CENTRO UNIVERSITÁRIO UNIVATES
CURSO DE SISTEMAS DE INFORMAÇÃO

MINERAÇÃO DE DADOS PARA DESCOBERTA DE CONHECIMENTO NA ÁREA DE ONCOLOGIA

Fabrício Scheunemann

Lajeado, dezembro de 2015

Fabrício Scheunemann

MINERAÇÃO DE DADOS PARA DESCOBERTA DE CONHECIMENTO NA ÁREA DE ONCOLOGIA

Trabalho de Conclusão de Curso II apresentado ao Centro de Tecnologia da Informação do Centro Universitário UNIVATES, como parte dos requisitos para a obtenção do título de bacharel em Sistemas de Informação.

Orientador: Prof. Me. Fabrício Pretto

Lajeado, dezembro de 2015

Fabrício Scheunemann

MINERAÇÃO DE DADOS PARA DESCOBERTA DE CONHECIMENTO NA ÁREA DE ONCOLOGIA

Este trabalho foi julgado adequado para a obtenção do título de bacharel em Sistemas de Informação do CETEC e aprovado em sua forma final pelo Orientador e pela Banca Examinadora.

Prof. Fabrício Pretto, UNIVATES
Mestre pela PUCRS – Porto Alegre, Brasil

Prof. Mouriac Halen Diemer, UNIVATES
Mestre pela UFRGS – Porto Alegre, Brasil

Prof. Vilson Cristiano Gärtner,
UNIVATES
Mestre pela UNISINOS – São Leopoldo,
Brasil

Lajeado, dezembro de 2015

Seja você quem for, seja qual for a posição social que você tenha na vida, a mais alta ou a mais baixa, tenha sempre como meta muita força, muita determinação e sempre faça tudo com muito amor e com muita fé em Deus, que um dia você chega lá. De alguma maneira você chega lá.

(Ayrton Senna)

AGRADECIMENTOS

Ao Centro Universitário UNIVATES onde busco adquirir o título de Bacharel em Sistemas de Informação.

Agradecimento aos professores e colegas do curso Sistemas de Informação do Centro Universitário UNIVATES.

Agradecimento aos membros da banca examinadora, pela disponibilidade de participar e pelas contribuições acerca do trabalho.

Cabe um agradecimento especial à minha irmã Fernanda Scheunemann e aos meus pais Adelio Scheunemann e Noimi Scheunemann, lamentando muito que meu pai já tenha falecido, não estando hoje aqui para ver mais esta conquista.

À minha esposa, Patricia Damian Miotto e meu filho Gustavo Miotto Scheunemann, pelas alegrias que me proporcionam, pelo amor, carinho e principalmente compreensão.

À Casa de Saúde no qual este estudo foi realizado e aos colaboradores que, de forma direta ou indireta, contribuíram para a realização deste estudo.

Em especial ao meu amigo, professor e orientador Fabrício Pretto, pelo privilégio de sua orientação, disposição e auxílios oferecidos durante o desenvolvimento deste trabalho.

RESUMO

No mercado competitivo da atualidade, as organizações buscam qualificar seu gerenciamento e tomada de decisão a partir da análise das informações. O simples fato de armazenar e recuperar esta informação já proporciona um grande benefício às organizações. Contudo, apenas resgatar a informação não propicia todas as vantagens possíveis. As técnicas de mineração de dados permitem que se explorem grandes conjuntos de dados a fim de estabelecer relações, associações e descobrir padrões úteis que tenham valor para a organização com o propósito de se entender o fenômeno gerador dos dados. O presente trabalho expõe os conceitos de metodologias, técnicas e algoritmos de mineração de dados como fundamento teórico, bem como a aplicação do algoritmo de classificação Árvore Aumentada do Naïve Bayes (TAN) com a descoberta não supervisionada. Utilizou-se a ferramenta de mineração de dados WEKA com o intuito de descobrir conhecimento útil da especialidade médica de oncologia na base de dados de uma Casa de Saúde.

Palavras-chave: Mineração de Dados. Descoberta de Conhecimento. Oncologia.

ABSTRACT

In today's competitive market, organizations seek to qualify their management and decision-making based on the analysis of information. Simply store and retrieve this information already provides a major benefit to organizations. However, only retrieve the information does not provide every possible advantage. Data mining techniques allow us to explore large sets of data to establish relationships, associations and discover useful patterns that have value to the organization in order to understand the phenomenon generating the data. This paper presents the concepts of methodologies, techniques and data mining algorithms and theoretical foundation as well as the application of Tree Augmented Naive Bayes (TAN) classification algorithm the discovery unsupervised. Was used the data mining workbench WEKA in order to discover useful knowledge from the medical specialty of oncology in the database of a home of health.

Keywords: Data Mining. Knowledge Discovery. Oncology.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ARFF: Attribute Relation File Format

CENEPE: Centro de Ensino e Pesquisa

CID: Cadastro Internacional de Doenças

DM: Data Mining

DW: Data Warehouse

INCA: Instituto Nacional do Câncer

IMC: Índice de Massa Corporal

KDD: Knowledge Discovery in Databases

RHC: Registros Hospitalares de Câncer

SIRHC: Sistema para Informatização dos dados de Registros Hospitalares de Câncer

SUS: Sistema Único de Saúde

TAN: Tree Augmented Naive Bayes

WEKA: Waikato Environment for Knowledge Analysis

SUMÁRIO

1 INTRODUÇÃO

O rápido avanço na tecnologia de coleta e armazenamento de dados permitiu que as organizações acumulassem vasta quantidade de dados, principalmente na área da saúde. Muitas vezes, ferramentas e técnicas tradicionais de análise de dados não podem ser usadas devido ao tamanho do conjunto de informações serem muito grande, tornando-se necessário o desenvolvimento de novos métodos para análises de dados.

A mineração de dados é uma tecnologia que combina métodos tradicionais de análise com algoritmos sofisticados para processar grandes volumes de dados, com o objetivo de estabelecer relações, associações e descobrir padrões úteis que poderiam permanecer ignorados.

Na mineração de dados, o processo geral de conversão de dados brutos em informações úteis, é chamado de descoberta de conhecimento em banco de dados (KDD – *Knowledge Discovery in Databases*). Este processo consiste de uma série de passos de transformação, do pré-processamento ao pós-processamento dos resultados da mineração de dados (TAN; STEINBACH; KUMAR, 2009).

Segundo Carvalho (2005), a mineração de dados pode ser realizada de três diferentes formas em função do nível de conhecimento que se tenha do problema estudado. Se há pouco conhecimento, faz-se a descoberta não supervisionada; se há suspeita de alguma relação interessante, faz-se a testagem em hipótese; se há muito conhecimento, faz-se a modelagem matemática da relação.

Qualquer uma das três possíveis metodologias de mineração de dados necessita basicamente das mesmas técnicas para a sua realização. As técnicas são de caráter genérico e podem ser implementadas através de ferramentas diferentes como Árvores de Decisão,

Algoritmos Estatísticos, Algoritmos Genéticos, Regras de Decisão, Redes Neurais Artificiais, Redes Bayesianas e Lógica Fuzzy (REZENDE et tal , 2003).

Ao longo do tempo, percebeu-se que a velocidade de armazenamento das informações no setor da saúde era muito maior do que a velocidade de análise, o que gera um problema e uma contradição, pois as organizações, por possuírem vasta quantidade de dados, possuem uma falsa sensação de que estão bem informadas, porém essas informações precisam ser analisadas de forma correta e em tempo hábil.

Diante disso, este trabalho de pesquisa propõe identificar e analisar os dados para mineração de dados na descoberta de conhecimento na área de oncologia em uma Casa de Saúde, o estudo aonde será desenvolvido é uma instituição filantrópica de direito privado, sendo referência em diversas especialidades nas regiões do Vale do Taquari e Rio Pardo. Está inscrita nos conselhos Municipal, Estadual e Federal, sendo reconhecida como de utilidade pública e de extrema importância para a população.

1.1 Definição do Problema

O presente estudo buscou analisar os dados do centro de oncologia da Casa de Saúde, que atualmente enfrenta algumas dificuldades na extração de informações úteis para desempenhar de forma eficiente sua gestão.

Pensando nisso e levando em consideração o fato de estar inserido em uma Casa de Saúde e poder usufruir dos recursos disponibilizados pela instituição, a proposta do estudo visa avaliar a vasta quantidade de dados de forma otimizada através de técnicas de mineração de dados, a fim de contribuir para a gestão da oncologia.

1.2 Delimitação do estudo

O estudo está limitado à área temática de Inteligência Artificial, na qual o assunto abordado será a mineração de dados para descoberta de conhecimento em uma Casa de Saúde de médio porte, tendo como objeto de estudo dados da área de oncologia da instituição.

Esta pesquisa foi realizada no ano de 2015, tendo envolvido praticamente todo segundo semestre para a coleta dos dados, geração de informações, observações e conclusões.

1.3 Objetivos

O objetivo geral do trabalho é analisar uma base de dados de uma Casa de Saúde de médio porte do Vale do Taquari, com o intuito de utilizar técnicas de mineração de dados para descobrir conhecimento útil entre atributos da especialidade médica de oncologia. Para atingir o objetivo geral, as seguintes etapas se fazem necessárias:

- Adquirir uma compreensão das interações e processos de serviços de saúde desenvolvidos no centro de oncologia;
- Realizar um estudo de metodologias e algoritmos de mineração de dados como fundamento teórico para a sua aplicação prática sobre dados reais;
- Identificar padrões que expliquem e magnifiquem o entendimento dos problemas abordados;
- Selecionar as técnicas de mineração de dados apropriadas para descoberta de conhecimento;
- Avaliar os resultados obtidos com o propósito de se entender o fenômeno gerador dos dados.

1.4 Justificativa

No Brasil os registros estatísticos sobre o câncer ainda são bastante falhos, e não retratam a realidade brasileira. Nos últimos anos houve uma tentativa de dar maior confiabilidade aos dados nacionais divulgados pelo INCA com o sistema SISRHC (INCA, 2002).

O câncer é a segunda maior causa de morte no Brasil (superadas apenas por doenças cardiovasculares - como infarto e hipertensão). Estima-se que no meio do século 21 o câncer já seja a principal causa de morte no Brasil. Os motivos que levam ao crescimento da incidência do câncer são o aumento da expectativa de vida da população em geral, associada à maior exposição a fatores de risco (INCA, 2013).

Este trabalho justifica-se pela necessidade de descobrir conhecimento implícito na base de dados da Casa de Saúde da especialidade médica de oncologia utilizando técnicas de mineração de dados.

Para a instituição pesquisada este estudo é relevante, pois vem ao encontro de suas necessidades em melhorar a gestão da oncologia, podendo desenvolver planos de ações para adequar à gestão conforme sua realidade.

Já para o acadêmico, a área temática de estudo é importante devido a sua afinidade com o assunto e os fatos vivenciados pelo Centro de Saúde em busca de informações para gestão. Além disso, será uma oportunidade para exercer a prática dos ensinamentos científicos estudados ao longo da graduação, podendo, desta forma, aprimorar seu conhecimento na área através das informações buscadas para a realização do presente estudo.

Este estudo ainda poderá ser relevante para a sociedade em geral, podendo ser utilizado como referência para demais acadêmicos, comunidade e empresas que possam se interessar pelo assunto.

Por fim, para a instituição de Ensino, o estudo mostra-se importante, pois através da qualificação profissional dos seus alunos vem desenvolvendo habilidades e competências, tornando-os capazes de identificar situações problemáticas e do mesmo modo encontrar a solução mais adequada.

1.5 Estrutura do Trabalho

O presente trabalho tem divisões estabelecidas na forma de capítulos, sendo, após essa introdução, apresentado no segundo capítulo o referencial teórico, cujo objetivo é explanar os conceitos utilizados para construção do conhecimento necessário para elaboração e entendimento do estudo de caso proposto. No terceiro capítulo é apresentada a metodologia utilizada para realização do presente trabalho. No quarto capítulo é apresentada a caracterização da empresa utilizada como estudo de caso. No quinto capítulo são apresentados os trabalhos relacionados do assunto proposto. No sexto capítulo são apresentados os resultados do estudo. Por fim, no sétimo capítulo são apresentadas as conclusões do estudo e as recomendações para trabalhos futuros.

2 REVISÃO DE LITERATURA

Nesta seção é apresentada a base teórica fundamentada nos principais conceitos relacionados à oncologia para mineração de dados com finalidade de oferecer um embasamento teórico para o estudo.

2.1 Oncologia

Segundo PONTES (2013), a oncologia é a especialidade médica que estuda os tumores, procura compreender como o câncer se desenvolve no organismo e qual o tratamento mais adequado para cada caso.

A oncologia prega um tratamento multidisciplinar, envolvendo (médicos oncologistas, cirurgiões, radiologistas, radioterapeutas, patologistas, enfermeiros, nutricionistas, fisioterapeutas, assistentes sociais, psicólogos entre outros) devido à enorme complexidade da doença e das mais diversas opções terapêuticas, visando ao bem estar do paciente em tratamento e de seus familiares.

Apesar da existência de protocolos médicos, o tratamento oncológico é constantemente individualizado, pois cada paciente, tumor e situação exigem uma abordagem terapêutica.

2.1.1 Câncer

O câncer ou neoplasia maligna é atribuído a um conjunto de mais de cem doenças que têm em comum o crescimento desordenado (maligno) de células que invadem os tecidos e

órgãos, podendo espalhar-se (metástase) para outras regiões do corpo (INCA, 2002; Pérez-Tamayo, 1987; Robbins, 1984).

Segundo INCA (2014), o câncer representa atualmente uma ameaça crescente para a saúde no mundo. É a doença que mais cresce e uma das maiores causas de morte, ao lado das doenças cardiovasculares. São mais de 8 milhões de casos novos a cada ano no mundo, um aumento de quase 40% nos últimos 20 anos.

No Brasil, a estimativa para o ano de 2015, aponta para a ocorrência de aproximadamente 576 mil casos novos de câncer, incluindo os casos de pele não melanoma. Sem considerar os casos de câncer de pele não melanoma, estimam-se 395 mil casos novos de câncer, 204 mil para o sexo masculino e 190 mil para sexo feminino (INCA, 2014).

Conforme Organização Mundial da Saúde (OMS) prevê que, em 2030, cerca de 22 milhões de pessoas, entre homens, mulheres e crianças, serão diagnosticadas com câncer por ano e 13 milhões morrerão da doença.

2.1.2 Causas do Câncer

Segundo INCA (2002), as causas de câncer são variadas,

> [...] podendo ser externas ou internas ao organismo, estando ambas inter-relacionadas. As causas externas relacionam-se ao meio ambiente e aos hábitos ou costumes próprios de um ambiente social e cultural. As causas internas são, na maioria das vezes, geneticamente pré-determinadas, estão ligadas à capacidade do organismo de se defender das agressões externas. Esses fatores causais podem interagir de várias formas, aumentando a probabilidade de transformações malignas nas células normais.

Ainda segundo INCA (2002), a menor parte dos casos é relacionada a fatores internos (herança genética), que tornam o organismo incapaz de se defender de uma ameaça. Entre 80% e 90% de todos os cânceres estão associados a fatores externos, dessa forma, a adoção de comportamentos e hábitos saudáveis pode minimizar muito o risco e a ocorrência do câncer.

2.1.3 Estadiamento Geral do Câncer

Segundo SASSE (2008), o estágio é um sistema de classificação baseado na extensão anatômica aparente do câncer. A classificação ajuda a definir o plano terapêutico e o

diagnóstico em cada paciente, sendo dividido em quatro estágios, sendo o estágio I o mais precoce e o estágio IV o mais avançado. Para definir o estágio do tumor utiliza-se a classificação TNM dividido em três componentes básicos:

- A letra (T) significa a extensão do tumor primária;
- A letra (N) significa ausência ou presença de metástases nos linfonodos;
- A letra (M) significa ausência ou presença de doença metastática à distância;

As neoplasias malignas hematológicas não são classificadas pelo sistema TNM. As leucemias são classificadas segundo o tipo celular e diferenciação, mas não são estagiadas.

2.1.4 Modalidades Terapêuticas

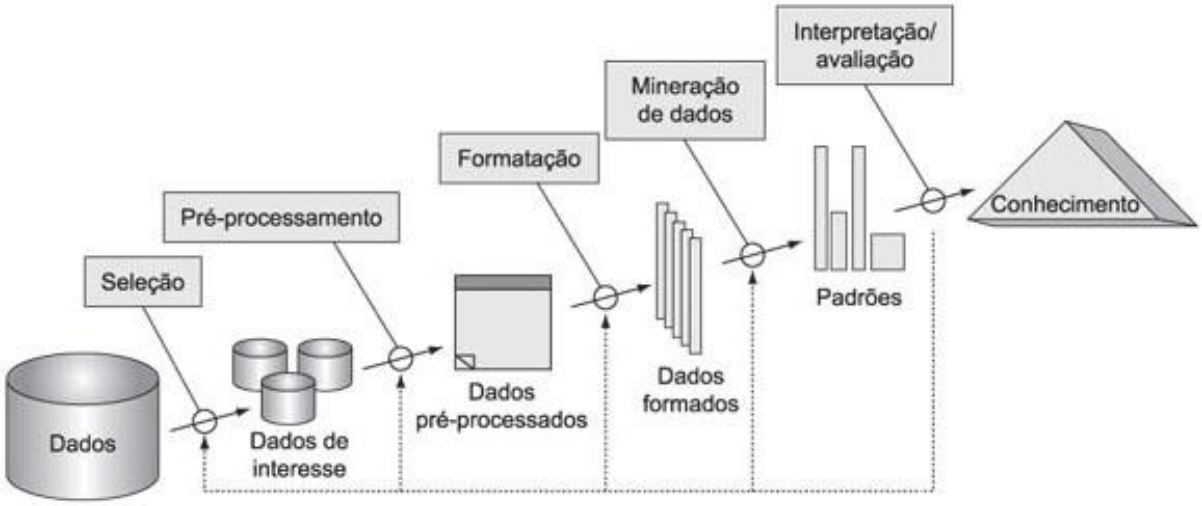
Segundo (INCA, 2002), o tratamento do câncer geralmente inclui a combinação de mais de um método terapêutico, visando obter índices maiores de cura, de modo geral estas modalidades terapêuticas procuram reduzir o número de células neoplásicas. As principais modalidades terapêuticas utilizadas são:

- **Cirurgia:** Geralmente é o tratamento mais importante, utilizado para tumores localizados, em circunstâncias anatômicas favoráveis, influenciando mais na cura do paciente;
- **Radioterapia:** Usualmente utilizado para tumores que costumam recidivar localmente após a cirurgia ou para tumores que não podem ser ressecados totalmente. A área a ser tratada é cuidadosamente identificada para que somente as células tumorais fiquem expostas a doses adequadas de irradiação;
- **Quimioterapia:** Frequentemente é usado para tratamento adjuvante após o tratamento cirúrgico, radioterápico, ou paliativo, em doenças mais avançadas, mas pode ser usada como tratamento principal nos casos de câncer de testículo, leucemias, linfomas;

2.2 Descoberta de Conhecimento

Segundo Fayyad (1996) a descoberta de conhecimento em banco de dados é um processo usado para a identificação de padrões válidos em análise de grandes conjuntos de dados, podendo descobrir informações úteis, conforme ilustrado na figura 1.

Figura 1 - Etapas do processo KDD



Fonte: Fayyad et at. (1996)

O processo KDD é um conjunto de atividades contínuas compostas por cinco etapas, serão citadas no processo de andamento do KDD:

- **Seleção:** nesta fase do KDD serão decididos quais os conjuntos de dados que serão relevantes pra a tarefa de análise da base de dados.

- **Pré-Processamento:** nesta fase acontece a limpeza dos dados e ajustes nas informações ausentes, errôneas e inconsistentes nas bases de dados, a fim de ter uma qualidade dos dados.

- **Formatação ou transformação:** nesta fase acontece a transformação dos dados, serão analisados os dados e reorganizá-los para que sejam interpretados por um software de mineração de dados.

- **Mineração de Dados:** nesta fase faz com que os meros dados sejam transformados em informações através de algoritmos.

- **Interpretação ou Avaliação:** nesta fase é onde as regras indicadas pela etapa de mineração serão interpretadas para a descoberta de conhecimento, após a interpretação poderão surgir padrões, relacionamentos e descoberta de novos fatos.

2.3 Data Warehouse

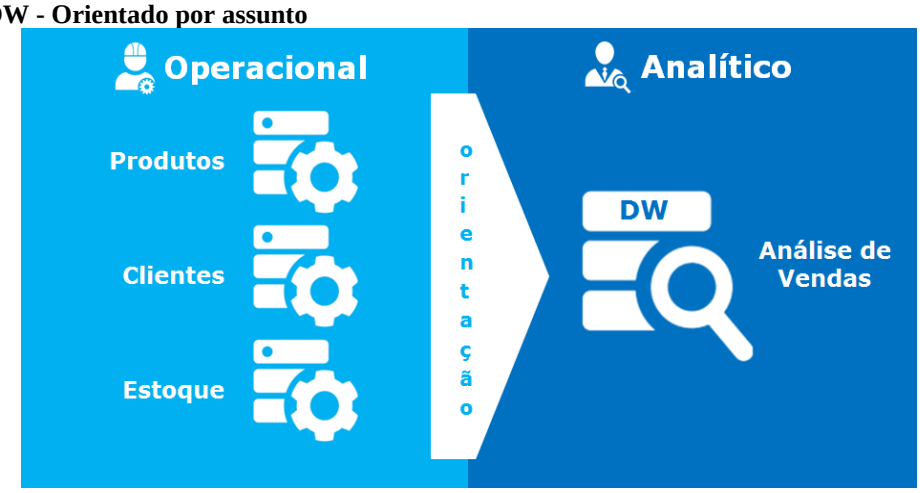
Segundo SILVEIRA (2014) o *Data Warehouse* (DW) é um repositório de informação que congrega os dados de origem operacional e transacional de uma organização e dados externos, que são filtrados, validados e carregados no DW, que passam a ser a fonte de informação para aplicações de análise.

Sua construção é um processo normalmente moroso e complexo, por diversos fatores, dentre os quais a grande quantidade de dados, diversas fontes de informações com bases heterogêneas e muitas vezes inconsistentes, sendo necessário o envolvimento de várias áreas da empresa para interpretação dos dados.

Embora o conceito de DW se aplique a grandes quantidades de dados, sua capacidade não é infinita, devendo ser utilizada sabiamente, apenas dados relevantes devem constar no DW. Segundo INMON (1997) e DATE (2004), o DW é uma coleção de dados para apoiar as decisões gerenciais com as seguintes características:

- Orientado por Assunto: contém informações importantes para o negócio da empresa, onde toda a modelagem do DW é orientada a partir dos principais assuntos da empresa. O exemplo ilustrado na figura 2 apresenta os dados importantes da empresa que são produtos, estoque e cliente, a fim de ter uma visão da análise de vendas.

