Vítor Hugo Antero de Melo

ANÁLISE DE DADOS PÚBLICOS DE SAÚDE COM MAPAS AUTO-ORGANIZÁVEIS DE KOHONEN (SOM)

Trabalho de Graduação

Universidade Federal de Pernambuco
secgrad@cin.ufpe.br
www.cin.ufpe.br/~secgrad

RECIFE
2015

Universidade Federal de Pernambuco
Centro de Informática
Graduação em Ciência da Computação

Vítor Hugo Antero de Melo

ANÁLISE DE DADOS PÚBLICOS DE SAÚDE COM MAPAS AUTO-ORGANIZÁVEIS DE KOHONEN (SOM)

Trabalho apresentado ao Programa de Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: *Germano Crispim Vasconcelos*

RECIFE
2015

Agradecimentos

Agradeço primeiramente aos meus pais e minha família, por sempre acreditarem e investirem na minha educação. Eles são peças chaves na minha formação pessoal e jamais titubearam ao oferecer apoio.

Meus agradecimentos seguem ao professor Germano Vasconcelos pela orientação e por ser sempre solícito durante a elaboração deste projeto.

Agradeço aos amigos, por ajudarem a suportar os momentos difíceis do curso e estarem sempre presentes durante as noites em claro no Centro de Informática, transformando-as muitas vezes em momentos de descontração.

À minha namorada, por ser sempre ouvidos sobre os desafios do curso, mesmo não sendo da área de computação.

Nós podemos ver pouco à frente, mas podemos ver que há muito a ser feito.

—ALAN TURING

Resumo

A quantidade de dados abertos para o público aumenta a cada dia, impulsionando grandes avanços na área de análise de dados. Aderindo à tendência, diferentes esferas governamentais no Brasil disponibilizam os dados de suas secretarias, ministérios e programas. Apesar desse contexto promissor, a gestão pública ainda é conhecida por uma ineficiência grave e frequente, o que acarreta problemas em toda a máquina pública e finalmente na sociedade.

Dados abertos muitas vezes não são classificados de qualquer forma, o que sugere que qualquer técnica que venha a ser utilizada neles deve ser própria para problemas que envolvem aprendizagem não-supervisionada. Abordagens não-supervisionadas podem ser utilizadas para detecção de padrões típicos de comportamento e de distorções no comportamento médio ou esperado dos dados. A aplicação e o estudo delas é importante tanto para fins técnicos quanto sociais.

Neste trabalho será utilizada a técnica de Mapas Auto-Organizáveis (SOM) para segmentar a base de dados de Autorização de Internações Hospitalares, disponibilizada pelo DATASUS. A rede SOM provê uma ótima maneira de reduzir a dimensionalidade dos dados, enquanto mantém as relações topológicas no conjunto de treinamento. Foi realizada uma análise experimental e foi observado que vários registros na base de dados não estão de acordo com a tabela de cobranças do SUS, necessitando maior atenção.

Palavras-chave: Dados Abertos, Sistema Único de Saúde, Mineração de Dados, Self-Organizing Map

Abstract

The amount of open data for the general public grows daily, driving major advances in the data analysis field. Following the trend, different levels of government in Brazil make the data of their departments and ministries available. Despite this promising context, public management still is known for its deep and frequent inefficiency, which causes problems for public administration and society.

Open data are seldom classified, meaning that any technique to be used therein must be suitable for problems involving unsupervised learning. Unsupervised learning approaches may be used for detecting typical behaviour patterns and distortions in the behaviour expected from the data. Deploying and researching such techniques is important for both academic and social purposes.

In this work the *Self-Organizing Map* (SOM) technique will be used to cluster Hospital Admissions Authorization database, made available by DATASUS. The SOM network provides a good way to reduce data dimensionality, whilst maintaining its topological relations. An experimental analysis was performed and it was observed many records in the database are not in agreement with the Unified Health System (SUS) collection table, requiring a more thorough analysis.

Keywords: Open Data, Healthcare System, Data Mining, Self-Organizing Map

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

AIH	Autorização de Internação Hospitalar
ANVISA	Agência Nacional de Vigilância Sanitária
CID	Classificação Internacional de Doenças
CRISP-DM	*Cross Industry Standard Process for Data Mining*
DATASUS	Departamento de Informática do SUS
OGP	Parceria para Governo Aberto
OMS	Organização Mundial de Saúde
PCA	*Principal Component Analysis*
PIB	Produto Interno Bruto
SIPAGEH	Sistema de Indicadores Padronizados para Gestão Hospitalar
SOM	Mapa Auto-Organizável
SUS	Sistema Único de Saúde

Sumário

1

Introdução

Os gastos com o setor de saúde brasileiro contabilizam aproximadamente 10% do Produto Interno Bruto (PIB) nacional, segundo dados da Organização Mundial de Saúde (WHO, 2015). Apesar de um gasto impressionante, o Brasil conquistou a penúltima colocação em um ranking divulgado pela Bloomberg sobre a eficiência de sistemas de saúde de países com mais de 5 milhões de habitantes (Bloomberg, 2014). Países como Argentina e México apresentam gastos per capita e porcentual do PIB menores que o brasileiro e são considerados mais eficientes. Mesmo o ranking não sendo um indicador definitivo da gestão de saúde pública brasileira, sabe-se que ele é sintomático de sua conhecida ineficiência.

O Sistema Único de Saúde (SUS) foi instituído pela Constituição Federal de 1988 com o objetivo de garantir que a população tenha acesso integral, universal e igualitário à saúde. Ainda assim, existem poucas propostas de modelos computacionais para avaliar o desempenho na gestão dos recursos públicos pelo sistema. Parte disso se deve à dificuldade de acesso a dados de setores públicos, mas o cenário vem mudando desde 2011, quando o Brasil tomou parte na criação da Parceria para Governo Aberto (OGP), iniciativa com a pretensão de incentivar práticas governamentais relacionadas à transparência dos governos e ao acesso à informação pública. Como fruto dessa iniciativa, foi lançado o Portal Brasileiro de Dados Abertos (PLANEJAMENTO, 2015), ferramenta para a disponibilização de dados públicos.

Através do Portal são disponibilizados os arquivos de Autorização de Internação Hospitalar (AIH), uma base valiosa que contém grande parte dos dados de internações hospitalares no país. Técnicas de mineração de dados são utilizadas para descobrir padrões em grandes conjuntos de dados, como esta base. Em um mundo com uma quantidade cada vez maior de dados disponíveis, tais técnicas mostram-se uma necessidade importante. Essa necessidade é evidenciada por trabalhos que empregam a mineração de dados na solução de problemas reais, como em (PHUA et al., 2010; BOLTON; HAND, 2002).

(LUBAMBO, 2008) utiliza um processo de mineração de dados para construir uma ferramenta de apoio à decisão a ser utilizada por órgãos públicos. Para tal, foram utilizados dados reais da Secretaria da Fazenda do Estado de Pernambuco, mostrando como o setor público pode se valer de meios que possibilitem uma melhor gestão de seus recursos.

1.1 Objetivos

A base de dados utilizada neste trabalho não possui nenhuma classificação dos dados, sob nenhum aspecto. Para tratar conjuntos de dados nesta situação, são utilizados algoritmos de aprendizagem não-supervisionada. Abordagens não-supervisionadas são próprias para uma análise exploratória de dados e podem ser utilizadas para detecção de *clusters*. Estas técnicas são chamadas de algoritmos de segmentação ou *clustering*. Em (WATTS; WORNER, 2009) os autores lançam mão de técnicas de *clustering* para estimar a invasão de regiões geográficas por diferentes espécies de insetos. Técnicas de segmentação consistem em agrupar elementos de um determinado conjunto de dados em grupos, de modo que objetos mais similares pertençam ao mesmo grupo e os menos similares fiquem em grupos distintos.

Em (KOHONEN, 1990) é descrita a rede neural Mapa Auto-Organizável (SOM), uma técnica de aprendizagem não-supervisionada bastante utilizada na visualização de dados. Esta rede é muito difundida por conseguir representar bem um espaço multidimensional em um espaço de baixa dimensão (muitas vezes apenas duas), chamado de mapa. Um mapa consiste de neurônios que têm associados a eles uma posição no mapa e vetores peso de mesma dimensão dos vetores de entrada. Para formar o mapa, os vetores peso competem entre si para saber qual deles é ativado para cada vetor de entrada, a fim de serem ajustados. A Figura 1.1 mostra como todos os neurônios de um SOM são conectados em paralelo à camada de entrada, o que permite a disputa entre eles.

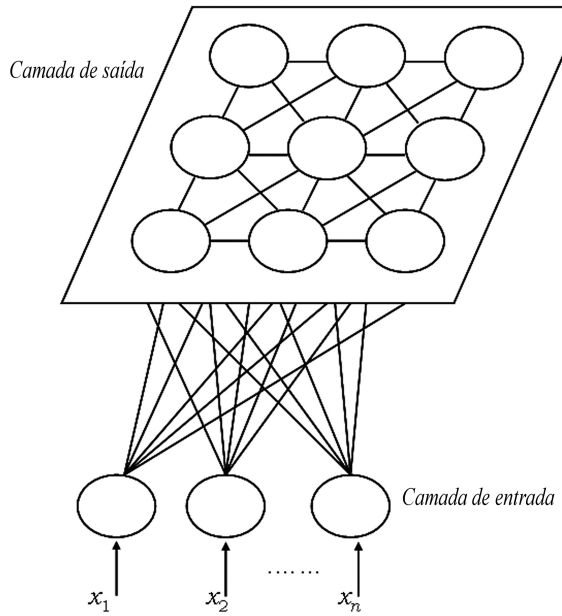


Figura 1.1: Visualização da rede SOM, onde a camada de saída forma um mapa alimentado pelos vetores da camada de entrada.

Para além da visualização de dados, em (KIANG, 2001) o autor mostra como um Mapa Auto-Organizável pode ser estendido para a segmentação daqueles. O output de uma rede SOM não provê de imediato os agrupamentos dos neurônios no mapa, a menos que se decida ter

cada neurônio representando um *cluster*, o que se torna impraticável quando se utiliza mapas extensos. Esta rede neural será utilizada neste trabalho com o objetivo de segmentar os dados do conjunto de AIHs. Para guiar o processo de mineração de dados, será utilizado o modelo de processos conhecido como CRISP-DM. Este modelo descreve abordagens utilizadas para resolver problemas por especialistas da área de mineração de dados.

A utilização dos modelos e técnicas citados busca responder as seguintes perguntas: Quais são as tendências de gastos na gestão pública de saúde? Elas são justificáveis?

1.2 Estrutura do trabalho

Este trabalho é composto por 6 capítulos. No Capítulo 2 é apresentado o modelo de processo CRISP-DM e todas as suas etapas. É discutida a importânca da implantação de um modelo como esse e como ele dialoga com o projeto.

O Capítulo 3 contextualiza a base de dados utilizada e são discutidas todas as etapas que envolvem o projeto antes do treinamento da rede. É mostrado como cada etapa do CRISP-DM anterior à modelagem é aplicada.

Após a base de dados ter sido analisada, processada e ter tido amostras coletadas, o Capítulo 4 detalha toda a fase de modelagem do processo. A rede SOM é explicada de forma detalhada e sua configuração e segmentação são abordadas. A fase de avaliação é elaborada no Capítulo 5, onde os resultados são mostrados e discutidos. Por fim, o Capítulo 6 traz as considerações finais e possíveis trabalhos futuros.

2

CRISP-DM

O *Cross Industry Standard Process for Data Mining* (CRISP-DM) (CHAPMAN et al., 2000) é um modelo de processo de mineração de dados criado por um consórcio de empresas da área, que descreve as melhores práticas para lidar com problemas de mineração de dados. Este modelo foi criado quando o mercado de mineração de dados ainda tomava forma e mostrava a necessidade de ter um processo padrão. Para bem atender a indústria, o consórcio de empresas definiu um processo criado pela comunidade e sem amarras a nenhuma ferramenta. Em um artigo recente do KDNuggets (PIATETSKY, 2014), um dos mais respeitados websites da área de *Data Science*, o autor divulga uma pesquisa feita entre pesquisadores da área que estabelece o CRISP-DM como a metodologia mais utilizada para projetos de mineração de dados.

Na Seção 2.1 é explicado como o processo funciona e em suas subseções são descritas suas diferentes fases.

2.1 O processo

É importante perceber que este modelo não menciona ferramentas ou técnicas específicas, dando importância ao processo analítico em si. O CRISP-DM é um processo iterativo e cíclico, *i.e.*, a sequência de suas fases não é estrita e é muito comum voltar ou adiantar fases. Em (MCCUE, 2014, Capítulo 4) é demonstrada a importância do processo ser iterativo quando se trata de problemas criminais ou fraudulentos. Nesses casos é comum que os crimes e criminosos mudem, o que acarreta mudanças nos padrões detectados. O autor menciona as mudanças frequentes na operação de mercados ilegais de drogas, associadas com a troca de quem vende ou da droga vendida. De maneira similar, os padrões detectados em nossa base de dados podem não ser os mesmos em um futuro próximo por várias razões, como troca de gestores nos hospitais, novas diretrizes do Ministério da Saúde ou algum surto epidêmico.

O modelo de processo do CRISP-DM é dividido em 6 passos: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implantação. A Figura 2.1 mostra como essas fases se comportam no ciclo de vida de um projeto. A fase de implantação não é percorrida neste trabalho, por fugir de seu escopo.

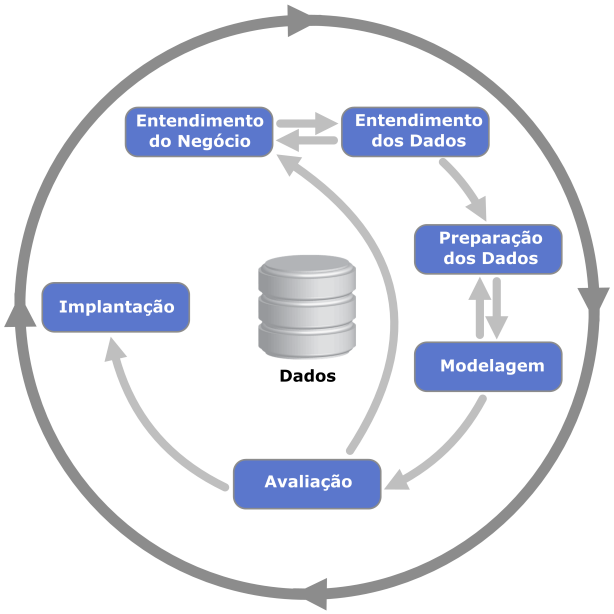


Figura 2.1: Diagrama do CRISP-DM. As setas indicam as dependências mais frequentes entre as fases.

2.1.1 Entendimento do negócio

A primeira fase do processo de mineração de dados consiste em compreender e determinar os objetivos gerais do projeto. Como o intuito é transformar este conhecimento em uma definição de um problema de mineração de dados e traçar um plano preliminar para atacá-lo, nesta etapa devem ser identificadas as metas principais, os recursos disponíveis e suas restrições, além de métricas específicas que auxiliem na avaliação de desempenho da empreitada. Mais especificamente, este estágio é dividido em quatro tarefas:

- Determinar os objetivos do negócio;
- Avaliar a situação;
- Determinar os objetivos da mineração de dados;
- Produzir plano de projeto.

Determinar os objetivos do negócio

Na primeira tarefa do processo deve ser definido o que o cliente deseja. É comum acontecer de o cliente ter interesses conflitantes ou restrições, que podem influenciar no decorrer do projeto. Espera-se no fim desta tarefa que o analista tenha informações sobre a situação do negócio. Além disso, seus objetivos também devem ser definidos aqui. Finalmente, deve-se indicar critérios de sucesso a serem utilizados no fim do processo.

Avaliar a situação
Ao avaliar a situação, deve-se colher mais informações a respeito de qualquer fator que possa influenciar no plano do projeto, como recursos disponíveis e restrições, por exemplo. Uma lista dos recursos disponíveis faz parte dos resultados desta etapa. Nela devem ser incluídos recursos humanos, os dados disponíveis e recursos computacionais, tanto de hardware quanto de software. Além disso, devem ser analisados os requisitos e possíveis restrições do projeto, não apenas em relação aos recursos disponíveis, como também questões práticas, por exemplo, o tempo necessário. Uma avaliação de custos e benefícios também é necessária, para se ter certeza que os primeiros não sobrepõem os últimos.

Determinar os objetivos da mineração de dados
De modo similar à primeira etapa, esta também trata de objetivos do processo. Ao contrário daquela, porém, esta etapa se atém a determinar objetivos técnicos para a mineração de dados. Isto é feito traduzindo os objetivos de negócio determinados anteriormente para perguntas mais diretas que possam ser respondidas por softwares. Assim, devem ser estabelecidos os outputs do projeto de modo que ajudem a atingir os objetivos de negócio.

