UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIENCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA
DISSERTAÇÃO DE MESTRADO

Eduardo A. Liboredo Ferreira

Mineração de Dados Aplicada à dados Médicos

Belo Horizonte
Junho
2015

Eduardo A. Liboredo Ferreira

Mineração de Dados Aplicada à dados Médicos

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Bioinformática do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais

Orientador: Prof. Dr. Sérgio Vale Aguiar Campos

Belo Horizonte

Junho

2015

Agradecimentos

Gostaria de Agradecer aos parceiros e colaboradores deste trabalho, que cederam os dados e as tão necessárias consultorias.

Agradeço aos meus amigos e companheiro de laboratório, sem ordem específica de prioridade: Fernando, Lucas, Jerônimo, Gabriel, Paulo, Vinícios, Gustavo, Phillippe, Juliano e Disso pelas horas infindáveis de companheirismo e diversão.

Agraço ao meu orientador, Sérgio Campos e a Alessandra por terem me recebido de braços abertos no laboratório. Agradeço novamente ao Sérgio, pela orientação no trabalho desenvolvido e, principalmente, pelos puxões de orelha, honestidade e lisura com que sempre me tratou.

Agradeço aos professores do programa da bioinformática pela dedicação e qualidade das disciplinas. Agradeço à secretaria do programa, especificamente à Sheila e Paula pela eficiência e cordialidade. Agradeço à coordenação do programa, que prima pela qualidade e excelência do programa.

Agradeço, sempre que posso, à minha esposa Raquel, por me apoiar e aturar ao longo de todos esses anos que passamos juntos. Agradeço ao meu filho, por mudar a minha perspectiva do mundo. Tem sido um privilégio acompanhar seu crescimento.

“The force is with you, young Skywalker,
but you are not a Jedi yet. ”
– Darth Vader.

Índice

Lista de figuras

Lista de tabelas

Resumo

A mineração de dados médicos é um processo desafiador. A falta de grandes bases de dados e a complexidade dos dados são alguns dos desafios. Isso é especialmente verdade para doenças raras e negligenciadas. Essas bases de dados são, em geral, relativamente pequenas, largas e esparsas, configurando bases bastante difíceis de se analisar. Restrições legais, éticas e sociais devido ao status especial da medicina agravam as dificuldades de se trabalhar com dados médicos. Podemos mencionar ainda a complexidade das bases médicas, que podem ser composta por várias fontes como dados de exames laboratoriais, entrevistas e imagens. Este trabalho apresenta dois estudos de caso, com uma proposta de metodologia para se lidar com esses desafios. O primeiro estudo de caso é a analise da base de dados da doença Paracoccidioidomicose (PCM). A PCM é uma doença tipicamente brasileira causada pelo fungo *Paracoccidioides brasiliencis*. Esta doença é um importante problema de saúde publica devido ao seu poder incapacitante e a altas taxas de mortes prematuras se não tratada. Sua forma primária é a pulmonar, podendo se disseminar para outros órgãos. Ela afeta principalmente homens na faixa dos 30 a 50 anos, causando também impacto econômico por afetar indivíduos em fase produtiva. A base de dados de PCM é uma base pequena, larga e esparsa (com muitos valores não preenchidos. São discutidos métodos de análise para ajudar a entender melhor tando a doença quanto este tipo de dados. O segundo estudo de caso é a análise de dados de eletrocardiograma. Segundo dados do governo brasileiro, doenças cardíacas são responsáveis por 30% das mortes no Brasil. O diagnóstico preciso e rápido é o primeiro passo para um tratamento eficaz. A eletrocardiografia (ECG) é um método de investigação do aparelho cardiovascular com valor diagnóstico e prognóstico, fácil realização e baixo custo, e de grande utilidade clínica. Para este estudo de caso analisamos o principal método de classificação do ECG, o código de Minnesota e testamos metodologias e técnicas para checar a viabilidade da criação de uma ferramenta automática que auxilie o médico a identificar se um exame é normal ou alterado. Para ambos os estudos de casos foram testados diversos métodos de tratamento de dados e algoritmos de mineração. Para PCM, apesar das dificuldades da analise, alguns resultados interessantes foram descobertos. Falhas no preenchimento dos prontuários, notadamente, 40% dos exames não tinham resultados sobre raio-x de tórax, um exame básico em doenças pulmonares. A classificação da forma cutânea da doença com 93% de precisão são alguns dos achados. Para o segundo estudo de caso o ECG, foi possível se classificar arritmias cardíacas com 95% de precisão, baseados no código de Minnesota tamém foi possível se classificar defeitos de condução ventricular com 92% de precisão.

Introdução

1. Introdução

1.1 Mineração de dados médicos

Mineração de dados aplicada à dados médicos é um processo desafiador. A falta de grandes bases de dados e a complexidade estrutural dessas bases são alguns dos desafios encontrados. Os protocolos de exames para diagnósticos são, em geral, complexos e possuem vários atributos diferentes. Exames e testes são pedidos de acordo com a experiencia pessoal do médico e disponibilidade de recursos. Muitas vezes, os pacientes não realizam os procedimentos e exames requisitados e deixam lacunas nos prontuários. As bases de dados de doenças, principalmente, provém de diversas fontes diferentes, como entrevista com o paciente, testes laboratoriais, resultados de equipamentos e exames diretos. Isso tende a produzir bases de dados altamente variadas e difíceis de serem analisadas, demandando o uso de diferentes técnicas e ferramentas para serem exploradas de maneira eficiente. Existem ainda restrições éticas, legais e sociais relativas à privacidade e a validação clínica dos achados. Os desafios da mineração de dados médicos são discutido em mais detalhes na seção a seguir.

Apesar de grandes avanços na área de informática e gerenciamento de dados da saúde no que se refere a estatísticas, gerenciamento de grandes hospitais e estudos de larga escala, há uma escassez de ferramentas analíticas para se extrair conhecimento desses dados (SHULKA et al., 2014). No entanto, a mineração de dados vem ganhando espaço na área da saúde. Ela pode ser usada pelas operadoras de planos de saúde para detectar fraudes e abusos, ajudar as organizações de saúde a tomar decisões de gerenciamento e de relação com clientes. Médicos podem identificar tratamentos eficazes e boas práticas clínicas (KOH et al., 2011). Pode ainda ajudar pesquisadores a identificar sintomas e características da doença para diagnóstico e tratamento (BREAULT et al, 2002).

Médicos tomam decisões diagnósticas e recomendam de tratamentos baseados no histórico do paciente, exames clínicos, e laboratoriais. Aplicar técnicas de mineração de dados nas bases médicas pode fornecer aos médicos ferramentas analíticas e preditivas que vão além do que é visto na superfície dos dados. Por exemplo, um médico pode ter acesso ao histórico de diagnósticos e tratamentos dados a casos

similares e seus resultados. Ferramentas preditivas podem aconselhar médicos em suas decisões diagnósticas. Apesar de existirem sistemas de apoio a decisão para diagnóstico(BRENER, 2007; GREENS, 2007), o uso de ferramentas computacionais para esse fim ainda é limitado.

A mineração de dados médicos é desafiadora e distinta de outras áreas. Para doenças raras e negligenciadas as dificuldades são agravadas. Apesar dos desafios, a mineração de dados médicos pode ser a mais recompensadora. Achar uma solução para uma pergunta médica relevante pode significar a melhoria da saúde e qualidade de vida de vários pacientes.

1.1.1 Desafios da Mineração de dados médicos

A dificuldade de se minerar dados médicos pode ser dividida em 4 esferas segundo CIOS e MOORE (2002):

- Heterogeneidade dos dados
- Questões éticas, legais e sociais
- Filosofia estatística
- Status especial da medicina.

a) Heterogeneidade

Os dados brutos provém de inúmeras fontes como exames clínicos, laboratoriais, exames baseados em imagens, entrevista com pacientes e as observações e interpretações do médico. Todos esses fatores influenciam no diagnóstico, prognóstico e tratamento do paciente. Entrevistas, prontuários e laudos médicos são particularmente desafiadores devido às diferentes maneiras de se referir a um mesmo achado clínico ou mesmo a diferentes interpretações por parte dos médicos.

b) Questões éticas, legais e sociais

Pelo fato dos dados coletados pertencerem a seres humanos existem uma série de restrições legais e éticas para seu uso. Cuidados especiais a respeito de privacidade e segurança dos dados tem de ser tomados.

c) Filosofia estatística

Estudos em medicina geralmente são amarrados a uma metodologia estatística, o que limita sua aplicação em bases heterogêneas. Mesmo métodos de mineração de dados, muitas vezes, não conseguem extrair informações da base de dados bruta. É necessária uma série de tratamentos e transformações para se extrair informações úteis. As particularidades de cada base de dados médicas dificulta muito a criação de uma metodologia unificada de mineração.

d) Status especial da medicina

A medicina tem status especial na ciência. Os eventos médicos são de vida ou morte e constituem uma necessidade social, um direito humano básico. A saúde do indivíduo muitas vezes depende da assistência médica e existe uma forte cobrança e atenção da mídia nesse sentido.

1.2 Mineração de Dados

Mineração de dados pode ser definida como o processo computacional de descobrimento de padrões em dados e apresentá-los de forma compreensível e útil [HASTIE et al., 2009]. O processo de descoberta de conhecimento de dados, conhecida como KDD (*knowledge discovery in databases*) envolve o uso de ferramentas computacionais complexas aplicados a esses dados para extrair informações ocultas aos métodos convencionais de analise. A quantidade de dados gerados atualmente é enorme e provém de varias fontes. O acumulo desses dados gerou a necessidade de se criar ferramentas para a análise por computador, pois à medida que o volume de informações cresce, se torna impraticável a análise manual desses dados(HASTIE et al., 2009; WITTEN et al., 2011).

A mineração de dados vem sendo aplicada em vários campos com bastante sucesso. Empresas do setor financeiro e seguradoras utilizam sistemas de apoio a decisão para calcular os riscos de empreendimentos. Empresas utilizam mineração de dados para fazer marketing personalizado, prevendo o comportamento de compra de

clientes [times2012, forbes2012]. Treinadores e organizações esportivas utilizam técnicas de mineração para otimizar treinamento e até mesmo ajudar a prever lesões em atletas.

1.3 Estudos de Caso

1.3.1 Paracoccidiodomicose – PCM

a) A doença

A paracoccidioidomicose (PCM) é uma micose sistêmica endêmica de grande interesse para os países da América Latina. Causada pelo fungo termo-dimórfico *Paracoccidioides brasiliensis,* apresenta distribuição heterogênea, havendo áreas de baixa e alta endemicidade (WANKE; AIDE, 2009). No adulto, a forma clínica predominante é a crônica, mas quando acomete crianças ou adolescentes apresenta-se na forma aguda ou subaguda. Quando não diagnosticada e tratada oportunamente, pode levar a formas disseminadas graves e letais, com rápido e progressivo envolvimento dos pulmões, tegumento, gânglios, baço, fígado e órgãos linfoides do tubo digestivo. (WANKE; AIDE, 2009; WANKE; LAZER; CAPONE, 2001)

b) Epidemiologia

Esta micose representa um importante problema de Saúde Pública devido ao seu alto potencial incapacitante e à quantidade de mortes prematuras que provoca, principalmente para segmentos sociais específicos, como os trabalhadores rurais, que além de tudo isso apresentam grandes deficiências de acesso e suporte da rede dos serviços de saúde favorecendo o diagnóstico tardio. A faixa etária mais acometida situa-se entre 30 e 50 anos de idade e mais de 90% dos casos são do sexo masculino. É infrequente abaixo dos 14 anos de idade, faixa na qual não existe predomínio de sexo.

Os indivíduos acometidos pela micose, usualmente encontram-se na fase mais produtiva da vida, sendo que a doença leva a impacto social e econômico (PANIAGO et al., 2003; SHIKANAI-YASUDA et al., 2006)

A paracoccidioidomicose é uma doença sem notificação compulsória e sem

dados precisos sobre sua incidência no Brasil. Acredita-se que a incidência anual em zonas rurais endêmicas varie de 3-4 novos casos/1.000.000 de habitantes até 1-3 novos casos/100.000 habitantes. É considerada a terceira causa de morte por doença infecciosa crônica, resultando em uma taxa de mortalidade de 1,65 casos/1.000.000 de habitantes (SHIKANAI-YASUDA et al., 2006).

Apesar de ser considerado um importante problema de Saúde Pública, dados de acompanhamentos clínicos de PCM são escassos. Ainda não existem bons parâmetros clínicos para se identificar a reincidência da doença ou estimar seu tempo de tratamento.

1.3.2 Eletrocardiograma - ECG

a) Doenças cardiovasculares no Brasil

Segundo dados oficiais do governo (www.brasil.gov.br), as doenças cardiovasculares são responsáveis por cerca de 30% de todas as mortes registradas no Brasil. Isso equivale a aproximadamente 308 mil mortes por ano. O diagnóstico preciso e rápido é o primeiro passo para um tratamento eficaz. A eletrocardiografia é um método de investigação do aparelho cardiovascular com valor diagnóstico e prognóstico, fácil realização e baixo custo, e de grande utilidade clínica. Utilizado desde a Unidade Básica de Saúde até o Centro de Tratamento Intensivo, o eletrocardiograma (ECG) é uma ferramenta básica para diversos profissionais da área da saúde. Com o custo crescente da medicina moderna se torna necessário o uso racional dos recursos, com priorização de técnicas e procedimentos de relação custo efetividade favorável, como o ECG (MACFARLANE et al., 2011; RIBEIRO et al., 2013). O eletrocardiograma é também amplamente utilizado em pesquisas populacionais, tento seu valor prognóstico sido descrito inúmeras vezes na literatura (GREENLAND et al., 2003; MACHADO et al., 2006; ZHANG; PRINEAS; EATON, 2010; ZHANG et al., 2012).

b) ELSA Brasil

O Estudo Longitudinal de Saúde do Adulto - ELSA Brasil - é uma investigação multicêntrica de coorte composta por 15 mil funcionários de seis instituições públicas

de ensino superior e pesquisa das regiões Nordeste, Sul e Sudeste do Brasil. A pesquisa tem o propósito de investigar a incidência e os fatores de risco para doenças crônicas, em particular, as cardiovasculares e o diabetes (AQUINO et al., 2012; RIBEIRO et al., 2013).

O ECG é um dos métodos realizados na linha de base do estudo ELSA-Brasil. Para uniformização dos dados foi criado de um Centro de Leitura de ECG (CL-ECG), de modo a garantir a qualidade dos registros e a uniformidade e comparabilidade da codificação (RIBEIRO et al., 2013). Os dados utilizados neste trabalho provém da primeira onda do projeto ELSA Brasil. No total foram analisados 11936 registros, já codificados pelo código de Minnesota e revisados por especialistas. Uma discussão mais detalhada sobre o código de Minnesota encontra-se na seção apropriada, Metodologia de ECG.

Apesar do código de Minnesota ser uma ferramente já estabelecida de auxílio de diagnóstico e estudos populacionais, ele é baseado em medidas morfológicas rígidas que não levam em consideração variações morfológicas intrínsecas da população, variações entre gêneros, idade ou histórico do paciente (KORS; VAN HERPEN, 2001. MACFARLANE; LATIF, 1996; MACFARLANE et al., 2000). Para o acompanhamento clínico do paciente, ou seja, fora de grandes estudos epidemiológicos, os médicos especialistas frequentemente ignoram os códigos para uma análise personalizada do exame. Para otimizar e agilizar o poder de diagnóstico do paciente é necessária a criação de uma ferramenta cujos parâmetros de auxílio à decisão levem em conta os fatores de variação intrínsecos da população e individuais.

II Objetivos

2. Objetivos

2.1 Objetivo Geral

Desenvolver e testar metodologias para mineração de dados médicos, aplicados a bases com características distintas.

Objetivos específicos:

- Testar a viabilidade de uso de técnicas de mineração de dados em bases médicas pequenas e largas
- Determinar qual algoritmo de classificação se comporta melhor para a base PCM
- Estabelecer correlações entre atributos de importância clínica de PCM
- Testar as correlações entre a recidiva da doença e seus atributos
- Aplicar técnicas de mineração de dados na base de dados de ECG do ELSA-Brasil
- testar a viabilidade de implementação de ferramenta de suporte a decisão que separe exames de ECG normais de alterados
- Testar o desempenho de diferentes algorítimos em bases de dados com características distintas

III Metodologia

3 Metodologia

3.1 Processo de descoberta de Conhecimento

Todo o processo de mineração e descoberta de conhecimento começa com a coleta dos dados. Para ser útil, os dados brutos precisam passar por transformações e pré-processamentos. A mineração de dados brutos dificilmente gera bons resultados (CIOS; MOORE, 2002; WITTEN; FRANK, 2011).

