Mineração de dados em triagem de risco de saúde

Article · May 2015

DOI: 10.5335/rbca.2015.4651

CITATIONS
2

READS
89

4 authors:

Thales Vaz Maciel
Universidade Federal do Rio Grande (FURG)
1 PUBLICATION **2** CITATIONS
SEE PROFILE

Vinicius Rosa Seus
Universidade Federal do Rio Grande (FURG)
17 PUBLICATIONS **6** CITATIONS
SEE PROFILE

Karina S. Machado
Universidade Federal do Rio Grande (FURG)
59 PUBLICATIONS **146** CITATIONS
SEE PROFILE

Eduardo Nunes Borges
Universidade Federal do Rio Grande (FURG)
21 PUBLICATIONS **34** CITATIONS
SEE PROFILE

Mineração de dados em triagem de risco de saúde

Thales Vaz Maciel[1]
Vinicius Rosa Seus[1]
Karina dos Santos Machado[1]
Eduardo Nunes Borges[1]

Resumo: Com a grande quantidade de dados gerados por sistemas de informação, utilizados em diversas áreas organizacionais, a necessidade de explorar esses dados torna-se evidente, com o objetivo de transformá-los em conhecimento interessante. O objetivo deste artigo é descrever, passo a passo, a aplicação de um processo de descoberta de conhecimento em banco de dados (KDD) no domínio da triagem médica. O foco do estudo está nas fases de pré-processamento e na mineração de dados, especificamente na tarefa de classificação. Discute-se a aplicação do algoritmo C4.5, denominado no software WEKA de J48, propõe-se uma abordagem sensível a custo e apresentam-se os resultados obtidos.

Palavras-chave: Descoberta de conhecimento em bases de dados. Mineração de dados. Triagem médica.

Abstract: *Large amount of data have been generated by information systems, applied in many organizational areas. Because of this it is necessary to explore such data in order to transform them into interesting knowledge. This paper aims at describing step by step, the application of knowledge discovery in databases (KDD) in the domain of medical triage. The study is focused on the pre-processing and data mining steps, specifically in a classification task. We discuss the application of the C4.5 algorithm, named J48 in the WEKA software, propose a cost-sensitive approach and present the obtained results.*

Keywords: *Data mining. Medical triage. Knowledge discovery from data.*

1 Introdução

De acordo com Han e Kamber [1], é possível observar um rápido crescimento na quantidade de dados coletada nas organizações. A descoberta de conhecimento em bases de dados, ou *knowledge discovery on databases* (KDD), consiste na atividade de transformar grandes quantidades de dados em informação que possa ser usada na prática, ou que produza relevância em determinada área de conhecimento [2].

Entende-se que o problema a ser tratado em KDD é o de transformar um conjunto de dados em um modelo utilizável, com determinado propósito, e que possibilite o entendimento humano. Um exemplo seria a possibilidade de predizer a ocorrência de uma importante situação em determinado domínio de conhecimento.

No caso do domínio da saúde, de acordo com Shama e Mansotra [4], essa é uma das áreas de maior importância para a aplicação de mineração de dados, podendo potencializar o auxílio no controle de infecções, diagnósticos e tratamento de várias doenças, além de gestão de recursos da saúde, gestão hospitalar e administração da saúde pública, por exemplo. Dentro desse contexto, Kohn [5] explica que a mineração de dados já tem sido utilizada de forma intensa e em larga escala por muitas organizações de saúde, o que torna essa ciência cada vez mais popular, senão essencial.

No mundo real, existem protocolos formais de atendimento que pautam as decisões humanas de triagem de risco, dentre os quais se destaca o Sistema de Triagem de Manchester (STM). O STM foi criado a partir dos estudos do Grupo de Triagem de Manchester (GTM), baseados na necessidade de enfermeiros e médicos

[1] Programa de Pós-Graduação em Engenharia de Computação, Furg, *Campus* Carreiros – Av. Itália, Km. 8 – Rio Grande (RS) – Brasil.
{thales.maciel,viniciuseus,karina.machado,eduardoborges}@furg.br

http://dx.doi.org/10.5335/rbca.2015.4651

obterem um consenso, fundamentado em evidências científicas que facilitem o processo de priorização de atendimento aos pacientes [8].

Esse sistema foi utilizado pela primeira vez em 1997, na cidade de Manchester, na Inglaterra [8]. Depois disso, outros hospitais da Europa passaram a utilizar o SMT, como, por exemplo, dois hospitais portugueses, que, após o primeiro uso de sistema, resolveram criar o seu próprio protocolo, chamado de Grupo Português de Triagem (GPT).

No Brasil, a organização dos sistemas de urgência foi dada por meio do ato de regulamento técnico dos sistemas de urgência. Dessa forma, a partir desse regulamento e impulsionadas pela Política Nacional de Humanização (PNH), foram realizadas capacitações sobre o protocolo de Manchester, ministradas pelo GPT, com o objetivo de implantar o protocolo no Brasil, sendo, assim, criado, em 2010, o Grupo Brasileiro de Classificação de Risco. A partir de 2010, todo o estado do Brasil já contava com o sistema de classificação de risco para o acolhimento de usuários de urgências e emergências.

Diante disso, o objetivo deste trabalho consiste em obter, por meio de um processo de descoberta de conhecimento, um modelo de classificação que possa ser aplicado no auxílio à triagem de risco de vida, na medicina. Entende-se que atividades de triagem têm o objetivo de identificar o risco de vida em pacientes mediante a análise de seus sinais vitais.

A partir desta introdução, o presente artigo está organizado em quatro seções. A seção 2 apresenta os materiais utilizados e métodos aplicados no estudo, o que inclui uma descrição detalhada do conjunto de dados e informações sobre as atividades de pré-processamento que foram realizadas. Resultados são apresentados na seção 3, onde as atividades de classificação e os modelos obtidos são descritos e discutidos. Por fim, a seção 4 conclui o estudo e apresenta possibilidades de continuidade em trabalhos futuros.

