UNIVERSIDADE FEDERAL DO TOCANTINS

CÂMPUS UNIVERSITÁRIO DE PALMAS

CURSO DE CIÊNCIA DA COMPUTAÇÃO

DESCOBERTA DE CONHECIMENTO NA BASE DE DADOS DO SINAN-TO PARA OS CASOS DE HANSENÍASE NO ESTADO DO TOCANTINS

DENILSON SANTOS SOBRINHO JÚNIOR

PALMAS (TO)

2021

DENILSON SANTOS SOBRINHO JÚNIOR

DESCOBERTA DE CONHECIMENTO NA BASE DE DADOS DO SINAN-TO PARA OS CASOS DE HANSENÍASE NO ESTADO DO TOCANTINS

Trabalho de Conclusão de Curso II apresentado à Universidade Federal do Tocantins para obtenção do título de Bacharel em Ciência da Computação, sob a orientação do(a) Prof.(a) Dr. Ary Henrique Morais de Oliveira .

Orientador: Dr. Ary Henrique Morais de Oliveira

PALMAS (TO)

2021

DENILSON SANTOS SOBRINHO JÚNIOR

DESCOBERTA DE CONHECIMENTO NA BASE DE DADOS DO SINAN-TO PARA OS CASOS DE HANSENÍASE NO ESTADO DO TOCANTINS

Trabalho de Conclusão de Curso II apresentado à UFT – Universidade Federal do Tocantins – Câmpus Universitário de Palmas, Curso de Ciência da Computação foi avaliado para a obtenção do título de Bacharel e aprovada em sua forma final pelo Orientador e pela Banca Examinadora.

Data de aprovação: 22 / 5 / 2021

Banca Examinadora:

Prof. Dr. Ary Henrique Morais de Oliveira

Prof. Dr. Warley Gramacho da Silva

Profa. Dra. Anna Paula de Sousa Parente Rodrigues

Bel. Valéria Perim da Cunha

UNIVERSIDADE FEDERAL DO TOCANTINS
CÂMPUS DE PALMAS
CURSO DE CIÊNCIA DA COMPUTAÇÃO
109 Norte Av. NS-15, ALCNO- 14, bloco II, Sal. 20 | 77001-090 | Palmas/TO
(63) 3232-8027 | www.uft.edu.br | comppalmas@uft.edu.br

ATA DE DEFESA DA DISCIPLINA DE PROJETO DE GRADUAÇÃO II

Ao **Vigésimo segundo dia** do mês de Maio de 2021 realizou-se a defesa de Projeto de Graduação, da disciplina de Projeto de Graduação II do discente **Denilson Santos Sobrinho Júnior** do curso de Ciência da Computação do Campus Universitário de Palmas da Universidade Federal do Tocantins (UFT), intitulado "**Descoberta de conhecimento na base de dados do Sinan para os casos de hanseníase no estado do Tocantins**", realizado sob a responsabilidade do Orientador(a) Prof(a). Dr(a) **Ary Henrique Morais de Oliveira**.

Atribuíram a Nota Final 8,5 (Oito vírgula cinco) pelo trabalho, tendo sido considerado **Aprovado**. Nada mais tendo a constar, assinam esta Ata os seguintes componentes da banca examinadora:

Prof(a). Dr(a). Ary Henrique Morais de Oliveira

Prof(a). Dr(a). Warley Gramacho da Silva

Prof(a). Dr(a). Anna Paula Parente Rodrigues

Bel. Valéria Perim da Cunha

DENILSON SANTOS SOBRINHO JÚNIOR

DESCOBERTA DE CONHECIMENTO NA BASE DE DADOS DO SINAN-TO PARA OS CASOS DE HANSENÍASE NO ESTADO DO TOCANTINS

Trabalho de Conclusão de Curso II apresentado à Universidade Federal do Tocantins para obtenção do título de Bacharel em Ciência da Computação, sob a orientação do(a) Prof.(a) Dr. Ary Henrique Morais de Oliveira .

Orientador: Dr. Ary Henrique Morais de Oliveira

PALMAS (TO)

2021

DENILSON SANTOS SOBRINHO JÚNIOR

DESCOBERTA DE CONHECIMENTO NA BASE DE DADOS DO SINAN-TO PARA OS CASOS DE HANSENÍASE NO ESTADO DO TOCANTINS

Trabalho de Conclusão de Curso II apresentado à UFT – Universidade Federal do Tocantins – Câmpus Universitário de Palmas, Curso de Ciência da Computação foi avaliado para a obtenção do título de Bacharel e aprovada em sua forma final pelo Orientador e pela Banca Examinadora.

Data de aprovação: 22 / 5 / 2021

Banca Examinadora:

Prof. Dr. Ary Henrique Morais de Oliveira

Prof. Dr. Warley Gramacho da Silva

Profa. Dra. Anna Paula de Sousa Parente Rodrigues

Bel. Valéria Perim da Cunha

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da Universidade Federal do Tocantins

J95d Júnior, Denilson Santos Sobrinho.

Descoberta de conhecimento na base de dados do Sinan para os casos de hanseníase no estado do tocantins. / Denilson Santos Sobrinho Júnior. – Palmas, TO, 2021.

68 f.

Monografia Graduação - Universidade Federal do Tocantins – Câmpus Universitário de Palmas - Curso de Ciências da Computação, 2021.

Orientador: Ary Henrique Morais de Oliveira

1. Hanseníase. 2. Descoberta de conhecimento em bases de dados. 3. Mineração de dados. 4. Associação. I. Título

CDD 004

AGRADECIMENTOS

Gostaria de agradecer a todos.

RESUMO

A hanseníase constitui-se como infecciosa e é considerada um problema de saúde pública por ser uma doença negligenciada, e por essa razão há um problema grande na prevenção e no combate da mesma. No estado do Tocantins a doença é considerada hiperendêmica de acordo com os parâmetros por 100.000 habitantes, sendo um dos estados que possuem mais casos novos no Brasil. Diante desse problema, busca-se através dos dados provenientes do Sistema de Notificação de Informação (Sinan) que realiza a coleta de notificação de doenças e agravos, a extração do conhecimento acerca dos casos de hanseníase no estado do Tocantins aplicando-se o processo de descoberta de conhecimento conhecido como *Knowledge Discover in Database (KDD)* e através da mineração de dados utilizando regras de associação para a caracterização das pessoas acometidas pela doença no estado. A metodologia utilizada foi do tipo descritivo e exploratório, de forma a buscar descrever os fatos e tentar explorar os dados, foi utilizado a análise exploratória de dados para a exploração dos atributos relevantes das características do paciente, da doença, do espaço e do tempo. Já na mineração de dados utilizando regras de associação foi responsável por fazer a associação desses atributos para saber o perfil dos pacientes que eram frequentes na base de dados de forma a caracterizá-los. Os resultados da pesquisa que foram encontrados são que as pessoas mais afetadas pela doença são pessoas que residentes em zonas urbanas, e as formas multibacilares são mais comuns no total, sendo a população masculina a mais afetada por elas. A escolaridade da população em sua maioria era de baixa escolaridade, sendo também relacionado a baixa renda das pessoas afetadas. Todos os anos do estudo a quantidade de casos foi hiperendêmica de acordo com os parâmetros nacionais e internacionais.

Palavra-chave: Descoberta de conhecimento em base de dados. Hanseníase. Saúde pública. Mineração de dados. Associação.

ABSTRACT

Leprosy is an infectious disease and is considered a public health problem because it is a neglected disease, and for that reason there is a big problem in preventing and combating it. In the state of Tocantins, the disease is considered hyperendemic according to the parameters per 100,000 inhabitants, being one of the states with the most new cases in Brazil. In view of this problem, it is sought through data from the Information Notification System (Sinan) that performs the collection of notification of diseases and conditions, the extraction of knowledge about leprosy cases in the state of Tocantins by applying the process of knowledge discovery known as *Knowledge Discover in Database (KDD)* and through data mining using association rules to characterize people affected by the disease in the state. The methodology used was of the descriptive and exploratory type, in order to seek to describe the facts and try to explore the data, exploratory data analysis was used to explore the relevant attributes of the patient's characteristics, the disease, the space and the time. In data mining using association rules, he was responsible for associating these attributes to know the profile of patients who were frequent in the database in order to characterize them. The results of the research that were found are that the people most affected by the disease are people who live in urban areas, and the multibacillary forms are more common in total, with the male population being the most affected by them. The education of the population was mostly low, and it was also related to the low income of the people affected. Every year of the study, the number of cases was hyperendemic according to national and international parameters.

Keywords: Knowledge discovery in databases. Leprosy. Public health. Data mining. Association.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O tratamento de informações em grande volume de dados é um tema de grande destaque devido a necessidade da concepção de métodos que otimizem o armazenamento e a recuperação de informações a partir de grandes volumes de dados. Alguns programas e equipamentos científicos produzem conjuntos de dados na escala de petabytes tornando a análise e avaliação dos resultados uma tarefa desafiadora (USAMA; GREGORY; PADHRAIC, 1997).

Para (USAMA; GREGORY; PADHRAIC, 1997):

> " Apesar do volume, a análise e extração de conhecimento destes dados devem ser viáveis e eficientes, caso contrário a produção dados não teria utilidade prática. A capacidade das ferramentas e métodos de extrair conhecimento útil dessas massas de dados é o que indica a relevância e importância dessas informações. "

A área da computação responsável pela extração de conhecimento útil em grandes bases de dados é conhecida por Descoberta de Conhecimento em Bases de Dados, do inglês *Knowledge Discovery Database (KDD)*. O KDD é um processo formado por várias etapas para a realização do processo de mineração de dados de forma a buscar novos relacionamentos, padrões e informações implícitas.

O KDD é considerado como uma atividade multidisciplinar, pois tem evoluído constantemente a partir da sua intersecção com outras áreas do conhecimento, tais como aprendizado de máquina, reconhecimento de padrões, banco de dados, estatística, inteligência artificial, sistemas especialistas, visualização de dados e computação de alto desempenho (USAMA; GREGORY; PADHRAIC, 1997).

Por ser multidisciplinar, pode ser utilizado em diversas áreas do conhecimento como por exemplo na saúde. Para (KOBUS, 2006):

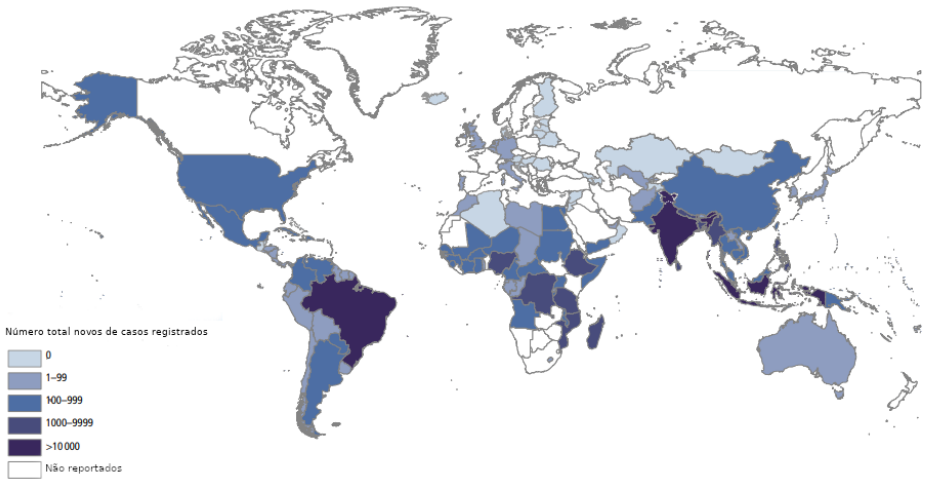
> "O uso de mineração de dados e análise epidemiológica, nos dados da saúde, proporcionaria o mapeamento da situação da saúde, assim como a produção de estatísticas de acordo com a incidência e prevalência de doenças e riscos a saúde, tanto atual, quanto futura dos usuários, baseado nessas estatísticas, os gestores seriam capacitados a tomar decisões relacionadas aos serviços a serem prestados a população, de acordo com suas reais necessidades, e ainda possibilitaria promover ações em promoção da saúde."

A hanseníse é uma doença crônica infecciosa, transmissível e de caráter crônico que ainda persiste como problema de saúde pública no Brasil. Seu agente etiolgico é o

Mycobacterium leprae, bacilo que afeta principalmente os nervos periféricos, olhos e pele. A doença atinge pessoas de qualquer sexo ou faixa etária, apresentando-se como uma evolução lenta e progressiva e quando não tratada, é passível de causar deformidades e incapacidades físicas, muitas vezes irreversíveis (BRASIL, 2016, 2017).

Em 2016, foram reportados a Organização Mundial de Saúde (OMS) 214.783 casos novos da doença em todo o mundo como mostra a figura 1. No continente americano foi reportado 27.356 novos casos sendo 25.218 (92%) dos novos casos pertencentes ao Brasil (WHO, 2017).

Figura 1 – Distribuição geográfica de novos casos de hanseníase no ano de 2016



Fonte: Organização Mundial da Saúde (OMS)

Diante desse cenário o Brasil é classificado com um país de alta carga da doença, ocupando o segundo lugar na relação de países com maior número de casos no mundo, atrás apenas da Índia (WHO, 2017).

Diante disso, há um sistema que realiza a coleta de dados de notificações e agravos em todo o território nacional, o nome desse sistema é o Sinan. O Sistema de Informação de Agravos de Notificação (Sinan) tem como objetivo coletar, transmitir e disseminar dados gerados rotineiramente pelo Sistema de Vigilância Epidemiológica das três esferas de governo, por intermédio de uma rede informatizada, para apoiar o processo de investigação e dar subsídios à análise das informações de vigilância epidemiológica das doenças de notificação compulsória (BRASIL, 2006).

Sua utilização efetiva permite a realização do diagnóstico dinâmico da ocorrência de um evento na população, podendo fornecer subsídios para explicações causais dos agravos de notificação compulsória, além de vir a indicar riscos aos quais as pessoas estão sujeitas, contribuindo assim, para a identificação da realidade epidemiológica de determinada área geográfica (BRASIL, 2006).

Portanto, sabendo-se que existe esse grande sistema que realiza a coleta de dados,

inclusive da hanseníase, torna-o um grande repositório para coleta, processamento e uso do processo de descoberta de conhecimento para a extração do conhecimento a partir dos dados existentes.

1.1 Justificativa

Apesar de grandes esforços no combate da doença, é notável que ainda há um grande número de casos da doença no estado fazendo-o ser considerado hiperendêmico e por consequência isso afeta no controle da doença e sua evolução. O problema de altos casos também pode estar relacionado a falta de informação da população em relação à existência da doença, de forma que só busquem o acesso às redes de saúde em casos mais avançados, assim contribuindo para a evolução da doença.

1.2 Descrição do problema

A hanseníase é uma doença causada pela infecção um bacilo e os suas consequências podem afetar de nervos periféricos causando danos irreversíveis, apesar de ser uma doença lenta e em muitos casos silenciosa pode provocar fenótipos clínicos resultando em preconceitos e baixa estima das pessoas atingidas por essa morbidade.

No estado do tocantins a doença é considerada hiperendêmica, possuindo um alto valor no número de casos novos. Apesar de a doença ser bastante antiga, a taxa de casos ainda é alta e há um enorme problema em realizar o combate da doença.

1.3 Hipóteses

Os grupos mais vulneráveis à doença não estão absorvendo adequadamente as informações necessárias para a prevenção, controle e redução da hanseníase, seja pela forma como as ações de saúde são executadas, ou pela maneira em que a comunicação é realizada, portanto, caracterizar as variáveis temporais associadas com o perfil do público e a região de domicílio e tratamento podem tornar o processo de redução e combate da doença mais efetivo, inclusive com inovações na forma de combate a doença.