Apesar de não possuir uma fórmula fixa, o processo de descoberta de conhecimento segue alguns passos que podem ser aplicados à maioria das bases de dado (WITTEN; FRANK, 2011). Esta seção visa resumir o processo de mineração, introduzindo e explicando os conceitos básicos. Detalhes individuais serão discutidos nas seções dos estudos de caso apropriadas. A figura 1 oferece um resumo visual e uma breve explicação das etapas do processo.

Figura 1: Representação do processo de mineração de dados. No pré-processamento são executadas ações que visam a preparação da base. Processos como união das fontes e padronização de dados, amostragem, remoção de campos irrelevantes/redundantes são executados nesta fase. O processo de transformação envolve mudanças mais profundas, como processos de normalização de dados, redução de dimensionalidade. A fase de análise envolve a aplicação dos algorítimos e técnicas para se identificar padrões nos dados. Técnicas como clusterização, classificação e seleção de atributos são utilizadas de acordo com os objetivos da análise. A última etapa, avaliação e interpretação, é a etapa que irá revelar se os padrões encontrados são úteis ou não. Todos os processos são interligados e descobertas/mudanças em uma delas frequentemente geram a necessidade de repetir os passos.

3.1.1 Extraction, Transformation and Load – ETL

O processo de coleta, pré processamento e armazenamento inicial é chamada de Extraction, Transformation and Load – ETL. Essa é uma das etapas mais importantes da mineração de dados e deve ser conduzida com cuidado (WITTEN; FRANK, 2011; HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Durante o ETL dados de múltiplas fontes são formatados, padronizados e erros menores como grafia despadronizada são corrigidos. Após a criação da base inicial outros métodos de refinamento podem ser aplicados.

3.1.2 Remoção de campos irrelevantes e redundantes

A remoção de atributos redundantes ou irrelevantes reduz o ruido e melhora a qualidade da analise (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; GUYON; ELISSEEFF, 2003). Atributos irrelevantes são campos que possuem um valor muito predominante nos registros ou que não agregam informação à analise. Exemplos são atributos como endereço ou localidade em estudos onde a distribuição não é necessária ou todos os indivíduos pertencem a mesma cidade. Atributos redundantes são atributos que possuem a mesmo valor informacional. A escolha dos atributos pode ser auxiliada pela analise de Distribuição de Frequência onde se visualiza a distribuição dos valores na base (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

3.1.3 Seleção de Atributos

A seleção de atributos é usada para selecionar um subconjunto dos dados que possuem relação com atributo escolhido. Essa técnica é auxiliada por algoritmos específicos e é utilizada quando se deseja separar um fragmento da base de dados para analises posteriores (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; GUYON; ELISSEEFF, 2003).

3.1.4 Classificadores

A construção de classificadores é indicada quando se quer prever um atributo baseado nos dados disponíveis. Essa técnica é direcionada ou assistida, ou seja, é necessária a escolha de um atributo alvo para a previsão de seu possível valor. Existem

várias classes de algoritmos classificadores. Algoritmos bayesianos ou probabilísticos, algoritmos de árvores de decisão, algoritmos de regras e algoritmos preguiçosos são alguns exemplos. (WITTEN; FRANK; HALL, 2011; QUINLAN, 1993).

3.2 Bases de dados

Foram utilizadas duas base de dados brutas para as análises com características bastante diferentes, descritas abaixo.

3.2.1 Paracoccidioidomicose - PCM

A base de Paracoccidioidomicose – PCM – foi cedida pelo Centro de Treinamento e Referência de Doenças Infecto-parasitárias da UFMG (CTR-DIP-UFMG). A base original é composta por registros de primeira consulta de 227 pacientes diagnosticados com PCM em formato proprietário SPSS. A base é composta por 314 atributos, sendo 9 atributos do tipo numérico e o restante do tipo nominal (*string*). A base é um dos maiores estudos de caso de PCM do mundo, tendo sido coletado ao longo de 35 anos em região de máxima endemicidade.

3.2.2 Eletrocardiograma – ECG

A base de dados de eletrocardiograma foi cedida pelo projeto ELSA-brasil. Os dados brutos consistem em 11938 exames de eletrocardiograma armazenados uma única planilha, pertencentes à primeira onda do projeto ELSA. A planilha possui 24 campos reservados aos códigos de Minnesota e 14 campos contendo as medições e campos com identificadores que, para a análise, são irrelevantes.

3.3 Ferramentas Utilizadas

a)WEKA

Existem várias ferramentas de analise multidimensional disponíveis. Para as análises de mineração de dados foi utilizada a ferramenta WEKA (HALL et al., 2009). WEKA é um programa desenvolvido pelo grupo de *Machine learning* da Universidade de Waikato, Nova Zelândia. Desenvolvido em Java, é multiplataforma e distribuído gratuitamente sob a licença GNU *general public licence*. Possui ferramentas de pré-processamento, classificação, regressão, clusterização, regras de associação e visualização. O programa ainda pode ser chamado por programas java ou diretamente pela linha de comando (bash), o que o que facilita rodar muitos testes em série. Essas características tornam o programa bastante versátil para mineração. A versão do software utilizada neste trabalho é a 3.6.10. O WEKA pode usar arquivos CSV (comma separated values) ou o formato nativo arff (*atribute related file format*). O formato arff não somente armazena sua base de dados, mas pode conter diversas informações, como resultados, modelos de testes/classificadores além de comentários. As figuras 2 e 3 mostram a diferença de estrutura de ambos.

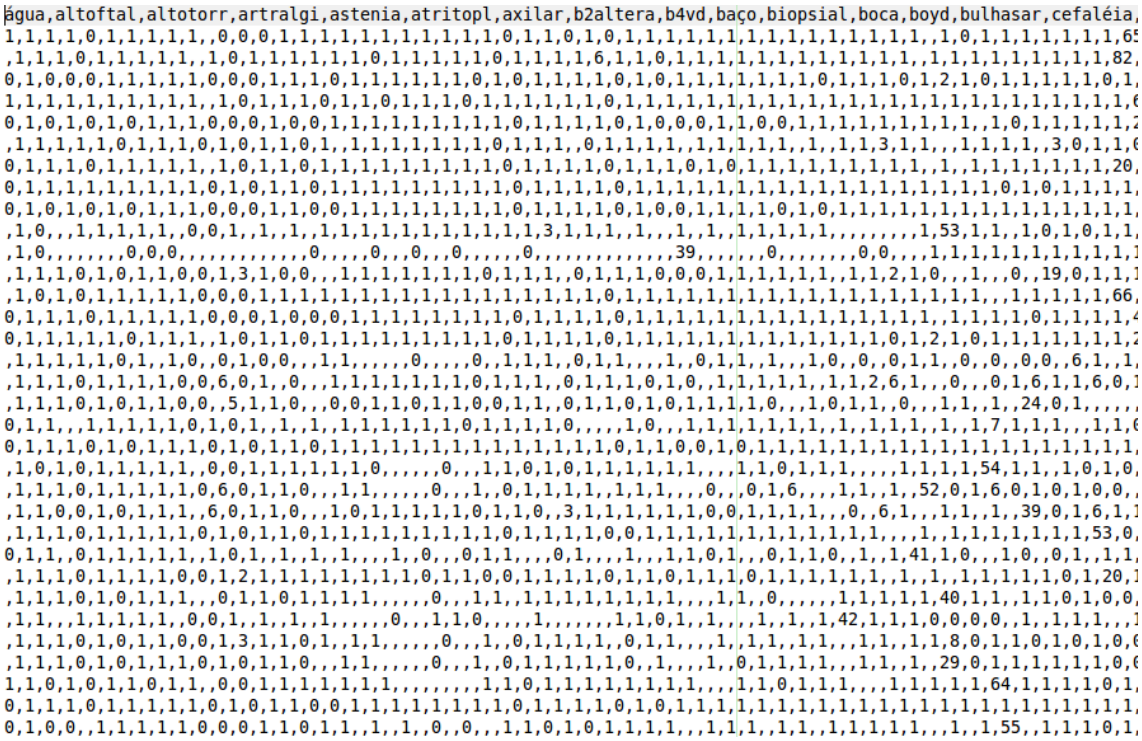

Figura 2: Estrutura de dados de um arquivo CSV. A primeira linha mostra os nomes das colunas, que se referem ao nome do atributo. Os valores são separados por vírgulas. Vírgulas em sequencia representam campos vazios

.

```
@attribute pcmpancr {0,1}
@attribute pcmpulmo {0,1}
@attribute pcmrenal {0,1}
@attribute ptlalter {0,1}
@attribute pulsmmii {0,1}
@attribute recidiva {0,1,6}
@attribute roncos {0,1}
@attribute rouquidã {0,1}
@attribute rxaltera {0,1}
@attribute sexo {0,1}
@attribute sibilos {0,1}
@attribute ssea {0,1}
@attribute ssim {0,1}
@attribute sudorese {0,1}
@attribute tabprevi {0,1}
@attribute tamanho {0,0.5,0.8,1,1.5,2,2.5,3,4,5,6,7}
@attribute tamanhof {0,1,2,3,4,6,7,8}
@attribute tevoluçã {0,1,2,3,4,5,6,7,8,9,10,11,12,15,16,18,19,20,21,24,36,42,48,60,72,84,264,336,552}
@attribute tinativa numeric
@attribute tonteira {0,1}
@attribute tosse {0,1}
@attribute trecidiv numeric
@attribute ttoanfo {0,1}
@attribute ttoitrac {0,1}
@attribute ttratame numeric

@data
1,1,1,1,0,1,1,1,1,1,?,0,0,0,1,1,1,1,1,1,1,1,1,1,1,0,1,1,0,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,?,1,0,1,1,1,1,1,1,65,1,1,1
?,1,1,1,0,1,1,1,1,1,1,?,1,0,1,1,1,1,1,1,1,0,1,1,1,1,1,1,0,1,1,1,1,6,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,?,1,1,1,1,1,1,1,1,82,1,?,1
0,1,0,0,0,1,1,1,1,1,1,0,0,0,1,1,1,1,0,1,1,1,1,1,1,1,0,1,0,1,1,1,1,1,0,1,0,1,1,1,1,1,1,1,0,1,1,1,0,1,2,1,0,1,1,1,1,1,1,0,1,46,1,1,1
1,1,1,1,1,1,1,1,1,1,1,?,1,0,1,1,1,1,0,1,1,1,0,1,1,1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,61,1,1,1
0,1,0,1,0,1,0,1,1,1,1,0,0,0,1,0,0,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,0,1,1,1,1,1,0,1,0,0,0,1,1,0,0,1,1,1,1,1,1,1,1,1,?,1,0,1,1,1,1,1,1,22,0,1,1
?,1,1,1,1,1,1,0,1,1,1,1,0,1,0,1,1,1,0,1,?,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,?,0,1,1,1,1,1,?,0,1,1,1,1,1,1,?,1,1,1,1,1,1,?,1,?,1,1,3,1,1,?,?,1,1,1,1,1,?,3,0,1,1,
0,1,1,1,0,1,1,1,1,1,1,?,1,0,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,0,1,1,1,1,0,1,0,1,0,1,1,1,1,1,1,1,1,1,?,1,?,1,1,1,1,1,1,1,1,20,1,1,1
0,1,1,1,1,1,1,1,1,1,1,0,1,0,1,1,0,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,0,1,1,1,1,34,1,1,1
0,1,0,1,0,1,0,1,1,1,1,0,0,0,1,1,0,0,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,0,1,0,0,1,1,1,1,1,0,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,45,1,1,1
?,1,0,?,?,1,1,1,1,1,1,?,0,0,1,?,1,?,1,?,1,1,1,1,1,1,1,1,1,1,1,1,3,1,1,1,?,1,?,?,1,?,1,?,1,?,1,1,1,1,1,1,?,?,?,?,?,?,?,?,?,1,53,1,1,?
?,1,0,?,?,?,?,?,?,?,?,0,0,0,?,?,?,?,?,?,?,?,?,?,?,?,?,?,0,?,?,?,?,0,?,?,0,?,?,0,?,?,0,?,?,?,?,?,?,0,?,?,?,?,?,?,?,?,?,?,?,?,?,?,39,?,?,?
?,1,1,1,0,1,0,1,1,0,0,1,3,1,0,0,?,?,1,1,1,1,1,1,1,1,0,1,1,1,?,0,1,1,1,0,0,0,1,1,1,1,1,1,?,1,1,2,1,0,?,?,1,?,1,?,?,0,?,19,0,1,1
? 1 0 1 0 1 1 1 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 66 1 1 1
```

Figura 3: Estrutura de dados de um arquivo arff. No arff é possível visualizar a estrutura dos dados sem a necessidade de um editor específico. Os atributos são precedidos por "@atribute" e seus possíveis valores aparecem a seguir. Se o campo tiver valores nominais estes aparecem entre chaves e campos numéricos com a denominação "numeric". O arquivo ainda guarda informações sobre filtros e pré-processamentos aplicado pelo WEKA, guardando o histórico de informações e é possível fazer anotações dentro do arquivo.

b) Outras Ferramentas

Para edições pontuais e conversão dos dados originais foram também utilizados editores de planilhas e para a conversão do arquivo SPSS para o formato CSV foi utilizado um programa de código aberto chamado PSPP.

3.3.1 Algoritmos Utilizados

A descrição e aplicabilidade dos algoritmos utilizados abaixo foi retirada da documentação do WEKA que acompanha o programa.

a) Seleção de Atributos – CsfSubsetEval

Avalia o mérito do subconjunto de dados considerando o valor individual de predição de cada atributo e leva em conta o valor de redundância entre eles. Os atributos selecionados são altamente correlacionados com o atributo alvo mantendo pouca

interrelação.

b) Árvores de decisão – J48 (C4.5)

O algoritmo de árvore escolhido foi o j48 que é a implementação do algoritmo C4.5 de construção de árvores. É um algoritmo amplamente utilizado e seu detalhes de funcionamento podem ser encontrados em QUINLAN(1993).

c) Regras – JRip

É a implementação do algoritmo RIPPER (COHEN, 1993) no WEKA. Detalhes de sua arquitetura podem ser encontrados na documentação do WEKA. Foi escolhido por ter os melhores resultados dentre as regras para as bases selecionadas.

d) Classificador probabilístico - NayveBayes

O algoritmo selecionado da classe probabilística foi o NayveBayes que, apesar se suas limitações e premissas tem sido utilizado com sucesso, inclusive em mineração de dados médicos(ABRAHAM; SIMHA; IYENGAR, 2006).

O algoritmo assume que todos os atributos são igualmente importantes para a decisão e também assume que os atributos são estatisticamente independentes, ou seja, saber o valor de um atributo não diz nada a respeito do valor de outro atributo.

3.4 Extraction, Tranformation and Load – ETL

O processo de extração, transformação e carregamento – ETL é o processo que prepara a base para analise. As bases de dados complexas frequentemente possuem campos irrelevantes ou redundantes que introduzem ruido na analise.

O ETL é um dos passos mais importantes na mineração de dados. Devido às particularidades dos algoritmos e perguntas a serem respondidas faz-se necessária a preparação direcionada da base.

Os métodos utilizados para cada uma das bases serão discutidos nas seções a seguir.

3.4.1 ETL – PCM

a) Pré-processamento

A transformação da base de dado de PCM seguiu os seguintes passos:

1. Conversão dos arquivos SPSS em arquivos CSV

2. Eliminação de campos irrelevantes para a análise, como informações pessoais, números de protocolo e identificadores

3. Padronização da grafia em campos aplicáveis. A figura 4 mostra um exemplo para a base PCM.

O processo de pré-processamento deixou 301 dos 314 atributos originais.