2 Materiais e métodos

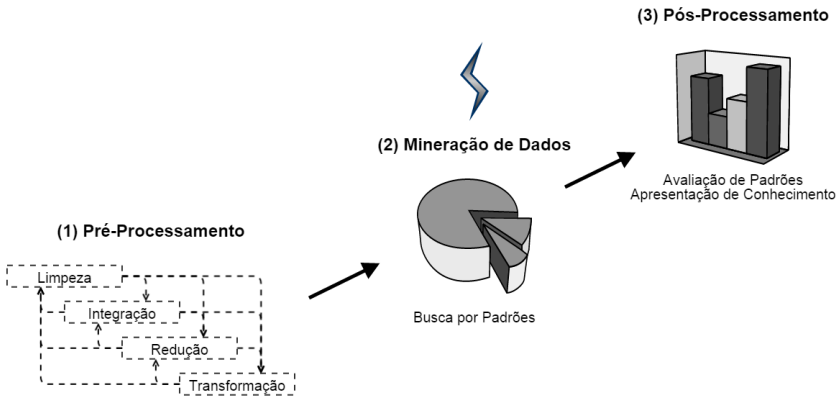
Para execução deste estudo, foi aplicado um processo de KDD (subseção 2.1) a partir do uso da ferramenta Weka, descrita em detalhe na subseção 2.2. A tarefa de mineração utilizada foi a Classificação, explicada na subseção 2.3. Foi utilizado um conjunto de dados referente à triagem de risco de vida em uma Unidade de Pronto Atendimento (UPA) do Sistema Único de Saúde (SUS), descrito detalhadamente na subseção 2.4. Os dados foram pré-processados a partir das técnicas relatadas na subseção 2.5, onde também é apresentada a problemática do desbalanceamento ocorrente no conjunto de dados em estudo.

2.1 KDD e mineração de dados

Segundo Han e Kamber [1] muitos pesquisadores tratam o termo "mineração de dados" como um sinônimo de KDD, porém, trata-se de uma etapa desse processo. O processo de KDD compreende três principais etapas: pré-processamento, mineração de dados e pós-processamento [2]. No pré-processamento, os dados, conforme disponibilizados, são preparados para as etapas seguintes. Ocorre a conversão dos dados, em uma forma bruta, para um formato mais apropriado, por meio da limpeza e remoção de dados ruidosos, por exemplo. Subsequentemente, ocorre a aplicação de técnicas de mineração de dados, com a possibilidade da execução de diversos tipos de algoritmos sobre um conjunto de dados, com o objetivo de produzir um modelo preditivo ou descritivo. No pós-processamento, o modelo obtido é analisado, avaliado e apresentado de maneira compreensível por humanos [3].

A Figura 1 representa a etapa de mineração de dados dentro do contexto do processo de descoberta de conhecimento.

Figura 1: Mineração de dados no contexto do processo de descoberta de conhecimento.



2.2 WEKA

A etapa de mineração de dados foi realizada com a utilização do software Waikato Environment for Knowledge Analysis (WEKA) [6], projeto desenvolvido pela Universidade de Waikato (NZ) na forma de uma coleção de algoritmos de aprendizado de máquina para a realização de tarefas de mineração de dados. Esses algoritmos podem ser executados diretamente sobre um conjunto de dados, ou utilizados dentro do contexto de aplicações escritas em linguagem Java, a mesma linguagem utilizada na implementação do sistema e sua Application Programming Interface (API).

O WEKA contém ferramentas para pré-processamento, tarefas de mineração em classificação, regressão, análise de agrupamentos (*clustering*), associação e, ainda, mecanismos para visualização de resultados na forma de dados estatísticos e artefatos gráficos, além da possibilidade de ser utilizado no desenvolvimento de novas abordagens ao aprendizado de máquina, considerando que todo pacote de software é *open source* (código fonte aberto) e disponibilizado sob a GNU General Public License.

2.3 Tarefa de classificação

A tarefa de mineração de dados utilizada foi a de classificação, que consiste na construção de um modelo para a predição de dados categóricos [1]. Inicialmente, foi aplicado um algoritmo baseado em árvores de decisão, o C4.5 [7], disponível no software WEKA sob a implementação J48. Também foi observada a aplicação de sensibilidade a custo no algoritmo de classificação descrito.

O método de aprendizado por classificação emprega um conjunto de instâncias já classificadas, as quais são utilizadas como base para inferir um modelo capaz de classificar novas instâncias. É comumente referido como aprendizado supervisionado, devido ao método de funcionamento, em que a atuação do algoritmo é supervisionada por meio do prévio conhecimento do resultado da classificação para cada instância utilizada no treinamento [3].

O sucesso na atividade pode ser medido, de forma objetiva, pela proporção de acertos sobre um conjunto de dados de teste. Contudo, muitas aplicações práticas de mineração de dados seguem critérios ainda subjetivos em relação à aceitação de um modelo no mundo real. O objetivo dessa tarefa é ser capaz de, após o treinamento com instâncias conhecidas, determinar a classe de instâncias novas.

2.3.1 Árvores de decisão

O algoritmo de árvores de decisão foi escolhido porque, segundo Freitas et al. [11], para pesquisa interdisciplinares, é muito importante a interpretação dos modelos obtidos por parte do especialista da área, o que, a partir da árvore de decisão, é muito mais fácil e produz mais conhecimento do que um modelo caixa preta como Support Vector Machines (SVM), por exemplo.

Uma série de melhoramentos ao algoritmo ID3 culminou em um novo sistema de indução de árvores de decisão chamado C4.5. Esses melhoramentos incluem métodos para o tratamento de atributos numéricos, valores ausentes, dados ruidosos e a geração de regras a partir de uma árvore [3].

O algoritmo C4.5 combina duas estratégias em seu funcionamento: a estratégia da divisão e conquista para aprendizado de árvores de decisão com a regra de cobertura sequencial para o aprendizado de regras [9]. A aproximação de divisão e conquista seleciona um atributo para ser nó raiz da árvore de decisão e divide a árvore pelos ramos obtidos para cada possível valor do atributo. O processo continua recursivamente para cada ramo, utilizando somente as instâncias que atingem cada ramo, respectivamente [3].

A estratégia de cobertura sequencial consiste em construir regras, sendo cada regra derivada de um ramo da árvore de decisão, abrangendo, assim, a maioria dos casos na base de dados. Instâncias que cada regra cobre são descartadas das recursões posteriores, e o algoritmo continua criando regras para instâncias remanescentes até que não haja mais nenhum caso sem classificação.

Embora essa abordagem seja mais lenta que outras, é mais segura [9], considerando que existe o processo de poda, capaz de eliminar classificações avaliadas como irrelevantes. O algoritmo C4.5 apresenta, também, uma abordagem de otimização sobre o processo de criação de janelas de amostragem, que funciona como uma técnica de amostragem eficiente para geração de árvores mais corretas. Nela, um subconjunto das instâncias de treinamento, com amostragem proporcional a cada classe (uma janela), é selecionado aleatoriamente e uma árvore de decisão é construída a partir dela. Essa árvore é utilizada para classificar exemplos que não foram incluídos na janela, geralmente apresentando novas instâncias que são classificadas incorretamente [10] e adicionando-as à janela inicial.

