modeling*) é a fase que seleciona a técnica de modelação, gera o projeto de teste e por fim constrói e avalia o modelo.

A fase avaliação (*evaluation*) é responsável pela avaliação dos resultados e revisão dos processos.

Por fim, a fase implementação (*deployment*) elabora o planeamento da implementação, produção de relatório final e projeto de revisão (Clifton e Thuraisingham, 2001).

As fases do ciclo de vida são flexíveis, pois em qualquer ponto do projeto poderá ser necessário recuar para fases anteriores. O resultado de cada fase determina qual a próxima fase a executar.

### 4.2.1. Compreensão do Negócio

Nesta fase pretende-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).

Na primeira etapa pretende-se a identificação inicial de fatores importantes que possam influenciar os resultados do projeto. No âmbito do processo de reestruturação hospitalar, foi publicado no Diário da República 1.ª série N.º 158 de 16 de agosto de 2012 e aprovado no Decreto -Lei n.º 234/2009 de 15 de setembro, a criação do HFAR enquanto hospital militar único, organizado em dois polos hospitalares, um em Lisboa e outro no Porto, como corolário do processo de reestruturação hospitalar nas Forças Armadas. O polo de Lisboa do HFAR é o resultado da fusão do HM, do HMP, do HMB e do HFA. Este novo hospital tem como missão principal a prestação de cuidados de saúde aos beneficiários da ADM (assistência doença dos militares), apresentando uma nova missão que é a promoção, cooperação e articulação com o SNS. Ao nível dos objetivos de negócio é apresentado como objetivo principal a previsão de tempos de permanência nos serviços de internamento. Para responder a este objetivo, será efetuada a descoberta de comportamentos e padrões existentes no processo de internamento hospitalar, com base em técnicas de DM. Este trabalho aborda o problema do agendamento da admissão hospitalar, em que a maior dificuldade será manter camas suficientes para pacientes oriundos do serviço de urgência e futuros pacientes internados, devido à fusão dos hospitais militares da região de Lisboa. 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, é apresentada informação mais detalhada sobre os recursos, constrangimentos, suposições e outros fatores a serem considerados, de modo a determinar uma meta para análise dos dados e plano de projeto.

Relativamente aos recursos disponíveis, e conforme referido no ponto 4.1, a instituição hospitalar dispõe de uma solução tecnológica para registo de informação clínica relevante, suportada numa BD relacional Oracle 10G. Ainda ao nível de recursos de *software*, serão utilizadas 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 no Departamento de Estatística da Universidade de Auckland, Nova Zelândia. Este sistema foi criado com intuito de permitir uma programação direcionada para a estatística e análise de dados (Costa, 2009) e a adoção da biblioteca *rminer* para a ferramenta R facilita o uso de algoritmos de DM nas tarefas de regressão e classificação, através de um conjunto reduzido e

coerente de funções (Cortez, 2010). Por exemplo, realiza uma seleção automática de modelos, ou seja, ajuste ideal dos hiper-parâmetros das ANN e SVM e permite ainda o cálculo de diversas métricas e gráficos que são úteis para o DM, incluindo procedimentos de análise de sensibilidade para extrair informação a partir de modelos treinados (Costa, 2009).

Através de manipulação de dados via *structured query language* (SQL), os dados serão extraídos para análise, mais propriamente o ficheiro internamento em formato **.xls*, **.csv*. Relativamente aos requisitos, premissas e restrições, estabeleceu-se setembro de 2013 como data de fim do projeto. Outro requisito será a segurança e proteção dos dados dos utentes, e por fim a apresentação de uma solução de BI que permita analisar os tempos de permanência nos serviços de internamento.

Os riscos que se apresentam como mais críticos para este projeto são a definição de objetivos, a possível extração deficiente dos dados e a definição incorreta de um modelo de previsão. Por último, esta subfase apresenta como necessário um glossário de terminologia de negócio e DM, com informação detalhada no Anexo A.

Na subfase seguinte pretende-se definir as metas de negócio e transpor os objetivos de negócio para a análise de dados. Deste modo, ao nível do negócio, estabeleceu-se nessa dissertação uma meta que seria desejável de obter: **o modelo deverá permitir efetuar previsões com uma margem de erro inferior a 20%.** A escolha recairá no modelo de regressão que produza valores mais próximos dos dados, pelo que o modelo de regressão ideal apresenta um valor próximo de 0%.

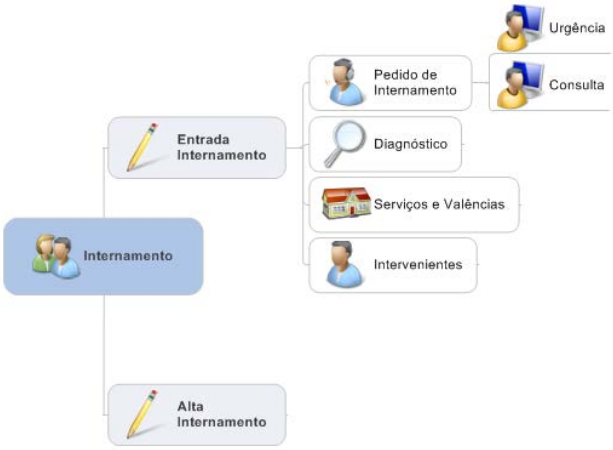
Por fim, na subfase plano de projeto pretende-se descrever o plano para alcançar as metas estabelecidas, através da especificação dos passos a executar e da seleção de ferramentas e técnicas. As fases, tarefas e precedências estão descriminadas com maior detalhe no Anexo A e para a fase de modelação foram escolhidas as técnicas de DM anteriormente referidas no capítulo 2.3, concretamente cinco técnicas de regressão: *Naive*, MR, DT, ANN, RF e SVM.

### 4.2.2. Compreensão dos Dados

Esta fase inicia-se com a aquisição, preparação, análise dos dados recolhidos, identificação dos problemas ao nível da qualidade dos dados (verificação da sua qualidade e procura de valores discrepantes), percepção inicial das relações entre os dados e detecção de subconjuntos interessantes (Chapman et al., 2000). De modo a se obter uma melhor percepção

do circuito de internamento existente, os episódios de pedido de internamento têm a sua origem em episódio de consulta, de urgência ou de plano operatório anteriormente registados, gerando um episódio de pré-internamento. Com a entrada do paciente é gerado um episódio de internamento associado 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 23.

**Figura 23: Diagrama do processo de internamento hospitalar**



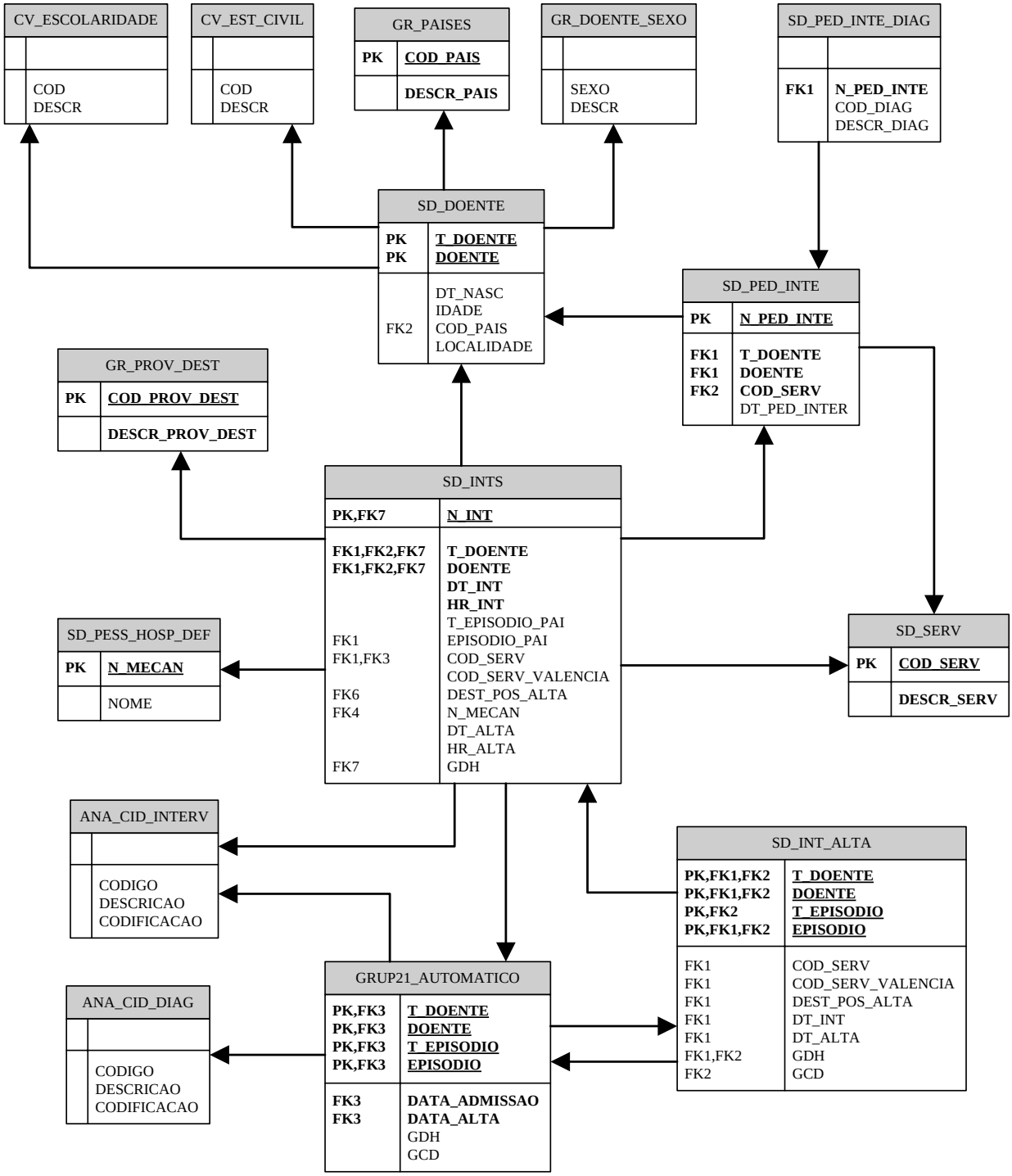

Na subfase seguinte procede-se à aquisição e leitura dos dados iniciais, enunciados no plano do projeto para a sua compreensão, mencionando o *dataset* adquirido, os métodos utilizados para sua aquisição e problemas encontrados. Após análise da estrutura da BD verificou-se o relacionamento entre as diversas tabelas associadas ao processo de internamento, representadas na Figura 24.

Relativamente aos dados dos internamentos, os atributos disponibilizados já se resumiam a um registo por utente e por n.º de processo de internamento, permitindo a sua transposição direta para o ficheiro de entrada às técnicas de DM. A manipulação de dados foi efetuada via ferramenta *SQL Navigator* 6.4, executado o *script* para aquisição dos dados iniciais e discriminados no Anexo B.

Durante a tarefa da descrição e análise dos dados adquiridos, verificou-se que os dados ocorreram entre Outubro de 2000 e Março de 2013 e durante este período foram efetuados cerca de 26462 episódios de internamento associados às diversas especialidades médicas.

Nesta fase crucial, foi obtido o ficheiro *internamento.xls*, peça chave para a continuação do restante trabalho. De referir que durante as diversas fases, foi utilizada a ferramenta *microsoft excel* para visualização e tratamento do conjunto de dados.

**Figura 24: Diagrama de entidades – Notação UML**



No Anexo B é apresentado uma breve descrição dos diversos atributos através do dicionário de dados, concluindo-se que o atributo a prever seria o número de dias de internamento do utente (N_Dias_Intern). Na sequência das tarefas anteriores, conclui-se que os recursos existentes adequam-se ao suporte da investigação em curso.

Assim, de acordo com a metodologia seguida, segue-se a tarefa de exploração dos dados e a análise da distribuição dos atributos fundamentais. Para análise dos atributos, recorreu-se à ferramenta *open source rattle.*

Esta ferramenta gráfica assenta no ambiente de programação *open source* R e permite uma análise mais sofisticada dos dados e a construção de gráficos simples. Numa segunda etapa utilizou-se a interface de linha de comandos da ferramenta R.

O R dispõe de um conjunto de ferramentas para a análise estatística dos dados, de diversas técnicas de DM a aplicar nesta investigação e permite uma enorme flexibilidade na importação dos dados de diversos formatos.

Após análise dos modelos existentes na literatura, estabeleceu-se uma relação com os atributos utilizados nos diversos trabalhos pesquisados, conforme informação disponível na Tabela 4, apontando-se em referência o nome o atributo utilizado na ferramenta R.

**Tabela 4: Relação dos atributos aqui adotados com os propostos pelo estado de arte**

| Bibliografia | Estado de Arte |
|---|---|
| (Merom, et al., 1998) | • Sexo **(Sexo)**<br>• Tipo de episódio de origem **(T_Episod_Origem)**<br>• Diagnóstico principal **(Diag_Principal)**<br>• Nº de dias de internamento **(N_Dias_Intern)** |
| (Abelha, et al., 2007) | • Idade **(Idade_Intern)**<br>• Sexo **(Sexo)**<br>• Procedimento cirúrgico principal **(Proc_Principal)**<br>• Nº de dias de internamento **(N_Dias_Intern)** |
| (Aranha, et al., 2009) | • Procedimento cirúrgico principal **(Proc_Principal)** |
| (Kalra, et al., 2010) | • Data de admissão no internamento **(Dt_Internamento)**<br>• Data de saída do internamento **(Dt_Alta_Intern)**<br>• Sexo **(Sexo)**<br>• Idade **(Idade_Intern)**<br>• Diagnóstico principal **(Diag_Principal)** |
| (Pena, et al., 2010) | • Idade **(Idade_Intern)**<br>• Sexo **(Sexo)** |
| (Oliveira, et al., 2010) | • Sexo **(Sexo)**<br>• Idade **(Idade_Intern)**<br>• Diagnóstico principal **(Diag_Principal)**<br>• Nº de dias de internamento **(N_Dias_Intern)**<br>• Data da alta do internamento **(Dt_Alta_Intern)** |
| (Azari et al., 2012) | • Especialidade médica **(Espec_Medica)**<br>• Nº de dias de internamento **(N_Dias_Intern)**<br>• Diagnóstico principal **(Diag_Principal)** |

| Bibliografia | Estado de Arte |
|---|---|
| (Bachouch et al., 2012) | • Sexo **(Sexo)**<br>• Nº de dias de internamento **(N_Dias_Intern)**<br>• Data do internamento **(Dt_Internamento)**<br>• Data alta do internamento **(Dt_Alta_Intern)** |
| (Castillo, 2012) | • Idade **(Idade_Intern)**<br>• Diagnóstico principal **(Diag_Principal)**<br>• Sexo **(Sexo)**<br>• Escolaridade **(Escolaridade)**<br>• Nº de dias de internamento **(N_Dias_Intern)**<br>• Tipo de episódio de internamento **(T_Episod_Intern)**<br>• Nº de internamentos anteriores **(N_Intern_Anterior)**<br>• Procedimento cirúrgico principal **(Proc_Principal)**<br>• Serviço de internamento **(Serv_Intern)** |
| (Freitas, et al., 2012) | • Ano do internamento **(Ano_Intern)**<br>• Idade **(Idade_Intern)**<br>• Tipo de episódio de internamento **(T_Episod_Intern)** |
| (Rufino et al., 2012) | • Idade **(Idade_Intern)**<br>• Sexo **(Sexo)**<br>• Nº de internamentos anteriores **(N_Intern_Anterior)**<br>• Escolaridade **(Escolaridade)** |
| (Sheikh-Nia, 2012) | • Nº de doente **(N_Doente)**<br>• Idade **(Idade_Intern)**<br>• Sexo **(Sexo)**<br>• Médico alta **(N_Med_Alta)**<br>• Ano **(Ano_Intern)**<br>• Especialidade médica **(Espec_Medica)**<br>• Nº de dias de internamento **(N_Dias_Intern)**<br>• Diagnóstico principal **(Diag_Principal)** |

Decorrente da fase anterior e de modo a confirmar os atributos apresentados no estado de arte, foi solicitado a um painel de especialistas de diversas especialidades médicas do hospital, a seleção dos atributos relevantes para previsão dos internamentos, tendo-se obtido a informação descrita na Tabela 5.

A expectativa criada com esta primeira validação técnica dos atributos era reduzir o conjunto de dados existentes, obtendo melhor desempenho na execução das técnicas de DM e melhores modelos.

**Tabela 5: Atributos validados pelo painel de especialistas**

| Atributo \ Especialista | João Mairos Médico Ginecologia | Silvia Sousa Médica Medicina Interna | Rafael Fernandes Médico Cirúrgia Geral | Reis Ferreira Médico Pneumologista | Manuel Domingos Médico Medicina Interna | Paulo Neves Médico Cirúrgia Plástica | Regina Ramos Médica Cirúrgia Geral | Eduardo Fazenda Médico Gastrenterologista | Rui Carvalho Médico Neurocirurgião |
|---|---|---|---|---|---|---|---|---|---|
| Sexo | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Dt_Nascimento | | | | | ✓ | | | | |
| Idade_Intern | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ |
| País | | ✓ | | | ✓ | | ✓ | | |
| Localidade | | ✓ | | | | ✓ | ✓ | ✓ | ✓ |
| Escolaridade | | ✓ | | | ✓ | ✓ | | | |
| Est_Civil | | ✓ | | | ✓ | ✓ | ✓ | | |
| T_Episod_Origem | ✓ | | ✓ | | ✓ | ✓ | ✓ | | |
| T_Episod_Intern | ✓ | | ✓ | | ✓ | ✓ | ✓ | | |
| Dt_Ped_Intern | | | | | | ✓ | ✓ | | |
| Serv_Intern | | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Espec_Medica | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Dest_Alta | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | |
| N_Med_Alta | ✓ | | | | ✓ | ✓ | ✓ | ✓ | |
| Tratamento | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Proc_Principal | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Diag_Principal | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| Diag_Inicial | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| GDH | | ✓ | | | ✓ | ✓ | | ✓ | ✓ |
| GCD | | ✓ | | | ✓ | | | | |
| Dt_Internamento | ✓ | | | | ✓ | ✓ | ✓ | | |
| Mes_Intern | | | | | | | ✓ | ✓ | |
| Ano_Intern | | | | | | | ✓ | | |
| Hora_Intern | | ✓ | | | ✓ | ✓ | ✓ | | |
| Dt_Alta_Intern | ✓ | | | | ✓ | ✓ | ✓ | | |
| Hora_Alta_Intern | | | | | | ✓ | ✓ | | |
| N_Intern_Anterior | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| N_Dias_Intern | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Na próxima iteração procedeu-se à análise estatística dos atributos, verificação da qualidade dos dados e a sua adequação para o estudo dos tempos de internamento, estando toda a informação disponível no Anexo B.

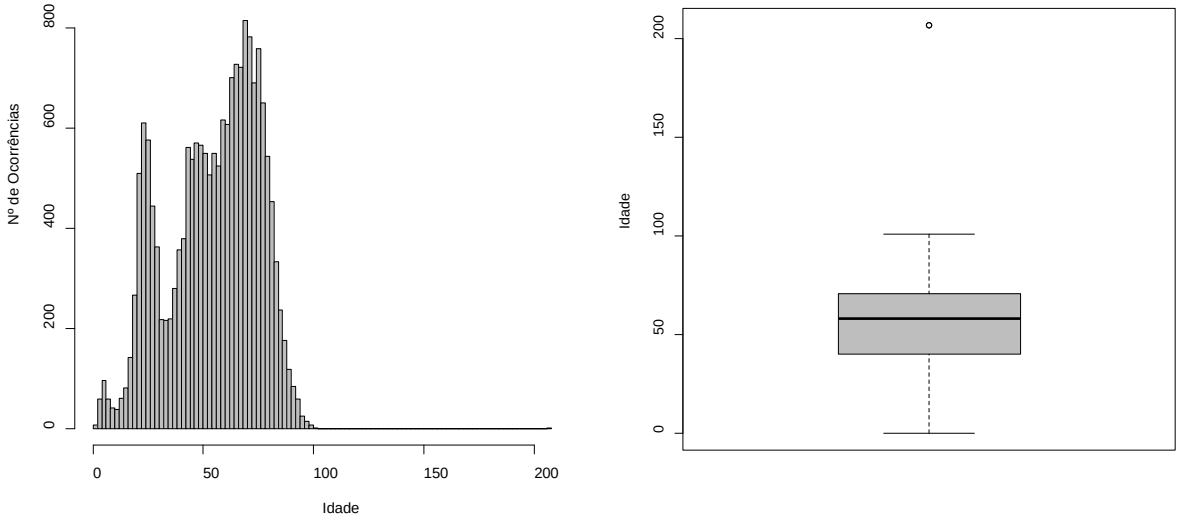
Neste ponto foi importante o uso da ferramenta *microsoft excel* para eliminação de carateres portugueses que não são compatíveis com a biblioteca *open source rminer*. Nesta análise, verificou-se que os atributos qualitativos representam cerca de 70% dos atributos selecionados e os quantitativos representam os restantes 30%. Existem valores omissos para alguns dos atributos, como o caso código de diagnóstico principal que apresenta 19268 valores em falta. Alguns dos atributos qualitativos apresentam um elevado número de valores

possíveis, representando uma dispersão elevada, podendo dificultar a utilização destes mesmos atributos pelas técnicas de DM escolhidas aquando da modelação. Apresenta-se como casos de exemplo as 11195 datas de nascimento possíveis e as 2436 localidades diferentes. Os gráficos gerados e escolhidos na análise subsequente dependeu do tipo de atributos a analisar e do número de valores possíveis.

Deste modo, para os atributos qualitativos com muitos valores possíveis, optou-se pelo diagrama de frequências do atributo. No caso dos atributos quantitativos, os gráficos mais utilizados foram o diagrama de frequências e o diagrama de extremos e quartis (*boxplot*). O diagrama de frequências permite obter uma representação gráfica em que um conjunto de dados é agrupado em classes uniformes, cuja base horizontal são as classes e seu intervalo, e a altura vertical representa a frequência com que os valores dessa classe estão presentes no conjunto de dados. O diagrama de extremos e quartis é uma representação gráfica sobre a forma como os dados se distribuem, nomeadamente a sua menor ou maior concentração, simetria e existência de valores discrepantes. Um valor é considerado discrepante quando não se encontra compreendido no intervalo entre a barreira inferior e barreira superior.

A ação de processamento para análise dos atributos encontra-se descrita no Anexo B, pois toda a análise é demasiado exaustiva. A título de exemplo, pode-se verificar na Figura 25 que o atributo associado à idade do utente apresenta valores acima dos 90 anos de idade. São considerados valores válidos à exceção da idade com o valor 207, tendo sido considerado valor errado e consequentemente eliminado.

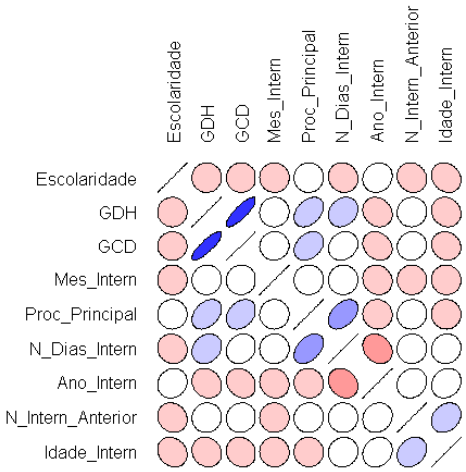
**Figura 25: Diagrama de frequência e *boxplot* para o atributo Idade_Intern**



Para os outros atributos foi realizada semelhante operação de análise gráfica, tendo em vista a validação dos diversos atributos, podendo ser visualizado no Anexo B. Na análise efetuada à caraterização dos utentes, identificou-se que ao nível da escolaridade, a maioria possui a formação básica. Pelo diagrama de extremos e quartis, verifica-se um enviesamento inferior, enquanto a parte central dos dados apresenta um enviesamento superior. O género masculino evidencia-se no atributo sexo e os utentes apresentam-se maioritariamente com o estado civil "casado(a)". Foi no mês de Janeiro que se obteve o maior número de internamentos, apresentando uma boa concentração dos dados, apesar do ligeiro enviesamento superior dos dados centrais. Segundo a figura no Anexo B relacionada com o tipo de episódio de internamento, foram realizados maioritariamente episódios de internamento, principalmente nos serviços físicos de especialidades e de cirurgia.

Ao nível da relação entre os atributos, verificou-se que com os dados fornecidos pelas ferramentas de DM utilizadas, pode-se facilmente aferir as relações entre alguns dos atributos. Um indicador estatístico importante para avaliação do grau de correlação (intensidade e direção) entre as variáveis é o coeficiente de correlação de *spearman*, representado na Figura 26. É uma medida de correlação não-paramétrica que permite usar variáveis ordinais, não requerendo que as variáveis sejam quantitativas.

**Figura 26: Gráfico correlação de *Spearman***




A análise de correlação devolve o valor da relação entre as duas variáveis, estando o seu valor compreendido entre -1 e 1. O sinal indica se a direção da correlação é negativa ou positiva e o valor indica a força da correlação. Sendo assim, neste conjunto de dados verifica-se que na maioria dos casos existe uma fraca correlação positiva ou negativa (0 a ± 0,20) entre os atributos. Os restantes casos apresentam uma correlação moderada positiva (0,20 a 0,7),

respetivamente a relação entre os atributos GDH e Proc_Principal (0,38), GDH e N_Dias_Intern (0,22), GCD e Proc_Principal (0,31), Proc_Principal e N_Dias_Intern (0,44), à excepção da correlação forte entre os atributos GDH e GCD (0,84)

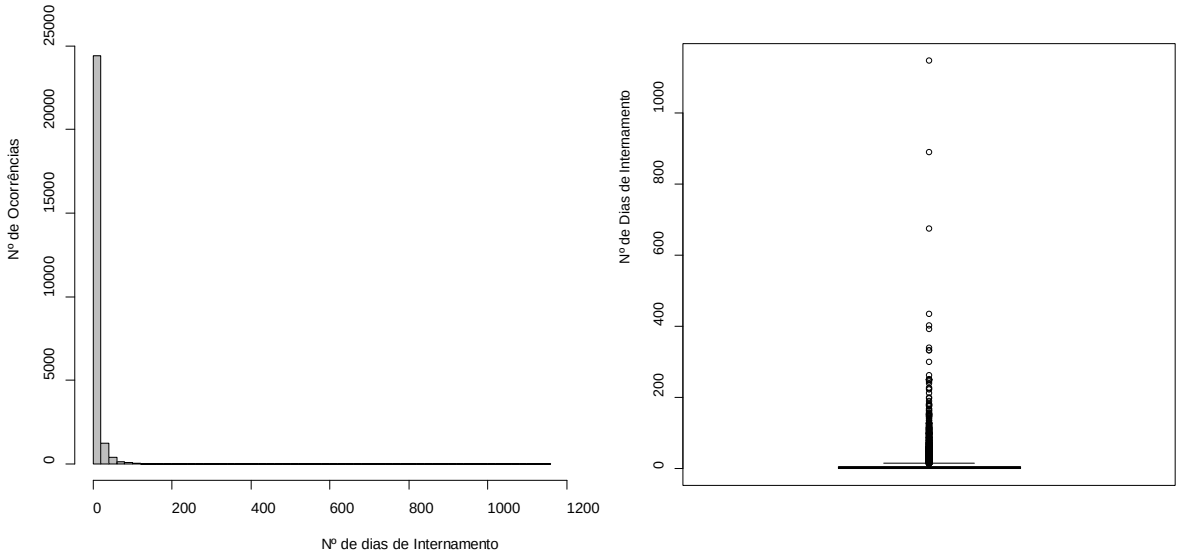
Conclui-se através desta análise a possível existência de dados redundantes ou um atributo ser derivado diretamente de outro (GDH e GCD).