1.4 Objetivos

1.4.1 Objetivos Gerais

Desenvolver mecanismos de visualização dos resultados do processo de extração de conhecimento da base de dados do Sinan-TO, aplicados para os casos de Hanseníase no Estado do Tocantins, a partir das tarefas de associação para complementar as informações adotadas no desenvolvimento de estratégias para o combate da doença.

1.4.2 Objetivos Específicos

Os objetivos específicos deste trabalho são os seguintes:

- Realizar um levantamento bibliográfico sobre os temas abordados no projeto, com destaque para saúde pública, hanseníase, mineração de dados para construção da fundamentação teórica e do atual estado da arte;
- Adotar o processo de descoberta de conhecimento em bases de dados identificando quais os métodos, técnicas e ferramentas de pré-processamento, mineração de dados e pós-processamento serão adotados para o desenvolvimento do trabalho.
- Realizar o pré-processamentos dos dados para a etapa de mineração a partir da base de dados do Sinan-TO, realizando a identificação dos dados, eliminação de ruídos e a devida transformação quando necessário.
- Executar o processamento a partir das tarefas de associação sob a base de dados pré-processada buscando padrões, relações e regularidades nos dados.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os conceitos relativos aos assuntos relacionados a este trabalho, citando conceito a respeito da Saúde Pública e no seu contexto principal orientado ao Brasil e mais em específico ao Tocantins. Explicar conceito acerca do Processo de Extração de Conhecimento em Bancos de Dados (KDD).

2.1 Saúde Pública

A saúde pública de acordo com (CEA, 1920) é "A ciência e a arte de prevenir doenças, prolongar a vida e promover a saúde através dos esforços organizados e escolhas informadas da sociedade, organizações, comunidades públicas e privadas e indivíduos" .

No geral, a saúde pública preocupa-se em proteger a saúde de populações inteiras (CONTROL; CDC, 2014). Essas populações podem ser tão pequenas quanto uma vizinhança local ou tão grandes quanto um país ou região inteira do mundo (CONTROL; CDC, 2014).

A saúde pública trabalha para rastrear surtos de doenças, prevenir lesões e esclarecer por que algumas pessoas têm maior probabilidade de sofrer de problemas de saúde do que outros (ALPHA, 2020). O desafio de preservar, manter e promover ativamente a saúde pública requer métodos especiais de coleta de informações (epidemiologia) e arranjos corporativos para agir sobre descobertas significativas e colocá-las em prática (RHODES; BRYANT, 2011).

De acordo com (RHODES; BRYANT, 2011) as estatísticas coletadas por epidemiologistas tentam descrever e explicar a ocorrência da doença em uma população, correlacionando fatores como dieta, ambiente, exposição à radiação ou tabagismo com a incidência e prevalência da doença.

Durante os últimos 150 anos, dois fatores moldaram o moderno sistema de saúde pública: primeiro, o crescimento do conhecimento científico sobre as fontes e meios de controle das doenças; segundo, o crescimento da aceitação pública do controle de doenças como uma possibilidade e uma responsabilidade pública (USA, 1988).

A ligação entre a ciência, o desenvolvimento de intervenções e a organização das autoridades públicas para empregar as intervenções aumentou a compreensão do público e o compromisso social com a melhoria da saúde. O crescimento de um sistema público de proteção à saúde dependeu tanto da descoberta científica quanto da ação social (USA, 1988).

2.1.1 Saúde Pública no Brasil

A história das políticas de saúde no Brasil está inserida no contexto da história do estado brasileiro, com início no período colonial, e do interesse em manter saudável a mão-de-obra, com grandes mudanças após a industrialização (FERREIRA, 2009).

A história dos cuidados com saúde do brasileiro passa, necessariamente, pela filantropia. Mais ainda pelo cunho filantrópico religioso, a caridade. As pessoas eram atendidas pelas instituições e médicos filantropos (CARVALHO, 2013). No Brasil Colônia, o acesso à saúde era precário e inexistente, basicamente sua organização sanitária espelhava a da metrópole (Portugal), sendo majoritariamente de responsabilidade militar (PAIM, 2015).

Uma das formas de cuidados com a saúde, principalmente os dos mais pobres foi através da primeira unidade de apoio à saúde de cunho católico, a Santa Casa, em Santos, São Paulo, no ano de 1543 (PAIM, 2015).

Já no final do século XVII, o enfoque é um pouco diferente, o poder colonial assume como um dos objetivos da própria administração, a questão da recuperação do estado de saúde de seus habitantes (GALVãO, 2010). O projeto de criação de um novo hospital que não era meramente estabelecido em bases filantrópicas, como ocorria com as antigas instituições existentes (GALVãO, 2010).

Dessa maneira ocorreu a regulamentação do ensino e da prática médica e a criação de hospitais públicos para atender doenças que exigiam maior controle do Estado, tais como as doenças mentais, a hanseníase e a tuberculose (TWF, 2005; JF, 1989).

A saúde emerge como questão social no Brasil apenas no início do século XX, no bojo da economia capitalista exportadora cafeeira, refletindo o avanço da divisão do trabalho, ou seja, a emergência do trabalho assalariado (BRAVO, 2009).

Em 1923 a luta trabalhista ganha seu primeiro triunfo a favor dos trabalhadores com a Lei Eloy Chaves, a qual consolida a base do sistema previdenciário brasileiro, com a criação da Caixa de Aposentadorias e Pensões aos empregados das empresas ferroviárias (ANGéLICA, 2020).

A partir da Revolução de 1930 com a extinção da República Velha da "política do café com leite" e ascensão do Governo Provisório do populista Getúlio Vargas, se inicia as primeiras formas de legislação social e de estímulo ao desenvolvimento industrial nacional (ANGéLICA, 2020).

Na década de 1950, o hospital tornou-se o principal ponto de referência para a busca de atendimentos de saúde (FERREIRA, 2009). A partir da década de 50 vários sistemas foram como o INPS em 1966 com a união do IAP e a centralização da previdência social, sendo que o mesmo ainda era restrito apenas ás pessoas que possuiam vínculo com o INPS, dessa forma deixando de fora do acesso à saúde (FERREIRA, 2009).

Após o fracasso do INPS, em 1974, foi criado o Instituto Nacional de Assistên-

cia Médica da Previdência Social (INAMPS), como uma autarquia federal vinculada ao Ministério da Previdência e Assistência Social e foi idealizado pelo Regime Militar pelo desmembramento do INPS (ANGéLICA, 2020).

As pressões por reformas na política de saúde possibilitaram algumas mudanças concretas ainda nos anos 1970, como a criação do INAMPS, mas ainda de forma incipiente e de acordo com os interesses do Estado. Entretanto, essas medidas favoreceram a construção de uma política de saúde mais universal, com prioridade para a extensão da oferta de serviços básicos (BAPTISTA, 2005; FALEIROS et al., 2006).

Nos anos 1980, o movimento da reforma sanitária na área da saúde indicava propostas de expansão de assistência médica na previdência social (RIBEIRO et al., 2009). Esse movimento criticava a mercantilização da medicina sob o comando da previdência social e buscava a universalização do direito à saúde, ampliando esse debate no Brasil (BAPTISTA, 2005).

Nesse cenário, ocorreu, em 1986, a VIII Conferência Nacional de Saúde, com a participação da comunidade e dos técnicos na discussão de uma política setorial. Nessa conferência foi aprovada a diretriz da universalização da saúde, sendo constituído o Sistema Unificado e Descentralizado da Saúde (SUDS), que se apresentou como estratégia-ponte para a construção do SUS (BAPTISTA, 2005; FALEIROS et al., 2006).

Em seguida o SUS foi finalmente aprovado na Assembleia Nacional Constituinte de 1987/88, sendo suas diretrizes estabelecidas e definidas na Constituição de 1988. Essa Constituição Federal deu nova forma à saúde no Brasil, estabelecendo-a como direito universal e concebendo-a de maneira integral, preventiva e curativa (FERREIRA, 2009).

Os principais marcos legais e normativos para a conformação do SUS, ressaltando a abrangência e a profundidade das mudanças propostas, foram a Constituição Federal de 1988 e as Leis Orgânicas da Saúde, de 1990. É no texto da Carta Magna, como você já deve ter levantado em sua pesquisa, que está explicitado:

> A saúde é direito de todos e dever do Estado, garantido mediante políticas sociais e econômicas que visem à redução do risco da doença e de outros agravos e ao acesso universal e igualitário às ações e aos serviços para a sua promoção, proteção e recuperação (Brasil, 1988, Art.196)."

Os princípios e diretrizes do SUS foram estabelecidos na Lei Orgânica da Saúde n° 8.080 de 1990:

- Universalização do direito à saúde;
- Descentralização com direção única para o sistema;
- Integralidade da atenção à saúde;
- Participação popular visando ao controle social.

A implantação do SUS tem início nos primeiros anos da década de 1990, após a promulgação da Lei Orgânica da Saúde (LOS) n. 8.080/90, de 19 de setembro de 1990, complementada pela Lei Orgânica da Saúde n° 142, de 28 de dezembro de 1990. Estas foram leis fundamentais que orientaram a operacionalização do sistema de saúde, visto que a primeira definiu os objetivos e atribuições do SUS, enquanto a segunda definiu as regras gerais para a participação popular e financiamento (CRUZ, 2011).

2.1.2 Sistema de Informação de Agravos de Notificação (Sinan)

O Sistema de Informação de Agravos de Notificação (Sinan) foi desenvolvido no início da década de 90, tendo como objetivo inicial a coleta e processamento dos dados sobre agravos de notificação em todo o território nacional (LSF, 1993).

É um sistema alimentado, principalmente, pela notificação e investigação de casos de doenças e agravos que constam da lista nacional de doenças de notificação compulsória de acordo com a portaria de Consolidação nº 4, de 28 de Setembro de 2017, anexo V - Capítulo I, mas é facultado a estados e municípios incluir outros problemas de saúde importantes em sua região (BRASIL, 2006).

2.1.2.1 Implantação

O Sinan foi implantado, de forma gradual, a partir de 1993. No entanto, esta implantação foi realizada de forma heterogênea nas unidades federadas e municípios, não havendo uma coordenação e acompanhamento por parte dos gestores de saúde, nas três esferas de governo (BRASIL, 2006).

De acordo com (LAGUARDIA et al., 2004):

> "O aplicativo Sinan foi concebido, originalmente, para armazenar, a partir de instrumentos e códigos de acesso padronizados em nível nacional, as informações das doenças de notificação compulsória, com suas respectivas fichas de notificação e investigação, sendo permitido às unidades federadas incluir notificações de outros agravos, adequando o sistema ao perfil epidemiológico de populações distintas. "

Somente a partir de 1998 é que o uso do Sinan foi regulamentado (BRASIL, 1998), tornando obrigatória a alimentação regular da base de dados nacional pelos municípios, estados e Distrito Federal, bem como designando a Fundação Nacional de Saúde (Funasa), por meio do Cenepi, como gestora nacional do Sistema (BRASIL, 2006).

O seu uso sistemático, de forma descentralizada, contribue para a democratização da informação, permitindo que todos os profissionais de saúde tenham acesso à informação e as tornem disponíveis para a comunidade (SINAN, 2016). Em razão disso é um instrumento relevante para auxiliar o planejamento da saúde, definir prioridades de intervenção, além de permitir que seja avaliado o impacto das intervenções (SINAN, 2016).

2.1.2.2 Objetivos

De acordo com (BRASIL, 2006), o Sistema de Informação de Agravos de Notificação (Sinan) possui o objetivo de:

> "Coletar, transmitir e disseminar dados gerados rotineiramente pelo Sistema de Vigilância Epidemiológica das três esferas de governo, por intermédio de uma rede informatizada, para apoiar o processo de investigação e dar subsídios à análise das informações de vigilância epidemiológica das doenças de notificação compulsória".

2.1.2.3 Funcionamento

O Sinan pode ser operacionalizado no nível administrativo mais periférico, ou seja, nas unidades de saúde, seguindo a orientação de descentralização do SUS (SINAN, 2016). A Ficha Individual de Notificação (FIN) é preenchida pelas unidades assistenciais para cada paciente quando da suspeita da ocorrência de problema de saúde de notificação compulsória ou de interesse nacional, estadual ou municipal.

Esse instrumento deve ser encaminhado aos serviços responsáveis pela informação e/ou vigilância epidemiológica das Secretarias Municipais, que devem repassar semanalmente os arquivos em meio magnético para as Secretarias Estaduais de Saúde (SES).

A comunicação das SES com a SVS deverá ocorrer quinzenalmente, de acordo com o cronograma definido pela SVS no início de cada ano (SINAN, 2016). Caso não ocorra nenhuma suspeita de doença, as unidades de saúde precisam preencher o formulário de notificação negativa, que tem os mesmos prazos de entrega (SINAN, 2016).

A tabela 1 mostra as doenças ou agravos que são notificados no Sinan.

Tabela 1 – Doenças ou agravos de notificação compulsória

Doenças ou agravos			
Acidente por animal peçonhento	Aids	Atendimento antirrábico	Botulismo
Cólera	Coqueluche	Dengue	Difteria
Doenças relacionadas ao trabalho	Epizootia	Esquistossomose	Febre Amarela
Febre de Chikungunya	Febre do Nilo	Febre Maculosa	Febre Tifóide
Gestante HIV	Hanseníase	Hantavirose	Hepatites virais
Influenza	Intoxicação Exógena	Leishmaniose	Leptospirose
Malária	Meningite	Poliomelite	Peste

Continua na próxima página

Tabela 1 – Continuação da tabela 7

Doenças ou agravos			
Raiva humana	Rotavírus	Rubéola	Sarampo
Síflis	Síndrome da Rubéola Congênita	Doenças Transmitidas por Alimentos	Tétano Acidental
Tétano Neonatal	Tracoma	Tuberculose	Violência Interpessoal /Autoprovocada
Zikavirus			

2.2 Hanseníase

2.2.1 Definição

A Hanseníase é uma doença infecto-contagiosa, crônica, causada pelo *Mycobacterium Leprae (M. Leprae)*, também conhecido como bacilo de Hansen (BH). Apresenta uma maior prevalência em áreas economicamente desfavorecidas, onde a população é submetida a fatores predispotentes como subalimentação, moléstias debilitantes e superpopulação (FARIA, 2003).

2.2.2 Histórico da Doença

A hanseníase é uma doença no qual era mundialmente conhecida como lepra, já tendo recebido várias denominações como morféia, mal de pele, doença lasarina, e cuja época exata do seu aparecimento não é bem conhecida. Os relatos mais antigos datam de 4266 a.C no Egito, 500 a 2000 a.C nos Livros Sagrados da Índia e 1100 a.C na China (TAVARES; MARINHO, 2007).

Nessas antigas civilizações, a doença era considerada como uma punição e os doentes eram obrigados a usar trajes especiais para o seu reconhecimento a distância eram regregados da sociedade. Essas atitudes bárbaras, associadas às deformidades e às multilações que a doença provocava, geraram o preconceito e a discriminação que ainda persiste até os dias de hoje. (TAVARES; MARINHO, 2007)

Devido ao preconceito e a discriminação, o termo lepra e seus derivados caíram em desuso no Brasil, por força da Lei 9.010 de 29/03/1995 sendo substituída por ”mal de Hansen”ou ”hanseníase”, nomenclatura proposta por Rothberg (TAVARES; MARINHO, 2007).

No Brasil, os primeiros casos da doença foram notificados no ano de 1600, na cidade do Rio de Janeiro (Yamanouchi et al, 1993), onde, anos mais tarde, seria criado o primeiro lazareto, local destinado a abrigar os doentes de Lázaro, lazarentos ou leprosos (Brasil, 1989).

2.2.3 Etiologia e Modo de Transmissão

O *Mycrobacterium Leprae*, classificado entre as *Mycobacteriaceae*, foi identificado por Hansen em 1874. A transmissão ocorre em humanos principalmente por meio da emissão de bacilos pelas vias aéreas superiores do doente com forma bacilífera e não tratado (CIMERMAN; CIMERMAN, 2003).

