UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
ESPECIALIZAÇÃO EM BANCO DE DADOS

RODRIGO JOSE FEUSER

MINERAÇÃO DE DADOS COM REGRAS DE ASSOCIAÇÃO APLICADA EM DADOS DE UNIDADE DE SAÚDE DE PRONTO ATENDIMENTO

MONOGRAFIA DE ESPECIALIZAÇÃO

PATO BRANCO
2017

RODRIGO JOSE FEUSER

MINERAÇÃO DE DADOS COM REGRAS DE ASSOCIAÇÃO APLICADA EM DADOS DE UNIDADE DE SAÚDE DE PRONTO ATENDIMENTO

Trabalho de Conclusão de Curso, apresentado ao II Curso de Especialização em Banco de Dados, da Universidade Tecnológica Federal do Paraná, campus Pato Branco, como requisito parcial para obtenção do título de Especialista.

Orientador: Prof. Dr. Richardson Ribeiro.

PATO BRANCO
2017

TERMO DE APROVAÇÃO

MINERAÇÃO COM REGRAS DE ASSOCIAÇÃO APLICADA EM DADOS DE UNIDADE DE SAÚDE DE PRONTO ATENDIMENTO.

por

RODRIGO JOSÉ FEUSER

Este Trabalho de Conclusão de Curso foi apresentado em 22 fevereiro de 2017 como requisito parcial para a obtenção do título de Especialista em Banco de Dados. O(a) candidato(a) foi arguido(a) pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

_____
Richardson Ribeiro
Prof.(a) Orientador(a)

_____
Dalcimar Casanova
Membro titular

_____
Marcelo Teixeira
Membro titular

"O Termo de Aprovação assinado encontra-se na Coordenação do Curso"

RESUMO

FEUSER. Rodrigo José. Mineração de dados com regras de associação aplicados em dados de unidade de saúde de pronto atendimento. 2017. 28 f. Monografia (II Curso de Especialização em Banco de Dados) - Universidade Tecnológica Federal do Paraná. Pato Branco, 2017.

Este trabalho tem por objetivo aplicar as etapas do processo da descoberta do conhecimento em um prontuário eletrônico de paciente oriundo de unidade de saúde da rede pública. A descoberta de doenças correlacionadas em prontuários, pode se tornar difícil ou demorada para o profissional da área da saúde se a quantidade de dados for imensa ou quando não houver a disponibilidade de sistemas computacionais especializados. Técnicas de mineração de dados, como por exemplo regras de associações, agrupamentos ou classificações, são alternativas para obter um sistema especializado que permite interpretar os dados usando grupos de pacientes. Para isso, nós usamos um conjunto de dados extraídos de um banco de dados de um prontuário eletrônico de usuários de sistema único de saúde, contendo 43.879 pacientes e 2.296.626 atendimentos. Resultados experimentais mostram que a associação de doenças e grupos de pacientes pode auxiliar os profissionais da área da saúde e gestores na aplicação de políticas de prevenção, podendo melhorar a qualidade de vida de uma região bem como a economia no tratamento de novos casos.

Palavras-chave: Mineração de Dados. Associação. Saúde.

ABSTRACT

FEUSER. Rodrigo José. Data mining with association rules applied to data from a health car unit. 2017. 28 f. Monography (II Specialization Course in Database) - Federal University of Technology - Parana. Pato Branco, 2017.

This work aims to apply the steps of the process of the discovery of knowledge in an electronic patient record from a public health unit. The discovery of correlated diseases in medical records can be difficult or time-consuming for the healthcare professional if the amount of data is immense or when there is no availability of specialized computer systems. Data mining techniques, such as rules of associations, clustering, or classifications, are alternatives for obtaining a specialized system that allows interpreting data using patient groups. To do this, we used a set of data extracted from a database of an electronic medical record of single health system users, containing 43,879 patients and 2,296,626 patients. Experimental results show that the association of diseases and groups of patients can help health professionals and managers in the application of prevention policies, which can improve the quality of life of a region as well as the economy in the treatment of new cases.

Palavras-chave: Data mining. Association. Health.

LISTA DE ABREVIATURAS E SIGLAS

ARFF	*Attribute-Relation File Format*
CID	Classificação Internacional de Doenças
CSV	*Comma-Separated Value*
DER	Diagrama de Entidade e Relacionamento
GPL	*General Public License*
JDBC	*Java Database Connectivity*
HTTP	*Hyper Text Transfer Protocol*
KDD	*Knowledge Discovery in Databases*
SGBD	Sistema Gerenciador de Banco de Dados
SQL	*Structured Query Language*
URL	*Universal Resource Locator*
UPA	Unidade de Pronto Atendimento
UTFPR	Universidade Tecnológica Federal do Paraná
WEB	*World Wide Web*
WEKA	*Waikato Environment for Knowledge Analysis*

LISTA DE FIGURAS

LISTA DE TABELAS

LISTAGENS DE CÓDIGOS

SUMÁRIO

1. INTRODUÇÃO

O uso dos prontuários eletrônicos de pacientes tem trazido as entidades de saúde a facilidade do gerenciamento dos dados, tais como informações dos pacientes, profissionais, atendimentos, etc. Atividades clínicas, como consultas, exames laboratoriais, prescrições médicas, diagnósticos, vacinações, entre outras, geram consequentemente, uma quantidade significativa de dados. Esses dados são normalmente usados para análises estatísticas, como quantidade de pacientes atendidos, tipos de convênios e controles financeiros.

Apesar da quantidade de sistemas computacionais disponíveis para o controle dessas atividades, poucos sistemas se preocupam com o uso de técnicas avançadas para melhorar a qualidade dos dados aos profissionais da saúde, como por exemplo, o uso de técnicas derivadas da mineração de dados.

Os dados dessas atividades quando agrupados, denotam um conjunto de dados geralmente na ordem de milhares. Uma possibilidade para obter o conhecimento implícito inerente a um relacionamento específico desses dados é aplicar as etapas do processo de descoberta de conhecimento (Knowledge Discovery in Databases - KDD), explorando principalmente os métodos de Mineração de Dados [9], [10]. Segundo [4], as principais aplicações da Mineração de Dados na saúde estão na efetividade de tratamentos médicos, gerenciamento de sistemas, além de detecção de uso indevido e ou fraudes de recursos destinados à saúde.

Na efetividade de tratamentos médicos, as aplicações de Mineração de Dados podem gerar informações que auxiliem o profissional da área da saúde nas tomadas de decisões como prescrições, exames e encaminhamentos. Ou ainda por meio da análise dos dados, pode-se chegar a conclusões sobre causas de uma doença, sintomas e tratamentos.

Neste trabalho, aplicamos as etapas do processo de descoberta de conhecimento em um prontuário eletrônico de pacientes de unidade de saúde da rede pública. O conjunto de dados extraídos do prontuário contêm 43.879 pacientes e 2.296.626 registros de atendimentos. Nessa base foram aplicadas as etapas como limpeza, seleção e transformação, mineração e interpretação e documentação dos dados. Na etapa de mineração de dados, foi utilizada a técnicas de regras de associações, por se tratar de uma técnica bem conhecida e adequada para o domínio de problema em questão.

Os resultados experimentais mostram que após realizar as etapas do KDD, usando um algoritmo gerador de regras por associação na etapa da mineração de dados, foi encontrado regras com elevado fator de confiança. Estas regras serão submetidas para análise de um especialista da área de saúde. Essa etapa não está inserida no escopo deste trabalho.

2. MINERAÇÃO DE DADOS E REGRAS DE ASSOCIAÇÃO

A mineração de dados ou data mining é o principal processo para realizar a descoberta de conhecimento em bancos de dados [9] [10]. Esse processo faz parte das etapas do KDD.

O processo de KDD [10] é iterativo e interativo, consistindo de nove passos [8] [10]. O processo é iterativo em cada etapa, o que significa que passos anteriores podem ser necessários, de modo que não existe uma definição completa de escolhas certas para cada aplicação. Assim é necessário compreender o processo e as diferentes necessidades e possibilidades para cada aplicação.