Uma segunda árvore, construída a partir desse conjunto maior, é testada sobre os exemplos remanescentes. Esse ciclo é repetido até que a árvore construída a partir de cada janela classifique corretamente os exemplos de treinamento que estão fora da janela. Geralmente, cada janela termina contendo apenas uma fração dos exemplos de treinamento.

Com a manutenção da aleatoriedade no processo de seleção da janela inicial, a geração de diversas árvores cria janelas iniciais diferentes e com exceções distintas. Esse potencial de múltiplas árvores provê a base para duas características: a seleção da árvore com a menor previsão de erro com a geração de regras a partir de todas as demais e a construção de um único classificador a partir de todas as regras disponíveis.

2.3.2 Sensibilidade a custo

Na ferramenta WEKA, a utilização da sensibilidade a custo consiste na aplicação de um metaclassificador em conjunto com um classificador base, tornando-o sensível a valores de custos que são manipuláveis [6].

Relaciona-se, fortemente, o desbalanceamento entre classes com a abordagem de aprendizado sensível ao custo, em classificação [1]. Entende-se, por exemplo, que, em diagnóstico médico, um caso de falso negativo seja deveras mais custoso do que um caso de falso positivo. Isso é explorado de forma mais abrangente na seção 3.

2.4 O conjunto de dados

O conjunto de dados utilizado neste estudo é referente a dados de sinais vitais que profissionais de medicina e enfermagem consideram, a fim de determinarem o grau de risco de vida em pacientes durante a atividade de triagem. Também está presente a categoria de risco indicada em cada instância.

Esses dados foram oriundos da projeção do banco de dados inerente ao sistema de informação utilizado em uma UPA, e o conjunto é composto por 21.821 registros de atividades de triagem individuais. Os atributos preditivos do conjunto de dados e suas respectivas descrições são apresentados na Tabela 1.

A determinação do valor de risco é o principal objetivo da triagem, sendo esse o atributo alvo dos modelos a serem inferidos. Esse campo possui variação de valores de 1 a 4, que representam, respectivamente, as classificações de risco Eletiva, Baixo, Médio e Alto.

2.5 O pré-processamento

Pré-processamento de dados é uma etapa importante no processo de mineração de dados e descoberta de conhecimento, em que as anomalias e inconsistências de dados são detectadas e corrigidas [1]. Durante o pré-processamento, limpeza de dados, integração, redução, transformação e outras medidas possíveis são executadas sobre os dados.

Tabela 1: Descrição dos atributos do conjunto de dados utilizados na pesquisa.

Atributo	Descrição
Id	Identificação do tratamento
Pas	Pressão arterial sistólica
Pad	Pressão arterial diastólica
Fc	Frequência cardíaca
Fr	Frequência respiratória
Temp	Temperatura
Peso	Peso
spo2	Pressão do oxigênio arterial
Ao	Abertura ocular
Mrm	Melhor resposta motora
Mrv	Melhor resposta verbal
Risco	Risco de morte: 1, 2, 3, 4
datahora	Data e horário da triagem
usuario	Profissional atuante na triagem

O conjunto de dados é apresentado na forma de um arquivo de valores separados por vírgulas (CSV), em que os valores estão entre aspas e as informações de cabeçalho não estão presentes. Inicialmente, observando-se os dados, foi possível verificar que os valores de vários atributos em algumas instâncias mantinham-se em zero.

No domínio de conhecimento em medicina, é entendido que um valor zero para qualquer sinal vital é uma indicação de ausência de valor válido ou de morte do paciente. Considerou-se que, nesse caso, o registro de morte do paciente não é uma possibilidade, devido ao processo do domínio de negócio. Portanto, o primeiro passo da limpeza de dados foi a conversão de valores iguais a zero em valores nulos, considerando que zero não é a representação mais adequada para eles, mas sim a sua não existência.

Além disso, notou-se que todas as instâncias têm registrados os mesmos valores para os campos *ao*, *mrm* e *mrv*, respectivamente, inteiros de 4, 5 e 6 (os mais altos níveis para cada indicador GCS). Isso indica que esses três campos de dados foram desconsiderados quando da verificação de sinais vitais durante a triagem, o que implica, diretamente, em tais dados não produzirem relevância estatística ou no cálculo das métricas de entropia, ou ganho de informação para fins de classificação em mineração de dados. Devido à irrelevância considerável, a remoção desses três campos foi realizada como parte da limpeza dos dados do conjunto.

Os campos denominados *id, datahora e usuario* também foram removidos do conjunto de dados a ser analisado, uma vez que foram considerados irrelevantes para efeitos de análise de risco de vida em triagem. Entende-se que, caso mantidos, poderiam gerar ramos de uma árvore de decisão que não produziriam qualquer relevância para o modelo, resultando em confusão e *overfitting*, uma situação indesejada em aprendizagem de máquina, quando um modelo é adequado fortemente a um conjunto de dados em específico apenas, e não se aplica na prática por esse motivo.

O conjunto de dados continha, inicialmente, 21.821 instâncias. Entretanto, 10.499 armazenavam dados nulos e foram descartadas do experimento. Trezentos e vinte e cinco registros continham dados discrepantes, ou seja, apresentavam valores inaceitáveis, extremamente maiores ou menores que valores considerados para humanos, e também foram eliminados. Assim, manteve-se 10.997 instâncias no conjunto de dados utilizado como entrada neste estudo.

Além da limpeza dos dados, para aplicar a tarefa de classificação, foi necessário converter o tipo de dado do campo risco para o tipo categórico, pois os valores estavam apresentados como numéricos e representam, na verdade, rótulos nominais para graus de risco de morte.

Para o subconjunto selecionado, há uma diferença considerável no total de casos que pertencem a cada classe de *risco* (atributo-alvo). Por exemplo, entre os pacientes que foram atribuídos à classificação de baixo risco e os que têm sido atribuídos à etiqueta de alto risco, existe uma diferença de, aproximadamente, 150 para 1 (Tabela 2).

Tabela 2: Descrição dos atributos do conjunto de dados utilizados na pesquisa.

Grau de risco	Total de instâncias
Eletivo	3802
Baixo	6173
Médio	981
Alto	41

Mesmo com essa diferença entre o total de instâncias em cada classe do atributo alvo, nenhuma técnica de reamostragem foi aplicada ao conjunto de dados estudado durante a fase de pré-processamento. Essa decisão ocorreu em detrimento da natureza da técnica de classificação aplicada neste estudo, chamada de aprendizagem sensível a custo [3]. A descrição do experimento e os resultados são apresentados na próxima seção.