É importante frisar que a maior parte da população é resistente ao M. Leapre e a ocorrência da hanseníase está relacionada a fatores genéticos e à resposta imunitária celular ao M. Leprae (CIMERMAN; CIMERMAN, 2003).

Quanto à influência de fatores ambientais na transmissão da hanseníase, a endemia permanece nos países mais tropicais, coincidindo com o subdesenvolvimento e na população menos favorecida. Entretanto, não se conhece ao certo a importância do estado nutricional, aglomeração domiciliar ou outras doenças concomitantes no desencadeamento da doença (TAVARES; MARINHO, 2007)

2.2.4 Diagnóstico Clínico

Visando a facilitar o emprego da poliquimioterapia e com o objetivo de diminuir o coeficiente de prevalência da hanseníase nos países endêmicos, a Organização Mundial da Saúde (OMS) recomenda a classificação como base no quadro clínico, levando-se em consideração o número de lesões cutâneas e o acometimento neural.

Classifica-se a hanseníase em paucibacilar (PB), quando o paciente tem menos que cinco lesões cutâneas e um tronco nervoso comprometido. Neste estão incluídas as formas indeterminadas e tuberculóide. Quando a o paciente apresentar mais de cinco lesões cutâneas e/ou mais de um tronco nervoso comprometido é classificado em hanseníase multibacilar (MB) e corresponde às formas virchowiana e dimorfa.

2.2.5 Formas Clínicas

Várias classificações foram propostas para a hanseníase ao longo dos anos, à medida que novos conhecimentos sobre a doença foram adquiridos. A classificação de Madri, estabelecida no Congresso Internacional da Hanseníase, realizado em Madri em 1953, segue o sistema polar definido em 1936 por Rabello Jr (LASTóRIA; ABREU, 2014).

Segundo essa classificação há quatro tipos de formas clínicas, sendo duas paucibacilares e duas sendo multibacilares. A hanseníase indeterminada (paucibacilar) é considerada a forma inicial e transitória, pode ocorrer de três a cinco anos após o contágio. A manifestação cutânea mais comum é o surgimento de manchas, únicas ou mútliplas, hipocômicas, acrômicas ou discretamente erimetosas, planas, com bordas geralmente imprecisos, porém algumas vezes mais nítidos (TAVARES; MARINHO, 2007).

A forma inicial pode expressar-se por distúrbio da sensibilidade, sem lesão cutânea visível. Essa forma da doença é encontrada em indivíduos de resposta imune não definida

diante do bacilo, usualmente crianças (TAVARES; MARINHO, 2007). E importante a detecção da doença ainda nesta fase, pois há a possibilidade de cura precoce e de prevenção da evolução a formas mutilantes (TAVARES; MARINHO, 2007).

A hanseníase tuberculóide (paucibacilar) é a forma da doença em que o sistema imune da pessoa consegue destruir os bacilos espontaneamente. Assim como na hanseníase indeterminada, a doença também pode acometer crianças (o que não descarta a possibilidade de se encontrar adultos doentes), tem um tempo de incubação de cerca de cinco anos, e pode se manifestar até em crianças de colo, onde a lesão de pele é um nódulo totalmente anestésico na face ou tronco (hanseníase nodular da infância) (SAúDE, 2017).

As lesões são únicas ou em pequeno número, com poucos ou nenhum bacilo, ressaltando a capacidade do hospedeiro de limitar a infecção (TAVARES; MARINHO, 2007). O comprometimento nervoso é precoce, traduz-se por marcada alteração da sensibilidade, tanto térmica como dolorosa e tátil, por espessamento de filetes nervosos próximos à lesão cutânea (TAVARES; MARINHO, 2007).

A hanseníase dimorfa (multibacilar) caracteriza-se, geralmente, por mostrar várias manchas de pele avermelhadas ou esbranquiçadas, com bordas elevadas, mal delimitadas na periferia, ou por múltiplas lesões bem delimitadas semelhantes à lesão tuberculóide, porém a borda externa é esmaecida (pouco definida). Há perda parcial a total da sensibilidade, com diminuição de funções autonômicas (sudorese e vasorreflexia à histamina) (SAúDE, 2017).

A hanseníase virchowiana (multibacilar) dentro do espectro imunológico, corresponde ao pólo de baixa resistência, portanto, multibacilar. Caracteriza-se pela cronicidade da sua evolução. Ocorre em pacientes com deficiência da resposta imune celular e predomínio da resposta imune corporal (TAVARES; MARINHO, 2007).

A hanseníase virchowiana pode advir da forma indeterminada ou eclodir já como forma virchowiana. Caracteriza-se pela infiltração difusa e progressiva da pele, mucosas, das vias aéreas superiores e nervos, sempre de forma simétrica. Pode acometer outros órgãos, como olhos, linfonodos, testículos, fígado e baço (TAVARES; MARINHO, 2007).

2.2.6 Estados Reacionais

A reações hansênicas representam episódios inflamatórios agudos ou subagudos que sem intercalam no curso crônico da doença, tanto nos paucibacilares quanto nos multibacilares. os estados reacionais ocorrem durante o tratamento com a PQT ou após a alta por cura, embora alguns pacientes já iniciem a doença com os episódios reacionais (TAVARES; MARINHO, 2007). Desta forma, é bastante importante o conhecimento das formas clínicas existentes de forma que o tratamento adequado seja realizado para o combate da doença.

As reações hansênicas são as principais causas de lesões neurais que levam às incapacidades físicas, por isso, é importante o diagnóstico precoce para a instituição da

terapêutica adequada, visando à prevenção dessas incapacidades e deformidades.[rotinas]

Frisa-se que há dois tipos de reações hansênicas, sendo a primeira do tipo 1 e a segundo do tipo 2. O processo inflamatório da Reação do tipo 1 envolve, principalmente, a pele e nervos invadidos pelo bacilo (FOSS et al., 2004).

Nesse tipo de reação há aparecimento de novas lesões dermatológicas (manchas ou placas), infiltração, alterações de cor e edema nas lesões antigas, com ou sem espessamento e neurite. (SERVIçOS, 2019). É de fundamental importância o rápido diagnóstico e manejo da reação, pois constituem a maior causa de lesão no nervo periférico e aumento das incapacidades. Não ocorrem em todos os pacientes, mas são frequentes, principalmente entre os pacientes multibacilares (SAúDE, 2017).

A reação tipo 2, é característica dos virchovianos polares, mas virchovianos subpolares e dimorfo-virchovianos, também podem manifestar esta reação. Em geral ocorre após o início do tratamento, mas muitos pacientes, a desenvolvem antes do tratamento, e neste caso, o diagnóstico da doença se faz durante este fenômeno reacional (URA, 2007).

É a expressão clínica mais frequente, cujo quadro inclui nódulos subcutâneos dolorosos, acompanhados ou não de febre, dores articulares e mal-estar generalizado, com ou sem espessamento e neurite (SERVIçOS, 2019).

2.2.7 Indicadores epidemiológicos

O Programa Nacional de Controle da Hanseníase (PNCH) assume como objetivo de saúde pública o controle da doença (WHO, 2008) e privilegia, nesse aspecto, o acompanhamento epidemiológico por meio do coeficiente de detecção de casos novos, em substituição ao indicador de prevalência pontual, optando pela sua apresentação por 100.000 habitantes, para facilitar a comparação com outros eventos (BRASIL, 2009).

O coeficiente de detecção em menores de 15 anos é prioridade da política atual de controle da hanseníase no país, por indicar focos de infecção ativos e transmissão recente (BRASIL, 2009).

O número de casos novos diagnosticado de hanseníase, por 100 mil habitantes, na população residente em determinado espaço geográfico, no ano considerado. A definição de caso de hanseníase baseia-se em critérios adotados pelo Ministério da Saúde para orientar as ações de vigilância epidemiológica e controle da doença em todo o país. A equação a seguir mostra o método de cálculo:

$$\frac{\begin{array}{c}\text{Número de casos novos residentes em determinado} \\ \text{local e diagnosticados no ano da avaliação}\end{array}}{\text{População total residente, no mesmo local e ano de avaliação}} \times 100.000$$

A tabela 2 ilustra a gravidade da taxa de incidência de casos gerais da hanseníase

por 100.000 habitantes.

Tabela 2 – Taxas de incidência da hanseníase por 100.000 habitantes

Taxa de Incidência	Gravidade
menos de 0,2	baixo
0,2 a 9,99	médio
10,0 a 19,99	alto
20,0 a 39,99	muito alto
40,0 ou mais	hiperendêmico

Fonte: Ministério da Saúde. Secretaria de Vigilância em Saúde (SVS)

Esse indicador é responsável por medir a força de morbidade, magnitude e tendência da endemia. Esse tipo de indicador também é definido para menores de 15 anos de acordo com a equação a seguir.

$$\frac{\begin{array}{c}\text{Número de casos novos em menores de 15 anos de idade} \\ \text{residentes em determinado local e diagnosticados no ano da avaliação}\end{array}}{\text{População de zero a 14 anos de idade, no mesmo local e ano de avaliação}} \times 100.000$$

Outra indicador importante é o indicador da proporção de casos novos de hanseníase com grau 2 de incapacidade física no momento do diagnóstico. A sua fórmula é dada por:

$$\frac{\begin{array}{c}\text{Número de casos novos com grau 2 de incapacidade física no} \\ \text{diagnóstico, residentes em determinado local e detectados no ano da avaliação}\end{array}}{\begin{array}{c}\text{Total de casos novos com grau de incapacidade física avaliados,} \\ \text{residentes no mesmo local e ano da avaliação}\end{array}} \times 100$$

Esse indicador possui o objetivo de avaliar a efetividade das atividades de detecção oportuna e/ou precoce de casos. Os parâmetros para esse indicador são definidos na tabela abaixo:

Tabela 3 – Taxas de detecção de casos novos de hanseníase em menores de 15 anos por 100.000 habitantes

Taxa de Incidência	Gravidade
menos de 5,0%	baixo
5% a 9,9%	médio
10,0% ou mais	hiperendêmico

Fonte: Ministério da Saúde. Secretaria de Vigilância em Saúde (SVS)

Já o indicador do total de novos casos de hanseníase com grau 2 de incapacidade

física no momento do diagnóstico é dado pela fórmula abaixo.

$$\frac{\text{Número de casos novos com grau 2 de incapacidade física no diagnóstico, residentes em determinado local e detectados no ano da avaliação}}{\text{População residente no mesmo local e ano da avaliação}}$$

Apesar de haver esse indicador para a avaliar as atividades da detecção, o mesmo não possui a tabela com os parâmetros relativos para a gravidade.

2.3 Processo de Descoberta de Conhecimento em Bases de Dados (KDD)

De acordo com (USAMA; GREGORY; PADHRAIC, 1997) o *Knowledge Discovery in Databases (KDD)* , é um processo não trivial para a identificação de padrões válidos, novos, potencialmente úteis e, em última análise, compreensíveis nos dados. O termo KDD foi cunhado pela primeira vez em 1989 para enfatizar que o conhecimento é o produto final da descoberta baseda em dados (FAYYAD; STOLORZ, 1997a).

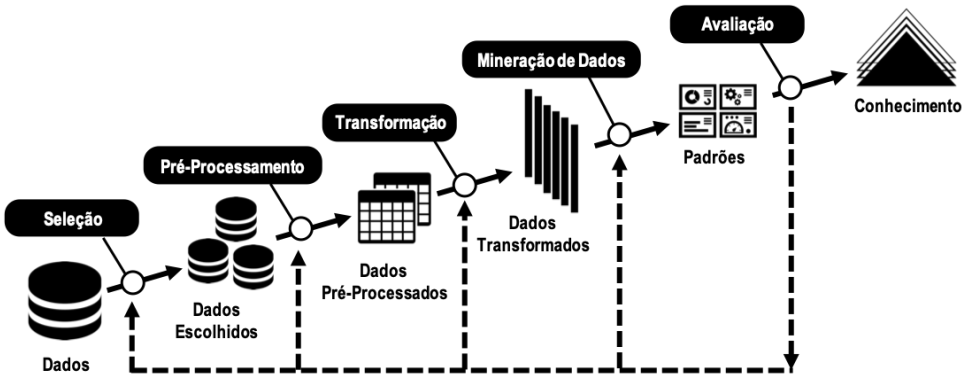
Algumas definições em relação ao KDD são explicadas por (USAMA; GREGORY; PADHRAIC, 1997):

> "Aqui, dados são um conjunto de fatos (por exemplo, casos em um banco de dados) e padrão é uma expressão em algum idioma que representa uma descrição parcimoniosa de um subconjunto dos dados ou de um modelo aplicável a esse subconjunto. Já o termo processo implica que o KDD é composto por várias etapas, que envolvem a preparação de dados, pesquisa de padrões, avaliação de conhecimento e refinamento, todos repetidos em várias iterações. Por não trivial, quer dizer que alguma pesquisa ou inferência está envolvida, ou seja, não é um cálculo direto de quantidades predefinidas, como calcular o valor médio de um conjunto de números. Os padrões descobertos devem ser válidos em novos dados com algum grau de certeza. Pode-se dizer que os padrões sejam novos e potencialmente úteis, ou seja, levem a algum benefício para o usuário / tarefa. Por fim, os padrões devem ser compreensíveis, se não imediatamente, depois de algum pós-processamento."

Portanto, para determinado uso, extrair um padrão também designa ajustar um modelo aos dados, encontrar estrutura a partir dos dados ou, em geral, qualquer descrição de alto nível de um conjunto de dados (FAYYAD; STOLORZ, 1997b). A figura 2 ilustra os passos essenciais do KDD.

As subseções seguintes serão apresentadas as etapas do KDD dividindo-se em três grandes etapas que são o pré-processamento de dados, mineração de dados e por fim o pós-processamento de dados.

Figura 2 – Etapas do KDD



Fonte: Fayyad, 1996

2.3.1 Pré-processamento de dados

A etapa de pré-processamento compreende as funções relacionadas à captação, à organização, ao tratamento e à preparação dos dados para a etapa de Mineração de Dados (GOLDSCHMIDT; PASSOS, 2005).

O pré-processamento de dados é uma etapa importante no processo de mineração de dados, porque se houver muitas informações irrelevantes e redundantes ou dados ruidosos e não confiáveis, a descoberta de conhecimento durante a fase de análise e treinamento será mais difícil (HAMAD; QADER, 2014). Essa etapa possui fundamental relevância no processo de descoberta de conhecimento.

As subseções seguintes apresentarão as principais técnicas de pré-processamento utilizados nesta etapa do KDD.

2.3.1.1 Seleção de Dados

A seleção de dados compreende em essência a identificação de quais informações dentre as bases de dados existentes, devem ser efetivamente consideradas durante o processo de KDD (GOLDSCHMIDT; PASSOS, 2005).

De acordo com (DUNKEL et al., 1997) "o primeiro processo de KDD deve ser a seleção de um conjunto apropriado de dados, e essa seleção deve ser especificamente para os propósitos da mineração de dados, ou deve ser usado em outros propósitos".

A seleção de dados tem um impacto significativo na qualidade de qualquer conhecimento descoberto e a maioria dos sistemas atuais pressupõe que os dados apropriados estão disponíveis para o uso (DUNKEL et al., 1997).

2.3.1.2 Limpeza de Dados

A fase da limpeza de dados envolve uma verificação consistência das informações, a correção de possíveis erros e o preenchimento ou a eliminação de valores desconhecidos

e redundantes, além da eliminação de valores não pertencentes ao domínio (GOLDSCHMIDT; PASSOS, 2005).