O código em R dos gráficos representados na fase compreensão dos dados é apresentado em detalhe no Anexo B. Deste modo, avançou-se para a fase de preparação dos dados, com um conjunto de 28 atributos escolhidos anteriormente pelo painel de especialistas.

### 4.2.3. Preparação dos Dados

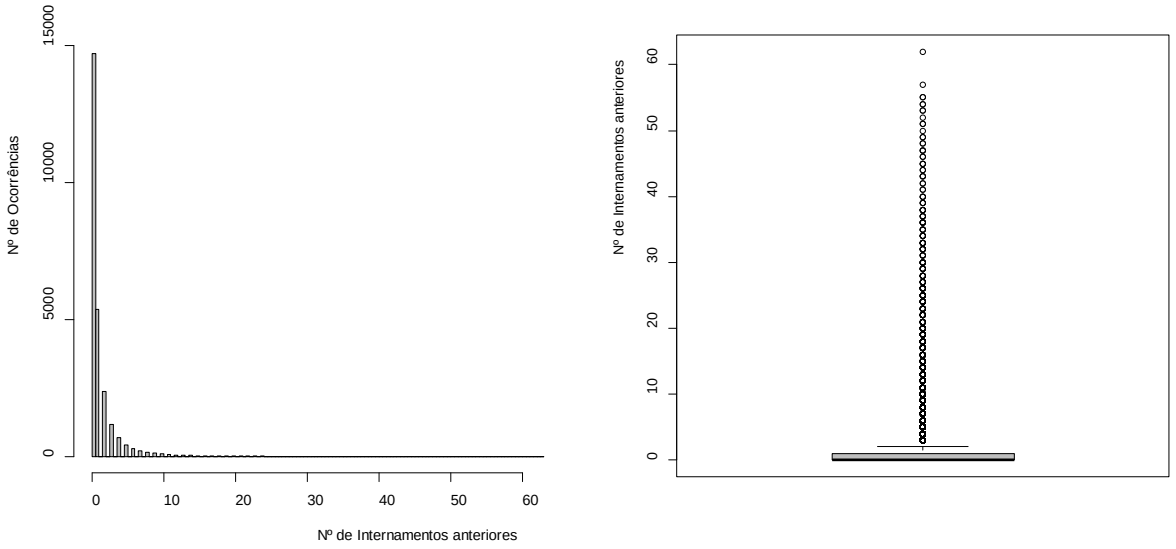
Esta tarefa pretende efetuar a seleção dos atributos, transformação e limpeza dos dados, abrangendo todas as atividades necessárias para construir o conjunto de dados final (Chapman et al., 2000). Durante a execução da seleção dos dados, optou-se por efetuar uma análise exploratória mais profunda dos atributos anteriormente selecionados no ponto 4.2.2 pelo painel de especialistas médicos. Assim sendo, na Figura 27 verifica-se que uma instância associada ao atributo número de dias de internamento foi considerada valor errado e posteriormente eliminada. Este caso refere-se a 2294 dias de internamento de um utente relativo a um episódio de ambulatório.

**Figura 27: Diagrama de frequência e *boxplot* para o atributo N_Dias_Intern**



Analisando a Figura 27 respeitante ao atributo N_Dias_Intern e a Figura 28 respeitante ao atributo N_Intern_Anterior, pode-se concluir que ambos os gráficos produzem um maior agrupamento de valores observados próximos da média (1,51 para a variável de número de internamentos anteriores e 6,92 para a variável número de dias de internamento) e menores frequências aquando do afastamento da média.

**Figura 28: Diagrama de frequência e *boxplot* do atributo N_Intern_Anterior**




Das representações gráficas anteriores ressalta imediatamente que existem vários valores discrepantes que se afastam muito do padrão, no entanto não foram nesta fase considerados valores errados.

Durante a tarefa de limpeza dos dados e para o caso apresentado, optou-se pela remoção do *dataset* os 29 episódios de internamento associados ao código de serviço 9, sendo este um serviço virtual para testes aplicacionais.

Outro procedimento efetuado para limpeza dos dados foi a verificação da correspondência entre os códigos dos vários atributos e seus respetivos atributos descritivos, de modo a eliminar um grande número de níveis existentes. Por exemplo, para o caso do atributo Escolaridade, os códigos 10, 31, 99, 999 foram substituídos por "NA", pois não apresentavam descritivo associado. No caso do atributo Est_Civil o valor 9 e "A" foram substituídos por "NA", pois apresentavam o descritivo "Desconhecido". Nos atributos Tratamento e Proc_Principal o valor 9999 foi substituído por "NA" por ser referente a um "Procedimento, Não operatório, NCOP".

A Tabela 6 apresenta os atributos redundantes excluídos, os que apresentam menos relevância em termos teóricos, os que possuem um elevado número de valores possíveis ou um elevado número de valores omissos.

**Tabela 6: Atributos excluídos do *dataset***

| Atributo | Motivo de exclusão |
|---|---|
| Dt_Nascimento | Existência de atributo para análise da idade (**Idade_Intern**). |
| País | 99,96% dos casos apontam como país de residência Portugal, enquanto os restantes 0,02% são países PALOP e 0,02% valores omissos. |
| Localidade | Moderado número (28,6%) de valores omissos = 7568/26462, onde 7568 é o número de *missings* e 26462 é a dimensão da amostra. |
| 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. |
| Dt_Alta_Intern | Existência de atributo para análise da hora e número de dias do internamento (**Hora_Alta_Intern** e **N_Dias_Intern**). |
| N_Med_Alta | Baixo número (19,1%) de valores omissos. Elevado número de níveis (156), código único de identificação de cada médico. O atributo **Espec_Medica** irá permitir efetuar agrupamentos por especialidades médicas associadas ao episódio de internamento. |
| Tratamento | Existência de atributo redundante para análise do procedimento principal (**Proc_Principal**). Elevado número (56%) de valores omissos. Substituição dos valores NA do atributo **Proc_Principal** pelos valores correspondentes do atributo Tratamento. |
| GDH | Moderado número (29,6%) de valores omissos. Elevado número de níveis (469), código único de classificação de doentes internados. O atributo **GCD** irá permitir efetuar agrupamentos por grandes categorias de diagnóstico. |

Nota: Dimensão da amostra – 26462 internamentos

Deste modo, e ainda englobando a fase de limpeza dos dados, avançou-se para a etapa de substituição dos valores omissos, com 26431 observações e 17 atributos selecionados.

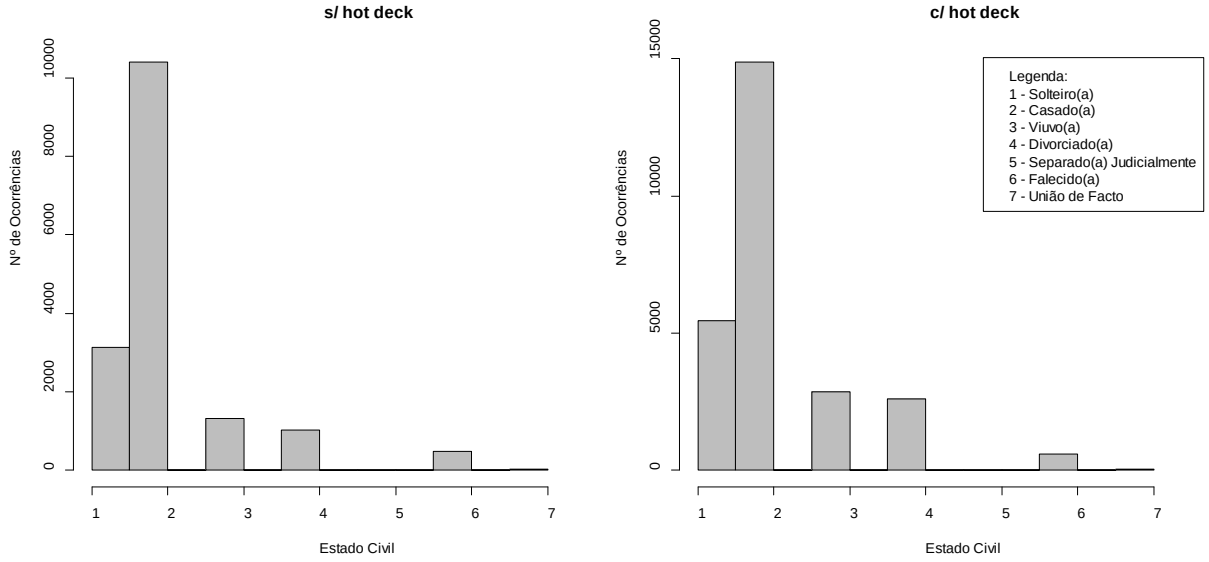
Conforme tema abordado no ponto 2.3, a Tabela 7 sumariza as ações desenvolvidas para tratar o problema dos valores omissos: ignorando os registos com valores omissos ou substituindo-os através de técnicas apropriadas, mais especificamente pela técnica *hot deck* que consiste em procurar o exemplo mais semelhante via *(1-neareast neighbor*) e posterior substituição de valores omissos pelo valor encontrado no exemplo mais próximo.

**Tabela 7: Tratamento dos valores omissos**

| Atributo | Técnica | Observação |
|---|---|---|
| Sexo | *Case deletion* | Eliminação de 12 registos indefinidos (valor N). |
| Escolaridade | *Hot deck* | Substituição de 11771 valores omissos. |
| Est_Civil | *Hot deck* | Substituição de 10046 valores omissos. |
| Proc_Principal | *Hot deck* | Substituição de 19407 valores omissos. |
| Diag_Principal | *Hot deck* | Substituição de 19268 valores omissos. |
| Diag_Inicial | *Hot deck* | Substituição de 16839 valores omissos. |
| GCD | *Hot deck* | Substituição de 7839 valores omissos. |
| Hora_Alta_Intern | *Hot deck* | Substituição de 5 valores omissos. |

O código em R para aplicação da técnica *hot deck* e os gráficos de frequência dos atributos alterados pela técnica referida anteriormente, encontram-se descritos no Anexo C. Como exemplo, é apresentado na Figura 29, o resultado obtido para o atributo Est_Civil. Verifica-se uma alteração do valor das frequências no eixo das ordenadas, concretamente o aumento do número de ocorrências no segundo gráfico.

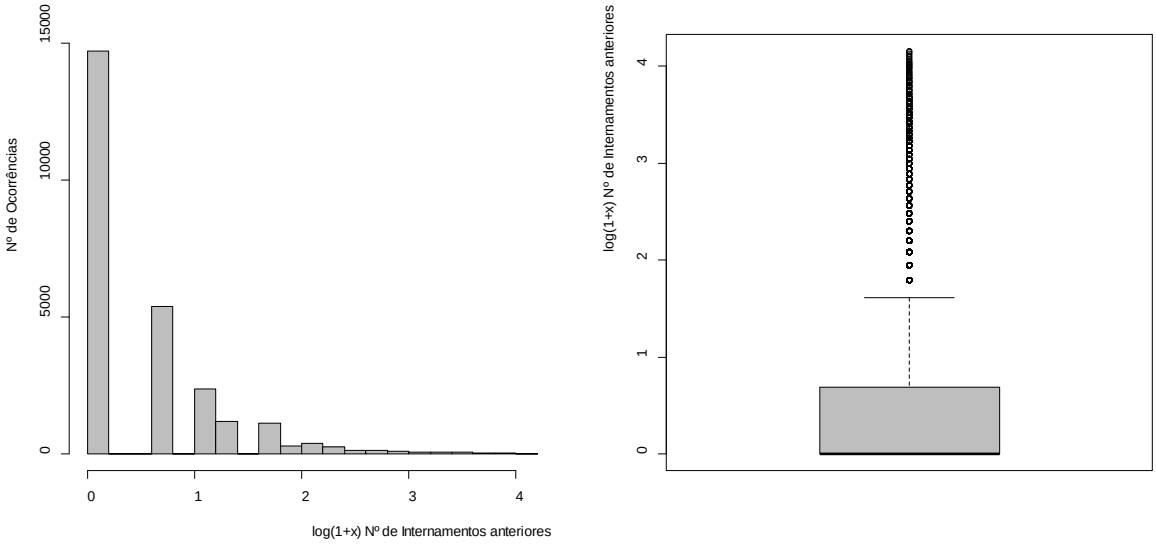
**Figura 29: Diagrama de frequência do atributo Est_Civil - aplicação do *hot deck***




Outra decisão tomada para tratamento do problema da qualidade dos dados foi a aplicação da técnica da transformação nas variáveis N_Intern_Anterior e N_Dias_Intern, concretamente a aplicação da função *log1p (x)* disponível na ferramenta R, que permite calcular o $\log\ (x + 1)$ com precisão.

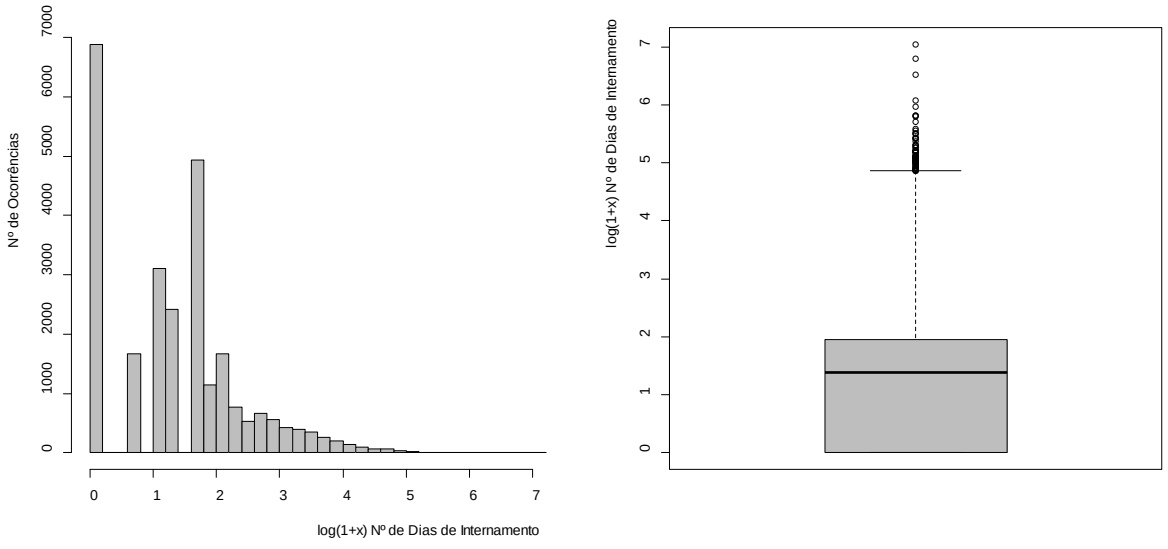
Trata-se de uma transformação que é muito comum quando uma variável quantitativa é muito enviesada para o estremo esquerdo do seu domínio de valores, sendo que esta transformação tende a facilitar a modelação das técnicas de aprendizagem. Com esta ação obteve-se os novos atributos transformados LG_N_Intern_Anterior (Figura 30) e LG_N_Dias_Intern (Figura 31).

**Figura 30: Diagrama de frequência e *boxplot* do atributo LG_N_Intern_Anterior**




O diagrama de frequências da Figura 30 apresenta uma distribuição assimétrica positiva. O diagrama de extremos e quartis apresenta um relativo enviesamento superior, ou seja, os dados estão menos concentrados na parte superior que na inferior. Verifica-se ainda a existência de valores que mesmo acima do intervalo superior, não são considerados valores discrepantes. No caso do atributo LG_N_Dias_Intern, o diagrama de frequência representado na Figura 31 apresenta uma assimetria positiva e o diagrama de extremos e quartis apresenta um enviesamento superior nos extremos e enviesamento inferior na zona central dos dados.

**Figura 31: Diagrama de frequência e *boxplot* do atributo LG_N_Dias_Intern**



Na iteração seguinte, ou seja a subfase construção dos dados, foi incluído um atributo derivado (atributos novos que são construídos de um ou mais atributos existentes no mesmo registo), que apresenta uma boa relevância para o objetivo proposto (Tabela 8).

**Tabela 8: Atributos incluídos no *dataset***

| Atributo | Razão Inclusão |
|---|---|
| DiaSemana_Intern | Nº do dia da semana referente ao internamento do utente, podendo-se verificar os valores 1 (Segunda) a 7 (Domingo). |

No passo seguinte, foram ainda criados valores transformados para os atributos Hora_Intern e Hora_Alta_Intern pois possuíam 746 e 728 níveis respetivamente. O seu formato foi alterado para "HH", obtendo no máximo 24 níveis possíveis.

Segundo a análise estatística da Tabela 28 do Anexo C, pode-se verificar que o atributo Escolaridade apresenta ainda muitos valores possíveis. O atributo foi simplificado de acordo com a Tabela 9 e o código em R para aplicação da função *delevels* é apresentado no Anexo C.

**Tabela 9: Recodificação do atributo Escolaridade**

| Código Antigo | Código Novo | Descrição | Frequências |
|---|---|---|---|
| 100; 200 | 1 | Sem habilitações | 1178 |
| 310 | 2 | Básico (1. Ciclo) | 10191 |
| 320 | 3 | Básico (2. Ciclo) | 3783 |
| 330 | 4 | Básico (3. Ciclo) | 6149 |
| 400; 620; 630; 700 | 5 | Secundário | 1740 |
| 810; 820; 830; 840; 850 | 6 | Superior | 3378 |

Analisando as Tabelas 30, 32 e 34 do Anexo C, pode-se verificar que os atributos Proc_Principal, Diag_Principal e Diag_Inicial apresentam ainda muitos valores possíveis, tornando computacionalmente difícil a criação do modelo na fase seguinte. O atributo Proc_Principal foi simplificado através da função *delevels*, reduzindo o número de niveis disponíveis, tendo-se obtido os valores descritos na Tabela 10. Posteriormente, os atributos Diag_Principal e Diag_Inicial foram simplificados de acordo com a Tabela 11.

**Tabela 10: Recodificação do atributo Proc_Principal**

| Código Antigo | Novo Cód. | Descrição | Freq. |
|---|---|---|---|
| 0 | 0 | Procedimentos e intervenções não classificadas | 0 |
| [01–05] [010–059] | 1 | Operações no sistema nervoso | 1 |
| [06–07] [060–079] | 2 | Operações sobre o sistema endócrino | 0 |
| [08–16] [080–169] | 3 | Operações no olho | 5467 |
| 17 [170–179] [1700–1799] | 3A | Outros procedimentos terapêuticos e de diagnósticos | 0 |
| [18–20] [180–209] [1800–2099] | 4 | Operações na orelha | 249 |
| [21–29] [210–299] [2100–2999] | 5 | Operações no nariz, boca e faringe | 1246 |

| Código Antigo | Novo Cód. | Descrição | Freq. |
|---|---|---|---|
| [30–34][300–349][3000–3499] | 6 | Operações sobre o sistema respiratório | 184 |
| [35–39][350–399][3500–3999] | 7 | Operações sobre o sistema cardiovascular | 649 |
| [40–41][400–419][4000–4199] | 8 | Operações no sistema sanguíneo e linfático | 24 |
| [42–54][420–549][4200–5499] | 9 | Operações sobre o sistema digestivo | 4212 |
| [55–59][550–599][5500–5999] | 10 | Operações no sistema urinário | 1178 |
| [60–64][600–649][6000–6499] | 11 | Operações nos órgãos genitais masculinos | 1708 |
| [65–71][650–719][6500–7199] | 12 | Operações nos órgãos genitais femininos | 889 |
| [72–75][720–759][7200–7599] | 13 | Procedimentos obstétricos | 1 |
| [76–84][760–849][7600–8499] | 14 | Operações no sistema músculo-esquelético | 5297 |
| [85–86][850–869][8500–8699] | 15 | Operações no sistema tegumentar | 2633 |
| [87–99][870–999][8700–9999] | 16 | Procedimentos diagnósticos e terapêuticos diversos | 2681 |

**Tabela 11: Recodificação dos atributos Diag_Principal e Diag_Inicial**

| Código Antigo | Novo Cód. | Descrição | Freq Diag_Principal | Freq Diag_Inicial |
|---|---|---|---|---|
| [001-139] [1000-1399] [10000-13999] | 1 | Doenças infecciosas e parasitárias | 41 | 20590 |
| [140–239] [1400–2399] [14000–23999] | 2 | Neoplasias | 11043 | 421 |
| [240–279] [2400–2799] [24000–27999] | 3 | Doenças endócrinas, nutricionais e metabólicas | 0 | 0 |
| [280–289] [2800–2899] [28000–28999] | 4 | Doenças do sangue e órgãos hematopoiéticos | 0 | 0 |
| [230–319] [2300–3199] [23000–31999] | 5 | Transtornos mentais | 1515 | 100 |
| [320–389] [3200–3899] [32000–38999] | 6 | Doenças do sistema nervoso e órgãos dos sentidos | 1767 | 1845 |
| [390–459] [3900–4599] [39000–45999] | 7 | Doenças do sistema circulatório | 367 | 187 |
| [460–519] [4600–5199] [46000–51999] | 8 | Doenças do sistema respiratório | 2378 | 498 |
| [520–579] [5200–5799] [52000–57999] | 9 | Doenças do sistema digestivo | 829 | 404 |
| [580–629] [5800–6299] [58000–62999] | 10 | Doenças do aparelho geniturinário | 488 | 256 |
| [630–679] [6300–6799] [63000–67999] | 11 | Complicações da gravidez, parto e puérpero | 2189 | 6 |
| [680–709] [6800–7099] [68000–70999] | 12 | Doenças da pele e tecido subcutâneo | 1724 | 551 |
| [710–739] [7100–7399] [71000–73999] | 13 | Doenças do sistema osteomuscular e do tecido conjuntivo | 1724 | 626 |
| [740–759] [7400–7599] [74000–75999] | 14 | Anomalias congênitas | 992 | 39 |
| [760–779] [7600–7799] [76000–77999] | 15 | Certas condições originadas no período perinatal | 0 | 0 |
| [780–799] [7800-7999] [78000-79999] | 16 | Sintomas, sinais e afeções mal definidas | 440 | 24 |
| [800-999] [8000-9999] [80000-99999] | 17 | Lesões e intoxicações | 207 | 120 |
| [V01-V8999] | 18 | Classificação suplementar de fatores que influenciam estado de saúde e o contato com os serviços de saúde | 2433 | 752 |
| [E800-E999] | 19 | Classificação suplementar de causas externas de lesões e envenenamentos | 0 | 0 |

O último atributo a ser transformado foi a Idade_Intern, pois apresentava ainda muitos valores possíveis, conforme informação estatística da Tabela 35 do Anexo C. O atributo Idade_Intern foi simplificado em grandes grupos etários de acordo com a Tabela 12.

**Tabela 12: Recodificação do atributo Idade_Intern**

| Código Antigo | Novo Código | Descrição | Frequências |
|---|---|---|---|
| [0 – 14] | 1 | <15 Anos | 534 |
| [15 – 44] | 2 | 15 – 44 Anos | 7518 |
| [45 – 64] | 3 | 45 – 64 Anos | 8155 |
| [65 – 84] | 4 | 65 – 84 Anos | 9186 |
| [85 – 150] | 5 | ≥85 Anos | 1026 |

Analisando todo o trabalho realizado nesta fase, é importante realçar alguns procedimentos efetuados para se obter uma boa modelação dos dados:

- Valores discrepantes (*outliers*) – Foram efetuados diversos diagramas de frequências e diagramas de extremos e quartis para análise da distribuição dos dados e identificação de valores discrepantes. Em alguns casos, os valores discrepantes resultaram da má introdução dos dados, como por exemplo a identificação de um utente com 207 anos ou o registo de 29 episódios de internamento associados ao código de serviço 9, sendo este um serviço virtual para testes aplicacionais. Em todos estes casos os dados foram excluídos da análise. Noutras situações os valores discrepantes foram considerados valores possíveis para o atributo em causa, optou-se pela sua inclusão na análise , sabendo que pode prejudicar a qualidade de alguns modelos que não lidam bem com os valores discrepantes.
- Atributos – Exclusão da análise de atributos redundantes, que apresentam pouca relevância, que possuem um elevado número de níveis ou um elevado número de valores omissos.
- Valores omissos – Para a identificação dos valores omissos procedeu-se à análise de frequências dos atributos em estudo e foram identificados vários atributos com valores omissos, como o caso do Sexo, Escolaridade e Est_Civil. Em todos os casos procedeu-se à substituição dos valores omissos através da técnica *hot deck*, à exceção da aplicação da técnica *case deletion* no atributo Sexo.
- Transformação – Aplicação de técnica de transformação em *log* nos atributos N_Intern_Anterior e N_Dias_Intern.
- Níveis – Diminuição dos níveis e recodificação dos atributos através da função *delevels*, como o caso do atributo Escolaridade.

Deste modo, avançou-se para a fase de modelação, com um conjunto de 19 atributos e 26419 registos.

### 4.2.4. Modelação

Esta fase trata da seleção e aplicação de técnicas para obtenção de modelos, ou seja, são selecionadas e aplicadas várias técnicas de modelação e os seus parâmetros são ajustados de forma a otimizar os resultados (Chapman et al., 2000). Neste estudo será utilizada uma abordagem de regressão, sendo que para o efeito foram testadas diferentes técnicas e métricas de regressão.

Na primeira iteração selecionou-se as várias técnicas de modelação, para posterior aplicação sobre o conjunto de dados. Neste estudo, foram testadas cinco técnicas de regressão: *Naive*, MR, DT, ANN, RF e SVM, apresentando-se de seguida uma pequena justificação para a sua escolha.

Conforme referido no ponto 2.3, as DT são um dos modelos mais utilizados em DM, apresentando como uma das principais vantagens a facilidade de compreensão e interpretação dos seus modelos pelo ser humano. As ANN apresentam como principais vantagens a não linearidade, flexibilidade e usabilidade. No entanto, são modelos de difícil compreensão e interpretação pelo ser humano, sendo mesmo conhecidas como *black box*. As SVM são adaptadas à regressão via função de custo. Por fim, a MR é um modelo que exige pouco processamento e de fácil interpretação pelo ser humano.

Segundo a tarefa de gestão de um projeto de testes, serão criados procedimentos para testar a qualidade e validade do modelo. Deste modo optou-se por um dos métodos utilizados para estimar a qualidade e desempenho de um modelo é o método de validação *holdout*.

O *holdout* permite dividir aleatoriamente os dados em dois conjuntos: conjunto de treino (para estimar os parâmetros do modelo (2/3)) e o conjunto de teste (para avaliar a precisão do modelo (1/3)).

Outro método utilizado é o *k-fold* com funcionamento semelhante ao anterior, mas 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. Um dos métodos mais conhecidos é o *kfold* (k=5) e em cada rotação é treinado um modelo, sendo que a estimativa global do modelo é dada pelo erro médio do teste das k rotações.