Raul S	Bom Jesus do Ga	MG	0	Bom Jesus do Galho	4	06.01.1982	20	1
	Sete Lagoas	MG			0	16.09.1973	34	0
	Laginha	MG		Lajinha	3	27.02.1954	45	0
	Ribeir?o das Neve	MG		Carlos Chagas	5	07.06.1946	53	0
	Belo Horizonte	mg		Bom Jesus do Galho	1	11.09.1943	39	0
	Acesita	MG		Caratinga	1	24.09.1971	19	0
	Ant?nio Dias	MG		Vit?ria	9	24.01.1910	66	0
Novo L	Belo Horizonte	MG		Belo Horizonte	0	30.07.1958	44	0
	Lagoa Santa	MG		Lagoa Santa	0	02.03.1980	23	0
	Bambu?	MG	0	Dores?polis	5	08.08.1957	0	1
	Santa Cruz	MG	0	Ponte Nova	8	08.05.1984	6	0
	BH	MG	0	BH	0	04.06.1964	24	0
	Justin?polis	mg	0	n?o mencionada	4	11.10.1994	7	1
	Caratinga	MG	0	Caratinga MG	5	0	20	1
	S?o Sebastiao Ma	MG	0	n?o mencionada	1	0	54	0
	Santana dos Mont	MG	0		0	0	0	0
	Belo Horizonte	MG	0	Entre Rios	4	01.02.1935	52	0
	Belo Horizonte	MG	0	?guas Belas-BA	9	0	39	0
	Ribeir?o das Neve	MG		Novo Cruzeiro	7	20.07.1939	53	0
06	Belo Horizonte	MG		Serrania	2	26.03.1960	41	0
	Te?filo Otoni	MG	0	Te?filo Otoni	5	02.02.1956	20	0
	Ara?ua?	Minas Gerais	0	Ara?ua?	7	13.09.1942	40	0
	Boa Esperan?a	Minas Gerais	0	n?o mencionada	6	0	42	0
	Cordisl?ndia	Minas Gerais	0	S?o Gon?alo do Sapuca?	6	12.08.1982	8	0
	Itapecerica	MG	0	Itapecerica	5	18.12.1958	29	0
	Itabira	Minas Gerais		Senhora do Carmo	4	24.02.1914	64	0
)/402	Belo Horizonte	MG		Taioberas	6	11.11.1957	44	0
ıl	Belo Horizonte	MG		Carana?ba	0	16.04.1952	55	0

Figura 4: Fragmento da base bruta de PCM mostrando os problemas de preenchimento de dados. No detalhe, a grafia para a cidade de Belo Horizonte aparece escrita de duas formas distintas. Ao lado pode-se vem a representação do estado de Minas Gerais escrita de 3 formas diferentes.

b) Redução de dimensionalidade

Como já mencionado, a redução de Dimensionalidade visa eliminar campos que

não agregam informações para a analise. Os campos Irrelevantes são campos que não pertencem ao escopo da análise ou que possuem atributos com valores repetidos com frequência muito alta, ou seja, se todos ou quase todos os atributos tem um mesmo valor. (WITTEN; FRANK; HALL, 2011; GUYON; ELISSEEFF, 2003). A redução da base de PCM ocorreu em duas etapas. A primeira, chamada de Análise de distribuição de Frequência, mostra quantas vezes um determinado valor aparece para o atributo e fornece uma visão geral da distribuição dos dados.

As figura 5 e 6 mostram como o WEKA fornece essa analise de distribuição. Além do WEKA, os programas R e editores de planilhas foram usados para calcular a distribuição de frequência dos dados.

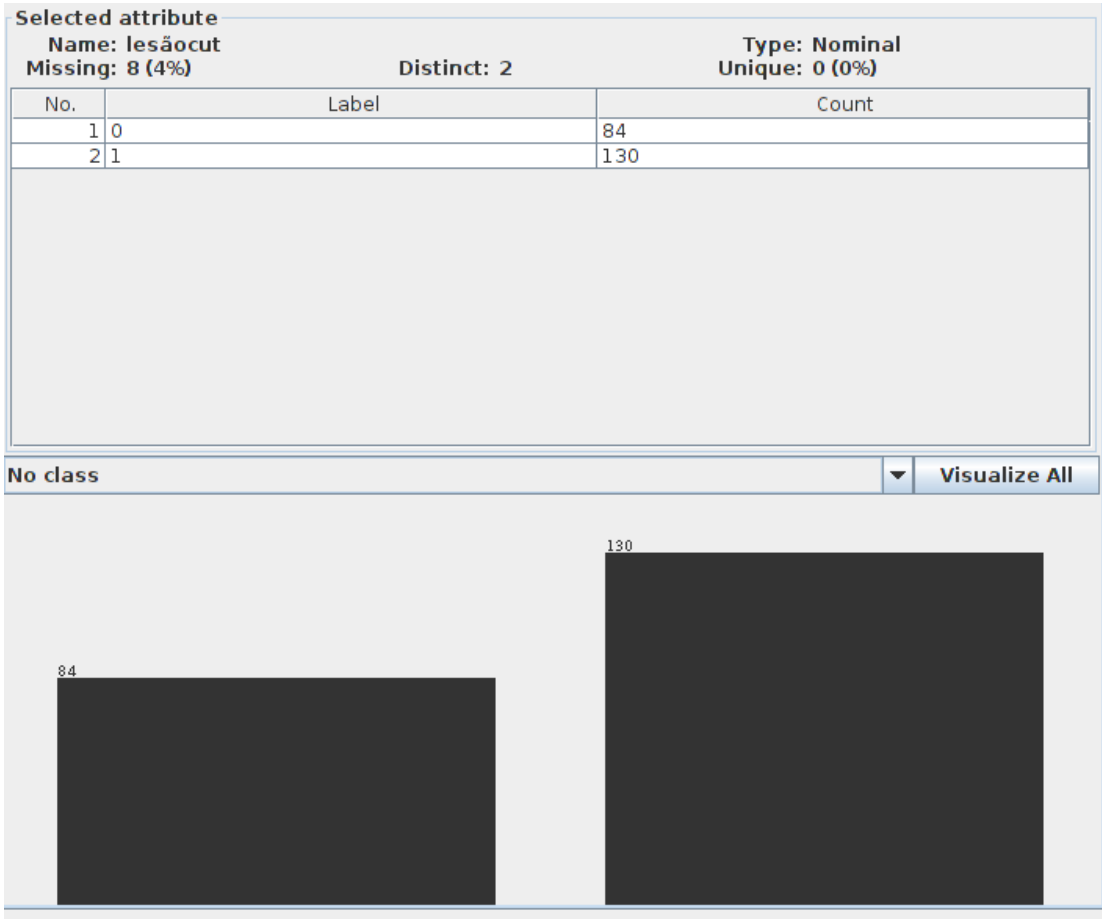

Figura 5: Tela mostrando análise de distribuição de frequência do WEKA para o atributo "lesão cutânea". Além da distribuição de frequência, o programa fornece o numero absoluto e a porcentagem relativa à quantidade de campos sem valor, no campo Missing. O campo Distinct se refere à quantos valores diferentes aparecem nos registros e o campo Unique se refere a quantos registros possuem valores únicos, ou seja, que não se repetem para o determinado atributo.

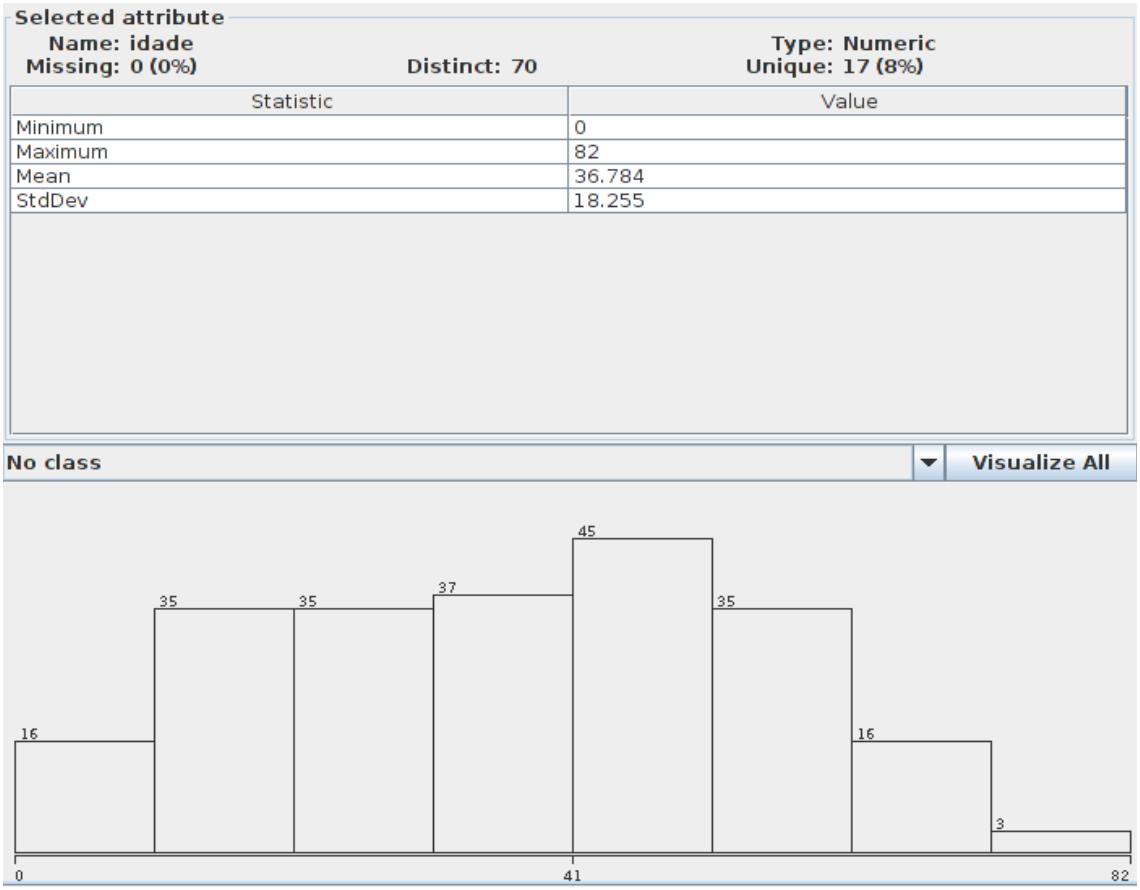

Figura 6: Analise de distribuição de frequência do atributo "idade". Para atributos do tipo numérico, a interface fornece ainda fornece analises como valor mínimo, máximo, média e desvio padrão.

Campos com uma alta concentração de um valor tendem a ser irrelevantes ou mesmo prejudiciais aos algoritmos de mineração pois esses campos geram ruido (WITTEN; FRANK; HALL, 2011). O valor mais frequente na base de dados de PCM foi o valor *não avaliado* ou *não preenchido*. Exigi-se um certo cuidado ao se analisar campos com valores não preenchidos ou ausentes. A falta de informação sobre um campo, no caso, um exame ou teste podem carregar informações diferentes dependendo do contexto. Um exame pode não ter sido preenchido pois pode depender do resultado de outro exame. Exames podem ser solicitados e não serem apresentados pelos pacientes no retorno ao médico, por exemplo. Essas informações precisam ser analisadas em seu contexto e abordagens simplistas como preencher os campos faltantes

com zeros não é uma boa prática e tende introduzir informações erradas na análise (WITTEN; FRANK; HALL, 2011; HASTIE; TIBSHIRANI; FRIEDMAN, 2009; HAN; KAMBER, 2001). O resultado negativo de um exame – valor 0 na base – traz uma informação totalmente diferente de um exame não realizado – *missing value*.

Atributos com mais de 70% de valores ausentes (*missing values)* foram descartados, ou seja, atributos com pelo menos 30% dos registros preenchidos foram mantidos. Com uma base de dados tão complexa como a de PCM, uma abordagem cuidadosa se fez necessária na hora de definir o limite de corte. Aproximadamente 29% da base de dados é composta de valores ausentes. Isso corresponde a 19.422 de um total de 66.822 campos. Um total de 109 atributos (36% dos 301 usados) possuem uma contagem de valores ausentes acima de 30%. Definir o limite de corte muito alto incorreria o risco de se eliminar muitos campos relevantes. É interessante ressaltar que dos 17 tipos de PCM testados pelo protocolo de exame, 11 não aparecem de maneira significativa na base de dados, como podemos observar na tabela 1. A tabela 2 mostra os atributos com maior concentração de valores ausentes da base de dados de PCM.

Tabela 1: Distribuição de frequência dos 11 tipos de PCM sem distribuição significativa

Tipo de PCM	Valor do atributo	Frequência (%)
PCM renal	normal	95.2
PCM pancreática	normal	95.2
PCM baço	normal	93
PCM genital	normal	94.7
PCM adrenal	normal	93.9
PCM intestinal	normal	93.9
PCM óssea	normal	93
PCM gastrica	normal	93.1
PCM hepática	normal	92.1
PCM limfática	normal	91.3
PCM neurologica	normal	90.4

Tabela 2: Distribuição de frequência dos 20 atributos com os maiores índices de valores ausentes.

Atributo	Valor dominante	Frequência(%)
Teste sorológico positivo	Não preenchido	98.69
Tamanho Boyd (cm)	Não preenchido	85.15
Raio X facial	não avaliado	93.5
Lactato alterado	não avaliado	92.1
CPK alterado	não avaliado	91.7
CK-MB alterado	não avaliado	90.8
HBeAg positivo	não avaliado	90.8
Anti-HBc positivo	não avaliado	90.5
Lipase alterada	não avaliado	88.2
CMV positivo	não avaliado	87.8
Anti-HIV (Elisa) positivo	não avaliado	87.3
Chagas positivo	não avaliado	87.3
Acido urico alterado	não avaliado	86.5
Anti-HAV positivo	não avaliado	86.34
Cultura alterada	não avaliado	86.4
Espirometria alterada	não avaliado	86.5
Anti-HCV positivo	não avaliado	86.3
Amilase alterada	não avaliado	85.6
Anti-HBs positivo	não avaliado	85.6
area lesão (cm^2)	Não preenchido	78.6

Após essa etapa foram removidos 52 atributos dos 314 atributos iniciais, restando 262. Depois do processo inicial de redução de dimensionalidade, uma segunda etapa foi aplicada para selecionar os atributos mais relevantes à análise, descrita a seguir.

c) Análise assistida por especialistas

A base de dados resultante do processamento inicial ainda é bastante larga e ruidosa. Uma segunda abordagem direcionada foi conduzida. Dos 262 atributos

restantes da etapa inicial, 23 foram selecionados pela equipe de especialistas como sendo relevantes para investigações de correlação. Esses atributos são: sexo, idade, recidiva, contato rural, tabagismo, etilismo e todas as 17 formas de PCM constantes no protocolo. Nesta fase foram conduzidas analises de seleção de atributos com o algoritmo csfSubsetEval do WEKA. Essa analise gera um subconjunto de atributos que se relacionam com a classe (atributo) selecionado, conforme mostrado na figura 7. Foram gerados 23 subconjuntos, um para cada atributo indicado pelos especialistas e esses subconjuntos foram agrupados na base final para os testes de classificação. Dos 262 atributos iniciais dessa etapa, 112 atributos tiveram correlação significativa com pelo menos um dos 23 atributos escolhidos pelos especialistas.

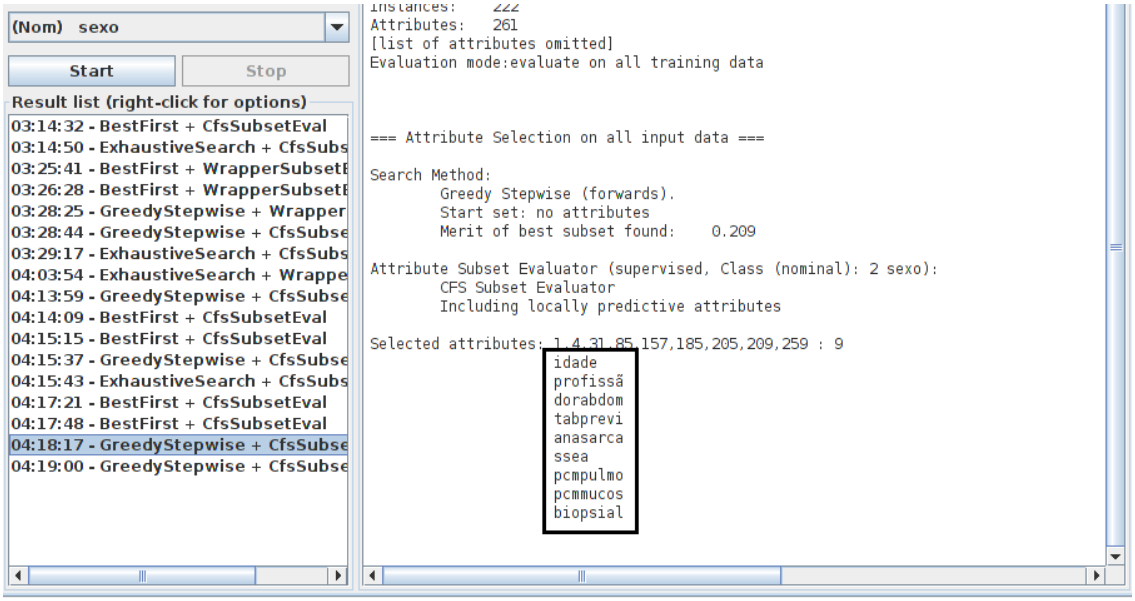

Figura 7: Captura de tela do WEKA para seleção de atributos relacionados à classe "sexo". No quadro direito estão listados o método de busca e o algorítimo utilizado. No destaque à direita encontra-se o subconjunto correlacionado ao atributo selecionado.