Produzir plano de projeto
A última tarefa desta fase prevê a produção de um plano para atingir todos os objetivos determinados previamente. São dois os documentos que se esperam nesta etapa: o plano de projeto em si, onde são listados todos os estágios a serem executados no projeto junto com seus requisitos, e uma avaliação das ferramentas e técnicas que serão utilizadas.

2.1.2 Entendimento dos dados

Na segunda fase do CRISP-DM os dados são coletados e o analista inicia sua exploração, ganhando familiaridade com a base. Verifica-se questões como a base estar completa, seus dados serem corretos, os dados incorretos serem frequentes caso existam e como valores em falta são tratados. Esta fase é também dividida em quatro tarefas:

- Coletar os dados iniciais;
- Descrever os dados;
- Explorar os dados;
- Verificar a qualidade dos dados.

Coletar os dados iniciais
Inicialmente os dados mencionados no plano de projeto devem ser coletados. Junto com os dados adquiridos devem ser reportados a localização, problemas enfrentados e as soluções realizadas. Os problemas com os dados podem ser vários, como dificuldade para adquiri-los ou

para visualizá-los, por exemplo.

Descrever os dados
A descrição dos dados é feita de maneira rasa, *i.e.*, suas propriedades intrínsecas (*e.g.* distribuições das variáveis) não são avaliadas aqui. Para esta tarefa, basta verificar o formato, quantidade e tipos dos dados e avaliar se são suficientes para alcançar os objetivos.

Explorar os dados
Durante a exploração dos dados, suas propriedades intrínsecas são trazidas à mesa. São analisadas suas variáveis, no sentido de entender como elas se comportam e suas relações entre si (uma a uma ou em pequenos grupos). Estas análises podem contribuir tanto para a descrição dos dados quanto para a verificação de sua qualidade.

Verificar a qualidade dos dados
Nesta etapa a qualidade dos dados é verificada focando em dois pontos essenciais: se os dados contêm erros e a frequência deles, e se os dados contêm valores não preenchidos e as frequências destes. Após a identificação desses problemas principais (e outros que façam sentido para a aplicação), devem ser propostas as soluções para atacá-los.

2.1.3 Preparação dos dados

Após a caracterização dos dados na fase anterior, eles devem ser tratados para serem utilizados na fase de modelagem. A preparação dos dados inclui uma limpeza da base, removendo variáveis que não agregam valor ou observações esdrúxulas. Algumas vezes os valores de certas variáveis podem ser recalculados baseado nas outras, o que previne a deleção de observações. Outras transformações possíveis são descritas e analisadas em cada uma das etapas que compõem esta fase:

- Selecionar os dados;
- Limpar os dados;
- Construir dados;
- Integrar dados;
- Formatar os dados.

Selecionar os dados
Como primeiro passo na preparação, os dados devem ser filtrados com base em critérios como relevância dos dados, qualidade e restrições técnicas como os tipos dos dados (*e.g.* linguagem natural). Além das decisões de quais dados farão parte ou não do conjunto final, as razões para

elas também devem ser explicitadas.

Limpar os dados
Nesta etapa os dados devem ser polidos levando em consideração os problemas apontados na tarefa de verificação da qualidade dos dados. Práticas comuns na limpeza dos dos dados envolvem a identificação de subconjuntos dos dados com melhor qualidade e a inserção de valores padrão para o caso de não preenchimento.

Construir dados
Esta tarefa ocorre principalmente no âmbito dos atributos, como a produção de novas variáveis a partir de outras existentes ou até mesmo a transformação de valores de atributos existentes. Além de manipular atributos individualmente ou pequenos grupos, o analista também pode sintetizar nos registros completamente, se servir ao seu propósito.

Integrar dados
A integração dos dados ocorre em projetos que lidam com várias tabelas diferentes entre si, em relação a seus atributos. Para integrar os dados, os atributos das variadas tabelas são combinados entre si, simplesmente concatenando-os ou criando novas variáveis a partir deles.

Formatar os dados
Ao contrário das outras etapas, esta não visa modificar o que os dados significam ou representam, mas apenas modificar sua sintaxe para facilitar a fase de modelagem. Técnicas comuns nesta etapa são normalização de variáveis, discretização, generalização e binarização.

2.1.4 Modelagem

Na fase de modelagem do projeto, são escolhidos os algoritmos específicos que serão utilizados. Esta escolha deve ser feita baseada na natureza do problema e do que se deseja extrair do processo. Abordagens não supervisionadas, por exemplo, são utilizadas quando deseja-se descobrir padrões ou relacionamentos em dados que não foram classificados previamente. Para o fim de escolher um modelo propício para o problema, esta fase é divida em seis etapas:

- Selecionar a técnica de modelagem;
- Confeccionar testes;
- Construir modelo;
- Avaliar modelo.

Selecionar a técnica de modelagem
A primeira tarefa da fase de modelagem é definir qual a técnica de modelagem que será utilizada no processo. Isto deve ser feito agora de uma maneira mais detalhada do que foi feito anteriomente no processo. Nesta etapa determina-se especificamente qual algoritmo será utilizado, documentando suas possíveis premissas, restrições e benefícios.

Confeccionar testes
Antes de um modelo ser construído, o analista deve gerar de antemão algum procedimento com o qual ele possa avaliar a qualidade do modelo. O teste é confeccionado baseado no modelo escolhido, o que significa que um teste feito para uma abordagem não-supervisionada não deveria ser utilizado em um problema de classificação, por exemplo. É uma prática comum dividir o conjunto de treinamento em dois: dados de treinamento e dados de teste. O modelo é construído utilizando os dados de treinamento e sua qualidade é avaliada com o os dados de teste.

Construir modelo
Na etapa onde finalmente são criados um ou mais modelos dos dados, a principal tarefa a ser feita é ajuste dos parâmetros. De uma maneira geral, as técnicas existentes possuem uma grande quantidade de parâmetros que podem ser configuráveis, e assim o analista deve informar quais foram os valores escolhidos para os parâmetros e sua justificativas.

Avaliar modelo
Diferente da próxima fase, a de avaliação, esta etapa preocupa-se apenas em avaliar o modelo produzido em si. Aqui, o modelo deve ser avaliado com base no que foi definido na confecção dos testes e os objetivos da mineração de dados determinados anteriormente. Nesta etapa deve-se aproveitar para listar as qualidades dos modelos gerados e ordená-los com base nos resultados dos testes. Com base nessas informações, os modelos têm seus parâmetros ajustados para serem reconstruídos.

2.1.5 Avaliação

Nesta etapa, ao contrário do que acontece durante a modelagem, onde apenas o modelo é avaliado, o processo como um todo é revisado. Aqui é avaliado em que nível o modelo atinge os objetivos descritos durante o entendimento do negócio, e procura-se alguma explicação de negócio que indique os problemas do modelo, caso seja deficiente. Para tal, esta fase é dividida em três tarefas:

- Avaliar resultados;
- Rever o processo;
- Estabelecer próximos passos.

Avaliar resultados
Nesta primeira etapa, os resultados obtidos são confrontados com os objetivos de negócio definidos anteriomente. Caso os objetivos não tenham sido atingidos, é dito que o modelo é deficiente, e faz parte desta tarefa investigar o porquê.

Rever o processo
Aqui uma revisão do processo é feita com o objetivo de identificar se possíveis fatores impactantes não foram levados em consideração anteriormente. Decisões feitas em fases anteriores são repensadas e, possivelmente, corrigidas. As conclusões desta tarefa influenciam fortemente na próxima etapa.

Estabelecer próximos passos
Com as conclusões das etapas anteriores em mãos, aqui decide-se qual rumo o projeto deve tomar. Antes de chegar à fase de implantação, talvez seja necessário voltar até mesmo para a fase de entendimento do negócio, caso tenha sido avaliado que o modelo não atingiu os objetivos por algum erro fundamental relativo ao problema.

2.1.6 Implantação

A última fase do CRISP-DM é a implantação do projeto desenvolvido. Esta etapa inclui a disseminação da informação, que pode ser feita de várias maneiras a depender do contexto. De maneira geral, esta fase consiste em implantar o sistema desenvolvido no lado do cliente, para que ele possa aplicá-lo diretamente em outros dados. Em outros casos, um relatório do processo pode ser suficiente para esta fase. De qualquer modo, esta fase é dividia em quatro etapas:

- Plano de implantação;
- Plano de monitoramento e manutenção;
- Relatório final;
- Rever projeto.

Plano de implantação
Este plano serve para guiar a estratégia de implantação do processo, com base nos resultados da avaliação feita na fase anterior. Devem ser incluídos neste plano, os passos necessários para implantação e explicações de como executá-los.

Plano de monitoramento e manutenção

Este plano é importante se for decidido que o modelo desenvolvido deve ser executado no lado do cliente. A estratégia montada no plano deve procurar evitar que os resultados da mineração de dados seja utilizada de forma errada por algum período de tempo.

Relatório final

Este relatório inclui todos os relatórios que foram produzidos anteriormente, organizando-os adequadamente. Este relatório é apenas uma síntese de todo o processo, visto que muitos detalhes foram lidados anteriormente.

Rever projeto

Esta etapa serve principalmente para o analista, pois ela consiste em uma revisão do que deu certo e errado durante o projeto. Com isso, os membros da equipe sabem o que precisa ser melhorado em um projeto futuro e o que realmente não funciona.

2.2 Considerações finais sobre o processo

É importante perceber, que devido à natureza deste trabalho, o processo sugerido pelo CRISP-DM não é seguido inteiramente à risca. Isso não se torna um empecilho ou entrave à sua utilização, devido ao fato de várias de suas etapas serem independentes de outras, o que permite sua modularização.

A primeira fase (Entendimento do Negócio) foi descrita na introdução deste trabalho. Prolongar esta fase fugiria ao escopo deste trabalho, por isso ela não será mais aprofundada.

O Capítulo 3 engloba as fases relativas aos dados no CRISP-DM. Destas duas fases, apenas a etapa de integração dos dados, descrita na Subseção 2.1.3, não será realizada, uma vez que a base utilizada é única.

A fase de modelagem é relatada quando se discute a rede SOM no Capítulo 4 e a fase de avaliação do processo tem seus passos seguidos no Capítulo 5.

A fase de implantação é a única que não será implementada neste trabalho, pois foge ao seu propósito. De certa forma, podemos dizer que a etapa do relatório final é constituída por este trabalho em si e a revisão do projeto é feita na conclusão.

3

Análise e tratamento dos dados

A base de dados utilizada neste trabalho foi divulgada pelo Departamento de Informática do SUS (DATASUS) através do Portal Brasileiro de Dados Abertos (PLANEJAMENTO, 2015). O *dataset* é populado por informações sobre as internações hospitalares do SUS, coletadas através dos formulários de AIH que são enviados aos gestores responsáveis - municipal (se em gestão plena) ou estadual.

No repositório constam arquivos mensais de todas as unidades da federação a partir de janeiro de 2008 até maio de 2014, totalizando 41 arquivos por estado. Devido à grande quantidade de dados, a ponto de tornar intratável a opção de lidar com a base completa, foi escolhido lidar apenas com um subconjunto da base. Serão utilizadas as informações pertencentes apenas ao estado de Pernambuco, pois além das razões geográficas, o estado de Pernambuco apresenta uma quantidade representativa de amostras com cerca de 45 mil mensais. A obtenção destes arquivos é a primeira etapa da fase de entendimento dos dados do CRISP-DM, mencionada na Subseção 2.1.2.

Neste capítulo será apresentado o conjunto de dados na Seção 3.1. O pré-tratamento realizado na base é discutido na Seção 3.2 e a Seção 3.3 aborda como os dados para treinamento foram gerados.

3.1 Conjunto de dados

Como discutido na Subseção 2.1.2, os dados precisam ser entendidos antes de se colocar a mão na massa. Independentemente de as tabelas serem disponibilizadas pelo mesmo órgão, elas podem conter diferenças entre si, devido a diferentes diretrizes que possam ser estabelecidas. A partir de janeiro de 2013 duas variáveis referentes a utilização de UCI foram adicionadas à base de dados, por exemplo. Além de tabelas com diferentes quantidades de colunas, existem múltiplos casos de variáveis com valores em desacordo com o dicionário de dados ou até mesmo que não são explicadas por ele. Tais fatores embaraçam a correta manipulação dos dados caso estes não sejam bem entendidos.

A lista completa de variáveis com uma breve descrição de cada uma pode ser encontrada

no Apêndice A.

Apesar de cada observação conter um número extenso de variáveis (95), muitas delas devem ser desconsideradas ou transformadas. Com o intuito de saber como cada variável pode contribuir para a distinção dos elementos da base, todas elas foram analisadas uma a uma. Tal análise individual das variáveis não apenas serviu para saber quais seriam descartadas, mas também apontou características que indicam tratamentos a serem feitos em um passo futuro.

3.1.1 Variáveis zeradas

Inicialmente, foi detectado que 19 variáveis não apresentam valor algum, *i.e.*, suas informações nunca são preenchidas na base de dados. Além destas, outras 15 variáveis também são consideradas ruins, pois todos seus valores são sempre iguais (ou muito próximo de 100%). Sabe-se que variáveis com esse tipo de distribuição não contribui para qualquer entendimento sobre a base, simplesmente por não influenciarem em nada na distinção entre os dados. Estas variáveis são listadas na Tabela 3.1 junto com a razão de cada uma para ser ignorada.

3.1.2 Variáveis redundantes

A presença de variáveis redundantes também é considerado um dos problemas desta base de dados. Existem 8 variáveis redundantes entre si, formando 4 pares:

- *NATUREZA* e *NAT_JUR*;
- *VAL_TOT* e *US_TOT*;
- *VALOR_UTI* e *VALOR_UCI*;
- *MARCA_UTI* e *MARCA_UCI*.

As variáveis *NATUREZA* e *NAT_JUR* representam a natureza jurídica do estabelecimento. É dito que a primeira deveria possuir valores apenas até maio de 2012, sendo substituída completamente pela última de lá em diante. Porém, pôde-se constatar que a variável *NATUREZA* continuou a ser preenchida e com valores correspondentes à *NAT_JUR*. A diferença entre as duas variáveis está na classificação utilizada por cada uma delas, a primeira utiliza a classificação de Regime e Natureza do SUS[1] e a segunda utiliza a classificação estabelecida pela Comissão Nacional de Classificação (CONCLA)[2], sendo esta muito mais detalhista que a anterior. Entretanto, é importante mencionar que a variável *NAT_JUR* possui uma taxa de *missing data* de 42,8%, que contrasta bastante com uma taxa de apenas 0,5% da outra variável.

As variáveis *VAL_TOT* e *US_TOT* representam o valor total da internação em real e em dólar, respectivamente. A análise de redundância destas duas variáveis foi feita com uma

[1] http://tabnet.datasus.gov.br/cgi/sih/cxdescr.htm - Acessado em julho de 2015

[2] http://concla.ibge.gov.br/ - Acessado em julho de 2015

Tabela 3.1: Variáveis que não apresentam valor de discriminação.

Índice	Nome do campo	Razão
12	UTI_MES_IN	Campo zerado
13	UTI_MES_AN	Campo zerado
14	UTI_MES_AL	Campo zerado
17	UTI_INT_IN	Campo zerado
18	UTI_INT_AN	Campo zerado
19	UTI_INT_AL	Campo zerado
27	VAL_SADT	Campo zerado
28	VAL_RN	Campo zerado
29	VAL_ACOMP	Campo zerado
30	VAL_ORTP	Campo zerado
31	VAL_SANGUE	Campo zerado
32	VAL_SADTSR	Campo zerado
33	VAL_TRANSP	Campo zerado
34	VAL_OBSANG	Campo zerado
35	VAL_PED1AC	Campo zerado
47	RUBRICA	Campo zerado
54	NACIONAL	Um valor em aprox. 100% da base
55	NUM_PROC	Campo zerado
57	TOT_PT_SP	Campo zerado
58	CPF_AUT	Campo zerado
60	NUM_FILHOS	Um valor em aprox. 100% da base
65	GESTRISCO	Um valor em aprox. 100% da base
68	CBOR	Um valor em aprox. 100% da base
69	CNAER	Um valor em aprox. 100% da base
70	VINCPREV	Um valor em aprox. 100% da base
74	GESTOR_DT	Um valor em aprox. 100% da base
77	INFEHOSP	Valores sempre iguais
85	ETNIA	Um valor em aprox. 100% da base
88	AUD_JUST	Um valor em aprox. 100% da base
89	SIS_JUST	Um valor em aprox. 100% da base
90	VAL_SH_FED	Valores sempre iguais
91	VAL_SP_FED	Valores sempre iguais
92	VAL_SH_GES	Valores sempre iguais
93	VAL_SP_GES	Valores sempre iguais

granularidade mensal. Foi verificado para cada mês os valores únicos de cotação de dólar, levando em consideração uma pequena margem de tolerância. Constatou-se que apesar da base de dados apresentarem cotações diferentes para meses diferentes, a cotação para cada mês é única.