O ambiente de programação escolhido para esta fase da modelação foi o ambiente R e a biblioteca *open source rminer* (Cortez, 2010). Devido à complexidade dos dados, esta abordagem de regressão iniciou-se com a exploração de modelos de previsão simples, como o caso da *Naive*, MR e DT. São técnicas simples devido às menores exigências computacionais, de modo a ganhar confiança para outros modelos mais complexos e validações mais robustas. A Tabela 13 descreve o modo como os modelos foram obtidos através da função *mining*, inicialmente utilizada para ajustes e previsões simples e uma execução da técnica.

**Tabela 13: Código para obtenção e teste do modelo *multiple regression* (MR)**

| *Multiple Regression* (MR) |
|---|
| library(rminer)<br>d<-read.table("internamento.csv", header=TRUE,sep=",")<br>M=**mining**(LG_N_Dias_Intern~.,data=d,Runs=1,method=c("holdout",2/3),model="mr")<br>**savemining**(M,"internamento_mr1.model")<br>print(mmetric(M,metric=c("R2","MAE","RMSE"),aggregate="no")) |

Da biblioteca *rminer* destacam-se as funções *mining* e *savemining*, sendo *mining* uma função poderosa que treina e testa um modelo ajustado às diversas execuções e métodos de validação. Resumidamente, 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 seguido pelo código acima descrito assume o procedimento *holdout* com 2/3 para treino e 1/3 para testes. O modelo é o resultado deste processo moroso e toda a informação e resultados obtidos são guardados pela função *savemining* para posterior análise.

Após a avaliação das experiencias iniciais demonstradas na Tabela 15, avançou-se para um maior número de execuções de forma a efetuar uma validação cruzada, robusta ao nível da confiança dos valores de erros médios apresentados.

A etapa seguinte consistia em experimentar modelos mais complexos com vinte execuções da técnica, como o caso da ANN, SVM e RF, obtendo-se os resultados expressos na Tabela 16.

Por fim, e após a execução de todas as iterações anteriores, foi implementado o método de validação cruzada *5-fold* para estimar a capacidade de generalização dos modelos. Devido à divisão aleatória para definir os 5 subconjuntos, serão aplicadas 20 execuções a cada procedimento *5-fold*, perfazendo o total de 20 x 5 = 100 experiências para cada teste. O código escrito em R para aplicação dos vários métodos existentes na biblioteca *rminer* encontra-se descrito no Anexo D.

### 4.2.5. Avaliação

Na fase de avaliação pretende-se avaliar o modelo construído, verificando o seu comportamento em ambiente de teste de dados assegurando que cumpre os objetivos de negócio (Chapman et al., 2000).

Esta fase está dividida na subfase de avaliação do modelo respetivamente com o grau de conhecimento dos objetivos de negócio e revisão do processo de DM de modo a verificar se ocorreu a omissão de algum fator ou tarefa importante. No final desta fase é tomada a decisão do uso dos resultados obtidos.

O modelo eleito deverá ser o que melhor generalize os dados treinados e o que melhor se identifique na aprendizagem de novos casos, os quais fazem parte do conjunto de teste (Silva, 2010). Conforme abordado no ponto 2.3, para avaliar os modelos produzidos foram selecionadas as métricas de regressão *R2, MAE* e *RMSE*, descritas na Tabela 14. A função *mmetric* pertencente à biblioteca *rminer* permite não só calcular as métricas de regressão, mas também avaliar a qualidade dos modelos. Todo o código executado em R para obtenção dos gráficos e resultados para análise estão descriminados com maior detalhe no Anexo E.

**Tabela 14: Descrição das métricas de regressão**

| Métrica | Descrição | Avaliação |
|---|---|---|
| R2 | *Coefficient of determination* | "Melhor " se valor superior, regressão, ]-Inf,1] |
| MAE | *Mean absolute error* | "Melhor" se valor inferior, regressão, [0,Inf[ |
| RMSE | *Root mean squared error* | "Melhor" se valor inferior, regressão, [0,Inf[ |

De referir que foi possível obter variáveis estatísticas viáveis, concretamente os valores médios e o desvio padrão, pelo que os resultados obtidos serão analisados no Capitulo 5.

### 4.2.6. Implementação

Este trabalho de investigação torna-se relevante pois permitirá o desenvolvimento de novos sistemas de cálculo de tempos de internamento com base em modelos gerados a partir de técnicas de DM. No entanto, os prazos estabelecidos para esta investigação não permitiram a execução desta fase.

# 5. Resultados e sua Discussão

Conforme referido no ponto 4.2.4, a experiência iniciou-se com uma execução da técnica, tendo-se obtido os resultados referidos na Tabela 15. Em todas as tabelas apresentadas neste capítulo e nos próximos, o modelo "Naive" do *rminer* representa a previsão da média dos valores da variável de saída nos exemplos de treino e o termo "ANN" designa o modelo "MLPE" do *rminer*.

**Tabela 15: Métricas obtidas dos testes de validação *holdout* e *runs=1***

| | | Métricas | | |
|---|---|---|---|---|
| Parâmetro | Modelo | R2 | MAE | RMSE |
| method=c("holdout",2/3), Runs=1 | Naive | -4,849e-05 | 0,861 | 1,085 |
| | MR | 0,649 | 0,440 | 0,638 |
| | DT | 0,614 | 0,430 | 0,674 |
| method=c("holdout",2/3) search="heuristic10", Runs=1 | ANN | 0,744 | 0,334 | 0,553 |
| | RF | **0,826** | **0,214** | **0,454** |
| | SVM | 0,745 | 0,297 | 0,545 |

Apesar de se tratar dos primeiros modelos, é de realçar o bom desempenho do modelo RF que obteve o resultado de 0,826 para a métrica R2.

A métrica R2 permite medir o grau de correlação entre as previsões e valores observados. Quanto mais próximos os valores de correlação estiverem do valor 1, melhor é a regressão.

Para se comparar dois modelos distintos torna-se necessário executar o maior número de simulações possíveis, para posteriormente se comparar as médias resultantes de cada algoritmo. Desse modo, efetuaram-se vinte execuções da técnica com a aplicação do método *holdout*, obtendo-se os resultados expressos na Tabela 16.

**Tabela 16: Métricas obtidas dos testes de validação *holdout* e *runs=20***

| | | Métricas | | |
|---|---|---|---|---|
| Parâmetro | Modelo | R2 | MAE | RMSE |
| method=c("holdout",2/3), Runs=20 | Naive | -0,056 ± 0,002 | 0,888 ± 0,003 | 1,111 ± 0,004 |
| | MR | 0,647 ± 0,003 | 0,442 ± 0,002 | 0,644 ± 0,003 |
| | DT | 0,619 ± 0,004 | 0,416 ± 0,004 | 0,67 ± 0,004 |
| method=c("holdout",2/3) search="heuristic10", Runs=20 | ANN | 0,741 ± 0,005 | 0,335 ± 0,004 | 0,554 ± 0,006 |
| | RF | **0,825 ± 0,003** | **0,215 ± 0,002** | **0,453 ± 0,005** |
| | SVM | 0,738 ± 0,003 | 0,305 ± 0,002 | 0,554 ± 0,003 |

Através da função *meanint* pertencente à biblioteca *rminer*, foi possível calcular a média e intervalo de confiança para as vinte execuções. Numa primeira análise, e verificados os

valores das métricas, o modelo RF obteve melhores resultados que os modelos ANN e SVM. Na fase final da avaliação aplicou-se o método *k-fold* (k=5) para todos os modelos, obtendo-se os resultados dos valores de erro médio para cada modelo expressos na Tabela 17.

**Tabela 17: Métricas obtidas dos testes de validação *k-fold* (k=5) e *runs*=20**

| | | Métricas | | |
|---|---|---|---|---|
| Parâmetro | Modelo | R2 | MAE | RMSE |
| method=c("kfold",5), Runs=20 | Naive | 0,000± 0,000 | 0,861 ± 0,000 | 1,085 ± 0,000 |
| | MR | 0,646 ± 0,000 | 0,442 ± 0,000 | 0,645 ± 0,000 |
| | DT | 0,622 ± 0,001 | 0,415 ± 0,001 | 0,667 ± 0,001 |
| method= c("kfold",5) search="heuristic10", Runs=20 | ANN | 0,742 ± 0,001 | 0,334 ± 0,001 | 0,551 ± 0,001 |
| | RF | **0,830± 0,000** | **0,209 ± 0,000** | **0,448 ± 0,001** |
| | SVM | 0,741 ± 0,000 | 0,301 ± 0,000 | 0,552 ± 0,000 |

Verifica-se na Tabela 17 que a métrica R2 apresenta os melhores resultados para os três últimos modelos (ANN, RF e SVM). Confirma-se que o resultado da métrica R2 do método *k-fold*, aplicado ao modelo RF (0,830) é superior ao resultado apresentado pelo método *holdout* para a correspondente métrica e modelo RF (0,825). Sendo o valor de R2 superior a 0,8 pode-se afirmar que se conseguiu gerar um bom modelo.

Posteriormente foi implementado o *student test (t-test)* de modo a verificar se as médias dos resultados de erro de dois modelos distintos são efetivamente diferentes entre si.

O *t-test* é um teste de hipótese que usa conceitos estatísticos para rejeitar ou não uma hipótese nula. A probabilidade de erro, isto é, a probabilidade do intervalo de confiança não conter o desconhecido valor do parâmetro denomina-se de *p-value*. Consequentemente, o nível de confiança é igual a 1-*p-value* e corresponde à probabilidade de o intervalo de confiança conter o valor do parâmetro, sendo 0,90 (90%), 0,95 (95%) e 0,99 (99%) os valores mais usuais.

Sendo este um teste paramétrico (variáveis quantitativas), apresenta como pressuposto a independência das amostras, a homogeneidade das variâncias e a distribuição normal das variáveis.

Neste teste bilateral, a hipótese nula defende que os dois modelos têm média igual e a hipótese alternativa de que as médias são diferentes. Estando os pressupostos verificados, as hipóteses são:

**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.

Neste trabalho, a confiança estatística foi obtida através da função *t-teste* com um intervalo de confiança de 0,95. O código em R para obtenção do *t-test* encontra-se descrito na Tabela 45 do Anexo E.

- $t_{(32,395)}$ = 388,296
- p = 5.403861e-61
- p ≤ 0,05 então rejeitar H0

A hipótese nula (H0) é rejeitada com um nível de confiança de 95% visto que existem evidências estatísticas para se afirmar que a diferença do resultado das médias dos modelos RF e SVM é diferente de 0 ($t_{(32,395)}$ = 388,296; p = 5.403861e-61).

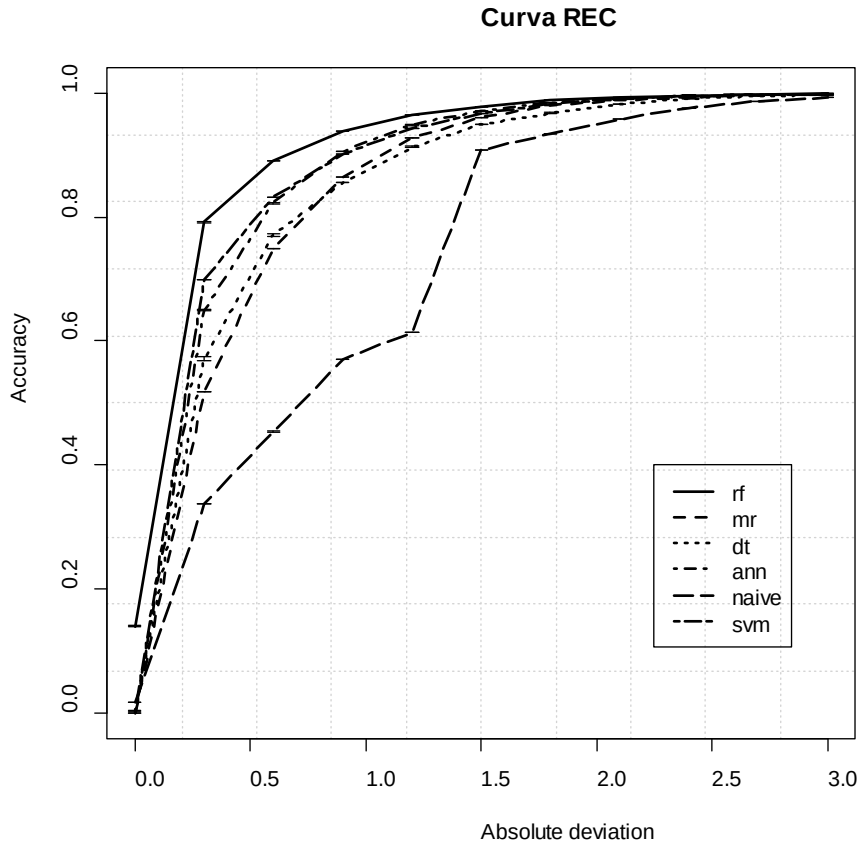
A rejeição da H0 significa que existem diferenças significativas entre as duas médias. Conclui-se que a média dos resultados obtidos pelo modelo RF é significativamente superior à média do modelo SVM.

Outra ferramenta utilizada para efetuar a comparação e avaliação dos resultados dos diversos modelos distintos é a análise via "*regression error characteristics* (REC)". A função *mgraph* da biblioteca *rminer* permite obter a curva REC conforme o gráfico expresso na Figura 32, apresentando-se de seguida o excerto de código executado para a sua obtenção:

***Regression Error Characteristic curve***

```
NV = loadmining ("internamento_naive20_5.model")
MR = loadmining ("internamento_mr20_5.model")
DT = loadmining ("internamento_dt20_5.model")
ANN = loadmining ("internamento_mlpe20_5.model")
RF = loadmining ("internamento_randomforest20_5.model")
SVM = loadmining ("internamento_svm20_5.model")

L = vector ("list",6); L[[1]] = RF; L[[2]] = MR; L[[3]] = DT; L[[4]] = ANN; L[[5]] = NV; L[[6]] = SVM;

Mgraph (L, graph = "REC", xva l= 3, leg = list (pos = c(2.25, 0.4), leg = c("rf", "mr", "dt", "ann", "naive",
"svm")), Grid = 10,  main = "Curva REC")
```

**Figura 32: Gráfico da curva REC para os modelos gerados**




Conforme abordado no ponto 2.3, as curvas REC mostram a taxa de acerto global no eixo das ordenadas, para diversos valores de tolerância de erro absoluto no eixo das abcissas. A precisão, ou taxa de acertos, é definida como a percentagem de pontos que se encaixam dentro da tolerância.

À exceção do modelo *Naive*, todos os modelos apresentam uma capacidade de previsão bastante boa, destacando-se como melhor modelo o RF, com uma curva bastante regular e superior à dos restantes modelos, sem propriamente pontos acentuados de mudança de comportamento. Para o estudo do caso torna-se importante definir um valor de tolerância de erro absoluto, verificado no eixo das ordenadas de uma curva REC.

A função *mmetric* possui uma métrica com a designação *normalized rec area* (NAREC) que nos devolve o valor preciso e a *tolerance* que permite obter o valor exato de percentagem de acertos mediante um valor de tolerância. Se o valor de tolerância for de 0,5 a taxa de acerto para o modelo RF será de 0,867 e a área normalizada apresentada pela curva REC é de 0,725, conforme os seguintes cálculos efetuados:

**Tolerância de 0,5 da curva REC**

```
M = loadmining ("internamento_randomforest20_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 = 0,867 ± 0,000**
**NAREC = 0,725 ±0,001**

Se a tolerância diminuir para 0,25 a taxa de acerto terá um valor inferior, mais concretamente o valor de 0,763 e a NAREC apresenta o valor 0,629. Concluindo, para o modelo RF e com uma tolerância de 0,5 consegue-se prever acertadamente 87% dos casos e com uma tolerância de 0,25 obtêm-se uma taxa de acertos de 76%.

De modo a se avaliar a qualidade dos resultados obtidos pelo melhor modelo (RF), recorreu-se ao gráfico de dispersão *regression scatter characteristics* (RSC), em que nos eixos das abcissas está representado os valores observados e no eixo das ordenadas os valores das previsões. Conforme excerto de código executado para a sua obtenção, implementou-se a função *mgraph* com o parâmetro *graph="RSC"* para obtenção dos gráficos descritos nas Figuras 33 e 34.

***Regression Scatter Characteristic curve***

```
T = 0.5 # valor para a tolerância a admitir


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 =
                "RSC: Gráfico de Dispersão")
        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 = "black", pch = 19, cex = 0.7)
        }
 }
```

**Figura 33: RSC: Gráfico dispersão do modelo RF – Tolerância 0,5**

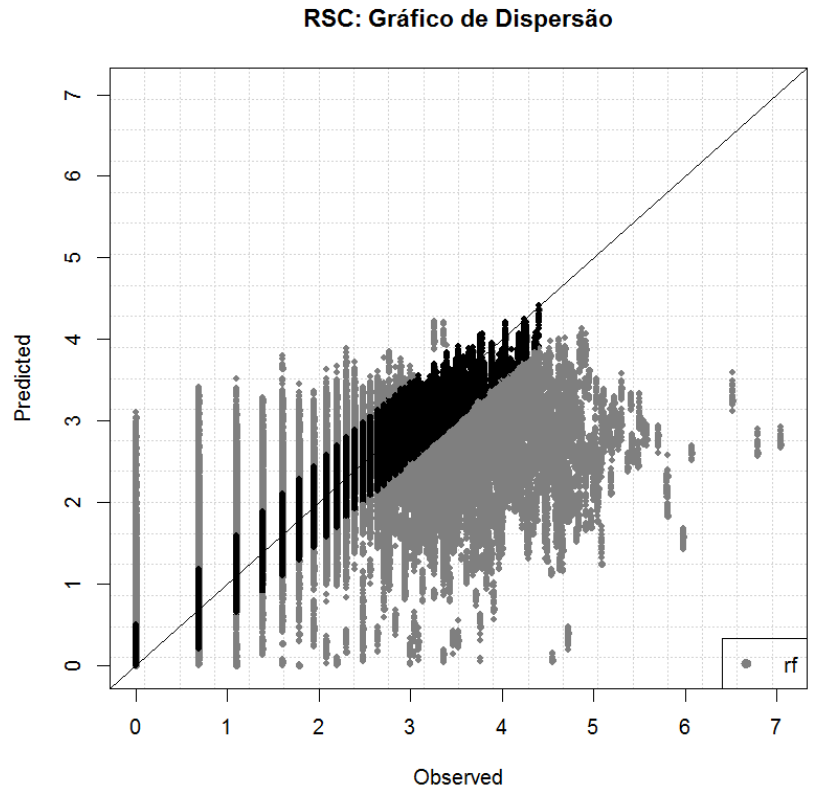
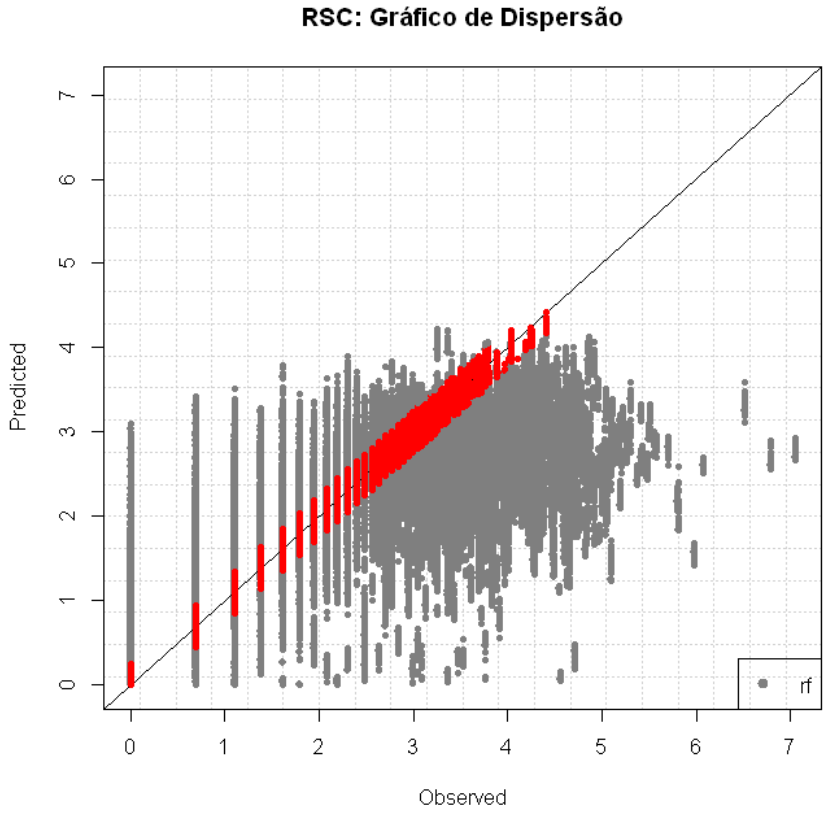



**Figura 34: RSC: Gráfico dispersão do modelo RF – Tolerância 0,25**



Nos gráficos demonstrados nas Figuras 33 e 34, constata-se que a maioria dos pontos se situa próximo da diagonal, pelo que quanto mais próximos os pontos da diagonal melhor o modelo de previsão. Os pontos com tolerância de 0,5 estão representados no gráfico pela cor preta e os pontos com tolerância 0,25 estão representados no gráfico pela cor vermelha. Verifica-se que no extremo do zero o erro máximo é obtido seguindo a linha vertical dos pontos negros/vermelhos, com tendência a subir em relação ao observado, enquanto no extremo superior os pontos negros/vermelhos vão até ao máximo, com tendência a descer em relação ao observado. Com base na informação descrita nas Figuras 33 e 34 e de modo a obter o erro máximo, apresenta-se um excerto dos cálculos para previsão dos erros.

***Erro Máximo nos Extremos Inferior e Superior***

T = 0.5 # valor para a tolerância a admitir

A – Calculo para o extremo inferior
B – Calculo para o extremo superior

As previsões são sobre uma escala logarítmica
y = log (x+1)
Função inversa:
x = exp (y)-1

**A:**
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:**
4.5 = log (x1+1)
4.0 = log (x2+1)

Valor mais alto é 4.5 e a tolerância é 0.5
log (x1+1) = 4.5 => aplicando exp (4.5)-1 => x1 = 89.01713
log (x2+1) = 4.0 => aplicando exp (4.0)-1 => x2 = 53.59815

x1 - x2 = 35.41898 => desvio em dias normais para B

Se para uma tolerância de 0,5 o modelo acerta em cerca de 87% dos tempos de internamento, tal significa que dá um erro máximo de 0,65 dias para o extremo inferior da escala (0) e um erro máximo aproximado de 35 dias no extremo superior da escala (4,5). Para uma tolerância de 0,25 o modelo acerta em cerca de 76% dos tempos de internamento, tal significa que dá um erro máximo de 0,28 dias para o extremo inferior da escala (0) e um erro máximo aproximado de 20 dias no extremo superior da escala (4,5).

Uma vez que se pretende um modelo explicativo para o negócio, pretendeu-se avaliar a importância de cada atributo de entrada no melhor modelo obtido com base em procedimentos de análise de sensibilidade (Cortez e Embrechts, 2013), ou seja, medir a importância relativa das entradas e extrair regras do modelo. Recorreu-se às funções *fit* e *importance* da biblioteca *rminer*, com aplicação na função *importance* do parâmetro *method="DAS"*, conforme exemplo apresentado:

***Relative Input Importance Barplot***

```
M = fit (LG_N_Dias_Intern~., data = d, model = "randomforest",  task = "reg", search = "heuristic10")
d <- read.table ("internamento.csv",  header = TRUE, sep = ",")

Imp = Importance (M, data = d, method = "DSA")
L = list (runs = 1, sen = t (Imp$imp), sresponses = Imp$sresponses)
mgraph (L, graph = "IMP", leg = names (d), col = "gray", Grid = 10)
```

A Figura 35 demonstra os resultados obtidos com a representação da importância dos atributos para a definição do modelo RF e a Tabela 18 apresenta os resultados obtidos para cada atributo.

**Figura 35: Importância dos atributos para a definição do modelo RF**

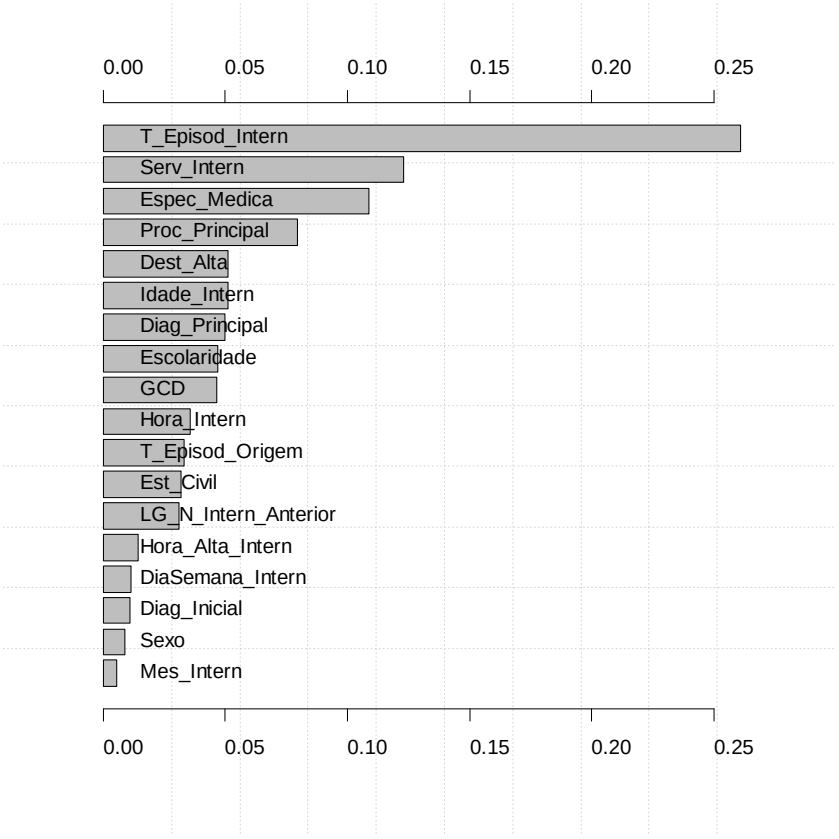


**Tabela 18: Resultados estatísticos da importância dos atributos no modelo RF**

| Atributo | Importância | Atributo | Importância |
|---|---|---|---|
| T_Episod_Intern | 0,261 | Hora_Intern | 0,036 |
| Serv_Intern | 0,123 | T_Episod_Origem | 0,033 |
| Espec_Medica | 0,109 | Est_Civil | 0,032 |
| Proc_Principal | 0,080 | LG_N_Intern_Anterior | 0,031 |
| Dest_Alta | 0,051 | Hora_Alta_Intern | 0,014 |
| Idade_Intern | 0,051 | DiaSemana_Intern | 0,011 |
| Diag_Principal | 0,050 | Diag_Inicial | 0,011 |
| Escolaridade | 0,047 | Sexo | 0,009 |
| GCD | 0,046 | Mes_Intern | 0,005 |

Pela análise gráfica da Figura 35 e pelos resultados estatísticos da Tabela 18, o atributo T_Episod_Intern destaca-se pela sua importância na definição do modelo, explicando 26,1% do modelo gerado com o objetivo de prever o número de dias de internamento. No entanto, outros dois atributos apresentam alguma importância, nomeadamente o atributo Serv_Intern (12,3%) e o atributo Espec_Medica (10,9%).

Na fase seguinte, procedeu-se a uma análise mais detalhada da influência dos valores de entrada mais importante na construção do modelo RF através da curva *variable effect curve* (VEC), proposta em (Cortez e Embrechts, 2013), conforme exemplo apresentado.

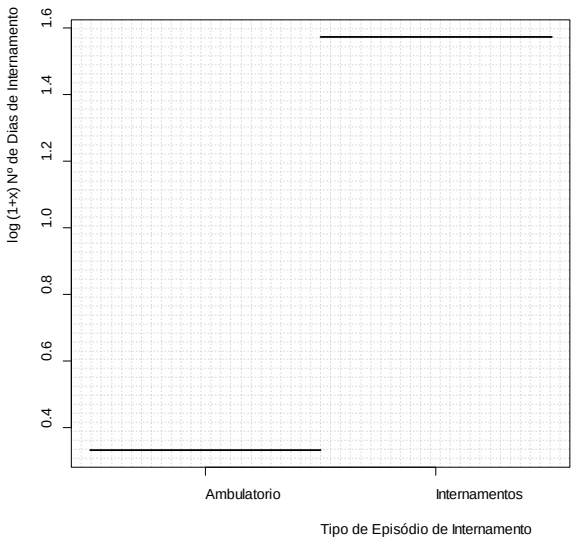
***Variable Effect Curve***