A execução dessa fase tem como objetivo, portanto, corrigir a base de dados, eliminando consultas desnecessárias que poderiam ser executadas futuramente pelos algoritmos de Mineração de Dados (GOLDSCHMIDT; PASSOS, 2005). Valores ausentes, ruído e inconsistências contribuem para dados imprecisos (GOLDSCHMIDT; PASSOS, 2005).

Algumas das principais técnicas de limpeza de dados utilizadas são as seguintes:

- Eliminação dos dados ruidosos;
- Eliminação de dados duplicados;
- Eliminação de dados redundantes;
- Substituição de dados faltantes com valores da média, moda ou alguma outra técnica estatística.

Geralmente, a limpeza de dados reduz erros e melhora a qualidade dos dados (RAHMAN et al., 2019). Portanto, é uma etapa bastante importante para melhorar a qualidade e a eficiência dos algoritmos que serão posteriormente utilizados no processo de mineração de dados.

2.3.1.3 Transformação de Dados

Na etapa de transformação dos dados, os dados devem ser codificados para ficarem em um formato no qual que possam ser usados como entrada dos algoritmos de mineração de dados (GOLDSCHMIDT; PASSOS, 2005).

Em muitos casos é necessário a transformação de dados em uma escala semelhante de maneira que não haja um peso muito grande para atributos no qual as suas escalas sejam muito grandes. Para esse tipo de transformação normalmente, utiliza-se duas técnicas para converter os atributos: Normalização e Padronização que são os métodos de transformação de dados mais populares e amplamente utilizados (BHATIA, 2019).

No caso de normalização, todos os atributos são convertidos em uma pontuação normalizada ou em um intervalo (0, 1). Entretanto, há problema da técnica de normalização que são os outliers presentes no conjunto de dados, ou seja, se houver um valor atípico, ele tenderá a reduzir todos os outros valores até o valor zero (BHATIA, 2019).

No caso de padronização, os valores são todos espalhados para que tenhamos um desvio padrão de 1. Usar a padronização tende a fazer com que os valores restantes para todos os outros atributos caiam em intervalos semelhantes, uma vez que todos os atributos terão o mesmo desvio padrão de 1 (BHATIA, 2019).

2.3.1.4 Enriquecimento de dados

A fase de enriquecimento consiste em conseguir agregar mais informações aos registros existentes para que estes forneçam mais elementos para o processo de descoberta de conhecimento. De acordo com (GOLDSCHMIDT; PASSOS, 2005) algumas técnicas utilizadas para o enriquecimento de dados são:

- Pesquisas: Nesta operação estão incluídas todas as iniciativas de enriquecimento que envolvem a captação de novas informações junto às fontes originais. Normalmente requerem a inclusão de novos atributos ou mesmo de novas tabelas nas bases de dados existentes;

- Consulta a bases externas: O processo de enriquecimento pode ser realizado mediante a incorporação de informações fornecidas por outros sistemas. É muito comum a importação de informações advindas de outras bases de dados.

2.3.2 Mineração de Dados

A fase de mineração de dados é uma fase do processo de Descoberta de Conhecimento em Banco de Dados (DCBD). Esta etapa é responsável pela aplicação dos algoritmos que são capazes de identificar e extrair padrões relevantes presente nos dados (GOLDSCHMIDT; PASSOS, 2005).

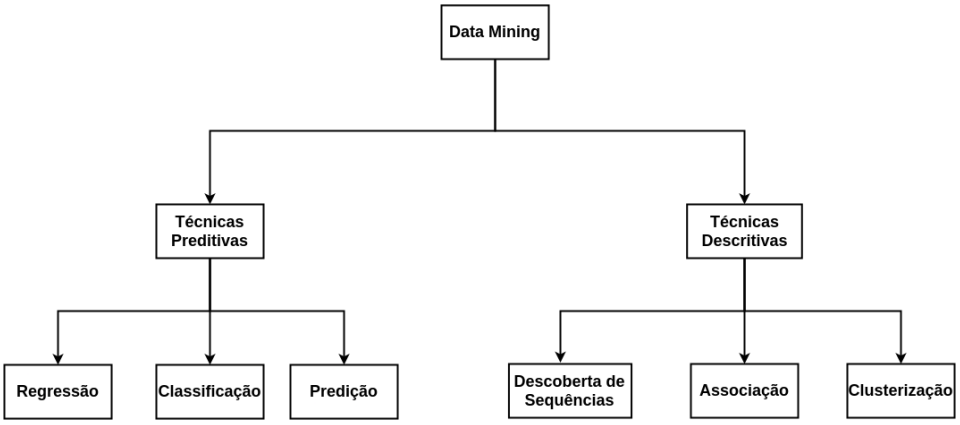
A mineração de dados é um processo essencial em que métodos inteligentes são aplicados para extrair padrões de dados. Ele procura padrões de interesse em uma forma representacional específica, ou um conjunto de tais representações, incluindo regras ou árvores de classificação, regressão, clustering, modelagem de sequência, dependência e assim por diante.

Nesta etapa é realizada a busca efetiva por conhecimentos úteis no contexto da aplicação de KDD. No processo de descoberta de conhecimento é considerado a principal etapa do processo de KDD (GOLDSCHMIDT; PASSOS, 2005). A figura 3 ilustra as tarefas de data mining.

Existem dois tipos principais de tarefas de mineração de dados que são a descrição e a predição. A tarefa de descrição consiste na descoberta automática de padrões previamente desconhecidos que descrevem as propriedades gerais dos dados existentes (BLAIEWICZ et al., 2003).

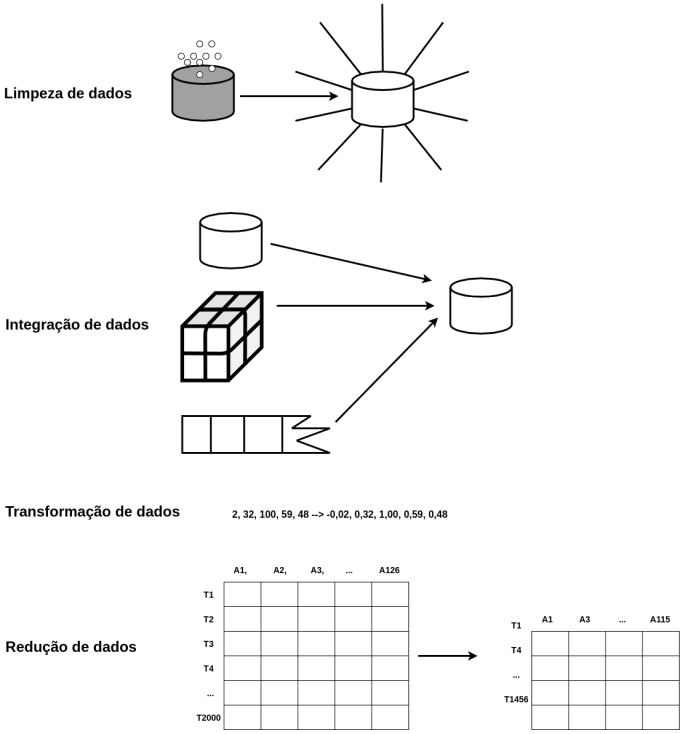
As tarefas de predição normalmente tentam fazer previsões de tendências e comportamentos com base na inferência dos dados disponíveis (BLAIEWICZ et al., 2003). O atributo a ser predito é chamado de alvo ou variável dependente, enquanto que o uso de atributos para fazer as predições são conhecidos como explicatórios ou variáveis dependentes (TAN; STEINBACH; KUMAR, 2006).

Figura 3 – Tarefas de data mining



Diversas técnicas podem ser utilizadas nessa etapa, e a escolha da técnica utilizada depende muitas vezes no tipo de tarefa de KDD a ser realizada. A seguir serão apresentadas as técnicas mais comuns da mineração de dados.

Figura 4 – Atividades de pré processamento



Fonte: Kamber (2012)

2.3.2.1 Descoberta de Associação

A mineração de regras de associação, é uma das técnicas mais importantes e bem pesquisadas de mineração de dados, foi introduzida pela primeira vez em (AGRAWAL; IMIELI;SKI; SWAMI, 1993). Possui o objetivo de extrair correlações interessantes, padrões frequentes, associações ou estruturas casuais entre conjuntos de itens nos bancos

de dados de transações ou outros repositórios de dados (KOTSIANTIS; KANELLOPOULOS, 2006).

A mineração frequente de conjuntos de itens leva à descoberta de associações e correlações entre itens em grandes conjuntos de dados transacionais ou relacionais. As regras de associação são amplamente usadas em diversas áreas, tais como redes de telecomunicações, gerenciamento de mercado e risco, controle de estoque (KOTSIANTIS; KANELLOPOULOS, 2006). A representação de uma base de dados com diversas transações é mostrada na tabela 4 abaixo.

Tabela 4 – Exemplo de uma tabela contendo algumas transações

TID	**Items**
1	Pão, Leite
2	Pão, Fraldas, Cerveja, Ovos
3	Leite, Fraldas, Cerveja, Refrigerante
4	Pão, Leite, Fraldas, Cerveja
5	Pão, Leite, Fraldas, Refrigerante

A tabela 4 ilustra a forma como uma tabela de transações é representada, entretanto a entrada da tabela para o algoritmo de mineração de regras de associação é mostrada a seguir por uma tabela booleana (tabela 5) chamada de tabela de *market basket* que é o modo de entrada de dados do algoritmo.

Tabela 5 – Exemplo de uma tabela no formato *market basket*

Transação	**Leite**	**Ovos**	**Cerveja**	**Pão**	**Refrigerante**	**Fraldas**
1	1	0	0	1	0	0
2	0	1	1	1	0	1
3	1	0	1	0	1	1
4	1	0	1	1	0	1
5	1	0	0	1	1	1

Considera-se $I = \{I_1, I_2, I_3...I_m\}$ como sendo um conjunto de literais, chamados de itens. Seja D um conjunto de transações, na qual cada transação T é um conjunto de itens tal que $T \subseteq I$. Associado com cada transação é um único identificador chamado de TID. Dizemos que uma transação T contém X, um conjunto de em I, se $X \subseteq T$. Uma regra de associação é uma implicação da forma $X \Longrightarrow Y$, onde $X \subset I, Y \subset I$ e $X \cap Y = \emptyset$.

Uma regra $X \Longrightarrow Y$ é mantida no conjunto de transações D com confiança c se $c\%$ das transações em D que contêm X também o conjunto de transações D se $s\%$ das transações em D contêm $X \cup Y$. Uma transação é o nome atribuído ao elemento de ligação existente em cada ocorrência de itens no banco de dados (GOLDSCHMIDT; PASSOS, 2005).

Normalmente, as regras de associação são consideradas interessantes se satisfizerem um limite mínimo de suporte e um limite mínimo de confiança. Esses limites podem ser definidos por usuários ou especialistas em domínio (HAN; KAMBER; PEI, 2012).

Em geral, um conjunto de itens (como o antecedente ou o consequente de uma regra) é chamado de conjunto de itens. O número de itens em um conjunto de itens é chamado de comprimento de um conjunto de itens. Os conjuntos de itens de algum comprimento k são referidos como *k-itemsets* (KOTSIANTIS; KANELLOPOULOS, 2006).

Geralmente, um algoritmo de mineração de regras de associação contém as seguintes etapas:

- O conjunto de conjuntos de itens-k candidatos é gerado por 1 extensões dos conjuntos de itens grandes (k-1) gerados na iteração anterior.
- Os suportes para os *k-itemsets* candidatos são gerados por uma passagem no banco de dados.
- Os conjuntos de itens que não têm o suporte mínimo são descartados e os demais são chamados de *k-itemsets* grandes.

Esse processo é repetido até que não sejam encontrados mais.

Dessa forma, dada um conjunto de transações, o problema de mineração de dados por regras de associação está em gerar todas as regras que contenham o suporte e confiança iguais ou maiores do que os valores mínimos determinados pelo usuário, referenciados como suporte mínimo e confiança mínima respectivamente. O suporte de uma gera $X \longrightarrow Y$, onde X e Y são um conjunto de itens, é dado pela seguinte fórmula:

$$Suporte(X \longrightarrow Y) = \frac{\text{Frequência de } (X \wedge Y)}{\text{(Total de T)}} \tag{1}$$

O numerador se refere ao número de transações em que X e Y ocorrem simultaneamente e o denominador ao total de transações.

Para a mineração de regras de associação há métricas básicas para a extração das regras de associação. São elas a confiança(c) a o suporte mínimo(sup) lift. A Confiança (equação 2) é definida como a medida de certeza ou confiabilidade associada a cada regra descoberta. Matematicamente, confiança é a porcentagem de transações que contêm X e Y de todas as transações que contêm X.

$$Confianca(X \longrightarrow Y) = \frac{Suporte(X \wedge Y)}{Suporte(X)} \tag{2}$$

Embora a confiança possa identificar as regras interessantes de todas as regras candidatas, ela vem com um problema. Dadas regras na forma de $X \longrightarrow Y$, a confiança considera apenas o antecedente (X) e a co-ocorrência de X e Y; não leva em consideração o consequente da regra (Y).

Portanto, a confiança não pode dizer se uma regra contém uma implicação verdadeira do relacionamento ou se a regra é pura coincidência. X e Y podem ser estatisticamente independentes e ainda assim receber um alto índice de confiança.

Outras medidas que apesar de não serem tão utilizadas, como elevação e *leverage* (alavancagem), são projetadas para resolver esse problema. Já a *leverage* é uma noção semelhante, mas, em vez de usar uma razão, a *Leverage* usa a diferença. A *leverage* mede a diferença na probabilidade de X e Y aparecerem juntos no conjunto de dados em comparação com o que seria esperado se X e Y fossem estatisticamente independentes um do outro de acordo com a equação 3.

$$Leverage(X \longrightarrow Y) = Suporte(X \wedge Y) - Suporte(X) * Suporte(Y) \tag{3}$$

Um relacionamento pode ser considerado interessante quando o algoritmo identifica o relacionamento com uma medida de confiança maior ou igual a um limite predefinido. Esse limite predefinido é chamado a confiança mínima.

O *lift* mede quantas vezes mais X e Y ocorrem juntos do que o esperado, se forem estatisticamente independentes um do outro. O *lift* é uma medida de como X e Y estão realmente relacionados, em vez de coincidentes acontecerem juntos de acordo com a equação seguinte.

$$Lift(X \longrightarrow Y) = \frac{Suporte(X \wedge Y)}{Suporte(X) * Suporte(Y)} \tag{4}$$

2.3.2.2 Classificação

Classificação é um aprendizado uma função que mapeia (classifica) um item de dados em uma das várias classes predefinidas (Weiss e Kulikowski 1991; Hand 1981).

No aprendizado de classificação, um classificador é apresentado como um conjunto de exemplos que já estão classificados e, a partir desses exemplos, o classificador aprende a atribuir exemplos não vistos (DIETRICH; HELLER; YANG, 2015).

É um método clássico usado por pesquisadores de aprendizado de máquina e estatísticos para prever o resultado de amostras desconhecidas e é usado para categorizar objetos (ou coisas) em um determinado número discreto de classes (BHATIA, 2019).

O processo de classificação de dados envolve aprendizagem e classificação. Na aprendizagem, os dados de treinamento são analisados por algoritmo de classificação. Na classificação, os dados de teste são usados para estimar a precisão das regras de classificação. Se a precisão for aceitável, as regras podem ser aplicadas às novas tuplas de dados (RAMAGERI, 2015).