Os outros dois pares de variáveis redundantes são referentes à utilização de UTI/UCI pelo paciente ou não, especificamente sobre o valor gasto e o tipo utilizado. Elas são ditas redundantes, pois não existem diferenças entre uma UTI e uma UCI de mesmo tipo, são apenas termos intercambiáveis. Foi verificado que apesar de as variáveis *MARCA_UTI* e *MARCA_UCI* apresentarem valores diferentes, eles representam os mesmos tipos de Unidades de Tratamento/Cuidados Intensivo(s). A Tabela 3.2 mostra quais são os valores diferentes que as variáveis apresentam. Esta tabela não representa todos os valores previstos para cada variável no dicionário, ela mostra apenas aqueles valores que estão presentes na base de dados aqui utilizada.

Tabela 3.2: Valores diferentes nas variáveis *MARCA_UTI* e *MARCA_UCI*.

Tipo	***MARCA_UTI***	***MARCA_UCI***
UTI adulto nível II	75	01
UTI neonatal nível III	82	03

3.1.3 Variáveis com valores errados

Existem 4 variáveis na base de dados que podem ser recalculadas de acordo com as informações presentes nela, elas são: *COD_IDADE*, *IDADE*, *QT_DIARIAS* e *DIAS_PERM*. As duas primeiras são relativas à idade do paciente e as duas últimas ao tempo de internação do paciente. É possível constatar que elas apresentam valores errados em grande parte da base de dados, pois esta contém as variáveis *NASC*, *DT_INTER* e *DT_SAIDA*, que são respectivamente as datas de nascimento, internação e alta do paciente.

Utilizando as datas de nascimento e internação do paciente foi recalculada a idade dos pacientes. Constatou-se que apenas no mês de janeiro de 2011, 88,7% dos registros não contêm a idade correta do paciente. A diferença dos erros quase nunca é pequena, com casos onde o paciente, segundo a tabela, tem 54 anos, mas na verdade ele tem apenas 5. A Figura 3.1 mostra como a distribuição das idades muda quando elas são corrigidas.

As variáveis referentes ao tempo de estadia do paciente foram recalculadas utilizando as variáveis *DT_INTER* e *DT_SAIDA*. De maneira semelhante às variáveis de idade, essas também possuem graves irregularidades, talvez até mais problemáticas, dado o que elas representam. Em alguns registros, a diferença entre o tempo de permanência informado e o tempo calculado de acordo com as datas de internação e alta chega ser superior a 2000 dias. Entretanto, a quantidade de casos onde existe qualquer diferença de dias é muito menor do que os registrados com as variáveis de idade, representando apenas 3,1% do total da base. A Figura 3.2 mostra a

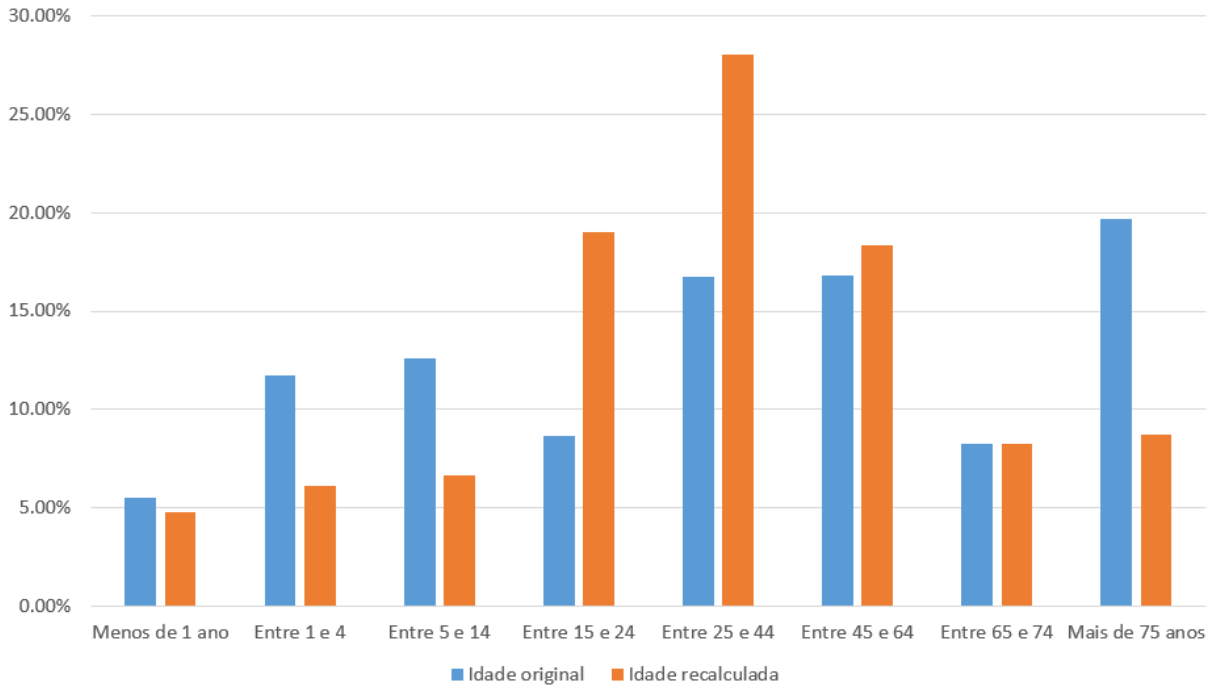


Figura 3.1: Distribuição das faixas etárias presentes no arquivo de janeiro de 2011 antes e após correção.

distribuição das diferenças absolutas de dias por faixas de valores, levando em consideração a variável *DIAS_PERM*. Esta figura é restrita aos registros onde a diferença é maior que zero.

Não é possível identificar as causas dos erros, se é má-fé, leniência ou simples erro de digitação de quem deve preencher os campos.

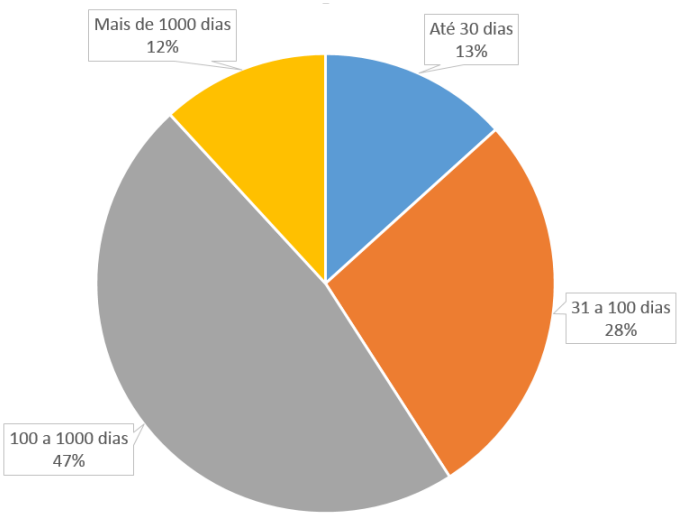


Figura 3.2: Distribuição das diferenças absolutas de dias calculadas com os valores da variável *DIAS_PERM*.

3.1.4 Valores em desacordo com o dicionário ou outras variáveis

A verificação dos valores individuais das variáveis também é um importante passo na análise dos dados. Foram identificadas duas situações onde algumas variáveis apresentam clara incongruência no conjunto.

A primeira situação diz respeito à variável *CONTRACEP1*, que indica qual método contraceptivo foi utilizado pelo paciente. Neste caso, existem alguns poucos registros onde o contraceptivo utilizado entra em desacordo com o sexo do paciente. Na totalidade das ocorrências, o paciente é do sexo masculino e os contraceptivos são de uso exclusivo feminino.

A segunda situação aponta inconsistências em alguns registros com a variável *ESPEC*, que indica a especialidade do leito do paciente. Neste caso, a variável continha valores que não eram previstos pelo dicionário de dados. Estas ocorrências representam menos de 0,5% da base.

3.2 Tratamento dos dados

O pré-processamento dos dados é um passo bastante relevante no processo de mineração de dados, como abordado na Subseção 2.1.3. O tratamento dos dados leva em consideração o entendimento extraído deles e debatido na seção anterior.

3.2.1 Exclusão de variáveis e registros

Inicialmente, todas as 34 variáveis presentes na Tabela 3.1 foram sumariamente excluídas da base. Além dessas, outras 10 variáveis foram deletadas da base também neste primeiro momento, que são:

- *N_AIH*;
- *SEQUENCIA*;
- *REMESSA*;
- *US_TOT*;
- *NAT_JUR*;
- *DIAG_SECUN*;
- *CONTRACEP2*;
- *MUNIC_RES*;
- *GESTOR_COD*;
- *GESTOR_CPF*.

As três primeiras variáveis representam informações burocráticas da internação. Elas são desconsideradas logo pois possuem uma grande quantidade de valores diferentes, o que dificultaria as suas representações. O campo *N_AIH*, que indica o número de uma AIH, possui um valor diferente para cada registro, por servir como um identificador da internação. Os outros dois campos, *SEQUENCIA* e *REMESSA*, apesar de não possuírem uma distribuição tão variada quanto a anterior, também possuem uma quantidade de valores categóricos que torna intratável a manipulação desses.

Decidiu-se por desconsiderar o campo *US_TOT* por ele ser uma variável redundante, como apresentado na Subseção 3.1.2. Este campo não traz ganho de informação pois seu valor é sempre diretamente relacionado ao da variável *VALOR_TOT*. De maneira semelhante, a variável *NAT_JUR* também é desconsiderada pelo fato de sua informação já ser transmitida pelo campo *NATUREZA*.

O campo *DIAG_SECUN* foi excluído devido à sua grande quantidade de *missing data* e também à alta similaridade com o campo *DIAG_PRINC*, quando comparando os primeiros caracteres dos dois. A mesma situação ocorre com as variáveis *CONTRACEP1* e *CONTRACEP2*, resultando na exclusão da última. *MUNIC_RES*, que representa o munícipio do paciente, foi preterida em relação ao CEP.

As últimas duas variáveis apresentam informações do gestor responsável por cada AIH. A variável *GESTOR_COD* especifica o motivo da autorização da internação pelo gestor e *GESTOR_CPF* informa seu CPF. O autor deste trabalho deliberou por não lidar com variáveis de cunho pessoal durante este processo.

Além das 10 variáveis mostradas anteriormente, outras 5 foram excluídas da base por motivos similares. Estas variáveis são: *CGC_HOSP*, *CID_NOTIF*, *SEQ_AIH5*, *CNPJ_MANT* e *CID_ASSO*.

Em relação à deleção de registros, vários foram removidos baseado nas incongruências relatadas na Subseção 3.1.4. Foram removidos aproxidamente 150 registros que apontavam inconsistências de acordo com a variável *SEXO*, por exemplo. Neste caso, foram removidos os registros que apontavam o paciente de sexo masculino utilizando métodos contraceptivos não condizentes com o gênero. Além desses, cerca 8500 linhas das tabelas foram removidas devido aos valores não reconhecidos da variável *ESPEC*.

Outros 130 mil registros foram excluídos da base de dados, por apresentarem graves inconsistências em relação ao tempo de duração da internação, como pode ser visto na Figura 3.2. Neste caso, foram excluídos os registros que apresentavam uma diferença absoluta com a quantidade de diárias ou os dias de permanência maior do que 1.

3.2.2 Transformações de variáveis

Além da exclusão de variáveis e registros da subseção anterior, alguns campos foram transformados e deixaram de existir. Os campos relativos à UTI e UCI foram mesclados entre si,

resultando no descarte das variáveis de UCI. A coluna *VAL_UCI* teve seus valores simplesmente somados aos de *VAL_UTI*. Em relação às variáveis que representam o tipo da UTI/UCI, O campo *MARCA_UCI* teve seus valores transformados para os códigos de *MARCA_UTI*. Foi adicionado um novo campo, *MARCA_DIFF*, que serve para indicar quando os campos anteriores não concordam um com o outro.

Outros 2 pares de variáveis sofreram modificações semelhantes à do par *MARCA_UTI* e *MARCA_UCI*.

As variáveis *UF_ZI* e *MUNIC_MOV* indicam o município gestor e o município do estabelecimento, respectivamente. Como é bastante comum que o município gestor de um estabelecimento seja o próprio município onde ele se situa, foi excluída a primeira variável em favor da segunda. Entretanto, foi adicionado outro campo que indica quando essas duas variáveis são iguais ou diferentes.

O par *PROC_SOLIC* e *PROC_REA* foi modificado como o anterior, prevalecendo o segundo sobre o primeiro. A razão disso é que o segundo representa qual foi o procedimento efetivamente realizado na internação, enquanto o primeiro diz qual foi o procedimento solicitado. Além disso, o campo *PROC_REA* teve seus valores reduzidos para apenas seus 4 primeiros dígitos originais, pois estes são os mais significativos de um total de 9[3]. Assim, a diversidade deste campo caiu de 1649 valores distintos para apenas 30.

Outras variáveis também tiveram suas quantidades de dígitos reduzidas, entre elas está o *CEP*. Apenas seus três primeiros dígitos são utilizados, o que representa um setor geográfico[4]. Além desta, as variáveis que representam algum código da Classificação Internacional de Doenças (CID) têm apenas o dígito mais à esquerda utilizado. Assim como nas variáveis de procedimento, o dígito mais à esquerda em códigos CID é o mais significativo, representando um capítulo ao qual a enfermidade pertence[5]. As variáveis que representam códigos CID são: *DIAG_PRINC* e *CID_MORTE*.

3.3 Geração dos subconjuntos de treinamento e teste

O subconjunto de treinamento compreende apenas dados até maio de 2013, enquanto o de testes engloba os dados a partir de junho de 2013 até maio de 2014. É importante observar que o conjunto de testes é apenas utilizado durante a avaliação da modelagem. Será verificado se a modelagem do conjunto de treinamento produziu uma representação consistente através da observação do comportamento deste modelo no conjunto de testes.

A geração destes subconjuntos ocorreu de forma aleatória e a quantidade de dados em cada um deles corresponde a 5% dos dados disponíveis. Para evitar perdas de características dos dados devido a aleatorizações ruins, foram escolhidas algumas variáveis julgadas importantes e

[3]http://sigtap.datasus.gov.br/ - Acessado em julho de 2015
[4]http://www.correios.com.br/ - Acessado em julho de 2015
[5]http://www.cid10.com.br/ - Acessado em julho de 2015

testes de bondade de ajusta foram aplicadas nelas. As variáveis escolhidas foram: *VAL_TOT*, *QT_DIARIAS*, *MORTE*, *CNES*, *RACA_COR* e *ESPEC*.

Os testes de hipótese para as variáveis categóricas foram feitas com o teste Chi-quadrado proposto por Pearson (PEARSON, 1900). Para a única variável contínua das seis selecionadas, *VAL_TOT*, foi utilizado o teste de Kolmogorov-Smirnov (MASSEY, 1951). Todos os testes foram realizados com um nível de significância de 5%. Caso o teste de qualquer uma das seis variáveis rejeitasse a hipótese nula, a amostra seria descartado e outra aleatorização seria realizada.

Após da amostragem dos conjuntos de treinamento e teste, estes tiveram seus dados reformatados para facilitar a utilização deles pela rede neural. Todas as variáveis contínuas, unicamente aquelas referentes a valores, foram normalizadas entre 0 e 1, normalização conhecida como MIN-MAX. A normalização não foi feita por variável, mas sim baseada nos valores de todas elas de modo que elas não perdessem as relações entre si.

Os campos relativos a número de dias ou idade do paciente foram divididos em faixas de valores. As variáveis *UTI_MES_TO* e *UTI_INT_TO* tiveram seus valores divididos em apenas 3 faixas, devido às suas baixas quantidades de valores distintos. Os números de dias não referentes à UTI foram divididos em 7 faixas diferentes, sendo que as 4 primeiras contêm apenas 1 dia (0, 1, 2 ou 3 dias) e as outras englobando valores até uma semana, até um mês ou mais de um mês. Para dividir as idades dos pacientes em faixas etárias, lançou-se mão de diretivas da OMS (WORLD HEALTH ASSEMBLY, 1948).

Após a divisão de algumas variáveis em faixas de valores, todas as variáveis categóricas foram binarizadas. A binarização de um variável categórica ocorre de maneira densa ou esparsa. A primeira ocorre quando uma variável possui uma baixa quantidade de valores distintos, tornando possível que em N bits cada um represente um desses valores distintos, *i.e.*, apenas um bit é setado por vez. A binarização densa é escolhida quando a quantidade de valores distintos é grande, o que faria com que uma representação esparsa gerasse um bitset muito extenso. Neste trabalho, as variáveis que continham até 5 valores distintos foram representas de forma esparsa e as outras de forma densa. Os conjuntos gerados nessa fase transformaram as observações em vetores de 130 posições, com valores entre 0 e 1.