```
M=fit(LG_N_Dias_Intern~.,data=d,model="randomforest", task="reg", search="heuristic10")
d<-read.table("internamento.csv", header=TRUE,sep=",")

Imp=Importance (M, data=d, method="DSA")

vecplot(Imp,graph="VEC",xval=6,Grid=50,main="", xlab="Tipo de Episódio de Internamento", ylab="log (1+x) Nº de Dias de Internamento")
vecplot(Imp,graph="VEC",xval=7,Grid=50,main="", xlab="Serviço de Internamento", ylab="log (1+x) Nº de Dias de Internamento")
vecplot(Imp,graph="VEC",xval=8,Grid=50,main="", xlab="Especialidade Médica", ylab="log (1+x) Nº de Dias de Internamento")
```

A curva VEC representa graficamente os valores de $x_a$ (eixo das abcissas) em compararação com os resultados (eixo das ordenadas). A função *mgraph* permite por exemplo obter a curva VEC para os três atributos mais importantes na construção do modelo.

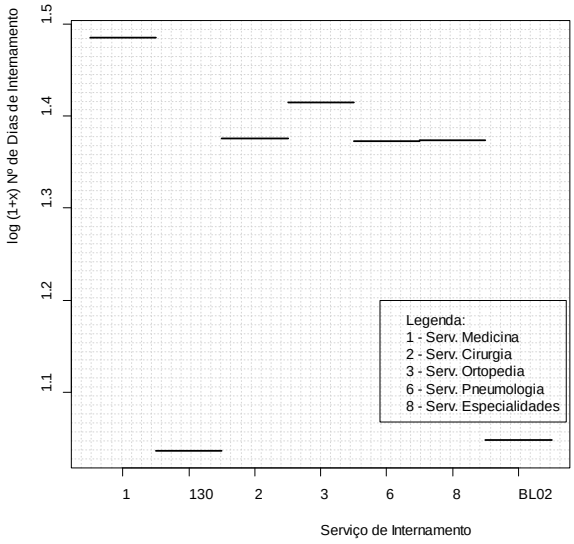
Na Figura 36 verifica-se que um episódio de internamento em regime de ambulatório corresponde a um número de dias baixo (0,2 na escala de transformação logarítmica), enquanto se for o caso de um regime de internamento o tempo de estadia no hospital aumenta (1,57 na escala transformada). De realçar que este tipo de resultado faz todo o sentido.

**Figura 36: Influência do tipo de episódio de internamento**




Relativamente ao atributo caraterizador do serviço de internamento, pode-se observar na Figura 37 um maior número de dias de estadia no internamento do serviço 1 (medicina) com um valor estimado de 1,49 na escala de transformação logarítmica, seguindo-se o serviço 3 (ortopedia) com um valor estimado de 1,42 e por último os serviços 2 (cirurgia), 8 (especialidades) e 6 (pneumologia) com um valor de 1,38.

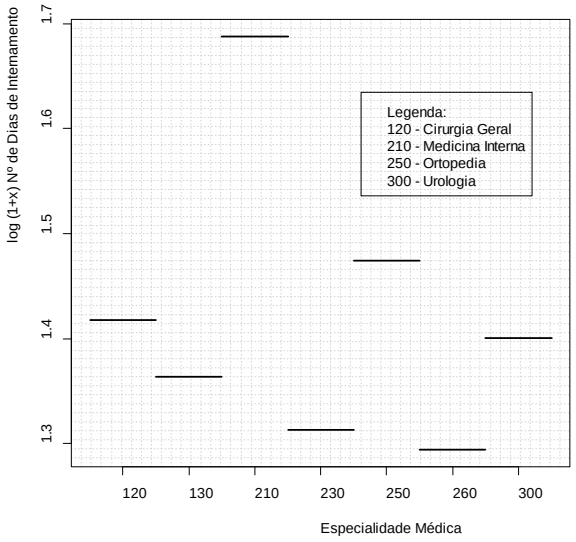
**Figura 37: Influência do serviço de internamento**




Na Figura 38 verifica-se que na escala de transformação logarítmica a especialidade médica com o código 210 (medicina interna) representa 1,69 dias de estadia no internamento,

seguida da especialidade 250 (ortopedia) com um valor aproximado de 1,48, a especialidade 120 (cirurgia geral) com 1,42 e por fim a especialidade 300 (urologia) com 1,40.

**Figura 38: Influência da especialidade médica**




Deste modo, a função *importance* possibilitou a compreensão do peso de cada atributo na construção do melhor modelo e a curva VEC permitiu entender a influência de cada atributo utilizado no modelo.

Confirma-se ainda que existe uma relação dos atributos mais influentes desta investigação e os atributos existentes em outros estudos relacionados. Apresenta-se o caso do atributo T_Episod_Intern abordado por Castillo (2012) e Freitas et al. (2012), o atributo Serv_Intern referenciado por Castillo (2012), o atributo Espec_Medica mencionado por Azari et al. (2012) e Sheikh-Nia (2012) e por último o atributo Proc_Principal abordado por Abelha et al. (2007), Aranha et al. (2009) e Castillo (2012).

# 6. Conclusões

## 6.1. Síntese do Trabalho Efetuado

Para este trabalho de investigação estabeleceu-se como objetivo principal a definição de um modelo preditivo para tempos de internamento de pacientes através da descoberta de comportamentos e padrões existentes no processo de internamento hospitalar, com base em técnicas de DM. A previsão de tempos de permanência nos serviços de internamento permitirá optimizar o número de camas livres e de recursos humanos nos serviços de internamento. Com a previsão do número de dias de internamento, pretende-se resolver o problema da lista de espera e reduzir o custo associado ao processo de internamento dos pacientes.

Dada a natureza dos atributos a modelar, optou-se por definir o objetivo de DM como sendo de regressão. Na fase de modelação foram utilizadas as seguintes técnicas: *Naive,* MR, DT, ANN, RF e SVM

Conforme demonstrado no ponto 4.2.5, os resultados obtidos apontam a RF (R2 de 0,830) como a técnica que no geral consegue obter os melhores resultados, seguida pela ANN (R2 de 0,742) e por fim pela SVM (R2 de 0,741). Confirma-se que com o resultado da métrica R2 aplicado ao modelo RF (superior a 0,8), pode-se afirmar que se conseguiu gerar um bom modelo. A SVM não deverá ser solução devido ao elevado esforço computacional verificado.

De forma a comparar os modelos e a considerar a RF efetivamente melhor que as restantes, procedeu-se à análise da curva REC, que demonstrou a elevada qualidade do modelo RF extraído. Todos os modelos apresentam uma capacidade de previsão bastante boa, destacando-se como melhor modelo o RF, com uma curva bastante regular e superior à dos restantes modelos e sem propriamente pontos acentuados de mudança de comportamento. Para o modelo RF com uma tolerância de 0,5 consegue-se prever aproximadamente 87% dos casos, e com uma tolerância de 0,25 consegue-se prever aproximadamente 76% dos casos.

Uma vez que se pretende um modelo explicativo para o negócio, pretendeu-se avaliar a importância de cada atributo de entrada no melhor modelo obtido com base em procedimentos de análise de sensibilidade, ou seja, medir a importância relativa das entradas e extrair regras do modelo. O atributo T_Episod_Intern destaca-se pela sua importância na definição do modelo, explicando 26,1% do modelo gerado com o objetivo de prever o número

de dias de internamento, seguido por outros dois atributos, nomeadamente o atributo Serv_Intern (12,3%) e o atributo Espec_Medica (10,9%).

Através da curva VEC, procedeu-se a uma análise mais detalhada da influência dos valores de entrada mais importante na construção do modelo RF. Um episódio de internamento em regime de internamento corresponde um tempo de estadia médio de 1,57 (na escala da transformação logarítmica). Relativamente ao atributo caraterizador do serviço de internamento, verificou-se um maior número de dias de estadia no internamento no serviço 1 (medicina) com um valor estimado de 1,49 (na escala de transformação logarítmica), seguindo-se o serviço 3 (ortopedia) com um valor estimado de 1,42 e por último os serviços 2 (cirurgia), 8 (especialidades) e 6 (pneumologia) com um valor de 1,38 dias. Por fim, a especialidade médica com o código 210 (medicina interna) representa 1,69 dias de estadia no internamento (na escala de transformação logarítmica), seguida da especialidade 250 (ortopedia) com um valor aproximado de 1,48, a especialidade 120 (cirurgia geral) com 1,42 e por fim a especialidade 300 (urologia) com 1,40.

No entanto, se a análise for efetuada conforme os valores na escala real, será aplicada a função inversa do $\log(x+1)$, ou seja, a função $e^{y}$- 1.

Deste modo, conclui-se que um episódio de internamento em regime de internamento apresenta um tempo de estadia médio de 3,81 dias. Ao nível dos serviços de internamento, o serviço 1 (medicina) apresenta uma estadia média de 3,44 dias, seguindo-se o serviço 3 (ortopedia) com 3,14 e por último os serviços 2 (cirurgia), 8 (especialidades) e 6 (pneumologia) com um valor de 2,98 dias. Por fim, a especialidade médica com o código 210 (medicina interna) representa 4,42 dias de estadia no internamento, seguida da especialidade 250 (ortopedia) com 3,39, a especialidade 120 (cirurgia geral) com 3,14 e por fim a especialidade 300 (urologia) com 3,10 dias.

## 6.2. Contributos

A conjuntura económica atual é marcada por uma crise financeira sem precedentes, pelo que a necessidade de reduzir o tempo de internamento, aumentar o número de camas disponíveis para novos internamentos e prestar melhores cuidados de saúde, têm sido objetivos difíceis de cumprir no contexto hospitalar. O estudo dos tempos de internamento é importante para a gestão hospitalar dada a sua clara relação com os custos hospitalares, permitindo obter uma melhor eficácia e eficiência. Os custos hospitalares tornaram-se um

fator decisivo, levando as instituições hospitalares a apostarem numa melhoria da eficiência dos seus processos internos e na extração de informação útil para apoiar a tomada de decisão.

É notório que o sistema atual apresenta uma inércia considerável, que torna visível que as suas ações nem sempre são coordenadas de modo a que se possam desenvolver de forma articulada e harmoniosa no tempo.

Esta dissertação tem como propósito o estabelecimento de princípios de otimização de forma a sensibilizar os profissionais que atuam nesta área e oferecer as melhores condições de controlo e bem-estar para os pacientes. Neste sentido, procura-se minimizar as dificuldades do atual sistema hospitalar em Portugal, pois considera-se muito importante que exista um processo metodológico que possa, ao nível da intervenção hospitalar, contribuir para a implementação de um desenvolvimento eficaz.

Esta dissertação aborda o problema do agendamento da admissão hospitalar, em que a maior dificuldade é manter camas suficientes para pacientes oriundos do serviço de urgência e futuros pacientes internados. 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.

A base desta investigação partiu de um estudo de caso referente ao fluxo de trabalho dos internamentos e apresenta como objetivo principal a definição de um modelo preditivo para tempos de internamento de pacientes numa instituição hospitalar através da descoberta de comportamentos e padrões existentes no processo de internamento hospitalar.

Ao nível da literatura existente, efetuou-se um levantamento bastante exaustivo da área de BI, concretamente a componente DM, apontando-se o padrão DM escolhido e as técnicas de regressão utilizadas.

Constatou-se que a ferramenta R e nomeadamente a biblioteca *rminer* foram bastante eficazes em todo o processo na obtenção dos diversos resultados, e sendo estas soluções *open source*. Numa fase inicial também foi usada a ferramenta *rattle*, que permite a um utilizador inexperiente iniciar-se em aplicações semelhantes.

Com a aplicação da metodologia CRISP-DM (não-proprietária e gratuita) obteve-se ao nível da previsão, diversos modelos de qualidade, revelando-se adequada aos objetivos propostos. As suas diversas fases permitiram escalonar as diversas atividades para obtenção do resultado final, desde a definição do objetivo de negócio, aquisição dos dados iniciais, exploração e verificação da qualidade dos dados, seleção dos dados, seleção da técnica de modelação, e por fim construção e avaliação do modelo. Apesar de todas as suas fases serem importantes, conclui-se que se deve dar a devida importância à seleção e tratamento dos

atributos, pois serão estes que permitirão obter o modelo de previsão adequado ao objetivo proposto.

Posteriormente efetuou-se uma comparação dos resultados das diversas técnicas com base em métricas de regressão para aferir os melhores modelos. Constata-se ainda que, todos os modelos apresentam na métrica R2 uma melhoria superior a 60%, quando comparados com a *Naive* (média dos valores da saída). Através de análise de sensibilidade, verificou-se que o processo de internamento dos pacientes deverá ter em conta os três atributos mais influentes na definição do modelo: a seleção adequada do tipo de episódio de internamento a efetuar, determinar assertivamente o serviço físico onde o doente será internado e a especialidade médica associada.

O melhor modelo obtido foi o RF e os seus resultados são úteis para o negócio. Deve pois, o novo processo de previsão de tempos de internamento evitar a possibilidade de que possam ocorrer erros ou desvios no planeamento dos internamentos. Contribuirá para o aumento da qualidade de serviço, gerindo de forma eficiente e simultaneamente promovendo a competitividade dos recursos disponíveis e benefícios desejáveis de alcançar. O modelo gerado no âmbito desta dissertação pode 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.

Convém frisar que esta dissertação diz respeito ao primeiro estudo efetuado sobre a aplicação de técnicas de DM sobre dados do atual HFAR . Nesse sentido, foi necessário aplicar quase todas as fases da metodologia CRISP-DM, desde a compreensão do negócio até a avaliação dos modelos de DM. Neste projeto de dissertação, realça-se aqui o elevado esforço que foi dedicado à coleta e tratamento de dados. Trata-se de um fenómeno que, embora seja muito comum em projetos de DM, dificultou a condução dos trabalhos desta dissertação. Ainda assim, considera-se que o trabalho apresentado é de qualidade e apresenta um potencial de utilidade, conforme aquilo que foi descrito nesta secção.

## 6.3. Limitações

Importa referir que existem algumas limitações nesta investigação, nomeadamente:

- Os testes realizados foram abordados somente segundo o objectivo de regressão, sendo que poderia ser definido em alternativa uma classificação ordinal;

- Foram utilizados os dados disponibilizados por uma instituição hospitalar. O estudo em causa não se pode generalizar para todos os hospitais, pois terá de haver preocupações com as fases de pré-processamento para cada uma das instituições hospitalares, englobando tarefas distintas de seleção de dados, transformação de variáveis, e a substituição dos valores omissos;
- Foram somente utilizadas três métricas de regressão para avaliação das técnicas e dos modelos gerados;
- A não implementação do melhor modelo, de modo a confrontar os resultados do modelo com a realidade;
- Não foram utilizados alguns dos atributos propostos pelo estado da arte.

## 6.4. Trabalho Futuro

Esta investigação proporciona diversas perspetivas de trabalho futuro. Foi possível determinar os atributos que mais influenciaram a definição do modelo extraído, sendo importante indicar possíveis futuros caminhos na área da previsão de tempos de internamento, nomeadamente:

- Implementação de um sistema de apoio à decisão com base no conhecimento adquirido, ou seja, implementação de sistema automático com base nos atributos mais influentes, possibilitando a atribuição automática de cama no serviço de internamento;
- Otimização das técnicas de DM com a implementação de novos parâmetros e aplicação de outras métricas de regressão;
- Uso de atributos relacionados com dados demográficos do utente para definição do modelo, pois o processo de internamento poderá estar por exemplo relacionado com a distância da morada do utente ao hospital;
- Carregamento de variáveis mencionadas pelo painel de especialistas de diversas especialidades médicas do ponto 4.2.2: fatores socioeconómicos, eventos adversos (segurança do doente), nutrição, fragilidade, doenças crónicas associadas, comorbidades e apoio familiar;
- Utilizar menos atributos para a construção do modelo, diminuir o tempo de execução do processo e obter igualmente resultados muito próximos dos verificados nesta investigação;

- Alterar o estudo para um problema de classificação ordinal: a variável alvo número de dias em escalão, podendo ser interessante em termos de planeamento e listas de espera. É necessário perceber junto dos gestores de negócio a sua utilidade, pois é importante definir corretamente os valores de corte e o número de níveis necessários;
- Por fim, estender o estudo a outros hospitais repetindo a experiência como uma outra base de dados de características similares, comparando os resultados dos modelos obtidos;
- Testar novas métricas de regressão que possibilitem a obtenção de melhores resultados;
- Comparação das funcionalidades da ferramenta *open source* com ferramenta proprietária, de forma a avaliar os resultados, vantagens e desvantagens de cada uma.

# Referências Bibliográficas

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.

Aranha, G. T., Vieira, R. W., Oliveira, P. P., Junior, O. P., Benze, B. G., Filho, L. d., et al. (2009). Identificação de um método estatístico como instrumento da qualidade: tempo da presença do doente na sala de operação. *Rev Bras Cir Cardiovasc, 24*(3), pp. 382-390.

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). IEEExplore Digital Library.

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW. *IADIS European Conference Data Mining*, (pp. 182-185).

Bachouch, R. B., Guinet, A., & Hajri-Gabouj, S. (2012). An integer linear model for hospital bed planning. (Elsevier, Ed.) *International Journal of Production Economics, 140*(2), 833-843.

Barrento, M., Neto, M., Martins, M. d., & Dias, S. (2010). Sistemas de Business Intelligence Aplicados à Saúde. In Á. Rocha, & E. U. Pessoa (Ed.), *Sistemas e Tecnologias de Informação na Saúde* (pp. 77-91). Edições Universidade Fernando Pessoa.

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

Bose, I., & Mahapatra, R. K. (2001). Business data mining — a machine learning perspective. (Elsevier, Ed.) *Information & Management, 39*(3), pp. 211-225.

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

Carriço, M. (2012). *Economia*. Obtido em 2013, de Jornal de Negócios: http://www.jornaldenegocios.pt/economia/detalhe/hospitais_ja_reduziram_1000_camas_desde_2011.html

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

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. *CRISP-DM Consortium.* Obtido de http://www.crisp-dm.org

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

Corrêa, A. M., & Sferra, H. H. (2003). Conceitos e Aplicações de Data Mining. *Revista de Ciencia & Tecnologia, 11*(12), 19-34.

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, 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. (Elsevier, Ed.) *Information Sciences, 225*, pp. 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.

Cunha, N. M. (2009). Metodologia de Selecção de Segmentações Diversificadas: Um Caso de Aplicação de Técnicas de Data Mining em Dados de Consumo para Avaliação de Portfólios de Cartões Bancários. *Dissertação de Mestrado.* Faculdade de Economia da Universidade do Porto, Porto.

Delen, D., Coqdell, D., & Kasapc, N. (2012). A comparative analysis of data mining methods in predicting NCAA bowl outcomes. (Elsevier, Ed.) *International Journal of Forecasting, 28*(2), 543-552.

Diário da Republica 1.ª série N.º 158 de 16 de agosto de 2012. (2012). Obtido de http://dre.pt/pdf1sdip/2012/08/15800/0449004492.pdf

Diário da República, 1.ª série — N.º 179 — 15 de Setembro de 2009. (2009). Obtido de http://www.emgfa.pt/documents/4xcs0fzgnm8h.pdf

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence*, 37-54.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). 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.

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*. Esposende.

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.

Gago, P., Santos, M. F., Silva, Á., Cortez, P., Neves, J., & Gomes, L. (2005). INTCare : a knowledge discovery based intelligent decision support system for intensive care medicine. (Lavoisier, Ed.) *Journal of Decision Systems, 14*(3), 241-259.

Gonçalves, D., Santos, M. Y., & Cruz, J. (2010). Implementação de um sistema de Business Intelligence para a análise da qualidade de vida pré e pós-operatória. In Á. Rocha, *Sistemas e Tecnologias de Informação na Saúde* (pp. 93-110). Edições Universidade Fernando Pessoa.

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

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

Kohavi, R., & Provost, F. (1998). Glossary of Terms. *Machine Learning, 30*, 271-274.

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.

Liao, S.-H., Chu, P.-H., & Hsiao, P.-Y. (2012). Data mining techniques and applications – A decade review from 2000 to 2011. (Elsevier, Ed.) *Expert Systems with Applications, 39*(12), 11303-11311.

Liu, G., Xu, Y., & Wu, Z. (2001). Total solution for structural mechanics problems. *Computer Methods in Applied Mechanics and Engineering, 191*(8-10), 989-1012.

Maimon, O., & Rokach, L. (2005a). *Data mining and knowledge discovery handbook.* New York Inc: Springer.

Maimon, O., & Rokach, L. (2005b). Decomposition Methodology For Knowledge Discovery And Data Mining. In H. Bunke, & P. Wang, *Series in Machine Perception and Artificial Intelligence* (Vol. 61). worldscientific.

Marshall, A., Vasilakis, C., & El-Darzi, E. (2005). Length of Stay-Based Patient Flow Models: Recent Developments and Future Directions. (Springer, Ed.) *Health Care Management Science, 8*(3), pp. 213-220.

Mazier, A., Xie, X., & Sarazin, M. (2010). Scheduling Inpatient Admission Under High Demand of Emergency Patients. *6th annual IEEE Conference on Automation Science and Engineering*, (pp. 792-797). Toronto.

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.

Meyfroidt, G., Güizab, F., Ramon, J., & Bruynooghe, M. (2009). Machine learning techniques to examine large patient databases. (Elsevier, Ed.) *Best Practice & Research Clinical Anaesthesiology, 23*(1), 127-143.

Michalewicz, Z., Schmidt, M., Michalewicz, M., & Chiriac, C. (2006). *Adaptive Business Intelligence* (1 ed.). (Springer, 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 Superior de Ciências do Trabalho e da Empresa, Lisboa.

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.

Oliveira, J. P. (2009). Identificação E Caracterização De Situações De "Churn" em Sistemas De Telecomunicações. *Tese de mestrado.* ISCTE – Instituto Superior de Ciências do Trabalho e da Empresa, Lisboa.

Olson, D. L., Delen, D., & Meng, Y. (2012). Machine Comparative analysis of data mining methods for bankruptcy prediction. (Elsevier, Ed.) *Decision Support Systems, 52*(2), pp. 464-473.

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.

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.

Preiss, B. R. (1998). *Data Structures and Algorithms with Object-Oriented Design Patterns in Java.* Waterloo: John Wiley & Sons.

Pyle, D. (1999). *Data Preparation for Data Mining.* Morgan Kaufmann.

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, M. F., & Azevedo, C. S. (2005). Data Mining – Descoberta de Conhecimento em Bases de Dados. FCA Editores.

Santos, M., & Portela, F. (2011). Enabling ubiquitous data mining in intensive care : features selection and data pre-processing. *ICEIS 2011 - Proceedings of the 13th International Conference on Enterprise Information Systems. 1*, pp. 261-266. Beijing: SciTePress.

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.

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.

Stitson, M. O., Weston, J. A., Gammerman, A., Vovk, V., & Vapnik, V. (1996). *Theory of Support Vector Machines.* London: University of London.

Suthummanon, S., & Omachonu, V. K. (2004). DRG-Based Cost Minimization Models: Applications in a Hospital Environment. In R. I. Field, *Health Care Regulation in America : Complexity, Confrontation, and compromise* (Vol. 3, pp. 197-205).

Tsumoto, S., & Hirano, S. (2010). Risk Mining in Medicine: Application of Data Mining to Medical Risk Management. *Fundamenta Informaticae, 98*(1), 107-121.

Venkatadri, M., & Reddy, L. C. (2010). A Comparative Study On Decision Tree Classification Algorithms In Data Mining. *2*(2), pp. 24-29.

Wikipedia. (2013). *Support vector Machine*. Obtido em 20 de 9 de 2013, de Wikipedia: http://en.wikipedia.org/wiki/File:Svm_max_sep_hyperplane_with_margin.png

Williams, T. A., Dobb, G. J., Finn, J. C., & Webb, S. A. (2005). Long-term survival from intensive care: a review. *Intensive Care Med., 31*(10), 1306-1315.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2 ed.). (Elsevier, Ed.) San Francisco: Morgan Kaufmann.

Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. *Circulation*, 654-657.

# Anexos

# A. Compreensão do Negócio

**Tabela 19: Glossário de terminologia de negócio e DM**

| Termo | Descrição |
|---|---|
| *Business Intelligence* | Recorre a técnicas, identificando, extraindo, e analisando dados empresariais. |
| *Data Mining* | Processo de exploração de grandes quantidades de dados à procura de padrões consistentes, através do relacionamento sistemático entre variáveis, detetando assim novos subconjuntos de dados. |
| *Rattle* | Ferramenta que apresenta resumos visuais e estatísticos dos dados. Ferramenta *open-source* assente no ambiente R a qual permite visualizar e transformar dados de forma a facilitar a modelação. |
| *Weka* | Instrumento gráfico que contém diversas ferramentas de visualização e algoritmos para análise dos dados e da modelação. |
| *CRISP-DM 1.0* | Modelo de processo de DM que descreve as tarefas usadas pelos especialistas em DM para resolução de problemas. |
| *Dataset* | Coleção de dados. |
| *Outlier* | Observações que apresentam um grande afastamento das restantes ou são inconsistentes. |
| Valores omissos | Variável observada sem valor armazenado, valor omisso. |

**Tabela 20: *Project Plan* – Fases e tarefas**

| Fase | Nº | Tarefa | Preced. |
|---|---|---|---|
| Compreensão do Negócio | 1 | Definição dos objetivos e requisitos da perspetiva do negócio. | |
| | 2 | Definição de um problema de DM. | 1 |
| | 3 | Definição de um plano. | 2 |
| | 4 | Definição de critérios de sucesso para alcançar os objetivos. | 1 |
| | 5 | Levantamento dos recursos disponíveis para o projeto. | 1 |
| | 6 | Levantamento de todas as exigências e constrangimentos do projeto. | 3 |
| | 7 | Registo dos riscos e eventos que afetam o projeto e elaboração de plano de contingência. | 3 |
| | 8 | Definição de metas de negócio. | 1 |
| | 9 | Produzir o plano de projeto. | 3 |
| Compreensão dos Dados | 10 | Aquisição inicial dos dados. | 6 |
| | 11 | Análise e exploração dos dados adquiridos. | 10 |
| | 12 | Examinar a qualidade dos dados e existência de *valores discrepantes.* | 11 |
| Preparação dos Dados | 13 | Seleção de dados revelantes para análise. | 12 |
| | 14 | Definição de técnicas adequadas para limpeza de dados. | 13 |

| Fase | Nº | Tarefa | Preced. |
|---|---|---|---|
| | 15 | Definição de atributos construtivos. | 14 |
| | 16 | Definição de métodos para criar novos registos ou valores. | 15 |
| Modelação | 17 | Selecionar a técnica de modelação. | 16 |
| | 18 | Elaboração de plano de treino para testar e avaliar o modelo. | 17 |
| | 19 | Construção e a descrição dos modelos resultantes. | 18 |
| | 20 | Interpretação dos modelos. | 19 |
| Avaliação | 21 | Avaliar o modelo e verificar o comportamento em ambiente de testes. | 20 |
| | 22 | Verificação do cumprimento dos objetivos de negócio. | 21 |
| | 23 | Revisão do processo de DM. | 22 |
| | 24 | Decisão de uso dos resultados de DM. | 23 |
| Implementação | 25 | Definição de uma estratégia para desenvolvimento do modelo. | 24 |
| | 26 | Definição de plano de monitorização e manutenção do modelo. | 24 |
| | 27 | Elaboração de relatório final, incluindo todos os anteriores entregáveis e resumo dos resultados obtidos. | 26 |
| | 28 | Avaliação final do projeto. | 27 |

**Tabela 21: *Project Plan* – Técnicas de DM**

| Técnica | Justificação |
|---|---|
| *Decision Trees (DT)* | Funcionam bem com grandes conjuntos de dados, muitas variáveis e com diferentes tipos de dados.<br>Fácil compreensão da sua estrutura, bom desempenho nos casos em que os atributos apresentam ruído e erros. |
| *Artificial Neural Networks (ANN)* | Robustas a erros na situação de treino dos dados, sendo adequadas para a aprendizagem a partir de exemplos ruidosos. |
| *Support Vector Machines (SVM)* | Complexidade computacional da SVM não depende diretamente da dimensionalidade do espaço de entrada; usam minimização do risco estrutural (menos propensas a *sobre ajustamento*) |

# B. Compreensão dos Dados

**Tabela 22: *Script sql* para aquisição do *dataset* inicial.**