2.6 Sistema de Triagem de Manchester

De acordo com [8], o GTM utiliza uma metodologia cuja tomada de decisão baseia-se em prioridades clínicas e não em diagnósticos médicos ou de enfermagem. Com a idealização do protocolo, o GTM tinha como objetivos o desenvolvimento de terminologias e definições comuns a todos os departamentos de emergência, bem como a criação de um programa capaz de capacitar os profissionais responsáveis para sua operação, além de um guia de auditoria para avaliar a aplicação do sistema.

Depois de avaliar outros sistemas de classificação de risco, o GTM definiu uma especificação da metodologia do STM, com a utilização de critérios de uma lista de 52 condições pré-definidas. Com essas predefinições sobre a condição do paciente, são utilizados discriminadores gerais e específicos.

Discriminadores são características que diferenciam o estado dos pacientes, para que, dessa forma, possam ser alocados em uma das cinco prioridades clínicas. Esse processo pode ser exemplificado da seguinte forma: a queixa do paciente leva a um fluxograma de apresentação composto de discriminadores, e as respostas positivas ou negativas a esses discriminadores levam a uma prioridade clínica definida por cores, que indica a gravidade e o tempo máximo que o paciente pode esperar por atendimento.

3 Resultados e discussão

A tarefa de classificação aplicada neste trabalho tem por objetivo determinar o risco que foi atribuído a cada instância de triagem, com base nos demais atributos preditivos, computacionalmente. O algoritmo de árvore de decisão J48 foi aplicado com os parâmetros número mínimo de instâncias por nó igual a 2 e fator de confiança mínima utilizada em podas igual a 0,25. O experimento foi configurado para testar o modelo obtido com uma abordagem de validação cruzada com 10 partições. O resultado da avaliação do modelo gerado, considerando a acurácia, foi de 59,33%.

A acurácia de um modelo é considerada, geralmente, como o principal indicador de sucesso da classificação e, para este experimento, a acurácia apresentada não teve um valor muito alto. Contudo, isso não foi visto como um problema para este estudo, pois a acurácia não foi considerada como o único critério de sucesso.

Verificou-se, na matriz de confusão (Tabela 3), resultante do teste do modelo gerado, que, dentre as ocorrências de erros de classificação, houve predições de risco para níveis mais altos e, também, mais baixos do que eram, na realidade. Além disso, observou-se que o modelo gerado não possibilitou a identificação de casos quaisquer da classe correspondente ao nível de risco alto, embora esteja explícita a existência de casos do tipo. Na matriz abaixo, os números de acertos na classificação podem ser observados na diagonal principal, enquanto os erros podem ser verificados fora da diagonal principal (Tabela 3).

Tabela 3: Matriz de confusão obtida com a aplicação do algoritmo J48 no conjunto de dados.

A	B	C	D	Classe
650	3127	25	0	A = Eletivo
438	5626	109	0	B = Baixo
34	698	249	0	C = Médio
229	10	0	0	D = Alto

Considerando a natureza da triagem de risco de vida, entende-se que nenhuma das características, inerentes ao modelo gerado, é aceitável. Isso acontece pela conveniência prática do domínio do negócio, em que se faz admissível que um caso de determinado nível de risco seja classificado como maior do que realmente é, enquanto o inverso não é aceitável [1].

Exemplos disso, na prática, seriam eventuais casos de risco baixo sendo triados como risco médio ou eventuais casos de risco médio sendo triados como risco alto, os quais podem ser aceitáveis, enquanto o contrário não pode ser permitido. Além disso, o fato de que esse modelo, obtido inicialmente, não atribuiu o mais alto nível de risco a qualquer instância o torna absolutamente inadequado para uso na prática.

Mecanismos de classificação em mineração de dados trabalham com a análise do custo em suas operações de tomada de decisão, e é possível fazer que essas operações sejam sensíveis a seus custos [3]. Em aplicação, isso significa que os valores de custo de classificação podem ser manipulados para induzir um algoritmo a considerar determinadas possibilidades de classificação.

No presente estudo, verificou-se que essa característica foi causada pela anteriormente referida diferença entre o número de instâncias pertencentes a cada classe, em que classes com o maior número de instâncias de treinamento apresentam menor custo de predição do que classes com menor número de instâncias de treinamento. Isso levou o algoritmo de classificação a propor classificações apenas para classes cuja quantidade de instâncias correspondentes era majoritária no conjunto de dados, especificamente, riscos eletivo e baixo, mas descartando, estatisticamente, classificações como riscos médio e alto.

Diante dos resultados não promissores apresentados a uma abordagem simples de classificação com o algoritmo J48, foi considerada uma nova, em que, por meio de um metaclassificador sensível a custo em associação com o algoritmo J48, é possível definir se a atividade será realizada por meio de avaliação sensível a custo ou aprendizado sensível a custo [6], sendo esta utilizada neste estudo.

Tal possibilidade permite que os profissionais definam uma matriz de custo, ou seja, artefato que tem a capacidade de definir, a certas áreas de uma matriz de confusão, um custo mais elevado do que o padrão em condições ordinárias. Isso possibilita que um algoritmo de classificação tente evitar a classificação de ocorrências nas áreas determinadas. Entende-se que tal medida tem o potencial de aumentar a precisão e/ou a relevância do modelo no domínio do negócio.

Utilizando o algoritmo J48 com o metaclassificador sensível a custo, para proporcionar aprendizagem sensível a custo com uma matriz de custo, conforme demonstrado na Tabela 4, o percentual de acurácia na classificação caiu para 56,56%, sendo a respectiva matriz de confusão apresentada na Tabela 5.

Tabela 4: Matriz de custo utilizada pelo algoritmo J48 sensível a custo.

0	38	0	0
38	0	0	0
84	84	0	0
99	99	99	0

Tabela 5: Matriz de confusão obtida com a aplicação do algoritmo J48 sensível a custo.

A	B	C	D	Classe
662	2979	157	4	A = Eletivo
456	5068	639	10	B = Baixo
37	453	480	11	C = Médio
2	13	16	10	D = Alto

Ao analisar as árvores de decisão resultantes de cada um dos experimentos descritos, foi observado que, em oposição ao modelo gerado pela utilização do algoritmo J48 sem sensibilidade a custo, em que nenhuma instância de triagem foi classificada como alto risco, o modelo gerado com a capacidade de sensibilidade a custo foi capaz de possibilitar a ocorrência de classificações em todas as classes.

A matriz de confusão disposta na Tabela 5 também evidencia esse resultado. Um exemplo disso pode ser notado bem ao início da árvore de decisão gerada pela utilização do modelo, na qual supostas triagens de pacientes com os sinais vitais de pressão arterial sistólica entre 128 mmHg e 169 mmHg, temperatura menor ou igual a 37,1 ºC, oxigenação entre 84 SpO2 e 88 SpO2 e frequência cardíaca maior do que 146 BPM seriam classificadas como de alto risco.