2.3.2.3 Clusterização

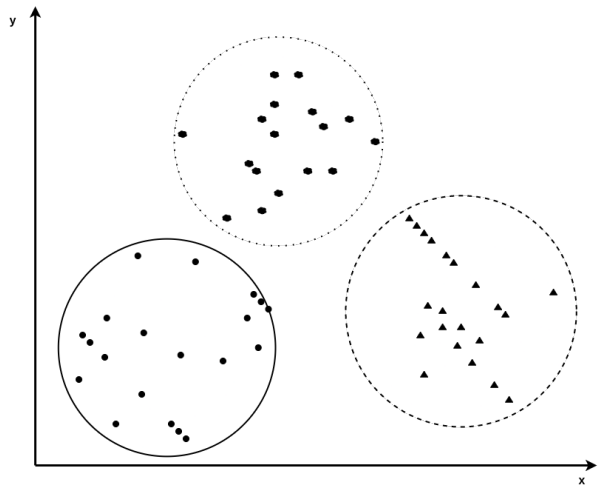
A tarefa de clusterização, também chamada de Agrupamento, é usada para particionar os registros de uma base de dados em subconjuntos ou clusters, de tal forma que elementos em um cluster compartilhem um conjunto de propriedades comuns que os distingam dos elementos de outros clusters (GOLDSCHMIDT; PASSOS, 2005).

As técnicas de cluster são do tipo não supervisionadas no sentido de que o cientista de dados não determina, com antecedência, os rótulos a serem aplicados aos clusters. A estrutura dos dados descreve os objetos de interesse e determina a melhor forma de agrupar os objetos (DIETRICH; HELLER; YANG, 2015).

Algumas das aplicações de clusterização de dados são as seguintes:

- Analisar o histórico de crédito de clientes de bancos para identificar se seria arriscado ou seguro conceder empréstimos a eles;
- Analisar o histórico de compras dos clientes de um shopping para prever se irão comprar determinado produto ou não;
- O clustering também é útil em aplicativos de detecção de outliers, como detecção de fraude de cartão de crédito.

Figura 5 – Clusterização



2.3.2.4 Regressão

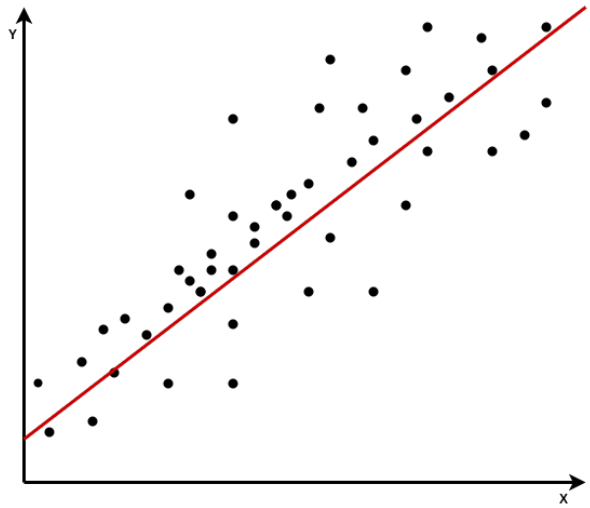
A regressão é uma técnica de mineração de dados (aprendizado de máquina) usada para ajustar uma equação a um conjunto de dados. A forma mais simples de regressão é a regressão linear (HAN; KAMBER; PEI, 2012), usa a fórmula de uma equação de uma

linha reta $y = ax + b$ e determina os valores apropriados para a e b para prever o valor de y com base em um determinado valor de x.

$$y = \alpha + \beta x \tag{5}$$

A regressão também é conhecida por predição funcional, predição de valor real, função de aproximação, ou ainda, aprendizado de classes contínuas (USUALDO, 2003). Enquanto que a classificação prevê os dados categóricos, a regressão é aplicada a valores numéricos, tendo como principal propósito a previsão de dados históricos existentes em uma base de dados (MICHIE; TAYLOR, 1994).

Figura 6 – Regressão linear



Técnicas avançadas, tal qual as regressões múltiplas, predizem a relação entre múltiplas variáveis. A adição de mais variáveis aumenta consideravelmente a complexidade da predição.

Os métodos de regressão podem ser utilizados em diversas áreas de conhecimento como, por exemplo, para previsão da economia nacional com base em certas informações (tais como níveis de renda e investimentos), para a verificação de quais fatores ajudam a manter a qualidade dos serviços oferecidos ou na medida de viabilidade de um novo produto. (Abricom).

2.3.2.5 Apriori

Apriori é um algoritmo proposto por R. Agrawal e R. Srikant em 1994 para mineração de conjuntos de itens frequentes para regras de associação booleana (HAN; KAMBER; PEI, 2012). O algoritmo emprega busca em profundidade e gera conjuntos de itens candidatos (padrões) de k elementos a partir de conjuntos de itens de $k-1$ elementos [xxx].

Este algoritmo consiste em duas fases. A primeira passagem do algoritmo simplesmente conta a ocorrência de itens para determinar grandes *1-itemsets*. Já a passada

subsequente, é dita como passagem k, consiste em duas fases. Na primeira passagem os grandes conjuntos de L_{k-1} encontrados na (k)-ésima passagem são usados para gerar os conjuntos de itens candidatos C_k usando a função chamada de *apriori-gen*, para a geração de candidatos (AGRAWAL; SRIKANT, 1994). O pseudocódigo 1, localizado abaixo representa o algoritmo apriori.

Algoritmo 1 - Algoritmo Apriori

```
1: procedure APRIORI(D)
2:     L_1 ← {1-itemsets grande}
3:     for (k=2; L_{k-1} != 0; k++) do
4:         C_k ← apriori-gen(L_{k-1});
5:         for each transactions t ∈ D do
6:             C_t ← subset(C_k, t);
7:             for each candidates c ∈ C_t do
8:                 c.count++;
9:         L_k ← {c ∈ C_k | c.count >= minsup};
10: Answer ← U_k L_k;
```

Fonte: (AGRAWAL; IMIELI;SKI; SWAMI, 1993)

Em seguida, o banco de dados é verificado e o suporte dos candidatos em C_k é contado. Para uma contagem rápida, precisa-se determinar determinar com eficiência os candidatos em C_k em que estão contidos em uma determinada transação t (AGRAWAL; SRIKANT, 1994). As notações e definições dos conjuntos de dados são mostrados na tabela 6.

Tabela 6 – Definições das notações

k-itemset	**Um *itemset* contendo k items**
L_k	Conjunto de *k-items* grandes L_k (aqueles com suporte mínimo). Cada membro deste conjunto possui dois campos, sendo o *itemset* e a contagem de suporte.
C_k	Conjunto de *k-items* candidatos C_k (conjunto de itens potencialmente grandes). Cada membro deste conjunto possui dois campos, sendo o conjunto de itens e a contagem de suporte.
$\overline{C_k}$	Conjunto de *k-itemsets* candidatos quando os TIDs das transações geradoras são mantidos associados aos candidatos.

Fonte: (AGRAWAL; SRIKANT, 1994)

A geração dos itemsets candidatos, de antemão, toma como argumento L_{k-1}, o conjunto de todos (k1)-itemsets frequentes. Para tal, utiliza-se a função *apriori_gen*, que retorna um superconjunto de todos os k-itemsets freqüentes. A intuição por trás desse procedimento é que se um itemset X tem suporte mínimo, todos os seus subconjuntos também terão (AGRAWAL; SRIKANT, 1994). A função, em um primeiro estágio, une L_{k-1} com L_{k-1}. No estágio seguinte, são eliminados os itemsets $c_k \in C_k$, desde que um dado (k-1)-subset de c_k não pertença a L_{k-1}.

2.3.3 Pós processamento

A última fase do processo é a análise e interpretação das informações descobertas. Ao longo dos anos, muitos esforços se concentraram em melhorar o desempenho algorítmico (em termos de tempo de execução e consumo de memória), mas essa fase foi surpreendentemente negligenciada (RUIZ; KAMSU-FOGUEM; GRABOT, 2014).

O pós-processamento dos resultados está se tornando cada vez mais importante nas empresas, a fim de encontrar e validar as regras mais interessantes para cada problema específico (P.A.; B., 2014).

O pós processamento é um componente importante do KDD que consiste em variados procedimentos e métodos que podem ser categorizados nos seguintes grupos de acordo com (BRUHA; FAMILI, 2000):

- Filtragem de conhecimento: Truncamento e pós-poda de regra. Se os dados de treinamento são ruidosos, o algoritmo indutivo gera folhas de uma árvore de decisão ou regras de decisão que cobrem um número muito pequeno de objetos de treinamento. Isso acontece porque o algoritmo indutivo (aprendizado) tenta dividir subconjuntos de objetos de treinamento em subconjuntos ainda menores que seriam genuinamente consistentes. Para superar esse problema, uma árvore ou um conjunto de regras de decisão deve ser reduzido, por pós-poda (árvores de decisão) ou truncamento (regras de decisão);

- Interpretação e explicação: Utiliza-se o conhecimento adquirido para previsões ou para utilização em um módulo para um sistema especialista como base de conhecimento. A visualização do conhecimento também é utilizada de forma que seja compreensível ao usuário final. Nesta etapa, também pode resumir regras e combiná-las com um conhecimento específico do domínio fornecido para a tarefa dada.

- Avaliação: Depois que um sistema de aprendizagem induz as hipóteses de modelos a partir de um conjunto de treinamento, sua avaliação deve ocorrer;

- Integração de conhecimento: Os sistemas tradicionais de tomada de decisão dependem de uma única técnica, estratégia, modelo. Novos sistemas sofisticados de apoio

à decisão combinam ou refinam os resultados obtidos a partir de vários modelos, geralmente produzidos por métodos diferentes.

Logo, essa etapa possui bastante importância porque os modelos validam os dados analisados trazem novos conhecimentos aos usuários e são úteis para o apoio na tomada de decisão por dar uma compreensão maior para o tomador de decisão (RUIZ; KAMSU-FOGUEM; GRABOT, 2014).

3 METODOLOGIA

Este capítulo é responsável pela apresentação da metodologia utilizada neste trabalho.

3.1 Tipo de Estudo

Este estudo baseia-se em uma estratégia quantitativa de pesquisa, de caráter descritivo e explicativo. A pesquisa do tipo descritivo é quando o pesquisador apenas registra e descreve os fatos observados sem interferir neles, visa a descrever as características de determinada população ou fenômeno ou o estabelecimento de relações entre variáveis (PRODANOV; FREITAS, 2013).

Tal pesquisa observa, registra, analisa e ordena dados, sem manipulá-los, isto é, sem interferência do pesquisador. Procura descobrir a frequência com que um fato ocorre, sua natureza, suas características, causas e relações com outros fatos (PRODANOV; FREITAS, 2013).

Nesta seção, pretende-se demonstrar os procedimentos metodológicos do tipo de pesquisa utilizado por meio da abordagem quantitativa através da aplicação de técnicas de estatística descritiva na base de dados do estudo.

Em relação a técnica descritiva a análise da base de dados para explicar os eventos ocorridos com base nas características apresentada de forma explícita e de forma implícita ao utilizar as técnicas de mineração de dados para a extração de conhecimento implícito para a explicação de um fenômeno.

3.2 Local de Estudo

O local de estudo deste trabalho corresponde às notificações do estado do Tocantins reportadas pelo Sistema de Notificação de Agravos (Sinan) notificadas no Estado do Tocantins. O Tocantins é um estado localizado na região norte com população de 1.590.248 habitantes de acordo com a projeção do IBGE em 2020 e de uma área de 277.423,630 km^2 sendo composto por 139 municípios (IBGE, 2020).

Figura 7 – Localização geográfica do estado do Tocantins

3.3 Revisão bibliográfica

Como uma primeira etapa de um estudo apresentamos o levantamento bibliográfico, que tem por finalidade levantar todas as referências encontradas sobre um determinado tema (CERVO; BERVIAN, 2002). Essas referências podem estar em qualquer formato, ou seja, livros, sites, revistas, vídeo, enfim, tudo que possa contribuir para um primeiro contato com o objeto de estudo investigado.

Para a realização do trabalho presente foi realizado a consulta bibliográfica em relação às áreas referentes a metodologia de descoberta de conhecimento chamado de *Knowledge Discovery in Databases (KDD)*, com uma atenção maior na etapa referente à mineração de dados e suas tarefas mais importantes, em especial a tarefa a ser utilizada neste trabalho.

A realização da busca bibliográfica em conceitos básicos de estatística descritiva foram realizados com base na análise exploratória de dados a fim de descrever os atributos e explorá-los de forma que haja uma maior compreensão dos dados presentes na base de dados. Por fim, realizou-se a busca referencial às referências na área da saúde sendo mais específico na área na doença hanseníase.

3.4 Ferramentas

Para a realização dos trabalhos nessa pesquisa serão utilizadas diversas ferramentas em diferentes módulos, sendo eles os módulos de realização da análise exploratória, responsável pela geração dos gráficos estatísticos estáticos, o módulo de geração de regras utilizando-se algoritmos de mineração de dados, responsável por aplicar os algoritmos de mineração de dados e gerar os resultados.

Para a análise de dados será utilizada a linguagem de programação Python na versão 3.8 utilizando-se as bibliotecas Numpy e Pandas para importação e processamento dos dados e por fim a biblioteca Matplotlib para a visualização dos dados explorados.

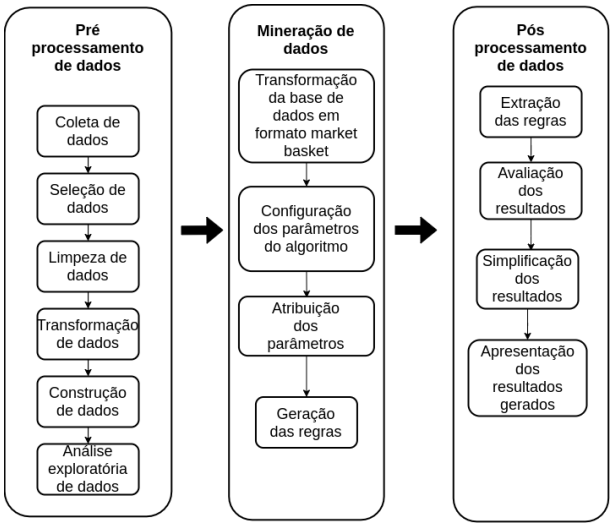
Para a aplicação dos algoritmos de mineração de dados foi utilizado uma biblioteca para geração de regras de associação para a linguagem Python, ela é responsável pela

transformação da base de dados com o atributos categóricos para a uma tabela de *market basket* e gerar as regras de associação para a análise.

3.5 Procedimentos

Os procedimentos utilizados neste trabalho seguirão as mesmas bases de procedimentos utilizados nas etapas do KDD apresentados na seção 2.3 para a realização da extração de conhecimento. O fluxograma a seguir ilustra melhor os procedimentos a serem utilizados neste trabalho.

Figura 8 – Fluxograma da metodologia



3.5.1 Pré-processamento de dados

A etapa de pré-processamento de processamento de dados foi realizada seguindo-se as etapas do KDD, apresentadas em (GOLDSCHMIDT; PASSOS, 2005). As subetapas dos pré processamento realizadas neste trabalho são as seguintes: coleta de dados, seleção de dados, limpeza de dados, transformação de atributos, construção de atributos, e a análise exploratória de dados.

3.5.1.1 Coleta de dados

A coleta de dados foi viabilizada através da solicitação da base de dados ao Sistema de Informação de Agravos de Notificação (Sinan-TO) para os casos de hanseníase no Estado do Tocantins. O seu devido uso foi aprovado pelo comitê de ética da Universidade Federal do Tocantins por se tratar de pesquisas envolvendo seres humanos.

O arquivo recebido encontrava-se no formato .dbf, que consiste em uma planilha eletrônica. Para que fosse possível realizar as funções posteriores, foi necessário a realização da importação da base de dados para um Sistema de Gerenciamento de Bancos de

Dados (SGBDs) postgreSQL para que fosse realizado os posteriores tratamentos da base de dados.

O dicionário de dados da base dados é disponibilizado no site do Sinan-TO, nele contém as informações relevantes dos atributos existentes na base de dados. A base de dados é fruto dos dados contidos no dicionário de dados da ficha de notificação individual de agravos e da ficha de notificação de agravos da hanseníase.