As tarefas de aprendizado descritivo, ou não supervisionado, se referem a identificação de informações relevantes nos dados sem a presença de um elemento externo para guiar o aprendizado [8]. Essas tarefas podem ser divididas em sumarização, que tem por objetivo encontrar uma descrição simples e compacta dos dados; em associação, que realiza a busca de padrões frequentes de associações entre atributos de um conjunto de dados e o

agrupamento, que lida com a identificação de grupos nos dados de acordo com a similaridade entre objetos.

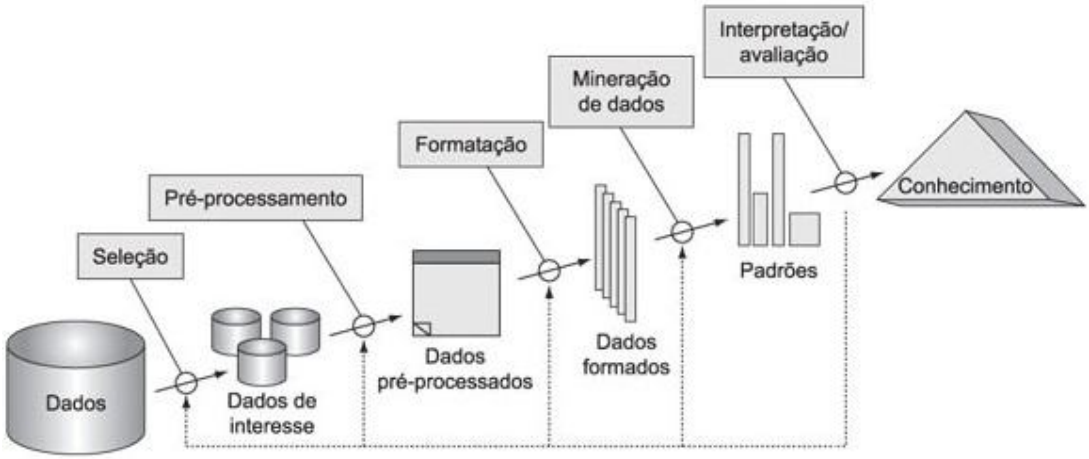


Figura 1. Visão dos passos de KDD.

As tarefas de aprendizado descritivo, ou não supervisionado, se referem a identificação de informações relevantes nos dados sem a presença de um elemento externo para guiar o aprendizado [8]. Essas tarefas podem ser divididas em sumarização, que tem por objetivo encontrar uma descrição simples e compacta dos dados; em associação, que realiza a busca de padrões frequentes de associações entre atributos de um conjunto de dados e o agrupamento, que lida com a identificação de grupos nos dados de acordo com a similaridade entre objetos.

As regras de associação devem evidenciar não apenas os conjuntos triviais de dados, mas também algumas relações que não estão aparentes, facilitando a tomada de decisão. A técnica de regras por associação procura itens que ocorrem de forma frequente no conjunto de dados [17].

Descobrindo-se dessa forma uma relação já consolidada ou no mínimo uma tendência. O objetivo, então, é encontrar associações relevantes entre os itens, do tipo *X* (antecedente) → *Y* (consequente) [7]. O grau de interesse é representado pela confiança das regras, dada pela equação da Listagem 1, que consiste na probabilidade de ocorrer um conjunto de termos dado que ocorreu um outro conjunto [8].

$$\text{confiança}(A \rightarrow B) = \frac{P(A \cup B)}{P(A)} = \frac{\text{suporte}(A \cup B)}{\text{suporte}(A)}$$

Listagem 1. Equação Confiança Algoritmo Apriori

O algoritmo Apriori, introduzido inicialmente por [2], foi o primeiro algoritmo para mineração de *itemsets* e regras de associação. Cada conjunto de dados é identificado com um número e é chamado de transação, onde o resultado do algoritmo é um conjunto de regras que informam a frequência com que os itens são relacionados no conjunto de dados.

Esse algoritmo procura identificar relações e dependências significativas entre os atributos. Trabalhando com as medidas de confiança e suporte, ele consegue classificar as melhores regras.

O algoritmo Apriori faz recursivas buscas no banco de dados à procura dos conjuntos frequentes (conjuntos que satisfazem um suporte mínimo estabelecido) [1]. Possui diversas propriedades que otimizam o seu desempenho, como por exemplo a propriedade de antimonotonia da relação, que diz que para um *itemset* ser frequente todo o seu subconjunto também devem ser, além de utilizar recursos da memória principal e estrutura *hash*, que fazem uma busca rápida e obtém um valor desejado.

2.1 REGRAS DE ASSOCIAÇÕES EM DADOS DA SAÚDE

Os algoritmos utilizados em sistemas de aprendizagem são normalmente baseados em diferentes áreas do conhecimento (e.g., matemática e estatística), e trabalham por meio de agrupamentos, classificação ou associação, possibilitando a criação de regras sobre os dados. O conhecimento é gerado a partir da interpretação dessas regras [4], exploradas para diversos fins, não apenas na saúde, mas também no comércio e indústria de bens de consumo.

A utilização de regras de associação em [6] disponibiliza o uso de técnicas de mineração de dados para a identificação automática de beneficiários propensos ao diabetes mellitus tipo 2, onde identifica que a quantidade de dados gerados pelas instituições de saúde pode conter um cenário generoso para a pesquisa de padrões de doenças e enfermidades que afetam o ser humano.

As técnicas de mineração de dados, se aplicadas em programas de promoção e prevenção da saúde podem auxiliar na identificação de necessidades da saúde dos usuários, bem como a organização dos serviços de saúde necessários para supri-las, melhorando a qualidade de vida da população, explicitado em [14] com o uso do data mining na promoção de saúde.

Para [13], que aplicou a mineração de dados com regras de associação em um protótipo de programa especialista na tentativa de descoberta do conhecimento associando informações de beneficiários á procedimentos médicos, clinica médica e caráter de internação em dados da Secretaria Municipal de Saúde de Londrina - PR, considera que os dados no seu formato original possuem benefício para os seus sistemas de origem, mas seu grande valor está na possibilidade de se extrair informações úteis para suporte de decisão ou exploração e compreensão do fenômeno de gerenciamento de fonte de dados e que dados analisados podem prover conhecimento adicional sobre um negócio, indo explicitamente além dos armazenados, para derivar conhecimento.

Os trabalhos de mineração de dados utilizam diferentes algoritmos para avaliar resultados, como os utilizados em [11], que verifica as vantagens ou não de modelos preditivos sobre os benefícios de diferentes tarefas de aprendizado supervisionado em tratamento a pacientes internados em Unidades de Terapia Intensiva, discutindo e analisando os resultados de árvores de decisão, Regras de Associação, *Random Forests*, Redes Neurais, redes Bayesianas, *Support Vector Machines* e Processo Gaussiano.

Alguns estudos como em [1], Identificação de Padrões em Registros de Doenças com Técnicas de Mineração de Dados utilizando regras de associação, contou com o apoio do Instituto Oswaldo Cruz, situado no município de Bambuí, em Minas Gerais, responsável pelo aprofundamento do conhecimento científico da Doença de Chagas, onde realizou o mapeamento, a partir do ano de 1986, de pacientes que apresentam a doença, o que permitiu ter várias informações consistentes, para que os padrões de doenças e áreas de risco sejam automatizados, aumentando, assim, a eficiência de auxílio e suporte a todos os envolvidos.

Ainda há outras aplicações para mineração de dados com regras de associação como em [17] Aplicação de Regras de Associação para Mineração de Dados na Web e [5] Mineração de Dados em Triagem de Risco de Saúde.

Embasado por estas considerações e na tentativa de elucidar o escopo deste trabalho de forma eficiente e objetiva, verificar-se-á a eficácia das regras de associação em *data mining* sobre a base de dados de atendimentos emergenciais em Unidade de Pronto Atendimento. Diversamente da bibliografia apresentada, pretende-se encontrar regras que possam identificar padrões de atendimentos realizados com base em prontuários eletrônicos preenchidos de forma correta e com classificação da doença identificada com base na Classificação Internacional de Doenças e Problemas Relacionados a Saúde [12] da Organização Mundial da Saúde (OMS), definida pela sigla CID.