3.4.2 ETL – ECG

A base de dados de ECG-ELSA consiste de uma planilha única com 11938 registros codificados pelo código de Minnesota e revisados por especialistas. Utilizando-se um editor de planilhas foram retiradas as colunas do arquivo relativas aos códigos identificadores dos pacientes, data dos exames, colunas vazias e outras colunas

irrelevantes. Após a limpeza inicial, restaram 24 colunas que recebem os código de Minnesota aplicáveis e 14 colunas com medições dos exames. O código de Minnesota é um sistema de classificação morfológica do ECG e uma descrição mais detalhada se encontra na seção seguinte.

As 24 colunas do código de Minnesota são preenchidas de maneira esparsa, podendo haver mais de um código por registro e códigos repetidos como mostra a figura 8. Para uma classificação adequada, os códigos de Minnesota fora isolados em arquivos separados com o auxílio de programas escritos pra esta finalidade. No total foram encontrados 66 códigos classificados na base. As códigos então foram isolados em arquivos contendo todos os dados da base original, mas somente a classificação de um dos possíveis códigos atribuídos, como mostrado na figura 9.

			5-4-0				
	5-3-0	5-3-0	5-3-0	6-3-0			
					7-6-0		
					7-6-0		8-8-0
					7-6-0		
							8-8-0
							8-8-0
							8-8-0
							8-8-0
					7-5-0		
					7-5-0		
							8-8-0
	5-3-0		5-3-0				8-4-1

Figura 8: Fragmento da matriz de dados original, mostrando linhas contento mais de um código por registro, códigos repetidos na mesma linha e campos sem preenchimento.

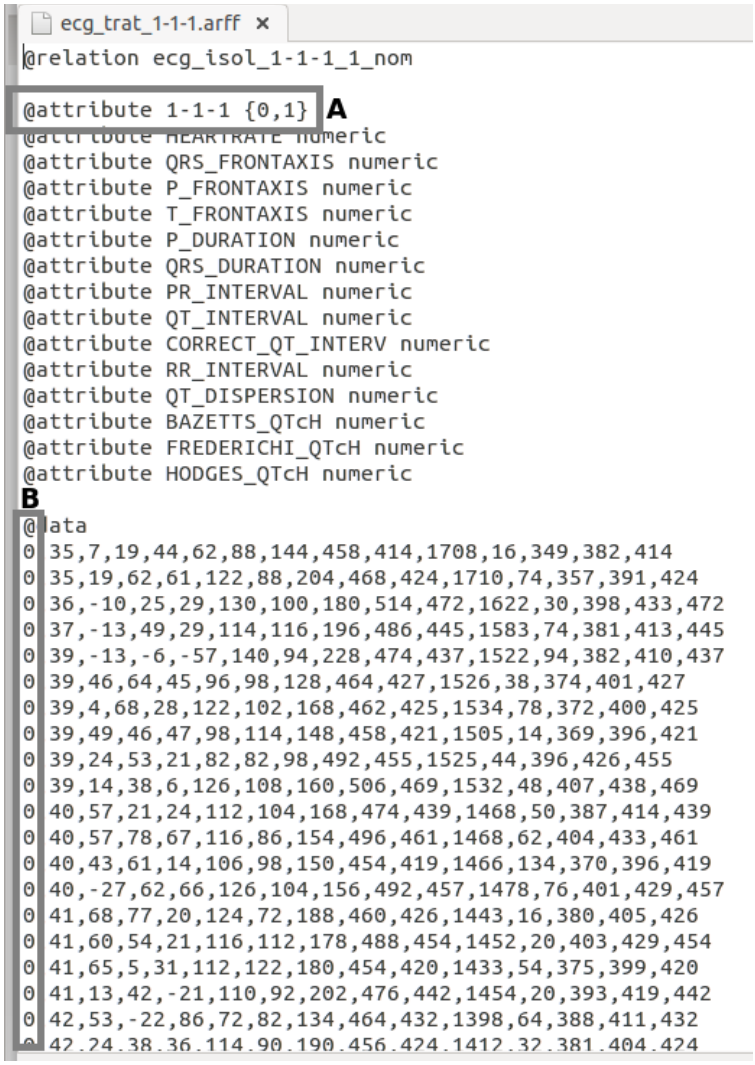

Figura 9: Exemplo de arquivo com o código isolado. (A) destaque do atributo com o código escolhido para ser isolado. (B) valores atribuídos para o atributo código. "1" quando o código estiver presente no registro e "0" quando o exame não tiver recebido o código correspondente.

A maioria dos 66 códigos encontrados tem baixa representatividade na base. Apenas 5 códigos classificados possuem mais de mil registos únicos, sendo eles: 5-3-0, 7-6-0, 8-8-0, 9-4-1 e 9-4-2. Do total, 45 códigos aparecem em menos de 100 registros. A tabela 3 mostra a distribuição de frequência dos códigos encontrados. Curiosamente, apenas seis dos 11938 registros não receberam nenhum código.

Tabela 3: Distribuição de Frequência dos códigos de Minnesota encontrados na base. O total de registros é de 11938.

Código	Nº Ocorrências	Código	Nº Ocorrências	Código	Nº Ocorrências
1-1-1	31	2-2-0	9	7-2-1	200
1-1-2	18	2-3-0	84	7-3-0	195
1-1-3	5	2-4-0	1	7-4-0	74
1-1-4	15	2-5-0	24	7-5-0	473
1-1-5	10	3-1-0	375	7-6-0	1439
1-1-6	12	3-2-0	7	7-7-0	37
1-1-7	5	3-3-0	207	7-8-0	1
1-2-1	27	4-1-1	4	8-1-1	83
1-2-2	60	4-1-2	17	8-1-2	13
1-2-3	11	4-2-0	126	8-3-1	28
1-2-4	57	4-3-0	349	8-3-2	11
1-2-5	6	4-4-0	6	8-4-1	48
1-2-6	45	5-1-0	21	8-7-0	30
1-2-7	4	5-2-0	409	8-8-0	3195
1-2-8	28	5-3-0	1329	9-1-0	174
1-3-1	168	5-4-0	481	9-2-0	447
1-3-2	25	6-2-3	1	9-3-0	88
1-3-3	82	6-3-0	190	9-4-1	8923
1-3-4	186	6-4-1	7	9-4-2	2644
1-3-5	21	6-5-0	297	9-5-0	53
1-3-6	31	6-8-0	6	9-8-1	2
2-1-0	392	7-1-1	57	9-8-2	4

Para uma segunda análise foram utilizados grupos de códigos. O isolamento dos códigos foi feito com base no primeiro número do código, ou seja, registros que receberam qualquer um dos códigos começando com o numero 7 por exemplo, que representa alterações na condução ventricular, foram agrupados. Esse agrupamento foi efetuado para todos os 9 tipos de código.

Esse agrupamento tem como objetivo testar a capacidades dos algoritmos de classificar as categorias de alteração, utilizando parâmetros mais abrangentes.

3.4.3 – Código de Minnesota

O Código de Minnesota é o sistema de codificação de ECG mais frequentemente usado em estudos epidemiológicos e populacionais, além de estudos clínicos. Ele é dividido em nove grupos de classificação, com critérios rigidamente definidos (RIBEIRO et al., 2013). O código sofreu poucas modificações desde sua descrição inicial em 1960 (PRINEAS; CROW; ZHANG; 2010). Um manual auxilia na aplicação do código com explicações detalhadas sobre as mensurações e como obter a codificação final. O código de Minnesota foi utilizado na maioria dos grandes estudos epidemiológicos populacionais e tem seu valor prognóstico estabelecido. MACHADO et al., 2006; ZHANG; PRINEAS; EATON 2010; ZHANG et al., 2012). Nos últimos anos, os progressos da tecnologia da informação e dos conhecimentos sobre a interpretação automática dos ECG tornaram possível o desenvolvimento de programas de codificação automática dos traçados pelo código de Minnesota (KORS et al., 2000; KORS; HERPEN, 2001; MACFARLANE, 1996; PRINEAS; CROW; ZHANG, 2010).

Tais métodos mostraram-se particularmente atraentes para o uso em estudos populacionais, que passaram a contar com amostras cada vez maiores. Comparações entre os métodos manual e automático mostraram que a codificação automática apresenta menor variabilidade e maior acuidade e tem sido recomendada como de escolha para estudos epidemiológicos(KORS et al., 2000; KORS; HERPEN, 2001; PRINEAS; CROW; ZHANG, 2010). O código de Minnesota não produz uma interpretação do ECG, apenas classifica a morfologia eletrocardiográfica, deixando a interpretação e diagnóstico para o médico responsável (MACFARLANE, 2000). Os códigos são divididos em tríades de números que indicam o tipo de alteração e gravidade. A tabela 4 exemplifica o código.

Tabela 4: Classificações usadas pelo código de Minnesota.

Código de Minnesota	Anormalidade no ECG
1-1-1 1-3-6	Ondas Q
2-1 2-5	Desvio no eixo QRS
3-1 3-3	Ondas R de alta amplitude
4-1-1 4-4	Junção ST (J) e depressão do segmento

5-1 5-4	Itens da onda T
6-1 6-8	Defeito na condução A-V
7-1-1 7-8	Defeito na Condução Ventricular
8-1-1 8-9	Arritmias
9-1 9-8-2	Alterações gerais, incluindo elevação do segmento ST (9-2)

Adaptado de MACFARLANE, 2000

3.5 - Classificadores

Os classificadores são uma ferramente fundamental da mineração de dados. Os algoritmos de classificação conferem poder de predição dos dados, podendo assim designar uma classificação ao registro. Os métodos testados nas bases de dados de PCM e ECG foram árvores de classificação, classificadores bayesianos (modelos probabilísticos) e classificadores de regras. Os algoritmos utilizados serão discutidos a seguir para cada um dos estudos de caso.

3.5.1 – Classificadores para PCM

Para a análise de classificação da base de dados filtrada, o método escolhido foi a construção de árvores de decisão, ou árvores de classificação. É importante salientar que outros métodos foram testados para a base de dados, sendo eles: redes bayesianas com os algoritmos NaiveBayes e BayesNet implementados no WEKA e classificadores de regras com o algoritmo JRIP. Os resultados alcançados com as árvores foram superiores aos demais métodos e serão descritos mais detalhadamente.

a) Árvores de classificação

Árvores de decisão são uma tecnologia de mineração indicados para prever e classificar atributos. Essas árvores podem gerar resultados à partir de diferentes variáveis por repetição que por sua vez podem ser usadas para analisar características, similaridades e diferenças nos dados (QUINLAN, 1993). É uma abordagem direcionada, ou seja, o algoritmo precisa de um atributo alvo definido pelo usuário que terá então sua classificação prevista.

O algoritmo utilizado para a análise foi o j48, nativo do WEKA. Esse algoritmo

é uma implementação do amplamente conhecido e utilizado C4.5 (para detalhes do seu funcionamento ver QUINLAN, 1993). Esse algoritmo foi escolhido por sua versatilidade. Ele classifica atributos nominais, é capaz de utilizar atributos numéricos e é também capaz de lidar com dados ausentes evitando assim a necessidade de se discretizar variáveis numéricas e imputar valores ausentes. Esse algoritmo é uma ótima alternativa para uma base complexa como a PCM que é uma mistura de campos nominais, numéricos e ausentes.

Todos os testes foram feitos utilizando a opção de *10 fold cross-validation*. Nesse método de teste, a base de dados é dividida em 10 partes onde se usa 9 para treino (aprendizado) e 1 para teste. Após a conclusão do teste usa-se um conjunto diferente de 9 partes para treino e 1 para teste até que todas as 10 partes tenham sido usadas para testes. O resultado final é a média dos 10 testes. A figura 10 exemplifica um *10 fold cross-validation*. Esse método fornece uma estimativa melhor do uso do conjunto de regras da árvore em situações reais. O uso do *dataset* inteiro pode produzir uma estimativa muito otimista do seu classificador (WITTEN; FRANK; HALL, 2011).

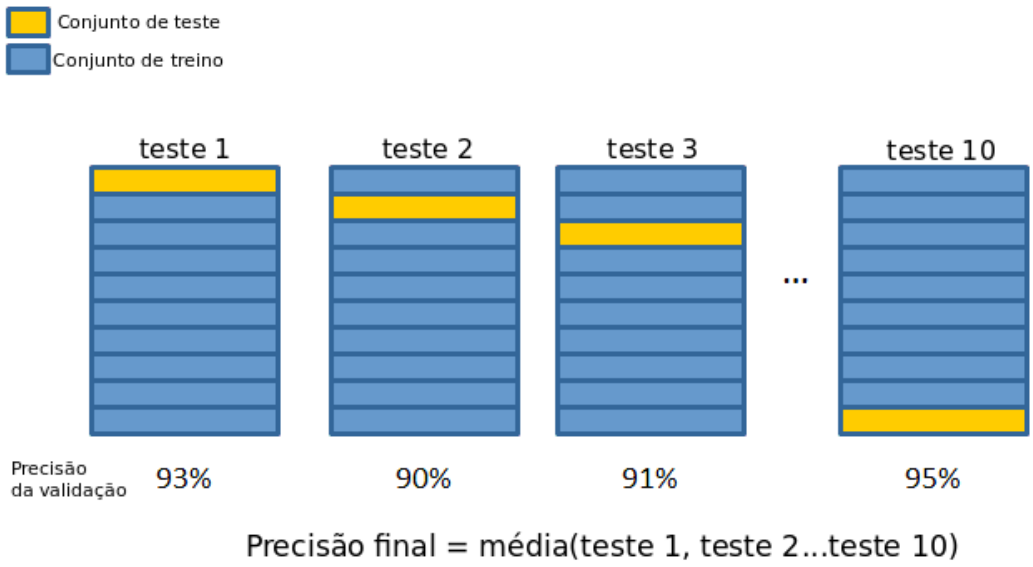

Figura 10: Exemplo de 10 fold cross-validation. Nesse método, a base de dados é dividida em 10 partes iguais onde nove são utilizadas para treino e 1 para teste. Após o teste, a décima parte seguinte é usada para teste e o restante da base para treino. O ciclo se repete até que todas as 10 partes da base tenham sido utilizadas para teste. Note que nem a divisão da base nem a escolha da porção é aleatória. Os ciclos são sequenciais com a ordem definida pelo próprio algoritmo

3.4.2 – Classificadores para ECG

Os mesmos algoritmos de classificação foram testados para ECG. Tendo sido escolhidos os algoritmos de árvore – j48 – e de regras – JRip – para esta base.

Diferentemente de PCM onde as árvores foram amplamente superiores aos demais, para ECG os algoritmos obtiveram desempenhos muito semelhantes em vários casos. Tanto as árvores como as regras produzem modelos de classificação que são legíveis e facilmente modificáveis. Para a base ECG o algoritmo de regras JRip criou modelos de classificação menores que o j48 (árvore). A figura 11 exibe os exemplos.