Devido à dimensão da árvore gerada pelo experimento de classificação sensível a custo, a ilustração bidimensional de sua estrutura foi particionada em seções mais apropriadas para visualização. As Figuras 2, 3, 4, 5, 6, 7, 8 e 9 ilustram parcialmente, cada uma, trechos da árvore de decisão obtida. Os marcadores no formato de estrela de dezesseis pontas com um número no centro indicam o ponto de continuação da estrutura em outra figura.

Figura 2: Ilustração do tipo de árvore de decisão gerada nos resultados.

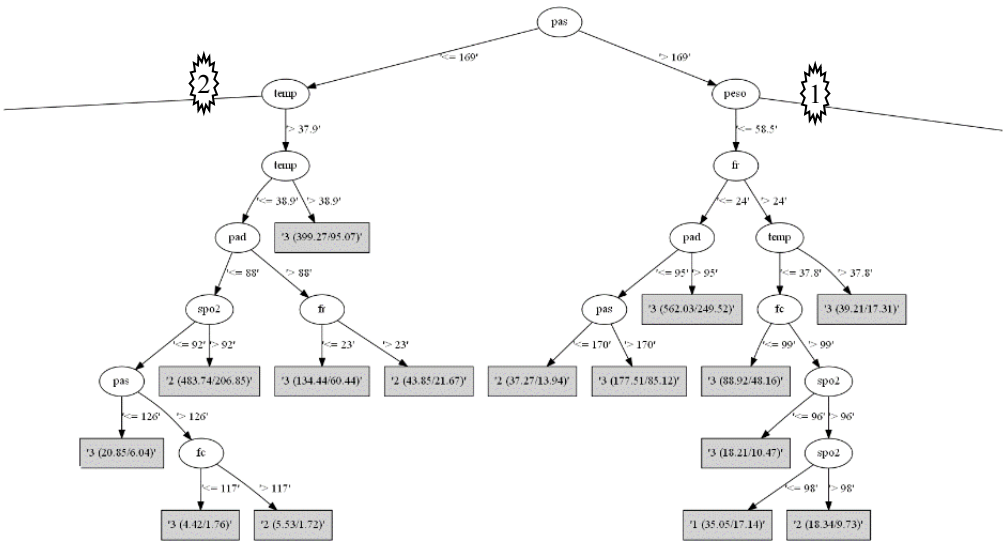


Figura 3: Ilustração parcial da árvore de decisão gerada nos resultados.

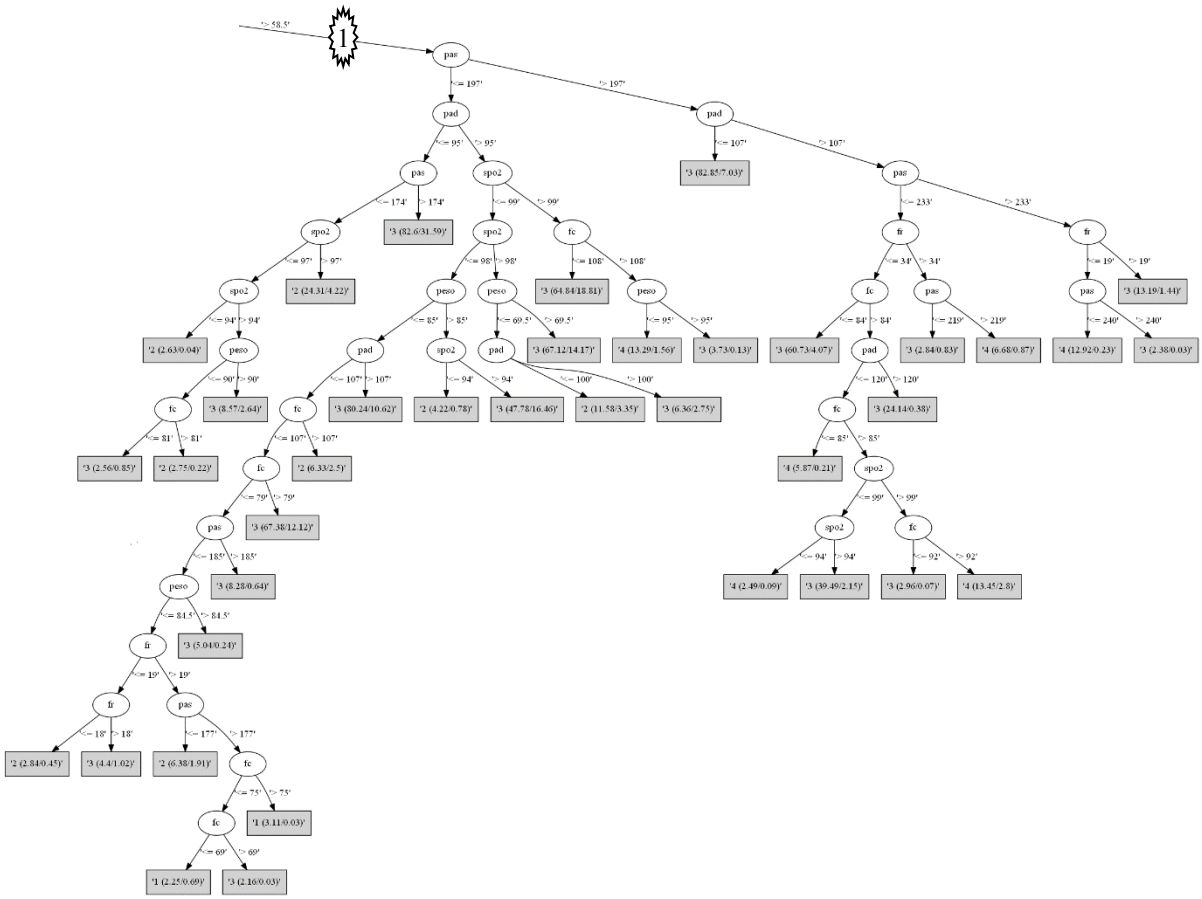


Figura 4: Ilustração parcial da árvore de decisão gerada nos resultados.