3. METODOLOGIA

As ferramentas computacionais especializadas que foram utilizadas nesse projeto são: o sistema gerenciador de Banco de Dados PostgresSQL[1] [15], e o WEKA[2] [16].

O WEKA é um software que possui uma coleção de algoritmos de aprendizado de máquina que podem ser utilizados diretamente sobre um conjunto de dados [4] [6] [7] [14]. Aplicativo em tecnologia Java, foi desenvolvido pela Universidade de Waikato na Nova Zelândia e sob licença em padrão *GLP* é uma ferramenta *open source*.

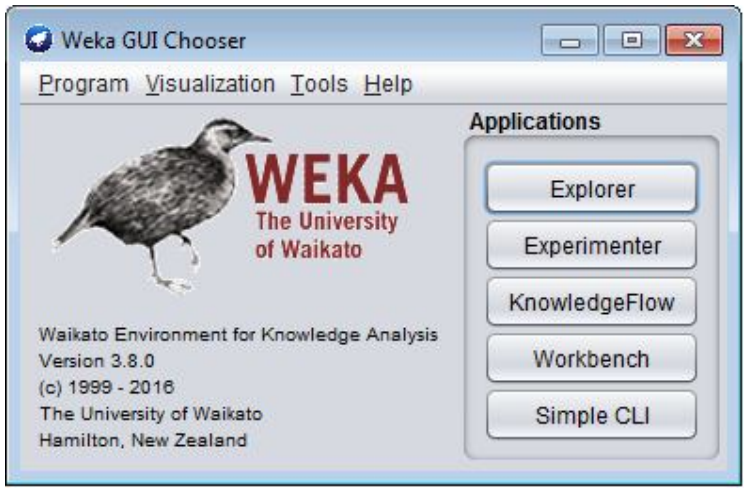


Figura 2. Software WEKA.

Os dados do prontuário eletrônico de pacientes utilizados neste trabalho foram disponibilizados por um projeto com a participação da Secretária Municipal de Saúde de Pato Branco, a Secretaria de Ciência e Tecnologia, a Universidade Tecnológica Federal do Paraná, Campus de Pato Branco, e a empresa desenvolvedora do sistema de prontuário (IDS Desenvolvimento de Software e Assessoria Ltda.). Dados pessoais, como nome, RG, CPF, telefone, e outros, não fazem parte do escopo desse projeto. Portanto, em momento algum, os pesquisadores desse projeto souberam a identificação de pacientes ou de seus responsáveis.

A base de dados foi entregue pelo fornecedor de software da Secretaria de Saúde em um backup para banco de dados PostgreSQL 9.5 64 bits. O arquivo foi restaurado utilizando a linha de comando da Listagem 1.

```
psql -h localhost -p 5432 -u "postgres" -f "file.backup" --d winsaude
```

Listagem 2. Código de linha de comando para restaurar o banco de dados.

[1] PostgreSQL is released under the PostgreSQL License, a liberal Open Source license, similar MIT licenses.
[2] Weka is open source software issued under the GNU General Public License.

A restauração da base de dados gerou o banco de dados winsaude com dois esquemas: público e winsaude. Dentro destes dois esquemas, apenas estavam presentes as tabelas e os índices utilizados pelas consultas do programa de manipulação dos dados que não foi disponibilizado para este trabalho.

A divisão de tabelas nos dois esquemas: público com 41 tabelas, incluindo 3 tabelas com o dicionário de dados do banco, ddcampos, ddtabela e ddcodfix, a fim de identificar os campos e tabelas. No Apêndice C, está o diagrama de entidade relacionamento (DER), mostrando o relacionamento entre as tabelas utilizadas na mineração.

O PostgreSQL fornece uma ferramenta para administração dos bancos de dados existentes, possibilitando realizar todo o tipo de tarefas desde a criação de *backups*, criação de bancos, funções, *triggers* e tabelas até a realização de simples consultas SQL.

Pela console de administração do SGBD PostgreSQL, foram realizadas todas as operações dentro do banco de dados restaurado.

A lista de tabelas utilizadas neste trabalho, foram selecionadas pela instrução da Listagem 3.

```
select * from ddtabela
where tabcodigo in (674,157,103,107,108);
```

Listagem 3. Instrução SQL para selecionar as tabelas utilizadas na mineração.

tabcodigo smallint	tabnomfis character varying(9)	tabdescri character varying(50)	tabchprim character varying(80)	tabsigla character varying(3)	tabforcad smallint
108	SASEXOS	Sexos	SEXCODIGO	SEX	0
107	SAUSUARI	Usuários	USUCODIGO	USU	1
103	SABAIDIS	Bairros e Distritos	BDICODIGO	BDI	1
157	SAFICATE	Fichas de Atendimentos	USUCODIGO;FATCODIGO	FAT	0
674	SASUSUSU	Informações de e-SUS dos Usuários	USUCODIGO	ESU	0

Tabela 1. Relação retornada pela consulta SQL da Listagem 3.

As tabelas utilizadas estão listadas na Tabela 1. Segue uma breve descrição de cada uma abaixo:

SAUSUARI: A tabela de usuários, possui os campos relativos aos pacientes, ou seja, usuários do sistema de saúde SUS. Possui relação de um para muitos atendimentos, ou prontuários, com a tabela SAFICATE. Já com as tabelas de Informações do SUS, bairros, distritos e sexo, a cardinalidade de um registro é um para um.

SABAIDIS: Traz as informações dos bairros e localidades dos usuários do sistema único de saúde. Os logradouros, ruas e numeração de residências não serão utilizados para não identificar o local de moradia dos pacientes, possui a relação de um para um com a tabela SAUSUARI. Neste caso os atendimentos estarão sempre com o último endereço do usuário, ou seja, relações de sintomas ou doenças não podem ser relacionadas ao local de moradia, pois as mudanças de endereço do usuário distorcem as estatísticas.

SASEXOS: Informa três tipos de situações para sexo dos usuários, masculino, feminino e indiferente.

SASUSUSU: As informações de e-SUS, que é um questionário de situação sócio-econômica do paciente, além de históricos de doenças e hábitos de vida dos usuários que se utilizam e do sistema de saúde no município.

SAFICATE: Esta tabela possui o prontuário eletrônico do paciente, usuário. Nela estão todos os detalhes do atendimento, desde a triagem pelos enfermeiros, passando pelo atendimento médico, medicação e alta ou encaminhamento para casa hospitalar.

4. RESULTADOS E DISCUSSÃO

A sequência utilizada de processamento e limpeza dos dados em estudo foram:

1. Criar uma tabela para receber os dados de atendimentos médicos, SAFICATE, com os campos identicos aos da tabela original identificada como SAFICATE2;

```
CREATE TABLE SAFICATE2 AS
SELECT * FROM SAFICATE;
```

Listagem 4. Instrução SQL para criar tabela SAFICATE2.

2. Inserir todos os dados de SAFICATE em SAFICATE2, (2.296.626 registros);

```
Insert into SAFICATE2 From (Select * from Saficate);
```

Listagem 5. Instrução SQL para inserir registros na tabela SAFICATE2.

3. Deletado da tabela SAFICATE2 os registros com CID NULL, sem identificação de doença, (1.639.420 linhas descartadas);

```
Delete From winsaude.saficate2 where cidprinci  is null;
```

Listagem 6. Instrução SQL para excluir registros com CID não informado.

4. Deletado da tabela SAFICATE2 os registros com CID Z00, doenças não identificadas, apenas sinalizam exames gerais, (231.923 linhas descartadas);

```
Delete From winsaude.saficate2 where cidprinci  = 'Z000';
```

Listagem 7. Instrução SQL para excluir registros com CID de investigação de dor.

5. Verificado registros com situação de inativos, (84.215 registros);

```
Delete From winsaude.saficate2 where fatsituac = 1;
```

Listagem 8. Instrução SQL para excluir registros inativos.

6. Verificado quantidade de registros na tabela SAFICATE2, (593.135 registros);

```
Select count(*) From winsaude.saficate2;
```

Listagem 9. Instrução SQL para verificar quantidade de registros na tabela.