A

```
QRS_DURATION <= 98: 0 (9676.0/96.0)
QRS_DURATION > 98
|   QRS_DURATION <= 118
|   |   HEARTRATE <= 80
|   |   |   QRS_DURATION <= 110: 1 (1578.0/447.0)
|   |   |   QRS_DURATION > 110
|   |   |   |   QT_INTERVAL <= 440: 1 (199.0/64.0)
|   |   |   |   QT_INTERVAL > 440
|   |   |   |   |   P_FRONTAXIS <= 36
|   |   |   |   |   |   P_FRONTAXIS <= 25
|   |   |   |   |   |   |   QRS_DURATION <= 114
|   |   |   |   |   |   |   |   HEARTRATE <= 56: 1 (3.14/0.14)
|   |   |   |   |   |   |   |   HEARTRATE > 56: 0 (2.0)
|   |   |   |   |   |   |   QRS_DURATION > 114: 0 (3.0)
|   |   |   |   |   |   P_FRONTAXIS > 25: 0 (8.14)
|   |   |   |   |   P_FRONTAXIS > 36: 1 (43.73/18.73)
|   |   HEARTRATE > 80
|   |   |   PR_INTERVAL <= 128: 1 (5.2/0.07)
|   |   |   PR_INTERVAL > 128
|   |   |   |   P_FRONTAXIS <= 54
|   |   |   |   |   QT_DISPERSION <= 58
|   |   |   |   |   |   QT_DISPERSION <= 32
|   |   |   |   |   |   |   QRS_FRONTAXIS <= 5
|   |   |   |   |   |   |   |   P_FRONTAXIS <= 39: 0 (2.0)
|   |   |   |   |   |   |   |   P_FRONTAXIS > 39: 1 (3.0)
|   |   |   |   |   |   |   QRS_FRONTAXIS > 5: 0 (7.35)
|   |   |   |   |   |   QT_DISPERSION > 32: 1 (5.7/1.0)
|   |   |   |   |   QT_DISPERSION > 58: 0 (9.0)
|   |   |   |   P_FRONTAXIS > 54
|   |   |   |   |   QT_INTERVAL <= 404
|   |   |   |   |   |   PR_INTERVAL <= 154
|   |   |   |   |   |   |   CORRECT_QT_INTERV <= 425
|   |   |   |   |   |   |   |   CORRECT_QT_INTERV <= 401: 1 (2.0)
|   |   |   |   |   |   |   |   CORRECT_QT_INTERV > 401: 0 (9.22/1.0)
|   |   |   |   |   |   |   CORRECT_QT_INTERV > 425: 1 (4.44/1.0)
|   |   |   |   |   |   PR_INTERVAL > 154
|   |   |   |   |   |   |   P_DURATION <= 120
|   |   |   |   |   |   |   |   PR_INTERVAL <= 180: 1 (16.7/2.23)
|   |   |   |   |   |   |   |   PR_INTERVAL > 180: 0 (3.13/0.09)
|   |   |   |   |   |   |   P_DURATION > 120: 1 (6.26/0.09)
|   |   |   |   |   QT_INTERVAL > 404: 0 (3.0)
|   QRS_DURATION > 118: 0 (346.0/9.0)

Number of Leaves  :     22

Size of the tree :      43
```

B

```
JRIP rules:
===========

(QRS_DURATION >= 100) and (QRS_DURATION <= 114) => 7-6-0=1 (1825.0/543.0)
(QRS_DURATION >= 116) and (QRS_DURATION <= 118) and (QT_DISPERSION >= 40) and (P_DURATION >= 120) => 7-6-0=1 (20.0/2.0)
 => 7-6-0=0 (10091.0/139.0)

Number of Rules : 3
```

Figura 11: Modelos de classificação gerados pelos algoritmos j48 e JRip. Ambos os conjuntos de regras/passos obtiveram taxa de acerto global de 94%. (A) árvore de classificação criada para o código 7-6-0. (B) Regras criadas pelo algoritmo JRip para a mesma base. Note que o algoritmo criou apenas 3 regras.

IV Resultados e Discussão

4 Resultados

4.1 – Resultados PCM

Os resultados de classificação de atributos estão descritos a seguir. Para a análise de PCM foram criadas árvores de decisão podadas (*pruned trees*). As árvores de decisão podadas são versões simplificadas de árvores de decisão. São descartados os ramos menos relevantes, com menor poder preditivo, e isso tende a reduzir o ruido na análise. Essas árvores provém também um modelo mais amigável o compreensível para a vizualização da árvore. A figura 12 exemplifica uma árvore podada.

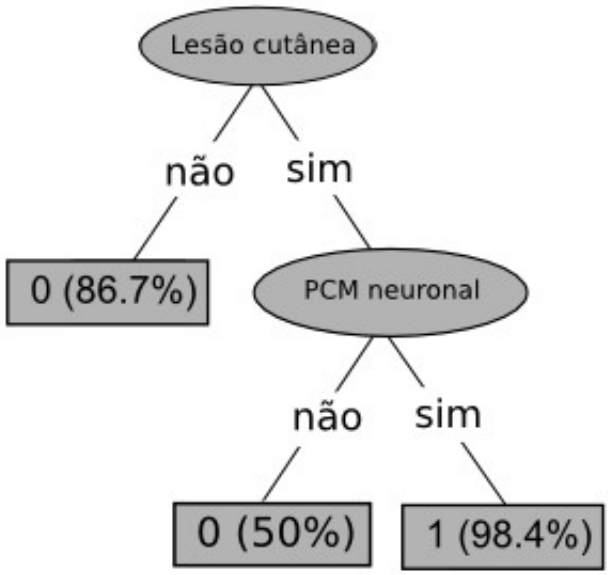

Figura 12: Exemplo de árvore de decisão. A árvore gerada para a classificação do Atributo PCM cutânea. Entre parênteses está a taxa de acerto de cada um dos passos da árvore. A taxa global do algorítimo é uma média ponderada entre elas. Neste caso a precisão do algoritmo foi de aproximadamente 93% de acerto.

4.1.1 – Árvore de classificação: atributo PCM cutânea

A tabela 5 apresenta os resultados de classificação para PCM cutânea. A primeira parte da tabela apresenta as instancias corretamente classificadas, incorretamente classificadas e suas respectivas porcentagens. A segunda parte da tabela mostra números

mais detalhados sobre a real qualidade do classificador. Ela nos mostra uma taxa de tp (*true positives*), positivos verdadeiros de 0,904 para a classe 0 (negativo para PCM cutânea). Isso quer dizer que as instancias foram corretamente classificadas como 0 em 93,2% dos casos. A taxa de fp (*false positives*) nos indica que 7,2% das instancias foram classificadas como 0 erroneamente. O cenário ideal seria a taxa de tp=1 e a taxa de fp=0. A terceira parte da tabela, *confusion matrix,* indica o numero de instancias classificadas corretamente e incorretamente. A primeira linha mostra que 65 instancias foram classificadas corretamente como negativas e 5 foram classificadas de maneira errada como negativas. A segunda linha nos indica que 128 instancias foram corretamente classificadas como positivas para PCM cutânea e 10 foram incorretamente classificadas como positivas.

Esses resultados nos mostram que é possível prever a incidência de PCM cutânea com alta taxa de precisão. Seria possível, por exemplo, construir um sistema de suporte a decisão para esse atributo, auxiliando no diagnóstico e tratamento da doença.

Tabela 5: Árvore podada criada com o algoritmo J48 para a classificação do atributo PCM cutânea. Coluna TP = taxa de positivos verdadeiros (true positive). FP = taxa de falsos positivos A matriz de confusão explicita quantas instancias foram corretamente classificadas para cada classe.

10 fold cross-validation para PCM cutânea		
Atributos corretamente classificados	196	92,9%
Atributos incorretamente classificados	15	7,1%

tp rate	fp rate	precision	Area ROC	class
0,932	0,072	0,872	0,936	0
0,928	0,068	0,962	0,918	1
0,929	0,07	0,931	0,924	Média ponderada

Confusion matrix		
correto	incorreto	Classificado como
65	5	negativo
128	10	positivo

4.1.2 Árvore classificação: Atributo sexo

O sexo do indivíduo é um importante fator clínico para PCM. O efeito dos hormônios no desenvolvimento da doença em mulheres em idade reprodutiva já foi descrito na literatura (DE MOURA, 2008; SHIKANAI-YASUDA et al., 2006). A tabela 6 exibe os resultados para a classificação do gênero do paciente, baseado nas características da doença. A taxa de acerto geral é de quase 82%. Apesar do número relativamente alto, a precisão de classificação das classes 0 (masculino) e 1 (feminino) é bem diferente. A precisão para a classe 0 é de 0,86 e precisão para a classe 1 cai para 0,47 e a taxa de tp de apenas 0,263 para a classe 1. Essa diferença pode ser explicada pela composição da base de dados: 177 homens e apenas 38 mulheres. Atributos assimétricos podem sofrer distorções na classificação por 10 fold cross-validation, já que a base é fatiada e não há garantia de que há uma distribuição homogênea em cada um dos testes. Como base de comparação, o mesmo algoritmo melhorou seu desempenho quando aplicado à toda a base de teste. Sua taxa de acerto geral subiu para 87% e a precisão para classificar homem subiu para 0,92 e o de mulheres subiu para 0,65. Curiosamente, a árvore não podada para a classificação do sexo teve uma taxa de acerto de 94,4%, como mostra a tabela 7. De todos os testes realizados com PCM, esse foi o único caso notável de melhora significativa na classificação utilizando-se uma árvore completa (*unpruned tree*). A árvore completa possui 55 folhas e um tamanho total de 72. A árvore podada é bem mais simples, contento 6 folhas e um tamanho total de 11.

Além disso, os principais atributos ligados à classificação do gênero foram “PCM mucosa”, “PCM pulmonar”, “PCM linfonodal”, “PCM cutânea”, “inchaço linfonodal”, “sopro sistólico aórtico” e “tempo de evolução. O tempo de evolução significa a quanto tempo o paciente tem a doença e ela ainda está progredindo. É importante notar que ambas as árvores podadas – base inteira de teste e a 10 fold cross-validation, usaram as mesmas regras de classificação. Esses resultados estão de acordo com o encontrado na literatura para diferença de apresentação da doença em homens e mulheres.

Tabela 6: Árvore podada criada com o algoritmo J48 para a classificação do atributo Sexo. Coluna TP = taxa de positivos verdadeiros (true positive). FP = taxa de falsos positivos A matriz de confusão explicita quantas instancias foram corretamente classificadas para cada classe.

10 fold cross-validation para Sexo		
Atributos corretamente classificados	176	81,9%
Atributos incorretamente classificados	39	18,1%

tp rate	fp rate	precision	Area ROC	class
0,938	0,737	0,856	0,653	0
0,263	0,062	0,476	0,653	1
0,819	0,618	0,789	0,653	Média ponderada

Confusion matrix		
correto	incorreto	Classificado como
166	11	negativo
10	28	positivo

Tabela 7: Árvore completa criada com o algoritmo J48 para a classificação do atributo Sexo. Coluna TP = taxa de positivos verdadeiros (true positive). FP = taxa de falsos positivos A matriz de confusão explicita quantas instancias foram corretamente classificadas para cada classe.

Full trainig set para Sexo		
Atributos corretamente classificados	203	94,4%
Atributos incorretamente classificados	12	5,6%

tp rate	fp rate	precision	Area ROC	class
0,989	0,263	0,946	0,971	0
0,737	0,011	0,933	0,971	1
0,944	0,219	0,944	0,971	Média ponderada

Confusion matrix		
correto	incorreto	Classificado como
175	2	negativo
28	10	positivo

4.1.3 Árvore classificação: Atributo PCM recidiva

Como descrito anteriormente, atualmente não existem parâmetros clínicos para se prever a recidiva da doença. Esse foi outro atributo de interesse clínico testado que mostrou resultados promissores. A tabela 8 exibe os resultados. A taxa de acerto geral foi de 73,8%. Os positivos verdadeiros – *tp rate* – para recidiva deram uma taxa de 0,932 e a precisão geral foi de 0,71. Como era esperado o algoritmo melhorou seu desempenho ao ser aplicado à base de dados completa. A taxa de acerto geral subiu para 83%, a precisão geral subiu para 0,86. Mais uma vez, as regras de classificação geradas para a base completa e a 10 fold cross-validation foram iguais. Os atributos utilizados para a classificação de recidiva foram: PCM intestinal; Contagem global de leucócitos; Tempo de tratamento; Tratamento com anfotericina b; Raio X tórax e PCM disseminada. Esses são atributos de importância clínica ligados à progressão e tratamento da doença. Esses achados podem estimular e possivelmente guiar estudos de validação clínica. A coleta de dados ainda está em andamento e pesquisas futuras podem validar os parâmetros para uma predição segura da recidiva da doença.

Tabela 8: Árvore podada criada com o algoritmo J48 para a classificação do atributo Recidiva. Coluna TP = taxa de positivos verdadeiros (true positive). FP = taxa de falsos positivos A matriz de confusão explicita quantas instancias foram corretamente classificadas para cada classe.

10 fold cross-validation para Recidiva		
Atributos corretamente classificados	152	73,8%
Atributos incorretamente classificados	54	26,2%

tp rate	fp rate	precision	Area ROC	class
0,241	0,068	0,583	0,619	0
0,932	0,759	0,758	0,615	1
0,738	0,564	0,709	0,616	Média ponderada

Confusion matrix		
correto	incorreto	Classificado como
14	44	negativo
138	10	positivo

4.1.3 Árvore classificação: Atributo RaioX Alterado

Uma das surpresas deste trabalho foi a constatação de que 40% dos pacientes já diagnosticados com alguma forma de PCM não possuíam informações sobre os resultados da radiografia pulmonar. Essa informação foi recebida com surpresa pela equipe de especialistas pois esse é um exame primário do protocolo de diagnóstico da doença. A tabela 9 mostra os resultados de classificação para esse atributo. Com uma taxa global de acerto de 93% e baixas taxas de fp (falsos positivos) notamos que é possível prever a alteração do raio-x de tórax dos pacientes baseados em outros parâmetros clínicos. Esses resultados podem servir para amenizar o problema da falta de informação sobre esses exames. Muitas vezes os pacientes não realizam os exames pedidos ou não os apresentam no retorno da consulta, o que deixa lacunas no histórico.

Tabela 9: Árvore podada criada com o algoritmo J48 para a classificação do atributo RxAltera. Coluna TP = taxa de positivos verdadeiros (true positive). FP = taxa de falsos positivos A matriz de confusão explicita quantas instancias foram corretamente classificadas para cada classe.

Classificador para RxAltera		
Atributos corretamente classificados	120	93,0%
Atributos incorretamente classificados	9	7%

tp rate	fp rate	precision	Area ROC	class
0,969	0,182	0,939	0,763	0
0,818	0,031	0,9	0,903	1
0,93	0,143	0,929	0,799	Média ponderada

Confusion matrix		
correto	incorreto	Classificado como
93	3	negativo
27	6	positivo

4.2 Resultados ECG-ELSA

Os testes iniciais foram a classificações dos códigos de Minnesota para se testar o comportamento dos algoritmos e selecionar os melhores para testes mais complexos.

Na etapa de classificação inicial os algoritmos testados obtiveram desempenho semelhante em precisão. A escolha dos algoritmos então foi feita com base nas características dos modelos de classificação criados e formato de apresentação dos resultados. Os algoritmos de árvores e de regras criam modelos de classificação facilmente compreensíveis (figura 11, pag. 31), o que facilita a analise das regras por especialistas clínicos. Além disso, a fácil visualização e compreensão das regras pode facilitar a personalização e alteração das mesmas, conferindo ainda mais flexibilidade ao modelo.

Na grande maioria dos casos o algoritmo JRip de regras gerou regras de classificação mais simples, sem sacrificar desempenho, sendo escolhido como a melhor alternativa inicial.

O WEKA registra ainda a probabilidade individual de cada uma das instancias classificadas pelo algoritmo, quando cabível (figura 14, pag. 41). Os algoritmos utilizados dispões de validação interna, que gera um escore de probabilidade de acerto da classificação. Esse traço nos permite diferenciar classificações com alto grau de certeza de casos limítrofes ou menos confiáveis. As probabilidades individuais das classificações nos permitem abordagens ainda mais conservadoras, com possibilidade de redução maior de possíveis erros. Isso é essencial para se trabalhar com classificação de dados sensíveis como dados de diagnóstico médico, onde a redução de erro é prioridade. Todos os 66 códigos foram testados. Abaixo estão detalhados os melhores resultados.

4.2.1 Classificadores para o Código 8-8-0

A classificação do código 8-8-0 obteve desempenho muito semelhante entre a árvore(J48) e as regras(JRip). A precisão geral de ambos ficou acima de 96%, sendo o J48 levemente inferior, 96,35% ante a 96,5% do JRip. A tabela 10 a seguir mostra os resultados para o classificador de regras JRIP.

Tabela 10: Árvore podada criada com o algoritmo JRip para a classificação do código 5-3-0. Coluna TP = taxa de positivos verdadeiros (true positive). FP = taxa de falsos positivos A matriz de confusão explicita quantas instancias foram corretamente classificadas para cada classe.