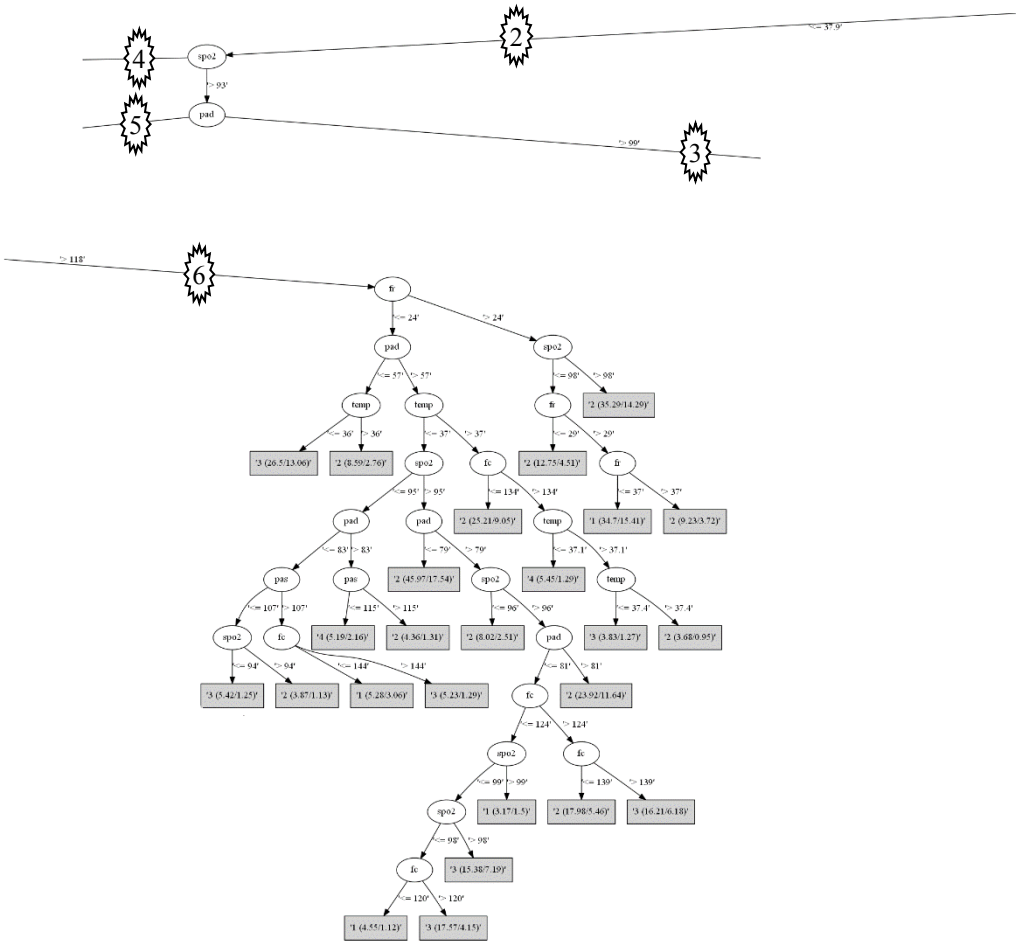


Figura 5: Ilustração parcial da árvore de decisão gerada nos resultados.

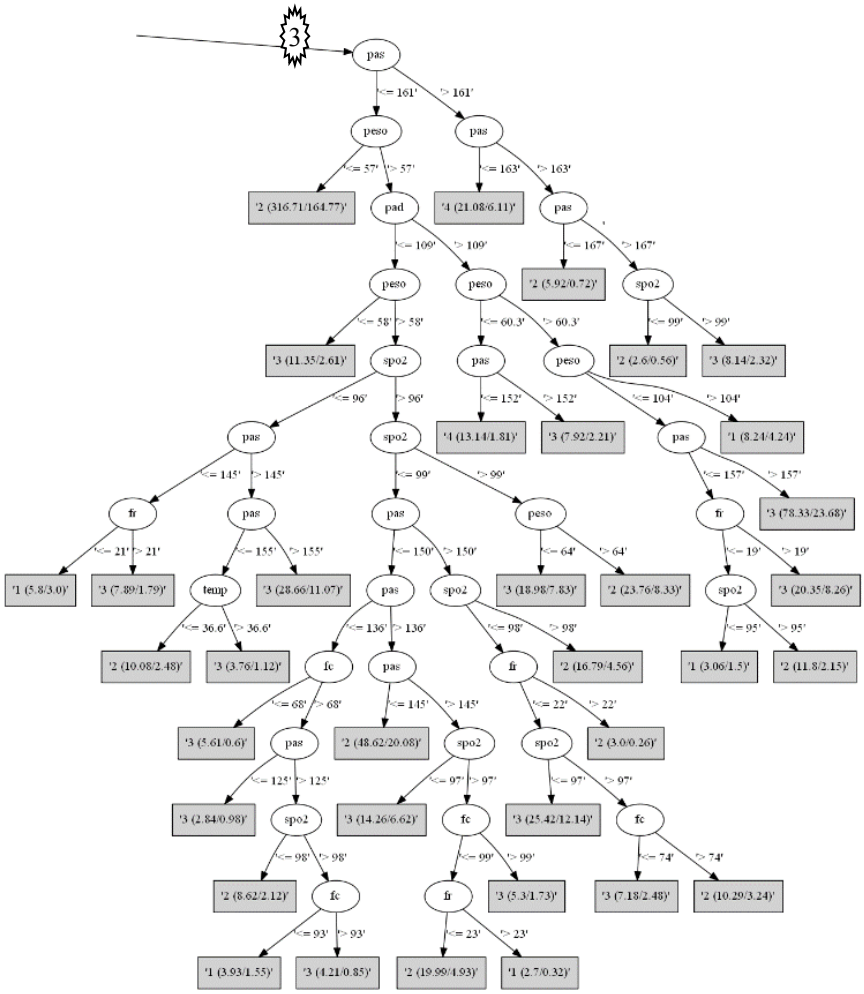


Figura 6: Ilustração parcial da árvore de decisão gerada nos resultados.