Nesta fase do processamento foram encontrados 593.135 registros de atendimentos na Unidade de Pronto Atendimento. Ao agrupar os prontuários pela classificação internacional

de doenças foi encontrado 5.404 tipos de CID. Para ser possível correlacionar todas as doenças que o paciente foi atendido, o Postgres precisaria ter tabelas ou consultas com todas estas colunas, mas a limitação deste banco de dados é de 1.600 colunas para o tipo de dados String. Esta situação levou a criar um campo chamado cidgrupo, possibilitando agrupar os 5.404 cids em 256 grupos de patologias. Depois será possivel especificar os grupos através de instruções sql para melhorar a granularidade e poder descobrir situações mais especificas.

7. Realizar criação do campo CIDGRUPO na tabela SAFICATE2;

```
ALTER TABLE winsaude.saficate2 ADD COLUMN cidgrupo character varying(5);
```

Listagem 10. Instrução SQL para criar campo do grupo de CID.

8. Download da tabela dos grupos de CID10 para fazer a carga no campo CIDGRUPO, a relação dos grupos de CID constam no Anexo A;

```
http://www.sboc.org.br/downloads/CID-10-GRUPOS.xls;
```

Listagem 11. Endereço HTML para download dos grupos de CID.

9. Criação de função para ler os CID's do campo CIDPRINCI e encontrar o grupo correto para inserir no campo CIDGRUPO. Os CID's do prontuário estão inseridos em grupos que generalizam um pouco mais a doença, sem especifica-la;

```
CREATE OR REPLACE FUNCTION public.grupocid(cidprinci character varying)
    RETURNS character varying AS$BODY$
DECLARE
    i integer;
    grupo varchar(5);
BEGIN
     i = cast(substr(cidprinci,2,2) as integer);
    grupo = substr(cidprinci,1,1);
        if grupo = 'A' then
        if i < 15  then return 'A00'; end if;
                if i between 15 and 19 then return 'A15'; end if;
               if i between 20 and 29 then return 'A20'; end if;
                if i between 30 and 49 then return 'A30'; end if;
                if i between 50 and 64 then return 'A50'; end if;
                if i between 65 and 69 then return 'A65'; end if;
                if i between 70 and 74 then return 'A70'; end if;
                if i between 75 and 79 then return 'A75'; end if;
                if i between 80 and 89 then return 'A80'; end if;
                if  89 < i  then return 'A90'; end if;
            end if;
    if grupo = 'B' ... (repete todos os grupos);
```

Listagem 12. Função para identificar o grupo do CID.

10. Carga do campo CIDGRUPO;

```
Select grupocid(cidprinci) From SAFICATE2;
```

Listagem 13. Chamada da Função para inserir o grupo do CID no registro.

11. Criação dos índices da tabela Saficate2 para campos cidgrupo, código de usuário e código da unidade de atendimento;

```
CREATE INDEX idxwinfkcidgrupo
  ON winsaude.saficate2  USING btree
(cidgrupo COLLATE pg_catalog."default");

CREATE INDEX idxwinfkusafat2
  ON winsaude.saficate2 USING btree
  (usacodigo);

CREATE INDEX idxwinfkusufat2
  ON winsaude.saficate2USING btree
(usucodigo);
```

Listagem 14. Instrução SQL para criação de Índices na tabela.

O término do pré-processamento foi com a criação dos índices da tabela para acelarar as consultas e a geração dos dados nos campos críticos da tabela, realizando assim uma espécie de tunning na tabela SAFICATE2.

Para executarmos o próximo passo do KDD será necessário transformar os dados pré-processados em um arquivo em formato csv, com os dados separados por vírgulas, capaz de ser lido pelo software WEKA [16], realizar através do algoritmo apriori as associações que serão a base das regras para a descoberta do conhecimento.

O Postgresql [15], pode gerar qualquer consulta de dados em arquivos de texto com formatação específica, inclusive no formato csv. Foi elaborado uma consulta SQL com os campos das tabelas selecionadas que possuem dados relevantes e capazes de serem descricionados para uma boa análise associativa.

A consulta que gerou o arquivo csv, consta em parte no Apêndice A, com 9.717 registros e 352 atributos. No Apêndice B está parte do arquivo csv. Contendo faixa etária, sexo, alguns dados de senso do e-SUS, o bairro e o código do grupo de CID's de todas as doenças atendidas na unidade de atendimento de saúde, agrupada por usuário, paciente. O conjunto de atributos do arquivo gerado precisa ser descricionado em 's' , 'n' ou '?' para campos com valores indeterminados, assim atende a necessidade do algoritmo Apriori na identificação dos itens frequentes no conjunto de dados [13], [16] e [7].

No software WEKA, após selecionar a opção Explorer, é necessário carregar o arquivo csv com os dados gerados pelo PostgreSQL utilizando o botão *Open file*. Em seguida foi selecionado a opção *Associate* para selecionar o algoritmo de associação *Apriori*.

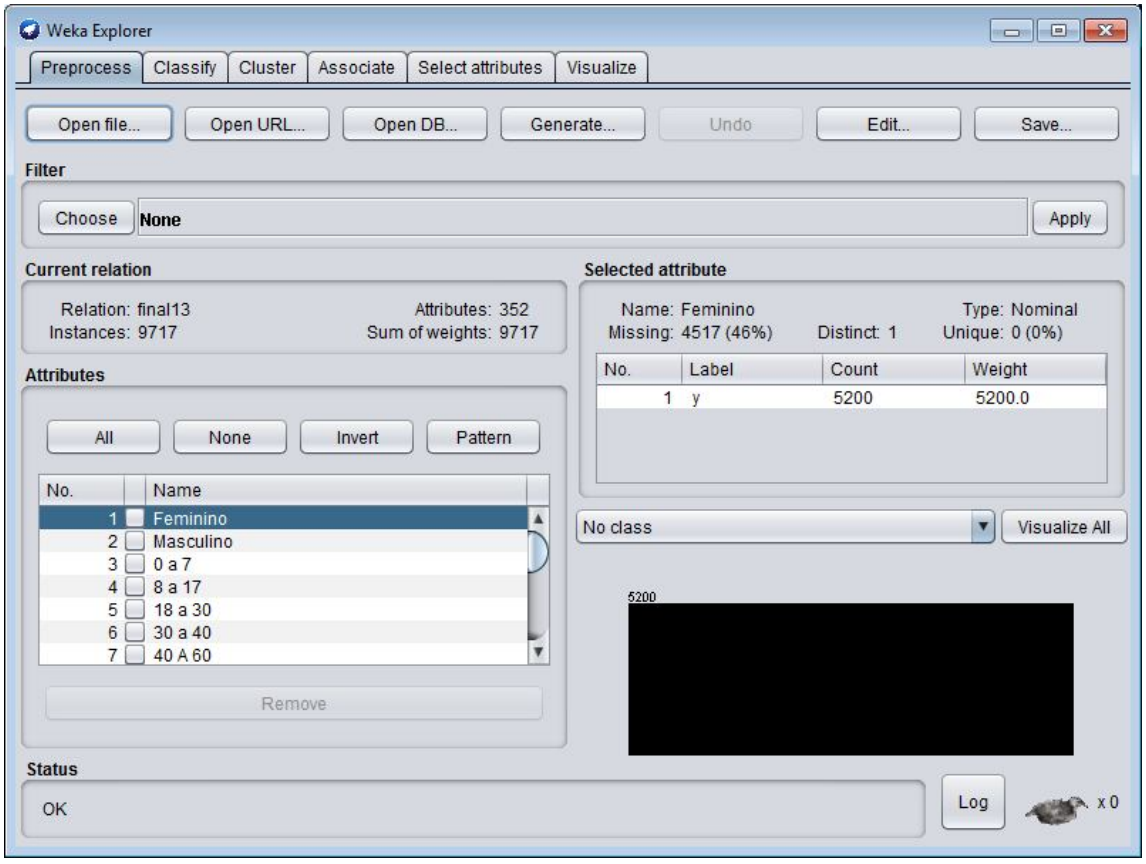


Figura 3. Carregar arquivo csv para análise do algoritmo de associação.

