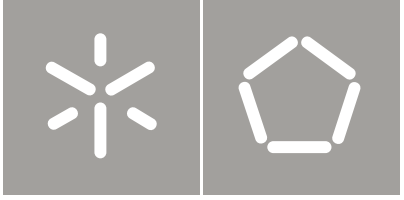
Universidade do Minho
Escola de Engenharia

João Manuel de Campos Gonçalves

Utilização de técnicas de data mining na previsão do plano terapêutico em medicina intensiva

Outubro de 2012



Universidade do Minho
Escola de Engenharia

João Manuel de Campos Gonçalves

Utilização de técnicas de data mining na previsão do plano terapêutico em medicina intensiva

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

Trabalho efetuado sob a orientação do
Professor Doutor Manuel Filipe Santos
Mestre Carlos Filipe Portela

Outubro de 2012

Universidade do Minho, outubro de 2012

Assinatura: _____

Agradecimentos

Ao orientador, Professor Manuel Filipe Santos, pelas suas orientações e ensinamentos ao longo desde trabalho.

Ao coorientador, Mestre Filipe Portela, pelos esclarecimentos, orientações e disponibilidade demonstrada.

Ao Dr. Álvaro Silva, médico do Hospital de Santo António, pelos esclarecimentos e orientações na área médica.

Ao colega José Miguel Nunes pela amizade e inestimável apoio.

À Neide, deixo um agradecimento muito especial pela presença e incentivo fundamentais no desenvolvimento deste projeto.

Ao meu pai (in memoriam) pelo incentivo durante toda a minha vida e em especial na fase inicial deste projeto.

Por último, mas não menos importante, quero deixar um agradecimento especial à minha mãe, irmãs, afilhada Sara e cunhado Jorge, pelo apoio incondicional e por estarem sempre presentes.

Utilização de técnicas de *data mining* na previsão do plano terapêutico em medicina intensiva

Resumo

Um dos principais dilemas existentes na medicina intensiva prende-se com o plano terapêutico, mais concretamente, que medicamentos e quando é que estes devem ser administrados a um doente. No plano terapêutico, a interpretação rápida e avaliação precisa de dados fisiológicos, são cruciais para uma tomada de decisão mais eficiente e eficaz por parte dos médicos.

No sentido de apoiar a decisão dos médicos, este trabalho tem como objetivo prever o nível de sépsis e a melhor terapêutica para doentes com problemas microbiológicos, baseados nos níveis de sépsis. Para isso, foi desenvolvido um conjunto de modelos de *Data Mining* (DM), utilizando técnicas de previsão e modelos classificação, que irão possibilitar o médico decidir qual a terapêutica adequada a aplicar, bem como aquela que apresente uma elevada taxa de sucesso. Os dados utilizados nos modelos de DM foram recolhidos no Serviço de Cuidados Intensivos do Hospital de Santo António, Porto, Portugal.

Nesta dissertação, foi utilizada a tarefa de previsão, o modelo de classificação, o método de aprendizagem supervisionada e os algoritmos: Árvores de Decisão, Máquinas de Vetores de Suporte e o classificador *Naïve Bayes*, para prever o nível de sépsis e o plano terapêutico de doentes com sépsis. Relativamente à avaliação, utilizaram-se a Matriz de Confusão, incluindo as métricas associadas e a *Cross-validation*. De entre as métricas associadas na análise, foram utilizadas: a taxa de erro total, a sensibilidade, a especificidade e a acuidade, que permitiram identificar quais as medidas mais relevantes para a previsão do nível da sépsis e do plano terapêutico em estudo.

Concluindo, foi possível prever com grande acuidade o nível de sépsis, no entanto, o mesmo já não é possível dizer no que diz respeito à medicação. Apesar de os modelos da sépsis terem bons resultados, o plano terapêutico não apresenta o mesmo nível de acuidade. Os resultados provam que de uma forma geral existe uma fraca correlação entre o nível de sépsis e o plano terapêutico, referente ao grupo de medicamentos. No entanto, é de salientar que para alguns grupos de medicamentos, os modelos tiveram uma bom desempenho (nível de acertos em algumas classes foi superior a 80%).

Palavras-Chave: *Data mining*; Modelos de classificação; Cuidados intensi-

vos; Plano terapêutico; Sépsis; CRISP-DM

Using data mining techniques to predict the therapeutic plan in intensive care medicine

Abstract

One of the main problems existing in intensive medicine is related to the therapeutic plan, particularly what and when drugs must be administered to a patient. In the therapeutic plan it is crucial to make a rapid interpretation and accurate assessment of physiological data for efficient and effective decision-making by doctors.

The present investigation aims to support doctor's decision-making on predicting sepsis level and the best treatment for patients with microbiological problems based on sepsis levels. Thus, a set of Data Mining (DM) models was developed using forecasting techniques and classification models which will enable a doctor's decision about appropriate therapy to apply, as well as the most successful one. The data used in DM models were collected at the Department of Intensive Care of the Hospital de Santo António, in Oporto, Portugal.

Classification DM models where considered to predict sepsis level and therapeutic plan for patients with sepsis in a supervised learning approach. Models were induced making use of the following algorithms: Decision Trees, Support Vector Machines and Naïve Bayes classifier. Confusion Matrix, including associated metrics, and Cross-validation were used for the evaluation. Analysis of the total error rate, sensitivity, specificity and accuracy were the associated metrics used to identify the most relevant measures to predict sepsis level and treatment plan under study.

In conclusion, it was possible to predict with great accuracy the sepsis level, but not the medication. Although the good sepsis models results attained, therapeutic plan does not present the same level of accuracy. The results have showed that in general there is a small correlation between sepsis level and therapeutic plan, considering the drugs group. However, for some drugs groups models the results are interesting (some classes exceeded 80% in terms of the accuracy level).

Keywords: Data mining; Classification models; Intensive care; Therapeutic plan; Sepsis; CRISP-DM

Conteúdo

Lista de Acrónimos

AD Árvores de Decisão

AG Algoritmos Genéticos

APACHE Acute Physiology and Chronic Health Evaluation

AR Action Research

BD Base de dados

BI Business Intelligence

CRISP-DM CRoss-Industry Standard Process for Data Mining

CV Cross-validation

DCBD Descoberta de Conhecimento em Bases de Dados

DM Data Mining

DW Data Warehouses

ECG Escala de Coma Glasgow

HSA Hospital de Santo António

MC Matriz de Confusão

MI Medicina Intensiva

MVS Máquinas de Vetores de Suporte

ODM Oracle Data Mining

OLAP On-Line Analytical Processing

RNA Redes Neurais Artificiais

ROC Receiver Operating Characteristic

RS Rough Sets

SAD Sistema de Apoio à Decisão

SAPS Simplified Acute Physiology Score

SCI Serviço de Cuidados Intensivos

SOFA Sequential Organ Failure Assessment

SSC Surviving Sepsis Campaign

TI Tecnologias de Informação

TIC Tecnologias de Informação e Comunicação

UCI Unidade de Cuidados Intensivos

Lista de Figuras

Lista de Tabelas

1 Introdução

O interesse nas áreas da Descoberta de Conhecimento em Bases de Dados (DCBD) e de *Data Mining* (DM) emergiu devido aos avanços das Tecnologias de Informação e Comunicação (TIC), sendo que hoje em dia é fácil coletar, armazenar, processar e partilhar dados (Cruz and Cortez, 2009). Assim, tem-se assistido a um crescimento exponencial da quantidade de dados armazenados, sendo que muitos destes dados contêm informação valiosa (padrões ou tendências) que podem auxiliar a tomada de decisão (Turban et al., 2008).

Para Turban et al. (2008), o DM tornou-se numa ferramenta popular, muito bem-sucedida e útil em muitas áreas, como é o caso dos cuidados de saúde. Esta pode ser útil na identificação de novos padrões, melhorar a capacidade de sobrevivência dos doentes, descobrir as relações entre sintomas e doenças (como doenças e tratamentos bem-sucedidos). Esta ferramenta é necessária para ajudar os profissionais médicos a tomar decisões informadas e corretas em tempo útil (Turban et al., 2011).

Devido à limitação dos especialistas humanos, é necessário recorrer a ferramentas (semi)automatizadas de DM para analisar os dados em estado bruto e extrair informações de alto nível para os decisores (Turban et al., 2008).

Neste projeto, pretende-se mostrar de que forma é que a área do DM pode ser útil na Medicina Intensiva (MI).

1.1 Enquadramento

Ao longo dos últimos 100 anos, grandes avanços têm sido feitos relativamente à sépsis[1], nomeadamente na terapêutica, mas apesar desses avanços, a taxa de mortalidade permanece inaceitavelmente alta (Vincent and Abraham, 2006).

Num estudo feito em Portugal, em dezassete Unidades de Cuidados Intensivos (UCI) (i.e. 41% das UCIs), onde participaram todos os doentes adultos internados entre dezembro de 2004 e novembro de 2005, concluiu-se que 22% dos internamentos em UCI são devidos a Sépsis (Póvoa et al., 2009). Assim, 9% dos doentes apresentaram sépsis, 40% sépsis grave e 51% por choque séptico, dando origem a uma taxa de mortalidade de 30%, numa taxa hospitalar global de 38% (Póvoa et al., 2009).

Dados recentes, dos Estados Unidos da América (EUA), indicam que a incidência da sépsis representa uma estimativa de 751000 casos/ano, com

[1] Ver definição em 2.2.1.

aumentos anuais de pelo menos 1,5% (Angus et al., 2001).

Para reduzir a mortalidade da sépsis, existe um conjunto de procedimentos a serem tomados, realizados numa fase precoce, porque sabemos hoje, por exemplo, que por cada hora de atraso no início da cada antibioterapia efetiva, está associado uma diminuição média de sobrevivência de 7,6% (Kumar et al., 2006).

A elaboração de um plano terapêutico para a sépsis pode resultar não só na diminuição da mortalidade, mas também, na redução substancial de custos para as instituições, devido à possível melhoria na utilização dos recursos existentes (Shorr et al., 2007).

Nos dias de hoje, há uma infinidade de tecnologia nos hospitais, em particular, nas UCI. Na prática, as UCI, estão sempre na vanguarda no uso das Tecnologias de Informação (TI), mas, ainda existe um número reduzido que não dispõe de um sistema de gestão de dados (Meyfroidt, 2009). Os dados clínicos produzidos todos os dias podem ser integrados num sistema de apoio à decisão (SAD) em tempo real para melhorar a qualidade do atendimento aos doentes (Portela et al., 2011). No entanto, há muitos aspetos sensíveis que devem ser levados em conta, principalmente, qualidade dos dados e a integração de fontes de dados heterogéneas (Portela et al., 2011).

Um dos principais problemas existentes na MI prende-se com as terapêuticas, isto é, quando estas devem ser administradas a um doente. As TI podem ter um papel muito importante no apoio à qualidade e eficiência dos cuidados de saúde, fornecendo a informação certa no momento certo à pessoa certa (Handel and Hackman, 2010 & Nowinski et al., 2007). Num ambiente sensível como a UCI, onde a condição clínica dos doentes é fundamental, com vidas em risco, há uma forte pressão para a rápida tomada de decisões, estando muitas vezes a informação incompleta (De Turck et al., 2007).

Devido à condição crítica dos doentes e da enorme quantidade de dados, pode ser difícil para os médicos, decidirem sobre o melhor procedimento a ter, com a finalidade de fornecer-lhes os melhores cuidados de saúde possíveis (Pereira et al., 2007). Por outro lado, o fator humano pode levar a erros no processo decisório, não havendo tempo suficiente para analisar a situação, devido a circunstâncias stressantes (Pereira et al., 2007). Além disso, não é possível analisar e memorizar de forma contínua, todos os dados com a frequência da sua recolha (Pereira et al., 2007).

Assim, no plano terapêutico, a interpretação rápida e avaliação precisa

de dados fisiológicos, na monitorização do estado dos doentes nos cuidados intensivos, vão ser cruciais para uma mais eficiente e eficaz tomada de decisão por parte dos médicos.

Com o objetivo de apoiar a decisão dos clínicos, este trabalho teve como foco, a previsão da sépsis e a previsão do plano terapêutico para doentes com problemas microbiológicos, baseados nos níveis de sépsis.

1.2 Metodologia de investigação

Durante algum tempo, os investigadores dos Sistemas de Informação (SI), foram encorajados a considerar a metodologia de investigação *Action Research* (AR) como uma abordagem adequada de investigação de entre as várias metodologias adotadas pelos SI (West et al., 1985). Segundo Avison et al. (1999), a AR tem características ideais para o estudo de sistemas de informação com um impacto significativo na área.

Neste trabalho, optou-se por esta metodologia por se enquadrar no desenvolvimento do projeto e por utilizar um método cíclico sistemático de planeamento: ação, observação, reflexão e avaliação (O'brien, 1998 & McNiff, 2002).

Com o objetivo de desenvolver um conjunto de modelos de decisão que possibilitarão o médico decidir a melhor terapêutica a aplicar ao doente, houve necessidade de fazer o reconhecimento e levantamento de factos. Numa abordagem para descobrir mais sobre a natureza, contexto, relevância e resolução do problema, iniciou-se a implementação de uma série de medidas de ação.

Na fase inicial, após se ter decidido sobre a área de interesse, foi elaborada a questão de investigação onde se baseou a revisão de literatura relevante. Com o propósito de encontrar respostas para a questão de investigação, a pesquisa de trabalhos relacionados foi muito importante para a análise do atual estado da arte. Posteriormente, foi feito o estudo e preparação dos dados, depois foi elaborada uma proposta de modelo teórico que permitiu responder ao problema apresentado. Foi também efetuado um levantamento dos indicadores necessários para o desenvolvimento dos modelos. Depois foram estudadas e definidas as variáveis que influenciam a terapêutica. Desenvolveram-se modelos de DM para a previsão e descoberta de padrões de tratamento de sépsis. Finalmente, foram realizados testes em modelos com base em dados reais, que foram recolhidos em tempo real e tratados *online*.

1.3 Âmbito da revisão bibliográfica

Com o propósito de aprofundar a aprendizagem dos conceitos para trazer maior clareza ao estudo, inicialmente foi feita uma revisão de literatura sobre o tema, pois é através dela que se situa o trabalho, sendo muito importante contextualizá-lo através do atual estado da arte. Houve necessidade de situar o trabalho numa linha de pesquisas onde este se insere, com base nos tópicos, autores e data, para a revisão de literatura.

A revisão bibliográfica foi iniciada através da consulta no portal web da *Surviving Sepsis Campaign* (SSC, 2010), que é atualmente o local onde estão escritas as orientações internacionais para o tratamento da sépsis, por forma a enquadrar o melhor possível o problema a investigar.

No entanto, houve necessidade de ler aquilo que há de mais atual sobre o assunto, para isso foi feita uma pesquisa nas fontes secundárias, por tópico ou palavras-chave, nos seguintes serviços, pela ordem indicada:

Google scholar - http://scholar.google.pt/

RCAAP - http://www.rcaap.pt/

Web of Knowledge - http://webofknowledge.com

SCOPUS - http://www.scopus.com/home.url

NDLTD - http://www.ndltd.org/

Driver - http://www.driver-repository.eu/

Feita a pesquisa, através do tópico ou palavras-chave: *Data Mining*; *Classification models*; *Intensive care*; *Therapeutic plan* e *Sepsis*. Foram encontrados artigos em revistas periódicas, teses de doutoramento e dissertações de mestrado.

Os conceitos relevantes foram: Técnicas de *Data Mining*; Modelos de classificação sépsis; Plano terapêutico; Cuidados intensivos e Definições de sépsis.

Os critérios de seleção das referências bibliográficas, foram: o número de citações e as citações mais atuais, embora houvesse necessidade de sustentar algum conceito com referências mais antigas.

1.4 Estrutura do documento

Em conformidade com a estrutura referencial disponibilizada pela direção de mestrado, esta dissertação possui nove capítulos:

Introdução Neste capítulo introdutório é apresentado todo o enquadramento teórico e relevância do trabalho, bem como as motivações que levaram ao desenvolvimento da investigação. É também apresentada a metodologia de investigação utilizada e uma breve descrição sobre o âmbito da revisão bibliográfica.

Problema estudado Neste capítulo é apresentada uma revisão das definições dos três níveis de sépsis e dos conceitos sobre sépsis, conforme as diretrizes internacionais de consenso para o seu tratamento.

Revisão de literatura Neste capítulo é apresentado o sistema de classificação PIRO e ainda trabalhos e casos de estudo em que foram utilizados métodos de classificação, com o objetivo de melhorar a deteção da sépsis.

Conceptualização do problema a estudar Neste capítulo é apresentada uma revisão dos principais conceitos sobre Descoberta de Conhecimento em Bases de Dados, *Business Intelligence e Data Mining*, explicando as várias etapas do processo e técnicas utilizadas. É também apresentada a metodologia CRISP-DM, referencial necessário para todo o processo de DM, descrevendo as fases que a compõem e os documentos por ela produzidos.

Objetivos Neste capítulo, são descritos os objetivos que se pretenderam atingir.

Descrição do estudo Neste capítulo é descrito passo-a-passo o trabalho realizado no âmbito do projeto de mestrado, incluindo a descrição da ferramenta utilizada, materiais e métodos. É também descrito todo o trabalho realizado de acordo com a metodologia CRISP-DM.

Resultados Neste capítulo são apresentados os resultados obtidos de forma detalhada, bem como as implicações desses resultados de forma a dar suporte às ideias contidas nos capítulos seguintes: discussão e conclusões.

Discussão Neste capítulo são discutidos e interpretados os resultados obtidos.

Conclusões Neste último capítulo, é feita uma síntese do projeto realizado, com as conclusões mais relevantes do trabalho desenvolvido e, por último, são apresentadas as recomendações para trabalho futuro.

2 Contextualização

2.1 Introdução

Apesar dos avanços recentes no diagnóstico e tratamento relativamente à sépsis, a taxa de mortalidade permanece inaceitavelmente elevada (Vincent et al., 2006).

É necessário um sistema de deteção útil e preciso para a estratificação dos doentes com sépsis, tanto pelo risco de um evento adverso, como pelo potencial de resposta à terapia (Levy et al., 2003a). Um sistema bem caracterizado que permita também a previsão do *outcome*[2], beneficiará largamente os profissionais da MI com um sistema de apoio à decisão (Levy et al., 2003a). O doente também beneficiará com o estabelecimento dos riscos menos baseadas nas decisões das crenças pessoais e mais na evidência científica (Levy et al., 2003a).

Um sistema de deteção da sépsis focado na predisposição, na infeção, na resposta do hospedeiro e na falência orgânica pode fornecer uma base útil para a estratificação do risco e permitir uma abordagem terapêutica mais individualizada (Rosolem et al., 2010).

Para Rosolem et al. (2010), uma das possíveis razões para a falha de diversos estudos é a heterogeneidade dos grupos de doentes estudados, o que pode dissimular qualquer potencial benefício em subgrupos específicos de doentes. Visto isso, podem-se obter propostas de melhoria, na seleção dos alvos para as intervenções, através de uma melhor caracterização dos doentes com sépsis (Vincent et al., 2009).

Espera-se que, definindo o processo séptico através de uma análise detalhada de cada um dos seus componentes, o desenvolvimento de sépsis seja melhor compreendido, contribuindo no futuro para a melhoria das intervenções terapêuticas para a sépsis (Opal, 2005).

2.2 *Surviving Sepsis Campaign*

A *Surviving Sepsis Campaign* (SSC) fornece as orientações internacionais para o tratamento da sépsis, sépsis grave e choque séptico (Dellinger et al., 2008).

A SSC é um programa liderado pela ESICM (*European Society of Intensive Care Medicine*), ISF (*International Sepsis Forum*) e SCCM (*Society of Critical Care Medicine*), que visa melhorar o diagnóstico de sobrevivência e gestão de doentes com sépsis, abordando os desafios a ela associados, tendo como missão

[2] Alta hospitalar.

(SSC, 2010):

- Aumentar a consciencialização, compreensão e conhecimento;
- Alterar as perceções e comportamentos;
- Aumentar o ritmo das mudanças nos padrões de cuidados;
- Definir padrões de cuidados a ter com a sépsis grave;
- Reduzir a mortalidade em 25%, associada à sépsis, nos próximos 5 anos;
- Trabalhar com as partes interessadas para melhorar a gestão da sépsis através de iniciativas específicas.