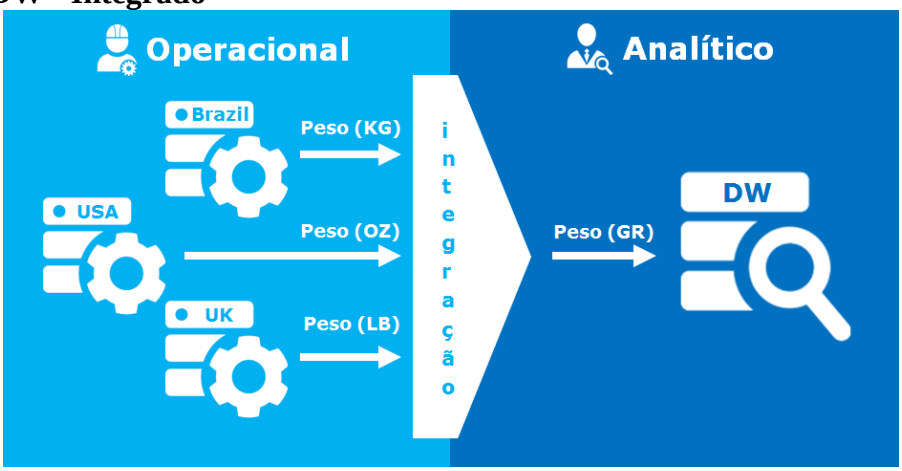
Figura 2 - DW - Orientado por assunto



Fonte: Data Warehouse – Repositório da Intuitivus

- Integrado: os dados são reunidos no DW visando padronizar os dados de origem em uma única representação. O exemplo ilustrado na figura 3 apresenta diferentes unidades de medidas originadas de distintos países, para serem padronizados em uma única unidade de medida na base de dados do DW.

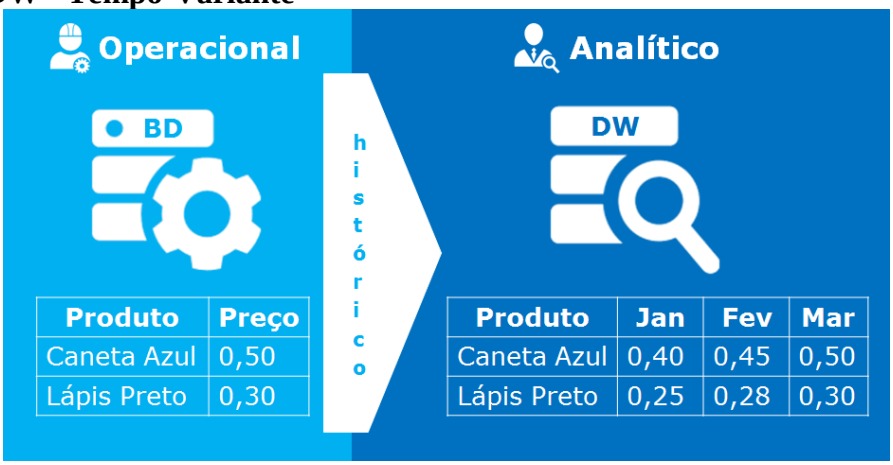
Figura 3 - DW - Integrado



Fonte: Data Warehouse – Repositório da Intuitivus

- **Tempo-variante:** as informações estão relacionadas a um determinado período do tempo, o que proporciona o armazenamento do histórico dos dados. O exemplo ilustrado na figura 4 apresenta dois produtos (caneta azul e lápis preto) que armazenado no DW se tem um histórico destes produtos.

Figura 4 - DW - Tempo-Variante

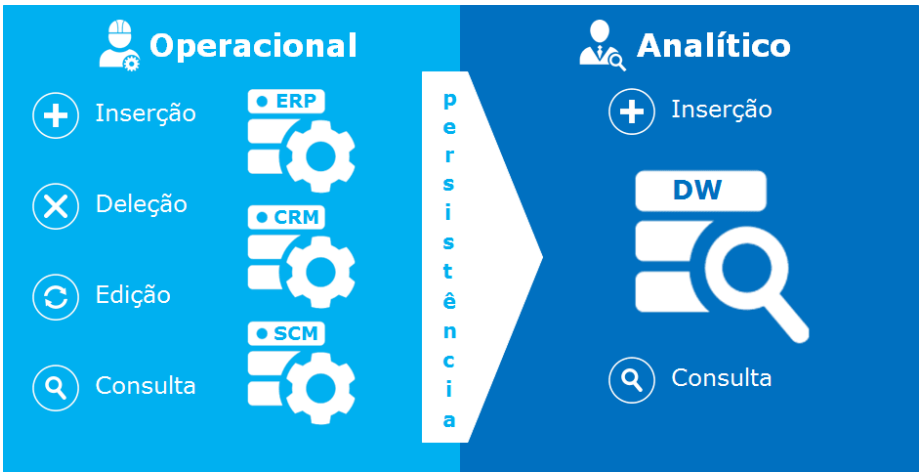


Fonte: Data Warehouse – Repositório da Intuitivus

- **Não volátil:** Nos sistemas transacionais os dados sofrem diversas alterações como inserção, deleção, edição e consulta. No DW, mais dados podem ser adicionados, mas

nunca editados e removidos, isto capacita ao gerenciamento, uma visão consistente dos negócios ilustrada na figura 5.

Figura 5 - DW - Não volátil



Fonte: Data Warehouse – Repositório da Intuitivus

2.3.1 Preparação dos dados

Segundo CARVALHO (2005), os dados são o centro da técnica de *Data Mining* (DM) e do processo de DW, podendo ser classificado de diversas formas, sendo a mais importante à abstração, pois quanto maior a abstração do dado, menor será seu volume, conforme ilustra Figura 6.

Figura 6 - Hierarquia da Abstração

Fonte: Adaptado pelo autor com base em Carvalho (2005, p. 183).

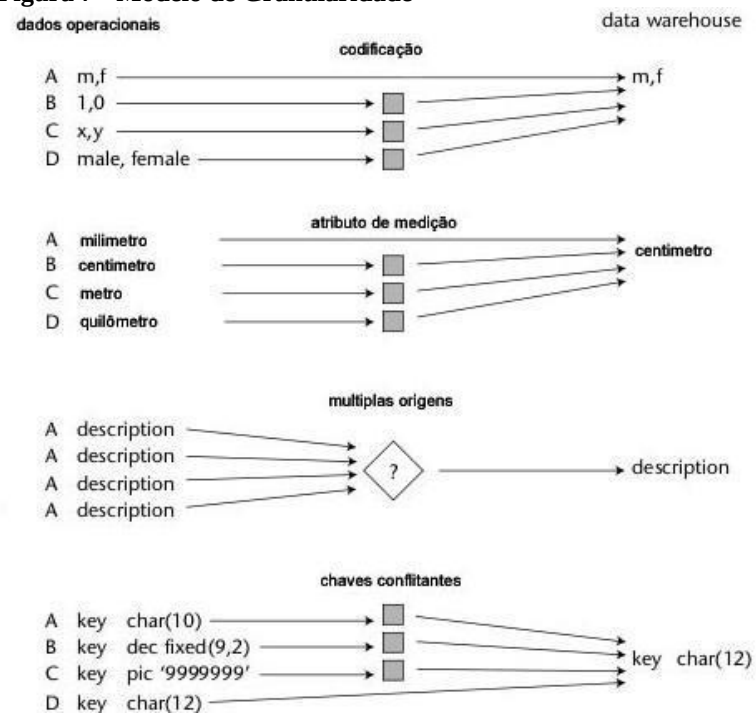
Na camada inferior desta hierarquia encontra-se o dado operacional em grandes quantidades, que por sua vez é a maior dificuldade do processo DW, pois será necessário unificar os dados operacionais com diferentes plataformas, propósitos e softwares. O próximo nível é o dado resumido, onde as informações são reunidas de forma condensada. Após determina-se o modelo de dados que tem por finalidade informar as relações com outros dados, formas e tipos. Em seguida, define-se o metadado, que é um modelo lógico com suas entidades, seus atributos e relações significativas para o analista que realiza a mineração de dados.

Finalizando, o nível mais elevado de abstração determinam-se as especificações, que são diretrizes de caráter genérico, as quais envolvem as necessidades da organização para a realização de suas atividades (CARVALHO, 2005).

2.3.2 Granularidade

INMON (1997), classifica a granularidade como o aspecto mais importante do projeto de um DW, pois ela afeta profundamente o volume de dados que residem no DW, e ao mesmo tempo, afeta o tipo de consulta que pode ser atendida. Na figura 7 são ilustradas as várias formas de universalização da informação dos dados operacionais para o DW.

Figura 7 - Modelo de Granularidade



Fonte: (INMON, 2005)

2.4 Mineração de dados

Segundo CARVALHO (2005), “[...] DM como o uso de técnicas automáticas de exploração de grandes quantidades de dados de forma a descobrir novos padrões e relações que, devido ao volume de dados, não seriam facilmente descobertas a olho nu pelo ser humano”.

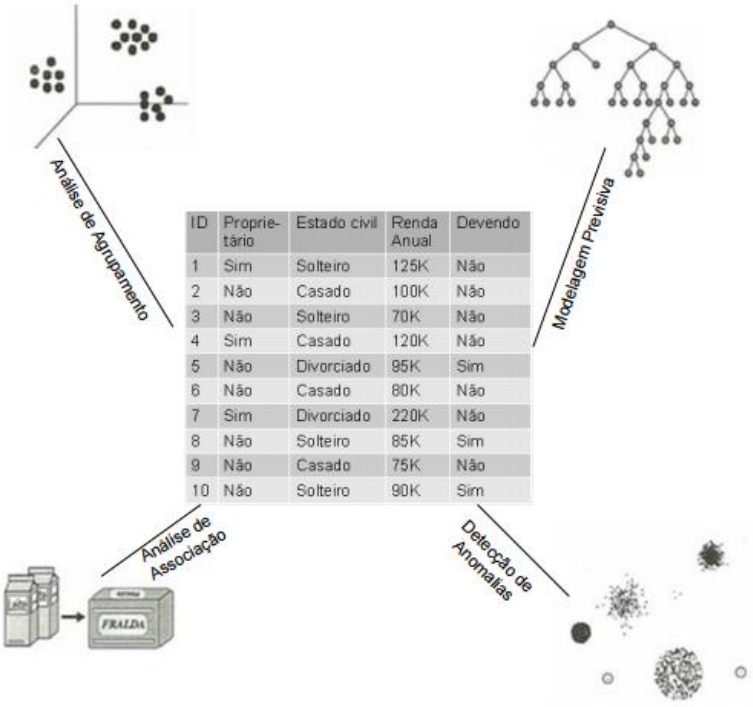
Segundo TAN, STEINBACH e KUMAR (2009), “[...] a mineração de dados é o processo de descoberta automático de informações uteis em grandes depósitos de dados”.

Nesta etapa é definida a tarefa de mineração de dados e as técnicas a serem utilizadas. De acordo com Amo (2004), é importante distinguir o que é uma tarefa e o que é uma técnica de mineração de dados.

2.4.1 Tarefas de mineração de dados

A tarefa de mineração de dados consiste na especificação do que se pretende fazer com os dados, ou seja, qual o objetivo do processo. A figura 8 ilustra quatro das tarefas centrais da mineração de dados.

Figura 8 - Quatro das tarefas centrais da MD



ID	Proprie-tário	Estado civil	Renda Anual	Devendo
1	Sim	Solteiro	125K	Não
2	Não	Casado	100K	Não
3	Não	Solteiro	70K	Não
4	Sim	Casado	120K	Não
5	Não	Divorciado	95K	Sim
6	Não	Casado	80K	Não
7	Sim	Divorciado	220K	Não
8	Não	Solteiro	85K	Sim
9	Não	Casado	75K	Não
10	Não	Solteiro	90K	Sim

Fonte: TAN, STEINBACH e KUMAR (2009, pág 9).

A seguir uma breve descrição para cada uma dessas tarefas segundo TAN, STEINBACH e KUMAR (2009):

- **Modelagem Previsiva:** refere-se à tarefa de construir um modelo para variável alvo como uma função das variáveis explicativas. Há dois tipos de tarefas de modelagem de previsão: classificação, a qual é usada para variáveis discretas; e regressão, que é usada para variáveis contínuas;

- **Detecção de anomalias:** É a tarefa de identificar observações cujas características sejam significativamente diferentes do resto dos dados;

- **Análise de associação:** É usada para descobrir padrões que descrevam características altamente associadas dentro dos dados;

- **Análise de Agrupamentos (Cluster):** Procura encontrar grupos de observações intimamente relacionadas de modo que observações que pertençam ao mesmo grupo sejam mais semelhantes entre si do que com as que pertençam a outros grupos.

Para cada tarefa de mineração de dados há um conjunto de algoritmos específicos a ser aplicado, e estes utilizam determinadas técnicas de mineração de dados que são abordados a seguir.

2.4.2 Técnicas de mineração de dados

A técnica consiste na escolha de métodos ou algoritmos que permitam que esses objetivos sejam alcançados.

Segundo CARVALHO (2005), a mineração de dados pode ser realizada de três diferentes formas em função do nível de conhecimento que se tenha do problema estudado. Se há pouco conhecimento, faz-se a descoberta não supervisionada; se há suspeita de alguma relação interessante, faz-se a testagem em hipótese; se há muito conhecimento, faz-se a modelagem matemática da relação.

Ao utilizar-se uma tarefa de mineração de dados, algumas técnicas são usadas para concretizá-las. Neste capítulo serão abordados os algoritmos de Árvores de Decisão, Classificação Naïve Bayes e Redes Neurais, os demais algoritmos não serão abordados por não serem objeto comum de pesquisa na área da saúde, de acordo com os artigos estudados.

2.4.2.1 Árvores de Decisão

SANTOS; RAMOS (2009, p. 132) definem essa técnica da seguinte maneira,

> [...] as árvores de decisão, como o próprio nome indica, são constituídas por estruturas em árvores que representam um conjunto de decisões. Os algoritmos dessa técnica permitem gerar regras de classificação dos dados, baseados nas informações armazenadas na base de dados.

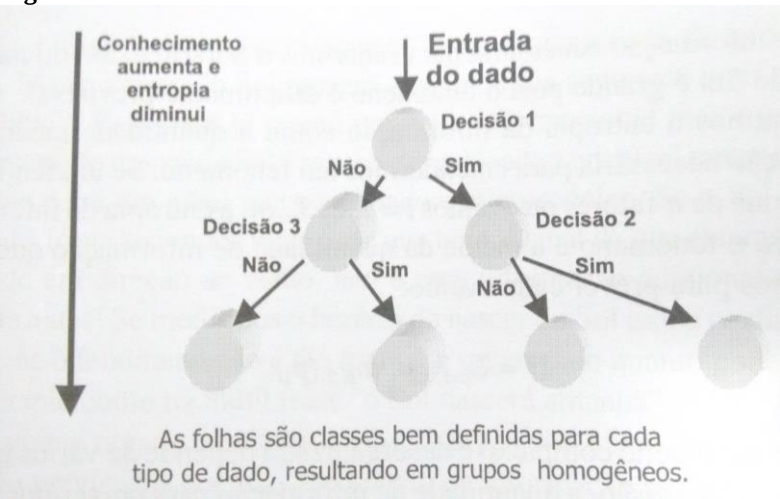
O sucesso das árvores de decisão se dá pelo fato de ser uma técnica extremamente simples e eficiente, não necessita de parâmetros de configuração e geralmente tem um bom grau de assertividade, mas é necessária uma análise detalhada dos dados que serão usados para garantir bons resultados.

Para GARCIA (2000), as árvores de decisão também consistem de: nodos (nós), que representam os atributos, e de arcos (ramos), provenientes desses nodos e que recebem os valores possíveis para esses atributos (cada ramo descendente corresponde a um possível valor desse atributo). Nas árvores existem nodos folha (folha da árvore), que representam as diferentes classes de um conjunto de treinamento, ou seja, cada folha está associada a uma classe. Cada percurso na árvore (da raiz à folha) corresponde a uma regra de classificação.

Conforme CARVALHO (2005), a cada nível da árvore de decisão, é preciso definir regras heurísticas que separem os dados em subgrupos que será mais homogêneo e mais óbvio em seu padrão de comportamento, assim sua entropia[1] será menor, ilustrado na figura 9.

[1] Na Ciência da Computação, a entropia é a falta de conhecimento no presente que deve ser suprida no futuro.

Figura 9 - Árvore de Decisão



Fonte: CARVALHO (2005, pág 158).

2.4.2.2 Classificação Naïve Bayes

Segundo COELHO (2002), a classificação bayesiana consiste em se obter uma distribuição de probabilidade, associada aos diferentes valores que o parâmetro de interesse pode assumir, de modo a representar o grau de credibilidade associado a cada um deles, dado o conjunto de dados observados.

Para Han & Kamber (2005), este tipo de classificação é baseada no teorema de Bayes. Este teorema é um modelo estatístico que permite determinar a probabilidade de hipóteses ocorrerem em um determinado conjunto de registros. Seja P(H|X) a probabilidade da hipótese H estar correta dado X (também chamada de probabilidade a posteriori); P(X|H) a probabilidade de X ocorrer, dada a hipótese; P(H) a probabilidade da hipótese ocorrer; e P(X) a probabilidade de X ocorrer, o teorema de Bayes é definido por

$$\mathbf{P(H|X) = \frac{P(X|H)P(H)}{P(X)}} \qquad (1)$$

Para a classificação utilizando o teorema de Bayes, a hipótese é que um registro com determinado atributo X pertença à classe C. P(H) é a probabilidade a priori de um registro qualquer pertencer à classe C. Os classificadores baseados no teorema de Bayes buscam maximizar a probabilidade a posteriori de X (por isso chamados maximum a posteriori), retornando como resultado a classe com maior probabilidade. Alternativamente, pode-se obter como resultado uma distribuição de probabilidades para cada classe.

Uma característica dos classificadores Bayesianos é que o modelo pode ser facilmente atualizado com um novo registro de treino, bastando atualizar as probabilidades correspondentes ao novo registro.

O classificador Bayesiano mais simples é o naïve Bayes (ou Bayes ingênuo). Ele possui esse nome porque desconsidera que possam ocorrer correlações entre atributos, ou seja, dados quaisquer atributos A e B, A é condicionalmente independente de B. O Naïve Bayes é computacionalmente barato, pois apenas mantém contadores para cada atributo, e necessita executar operações aritméticas básicas. Ao utilizar Naïve Bayes, percebe-se que não há busca explícita por uma hipótese, a hipótese simplesmente é formada pela contagem de frequências (NEAPOLITAN, 2004).

Na (Tabela 1), tem-se um exemplo de entrada de dados para o método de Naïve Bayes, referentes às condições climáticas que determinam ou não a ocorrência de um jogo. No cabeçalho da tabela, estão identificadas as classes e nos demais dados são os atributos

Tabela 1 - Exemplo Naïve Bayes

Perspectiva	Temperatura	Umidade	Vento	Jogo
Ensolarado	Quente	Alta	Fraco	Não
Ensolarado	Quente	Alta	Forte	Não
Nublado	Quente	Alta	Fraco	Sim
Chuvoso	Moderada	Alta	Fraco	Sim
Chuvoso	Frio	Normal	Fraco	Sim
Chuvoso	Frio	Normal	Forte	Não
Nublado	Frio	Normal	Forte	Sim
Ensolarado	Moderada	Alta	Fraco	Não
Ensolarado	Frio	Normal	Fraco	Sim
Chuvoso	Moderada	Normal	Fraco	Sim
Ensolarado	Moderada	Normal	Forte	Sim
Nublado	Moderada	Alta	Forte	Sim
Nublado	Quente	Normal	Fraco	Sim
Chuvoso	Moderada	Alta	Forte	Não

Fonte: Adaptado de Witten, Frank, 2005.

A (Tabela 2) mostra os dados resumidos do jogo, referente às condições climáticas representando se haverá ou não jogo.

Tabela 2- Dados resumidos do exemplo Naïve Bayes

Perspectiva	S	N	Temperatura	S	N	Umidade	S	N	Vento	S	N	Jogo	S	N
Ensolarado	2	3	Quente	2	2	Alta	3	4	Fraco	6	2	Não	-	5
Nublado	4	0	Moderada	4	2	Normal	6	1	Forte	3	3	Sim	9	-
Chuvoso	3	2	Frio	3	1									

Fonte: Adaptado de Witten, Frank, 2005.

Na (Tabela 3) serão apresentados os atributos da (TABELA 2), verificando a ocorrência do atributo jogo relacionado aos outros atributos.

Tabela 3 - Ocorrência do atributo jogo relacionado aos demais atributos do exemplo Naïve Bayes

Perpect.	S	N	Temp.	S	N	Umid.	S	N	Vento	S	N	Jogo	S	N
Ensolarado	2/9	3/5	Quente	2/9	2/5	Alta	3/9	4/5	Fraco	6/9	2/5	Não	-	5/14
Nublado	4/9	0/5	Moderada	4/9	2/5	Normal	6/9	1/5	Forte	3/9	3/5	Sim	9/14	-
Chuvoso	3/9	2/5	Frio	3/9	1/5									

Fonte: Adaptado de Witten, Frank, 2005.

Na (Tabela 4) foi gerada uma nova combinação de valores com o atributo jogo = (?), sendo o ponto de interrogação a perspectiva que deseja-se predizer, verificando qual a probabilidade de ocorrer ou não o jogo, de acordo com a situação climática apresentada pelas variáveis (Perspectiva = Sol, Temperatura = boa, Umidade = alta, Vento = forte).

Tabela 4 - Nova condição de jogo do exemplo Naïve Bayes

Perspectiva	Temperatura	Umidade	Vento	Jogo
Ensolarado	Frio	Alta	Forte	?

Fonte: Adaptado de Witten, Frank, 2005.

Segundo WITTEN (2005), para descobrir a probabilidade de jogo = sim e jogo = não, deve-se multiplicar os valores de acordo com os dados apresentados na (TABELA 3). Por exemplo, para calcular a probabilidade de jogo = sim, devem-se obter os valores na (TABELA 3), e multiplicá-los. Desta forma, o cálculo para esta situação é:

$$\text{Sim} = 2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0{,}0053 \quad (2)$$

Onde 2/9 se refere ao atributo ensolarado, 3/9 a frio, 3/9 a alta, 3/9 a verdadeiro e 9/14 ao total de vezes que jogo é igual a 'sim' da (TABELA 1), onde 9 é o valor total de ocorrências 'sim' e 14 é o total de ocorrências. Do mesmo modo, calcula-se o valor de jogo = 'não':

$$\text{Não} = 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0{,}0206 \quad (3)$$

Ainda segundo WITTEN (2005), os valores acima obtidos podem ser normalizados de acordo com a fórmula 4, que consiste em dividir o valor obtido na ocorrência pelo valor total de ocorrências. Na probabilidade sim, o valor de ocorrências é de 0,0053 e o valor total de ocorrências neste exemplo é a soma do valor de (Sim) mais o valor de (Não), de acordo com a coluna jogo da (TABELA 1).

$$\text{Sim} = \frac{0{,}0053}{0{,}0053 + 0{,}0206} = 20{,}5\% \quad (4)$$

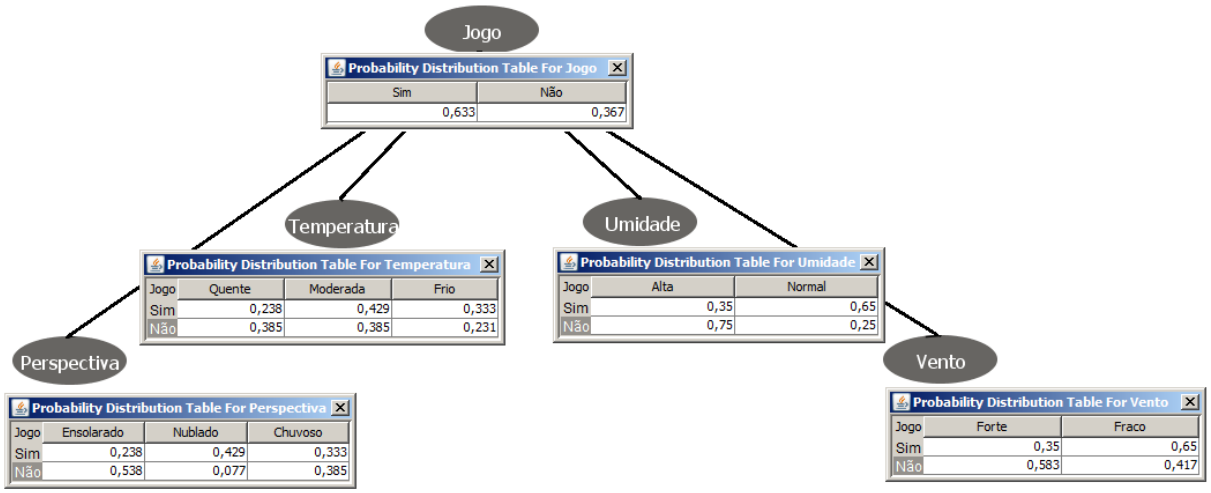
$$\text{Não} = \frac{0{,}0206}{0{,}0053 + 0{,}0206} = 79{,}5\ \% \quad (5)$$

Normalizando tem-se que a probabilidade de 79,5% de não ocorrer o jogo e de 20,5% de ocorrer o jogo.