As primeiras operações realizadas na base de dados foram a contabilização do total de atributos no qual foi contabilizado um total de 90 atributos e 21.952 registros na base de dados original, sendo divididos em dados do paciente, doença, procedimentos, acompanhamento, localização e por fim temporal.

Em vista disso, foi criado uma tabela com o nome do atributo, frequência absoluta e relativa de preenchimento, com o intuito de se saber a porcentagem de preenchimento de forma a saber quais dados futuramente serão selecionados para os algoritmos de mineração de dados, como uma das formas de parâmetro para a seleção de dados significativos no processo de mineração de dados.

3.5.1.2 Seleção de Dados

A etapa de seleção de dados neste trabalho foi orientado a partir da consulta prévia do dicionário de dados e da análise da relevância que os atributos poderiam representar para o estudo posterior. Para a seleção de dados foi verificado as características dos pacientes, as formas de tratamento, as características da doença, os procedimentos e por fim as variáveis temporais.

Atributos no qual era possível identificar o paciente, como por exemplo o nome do paciente, nome da mãe foram desconsiderados do estudo pelo fato de identificarem as pessoas sendo que o mesmo é um dado sensível e ainda irrelevante para o estudo, atributos internos do sistema ou atributos que possuem uma constante única também foram desconsiderados.

Atributos textuais também foram desconsiderados do processo de seleção por não ser possível a realização da mineração de dados e descoberta de padrões posteriormente. A tabela 7 lista os atributos que foram selecionados para o uso posterior para a análise exploratória de dados e para a mineração de dados.

Tabela 7 – Dicionário de dados dos atributos selecionados da base de dados

Atributo	Descrição	Categoria	DBF
Sexo	Sexo do paciente	- Masculino - Feminino	CS_SEXO

Continua na próxima página

Tabela 7 – Continuação da tabela 7

Atributo	Descrição	Categoria	DBF
Raça	Raça do paciente	- Branca - Preta - Parda - Amarela - Indígena - Ignorada	CS_RACA
Idade	Idade do paciente na data do diagnóstico		NU_IDADE_N
Escolaridade	Escolaridade do paciente	- Analfabeto - Ensino Fundamental Incompleto - Ensino Fundamental Completo - Ensino Médio Incompleto - Ensino Médio Completo - Ensino Superior Incompleto - Ensino Superior Completo - Não se aplica - Ignorado	CS_ESCOL_N
Classificação operacional	Classificação operacional	- Paucibacilar - Multibacilar	CLASSOPERA
Forma clínica	Forma clínica	- Indeterminada - Tuberculóide - Dimorfa - Virchowiana - Não avaliado	FORMACLINI
Avaliação de incapacidade física	Avaliação de incapacidade física no momento do diagnóstico	- Grau 0 - Grau I - Grau II - Não avaliado	AVALIA_N

Continua na próxima página

Tabela 7 – Continuação da tabela 7

Atributo	Descrição	Categoria	DBF
Modo de entrada	Modo de entrada do paciente no sistema	- Caso Novo - Transferência de mesmo município - Tranferência de outro município - Transferência de outro estado - Transferência de outro país - Recidiva - Outros reingressos - Ignorado	MODOENTR
Tipo de Saída	Tipo de saída do paciente	- Cura - Transferência para mesmo município - Transferência para outro estado - Transferência para outro país - Óbito - Abandono - Erro diagnóstico - Transferência não especificada	TPALTA_N
Município de residência atual	Município de residência do paciente		MUNIRESAT

Numa análise posterior foi realizado uma seleção dos dados na forma de redução vertical no qual resultou na exclusão de registros, como no caso do exemplo dos anos de notificação que anteriormente se iniciava nos anos de 1998 e terminava no ano de 2016.

Pelo fato de existir uma quantidade de apenas 12 registros, uma frequência demasiadamente pequena de pacientes no período compreendido dos anos de 1998 a 2000 em relação aos anos posteriores, a exclusão desses registros foi executada para a evitar a presença de *outliers*.

3.5.1.3 Limpeza de Dados

Os procedimentos realizados neste trabalho foram diversos, tais como a correção de dados preenchidos de maneira incorreta ou que possuíam valores faltantes.

Os que estavam faltantes foram corrigidos para categorias desconhecidas para os atributos que já existiam a mesma. Atributos com escalas incorretas foram corrigidas, tais como a variável idade na qual não estavam num valor aceitável para uma variável de idade em anos, logo a escala do valor para este atributo foi corrigida para o valor em anos.

3.5.1.4 Transformação de dados

No que tange à transformação de dados, foi realizado a transformação no atributo NU_IDADE_N, que corresponde a idade do paciente, o mesmo se encontrava-se em uma forma de representação diferente da representação em anos. A tabela 8 mostra o esquema utilizado para o registro da idade do paciente para as notificações, sendo que os algarismos em negrito são os dígitos significantes e os algarismos de milhar são responsáveis por mudar o contexto da idade (dias, semanas, meses e anos).

Tabela 8 – Exemplo da representação da idade do paciente na base de dados

Representação da idade	Descrição
10**15**	15 dias
200**3**	3 semanas
300**5**	5 meses
40**45**	45 anos

Na tabela 8, podemos pegar como um exemplo a representação 3005 presente na linha 3, sendo a representação da idade de um paciente de 5 meses. A transformação dessas representações foram extraídas para cada contexto e transformada em uma quantidade em anos, sendo que idades inferiores a 0 anos foram convertidas para idade igual a 0. As idades que após a conversão ficaram com número com casas decimais foram convertidos para o valor inteiro.

Poucos registros apresentavam representações de idade inferior a 4.000, logo os mesmos foram convertidos para a escala em anos. Já a maioria dos pacientes que tinham a representação da idade com valor maior que 4.000 apenas uma operação de subtração 4.000 do seu valor de idade foi calculada para chegar ao seu número de idade em anos.

3.5.1.5 Construção de atributos

A construção de atributos basicamente foi responsável pela criação de um novo atributo a partir de mesmo já existente na base de dados, sem necessariamente alterá-lo. O atributo em questão foi novamente o NU_IDADE_N da idade, que anteriormente havia sido transformado para a representação em anos. O mesmo foi utilizado para criação de

um novo atributo chamado de FAIXA_IDADE, sendo que o mesmo agora é caracterizado como de um tipo categórico sendo a representação de intervalos de idade de acordo sendo representado pelos valores de [0 a 9 anos, 10 a 14 anos, 15 a 29 anos, 30 a 49 anos, 50 a 74 anos e 75 a 100 anos].

3.5.2 Análise Exploratória de Dados

A partir dos dados pré processados nas etapas anteriores foi realizado a análise exploratória de dados orientada aos atributos categóricos do paciente, da doença e de atributos temporais e de localidade.

A representação visual foi gerada a partir da criação de de gráficos tais como os gráficos de barra para as variáveis categóricas, histogramas para as variáveis de idade do paciente.

3.5.3 Processamento de Dados

Esta etapa será responsável pela utilização dos algoritmos de mineração de dados propostos de acordo com os objetivos esperados.

A partir dos objetivos esperados foi realizado o estudo do algoritmo a ser utilizado de acordo com as limitações da base de dados e portanto para esse tipo de tarefa foi escolhido o algoritmo apriori para geração de regras de associação. O mesmo foi escolhido por se tratar de um algoritmo muito importante descrito na literatura e muito utilizado em pesquisas.

De forma que fosse necessário a executar o algoritmos para geração de regras de associação foi necessário a formatação da base de dados de maneira que a mesma se adequasse a estrutura de uma base *market basket* para a utilização do algoritmo apriori, para isso realizado a seleção e formatação dos atributos categóricos que especificam as características da doença e do paciente.

Na próxima etapa foi realizado a configuração dos parâmetros utilizados para a geração das regras, esses parâmetros são o suporte, a confiança, ambos apresentados na seção 2.3.2.1.

A escolha dos mesmo se torna muito importante pois de acordo com os seus valores o resultado pode acarretar na geração de muitas regras sendo que muitas não terão muita importância para o estudo, sendo assim um bom limiar garante uma chance maior de se conseguir regras com mais confiança e relevância.

3.5.4 Pós-Processamento de Dados

O pós processamento de dados foi realizado a etapa de validação dos resultados obtidos através da consulta com o especialista da área. Os dados obtidos são analisados pelo especialista e devidamente validados, como forma de geração de novo conhecimento.

A geração das regras de associação são filtradas em conjunto com o especialista para a seleção apenas das regras que possuam um conhecimento explícito e válido.

Nesta etapa também é realizada a visualização dos resultados gerados de forma que o conhecimento técnico seja repassado de forma mais simples para a interpretação de quem for analisar os resultados.

4 RESULTADOS

Esta seção é responsável pela apresentação dos resultados obtidos através dos métodos propostos na metodologia presente na seção anterior. Esta seção está dividida entre os resultados gerados através da análise exploratória de dados e da mineração de dados, através da geração das regras de associação.

4.1 Análise Exploratória de Dados

Os primeiros resultados gerados são os referentes a análise exploratória de dados na qual foi realizada a mensura e exploração dos dados e a geração de gráficos para a visualização dos dados de forma mais facilitada.

4.1.1 Análise unidimensional

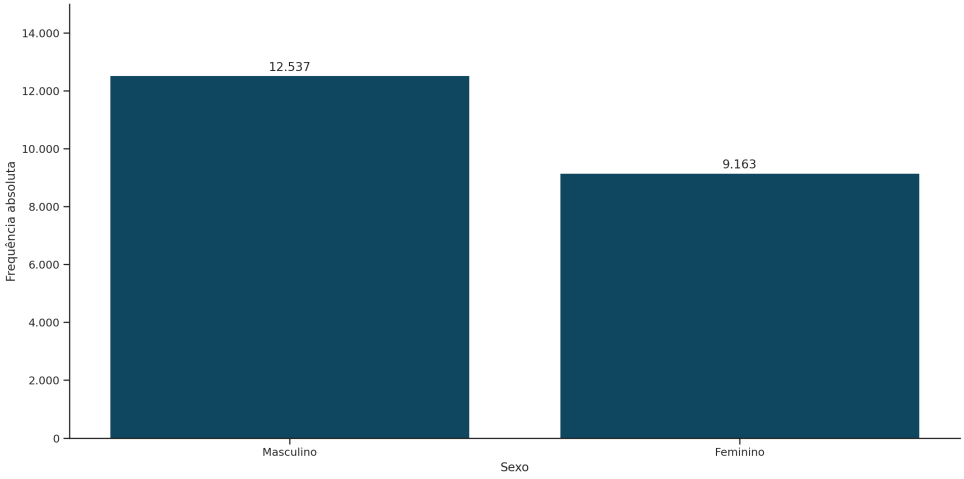
Na análise unidimensional é responsável apenas pela análise dos atributos simples, na qual mostra-se apenas os mesmos, sem estabelecer qualquer relação com outras variáveis. Nesta etapa foi realizado a análise das características do paciente, da doença as geográficas e por fim temporais no período de 2001 a 2016.

4.1.1.1 Características dos pacientes: Sexo, raça, idade, escolaridade

Neste espaço analisamos os atributos dos pacientes de forma a analisar a distribuição das características dos pacientes afetados pela doença.

A figura abaixo (figura 9) apresenta a distribuição de casos pela variável sexo do paciente.

Figura 9 – Casos de hanseníase pela variável sexo

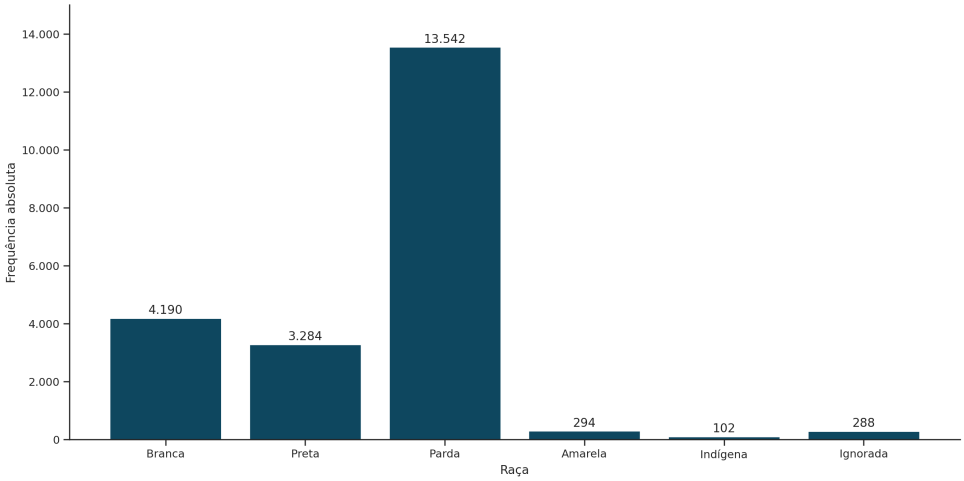


Fonte: Compilação do autor

Analisando-se a figura 9, vê-se que os casos concentram-se na população masculina 12.537 casos (57.77%), já na população feminina, concentram-se 9.163 (42,33%) sendo que a população masculina possui 15,44% a mais que ocorre na população feminina.

Um outro atributo característico do paciente é raça e também é ilustrado de forma a verificar se a quantidade de casos da doença possui alguma relação com a raça do paciente. A figura 10 ilustra a quantidade de casos para a variável raça.

Figura 10 – Casos de hanseníase pela variável raça

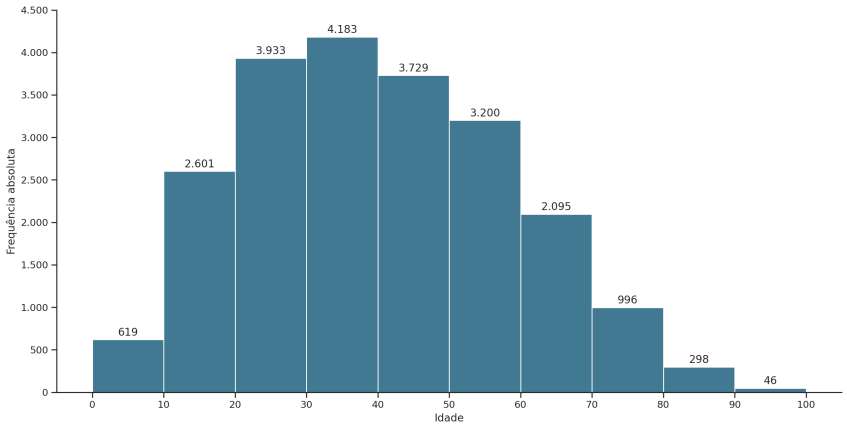


Fonte: Compilação do autor

Pode-se ver pela figura 10 que a raça parda foi a predominante sendo responsável por 13.542 de novos casos (62,1%). No entanto, apesar de um alto valor em relação as outras raças é verifica-se que a raça da população brasileira, a parda é a mais frequente, logo há maior chance de uma pessoa de cor parda ser frequente.

Como forma de verificar a caracterização da população afetada a idade importa bastante por se verificar se a doença atinge uma determinada população pela sua idade. A figura 11 mostrará a distribuição de casos pela faixa de idade dos pacientes.