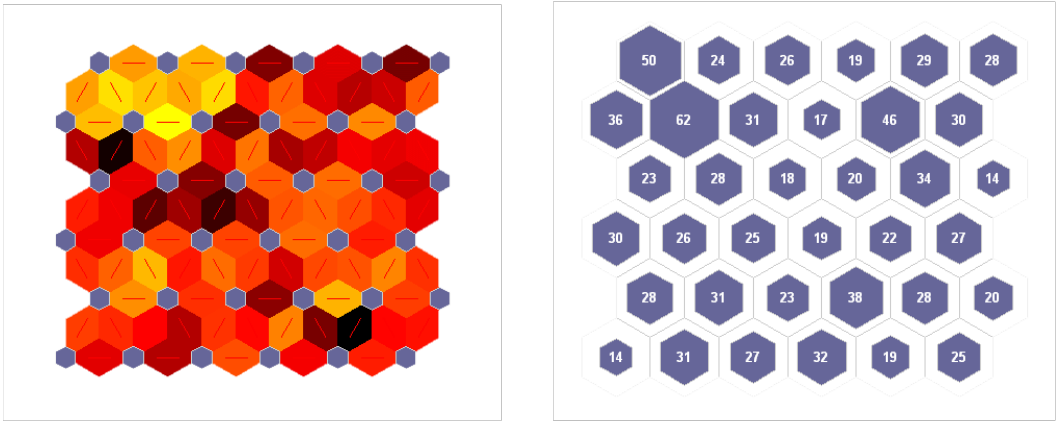
4

Treinamento do Mapa Auto-Organizável

O Mapa Auto-Organizável (*Self-Organizing Map* em inglês) é inspirado na habilidade do cérebro de ativar diferentes regiões a depender do estímulo recebido (KOHONEN, 1990). Kohonen afirma que essa habilidade indica uma organização espacial na representação interna das informações do cérebro. Assim, a rede SOM foi concebida como um mapa que deve apresentar essa mesma característica.

De modo a simular o comportamento de um mapa cerebral, o treinamento de uma rede SOM agrupa os neurônios da rede em vizinhanças similares entre si. Assim, o mapa aprende não apenas a distribuição, mas também a topologia dos dados usados para treiná-lo.

Outro grande atributo do SOM é a facilidade de visualizar os resultados do treinamento. A grade de neurônios pode ser facilmente utilizada para visualizar as distintas regiões no mapa, como pode ser visto na Figura 4.1.



(a) U-Matrix (b) Mapa de hits

Figura 4.1: U-Matrix e mapa de hits de uma mesma rede SOM. (a) Nesta U-Matrix os pontos azuis representam os neurons, linhas vermelhas conectam vizinhos e as cores claras representam vizinhos mais próximos. (b) O mapa de hits mostra quantos vetores de entrada estão associados a cada neuron do grid.

Essas características fazem da rede SOM uma das redes neurais mais utilizadas na

análise de *clusters*, para os mais variados problemas. Em (BROCKETT; XIA; DERRIG, 1998) os autores utilizam a rede SOM junto com uma rede de *backpropagation* para analisar fraudes em reivindicações de compensações por acidentes de carro. (YANG; OUYANG; SHI, 2012) propõem um método para segmentar as regiões de um mapa e testam a abordagem em dados de terremotos e imagens.

Este capítulo trata da fase de modelagem do CRISP-DM. A etapa de avaliação do modelo não é executada aqui. Esta etapa será relatada junto com a avaliação do processo no próximo capítulo.

A próxima seção descreve o algoritmo de maneira mais detalhada. A Seção 4.2 descreve os parâmetros para o treinamento da rede escolhidos neste trabalho e a Seção 4.3 mostra como o mapa é segmentado para a geração dos *clusters*.

4.1 O Mapa Auto-Organizável

Um mapa auto-organizável é composto por um conjunto de neurônios (ou nós) distribuídos em uma única camada N-dimensional, sendo sua representação mais usual bidimensional. Os neurônios do mapa podem apresentar arranjos topológicos variados, sendo os mais comuns hexagonal ou retangular (Figura 4.2). Cada neurônio do mapa é associado a um vetor peso, de mesma dimensão que os vetores da camada de entrada da rede.

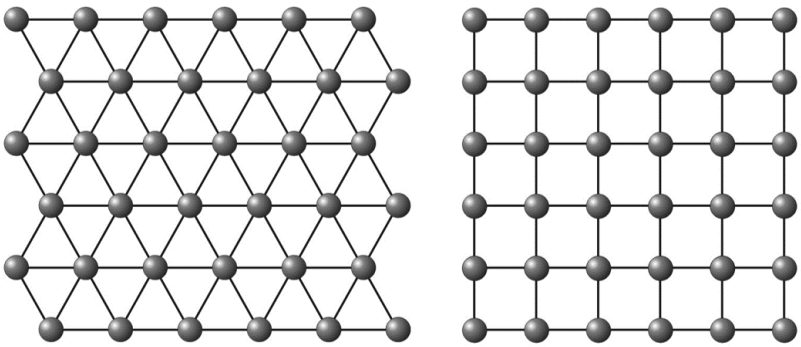


Figura 4.2: Grades hexagonal e retangular para mapas bidimensionais.

Para um determinado vetor de entrada $\mathbf{p}$, o neurônio vencedor $\mathbf{j}$ é aquele cujo vetor peso $\mathbf{w_j}$ associado possui a menor distância $\mathbf{d_j}$ para aquele vetor $\mathbf{p}$. Nas redes SOM, não apenas o neurônio vencedor é atualizado de acordo com $\mathbf{p}$, mas também todos os neurônios que estão em sua vizinhança $N_{\mathbf{j}}(t)$, como mostra a Equação 4.1. Nesta equação, $\mathbf{w_i}$ representa o vetor peso que é atualizado no instante t+1, $\eta(t)$ é a taxa de aprendizagem no instante t e $\mathbf{p}$ é o vetor de entrada. A vizinhança $N_{\mathbf{j}}(t)$ contém os índices de todos os nós $\mathbf{i}$ cuja distância ao nó $\mathbf{j}$ é menor que d no instante t e é mostrada na Equação 4.2.

$$\mathrm{w_i}(t+1) = \mathrm{w_i}(t) + \eta(t)[\mathrm{p} - \mathrm{w_i}(t)], \mathrm{i} \in N_{\mathrm{j}}(t) \tag{4.1}$$

$$N_{\mathrm{j}}(t) = \{\, \mathrm{i} \mid d_{\mathrm{ij}} \leq d(t) \,\} \tag{4.2}$$

Assim, após várias atualizações na rede, os vizinhos de um neurônio tendem a ser similares a ele. É importante saber que neste trabalho é utilizada a aprendizagem em lotes, ou seja, os neurônios do mapa não são atualizados a cada novo vetor **p** que é apresentado a ele. Todo o conjunto de dados é apresentado ao mapa antes dos pesos serem atualizados, de modo que para cada um dos vetores de entrada seja lembrado quem é o neurônio vencedor. Assim, cada vetor peso é atualizado levando em consideração a média de todos os vetores de entrada para os quais seu neurônio é o vencedor.

Pode-se perceber nas equações anteriores, tanto o tamanho da vizinhança de influência de cada neurônio quanto a taxa de aprendizagem variam a cada instante. O modo como essas duas variáveis se modificam durante o treinamento do mapa define duas fases neste processo, explicadas nas próximas subseções.

4.1.1 Fase de ordenação

A primeira fase do treinamento de um mapa auto-organizável é a fase de ordenação topológica dos pesos. Esta fase ordena o mapa de forma grosseira, com o objetivo de definir suas vizinhanças. A duração da fase de ordenação é geralmente muito menor que a da fase de convergência e é de T passos (ou iterações). A Equação 4.2, que define quem são os vizinhos de um neurônio **j** em um instante t depende da função $d(t)$. Esta é a função de de tamanho da vizinhança e é definida por:

$$d(t) = 1 + (N_0 - 1)(1 - \frac{t}{T}) \tag{4.3}$$

Esta equação mostra que o tamanho da vizinhança cai linearmente de um valor inicial N_0 até chegar em 1, uma vez que $t \leq T$. De maneira similar, a taxa de aprendizagem $\eta(t)$ também cai linearmente de um valor inicial η_0 até um final η_T, de acordo com a Equação 4.4.

$$\eta(t) = \eta_T + (\eta_0 - \eta_T)(1 - \frac{t}{T}) \tag{4.4}$$

Os valores iniciais N_0 e η_0 costumam ser bastante altos para assim os vetores peso se moverem a passos largos em direção aos vetores de entrada.

4.1.2 Fase de convergência

A fase de convergência realiza ajustes finos nos vetores peso dos neurônios do mapa. Esta é a fase mais demorada e mantém fixos os valores de tamanho da vizinhança e taxa de aprendizagem utilizados, geralmente muito baixos. A vizinhança normalmente varia entre apenas o próprio neurônio (tamanho 0) ou seus vizinhos mais próximos (tamanho 1).

Nesta etapa, os vetores peso tendem a aprender a distribuição dos dados representados no espaço de entrada, enquanto mantém a ordenação aprendida na fase anterior. A Figura 4.3, extraída de (KOHONEN, 1990), mostra como as posições dos vetores peso de uma rede SOM são ajustadas ao longo de todo o processo de treinamento.

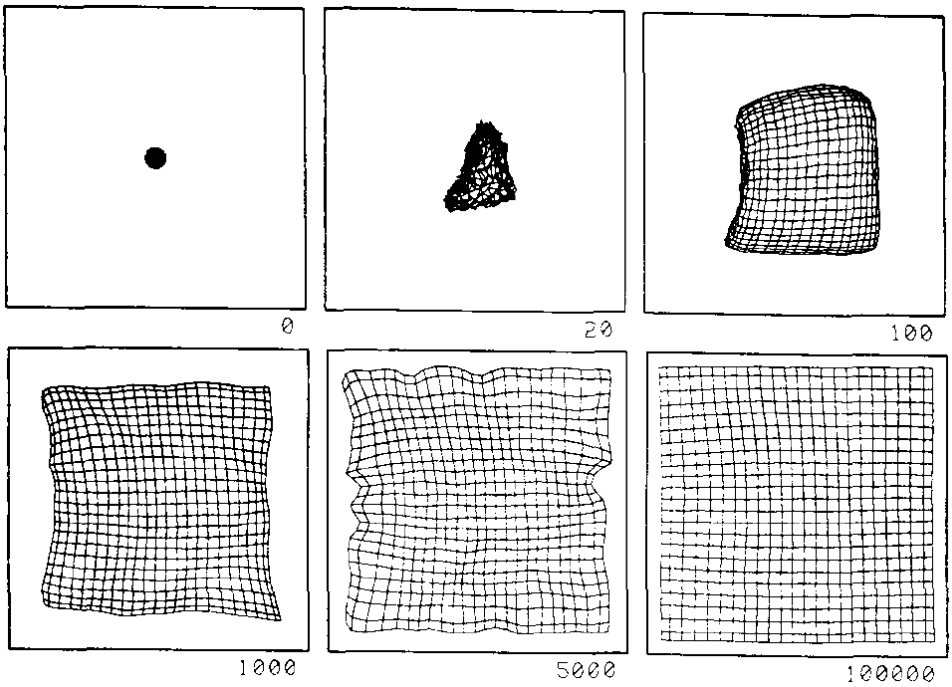


Figura 4.3: Posições dos neurônios de uma rede SOM em vários pontos no tempo ao longo do processo de treinamento.

4.2 Configuração da rede

Diferentes configurações para a rede foram testadas antes de se decidir qual seria utilizada neste trabalho. Inicialmente, foram criados mapas de topologia hexagonal com tamanhos 10x10, 15x15 e 20x20. Para o processo de treinamento não demorar muito, decidiu-se usar uma baixa quantidade de épocas, variando entre os valores 200, 500 e 1000. A quantidade de épocas utilizada na fase de ordenação foi sempre 100.

Tanto as redes com lados de tamanhos 15 e 20 apresentaram resultados satisfatórios nestes treinamentos preliminares. No fim, foi escolhido treinar um mapa de tamanho 15x15 por limitações de tempo, uma vez que as redes de tamanho 20x20 demoraram aproximadamente o dobro do tempo para serem treinadas.

(KOHONEN, 1990) faz sugestões práticas sobre o treinamento de um mapa auto-organizável. Algumas foram seguidas, mas outras não foram devido a limitações da natureza deste trabalho (tempo foi o fator crucial). A primeira dica diz que a quantidade de épocas utilizada no treinamento deve ser de cerca de 500 vezes a quantidade de neurônios presentes no mapa. Isso acarreteria em mais de 110 mil épocas para um mapa de tamanho 15x15, o que

inviável para este contexto. Assim, foi decidido utilizar apenas 10 mil épocas, o que já fez com que o treinamento demorasse 2 dias. Destas 10 mil épocas, mil foram utilizadas na primeira fase.

A fase de ordenação iniciou os valores de tamanho da vizinhança igual a 5 e taxa de aprendizagem 0,9. Durante esta etapa, os dois valores caíram linearmente para 1 e 0,02, respectivamente. Todos os valores da configuração podem ser vistos na Tabela 4.1.

Tabela 4.1: Valores de configuração da rede SOM.

Configuração	**Valor**
Topografia	Hexagonal
Tamanho do mapa	15 x 15
Qtd. de épocas total	10.000
Qtd. de épocas na ordenação	1.000
Vizinhança no início da ordenação	5
Taxa de aprendizagem no início da ordenação	0,9
Vizinhança no início da fase de ajustes	1
Taxa de aprendizagem no início da fase de ajustes	0,02

4.2.1 Inicialização dos vetores peso

A escolha dos vetores peso iniciais da rede é um assunto que já recebeu bastante atenção. Normalmente ela é feita de duas maneiras: aleatória ou baseada nos dados conhecidos. O algoritmo original considerava uma simples escolha aleatória dentre os vetores do espaço de entrada. Outra forma de inicialização aleatória pode ser feita causando perturbações nos valores do vetor média dos vetores de entrada. (SU; LIU; CHANG, 2002) utiliza interpolações lineares de centróides obtidos após a segmentação do espaço de entrada com a técnica *k-means*. (KOHONEN, 2001; SCHATZMANN; GHANEM, 2003) mostram como interpolar componentes a partir de *Principal Component Analysis* (PCA) da base, para serem utilizados como os vetores iniciais. (ATTIK; BOUGRAIN; ALEXANDRE, 2005) faz uma projeção linear dos dados e cria um grid irregular (com retângulos incongruentes) de modo que eles estejam distribuídos igualmente em cada célula. Os vetores presentes em cada célula são combinados para representar um vetor peso inicial.

Neste trabalho foi utilizada a inicialização dos vetores mostrada por (SCHATZMANN; GHANEM, 2003). Mais especificamente, o mapa é inicializado com o vetor média do *dataset* em seu centro. No caso, o vetor média é a média de todos os vetores da base. Os dois componentes principais com maiores auto valores associados são escolhidos para interpolar o vetor média nas duas dimensões, utilizando valores no intervalo [-*std*;*std*] como peso, onde *std* é o desvio padrão dos dados.

4.3 Segmentação do mapa

O fim do treinamento não marca o fim da análise. Quando uma rede SOM é treinada, seus vetores peso (também chamados de *proto-clusters* ou protótipos) precisam ser segmentados entre si para assim se obter os *clusters* do conjunto de dados. (VESANTO; ALHONIEMI, 2000) ataca este problema com uma segmentação hierárquica dos protótipos. Essa abordagem foi escolhida para este trabalho.

Na Figura 4.4 pode ser visto o dendrograma dos *proto-clusters*. O dendrograma de um conjunto de dados mostra como estes dados foram agrupados entre si. Este agrupamento dos dados é feito com funções de linkagem, que leva em consideração as distâncias entre eles. Para calcular a distância entre os vetores foi escolhida a distância euclidiana e o método de linkagem foi o método de Ward (WARD, 1963). Este método foi escolhido, pois ele minimiza a variância intra-*cluster* total, o que gerou resultados satisfatórios.