A classificação de Naïve Bayes representa apenas uma distribuição simples, que também pode ser representada por árvores de decisão, possuindo uma desvantagem, pois fragmenta um conjunto de treinamento em pedaços menores, gerando estimativas de probabilidade menos confiáveis.

Os valores das tabelas dos nodos foram obtidos de acordo com o cálculo de probabilidade de Naïve Bayes, visto anteriormente onde os dados foram calculados a partir da (TABELA 1). O grafo apresentado na figura 10 foi gerado a partir da classificação de Naïve Bayes, onde os atributos não possuem ligação entre eles, sendo exibidas nas tabelas apenas as informações dos atributos referentes ao próprio nodo.

Figura 10 - Grafo detalhado do exemplo Naïves Bayes



Fonte: Do autor (2015) adaptado de Witten, Frank, 2005.

Para o mesmo problema pode ser construída as Redes Bayesiana que representa a mesma distribuição de probabilidade. As Redes Bayesianas são baseadas na estatística (independência condicional) em que as distribuições são representações gráficas desenhadas

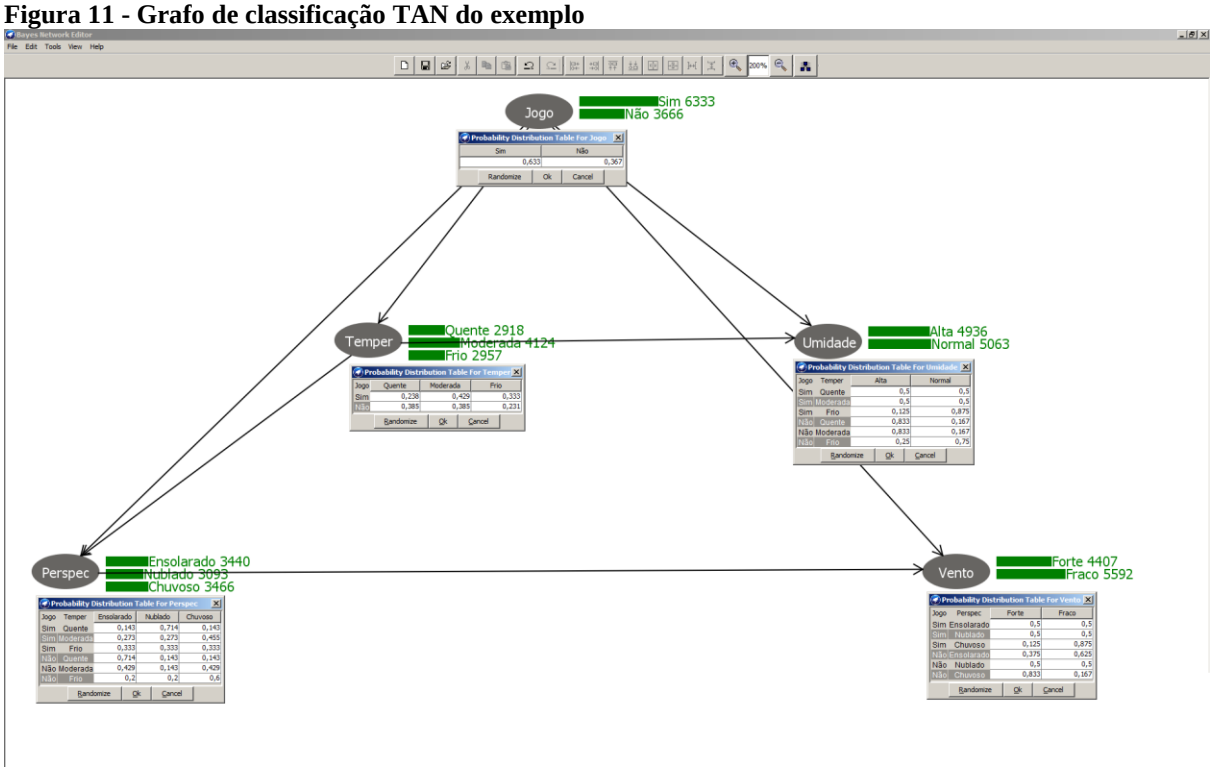
como uma rede de nós, sendo um nó para cada atributo e conectado por arestas de tal maneira a não existirem ciclos, formando um grafo acíclico dirigidas (WITTEN, FRANK, 2005).

A aprendizagem em redes Bayesianas consiste na indução de dois componentes distintos, a estrutura gráfica das dependências condicionais (seleção do modelo) e a distribuição condicional quantificando a estrutura de dependências (estrutura de parâmetros).

Existem vários algoritmos de classificação e criação de redes Bayesianas, o algoritmo que será utilizado no estudo é o *tree augmented naïve Bayes* (TAN) o que traduzido textualmente significa árvore aumentada do naïve Bayes, no entanto a tradução fica apenas o esclarecimento. O classificador TAN, criado por Friedman e Goldszmidt, com o objetivo de melhorar o Naïve Bayes, sendo uma estrutura parecida, porém permite a dependência entre os atributos.

Da mesma forma que em Naïve Bayes, os atributos são dependentes condicionais da classe, mas também podem depender condicionalmente de outros atributos, ilustrado na figura 11.

Figura 11 - Grafo de classificação TAN do exemplo



Fonte: Do autor (2015) adaptado de Witten, Frank, 2005.

Para descobrir a dependência entre atributos no método TAN, é utilizado o algoritmo Chow e Lui, onde cada nodo pode ter no máximo um pai e deve-se encontrar os atributos que

tenha maior correlação. O algoritmo realizado o cálculo de relação de acordo com os valores obtidos em X e Y:

$$Ip(X,Y) = \sum_{x,y} P(x,y) \frac{P(x,y)}{P(x)P(y)} \quad (6)$$

Na fórmula acima, o valor Ip(X;Y) é a informação que X exerce sobre Y ou vice-versa, sendo esta informação calculada para todos os pares de atributos. A partir do algoritmo de Chow e Lui, Friedman adaptou-a para que todos os atributos sejam dependentes da classe, o objetivo é obter a árvore de dependências que maximize o peso das informações mútuas entre os atributos.

$$Ip(X,Y \mid C) = \sum_{x,y,c} P(x,y,c) \frac{P(x,y \mid c)}{P(x \mid c)P(y \mid c)} \quad (7)$$

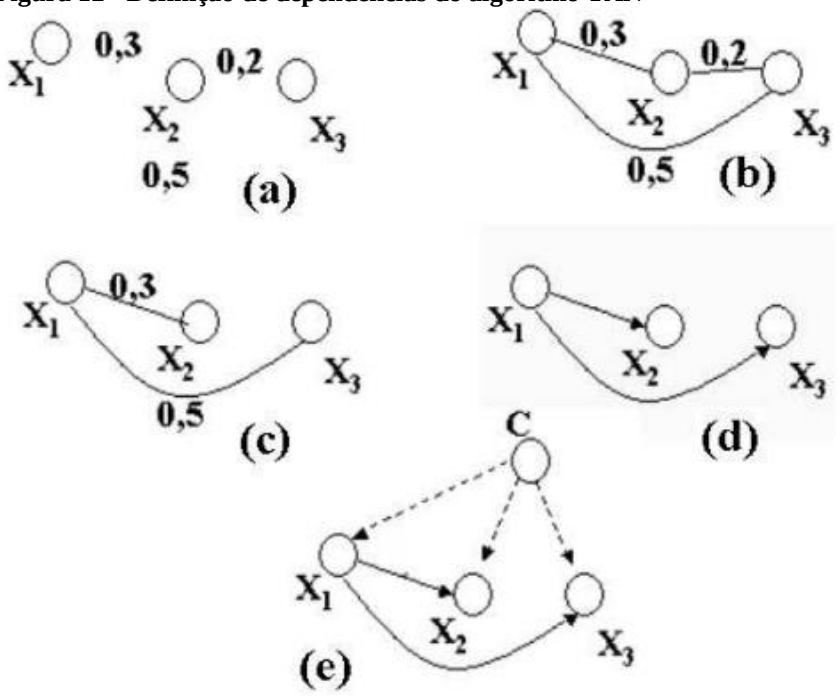
A construção da árvore de dependências é o que diferencia o TAN do Naïve Bayes, pois teoricamente, é devido à dependência entre os atributos que o TAN melhora o desempenho em relação ao Naïve Bayes.

Para construir o grafo de dependências baseado no método TAN, deve-se utilizar a fórmula Chow e Lui adaptada por Friedman, que serão citados em cinco passos:

- 1º Passo: Obtém a informação mútua entre cada par de nodos;
- 2º Passo: Desenha o grafo com os nós e as ligações entre eles, verificando o custo de cada ligação;
- 3º Passo: Calcula o grafo que maximiza a informação mútua entre os atributos de forma acíclica;
- 4º Passo: Define o nodo raiz com as informações mútuas mais altas;
- 5º Passo: Adiciona a classe pai de todos os atributos;

Após estes cinco passos, o grafo utilizando o método TAN é criado, conforme ilustrado na figura 12, onde existem as relações entre o nodo pai com os atributos, conforme método de Naïve Bayes e entre os atributos de acordo com método TAN.

Figura 12 - Definição de dependências do algoritmo TAN



Fonte: Adaptado de Witten, Frank, 2005.

Por isso, considera-se que o algoritmo TAN, é uma extensão de Naïve Bayes, pois mantém a mesma estrutura, porém relacionando os atributos entre si.

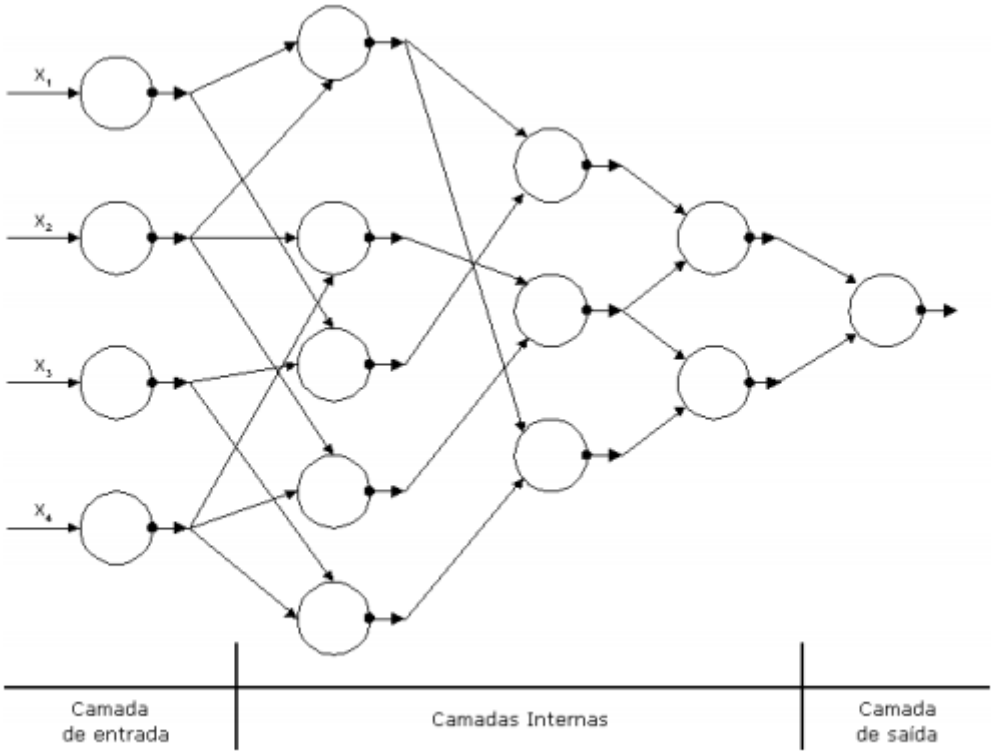
2.4.2.3 Redes Neurais

Segundo TAFNER, citado por LEMOS (2003):

> [...] numa Rede Neural Artificial as entradas, simulando uma área de captação de estímulos, podem ser conectadas em muitos neurônios, resultando, assim, em uma série de saídas, onde cada neurônio representa uma saída. Essas conexões, em comparação com o sistema biológico, representam o contato dos dentritos com outros neurônios, formando assim as sinapses. A função da conexão em si é tornar o sinal de saída de um neurônio em um sinal de entrada de outro, ou ainda, orientar o sinal de saída para o mundo externo (mundo real). As diferentes possibilidades de conexões entre as camadas de neurônios podem ter, em geral, n números de estruturas diferentes (TAFNER, 1998).

Tal qual um neurônio que se comunica com outros formando sinapses, as redes neurais recebem e retransmitem informação, como ilustrado na figura 13.

Figura 13 - Rede Neural



Fonte: SANDRI (2009)

A complexidade dessa forma de comunicação torna necessária a classificação da rede de neurônios em três camadas, composta por tipos de neurônios, segundo sua função:

- **Camada de Entrada:** É a camada inicial da rede de neurônios, onde ficam os neurônios responsáveis pela apresentação de padrões à rede;

- **Camadas Intermediárias (Internas) ou Ocultas:** É a camada onde os neurônios responsáveis pela maior parte do processamento, por meio de conexões ponderadas, extraem as características;

- **Camada de Saída:** É a camada onde se localizam os neurônios responsáveis pelo resultado final.

Pela sua complexidade, as Redes Neurais têm algumas desvantagens quando aplicadas no processo de KDD, pois geram resultados de difícil compreensão ao usuário. Entretanto, é aplicável de forma eficaz em dados que contêm ruído.

2.5 Qualidade dos dados

Segundo TAN, STEINBACH e KUMAR (2009), [...] a qualidade dos dados muitas vezes estão longe da perfeição. Embora a maior parte das técnicas de mineração de dados pode tolerar algum nível de imperfeição nos dados.

A seguir serão detalhados problemas de qualidade de dados, conforme TAN, STEINBACH e KUMAR (2009):

- **Valores Faltando:** Há diversas estratégicas para lidar com dados faltando, cada uma das quais pode ser apropriada em determinadas circunstâncias, como (eliminar objetos ou atributos de dados, eliminar valores faltando, ignorar valores faltando durante a análise);

- **Valores Inconsistentes:** Alguns tipos de inconsistências são fáceis de detectar, porém outros casos pode ser necessário consultar uma fonte externa de informação, uma vez que uma inconsistência tenha sido detectada, às vezes é possível corrigir;

- **Dados Duplicados:** Para detectar e eliminar tais duplicatas, caso houver dois objetos que realmente representem um único, então os valores dos atributos correspondentes podem diferir e estes valores inconsistentes devem ser resolvidos, caso contrário deve ser tomado cuidado para evitar combinar acidentalmente objetos de dados que sejam semelhantes, mas não duplicados.

2.6 Principais Softwares para mineração de dados

Existem disponíveis no mercado ferramentas gratuitas e pagas para mineração de dados. Essas ferramentas são capazes de executar as etapas do processo de mineração. Na (Tabela 5) são apresentadas as principais ferramentas com suas tarefas disponíveis para mineração de dados:

Tabela 5 - Principais ferramentas para mineração de dados

Ferramenta	Fornecedor	Tarefas	Licença
Business Objects	SAP AG.	Classificação, Regras de Associação, Clusterização e Sumarização.	Software Pago
Clementine	SPSS Inc.	Classificação, Regras de Associação, Clusterização, Sequência e Detecção de Desvios.	Software Pago
Darwin	Thinking Machines	Classificação.	Software Pago

DBMiner	DBMiner Technology Inc.	Classificação, Regras de Associação e Clusterização.	Software Pago
Gemanics Expression	Gemanics Developer	Análise de Sequências.	Software Pago
Intelligent Miner	IBM Corp.	Classificação, Regras de Associação, Clusterização e Sumarização.	Software Pago
Microsoft Data Analyser	Microsoft Corp.	Classificação e Clusterização.	Software embarcado na licença do SQL Server
MineSet	Silicon Graphics Inc.	Classificação, Regras de Associação e Análise Estatística.	Software Pago
Oracle Data Miner	Oracle	Classificação, Regressão, Associação, Clusterização e Mineração de Textos.	Software embarcado na licença do Oracle
SAS Enterprise Miner Suite	SAS Inc.	Classificação, Regras de Associação, Regressão e Sumarização.	Software Pago
SAS Text Miner	SAS Inc.	Mineração de Textos.	Software Pago
WEKA	University of Waikato	Classificação, Regressão e Regras de Associação.	Software Livre

Fonte: Do autor (2015)

Ao observar a (TABELA 5), percebe-se que a maioria das ferramentas analisadas possuem técnicas de classificação e associação, porém apenas duas possuem técnicas mineração de textos e uma técnica possui detecção de desvios. Assim, pode-se concluir que as técnicas menos utilizadas nas ferramentas analisadas são as técnicas de mineração de textos e detecção de desvios.

O software de mineração de dados a ser utilizado neste estudo será o WEKA (*Waikato Environment for Knowledge Analysis* ou Ambiente para a Análise do Conhecimento), por ser um software de distribuição gratuita, desenvolvido em Java, que se consolidou como a ferramenta de mineração de dados mais utilizada no meio acadêmico. Grande parte de seus componentes de software são resultantes de teses e dissertações de grupos de pesquisa da Universidade de Waikato, Nova Zelândia.

Através de sua interface gráfica, conhecida como WEKA Explorer, é possível conduzir processos de mineração de dados de forma simples, realizando a avaliação dos resultados obtidos e a comparação de algoritmos.

3 METODOLOGIA

Este capítulo tem como finalidade detalhar os procedimentos que norteiam o desenvolvimento desta pesquisa.

3.1 Tipo de Pesquisa

Primeiramente será definido o tipo de pesquisa quando aos seus objetivos, natureza da abordagem, procedimentos técnicos e a unidade de análise. Em seguida serão apresentados a amostra, o plano de coleta de dados, o plano de tratamento dos dados e os procedimentos éticos.

3.1.1 Quanto aos objetivos

Considerando os objetivos desta pesquisa a mesma pode ser caracterizada como exploratória. Para GIL (1996, p.45) o objetivo deste tipo de estudo é proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explícito ou de maneira a possibilitar a construção de hipóteses. Tem como objetivo principal o aprimoramento de ideias ou a descoberta de intuições.

Para a realização deste trabalho de pesquisa, será utilizada a pesquisa exploratória, com o intuito de descobrir conhecimento útil da especialidade médica de oncologia na base de dados de uma Casa de Saúde.

3.1.2 Quanto à natureza de abordagem

O método desta pesquisa, quanto à natureza de abordagem, pode ser classificada como estudo de caso. Segundo Malhotra (2001), o estudo de caso é um processo que procura descrever e analisar alguma entidade em termos qualitativos, complexos e compreensivos e, não invariavelmente, como ele se desdobra em um período de tempo. O estudo de caso caracteriza-se por ser um método qualitativo, devido ao fato de que as inferências a partir dos resultados obtidos não são estatísticas (BARBETTA, 1994). Consiste na análise intensiva de uma ou poucas situações, sendo priorizada a descrição completa e o entendimento dos fatores de cada situação (BOYD & STASCH, apud BARBETA, 1994).

O presente estudo caracteriza-se por apresentar a abordagem qualitativa na sua parte exploratória, pois visa averiguar os dados coletados através de técnicas de mineração de dados sobre uma vasta quantidade de dados, buscando encontrar padrões entre os atributos da área de oncologia da Casa de Saúde.

3.1.3 Quanto aos procedimentos técnicos

O procedimento técnico adotado para esse estudo foi à pesquisa bibliográfica. Para Marconi e Lakatos (2010) é o levantamento da bibliografia já publicada que tenha relação com o tema. Tem como objetivo colocar o pesquisador em contato direto com tudo o que foi escrito sobre determinado assunto, como livros, revistas, artigos e jornais. Para os autores Cervo; Bervian e Silva (2007), a pesquisa bibliográfica pretende explicar uma situação através de contribuições culturais ou científicas já divulgadas sobre o tema ou assunto e com isso contribuir para o desenvolvimento do estudo.

3.2 Unidade de análise

Neste estudo a unidade de análise são as técnicas de mineração de dados empregadas para descobrir informações implícitas na especialidade médica de oncologia em uma Casa de Saúde.

3.3 Amostra

Participarão desta amostra, pacientes com diferentes diagnósticos de doenças oncológicas em tratamento na área de Oncologia de uma Casa de Saúde, maiores e iguais a 15 anos, de ambos os sexos, no período de 2011 a 2014. Para compor esta amostra, os pacientes deverão ter diagnóstico conclusivo de doença oncológica.

3.4 Plano de coleta de dados

O estudo foi iniciado entrando em contato com o responsável técnico oncologista, a coordenadora e o enfermeiro do centro de oncologia da Casa de Saúde, que deram seu consentimento para a realização desta pesquisa.

Realizadas estas medidas, o projeto foi submetido para apreciação do CENEPE (Centro de Ensino e Pesquisa) da Casa de Saúde, onde foi aprovado. A coleta de dados iniciou no segundo semestre de 2015 utilizando dados do sistema SISRHC (Sistema para Informatização dos dados de Registros Hospitalares de Câncer), a fim de estabelecer o enriquecimento do experimento utilizou-se outra fonte de dados, o sistema de gestão TASY.

O SISRHC é um sistema desenvolvido pelo INCA (Instituto Nacional do Câncer), com o objetivo de consolidar as informações hospitalares provenientes dos RHC (Registros Hospitalares de Câncer) de todo o Brasil.

O TASY é um sistema desenvolvido pela PHILIPS, com objetivo de integrar todas as informações hospitalares em uma plataforma única que possa facilitar o fluxo das informações entre todos os setores da empresa.