```
SELECT
sd_ints.t_doente AS "Tipo_Doente",
sd_ints.doente AS "N_Doente",
sd_doente.sexo AS "Sexo",
gr_doente_sexo.descr AS "Descr_Sexo",
sd_doente.dt_nasc AS "Dt_Nascimento",
ROUND(TO_NUMBER((TO_DATE('01-04-2013','DD-MM-YYYY') - sd_doente.dt_nasc)/360), 0) AS "Idade",
ROUND(TO_NUMBER((sd_ints.dt_int - sd_doente.dt_nasc)/360), 0) AS "Idade_Intern",
NVL(sd_doente.cod_pais, 'NULL') AS "Pais",
NVL(gr_paises.descr_pais, 'NULL') AS "Descr_Pais",
NVL(sd_doente.localidade, 'NULL') AS "Localidade",
NVL(sd_doente.escolaridade, 'NULL') AS "Escolaridade",
NVL(cv_escolaridade.descr, 'NULL') AS "Descr_Escolaridade",
NVL(sd_doente.estado_civil, 'NULL') AS "Est_Civil",
NVL(cv_est_civil.descr, 'NULL') AS "Descr_Est_Civil",
NVL(sd_ints.t_episodio_pai, 'NULL') AS "T_Episod_Origem",
NVL(sd_ints.episodio_pai, 'NULL') AS "Episod_Origem",
sd_int_alta.t_episodio AS "T_Episod_Intern",
sd_ints.n_int AS "N_Intern",
NVL(TO_CHAR(sd_ped_inte.dt_ped_inter), 'NULL') AS "Dt_Ped_Intern",
sd_ints.cod_serv AS "Serv_Intern",
a.descr_serv AS "Descr_Serv_Intern",
sd_ints.cod_serv_valencia AS "Espec_Medica",
b.descr_serv AS "Descr_Espec_Medica",
NVL(sd_int_alta.dest_pos_alta, 'NU') AS "Dest_Alta",
NVL(gr_prov_dest.descr_prov_dest, 'NULL') AS "Descr_Dest_Alta",
NVL(sd_ints.n_mecan, 'NULL') AS "N_Med_Alta",
NVL(sd_pess_hosp_def.nome, 'NULL') AS "Nome_Med_Alta",
NVL(sd_ints.cod_trat_oper, 'NULL') AS "Tratamento",
NVL(c.descricao, 'NULL') AS "Descr_Tratamento",
NVL(grup21_automatico.proc_princ, 'NULL') AS "Proc_Principal",
NVL(d.descricao, 'NULL') AS "Descr_Proc_Principal",
NVL(grup21_automatico.diagn_princ, 'NULL') AS "Diag_Principal",
NVL(ana_cid_diag.descricao, 'NULL') AS "Descr_Diag_Principal",
NVL(sd_ped_inte_diag.cod_diag, 'NULL') AS "Diag_Inicial",
NVL(sd_ped_inte_diag.descr_diag, 'NULL') AS "Descr_Diag_Inicial",
```

```
NVL(TO_CHAR(sd_int_alta.gdh), 'NULL') AS "GDH",
NVL(TO_CHAR(sd_int_alta.gcd), 'NULL') AS "GCD",
sd_ints.dt_int AS "Dt_Internamento", --sd_int_alta.dt_int,
SUBSTR(sd_ints.dt_int, 4, 3) AS "Mes_Intern",
SUBSTR(sd_ints.dt_int, 8, 4) AS "Ano_Intern",
SUBSTR(TO_CHAR(sd_ints.hr_int, 'DD-MM-YYYY HH24:MI:SS'), 12, 8) AS "Hora_Intern",
sd_ints.dt_alta AS "Dt_Alta_Intern", --sd_int_alta.dt_alta,
NVL(SUBSTR(TO_CHAR(sd_ints.hr_alta, 'DD-MM-YYYY HH24:MI:SS'), 12, 8), 'NULL') AS "Hora_Alta_Intern",
ROUND(TO_NUMBER(sd_ints.dt_alta - sd_ints.dt_int), 1) AS "N_Dias_Intern"
FROM sd_ints
INNER JOIN sd_doente ON (sd_ints.doente = sd_doente.doente)
LEFT JOIN gr_doente_sexo ON (sd_doente.sexo = gr_doente_sexo.sexo)
LEFT JOIN gr_paises ON (sd_doente.cod_pais = gr_paises.cod_pais)
left JOIN cv_est_civil ON (sd_doente.estado_civil = TRIM(cv_est_civil.cod))
LEFT JOIN cv_escolaridade ON (sd_doente.escolaridade = cv_escolaridade.cod)
LEFT JOIN gr_prov_dest ON (sd_ints.dest_pos_alta = gr_prov_dest.cod_prov_dest)
LEFT JOIN grup21_automatico ON (sd_ints.n_int = grup21_automatico.episodio)
LEFT JOIN ana_cid_interv c ON (sd_ints.cod_trat_oper = c.codigo AND c.codificacao = 'CID09')
LEFT JOIN ana_cid_interv d ON (grup21_automatico.proc_princ = d.codigo AND d.codificacao = 'CID09')
LEFT JOIN ana_cid_diag ON (grup21_automatico.diagn_princ = ana_cid_diag.codigo AND
ana_cid_diag.codificacao = 'CID09')
INNER JOIN sd_serv a ON (sd_ints.cod_serv = a.cod_serv)
INNER JOIN sd_serv b ON (sd_ints.cod_serv_valencia = b.cod_serv)
LEFT JOIN sd_pess_hosp_def ON ( sd_ints.n_mecan = sd_pess_hosp_def.n_mecan)
INNER JOIN sd_int_alta ON (sd_ints.n_int = sd_int_alta.episodio)
LEFT JOIN sd_ped_inte ON ( sd_ints.episodio_pai = sd_ped_inte.n_ped_inte)
LEFT JOIN sd_ped_inte_diag ON (sd_ped_inte_diag.dt_cri < to_date('01042013','ddmmyyyy') AND
sd_ints.episodio_pai = sd_ped_inte_diag.n_ped_inte)
WHERE sd_ints.dt_alta < to_date('01042013','ddmmyyyy')
ORDER BY sd_ints.n_int;
```

**Tabela 23: Dicionário de dados**

| Nome do Atributo | Descrição | Valores Possíveis | Classificação Original | Tipo de Tratamento |
|---|---|---|---|---|
| Tipo_Doente | Tipo de utente | EXT - Externo<br>FAP – Força Aérea | Nominal | Nominal |
| N_Doente | Número de utente | EXT - Número sequencial<br>FAP - Alfanumérico | Nominal | Nominal |
| Sexo | Código do sexo | N; M; F | Nominal | Nominal |
| Descr_Sexo | Descrição do código do sexo | Não Definido; Masculino; Feminino | Nominal | Nominal |

| Nome do Atributo | Descrição | Valores Possíveis | Classificação Original | Tipo de Tratamento |
|---|---|---|---|---|
| Dt_Nascimento | Data nascimento do utente | Formato: DD-MM-YYYY | Data | Nominal |
| Idade | Idade do utente, resultado da diferença entre a data 31-01-2013 e a data de nascimento | Vários valores compreendidos entre os 3 e os 216 anos. | Numérica | Numérica |
| Idade_Intern | Idade do utente à data do internamento | Vários valores compreendidos entre os 0 e os 207 anos. | Numérica | Numérica |
| Pais | Código do país de residência | 5420; 5472; P | Nominal | Nominal |
| Descr_Pais | Descrição do país de residência | Guiné-Bissau; S. Tomé e Príncipe; Portugal | Nominal | Nominal |
| Localidade | Morada do utente | Vários valores. | Nominal | Nominal |
| Escolaridade | Código de escolaridade do utente | Vários valores compreendidos entre o 10 e o 999. | Ordinal | Numérica |
| Descr_Escolaridade | Descrição de escolaridade do utente | Vários valores compreendidos entre as desconhecidas e o doutoramento | Ordinal | |
| Est_Civil | Código do estado civil do utente | [1 :9]; A | Nominal | Nominal |
| Descr_Est_Civil | Descrição do estado civil do utente | Vários valores. Ex.: Casado(a); Solteiro(a) | Nominal | Nominal |
| T_Episod_Origem | Tipo de episódio de origem do internamento | Vários valores. Ex: Pré-Internamento; Consulta | Nominal | Nominal |
| Episod_Origem | Número do episódio de origem do internamento | Vários valores. | Nominal | Nominal |
| T_Episod_Intern | Tipo de episódio de internamento | Internamento; Ambulatório | Nominal | Nominal |
| N_Intern | Número de registo de internamento | Vários valores compreendidos entre o 181 e o 28865. | Numérica | Numérica |
| Dt_Ped_Intern | Data do pedido de internamento | Formato: DD-MM-YYYY | Data | Nominal |
| Serv_Intern | Código do serviço de internamento ou ambulatório | Vários valores. Ex.: 1; 2; 6; 8 | Nominal | Nominal |
| Descr_Serv_Intern | Descrição do serviço de internamento | Vários valores. Ex.: Serviço de Medicina; Serviço de Cirurgia | Nominal | Nominal |
| Espec_Medica | Código da especialidade médica associada ao internamento | Vários valores. Ex.: 120; 250 | Nominal | Nominal |
| Descr_Espec_Medica | Descrição da especialidade médica associada ao internamento | Vários valores. Ex.: Cirurgia Geral; Ortopedia | Nominal | Nominal |
| Dest_Alta | Código do destino do utente após a alta clinica | Vários valores. Ex.: C; H | Nominal | Nominal |
| Descr_Dest_Alta | Descrição do destino do utente após a alta clinica | Vários valores. Ex.: Clinica; Transferido para outro hospital | Nominal | Nominal |
| N_Med_Alta | Número mecanográfico do médico responsável/associado | Vários valores. | Nominal | Nominal |

| Nome do Atributo | Descrição | Valores Possíveis | Classificação Original | Tipo de Tratamento |
|---|---|---|---|---|
| | ao internamento | | | |
| Nome_Med_Alta | Nome do médico responsável/associado ao internamento | Vários valores. | Nominal | Nominal |
| Tratamento | Codificação clinica para procedimentos, tratamentos e doenças | Vários valores.<br>Ex.: 7936 | Ordinal | Numérica |
| Descr_Tratamento | Descrição da codificação clinica para procedimentos, tratamentos e doenças | Vários valores.<br>Ex.: Redução aberta de fractura da tíbia e perónio, c/fixação interna | Nominal | Nominal |
| Proc_Principal | Código do procedimento principal | Vários valores.<br>Ex.: 2188 | Ordinal | Numérica |
| Descr_Proc_Principal | Descrição do procedimento principal | Vários valores.<br>Ex.: Septo plastias ncop | Nominal | Nominal |
| Diag_Principal | Código do diagnóstico principal | Vários valores.<br>Ex.: 470 | Ordinal | Numérica |
| Descr_Diag_Principal | Descrição do diagnóstico principal | Vários valores.<br>Ex.: Desvio do septo nasal (adquirido) | Nominal | Nominal |
| Diag_Inicial | Código do diagnóstico inicial no pedido de internamento | Vários valores.<br>Ex.: 3669 | Ordinal | Numérica |
| Descr_Diag_Inicial | Descrição do diagnóstico inicial no pedido de internamento | Vários valores.<br>Ex.: Catarata não especificada | Nominal | Nominal |
| GDH | Código Grupo Diagnóstico Homogéneo | Vários valores. | Ordinal | Numérica |
| GCD | Código Grande Categoria Diagnóstico | Vários valores. | Ordinal | Numérica |
| Dt_Internamento | Data do internamento do utente | Formato: DD-MM-YYYY | Data | Nominal |
| DiaSemana_Intern | Dia da semana do internamento do utente | [1 (Segunda) a 7 (Domingo)] | Ordinal | Numérica |
| Mes_Intern | Mês do internamento do utente | [1(janeiro) a 12(Dezembro)] | Ordinal | Numérica |
| Trimestre_Intern | Trimestre do internamento do utente | [1 – 4] | Ordinal | Numérica |
| Ano_Intern | Ano do internamento do utente | Vários valores compreendidos entre 2000 e 2013. | Ordinal | Numérica |
| Hora_Intern | Hora do internamento do utente | Formato: HH:MM:SS | Data | Nominal |
| Dt_Alta | Data de alta do utente | Formato: DD-MM-YYYY | Data | Nominal |
| Hora_Alta_Intern | Hora de alta do utente | Formato: HH:MM:SS | Data | Nominal |
| N_Intern_Anterior | Nº de internamentos anteriores pelo utente | Vários valores compreendidos entre 0 e 63. | Numérica | Numérica |
| N_Dias_Intern | Nª de dias de internamento do utente | Vários valores compreendidos entre 0 e 1148. | Numérica | Numérica |

**Tabela 24: Frequências dos atributos**

| Categorias de Entrada | Atributo | Níveis | Característica | n | % Válidos | % Totais |
|---|---|---|---|---|---|---|
| Características do paciente | Sexo | 3 | M | 15234 | 57,57 | 57,57 |
| | | | F | 11214 | 42,38 | 42,38 |
| | | | N | 14 | 0,05 | 0,05 |
| | | | Missing | 0 | 0,00 | 0,00 |
| | | | Total | 26462 | 100,00 | 100,00 |
| | Dt_Nascimento | 11195 | Missing | 0 | 0,00 | 0,00 |
| | | | Total | 26462 | 100,00 | 100,00 |
| | Idade_Intern | 102 | 70 | 625 | 2,36 | 2,36 |
| | | | 69 | 562 | 2,12 | 2,12 |
| | | | 72 | 558 | 2,11 | 2,11 |
| | | | Outros Níveis Total | 24717 | 93,41 | 93,41 |
| | | | Missing | 0 | 0,00 | 0,00 |
| | | | Total | 26462 | 100,00 | 100,00 |
| | Pais | 3 | P | 26451 | 99,98 | 99,96 |
| | | | 5472 | 3 | 0,01 | 0,01 |
| | | | 5420 | 2 | 0,01 | 0,01 |
| | | | Missing | 6 | 0,02 | 0,02 |
| | | | Total | 26419 | 100,00 | 100,00 |
| | Localidade | 2436 | Lisboa | 1320 | 6,99 | 4,99 |
| | | | MONTIJO | 383 | 2,03 | 1,45 |
| | | | ODIVELAS | 365 | 1,93 | 1,38 |
| | | | Outros Níveis Total | 16827 | 89,06 | 63,59 |
| | | | Missing | 7567 | 40,05 | 28,60 |
| | | | Total | 2646 | 100,00 | 100,00 |
| | Escolaridade | 33 | 330 | 2636 | 17,94 | 9,96 |
| | | | 999 | 1894 | 12,89 | 7,16 |
| | | | 310 | 1721 | 11,71 | 6,50 |
| | | | Outros Níveis Total | 8440 | 54,45 | 31,89 |
| | | | Missing | 11771 | 80,12 | 44,48 |
| | | | Total | 26462 | 100,00 | 100,00 |
| | Est_Civil | 9 | 2 | 10409 | 63,41 | 39,34 |
| | | | 1 | 3145 | 19,16 | 11,88 |
| | | | 3 | 1319 | 8,03 | 4,98 |
| | | | Outros Níveis Total | 1543 | 9,40 | 5,83 |
| | | | Missing | 10046 | 61,20 | 37,96 |
| | | | Total | 26462 | 100,00 | 100,00 |
| Processo administrativo Internamento do paciente | T_Episod_Origem | 12 | Pre-Intern | 12547 | 47,42 | 47,42 |
| | | | Ficha-ID | 12547 | 45,45 | 45,45 |
| | | | Consultas | 1374 | 5,19 | 5,19 |
| | | | Outros Níveis Total | 516 | 1,94 | 1,94 |
| | | | Missing | 0 | 0,00 | 0,00 |
| | | | Total | 26462 | 100,00 | 100,00 |
| | T_Episod_Intern | 2 | Internamentos | 19399 | 73,30 | 73,31 |
| | | | Ambulatório | 7063 | 26,70 | 26,69 |
| | | | Missing | 0 | 0,00 | 0,00 |
| | | | Total | 26462 | 100,00 | 100,00 |
| | Dt_Ped_Intern | 2650 | Missing | 12687 | 92,10 | 47,94 |
| | | | Total | 26462 | 100,00 | 100,00 |
| | Serv_Intern | 21 | 2 | 7283 | 27,52 | 27,52 |
| | | | 8 | 7065 | 26,70 | 26,70 |
| | | | 1 | 3913 | 26,70 | 26,70 |
| | | | Outros Níveis Total | 8201 | 30,91 | 30,91 |
| | | | Missing | 0 | 0,00 | 0,00 |
| | | | Total | 26462 | 100,00 | 100,00 |

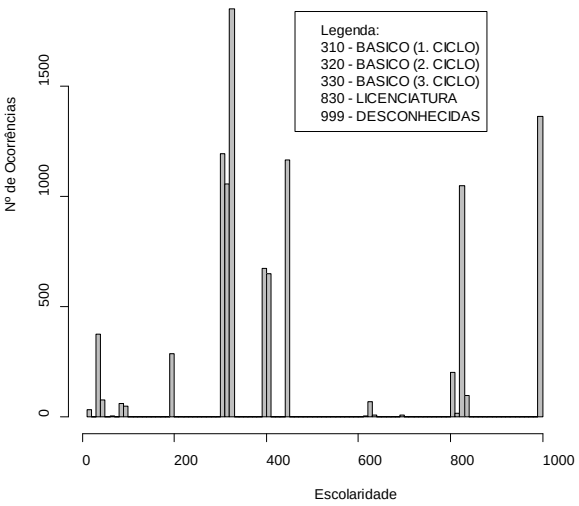
| Categorias de Entrada | Atributo | Níveis | Característica | n | % Válidos | % Totais |
|---|---|---|---|---|---|---|
| | Espec_Medica | 24 | 120 | 5360 | 20,26 | 20,26 |
| | | | 130 | 3627 | 13,71 | 13,71 |
| | | | 230 | 2942 | 11,12 | 11,12 |
| | | | Outros Níveis Total | 14533 | 54,92 | 54,92 |
| | | | Missing | 0 | 0,00 | 0,00 |
| | | | Total | 26462 | 100,00 | 100,00 |
| | N_Med_Alta | 161 | 079117 | 1891 | 8,84 | 7,15 |
| | | | 079116 | 1742 | 8,15 | 6,58 |
| | | | 096663 | 1091 | 5,10 | 4,12 |
| | | | Outros Níveis Total | 16658 | 77,91 | 62,95 |
| | | | Missing | 5080 | 23,76 | 19,20 |
| | | | Total | 26462 | 100,00 | 100,00 |
| | Dest_Alta | 8 | C | 25819 | 97,57 | 97,57 |
| | | | O | 414 | 1,56 | 1,56 |
| | | | H | 130 | 0,49 | 0,49 |
| | | | Outros Níveis Total | 99 | 0,37 | 0,37 |
| | | | Missing | 0 | 0,00 | 0,00 |
| | | | Total | 26462 | 100,00 | 100,00 |
| | GCD | 25 | 2 | 2656 | 14,26 | 10,04 |
| | | | 3 | 2088 | 11,21 | 7,89 |
| | | | 8 | 2051 | 11,01 | 7,75 |
| | | | Outros Níveis Total | 11828 | 63,51 | 44,67 |
| | | | Missing | 7839 | 42,09 | 29,62 |
| | | | Total | 26462 | 100,00 | 100,00 |
| | GDH | 469 | 39 | 1709 | 9,18 | 6,46 |
| | | | 270 | 757 | 4,06 | 2,86 |
| | | | 55 | 666 | 3,58 | 2,52 |
| | | | Outros Níveis Total | 15491 | 83,18 | 58,54 |
| | | | Missing | 7839 | 42,09 | 29,62 |
| | | | Total | 26462 | 100,00 | 100,00 |
| | Dt_Alta | 3629 | Missing | 0 | 0,00 | 0,00 |
| | | | Total | 26462 | 100,00 | 100,00 |
| | Ano_Intern | 15 | 2012 | 3686 | 13,93 | 13,93 |
| | | | 2011 | 2699 | 10,20 | 10,20 |
| | | | 2004 | 2527 | 9,55 | 9,55 |
| | | | Outros Níveis Total | 17550 | 66,32 | 66,32 |
| | | | Missing | 0 | 0,00 | 0,00 |
| | | | Total | 26462 | 100,00 | 100,00 |
| | Mes_Intern | 12 | 1 | 2740 | 10,35 | 10,35 |
| | | | 3 | 2625 | 9,92 | 9,92 |
| | | | 10 | 2467 | 9,32 | 9,32 |
| | | | Outros Níveis Total | 18630 | 70,40 | 70,40 |
| | | | Missing | 0 | 0,00 | 0,00 |
| | | | Total | 26462 | 100,00 | 100,00 |
| | Dt_Internamento | 3411 | Missing | 0 | 0,00 | 0,00 |
| | | | Total | 26462 | 100,00 | 100,00 |
| | Hora_Intern | 13178 | 08:30:00 | 1769 | 6,69 | 6,69 |
| | | | 09:00:00 | 401 | 1,52 | 1,52 |
| | | | 10:00:00 | 350 | 1,32 | 1,32 |
| | | | Outros Níveis Total | 23942 | 90,48 | 90,48 |
| | | | Missing | 0 | 0,00 | 0,00 |
| | | | Total | 26462 | 100,00 | 100,00 |
| | Hora_Alta_Intern | 13026 | 15:00:00 | 1804 | 6,82 | 6,82 |
| | | | 12:00:00 | 886 | 3,35 | 3,35 |
| | | | 16:00:00 | 562 | 2,12 | 2,12 |
| | | | Outros Níveis Total | 23205 | 87,69 | 87,69 |

| Categorias de Entrada | Atributo | Níveis | Característica | n | % Válidos | % Totais |
|---|---|---|---|---|---|---|
| | | | Missing | 5 | 0,02 | 0,02 |
| | | | Total | 26462 | 100,00 | 100,00 |
| | N_Intern_Anterior | 64 | 0 | 14735 | 55,68 | 55,68 |
| | | | 1 | 5382 | 20,34 | 20,34 |
| | | | 2 | 2382 | 9,00 | 9,00 |
| | | | Outros Níveis Total | 3963 | 14,98 | 14,98 |
| | | | Missing | 0 | 0,00 | 0,00 |
| | | | Total | 26462 | 100,00 | 100,00 |
| Processo clinico Internamento do paciente | Proc_Principal | 594 | 1341 | 922 | 13,07 | 3,48 |
| | | | 863 | 486 | 6,89 | 1,84 |
| | | | 149 | 415 | 5,88 | 1,57 |
| | | | Outros Níveis Total | 5232 | 74,16 | 19,77 |
| | | | Missing | 19407 | 275,08 | 73,34 |
| | | | Total | 26462 | 100,00 | 100,00 |
| | Tratamento | 802 | 1341 | 1800 | 14,73 | 6,80 |
| | | | 8026 | 704 | 5,76 | 2,66 |
| | | | 2188 | 498 | 4,08 | 1,88 |
| | | | Outros Níveis Total | 9216 | 34,83 | 75,43 |
| | | | Missing | 14244 | 116,58 | 53,83 |
| | | | Total | 26462 | 100,00 | 100,00 |
| | Diag_Principal | 1059 | 3669 | 665 | 9,24 | 2,51 |
| | | | 470 | 457 | 6,35 | 1,73 |
| | | | 36250 | 253 | 3,52 | 0,96 |
| | | | Outros Níveis Total | 5819 | 80,89 | 21,99 |
| | | | Missing | 19268 | 267,83 | 72,81 |
| | | | Total | 26462 | 100,00 | 100,00 |
| | Diag_Inicial | 657 | 3669 | 908 | 9,44 | 3,43 |
| | | | 13 | 603 | 6,27 | 2,28 |
| | | | 10 | 555 | 5,77 | 2,10 |
| | | | Outros Níveis Total | 7557 | 78,53 | 28,56 |
| | | | Missing | 16839 | 174,99 | 63,63 |
| | | | Total | 26462 | 100,00 | 100,00 |
| Variável alvo | N_Dias_Intern | 186 | 0 | 6908 | 26,11 | 26,11 |
| | | | 2 | 3115 | 11,77 | 11,77 |
| | | | 4 | 2958 | 11,18 | 11,18 |
| | | | Outros Níveis Total | 13481 | 50,94 | 50,94 |
| | | | Missing | 0 | 0,00 | 0,00 |
| | | | Total | 26462 | 100,00 | 100,00 |

**Tabela 25: Sumário estatístico descritivo dos atributos**

| Atributo | N Válidos | Média | Desvio Padrão | Min | P25 | Mediana | P75 | Máximo |
|---|---|---|---|---|---|---|---|---|
| Idade_Intern | 26462 | 54,68 | 20,47 | 0,00 | 40,00 | 58,00 | 71,00 | 207,00 |
| N_Intern_Anterior | 26462 | 1,51 | 3,97 | 0,00 | 0,00 | 0,00 | 1,00 | 63,00 |
| N_Dias_Intern | 26462 | 7,14 | 23,77 | 0,00 | 0,00 | 3,00 | 6,00 | 2294,00 |
| GCD | 18623 | | | 0,00 | 3,00 | 7,00 | 9,00 | 25,00 |
| GDH | 18623 | | | 2,00 | 61,00 | 222,00 | 356,00 | 901,00 |
| Mes_Intern | 26462 | | | 1,00 | 3,00 | 6,00 | 9,00 | 12,00 |
| Ano_Intern | 26462 | | | 2000,00 | 2004,00 | 2007,00 | 2011,00 | 2013,00 |

**Figura 39: Diagrama de frequência e *boxplot* do atributo Escolaridade**



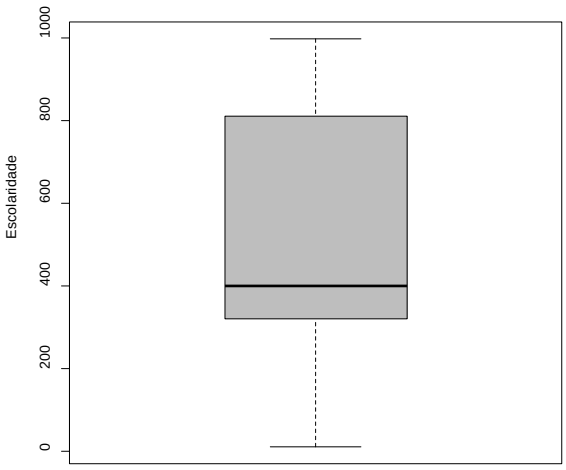
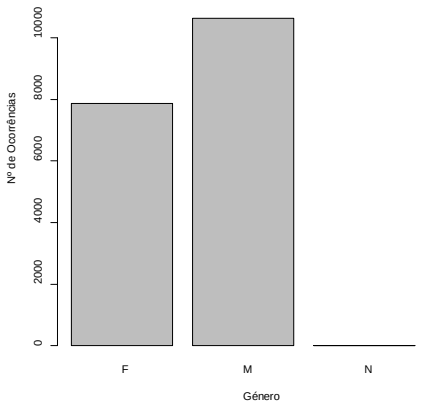




**Figura 40: Diagrama de frequência do atributo Género**

**Figura 41: Diagrama de frequência do atributo Est_Civil**



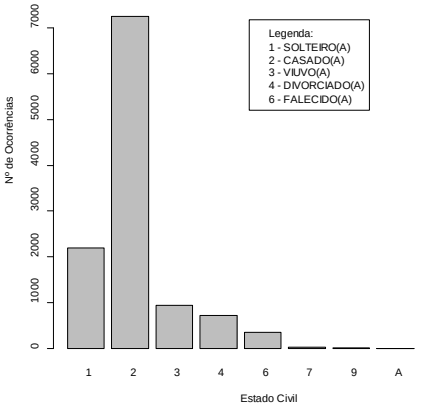
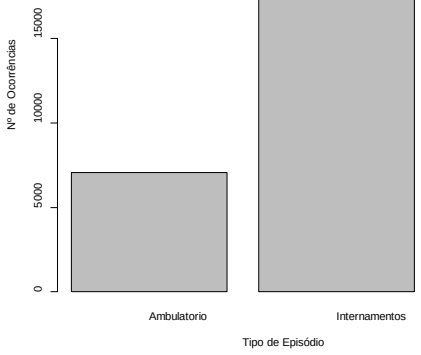




**Figura 42: Diagrama de frequência do atributo T_Episod_Intern**

**Figura 43: Diagrama de frequência do atributo Serv_Intern**



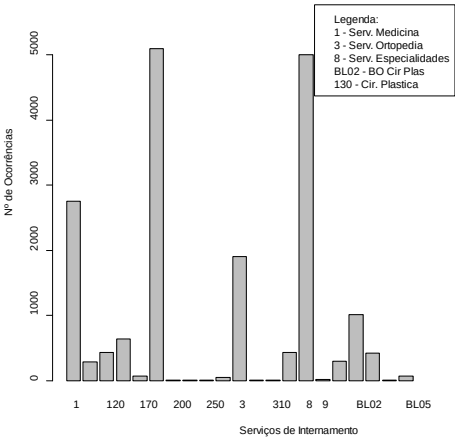



**Figura 44: Diagrama de frequência e *boxplot* do atributo Mês_Intern**

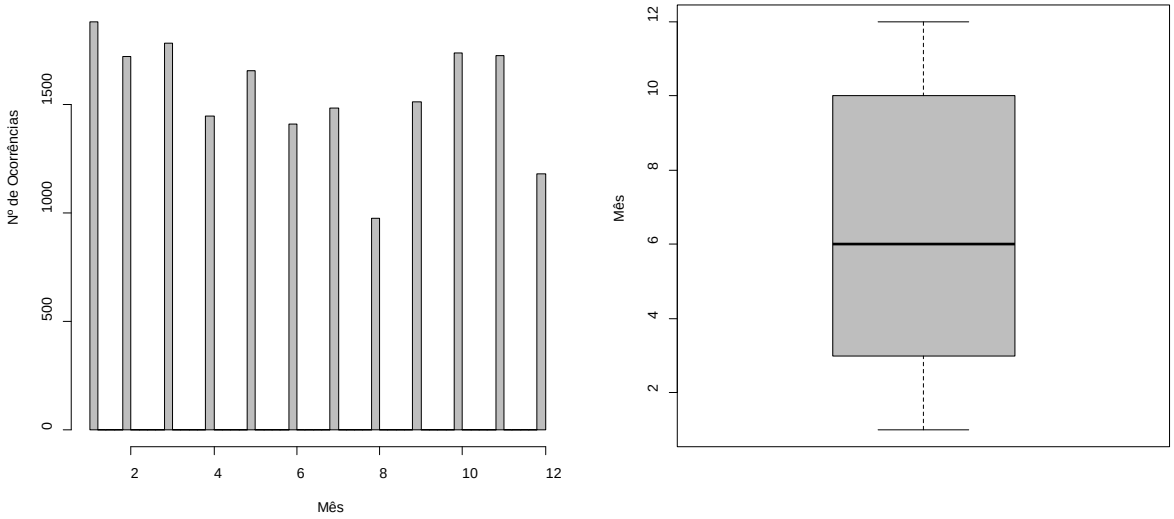



**Tabela 26: Código R dos gráficos representados na fase compreensão dos dados**

**Diagrama de frequência para o atributo Idade_Intern**
hist(d$Idade_Intern, plot = TRUE, main="", xlab="Idade", ylab="Nº de Ocorrências", breaks=100, col="gray")

**Diagrama de extremos e quartis do atributo Idade_Intern**
boxplot(d$Idade_Intern, main="", xlab="", ylab="Idade", col="gray")

**Diagrama de frequência do atributo Escolaridade**
hist(d$Escolaridade, plot = TRUE, main="", xlab="Escolaridade", ylab="Nº de Ocorrências", breaks=100, col="gray")
legend(locator(1), xpd=TRUE, legend=c("Legenda:", "310 - BASICO (1. CICLO)", "320 - BASICO (2. CICLO)", "330 - BASICO (3. CICLO)", "830 - LICENCIATURA", "999 - DESCONHECIDAS"))

**Diagrama de extremos e quartis do atributo Escolaridade**
boxplot(d$Escolaridade, main="", xlab="", ylab=" Escolaridade ", col="gray")

**Diagrama de frequência do atributo Género**
plot(d$Sexo, main = "", xlab = "Género", ylab = "Nº de Ocorrências", col = "gray")

**Diagrama de frequência do atributo Est_Civil**
plot(d$Est_Civil, main = "", xlab = "Estado Civil", ylab = "Nº de Ocorrências", col = "gray")
legend(locator(1), xpd=TRUE, legend=c("Legenda:", "1 - SOLTEIRO(A)", "2 - CASADO(A)", "3 - VIUVO(A)", "4 - DIVORCIADO(A)", "6 - FALECIDO(A)"))

**Diagrama de frequência do atributo Mês_Intern**
hist(d$Mes_Intern, plot = TRUE, main="", xlab="Mês", ylab="Nº de Ocorrências", breaks = 80, col="gray")

**Diagrama de extremos e quartis do atributo Mês_Intern**
boxplot(d$Mes_Intern, main="", xlab=" ", ylab=" Mês ", col="gray")

**Diagrama de frequência do atributo T_Episod_Intern**
plot(d$T_Episod_Intern, main = "", xlab = "Tipo de Episódio", ylab = "Nº de Ocorrências", col = "gray")

**Diagrama de frequência do atributo Serv_Intern**
plot(d$Serv_Intern, main = "", xlab = "Serviços de Internamento", ylab = "Nº de Ocorrências", col = "gray")
legend(locator(1), xpd=TRUE, legend=c("Legenda:", "1 - Servico de Medicina", "3 - Servico de Ortopedia", "8 - Servico de Especialidades", "BL02 - Bloco Operat Ambulat Cir Plas", "130 - Cirurgia Plastica"))

# C. Preparação dos Dados

**Tabela 27: Código R para tratamento dos valores omissos**

**Fase de pré-processamento da BD**