2.2.1 Definições sépsis

A sépsis é uma infeção geral grave, difícil de definir, diagnosticar e tratar (SSC, 2010). Está associada a uma série de condições clínicas causadas por uma resposta inflamatória sistémica do organismo a uma infeção, que se desenvolve na sépsis grave, sendo acompanhada por disfunção orgânica simples, múltipla ou falha, levando à morte (SSC, 2010). É uma das principais causas de morte, matando diariamente cerca de 1.400 pessoas no mundo (Bone et al., 1992). Embora se saiba que não existe uma definição clínica clara que possa ser facilmente comunicada e adotada na globalidade. A sua ausência torna o diagnóstico e terapêutica da sépsis num desafio clínico (SSC, 2010).

Nas tabelas 1, 2 e 3, apresentam-se as variáveis associadas aos níveis de sépsis, necessárias para o desenvolvimento dos modelos de classificação.

2.2.2 Desafios

Os profissionais dos cuidados intensivos (médicos e enfermeiros) consideram a sépsis como uma das condições mais desafiadoras e difíceis de gerir, porque o curso da sépsis varia muito de doente para doente e pode-se desenvolver como resultado de várias circunstâncias (SSC, 2010).

A gestão de doentes com sépsis, envolve uma variedade de intervenções terapêuticas, sendo a eficácia do tratamento mais provável se a sépsis grave for evitada, através de cuidados apropriados e precoces. Uma vez diagnosticada, o objetivo da terapia é eliminar as infeções subjacentes com antibióticos (SSC, 2010). Devido aos desafios de diagnosticar e tratar a condição complexa em que

Variáveis	**A sépsis é definida como uma suspeita infeção com uma ou mais das seguintes condições**
Gerais	Febre >38.3°C; Hipotermia <36°C; Frequência cardíaca >90 min^{-1} ou >2 SD acima do valor normal para a idade; Taquipeneia; Estado mental alterado; Edema significativo ou balanço hídrico positivo (>20 mL/kg over 24 hrs); Hiperglicémia (plasma glucose >120 mg/dL) na ausência de diabetes.
Inflamatórias	Leucócitos (Contagem de glóbolos brancos >12,000 μL^{-1}); Leucopenia (Contagem de glóbolos brancos <4000 μL^{-1}); Contagem de glóbolos brancos normal com >10% de formas imaturas; Plasma C-proteína riativa >2 SD acima do valor normal; Plasma procalcitonina >2 SD acima do valor normal.

Tabela 1: Definição de sépsis
Adaptado de (Dellinger et al., 2008 & Levy et al., 2003b)

Variáveis	**A sépsis grave é definida como sépsis associada à disfunção orgânica, hipoperfusão ou hipotensão**
Disfunção orgânica	Hipoxemia arterial (PaO2/FIO2 <300); Oligúria aguda (produção de urina <0.5 $mL \cdot kg^{-1} \cdot hr^{-1}$ ou 45 mmol/L há pelo menos 2 hrs); Creatinina > 2.0 mg/dL; Alterações da coagulação (INR >1.5 ou aPTT >60 segs); Trombocitopenia (Contagem de plaquetas <100,000 μL^{-1}); Hiperbilirrubinemia (plasma de bilirrubina total > 2.0 mg/dL ou 35 mmol/L).
Perfusão tecidual	Hiperlactatemia (>2 mmol/L)
Hemodinâmica	Hipotensão arterial(SBP <90 mm Hg, MAP <65 mm Hg, ou reduzir SBP >40 mm Hg)

Tabela 2: Definição de sépsis grave
Adaptado de (Dellinger et al., 2008 & Levy et al., 2003b)

O choque séptico é definido como uma insuficiência circulatória aguda inexplicável por outras causas
Insuficiência circulatória aguda é definida como hipotensão arterial persistente (SBP <90 mmHg, MAP <60, ou a redução SBP >40 mm Hg desde o início, apesar de reposição de volume adequado)

Tabela 3: Definição de choque séptico
Adaptado de (Dellinger et al., 2008 & Levy et al., 2003b)

se encontram, aproximadamente 10% dos doentes com sépsis não recebem no prazo adequado o tratamento de antibióticos, o que aumenta a mortalidade em 10-15% (Lyseng-Williamson and Perry, 2002).

2.2.3 Custos

Relativamente aos custos, a sépsis impõe uma carga significativa sobre os recursos de saúde, respondendo por 40% das despesas totais da UCI e que custam anualmente até 7,6 bilhões de dólares na Europa e $16,7 bilhões de dólares nos Estados Unidos em 2000, sendo o custo médio por caso de aproximadamente $22.000 (SSC, 2010).

2.3 Sumário

Sendo aceite cientificamente que uma intervenção atempada e adequada, pode melhorar significativamente o prognóstico dos doentes com sépsis grave e choque séptico, é indispensável a implementação de mecanismos que permitam a sua rápida identificação e criação antecipada de terapêutica otimizada (SSC, 2010).

3 Revisão de literatura

3.1 Introdução

Apesar das definições de sépsis, sépsis grave e choque séptico, da última conferência de consenso da SSC (2010), estes termos não permitem a caracterização precisa e a deteção dos doentes com esta condição nem a previsão do seu *outcome* (Levy et al., 2003a).

De entre vários estudos que abordam a questão dos fatores que afetam o resultado dos doentes com sépsis, o primeiro estudo que aborda em conjunto os três níveis de Predisposição, Infeção e Resposta foi publicado por Moreno et al. (2008).

Nesta secção são abordados outros estudos e trabalhos efetuados por outros investigadores, de forma a indicar o que foi feito de mais relevante, relativamente à área do problema estudado.

3.2 Estado da arte

Num trabalho elaborado por Levy et al. (2003a), aborda o aparecimento do conceito PIRO que é um sistema de classificação sugerido por John Marshall. Este sistema tem como finalidade estratificar o estado dos doentes em função da: (P) Predisposição, (I) Infeção, (R) Resposta do hospedeiro e grau de disfunção de (O) Órgão concomitante (Levy et al., 2003a; Opal, 2005 & Rabello et al., 2009). Este conceito foi proposto pela segunda conferência de consenso internacional sobre definições sépsis em 2001, que reuniu simultaneamente as mais importantes sociedades médicas[3] (Levy et al., 2003a). Devido à imprecisão e heterogeneidade da população definida como doentes com sépsis, levou à introdução deste novo sistema de classificação sépsis conhecido como PIRO (Opal, 2005).

O conceito PIRO foi proposto com o objetivo de melhorar a deteção da sépsis (Rabello et al., 2009). Além disto, a organização destes doentes em grupos mais homogéneos pode ajudar a melhorar a prática clínica, determinar o prognóstico e ajudar à inclusão destes doentes em estudos clínicos (Levy et al., 2003a).

O conceito PIRO permite, segundo (Rosolem et al., 2010): uma estratificação mais adequada dos doentes em diferentes grupos de gravidade; delinear estudos clínicos para avaliar estratégias terapêuticas em doentes com sépsis grave

[3]SCCM - *Society of Critical Care Medicine*; ESICM - *European Society of Intensive Care Medicine*; ACCP - *American College of Chest Physicians*; ATS - *American Thoracic Society*; SIS - *Surgical Infection Society*.

e; utilizar esta ferramenta para análise dos desfechos. O PIRO é utilizado para classificação diária do grau de disfunção/falha orgânica de um paciente (Vincent et al., 1996). O SOFA considera o pior valor ocorrido ao longo do dia, para o cálculo de uma classificação de 0 (sem falha) a 4 (falha grave) para cada órgão: respiratório, renal, cardiovascular, neurológico, coagulação e hepático (Strand and Flaatten, 2008).

É importante salientar que o conceito PIRO é rudimentar e surge como uma proposta de investigação e um conceito a desenvolver, sendo necessários testes e aperfeiçoamentos antes que possa ser considerado pronto para aplicação rotineira na prática clínica (Levy et al., 2003a). A sua elaboração exige extensa avaliação da história natural da sépsis, para definir as variáveis que predizem não só um resultado adverso, mas também o potencial de resposta à terapia (Levy et al., 2003a).

Num estudo elaborado por Ribas et al. (2011), foram usadas árvores de decisão para a previsão de resultados com sépsis em doentes tratados com antibióticos. Foi apresentado como método de previsão, uma aplicação polinomial de um modelo estatístico algébrico para avaliar a dependência entre o uso antecipado de antibióticos e os resultados obtidos na UCI (Ribas et al., 2011). Para Ribas et al. (2011), dada a disfunção orgânica, este método revelou uma clara dependência entre o tratamento antecipado com antibióticos e o *outcome* da UCI, respetivamente medidos com o APACHE II e o SOFA em doentes com sépsis grave. O APACHE II (*Acute Physiology and Chronic Health Evaluation*) é um sistema de classificação de gravidade da doença que usa os princípios básicos fisiológicos para estratificar doentes com doença aguda em prognóstico de risco de morte (APACHE, 1985). *O SOFA (Sequential Organ Failure Assessment)* é um sistema de avaliação sequencial de falência orgânica. A avaliação é baseada em seis pontos diferentes para cada um dos problemas: respiratório, hepático, cardiovascular, coagulação, renal e neurológico (Vincent et al., 1996).

Neste trabalho, o efeito dos antibióticos foi ainda estudado com a utilização de árvores de regressão (Ribas et al., 2011). A principal conclusão deste estudo é que esse efeito protetor tem mais importância para falhas multiorgânicas graves acompanhadas por altas pontuações medidas com o APACHE II (mostrando uma diminuição da taxa de mortalidade de cerca de 10%) (Ribas et al., 2011). Neste estudo prospetivo, os dados de teste são de doentes internados na UCI do hospital Vall d'Hebron (Barcelona, Espanha).

Num outro estudo, também feito por (Ribas et al., 2012), os autores

utilizaram a regressão logística para obter indicadores ocultos do modelo, através de análise fatorial. Esses indicadores extraídos são usados na previsão de mortalidade por sépsis grave, utilizando a regressão logística. Os resultados relatados mostram que o método proposto melhora os resultados obtidos com o valor de mortalidade padrão atual, que é baseada na classificação APACHE II (Ribas et al., 2012).

No trabalho de Gwadry-Sridhar et al. (2011), são observados os dados de uma grande estudo observacional de doentes (*coorte*), com variáveis recolhidas em diferentes períodos de tempo, a fim de determinar se desenvolve sépsis ou não. É abordada a análise de *cluster* para formar grupos de pontos de dados correlacionados. O resultado é aplicado com o modelo de *Markov*[4], que pode estimar com precisão a probabilidade de um doente desenvolver sépsis (Gwadry-Sridhar et al., 2011).

Chan and Ting (2011) desenvolveram um trabalho que aborda a construção de um novo modelo de previsão de mortalidade com o teorema de *Bayes* e Algoritmo Genético (AG). Para este mesmo autor, o SAPS II (*Simplified Acute Physiology Score*) é um dos sistemas mais populares de classificação de mortalidade atualmente disponíveis. O SAPS II fornece uma estimativa do risco de morte sem ter que especificar um diagnóstico primário (Strand and Flaatten, 2008). As classificações do SAPS II são convertidas numa probabilidade de mortalidade hospitalar (Le Gall et al., 1993), utilizando a análise de regressão logística (Strand and Flaatten, 2008). Usa 12 variáveis fisiológicas mais a idade, tipo de admissão, presença de metástases, cancro hematológico ou SIDA (Strand and Flaatten, 2008).

Este estudo recolheu dados de 496 doentes internados na UCI entre o ano 2000 e 2001. A idade média dos doentes foi 59,96 anos e 23,8% dos doentes morreram antes da alta. Estes dados foram usados como dados de treino para previsão da mortalidade e foi construído um modelo exponencial de *Bayes* combinado com o Modelo Estatístico de *Bayes* (MEB) e AG (Chan and Ting, 2011). Os pesos ótimos e os parâmetros foram determinados com o AG. Além disso, foram recolhidos dados sobre 142 doentes para testar o novo modelo. A idade média dos doentes para este grupo foi 57,80 anos e 21,8% dos doentes morreram antes da alta. O poder da previsão da mortalidade do novo modelo

[4]A Técnica estatística Análise de *Markov* é utilizada na previsão do comportamento futuro de uma variável cujo estado ou comportamento atual não depende do estado ou comportamento passado, ou seja, é aleatório(Markov, 2011).

foi melhor do que SAPS II (p <0,001). O novo modelo que combina MEB e AG tanto pode gerir dados binários como dados contínuos (Chan and Ting, 2011). Neste modelo, a mortalidade está prevista ser elevada se o doente estiver com a Escala de Coma *Glasgow* (ECG) inferior a 5. A ECG descreve o nível do consciência do doente, entre 3 e 15, onde 3 é o pior valor e 15 o melhor. O cálculo pode ser visto várias vezes ao longo do dia e é composto por três parâmetros: melhor resposta visual, melhor resposta verbal e melhor resposta motora (Jones, 1979). Embora o MEB seja amplamente utilizado para a tomada de decisão médica, tem algumas limitações no que diz respeito à gestão de dados contínuos (Chan and Ting, 2011).

3.3 Sumário

Relativamente ao sistema PIRO, o modo como as categorias de prognóstico interagem dependerá certamente de vários outros fatores, tais como o tipo e temporização de eventuais intervenções terapêuticas (Moreno et al., 2008).

Uma vez que a disfunção/falha de órgãos está atualmente definida no SOFA por uma combinação de variáveis fisiológicas e terapêuticas (por exemplo, pressão arterial e/ou utilização de agentes vasoativos), é neste momento extremamente difícil dissociar as consequências da lesão a partir da resposta de intervenções terapêuticas (Moreno et al., 2008).

Dos estudos analisados, verificou-se uma tendência para modelos de previsão de mortalidade com sépsis, onde não estão presentes modelos de previsão terapêutica, com a exceção do PIRO, que aborda estratégias terapêuticas, mas segundo Moreno et al. (2008), precisam ainda de ser trabalhadas.

Também se pode realçar que, dos trabalhos analisados, os modelos apresentados são sustentados pelos sistemas de classificação APACHE II, SOFA e SAPS II.

De acordo com a literatura, ainda não foram desenvolvidos modelos de classificação do nível de sépsis e terapêutica. Nesse sentido, este projeto de dissertação pretende colmatar essas falhas de conhecimento.

4 Conceptualização do problema

4.1 Introdução

Atualmente a sociedade, é inundada por dados provenientes de diversas fontes existentes em variadíssimas áreas. A sua análise é cada vez mais pertinente para garantir o sucesso das organizações.

O termo Descoberta de Conhecimento em Bases de Dados (DCBD), surgiu em 1989 para se referir ao amplo processo de descoberta de conhecimento em dados e, enfatizar a "alto nível" da aplicação de determinados métodos de *data mining* (DM) (Fayyad et al., 1996b). Fayyad et al. (1996b) considera DM como uma das fases do processo de DCBD. Este processo, depende de técnicas de análise de dados, dentro das quais se encontra o DM, que é um conjunto de técnicas que efetuam a extração do conhecimento (Santos and Azevedo, 2005). A DCBD pode ser definida como o processo de identificar padrões e/ou modelos, a partir de dados em bruto, que sejam novos, potencialmente úteis e compreensíveis (Fayyad et al., 1996b).

As aplicações de DCBD integram teorias, métodos e algoritmos provenientes das áreas de inteligência artificial, aprendizagem automática, reconhecimentos de padrões, estatística, base de dados e outras, tendo como objetivo a extração de conhecimento a partir de grandes bases de dados (Fayyad et al., 1996a).

Os algoritmos utilizados para procurar padrões nos dados são denominados de algoritmos de *data mining* (DM). O processo global de DCBD, que se desenvolve em várias fases, inclui a utilização de algoritmos de DM e a interpretação de padrões encontrados pelos mesmos, os quais são posteriormente utilizados no suporte à tomada de decisão (Santos and Ramos, 2006).

4.2 *Business Intelligence*

O termo *Business Intelligence* (BI) é abrangente, combina arquiteturas, bases de dados, ferramentas analíticas, aplicações e metodologias (Turban et al., 2011). O principal objetivo do BI é disponibilizar o acesso interativo (por vezes em tempo real) de dados, para permitir a sua manipulação, dando aos gestores de negócios e analistas a capacidade de realizar uma análise adequada (Turban et al., 2011). Ao analisar dados históricos e atuais, os decisores obtêm informações valiosas que lhes permitem tomar melhores e mais informadas decisões. O processo de BI baseia-se na transformação dos dados em informação, para em seguida dar

origem à decisão e finalizar com a ação (Turban et al., 2011).

Os sistemas de BI combinam dados com ferramentas analíticas, de forma a disponibilizar informação relevante para a tomada de decisão. O objetivo destes sistemas é melhorar a disponibilidade e qualidade desta informação (Cody et al., 2002). Estes sistemas têm aplicado: a funcionalidade, a escalabilidade e a segurança dos atuais sistemas gestores de bases de dados para construir *Data Warehouses* (DW) que são analisados com técnicas de *On-Line Analytical Processing* (OLAP) e de DM (Santos and Ramos, 2006).

4.3 *Data Mining*

O DM é um processo que usa técnicas estatísticas, matemáticas e de inteligência artificial, para extrair e identificar informação útil e subsequente conhecimento de grandes bases de dados. Isto é conseguido através da descoberta de padrões matemáticos, que podem ser regras, afinidades, correlações, tendências, ou modelos de previsão (Nemati and Barko, 2001).

Na prática, as tarefas associadas ao DM, podem ser divididas em três grupos: previsão, associação e *clustering*. Com base na forma como os padrões são extraídos a partir dos dados históricos, os algoritmos dos métodos de aprendizagem em DM podem ser classificados como supervisionados ou não supervisionados (Turban et al., 2011). Nos algoritmos de aprendizagem supervisionados, os dados de aprendizagem incluem tanto os atributos descritivos (i.e. variáveis independentes ou variáveis de decisão), bem como o atributo de classe (i.e. variável de saída ou variável resultado). Em contraste, com a aprendizagem não supervisionada, os dados de aprendizagem incluem apenas os atributos descritivos (Turban et al., 2011).

No grupo da previsão, as tarefas de DM incluem dois grandes modelos supervisionados: modelos de **Classificação** e modelos de **Regressão**. Os modelos de regressão são utilizados sempre que se pretender prever uma variável com valores contínuos (Santos and Ramos, 2006). Por exemplo, um regressor pode prever, de acordo com os resultados das análises, a probabilidade que um doente tem de viver. Relativamente aos classificadores, são utilizados sempre que se pretende organizar os dados de entrada em classes predefinidas, correspondente a encontrar uma função que associa um caso a uma classe dentro de diversas classes (Santos and Azevedo, 2005). Por exemplo, classificações das terapêuticas nas unidades de cuidados intensivos, como é o caso deste trabalho de investigação.

A classificação aprende com as características das variáveis independentes e de saída, através de um processo de aprendizagem supervisionada, em que ambos os tipos de variáveis são apresentados ao algoritmo (Turban et al., 2011).

Será útil verificar as tarefas de DM, através da Figura 1, para se perceber toda a taxonomia, incluindo os métodos de aprendizagem e os algoritmos para cada uma das respetivas tarefas.

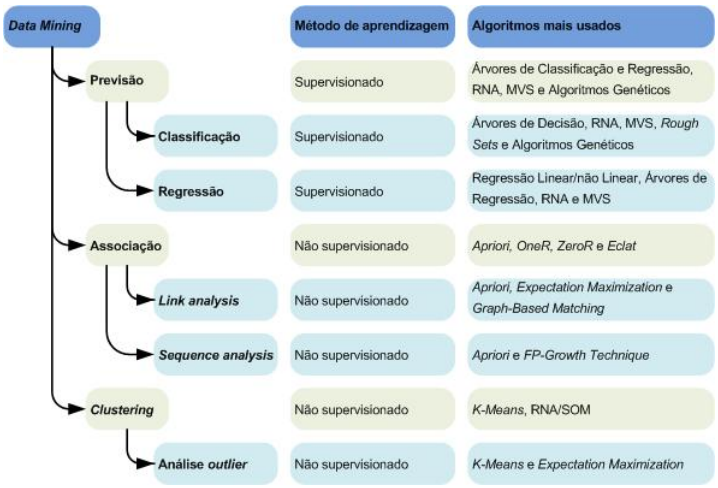


Figura 1: Taxonomia para tarefas de *Data Mining* adaptado de Turban et al. (2011)

O objetivo da classificação é analisar os dados armazenados em bases de dados e gerar automaticamente um modelo que possa prever o comportamento futuro (Turban et al., 2011). A classificação corresponde a uma técnica de aprendizagem de máquina, que aprende padrões de dados, a fim de colocar as novas instâncias nas respetivas classes, de acordo com o modelo de classificação (Santos and Azevedo, 2005). Este modelo induzido sobre os registos, consiste num conjunto de dados de treino com exemplos pré-classificados, que ajudam a distinguir as classes predefinidas (Turban et al., 2011). Espera-se que o modelo possa então ser usado para prever as classes de outros registos não classificados e, mais importante, para prever com precisão futuros eventos reais.