Atualmente estes registros da Casa de Saúde representam apenas dados e não conhecimento, visando transformar estes dados em conhecimento, utilizou-se o processo de descoberta de conhecimento em banco de dados (KDD), visto na seção 2.2, na revisão de literatura. Após foi analisado os resultados do algoritmo TAN escolhido na mineração de dados sobre as informações de forma a identificar padrões.

Após uma avaliação criteriosa dos dados minerados, pode-se afirmar que o objetivo foi alcançado ou eventualmente, pode-se concluir que as soluções propostas não são eficientes. Sendo assim, deverá ser aplicado outro algoritmo que busque alcançar o objetivo.

3.5 Procedimentos éticos

Após a finalização do estudo será entregue uma cópia ao CENEPE, sendo disponibilizado o e-mail do pesquisador para aqueles que demonstrarem interesse em conhecer os dados da pesquisa.

4 CARACTERIZAÇÃO DA EMPRESA

Fundada em 1931, a Casa de Saúde é uma instituição filantrópica de direito privado, sendo referência em diversas especialidades nas regiões do Vale do Taquari e Rio Pardo. Está inscrita nos conselhos Municipal, Estadual e Federal, sendo reconhecida como de utilidade pública e de extrema importância para a população.

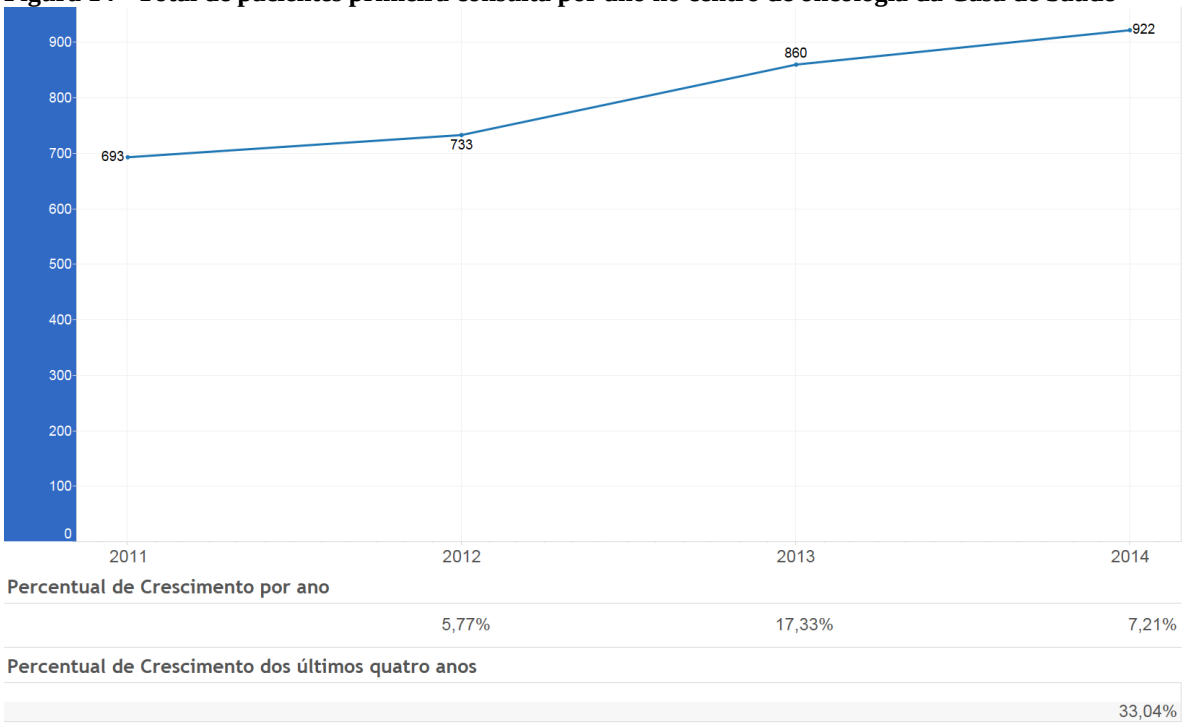
Os serviços da Casa de Saúde estão distribuídos em uma área construída de 27,2 mil m², comprovando a evolução da instituição, se comparado aos 2,6 mil m² da época de sua fundação. A amplitude das instalações consolida a Casa de Saúde como sendo uma das mais completas em oferecer serviços no interior do Rio Grande do Sul.

Dentre os serviços da Casa de Saúde foi selecionado o centro de oncologia por possuir vasta quantidade de dados, mas com dificuldade na extração de informações úteis relacionadas à gestão da oncologia, surgindo à oportunidade de aplicar técnicas de mineração de dados na descoberta de conhecimento.

O centro de oncologia prega um tratamento multidisciplinar conforme citado na seção 2.1 atendendo a pacientes do sistema único de saúde (SUS) e de convênios. O levantamento dos últimos quatro anos do centro de oncologia da Casa de Saúde indica crescimento no número de primeiras consultas, de 2011 a 2014 foram registrados no hospital 3.208 pacientes primeira consulta.

Comparado o número de pacientes primeira consulta dos últimos quatro anos registra-se um crescimento na quimioterapia de 33,04%. Em 2011, passaram por primeira consulta 693 pacientes, os dados de 2014 apontam 922 pacientes, ilustrado na figura 14.

Figura 14 - Total de pacientes primeira consulta por ano no centro de oncologia da Casa de Saúde



Fonte: Do autor (2015) com base nos dados do SISRHC

O levantamento assinala ainda um incremento no número de primeiras consultas de 17,33% no ano 2013, em relação ao ano anterior, uma das explicações é devido à nova instalação do Serviço de Quimioterapia, que pode comportar um número considerável de primeiras novas consultas e tratamentos de neoplasias.

A nova sede da Oncologia foi inaugurada em dezembro de 2012 e iniciando suas atividades em janeiro de 2013, a sede está localizada no quarto andar do Centro de Tecnologia Avançada da Casa de Saúde, o espaço foi ampliado de 407m² para 1.422m², o local se tornou o maior serviço de Quimioterapia do Estado, em termos de dimensão, podendo atender até 60 pacientes simultaneamente.

5 TRABALHOS RELACIONADOS

Este capítulo destina-se a abordar trabalhos cujo foco seja a aplicação de técnicas e algoritmos de mineração de dados na área da saúde. Existem vários trabalhos aplicando mineração de dados no contexto da saúde, entretanto, neste capítulo serão abordados apenas trabalhos envolvendo técnicas de mineração de classificadores como Naïve Bayes e algoritmos de classificação TAN.

ABICALAFFE (2000) apresentou em seu artigo na Pontifícia Universidade Católica do Paraná (PUC/PR) uma proposta de desenvolvimento de um software aplicando a rede bayesiana na prevenção da gestação de alto risco. Os resultados deste estudo apresentaram um grau de assertividade na análise em 90% dos casos de prematuridade e de recém-nascido baixo peso e estima-se uma redução de pelo menos 30% dos casos de prematuridade e de trabalho de parto prematuro além de mais de 40% dos casos de recém-nascido de baixo peso.

FILHO (2006) apresentou em sua monografia na Universidade Católica de Goiás uma proposta de desenvolvimento de uma ferramenta para o auxílio no diagnóstico de anomalias cromossômicas para a Síndrome de Turner. Os resultados deste estudo apresentam maior acurácia com classificador Naïve Bayes em uma amostra de 84 pacientes, foram identificados 63 pacientes com Síndromes de Turner, aproximadamente 89% dos pacientes, também foram efetuado teste com os modelos Árvore de Decisão, TAN e BAN e Rede Neural, porém não foram tão satisfatórias quanto o Naïve Bayes.

REDEKER (2010) apresentou em sua monografia no Centro Universitário UNIVATES, um estudo de descoberta de conhecimento na área de cardiologia de uma Casa de Saúde utilizando o algoritmo TAN com intuito de descobrir relações entre características de pacientes. Os resultados deste estudo apresentou que pacientes do sexo masculino tem maior probabilidade de sofrerem de algum problema cardíaco, já pacientes do sexo feminino

possuem maior probabilidade de ir a óbito. O principal diagnóstico de cardiopatias são as doenças isquêmicas do coração, com destaque para a angina instável ocorrendo principalmente em pacientes do sexo masculino.

SARABANDO (2010) apresentou em sua monografia na Universidade do Porto, um estudo da aplicação de redes bayesianas ao prognóstico da sobrevivência no cancro de próstata. Os resultados desde estudo mostram que as redes geradas automaticamente comportam-se tão bem, ou melhor, que a rede construída manualmente, além de apresentarem relações causais não existentes na rede gerada manualmente. Entres as ferramentas utilizadas, a ferramenta WEKA apresentou a melhor confiança, em que a história familiar de cancro da próstata pode influenciar o prognóstico do doente, mas não é condição para que esse prognóstico seja a morte.

PACHIAROTTI (2012) apresentou em sua monografia na Universidade Vila Velha, um projeto implementando técnicas de mineração de dados como classificação (algoritmo J48 de árvores de decisão, Redes Neurais e Naïve Bayes), associação (algoritmo Apriori), clusterização (k-means), em um cenário de atendimento médico para a descoberta de padrões e comportamentos que possibilitem uma melhor tomada de decisões para os gestores de operadoras médicas, de forma a otimizar recursos e prover maior qualidade no atendimento ao público. Os resultados desde estudo apresentaram que a maior parte das marcações se distribui entre os dias da semana de segunda a quarta e estão concentradas no início da segunda quinzena do mês e no final da primeira quinzena do mês. Mais da metade dos pacientes não compareceram às consultas, a maioria das marcações é de caráter eletivo e o número de encaixes é baixo em proporção aos horários de atendimento. Grande quantidade de agendamentos é realizada com antecedência de 15 a 30 dias. Foi sugerido aos gestores duas ações para evitar ao máximo faltas, não permitir o agendamento para mais de 10 dias de antecedência e procurar confirmar todos os agendamentos com o paciente, preferentemente até dois dias antes da consulta.

Conforme o estudo realizado sobre os trabalhos relacionados, verificou-se que minerações de dados envolvendo técnica de mineração de classificadores como Naïve Bayes e algoritmos de classificação TAN estão sendo bastante empregadas, uma vez que esta técnica possui grande aplicabilidade à área da saúde, devido a sua natureza de apoio a decisão, além de ser facilmente validada junto ao especialista e aos usuários envolvidos.

6 RESULTADOS

Partindo de uma das ideias de trabalho futuro citado por REDEKER (2010), em utilizar outras especialidades médicas, foi fator determinante a especialidade médica de oncologia, para aplicação de técnicas de mineração de dados na descoberta de conhecimento útil, pela necessidade em melhorar a gestão do Centro de Oncologia. Para a descoberta de conhecimento foi utilizado o processo KDD, conforme as cinco etapas que serão citadas nas subseções que seguem:

6.1 Seleção dos dados

O processo para descoberta de conhecimento foi iniciado com a etapa de seleção dos dados. Após o reconhecimento dos atributos dos dados da base do SISRHC, foi gerado um modelo de dados no formato de planilha, exportando os registros relevantes para o experimento de pacientes de primeira consulta oncológica entre o período de 2011 a 2014, totalizando 3267 registros e 10 atributos conforme (Tabela 6).

Tabela 6 - Atributos de identificação do arquivo do SISRHC na etapa seleção dos dados

ATRIBUTO	DESCRIÇÃO
SEXO	Indica o sexo do paciente
RACA	Indica a raça do paciente
IDADENEO	Indica a idade do paciente quando constatada a neoplasia
DTPRICON	Indica o ano da primeira consulta
CIRURGIA	Indica se o paciente fez cirurgia
EMTRAT	Indica quanto tempo está em tratamento o paciente
HISTFAM	Indica se há histórico familiar de câncer
ALCOOL	Indica se há histórico de consumo de bebida alcoólica
TABAG	Indica se há histórico de consumo de tabaco
CID	Indica o tumor conforme CID-O (Cadastro Internacional de Doenças para Oncologia)

Fonte: Do autor (2015)

A fim de estabelecer o enriquecimento do experimento, utilizou-se outra fonte de dados, o sistema de gestão TASY, para tanto, foi realizada uma consulta SQL na base de dados do TASY, separando as informações relacionadas com os pacientes do SISRHC complementando os dados com mais 6 atributos conforme (Tabela 7).

Tabela 7 - Atributos de identificação do arquivo do TASY na etapa seleção dos dados

ATRIBUTO	DESCRIÇÃO
PESO	Indica o peso do paciente
ALTURA	Indica a altura do paciente
OBITO	Indica se o paciente foi a óbito
PESQUISA	Indica se o paciente participa do protocolo de pesquisa nacional e/ou internacional
MUNIBGE	Indica o código do município IBGE
COORDEN	Indica a microrregião do município IBGE

Fonte: Do autor (2015)

A ideia desse enriquecimento foi complementar os dados, pois o SISRHC não possuía tais informações, ou se possuía, não eram completas ou confiáveis.

As informações das duas fontes de dados foram relacionadas conforme o número do prontuário do paciente, campo que ambas as fontes de dados possuíam. Ressaltando que o atributo prontuário, não será considerado na análise para preservar a privacidade dos pacientes com neoplasias e não ter utilidade para análise.

6.2 Pré-Processamento

Nesta etapa de pré-processamento foram verificados os atributos que necessitavam de correção e ajustes de formatação nos dados, a fim de eliminar problemas tornando mais adequados para uso dos algoritmos na mineração de dados.

Alguns erros ocorrem por falha humana, como o atributo SEXO, que primeiramente havia sido selecionado da fonte de dados do SISHC, porém como foram encontrados 64 registros de dados incorretos, optou-se em cruzar os valores das duas fontes de dados para obter a real informação deste atributo. No final, foi constatado que a fonte de dados do TASY estava completamente correta, por isto foi utilizada o atributo SEXO desta fonte de dados.

Alguns erros são sistemáticos e mais fáceis de detectar e corrigir, como o atributo ALTURA, que não se tinha um padrão no TASY sendo cadastrado em alguns momentos em

centímetro e em outros em metro. Assim, foram convertidos todos os valores de centímetro para metro.

Nos atributos PESO e ALTURA foi constatado que havia 26 registros com ausência de valores, assim foi conversado com a equipe da oncologia e constatado que estes registros poderiam ser resgatados nos prontuários manuais, devido ao fato de ser necessário para o tratamento oncológico esta informação, assim foram separadas as pastas dos pacientes e preenchidos para não eliminarmos os pacientes por falta de registro.

O atributo OBITO foi necessário extrair da fonte de dados TASY, devido à falha no processo de lançamento da informação no SISRHC. Como no TASY a informação estava em formato de data, foi convertido caso estivesse preenchido no valor (S), caso contrário no valor (N).

O atributo RACA foi preciso extrair da fonte de dados TASY, primeiramente havia sido selecionado da fonte de dados do SISHC, porém como foram encontrados 122 registros de dados incorretos, optou-se em cruzar os valores das duas fontes de dados para termos a real informação deste atributo. Como ocorreu com o atributo SEXO, foi constatado que a fonte de dados do TASY estava completamente correta, por isto foi utilizada o atributo RACA desta fonte de dados.

No final da etapa de pré-processamento, do montante de 3267 pacientes foram desconsiderados 59 pacientes por apresentarem inconsistência nos dados, totalizando 3208 pacientes com registros íntegros. Na (Tabela 8) estão os atributos atualizados na etapa do pré-processamento da fonte de dados do SISRHC.

Tabela 8 - Atributos de identificação do arquivo do SISRHC na etapa pré-processamento

ATRIBUTO	DESCRIÇÃO
IDADENEO	Indica a idade do paciente quando constatada a neoplasia
DTPRICON	Indica o ano da primeira consulta
CIRURGIA	Indica se o paciente fez cirurgia
EMTRAT	Indica quanto tempo está em tratamento o paciente
HISTFAM	Indica se há histórico familiar de câncer
ALCOOL	Indica se há histórico de consumo de bebida alcoólica
TABAG	Indica se há histórico de consumo de tabaco
CID	Indica o tumor conforme CID-O (Cadastro Internacional de Doenças para Oncologia)

Fonte: Do autor (2015)

Na (Tabela 9) estão atributos que foram atualizados na etapa do pré-processamento da fonte de dados do TASY.

Tabela 9 - Atributos de identificação do arquivo do TASY na etapa pré-processamento

ATRIBUTO	DESCRIÇÃO
SEXO	Indica o sexo do paciente
RACA	Indica a raça do paciente
PESO	Indica o peso do paciente
ALTURA	Indica a altura do paciente
OBITO	Indica se o paciente foi a óbito
PESQUISA	Indica se o paciente participa do protocolo de pesquisa nacional e/ou internacional
MUNIBGE	Indica o código do município IBGE
COORDEN	Indica a microrregião do município IBGE

Fonte: Do autor (2015)

6.3 Formatação

Após serem selecionados, limpos e pré-processados, os dados necessitam ser armazenados e formatados adequadamente para que os algoritmos possam ser aplicados.

Para os atributos IDADENEO, PESO e ALTURA optou-se em normalizar os dados, que consiste em ajustar a escala dos valores de cada atributo de forma que os valores fiquem em pequenos intervalos.

O atributo (IDADENEO) foi transformado em dois novos atributos, o atributo (FETARIA) que agrupa a idade de quatro em quatro anos utilizando como referência a faixa etária IBGE conforme Anexo A, e o atributo (EETARIA) que agrupa a idade em três estruturas etárias conforme Anexo B (IBGE, 2000).

Os atributos (PESO e ALTURA) foram utilizados para compor um único atributo índice de massa corporal (IMC), aplicando a fórmula:

$$IMC = \frac{PESO\ (QUILOS)}{ALTURA^2\ (METROS)} \quad (8)$$

Desta forma, com o valor do IMC foi possível relacionar os resultados aos pacientes conforme (Tabela 10).

Tabela 10 – Resultados do Índice de Massa Corporal

IMC	Resultado
Abaixo 18,49	Subnutrido
Entre 18,5 e 24,99	Peso saudável
Entre 25 e 29,99	Sobrepeso
Entre 30 e 34,99	Obesidade I
Entre 35 e 39,99	Obesidade II
Acima de 40	Obesidade III

Fonte: Adaptado pelo autor (2015), conforme (OMS, 1997)

As informações modificadas foram agrupadas em um repositório único no formato de planilha com 16 atributos de ambas as fontes de dados conforme (Tabela 11).

Tabela 11 - Atributos de identificação do arquivo de ambas as fontes

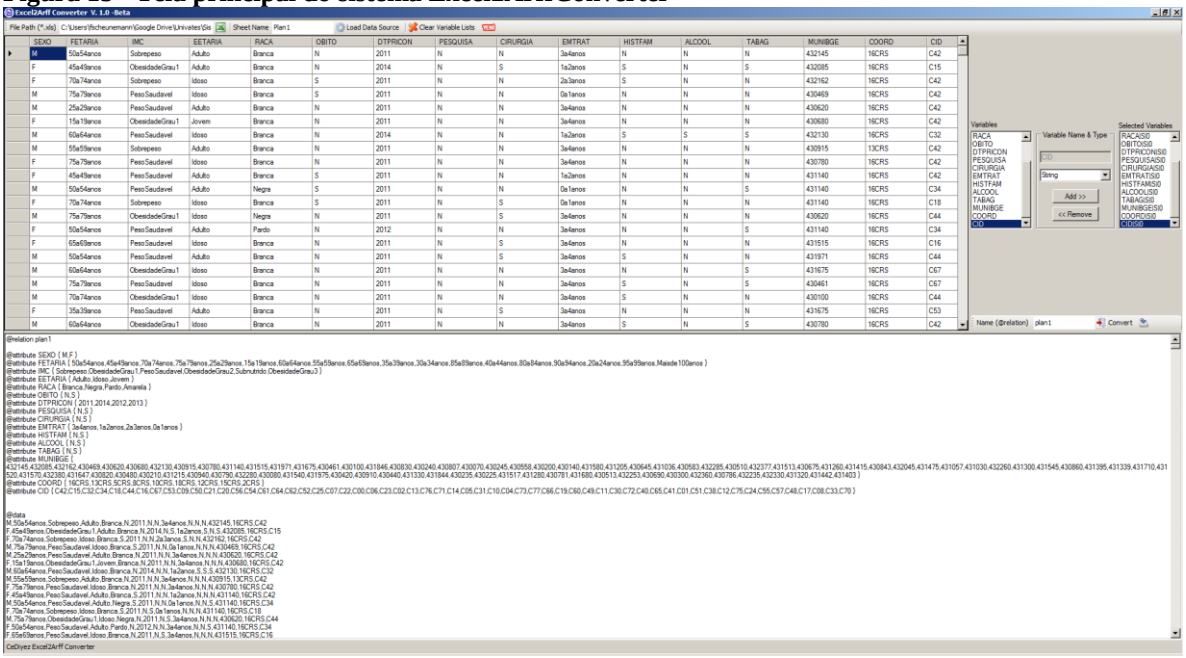
ATRIBUTO	DESCRIÇÃO	VALORES
SEXO	Indica o sexo do paciente	M, F
FETARIA	Indica a faixa etária do paciente conforme Faixa etária IBGE	(ANEXO A)
EETARIA	Indica a estrutura etária	Jovem, Adulto, Idoso
IMC	Indica o resultado do IMC	Subnutrido, PesoSaudavel, Sobrepeso, Obesidade Grau I, Obesidade Grau II, Obesidade Grau III
RACA	Indica a raça do paciente	Branca, Negra, Pardo e Amarela
OBITO	Indica se o paciente foi a óbito	S, N
DTPRICON	Indica o ano da primeira consulta	2011, 2012, 2013, 2014
PESQUISA	Indica se o paciente participa do protocolo de pesquisa nacional e/ou internacional	S,N
CIRURGIA	Indica se o paciente fez cirurgia	S,N
EMTRAT	Indica quanto tempo está em tratamento o paciente	0a1 ano, 1a2anos, 2a3anos, 3a4anos
HISTFAM	Indica se há histórico familiar de câncer	S,N
ALCOOL	Indica se há histórico de consumo de bebida alcoólica	S,N
TABAG	Indica se há histórico de consumo de tabaco	S,N
CID	Indica o CID (Cadastro Internacional de Doenças para oncologia)	(ANEXO C)
MUNIBGE	Indica o código do município IBGE	(ANEXO D)
COORD	Indica a microrregião do município IBGE	(ANEXO D)

Fonte: Do autor (2015).

Nesta etapa, foi gerado um arquivo com a extensão ARFF (*Attribute Relation File Format*) um dos padrões aceitáveis pelo WEKA, com o auxílio do sistema open source Excel2ArffConverter obtido na Sourceforge.

Este software converte arquivos de planilha no formato XLS para ARFF, onde os dados são armazenados em duas seções distintas. A primeira seção contém informações de cabeçalho, enquanto que a segunda, informação dos dados, ilustrado na figura 15.