Algumas configurações com o algoritmo Apriori, como o suporte mínimo e a confiança mínima, podem ser necessários. O algoritmo realizará a busca de itens frequentes dentro de uma faixa de valores aceitável para a proposta.

O algoritmo Apriori originalmente em [3][14], não previa todos parâmetros de entrada utilizados pelo WEKA, nos interessando alguns apenas para este trabalho. São eles:

delta: que é valor inicial que será decrescido até atingir o suporte mínimo aceito em lowerBoundMinSupport, ira de 0,05 até 0,1, ou seja 10%.

metricType*:* O tipo de métrica a qual serão geradas as regras será *Confidence* ou confiança.

minMetric: O valor mínimo mais aceito na maioria dos trabalhos para regras de aceitação em um nível aceitável é 70% de confiança. Convencionou-se aceitar regras acima deste valor para a análise de um especialista validar a descoberta ou não de conhecimento.

numRules: o máximo de 100 regras, ou até o limite do percentual de 70% escolhido como regra aceitável.

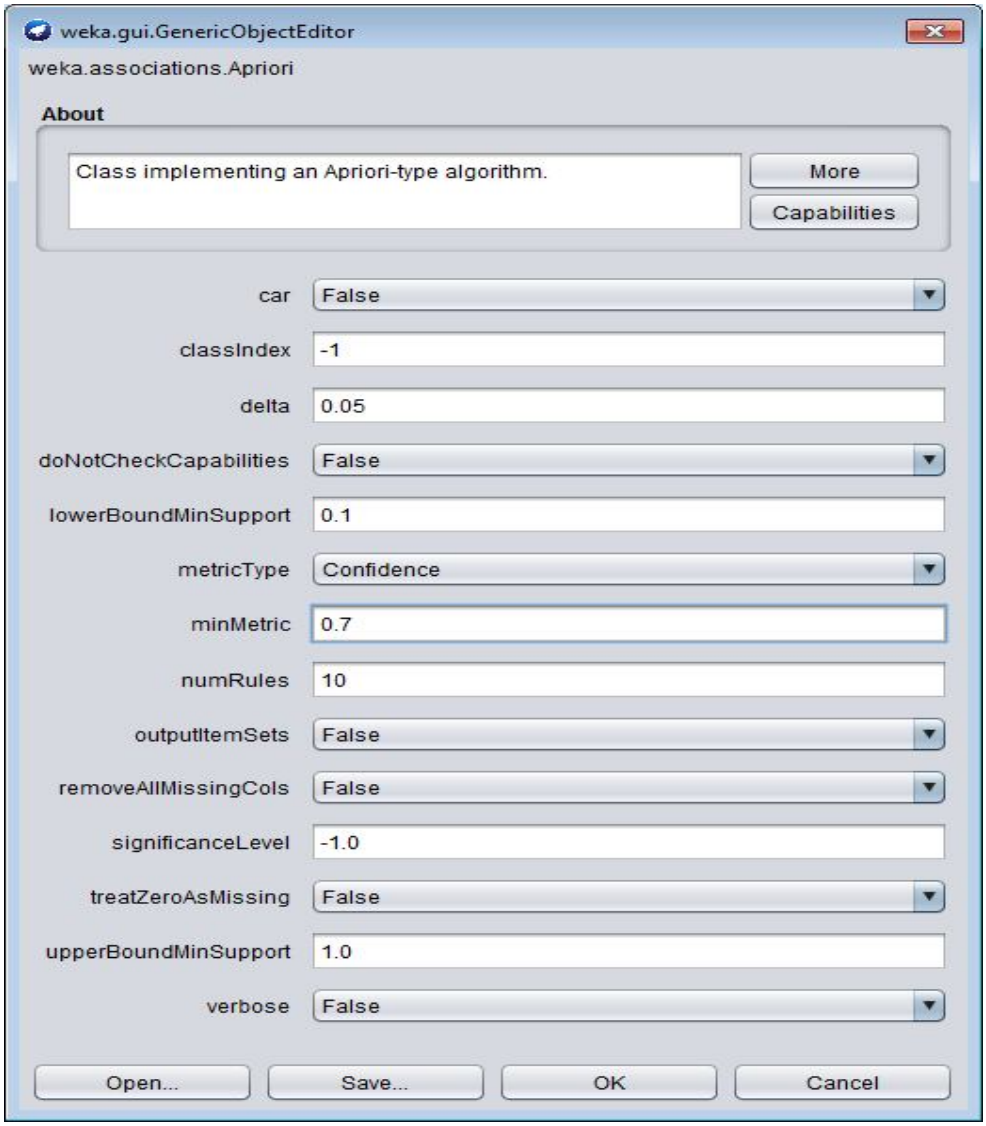


Figura 4. Configuração de parâmetros do algoritmo Apriori.

Após a configuração dos parâmetros do algoritmo Apriori, para dar inicio a busca das regras e dar continuidade ao processo do KDD, foi selecionado o botão *Start.*

Na sequência o resultado almejado foi encontrado pelo algoritmo Apriori, 25 (vinte e cinco) regras de associação com confiança entre 92% e 70%, após percorrer o arquivo csv por 18 ciclos, em 9.717 instâncias (linhas) e 352 atributos (colunas). As quais serão o objeto de discussão deste trabalho.

As regras geradas foram relacionadas na Tabela 2, em ordem decrescente de percentual de confiança, ou seja, da regra de maior confiança para a menor, até o limite estipulado de 70%, que segundo os trabalhos analisados durante a pesquisa trariam regras com maior probabilidade de trazer um novo conhecimento, como proposto pela mineração de dados.

Serão analisadas apenas as cinco primeiras regras, pois este trabalho sugere em seu escopo que as regras sejam validadas por um especialista da área de saúde.

Nº Regra	Antecedente	Consequente	Confiança
1	J20=y R50=y 1166	==> J00=y 1078	92%
2	Frequenta Escola=y R50=y 1181	==> J00=y 1041	88%
3	Masculino=y J20=y 1195	==> J00=y 1019	85%
4	0 a 7=y J20=y 1334	==> J00=y 1136	85%
5	J20=y 2338	==> J00=y 1965	84%
6	A00=y R50=y 1346	==> J00=y 1127	84%
7	H65=y 1436	==> J00=y 1201	84%
8	A00=y R10=y 1191	==> J00=y 990	83%
9	Feminino=y 0 a 7=y 1540	==> J00=y 1243	81%
10	Q00=y R10=y 1441	==> J00=y 1159	80%
11	0 a 7=y 3199	==> J00=y 2569	80%
12	Masculino=y 0 a 7=y 1659	==> J00=y 1326	80%
13	Q00=y R50=y 1830	==> J00=y 1443	79%
14	8 a 17=y 1890	==> J00=y 1490	79%
15	A00=y 2378	==> J00=y 1868	79%
16	Frequenta Escola=y 2299	==> J00=y 1804	78%
17	J09=y R50=y 1476	==> J00=y 1130	77%
18	R10=y R50=y 2228	==> J00=y 1676	75%
19	Q00=y R10=y 1441	==> R50=y 1069	74%
20	Q00=y 2811	==> J00=y 2085	74%
21	Feminino=y Q00=y 1540	==> J00=y 1128	73%
22	Masculino=y R10=y 1421	==> J00=y 1036	73%
23	Feminino=y R10=y R50=y 1399	==> J00=y 1019	73%
24	J09=y 2318	==> J00=y 1660	72%
25	Masculino=y R50=y 1940	==> J00=y 1363	70%

Tabela 2. Regras geradas pelo Algoritmo Apriori para Confiança 70%

O resultado da primeira regra evidencia que o Grupo de CID J20 (Outras infecções agudas das vias aéreas inferiores, que constam na relação do Anexo A), associado com o Grupo R50 (sinais como febre de origem desconhecida) em 1.166 registros, encontraram com confiança de 92% como consequente em 1.078 das vezes o grupo J00 (Infecções agudas das vias aéreas superiores).