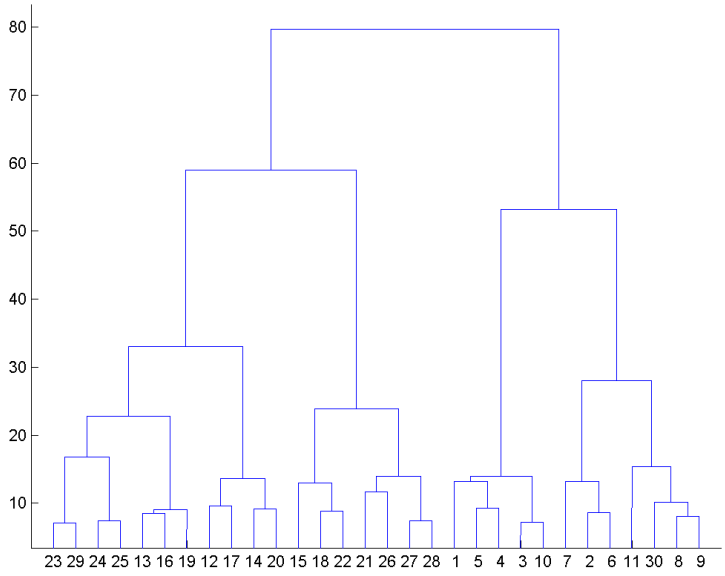
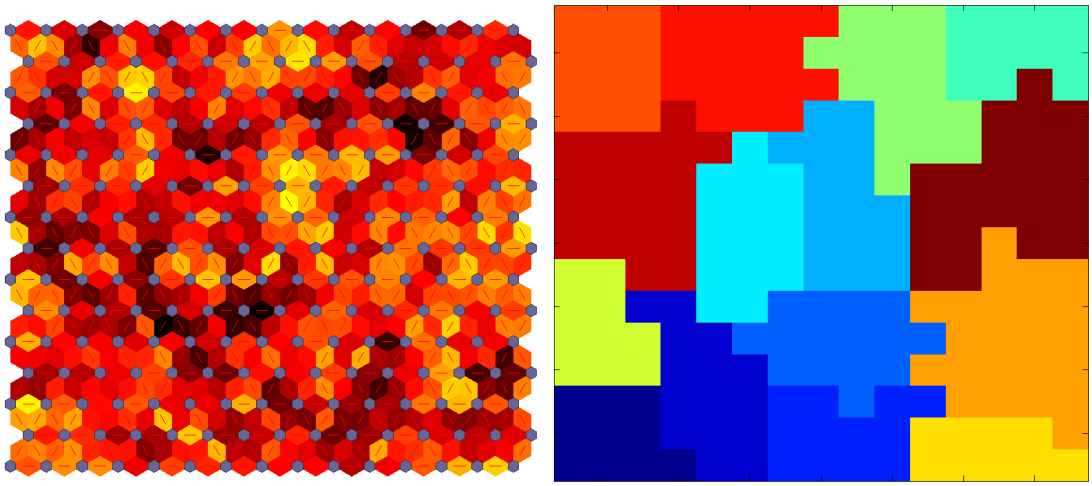


Figura 4.4: Dendrograma dos vetores peso após o treinamento. As linhas horizontais agrupam elementos e as linhas verticais medem a distância entre eles.

Pode-se criar grupos a partir de dendrogramas traçando linhas horizontais neles, sendo que os grupos são os ramos da árvore abaixo da linha. Observando a Figura 4.4, 2 *clusters* podem ser obtidos ao tracejar uma linha onde o eixo vertical marca 70. Ou então, traçar uma outra linha onde o eixo vertical marca 40 gera 4 *clusters*, por exemplo.

A Figura 4.5 mostra a representação U-Matrix e uma segmentação com 15 regiões do mapa partir do dendrograma. Pode-se verificar que a escolha de 15 grupos é visualmente agradável.

Para definir a quantidade de *clusters* que seriam gerados a partir do dendrograma, foi utilizado o índice de Davies-Bouldin (DAVIES; BOULDIN, 1979). Este critério leva em consideração tanto as distâncias intra-*cluster* dos dados, quanto as distâncias inter-*clusters* de seus



(a) U-Matrix (b) 15 clusters

Figura 4.5: U-Matrix e segmentação do mapa treinado.

centróides. Como o índice é calculado sobre o resultado de uma segmentação, o dendrograma foi particionado em várias quantidades de grupos diferentes. Especificamente, foram geradas divisões contendo de 2 a 15 grupos. O índice foi calculado para cada uma das 14 partições e foi escolhido um resultado baixo e que gerasse a maior quantidade de *clusters* possível. Os valores do índice para cada partição são mostrados na Tabela 4.2. A partição com 15 *clusters* foi escolhida, pois apesar de não ter gerado o menor índice, seu índice não difere muito do menor. Perceba que os valores do índice não respeitam nenhuma escala, o que quer dizer que eles apenas fazem sentido quando comparados entre si, não fazendo sentido comparar índices em bases de dados diferentes.

Tabela 4.2: Índices de Davies-Bouldin para diferentes partições do mapa.

Quantidade de *clusters*	**Índice de Davies-Bouldin**
2	4,61
3	4,68
4	4,80
5	4,40
6	4,02
7	4,12
8	4,21
9	3,86
10	3,72
11	3,70
12	3,77
13	3,74
14	3,71
15	3,78

5

Resultados e discussões

O capítulo anterior mostrou como a rede neural foi treinada para a base de dados de AIHs e segmentou seu resultado para uso posterior. Neste capítulo serão gerados resultados baseado nesta segmentação e estes serão analisados.

Este capitulo compreende duas etapas de fases diferentes do CRISP-DM. A primeira seção deste capítulo mostra quais indicadores de gestão hospitalar serão utilizados para avaliar o resultado do treinamento e faz esta avaliação, sendo um frumto da etapa de avaliação do modelo na fase de modelagem. A segunda seção estabelece qual é o *cluster* a ser analisado e a a Seção 5.3 apresenta diferentes análises nos dados desse grupo. Estas seções já fazem parte da fase de avaliação do CRISP-DM, mais especificamente a etapa de avaliação dos resultados obtidos na fase de modelagem.

5.1 Indicadores de gestão hospitalar e análise do treinamento

De modo a aferir a qualidade do processo de modelagem do problema, alguns indicadores de gestão hospitalar foram utilizados para comparar a segmentação dos conjuntos de treinamento e teste. O uso de indicadores é interessante, pois eles auxiliam a reter informações importantes do objeto analisado, tanto de forma quantitativa quanto qualitativa.

Existem várias propostas de indicadores que poderiam ser utilizadas. Este trabalho utilizou indicadores estabelecidos pela Agência Nacional de Vigilância Sanitária (ANVISA)[1] e pelo projeto SIPAGEH[2]. Foram escolhidos 5 indicadores:

- Taxa de mortalidade;
- Taxa de cesáreas;
- Média de permanência;
- Média de valores gastos em UTI;

[1] http://www.anvisa.gov.br/servicosaude/avalia/indicadores/index.htm - Acessado em julho de 2015
[2] http://www.projeto.unisinos.br/sipageh/ - Acessado em julho de 2015

- Média de valores totais das internações.

A taxa de mortalidade T_m mede a porcentagem de pacientes que chegaram a óbito em um determinado conjunto de internações. A taxa de cesáreas T_c mede a proporção de procedimentos obstetrícios cesarianos em relação a todos os partos realizados. Ela são medidas como:

$$T_m = 100 \times \frac{\textit{Número de óbitos}}{\textit{Número de altas}} \tag{5.1}$$

$$T_c = 100 \times \frac{\textit{Número de cesáreas}}{\textit{Número de partos}} \tag{5.2}$$

Os outros indicadores tem uma associação direta com os valores de atributos da tabela. Para esses são calculadas as médias dos valores apresentados pelas variáveis *DIAS_PERM*, *VAL_UTI* e *VAL_TOT*. As Figuras 5.1 - 5.5 mostram como esses indicadores se comportam para cada *cluster* nos conjuntos de treinamento e teste. As barras verticais presentes nas Figuras 5.3 - 5.5 marcam uma distância equivalente ao desvio padrão daquela variável em cada um dos grupos.

Pode-se verificar a partir destes gráficos que os *clusters* nos conjuntos de treinamento e teste têm comportamentos parecidos na maioria dos casos. Apenas em algumas situações os valores eram muito diferentes. O caso mais grave aparenta ser a taxa de cesáreas para o segmento de número 14. Enquanto no conjunto de treinamento ele apresenta uma taxa de 100%, no conjunto de testes essa taxa é de 0%. Isso acontece porque esse grupo de dados conta com apenas 1 procedimento de parto (que é cesariano) no conjunto de treinamento, enquanto no conjunto de testes não existe nenhum procedimento de parto. Ainda assim, resultados deste tipo não minimizam a qualidade do mapa, evidenciada pela comparação dos outros indicadores.

5.2 Escolha do *cluster* de interesse para análise

Após a verificação da boa qualidade do treinamento realizado, deve ser escolhido um *cluster* para ser analisado. (DEMŠAR, 2006) apresenta técnicas para comparar resultados de diversos algoritmos classificadores executados em múltiplos *datasets* diferentes. Uma dessas técnicas é o teste de Friedman (FRIEDMAN, 1937), que ordena os algoritmos para cada conjunto de dados separadamente, assinalando *ranks* menores para aqueles algoritmos com melhor desempenho. Em casos de empates, um rank médio é atribuído aos algoritmos em questão. O teste verifica se diferentes grupos de dados pertencem ao mesmo universo em relação a um conjunto de variáveis (ou medições delas). Analogamente, neste trabalho os *clusters* (representando os classificadores) serão ordenados de acordo com os valores que os indicadores previamente propostos (representando os *datasets*) assumem em cada um destes grupos.

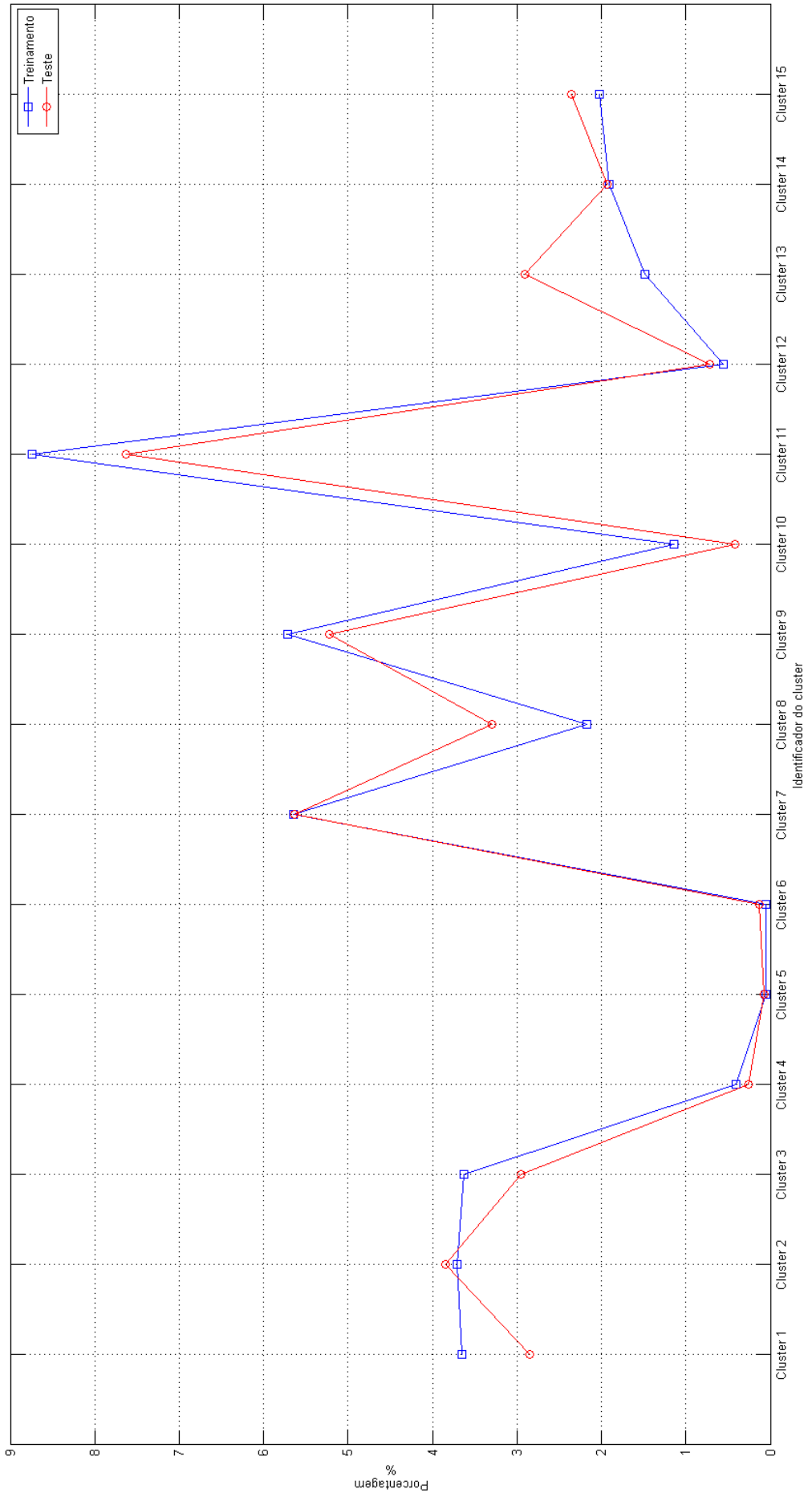


Figura 5.1: Indicador taxa de mortalidade.

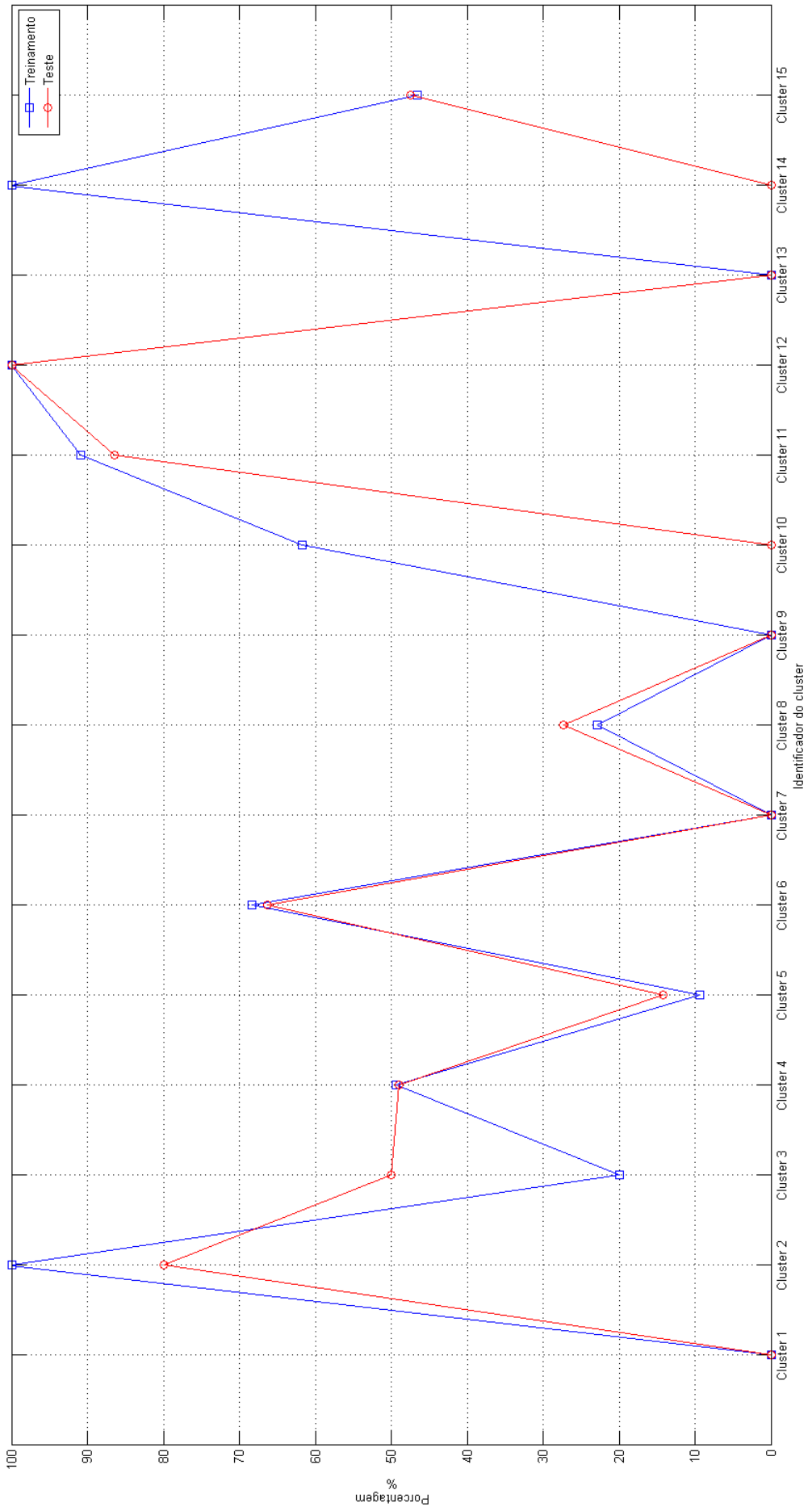


Figura 5.2: Indicador taxa de cesáreas.