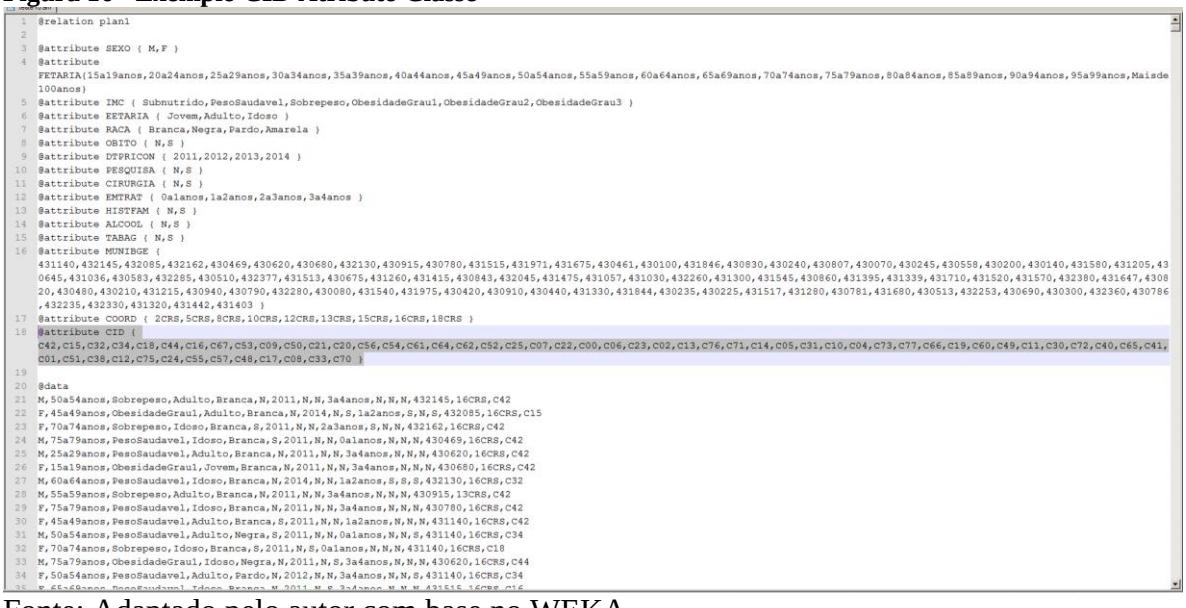
Figura 15 - Tela principal do sistema Excel2ArffConverter



Fonte: Adaptado pelo autor com base no Excel2ArffConverter

Para que o nodo seja escolhido como nodo classe, em suma, representa o atributo mais significativo entre todos os outros atributos. No arquivo com extensão ARFF, o atributo classe deve estar na última linha dos atributos, ilustrado na figura 16.

Figura 16 - Exemplo CID Atributo Classe



```
@relation plan1

@attribute SEXO { M,F }
@attribute
FETARIA{15a19anos,20a24anos,25a29anos,30a34anos,35a39anos,40a44anos,45a49anos,50a54anos,55a59anos,60a64anos,65a69anos,70a74anos,75a79anos,80a84anos,85a89anos,90a94anos,95a99anos,Maisde
100anos}
@attribute IMC { Subnutrido,PesoSaudavel,Sobrepeso,ObesidadeGrau1,ObesidadeGrau2,ObesidadeGrau3 }
@attribute EETARIA { Jovem,Adulto,Idoso }
@attribute RACA { Branca,Negra,Pardo,Amarela }
@attribute OBITO { N,S }
@attribute DTPRICON { 2011,2012,2013,2014 }
@attribute PESQUISA { N,S }
@attribute CIRURGIA { N,S }
@attribute EMTRAT { 0a1anos,1a2anos,2a3anos,3a4anos }
@attribute HISTFAM { N,S }
@attribute ALCOOL { N,S }
@attribute TABAG { N,S }
@attribute MUNIBGE {
431140,432145,432085,432162,430469,430620,430680,432130,430915,430780,431515,431971,431675,430461,430100,431846,430830,430240,430807,430070,430245,430558,430200,430140,431580,431205,43
0645,431036,430583,432285,430510,432377,431513,430675,431260,431415,430843,432045,431475,431057,431030,432260,431300,431545,430860,431395,431339,431710,431520,431570,432380,431647,4308
20,430480,430210,431215,430940,430790,432280,430080,431540,431975,430420,430910,430440,431330,431844,430235,430225,431517,431280,430781,430680,430513,432253,430690,430300,432360,430786
,432235,432330,431320,431442,431403 }
@attribute COORD { 2CRS,5CRS,8CRS,10CRS,12CRS,13CRS,15CRS,16CRS,18CRS }
@attribute CID {
C42,C15,C32,C34,C18,C44,C16,C67,C53,C09,C50,C21,C20,C56,C54,C61,C64,C62,C52,C25,C07,C22,C00,C06,C23,C02,C13,C76,C71,C14,C05,C31,C10,C04,C73,C77,C66,C19,C60,C49,C11,C30,C72,C40,
C01,C51,C38,C12,C75,C24,C55,C57,C48,C17,C08,C33,C70 }

@data
M,50a54anos,Sobrepeso,Adulto,Branca,N,2011,N,N,3a4anos,N,N,N,432145,16CRS,C42
F,45a49anos,ObesidadeGrau1,Adulto,Branca,N,2014,N,S,1a2anos,S,N,S,432085,16CRS,C15
F,70a74anos,Sobrepeso,Idoso,Branca,S,2011,N,N,2a3anos,S,N,N,432162,16CRS,C42
M,75a79anos,PesoSaudavel,Idoso,Branca,S,2011,N,N,0a1anos,N,N,N,430469,16CRS,C42
M,25a29anos,PesoSaudavel,Adulto,Branca,N,2011,N,N,3a4anos,N,N,N,430620,16CRS,C42
F,15a19anos,ObesidadeGrau1,Jovem,Branca,N,2011,N,N,3a4anos,N,N,N,430680,16CRS,C42
M,60a64anos,PesoSaudavel,Idoso,Branca,N,2014,N,N,1a2anos,S,S,S,432130,16CRS,C32
M,55a59anos,Sobrepeso,Adulto,Branca,N,2011,N,N,3a4anos,N,N,N,430915,13CRS,C42
F,75a79anos,PesoSaudavel,Idoso,Branca,N,2011,N,N,3a4anos,N,N,N,430780,16CRS,C42
F,45a49anos,PesoSaudavel,Adulto,Branca,S,2011,N,N,1a2anos,N,N,N,431140,16CRS,C42
M,50a54anos,PesoSaudavel,Adulto,Negra,S,2011,N,N,0a1anos,N,N,S,431140,16CRS,C34
F,70a74anos,Sobrepeso,Idoso,Branca,S,2011,N,S,0a1anos,N,N,N,431140,16CRS,C18
M,75a79anos,ObesidadeGrau1,Idoso,Negra,N,2011,N,S,3a4anos,N,N,N,430620,16CRS,C44
F,50a54anos,PesoSaudavel,Adulto,Pardo,N,2012,N,N,3a4anos,N,N,S,431140,16CRS,C34
F,65a69anos,PesoSaudavel,Idoso,Branca,N,2011,N,S,3a4anos,N,N,N,431515,16CRS,C16
```

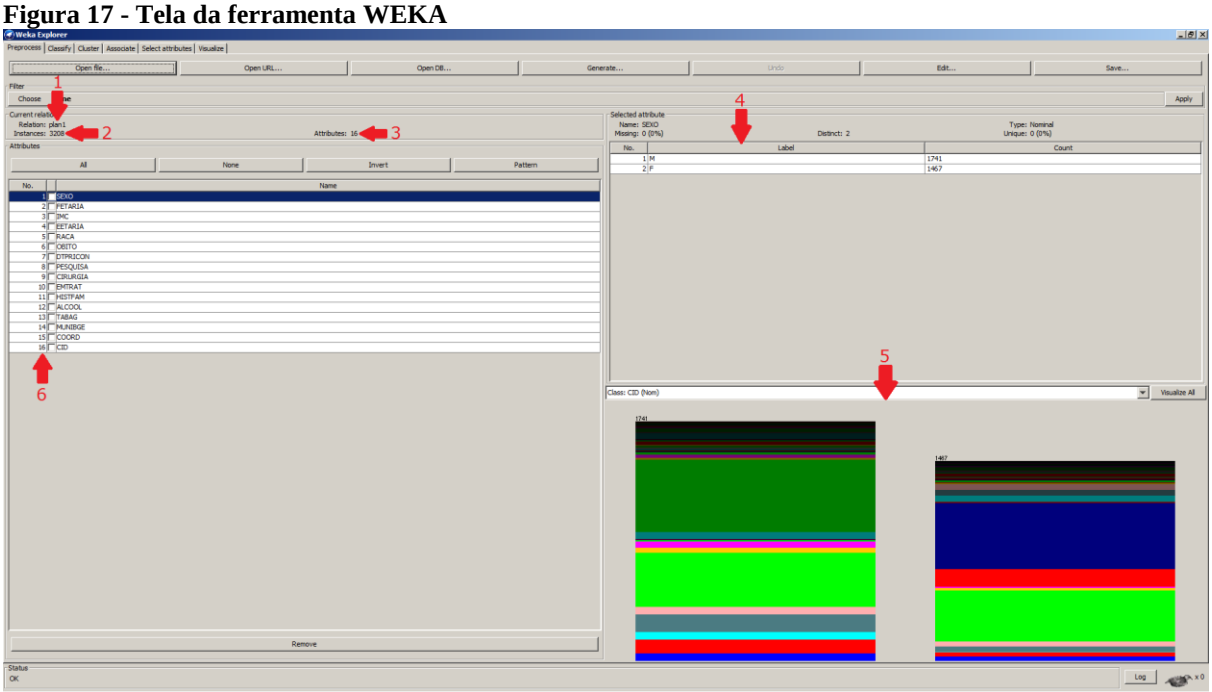
Fonte: Adaptado pelo autor com base no WEKA.

No cabeçalho deste arquivo são listados os atributos de identificação do arquivo, sempre iniciados por "@attribute", após informado o nome do atributo correspondente e por fim, o conjunto de dados identificadores associados ao mesmo.

Abaixo do cabeçalho de atributos, iniciada pela marcação “@data” estão listados os 3208 registros. Durante o processo de importação, o WEKA realizada um pré-processamento das informações garantindo a integridade, desta forma, o sistema verifica se para todas as informações, existe um atributo cadastrado, caso não exista, o sistema retorna um erro e não realiza a importação dos dados.

Na figura 17, após a importação de um arquivo ARFF é possível visualizar os atributos na tela da ferramenta WEKA. No campo *Relation* (1) é informado o nome do arquivo importado. No campo *Instances* (2) são informados os números de instâncias válidas localizadas do arquivo, onde cada linha de dado do arquivo é considerada uma instância. No campo *Attributes* (3) é informado o número de atributos do arquivo, com informações dos atributos cadastrados listados em forma de tabela. Na tabela *Select attribute* (4) são listados todos os dados de cada atributo selecionado na tabela anterior. No gráfico (5) apresenta um histograma das quantidades do atributo selecionado na tabela de atributos. Por fim, a tabela (6) apresenta os atributos.

Figura 17 - Tela da ferramenta WEKA

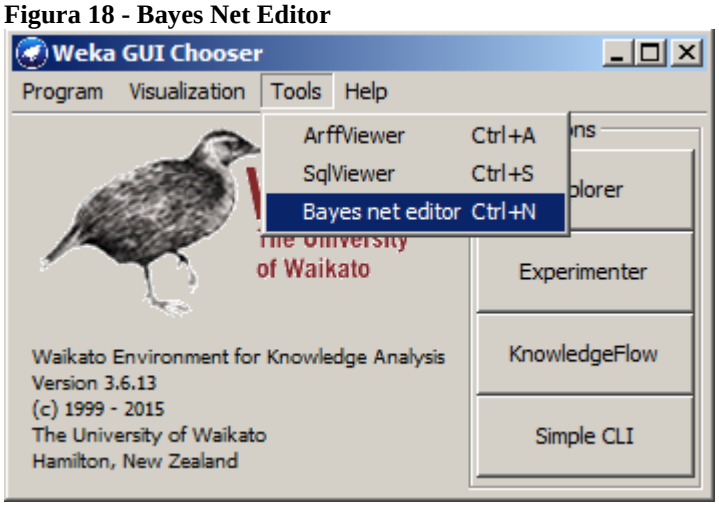


Fonte: Adaptado pelo autor com base no WEKA

6.4 Mineração de Dados

Após importado os dados para a ferramenta WEKA é possível iniciar a etapa de mineração de dados. O WEKA possui um recurso chamado Bayes Net Editor, onde podem ser geradas minerações de dados com base na classificação de Bayes, ilustrado na figura 18.

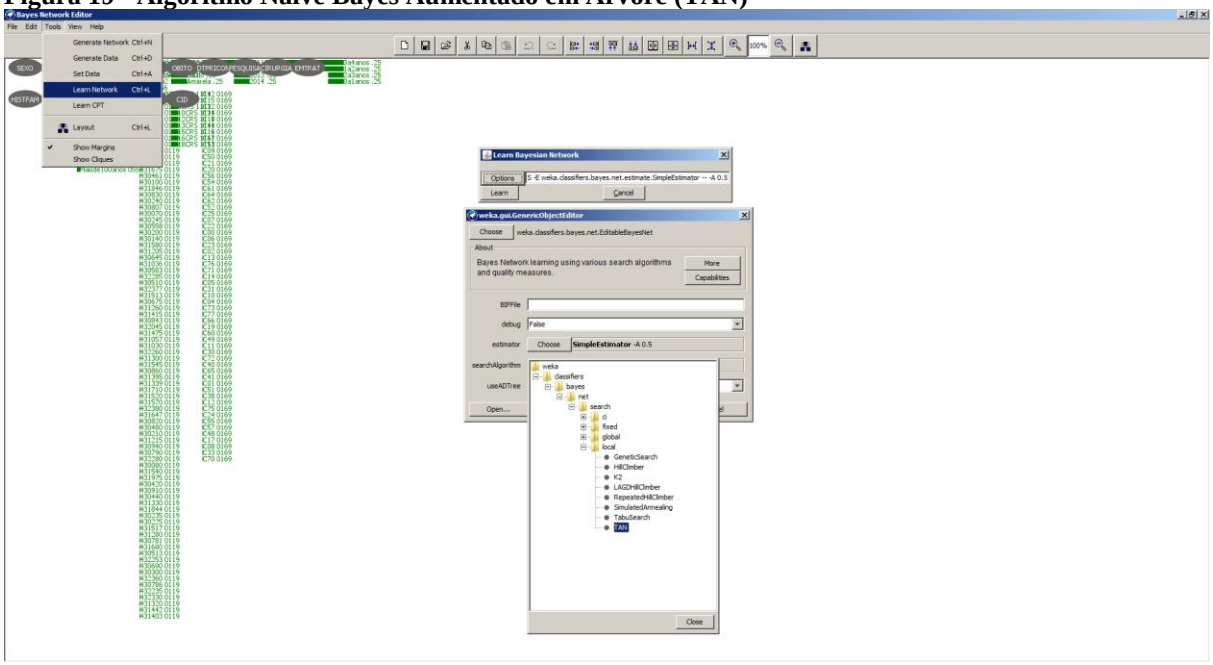
Figura 18 - Bayes Net Editor



Fonte: Adaptado pelo autor com base no WEKA.

Para obter a estrutura da rede deve ser definido o algoritmo de mineração de dados. No estudo foi parametrizado o algoritmo TAN, que permite obter uma rede com uma estrutura de melhor visualização das relações entre atributos, ilustrado na figura 19.

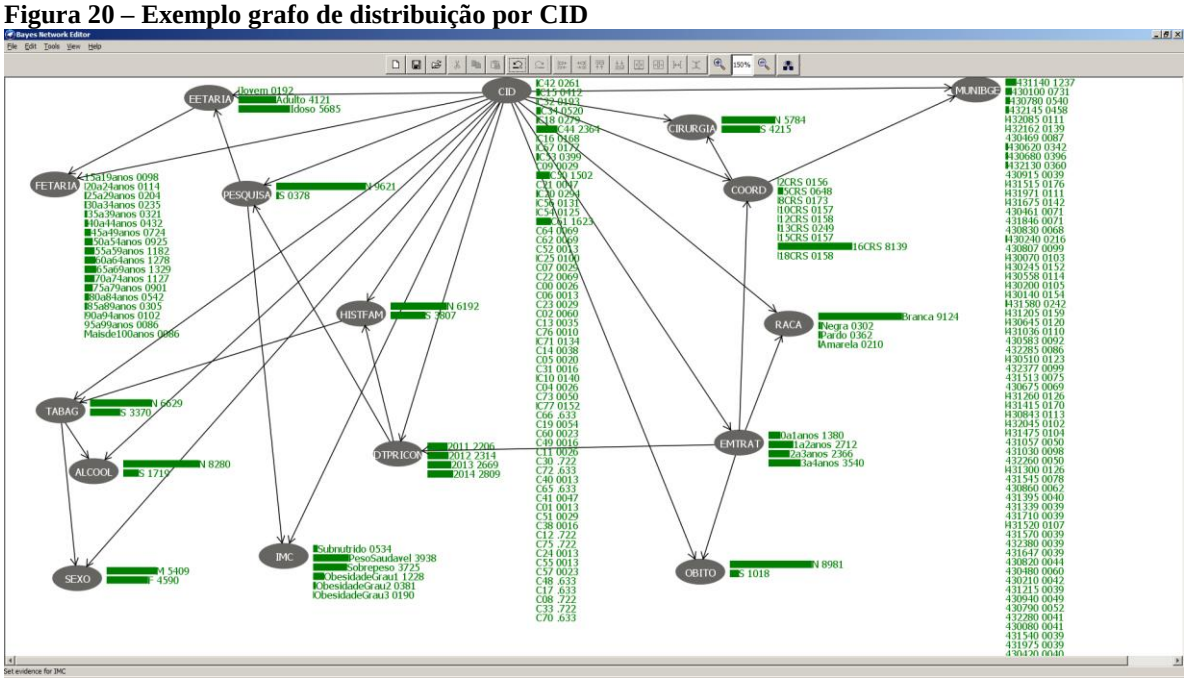
Figura 19 - Algoritmo Naïve Bayes Aumentado em Árvore (TAN)



Fonte: Adaptado pelo autor com base no Bayes Net Editor da ferramenta WEKA.

Após a escolha do algoritmo TAN, foi selecionado a opção *Learn*, onde será exibido um grafo com as relações encontradas e as probabilidades dos valores de cada atributo, ilustrado na figura 20.

Figura 20 – Exemplo grafo de distribuição por CID



Fonte: Adaptado pelo autor com base no Bayes Net Editor da ferramenta WEKA.

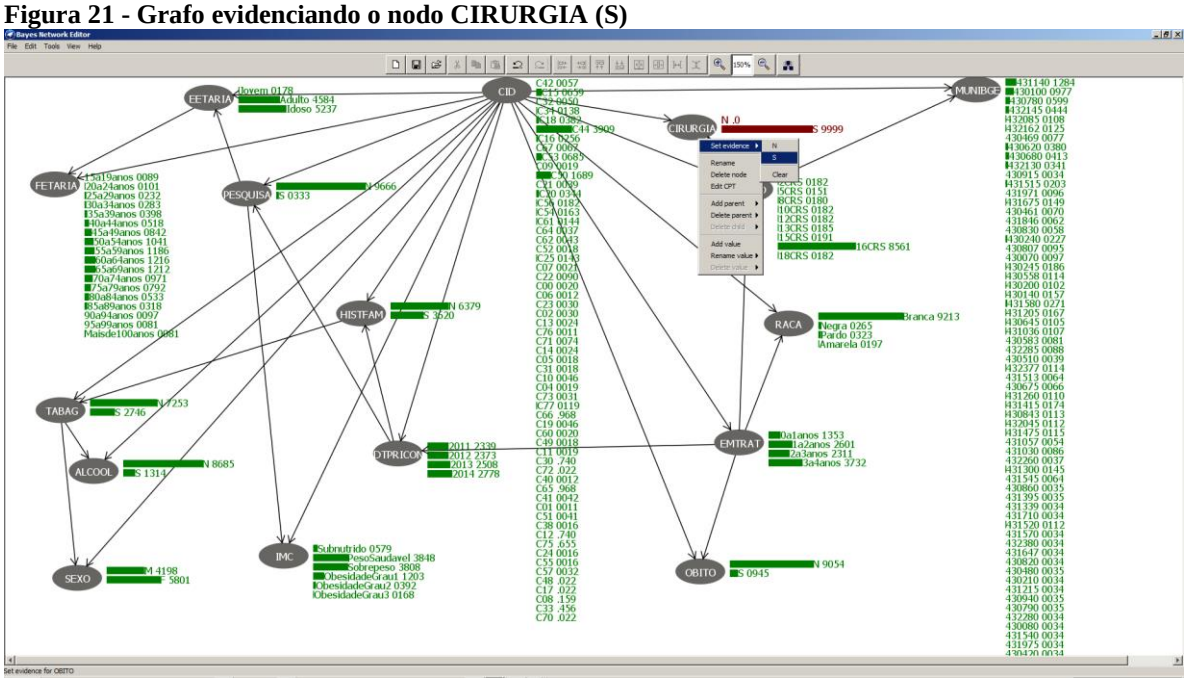
O software WEKA permite que sejam geradas simulações, incluindo ou excluindo relações (arestas) entre os atributos do arquivo.

Além disto, é possível evidenciar valores, essa ação implica em selecionar um atributo (nodo) e determinar que apenas valores iguais ou diferentes a um certo critério devem ser exibidos. Na mineração foi utilizada a opção *Set evidence*, que filtra um único valor no nodo escolhido, em consequência os valores dos nodos relacionados serão atualizados automaticamente, calculando os pesos (percentuais) de todos os nodos.

O atributo classe CID foi executado em todos os experimentos e minerado com o algoritmo TAN, os resultados da mineração serão apresentados em grafos pela ferramenta WEKA.

O primeiro experimento conclusivo foi iniciado evidenciado o nodo CIRURGIA com o valor (S), ao analisar os resultados apresentados, baseado nas afirmações abaixo que os tipos de câncer que os pacientes mais buscam por tratamento cirúrgico são os cânceres de pele (C44) com 39,09% dos casos, mama (C50) com 16,89% dos casos, colo de útero (C53) com 6,85% dos casos e esôfago (C15) com 6,59% dos casos, ilustrado na figura 21.