A primeira regra mostra que o CID mais comumente tratado e assinalado nos prontuários eletrônicos, associam como consequente as doenças do grupo J00, que de modo geral engloba infecções agudas das vias aéreas superiores, o que corresponde ao tratamento de sinusite, faringite, amigdalite e laringite como as maiores causas de atendimento desta Unidade de Pronto Atendimento.

Na segunda regra encontramos correlação de 1181 ocorrências de pessoas que frequentam escola e diagnosticadas com CID do grupo R50 incorreram 1041 vezes em CID do Grupo J00 com 88% de confiança. Deixa evidente que pessoas com estes sintomas contaminam outras pessoas que compartilham o mesmo ambiente, sendo a escola um local de convivência por tempo prolongado facilitando a propagação de gripes, resfriados que podem levar a febre e infecções agudas das vias aéreas superiores.

Na terceira regra em 1195 incidências pessoas do sexo masculino que tiveram diagnóstico de CID do grupo J20 tiveram como consequentes com 85% de confiança em 1019

vezes doenças do grupo J00. Pessoas do sexo masculino são mais afetados que pessoas do sexo feminino neste grupo de indivíduos atendidos pelo pronto atendimento municipal.

Na regra 4, em 1334 vezes crianças em idade de zero a sete anos com sinais de febre são atingidas em 1136 vezes com as doenças do grupo J00, isso dá uma medida de 85% de confiança. Percebe-se que crianças nesta fase da vida estão mais suscetíveis a este tipo de doença, aumentando os atendimentos de Pediatria.

Na regra 5, com 84% de confiança, as doenças do grupo J20 em 2338 vezes tiveram como consequentes as doenças do grupo J00 em 1965 vezes. Esta regra pode não agregar novo conhecimento, pois ter infecções das vias aéreas inferiores, pode levar a ter também infecções aéreas nas vias superiores.

As próximas regras serão analisadas com o auxílio de um especialista para detectar a consistência ou não de todas as regras em um trabalho futuro.

4.1 RESULTADOS AUXILIARES

Para um segundo cenário, diminuiu-se a confiança para 50%, na tentativa de encontrar regras que gerem conhecimento em outras correlações de doenças que podem estar associadas. Já que os atendimentos desta unidade de pronto atendimento demonstram percentual de confiança mais baixo para sintomas de uma outra enfermidade que não seja as do grupo J00.

Com a confiança reduzida para 50%, o algoritmo Apriori encontrou para o conjunto de dados, 82 regras com associações talvez menos triviais que as regras com maior confiança, mas que para um especialista podem ser mais interessantes. Estas regras são apresentadas na Tabela 3.

No escopo deste trabalho foi proposto encontrar possíveis correlações entre grupos de doenças que podem ou não estar associadas, mas também buscou-se encontrar regras associando os grupos de CID de doenças com os dados de sexo, dados de senso do e-SUS e o bairro de moradia.

O algoritmo Apriori identificou nas regras da Tabela 3, que algumas vezes parecem ser triviais, recorrentes e até controversas, a regra 23: pessoas do sexo feminino com CID R10 (Sintomas e sinais relativos ao aparelho digestivo e ao abdome), com R50 (Febre) podem ter como consequente grupo J00 (Infecções agudas das vias aéreas superiores). Uma dor no abdômen poderia ser uma pneumonia.

Já a regra 36 identificou que dos 1890 usuários de 8 a 17 anos atendidos na unidade, 1222 deles frequentam a escola.

A regra 28, infere situações como J00 e Q00 (Malformações congênitas do sistema nervoso) em 2085 vezes, encontraram em 1443 consequentes no grupo R50.

Estas análises precisam ser verificadas por um especialista de saúde em atendimentos de unidades de pronto atendimento para verificar a validade das regras e se geram ou não alguma forma de conhecimento e aplicabilidade.

Como trabalho futuro sugestiona-se alterar a métrica do algoritmo Apriori de medida de confiança para Lift, levando em consideração a divisão da confiança da regra pelo suporte do consequente, não ficando tão implícito a quantidade de correlações existentes entre o antecedente e o consequente, mas sim se ocorre com alguma frequencia.

26	R50=y 4457	==> J00=y 3114	70%
27	Feminino=y R50=y 2517	==> J00=y 1751	70%
28	J00=y Q00=y 2085	==> R50=y 1443	69%
29	R10=y 3692	==> J00=y 2555	69%
30	J00=y J09=y 1660	==> R50=y 1130	68%
31	Feminino=y Q00=y 1540	==> R50=y 1042	68%
32	Feminino=y J00=y R10=y 1519	==> R50=y 1019	67%
33	Feminino=y R10=y 2271	==> J00=y 1519	67%
34	J00=y R10=y 2555	==> R50=y 1676	66%
35	Q00=y 2811	==> R50=y 1830	65%
36	8 a 17=y 1890	==> Frequenta Escola=y 1222	65%
37	J09=y 2318	==> R50=y 1476	64%
38	M50=y 1799	==> Feminino=y 1134	63%
39	R10=y R50=y 2228	==> Feminino=y 1399	63%
40	Feminino=y 5200	==> J00=y 3244	62%
41	Masculino=y 4517	==> J00=y 2817	62%
42	M50=y 1799	==> R50=y 1109	62%
43	Feminino=y R10=y 2271	==> R50=y 1399	62%
44	R10=y 3692	==> Feminino=y 2271	62%
45	J00=y R10=y R50=y 1676	==> Feminino=y 1019	61%
46	R10=y 3692	==> R50=y 2228	60%
47	A00=y J00=y 1868	==> R50=y 1127	60%
48	J00=y R10=y 2555	==> Feminino=y 1519	59%
49	Q00=y R50=y 1830	==> R10=y 1069	58%
50	Feminino=y J00=y R50=y 1751	==> R10=y 1019	58%
51	J09=y 2318	==> Feminino=y 1343	58%
52	8 a 17=y 1890	==> R50=y 1094	58%
53	J00=y J20=y 1965	==> 0 a 7=y 1136	58%
54	Frequenta Escola=y J00=y 1804	==> R50=y 1041	58%
55	J20=y 2338	==> 0 a 7=y 1334	57%
56	Q00=y R50=y 1830	==> Feminino=y 1042	57%
57	A00=y 2378	==> R50=y 1346	57%
58	R50=y 4457	==> Feminino=y 2517	56%
59	J00=y R50=y 3114	==> Feminino=y 1751	56%
60	J00=y Q00=y 2085	==> R10=y 1159	56%
61	Feminino=y R50=y 2517	==> R10=y 1399	56%
62	J00=y J20=y 1965	==> R50=y 1078	55%
63	Q00=y 2811	==> Feminino=y 1540	55%
64	J00=y Q00=y 2085	==> Feminino=y 1128	54%
65	Feminino=y J00=y 3244	==> R50=y 1751	54%
66	J00=y R50=y 3114	==> R10=y 1676	54%
67	J00=y 6061	==> Feminino=y 3244	54%
68	Frequenta Escola=y 2299	==> 8 a 17=y 1222	53%
69	A00=y J00=y 1868	==> R10=y 990	53%
70	8 a 17=y 1890	==> R10=y 985	52%
71	0 a 7=y 3199	==> Masculino=y 1659	52%
72	J00=y J20=y 1965	==> Masculino=y 1019	52%
73	0 a 7=y J00=y 2569	==> Masculino=y 1326	52%
74	J00=y 6061	==> R50=y 3114	51%
75	Frequenta Escola=y 2299	==> R50=y 1181	51%
76	Q00=y 2811	==> J00=y R50=y 1443	51%
77	Q00=y 2811	==> R10=y 1441	51%
78	J20=y 2338	==> Masculino=y 1195	51%
79	A00=y 2378	==> Feminino=y 1210	51%
80	Frequenta Escola=y 2299	==> Feminino=y 1157	50%
81	A00=y 2378	==> R10=y 1191	50%
82	J09=y 2318	==> R10=y 1159	50%

Tabela 3. Regras geradas pelo Algoritmo Apriori para Confiança 50%

5. CONCLUSÕES

Os processos de data mining estão cada vez mais em pratica na área da saúde, evidenciando sua necessidade de utilização, pois as massas de dados gerados neste setor são grandiosas e crescem exponencialmente ano a ano. A maior dificuldade é o acesso a pesquisa destes bancos de dados. Ainda não se dispõe de legislação ampla sobre a disponibilidade de prontuários eletrônicos padronizados, ou centralizados, o que facilitaria os trabalhos de mineração, mas também o acesso ao histórico de tratamento do paciente por outras unidades de saúde e médicos.