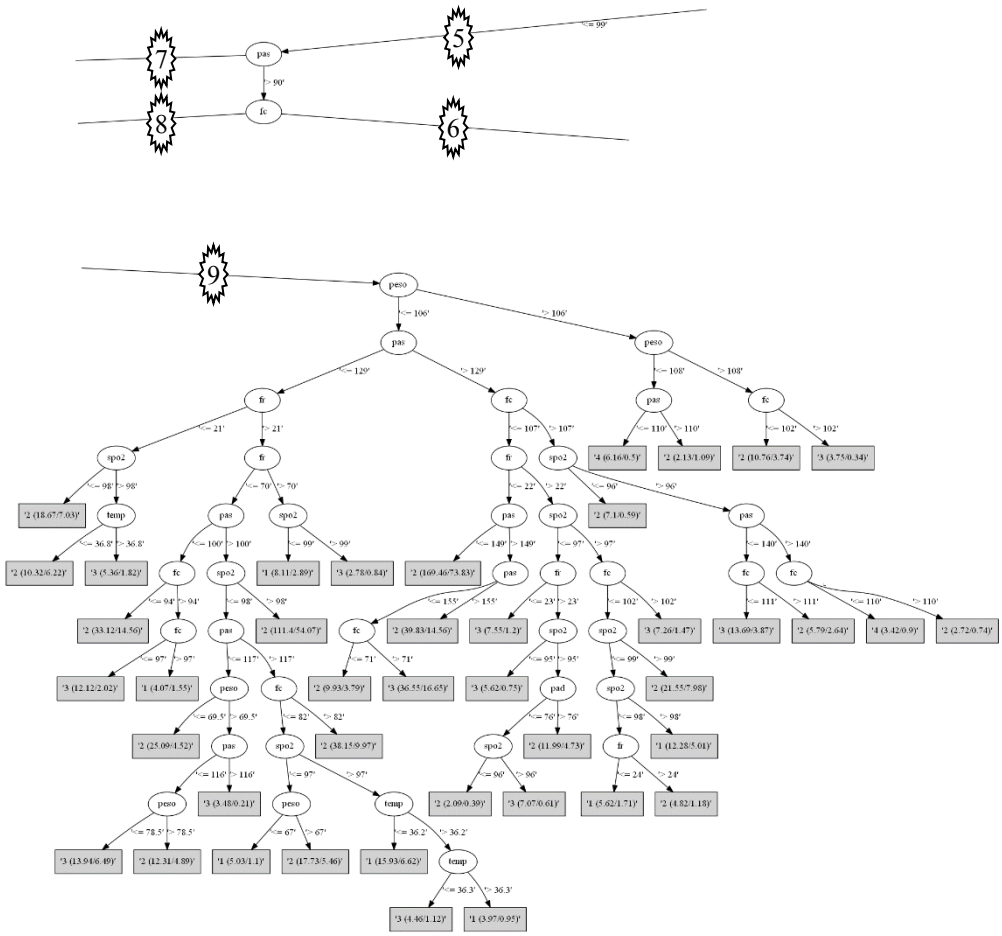


Figura 7: Ilustração parcial da árvore de decisão gerada nos resultados.

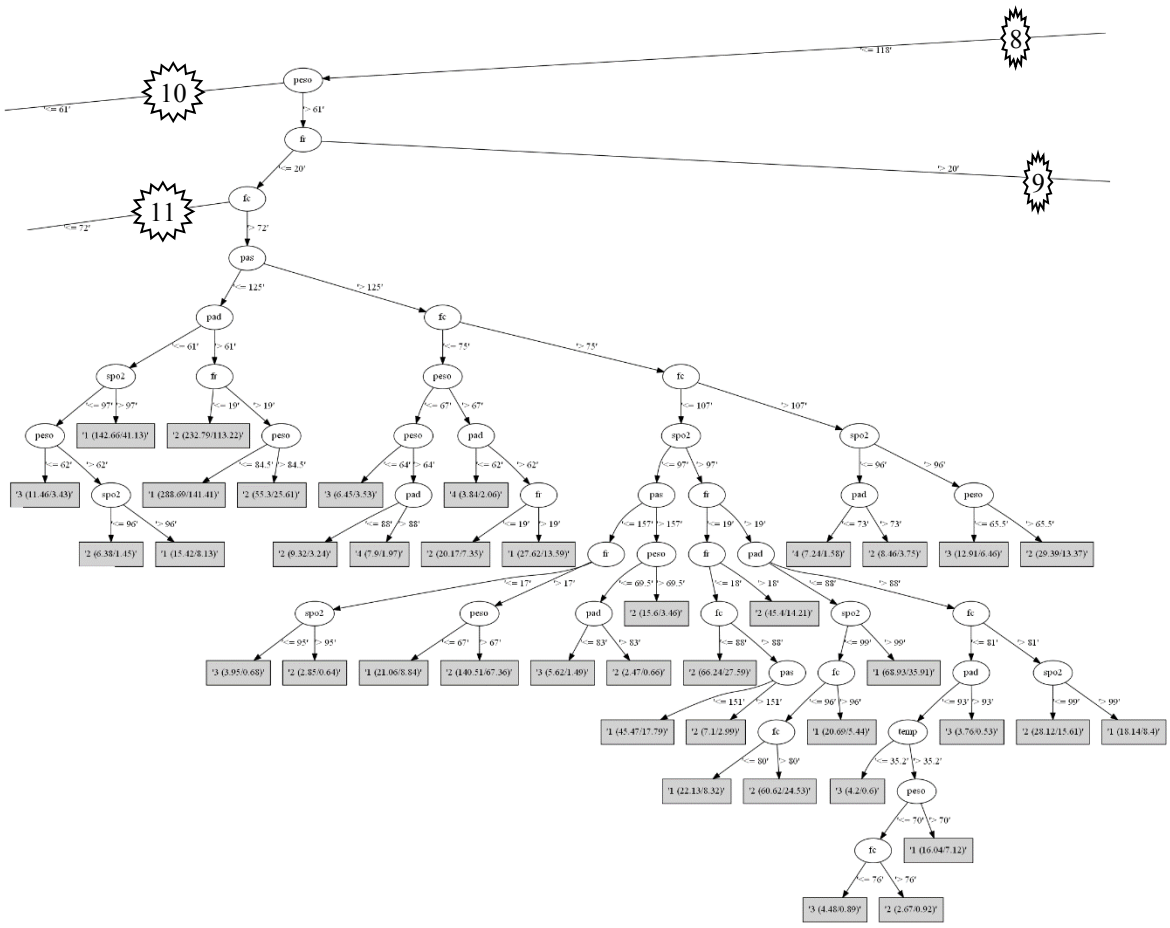


Figura 8: Ilustração parcial da árvore de decisão gerada nos resultados.