Figura 21 - Grafo evidenciando o nodo CIRURGIA (S)

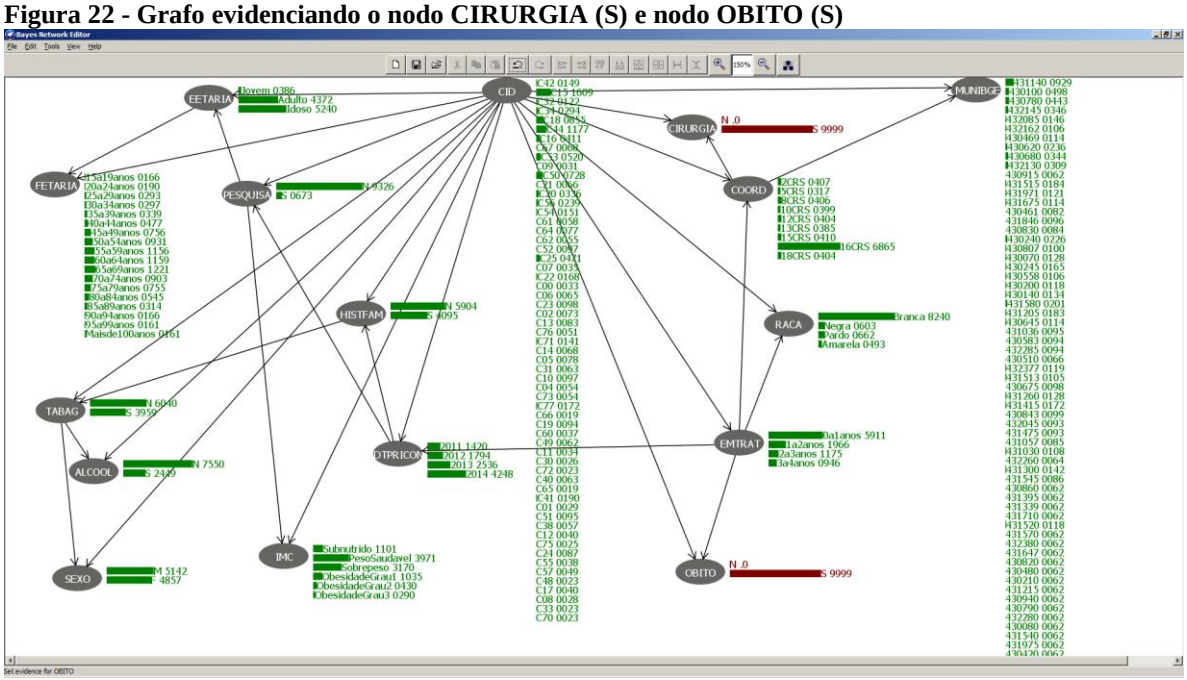


Fonte: Adaptado pelo autor com base no Bayes Net Editor da ferramenta WEKA.

Analisando o nodo EETARIA, observa-se que pacientes que buscam por tratamento cirúrgico estão na estrutura etária (idoso) representando 52,37%, seguidos apenas por pacientes na estrutura etária (adulto) representando 45,84%. Analisando o nodo SEXO, a maior ocorrência em pacientes do sexo feminino representando 58,01%. Analisando o nodo IMC, destacam-se pacientes com resultado (peso saudável) representando 38,48% e pacientes com (sobrepeso) representando 38,08%. Analisando o nodo OBITO observa-se que pacientes que realizaram algum procedimento cirúrgico possuem uma probabilidade de 9,45% de ir a óbito.

Ao evidenciar o novo OBITO com valor (S), a mineração mostra maior probabilidade de óbito para pacientes que realizam cirurgias com câncer de esôfago (C15), ilustrado na figura 22.

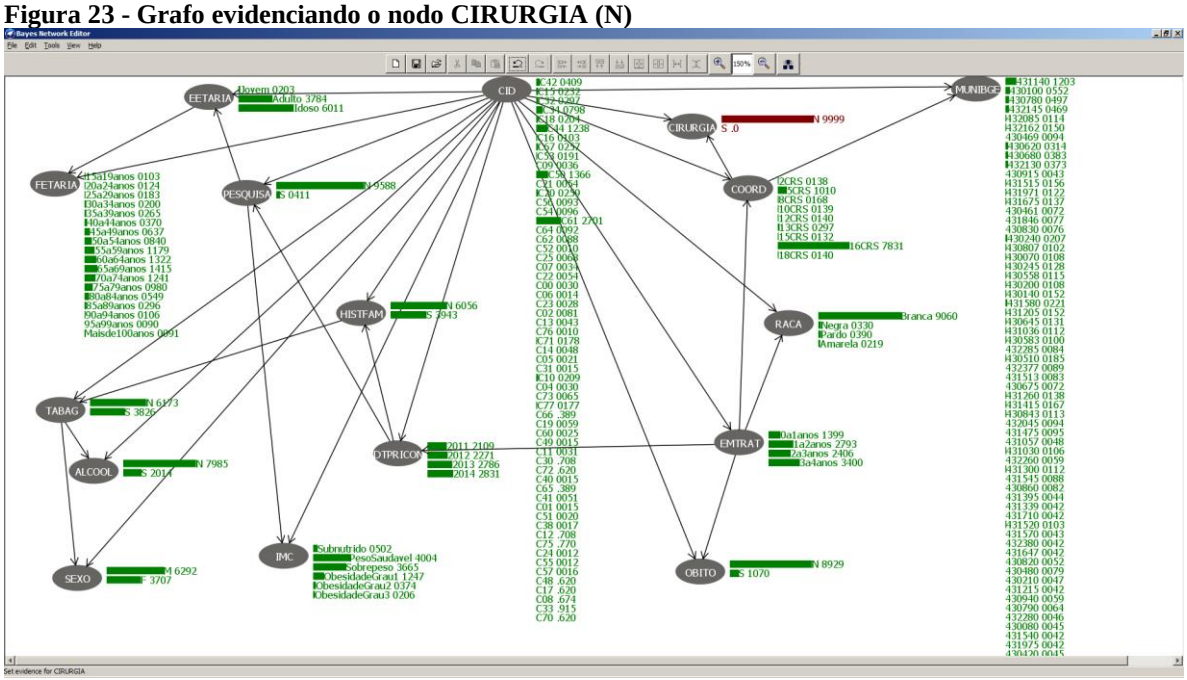
Figura 22 - Grafo evidenciando o nodo CIRURGIA (S) e nodo OBITO (S)



Fonte: Adaptado pelo autor com base no Bayes Net Editor da ferramenta WEKA.

Para finalizar o primeiro experimento foi evidenciado o nodo CIRURGIA com o valor (N) e observa-se no nodo OBITO que os pacientes que não realizaram procedimento cirúrgico têm 10,70% de probabilidade de ir a óbito, percentual um pouco acima dos pacientes que realizam procedimento cirúrgico, ilustrado na figura 23.

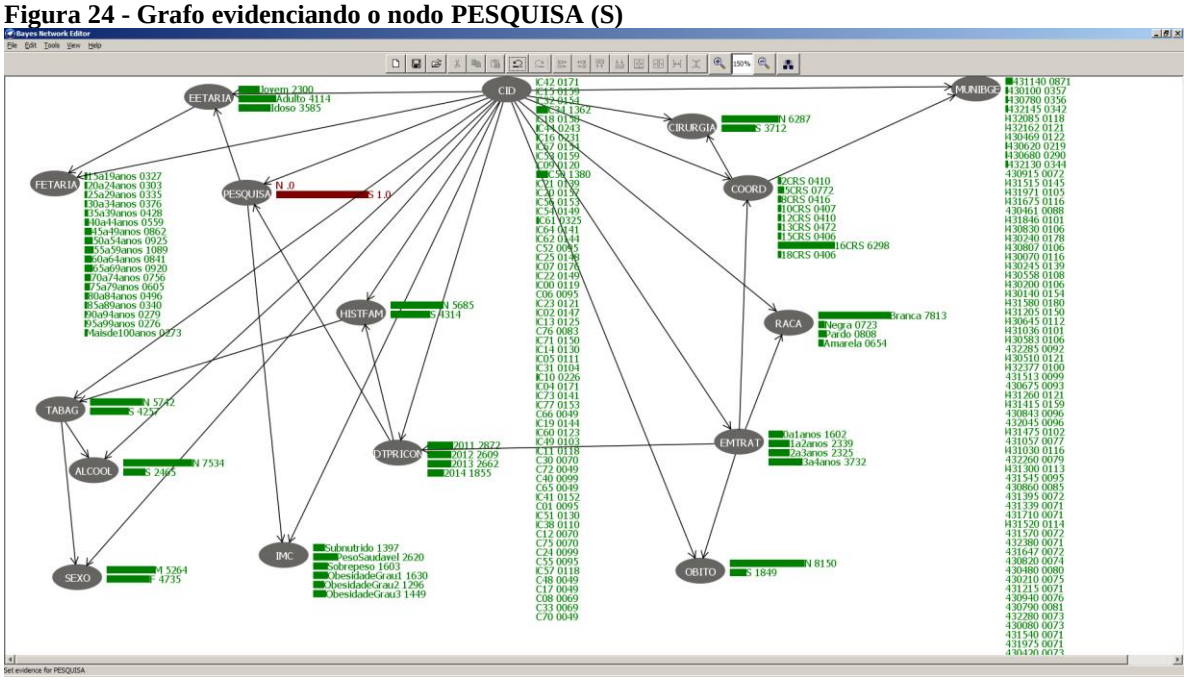
Figura 23 - Grafo evidenciando o nodo CIRURGIA (N)



Fonte: Adaptado pelo autor com base no Bayes Net Editor da ferramenta WEKA.

O segundo experimento conclusivo foi iniciado evidenciando o nodo PESQUISA com o valor (S), ao analisar os resultados apresentados, nota-se que o protocolo de pesquisa está focado em dois tipos de cânceres, o câncer de mama (C50) e o câncer de brônquios e pulmões (C34). Analisando o nodo IMC, há uma probabilidade maior em pacientes com peso saudável representando 26,20% dos casos de pacientes do protocolo de pesquisa, ilustrados na figura 24.

Figura 24 - Grafo evidenciando o nodo PESQUISA (S)

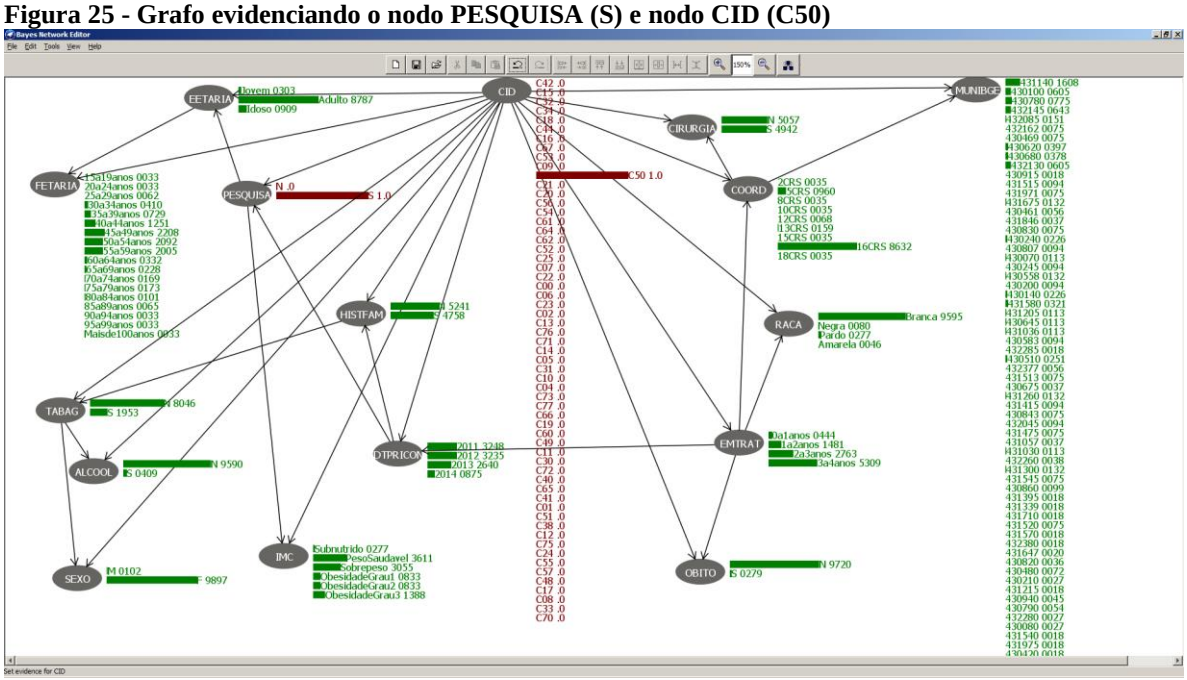


Fonte: Adaptado pelo autor com base no Bayes Net Editor da ferramenta WEKA.

Após foi evidenciado o nodo CID com o valor (C50) e analisado os resultados do nodo OBITO e constatou-se a probabilidade de 97,20% de não ir a óbito pacientes de pesquisa com câncer de mama (C50).

Analisando o nodo EETARIA, observa-se que os pacientes de pesquisa com estrutura etária (adulto) representam 87,87% nos casos de câncer de MAMA (C50). Analisando o nodo FETARIA, destacam-se três faixas etárias dos (45 a 49 anos, 50 a 54anos e 55 a 59 anos), ilustrado na figura 25.

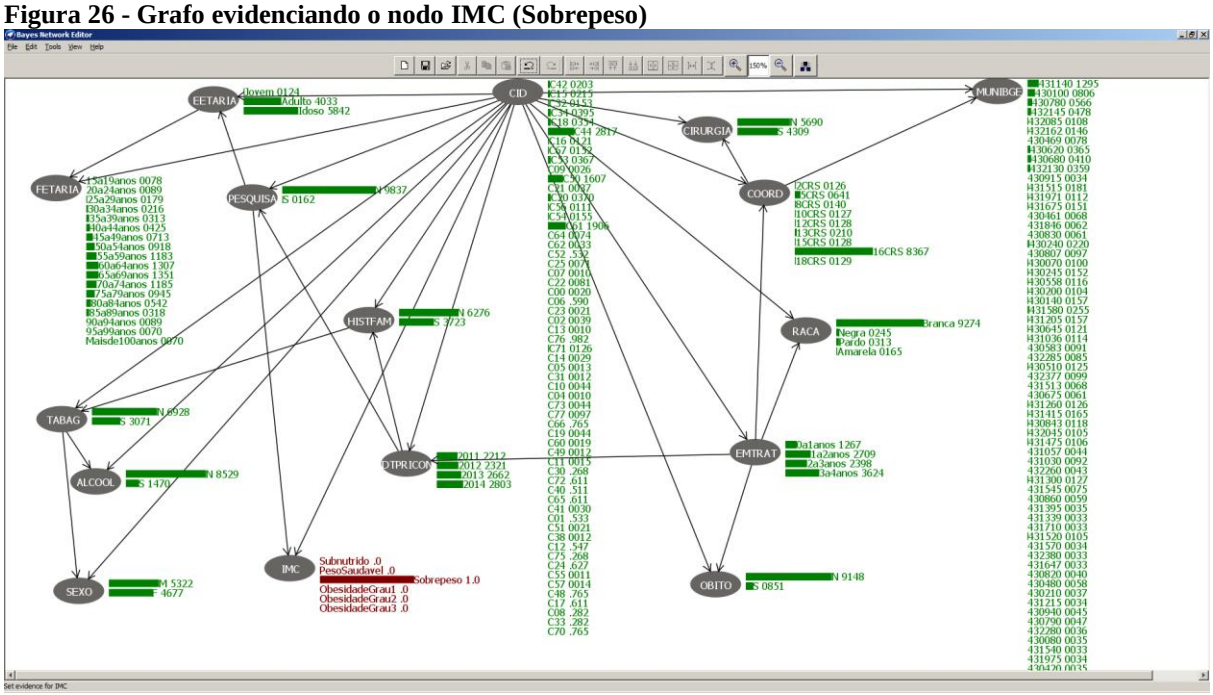
Figura 25 - Grafo evidenciando o nodo PESQUISA (S) e nodo CID (C50)



Fonte: Adaptado pelo autor com base no Bayes Net Editor da ferramenta WEKA.

O terceiro experimento conclusivo foi iniciado evidenciando o nodo IMC com o valor (Sobrepeso), ao analisar os resultados apresentados, conclui-se no nodo MUNIBGE (431140) possui uma predominância maior que correspondente ao município IBGE de Lajeado e os cânceres que se destacam são os de pele (C44) com 28,17%, próstata (C61) com 19,06% e mama (C50) com 16,07%, ilustrado na figura 26.

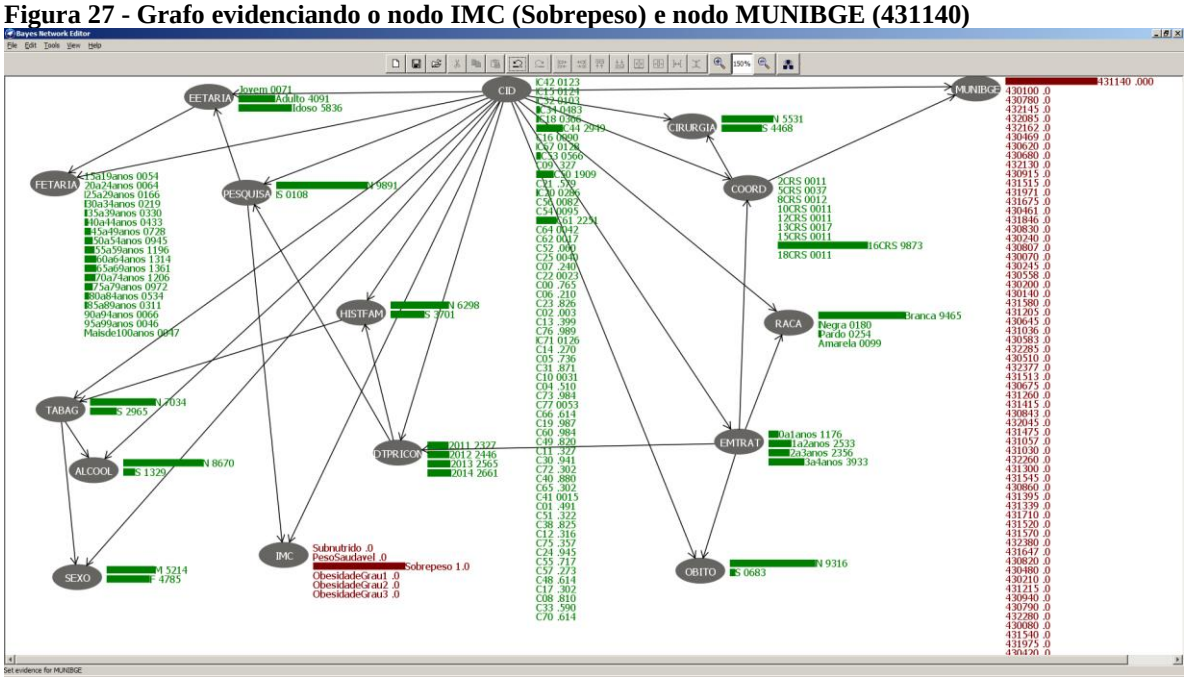
Figura 26 - Grafo evidenciando o nodo IMC (Sobrepeso)



Fonte: Adaptado pelo autor com base no Bayes Net Editor da ferramenta WEKA.

Após foi evidenciado o nodo MUNIBGE com o valor (431140) observa-se que os cânceres de pele (C44), próstata (C61) e mama (C50) continuaram se destacando. Analisando o nodo EETARIA, os pacientes com estrutura etária (idoso) representam 58,36%. Analisando o nodo FETARIA, destacam-se duas faixas etárias dos (60 a 64 anos e 65 a 69 anos), ilustrado na figura 27.

Figura 27 - Grafo evidenciando o nodo IMC (Sobrepeso) e nodo MUNIBGE (431140)

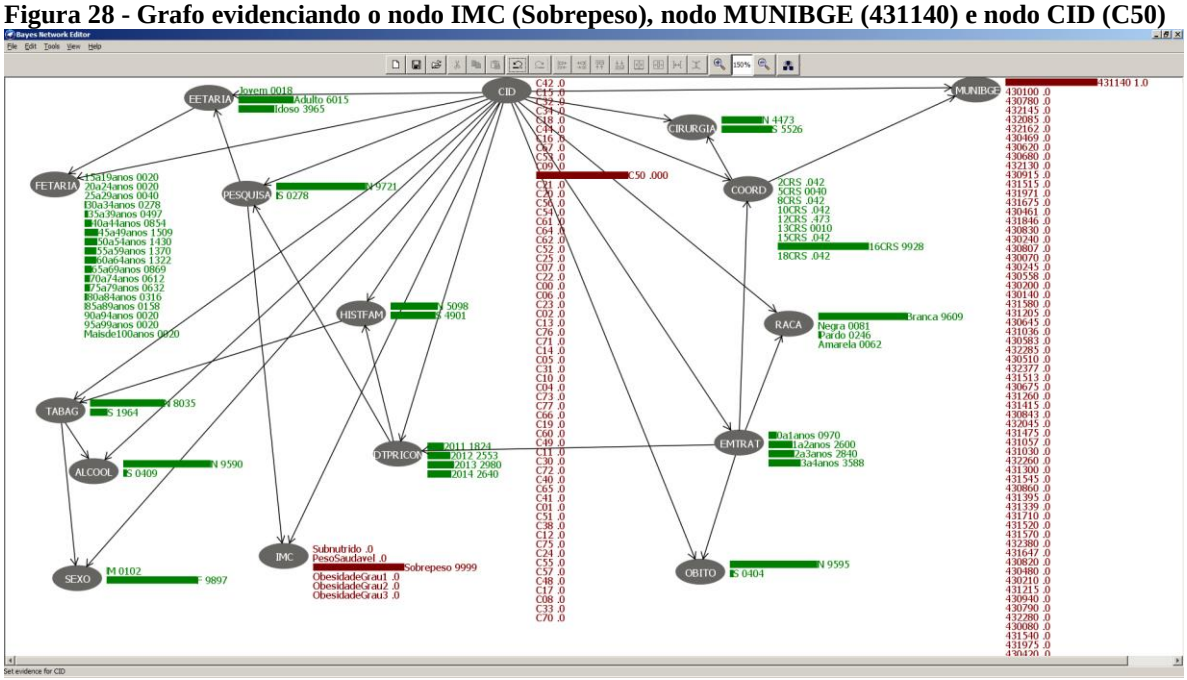


Fonte: Adaptado pelo autor com base no Bayes Net Editor da ferramenta WEKA.

A fim de compreender melhor as informações de cada um dos três tipos de cânceres de pele (C44), próstata (C61) e mama (C50) neste experimento, foi colocado em evidência cada um deles, permanecendo as configurações dos nodos IMC e MUNIBGE.

O primeiro teste foi evidenciado o nodo CID com o valor (C50) e observa-se que no nodo EETARIA destaca-se a estrutura etária (adulto) que corresponde a 60,15%, ilustrado na figura 28.