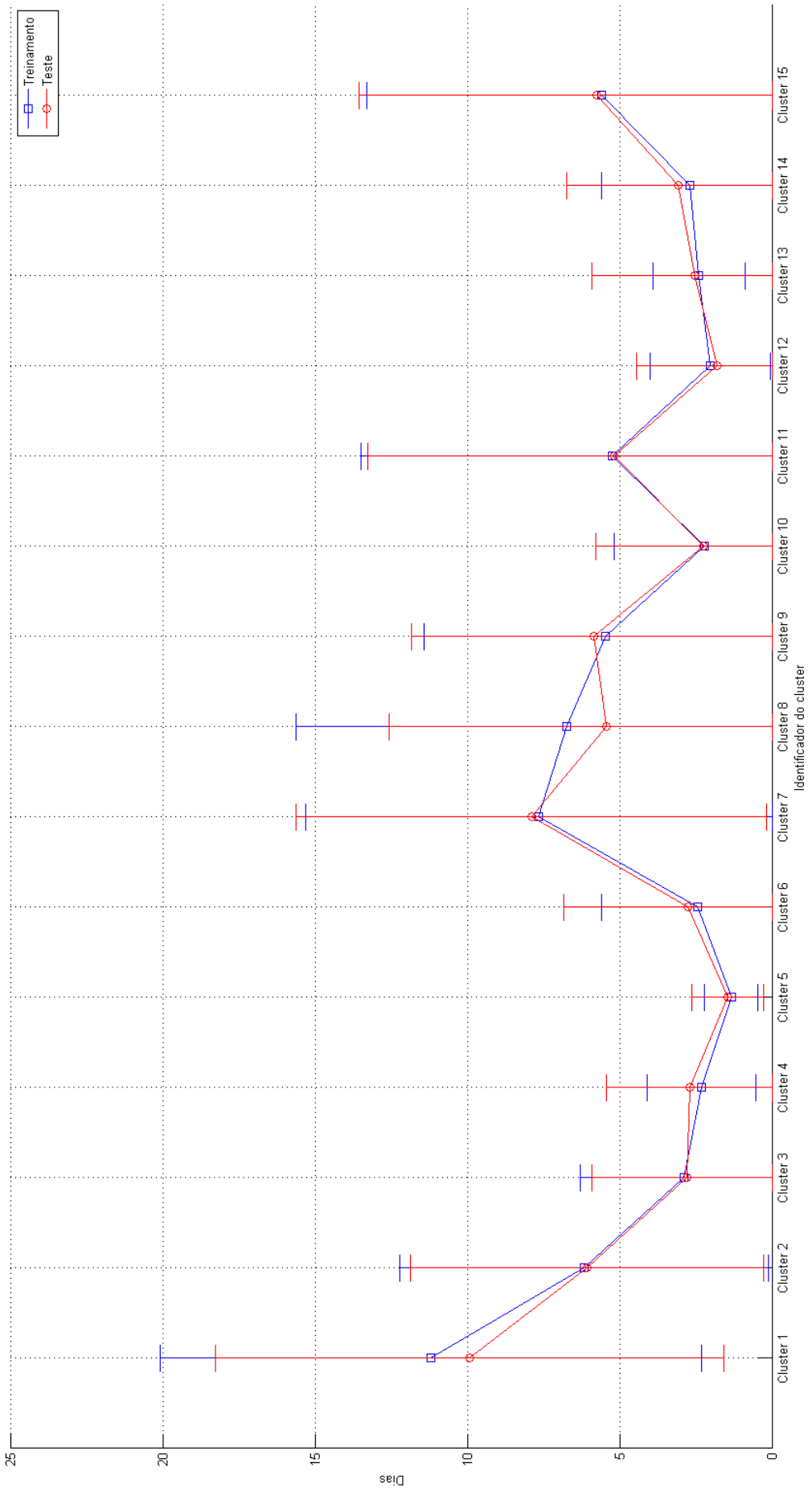


Figura 5.3: Indicador média de permanência.

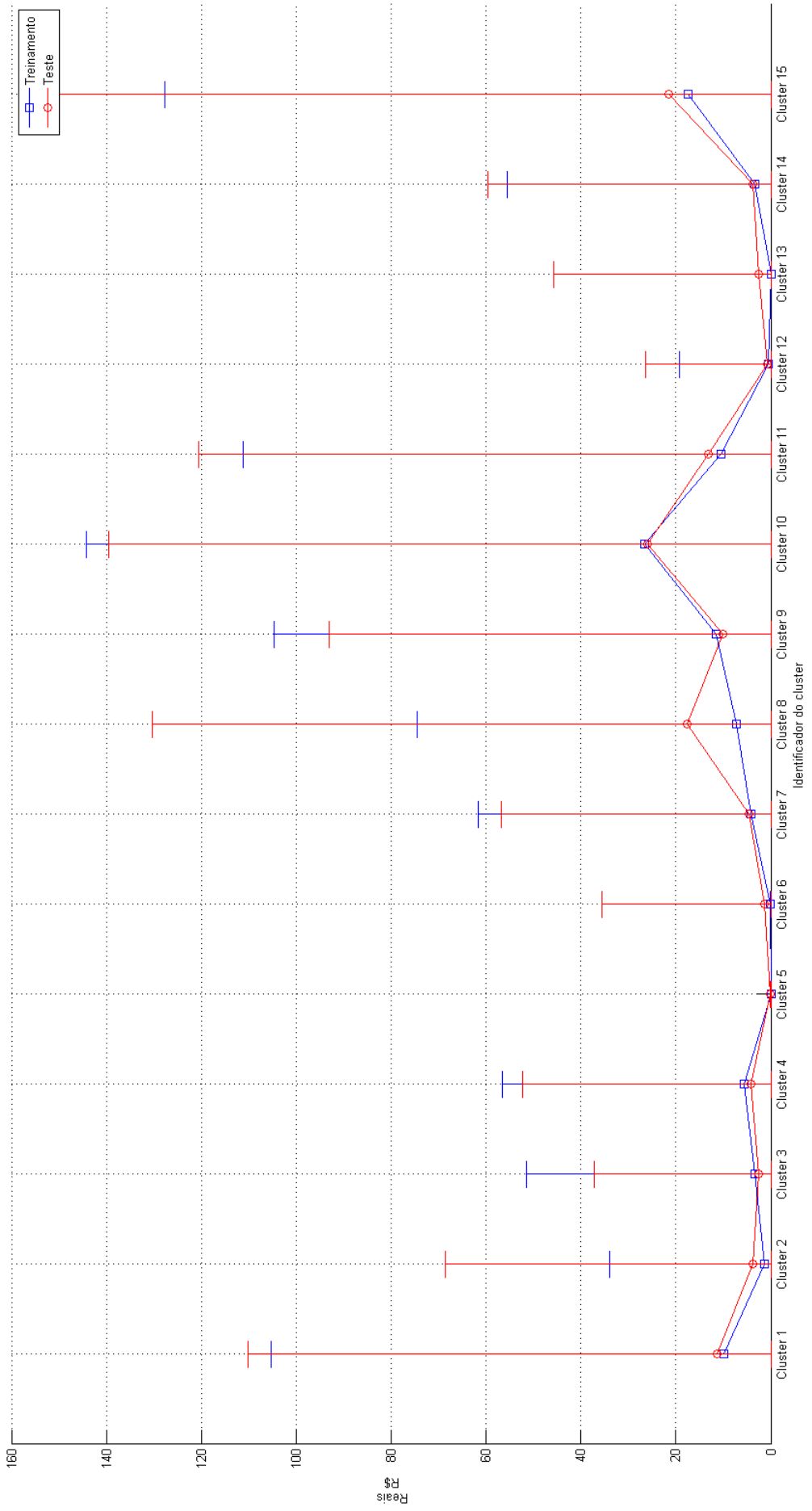


Figura 5.4: Indicador valores de UTI.

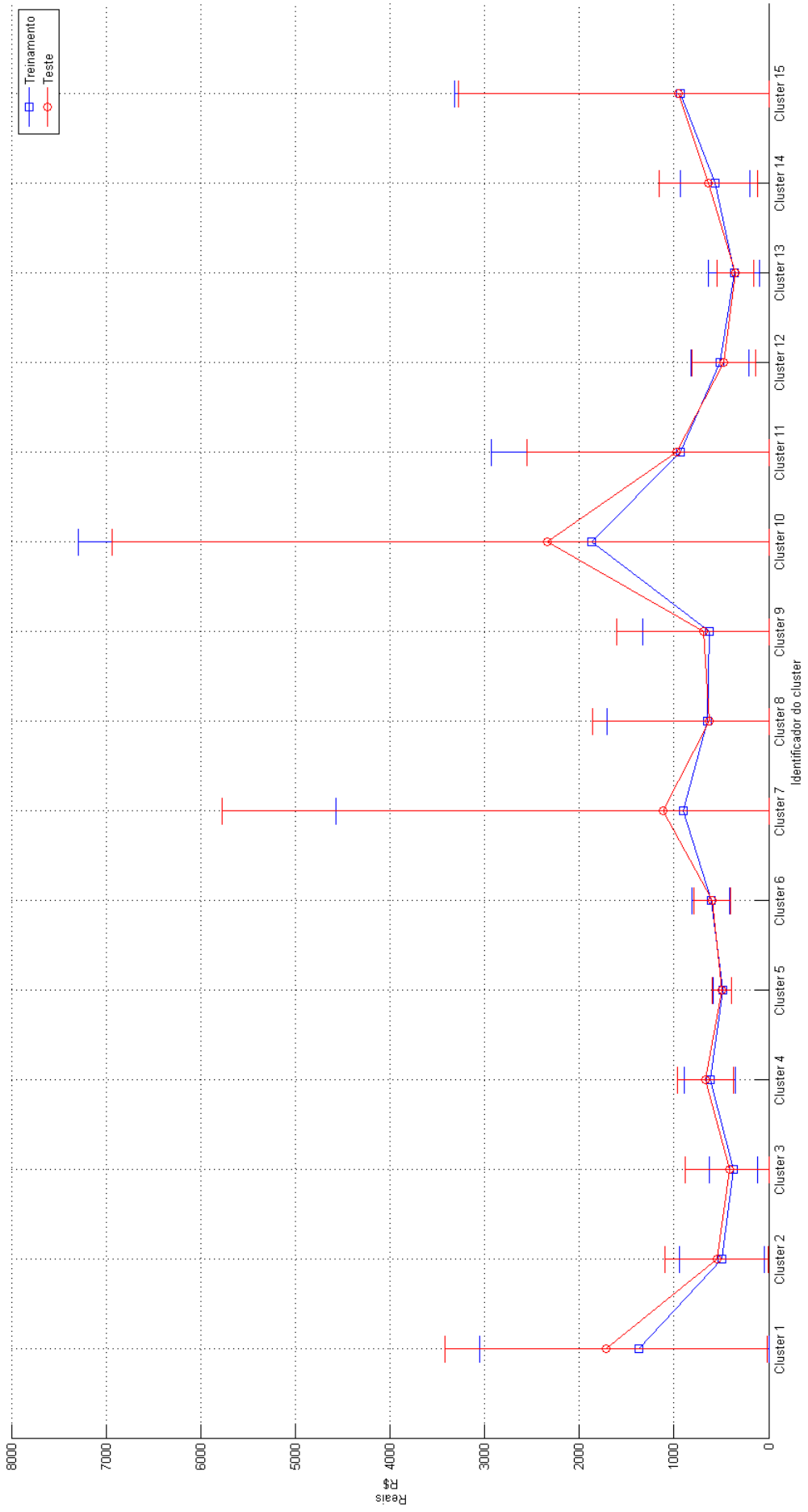


Figura 5.5: Indicador valores totais.

Será escolhida apenas uma região da segmentação para análise, e para isso, apenas o processo intermediário de criar uma tabela de ranks dos grupos de dados interessa. Todos os indicadores foram utilizados para gerar esta ordenação. Além deles foram utilizados os desvios padrões das variáveis representadas por indicadores de média (*i.e. DIAS_PERM*, *VAL_UTI* e *VAL_TOT*) e também o tamanho porcentual de cada *cluster* em relação à base de dados completa. Apenas a variável referente ao tamanho de cada *cluster* foi ordenada de forma ascendente. Todas as outras foram ordenadas de forma descendente. A Tabela 5.1 mostra o resultado da ordenação com todas as variáveis na segunda coluna. Para verificar a robustez da classificação original, foram feitas outras 9 ordenações, sempre desconsiderando exatamente uma variável. Essas outras ordenações podem ser vistas nas colunas 3-11 da mesma tabela.

Tabela 5.1: Ordenação dos clusters. A segunda coluna representa a ordenação com todas as variáveis e as colunas subsequentes são ordenações feitas retirando sempre uma variável. A variável retirada em cada coluna é explicitada em seu topo. Variáveis com uma barra horizontal em cima representam a média, a função sigma representa o desvio padrão e a última variável se refere ao tamanho porcentual de cada cluster.

Ordem	Original	Cesáreas	Mortalidade	$\overline{VAL\_TOT}$	$\sigma(VAL\_TOT)$	$\overline{VAL\_UTI}$	$\sigma(VAL\_UTI)$	$\overline{DIAS\_PERM}$	$\sigma(DIAS\_PERM)$	%
1	CL 11	CL 1	CL 10	CL 11	CL 11	CL 7	CL 7	CL 10	CL 10	CL 11
2	CL 1	CL 7	CL 1	CL 7	CL 1	CL 11	CL 11	CL 11	CL 11	CL 15
3	CL 10	CL 11	CL 15	CL 1	CL 7	CL 1	CL 1	CL 15	CL 7	CL 1
4	CL 7	CL 9 (4)	CL 11	CL 10 (4)	CL 10	CL 10	CL 10	CL 1 (4)	CL 1	CL 10
5	CL 15	CL 10 (4)	CL 7	CL 15 (4)	CL 15	CL 15	CL 15	CL 7 (4)	CL 15	CL 8
6	CL 9	CL 15	CL 8	CL 9	CL 9	CL 8	CL 9	CL 9	CL 9	CL 7
7	CL 8	CL 8	CL 9	CL 8	CL 8	CL 9	CL 8	CL 8	CL 8	CL 9
8	CL 2	CL 2	CL 2	CL 2	CL 2	CL 2	CL 2	CL 2	CL 2	CL 2
9	CL 3	CL 3	CL 14	CL 3	CL 3	CL 3	CL 3	CL 14	CL 14	CL 14
10	CL 14	CL 14	CL 4	CL 14	CL 14	CL 14	CL 14	CL 3	CL 3	CL 3
11	CL 4	CL 4	CL 3	CL 4	CL 4	CL 12	CL 4	CL 4	CL 4	CL 4
12	CL 12	CL 12	CL 12	CL 12	CL 12	CL 4	CL 12	CL 12	CL 12	CL 12
13	CL 6	CL 6	CL 6	CL 6	CL 6	CL 6	CL 6	CL 6	CL 6	CL 6
14	CL 13	CL 13	CL 5	CL 13	CL 5	CL 13	CL 13	CL 5	CL 13	CL 13
15	CL 5	CL 5	CL 13	CL 5	CL 13	CL 5	CL 5	CL 13	CL 5	CL 5

A análise desta tabela mostra que as ordenações não sofrem muita alteração quando desconsiderada uma variável qualquer. Por exemplo, observando as 5 últimas linhas da tabela, percebe-se que os *clusters* 4 e 12 se alternam entre a 11ª e 12ª posições, o *cluster* 6 mantém sempre a 13ª colocação e as posições 14 e 15 são ocupadas sempre pelos grupos 13 e 5.

De maneira similar, se as melhores colocações da tabela forem analisadas, percebe-se que 4 *clusters* específicos apresentam sempre bons resultados. São eles: 1, 7, 10 e 11. A escolha do grupo mais apropriado pode agora ser restrita a eles. Destes, os *clusters* 1 e 11 foram descartados devido a seus tamanhos, representando 6,1% e 12,7% da base respectivamente. Para escolher entre os *clusters* 7 e 10 foi visualizado os gráficos de seus indicadores, como nas Figuras 5.1 - 5.5. Porém, desta vez foram comparados os indicadores no conjunto de treinamentos e na base completa. Os dois apresentaram a mesma distribuição dos indicadores para os dois conjuntos de dados, com exceção do *cluster* 7 quando analisado o indicador taxa de cesáreas (Figura 5.6). Assim, ele foi preterido em relação ao *cluster* 10.