A relação de atendimentos de unidades de pronto atendimento, são muito peculiares devido ao fato de ser em geral procurado pelos usuários para tratamento de dores agudas, acidentes ou males repentinos. Neste caso as correlações de padrões encontradas pelo algoritmo Apriori, para doenças podem ser afetadas por vários fatores, mas podem inevitavelmente demonstrar fatores ainda não detectados em outras formas de visualizações.

As doenças do grupo J00, foram encontradas com o maior grau de confiança, devido a características e fatores do local da unidade de pronto atendimento que teve os dados analisados, como a localização geográfica, o clima, a situação sócio econômica dos usuários, período da analise dos dados, modo de vida da população, etc. Outras unidades provavelmente retornariam outras regras associadas, permitindo a análise das ocorrências locais e uma prevenção mais específica por localidades no combate das doenças.

A alta detecção de CID's do grupo J20 e J00 evidencia aparentemente que estes casos possuem maiores ocorrências e ficam evidentes quando utilizados valores de confiança elevado na parametrização das regras do algoritmo Apriori. Neste caso especifico desta Unidade de Pronto Atendimento, seria necessário baixar o grau de confiança ou utilizar formas diferentes de métricas para encontrar regras de associação com outras doenças associadas.

Uma das medidas que pode vir a ser utilizadas pelo algoritmo Apriori, seria o Lift que considera dividir o valor de confiança da regra pelo suporte do consequente. A geração destas regras podem melhorar a busca de conhecimento para casos de doenças com menos incidências. Essa verificação será objeto de estudo em um trabalho futuro.

Nas fases de preparação e transformação dos dados para aplicação do algoritmo, ficou claro a falta de padronização do preenchimento dos campos do prontuário eletrônico do paciente pelos operadores do sistema, causando uma eliminação muito grande de registros para a mineração, como registros sem o CID da doença. Uma causa pode ter sido, como visto em [6], em 2007, uma determinação do Conselho Federal de Medicina desobrigou a notificação do CID - Código Internacional de Doenças [12] nas guias dos atendimentos ambulatoriais.

Por outro lado o governo está obrigando as Unidades Básicas de Saúde a utilizarem o Prontuário Eletrônico de Paciente como forma de padronização na entrada e envio dos dados de atendimentos médicos. Esta medida pretende melhorar a fiscalização e em pouco tempo prover um banco de dados capaz de ser explorado para dar maior precisão a gestão dos recursos e as políticas de prevenção de doenças em todo o território nacional.

O próximo passo seria apresentar o resultado para um especialista em atendimento de unidade de saúde definir a veracidade das regras encontradas e assim terminar o ciclo do KDD para a descoberta de conhecimento. Isso poderia expandir este trabalho para outros locais de atendimentos de emergência, como: hospitais, ambulatórios e ou clinicas médicas.

6. REFERÊNCIAS

[1] 12th International Conference on Information Systems & Technology Management - CONTECSI. Identificação de Padrões em Registros de Doenças com Técnicas de Mineração de Dados. Disponível em *www.contecsi.fea.usp.br/envio/index.php/contecsi/12CONTECSI* Acesso em 22/01/2017.

[2] Agrawal, R. Imielinski, T. e Swami, A. Mining Association Rules between Sets of Items in Large Databases. In: ACM SIGMOD CONFERENCE ON MANEGEMENTOF DATA, Washington, DC, USA, 1993. ACM Press - New York, NY, USA.

[3] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. e Verkano, A. I. Advances in Knowledge discovery & data mining. Fast Discovery of Association Rules. AAAI/MIT, 1996.

[4] Borges, A. P., de Almeida Jr., J. L., Guisi, D., Lapa, H. S., Ribeiro, R., Souza, G. G. e Teixeira, M. Técnicas de Classificação em Problemas Relacionados a Doenças Cardíacas. WPCCG' 16, setembro 28, 2016, Ponta Grossa, Paraná, Brasil.

[5] Borges, E. N., Machado, K. S., Macieli, T. V. e Seus, V. R. Mineração de dados em triagem de risco de saúde. Revista Brasileira de Computação Aplicada (ISSN 2176-6649), Passo Fundo, v. 7, n. 2, p. 26-40, mai. 2015 26.

[6] Carvalho, D. R., Dallagassa M. R. e da Silva, S. H. Uso de técnicas de mineração de dados para a identificação automática de beneficiários propensos ao diabetes mellitus tipo 2. Inf., Londrina, v. 20, n. 3, p. 274 - 296, set./dez 2015. http://www.uel.br/revistas/informação/.

[7] Costa, R. B. R. Aplicação do Processo de Mineração de Dados para Auxílio à Gestão do Pronto-Socorro de Clínica Médica do Hospital Universitário de Brasília. UnB, 2007.82

[8] de Carvalho, A.C.P.L.F., Gama, J., Lorena, A. C. e Faceli, K. (2011) Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina. - Rio de Janeiro : LTC, 2011.

[9] Fayyad, U., Piatetsky-Shapiro, G. e Smyth, P. From Data Mining to Knowledge Discovery in Databases. Disponível em http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf. AI Magazine, American Association for Artificial Intelligence. Boston, 1996.Acessoem 13/01/17.

[10] Maimon, O. e Rokach, L. Introduction to Knowledge Discovery in Databases. Disponível em http://www.ise.bgu.ac.il/faculty/liorr/hbchap1.pdf. Dep. of Industrial Engineering. Tel-Aviv University. Israel 2009. Acesso em 13/01/2017.

[11] Meyfroidt Geert et al. Machine learning techniques to examine large patient databases. Amsterdam, v. 1, n. 23, p 127-143, 2009.

[12] Organização Mundial da Saúde (OMS). Centro Colaborador da OMS para a Classificação de Doenças em Português (CBCD). Classificação estatística Internacional de Doenças e Problemas Relacionados a Saúde - CID - 10. 2008. Disponível em http://www.datasus.gov.br/cid10/V2008/cid10.htm. Acesso em 02/02/2017.

[13] Silva, G. C. Mineração de Regras de Associação Aplicada a Dados da Secretaria Municipal de Saúde de Londrina -PR/ Glauco Carlos Silva – Porto Alegre: Programa de Pós-Graduação em Ciência Computação, 2005.

[14] Souza, A. M. P. e Zaia, J. E. O uso do Data Mining na Promoção de Saúde. Asa, São Paulo, v. 3, n. 2, p.12-21, Jan./Abr. 2015.

[15] The PostgreSQL Global Development Group. Disponível em:
https://www.postgresql.org/about/. Acesso em 22/01/2017.

[16] Weka 3: Data Mining Software in Java.
Disponível em http://www.cs.waikato.ac.nz/ml/weka/. Acesso em 22/01/2017.

[17] Vasconcelos, M. R. e Carvalho C. L. Aplicação de Regras de Associação para Mineração Technical Report - RT-INF_004-04 - Relatório Técnico Novembro - 2004.