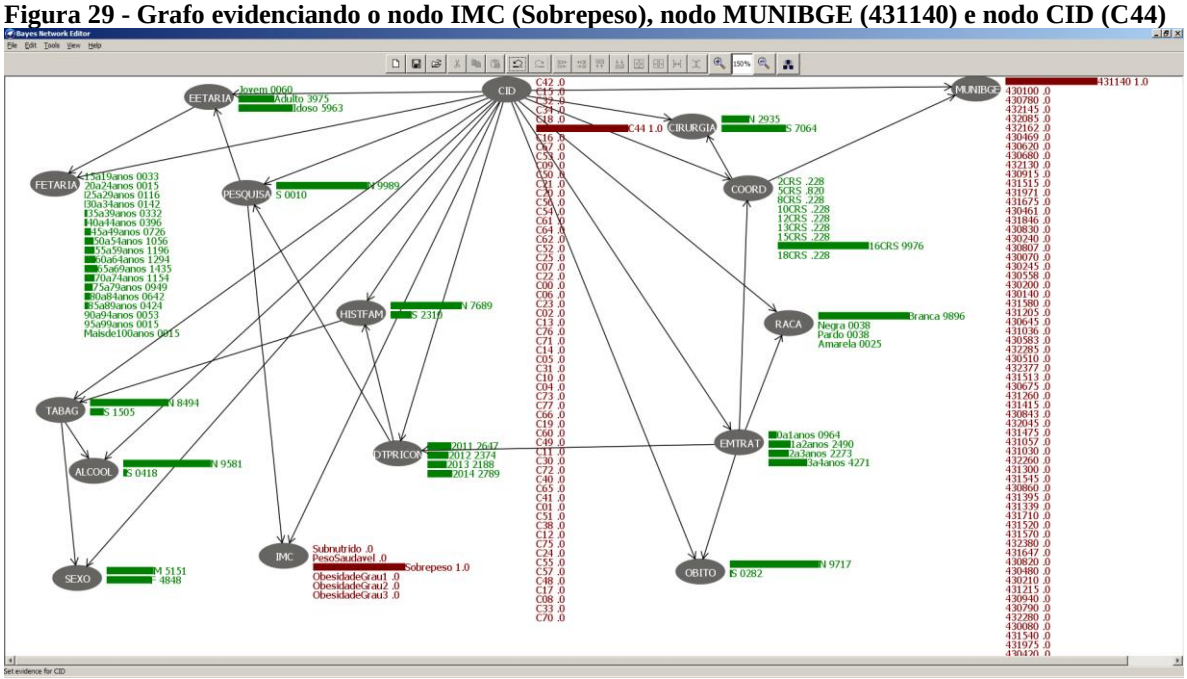
Figura 28 - Grafo evidenciando o nodo IMC (Sobrepeso), nodo MUNIBGE (431140) e nodo CID (C50)



Fonte: Adaptado pelo autor com base no Bayes Net Editor da ferramenta WEKA.

O segundo teste foi evidenciado o nodo CID com o valor (C44) e observa-se que os no nodo EETARIA destaca-se a estrutura etária (idosos) que corresponde 59,63%, ilustrado na figura 29.

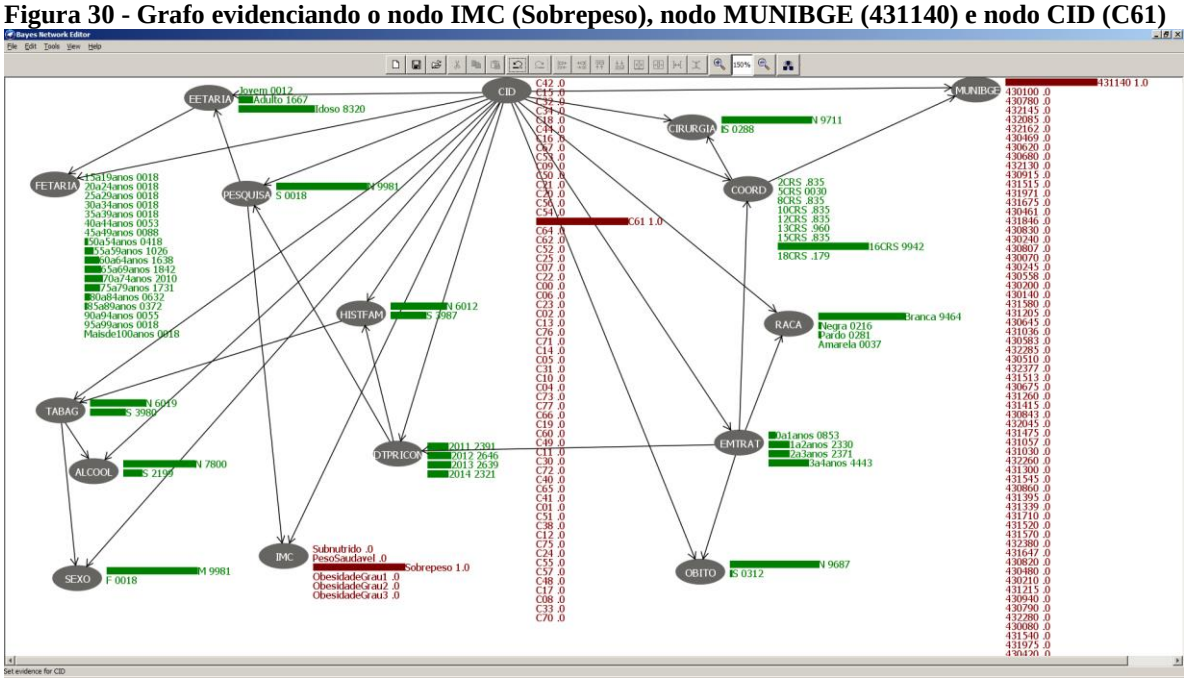
Figura 29 - Grafo evidenciando o nodo IMC (Sobrepeso), nodo MUNIBGE (431140) e nodo CID (C44)



Fonte: Adaptado pelo autor com base no Bayes Net Editor da ferramenta WEKA.

Por fim, foi evidenciado o nodo CID com o valor (C61) e observa-se que o nodo EETARIA destaca-se a estrutura etária (idosos) que corresponde 83,20%, ilustrado na figura 30.

Figura 30 - Grafo evidenciando o nodo IMC (Sobrepeso), nodo MUNIBGE (431140) e nodo CID (C61)



Fonte: Adaptado pelo autor com base no Bayes Net Editor da ferramenta WEKA.

6.5 Interpretação

A quinta e última etapa foi à interpretação e avaliação dos resultados, esta etapa consiste em validar o conhecimento extraído na etapa mineração de dados.

6.5.1 Primeiro Experimento

O planejamento cirúrgico deve incluir todos os cuidados referentes aos princípios gerais da cirurgia e ao preparo do paciente e seus familiares. Através do tratamento cirúrgico, o câncer, em sua fase inicial, pode ser controlado e/ou curado, quando este é o tratamento indicado para o caso.

De acordo com os dados minerados para este estudo, o resultado que magnifica este experimento é que os cânceres de esôfago (C15), pele (C44), mama (C50) e colo de útero (C53) são os tipos de câncer que os pacientes mais buscam por tratamento cirúrgico, dentre estes tipos a maior probabilidade de ir a óbito são pacientes que realizam cirurgias com câncer

de esôfago (C15) uma das hipóteses citadas pela equipe de especialistas, por ser uma cirurgia geralmente de grande porte e potencialmente contaminada.

Foram selecionadas também informações importantes para este experimento extraídas na mineração de dados:

- Pacientes oncológicos primeira consulta apresentam 42,15% de probabilidade de realizar um procedimento cirúrgico;
- Maior ocorrência em pacientes do sexo feminino com 58,01% dos casos analisados;
- Maior ocorrência em pacientes com peso saudável e sobrepeso;
- Maior ocorrência em pacientes que buscam tratamento cirúrgico da estrutura etária (Idoso), seguidos apenas pelos pacientes com estrutura etária (Adulto);
- Pacientes que realizaram algum procedimento cirúrgico possuem uma probabilidade de 9,45% de ir a óbito, já os pacientes que não realizaram procedimento cirúrgico tem o percentual de 10,70% de ir a óbito;

Do ponto de vista dos gestores, as informações apresentadas sobre o primeiro experimento são relevantes, pois trazem uma informação consolidada extraída de dois sistemas (TASY e SISRHC), de pacientes oncológicos que optaram por tratamento cirúrgico do Centro de Oncologia, que a partir destas informações poderão ser trabalhadas em conjunto com o Centro Cirúrgico para tomada de decisão visando o bem estar do paciente.

6.5.2 Segundo Experimento

O desafio dos profissionais da área oncológica consiste em encontrar a maneira mais eficaz de tratar a doença com o mínimo de efeitos colaterais para o paciente. Além dos tratamentos convencionais de quimioterapia e radioterapia, a Casa de Saúde oferece acesso ao programa de tratamento com novas drogas, que permite a participação de pacientes em protocolos de pesquisa nacionais e/ou internacionais.

De acordo com os dados minerados para este estudo, o resultado que magnifica este experimento é que os pacientes com câncer de mama (C50) que optaram pelo medicamento

do protocolo de pesquisa apresentaram a probabilidade de 97,20% de eficiência, uma das hipóteses que o tratamento de pesquisa pode proporcionar uma ocasião única de mudar o curso da doença, quando já não respondem aos tratamentos tradicionais.

Foram selecionadas também informações importantes para este experimento extraídas na mineração de dados:

- O protocolo de pesquisa está focado em dois tipos de cânceres no câncer de mama (C50) e o câncer de brônquios e pulmões (C34);
- Maior ocorrência do protocolo de pesquisa em pacientes com peso saudável;
- Pacientes com estrutura etária (adulto) representam 87,87% dos tratamentos de pesquisa nos casos de neoplasia de mama (C50);
- Pacientes que participaram do protocolo de mama (C50) de pesquisa estão entre três faixas etárias dos (45 a 49 anos, 50 a 54anos e 55 a 59 anos);

Do ponto de vista dos gestores, as informações apresentadas sobre o segundo experimento, confirmam suas deduções e trazem uma riqueza de informações para a Pesquisa Clínica em Oncologia.

6.5.3 Terceiro Experimento

Cada vez mais a alimentação tem sido associada a um maior risco no desenvolvimento do câncer, algumas mudanças nos nossos hábitos alimentares podem nos ajudar a reduzir os riscos.

De acordo com os dados minerados para este estudo, o resultado que magnifica este experimento é que pacientes com o resultado IMC (sobrepeso) destacam-se em três tipos de cânceres, o câncer de pele (C44), próstata (C61) e mama (C50) com maior predominância no município IBGE (431140) correspondente a Lajeado, umas das hipóteses é de que as pessoas têm optado por alimentos práticos, como comidas semiprontas, que por sua vez, é uma alimentação pobre em nutrientes vitais ao nosso organismo.

Estes alimentos práticos são facilmente encontrados nos centros de comidas de fast-food, que estão em alta em Lajeado, porém estes alimentos são pobres em fibras, com altos

teores de gorduras e altos níveis calóricos estando relacionada a um maior risco para o desenvolvimento destes tipos de cânceres (INCA, 2013).

Foram selecionadas também informações importantes para este experimento extraídas na mineração de dados:

- O câncer de mama (C50) destaca-se em pacientes com estrutura etária (adulto) e os cânceres de pele (C44) e próstata (C61) em pacientes com estrutura etária (idosos) no município IBGE (431140) correspondente a Lajeado.

Do ponto de vista dos gestores, as informações apresentadas sobre o terceiro experimento, confirmam o que esta sendo comentado nos últimos tempos sobre as desvantagens dos alimentos práticos. Uma sugestão vinda do autor do estudo, para realização de um projeto interno em parceria com a Nutrição e o Centro de Oncologia, na formulação de uma cartilha aos colaboradores da instituição sobre os riscos da má alimentação, podendo expandir para a comunidade em geral.

7 CONCLUSÕES

Com o desenvolvimento deste estudo, foi possível destacar a importância do uso de técnicas de mineração de dados para a descoberta de conhecimento, especialmente quando aplicadas no caso proposto.

Através do problema de pesquisa apresentado e dos objetivos iniciais, foi possível identificar que todas as etapas foram cumpridas com sucesso, podendo ser aproveitada para uso na gestão do Centro de Oncologia da Casa de Saúde e ser relevante para a sociedade em geral.

A revisão bibliográfica apresentada evidenciou a importância e validade dos conceitos relacionados à mineração de dados que podem ser expandidos para outras especialidades médicas na Casa de Saúde.

Com a análise dos dados coletados, percebe-se que o algoritmo TAN contribuiu para a descoberta de conhecimento, pois além de relacionar o nodo classe com os demais nodos do grafo, a relação entre os nodos ajudou a verificar a característica dos pacientes oncológicos.

Com a análise foi possível apurar no primeiro experimento que os cânceres de esôfago (C15), pele (C44), mama (C50) e colo de útero (C53) são os tipos de câncer que os pacientes mais buscam por tratamento cirúrgico, dentre estes tipos a maior probabilidade de ir a óbito são pacientes que realizam cirurgias com câncer de esôfago (C15). Através do tratamento cirúrgico, o câncer, em sua fase inicial, pode ser controlado e/ou curado, quando este é o tratamento indicado para o caso, mas irá depender de diversos fatores, que serão discutidos entre o médico e o paciente.

No segundo experimento foi identificado que os pacientes com câncer de mama (C50) que optaram pelo medicamento do protocolo de pesquisa apresentaram um grande potencial de eficiência, pois o tratamento de pesquisa pode proporcionar uma ocasião única de mudar o curso da doença, quando já não respondem aos tratamentos tradicionais.

No terceiro experimento foram identificados três tipos de cânceres de pele (C44), próstata (C61) e mama (C50) estando relacionados a pacientes com sobrepeso na cidade de Lajeado, estes tipos de cânceres podem estar associadas à má alimentação, que tem sido um dos maiores riscos no desenvolvimento do câncer, algumas mudanças nos nossos hábitos alimentares podem ajudar a reduzir os riscos.

Sendo assim, conclui-se que os resultados alcançados no estudo atendem os objetivos do trabalho, e que o método de descoberta não supervisionada e a ferramenta WEKA mostraram-se eficientes e preciso na apuração das informações.

7.1 Trabalhos Futuros

Com o estudo formalizado e avaliado, trabalhos futuros podem focar na descoberta de conhecimento em outras especialidades médicas, como Redeker (2010) desenvolveu seu estudo em pacientes da especialidade de cardiologia e o presente estudo desenvolvido em pacientes da especialidade de oncologia.

Durante o período de visitas no Centro de Oncologia, percebeu-se a falta de formalização de registro das atividades desempenhadas sistematicamente, pois as mesmas são executadas conforme a compreensão dos colaboradores, não havendo um fluxo a ser seguido. Uma sugestão é uma reavaliação do processo efetuado nos sistemas do Centro de Oncologia, onde as atividades possam ser mapeadas, posteriormente exploradas pelos analistas de sistemas da instituição e descritas no manual de tarefas, para haja uma qualidade nas informações em todas as fontes de dados e para que os colaboradores possam saber do fluxo correto dos processos do Centro de Oncologia.

Em 2016, a Casa de Saúde tem programada a migração do banco de dados Oracle para a versão 11g, onde esta versão é dotada nativamente de recursos para a utilização de técnicas de mineração de dados, que poderão ser utilizadas em um futuro mestrado pelo autor.

REFERÊNCIAS

ABICALAFFE, C. L.L.; AMARAL, V. F.; DIAS, J. S. **Aplicação Da Rede Bayesiana na Prevenção da Gestação de Alto Risco.** Paraná: Dissertação de Pós-Graduação da Pontifícia Universidade Católica do Paraná, 2000.

AMO, S.. **Técnicas de mineração de dados.** XXIV Congresso da Sociedade Brasileira de Computação, 2004.

BARBETTA, P. A. **Estatísticas aplicadas às Ciências Sociais.** 1ª Ed. Florianópolis: UFSC, 1994.

BRASIL, IBGE, **Distribuição da população segundo os grupos de idade**. Obtida via internet. Disponível em: <www.ibge.gov.br> Acesso em: 02 out. 2015.

BERKA, Pert, RAUCH, Jan, ZIGHED Djamel A., **DM and Medical Knowledge Management: Cases and Applications.** New York: Hershey, 2009.

BOYD, H. W., WESTFALL, R. STASCH, S. F. **Markting Research: text and cases.** Illinois: Richard D. Irwin, 1989.

CARDOSO O. N. P.; MACHADO R. T. M. **Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras.** Obtida via internet. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-76122008000300004>. Acesso em: 19 mai. 2015.

CARVALHO, L. A. V. - **Datamining: a mineração de dados no marketing, medicina, economia, engenharia e administração** - Ciência Moderna - RJ, 2005.

CERVO, Amado Luiz; BERVIAN, Pedro Alcino; SILVA, Roberto da. **Metodologia científica.** 6. ed. São Paulo: Pearson Prentice Hall, 2007. E-book. Disponível em: <http://univates.bv3.digitalpages.com.br/users/publications/9788576050476/pages/_1>. Acesso em: 21 maio 2015.

COLLIS, Jill; HUSSEY, Roger**. Pesquisa em Administração: Um guia prático para alunos de graduação e pós-graduação**. 2. ed. Porto Alegre: Bookman, 2005.

DATE, C. J. **Introdução a Sistemas de Bancos de Dados.** 8ª Ed., Rio de Janeiro: Campus, 2004.

DELGIGLIO, Auro; HOLTZ, Luciana; KALIKS, Rafael. **Os diferentes tipos de tratamento.** Obtida via internet. Disponível em: < http://www.pacientecomcancer.com/capitulo/5/>. Acesso em: 19 mai. 2015.

FADUL, Sergio; ÉBOLI, Evandro. **'País terá meio milhão de casos de câncer em 2014 e 2015', diz presidente do Inca.** Obtida via internet. Disponível em: <http://oglobo.globo.com/sociedade/saude/pais-tera-meio-milhao-de-casos-de-cancer-em-2014-2015-diz-presidente-do-inca-14694793>. Acesso em: 06 jun. 2015.

FAYYAD, Usama, PIATETSKY-SHAPIRO, Gregory e SMYTH, Padhraic. **From Data Mining to Knowlegde discovery**. American Association for Artificial Intelligence. 1996.

FILHO, H. P.F. **Aplicabilidade de Memória Lógica como Ferramenta Coadjuvante no Diagnóstico das Doenças Genéticas.** Goiás: Dissertação de Mestrado da Universidade Católica de Goiás, 2006.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa.** 5ª Ed. São Paulo: Atlas, 2010.

HAN, J.; KAMBER, M.; Pei, J**. Data mining: Concepts and techniques.** San Francisco, USA: Morgan Kaufmann Publishers Inc., 2005.

JUNG, Carlos Fernando**. Metodologia Científica: Ênfase em Pesquisa Tecnológica.** 3 ª Ed. rev. e ampl. 2003. Disponível em:< http://www.jung.pro.br>. Acesso em: 02 mai. 2015.

LAKATOS, Eva Maria; MARCONI, Marina de Andrade. **Fundamentos de metodologia científica.** 7 ª Ed. São Paulo: Atlas, 2010.

LEMOS, Elaine P. **Análise de Crédito Bancário com o uso de Data Mining: Redes Neurais e Árvores de Decisão.** Curitiba: Dissertação de Mestrado da Universidade Federal do Paraná, 2003.

MALHOTRA, N. K. **Pesquisa de Marketing: uma orientação aplicada.** 3ª Ed. Porto Alegre: Bookman, 2001.

MATTAR, F. N. **Pesquisa de marketing**. 3ª Ed. São Paulo: Atlas, 2001.

NASCIMENTO, R.J.O. **Mineração e Análise de Dados em SQL.** Obtida via internet. Disponível em: <http://www.devmedia.com.br/mineracao-e-analise-de-dados-em-sql/29337>. Acesso em: 22 mai. 2015.

NEAPOLITAN, R. E.,**Learning Bayesian Networks.** New Jersey, USA: Prentice Hall, 2004.

ORGANIZAÇÃO MUNDIAL DA SAÚDE - OMS. **Obesidade. Impedindo e controlando a epidemia global.** Genebra, 1997.

PACHIAROTTI, J.F.B. **Aplicação de Técnicas de Mineração de Dados no aprimoramento do Atendimento Médico em um Cenário de Plano de Saúde.** Espírito Santo. Dissertação de graduação da Universidade Vila Velha, 2012.

PHILIPS CLINICAL INFORMATICS. **Tasy Prestador**. Obtida via internet. Disponível em: <http://www.cilatam.philips.com.br/solucoes/13/tasy-prestador/>. Acesso em: 02 out. 2015.

PONTES, L. B.. **O que é oncologia.** Disponível em:<http://www.einstein.br/einstein-saude/em-dia-com-a-saude/Paginas/o-que-e-oncologia.aspx> Acesso em: 19 mai. 2015.

RAMOS, D. R. **Data Warehouse.** Disponível em:<http://www.devmedia.com.br/data-warehouse/12609> Acesso em: 02 out. 2015.

REDEKER, G. A.. **Descoberta de Conhecimento na área de Cardiologia.** Lajeado. Dissertação de graduação do Centro Universitário UNIVATES, 2010.

REZENDE, S.O. ; PUGLIESI, J. B. ; MELANDA, E. A.; PAULA, M. F., **Mineração de Dados. In Solange Oliveira Rezende. (Org.). Sistemas Inteligentes – Fundamentos e Aplicações.** Barueri, SP. Editora Ltda, 2003.

RIBEIRO, T. L. **Índice de Massa Corporal (IMC)** Obtida via internet. Disponível em: < http://www.indicedemassacorporal.com/index.html> Acesso em: 03 out. 2015.

RICHARDSON, R. J. **Pesquisa social: métodos e técnicas.** 3 ª Ed. São Paulo: Atlas, 1999.

ROSSO, Leandro – **Oncomining - protótipo utilizando técnicas Data Mining.** Obtida via internet. Disponível em: <http://www.unochapeco.edu.br/saa/tese/2502/TCC%202%20-%20LEANDRO%20ROSSO.pdf> Acesso em: 19 mai. 2015.

SANTOS, M.Y. ; RAMOS, R. **Business Intelligence – Tecnologias da Informação na Gestão de Conhecimento**, FCA, 2a Edição, 2009.

SASSE, André. **E-Cancer Informações em Câncer e Oncologia.** Obtida via internet. Disponível em: <http://andre.sasse.com/oncologia>. Acesso em: 19 mai. 2015.

SARABANDO, A. C. L.; **Um estudo do comportamento de Redes Bayesianas no prognóstico da sobrevivência no cancro da próstata.** Porto: Dissertação de Mestrado da Universidade do Porto, 2010.

SILVEIRA, Vinicius. **O que é Data Warehouse?** Obtida via internet. Disponível em: <http://blog.intuitivus.com.br/pt/o-que-e-data-warehouse/>. Acesso em: 06 jun. 2015.

SIRUGI, Fernando**. Estrutura Etária da População**. Obtida via internet. Disponível em: <http://www.infoescola.com/geografia/estrutura-etaria-da-populacao/> Acesso em: 03 out. 2015.

SOURCEFORGE. **Excel2ArffConverter.** Obtida via internet. Disponível em: <http://sourceforge.net/projects/exceltoarffconv/>. Acesso em: 01 ago. 2015.

SOUZA, V.D.M.; CARVALHO, D.R. **Avaliando a relação entre quão corretas e interessantes podem ser as regras de associação descobertas.** Obtida via internet. Disponível em: <http://www.utp.br/tuiuticienciaecultura/TCC_online/Avaliando%20a%20rela%C3%A7%C3%A3o/avaliando.pdf>. Acesso em: 19 mai. 2015.

TAN, P-N; STEINBACH, M.; e KUMAR, V. - **Introdução ao Data Mining - Mineração de Dados.** – Ciência Moderna Ltda - RJ, 2009.

WITTEN, Ian H., FRANK, Eibe. **DM: Practical Machine Learning Tools and Techniques, Second Edition.** San Francisco: Morgan Kaufmann Publishers, 2005.