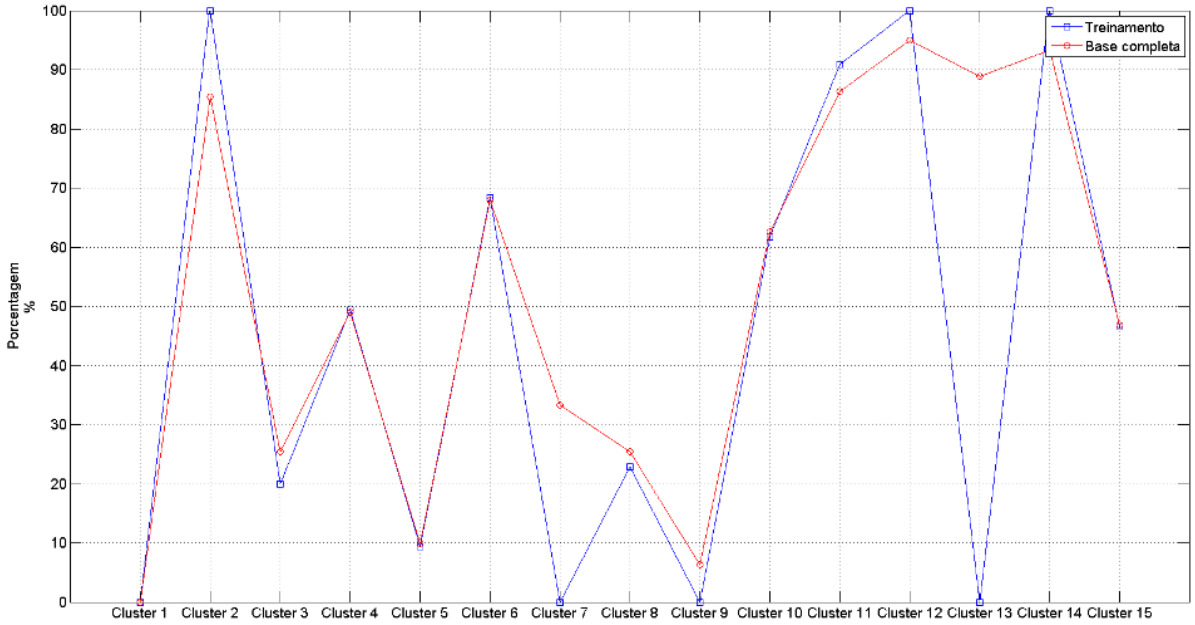


Figura 5.6: Indicador taxa de cesáreas para o conjunto de treinamento (azul) e a base completa (vermelho).

5.3 Análise do cluster

A análise do grupo 10 se baseou nos valores de procedimentos estabelecidos pela tabela de valores do SUS[3]. As Subseções 5.3.1 e 5.3.2 apresentam diferentes análises do mesmo conjunto de dados, mas seguindo caminhos diferentes.

5.3.1 Análise A

Nesta primeira análise, verificou-se a existência de 360 procedimentos de parto cesariano (código 0411010034) e nenhum destes registros possui o valor total em acordo com a tabela do SUS (Tabela 5.2). Foi constatado que este procedimento foi atualizado em junho de 2015, mas seus valores estão mantidos desde janeiro de 2008.

Tabela 5.2: Valores de partos cesarianos estabelecidos pelo SUS e os extremos encontrados no cluster 10.

Tipo	Valor do SUS	Valor mínimo	Valor Máximo
Serviço Hospitalar	R$395,68	R$396,68	R$452,68
Serviço Profissional	R$150,05	R$160,05	R$321,63
Total Hospitalar	R$545,73	R$556,73	R$774,31

As diferenças entre os valores cobrados e os da tabela chamaram a atenção. Todos aqueles 360 procedimentos foram feitos por apenas 2 hospitais, que serão chamados de Hospital

[3]http://sigtap.datasus.gov.br/tabela-unificada/app/sec/inicio.jsp - Acessado em julho de 2015

A1 e Hospital A2. Esses hospitais possuem 22 procedimentos cirúrgicos em comum (incluindo o parto cesariano), que também foram analisados em relação à tabela do SUS. Apesar dos dois hospitais cobrarem valores mais caros em vários desses procedimentos, as diferenças são geralmente baixas. Nos registros do Hospital A1 no *cluster* 10, o sobrepreço de procedimentos é pouco frequente. Apenas 3 procedimentos apresentaram uma diferença média de valor superior a R$50,00 (quando havia diferença). A Tabela 5.3 mostra a quantidade de ocorrências de cada procedimento e quantas vezes a diferença é superior a 0 real. A terceira coluna informa o montante dos sobrepreços e a última coluna mostra a diferença média de valor para cada procedimento quando maior que zero.

Tabela 5.3: Diferenças em valores de procedimentos para o Hospital A1 no cluster 10.

Cód. Procedimento	**#DIFF>0 / #Total**	**Σ DIFF**	$\overline{\textbf{DIFF > 0}}$
0401020088	1 / 8	R$16,00	R$16,00
0401020100	0 / 1	R$0,00	R$0,00
0406020574	1 / 26	R$20,06	R$20,06
0407020284	1 / 47	R$160,48	R$160,48
0407030026	6 / 103	R$344,84	R$57,47
0407040064	1 / 20	R$20,06	R$20,06
0407040080	0 / 1	R$0,00	R$0,00
0407040102	5 / 108	R$104,12	R$20,82
0407040129	2 / 54	R$32,00	R$16,00
0408060123	0 / 1	R$0,00	R$0,00
0409040070	0 / 1	R$0,00	R$0,00
0409040215	1 / 8	R$8,00	R$8,00
0409050083	3 / 52	R$308,90	R$102,97
0409060020	0 / 1	R$0,00	R$0,00
0409060100	0 / 1	R$0,00	R$0,00
0409060135	0 / 1	R$0,00	R$0,00
0409060194	0 / 1	R$0,00	R$0,00
0409060216	0 / 1	R$0,00	R$0,00
0409060232	0 / 1	R$0,00	R$0,00
0409070050	3 / 55	R$48,00	R$16,00
0411010034	205 / 205	R$25.215,72	R$123,00
0411020013	39 / 47	R$111,12	R$2,85

No Hospital A2, apenas o parto cesariano chama a atenção desses 22 procedimentos. Quando analisada a base completa dos dois hospitais, o parto cesariano gera um sobrepreço de R$122.507,55 no Hospital A1 e de R$28.819,52 no Hospital A2. Nenhum outro procedimento cirúrgico gera um sobrepreço maior que 1 mil reais nos 41 meses analisados. Com isto em mente, foram analisados os 3 procedimentos clínicos mais frequentes em cada hospital. Para o Hospital A1, eles são:

- **0303030046** - Tratamento de distúrbios metabólicos;
- **0303130016** - Atendimento a paciente sob cuidados prolongados devido à causas externas;
- **0303140046** - Tratamento das doenças crônicas das vias aéreas inferiores.

O tratamento de distúrbios metabólicos tem o campo *Quantidade Máxima* igual a 1, o que indica que sua cobrança não é afetada pelo tempo de permanência do paciente. Considerando o valor total de R$139,42 este procedimento apresenta um sobrepeço em aproximadamente 54% de seus registros no Hospital A1. No entanto, isso representa um sobrepreço de apenas R$3.683,27 durante o período analisado, principalmente devido a sua baixa quantidade de ocorrências (apenas 247 na base completa).

O segundo procedimento possui 841 ocorrências na base completa do Hospital A1. Ao contrário do procedimento anterior, este apresentar um valor de 45 como *Quantidade Máxima*. Isto significa que seu valor total pode ser cobrado até 45 vezes durante uma única internação. No caso, esta quantidade de cobranças é limitada pela quantidade de diárias do paciente, fazendo com que esse número deva ser utilizado para normalizar o valor total. Feito isto, verifica-se que alguns registros apresentam um preço por dia maior do que R$66,47 (estabelecido pelo SUS). No entanto, o que surpreende é que existem muito mais ocorrências de uma cobrança menor do que a esperada, fazendo na verdade com que o montante cobrado por esse procedimento seja negativo em R$4.641,90 no período de 2011 a 2014.

Este último procedimento é semelhante ao primeiro, no que diz respeito a como sua cobrança deve ser feita e a quantidade de registros com sobrepreço. Apesar de contar com 62% das ocorrências com preços diferentes da tabela, este procedimento também possui um montante baixo de diferença de preços, devido a conter apenas 78 registros na base de dados. Isto gera uma diferença ao longo do período de análise que pode ser considerada irrisória, de apenas R$1.964,60.

O Hospital A2 não possui nenhum procedimento clínico em comum o Hospital A1, quando se leva em consideração os 3 mais frequentes. Eles são:

- **0303140151** - Tratamento de pneumonias ou influenza (gripe);
- **0303010061** - Tratamento de doenças infecciosas e intestinais;
- **0310010039** - Parto normal.

O tratamento de pneumonias é outro procedimento que deve ser cobrado apenas 1 vez. Este procedimento tem um custo de R$582,42 e foi executado 300 pelo Hospital A2 ao longo de 41 meses. Porém, este procedimento foi sobretaxado em 63% das ocorrências, o que gerou um montante de aproximadamente R$6.700,00 cobrados a mais pelo hospital no período.

De maneira similar ao procedimento anterior, o tratamento de doenças infecciosas e intestinais também tem cobrança única e foi sobretaxado em grande parte de suas ocorrências no Hospital A2. Com 265 registros, sendo 71% deles mais caros do que deveriam, este procedimento teve um sobrepreço de aproximadamente R$5.500,00.

Por fim, o parto normal é outro procedimento que deve ter uma cobrança única, o que significa que a quantidade de diárias não deve impactar no valor total do procedimento. Dentro dos 3 procedimentos mais frequentes no neste grupo, este é o mais frequente quando se trata da

base completa, com 379 registros. Desses, menos de 4% não apresentam uma cobrança superior à estabelecida pelo SUS. No período analisado, o Hospital A2 cobrou cerca de R$13.750,00 a mais pelo parto normal.

As Tabelas 5.4 e 5.5 sumarizam os sobrepreços cobrados pelos hospitais A e B relatados acima. No entanto, essa análise não é conclusiva sobre a índole dos hospitais por dois motivos:

1. Apesar dos preços diferentes não serem explicados na base de dados e não serem corroborados pela tabela do SUS, eles podem ser legais por outros motivos. O caso dos partos cesarianos, por exemplo, é de se estranhar porque não existe um único caso onde seja cobrado o valor previsto, em nenhum dos dois hospitais;

2. Apenas um amostra de todos os procedimentos realizados por esses hospitais foi analisada aqui. A prática de sobretaxar o SUS pode ser sistêmica ou não.

Tabela 5.4: Diferenças em valores de procedimentos específicos para o Hospital A1 na base completa.

Código do Procedimento	Montante de sobrepreço
0411010034	R$122.507,55
0303030046	R$3.683,27
0303130016	-R$4.641,90
0303140046	R$1.964,60
TOTAL	**R$123.513,52**

Tabela 5.5: Diferenças em valores de procedimentos específicos para o Hospital A2 na base completa.

Código do Procedimento	Montante de sobrepreço
0411010034	R$28.819,52
0303140151	R$6.696,60
0303010061	R$5.503,15
0310010039	R$13.759,04
TOTAL	**R$54.778,31**

5.3.2 Análise B

Diferente da análise anterior, esta não tem um procedimento específico em mente. Em primeiro lugar foi verificado quais hospitais fazem parte do *cluster* 10 e quantos procedimentos realizados nele cada um apresenta. Constatou-se que existem apenas 12 hospitais diferentes sendo representados neste *cluster*, mas apenas 6 contêm mais de 10 procedimentos realizados. Estes foram nomeados aqui de Hospital B1 até Hospital B6 e a Tabela 5.6 apresenta algumas informações deles, como quantidade de procedimentos e valores médios cobrados.

Tabela 5.6: Estatísticas de cada hospital no cluster 10. A segunda e terceira colunas informam o total de procedimentos realizados e quantidade de procedimentos únicos. As duas últimas mostram o valor médio e o desvio padrão de todos os procedimentos (deste cluster) em cada hospital.

Hospital	Procedimentos realizados	Procedimentos distintos	Valor médio	Desvio padrão
B1	1.606	228	R$1.523,54	R$4.648,19
B2	217	18	R$1.578,04	R$1.025,14
B3	611	114	R$1.397,93	R$1.357,80
B4	96	20	R$9.429,54	R$2.762,60
B5	81	36	R$445,57	R$303,76
B6	154	37	R$506,80	R$248,42

Dois hospitais chamam a atenção quando se observa a tabela acima. Apesar do Hospital B1 possuir uma média dos valores similar a de outros dois (B2 e B3), ele possui um desvio padrão nos valores dos procedimentos bastante elevado. Isso acontece devido à sua quantidade de procedimentos distintos ser muito superior a dos outros hospitais, o que acaba por incluir procedimentos de complexidades variadas neste grupo. O segundo hospital que chama atenção é o B4, por possuir um valor médio dos procedimentos muito superior a dos demais. Aliado a isto, o desvio padrão dos mesmos valores também destoa, principalmente quando se leva em consideração sua baixa quantidade de procedimentos distintos. Os hospitais B5 e B6 possuem valores muito abaixo dos outros por não serem hospitais da capital, que lidam com casos de maior complexidade.

As Figuras 5.7 e 5.8 mostram informações detalhadas de dois procedimentos que destoam dos outros realizados pelo Hospital B4.

Figura 5.7: Informações do procedimento 0406010692.

Dos 20 procedimentos distintos realizados pelo Hospital B4, a maior parte possui poquíssimias ocorrências. No entanto, esses dois procedimentos se diferem bastante dos outros porque são os únicos que possuem mais de 10 ocorrências registradas no neste grupo.

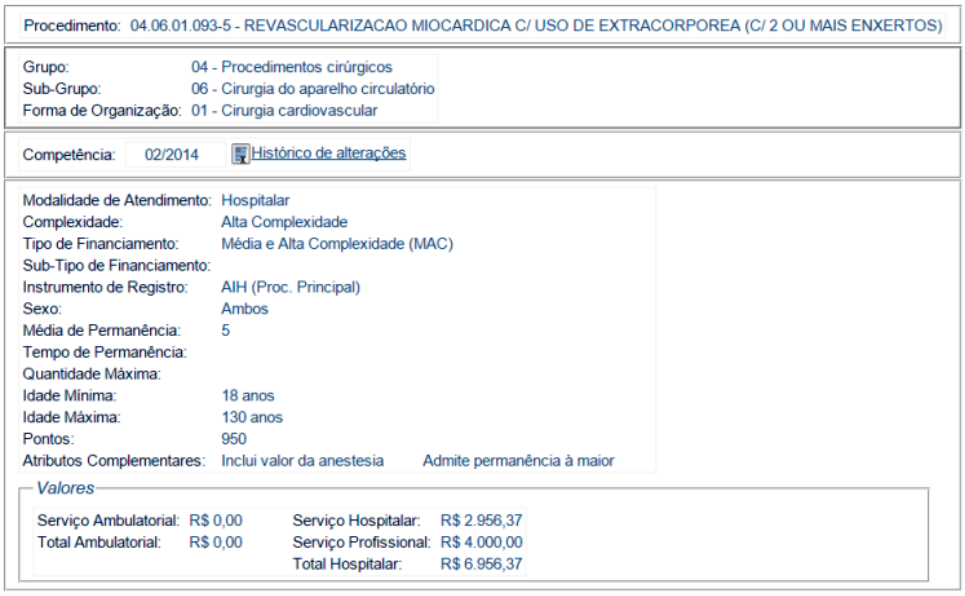


Figura 5.8: Informações do procedimento 0406010935.

Nota-se que apesar destes dois procedimentos apresentarem valores elevados, o valor médio dos procedimentos do Hospital B4 ainda é muito superior a eles. O procedimento 0406010692 possui um valor total tabelado pelo SUS de R$6.321,74 e o outro procedimento deveria custar R$6.956,37, mas apresentam valores médios de R$10.456,10 e R$9.810,45 respectivamente. Para entender melhor de onde vem essas diferenças nos valroes totais, deve-se observar também as diferenças nos valores de serviços hospitalares (SH) e serviços profissionais (SP) de cada procedimento.

A tabela do SUS prevê valores de serviços hospitalares e profissionais para o procedimento 0406010692 de R$2.956,37 e R$3.365,37 respectivamente. No entanto, os valores médios cobrados pelo hospital B4 são de R$6.962,93 (SH) e R$3.493,17 (SP). De maneira semelhante, os valores previstos para o procedimento 0406010935 são de R$2.956,37 para o serviço hospitalar e de R$4.000,00 para o profissional, enquanto são cobrados em média R$5.686,52 e R$4.123,93 respectivamente. Com as informações sumarizadas na Tabela 5.7, é fácil perceber que as diferenças significativas são relacionadas aos serviços hospitalares.

Quando se observa todos as ocorrências destes procedimentos na base completa, nota-se a predominância do Hospital B4 em suas execuções. O procedimento de código 0406010692 possui 526 ocorrências registradas, sendo que 86,7% são do Hospital B4. De maneira semelhante, o mesmo hospital é responsável por 86,1% dos 885 registros feitos do segundo procedimento. No entanto, todos os outros hospitais que realizam estes mesmos procedimentos apresentam valores similares, como pode ser visto na Tabela 5.8.

Mesmo que essas diferenças nos valores dos procedimentos sejam válidas, elas não

Tabela 5.7: Diferenças para valores de procedimentos no Hospital B4.