JRip 10 fold cross-validation para 8-8-0		
Atributos corretamente classificados	11521	96,5%
Atributos incorretamente classificados	415	3,5%

tp rate	fp rate	precision	Area ROC	class
0,976	0,064	0,977	0,957	0
0,936	0,024	0,934	0,957	1
0,965	0,053	0,965	0,957	Média ponderada

Confusion matrix		
correto	incorreto	Classificado como
8529	212	negativo
2992	203	positivo

É interessante ressaltar que os parâmetros escolhidos por ambos os classificadores foi bastante semelhante. O código de início 8 é ligado à arritmias cardíacas e as regras encontradas pelos algoritmos estão ligadas à frequência e ao ritmo cardíaco, em concordância com a classificação do código de Minnesota. Isso mostra que esses algoritmos são adequados para se fazer esse tipo de analise e as regras descobertas estão de acordo com as convenções mais utilizadas. A figura 13 mostra o conjunto de regras selecionados pelo algoritmo JRip e para o J48 para a classificação do código 8-8-0. Isso demonstra potencial para se adaptar o conjunto de regras descobertas para refletir as variações e particularidades intrínsecas populacionais, de gêneros e pacientes com histórico de doenças cardíacas.

```
J48 pruned tree
------------------

HEARTRATE <= 59
|   P_FRONTAXIS <= -19
|   |   P_FRONTAXIS <= -36: 0 (13.07/0.0)
|   |   P_FRONTAXIS > -36
|   |   |   P_DURATION <= 98: 1 (7.05/1.04)
|   |   |   P_DURATION > 98
|   |   |   |   HEARTRATE <= 51: 1 (2.02/0.02)
|   |   |   |   HEARTRATE > 51: 0 (17.07/3.0)
|   P_FRONTAXIS > -19: 1 (3199.79/216.81)
HEARTRATE > 59: 0 (8697.0/201.0)

Number of Leaves  :     6

Size of the tree :  11
```

A

```
JRIP rules:
===========

(HEARTRATE <= 59) and (P_FRONTAXIS >= -17) => 8-8-0=1 (3175.0/199.0)
(RR_INTERVAL >= 1003) and (P_FRONTAXIS >= -30) => 8-8-0=1 (29.0/13.0)
 => 8-8-0=0 (8732.0/203.0)

Number of Rules : 3
```

B

Figura 13: Regras de classificação para o código 8-8-0. (A) Mostra as regras utilizadas pelo algoritmo J48 e (B) as regras utilizadas pelo JRip.

```
Time taken to build model: 0.34 seconds

=== Predictions on training set ===

inst#,     actual, predicted, error, probability distribution
     1        2:1        2:1          0.063 *0.937
     2        2:1        2:1          0.063 *0.937
     3        2:1        2:1          0.063 *0.937
     4        2:1        2:1          0.063 *0.937
     5        2:1        2:1          0.063 *0.937
     6        2:1        2:1          0.063 *0.937
     7        2:1        2:1          0.063 *0.937
     8        2:1        2:1          0.063 *0.937
     9        2:1        2:1          0.063 *0.937
    10        2:1        2:1          0.063 *0.937
    11        2:1        2:1          0.063 *0.937
    12        2:1        2:1          0.063 *0.937
    13        2:1        2:1          0.063 *0.937
    14        2:1        2:1          0.063 *0.937
    15        2:1        2:1          0.063 *0.937
    16        2:1        2:1          0.063 *0.937
    17        2:1        2:1          0.063 *0.937
    18        2:1        2:1          0.063 *0.937
    19        2:1        2:1          0.448 *0.552
    20        1:0        2:1      +   0.063 *0.937
    21        1:0        2:1      +   0.063 *0.937
```

Figura 14: Probabilidades de classificação individual. A coluna "*actual*" mostra a classificação do atributo. Classe 2, valor 1 (2:1) e Classe 1 valor 0 (1:0). A coluna "*predicted*" é o valor achado pelas regras do algoritmo. As classificações erradas são apontadas na coluna "*error*" com o símbolo +. A coluna "probability distribution" mostra a probabilidade de classificação das classes. O primeiro valor é a probabilidade para a classe 1, seguido pela probabilidade para a classe 2. No destaque encontra-se um caso onde a probabilidade de classificação é bem inferior à precisão geral do algoritmo.

4.2.2 Classificadores para o Código 5-3-0

Os resultados para o código 5-3-0 estão apresentados na tabela 11. A taxa global de acerto está acima de 90% porém com altas taxas de falsos negativos (atributos positivos para o código, classificados como negativo). Apesar da parente alta taxa de acerto a classificação deste código é inadequada para os objetivos propostos. Com o objetivo de testar a possibilidade de criar um classificador que separe exames alterados de normais, o foco dos resultados é a redução dos falsos negativos. Numa situação hipotética de implementação, um exame classificado como positivo ou alterado seguiria para revisão minuciosa de especialistas, minimizando o impacto do erro. Um exame classificado como negativo teria sua prioridade de revisão reduzida podendo gerar repercussões ao paciente. Para este código em particular, temos uma taxa de apenas 38% de positivos verdadeiros e ainda uma taxa de mais de 60% de falsos negativos, ou seja, 60% dos registro positivos foram classificados como negativos.

Tabela 11: Resultados do algoritmo JRip para a classificação do código 5-3-0. Coluna TP = taxa de positivos verdadeiros (true positive). FP = taxa de falsos positivos A matriz de confusão explicita quantas instancias foram corretamente classificadas para cada classe.

10 fold cross-validation para 5-3-0		
Atributos corretamente classificados	10807	90,5%
Atributos incorretamente classificados	1129	9,5%

tp rate	fp rate	precision	Area ROC	class
0,97	0,612	0,927	0,679	0
0,388	0,03	0,62	0,675	1
0,905	0,547	0,893	0,676	Média ponderada

Confusion matrix		
correto	incorreto	Classificado como
10291	316	negativo
516	813	positivo

4.2.3 Classificadores para o Código 7-6-0

Os resultados para o código 7-6-0 estão representados na tabela 12. A taxa de

acerto de global obtida está acima de 94% com baixas taxas de falsos positivos. A análise deste código revelou uma grande disparidade de resultados entre os algoritmos escolhidos. Apesar do desempenho de classificação ter sido estatisticamente igual, aproximadamente 94%, o conjunto de regras e parâmetros variou consideravelmente. Enquanto o algoritmo de árvores criou uma árvore com 22 folhas e tamanho total de 43, o algoritmo de regras utilizou apenas 3 regras para a classificação. Tanto a árvore quanto as regras são mostradas na figura 12 (pag. 31).

Tabela 12: Resultados do algoritmo JRip para a classificação do código 7-6-0. Coluna TP = taxa de positivos verdadeiros (true positive). FP = taxa de falsos positivos A matriz de confusão explicita quantas instancias foram corretamente classificadas para cada classe.

10 fold cross-validation para 7-6-0		
Atributos corretamente classificados	11252	94,3%
Atributos incorretamente classificados	684	5,7%

tp rate	fp rate	precision	Area ROC	class
0,945	0,075	0,989	0,93	0
0,925	0,055	0,698	0,93	1
0,943	0,073	0,954	0,93	Média ponderada

Confusion matrix		
correto	incorreto	Classificado como
9921	576	negativo
1331	108	positivo

4.2.4 Classificadores para o Agrupamento 8

Comparando os resultados do código 8-8-0 com seu respectivo agrupamento, podemos notar que houve um decréscimo de apenas 1% na taxa de acerto global. Os parâmetros de qualidade ainda indicam um bom classificador, com baixas taxas de falsos positivos e principalmente, baixa taxa de falsos negativos. A tabela 13 detalha os resultados. O algoritmo gerou 3 regras para o código 8-8-0 e para o agrupamento foram geradas 8 regras. O aumento das regras também era esperado devido à maior

complexidade dos dados. A figura 15 mostra o comparativo das regras geradas para o código e seu agrupamento.

Tabela 13: Resultados do algoritmo JRip para a classificação do agrupamento dos códigos 8. Coluna TP = taxa de positivos verdadeiros (true positive). FP = taxa de falsos positivos A matriz de confusão explicita quantas instancias foram corretamente classificadas para cada classe.

10 fold cross-validation para 8		
Atributos corretamente classificados	11396	95,5%
Atributos incorretamente classificados	540	4,5%

tp rate	fp rate	precision	Area ROC	class
0,973	0,091	0,964	0,936	0
0,909	0,027	0,93	0,936	1
0,955	0,073	0,955	0,936	Média ponderada

Confusion matrix		
correto	incorreto	Classificado como
8324	233	negativo
3072	307	positivo

A

```
JRIP rules:
===========

(HEARTRATE <= 59) and (P_FRONTAXIS >= -17) => 8-8-0=1 (3175.0/199.0)
(RR_INTERVAL >= 1003) and (P_FRONTAXIS >= -30) => 8-8-0=1 (29.0/13.0)
 => 8-8-0=0 (8732.0/203.0)

Number of Rules : 3
```

B

```
JRIP rules:
===========

(T_FRONTAXIS >= 72) and (T_FRONTAXIS >= 81) and (QRS_DURATION <= 118) => 5=1 (332.0/44.0)
(T_FRONTAXIS <= -9) and (T_FRONTAXIS <= -21) and (QRS_DURATION <= 116) => 5=1 (283.0/39.0)
(QT_DISPERSION >= 50) and (QRS_FRONTAXIS <= 28) and (T_FRONTAXIS >= 54) and (QRS_DURATION <= 116) => 5=1 (301.0/129.0)
(QRS_FRONTAXIS <= 41) and (T_FRONTAXIS >= 47) and (T_FRONTAXIS >= 60) and (QRS_DURATION <= 112) and (T_FRONTAXIS >= 71) and
(P_FRONTAXIS <= 59) => 5=1 (53.0/18.0)
(QRS_FRONTAXIS <= 41) and (T_FRONTAXIS >= 45) and (QRS_FRONTAXIS <= 20) and (QRS_DURATION <= 114) and (FREDERICHI_QTcH >= 416)
and (P_FRONTAXIS <= 57) and (QT_DISPERSION >= 52) => 5=1 (50.0/22.0)
(QRS_FRONTAXIS <= 41) and (T_FRONTAXIS >= 47) and (T_FRONTAXIS >= 57) and (QRS_DURATION <= 118) and (QRS_FRONTAXIS <= 28) and
(BAZETTS_QTcH >= 429) and (BAZETTS_QTcH <= 443) => 5=1 (60.0/25.0)
(T_FRONTAXIS <= 7) and (T_FRONTAXIS <= -7) and (QRS_DURATION <= 94) and (QRS_FRONTAXIS >= 13) => 5=1 (78.0/28.0)
 => 5=0 (10779.0/961.0)

Number of Rules : 8
```

Figura 15: Comparação das regras geradas para o código 8-8-0 e para o agrupamento 8. Em (A) regras do código 8-8-0. Em (B) regras do agrupamento.

4.2.5 Classificadores para o Agrupamento 7

Comparando-se os resultados de classificação entre o código 7-6-0 e seu

agrupamento, a taxa de acerto global caiu levemente de 94% para 92%. No entanto, os resultados do agrupamento mostram uma redução na taxa de falsos negativos (testes positivos classificados como negativos) e em contra partida, um aumento na taxa de falsos positivos (testes negativos classificados como positivos). Numa abordagem conservadora, onde se deseja reduzir a quantidade de erros, principalmente de falsos negativos, esses resultados são promissores. Um exame classificado como alterado sendo normal tem impacto muito menor que um exame alterado classificado como normal dentro dos objetivos deste trabalho. A queda de desempenho dos classificadores já era esperado, devido à maior complexidade dos dados.

Tabela 14: Resultados do algoritmo JRip para a classificação do agrupamento dos códigos 7. Coluna TP = taxa de positivos verdadeiros (true positive). FP = taxa de falsos positivos A matriz de confusão explicita quantas instancias foram corretamente classificadas para cada classe

10 fold cross-validation para o agrupamento 7		
Atributos corretamente classificados	10979	92%
Atributos incorretamente classificados	684	8%

tp rate	fp rate	precision	Area ROC	class
0,958	0,226	0,942	0,86	0
0,774	0,042	0,828	0,86	1
0,92	0,188	0,918	0,86	Média ponderada

Confusion matrix		
correto	incorreto	Classificado como
9062	398	negativo
1917	559	positivo

V Conclusões e Perspectivas Futuras

5. Conclusões e Perspectivas Futuras

5.1 Estudo de Caso PCM

Apesar das dificuldades de se trabalhar com uma base complexa como a PCM, as analises mostraram resultados promissores.

A classificação da forma cutânea da doença com 92,9% de precisão e uma baixa taxa de falsos positivos (7%) demonstra que é possível se aplicar com sucesso mineração de dados em bases relativamente pequenas e complexas.

As diferenças da manifestação da doença entre gêneros descritas na literatura foram observadas neste estudo. Os atributos ligados à diferenciação de gênero foram: “PCM mucosa”, “PCM pulmonar”, “PCM linfonodal”, “PCM cutânea”, “inchaço linfonodal”, “sopro sistólico aórtico” e “tempo de evolução”. Esses atributos explicitam as relações entre manifestação e progressão da doença e gênero, descritas na literatura (de MOURA, 2008; SHIKANAI-YASUDA et al., 2006; , SANTOS et al., 2003)

O modelo para classificação da recidiva ficou acima de 70%. e os atributos ligados à classificação foram: PCM intestinal; Contagem global de leucócitos; Tempo de tratamento; Tratamento com anfotericina b; Raio-X tórax e PCM disseminada. Esses achados podem guiar pesquisas futuras e ajudar a validar os parâmetros necessários para se prever a recidiva da doença.

Dentre os algorítimos utilizados nos testes, o j48 para construção de árvores de classificação/decisão foi que obteve os melhores resultados para a base PCM.

Os principais achados deste estudo de caso foram publicados na IEEE Healthcom 2014, sob o titulo: Medical Data Mining: a case study of a Paracoccidioidomycosis Patient’s Database. DOI:10.1109/HealthCom.2014.7001854, sendo selecionado para apresentação em conferência.

5.2 Estudo de caso ECG

A classificação com alta taxa de acerto dos códigos de Minnesota notadamente 7-6-0 e 8-8-0, e seus respectivos agrupamentos, indica que é possível se criar uma ferramenta de classificação para auxiliar médicos em grandes centros a classificar exames de ECG como normal ou alterado com eficácia. Os algoritmos geraram conjuntos de regras que, no geral são razoavelmente simples, o que facilitaria as possíveis adaptações necessárias para se representar variações intrínsecas populacionais, de gênero ou mesmo alterações individuais permanentes em paciente com histórico de doenças cardíacas. As regras encontradas estão de acordo com as características morfológicas classificadas pelo código de Minnesota.

Os dados disponíveis são escassos para vários dos códigos e sua classificação adequada pode depender de novas coletas.

A possibilidade de mostrar o grau de certeza do algorítimo para a classificação individual de cada exame nos permite a avaliação não somente do classificador como um todo, mas do exame testado. Esse traço pode ajudar na redução de erros de classificação, principalmente em abordagens mais conservadoras, evidenciando casos limítrofes, onde a certeza da classificação pode ser inferior à média do classificador.

O projeto ELSA é um projeto de acompanhamento de longo prazo. O incremento de dados gerados pelas ondas subsequentes tende a melhorar as analises. Os modelos gerados em uma onda podem ser testados e validados pelas ondas seguintes, gerando um grande potencial de pesquisas futuras.

Para o aprofundamento das analises é necessário estabelecer os parâmetros de normalidade populacional com os especialistas, bem como as variações aceitáveis entre gêneros e pacientes com históricos de eventos cardiovasculares para que os modelos de classificação sejam adaptados.

VI Referências

6. Referências

AQUINO EM, BARRETO SM, BENSENOR IM, CARVALHO MS, CHOR D, DUNCAN BB, et al. Brazilian Longitudinal Study of Adult Health (ELSA-Brasil): objectives and design. Am J Epidemiol. 2012;175(4):315-24. DOI:10.1093/aje/kwr294

BERMAN, J.J "Confidentiality Issues for Medical Data Miners," Artif Intell Med. Pp :25-36.,2002.

BERNER, E., "Clinical Decicion Support Systems". Springer Science+Business Media, 2007 .

BLACKBURN H, KEYS A, SIMONSEN E, et al. The electrocardiogram in population studies. A classification system. Circulation 1960;21:1160-75.

BREAULT, J.L, GOODALL, C.R. & FOS, P.J. (2002). Data mining a diabetic data warehouse. Artificial Intelligence in Medicine, 26(1), 37-54.

CIOS, K.J., MOORE, G.W. "Uniqueness of Medical Data Mining," Artif Intell Med. 26(1-2), 1-24,2002

COHEN, W, W: Fast Effective Rule Induction. In: Twelfth International Conference on Machine Learning, 115-123, 1995.

DE MOURA, A. C. L. (2008). Estudo Clinico e Imunologico de Controle de Cura de Paracoccidioidomicose Cronica. PhD thesis, Universidade Federal de Minas Gerais.

GREENLAND P, XIE X, LIU K, COLANGELO L, LIAO Y, DAVIGLUS ML, et al. Impact of minor electrocardiographic ST-segment and/or T-wave abnormalities on cardiovascular mortality during long- term follow-up. Am J Cardiol. 2003;91(9):1068-74.

GREENS, R., "Clinical Decision Support". Elsevier Inc., 2007.

GUYON, I. & ELISSEEFF, A. (2003). An introduction to variable and feature selection. J. Mach. Learn. Res., 3:1157--1182.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P. \& WITTEN, I. H. (2009). The weka data mining software: An update. SIGKDD Explorations, 11.