A classificação é a tarefa de DM mais comum, sendo as Redes Neurais Artificiais (RNA), Árvores de Decisão (AD), Máquinas de Vetores de Suporte (MVS), Algoritmos Genéticos (AG) e *Rough Sets* (RS) as técnicas mais aplicadas (Turban et al., 2011).

4.4 Metodologia CRISP-DM

4.4.1 Introdução

Neste capítulo, é apresentada e caracterizada a metodologia CRISP-DM (*CRoss-Industry Standard Process for Data Mining*), descrevendo-se as fases que a compõe, bem como as técnicas utilizadas e os documentos por ela produzidos.

Segundo Groth (2000), o processo de *data mining* (DM), enquadrado no contexto de uma metodologia, torna-se mais fácil de compreender, implementar e desenvolver.

A fim de sistematizar a realização deste projeto de DM, com base nas melhores práticas, decidiu-se adotar a metodologia CRISP-DM.

A metodologia CRISP-DM foi desenvolvida pelo consórcio formado pelas empresas *NCR - Systems Engineering Copenhagen* (EUA e Dinamarca), *DaimlerChrysler AG* (Alemanha), *SPSS Inc* (EUA) e *OHRA Verzekeringen en Bank Groep* (Chapman et al., 2000). O seu desenvolvimento teve origem no interesse crescente e generalizado, por um lado pelo mercado de DM e, por outro, pelo consenso de que a indústria necessitava de um processo padronizado (Wirth, 2000).

O modelo definido na Figura 2, retrata o ciclo de vida de projetos de DM, representado por fases com as tarefas detalhadas em cada fase. No Anexo A, indica o fluxo de todo o processo e as dependências existentes entre as fases:

4.4.2 Compreensão do negócio

Nesta fase, pretende-se compreender o negócio, com foco nos objetivos do projeto e nos requisitos do ponto de vista do negócio, convertendo depois os objetivos do negócio em objetivos de DM. Depois, deve ser traçado um plano que valide a satisfação dos objetivos (Chapman et al., 2000).

4.4.3 Compreensão dos dados

Nesta fase, dá-se início à recolha e consequente exploração dos dados, com vista à sua compreensão, análise e identificação de problemas de qualidade dos mesmos. Segue-se a identificação de relações entre os dados ou, a deteção de subconjuntos interessantes destes, a fim de serem analisados posteriormente de forma a permitir identificar conhecimento oculto (Chapman et al., 2000).

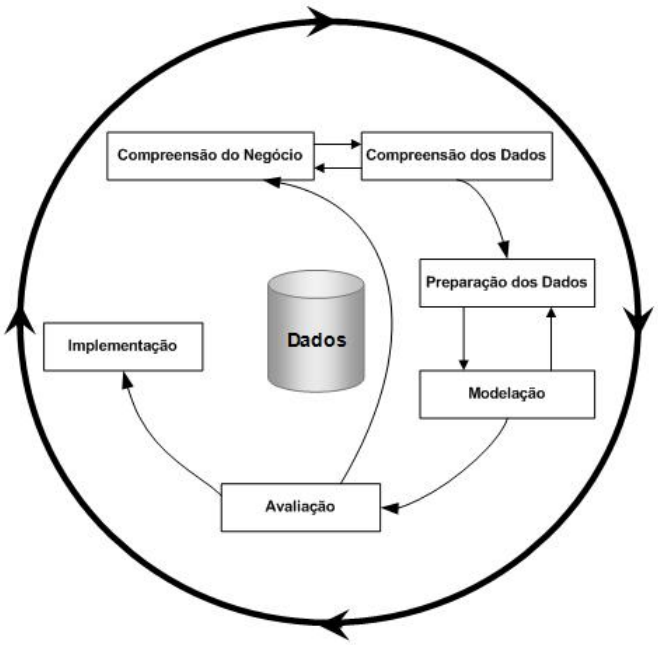


Figura 2: Fases do CRISP-DM
(Chapman et al., 2000)

4.4.4 Preparação dos dados

Esta fase envolve todas as atividades necessárias para a construção do conjunto final de dados. Estes dados serão usados pelas ferramentas de modelação para posteriormente serem analisados pelos algoritmos de DM. As tarefas de preparação dos dados incluem a seleção de tabelas, atributos e registos, bem como a transformação e limpeza dos dados, com vista à sua posterior análise pelas ferramentas de modelação (Chapman et al., 2000).

4.4.5 Modelação

Nesta fase, são selecionadas várias técnicas de modelação e os seus parâmetros são ajustados de forma a otimizar os resultados. Normalmente, existem várias técnicas para o mesmo tipo de problema de DM, sendo que algumas têm requisitos específicos sobre a forma como os dados são apresentados, por isso, pode ser necessário voltar à fase de preparação de dados (Chapman et al., 2000).

No âmbito deste trabalho, optou-se por adotar as seguintes técnicas: Árvores de Decisão (AD), Máquinas de Vetores de Suporte (MVS) e o classificador *Naïve*

Bayes, como proposta de sistema de classificação.

Árvores de Decisão para Turban, as ADs servem para classificar os dados, num número finito de classes, com base nos valores de entrada. As ADs são uma simples representação do conhecimento, muito eficiente na construção de classificadores que preveem classes baseadas nos valores de atributos de um conjunto de dados (Turban et al., 2008 and Turban et al., 2011).

As AD são, essencialmente compostas por uma hierarquia de declarações "se-então" e são, portanto, significativamente mais rápidas, exigindo menor esforço computacional, relativamente às Redes Neuronais Artificiais (RNA) (Turban et al., 2008 and Turban et al., 2011). Estes algoritmos são mais apropriados para dados discretos ou divididos em intervalos. Portanto, incorporar as variáveis contínuas numa estrutura de AD requer discretização, isto é, converter variáveis contínuas em categorias de gamas numéricas (Turban et al., 2008 and Turban et al., 2011).

Máquinas de Vetores de Suporte é uma técnica introduzida por Cortes and Vapnik (1995), como uma nova técnica para resolver problemas de reconhecimento de padrões. Segundo a teoria das MVSs (Cortes and Vapnik, 1995), as técnicas tradicionais de reconhecimento de padrões são baseados na minimização do *risco empírico* – isto é, na tentativa de otimizar o desempenho do conjunto de treino –, as MVSs minimizam o *risco estrutural* – isto é, a probabilidade de classificar mal padrões de dados desconhecidos segundo uma distribuição de probabilidade. O que faz com que as MVSs sejam atrativas, devido à capacidade de condensar a informação contida no conjunto de treino e o uso de famílias de decisão de relativa baixa dimensão (Pontil and Verri, 1998).

Naïve Bayes é um algoritmo de classificação, baseado no teorema de *Bayes*, que prevê a probabilidade de um conjunto de dados pertencer a uma determinada classe (Langley and Sage, 1994).

O classificador *Naïve Bayes* supõe independência de atributos dentro de cada classe, o que permite que esta use a igualdade

$$p(x|wj) = \prod_{i=1}^{d} p(xi|wj)$$

onde os valores de $p(xi|wj)$ representam as probabilidades condicionais arma-

zenados em cada classe. Esta abordagem simplifica o cálculo das probabilidades de classe para uma dada observação (Langley and Sage, 1994).

Para Langley and Sage (1994), a aprendizagem do classificador *Naïve Bayes*, baseia-se em incrementos simples de contagem cada vez que encontra uma nova instância, juntamente com uma contagem separada para uma classe, cada vez que encontra uma instância da classe. Estas contagens deixam a estimativa no classificador $p(wj)$ para cada classe wj, sendo que, para cada valor nominal, o algoritmo atualiza a contagem para esse par classe-valor (Langley and Sage, 1994). A segunda contagem deixa a estimativa no classificador $p(xi|wj)$, sendo que, para cada atributo numérico, o método mantém duas quantidades, a soma e a soma dos quadrados, que permite calcular a média e a variância para uma curva normal usada para encontrar $p(xi|wj)$ (Langley and Sage, 1994).

O classificador *Naïve Bayes* é um método simples e leve, amplamente testado para a indução probabilística, é particularmente eficiente quando a dimensão dos dados de entrada é alta (Langley, 1993). O *Naïve Bayes* apresenta um melhor desempenho relativamente a muitos métodos de classificação sofisticados (Friedman and Goldszmidt, 1996).

4.4.6 Avaliação

Esta fase tem como finalidade avaliar a utilidade do(s) modelo(s). Antes de proceder à implementação final do(s) modelo(s), é importante avaliá-lo(s) cuidadosamente, rever os passos executados na sua construção, de forma a ter a certeza que se atingiram os objetivos do negócio, assim como, avaliar se alguma questão importante para o negócio não tenha sido considerada (Chapman et al., 2000).

No âmbito deste trabalho, optou-se por adotar as seguintes técnicas de avaliação: Matriz de Confusão (MC) e *Cross-validation* (CV). As métricas de avaliação para a sépsis foram: a taxa de erro total, a acuidade e sensibilidade e especificidade; para a terapêutica foi a acuidade.

Matriz de Confusão de um classificador indica o número de classificações corretas *versus* as previsões efetuadas para cada classe, sobre um conjunto de exemplos (Kohavi and Provost, 1998). A MC é uma das técnicas de avaliação mais utilizadas em problemas de classificação. No caso binário, cada exemplo é previsto como sendo Positivo ou Negativo (Santos and Azevedo, 2005). Na Tabela 4 mostra quatro valores possíveis (2x2).

Tabela 4: Matriz de confusão de duas classes
Adaptado de (Santos and Azevedo, 2005)

Classe	Previsão C+	Previsão C-
Real C+	Verdadeiros Positivos (VP)	Falsos Negativos (FN)
Real C-	Falsos Positivos (FP)	Verdadeiros Negativos (VN)

Verdadeiros Positivos designados por VP, correspondem ao número de exemplos positivos corretamente classificados;

Verdadeiros Negativos designados por VN, correspondem ao número de exemplos negativos efetivamente classificados como negativos;

Falsos Positivos corresponde ao número de exemplos positivos classificados como negativos (i.e. mal classificados), representados por FP;

Falsos Negativos número de exemplos negativos classificados como positivos (i.e. mal classificados), representados por FN.

Desta matriz podem ser derivadas muitas outras medidas, tais como (Santos and Azevedo, 2005): Taxa de erro total (1), sensibilidade (2), especificidade (3) e acuidade da previsão (4).

$$erro = \frac{FP + FN}{n} x100(\%) \tag{1}$$

$$sens = \frac{VP}{VP + FN} x100(\%) \tag{2}$$

$$espec = \frac{VN}{VN + FP} x100(\%) \tag{3}$$

$$acuip = \frac{VP + VN}{n} x100(\%) \tag{4}$$

Cross-validation a fim de utilizar todos os casos disponíveis e comparar a precisão da previsão, pode-se usar o método *Cross-validation* (CV). No CV o conjunto de dados é aleatoriamente dividido em subconjuntos mutuamente exclusivos de *k* tamanhos aproximadamente iguais. O modelo de classificação é treinado e testado *k* vezes. Cada uma das vezes é treinado com todos, mas só

um *fold* é então testado com os restantes. A estimativa global da precisão da validação cruzada de um modelo é calculada pela média das *k* medidas individuais de precisão, como mostrado na seguinte equação: onde PVC é a precisão da validação cruzada, o *k* é o número de *folds* utilizados, e *A* é a medida de precisão (por exemplo, taxa de sucesso, sensibilidade, especificidade) de cada *folds* (Turban et al., 2011).

$$PVC = \frac{1}{k} \sum_{i=1}^{k} Ai$$

4.4.7 Implementação

A criação do(s) modelo(s) não marca o fim do projeto. Mesmo que o objetivo dos modelos seja aumentar o conhecimento sobre os dados, a informação obtida tem que ser organizada e apresentada para que o utilizador a possa utilizar. No final do projeto, será realizado um relatório final. Dependendo do plano de implementação, o relatório pode ser apenas um resumo do projeto ou pode ser uma apresentação final e abrangente do resultado de todo o processo de DM (Chapman et al., 2000).

4.5 Questão de investigação

Para dar suporte à dissertação, foi elaborada a questão de investigação, onde se pretende saber:

qual a viabilidade de prever o plano terapêutico de doentes com sépsis, utilizando modelos de data mining para classificação?

4.6 Sumário

Os sistemas de BI oferecem vantagens às organizações, pois permitem a análise de dados provenientes de várias fontes, através de ferramentas de análise, com o objetivo de disponibilizar informação para a tomada de decisão. Estes sistemas combinam dados com ferramentas analíticas, tais como: Data Warehouse, sistemas OLAP ou técnicas de DM para a extração de conhecimento. Através das ferramentas de DM podem-se encontrar padrões nos dados, inferindo regras a partir destes. A descoberta de padrões de conhecimento divide-se em previsão, associação e *clustering*. A previsão pode ser conseguida através de métodos de classificação ou regressão.

Nesta dissertação, será o método de classificação que fará parte do desenvolvimento dos modelos.

A etapa de DM torna-se numa tarefa complexa, daí que se tenha decidido adotar a metodologia CRISP-DM a fim de sistematizar a realização deste projeto, com base nas melhores práticas. A metodologia CRISP-DM é descrita em termos de um processo hierárquico, com um ciclo de vida que se desenvolve em seis fases: Compreensão do Negócio, Compreensão dos Dados, Preparação dos Dados, Modelação, Avaliação e Implementação. Na fase de Modelação foram descritas algumas técnicas a utilizar neste trabalho: AD, MVS e NB.

Existem diversas métricas para a avaliação dos modelos criados pelas técnicas de DM. Na Avaliação dos modelos, optou-se por adotar as seguintes técnicas de avaliação: a *Cross-validation* e a Matriz de Confusão, incluíndo as métricas associadas para se determinar a taxa de erro total, a sensibilidade, a especificidade e a acuidade da previsão.

5 Objetivos

Com o propósito de responder à questão de investigação e apoiar a decisão dos médicos, este trabalho focou-se na melhor terapêutica para doentes com problemas microbiológicos, tendo como principal ponto de partida o nível de sépsis.

Descrevem-se os seguintes objetivos para este projeto de dissertação:

- Perceber até que ponto é possível prever, com elevado grau de acuidade, o nível de sépsis e o plano terapêutico de doentes com sépsis;
- Estudar e definir as variáveis que influenciam a terapêutica;
- Desenvolver e testar um conjunto de modelos de classificação, baseados em DM, que possibilitarão o médico decidir sobre a melhor terapêutica a aplicar, adequada aos problemas do doente;
- Testar os modelos de classificação com base em dados reais.

6 Descrição do estudo

6.1 Introdução

Como mencionado anteriormente, o processo de *data mining* (DM) usa técnicas para, a partir de grandes volumes de dados, extrair conhecimento, utilizando modelos de previsão.

Nesta dissertação, foi utilizada a tarefa de previsão, o modelo de classificação, o método de aprendizagem supervisionada e os algoritmos: Árvores de Decisão (AD), Máquinas de Vetores de Suporte (MVS) e o classificador *Naïve Bayes* (NB), para prever o nível de sépsis e o plano terapêutico de doentes com sépsis. Relativamente à avaliação, utilizaram-se a Matriz de Confusão (MC), incluindo as métricas associadas e a *Cross-validation* (CV). De entre as métricas associadas foram utilizadas a análise da taxa de erro total, a sensibilidade, a especificidade e a acuidade que permitiram identificar quais as medidas mais relevantes para a previsão do nível da sépsis e do plano terapêutico em estudo.

A ferramenta utilizada foi o *SQL Developer* da *Oracle* para análise de dados e o *Oracle Data Mining* para desenvolvimento dos modelos.

Nesta secção, todo o processo de DM é conduzido pelas diferentes fases da metodologia CRISP-DM, referencial adotado para esta dissertação. A metodologia adotada retrata o ciclo de vida de todo o projeto de DM, representado por fases na Figura 2. No Anexo A, estão detalhadas as tarefas, o fluxo de todo o processo e as dependências existentes entre as fases.

6.2 Compreensão do Negócio

Os dados utilizados nos modelos de DM foram recolhidos no Serviço de Cuidados Intensivos (SCI), designado por UCI, do Centro Hospitalar do Porto, Hospital de Santo António (HSA), Porto, Portugal. São dados relativos à unidade de cuidados intensivos no período compreendido entre 19/8/2011 e 4/7/2012, correspondendo a 305 dias de internamento de 394 doentes em 12 camas.

Os sistemas utilizados para a obtenção dos dados foram:

- 12 monitores de sinais vitais;
- 10 ventiladores;
- Sistema farmacêutico;
- Laboratório de exames.

Nas UCIs existem doentes em condições graves, possivelmente com falência orgânica ou em risco de vida. Os profissionais da UCI (enfermeiros e médicos) são responsáveis pelo tratamento do doente, evitando a falência orgânica e revertendo o doente para uma situação clínica anterior à entrada na UCI (Silva et al., 2008).

Neste projeto realizou-se um estudo empírico, recorrendo às definições da sépsis, a bases de dados e a técnicas de DM, com o objetivo de construir modelos de classificação capazes de prever, com elevado grau de acuidade, o nível de sépsis e o plano terapêutico de doentes com sépsis.

Os objetivos do negócio são:

- Compreender a condição atual do doente com nível de sépsis, sendo critério de sucesso, tirar uma conclusão clínica sobre o estado do doente;
- Perceber que medicamentos são utilizados para o tratamento da sépsis, sendo critério de sucesso, ajudar o processo de decisão clínica, dando informações sobre o plano terapêutico.

Os objetivos de *data mining* são:

- Determinar o nível de sépsis de um doente, com critério de sucesso de 95% de acuidade;
- Prever o grupo de medicamento a administrar a um doente com sépsis, com critério de sucesso de 80% de acuidade ou demonstrar que não existe correlação entre o nível de sépsis e o medicamento administrado.

Quanto às ferramentas e técnicas, todo o projeto foi elaborado, recorrendo a tecnologia *Oracle*, no ambiente de desenvolvimento *SQL Developer* da *Oracle.* Trata-se de uma aplicação multiplataforma (e.g. Windows, Linux, Mac OS) para análise de dados, que inclui também o *Oracle Data Mining* (ODM) para *data mining* (DM). O *SQL Developer*, permite desenvolver *queries*, *procedures*, *functions*, *triggers*, *views* e outras instâncias de desenvolvimento. É uma aplicação que proporciona uma fácil manipulação dos dados, para além de ser orientada também para DM, incluindo uma grande variedade de algoritmos.

Num estudo realizado pelo KDnuggets em 2009, sobre as ferramentas comerciais de DM mais utilizadas em projetos reais, atribui o quarto lugar (29%) à ODM (Piatetsky-Shapiro, 2009).

Neste projeto foi utilizada uma base de dados *Oracle*, instalada num servidor com a versão de *linux CentOS 5.8*.

No sentido de apresentar o plano do projeto, foi disponibilizado no Anexo B o diagrama onde estão definidas as atividades, bem como os tempos de duração de todas as tarefas e o seu encadeamento.

6.3 Compreensão dos Dados

Neste projeto, tal como já foi anteriormente mencionado, foram utilizados dados recolhidos pela UCI do HSA, relativos ao período compreendido entre 19/8/2011 e 4/7/2012, referentes a 305 dias de internamento de 394 doentes.

As tabelas disponibilizadas pelo HSA para este projeto, foram geradas a partir de dados reais, recolhidos em tempo real e tratados *online*, recorrendo às seguintes fontes de dados:

- Processo Clínico Eletrónico (PCE);
- Monitor dos Sinais Vitais (MSV);
- Laboratório (LAB);
- Gestão Hospitalar de Armazém e Farmácia (GHAF).

Para se perceber a proveniência dos dados, foram disponibilizadas as seguintes tabelas referentes à base de dados (BD) do HSA, contendo os dados necessários ao desenvolvimento do projeto:

- Medicamentos, com 1 010 registos e os seguintes atributos:
  - PID (número de identificação do doente)
  - Designação do medicamento;
  - Grupo de medicamento (1º nível);
  - Subgrupo de medicamento (2º nível);
  - Subsubgrupo de medicamento (3º nível).
- Análises, com 372 989 registos e os seguintes atributos:
  - PID;
  - Hora da validação dos dados;
  - Data de validação;
  - Bilirrubina total;

  - Creatinina;
  - Glicose;
  - Leucócitos;
  - Plaquetas.

- Medicação, com 88 346 registos e os seguintes atributos:
  - PID;
  - Medicamento;
  - Data da toma;
  - Custo.

- Sinais vitais, com 6 739 927 registos e os seguintes atributos:
  - PID;
  - Data da inserção dos dados;
  - Frequência Cardíaca (FC);
  - Pressão Arterial Sistólica (PAS);
  - Pressão Arterial Média (PAM);
  - Temperatura (TEMP).