```
library(rminer) #executar a biblioteca RMINER
d<-read.table("internamento.csv", header=TRUE,sep=";")

summary(d)
```

**Substituição de valores omissos pelo valor encontrado no exemplo mais próximo**

```
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)
A<-imputation(imethod = "hotdeck", d, Attribute = "Hora_Alta_Intern", Missing = NA, Value = 1)
```

**Gravar a base de dados com as alterações realizadas**

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

**Criar os histogramas dos atributos alterados via *hot deck***

```
d2<-read.table("internamento_comNA.csv", header=TRUE,sep=";")
d3<-read.table("internamento_semNA.csv", header=TRUE,sep=";")

par(mfrow=c(1,2)) # colocar dois gráficos lado a lado
```

**Diagrama de frequências do atributo Escolaridade**

```
hist(d2$Escolaridade, plot = TRUE, main= "s/ hot deck", xlab="Escolaridade", ylab="Nº de Ocorrências", col="red")
hist(d3$Escolaridade, plot = TRUE, main= "c/ hot deck", xlab= "Escolaridade", ylab="Nº de Ocorrências", col="blue")
```

**Diagrama de frequências do atributo Est_Civil**

```
hist(d2$Est_Civil, plot = TRUE, main = "s/ hot deck", xlab = "Estado Civil", ylab = "Nº de Ocorrências", col = "red")
hist(d3$Est_Civil, plot = TRUE, main = "c/ hot deck", xlab = "Estado Civil", ylab = "Nº de Ocorrências", col = "blue")
```

**Diagrama de frequências do atributo Proc_Principal**

```
hist(d2$Proc_Principal, plot = TRUE , main = "s/ hot deck", xlab="Procedimento Principal", ylab="Nº de Ocorrências", col="red")
hist(d3$Proc_Principal, plot = TRUE , main= "c/ hot deck", xlab="Procedimento Principal", ylab="Nº de Ocorrências", col="blue")
```

**Diagrama de frequências do atributo Diag_Principal**

```
plot(d2$Diag_Principal, main= "s/ hot deck", xlab="Diagnóstico Principal", ylab="Nº de Ocorrências", col="red")
plot(d3$Diag_Principal, main="c/ hot deck", xlab= "Diagnóstico Principal", ylab="Nº de Ocorrências", col="blue")
```

**Diagrama de frequências do atributo Diag_Inicial**

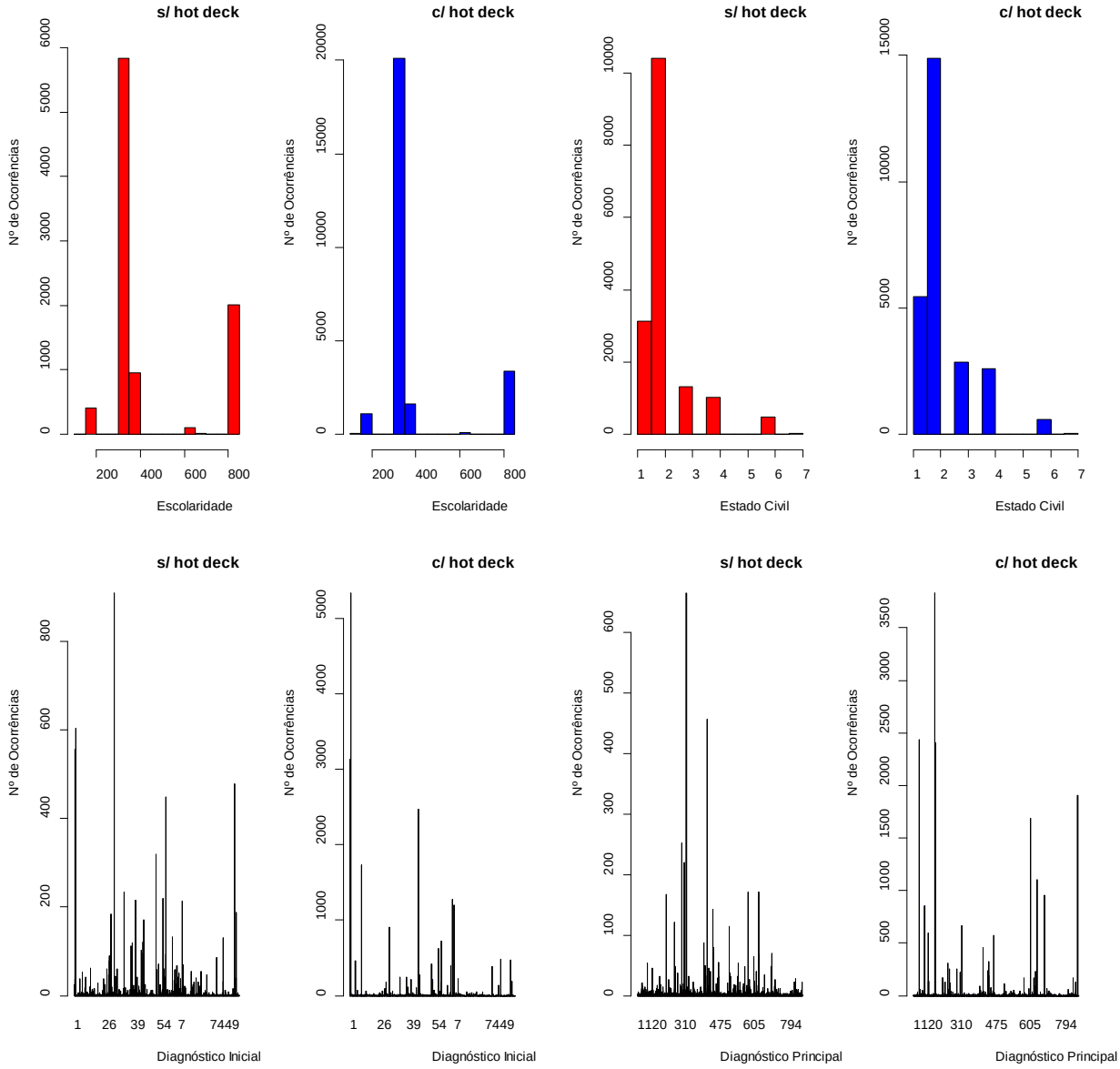
```
plot(d2$Diag_Inicial, main = "s/ hot deck", xlab="Diagnóstico Inicial", ylab="Nº de Ocorrências", col="red")
plot(d3$Diag_Inicial, main = "c/ hot deck", xlab="Diagnóstico Inicial", ylab="Nº de Ocorrências", col="blue")
```

**Diagrama de frequências do atributo GCD**

```
hist(d2$GCD, plot = TRUE , main="s/ hot deck", xlab="Grande Categoria Diagnóstico", ylab="Nº de
```

```
Ocorrências", col="red")
hist(d3$GCD, plot = TRUE , main = "c/ hot deck",  xlab="Grande Categoria Diagnóstico",  ylab="Nº de Ocorrências", col="blue")

Diagrama de frequências do atributo Hora_Alta_Intern
plot(d2$Hora_Alta_Intern, main = "s/ hot deck",  xlab = "Hora de Alta",  ylab = "Nº de Ocorrências", col = "red")
plot(d3$Hora_Alta_Intern, main = "s/ hot deck",  xlab = "Hora de Alta",  ylab = "Nº de Ocorrências", col = "blue")
```

**Figura 45: Diagrama de frequência dos atributos alterados via *hot deck***



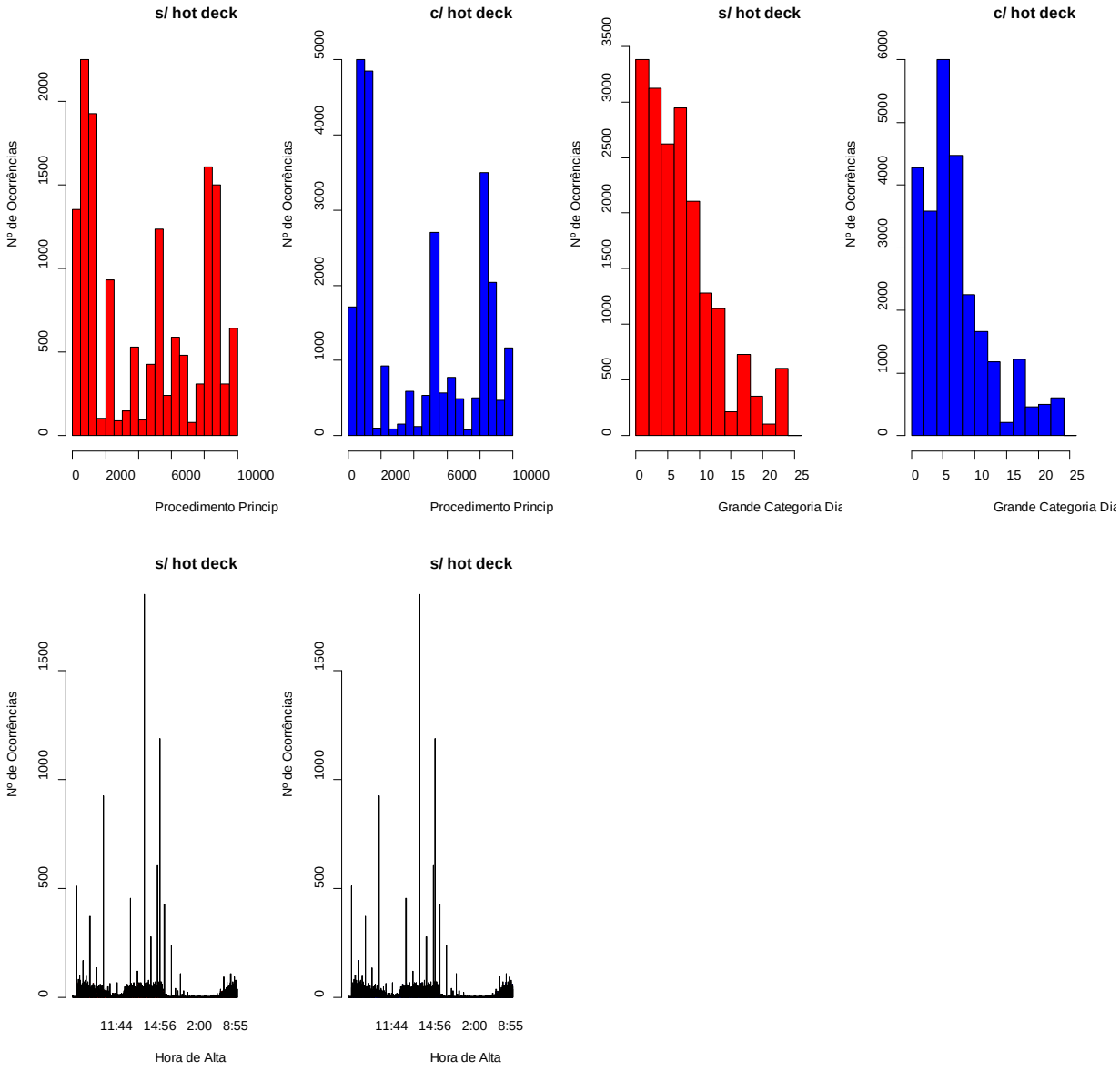



**Figura 46: Diagrama de frequência e *boxplot* do atributo DiaSemana_Intern**

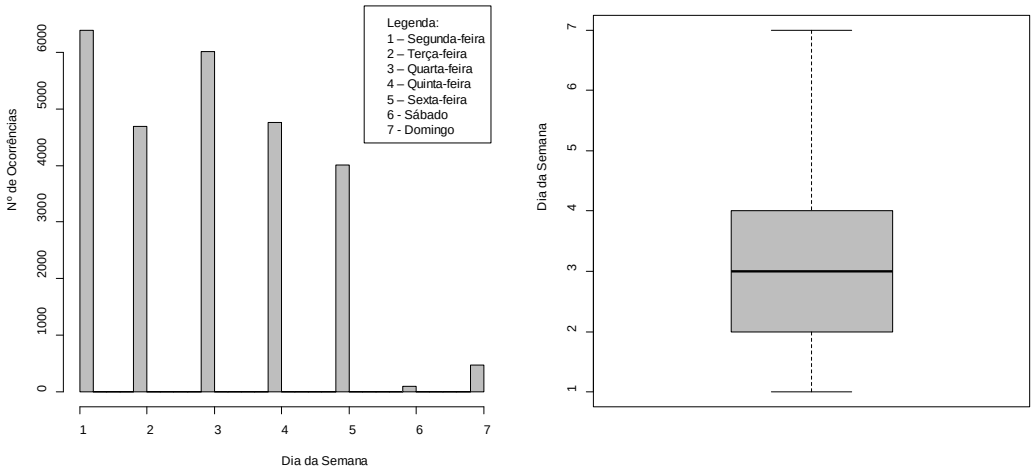


**Figura 47: Diagrama de frequência e *boxplot* do atributo transformado Hora_Intern**

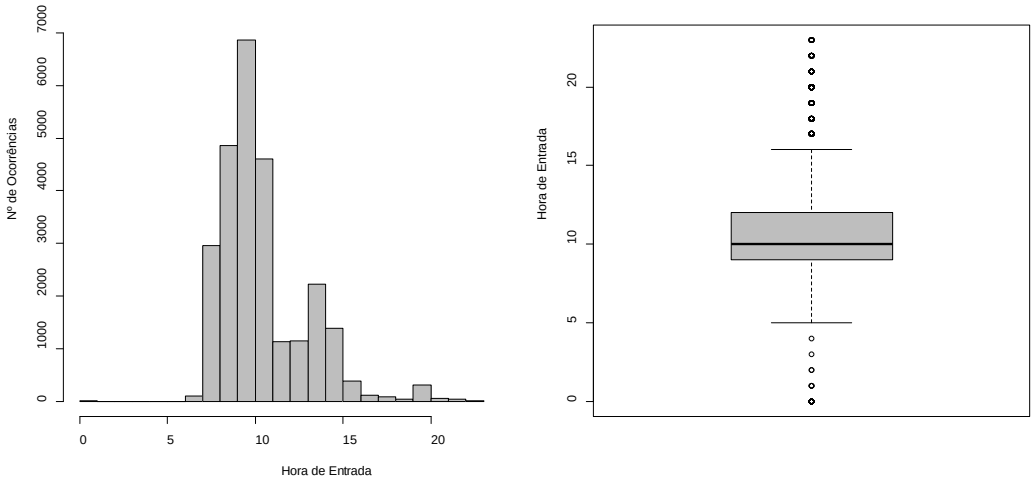



**Figura 48: Diagrama de frequência e *boxplot* do atributo transformado Hora_Alta_Intern**

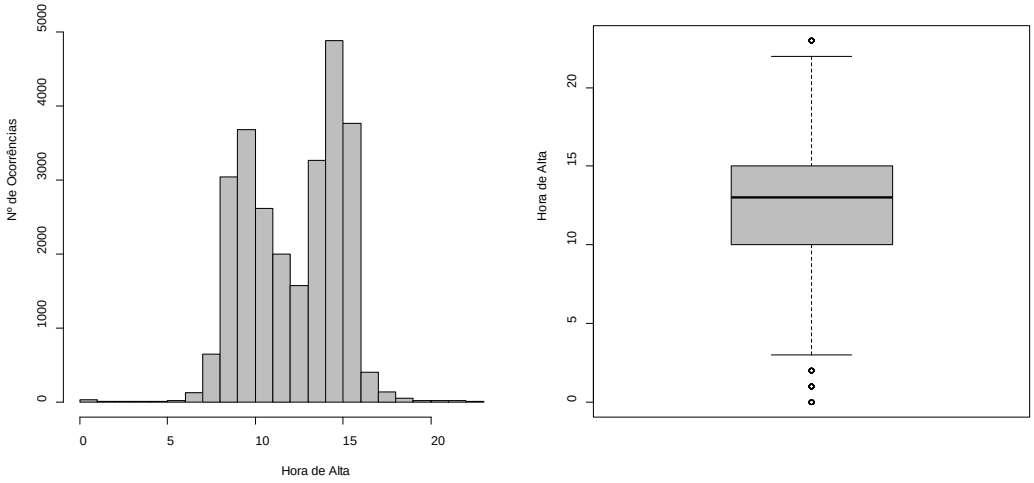



**Tabela 28: Análise estatística do atributo Escolaridade**

| Código | Descrição | Frequências |
|---|---|---|
| 100 | Habilitações s/ equivalência pelo me | 66 |
| 200 | Sem habilitações | 1112 |
| 310 | Básico (1. Ciclo) | 10191 |
| 320 | Básico (2. Ciclo) | 3783 |
| 330 | Básico (3. Ciclo) | 6149 |
| 400 | Secundário | 1620 |
| 620 | Técnico-profissional (nível 2) | 8 |
| 630 | Técnico-profissional (nível 3) | 98 |
| 700 | Medio | 14 |
| 810 | Bacharelato | 1514 |
| 820 | Estudos superiores especializados | 23 |
| 830 | Licenciatura | 1692 |
| 840 | Mestrado | 145 |
| 850 | Doutoramento | 4 |

**Figura 49: Diagrama de frequência e *boxplot* do atributo transformado Escolaridade**

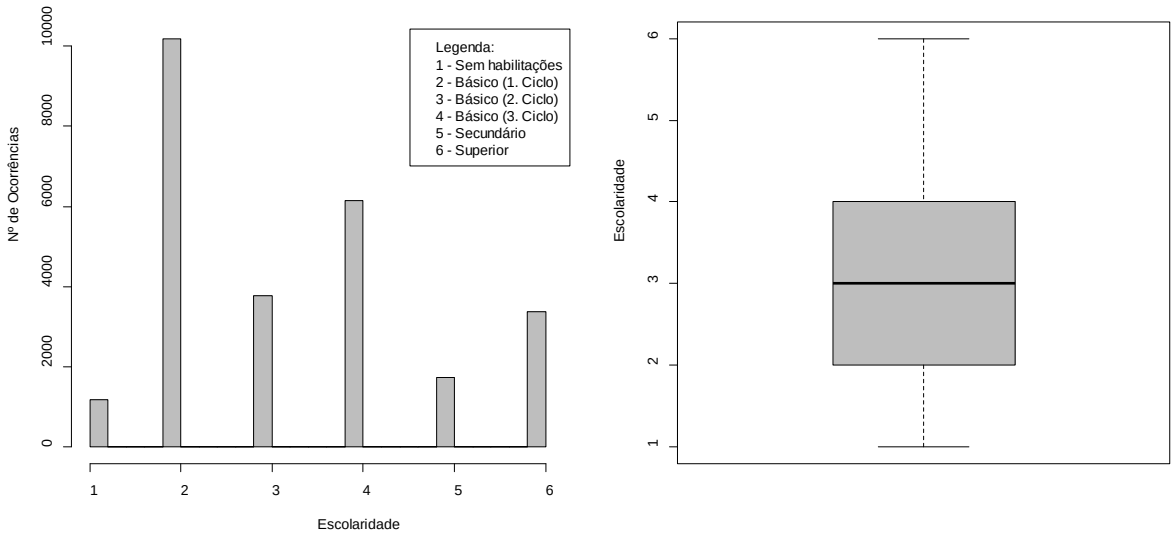



**Tabela 29: Código R para agrupamento das classes de Escolaridade**

| |
|---|
| **Fase de pré-processamento da BD**<br>library(rminer) #executar a biblioteca RMINER<br>d<-read.table("internamento.csv", header=TRUE,sep=";") summary(d$Escolaridade)<br><br>table(d$Escolaridade) |
| **Substituição do valor da classe pelo novo valor**<br>f=factor(d$Escolaridade)<br><br>B<- delevels(f, "100", "1")<br>B<- delevels(f, "200", "1")<br>B<- delevels(f, "310", "2")<br>B<- delevels(f, "320", "3")<br>B<- delevels(f, "330", "4")<br>B<- delevels(f, "400", "5")<br>B<- delevels(f, "620", "5")<br>B<- delevels(f, "630", "5")<br>B<- delevels(f, "700", "5")<br>B<- delevels(f, "810", "6")<br>B<- delevels(f, "820", "6")<br>B<- delevels(f, "830", "6")<br>B<- delevels(f, "840", "6")<br>B<- delevels(f, "850", "6") |
| **Gravar a base de dados com as alterações realizadas**<br>write.table(B,file=" internamento.csv",row.names=FALSE,col.names=TRUE,sep=";") |
| **Criar histograma do atributo**<br>d<-read.table("internamento.csv", header=TRUE,sep=";")<br><br>hist(d$Escolaridade, plot = TRUE , main = "", xlab = "Escolaridade", ylab = "Nº de Ocorrências", col = "gray") |

**Tabela 30: Análise estatística do atributo Proc_Principal**

| Código | Descrição | Frequências |
|---|---|---|
| 1341 | Facoemulsificacao e aspiração de catarata | 4725 |
| 8026 | Astroscopia do joelho | 1341 |
| 5304 | Reparação unilateral. Hérnia inguin. indirecta c/enxerto ou prótese | 819 |
| 821 | Excisão de chalazio | 705 |
| 6011 | Biopsia fechada [percutanea] [agulha] da próstata | 610 |
| | Outros Níveis Total | 18219 |

**Figura 50: Diagrama de frequência e *boxplot* do atributo transformado Proc_Principal**

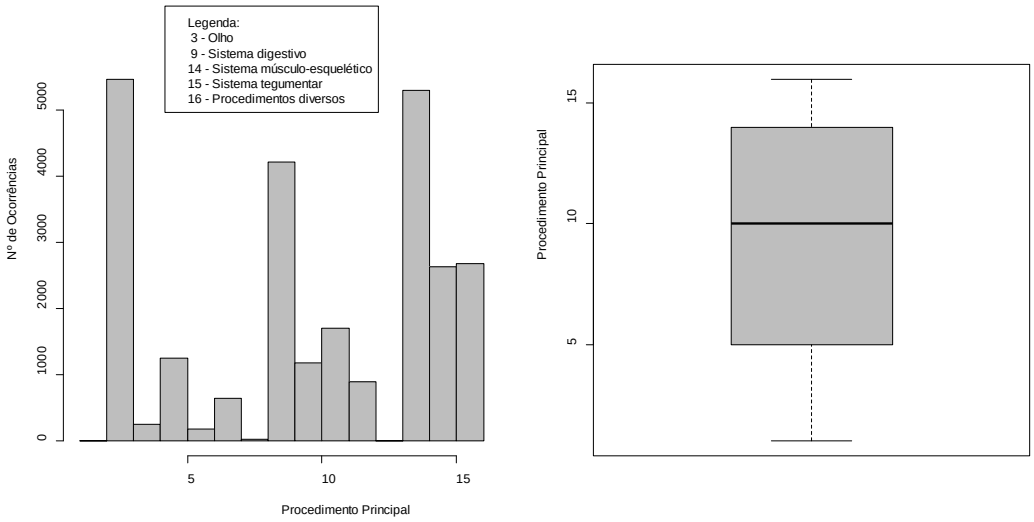



**Tabela 31: Código R para agrupamento das classes do Proc_Principal**

**Fase de pré-processamento da BD**