HALL, M.A., "Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning," Proc. 17th Int',l Conf. Machine Learning, pp. 359-366, 2000.

HAN, J. & KAMBER, M. (2001). Data mining: concepts and techniques. Morgan Kauf-mann Publishers, second edition

HASTIE, T.; TIBSHIRANI, R. & FRIEDMAN, J (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, second edition

KOH, H C, & TAN, G. "Data mining applications in healthcare." Journal of Healthcare Information Management. Vol 19.2 (2011): 65.

KORS JA, CROW RS, HANNAN PJ, RAUTAHARJU PM, FOLSOM AR. Comparison of computer-assigned Minnesota Codes with the visual standard method for new coronary heart disease events. Am J Epidemiol. 2000;151(8):790-7.

KORS JA, VAN HERPEN G. The coming of age of computerized ECG processing: can it replace the cardiologist in epidemiological studies and clinical trials? Stud Health Technol Inform. 2001;84(Pt 2):1161-7. DOI:10.3233/978-1-60750-928-8-1161

MACFARLANE PW, LATIF S. Automated serial ECG comparison based on the Minnesota code. J Electrocardiol. 1996;29(Suppl):29-34.

MACFARLANE PW, Minnesota coding and the prevalence of ECG abnormalities, Editorial, Heart 2000;84:6 582-584

MACFARLANE PW, OOSTEROM A, PAHLM O, KLIGFIELD P, JANSE M, CAMM A, editors. Comprehensive electrocardiology. 2.ed. London: Springer; 2011.

MACHADO DB, CROW RS, BOLAND LL, HANNAN PJ, TAYLOR JR HA, FOLSOM AR. Electrocardiographic fi ndings and incident coronary heart disease among participants in the Atherosclerosis Risk in Communities (ARIC) study. Am J Cardiol. 2006;97(8):1176-81.

PANIAGO AM, AGUIAR JI, AGUIAR ES, CUNHA RV, PEREIRA GR, LONDERO AT, et al. Paracoccidioidomicose: estudo clínico e epidemiológico de 422 casos observados no Estado do Mato Grosso do Sul. Rev Soc Bras Med Trop. 2003;36(4):455-459.

Portal da Saude - Ministerio da Saude - Governo Federal - Brazil http://portalsaude.saude.gov.br

PRINEAS RJ, CROW RS, ZHANG ZM. The Minnesota code manual of electrocardiographic findings. 2.ed. London: Springer; 2010.

QUINLAN, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

RANJIT ABRAHAM, JAY B.SIMHA, IYENGAR (n.d). A comparative analysis of discretization methods for Medical Datamining with Naive Bayesian classifier. 2006 IEEE. P1-2 DOI: 10.1109/ICIT.2006.5

RIBEIRO AL, PEREIRA SV, BERGMANN K, LADEIRA RM, OLIVEIRA RA, LOTUFO PA, MILL JG, BARRETO SM. Desafios à implantação do centro de leitura de eletrocardiografia no ELSA-Brasil / Challenges to implementation of the ECG reading center in ELSA-Brasil. Rev. Saúde Pública vol.47 supl.2 São Paulo, 2013

SHIKANAI-YASUDA MA, TELLES FILHO FQ, MENDES RP, COLOMBO AR, MORETTI MA. Consenso de paracoccidioidomicose. Rev Soc Bras Med Trop. 2006;39:297-310.

SHIKANAI-YASUDA, M. A.; TELLES FILHO, F. D. Q.; MENDES, R. P.; COLOMBO, A. L. \& MORETTI, M. L. (2006). Consenso em paracoccidioidomicose. Revista da Sociedade Brasileira de Medicina Tropical, 39:297 - 310.

SHULKA, D. P., PATEL, S. B. P & SEN, A. K "A literature review in health

informatics using data mining techniques." Int. J. Softw. Hardware Res. Eng. IJOURNALS (2014).

WANKE B, LAZER MS, CAPONE D. Paracoccidioidomicose. In: Sociedade de Pneumologia e Tisiologia do Estado do Rio de Janeiro, Aidé MA, editors. Pneumologia aspectos práticos e atuais. Rio de Janeiro: Revinter; 2001. p. 147-52.

WANKE, Bodo; AIDE, Miguel Abidon. Capítulo 6 - Paracoccidioidomicose. J. bras. pneumol., São Paulo, v. 35, n. 12, Dec. 2009 .

WITTEN.I; FRANK.E \& HALL.M (2011). Practical Machine Learning Tools and Techniques.Morgan Kaufmann Publishers, third edition

ZHANG ZM, PRINEAS RJ, EATON CB. Evaluation and comparison of the Minnesota Code and Novacode for electrocardiographic Q-ST wave abnormalities for the independent prediction of incident coronary heart disease and total mortality (from the Women's Health Initiative). Am J Cardiol. 2010;106(1):18-25. DOI:10.1016/j.amjcard.2010.02.007

ZHANG ZM, PRINEAS RJ, SOLIMAN EZ, BAGGETT C, HEISS G. Prognostic significance of serial Q/ST-T changes by the Minnesota Code and Novacode in the Atherosclerosis Risk in Communities (ARIC) study. Eur J Prev Cardiol. 2012;19(6):1430-6. DOI:10.1177/1741826711426091

ANEXO I –

Artigo Publicado:

Medical Data Mining: a case study of a Paracoccidioidomycosis Patient's Database

2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom

Medical Data Mining: a case study of a Paracoccidioidomycosis Patient's Database

Eduardo Liboredo Ferreira
and Herbert Rausch
and Sergio Campos
Department of Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais,
Brasil, 31270-901
Telephone: +55 (31) 3409-5566
Email: eduardoferreira@ufmg.br,
scampos@dcc.ufmg.br

Alessandra Faria-Campos
INMETRO
Xérem-Duque de Caxias, Rio de Janeiro,
Brasil, 25250-020

Enio Pietra
and Lilian da Silva Santos
Medical School
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais,
Brasil, 30130-100
Telephone: +55 (31) 3409-9300

***Abstract*—Data mining applied to medical databases is a challenging process. The unavailability of large sources of data and data complexity are some of the difficulties encountered. This is especially true for rare and neglected diseases. Those databases are, in general, relatively small, wide and sparse, making them very challenging to analyze. There are also ethical, legal and social issues regarding privacy and clinical validation of the findings. This work proposes a way of dealing with this challenge with a case study of data mining applied in a Paracoccidioidomycosis (PCM) patients database. Paracoccidioidomycosis (PCM) is a typical Brazilian disease, caused by the yeast *Paracoccidioides brasiliensis*. This disease represents an important Public Health issue, due to its high incapacitating potential and the amount of premature deaths it causes if untreated. This paper discusses methods for the analysis of this complex dataset, to help increase the understanding of both the disease and this type of data. Despite the challenges of the dataset, some interesting findings were made being: flaws in form filling protocols, notably the lack of chest X-ray in 40% of the records; the discovery of a possible new relation between smoking habits and PCM evolution time. The average evolution time for smoking patients was 2.8 times longer; the successful classification/prediction of the cutaneous form of the disease with a 93% precision rate are some of the discoveries made.**

I. Introduction

Data mining applied to medical databases is a challenging process. The unavailability of large sources of data and data complexity are some of the difficulties encountered. The examination protocols are, in general, complex and have several attributes. Different tests and exams are requested based on doctors personal experience and resource availability. Patients often fail to comply with the follow up procedures leaving the medical records incomplete. This tends to produce datasets that are difficult to analyze and require the use of multiple tools and techniques to be efficiently explored. There are also ethical, legal and social issues regarding privacy and clinical validation of the findings. This is especially true for rare and neglected diseases. Data mining use is, however, increasing important in clinical and health care fields. It can help health care insurers detect fraud and abuse, health care organizations make customer relationship management decisions, physicians identify effective treatments and best practices [7] and help researchers identify important disease traits for diagnose and treatment [2]. This work presents a case study of a Paracoccidioidomycosis database. This is one of the largest case studies for the disease in the world, the patient data corresponds to 35 years of data collection in the maximum endemicity region of Brazil.

Paracoccidioidomycosis (PCM) is a typical Brazilian disease, caused by the fungus *Paracoccidioides brasiliensis*. The most common form of infection is through inhalation of the mycelial form. It causes infection of the lung epithelium and can spread to other organs. The disease mainly affects farm workers who are exposed to contaminated soil during labor, but it is known to affect non farm workers as well [13]. This disease represents an important Public Health issue, due to its high incapacitating potential and the amount of premature deaths it causes if untreated [10]. The analysis and management of PCM related data presents several challenges. One of the challenges is related to data acquisition during patient evaluation and diagnosis. The Center of Training and Reference on Infeccious-Parasitary Diseases from the Federal University of Minas Gerais (CTR-DIP-UFMG) has developed a protocol for clinical analysis of PCM patients. This protocol includes a large number of clinical variables that are assessed in each medical examination, including x-ray and serology tests, which are also used in tracking the disease progression. Currently, there are no reliable clinical parameters for PCM to establish the treatment duration nor predict the disease relapse. The use of data mining techniques and Business Intelligence (BI), can help to find patterns and useful information otherwise invisible.

Several tools are available for data analysis. Some specialize in multi-dimensional analysis, others in data mining and others in statistical analysis. Researchers must often use different tools, under different environments and platforms (web, desktop) to obtain all the relevant data. In this work, we have used the tools Mondrian [8], WEKA [4] and R [12] to filter, select and use clustering and classification techniques on the information available to explore the correlation between its

978-1-4799-6644-8/14/$31.00 ©2014 IEEE

variables. The records used are composed of first examination data of patients diagnosed with PCM. A total of 227 patients charts with 314 items each compose the database.

II. Data Mining

Data mining can be defined as the computational process of discovering patterns in data and present them in an understandable and useful way [6]. It aims to uncover patterns and relations that are invisible by manual processing. In order to be useful, the raw data must be collected, pre-processed and stored in digital format. This section summarizes the data mining process used for the PCM database, introducing and explaining basic concepts. Details of the individual methods used are described in more details on each section following.

The collection of data, pre-process and storage is called Extraction, Transformation and Load (ETL). The ETL is one the most important steps in data mining [16], [6]. During the ETL, data from multiple sources are formated, standardized, and minor issues and human error when collecting data, such as the same value with different spellings, are corrected. Section *Database and ETL* describes the details of the process applied to the PCM dataset.

To optimize analysis, techniques for reducing the dimensionality of the database were employed. Through this process, irrelevant and redundant fields are eliminated, reducing computational costs and generally improves the quality of the analysis [6], [3]. To perform the reduction two methods have been used, *Frequency Distribution Analysis* and *Attribute Selection*.

For the analysis of frequency distribution, the tool Mondrian was used. Mondrian is an open source On line Analytical Processing server (OLAP) [8]. It is a versatile tool, offering a simple and intuitive interface and allows the user to easily visualize and navigate through the data. The frequency distribution shows the most common value of an attribute. It is important to retrieve this information because fields that have a highly dominant value tend to be irrelevant for clustering and classification, often introducing noise in the analysis [5]. The details are discussed in the subsection *Reduction of Dimensionality*.

The next step for dimensionality reduction was Attribute or Feature Selection. It is a process for finding the best subset of relevant features related to a class(attribute) for model construction. The models used in this work were clustering and classification trees, discussed later on. Two different approaches have been taken for this process . The first method was unassisted, with a progressive selection based on probability [9], to select the most relevant attributes on the dataset based on a statistical score. The results and detail are discussed in the respective subsection *Attribute Selection*. The models for the unassisted analysis were built using clustering algorithm k-means.

Clustering algorithms divides the instances, patients in this case, into natural groups, presumably revealing important attributes that separates them from the other groups [16]. This technique applies when there is no class or attribute to be predicted and it can highlight the most relevant attributes [5], [4]. The details and results are described in the section *Unassisted Analysis: Clustering*.

The second method for attribute selection is an assisted method. A set of attributes, called base attributes, have been selected by specialists as clinically relevant and used to further refine the dataset for analysis. Details on methodology and results are discussed in the section *Expert Assisted Analysis* and following subsection. For the assisted analysis, the construction of classification models has been performed, more specifically, classification trees were built.

Classification trees are a set of rules and steps that leads to the prediction of the possible value of a target attribute [16], [11]. A graphical representation of a decision tree is shown in III. The fundamental difference between these methods – clustering and classification trees — is that clustering is an undirected method for grouping and classification while the decision tree is directed, predicting the value of the intended, or target, attribute.

III. Database and ETL

One of the most important steps in data mining is the process called Extraction, Transformation and Load (ETL). During the ETL, the data is gathered, filtered and formated for the intended analytical tools. The raw database for this study contains the clinical data from 227 PCM patients made available by CTR-DIP-UFMG in SSPS format. The database consists of 314 attributes per patient, with 09 numeric type attributes (age, time of progression, RCD size, AX size, boyd size, lesion size, lesion area, inactivation time, treatment duration) and the remaining attributes are nominal.

Two databases have been prepared, one for the On Line Analytical Processing (OLAP) analysis and the second for data mining techniques, which includes attribute selection, clustering and building classification trees. The ETL process of the raw data has followed these steps: 1. Conversion of the SSPS files into CSV files; 2. Elimination of irrelevant fields to the analysis, such as patient personal information and protocol number; 3. Standardization of spelling in applicable fields. 4. Uploading the CSV into a MySQL table. This processes left the database with 301 attributes.

A. Reduction of Dimensionality

The Reduction of the dimensionality of the data by deleting unsuitable attributes improves the performance of learning algorithms and, more important, yields a more compact and easily interpretable representation of the target concept, focusing users attention on the most relevant variables [16], [3]. To reduce the dataset, elimination of redundant or irrelevant features have been conducted. Redundant features are those which provide no more information than the currently selected features and irrelevant features are those that provide no useful information in any context. The methods used to reduce dimensionality were *Frequency Distribution Analysis* and *Attribute Selection* [6], [3], discussed in more details in following sessions.

1) Frequency Distribution: The first step in reducing dimensionality was a frequency distribution analysis of the data, using the OLAP tool Mondrian [8]. The distribution of frequency analysis shows how many times a value is present in the records and gives a general picture of data distribution. Fields with a high concentration of a single value tend to be

TABLE I. FREQUENCY DISTRIBUTION: ATTRIBUTES, RESPECTIVE MOST COMMON VALUE AND IS FREQUENCY.

Attribute	Value	frequency %
PCM renal	normal	95.2
PCM pancreatic	normal	95.2
PCM splenic	normal	93
PCM genital	normal	94.7
PCM adrenal	normal	93.9
PCM intestinal	normal	93.9
PCM bone	normal	93
PCM gastric	normal	93.1
PCM liver	normal	92.1
PCM limphatic	normal	91.3
PCM neurologic	normal	90.4
positive Serological test	Unfilled	98.69
Boyd size (cm)	Unfilled	85.15
lesion area (cm^2)	Unfilled	78.6
facial X-ray	Unevaluated	93.5
altered Lactate	Unevaluated	92.1
altered CPK	Unevaluated	91.7
altered CK-MB	Unevaluated	90.8
positive HBeAg	Unevaluated	90.8
positive Anti-HBc	Unevaluated	90.5
altered Lipase	Unevaluated	88.2
positive CMV	Unevaluated	87.8
positive Anti-HIV (Elisa)	Unevaluated	87.3
positive Chagas	Unevaluated	87.3
altered Uric acid	Unevaluated	86.5
positive Anti-HAV	Unevaluated	86.34
altered Culture	Unevaluated	86.4
altered Spirometry	Unevaluated	86.5
positive Anti-HCV	Unevaluated	86.3
altered Amylase	Unevaluated	85.6
Anti-HBs positiv	Unevaluated	85.6

irrelevant or even detrimental to mining algorithms as they create noise [16]. Fields with more than 70% of missing values have been discarded. Different frequency ratios have been tested for elimination and the 70% threshold obtained the most relevant results. With a dataset as complex as the PCM, a careful approach while discarding fields had to be taken. Approximately 29% of the database used is composed of missing values. This accounts for 19422 of 66822 total fields. A total of 109 attributes (36% of 301 used) have missing values count above 30%. Setting the threshold too high for missing values would incur the risk of eliminating too many relevant fields.

An interesting observation is that of the 17 types of PCM tested in the examination protocol, 11 types are not significantly observed in the dataset as shown in table I. The table also shows other attributes with high concentration of a single value. From the 314 initial attributes, 52 were discarded, 262 remained for analysis. Only fields with missing values were discarded from the dataset. All 17 types of PCM were maintained.