	Proc. 0406010692	**Proc. 0406010935**
Valor SH	R$6.962,93	R$5.686,52
Valor SP	R$3.493,17	R$4.123,93
Valor Total	R$10.456,10	R$9.810,45
Diferença total média	R$3.655,64	R$2.403,52
Soma das diferenças	R$87.735,31	R$81.719,70

são explicadas pelos dados e nem pela tabela do SUS. Não cabe também a explicação de que ocorreram aumentos com o passar do tempo, pois seus valores na tabela do SUS não são atualizados desde 2008 e os registros também não apontam aumentos quando analisados em uma frequência anual. As médias de cada ano no período de 2011 até 2014 é muito similar, estando sempre entre R$10.500,00 e R$10.600,00.

Tabela 5.8: Valores médios de cada procedimento em outros hospitais.

Hospital	**Proc. 0406010692 (Qtd.)**	**Proc. 0406010935 (Qtd.)**
C1	R$12.944,42 (2)	R$8.308,91 (4)
C2	R$9.504,28 (10)	R$11.180,76 (36)
C3	R$10.260,55 (33)	R$10.308,12 (43)
C4	R$10.162,36 (15)	R$10.019,56 (27)
C5	R$10.936,38 (4)	R$0,00 (0)
C6	R$9.891,86 (1)	R$11.245,88 (4)
C7	R$9.472,55 (5)	R$9.677,68 (9)

Portanto, pode-se dizer que os dados apresentados e encontratos nestas duas análises indicam desacordos com tabela de cobranças divulgada pelo SUS. Essas informações sugerem que algum tipo de tabela paralela é utilizada na prática pelos hospitais conveniados ao sistema. Esta hipótese só poderia ser verificada com a disponibilização de mais informações por parte do próprio SUS ou até mesmo uma investigação *in loco*.

6

Conclusão

Este trabalho abordou a importância de dados públicos serem abertos e mostrou uma forma de se trabalhar com eles, especificamente aplicando técnicas de mineração de dados. Todo o trabalho foi feito tendo seguindo diretrizes do modelo de processo CRISP-DM, passando por etapas de extrema importância, como a análise dos dados e a modelagem. Assim, este capítulo pode ser entendido como os resultados das etapas de revisão do processo e estabelecimento de próximos passos na fase de avaliação do CRISP-DM.

Um fruto muito importante deste trabalho é a visão geral feita sobre a base de dados de Autorizações de Internações Hospitalares do SUS. Muitas de suas 95 variáveis foram esmiuçadas e contextualizadas. Como parte desta análise, foram sugeridas ações a serem tomadas com esta base de dados, indicando problemas e soluções encontradas.

A análise investigativa dos dados foi possibilitada com o uso de Mapa Auto-Organizáveis na modelagem dos dados. A rede neural foi treinada com um conjunto de treinamento que consistia de dados até maio de 2013 e testada com dados do período de junho de 2013 a maio de 2014. Com a devida validação do treinamento da rede, ela foi segmentada e foi escolhido um *cluster* de interesse para análise. A escolha deste *cluster* foi feita com base em indicadores de gestão hospitalar sugeridos pelo Ministério da Saúde e outra instituição.

Após a devida segmentação do mapa, a base de dados completa, com aproximadamente 1,6 milhão de registros, foi executada nele. Nesta etapa foi possível identificar vários registros de internações hospitalares que necessitam de mais atenção, como hospitais que cobram sistematicamente a mais por procedimentos de preços tabelados.

A próxima seção deste capítulo fala de sugestões de trabalhos futuros, indicações de onde este trabalho pode melhorar.

6.1 Trabalhos futuros

Alguns pontos no curso do processo feito aqui podem ser aprimorados. Algo que ajudaria em várias partes do processo deste trabalho é obter a opinião de especialistas sobre o assunto. Por especialistas deve-se entender gestores de hospitais. O autor requisitou ajuda de médicos

durante a execução deste projeto, mas o conhecimento deles estava restrito à medicina. Ou seja, não puderam ajudar em relação às práticas de gestão utilizadas pelos estabelecimentos onde trabalham. A consultoria de um gestor traria benefícios gerais e específicos, dentre eles:

- Análise dos objetivos e conquistas do projeto;
- Reavaliação das variáveis utilizadas;
- Utilização de melhores indicadores.

Estes três itens são de suma importância para este projeto. Mais especificamente, os dois últimos são partes práticas que podem ser continuamente refinadas. A reavaliação das variáveis utilizadas reclassificaria a importância de cada variável, redefinindo quais ficariam no conjunto final, quais ficariam de fora e quais seriam levadas em consideração para a geração dos conjuntos de treinamento e teste. A utilização de melhores indicadores faria na verdade com que o trabalho tomasse um rumo mais específico, a depender de outros indicadores escolhidos. Esta parte mudaria consideravelmente a etapa de análise dos resultados, possibilitando a ocorrência de novas descobertas.

Outro ponto onde este trabalho tem espaço para melhorar é na etapa de modelagem. A utilização de Mapa Auto-Organizáveis funcionou muito bem e rendeu resultados satisfatórios. Contudo, suas configurações devem ser pensadas novamente, principalmente o tamanho do mapa e a quantidade de épocas. Mapas de tamanhos maiores devem ser testados e a quantidade de épocas deve acompanhar esse crescimento. Foi mencionado no Capítulo 4 que a quantidade de épocas utilizada para treinar uma rede SOM deve ser por volta de 500 vezes a quantidade de neurônios no mapa. Neste trabalho, a quantidade de épocas era menos de 50 vezes a quantidade de neurônios do mapa. As escolhas desses dois parâmetros foram restringidas pelo tempo necessário para treinar a rede. Com um mapa de tamanho 15x15 e treinamento durando 10 mil épocas, todo o processo de treinamento da rede durou 2 dias para ser completado. Isto indica que além de modificar os parâmetros, implementações mais rápidas do algoritmo devem ser buscadas. Implementações mais rápidas poderiam ser em outra linguagem, com paralelização ou até mesmo de versões modificadas do algoritmo.

Referências

ATTIK, M.; BOUGRAIN, L.; ALEXANDRE, F. Self-organizing map initialization. In: **Artificial Neural Networks**: Biological Inspirations–ICANN 2005. [S.l.]: Springer, 2005. p.357–362.

Bloomberg. **Most Efficient Health Care 2014**: Countries. Acessado em 30 de junho de 2015, `http://www.bloomberg.com/infographics/2014-09-15/most-efficient-health-care-around-the-world.html`.

BOLTON, R. J.; HAND, D. J. Statistical fraud detection: A review. **Statistical science**, [S.l.], p.235–249, 2002.

BROCKETT, P. L.; XIA, X.; DERRIG, R. A. Using Kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud. **Journal of Risk and Insurance**, [S.l.], p.245–274, 1998.

CHAPMAN, P. et al. **CRISP-DM 1.0 Step-by-step data mining guide**. 2000.

DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, [S.l.], n.2, p.224–227, 1979.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. **The Journal of Machine Learning Research**, [S.l.], v.7, p.1–30, 2006.

FRIEDMAN, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. **Journal of the American Statistical Association**, [S.l.], v.32, n.200, p.675–701, 1937.

KIANG, M. Y. Extending the Kohonen self-organizing map networks for clustering analysis. **Computational Statistics & Data Analysis**, [S.l.], v.38, n.2, p.161–180, 2001.

KOHONEN, T. The self-organizing map. **Proceedings of the IEEE**, [S.l.], v.78, n.9, p.1464–1480, 1990.

KOHONEN, T. **Self-Organizing Maps**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001.

LUBAMBO, S. W. **Processo de mineração de dados como apoio à decisão no controle de gastos públicos**. 2008. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal de Pernambuco.

MASSEY, F. J. The Kolmogorov-Smirnov test for goodness of fit. **Journal of the American statistical Association**, [S.l.], v.46, n.253, p.68–78, 1951.

MCCUE, C. **Data mining and predictive analysis**: Intelligence gathering and crime analysis. [S.l.]: Butterworth-Heinemann, 2014.

PEARSON, K. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, [S.l.], v.50, n.302, p.157–175, 1900.

PHUA, C. et al. A comprehensive survey of data mining-based fraud detection research. **arXiv preprint arXiv:1009.6119**, [S.l.], 2010.

PIATETSKY, G. **CRISP-DM, still the top methodology for analytics, data mining, or data science projects**. Acessado em 19 de junho de 2015, http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html.

PLANEJAMENTO, M. do. **Portal brasileiro de Dados Abertos**. Acessado em 19 de junho de 2015, http://dados.gov.br/.

SCHATZMANN, J.; GHANEM, M. Using self-organizing maps to visualize clusters and trends in multidimensional datasets. **Department of Computing Data Mining Group, Imperial College, London**, [S.l.], p.132, 2003.

SU, M. C.; LIU, T. K.; CHANG, H.-T. Improving the self-organizing feature map algorithm using an efficient initialization scheme. **Tamkang Journal of Science and Engineering**, [S.l.], v.5, n.1, p.35–48, 2002.

VESANTO, J.; ALHONIEMI, E. Clustering of the self-organizing map. **IEEE Transactions on Neural Networks**, [S.l.], v.11, n.3, p.586–600, 2000.

WARD, J. H. Hierarchical grouping to optimize an objective function. **Journal of the American statistical association**, [S.l.], v.58, n.301, p.236–244, 1963.

WATTS, M. J.; WORNER, S. Estimating the risk of insect species invasion: Kohonen self-organising maps versus k-means clustering. **Ecological Modelling**, [S.l.], v.220, n.6, p.821–829, 2009.

WHO. **Global Health Expenditure Database**. Acessado em 30 de junho de 2015, http://apps.who.int/nha/database.

WORLD HEALTH ASSEMBLY, . **Regarding Age-Grouping to be used in the Tabulation of Health Statistics**. URL: http://apps.who.int/iris/handle/10665/86051.

YANG, L.; OUYANG, Z.; SHI, Y. A modified clustering method based on self-organizing maps and its applications. **Procedia Computer Science**, [S.l.], v.9, p.1371–1379, 2012.

Apêndice

A

Descrição das variáveis da base

Tabela A.1: Parte 1 do layout dos arquivos do DATASUS.

Índice	Nome do campo	Tipo	Descrição
1	UF_ZI	char(6)	Município gestor
2	ANO_CMPT	char(4)	Ano de processamento da AIH, no formato aaaa
3	MES_CMPT	char(2)	Mês de processamento da AIH, no formato mm
4	ESPEC	char(2)	Especialidade do leito
5	CGC_HOSP	char(14)	CNPJ do Estabelecimento
6	N_AIH	char(13)	Número da AIH
7	IDENT	char(1)	Identificação do tipo da AIH
8	CEP	char(8)	CEP do paciente
9	MUNIC_RES	char(6)	Município de residência do paciente
10	NASC	char(8)	Data de nascimento do paciente (aaaammdd)
11	SEXO	char(1)	Sexo do paciente
12	UTI_MES_IN	numeric(2)	Zerado
13	UTI_MES_AN	numeric(2)	Zerado
14	UTI_MES_AL	numeric(2)	Zerado
15	UTI_MES_TO	numeric(3)	Quantidade de dias de UTI no mês
16	MARCA_UTI	char(2)	Indica qual o tipo de UTI utilizada pelo paciente
17	UTI_INT_IN	numeric(2)	Zerado
18	UTI_INT_AN	numeric(2)	Zerado
19	UTI_INT_AL	numeric(2)	Zerado
20	UTI_INT_TO	numeric(3)	Quantidade de diárias em unidade intermediaria
21	DIAR_ACOM	numeric(3)	Quantidade de diárias de acompanhante
22	QT_DIARIAS	numeric(3)	Quantidade de diárias
23	PROC_SOLIC	char(10)	Procedimento solicitado
24	PROC_REA	char(10)	Procedimento realizado
25	VAL_SH	numeric(13,2)	Valor de serviços hospitalares
26	VAL_SP	numeric(13,2)	Valor de serviços profissionais
27	VAL_SADT	numeric(13,2)	Zerado
28	VAL_RN	numeric(13,2)	Zerado
29	VAL_ACOMP	numeric(13,2)	Zerado
30	VAL_ORTP	numeric(13,2)	Zerado

Índice	Nome do campo	Tipo	Descrição
31	VAL_SANGUE	numeric(13,2)	Zerado
32	VAL_SADTSR	numeric(11,2)	Zerado
33	VAL_TRANSP	numeric(13,2)	Zerado
34	VAL_OBSANG	numeric(11,2)	Zerado
35	VAL_PED1AC	numeric(11,2)	Zerado
36	VAL_TOT	numeric(14,2)	Valor total da AIH
37	VAL_UTI	numeric(8,2)	Valor de UTI
38	US_TOT	numeric(10,2)	Valor total, em dólar
39	DT_INTER	char(8)	Data de internação no formato aaammdd.
40	DT_SAIDA	char(8)	Data de saída no formato aaaammdd.
41	DIAG_PRINC	char(4)	Código do diagnóstico principal (CID10)
42	DIAG_SECUN	char(4)	Código do diagnóstico secundário (CID10)
43	COBRANCA	char(2)	Motivo de saída/permanência
44	NATUREZA	char(2)	Natureza jurídica do hospital (com conteúdo até maio/12). Era utilizada a classificação de Regime e Natureza
45	NAT_JUR	char(4)	Natureza jurídica do estabelecimento, conforme a Comissão Nacional de Classificação - CONCLA
46	GESTAO	char(1)	Indica o tipo de gestão do hospital
47	RUBRICA	numeric(5)	Zerado
48	IND_VDRL	char(1)	Indica exame VDRL
49	MUNIC_MOV	char(6)	Município do estabelecimento
50	COD_IDADE	char(1)	Unidade de medida da idade
51	IDADE	numeric(2)	Idade
52	DIAS_PERM	numeric(5)	Dias de permanência
53	MORTE	numeric(1)	Indica óbito
54	NACIONAL	char(2)	Código da nacionalidade do paciente
55	NUM_PROC	char(4)	Zerado
56	CAR_INT	char(2)	Caráter da internação
57	TOT_PT_SP	numeric(6)	Zerado
58	CPF_AUT	char(11)	Zerado
59	HOMONIMO	char(1)	Indicador se o paciente da AIH é homônimo do paciente de outra AIH
60	NUM_FILHOS	numeric(2)	Número de filhos do paciente
61	INSTRU	char(1)	Grau de instrução do paciente
62	CID_NOTIF	char(4)	CID de notificação
63	CONTRACEP1	char(2)	Tipo de contraceptivo utilizado
64	CONTRACEP2	char(2)	Segundo tipo de contraceptivo utilizado
65	GESTRISCO	char(1)	Indicador se é gestante de risco
66	INSC_PN	char(12)	Número da gestante no pré-natal
67	SEQ_AIH5	char(3)	Sequencial de longa permanência (AIH tipo 5)
68	CBOR	char(3)	Ocupação do paciente, segundo a Classificação Brasileira de Ocupações - CBO
69	CNAER	char(3)	Código de acidente de trabalho
70	VINCPREV	char(1)	Vínculo com a Previdência

Índice	Nome do campo	Tipo	Descrição
71	GESTOR_COD	char(3)	Motivo de autorização da AIH pelo gestor
72	GESTOR_TP	char(1)	Tipo de gestor
73	GESTOR_CPF	char(11)	Número do CPF do gestor
74	GESTOR_DT	char(8)	Data da autorização dada pelo Gestor (aaaammdd)
75	CNES	char(7)	Código CNES do hospital
76	CNPJ_MANT	char(14)	CNPJ da mantenedora
77	INFEHOSP	char(1)	Status de infecção hospitalar
78	CID_ASSO	char(4)	CID causa
79	CID_MORTE	char(4)	CID da morte
80	COMPLEX	char(2)	Complexidade
81	FINANC	char(2)	Tipo de financiamento
82	FAEC_TP	char(6)	Subtipo de financiamento FAEC
83	REGCT	char(4)	Regra contratual
84	RACA_COR	char(4)	Raça/Cor do paciente
85	ETNIA	char(4)	Etnia do paciente, se raça cor for indígena.
86	SEQUENCIA	numeric(9)	Sequencial da AIH na remessa
87	REMESSA	char(21)	Número da remessa
88	AUD_JUST	char (50)	Justificativa do auditor para aceitação da AIH sem o número do Cartão Nacional de Saúde
89	SIS_JUST	char (50)	Justificativa do estabelecimento para aceitação da AIH sem o número do Cartão Nacional de Saúde
90	VAL_SH_FED	numeric (10, 2)	Valor do complemento federal de serviços hospitalares. Está incluído no valor total da AIH
91	VAL_SP_FED	numeric (10, 2)	Valor do complemento federal de serviços profissionais. Está incluído no valor total da AIH
92	VAL_SH_GES	numeric (10, 2)	Valor do complemento do gestor (estadual ou municipal) de serviços hospitalares. Está incluído no valor total da AIH
93	VAL_SP_GES	numeric (10, 2)	Valor do complemento do gestor (estadual ou municipal) de serviços profissionais. Está incluído no valor total da AIH
94	VAL_UCI	numeric (10, 2)	Valor de UCI
95	MARCA_UCI	char (2)	Tipo de UCI utilizada pelo paciente