ANEXO A

Tabela com a Relação de Grupos de CID

CID Inicial	CID Final	Descrição Abreviada
A00	A09	Doenças infecciosas intestinais
H65	H75	Doenças do ouvido médio e da mastóide
J00	J06	Infecções agudas das vias aéreas superiores
J09	J18	Influenza [gripe] e pneumonia
J20	J22	Outras infecções agudas das vias aéreas inferiores
J30	J39	Outras doenças das vias aéreas superiores
J40	J47	Doenças crônicas das vias aéreas inferiores
M40	M54	Dorsopatias
M40	M43	Dorsopatias deformantes
M45	M49	Espondilopatias
M50	M54	Outras dorsopatias
M60	M79	Transtornos dos tecidos moles
M60	M63	Transtornos musculares
N25	N29	Outros transtornos do rim e do ureter
P70	P74	Transt endocr e metabol trans espec feto e recem-nascido
P75	P78	Transt aparelho digestivo do feto ou recem-nascido
P80	P83	Afecc comprom tegument e reg term fet e recem-nascido
P90	P96	Outros transtornos originados no período perinatal
Q00	Q07	Malformações congênitas do sistema nervoso
Q10	Q18	Malform congênitas do olho, ouvido, face e pescoço
Q20	Q28	Malformações congênitas do aparelho circulatório
Q30	Q34	Malformações congênitas do aparelho respiratório
Q35	Q37	Fenda labial e fenda palatina
Q38	Q45	Outras malformações congênitas aparelho digestivo
Q50	Q56	Malformações congênitas dos órgãos genitais
Q60	Q64	Malformações congênitas do aparelho urinário
Q65	Q79	Malform e deformação congênita do sistema osteomuscular
Q80	Q89	Outras malformações congênitas
Q90	Q99	Anomalias cromossômicas NCOP
R00	R09	Sintomas e sinais relat aparelho circulatório e respiratório
R10	R19	Sintomas e sinais relat ao aparelho digest e abdome
R20	R23	Sintomas e sinais relat a pele e tecido subcutâneo
R40	R46	Sintomas e sinais rel cogn percep estado emoc e compor
R47	R49	Sintomas e sinais relativos a fala e a voz
R50	R69	Sintomas e sinais gerais
R70	R79	Achados anormais exames de sangue, sem diagnostico
R90	R94	Achados anormais exames diag imagem estud função, s/ diag
R95	R99	Causas mal definidas e desconhecidas de mortalidade
U99	U99	CID 10ª Revisão não disponível

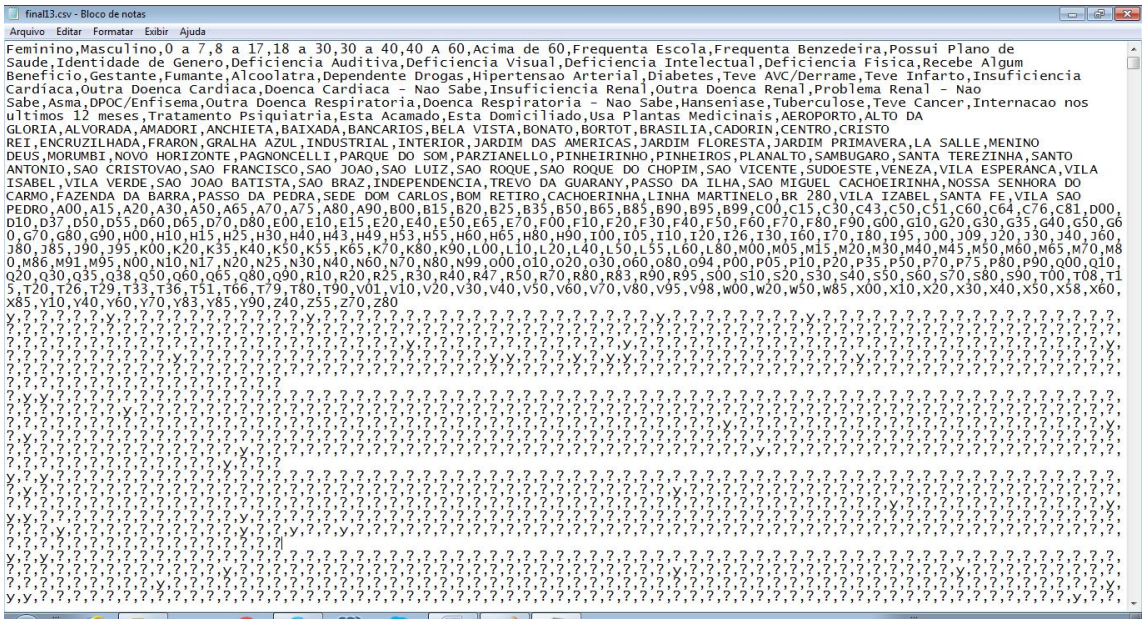
APENDICE A

Trecho da Consulta SQL que gerou o arquivo CSV.

```
Copy (
 SELECT
 --DEFINIÇÃO DE SEXO
CASE WHEN TIPOSEXO(F.USUCODIGO) = 'F' THEN 'y' ELSE '?' END  "Feminino",
CASE WHEN TIPOSEXO(F.USUCODIGO) = 'M' THEN 'y' ELSE '?' END  "Masculino",
 --FAIXA ETÁRIA
CASE WHEN FAIXAETARIA(F.USUCODIGO) = '0' THEN 'y' ELSE '?' END "0 a 7",
CASE WHEN FAIXAETARIA(F.USUCODIGO) = '1' THEN 'y' ELSE '?' END "8 a 17",
CASE WHEN FAIXAETARIA(F.USUCODIGO) = '2' THEN 'y' ELSE '?' END "18 a 30",
CASE WHEN FAIXAETARIA(F.USUCODIGO) = '3' THEN 'y' ELSE '?' END "30 a 40",
CASE WHEN FAIXAETARIA(F.USUCODIGO) = '4' THEN 'y' ELSE '?' END "40 A 60",
CASE WHEN FAIXAETARIA(F.USUCODIGO) = '5' THEN 'y' ELSE '?' END "Acima de 60",
 --CENSO
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO AND USU.ESUFREESC > 0) "Frequenta Escola",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO AND USU.ESUFRECUR > 0) "Frequenta Benzedeira",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO AND USU.ESUPLASAU > 0) "Possui Plano de Saude",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO AND USU.ESUORISEX > 0) "Identidade de Genero",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO AND USU.ESUDEFAUD > 0) "Deficiencia Auditiva",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO AND USU.ESUDEFVIS > 0) "Deficiencia Visual",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO AND USU.ESUDEFINT > 0) "Deficiencia Intelectual",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO AND USU.ESUDEFFIS > 0) "Deficiencia Fisica",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO AND USU.ESURECBEN > 0) "Recebe Algum Beneficio",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO AND USU.ESUGESTAN > 0) "Gestante",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO AND USU.ESUPESO   > 0) "Peso",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO AND USU.ESUFUMANT > 0) "Fumante",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO AND USU.ESUALCOOL > 0) "Alcoolatra",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO AND USU.ESUOUTDRO > 0) "Dependente Drogas",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO AND USU.ESUHIPART > 0) "Hipertensao Arterial",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO AND USU.ESUDIABET > 0) "Diabetes",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO AND USU.ESUAVCDER > 0) "Teve AVC/Derrame",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO AND USU.ESUINFART > 0) "Teve Infarto",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO AND USU.ESUINSCAR > 0) "Insuficiencia Cardiaca",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO AND USU.ESUOUTCAR > 0) "Outra Doenca Cardiaca",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'W50' ) "W50",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'W65' ) "W65",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'W75' ) "W75",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'W85' ) "W85",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'X00' ) "X00",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'X10' ) "X10",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'X20' ) "X20",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'X30' ) "X30",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'X40' ) "X40",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'X50' ) "X50",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'X58' ) "X58",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'X60' ) "X60",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'X85' ) "X85",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'Y10' ) "Y10",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'Y35' ) "Y35",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'Y40' ) "Y40",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'Y60' ) "Y60",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'Y70' ) "Y70",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'Y83' ) "Y83",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'Y85' ) "Y85",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'Y90' ) "Y90",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'Z40' ) "Z40",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'Z55' ) "Z55",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'Z70' ) "Z70",
(SELECT CASE WHEN COUNT(*) > 0 THEN 'y' ELSE '?' END FROM winsaude.safícate2 s where s.usucodigo = f.usucodigo AND  CIDGRUPO = 'Z80' ) "Z80"
    from winsaude.safícate2 f
    where f.usucodigo IN (select u.usucodigo FROM SASUSUSU u)
    group by  f.usucodigo
 )
 TO 'C:/final13.csv'
 DELIMITER ','
 CSV HEADER
```

APENDICE B

Trecho do Arquivo csv para importação no software WEKA.



APÊNDICE C

Diagrama de Entidade Relacionamento.