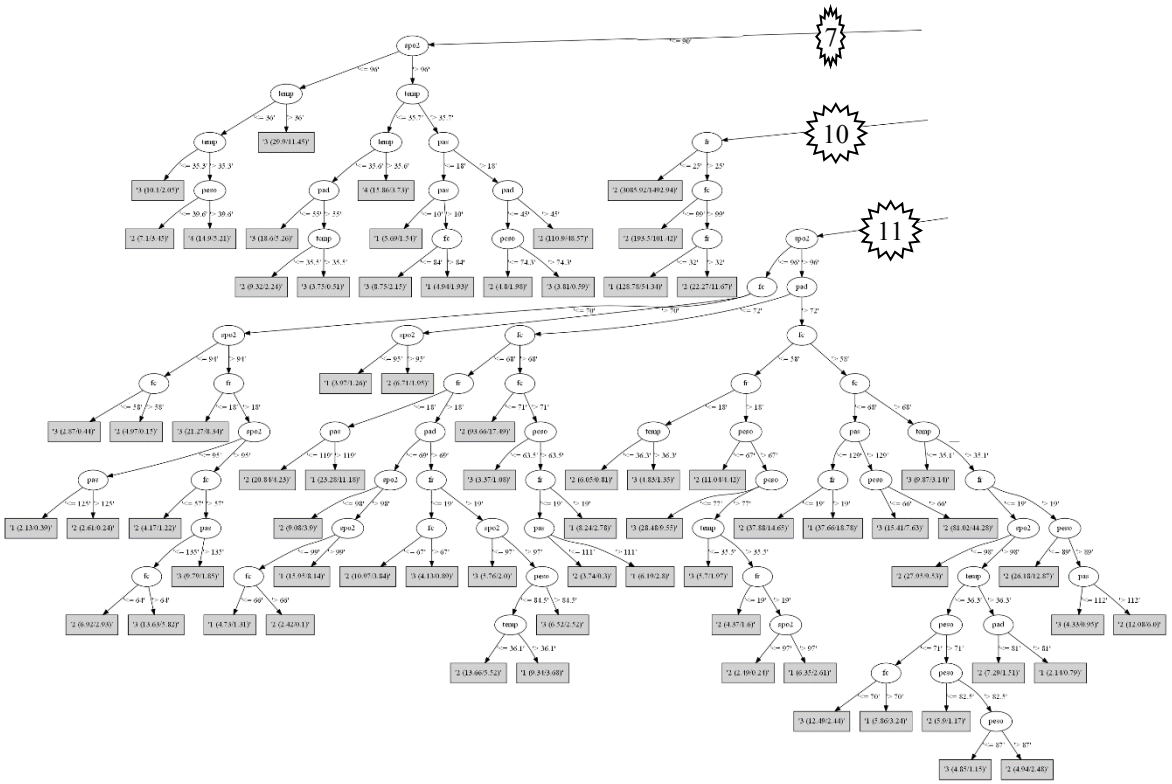
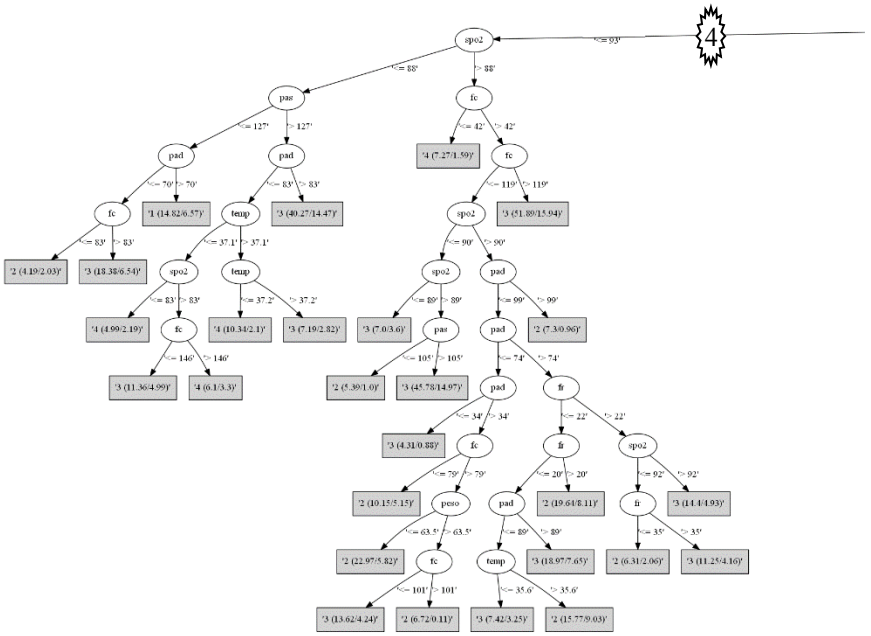


Figura 9: Ilustração parcial da árvore de decisão gerada nos resultados.



4 Conclusão

Mesmo que, associando o algoritmo de classificação J48 com um metaclassificador para aprendizado sensível a custo, a atividade tenha sido desempenhada com uma diferença de 2,77% na acurácia de classificação, o conhecimento sobre o domínio do negócio inerente contribuiu para a descoberta de que, em determinados casos (e neste específico), essa não deve ser considerada a única métrica de sucesso.

As matrizes de confusão apresentadas em ambos os experimentos mostraram que, por meio da associação de aprendizado sensível a custo ao classificador, foi possível obter um modelo gerado por aprendizado de máquina que seja aplicável no domínio de negócio de triagem médica e passível de avaliação na prática.

Portanto, com base no estudo realizado, conclui-se que a aplicação de um processo de KDD utilizando classificação com árvores de decisão a dados de triagem pode ser muito importante para o entendimento de que tipo de característica (atributo) é mais determinante para cada uma das classes de risco, assim como os intervalos de valores de cada atributo. Esse conhecimento seria muito difícil de ser obtido a partir somente de análises visuais e consultas simples a esses dados de triagem.

Referências

[1] Han, Jiawei; Kamber, Micheline. Data Mining. Concepts and Techniques. Second edition. The Morgan Kaufmann Series in Data Management Systems. Elsevier Inc., 2006.

[2] Tan, Pang-Ning; Steinbach, Michael; Kumar, Vipin. Introduction to Data Mining 2005.

[3] Witten, Ian H.; Frank, Eibe; Hall, Mark A. Data Mining. Pratical Machine Learning Tools and aTechniques. Third edition. The Morgan Kaufmann Series in Data Management Systems. Elsevier Inc., 2011.

[4] Shama A. Mansotra V. Emerging applications of data mining for healthcare management - A critical review. 2014 International Conference on Computing for Sustainable Global Development (INDIACom), New Dehli. pages 377-382. 2014.

[5] Kohn, H C and Tan G. Data Mining Applications in Healthcare. Journal of Healthcare Information Management - Vol. 19, No. 2. Pages 64-72. 2005.

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[7] Kohavi, R; Quinlan, J. R. Data mining tasks and methods: Classification: decision-tree discovery. Handbook of data mining and knowledge discovery, pages 267-276, 2002.

[8] Quinlan, J. R. C4.5: Programs for Machine Learning. 1993. Morgan Kauffman.

[9] Alvez, F. Aplicação de Técnicas de Mineração de Dados a Uma Base de Um Sistema Gerenciador de Informações para UTI. 2005. Dissertação de Mestrado. PUCPR.

[10] Enembreck, F. Um Sistema Paraconsciente para Verificação Automática de Assinaturas Manuscritas. 1999.

[11] Freitas A., Wieser D., Apweiler R. On the Importance of Comprehensible Classification Models for Protein Function Prediction. IEEE Computer Society Press. 2010.