Na tabela 5 apresentam-se os atributos, descrição e respetiva fonte de dados, referentes à seleção inicial dos atributos.

Na exploração dos dados, procedeu-se à análise estatística das variáveis das análises clínicas e dos sinais vitais em uso, antes da transformação. No anexo C, estão descritas como análise estatística: a percentagem de nulos, percentagem de valores distintos, moda, média, mediana, mínimo, máximo, desvio padrão e variância dos dados.

Para pré-validação dos valores, conforme tabela 6, foi considerado o intervalo definido pelos médicos dos cuidados intensivos para os resultados dos sinais vitais recolhidos automaticamente.

Na tabela 7, estão presentes os valores máximo e mínimo dos sinais vitais e análises clínicas, com base nos limites da sépsis (tabelas 1, 2 e 3).

Atributos	Descrição	Fontes de dados
BILIRRUBINA	Bilirrubina total	LAB
CREATININA	Creatinina	LAB
CUSTO	Custo farmacêutico do medicamento	GHAF
DESIGNACAO_MED	Designação do medicamento	GHAF
FC	Frequência cardíaca	MSV
GLICOSE	Glicose	LAB
HORA_RECOLHA	Hora em que o exame clínico foi recolhido	LAB/MSV
LEUCOCITOS	Leucócitos	LAB
N_EPISODIO	Número do episódio associado ao doente	PCE/MSV/LAB/GHAF
DATA_VALOR	Data de registo/recolha do valor	PCE/MSV/LAB/GHAF
PAM	Pressão arterial média	MSV
PAS	Pressão arterial sistólica	MSV
PLAQUETAS	Plaquetas	LAB
TEMP	Temperatura	MSV
GRUPO_I	Grupo do medicamento de 1º nível	GHAF
GRUPO_A	Grupo do medicamento de 2º nível	GHAF
SUBGRUPO_I	Grupo do medicamento de 3º nível	GHAF

Tabela 5: Atributos, descrição e fonte de dados

Descrição	Valor mínimo	Valor máximo
Frequência Cardíaca (FC)	0	250
Pressão Arterial Média (PAM)	0	200
Pressão Arterial Sistólica (PAS)	0	300
Temperatura (TEMP)	34	45

Tabela 6: Intervalo de valores dos sinais vitais

Variáveis	Mínimo	Máximo
PAM	65	9999
PAS	90	9999
TEMP	36	38,3
FC	0	90
BILIRRUBINA	0	2
CREATININA	0	2
GLICOSE	0	120
LEUCOCITOS	4000	12000
PLAQUETAS	100	9999

Tabela 7: Tabela com os valores máximo e mínimo dos sinais vitais e análises clínicas definidos para a sépsis

Após a exploração dos dados, verificou-se a sua qualidade, conforme tabela

8, que mostra a percentagem de valores nulos em todo os atributos selecionados.

Atributos	Valores nulos
BILIRRUBINA	0,0368%
CREATININA	0,0036%
CUSTO	0
DATA_VALOR	0
DESIGNACAO_MED	0
FC	0
GLICOSE	0,0006%
HORA_VALIDACAO	0
HORA_VALOR	0
LEUCOCITOS	0,0034%
N_EPISODIO	0
PAM	0
PAS	0
PLAQUETAS	0,0058%
TEMP	0
GRUPO_A	0
GRUPO_I	0
SUBGRUPO_I	0

Tabela 8: Qualidade dos dados

6.4 Preparação dos Dados

No processo de preparação dos dados, procedeu-se à análise estatística das variáveis selecionadas antes da transformação. No anexo D, estão descritas como análise estatística: a percentagem de nulos, percentagem de valores distintos, moda, média, mediana, mínimo, máximo, desvio padrão e variância dos dados a serem utilizados pelos modelos de DM.

A percentagem de nulos é elevada, porque em certas alturas os valores não são recolhidos à mesma hora, por exemplo, se a bilirrubina for recolhida às 10h e a creatinina às 11h, os valores vão aparecer em dois registos horários diferentes, sendo o valor recolhido, válido numa hora e nulo na outra. Se para um novo registo de um doente, à mesma hora, forem recolhidos todos os dados, esse mesmo registo ficará completo, evitando-se assim os valores nulos.

Relativamente à seleção dos dados, iniciou-se o processo com a criação de duas vistas de dados, em que uma reúne os atributos da tabela das análises clínicas e a outra da tabelas dos sinais vitais. Os atributos que resultaram na vista de dados da tabela das análises são: N_EPISODIO, EXAME, DATA_VALIDACAO,

HORA_VALIDACAO, RESULTADO e SEPSIS. Os atributos que resultaram na vista de dados da tabela dos sinais vitais são: N_EPISODIO, CATEGORIA, DATA_VALOR, HORA_VALOR, VALOR_MIN, VALOR_MAX e SEPSIS. Ao criar esta vista, para garantir que os dados ficassem aptos para o processo de modelação, procedeu-se à eliminação dos nulos na seleção destes, através da clausula *is not null* para todos os atributos.

Após a seleção e eliminação dos dados, verificou-se ainda que alguns dados estavam fora do intervalo. A tabela 9 mostra a percentagem desses valores definidos para os sinais vitais. De seguida procedeu-se à sua limpeza.

Atributos	Valores fora do intervalo
FC	**5,53%**
PAM	**3,19%**
PAS	**1,65%**
TEMP	**1,64%**

Tabela 9: Valores fora do intervalo

Na tabela 10 são apresentados os dados obtidos após transformação, derivando assim, em novos atributos e consequentemente novos registos. Com a exceção da variável SEPSIS_FINAL, todas as restantes têm atribuído o valor 0 para os dados que estão dentro do intervalo e o valor 1 para os que estão fora dele. Para esse efeito, foram criados dois procedimentos que atribuem o valor 1 ou 0 ao atributo SEPSIS da respetiva vista.

Atributos	Dados
SEPSIS_BILIRRUBINA_TOTAL	**{0;1}**
SEPSIS_CREATININA	**{0;1}**
SEPSIS_FC	**{0;1}**
SEPSIS_GLICOSE	**{0;1}**
SEPSIS_LEUCOCITOS	**{0;1}**
SEPSIS_PAM	**{0;1}**
SEPSIS_PAS	**{0;1}**
SEPSIS_PLAQUETAS	**{0;1}**
SEPSIS_TEMP	**{0;1}**
SEPSIS_FINAL	**{0;2;3}**

Tabela 10: Dados obtidos após transformação

O procedimento da vista de dados das análises clínicas, compara o valor do atributo RESULTADO através da verificação dos valores do intervalo mínimo e

máximo da tabela 7, colocando-o a 0 se estiver dentro do intervalo e 1 se estiver fora deste. Sendo que o valor 1 representa sépsis e o 0 representa normalidade.

O procedimento da vista de dados dos sinais vitais, compara o valor dos atributos VALOR_MIN e VALOR_MAX através da verificação dos valores do intervalo mínimo e máximo da tabela 7, colocando-o a 0 se estiver dentro de intervalo e 1 se estiver fora deste.

Relativamente à variável SEPSIS_FINAL, é atribuído o valor 0 para doentes sem sépsis, o valor 2 para doentes com sépsis grave e o valor 3 para doentes com choque séptico. Recorrendo ao valor 1 dos restantes atributos da tabela 10 e em conformidade com as definições da sépsis das tabelas 1, 2 e 3, o procedimento calcula para cada registo, o respetivo nível de sépsis. Visto isso, o procedimento atribui 0, 2 ou 3, através de uma simples verificação, conforme definição da sépsis das referidas tabelas (1, 2 e 3).

Na variável SEPSIS_FINAL não foi considerado o estado do doente com sépsis (referente às variáveis da tabela 1), pois caso se viesse a verificar teria que ser classificado com o valor 1. Por esse facto não se pode atribuir o valor 1 à variável SEPSIS_FINAL, porque se verificou que os valores são facilmente confundidos com o estado do doente sem sépsis.

No processo de integração dos dados, para se efetuar a junção destes, foi criada uma vista de dados para a totalidade dos dados, reunindo os atributos da tabela 5 e os que sofreram transformação da tabela 10.

Por último, houve necessidade de converter os dados numéricos contínuos em intervalos de classes. Para isso, os intervalos de valores foram criados usando uma escala de 7 pontos adaptada pela *Clinical Global Impression - Severity scale* (CGI-S) (Guy, 1976). O objetivo da CGI-S é o de permitir que o médico possa avaliar a gravidade da doença (Guy, 2000). Neste sentido, foi utilizada a técnica de agrupamento *Bin Quantile Range*, sendo o *Range* (ODM, 2012) o número de classes, neste caso 7. As classes foram criadas tendo em conta o quantile dos valores.

6.5 Modelação

Com o propósito de obter modelos, a partir dos dados, que pudessem classificar de modo satisfatório, o nível de sépsis e o plano terapêutico de doentes com sépsis, tornou-se indispensável converter esse conhecimento numa tarefa de DM capaz de traduzir os objetivos do negócio e do DM.

Após a preparação dos dados e a sua transformação, chega a fase da

modelação a partir dessa transformação. A figura 3 representa os modelos de classificação da sépsis e a figura 4 representa os modelos de classificação do plano terapêutico.

Figura 3: Construção dos modelos de classificação para a sépsis

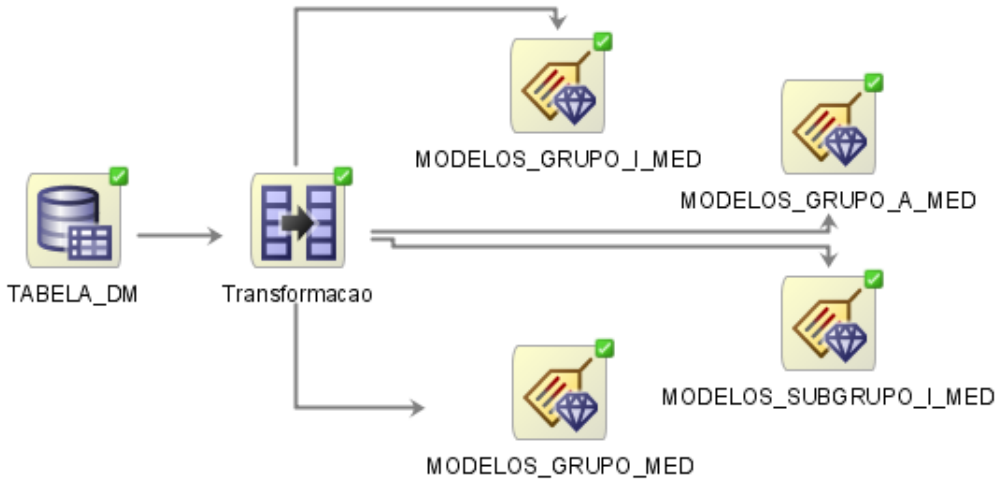


Figura 4: Construção dos modelos de classificação do plano terapêutico

Uma vez que as variáveis correspondem a valores contínuos e discretos, optou-se por utilizar os modelos de classificação, distinguindo o conjunto das variáveis independentes (análises e sinais vitais) das variáveis dependentes (sépsis final e medicamento).

As técnicas de modelação adotadas, como já foi referido na introdução desta secção, são as AD, MVS e NB, as suas características estão descritas na tabela 11.

Para a conceção/avaliação dos modelos, foram utilizados 70% dos dados para treino e os restantes 30% para teste.

Na construção dos modelos, foram efetuados alguns ajustes nos parâmetros.

Os valores contínuos e os valores discretos com mais de 10 classes, à exceção do número de episódio e da data, foram classificados usando o *Bin Quantile Range*, os restantes valores não sofreram alteração.

Descrição	Valores
Algorithm Name	**Support Vector Machine**
Active Learning	**yes**
Automatic Preparation	**on**
Complexity Factor	**0.068966**
Kernel Function	**Linear**
Tolerance	**.001**
Algorithm Name	**Naive Bayes**
Automatic Preparation	**on**
Pairwise Threshold	**0**
Singleton Threshold	**0**
Algorithm Name	**Decision Tree**
Automatic Preparation	**on**
Criteria For Splits	**20**
Criteria For Splits(%)	**.1**
Maximum tree depth	**7**
Minimum Child Record Count	**10**
Tree Impurity Metric	**Gini**

Tabela 11: Características das técnicas de modelação

Para gerar os modelos, o processo passou por duas etapas:

- 1ª etapa é constituída pelos modelos M1, M2, M3 e M4, para o nível de sépsis e para o plano terapêutico, em que os atributos do M1, M2 e M3 são selecionados manualmente e o M4 automaticamente;
- 2ª etapa é constituída pelos modelos M5, M6, M7, M8, M9 e M10, para o plano terapêutico, em que os atributos são todos selecionados manualmente.

Na 1ª etapa, foram desenvolvidos os modelos (M1 a M4), cada um deles com as três técnicas (AD, MVS e NB), referentes à condição do nível de sépsis. Ainda na 1ª etapa, foram de igual forma desenvolvidos mais quatro modelos (M1 a M4), cada um deles com as mesmas três técnicas (AD, MVS e NB), embora estes sejam referentes ao tratamento da sépsis, conforme descrição abaixo, para ambos os casos.

Os atributos de entrada (variáveis independentes) são:

$$CaseMix = \{PID, Data, Hora, Custo\}$$

$$Sepsis = \{BilirrubinaTotal; Creatinina; FCMax; FCMin; Glicose; Leucocitos\} +$$

$$\{PAMMax; PAMMin; PASMax; PASMin; Plaquetas; TempMax; TempMin\}$$

$$VarSepsis = \{SepsisBilirrubinaTotal; SepsisCreatinina; SepsisFC; SepsisTemp\} +$$

$$\{SepsisGlicose; SepsisLeucocitos; SepsisPAM; SepsisPAS; SepsisPlaquetas\}$$

Os atributos alvo (variáveis dependentes) são:

$$Alvo1 = SepsisFinal$$

$$Alvo2 = Medicamento$$

Cada modelo pode ser representado da seguinte forma:

$$Modelo_{\{M1\}*\{MVS;AD;NB\}*\{Alvo1;Alvo2\}} = \{CaseMix, Sepsis\}$$

$$Modelo_{\{M2\}*\{MVS;AD;NB\}*\{Alvo1;Alvo2\}} = \{CaseMix, VarSepsis\}$$

$$Modelo_{\{M3\}*\{MVS;AD;NB\}*\{Alvo1;Alvo2\}} = \{CaseMix; Sepsis, VarSepsis\}$$

$$Modelo_{\{M4\}*\{MVS;AD;NB\}*\{Alvo1;Alvo2\}} = \{Automático\}$$

Por exemplo, o modelo M1 referente às variáveis independentes (*CaseMix* e *Sepsis*), pode ser representado da seguinte forma:

$$M1_A = MVS * A1 * \{CaseMix, Sepsis\}$$

$$M1_B = AD * A1 * \{CaseMix, Sepsis\}$$

$$M1_C = NB * A1 * \{CaseMix, Sepsis\}$$

$$M1_D = MVS * A2 * \{CaseMix, Sepsis\}$$

$$M1_E = AD * A2 * \{CaseMix, Sepsis\}$$

$$M1_F = MB * A2 * \{CaseMix, Sepsis\}$$

Resultando em:

$$4modelos * 3tecnicas * 2alvos = 24cenários$$

Para um melhor esclarecimento, encontra-se no anexo E uma tabela onde é apresentada uma outra descrição para os mesmos modelos.

Devido à fraca qualidade dos modelos referentes ao plano terapêutico na 1ª fase, decidiu-se fazer uma revisão aos parâmetros e acrescentar os seis modelos referidos (M5 a M10).

Para a configuração dos seis modelos de classificação da terapêutica, foram utilizados os atributos mencionados no anexo F, resultando em:

$$6modelos * 3tecnicas * 3alvos = 64cenários$$

Relativamente à conversão dos dados numéricos contínuos em intervalos de classes, foi utilizada a técnica de agrupamento *Bin Quantile Range*. Nas figuras seguintes são apresentadas as classes criadas tendo em conta o quantile 7.

De seguida é feita uma análise exploratória e explicativa para os atributos de entrada e atributos alvo.

No atributo Bilirrubina, verifica-se a ocorrência em 6 classes, tendo uma distribuição aproximada de 20% nas 5 primeiras classes e de 0,81% na classe (>24.63).

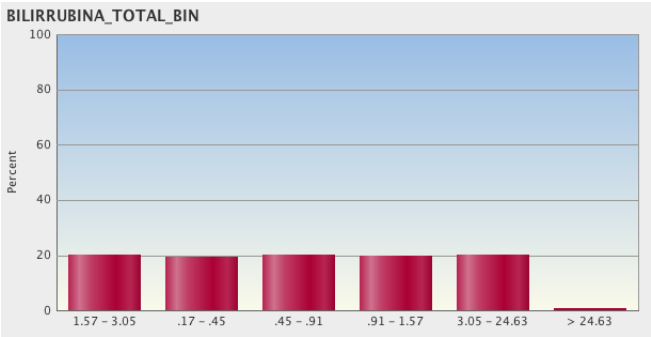


Figura 5: Classes do atributo Bilirrubina

No atributo Creatinina, verifica-se a ocorrência em 6 classes, tendo uma distribuição aproximada de 20% nas 5 primeiras classes e de 0,15% na classe (>6.45).

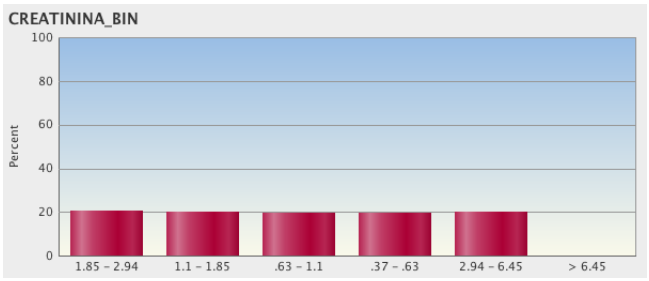


Figura 6: Classes do atributo Creatinina

No atributo Glicose, verifica-se a ocorrência em 6 classes, destacando-se a classe (108-124) com 23,73%.

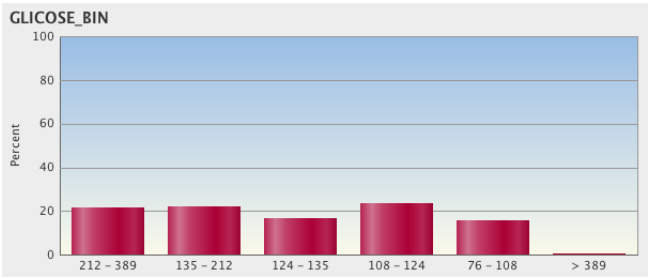


Figura 7: Classes do atributo Glicose

No atributo Leucócitos, verifica-se a ocorrência em 6 classes, destacando-se a classe (12.04-16.37) com 24,34%.

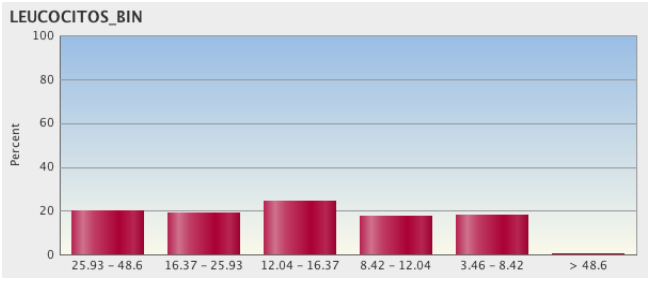


Figura 8: Classes do atributo Leucócitos

No atributo Plaquetas, verifica-se a ocorrência em 6 classes, destacando-se a classe (82-202) com 23,07%.

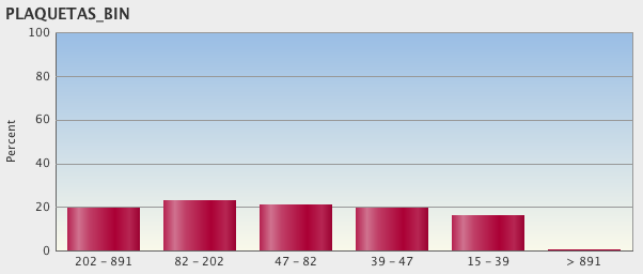


Figura 9: Classes do atributo Plaquetas

No atributo Custo, verifica-se a ocorrência em 4 classes, destacando-se a classe (0-.053) com 59,94%.