```
library(rminer) #executar a biblioteca RMINER
d<-read.table("internamento.csv", header=TRUE,sep=";")

summary(d$ Proc_Principal)
table(d$ Proc_Principal)
```

**Exemplo de substituição do valor da classe pelo novo valor**

```
f=factor(d$Proc_Principal)

C=delevels(f,c("60","61","62","63","64","600","602","604","605","612","613","622","623","624","625","631","
632","633","637","639","640","642","644","6011","6013","6015","6093","6191","6211","6373","6411","6442",
"6494","6496"),"11")

C=delevels(f,c("7592"),"13")

print(table(C))
```

**Gravar a base de dados com as alterações realizadas**

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

**Criar histograma do atributo**

```
d<-read.table("internamento.csv", header=TRUE,sep=";") # ler a base de dados

hist(d$Proc_Principal, plot = TRUE , main = "",  xlab = "Procedimento Principal",  ylab = "Nº de Ocorrências",
breahs=32, col = "gray")
```

**Tabela 32: Análise estatística do atributo Diag_Principal**

| Código | Descrição | Frequências |
|---|---|---|
| 2180 | Leiomioma submucoso do útero | 3829 |
| 1570 | Neoplasia maligna do retro peritoneu | 2433 |
| 2189 | Leiomioma uterino, não especificado | 2405 |
| V7285 | Exame especificado, não classificável em outra parte | 1904 |
| 7019 | Afeções hipertróficas ou atróficas da pele não especificadas | 1689 |
| | Outros Níveis Total | 14159 |

**Figura 51: Diagrama de frequência e *boxplot* do atributo transformado Diag_Principal**

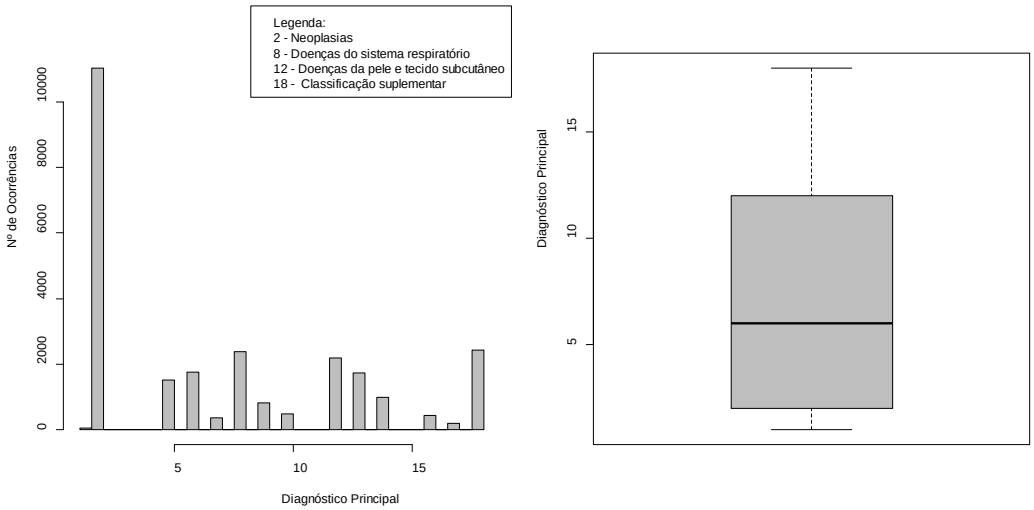



**Tabela 33: Código R para agrupamento das classes do Diag_Principal e Diag_Inicial**

**Fase de pré-processamento da BD**

```
library(rminer) #executar a biblioteca RMINER
d<-read.table("internamento.csv", header=TRUE,sep=";")

summary(d$ Diag_Principal)
summary(d$Diag_Principal)

table(d$ Diag_Principal)
table(d$Diag_Principal))
```

**Exemplo de substituição do valor da classe pelo novo valor**

```
f1=factor(d$Diag_Principal)
f2=factor(d$Diag_Principal)

D=delevels(f1,c("V1005","V5413","V5789","V5889","V671","V7285"),"18")
print(table(D))

E=delevels(f2,c("V431","V643","V615","V5881","V5841","V5830","V5401","V524","V518","V509","V508","
V502","V501","V4579","V4571","V4569","V442","V4361","V430","V2540","V140","V1083","V1040","V100
2"),"18")
print(table(D))
```

**Gravar a base de dados com as alterações realizadas**

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

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

**Criar histograma do atributo**

```
d<-read.table("internamento.csv", header=TRUE,sep=";")

hist(d$Diag_Principal, plot = TRUE , main = "",  xlab = "Diagnóstico Principal",  ylab = "Nº de Ocorrências", breaks=36, col = "gray")

hist(d$Diag_Inicial, plot = TRUE , main = "",  xlab = "Diagnóstico Inicial",  ylab = "Nº de Ocorrências", col = "gray")
```

**Tabela 34: Análise estatística do atributo Diag_Inicial**

| Código | Descrição | Frequências |
|---|---|---|
| 13 | Membro inferior - pé | 5329 |
| 10 | Membro inferior - joelho | 3128 |
| 49 | Ânus | 2469 |
| 2 | Mão e punho | 1736 |
| 68 | Útero | 1272 |
| | Outros Níveis Total | 12485 |

**Figura 52: Diagrama de frequência e *boxplot* do atributo transformado Diag_Inicial**

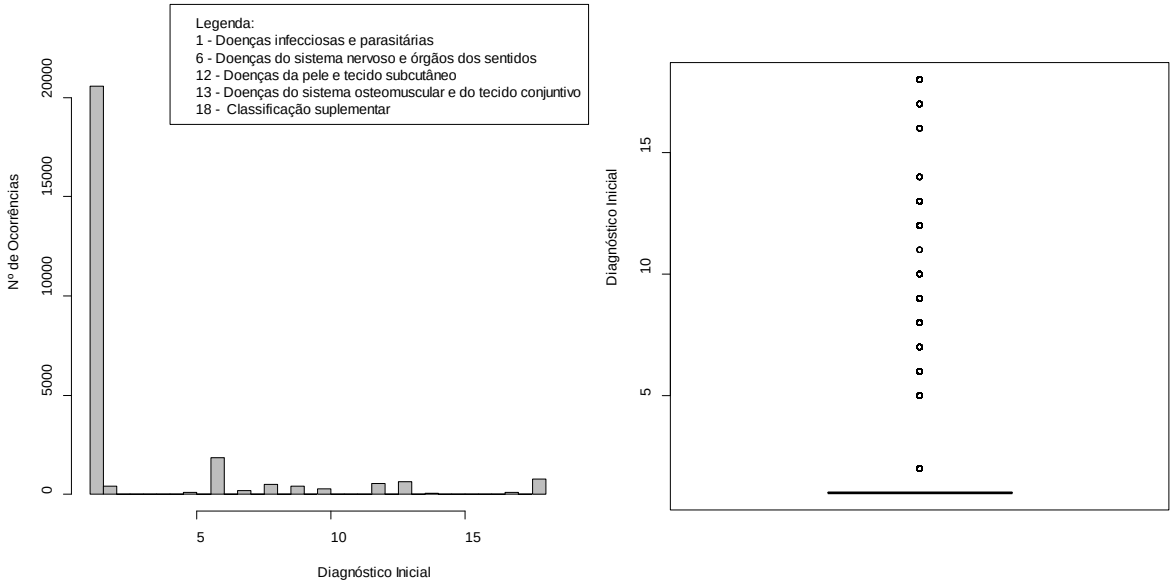



**Tabela 35: Análise estatística do atributo Idade_Intern**

| Código | Descrição | Frequências |
|---|---|---|
| 70 | Setenta anos | 625 |
| 69 | Sessenta e nove anos | 559 |
| 72 | Setenta e dois anos | 553 |
| 68 | Sessenta e oito anos | 538 |
| 71 | Setenta e um anos | 536 |
| | Outros Níveis Total | 23608 |

**Figura 53: Diagrama de frequência e *boxplot* do atributo transformado Idade_Intern**

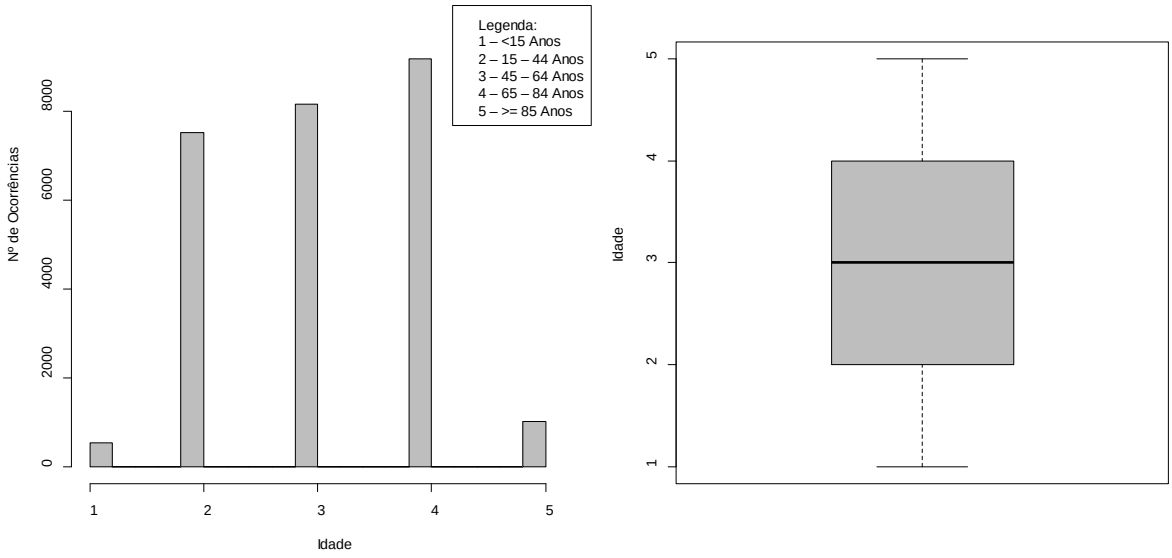



**Tabela 36: Código R para agrupamento das classes da Idade_Intern**

**Fase de pré-processamento da BD**

```
library(rminer) #executar a biblioteca RMINER
d<-read.table("internamento.csv", header=TRUE,sep=";") # ler a base de dados

summary(d$Idade_Intern)
table(d$Idade_Intern)
```

**Exemplo de substituição do valor da classe pelo novo valor**

```
f=factor(d$Idade_Intern)

F=delevels(f,c("0","1","2","3","4","5","6","7","8","9","10","11","12","13","14"),"1")
print(table(F))
```

**Gravar a base de dados com as alterações realizadas**

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

**Criar histograma do atributo**

```
d<-read.table("internamento.csv", header=TRUE,sep=";") # ler a base de dados

hist(d$Idade_Intern, plot = TRUE , main = "",  xlab = "Idade",  ylab = "Nº de Ocorrências", breaks=16, col = "gray")
```

**Tabela 37: Frequências dos atributos em estudo**

| Atributo | Níveis | Característica | n | % | Escala da Variável |
|---|---|---|---|---|---|
| Sexo | 2 | M (Masculino) | 15223 | 57,62 | Nominal |
| | | F (Feminino) | 11196 | 42,38 | |
| | | Total | 26419 | 100,00 | |
| Idade_Intern | 3 | 4 (65 – 84 Anos) | 9186 | 34,77 | Ordinal |
| | | 3 (45 – 64 Anos) | 8155 | 30,87 | |
| | | 2 (15 – 44 Anos) | 7518 | 28,46 | |
| | | Outros Níveis Total | 1560 | 5,90 | |

| Atributo | Níveis | Característica | n | % | Escala da Variável |
|---|---|---|---|---|---|
| | | Total | 26419 | 100,00 | |
| Escolaridade | 6 | 2 (Básico 1º Ciclo) | 10191 | 38,57 | Ordinal |
| | | 4 (Básico 3º Ciclo) | 6149 | 23,27 | |
| | | 3 (Básico 2º Ciclo) | 3783 | 14,32 | |
| | | Outros Níveis Total | 6296 | 23,83 | |
| | | Total | 26419 | 100,00 | |
| Est_Civil | 7 | 2 (Casado(a)) | 14891 | 56,36 | Ordinal |
| | | 1 (Solteiro(a)) | 5461 | 20,67 | |
| | | 3 Viúvo(a)) | 2860 | 10,82 | |
| | | Outros Níveis Total | 3207 | 12,14 | |
| | | Total | 26419 | 100,00 | |
| T_Episod_Origem | 12 | Pre-Intern | 12532 | 47,44 | Nominal |
| | | Ficha-ID | 12003 | 45,43 | |
| | | Consultas | 1372 | 5,19 | |
| | | Outros Níveis Total | 512 | 1,94 | |
| | | Total | 26419 | 100,00 | |
| T_Episod_Intern | 2 | Internamentos | 19366 | 73,30 | Nominal |
| | | Ambulatório | 7053 | 26,70 | |
| | | Total | 26419 | 100,00 | |
| Serv_Intern | 20 | 2 (Serviço de Cirurgia) | 7280 | 27,56 | Nominal |
| | | 8 (Serviço de Especialidades) | 7059 | 26,72 | |
| | | 1 (Serviço de Medicina) | 3913 | 14,81 | |
| | | Outros Níveis Total | 8167 | 30,91 | |
| | | Total | 26419 | 100,00 | |
| Espec_Medica | 23 | 120 (Cirurgia Geral) | 5353 | 20,26 | Nominal |
| | | 130 (Cirurgia Plástica) | 3619 | 13,70 | |
| | | 230 (Oftalmologia) | 2940 | 11,13 | |
| | | Outros Níveis Total | 14507 | 54,91 | |
| | | Total | 26419 | 100,00 | |
| Dest_Alta | 7 | C (Clinica) | 25820 | 97,73 | Nominal |
| | | O (Óbito) | 414 | 1,57 | |
| | | H (Transferido para outro Hospital) | 129 | 0,49 | |
| | | Outros Níveis Total | 56 | 0,21 | |
| | | Total | 26419 | 100,00 | |
| Proc_Principal | 15 | 3 (Operações no olho) | 5467 | 20,69 | Ordinal |
| | | 14 (Operações no sistema músculo-esquelético) | 5297 | 20,05 | |
| | | 9 (Operações sobre o sistema digestivo) | 4212 | 15,94 | |
| | | Outros Níveis Total | 11443 | 43,31 | |
| | | Total | 26419 | 100,00 | |
| Diag_Principal | 15 | 2 (Neoplasias) | 11043 | 41,80 | Ordinal |
| | | 18 (Classificação suplementar de fatores) | 2433 | 9,21 | |
| | | 8 (Doenças do sistema respiratório) | 2378 | 9,00 | |
| | | Outros Níveis Total | 10565 | 40,00 | |
| | | Total | 26419 | 100,00 | |
| Diag_Inicial | 15 | 1(Doenças infecciosas e parasitárias) | 20590 | 77,94 | Ordinal |
| | | 6 (Doenças do sistema nervoso e órgãos dos sentidos) | 1845 | 6,98 | |
| | | 18 (Classificação suplementar de fatores) | 752 | 2,85 | |
| | | Outros Níveis Total | 3232 | 12,23 | |
| | | Total | 26419 | 100,00 | |
| GCD | 25 | 8 (Doenças e Perturbações do Sistema Músculo-esquelético e Tecido Conjuntivo) | 3546 | 13,42 | Ordinal |
| | | 2 (Doenças e Perturbações do Olho) | 3496 | 13,23 | |
| | | 6 (Doenças e Perturbações do Aparelho | 3312 | 12,54 | |

| Atributo | Níveis | Característica | n | % | Escala da Variável |
|---|---|---|---|---|---|
| | | Digestivo) | | | |
| | | Outros Níveis Total | 16065 | 60,81 | |
| | | Total | 26419 | 100,00 | |
| DiaSemana_Intern | 7 | 1 (Segunda-feira) | 6390 | 24,19 | Ordinal |
| | | 3 (Quarta-feira) | 6005 | 22,73 | |
| | | 4 (Quinta-feira) | 4762 | 18,02 | |
| | | Outros Níveis Total | 9262 | 35,06 | |
| | | Total | 26419 | 100,00 | |
| Mes_Intern | 12 | 1 (Janeiro) | 2739 | 10,37 | Ordinal |
| | | 3 (Março) | 2622 | 9,92 | |
| | | 2 (Fevereiro) | 2458 | 9,30 | |
| | | Outros Níveis Total | 18600 | 70,40 | |
| | | Total | 26419 | 100,00 | |
| Hora_Intern | 23 | 10 (10 horas) | 6866 | 25,99 | Ordinal |
| | | 9 (9 horas) | 4854 | 18,37 | |
| | | 11 (11 horas) | 4602 | 17,42 | |
| | | Outros Níveis Total | 10097 | 38,22 | |
| | | Total | 26419 | 100,00 | |
| Hora_Alta_Intern | 24 | 15 (15 horas) | 4885 | 18,49 | Ordinal |
| | | 16 (16 horas) | 3769 | 14,27 | |
| | | 10 (10 horas) | 3685 | 13,95 | |
| | | Outros Níveis Total | 14080 | 53,29 | |
| | | Total | 26419 | 100,00 | |
| LG_N_Intern_Anterior | 64 | 0 | 14717 | 55,71 | Númerica |
| | | 0,693 | 5375 | 20,35 | |
| | | 1,099 | 2378 | 9,00 | |
| | | Outros Níveis Total | 3949 | 14,95 | |
| | | Total | 26419 | 100,00 | |
| LG_N_Dias_Intern | 179 | 0 | 6892 | 26,09 | Númerica |
| | | 0,693 | 1672 | 6,33 | |
| | | 1,099 | 3114 | 11,79 | |
| | | Outros Níveis Total | 14741 | 55,79 | |
| | | Total | 26419 | 100,00 | |

**Tabela 38: Sumário estatístico descritivo dos atributos quantitativos**

| Atributo | Média | Desvio Padrão | Min | P25 | Mediana | P75 | Máximo |
|---|---|---|---|---|---|---|---|
| LG_N_Intern_Anterior | 0,53 | 0,73 | 0,00 | 0,00 | 0,00 | 0,69 | 4,16 |
| LG_N_Dias_Intern | 1,37 | 1,08 | 0,00 | 0,00 | 1,39 | 1,95 | 7,05 |

**Tabela 39: Código R dos gráficos representados na fase preparação dos dados**

**Diagrama de frequências do atributo Est_Civil (Antes da aplicação do *hot deck*)**

```
hist(d2$Est_Civil, plot = TRUE, main = " s/ hot deck", xlab = "Estado Civil", ylab = "Nº de Ocorrências", col = "gray")
```

**Diagrama de frequências do atributo Est_Civil (Após a aplicação do *hot deck*)**

```
hist(d3$Est_Civil, plot = TRUE, main = " c/ hot deck", xlab = "Estado Civil", ylab = "Nº de Ocorrências", col = "gray")
legend(locator(1), xpd=TRUE, legend=c("Legenda:", "1 - Solteiro(a)", "2 - Casado(a)", "3 - Viuvo(a)", "4 - Divorciado(a)", "5 - Separado(a) Judicialmente", "6 - Falecido(a)", "7 - União de Facto"))
```

**Diagrama de frequências do atributo transformado Escolaridade**

```
hist(d$Escolaridade, plot = TRUE , main = "", xlab = "Escolaridade", ylab = "Nº de Ocorrências", col = "gray")
```

legend(locator(1), xpd=TRUE, legend=c("Legenda:", "1 - Sem habilitações", "2 - Básico (1. Ciclo)", "3 - Básico (2. Ciclo)", "4 - Básico (3. Ciclo)", "5 - Secundário", "6 - Superior"))

**Diagrama de extremos e quartis do atributo transformado Escolaridade**
boxplot(d$Escolaridade, main="", xlab="", ylab="Escolaridade", col="gray")

**Diagrama de frequências do atributo transformado Proc_Principal**
hist(d$Proc_Principal, plot = TRUE , main = "", xlab = "Procedimento Principal", ylab = "Nº de Ocorrências", col = "gray")
legend(locator(1), xpd=TRUE, legend=c("Legenda:", "3 - Olho", "9 - Sistema digestivo", "14 - Sistema músculo-esquelético", "15 - Sistema tegumentar", "16 - Procedimentos diversos"))

**Diagrama de extremos e quartis do atributo transformado Proc_Principal**
boxplot(d$Proc_Principal, main="", xlab="", ylab="Procedimento Principal", col="gray")

**Diagrama de frequências do atributo transformado Diag_Principal**
hist(d$Diag_Principal, plot = TRUE , main = "", xlab = "Diagnóstico Principal", ylab = "Nº de Ocorrências", breaks=36, col = "gray")
legend(locator(1), xpd=TRUE, legend=c("Legenda:", "2 - Neoplasias", "8 - Doenças do sistema respiratório", "12 - Doenças da pele e tecido subcutâneo", "18 - Classificação suplementar"))

**Diagrama de extremos e quartis do atributo transformado Diag_Principal**
boxplot(d$Diag_Principal, main="", xlab="", ylab="Diagnóstico Principal", col="gray")

**Diagrama de frequências do atributo transformado Diag_Inicial**
hist(d$Diag_Inicial, plot = TRUE , main = "", xlab = "Diagnóstico Inicial", ylab = "Nº de Ocorrências", breaks=36, col = "gray")
legend(locator(1), xpd=TRUE, legend=c("Legenda:", "1 - Doenças infecciosas e parasitárias", "6 - Doenças do sistema nervoso e órgãos dos sentidos", "12 - Doenças da pele e tecido subcutâneo", "13 - Doenças do sistema osteomuscular e do tecido conjuntivo", "18 - Classificação suplementar"))

**Diagrama de extremos e quartis do atributo transformado Diag_Inicial**
boxplot(d$Diag_Inicial, main="", xlab="", ylab="Diagnóstico Inicial", col="gray")

**Diagrama de frequências do atributo DiaSemana_Intern**
hist(d$DiaSemana_Intern, plot = TRUE , main = "", xlab = "Dia da Semana", ylab = "Nº de Ocorrências", breaks=22, col = "gray")
legend(locator(1), xpd=TRUE, legend=c("Legenda:", "1 – Segunda-feira", "2 – Terça-feira", "3 – Quarta-feira", "4 – Quinta-feira", "5 – Sexta-feira", "6 - Sábado", "7 - Domingo"))

**Diagrama de extremos e quartis do atributo DiaSemana_Intern**
boxplot(d$DiaSemana_Intern, main="", xlab="", ylab="Dia da Semana", col="gray")

**Diagrama de frequências do atributo transformado Hora_Intern**
hist(d$Hora_Intern, plot = TRUE , main = "", xlab = "Hora de Entrada", ylab = "Nº de Ocorrências", breaks=24, col = "gray")

**Diagrama de extremos e quartis do atributo transformado Hora_Intern**
boxplot(d$Hora_Intern, main="", xlab="", ylab="Hora de Entrada", col="gray")

**Diagrama de frequências para o atributo transformado Hora_Alta_Intern**
hist(d$Hora_Alta_Intern, plot = TRUE , main = "", xlab = "Hora de Alta", ylab = "Nº de Ocorrências", breaks=24, col = "gray")

**Diagrama de extremos e quartis do atributo transformado Hora_Alta_Intern**
boxplot(d$Hora_Alta_Intern, main="", xlab="", ylab="Hora de Alta", col="gray")

**Diagrama de frequências para o atributo N_Dias_Intern**
hist(d$N_Dias_Intern, plot = TRUE, main="", xlab="Nº de Dias de Internamento", ylab="Nº de Ocorrências", breaks=80, col="gray")

**Diagrama de extremos e quartis do atributo N_Dias_Intern**
boxplot(d$N_Dias_Intern, main="", xlab="", ylab="Nº de Dias de Internamento", col="gray")

**Diagrama de frequências do atributo N_Intern_Anterior**
hist(d$N_Intern_Anterior, plot = TRUE, main="", xlab="Nº de Internamentos anteriores", ylab="Nº de Ocorrências", breaks=200, col="gray")

**Diagrama de extremos e quartis do atributo N_Intern_Anterior**
boxplot(d$N_Intern_Anterior, main="", xlab=" ", ylab="Nº de Internamentos anteriores", col="gray")

**Diagrama de frequências do atributo LG_N_Intern_Anterior**
hist(d$LG_N_Intern_Anterior, plot = TRUE, main="", xlab="log(1+x) Nº de Internamentos anteriores", ylab="Nº de Ocorrências", breaks=15, col="gray")

**Diagrama de extremos e quartis do atributo LG_N_Intern_Anterior**
boxplot(d$LG_N_Intern_Anterior, main="", xlab="", ylab="log(1+x) Nº de Internamentos anteriores", col="gray")

**Diagrama de frequências do atributo LG_N_Dias_Intern**
hist(d$LG_N_Dias_Intern, plot = TRUE, main="", xlab="log(1+x) Nº de Dias de Internamento", ylab="Nº de Ocorrências", breaks=30, col="gray")

**Diagrama de extremos e quartis do atributo LG_N_Dias_Intern**
boxplot(d$LG_N_Dias_Intern, main="", xlab="", ylab="log(1+x) Nº de Dias de Internamento", col="gray")

**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")

# D. Modelação

**Tabela 40: Código R utilizado durante a fase de modelação**

**Fase de pré-processamento da BD**