ANEXOS

ANEXO A – FAIXA ETÁRIA IBGE

Faixa etária IBGE	Idade Inicial	Idade Final
Mais de 100 anos	100 anos	-
95 a 99 anos	95 anos	99 anos e 11 meses
90 a 94 anos	90 anos	94 anos e 11 meses
85 a 89 anos	85 anos	89 anos e 11 meses
80 a 84 anos	80 anos	84 anos e 11 meses
75 a 79 anos	75 anos	79 anos e 11 meses
70 a 74 anos	70 anos	74 anos e 11 meses
65 a 69 anos	65 anos	69 anos e 11 meses
60 a 64 anos	60 anos	64 anos e 11 meses
55 a 59 anos	55 anos	59 anos e 11 meses
50 a 54 anos	50 anos	54 anos e 11 meses
45 a 49 anos	45 anos	49 anos e 11 meses
40 a 44 anos	40 anos	44 anos e 11 meses
35 a 39 anos	35 anos	39 anos e 11 meses
30 a 34 anos	30 anos	34 anos e 11 meses
25 a 29 anos	25 anos	29 anos e 11 meses
20 a 24 anos	20 anos	24 anos e 11 meses
15 a 19 anos	15 anos	19 anos e 11 meses
10 a 14 anos	10 anos	14 anos e 11 meses
5 a 9 anos	5 anos	9 anos e 11 meses
0 a 4 anos	0 ano	4 anos e 11 meses

Fonte: Do autor (2015), com base nos dados IBGE

ANEXO B – ESTRUTURA ETÁRIA

Estrutura Etária	Idade Inicial	Idade Final
Idoso	60 anos	-
Adulto	20 anos	59 anos e 11 meses
Jovem	0 ano	19 anos e 11 meses

Fonte: Do autor (2015), com base nos dados IBGE

ANEXO C - CID-O - 3ª-EDIÇÃO (CLASSIFICAÇÃO INTERNACIONAL DE DOENÇAS PARA ONCOLOGIA – TERCEIRA EDIÇÃO)

CID-O	LOCALIZAÇÃO DO TUMOR
C00	Lábio
C01	Base da língua
C02	Outras partes e partes não especificadas da língua
C04	Assoalho da boca
C05	Palato
C06	Outras partes e partes não especificadas da boca
C07	Glândula parótida
C08	Outras glândulas salivares maiores e as não específicadas
C09	Amígdala
C10	Orofaringe
C11	Nasofaringe
C12	Seio periforme
C13	Hipofaringe
C14	Outras localizacoes, e as mal definidas, do lábio, cavidade oral e faringe
C15	Esôfago
C16	Estômago
C17	Intestino delgado
C18	Cólon
C19	Junção retossigmoidiana
C20	Reto
C21	Ânus e canal anal
C22	Fígado e vias biliares intra-hepáticas
C23	Vesícula biliar
C24	Outras partes e partes não-especificadas das vias biliares
C25	Pâncreas
C30	Cavidades nasal e ouvido médio
C31	Seios da face
C32	Laringe
C33	Traquéia
C34	Brônquios e pulmões
C38	Coração, mediastino e pleura
C40	Ossos, articulações e cartilagens articulares dos membros
C41	Neoplasia maligna dos ossos e das cartilagens articulares de outras localizações não-especificadas
C42	Sistemas hematopoético e reticuloendotelial
C44	Pele
C48	Retroperitônio e peritônio
C49	Tecido conjuntivo, subcutâneo e outros tecidos moles
C50	Mama
C51	Vulva
C52	Vagina
C53	Colo do útero
C54	Corpo do útero
C55	Útero
C56	Ovário
C57	Outros órgãos genitais femininos e os não especificados
C60	Pênis
C61	Próstata
C62	Testículo
C64	Rim
C65	Pelve renal
C66	Ureteres
C67	Bexiga
C70	Meninges
C71	Encéfalo
C72	Medulal espinhal, nervos cranianos e ouras partes do sistema nervoso central

C73	Glândula tiróide
C75	Outras glândulas endócrinas e estruturas relacionadas
C76	Outras localizações e localizações mal definidas
C77	Linfonodos (gânglios linfáticos)
C80	Localização primária desconhecida

Fonte: Do autor (2015), com base nos dados do SISRHC

ANEXO D – MUNICÍPIO IBGE COM A MICRORREGIÃO

MUNICÍPIO IBGE (CÓDIGO)	MUNICÍPIO IBGE (LOCALIDADE)	MICRORREGIÃO
430060	Alvorada	1º CRS
430087	Ararica	1º CRS
430310	Cachoeirinha	1º CRS
430390	Campo Bom	1º CRS
430460	Canoas	1º CRS
430640	Dois Irmaos	1º CRS
430760	Estancia Velha	1º CRS
430770	Esteio	1º CRS
430905	Glorinha	1º CRS
430920	Gravatai	1º CRS
431080	Ivoti	1º CRS
431162	Lindolfo Collor	1º CRS
431247	Morro Reuter	1º CRS
431306	Nova Hartz	1º CRS
431337	Nova Santa Rita	1º CRS
431340	Novo Hamburgo	1º CRS
431480	Portao	1º CRS
431490	Porto Alegre	1º CRS
431514	Presidente Lucena	1º CRS
431695	Santa Maria do Herval	1º CRS
431870	Sao Leopoldo	1º CRS
431990	Sapiranga	1º CRS
432000	Sapucaia do Sul	1º CRS
432300	Viamao	1º CRS
430085	Arambare	2º CRS
430110	Arroio dos Ratos	2º CRS
430165	Barao	2º CRS
430175	Barao do Triunfo	2º CRS
430190	Barra do Ribeiro	2º CRS
430265	Brochier	2º CRS
430270	Butia	2º CRS
430350	Camaqua	2º CRS
430360	Cambara do Sul	2º CRS
430468	Capela de Santana	2º CRS
430517	Cerro Grande do Sul	2º CRS
430535	Charqueadas	2º CRS
430544	Chuvisca	2º CRS
430650	Dom Feliciano	2º CRS
430676	Eldorado do Sul	2º CRS
430880	General Camara	2º CRS
430930	Guaiba	2º CRS
430955	Harmonia	2º CRS
431010	Igrejinha	2º CRS
431179	Marata	2º CRS
431198	Mariana Pimentel	2º CRS
431225	Minas do Leao	2º CRS
431240	Montenegro	2º CRS
431403	Pareci Novo	2º CRS
431405	Parobe	2º CRS
431575	Riozinho	2º CRS
431600	Rolante	2º CRS
431650	Salvador do Sul	2º CRS
431820	Sao Francisco de Paula	2º CRS
431840	Sao Jeronimo	2º CRS
431848	Sao Jose do Hortencio	2º CRS

431861	Sao Jose do Sul	2º CRS
431935	Sao Pedro da Serra	2º CRS
431950	Sao Sebastiao do Cai	2º CRS
432035	Sentinela do Sul	2º CRS
432055	Sertao Santana	2º CRS
432110	Tapes	2º CRS
432120	Taquara	2º CRS
432170	Tres Coroas	2º CRS
432200	Triunfo	2º CRS
432225	Tupandi	2º CRS
430063	Amaral Ferrador	3º CRS
430107	Arroio do Padre	3º CRS
430130	Arroio Grande	3º CRS
430450	Cangucu	3º CRS
430466	Capao do Leao	3º CRS
430512	Cerrito	3º CRS
430543	Chui	3º CRS
430605	Cristal	3º CRS
430710	Herval	3º CRS
431100	Jaguarao	3º CRS
431245	Morro Redondo	3º CRS
431417	Pedras Altas	3º CRS
431420	Pedro Osorio	3º CRS
431440	Pelotas	3º CRS
431450	Pinheiro Machado	3º CRS
431460	Piratini	3º CRS
431560	Rio Grande	3º CRS
431730	Santa Vitoria do Palmar	3º CRS
431700	Santana da Boa Vista	3º CRS
431850	Sao Jose do Norte	3º CRS
431880	Sao Lourenco do Sul	3º CRS
432232	Turucu	3º CRS
430010	Agudo	4º CRS
430290	Cacequi	4º CRS
430465	Capao do Cipo	4º CRS
430637	Dilermando de Aguiar	4º CRS
430670	Dona Francisca	4º CRS
430800	Faxinal do Soturno	4º CRS
430840	Formigueiro	4º CRS
431053	Itaara	4º CRS
431075	Ivora	4º CRS
431110	Jaguari	4º CRS
431113	Jari	4º CRS
431120	Julio de Castilhos	4º CRS
431210	Mata	4º CRS
431303	Nova Esperanca do Sul	4º CRS
431310	Nova Palma	4º CRS
431402	Paraiso do Sul	4º CRS
431447	Pinhal Grande	4º CRS
431532	Quevedos	4º CRS
431550	Restinga Seca	4º CRS
431690	Santa Maria	4º CRS
431740	Santiago	4º CRS
431810	Sao Francisco de Assis	4º CRS
431843	Sao Joao do Polesine	4º CRS
431912	Sao Martinho da Serra	4º CRS
431940	Sao Pedro do Sul	4º CRS
431960	Sao Sepe	4º CRS
431980	Sao Vicente do Sul	4º CRS
432065	Silveira Martins	4º CRS

432149	Toropi	4º CRS
432237	Unistalda	4º CRS
432345	Vila Nova do Sul	4º CRS
430057	Alto Feliz	5º CRS
430080	Antonio Prado	5º CRS
430210	Bento Goncalves	5º CRS
430225	Boa Vista do Sul	5º CRS
430230	Bom Jesus	5º CRS
430235	Bom Principio	5º CRS
430367	Campestre da Serra	5º CRS
430440	Canela	5º CRS
430480	Carlos Barbosa	5º CRS
430510	Caxias do Sul	5º CRS
430593	Coronel Pilar	5º CRS
430595	Cotipora	5º CRS
430740	Esmeralda	5º CRS
430786	Fagundes Varela	5º CRS
430790	Farroupilha	5º CRS
430810	Feliz	5º CRS
430820	Flores da Cunha	5º CRS
430860	Garibaldi	5º CRS
430910	Gramado	5º CRS
430925	Guabiju	5º CRS
430940	Guapore	5º CRS
431043	Ipe	5º CRS
431112	Jaquirana	5º CRS
431164	Linha Nova	5º CRS
431237	Monte Alegre dos Campos	5º CRS
431238	Monte Belo do Sul	5º CRS
431261	Muitos Capoes	5º CRS
431280	Nova Araca	5º CRS
431290	Nova Bassano	5º CRS
431308	Nova Padua	5º CRS
431320	Nova Petropolis	5º CRS
431330	Nova Prata	5º CRS
431335	Nova Roma do Sul	5º CRS
431400	Parai	5º CRS
431442	Picada Cafe	5º CRS
431446	Pinhal da Serra	5º CRS
431517	Protasio Alves	5º CRS
431725	Santa Tereza	5º CRS
431844	Sao Jorge	5º CRS
431862	Sao Jose dos Ausentes	5º CRS
431900	Sao Marcos	5º CRS
431975	Sao Vendelino	5º CRS
432235	Uniao da Serra	5º CRS
432250	Vacaria	5º CRS
432254	Vale Real	5º CRS
432280	Veranopolis	5º CRS
432330	Vila Flores	5º CRS
432360	Vista Alegre do Prata	5º CRS
430005	Agua Santa	6º CRS
430047	Almirante Tamandare do Sul	6º CRS
430055	Alto Alegre	6º CRS
430066	Andre da Rocha	6º CRS
430180	Barracao	6º CRS
430320	Cacique Doble	6º CRS
430355	Camargo	6º CRS
430410	Campos Borges	6º CRS
430462	Capao Bonito do Sul	6º CRS

430470	Carazinho	6º CRS
430490	Casca	6º CRS
430495	Caseiros	6º CRS
430550	Ciriaco	6º CRS
430585	Coqueiros do Sul	6º CRS
430597	Coxilha	6º CRS
430630	David Canabarro	6º CRS
430705	Ernestina	6º CRS
430750	Espumoso	6º CRS
430885	Gentil	6º CRS
430980	Ibiaca	6º CRS
430990	Ibiraiaras	6º CRS
430995	Ibirapuita	6º CRS
431127	Lagoa dos Tres Cantos	6º CRS
431130	Lagoa Vermelha	6º CRS
431125	Lagoao	6º CRS
431170	Machadinho	6º CRS
431180	Marau	6º CRS
431213	Mato Castelhano	6º CRS
431220	Maximiliano de Almeida	6º CRS
431235	Montauri	6º CRS
431242	Mormaco	6º CRS
431262	Muliterno	6º CRS
431265	Nao-Me-Toque	6º CRS
431267	Nicolau Vergueiro	6º CRS
431275	Nova Alvorada	6º CRS
431360	Paim Filho	6º CRS
431410	Passo Fundo	6º CRS
431477	Pontao	6º CRS
431660	Sananduva	6º CRS
431673	Santa Cecilia do Sul	6º CRS
431755	Santo Antonio do Palma	6º CRS
431775	Santo Antonio do Planalto	6º CRS
431795	Santo Expedito do Sul	6º CRS
431805	Sao Domingos do Sul	6º CRS
431842	Sao Joao da Urtiga	6º CRS
431860	Sao Jose do Ouro	6º CRS
432040	Serafina Correa	6º CRS
432050	Sertao	6º CRS
432080	Soledade	6º CRS
432090	Tapejara	6º CRS
432100	Tapera	6º CRS
432146	Tio Hugo	6º CRS
432215	Tunas	6º CRS
432218	Tupanci do Sul	6º CRS
432255	Vanini	6º CRS
432320	Victor Graeff	6º CRS
432335	Vila Langaro	6º CRS
432340	Vila Maria	6º CRS
430003	Acegua	7º CRS
430160	Bage	7º CRS
430435	Candiota	7º CRS
430660	Dom Pedrito	7º CRS
430965	Hulha Negra	7º CRS
431150	Lavras do Sul	7º CRS
430120	Arroio do Tigre	8º CRS
430280	Cacapava do Sul	8º CRS
430300	Cachoeira do Sul	8º CRS
430513	Cerro Branco	8º CRS
430690	Encruzilhada do Sul	8º CRS

430781	Estrela Velha	8º CRS
430975	Ibarama	8º CRS
431123	Lagoa Bonita do Sul	8º CRS
431339	Novo Cabrais	8º CRS
431406	Passa Sete	8º CRS
432026	Segredo	8º CRS
432070	Sobradinho	8º CRS
430222	Boa Vista do Cadeado	9º CRS
430223	Boa Vista do Incra	9º CRS
430560	Colorado	9º CRS
430610	Cruz Alta	9º CRS
430845	Fortaleza dos Valos	9º CRS
431000	Ibiruba	9º CRS
431087	Jacuizinho	9º CRS
431535	Quinze de Novembro	9º CRS
431643	Saldanha Marinho	9º CRS
431645	Salto do Jacui	9º CRS
431670	Santa Barbara do Sul	9º CRS
432030	Selbach	9º CRS
432220	Tupancireta	9º CRS
430040	Alegrete	10º CRS
430187	Barra do Quarai	10º CRS
431060	Itaqui	10º CRS
431171	Macambara	10º CRS
431175	Manoel Viana	10º CRS
431530	Quarai	10º CRS
431640	Rosario do Sul	10º CRS
431697	Santa Margarida do Sul	10º CRS
431710	Santana do Livramento	10º CRS
431830	Sao Gabriel	10º CRS
432240	Uruguaiana	10º CRS
430250	Bossoroca	12º CRS
430330	Caibate	12º CRS
430520	Cerro Largo	12º CRS
430635	Dezesseis de Novembro	12º CRS
430693	Entre-Ijuis	12º CRS
430783	Eugenio de Castro	12º CRS
430865	Garruchos	12º CRS
430950	Guarani das Missoes	12º CRS
431055	Itacurubi	12º CRS
431217	Mato Queimado	12º CRS
431455	Pirapo	12º CRS
431510	Porto Xavier	12º CRS
431595	Rolador	12º CRS
431630	Roque Gonzales	12º CRS
431647	Salvador das Missoes	12º CRS
431750	Santo Angelo	12º CRS
431770	Santo Antonio das Missoes	12º CRS
431800	Sao Borja	12º CRS
431890	Sao Luiz Gonzaga	12º CRS
431915	Sao Miguel das Missoes	12º CRS
431920	Sao Nicolau	12º CRS
431937	Sao Pedro do Butia	12º CRS
432057	Sete de Setembro	12º CRS
432234	Ubiretama	12º CRS
432375	Vitoria das Missoes	12º CRS
430420	Candelaria	13º CRS
430915	Gramado Xavier	13º CRS
430957	Herveiras	13º CRS
431215	Mato Leitao	13º CRS

431395	Pantano Grande	13º CRS
431407	Passo do Sobrado	13º CRS
431570	Rio Pardo	13º CRS
431680	Santa Cruz do Sul	13º CRS
432067	Sinimbu	13º CRS
432253	Vale do Sol	13º CRS
432252	Vale Verde	13º CRS
432260	Venancio Aires	13º CRS
432270	Vera Cruz	13º CRS
430030	Alecrim	14º CRS
430045	Alegria	14º CRS
430220	Boa Vista do Burica	14º CRS
430370	Campina das Missoes	14º CRS
430430	Candido Godoi	14º CRS
430673	Doutor Mauricio Cardoso	14º CRS
430900	Girua	14º CRS
430960	Horizontina	14º CRS
431040	Independencia	14º CRS
431301	Nova Candelaria	14º CRS
431342	Novo Machado	14º CRS
431500	Porto Lucena	14º CRS
431505	Porto Maua	14º CRS
431507	Porto Vera Cruz	14º CRS
431720	Santa Rosa	14º CRS
431790	Santo Cristo	14º CRS
431849	Sao Jose do Inhacora	14º CRS
431930	Sao Paulo das Missoes	14º CRS
432032	Senador Salgado Filho	14º CRS
432180	Tres de Maio	14º CRS
432210	Tucunduva	14º CRS
432230	Tuparendi	14º CRS
430195	Barra Funda	15º CRS
430215	Boa Vista das Missoes	15º CRS
430260	Braga	15º CRS
430515	Cerro Grande	15º CRS
430530	Chapada	15º CRS
430580	Constantina	15º CRS
430590	Coronel Bicaco	15º CRS
430642	Dois Irmaos das Missoes	15º CRS
430692	Engenho Velho	15º CRS
430912	Gramado dos Loureiros	15º CRS
431085	Jaboticaba	15º CRS
431142	Lajeado do Bugre	15º CRS
431230	Miraguai	15º CRS
431295	Nova Boa Vista	15º CRS
431349	Novo Barreiro	15º CRS
431346	Novo Xingu	15º CRS
431370	Palmeira das Missoes	15º CRS
431540	Redentora	15º CRS
431610	Ronda Alta	15º CRS
431620	Rondinha	15º CRS
431642	Sagrada Familia	15º CRS
431845	Sao Jose das Missoes	15º CRS
431936	Sao Pedro das Missoes	15º CRS
432010	Sarandi	15º CRS
432185	Tres Palmeiras	15º CRS
432195	Trindade do Sul	15º CRS
430070	Anta Gorda	16º CRS
430100	Arroio do Meio	16º CRS
430140	Arvorezinha	16º CRS

430200	Barros Cassal	16º CRS
430240	Bom Retiro do Sul	16º CRS
430245	Boqueirao do Leao	16º CRS
430461	Canudos do Vale	16º CRS
430469	Capitao	16º CRS
430558	Colinas	16º CRS
430583	Coqueiro Baixo	16º CRS
430620	Cruzeiro do Sul	16º CRS
430645	Dois Lajeados	16º CRS
430675	Doutor Ricardo	16º CRS
430680	Encantado	16º CRS
430780	Estrela	16º CRS
430807	Fazenda Vilanova	16º CRS
430830	Fontoura Xavier	16º CRS
430843	Forquetinha	16º CRS
431030	Ilopolis	16º CRS
431036	Imigrante	16º CRS
431057	Itapuca	16º CRS
431140	Lajeado	16º CRS
431205	Marques de Souza	16º CRS
431260	Mucum	16º CRS
431300	Nova Brescia	16º CRS
431415	Paverama	16º CRS
431475	Poco das Antas	16º CRS
431515	Progresso	16º CRS
431520	Putinga	16º CRS
431545	Relvado	16º CRS
431580	Roca Sales	16º CRS
431675	Santa Clara do Sul	16º CRS
431846	Sao Jose do Herval	16º CRS
431971	Sao Valentim do Sul	16º CRS
432045	Serio	16º CRS
432085	Tabai	16º CRS
432130	Taquari	16º CRS
432145	Teutonia	16º CRS
432162	Travesseiro	16º CRS
432285	Vespasiano Correa	16º CRS
432377	Westfalia	16º CRS
431513	Pouso Novo	16º CRS
430020	Ajuricaba	17º CRS
430150	Augusto Pestana	17º CRS
430258	Bozano	17º CRS
430400	Campo Novo	17º CRS
430500	Catuipe	17º CRS
430540	Chiapeta	17º CRS
430570	Condor	17º CRS
430587	Coronel Barros	17º CRS
430600	Crissiumal	17º CRS
431020	Ijui	17º CRS
431041	Inhacora	17º CRS
431115	Joia	17º CRS
430970	Mucum	17º CRS
431333	Nova Ramada	17º CRS
431390	Panambi	17º CRS
431430	Pejucara	17º CRS
431780	Santo Augusto	17º CRS
431910	Sao Martinho	17º CRS
431973	Sao Valerio do Sul	17º CRS
432023	Sede Nova	17º CRS
430105	Arroio do Sal	18º CRS

430163	Balneario Pinhal	18º CRS
430463	Capao da Canoa	18º CRS
430467	Capivari do Sul	18º CRS
430471	Caraa	18º CRS
430545	Cidreira	18º CRS
430655	Dom Pedro de Alcantara	18º CRS
431033	Imbe	18º CRS
431065	Itati	18º CRS
431173	Mampituba	18º CRS
431177	Maquine	18º CRS
431244	Morrinhos do Sul	18º CRS
431250	Mostardas	18º CRS
431350	Osorio	18º CRS
431365	Palmares do Sul	18º CRS
431760	Santo Antonio da Patrulha	18º CRS
432135	Tavares	18º CRS
432143	Terra de Areia	18º CRS
432150	Torres	18º CRS
432160	Tramandai	18º CRS
432166	Tres Cachoeiras	18º CRS
432183	Tres Forquilhas	18º CRS
432380	Xangri-La	18º CRS
430050	Alpestre	19º CRS
430064	Ametista do Sul	19º CRS
430185	Barra do Guarita	19º CRS
430237	Bom Progresso	19º CRS
430340	Caicara	19º CRS
430607	Cristal do Sul	19º CRS
430632	Derrubadas	19º CRS
430730	Erval Seco	19º CRS
430745	Esperanca do Sul	19º CRS
430850	Frederico Westphalen	19º CRS
431050	Irai	19º CRS
431160	Liberato Salzano	19º CRS
431270	Nonoai	19º CRS
431344	Novo Tiradentes	19º CRS
431380	Palmitinho	19º CRS
431445	Pinhal	19º CRS
431449	Pinheirinho do Vale	19º CRS
431470	Planalto	19º CRS
431555	Rio dos Indios	19º CRS
431590	Rodeio Bonito	19º CRS
432020	Seberi	19º CRS
432132	Taquarucu do Sul	19º CRS
432140	Tenente Portela	19º CRS
432147	Tiradentes do Sul	19º CRS
432190	Tres Passos	19º CRS
432310	Vicente Dutra	19º CRS
432350	Vista Alegre	19º CRS
432370	Vista Gaucha	19º CRS

Fonte: Do autor (2015), com base nos dados do TASY.