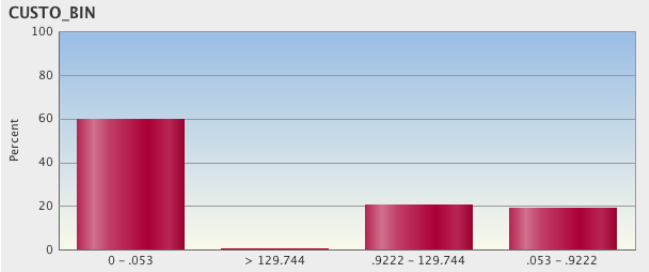


Figura 10: Classes do atributo Custo

No atributo Frequência Cardíaca Máxima, verifica-se a ocorrência em 6 classes, tendo uma distribuição aproximada de 20% nas 4 primeiras classes, 0,55% na classe (>229) e 18,51% na classe (49-80).

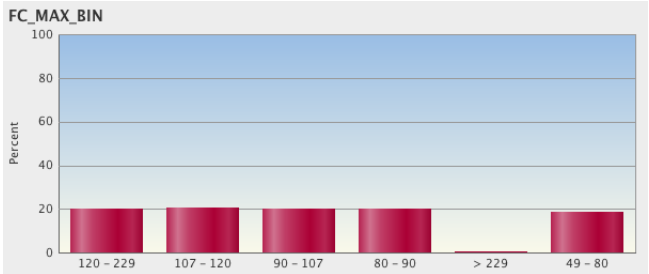


Figura 11: Classes do atributo Frequência Cardíaca Máxima

No atributo Frequência Cardíaca Mínima, verifica-se a ocorrência em 6 classes, tendo uma distribuição aproximada de 20% em 5 classes e 0,20% na classe (>157).

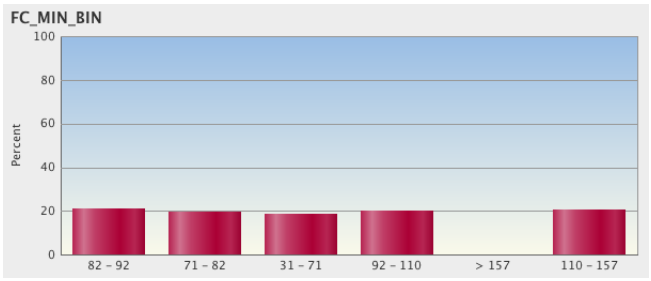


Figura 12: Classes do atributo Frequência Cardíaca Mímina

No atributo Pressão Arterial Média Máxima, verifica-se a ocorrência em 6 classes, tendo uma distribuição aproximada de 20% nas 5 primeiras classes e 0,10% na classe (>320).

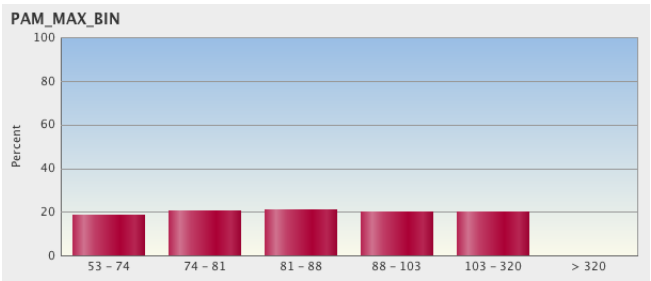


Figura 13: Classes do atributo Pressão Arterial Média Máxima

No atributo Pressão Arterial Média Mínima, verifica-se a ocorrência em 6 classes, tendo uma distribuição aproximada de 20% nas 5 primeiras classes e 0,10% na classe (>138).

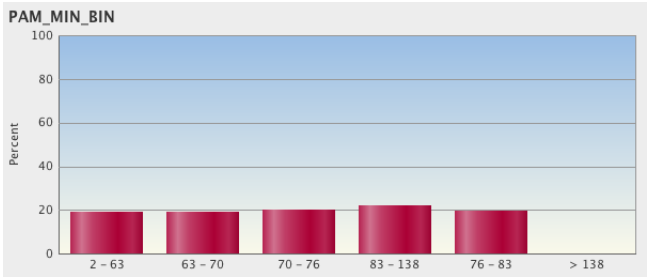


Figura 14: Classes do atributo Pressão Arterial Média Mínima

No atributo Pressão Arterial Sistólica Máxima, verifica-se a ocorrência em 6 classes, tendo uma distribuição aproximada de 20% em 5 classes e 0,20% na classe (>320).

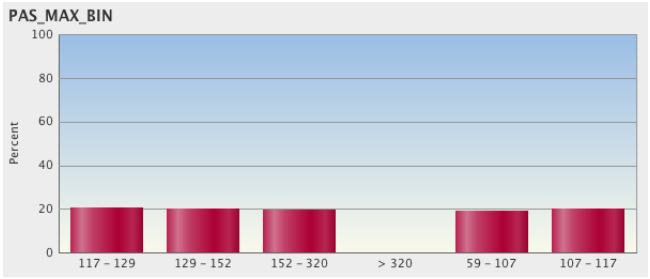


Figura 15: Classes do atributo Pressão Arterial Sistólica Máxima

No atributo Pressão Arterial Sistólica Mínima, verifica-se a ocorrência em 6 classes, tendo uma distribuição aproximada de 20% nas 5 primeiras classes e 0,10% na classe (>178).

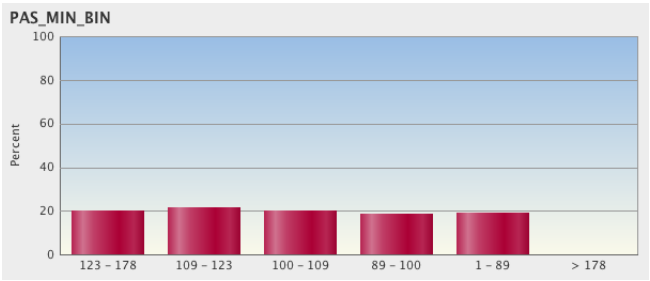


Figura 16: Classes do atributo Pressão Arterial Sistólica Mínima

No atributo Temperatura Máxima, verifica-se a ocorrência em 5 classes, destacando-se a classe (37-40) com 50,81%.

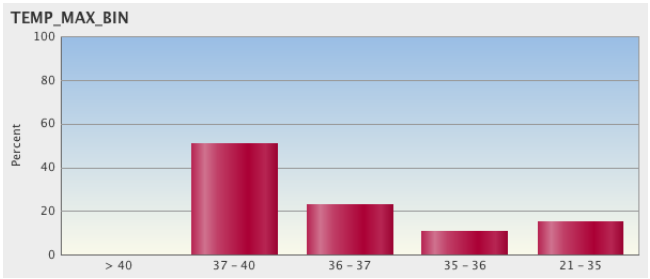


Figura 17: Classes do atributo Temperatura Máxima

No atributo Temperatura Mínima, verifica-se a ocorrência em 6 classes, destacando-se a classe (37-39) com 33,37%.

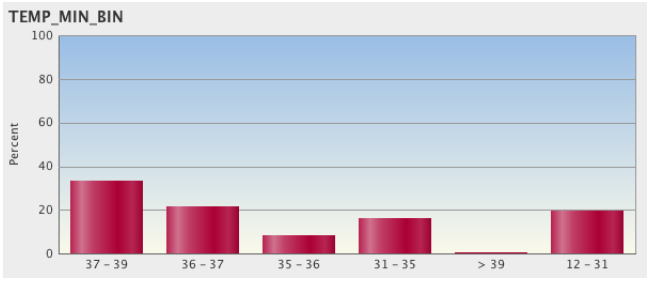


Figura 18: Classes do atributo Temperatura Mínima

No atributo Sépsis Bilirrubina, verifica-se a ocorrência de duas classes referentes aos valores 1 com 28,40% e 0 com 71,60%.

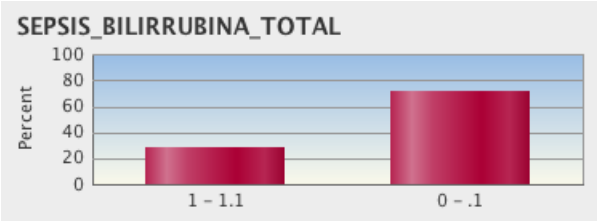


Figura 19: Classes do atributo Sépsis Bilirrubina

No atributo Sépsis Creatinina, verifica-se a ocorrência de duas classes referentes aos valores 1 com 32,76% e 0 com 67,24%.

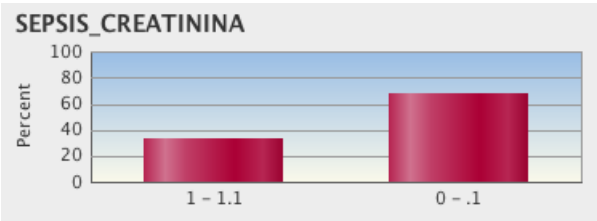


Figura 20: Classes do atributo Sépsis Creatinina

No atributo Sépsis Frequência Cardíaca, verifica-se a ocorrência de duas classes referentes aos valores 1 com 59,48% e 0 com 40,52%.

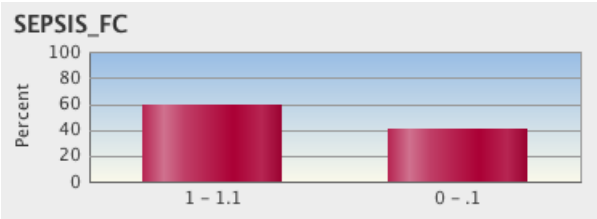


Figura 21: Classes do atributo Sépsis Frequência Cardíaca

No atributo Sépsis Glicose, verifica-se a ocorrência de duas classes referentes aos valores 1 com 63,69% e 0 com 36,31%.

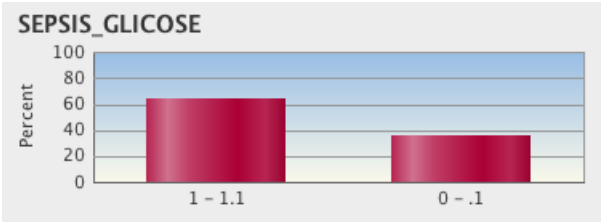


Figura 22: Classes do atributo Sépsis Glicose

No atributo Sépsis Leucócitos, verifica-se a ocorrência de uma classe referentes aos valores 1 com 100%.

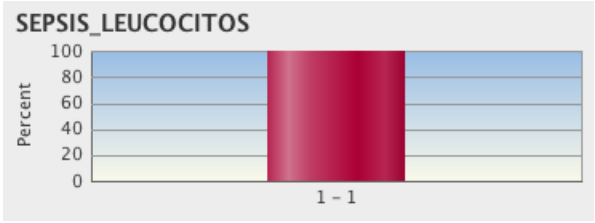


Figura 23: Classes do atributo Sépsis Leucócitos

No atributo Sépsis Pressão Arterial Média, verifica-se a ocorrência de duas classes referentes aos valores 1 com 24,29% e 0 com 75,71%.

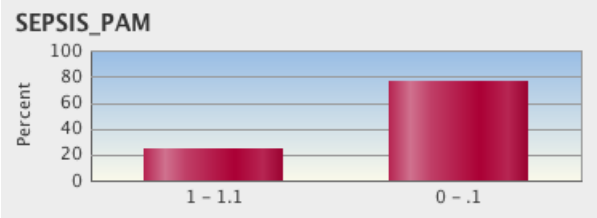


Figura 24: Classes do atributo Sépsis Pressão Arterial Média

No atributo Sépsis Pressão Arterial Sistólica, verifica-se a ocorrência de duas classes referentes aos valores 1 com 20,64% e 0 com 79,36%.

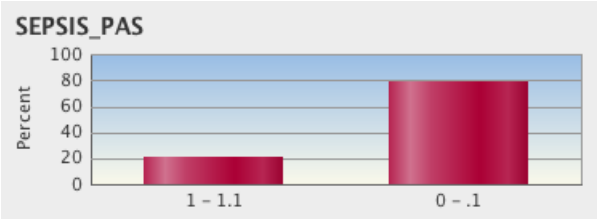


Figura 25: Classes do atributo Sépsis Pressão Arterial Sistólica

No atributo Sépsis Plaquetas, verifica-se a ocorrência de duas classes referentes aos valores 1 com 64,60% e 0 com 35,40%.

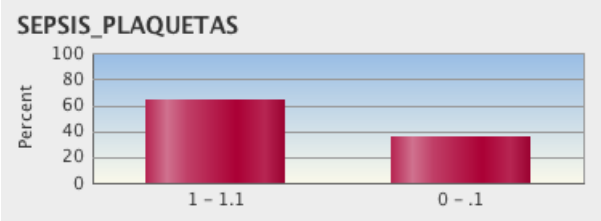


Figura 26: Classes do atributo Sépsis Plaquetas

No atributo Sépsis Temperatura, verifica-se a ocorrência de duas classes referentes aos valores 1 com 54,77% e 0 com 45,23%.

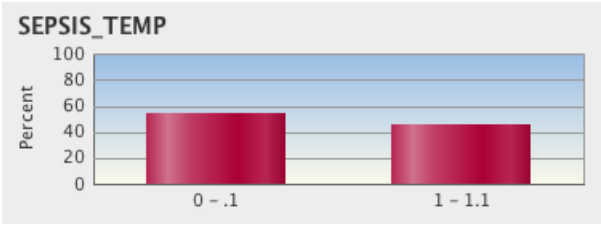


Figura 27: Classes do atributo Sépsis Temperatura

No atributo alvo Sépsis Final, verifica-se a ocorrência de duas classes referentes aos valores 2 com 19,52% e 3 com 80,48%.

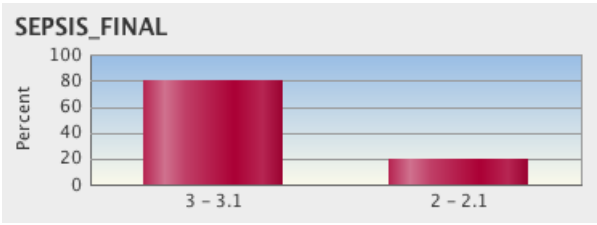


Figura 28: Classes do atributo Sépsis Final

No atributo alvo Grupo de Medicamento, verifica-se a ocorrência de onze classes, destacando-se a classe "Sistema Nervoso Central".

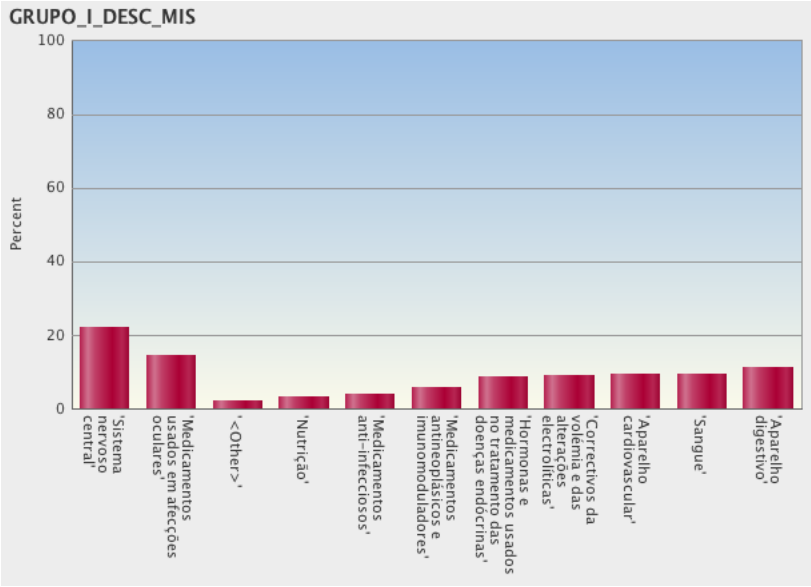


Figura 29: Classes do alvo Grupo de Medicamento

No atributo alvo Subgrupo de Medicamento, verifica-se a ocorrência de onze classes, destacando-se a classe "Outros".

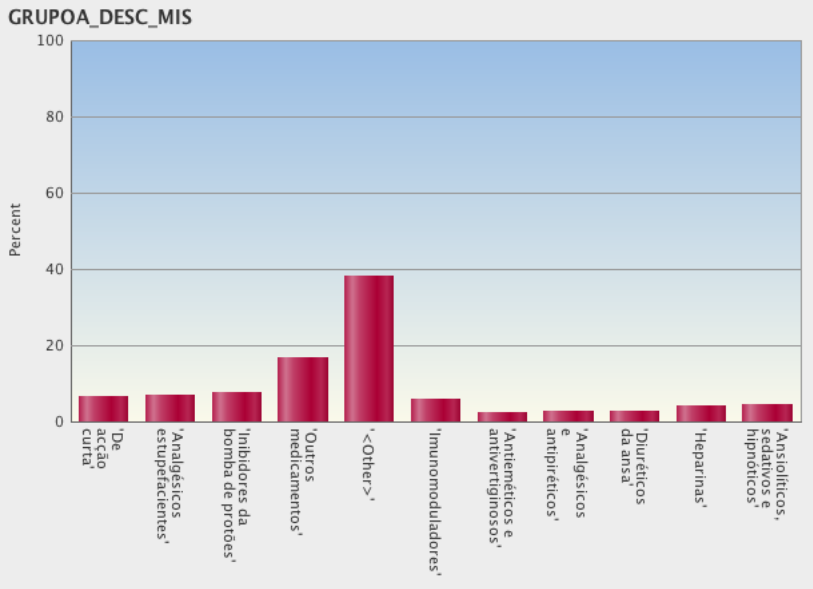


Figura 30: Classes do alvo Subgrupo de Medicamento

No atributo alvo Subsubgrupo de Medicamento, verifica-se a ocorrência de onze classes, destacando-se a classe "Outros".

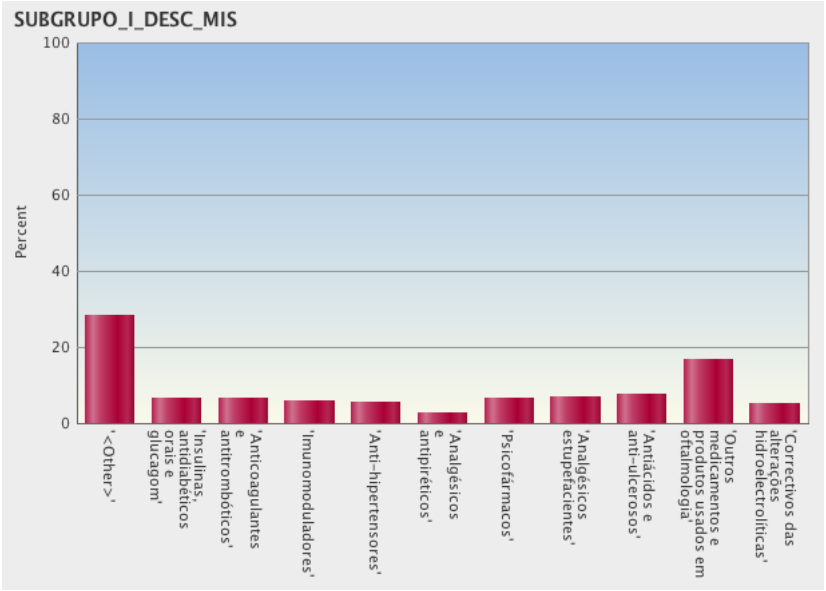


Figura 31: Classes do alvo Subsubrupo de Medicamento

6.6 Avaliação

A fim de utilizar todos os casos disponíveis e comparar a precisão da previsão, foi utilizado o método *cross-validation* (CV), para estimar a capacidade de síntese dos modelos de classificação.

Como o ODM implementa validação cruzada nas suas técnicas (Oracle, 2012), foi possível testar a precisão dos modelos sobre os dados que foram utilizados.

Visto que os resultados podem depender da divisão aleatória dos subconjuntos mutuamente exclusivos, foram aplicadas 10 execuções a cada subconjunto *10-fold*, num total de 10x10 = 100 resultados para cada configuração de teste.

Na avaliação dos resultados obtidos de forma a comparar os modelos de classificação referentes ao nível de sépsis, foi feita uma análise via curvas ROC (Bi and Bennett, 2003). As curvas ROC (*Receiver Operating Characteristic*) são frequentemente utilizadas na área médica para avaliar modelos computacionais de suporte à decisão, diagnóstico e prognóstico (Lasko et al., 2005).

De forma a comprovar todos os resultados atingidos na totalidade dos modelos desenvolvidos para a sépsis, na figura 32, pode-se observar através da ordenação crescente, os valores da acuidade para cada modelo de classificação e respetiva técnica. De recordar que os modelos M4 de cada técnica são automáticos e os restantes são provenientes da seleção manual dos atributos.

Dos doze cenários que dispomos, referentes ao nível de sépsis, foram selecionados para análise, os três melhores modelos não automáticos, um de cada técnica, sendo: SEPSIS_MVS_M3, SEPSIS_AD_M3 e SEPSIS_NB_M1, conforme figura 32.