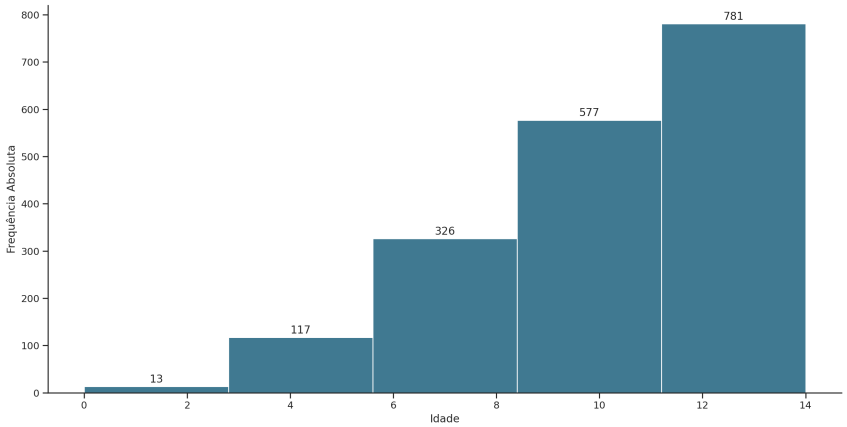
Figura 11 – Histograma da idade dos pacientes no período de 2001 a 2016



Fonte: Compilação do autor

Podemos ver que os casos aumentam e são maiores nas faixas de idade da fase adulta, sendo a mesma a fase economicamente ativa. Entretanto, quando se trata da doença hanseníase deve-se também se atentar aos casos em pessoas menores de 15 anos para saber se os mesmos também são afetados pela doença. A figura 12 mostra a distribuição de casos para pacientes com idade inferior a 15 anos.

Figura 12 – Histograma da idade dos pacientes menores de 15 anos

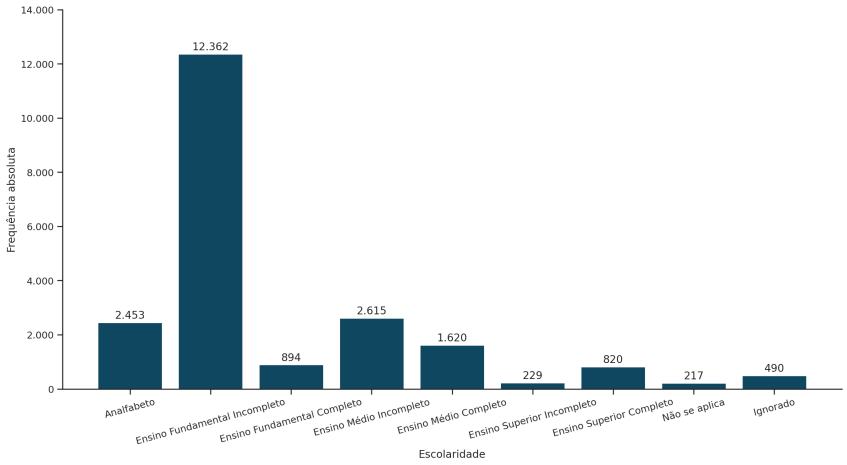


Fonte: Compilação do autor

Pode-se ver pela figura 12 que os casos da doença são maiores à medida que a idade do paciente é maior, sendo os pacientes de 11 a 14 anos (781 casos) possuindo 43,05% dos casos pela população menor de 15 anos e possuindo 3,6% dos casos gerais da hanseníase.

O atributo escolaridade também é um atributo bastante importante para a análise pois, através do mesmo podemo saber se a doença atinge pessoas com maior escolaridade ou não e se isso pode ter relação também com a vulnerabilidade econômica da população, a figura 13 mostra a distribuição pela escolaridade do paciente.

Figura 13 – Casos de hanseníase pela variável escolaridade

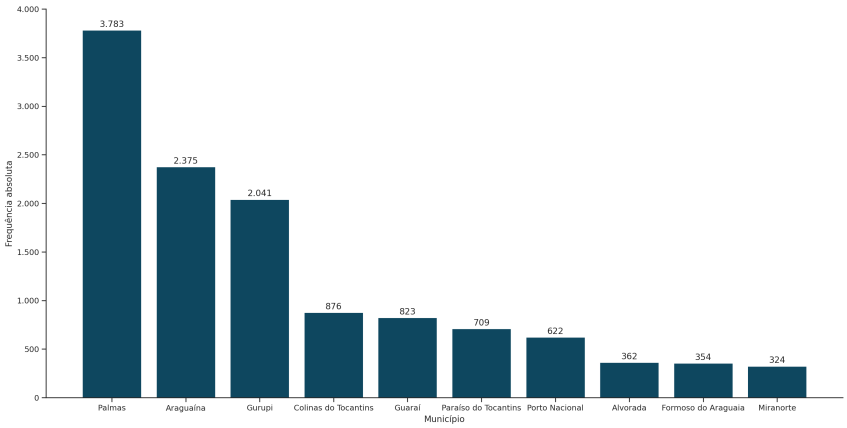


Fonte: Compilação do autor

4.1.1.2 Atributos de localização geográfica

A análise dos casos nos municípios é importante para saber em quais municípios os mesmos são mais afetados pela doença. Na figura 14, é possível ver os 10 municípios com a maior quantidade de casos novos de hanseníase nos períodos de 2001 a 2016.

Figura 14 – Casos de hanseníase nos 10 municípios com maiores número de casos no período de 2001 a 2016

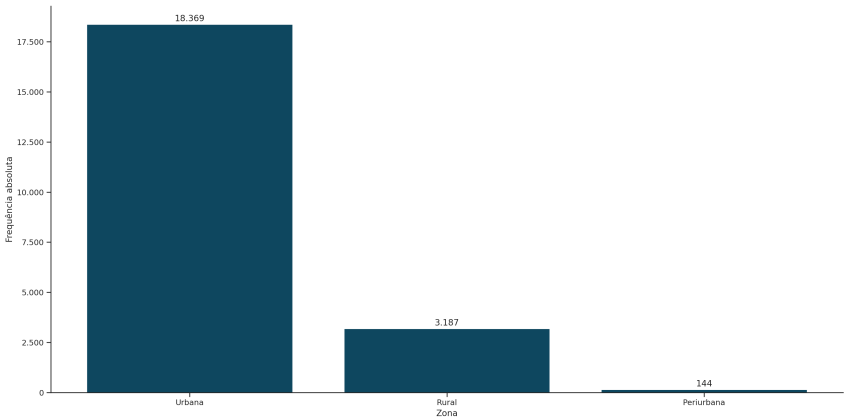


Fonte: Compilação do autor

Pode-se ver que as maiores quantidades de casos da doença estão localizados nas cidades com as maiores populações do estado, sendo o município de Palmas, como sendo a que mais possui aproximadamente (17.43%) dos casos novos.

Ao analisar os municípios que mais possuem número de casos, devemos ver também em quais tipos de zonas os pacientes são mais frequentes, e para isso na figura 15.

Figura 15 – Casos de hanseníase por zona



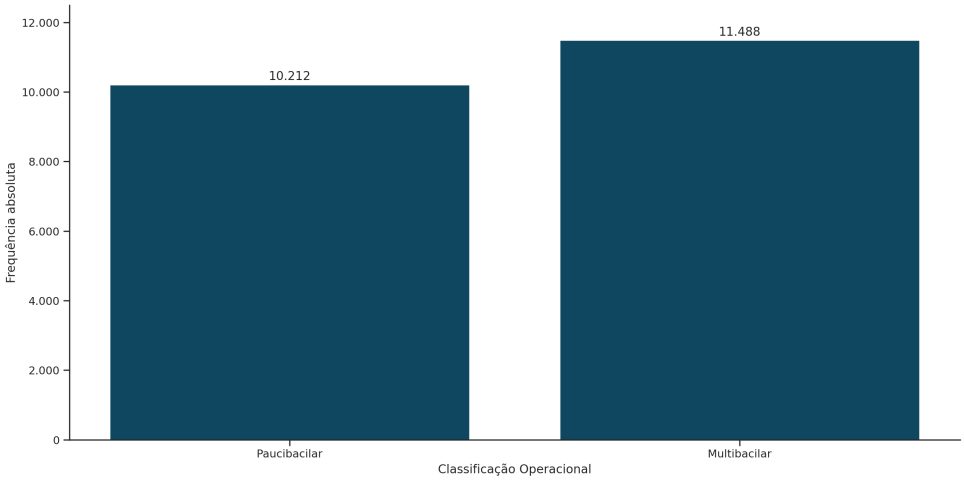
Fonte: Compilação do autor

Através da figura 15 é possível ver que a predominância de casos doença está localizada em sua maioria nas zonas urbanas na qual é responsável por 18.369 dos casos totais.

4.1.1.3 Atributos da doença

A classificação operacional dos casos de hanseníase são divididos entre paucibacilar e multibacilar, a imagem abaixo (figura 16) ilustra a quantidade de casos da doença entre essas duas classificações operacionais no período de 2001 a 2016.

Figura 16 – Classificação Operacional



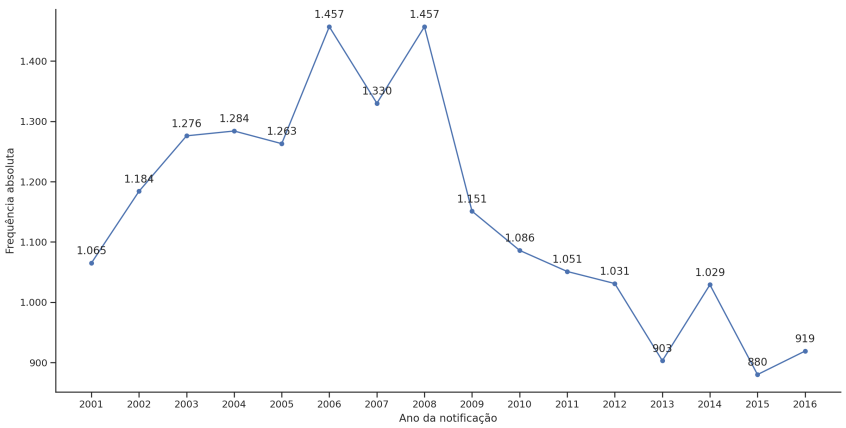
Fonte: Compilação do autor

Pode-se ver que o número de casos da doença é maior para a classificação operacional multibacilar, sendo composto por 11.488 novos casos (52,94%), ao passo que os casos paucibacilares são responsáveis por 47,06% dos casos.

4.1.1.4 Atributos temporais: ano do diagnóstico

Os casos da doença como já explicados anteriormente foram detectados com início de 2001 até o ano de 2016, logo deve-se expor a quantidade de casos referentes a esses anos. A figura 17 mostra o número de casos referentes a esses anos.

Figura 17 – Casos novos de hanseníase no período de 2001 a 2016



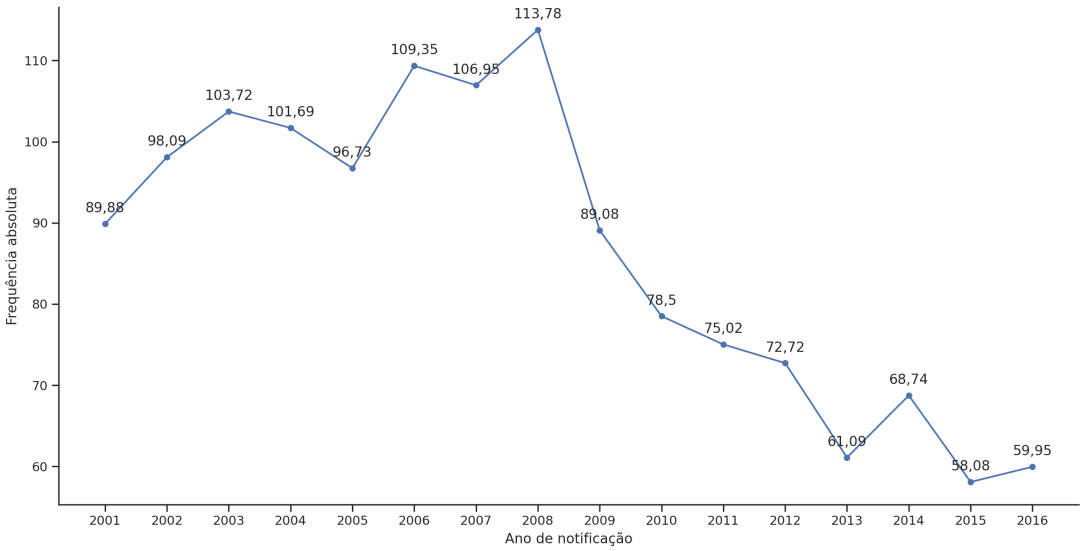
Fonte: Compilação do autor

É possível observar que os casos são maiores nos anos de 2006 e 2008, com 1.653 1.647 casos. Apesar de a exposição dos casos absolutos serem importantes para se ter uma noção da gravidade da situação em alguns casos ainda não é o suficiente para saber a gravidade das detecções em relação a população geral do estado.

Desta maneira, é necessário uma visualização da quantidade de casos através dos indicadores na qual calculam-se as taxas de detecção da hanseníase, de forma a verificar a gravidade da doença em razão da detecção da mesma. O processo foi realizado no período de tempo deste estudo que compreende aos anos de 2001 a 2016.

Os cálculos foram realizados de acordo com os parâmetros utilizados pela Organização Mundial da Saúde (OMS) e também pelo Ministério da Saúde. Os parâmetros para a detecção dos casos por ano são referentes às taxas de incidência de novos casos por 100.000 habitantes, como objetivo de medir a força da detecção de novos casos. A figura 18 apresenta a incidência dos novos casos de hanseníase por 100.000 habitantes para a população de idade em geral.

Figura 18 – Taxa de incidência para casos novos da hanseníase na população geral de por 100.000 habitantes no período de 2001 a 2016

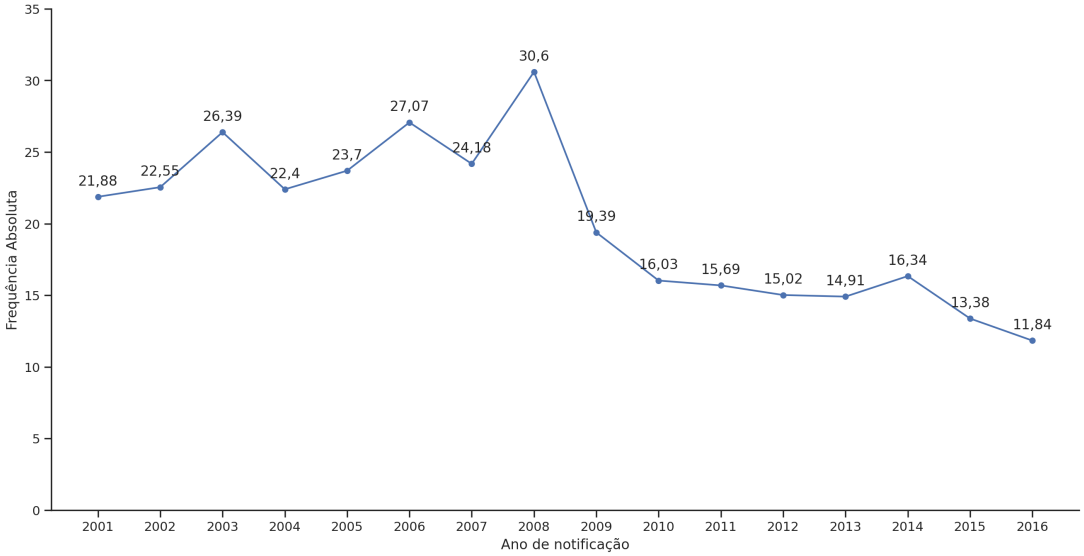


Fonte: Compilação do autor

Percebe-se que considerando os parâmetros da OMS, em todos do período estudado são considerados hiperendêmicos para os casos novos considerando os casos gerais (todas as idades), sendo a taxa de incidência superior a 40,0 (vide tabela 2).

Em se tratando da detecção de casos novos, é importante medir também a incidência de casos novos para pacientes de idades inferiores a 15 anos. A figura 19 ilustra as taxas de incidência da doença nos anos de estudo.