2) Attribute Selection: Attribute selection or feature selection is a process for finding the best subset of relevant features related to a class(attribute) of the model. The feature selection algorithm selects a group of features that are closely related to a selected class and contains relevant information. Two different approaches have been taken for this process. The initial attribute selection was conducted with a progressive selection based on probability [9], to select the most relevant attributes on the dataset based on a statistical score. It is an unassisted method, one of the reasons why it has been chosen as the starting point. The second approach consisted in selecting attributes based on specialists opinions on which attributes were clinically relevant. This method is discussed in the session *Expert Assisted Analysis*.

TABLE II. CLUSTERING OF RELEVANT FIELDS FROM DATA SUBSET "LESION MUCOSA". FIELDS MARKED WITH "-" ARE UNEVALUATED, FIELDS MARKED WITH "Y" ARE ALTERED TESTS OR PRESENT SYMPTOM AND MARKED WITH "N" ARE ABSENT OR NORMAL.

Attribute	#0	#1	#2	#3	#4
	41(18%)	67(30%)	19(8%)	47(21%)	53(23%)
evolution time	11.8537	10.6567	17.7368	36.8298	5.2264
skin lesion	Y	N	Y	N	N
smoking	Y	Y	-	Y	N
pcm mucosa	Y	Y	Y	N	N
mucosa lesion	Y	Y	Y	N	N

Based on the initial unassisted selection, four attributes have been selected as the most relevant: "Gender", "Vomiting", "Skin Lesion" and "Mucosa lesion". Further attribute selection has been conducted using this four attributes as base to create a subset for cluster analysis. The data subset consists of the attribute used as base for the selection and the attributes given by the selection algorithm. For example, the complete subset for the "Mucosa lesion" is : "mucosa lesion" "evolution time", "skin lesion", "nasal obstruction", "difficulty swallowing", "smoking habits", "blood pressure while standing", "mouth lesion" and "pcm mucosa". This subset was used for clustering as shown in table II. The algorithm used for attribute selection was the Correlation-Based Feature Selection (CfsSubsetEval) [1] implemented on Weka. From this point on, all relevant findings have been sent to specialists for further analysis.

IV. ANALYSIS

A. Unassisted Analysis: Clustering

Clustering techniques apply when there is no class to be predicted but the instances are to be divided into natural groups [16], [4], [5]. In this work each subset of selected attributes was clustered using k-means, available in Weka. K-means is a well know algorithm that requires a definition of the number of clusters (K value) *a priori*. The initial value of K was obtained using the Expectation Maximization (EM) algorithm, which can give a good estimation of the optimal number of clusters [16], [6]. Further filtering was made in the initial clusters, to identify and remove irrelevant fields and optimize the k value. Some clusters showed correlation between attributes, which have been submitted to specialists for analysis. The clusters for the subsets "Gender", "Vomiting" and "Skin lesion" showed no relevant results. Other algorithms were used for clustering the subsets, DBScan and hierarchical clustering, but showed no relevant results.

Table II presents the most relevant correlation in the initial analysis. It shows a reduced scenario, considering only the five relevant fields, which are evolution time, skin lesion, previous smoking habits, PCM mucosa and mucosa lesion. It shows the difference in disease evolution time between smoking and non-smoking patients. The optimal number of clusters found (k value was 5).

The average evolution time for non-smoking patients(clusters 2 and 4) was 7.04 months and smoking patients got an average time of 19.83 months(clusters 0, 1 and 3), 2.81 times longer. Analyzing patients with PCM mucosa the relation is 7.36 months for non-smoking and 12.55 for smoking. There

is a described relation between PCM and smoking in [14], [15] but not directly to its progression. The field ”evolution time” represents for how long the patient contracts the disease and it continues to progress. As described before, the most common infection method for PCM is the inhaling of contaminated soil. It starts mostly as the pulmonary version of the disease. If untreated, it can spread to other organs, like liver, limphonodes and even the nervous system. This study suggests that smoking habits helps the disease progression. Further investigation is needed to determine the nature of the influence.

Although no other subset produced relevant results, one of the clusters of the analysis had, as one of the main traits, the lack of X-ray test results. In total, 40% of the patients had ”unevaluated“ for the chest X-ray parameter. This realization came as a surprise to specialists since the chest X-ray is one of the primary tests in examination protocol.

V. Expert Assisted Analysis

After the initial analysis, a directed approach was taken. Of the 262 attributes left after the frequency analysis, 23 have been chosen by the specialist staff as relevant attributes for further correlation investigations. Gender, age, disease relapse, rural contact, smoking habits, drinking habits and the 17 forms of the disease have been indicated. A similar method to the previous analysis has been performed, consisting of attribute selection using the indicated attributes as base and discarding the attributes considered irrelevant. A total of 112 attributes, each showed significant correlation to at least one of the 23 base attributes, remained for the analysis. The filtered database was then used to build decision or classification trees. It is worth to note that other approaches for classification were tested. All attributes were also classified using bayesian networks, notably the NaiveBayes and BayesNet algorithms implemented in Weka. The results for the trees were largely superior and are described in the following session.

A. Classification trees

The decision tree is a data mining technology suitable for performing classification and prediction. The decision tree can produce results according to different variables by repetition that can thus be used to analyze the characteristics, similarities and differences in data [11]. As mentioned before, classification trees are a directed approach used for the prediction of the possible value of a target attribute [16], [11] In this work, Weka’s integrated j48 tree algorithm has been used for the classifiers. It is based on the widely used C4.5 algorithm (for details, reffer to [11]). The j48 algorithm was chosen for its versatile characteristics. It classifies nominal attributes and handles numeric attributes as wells as missing values, avoiding the need to discretize numeric fields and impute data. The trees were built with the 10 fold cross-validation option, which consists in dividing the dataset in 10 parts, using 9 parts for learning and 1 part for testing, then rotating and using a different part for testing and a different set of 9 parts for learning, until every one of the 10 parts has been used for testing. The cross-validation method gives a better estimation of real world use of the rules set for building the tree. Using the whole data set for learning may give an overly optimistic performance indicator [16]. Figure 1 shows the graphical representation of the pruned classification tree used for the attribute PCM skin (cutaneous form of the disease). Pruned trees are simplified versions of the classification tree. They discard the branches that are less relevant, with less predictive power, and tend to reduce the noise of the analysis. They also provide a more readable and friendly visualization of the rules for classification. For comparison, the tree in the figure 1 has 4 leaves and total size of 7. The unpruned version of the same tree has 19 leaves and a total size of 31.

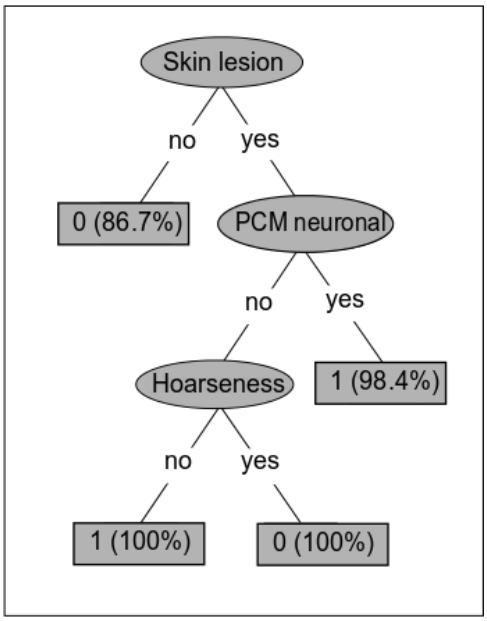

Fig. 1. Graphical representation of a pruned classification tree, built for the attribute PCM skin. The numbers that appear in the rectangles are the probable classification for the attribute PCM skin and its respective probability. 0 is negative for PCM skin and 1 is positive.

B. Classification Results

1) PCM skin: Table III shows the results for the pruned tree j48 tree, classifying the attribute PCM Skin(cutaneous). The first part shows the total number of correctly and incorrectly classified instances. The second part of the table displays more detailed information on the real quality of the classifier. In this case, it shows a 0.932 of true positives (tp rate) for class 0. This means that the instance was correctly classified as 0 in 93.2% of the tests. The false positives (fp rate) for class 0 is 0.072. This means that in 7.2% of the classification attempts, the instance was incorrectly classified as 0. The ideal scenario is a rate of 1 for tp and 0 for fp. The third part of the table is the *confusion matrix*. The matrix displays the number of correctly classified instances by class. The line for class ”negative” shows that 65 instances were correctly classified as negative for cutaneous PCM, and 5 were incorrectly classified as positive for cutaneous PCM. The line for class ”positive” shows that 128 instances were correctly classified as positive for pcm skin, and 10 were incorrectly classified as negative for cutaneous PCM. This results show that it is possible to predict the cutaneous form of PCM with high accuracy making it possible, for example, to build decision support system for this attribute, helping in the diagnosis and treatment of the disease.

TABLE III. PRUNED TREE USING THE J48 ALGORITHM FOR CLASSIFYING THE PCM SKIN ATTRIBUTE. THE CONFUSION MATRIX EXPLICITS THE INCORRECTLY CLASSIFIED INSTANCES AND ITS DISTRIBUTION. TP = RATIO OF TRUE POSITIVES; FP = RATIO OF FALSE POSITIVES.

10 fold cross-validation for attribute PCM skin		
correctly classified attributes	196	92.9%
incorrectly classified attributes	15	7.1%

tp rate	fp rate	precision	class
0.932	0.072	0.872	0
0.928	0.068	0.962	1
0.929	0.07	0.931	Weighted avg.

confusion Matrix		
correct	incorrect	classified as
65	5	negative
128	10	positive

2) Gender: Gender is an important attribute for PCM attribute correlation. The impact of female hormones in the progress of the disease, providing protection to women in reproductive age has been described in the literature [9], [15], [14]. Table IV shows the results for the pruned tree j48 tree, classifying the attribute *Gender*. The overall accuracy for gender is approximately 82%. But the accuracy for individual classes, 0 (males) and 1 (females), are very different. The precision for correctly classifying gender as 0 is 0.86 and the precision for classifying gender as 1 falls to 0.47 and the true positive rate to only 0.263. One of the factors that contributes to this variance is the composition of the dataset: 177 males and 38 females. In this case, using the full dataset as training improved the overall accuracy to almost 87% and the precision for classifying instances as 0(male) to 0.92 and 1(female) jumped to 0.625. The attributes selected by the classification algorithm for Gender classification rules are: "PCM mucosa", "PCM lungs" (pulmonary), "PCM limphonodes", "PCM skin", "swollen limphonodes" and "aortic systolic murmur". These results explicits the relation between disease type and manifestation and gender described in literature. It is important to note that both trees, 10-fold cross validation and full training set, used the same rules for classification.

TABLE IV. PRUNED TREE USING THE J48 ALGORITHM FOR CLASSIFYING THE GENDER ATTRIBUTE. THE CONFUSION MATRIX EXPLICITS THE INCORRECTLY CLASSIFIED INSTANCES AND ITS DISTRIBUTION. TP IS THE RATIO OF TRUE POSITIVES, FP FALSE POSITIVES.

10 fold cross-validation for attribute Gender		
correctly classified attributes	176	81.9%
incorrectly classified attributes	39	18.1%

tp rate	fp rate	precision	class
0.938	0.737	0.856	0
0.263	0.062	0.476	1
0.819	0.618	0.789	Weighted avg.

confusion Matrix		
correct	incorrect	classified as
166	11	male
10	28	female

3) Relapsing PCM: As previously discussed, there are no current clinical parameters to help predict the possible relapse of PCM. This is another attribute of clinical interest that showed promising results. Table V shows the correct classification was, overall, at 73.8%. The true positives for relapsing PCM (instance classified as 1) show a ratio of 0.932 and the overall precision at 0.71.

The attributes used as rules for classifying the relapsing disease are "intestinal PCM", "global leukocyte count", "treatment time", "treatment with amphotericin b", "chest X-ray" and "disseminated PCM". These attributes are important clinical features linked with disease progression and treatment, encouraging and possibly guiding further research for clinical validation. Since the data collection is still in progress, future studies on the subject may validate the parameters to safely predict a relapsing PCM.

TABLE V. PRUNED TREE USING THE J48 ALGORITHM FOR CLASSIFYING THE RECURRENCE OF PCM ATTRIBUTE. THE CONFUSION MATRIX EXPLICITS THE INCORRECTLY CLASSIFIED INSTANCES AND ITS DISTRIBUTION. TP = RATIO OF TRUE POSITIVES, FP = RATIO OF FALSE POSITIVES.

10 fold cross-validation for attribute Relapsing		
correctly classified attributes	152	73.8%
incorrectly classified attributes	54	26.2%

tp rate	fp rate	precision	class
0.241	0.068	0.583	0
0.932	0.759	0.758	1
0.738	0.564	0.709	Weighted avg.

confusion Matrix		
correct	incorrect	classified as
14	44	negative
138	10	positive

VI. CONCLUSION

Despite the difficulties in working with a complex, relatively small and sparse dataset, the results obtained by our analysis are promising. Both approaches, assisted and unassisted, revealed useful information. The discovery of a possible relation between smoking habits and disease progression encourages further research. The average evolution time for the disease progression was 2.8 times higher for smoking patients. During the clustering analysis, the lack of chest X-ray test appeared as a grouping attribute. This is a major test and 40% of the patients in the database do not have results registered. This information surprised and warned the specialists about errors in protocols and form filling procedures.

The successful classification of cutaneous PCM (PCM skin) with a high precision, 92.9%, and low false positive rates, 7%, demonstrate that data mining can be successfully applied on database with this characteristics.

The differences in the disease manifestation regarding gender described in the literature were also observed in this work. The main attributes for gender differentiation were "PCM mucosa", "PCM lungs" (pulmonary), "PCM limphonodes", "PCM skin", "swollen limphonodes" and "aortic systolic murmur". This attributes explicit the relation between disease type and gender described in literature [9], [15], [14] The model for classification of the relapsing PCM showed a precision above 70%. The attributes linked with relapsing are: "intestinal PCM", "global leukocyte count", "treatment time", "treatment with amphotericin b", "chest X-ray" and "disseminated PCM". These are important clinical features related to disease progression and treatment. These findings may help to guide further research, helping to uncover and validate the parameters for relapsing prediction. Data collection for PCM is still in progress. Despite the difficulties and complexity of the database, relevant information was obtained and we hope to encourage future research on the subject.

ACKNOWLEDGMENT

The authors would like to thank the medical staff of the Hospital das Clínicas - UFMG - for the cooperation and attention during this study.

REFERENCES

[1] M.A. Hall, "Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning," Proc. 17th Int',l Conf. Machine Learning, pp. 359-366, 2000.

[2] Breault, J.L, Goodall, C.R. & Fos, P.J. (2002). Data mining a diabetic data warehouse. Artificial Intelligence in Medicine, 26(1), 37-54.

[3] Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. J. Mach. Learn. Res., 3:1157–1182.

[4] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P. & Witten, I. H. (2009). The weka data mining software: An update. SIGKDD Explorations, 11.

[5] Han, J. & Kamber, M. (2001). Data mining: concepts and techniques. Morgan Kauf-mann Publishers, second edition

[6] Hastie, T.; Tibshirani, R. & Friedman, J (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction Springer, second edition

[7] Koh, Hian Chye, and Gerald Tan. "Data mining applications in healthcare." Journal of Healthcare Information ManagementVol 19.2 (2011): 65.

[8] Mondrian (2012). Mondrian. http://mondrian.pentaho.com/.

[9] de Moura, A. C. L. (2008). Estudo Clinico e Imunologico de Controle de Cura de Paracoccidioidomicose Cronica. PhD thesis, Universidade Federal de Minas Gerais.

[10] Portal da Saude - Ministerio da Saude - Governo Federal - Brazil http://portalsaude.saude.gov.br

[11] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[12] R (2012). R project. http://www.r-project.org/

[13] Restrepo, A.; McEwen, J. & Castaneda, E. (2001). The habitat of Paracoccidioides brasiliensis: how far from solving the riddle? Med. Mycol., 39:233241.

[14] Santos, W. A. d.; Silva, B. M. d.; Passos, E. D.; Zandonade, E. & Falqueto, A. (2003). Associação entre tabagismo e paracoccidioidomicose: um estudo de caso-controle no estado do espírito santo, brasil. Cadernos de Saúde de Pública, 19:245 253.

[15] Shikanai-Yasuda, M. A.; Telles Filho, F. d. Q.; Mendes, R. P.; Colombo, A. L. & Moretti, M. L. (2006). Consenso em paracoccidioidomicose. Revista da Sociedade Brasileira de Medicina Tropical, 39:297 310.

[16] Witten.I; Frank.E & Hall.M (2011). Practical Machine Learning Tools and Techniques.Morgan Kaufmann Publishers, third edition