Models

Name	Average Accuracy %	Algorithm
SEPSIS_MVS_M3	100	Support Vector Machine
SEPSIS_MVS_M1	100	Support Vector Machine
SEPSIS_MVS_M4	100	Support Vector Machine
SEPSIS_AD_M1	100	Decision Tree
SEPSIS_AD_M3	100	Decision Tree
SEPSIS_NB_M1	99,8856	Naive Bayes
SEPSIS_NB_M3	99,682	Naive Bayes
SEPSIS_NB_M4	99,5406	Naive Bayes
SEPSIS_MVS_M2	99,5053	Support Vector Machine
SEPSIS_AD_M2	99,5053	Decision Tree
SEPSIS_AD_M4	84,5427	Decision Tree
SEPSIS_NB_M2	49,8069	Naive Bayes

Figura 32: Totalidade dos modelos desenvolvidos para a sépsis

Através da análise da curva ROC, pode-se observar o elevado desempenho do modelo MVS_M3 para a determinação da sépsis, tal como se constata na Figura 33.

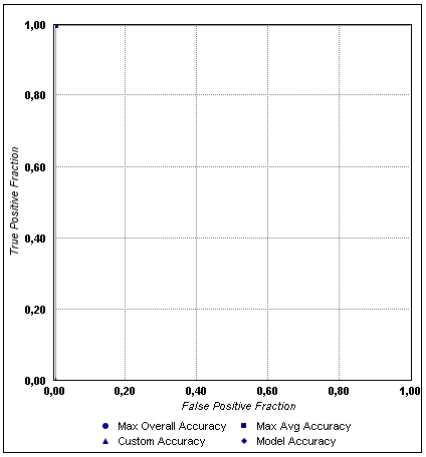


Figura 33: Curva ROC

Na figura 34 pode-se observar a árvore de decisão obtida a partir do modelo AD_M3.

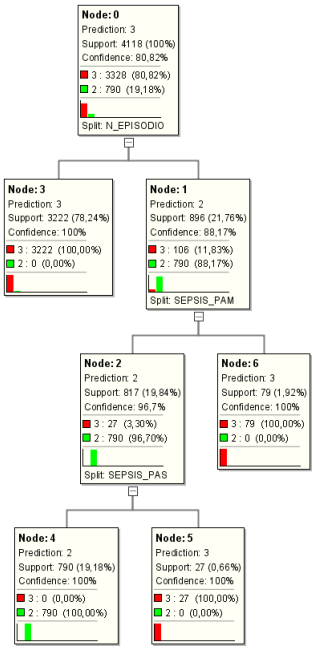


Figura 34: Árvore de decisão

As previsões dos melhores modelos de classificação foram analisadas e observou-se que a AD classificou as previsões do modelo de classificação AD_M3 com uma acuidade de 100%.

Ainda referente à avaliação dos resultados obtidos da classificação do nível de sépsis, foi feita uma análise através da matriz de confusão, da percentagem de erro total, sensibilidade, especificidade e acuidade dos três modelos referidos anteriormente. De forma a demonstrar a qualidade dos resultados atingidos, verifica-se através das tabelas 12, 13 e 14, as melhores previsões para o nível de sépsis fornecidas pelos modelos MVS_M3, AD_M3 e NB_M1 respetivamente.

Modelo MVS_M3	**Sépsis grave**	**Choque séptico**	Total	Corretos
Sépsis grave	334	0	334	100%
Choque séptico	0	1.415	1.415	100%
Total	334	1.415	**1.749**	
Corretos	100%	100%		

Erro total	**Sensibilidade**	**Especificidade**	**Acuidade**
0%	100%	100%	100%

Tabela 12: Matriz de confusão do modelo MVS_M3

Modelo AD_M3	**Sépsis grave**	**Choque séptico**	Total	Corretos
Sépsis grave	334	0	334	100%
Choque séptico	0	1.415	1.415	100%
Total	334	1.415	**1.749**	
Corretos	100%	100%		

Erro total	**Sensibilidade**	**Especificidade**	**Acuidade**
0%	100%	100%	100%

Tabela 13: Matriz de confusão do modelo AD_M3

Modelo NB_M1	**Sépsis grave**	**Choque séptico**	Total	Corretos
Sépsis grave	109	0	109	100%
Choque séptico	1	436	437	99,77%
Total	110	436	**546**	
Corretos	99,09%	100%		

Erro total	**Sensibilidade**	**Especificidade**	**Acuidade**
0,18%	100%	99,09%	99,82%

Tabela 14: Matriz de confusão do modelo NB_M1

Na avaliação dos resultados obtidos, de forma a comparar os modelos de classificação referentes ao plano terapêutico da sépsis, foi feita uma análise para

o melhor modelo por alvo, onde são apresentadas as respetivas classes. Como se verifica na tabela 15, na parte superior estão representados os quatro modelos desenvolvidos através da seleção manual dos atributos e na parte inferior da mesma tabela, os quatro modelos automáticos.

Alvos	Classes	Modelos	Acuidade
GRUPOMED_I (1º nível)	13	AD_M8	48,90%
GRUPOMED_A (2º nível)	49	NB_M8	45,30%
SUBGRUPOMED_I (3º nível)	37	AD_M6	45,19%
MEDICAMENTO	75	AD_M8	34,11%
GRUPOMED_I (1º nível)	13	AD_M4	62,84%
GRUPOMED_A (2º nível)	49	MVS_M4	44,00%
SUBGRUPOMED_I (3º nível)	37	MVS_M4	55,23%
MEDICAMENTO	75	MVS_M4	36,24%

Tabela 15: Melhor modelo por alvo, referente à terapêutica

De forma a comprovar todos os resultados atingidos na totalidade dos modelos desenvolvidos para o plano terapêutico da sépsis, no anexo G, nas figuras 35, 36, 37 e 38, pode-se observar os valores da acuidade para cada modelo de classificação e respetiva técnica. Dos sessenta e quatro cenários disponíves, referentes ao plano terapêutico da sépsis, foi selecionado para análise, o melhor modelo não automático e o melhor automático, sendo: GRUPOMED_I_AD_M8, GRUPOMED_A_NB_M8, SUBGRUPOMED_I_AD_M6, GRUPOMED_AD_M8, GRUPOMED_I_AD_M4, GRUPOMED_A_MVS_M4, SUBGRUPOMED_I_MVS_M4 e GRUPOMED_MVS_M4, conforme anexo G, figuras 35, 36, 37 e 38.

A importância relativa dos melhores modelos referentes à medicação, é apresentada pela acuidade na tabela 15. Relativamente aos resultados obtidos da classificação do plano terapêutico, foi feita uma análise através da matriz de desempenho dos melhores modelos referidos anteriormente. De forma a demonstrar os resultados atingidos, verifica-se através das tabelas 16, 17, 18 e 19, do anexo H, os baixos resultados para o plano terapêutico fornecidos pelos modelos da tabela 15.

Nas tabelas 16, 17 e 18 do anexo H, são apresentados apenas alguns resultados, devido à grande quantidade de dados. Nesse sentido foi feito um *snapshot* à totalidade de cada tabela, mostrando apenas os resultados com uma taxa de acerto superior a 80%. Nesses resultados, verifica-se que para alguns grupos de medicamentos, os níveis de assertividade são bastante aceitáveis.

6.7 Implementação

Este projeto de dissertação é relevante na medida em que foram desenvolvidos novos modelos de classificação do plano terapêutico de doentes com sépsis, com base nas várias técnicas de DM.

Com base nos bons resultados dos modelos de classificação, espera-se que este estudo possa servir de base à implementação de um Sistema de Apoio à Decisão (SAD) para operar em ambiente real e em tempo real.

Para a implementação do SAD, este depende de certas especificações técnicas, tais como: sistema com bons requisitos de rede; acesso contínuo à base de dados e a operar em tempo real; o servidor deverá ter boa capacidade de processamento e memória RAM.

6.8 Sumário

A elaboração da parte prática deste trabalho, guiada pelo referencial da metodologia CRISP-DM, permitiu o correto acompanhamento deste projeto de DM.

Devido ao tratamento *online* da grande quantidade de dados e dos modelos desenvolvidos, a adoção da ferramenta ODM contribuiu satisfatoriamente para o bom desempenho de todo o processo. O principal objetivo para este projeto foi a modelação de um conjunto de modelos de classificação, baseados em DM, onde se pretendeu prever, com elevado grau de acuidade, o nível de sépsis e o plano terapêutico de doentes com sépsis.

Nos primeiros testes realizadas, para o alvo SEPSIS_FINAL, os modelos de classificação obtiveram resultados satisfatórios, sendo: o modelo M3 com a técnica AD obteve 100%; o M3 com a técnica MVS obteve 100% e; o M1 com a técnica NB obteve 99,89%. Por sua vez, estes modelos revelaram possuir uma acuidade elevada, revelando serem de grande interesse.

Nos testes realizados para os alvos: GRUPOMED_I (1º nível), GRUPOMED_A (2º nível), SUBGRUPOMED_I (3º nível) e MEDICAMENTO, revelam uma importância relativa, porque só se verificam bons níveis de assertividade em algumas classes de medicamentos/grupos de medicamentos.

7 Resultados

Foram gerados e testados um total de 2 (variáveis alvo) * 4 (cenários) * 3 (técnicas de DM) = 24 modelos de classificação para a previsão do nível de sépsis. As tabelas 12, 13 e 14, apresentam os melhores resultados da previsão para cada conjunto de variáveis selecionadas, referindo a técnica, o cenário correspondente e os resultados relativos a: erro total, sensibilidade, especificidade e acuidade dos modelos.

Dos resultados obtidos para os níveis de sépsis, na figura 33 está representada a curva ROC obtida a partir do modelo SEPSIS_MVS_M3, que representa o melhor modelo em termos de acuidade (100%). Na figura 34 pode-se observar a árvore de decisão obtida a partir do modelo SEPSIS_AD_M3 com uma acuidade de 100%. Ainda relativamente à sépsis, verifica-se nas tabelas 12, 13 e 14, através das matrizes de confusão, as melhores previsões para o nível de sépsis fornecidas pelos modelos MVS_M3, AD_M3 e NB_M1 respetivamente.

Neste contexto, para a previsão da terapêutica, a avaliação dos modelos foi orientada à acuidade. Foram gerados e testados um total de 3 (variáveis alvo) * 6 (cenários) * 3 (técnicas de DM) = 64 modelos de classificação para a medicação. Assim, a tabela 15 apresenta os melhores resultados para a previsão do plano terapêutico e as correspondentes técnicas e classes, em termos de acuidade. Os resultados da referida tabela revelam uma importância relativa, porque só se verificam bons níveis de assertividade em alguns medicamentos/grupos de medicamentos. A título de exemplo, o grupo de medicamentos "Expectorantes", da tabela 17 do anexo H, apresenta uma taxa de verdadeiros positivos de 100% e de falsos negativos de 66,7%.

Em jeito de conclusão, foi possível prever com grande acuidade o nível de sépsis, no entanto, o mesmo já não é possível dizer no que diz respeito à medicação. Apesar de os modelos da sépsis terem bons resultados, o plano terapêutico não apresenta o mesmo nível de acuidade. Relativamente à sépsis, os resultados da acuidade, especificidade e sensibilidade foi de 100% nos modelos MVS_M3 e AD_M3. O modelo NB_M1, também referente à sépsis, teve 100% de sensibilidade, 99,09% de especificidade e 99,82% de acuidade.

Os resultados provam que de uma forma geral existe uma fraca correlação entre o nível de sépsis e o plano terapêutico, referente ao grupo de medicamento, sendo os resultados da acuidade os seguintes: Grupo_I, modelo AD_M8 teve 48,90%; Grupo_A, modelo NB_M8 teve 45,30%; Subgrupo_I, modelo AD_M6

teve 45,19% e; Medicamento com o modelo AD_M8 teve 34,11%. No entanto é de salientar que para alguns grupos de medicamentos, os modelos tiveram um bom desempenho (nível de acertos de algumas classes foi superior a 80%).

Com base nestes valores, pode-se ainda concluir que para obter bons resultados no plano terapêutico, é necessário adicionar outras variáveis que não estejam diretamente relacionadas com a sépsis.

8 Discussão

Com o terminar do projeto, é chegada a altura de confrontar os resultados obtidos com os objetivos inicialmente previstos. Nesse sentido é fundamental optar por uma postura crítica relativamente a todas as limitações que entretanto se verificam. Visto isso, convém referir que o projeto teve um elevado grau de complexidade nos processos de aquisição, transformação dos dados e indução dos modelos. Esse facto deveu-se à existência de muitos dados (total de 7 202 272 registos) e desses serem vários os registos com valores nulos ou fora de intervalos. Após o processo de seleção, tratamento e transformação dos dados foi obtido um conjunto de dados, a ser utilizado como tabela de entrada dos modelos de DM, com um total de 193 122 registos.

Por sua vez, relativamente à previsão dos níveis de sépsis, conseguiu-se bons resultados com os modelos de classificação, que apresentam uma acuidade de 100%, o que é muito bom. Relativamente à previsão do plano terapêutico, o mesmo não se pode dizer. O objetivo foi atingido de forma parcial, porque apesar de não se ter conseguido bons resultados a nivel da acuidade, é possivel provar que a correlação existente entre o nível de sépsis e o plano terapeutico é fraca.

O resultado deste trabalho é assim relevante para o domínio dos sistemas de apoio à decisão, pois pode ser muito útil na ajuda aos profissionais de saúde ao prever corretamente o nível de sépsis, permitindo perceber mais rapidamente a real condição do doente, evitando assim juizos errados.

A abordagem proposta neste trabalho é baseada em dados reais, onde foram gerados modelos que podem ser integrados num sistema de apoio à decisão, o que permitirá auxiliar os médicos no seu processo de tomada de decisão.

9 Conclusões

9.1 Síntese

Esta dissertação apresenta modelos de classificação, com dados recolhidos em tempo real na UCI do HSA, Porto, Portugal.

Foi considerado um grande conjunto inicial de dados, que resultou após transformação, num total de 193 112 registos, referentes a episódios que ocorreram em 305 dias de internamento e relativos a 394 doentes.

Os sistemas de BI permitem combinar a recolha de dados com ferramentas de análise, com o principal objetivo de disponibilizar informações para a tomada de decisão. De entre estes sistemas fazem parte as técnicas de DM para a extração de conhecimento. O DM foi um processo de extrema importância para este trabalho, tendo sido comprovada essa importância pela aplicação de algoritmos de aprendizagem com vista à procura de padrões e subsequente descoberta de informações úteis.

Para auxiliar na condução deste trabalho, recorreu-se à metodologia CRISP-DM e às ferramentas *SQL Developer* da *Oracle* e ODM. Dada a natureza das variáveis a modelar, optou-se por definir o objetivo de DM como sendo de classificação. Na fase de modelação foram adotadas três técnicas de classificação: MVSs, ADs e NBs. Após uma análise detalhada dos resultados da classificação obtidos, através de medições do erro total, sensibilidade, especificidade e acuidade, foram obtidos muito bons resultados para o nível de sépsis. Por sua vez, após uma análise dos resultados de classificação obtidos, através da matriz de confusão e métricas associadas, concluiu-se que os resultados obtidos têm uma elevada acuidade. Relativamente à terapêutica, os resultados da acuidade não foram satisfatórios, mas apesar disso, verificaram-se bons níveis de assertividade em alguns medicamentos/grupos de medicamentos.

A avaliação do nível de sépsis é uma tarefa crucial em ambientes de cuidados intensivos, por isso, quanto mais rapidamente o risco for identificado, mais rapidamente poderá ser aplicado o melhor tratamento. O desenvolvimento de modelos de classificação para a sépsis pode resultar não só na diminuição da mortalidade, mas também, na redução substancial de custos para as instituições, diagnosticando o correto nível de sépsis e consequente antecipando o correto tratamento a aplicar, evitando assim custos com experiências e testes de medicação.

Espera-se que os modelos de previsão proporcionem a adequada decisão ao médico, em relação à terapêutica a aplicar ao doente, apresentando uma elevada

taxa de sucesso.

O desenvolvimento de modelos de classificação para a sépsis pode ser um contributo fundamental para o desenvolvimento de um sistema de apoio à decisão.

9.2 Trabalho futuro

Após concluído este projeto, importa indicar, como trabalho futuro, o seguinte:

- Determinar novas variáveis que possam ser integradas nos modelos de modo a que se possa obter melhores resultados;
- Construir novos modelos, de modo a que seja verificada uma maior correlação entre as variáveis de entrada e alvo, ao nível da terapêutica;
- Implementar um sistema de apoio à decisão com os modelos da sépsis desenvolvidos neste projeto.

Bibliografia

Angus, D. C., Linde-Zwirble, W. T., Lidicker, J., Clermont, G., Carcillo, J., Pinsky, M., et al. (2001). Epidemiology of severe sepsis in the united states: analysis of incidence, outcome, and associated costs of care. *Critical Care Medicine, Baltimore*, 29(7):1303–1310.

APACHE, I. (1985). Apache ii: a severity of disease classification system.

Avison, D., Lau, F., Myers, M., and Nielsen, P. (1999). Action research. *Communications of the ACM*, 42(1):94–97.

Bi, J. and Bennett, K. P. (2003). Regression error characteristic curves. In *ICML*, pages 43–50.

Bone, R. C., Balk, R. A., Cerra, F. B., Dellinger, R. P., Fein, A. M., Knaus, W. A., Schein, R. M., and Sibbald, W. J. (1992). Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. the accp/sccm consensus conference committee. american college of chest physicians/society of critical care medicine. *Chest*, 101(6):1644–1655.

Chan, C. and Ting, H. (2011). Constructing a novel mortality prediction model with bayes theorem and genetic algorithm. *Expert Systems with Applications*, 38(7):7924–7928.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). Crisp-dm 1.0. *CRISP-DM Consortium*.

Cody, W., Kreulen, J., Krishna, V., and Spangler, W. (2002). The integration of business intelligence and knowledge management. *IBM Systems Journal*, 41(4):697–713.

Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning*, 20(3):273–297.

Cruz, A. and Cortez, P. (2009). Data mining via redes neuronais artificiais e máquinas de vectores de suporte. *Revista de Estudos Politécnicos*, 7(12):099–118.

De Turck, F., Decruyenaere, J., Thysebaert, P., Van Hoecke, S., Volckaert, B., Danneels, C., Colpaert, K., and De Moor, G. (2007). Design of a flexible

platform for execution of medical decision support agents in the intensive care unit. *Computers in Biology and Medicine*, 37(1):97–112.

Dellinger, R. P., Levy, M. M., Carlet, J. M., Bion, J., Parker, M. M., Jaeschke, R., Reinhart, K., Angus, D. C., Brun-Buisson, C., Beale, R., et al. (2008). Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2008. *Intensive Care Medicine*, 34(1):17–60.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996a). The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors (1996b). *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press.

Friedman, N. and Goldszmidt, M. (1996). Building classifiers using bayesian networks. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1277–1284.

Groth, R. (2000). *Data Mining: Building Competitive Advantage*. Prentice Hall PTR.

Guy, W. (1976). *ECDEU Assessment Manual for Psychopharmacology: 1976*. National Institute of Mental Health.

Guy, W. (2000). Clinical global impressions (cgi) scale. modified from: Rush j, et al. psychiatric measures.

Gwadry-Sridhar, F., Bauer, M., Lewden, B., and Hamou, A. (2011). A markov analysis of patients developing sepsis using clusters. *Knowledge Representation for Health-Care*, pages 85–100.

Handel, D. A. and Hackman, J. L. (2010). Implementing electronic health records in the emergency department. *The Journal of Emergency Medicine*, 38(2):257–263.

Jones, C. (1979). Glasgow coma scale. *AJN The American Journal of Nursing*, 79(9):1551–1557.

Kohavi, R. and Provost, F. (1998). Glossary of terms. *Machine Learning*, 30(June):271–274.

Kumar, A., Roberts, D., Wood, K. E., Light, B., Parrillo, J. E., Sharma, S., Suppes, R., Feinstein, D., Zanotti, S., Taiberg, L., et al. (2006). Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Critical Care Medicine*, 34(6):1589.

Langley, P. (1993). Induction of recursive bayesian classifiers. In *Machine Learning: ECML-93*, pages 153–164. Springer.

Langley, P. and Sage, S. (1994). Induction of selective bayesian classifiers. Technical report, DTIC Document.

Lasko, T., Bhagwat, J., Zou, K., Ohno-Machado, L., et al. (2005). The use of receiver operating characteristic curves in biomedical informatics. *Journal of biomedical informatics*, 38(5):404–415.