```
library(rminer)
library(randomForest)
d<-read.table("internamento.csv", header=TRUE,sep=",")
```

**Técnica *naive bayes* (naive)**

```
M=mining(LG_N_Dias_Intern~.,data=d,Runs=20,method=c("holdout",2/3),model="naive")
savemining(M,"internamento_naive20.model")

M=mining(LG_N_Dias_Intern~.,data=d,Runs=20,method=c("kfold",5),model="naive")
savemining(M,"internamento_naive20_5.model")
```

**Técnica *multiple regression* (mr)**

```
M=mining(LG_N_Dias_Intern~.,data=d,Runs=20,method=c("holdout",2/3),model="mr")
savemining(M,"internamento_mr20.model")

M=mining(LG_N_Dias_Intern~.,data=d,Runs=20,method=c("kfold",5),model="mr")
savemining(M,"internamento_mr20_5.model")
```

**Técnica *decision tree* (dt)**

```
M=mining(LG_N_Dias_Intern~.,data=d,Runs=20,method=c("holdout",2/3),model="dt")
savemining(M,"internamento_dt20.model")

M=mining(LG_N_Dias_Intern~.,data=d,Runs=20,method=c("kfold",5),model="dt")
savemining(M,"internamento_dt20_5.model")
```

**Técnica *multilayer perceptrons* (mlpe)**

```
M=mining(LG_N_Dias_Intern~.,data=d,Runs=20,method=c("holdout",2/3),model="mlpe",
search="heuristic10")
savemining(M,"internamento_mlpe20.model")

M=mining(LG_N_Dias_Intern~.,data=d,Runs=20,method=c("kfold",5),model="mlpe", search="heuristic10")
savemining(M,"internamento_mlpe20_5.model")
```

**Técnica *randomforest* (rf)**

```
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")
```

**Técnica *support vector machine* (svm)**

```
M=mining(LG_N_Dias_Intern~.,data=d,Runs=20,method=c("holdout",2/3),model="svm",
search="heuristic10")
savemining(M,"internamento_svm20.model")

M=mining(LG_N_Dias_Intern~.,data=d,Runs=20,method=c("kfold",5),model="svm", search="heuristic10")
savemining(M,"internamento_svm20_5.model")
```

**Tabela 41: Resultados obtidos com validação *holdout* e 20 execuções**

| | NV | MR | DT | ANN | RF | SVM |
|---|---|---|---|---|---|---|
| R2 | -0,04669096 | 0,6524699 | 0,6267437 | 0,7409031 | 0,8286387 | 0,7336037 |
| | -0,05641880 | 0,6517805 | 0,6150640 | 0,7440663 | 0,8218465 | 0,7369618 |
| | -0,05500513 | 0,6457841 | 0,6249931 | 0,7477654 | 0,8253164 | 0,7379083 |
| | -0,05487783 | 0,6503327 | 0,6205509 | 0,7463249 | 0,8098628 | 0,7404934 |
| | -0,06356501 | 0,6486697 | 0,6180793 | 0,7426036 | 0,8383057 | 0,7348712 |
| | -0,05623535 | 0,6499161 | 0,6371987 | 0,7376666 | 0,8347857 | 0,7416861 |
| | -0,04837910 | 0,6432731 | 0,6138068 | 0,7406730 | 0,8292113 | 0,7382967 |
| | -0,05263133 | 0,6478438 | 0,6143398 | 0,7380286 | 0,8226925 | 0,7332156 |
| | -0,05438171 | 0,6477227 | 0,6253750 | 0,7360057 | 0,8311569 | 0,7355745 |
| | -0,04940847 | 0,6501575 | 0,6097738 | 0,7428264 | 0,8237934 | 0,7252574 |
| | -0,05691755 | 0,6476035 | 0,6136474 | 0,7375554 | 0,8333447 | 0,7396165 |
| | -0,06014100 | 0,6577497 | 0,5955199 | 0,7538443 | 0,8287830 | 0,7449708 |
| | -0,05891608 | 0,6489493 | 0,6286285 | 0,7108442 | 0,8281883 | 0,7352976 |
| | -0,05146668 | 0,6378933 | 0,6215258 | 0,7436937 | 0,8209866 | 0,7313900 |
| | -0,06548640 | 0,6385260 | 0,6209488 | 0,7520184 | 0,8201786 | 0,7478368 |
| | -0,05766173 | 0,6509563 | 0,6279362 | 0,7599162 | 0,8229844 | 0,7453616 |
| | -0,06246588 | 0,6415014 | 0,6143268 | 0,7253029 | 0,8298892 | 0,7444987 |
| | -0,05365034 | 0,6484247 | 0,6157865 | 0,7455841 | 0,8184862 | 0,7370364 |
| | -0,05671335 | 0,6357785 | 0,6174996 | 0,7319901 | 0,8205158 | 0,7399511 |
| | -0,06036102 | 0,6506672 | 0,6225606 | 0,7500785 | 0,8193573 | 0,7442787 |
| MAE | 0,8927108 | 0,4383278 | 0,4086104 | 0,3323438 | 0,2138436 | 0,3035733 |
| | 0,8862936 | 0,4365221 | 0,4146870 | 0,3396142 | 0,2134441 | 0,3067897 |
| | 0,8892353 | 0,4444972 | 0,4143128 | 0,3391393 | 0,2138913 | 0,3058119 |
| | 0,8858299 | 0,4414777 | 0,4198808 | 0,3369207 | 0,2185186 | 0,3001745 |
| | 0,8875850 | 0,4411055 | 0,4140276 | 0,3297131 | 0,2060970 | 0,3052046 |
| | 0,8877399 | 0,4449576 | 0,4037143 | 0,3337120 | 0,2084500 | 0,3033194 |
| | 0,8914424 | 0,4442714 | 0,4143152 | 0,3428179 | 0,2168393 | 0,2994787 |
| | 0,9031157 | 0,4403000 | 0,4162001 | 0,3274701 | 0,2220826 | 0,3183619 |
| | 0,8880876 | 0,4442485 | 0,4121116 | 0,3352975 | 0,2150833 | 0,3051810 |
| | 0,8773868 | 0,4425012 | 0,4343231 | 0,3299912 | 0,2169475 | 0,3113253 |
| | 0,8763068 | 0,4363931 | 0,4137121 | 0,3357340 | 0,2103773 | 0,3053424 |
| | 0,8824139 | 0,4325991 | 0,4430543 | 0,3265377 | 0,2138096 | 0,3046989 |
| | 0,8939933 | 0,4405917 | 0,4037176 | 0,3561030 | 0,2097875 | 0,3064671 |
| | 0,8888883 | 0,4496790 | 0,4280899 | 0,3335813 | 0,2132248 | 0,3100571 |
| | 0,8861883 | 0,4437960 | 0,4124882 | 0,3247242 | 0,2197694 | 0,3012330 |
| | 0,8815953 | 0,4389958 | 0,4140060 | 0,3225965 | 0,2194683 | 0,3006471 |
| | 0,8931200 | 0,4465224 | 0,4204299 | 0,3494597 | 0,2113967 | 0,3008730 |
| | 0,8919649 | 0,4405470 | 0,4117665 | 0,3293380 | 0,2207197 | 0,3056674 |
| | 0,8825445 | 0,4501825 | 0,4139102 | 0,3405061 | 0,2170731 | 0,3029439 |
| | 0,8878674 | 0,4402403 | 0,4134122 | 0,3278837 | 0,2223512 | 0,3022145 |
| RMSE | 1,113982 | 0,6372771 | 0,6647070 | 0,5514474 | 0,4523659 | 0,5577456 |
| | 1,112644 | 0,6342491 | 0,6717117 | 0,5557253 | 0,4626820 | 0,5577236 |
| | 1,114346 | 0,6484690 | 0,6653332 | 0,5510407 | 0,4486760 | 0,5560911 |
| | 1,107270 | 0,6431112 | 0,6727455 | 0,5533340 | 0,4766895 | 0,5452650 |
| | 1,115780 | 0,6480229 | 0,6694834 | 0,5496217 | 0,4327039 | 0,5534651 |
| | 1,107587 | 0,6459041 | 0,6556591 | 0,5558425 | 0,4362935 | 0,5536323 |
| | 1,115974 | 0,6523978 | 0,6735634 | 0,5566457 | 0,4476962 | 0,5483076 |
| | 1,127965 | 0,6401695 | 0,6730312 | 0,5518024 | 0,4522482 | 0,5627594 |
| | 1,112678 | 0,6406560 | 0,6648755 | 0,5588861 | 0,4488863 | 0,5502290 |
| | 1,094924 | 0,6435050 | 0,6707856 | 0,5519492 | 0,4549794 | 0,5700789 |
| | 1,098101 | 0,6427985 | 0,6698208 | 0,5580285 | 0,4408184 | 0,5558457 |
| | 1,105522 | 0,6295429 | 0,6963035 | 0,5403088 | 0,4509889 | 0,5453749 |
| | 1,125107 | 0,6442799 | 0,6509599 | 0,5926845 | 0,4457262 | 0,5593327 |
| | 1,106547 | 0,6550658 | 0,6649068 | 0,5480310 | 0,4605359 | 0,5629138 |
| | 1,108619 | 0,6475719 | 0,6723232 | 0,5440443 | 0,4663587 | 0,5441724 |
| | 1,106341 | 0,6470339 | 0,6639170 | 0,5302433 | 0,4560720 | 0,5461188 |
| | 1,113117 | 0,6482724 | 0,6833065 | 0,5790695 | 0,4447053 | 0,5448039 |

| | NV | MR | DT | ANN | RF | SVM |
|---|---|---|---|---|---|---|
| | 1,112187 | 0,6439842 | 0,6723340 | 0,5443133 | 0,4650146 | 0,5560943 |
| | 1,110039 | 0,6529848 | 0,6728133 | 0,5584503 | 0,4614197 | 0,5496469 |
| | 1,115217 | 0,6373186 | 0,6704098 | 0,5434460 | 0,4621922 | 0,5535021 |

**Tabela 42: Resultados obtidos com validação *k-fold* (k=5) e 20 execuções**

| | NV | MR | DT | ANN | RF | SVM |
|---|---|---|---|---|---|---|
| R2 | -1,197211e-04 | 0,6458022 | 0,6234532 | 0,7419552 | 0,8289481 | 0,7406241 |
| | -2,620294e-05 | 0,6460202 | 0,6200754 | 0,7478732 | 0,8303922 | 0,7404327 |
| | -8,367981e-05 | 0,6460419 | 0,6236132 | 0,7473261 | 0,8301434 | 0,7405086 |
| | -3,926603e-05 | 0,6463104 | 0,6203803 | 0,7417724 | 0,8302789 | 0,7413322 |
| | -1,196195e-04 | 0,6463270 | 0,6236578 | 0,7389801 | 0,8299879 | 0,7410859 |
| | -1,524299e-04 | 0,6461651 | 0,6234591 | 0,7414686 | 0,8295603 | 0,7414890 |
| | -4,247749e-05 | 0,6459972 | 0,6204441 | 0,7442879 | 0,8314140 | 0,7411627 |
| | -5,201262e-05 | 0,6462424 | 0,6209877 | 0,7372672 | 0,8293056 | 0,7400338 |
| | -2,961031e-04 | 0,6455099 | 0,6221898 | 0,7423304 | 0,8283974 | 0,7412038 |
| | -7,452337e-05 | 0,6462972 | 0,6231359 | 0,7418818 | 0,8289224 | 0,7406887 |
| | -5,972553e-05 | 0,6453764 | 0,6229096 | 0,7389496 | 0,8297018 | 0,7422008 |
| | -5,964306e-06 | 0,6458626 | 0,6231734 | 0,7408516 | 0,8290655 | 0,7412713 |
| | -1,081220e-04 | 0,6454153 | 0,6209586 | 0,7398082 | 0,8297690 | 0,7412740 |
| | -2,951489e-05 | 0,6462146 | 0,6182594 | 0,7447267 | 0,8299411 | 0,7415596 |
| | -1,365611e-04 | 0,6459041 | 0,6203370 | 0,7370739 | 0,8296175 | 0,7413101 |
| | -8,927121e-05 | 0,6463795 | 0,6205486 | 0,7413174 | 0,8281786 | 0,7403202 |
| | -4,420304e-05 | 0,6462785 | 0,6222395 | 0,7446640 | 0,8302546 | 0,7411442 |
| | -1,177770e-04 | 0,6461977 | 0,6235436 | 0,7450299 | 0,8302903 | 0,7420771 |
| | -6,675415e-06 | 0,6465413 | 0,6210030 | 0,7397944 | 0,8299746 | 0,7411983 |
| | -5,403874e-05 | 0,6454925 | 0,6234169 | 0,7427420 | 0,8278059 | 0,7414455 |
| MAE | 0,8610653 | 0,4423664 | 0,4133627 | 0,3333341 | 0,2108113 | 0,3008043 |
| | 0,8610264 | 0,4421625 | 0,4177911 | 0,3308932 | 0,2082247 | 0,3009155 |
| | 0,8610259 | 0,4422218 | 0,4134568 | 0,3302405 | 0,2079770 | 0,3013276 |
| | 0,8610151 | 0,4421219 | 0,4174872 | 0,3321595 | 0,2083956 | 0,3007210 |
| | 0,8610410 | 0,4422346 | 0,4126553 | 0,3367458 | 0,2090714 | 0,3006655 |
| | 0,8610704 | 0,4420142 | 0,4133865 | 0,3332152 | 0,2090753 | 0,3007770 |
| | 0,8610160 | 0,4422632 | 0,4175499 | 0,3313224 | 0,2069634 | 0,3005339 |
| | 0,8610243 | 0,4421988 | 0,4171637 | 0,3396737 | 0,2094289 | 0,3015329 |
| | 0,8611809 | 0,4423721 | 0,4136959 | 0,3335220 | 0,2104933 | 0,3007525 |
| | 0,8610440 | 0,4420557 | 0,4134825 | 0,3341892 | 0,2098407 | 0,3010862 |
| | 0,8610403 | 0,4427537 | 0,4133558 | 0,3345371 | 0,2086819 | 0,3019582 |
| | 0,8610055 | 0,4424390 | 0,4134793 | 0,3326710 | 0,2086501 | 0,3007289 |
| | 0,8610585 | 0,4423505 | 0,4177620 | 0,3350239 | 0,2086196 | 0,3010962 |
| | 0,8610121 | 0,4420534 | 0,4216429 | 0,3327127 | 0,2086657 | 0,3005380 |
| | 0,8610450 | 0,4423000 | 0,4174343 | 0,3372805 | 0,2088171 | 0,3008300 |
| | 0,8610349 | 0,4420822 | 0,4171997 | 0,3337543 | 0,2102386 | 0,3008288 |
| | 0,8610388 | 0,4421616 | 0,4134863 | 0,3317913 | 0,2087937 | 0,3009098 |
| | 0,8610669 | 0,4422066 | 0,4133101 | 0,3314603 | 0,2091375 | 0,3002200 |
| | 0,8610087 | 0,4420664 | 0,4174973 | 0,3358837 | 0,2086389 | 0,3002694 |
| | 0,8610348 | 0,4424002 | 0,4134159 | 0,3324083 | 0,2106591 | 0,3006173 |
| RMSE | 1,084847 | 0,6456030 | 0,6656595 | 0,5510490 | 0,4486489 | 0,5524684 |
| | 1,084797 | 0,6454043 | 0,6686384 | 0,5446934 | 0,4467510 | 0,5526722 |
| | 1,084828 | 0,6453845 | 0,6655180 | 0,5452841 | 0,4470785 | 0,5525913 |
| | 1,084804 | 0,6451397 | 0,6683701 | 0,5512441 | 0,4469002 | 0,5517137 |
| | 1,084847 | 0,6451245 | 0,6654786 | 0,5542164 | 0,4472831 | 0,5519764 |
| | 1,084865 | 0,6452721 | 0,6656542 | 0,5515682 | 0,4478452 | 0,5515465 |
| | 1,084805 | 0,6454253 | 0,6683139 | 0,5485526 | 0,4454032 | 0,5518944 |
| | 1,084811 | 0,6452017 | 0,6678351 | 0,5560320 | 0,4481797 | 0,5530966 |
| | 1,084943 | 0,6458694 | 0,6667752 | 0,5506482 | 0,4493705 | 0,5518506 |
| | 1,084823 | 0,6451517 | 0,6659399 | 0,5511272 | 0,4486826 | 0,5523996 |

| | NV | MR | DT | ANN | RF | SVM |
|---|---|---|---|---|---|---|
| | 1,084815 | 0,6459910 | 0,6661398 | 0,5542488 | 0,4476594 | 0,5507866 |
| | 1,084786 | 0,6455480 | 0,6659067 | 0,5522261 | 0,4484948 | 0,5517786 |
| | 1,084841 | 0,6459555 | 0,6678608 | 0,5533367 | 0,4475710 | 0,5517758 |
| | 1,084798 | 0,6452270 | 0,6702345 | 0,5480817 | 0,4473446 | 0,5514711 |
| | 1,084856 | 0,6455101 | 0,6684082 | 0,5562365 | 0,4477701 | 0,5517373 |
| | 1,084831 | 0,6450766 | 0,6682219 | 0,5517295 | 0,4496568 | 0,5527919 |
| | 1,084806 | 0,6451687 | 0,6667314 | 0,5481490 | 0,4469322 | 0,5519142 |
| | 1,084846 | 0,6452425 | 0,6655796 | 0,5477561 | 0,4468852 | 0,5509188 |
| | 1,084786 | 0,6449290 | 0,6678216 | 0,5533512 | 0,4473006 | 0,5518565 |
| | 1,084812 | 0,6458852 | 0,6656915 | 0,5502082 | 0,4501443 | 0,5515929 |

# E. Avaliação

**Tabela 43: Código R utilizado durante a fase de avaliação**

**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")
```

**Tabela 44: Código R para obtenção do *t.test***

***Student Test (t.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)

----- output
Welch Two Sample t-test

data:  mRF and mSVM
t = 388,2957, df = 32,395, p-value < 2,2e-16
alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:
0,08801539
0,08894324

sample estimates:
mean of x: 0,8295975
mean of y: 0,7411181

cat("p-value:",t.test(mRF,mSVM)$p.value)
p-value: 5.403861e-61
```

**Tabela 45: Código R para obtenção da curva REC**

***Regression Error Characteristic curve***

```
NV=loadmining("internamento_naive20_5.model")
MR=loadmining("internamento_mr20_5.model")
DT=loadmining("internamento_dt20_5.model")
ANN=loadmining("internamento_mlpe20_5.model")
RF=loadmining("internamento_randomforest20_5.model")
SVM=loadmining("internamento_svm20_5.model")

L=vector("list",6); L[[1]]=RF; L[[2]]=MR; L[[3]]=DT; L[[4]]=ANN; L[[5]]=NV; L[[6]]=SVM;
```

```
mgraph(L,graph="REC",xval=3,leg=list(pos=c(2.25,0.4),leg=c("rf","mr","dt","ann","naive","svm")), Grid=10,
main="Curva REC" )
```

**Tabela 46: Código R para definir valor de tolerância de erro absoluto**

**Tolerância da curva REC**
**Tolerância de 0,25**

```
M=loadmining("internamento_randomforest20_5.model")

print(mmetric(M,metric=c("TOLERANCE","NAREC"), val=0.25))
TOLERANCE   NAREC
0,7606040        0,6247462
0,7638322        0,6298921
0,7644232        0,6314455
0,7647059        0,6289666
0,7643049        0,6282899
0,7649221        0,6282307
0,7663927        0,6326756
0,7639925        0,6281477
0,7612080        0,6261203
0,7607975        0,6260234
0,7630837        0,6308912
0,7656766        0,6307469
0,7645978        0,6297126
0,7628523        0,6299837
0,7633224        0,6286532
0,7607031        0,6266708
0,7626277        0,6279528
0,7621123        0,6276205
0,7637524        0,6288242
0,7608558        0,6264278

Tol=meanint(mmetric(M,metric="TOLERANCE", val=0.25,aggregate = "no"))
Nar=meanint(mmetric(M,metric="NAREC", val=0.25,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= 0,763± 0,001**
**NAREC= 0,629± 0,001**

```
T=mmetric(M,metric="TOLERANCE",val=0.25)
cat(mean(T),"\n")
```

**TOLERANCE= 0,7632384**

**Tolerância de 0,5**

```
M=loadmining("internamento_randomforest20_5.model")

print(mmetric(M,metric=c("TOLERANCE","NAREC"), val=0.5))
TOLERANCE   NAREC
0,8650581        0,7220393
0,8659918        0,7259107
0,8681946        0,7268923
0,8672530        0,7258487
0,8662594        0,7250342
0,8666386        0,7252324
0,8671774        0,7279855
```

```
0,8660881       0,7248482
0,8654615       0,7231630
0,8664296       0,7239050
0,8657093       0,7259516
0,8668112       0,7269441
0,8672899       0,7260000
0,8667961       0,7259052
0,8671095       0,7254295
0,8676136       0,7235060
0,8668846       0,7250356
0,8668913       0,7245134
0,8665692       0,7255765
0,8662635       0,7232172

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= 0,867 ± 0,000**
**NAREC= 0,725 ±0,001**

```
T=mmetric(M,metric="TOLERANCE",val=0.5)
cat(mean(T),"\n")
```

**TOLERANCE= 0,8666245**

## Tabela 47: Código R para obtenção do gráfico RSC

***Regression Scatter Characteristic curve***
**TOLERANCE = 0,25**

```
T=0.25 # valor para a tolerancia a admitir
```

**TOLERANCE = 0,5**

```
T=0.5 # valor para a tolerancia a admitir

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="RSC: Gráfico
de Dispersão")
        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="red",pch=19,cex=0.7) -- em que T=0.25
                points(X[I],Y[I],col="black",pch=19,cex=0.7) – em que T=0.5
        }
 }
```

**Tabela 48: Previsão de erro máximo nos extremos**

***Erro Máximo nos Extremos Inferior e Superior***
**TOLERANCE = 0,5**

A – Calculo para o extremo inferior
B – Calculo para o extremo superior

As previsões são sobre uma escala logarítmica
y=log(x+1)

Função inversa:
x=exp(y)-1

**A:**
0.5=log(x1+1)-log(x2+1)

Notas:
- 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=0.64 => desvio em dias normais para A

**B:**de acordo com a informação visível na Figura 28:

4.5=log(x1+1)
4.0=log(x2+1)

Notas:
- Valor mais alto é 4.5 e a tolerância é 0.5
- log(x1+1)=4.5 =>aplicando exp(4.5)-1 =>x1=89.01713
- log(x2+1)=4.0 =>aplicando exp(4.0)-1 =>x2=53.59815

x1-x2=35.41898 => desvio em dias normais para B

**TOLERANCE = 0,25**

A – Calculo para o extremo inferior
B – Calculo para o extremo superior

As previsões são sobre uma escala logarítmica
y=log(x+1)

Função inversa:
x=exp(y)-1

**A:**
0.25=log(x1+1)-log(x2+1)

Notas:
- Erro máximo => 0.25
- log(x2+1)=0 => aplicando exp(0)-1 => x2=0
- log(x1+1)=0.25 => aplicando exp(0.25)-1 =>x1=0.2840254

significa que
x1-x2=0.64 => desvio em dias normais para A

**B:** de acordo com a informação visível na Figura 28:

4.5=log(x1+1)
4.0=log(x2+1)

Notas:
- Valor mais alto é 4.5 e a tolerância é 0.25
- log(x1+1)=4.5 => aplicando exp(4.5)-1 => x1=89.01713
- log(x2+1)=4.25 =>aplicando exp(4.25)-1 =>x2=69.10541

x1-x2=19.91172 => desvio em dias normais para B

**Tabela 49: Código R para obtenção do gráfico IMP**

***Relative Input Importance Barplot***

```
library(rminer)
library(randomForest)

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("internamento.csv", header=TRUE,sep=",") # ler a bd

Imp=Importance (M, data=d, method="DSA")
print(round(Imp$imp,digits=3))
 [1] 0.009 0.051 0.047 0.032 0.033 0.261 0.123 0.109 0.051 0.080 0.050 0.011
[13] 0.046 0.011 0.005 0.036 0.014 0.031 0.000

L=list(runs=1,sen=t(Imp$imp),sresponses=Imp$sresponses)
mgraph(L,graph="IMP",leg=names(d),col="gray",Grid=10)
```

**Tabela 50: Código R para obtenção do gráfico VEC**

***Variable Effect Curve***

```
library(rminer)
library(randomForest)

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("internamento.csv", header=TRUE,sep=",") # ler a bd

Imp=Importance (M, data=d, method="DSA")

vecplot(Imp,graph="VEC",xval=6,Grid=50,main="", xlab="Tipo de Episódio de Internamento", ylab="log (1+x) Nº de Dias de Internamento")
vecplot(Imp,graph="VEC",xval=7,Grid=50,main="", xlab="Serviço de Internamento", ylab="log (1+x) Nº de Dias de Internamento")
legend(locator(1), xpd=TRUE, legend=c("Legenda:", "1 - Serv. Medicina", "2 - Serv. Cirurgia", "3 - Serv. Ortopedia", "6 - Serv. Pneumologia", "8 - Serv. Especialidades"))
vecplot(Imp,graph="VEC",xval=8,Grid=50,main="", xlab="Especialidade Médica", ylab="log (1+x) Nº de Dias de Internamento")
legend(locator(1), xpd=TRUE, legend=c("Legenda:", "120 - Cirurgia Geral", "210 - Medicina Interna", "250 - Ortopedia", "300 - Urologia"))
```