Le Gall, J., Lemeshow, S., and Saulnier, F. (1993). A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *JAMA: the journal of the American Medical Association*, 270(24):2957–2963.

Levy, M., Fink, M., Marshall, J., Abraham, E., Angus, D., Cook, D., Cohen, J., Opal, S., Vincent, J., and Ramsay, G. (2003a). 2001 sccm/esicm/accp/ats/sis international sepsis definitions conference. *Critical Care Medicine*, 31(4):1250–1256. [Special Articles].

Levy, M., Fink, M., Marshall, J., Abraham, E., Angus, D., Cook, D., Cohen, J., Opal, S., Vincent, J., and Ramsay, G. (2003b). 2001 sccm/esicm/accp/ats/sis international sepsis definitions conference. *Intensive Care Medicine*, 29(4):530–538.

Lyseng-Williamson, K. A. and Perry, C. M. (2002). Drotrecogin alfa (activated). *Drugs*, 62(4):617–630.

Markov (2011). Análise de markov. http://www.markovanalysis.com/. [Página web acedida em 24 de Novembro de 2011].

McNiff, J. (2002). Action research for professional development. *Concise Advice for New Action Researchers*.

Meyfroidt, G. (2009). How to implement information technology in the operating room and the intensive care unit. *Best Practice & Research Clinical Anaesthesiology*, 23(1):1–14.

Moreno, R., Metnitz, B., Adler, L., Hoechtl, A., Bauer, P., and Metnitz, P. (2008). Sepsis mortality prediction based on predisposition, infection and response. *Intensive care medicine*, 34(3):496–504.

Nemati, H. R. and Barko, C. D. (2001). Issues in organizational data mining: A survey of current practices. *Journal of Data Warehousing*, 6(1):25–36.

Nowinski, C. J., Becker, S. M., Reynolds, K. S., Beaumont, J. L., Caprini, C. A., Hahn, E. A., Peres, A., and Arnold, B. (2007). The impact of converting to an electronic health record on organizational culture and quality improvement. *International Journal of Medical Informatics*, 76:S174–S183.

O'brien, R. (1998). An overview of the methodological approach of action research. *Unpublished paper to Professor Joan Cherry, Course LIS3005Y, Faculty of Information Studies, University of Toronto. April*, 17.

ODM (2012). Binning (discretization). http://docs.oracle.com. [Página web acedida em 20 de Setembro de 2012].

Opal, S. (2005). Concept of piro as a new conceptual framework to understand sepsis. *Pediatric Critical Care Medicine*, 6(3):S55.

Oracle (2012). Data mining with oracle database 11g release 2 competing on in-database analytics. http://www.oracle.com/us/products/database/options/advanced-analytics/039550.pdf?ssSourceSiteId=ocomkr. [Página web acedida em 20 de Setembro de 2012].

Pereira, M., Curra, A., Rivas, R., Pereira, J., Banos, G., Teueiro, J., and Pazos, A. (2007). Computer aided monitoring system of intensive care unit patients. *WSEAS Transactions on Information Science and Applications*, 4(1):78–84.

Piatetsky-Shapiro, G. (2009). Data mining tools used poll. http://www.kdnuggets.com/polls/2009/data-mining-tools-used.htm. [Página web acedida em 31 de Maio de 2012].

Pontil, M. and Verri, A. (1998). Properties of support vector machines. *Neural Computation*, 10(4):955–974.

Portela, F., Vilas-Boas, M., and Santos, M. F. (2011). Improvements in data quality for decision support in intensive care. In *Electronic Healthcare: Third International Conference, Ehealth 2010, Casablanca, Morocco, December 13-15, 2010, Revised Selected Papers*, volume 69, page 86. Springer.

Póvoa, P. R., Carneiro, A. H., Ribeiro, O. S., Pereira, A. C., et al. (2009). Influence of vasopressor agent in septic shock mortality. results from the portuguese community-acquired sepsis study (saciuci study)*. *Critical Care Medicine*, 37(2):410.

Rabello, L., Rosolem, M., Leal, J., Soares, M., Lisboa, T., and Salluh, J. (2009). Entendendo o conceito piro: da teoria à prática clínica - parte 1. *Revista Brasileira Terapia Intensiva*, 21(4):425–431.

Ribas, V., Lopez, J., Ruiz-Rodriguez, J., Ruiz-Sanmartin, A., Rello, J., and Vellido, A. (2011). On the use of decision trees for icu outcome prediction in sepsis patients treated with statins. pages 37–43. cited By (since 1996) 0.

Ribas, V., Vellido, A., Ruiz-Rodríguez, J., and Rello, J. (2012). Severe sepsis mortality prediction with logistic regression over latent factors. *Expert Systems with Applications*, 39(2):1937–1943. cited By (since 1996) 0.

Rosolem, M., Rabello, L., Leal, J., Soares, M., Lisboa, T., and Salluh, J. (2010). Entendendo o conceito piro: da teoria à prática clínica: parte 2. *Revista Brasileira Terapia Intensiva*, 22(1):64–8.

Santos, M. F. and Azevedo, C. S. (2005). *Data Mining: Descoberta de Conhecimento em Bases de Dados*. FCA-Editora de Informática.

Santos, M. Y. and Ramos, I. (2006). *Business Intelligence: Tecnologias da Informação na Gestão de Conhecimento*. FCA-Editora de Informática.

Shorr, A. F., Micek, S. T., Jackson Jr, W. L., and Kollef, M. H. (2007). Economic implications of an evidence-based sepsis protocol: Can we improve outcomes and lower costs?*. *Critical Care Medicine*, 35(5):1257.

Silva, Á., Cortez, P., Santos, M., Gomes, L., and Neves, J. (2008). Rating organ failure via adverse events using data mining in the intensive care unit.

SSC (2010). Surviving sepsis campaign. http://www.survivingsepsis.org/Introduction/Pages/default.aspx. [Página web acedida em 24 de Novembro de 2011].

Strand, K. and Flaatten, H. (2008). Severity scoring in the icu: a review. *Acta Anaesthesiologica Scandinavica*, 52(4):467–478.

Turban, E., Sharda, R., Aronson, J., and King, D. (2008). *Business Intelligence: a Managerial Approach*. Pearson Prentice Hall. 1st edition.

Turban, E., Sharda, R., Delen, D., and King, D. (2011). *Business Intelligence: a Managerial Approach*. Pearson Prentice Hall. 2nd edition.

Vincent, J., Fink, M., Marini, J., Pinsky, M., Sibbald, W., Singer, M., Suter, P., Cook, D., Pepe, P., and Evans, T. (2006). Intensive care and emergency medicine: Progress over the past 25 years. *Chest*, 129(4):1061–1067.

Vincent, J., Moreno, R., Takala, J., Willatts, S., De Mendonca, A., Bruining, H., Reinhart, C., Suter, P., and Thijs, L. (1996). The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive care medicine*, 22(7):707–710.

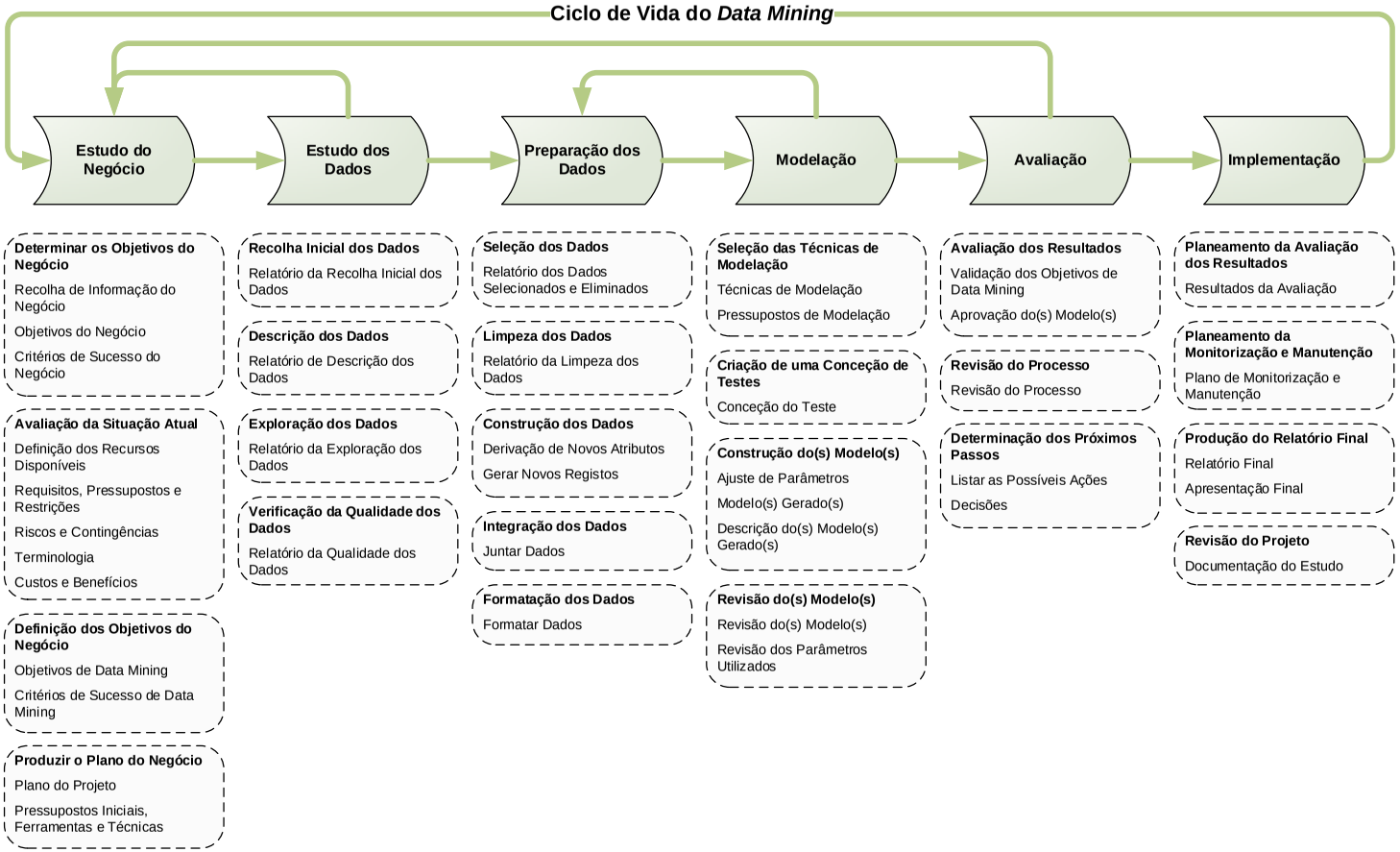
Vincent, J., Rello, J., Díaz, E., and Rodríguez, A. (2009). Management of sepsis: the piro approach. *Management of Sepsis: the PIRO Approach*.

Vincent, J. L. and Abraham, E. (2006). The last 100 years of sepsis. *American Journal of Respiratory and Critical Care Medicine*, 173(3):256.

West, D., Stowell, F., and Stansfield, M. (1985). Action research and information systems research. In Ellis, K., Gregory, A., Mears-Young, B., and Ragsdell, G., editors, *Critical Issues in Systems Theory and Practice*.

Wirth, R. (2000). Crisp-dm position statement. In *6th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, USA*.

Anexo A


Ciclo de Vida do *Data Mining*
Estudo do Negócio
Estudo dos Dados
Preparação dos Dados
Modelação
Avaliação
Implementação
Determinar os Objetivos do Negócio
Recolha de Informação do Negócio
Objetivos do Negócio
Critérios de Sucesso do Negócio
Avaliação da Situação Atual
Definição dos Recursos Disponíveis
Requisitos, Pressupostos e Restrições
Riscos e Contingências
Terminologia
Custos e Benefícios
Definição dos Objetivos do Negócio
Objetivos de Data Mining
Critérios de Sucesso de Data Mining
Produzir o Plano do Negócio
Plano do Projeto
Pressupostos Iniciais, Ferramentas e Técnicas
Recolha Inicial dos Dados
Relatório da Recolha Inicial dos Dados
Descrição dos Dados
Relatório de Descrição dos Dados
Exploração dos Dados
Relatório da Exploração dos Dados
Verificação da Qualidade dos Dados
Relatório da Qualidade dos Dados
Seleção dos Dados
Relatório dos Dados Selecionados e Eliminados
Limpeza dos Dados
Relatório da Limpeza dos Dados
Construção dos Dados
Derivação de Novos Atributos
Gerar Novos Registos
Integração dos Dados
Juntar Dados
Formatação dos Dados
Formatar Dados
Seleção das Técnicas de Modelação
Técnicas de Modelação
Pressupostos de Modelação
Criação de uma Conceção de Testes
Conceção do Teste
Construção do(s) Modelo(s)
Ajuste de Parâmetros
Modelo(s) Gerado(s)
Descrição do(s) Modelo(s) Gerado(s)
Revisão do(s) Modelo(s)
Revisão do(s) Modelo(s)
Revisão dos Parâmetros Utilizados
Avaliação dos Resultados
Validação dos Objetivos de Data Mining
Aprovação do(s) Modelo(s)
Revisão do Processo
Revisão do Processo
Determinação dos Próximos Passos
Listar as Possíveis Ações
Decisões
Planeamento da Avaliação dos Resultados
Resultados da Avaliação
Planeamento da Monitorização e Manutenção
Plano de Monitorização e Manutenção
Produção do Relatório Final
Relatório Final
Apresentação Final
Revisão do Projeto
Documentação do Estudo

Anexo B

	Task Name	Duration	Start	Finish
1	**Dissertação**	**256 days**	**Mon 07-11-11**	**Mon 29-10-12**
2	**Estudo do negócio**	**17 days**	**Mon 07-11-11**	**Tue 28-11-11**
3	Determinar os Objetivos do Negócio	5 days	Mon 07-11-11	Fri 11-11-11
4	Avaliação da Situação Atual	3 days	Mon 14-11-11	Wed 16-11-11
5	Definição dos Objetivos do Negócio	3 days	Thu 17-11-11	Mon 21-11-11
6	Produzir o Plano do Trabalho	4 days	Tue 22-11-11	Fri 25-11-11
7	Apresentação do Plano	1 day	Tue 29-11-11	Tue 29-11-11
8	Pesquisa e seleção de bibliografia	20 days	Mon 28-11-11	Fri 23-12-11
9	Revisão de literatura	20 days	Mon 02-01-12	Fri 27-01-12
10	Escrita do projeto de dissertação	20 days	Mon 30-01-12	Fri 24-02-12
11	**Estudo dos dados**	**30 days**	**Mon 27-02-12**	**Fri 06-04-12**
12	Recolha Inicial dos Dados	5 days	Mon 27-02-12	Fri 02-03-12
13	Descrição dos Dados	5 days	Mon 05-03-12	Fri 09-03-12
14	Exploração dos Dados	10 days	Mon 12-03-12	Fri 23-03-12
15	Verificação da Qualidade dos Dados	10 days	Mon 26-03-12	Fri 06-04-12
16	**Preparação dos dados**	**30 days**	**Mon 19-03-12**	**Fri 27-04-12**
17	Seleção dos Dados	6 days	Mon 19-03-12	Mon 26-03-12
18	Limpeza dos Dados	6 days	Tue 27-03-12	Tue 03-04-12
19	Construção dos Dados	6 days	Wed 04-04-12	Wed 11-04-12
20	Integração dos Dados	6 days	Thu 12-04-12	Thu 19-04-12
21	Formatação dos Dados	6 days	Fri 20-04-12	Fri 27-04-12
22	**Modelação**	**40 days**	**Mon 16-04-12**	**Fri 08-06-12**
23	Seleção das Técnicas de Modelação	10 days	Mon 16-04-12	Fri 27-04-12
24	Criação de uma Conceção de Testes	10 days	Mon 30-04-12	Fri 11-05-12
25	Construção do(s) Modelo(s)	10 days	Mon 14-05-12	Fri 25-05-12
26	Revisão do(s) Modelo(s)	10 days	Mon 28-05-12	Fri 08-06-12
27	**Avaliação**	**40 days**	**Mon 30-04-12**	**Fri 22-06-12**
28	Avaliação dos Resultados	13 days	Mon 30-04-12	Wed 16-05-12
29	Revisão do Processo	13 days	Thu 17-05-12	Mon 04-06-12
30	Determinação dos Próximos Passos	14 days	Tue 05-06-12	Fri 22-06-12
31	**Implementação (testes)**	**20 days**	**Mon 18-06-12**	**Fri 13-07-12**
32	Planeamento da Avaliação dos Resultados	5 days	Mon 18-06-12	Fri 22-06-12
33	Planeamento da Monitorização e Manutenção	5 days	Mon 25-06-12	Fri 29-06-12
34	Produção do Relatório Final	5 days	Mon 02-07-12	Fri 06-07-12
35	Revisão do Projeto	5 days	Mon 09-07-12	Fri 13-07-12
36	Escrita do relatório da dissertação	130 days	Mon 02-04-12	Fri 28-09-12
37	Revisão da Dissertação	20 days	Mon 01-10-12	Fri 26-10-12
38	Entrega da Dissertação	1 day	Mon 29-10-12	Mon 29-10-12

Anexo C

Atributos	Nulos	% Nulos	Valores distintos	% Valores distintos	Moda	Média	Mediana	Mínimo	Máximo	Desvio padrão	Variância
BILIRRUBINA	5	0.0368	9	0.0662	0 e 0,2	0.6	0.0	0	10	1.8	3.1
CREATININA	21	0.0036	838	0.1449	0.54	46.8	0.9	0.58	6370.2	234.5	54969.5
GLICOSE	2	0.0006	389	0.1103	0 e 1,14	142.3	130.0	0	1000	75.9	5767.1
LEUCOCITOS	20	0.0034	2235	0.3762	0 e -2	52.8	10.0	-25	53750	878.5	771719.3
PLAQUETAS	32	0.0058	656	0.1189	139	207.2	178.0	3	1701	146.9	21587.4
FC	0	0.0000	221	0.0005	85	86.5	86.0	0	250	19.4	375.3
PAM	0	0.0000	15753	0.0561	78.15	88.5	85.2	-39.66	319.96	23.8	564.5
PAS	0	0.0000	23047	0.0250	120.36	131.7	129.3	-40	320	30.0	897.9
TEMP	0	0.0000	2563	0.0056	36.73	34.0	36.2	-253.44	158.72	5.2	27.3
CUSTO	0	0.0000	2096	0.0237	0 e 0,70	7.11	0	0	2497.61	45.93	2110.17

Anexo D

Atributos	Nulos	% Nulos	Valores distintos	% Valores distintos	Moda	Média	Mínimo	Máximo	Desvio padrão	Variância
BILIRRUBINA	149232	78,47	206	46,71	0,39	1,96	0,12	27,02	3,93	15,46
CREATININA	139670	73,34	244	44,69	0,45	48,32	0,23	1996,6	229,63	52727,66
FC_MAX	2122	1,37	120	5,94	95	94,82	42	240	21,10	445,25
FC_MIN	2122	1,37	99	4,90	75	81,50	30	139	18,31	335,10
GLICOSE	164184	85,79	146	50,17	0	154,12	0	1000	83,95	7047,69
LEUCOCITOS	138486	70,56	468	77,61	0	57,36	0	12000	643,03	413486,11
PAM_MAX	60184	30,96	140	9,90	83	99,32	24	320	29,36	862,04
PAM_MIN	60184	30,96	99	7,00	72	79,39	7	239	16,93	286,74
PAS_MAX	10140	5,13	169	8,70	134	144,45	32	320	31,80	1011,36
PAS_MIN	10140	5,13	148	7,62	105	115,59	1	275	26,00	675,92
PLAQUETAS	140676	71,63	306	52,67	2	203,09	2	832	148,07	21923,93
TEMP_MAX	29907	14,40	22	1,25	37	35,97	19	44	3,16	10,01
TEMP_MIN	29907	14,40	26	1,48	37	33,83	12	40	5,31	28,24
SEPSIS_BILIRRUBINA	0	0,00	2	0,10	0	0,04	0	1	0,20	0,04
SEPSIS_CREATININA	0	0,00	2	0,10	0	0,06	0	1	0,24	0,06
SEPSIS_FC	0	0,00	2	0,10	1	0,57	0	1	0,50	0,25
SEPSIS_GLICOSE	0	0,00	2	0,10	0	0,10	0	1	0,30	0,09
SEPSIS_LEUCOCITOS	0	0,00	2	0,10	0	0,29	0	1	0,46	0,21
SEPSIS_PAM	0	0,00	2	0,10	0	0,12	0	1	0,32	0,10
SEPSIS_PAS	0	0,00	2	0,10	0	0,13	0	1	0,34	0,11
SEPSIS_PLAQUETAS	0	0,00	2	0,10	0	0,08	0	1	0,27	0,07
SEPSIS_TEMP	0	0,00	2	0,10	0	0,33	0	1	0,47	0,22
SEPSIS_FINAL	0	0,00	3	0,15	2	1,90	0	3	1,00	1,01

Anexo E

ATRIBUTOS	Entrada				Alvo
	M1	M2	M3	M4	
BILIRRUBINA		X	X	Auto	
CREATININA		X	X	Auto	
CUSTO	X	X	X	Auto	
DATA_VALOR	X	X	X	Auto	
FC_MAX		X	X	Auto	
FC_MIN		X	X	Auto	
GLICOSE		X	X	Auto	
HORA_VALIDACAO	X	X	X	Auto	
HORA_VALOR	X	X	X	Auto	
LEUCOCITOS		X	X	Auto	
N_EPISODIO	X	X	X	Auto	
PAM_MAX		X	X	Auto	
PAM_MIN		X	X	Auto	
PAS_MAX		X	X	Auto	
PAS_MIN		X	X	Auto	
PLAQUETAS		X	X	Auto	
TEMP_MAX		X	X	Auto	
TEMP_MIN		X	X	Auto	
SEPSIS_BILIRRUBINA	X		X	Auto	
SEPSIS_CREATININA	X		X	Auto	
SEPSIS_FC	X		X	Auto	
SEPSIS_GLICOSE	X		X	Auto	
SEPSIS_LEUCOCITOS	X		X	Auto	
SEPSIS_PAM	X		X	Auto	
SEPSIS_PAS	X		X	Auto	
SEPSIS_PLAQUETAS	X		X	Auto	
SEPSIS_TEMP	X		X	Auto	
SEPSIS_FINAL					X

Anexo F

ATRIBUTOS	Entrada										Alvo
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	
BILIRRUBINA		X	X	Auto			X	X		X	
CREATININA		X	X	Auto			X	X		X	
CUSTO	X	X	X	Auto							
DATA_VALOR	X	X	X	Auto							
FC_MAX		X	X	Auto			X	X		X	
FC_MIN		X	X	Auto			X	X		X	
GLICOSE		X	X	Auto			X	X		X	
HORA_VALIDACAO	X	X	X	Auto							
HORA_VALOR	X	X	X	Auto							
LEUCOCITOS		X	X	Auto			X	X		X	
N_EPISODIO	X	X	X	Auto	X	X	X	X			
PAM_MAX		X	X	Auto			X	X		X	
PAM_MIN		X	X	Auto			X	X		X	
PAS_MAX		X	X	Auto			X	X		X	
PAS_MIN		X	X	Auto			X	X		X	
PLAQUETAS		X	X	Auto			X	X		X	
TEMP_MAX		X	X	Auto			X	X		X	
TEMP_MIN		X	X	Auto			X	X		X	
SEPSIS_BILIRRUBINA	X		X	Auto	X	X	X	X	X		
SEPSIS_CREATININA	X		X	Auto	X	X	X	X	X		
SEPSIS_FC	X		X	Auto	X	X	X	X	X		
SEPSIS_GLICOSE	X		X	Auto	X	X	X	X	X		
SEPSIS_LEUCOCITOS	X		X	Auto	X	X	X	X	X		
SEPSIS_PAM	X		X	Auto	X	X	X	X	X		
SEPSIS_PAS	X		X	Auto	X	X	X	X	X		
SEPSIS_PLAQUETAS	X		X	Auto	X	X	X	X	X		
SEPSIS_TEMP	X		X	Auto	X	X	X	X	X		
SEPSIS_FINAL					X		X				
GRUPO_I											X
GRUPO_A											X
SUBGRUPO_I											X
DESIGNACAO											X

Anexo G

Models

Name	Average Accuracy %	Algorithm
GRUPOMED_I_AD_M4	62,8434	Decision Tree
GRUPOMED_I_MVS_M4	62,0409	Support Vector Machine
GRUPOMED_I_NB_M4	58,3492	Naive Bayes
GRUPOMED_I_AD_M8	48,9356	Decision Tree
GRUPOMED_I_AD_M6	48,8141	Decision Tree
GRUPOMED_I_AD_M7	45,8685	Decision Tree
GRUPOMED_I_AD_M5	45,6742	Decision Tree
GRUPOMED_I_AD_M10	44,1883	Decision Tree
GRUPOMED_I_NB_M5	42,7406	Naive Bayes
GRUPOMED_I_NB_M6	42,576	Naive Bayes
GRUPOMED_I_AD_M1	39,6579	Decision Tree
GRUPOMED_I_AD_M3	38,8335	Decision Tree
GRUPOMED_I_NB_M8	37,9355	Naive Bayes
GRUPOMED_I_NB_M7	37,8661	Naive Bayes
GRUPOMED_I_AD_M2	36,167	Decision Tree
GRUPOMED_I_NB_M10	33,2123	Naive Bayes
GRUPOMED_I_NB_M1	32,1747	Naive Bayes
GRUPOMED_I_NB_M2	31,4164	Naive Bayes
GRUPOMED_I_NB_M3	29,7901	Naive Bayes
GRUPOMED_I_AD_M9	28,7667	Decision Tree
GRUPOMED_I_MVS_M5	26,8084	Support Vector Machine
GRUPOMED_I_MVS_M8	22,6611	Support Vector Machine
GRUPOMED_I_MVS_M6	20,974	Support Vector Machine
GRUPOMED_I_MVS_M3	20,7348	Support Vector Machine
GRUPOMED_I_MVS_M7	18,7497	Support Vector Machine
GRUPOMED_I_MVS_M2	18,6453	Support Vector Machine
GRUPOMED_I_MVS_M1	17,9605	Support Vector Machine
GRUPOMED_I_MVS_M10	16,0293	Support Vector Machine
GRUPOMED_I_MVS_M9	13,8598	Support Vector Machine
GRUPOMED_I_NB_M9	8,3962	Naive Bayes

Figura 35: Totalidade dos modelos do grupo de medicamento

Models

Name	Average Accuracy %	Algorithm
GRUPOMED_A_NB_M8	45,3021	Naive Bayes
GRUPOMED_A_MVS_M4	44,0016	Support Vector Machine
GRUPOMED_A_NB_M4	42,3124	Naive Bayes
GRUPOMED_A_NB_M5	40,3908	Naive Bayes
GRUPOMED_A_AD_M6	39,8274	Decision Tree
GRUPOMED_A_AD_M5	39,1957	Decision Tree
GRUPOMED_A_AD_M10	37,342	Decision Tree
GRUPOMED_A_AD_M7	35,3721	Decision Tree
GRUPOMED_A_AD_M8	35,0616	Decision Tree
GRUPOMED_A_NB_M10	33,6049	Naive Bayes
GRUPOMED_A_AD_M4	33,0786	Decision Tree
GRUPOMED_A_NB_M1	25,5297	Naive Bayes
GRUPOMED_A_AD_M9	24,5316	Decision Tree
GRUPOMED_A_AD_M2	24,4964	Decision Tree
GRUPOMED_A_NB_M2	24,1324	Naive Bayes
GRUPOMED_A_NB_M9	22,3124	Naive Bayes
GRUPOMED_A_NB_M3	22,1334	Naive Bayes
GRUPOMED_A_AD_M3	18,7978	Decision Tree
GRUPOMED_A_MVS_M7	18,6527	Support Vector Machine
GRUPOMED_A_MVS_M10	16,4159	Support Vector Machine
GRUPOMED_A_MVS_M6	14,3463	Support Vector Machine
GRUPOMED_A_MVS_M5	13,6566	Support Vector Machine
GRUPOMED_A_MVS_M2	13,6187	Support Vector Machine
GRUPOMED_A_MVS_M3	13,5993	Support Vector Machine
GRUPOMED_A_MVS_M1	12,8104	Support Vector Machine
GRUPOMED_A_MVS_M9	8,7586	Support Vector Machine
GRUPOMED_A_AD_M1	3,2899	Decision Tree
GRUPOMED_A_NB_M6	1,7901	Naive Bayes
GRUPOMED_A_NB_M7	1,6559	Naive Bayes
GRUPOMED_A_MVS_M8	0,7876	Support Vector Machine

Figura 36: Totalidade dos modelos do subgrupo de medicamento

Models

Name	Average Accuracy %	Algorithm
SUBGRUPOMED_I_MVS_M4	55,2324	Support Vector Machine
SUBGRUPOMED_I_AD_M4	53,5712	Decision Tree
SUBGRUPOMED_I_NB_M4	51,997	Naive Bayes
SUBGRUPOMED_I_AD_M6	45,1928	Decision Tree
SUBGRUPOMED_I_AD_M5	44,8103	Decision Tree
SUBGRUPOMED_I_NB_M5	42,1301	Naive Bayes
SUBGRUPOMED_I_AD_M7	41,7004	Decision Tree
SUBGRUPOMED_I_NB_M10	41,611	Naive Bayes
SUBGRUPOMED_I_AD_M8	41,1454	Decision Tree
SUBGRUPOMED_I_NB_M8	38,2003	Naive Bayes
SUBGRUPOMED_I_AD_M2	32,41	Decision Tree
SUBGRUPOMED_I_AD_M3	31,046	Decision Tree
SUBGRUPOMED_I_NB_M1	29,548	Naive Bayes
SUBGRUPOMED_I_AD_M9	28,3163	Decision Tree
SUBGRUPOMED_I_NB_M2	28,1905	Naive Bayes
SUBGRUPOMED_I_NB_M3	26,6197	Naive Bayes
SUBGRUPOMED_I_NB_M9	23,1239	Naive Bayes
SUBGRUPOMED_I_MVS_M1	20,6803	Support Vector Machine
SUBGRUPOMED_I_MVS_M7	20,6758	Support Vector Machine
SUBGRUPOMED_I_MVS_M10	19,8272	Support Vector Machine
SUBGRUPOMED_I_MVS_M3	19,685	Support Vector Machine
SUBGRUPOMED_I_NB_M6	19,058	Naive Bayes
SUBGRUPOMED_I_MVS_M2	19,0253	Support Vector Machine
SUBGRUPOMED_I_MVS_M8	16,8033	Support Vector Machine
SUBGRUPOMED_I_MVS_M5	16,6401	Support Vector Machine
SUBGRUPOMED_I_MVS_M6	13,1371	Support Vector Machine
SUBGRUPOMED_I_MVS_M9	11,9903	Support Vector Machine
SUBGRUPOMED_I_NB_M7	4,3707	Naive Bayes
SUBGRUPOMED_I_AD_M1	2,7971	Decision Tree
SUBGRUPOMED_I_AD_M10	2,6719	Decision Tree

Figura 37: Totalidade dos modelos do subsubgrupo de medicamento

Models

Name	Average Accuracy %	Algorithm
GRUPOMED_MVS_M4	36,2449	Support Vector Machine
GRUPOMED_NB_M4	34,4868	Naive Bayes
GRUPOMED_AD_M8	34,1187	Decision Tree
GRUPOMED_AD_M7	34,1187	Decision Tree
GRUPOMED_AD_M4	31,8188	Decision Tree
GRUPOMED_NB_M5	30,9277	Naive Bayes
GRUPOMED_NB_M6	30,4185	Naive Bayes
GRUPOMED_AD_M10	29,9836	Decision Tree
GRUPOMED_AD_M6	29,8232	Decision Tree
GRUPOMED_AD_M5	29,8232	Decision Tree
GRUPOMED_NB_M8	26,4552	Naive Bayes
GRUPOMED_NB_M7	26,4483	Naive Bayes
GRUPOMED_NB_M10	24,0555	Naive Bayes
GRUPOMED_MVS_M8	19,7567	Support Vector Machine
GRUPOMED_NB_M9	18,9485	Naive Bayes
GRUPOMED_AD_M9	16,861	Decision Tree
GRUPOMED_AD_M3	16,4008	Decision Tree
GRUPOMED_NB_M2	15,7385	Naive Bayes
GRUPOMED_MVS_M10	15,5021	Support Vector Machine
GRUPOMED_NB_M1	15,3737	Naive Bayes
GRUPOMED_AD_M2	15,2605	Decision Tree
GRUPOMED_NB_M3	14,6203	Naive Bayes
GRUPOMED_MVS_M7	13,5795	Support Vector Machine
GRUPOMED_MVS_M2	12,928	Support Vector Machine
GRUPOMED_MVS_M5	12,4121	Support Vector Machine
GRUPOMED_MVS_M1	12,2221	Support Vector Machine
GRUPOMED_MVS_M3	11,4726	Support Vector Machine
GRUPOMED_MVS_M6	9,7255	Support Vector Machine
GRUPOMED_MVS_M9	6,4393	Support Vector Machine
GRUPOMED_AD_M1	0,0671	Decision Tree

Figura 38: Totalidade dos modelos dos medicamentos

Anexo H

	Aparelho cardiovascular	Aparelho digestivo	Aparelho geniturinário	Aparelho respiratório	Correctivos da volémia e das alterações electrolíticas	Hormonas e medicamentos usados no tratamento das doenças endócrinas	Medicamentos anti-infecciosos	Medicamentos antineoplásicos e imunomoduladores	Medicamentos usados em afecções cutâneas	Medicamentos usados em afecções oculares	Nutrição	Sangue	Sistema nervoso central	Correctos (%)	Total
Aparelho cardiovascular	6	0	0	41	39	0	23	0	0	0	0	0	0	3,6	168
Aparelho digestivo	10	30	0	22	24	43	19	0	0	0	0	0	0	13,9	216
Aparelho geniturinário	0	0	11	0	0	0	0	0	0	0	0	0	0	84,6	13
Aparelho respiratório	0	0	0	23	0	0	0	0	2	0	0	0	0	92,0	25
Correctivos da volémia e das alterações electrolíticas	8	0	0	17	77	0	0	5	0	41	0	6	0	50,0	154
Hormonas e medicamentos usados no tratamento das doenças endócrinas	0	0	0	6	29	56	8	51	0	12	0	0	0	34,6	162
Medicamentos anti-infecciosos	0	0	13	0	0	0	57	3	0	0	0	0	0	78,1	73
Medicamentos antineoplásicos e imunomoduladores	0	0	0	0	4	6	13	89	0	0	0	0	0	79,5	112
Medicamentos usados em afecções cutâneas	0	0	0	0	0	0	0	0	2	0	0	0	0	100,0	2
Medicamentos usados em afecções oculares	0	0	13	24	78	0	0	0	7	121	0	6	0	48,6	249
Nutrição	0	0	2	21	12	0	6	0	0	3	10	0	0	18,5	54
Sangue	0	0	14	10	26	21	36	0	0	0	6	47	0	29,4	160
Sistema nervoso central	0	34	29	39	24	12	2	93	3	87	7	30	13	3,5	373
Correctos (%)	25,0	46,9	11,7	11,3	24,6	40,6	34,8	36,9	10,5	38,3	19,6	38,8	100,0	542	
Total	24	64	94	203	313	138	164	241	19	316	51	121	13		

Tabela 16: Matriz do grupo de medicamento do modelo AD_M8

	Acção periférica	Alcalinizantes	Análogos nucleosídeos inibidores da transcriptase inversa (reversa)	Antiepilépticos e anticonvulsivantes	Anti-hemorroidários	Anti-infecciosos	Associações de vitaminas	Bloqueadores beta e alfa	Corticosteróides de aplicação tópica	Diuréticos poupadores de potássio	Expectorantes	Laxantes osmóticos	Potássio	Resinas permutadoras de catiões	Suplementos enzimáticos, bacilos lácteos e análogos	Correctos (%)	Total
Acção periférica	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11,1	72
Alcalinizantes	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	4,5	201
Análogos nucleosídeos inibidores da transcriptase inversa (reversa)	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	14,3	42
Antiepilépticos e anticonvulsivantes	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	25,0	8
Anti-hemorroidários	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	60,0	5
Anti-infecciosos	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	15,3	72
Associações de vitaminas	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	23,5	17
Bloqueadores beta e alfa	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	40,9	22
Corticosteróides de aplicação tópica	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	6,7	15
Diuréticos poupadores de potássio	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	16,7	48
Expectorantes	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	66,7	9
Laxantes osmóticos	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	44,8	29
Potássio	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	54,5	22
Resinas permutadoras de catiões	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	31,8	22
Suplementos enzimáticos, bacilos lácteos e análogos	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	20,8	24
Correctos (%)	100,0	90,0	85,7	100,0	100,0	100,0	80,0	81,8	100,0	100,0	100,0	81,3	85,7	100,0	83,3	378	
Total	8	10	7	2	3	11	5	11	1	8	6	16	14	7	6		

Tabela 17: Matriz do subgrupo de medicamento do modelo NB_M8

	Antidislipidémicos	Antiepilépticos e anticonvulsivantes	Anti-hemorroidários	Anti-inflamatórios intestinais	Antitússicos e expectorantes	Correctivos do equilíbrio ácido-base	Corticosteróides	Medicamentos captadores de iões	Medicamentos de aplicação tópica na vagina	Relaxantes musculares	Soluções para diálise peritoneal	Corretos (%)	Total
Antidislipidémicos	8	0	0	0	0	0	0	0	0	0	0	88,9	9
Antiepilépticos e anticonvulsivantes	0	2	0	0	0	0	0	0	0	0	0	100,0	2
Anti-hemorroidários	0	0	2	0	0	0	0	0	0	0	0	100,0	2
Anti-inflamatórios intestinais	0	0	0	31	0	0	0	0	0	0	0	83,8	37
Antitússicos e expectorantes	0	0	0	0	6	0	0	0	0	0	0	100,0	6
Correctivos do equilíbrio ácido-base	0	0	0	0	0	10	0	0	0	0	0	100,0	10
Corticosteróides	0	0	0	0	0	0	25	0	0	0	0	100,0	25
Medicamentos captadores de iões	0	0	0	0	0	0	0	7	0	0	0	100,0	7
Medicamentos de aplicação tópica na vagina	0	0	0	0	0	0	0	0	16	0	0	94,1	17
Relaxantes musculares	0	0	0	0	0	0	0	0	0	6	0	100,0	6
Soluções para diálise peritoneal	0	0	0	0	0	0	0	0	0	4	17	81,0	21
Corretos (%)	11,8	50,0	50,0	23,7	30,0	5,2	9,4	33,3	19,5	8,1	18,7	360	
Total	68	4	4	131	20	193	266	21	82	74	91		

Tabela 18: Matriz do subsubgrupo de medicamento do modelo AD_M6

	Complexo B	Dieta Entérica Completa Polimérica Enriquecida	Insulina Acção Intermédia	Levetiracetam	Ramipril	Resina Permutadora Catioes	Trazodona	Correct %	Total
Complexo B	3	0	0	0	0	0	0	100,0	3
Dieta Entérica Completa Polimérica Enriquecida	0	9	0	0	0	0	0	90,0	10
Insulina Acção Intermédia	0	0	14	0	0	0	0	93,3	15
Levetiracetam	0	0	0	2	0	0	0	100,0	2
Ramipril	0	0	0	0	9	0	0	90,0	10
Resina Permutadora Catioes	0	0	0	0	0	7	0	100,0	7
Trazodona	0	0	0	0	0	0	10	100,0	10
Correct %	18,8	81,8	15,2	16,7	40,9	41,2	47,6	317	
Total	16	11	92	12	22	17	21		

Tabela 19: Matriz dos medicamentos do modelo AD_M8

Glossário

- **Antibioterapia**: (antibiótico + terapia) tratamento efetuado com antibióticos
- **Antibióticos de amplo espectro**: Tratamento que tem como alvo um grande número de micro-organismos diferentes
- **Antibióticos**: Tratamento utilizado para infeções
- **Apneia:** Período sem respirar
- **Choque Séptico**: Falência de múltiplos órgãos resultantes de avanço da sépsis grave
- **Disfunção orgânica**: Órgãos que não funcionam como deveriam
- **Falência de órgãos**: Órgãos que não funcionam
- **Hiperventilação**: Níveis elevados de respiração rápida
- **Hipoglicemia**: Níveis baixos de açúcar no sangue
- **Hipotermia**: Baixa temperatura corporal
- **Infeção**: Fenómeno microbiológico caracterizado por uma resposta inflamatória à presença de microrganismos ou à invasão de um tecido normalmente estéril pelos mesmos
- **Intravenoso**: Administração rápida de grandes quantidades de fluidos rapidamente
- **O tratamento empírico**: Amplo tratamento antibiótico com base na experiência prévia do micro-organismo
- **Sépsis**: A resposta do organismo a uma infeção
- **Sépsis grave**: Avanço da sépsis nos órgãos mais afetados
- **Septicemia**: Emergência médica causada por bactérias e toxinas no sangue
- **Taquicardia**: Batimento cardíaco rápido