Figura 19 – Taxa de incidência para casos novos de hanseníase para idades menores de 15 anos por 100.000 habitantes no período de 2001 a 2016



Fonte: Compilação do autor

Já para os novos casos referentes aos pacientes com idades menores de 15 anos observa-se que todos os casos no período de 2001 a 2016 ultrapassam a taxa de incidência no valor de 10,0, logo, considera-se como hiperendêmico (vide tabela 3) também para essa população. É possível ver que a tendência de alta taxa de detecção é maior nos anos de 2006 e 2008, fato que também acontece quando se compara com a taxa de detecção geral de novos casos.

4.2 Mineração de dados por regras de associação

A metodologia para a mineração de dados através da geração de regras de associação foi realizado em pequenas etapas de forma a facilitar a metodologia de desenvolvimento. Primeiro a seleção dos atributos com características para a extração de padrões (tabela 9), transformação dos dados para entrada do algoritmo e por fim a filtragem das regras mais úteis verificando-se as restrições impostas.

4.2.1 Preparação inicial da base de dados

Para a realização da mineração de dados através das regras de associação havia a necessidade de selecionar apenas atributos categóricos referentes às características dos pacientes, da doença, e de localidade, logo apenas esses atributos serão utilizados no processo de geração de regras de associação, esses atributos são mostrados na tabela 9.

Tabela 9 – Atributos selecionados para mineração de regras de associação

Atributo	Descrição	Categoria	DBF
Sexo	Sexo do paciente	- Masculino - Feminino	CS_SEXO
Raça	Raça do paciente	- Branca - Preta - Parda - Amarela - Indígena - Ignorada	CS_RACA
Faixa de Idade	Idade do paciente na data do diagnóstico em intervalos de idade	- 0 a 9 anos - 10 a 14 anos - 15 a 29 anos - 30 a 49 anos - 50 a 74 anos - 75 a 100 anos	FAIXA_IDADE
Escolaridade	Escolaridade do paciente	- Analfabeto - Ensino Fundamental Incompleto - Ensino Fundamental Completo - Ensino Médio Incompleto - Ensino Médio Completo - Ensino Superior Incompleto - Ensino Superior Completo - Não se aplica - Ignorado	CS_ESCOL_N
Classificação operacional	Classificação operacional	- Paucibacilar - Multibacilar	CLASSOPERA

Continua na próxima página

Tabela 9 – Continuação da tabela 7

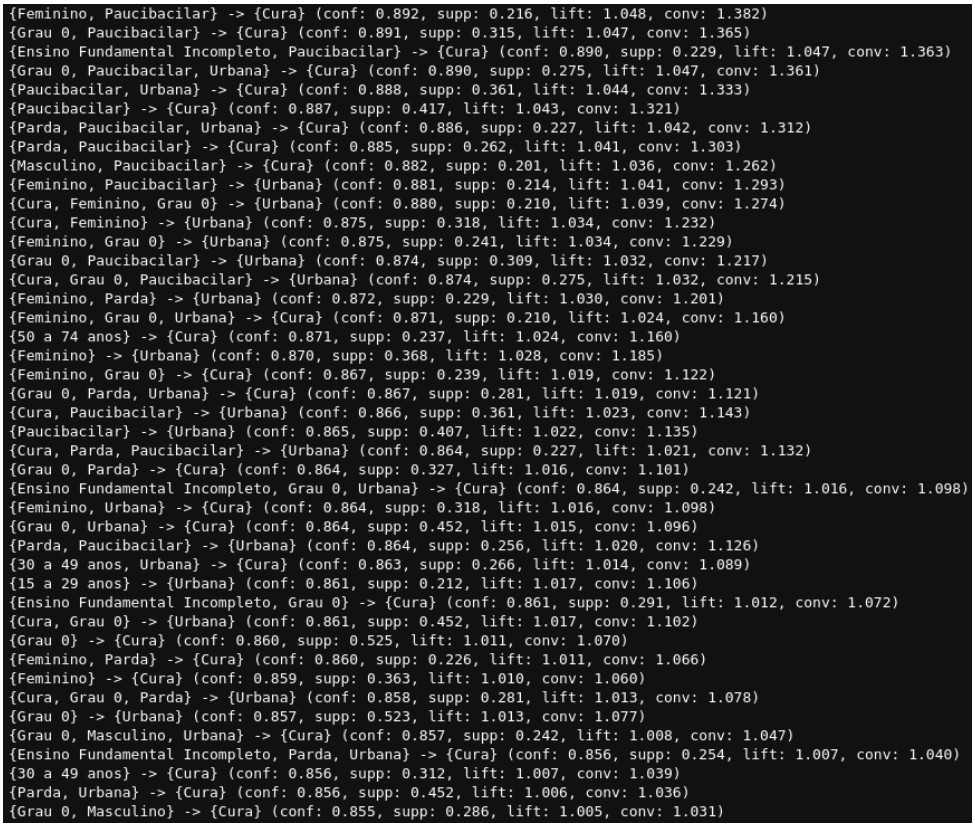
Atributo	Descrição	Categoria	DBF
Forma clínica	Forma clínica	- Indeterminada - Tuberculóide - Dimorfa - Virchowiana - Não avaliado	FORMACLINI
Avaliação de incapacidade física	Avaliação de incapacidade física no momento do diagnóstico	- Grau 0 - Grau I - Grau II - Não avaliado	AVALIA_N
Tipo de Saída	Tipo de saída do paciente	- Cura - Transferência para mesmo município - Transferência para outro estado - Transferência para outro país - Óbito - Abandono - Erro diagnóstico - Transferência não especificada	TPALTA_N

4.2.2 Extração das regras de associação

A etapa de geração das regras de associação é executada posteriormente a etapa de seleção dos atributos. Nesta etapa devem ser configurados os parâmetros das métricas que devem ser utilizados no algoritmo de forma a se obter os resultados de acordo com as necessidades esperadas.

As métricas utilizadas para a realização da tarefa são as seguintes: suporte mínimo 20% e confiança mínima de 50%, métricas já apresentadas na seção 2.3.2.1. Essa configuração foi realizada após a realização de alguns testes com valores de suporte e confiança maiores, sendo que nestes casos o número de regras se tornou bem menor. Após a configuração anterior dos parâmetros o algoritmo gerou um total de 111 regras, a figura 20 mostra o resultado das regras geradas pelo algoritmo apriori.

Figura 20 – Amostra da saída do algoritmo apriori



```
{Feminino, Paucibacilar} -> {Cura} (conf: 0.892, supp: 0.216, lift: 1.048, conv: 1.382)
{Grau 0, Paucibacilar} -> {Cura} (conf: 0.891, supp: 0.315, lift: 1.047, conv: 1.365)
{Ensino Fundamental Incompleto, Paucibacilar} -> {Cura} (conf: 0.890, supp: 0.229, lift: 1.047, conv: 1.363)
{Grau 0, Paucibacilar, Urbana} -> {Cura} (conf: 0.890, supp: 0.275, lift: 1.047, conv: 1.361)
{Paucibacilar, Urbana} -> {Cura} (conf: 0.888, supp: 0.361, lift: 1.044, conv: 1.333)
{Paucibacilar} -> {Cura} (conf: 0.887, supp: 0.417, lift: 1.043, conv: 1.321)
{Parda, Paucibacilar, Urbana} -> {Cura} (conf: 0.886, supp: 0.227, lift: 1.042, conv: 1.312)
{Parda, Paucibacilar} -> {Cura} (conf: 0.885, supp: 0.262, lift: 1.041, conv: 1.303)
{Masculino, Paucibacilar} -> {Cura} (conf: 0.882, supp: 0.201, lift: 1.036, conv: 1.262)
{Feminino, Paucibacilar} -> {Urbana} (conf: 0.881, supp: 0.214, lift: 1.041, conv: 1.293)
{Cura, Feminino, Grau 0} -> {Urbana} (conf: 0.880, supp: 0.210, lift: 1.039, conv: 1.274)
{Cura, Feminino} -> {Urbana} (conf: 0.875, supp: 0.318, lift: 1.034, conv: 1.232)
{Feminino, Grau 0} -> {Urbana} (conf: 0.875, supp: 0.241, lift: 1.034, conv: 1.229)
{Grau 0, Paucibacilar} -> {Urbana} (conf: 0.874, supp: 0.309, lift: 1.032, conv: 1.217)
{Cura, Grau 0, Paucibacilar} -> {Urbana} (conf: 0.874, supp: 0.275, lift: 1.032, conv: 1.215)
{Feminino, Parda} -> {Urbana} (conf: 0.872, supp: 0.229, lift: 1.030, conv: 1.201)
{Feminino, Grau 0, Urbana} -> {Cura} (conf: 0.871, supp: 0.210, lift: 1.024, conv: 1.160)
{50 a 74 anos} -> {Cura} (conf: 0.871, supp: 0.237, lift: 1.024, conv: 1.160)
{Feminino} -> {Urbana} (conf: 0.870, supp: 0.368, lift: 1.028, conv: 1.185)
{Feminino, Grau 0} -> {Cura} (conf: 0.867, supp: 0.239, lift: 1.019, conv: 1.122)
{Grau 0, Parda, Urbana} -> {Cura} (conf: 0.867, supp: 0.281, lift: 1.019, conv: 1.121)
{Cura, Paucibacilar} -> {Urbana} (conf: 0.866, supp: 0.361, lift: 1.023, conv: 1.143)
{Paucibacilar} -> {Urbana} (conf: 0.865, supp: 0.407, lift: 1.022, conv: 1.135)
{Cura, Parda, Paucibacilar} -> {Urbana} (conf: 0.864, supp: 0.227, lift: 1.021, conv: 1.132)
{Grau 0, Parda} -> {Cura} (conf: 0.864, supp: 0.327, lift: 1.016, conv: 1.101)
{Ensino Fundamental Incompleto, Grau 0, Urbana} -> {Cura} (conf: 0.864, supp: 0.242, lift: 1.016, conv: 1.098)
{Feminino, Urbana} -> {Cura} (conf: 0.864, supp: 0.318, lift: 1.016, conv: 1.098)
{Grau 0, Urbana} -> {Cura} (conf: 0.864, supp: 0.452, lift: 1.015, conv: 1.096)
{Parda, Paucibacilar} -> {Urbana} (conf: 0.864, supp: 0.256, lift: 1.020, conv: 1.126)
{30 a 49 anos, Urbana} -> {Cura} (conf: 0.863, supp: 0.266, lift: 1.014, conv: 1.089)
{15 a 29 anos} -> {Urbana} (conf: 0.861, supp: 0.212, lift: 1.017, conv: 1.106)
{Ensino Fundamental Incompleto, Grau 0} -> {Cura} (conf: 0.861, supp: 0.291, lift: 1.012, conv: 1.072)
{Cura, Grau 0} -> {Urbana} (conf: 0.861, supp: 0.452, lift: 1.017, conv: 1.102)
{Grau 0} -> {Cura} (conf: 0.860, supp: 0.525, lift: 1.011, conv: 1.070)
{Feminino, Parda} -> {Cura} (conf: 0.860, supp: 0.226, lift: 1.011, conv: 1.066)
{Feminino} -> {Cura} (conf: 0.859, supp: 0.363, lift: 1.010, conv: 1.060)
{Cura, Grau 0, Parda} -> {Urbana} (conf: 0.858, supp: 0.281, lift: 1.013, conv: 1.078)
{Grau 0} -> {Urbana} (conf: 0.857, supp: 0.523, lift: 1.013, conv: 1.077)
{Grau 0, Masculino, Urbana} -> {Cura} (conf: 0.857, supp: 0.242, lift: 1.008, conv: 1.047)
{Ensino Fundamental Incompleto, Parda, Urbana} -> {Cura} (conf: 0.856, supp: 0.254, lift: 1.007, conv: 1.040)
{30 a 49 anos} -> {Cura} (conf: 0.856, supp: 0.312, lift: 1.007, conv: 1.039)
{Parda, Urbana} -> {Cura} (conf: 0.856, supp: 0.452, lift: 1.006, conv: 1.036)
{Grau 0, Masculino} -> {Cura} (conf: 0.855, supp: 0.286, lift: 1.005, conv: 1.031)
```

Apesar de haver bastante regras as mesmas devem ser organizadas de forma que fique mais fácil e organizado de se analisar, dessa maneira a tabela 10 ilustra as 11 primeiras regras encontradas, de maneira mais visual mostrando apenas as regras do antecedente, consequente e das métricas de suporte e confiança.

Tabela 10 – Amostra das 10 primeiras regras de associação encontradas da figura 20

Regra	Antecedente	Consequente	Confiança	Suporte
0	Indeterminada, Urbana	Cura	88.34%	21.34%
1	Indeterminada	Cura	88.01%	24.93%
2	Cura, Feminino, Grau 0	Urbana	87.95%	21.0%
3	Cura, Feminino	Urbana	87.54%	31.76%
4	Feminino, Grau 0	Urbana	87.51%	24.1%
5	Feminino, Parda	Urbana	87.22%	22.9%
6	Feminino, Grau 0, Urbana	Cura	87.13%	21.0%
7	Feminino	Urbana	87.05%	36.76%
8	Feminino, Grau 0	Cura	86.69%	23.87%
9	Grau 0, Parda, Urbana	Cura	86.68%	28.05%
10	Grau 0, Parda	Cura	86.43%	32.71%

Percebe-se que há regras com um grau de confiança alto que foram responsáveis por atingir um valor máximo confiança de 88,00%, o que indica que existem regras de associação com confianças altas.

4.2.3 Filtragem das regras de associação

Normalmente, a geração de regras de associação podem acarretar na geração de muitas regras óbvias, mesmo que possuam um valor alto para a confiança e suporte. Dessa maneira é necessário fazer uma intervenção manual de forma a verificar quais as regras que possam ser mais relevantes.

Esta parte se refere ao pós processamento no qual a escolha das regras mais úteis serão responsáveis pela geração do conhecimento. Muitas dessas regras que são consideradas óbvias apesar de não serem responsáveis pela geração de conhecimento novo, também podem servir para a corroboração dos resultados gerados através da análise exploratória de dados, entretanto em primeiro lugar buscaremos as que possuem conhecimento novo em primeiro lugar.

De acordo com o trabalho da (KOBUS, 2006) muitas regras de associação possuíam valores mais baixos de suporte e confiança entretanto, para os especialistas que possuíam o conhecimento sobre o seu tema de trabalho, algumas dessas regras poderiam ser importantes, logo devemos não apenas nos ater ao valores dos parâmetros de suporte e confiança, mas sim a importância que uma regra pode ter em seu contexto.

Tabela 11 – Regras de associação significativas após a filtragem dos dados

Regra	Antecedente	Consequente	Confiança	Suporte
1	Urbana	Cura	85,40%	72,30%
2	Parda	Cura	85,20%	53,20%
3	Grau 0	Cura	86,00%	52,50%
4	Grau 0	Urbana	85,70%	52,50%
5	Masculino	Cura	85,00%	48,80%
6	Ensino Fundamental Incompleto	Cura	85,00%	48,40%
7	Multibacilar	Cura	81,88%	43,30%
8	Paucibacilar	Cura	88,70%	41,70%
9	Feminino	Cura	85,90%	36,30%
10	30 a 49 anos	Cura	85,90%	31,20%
11	30 a 49 anos	Urbana	84,70%	30,90%
12	50 a 74 anos	Cura	87,10%	23,70%
13	15 a 29 anos	Cura	82,40%	20,30%

As regras de associação mostradas na tabela 11, apesar de estarem tabuladas e mais organizadas precisam ser interpretadas, dessa maneira as mesmas são apresentada s desta maneira:

- Regra 1: 72,30% dos pacientes afetados moram na região ”Urbana”, desses 85,40% receberam a alta por ”Cura”;
- Regra 2: 53,20% dos pacientes afetados são da raça ”Parda”, desses 85,20% receberam a alta por ”Cura”;

- Regra 3: 52,50% dos pacientes afetados que foram diagnosticados com incapacidades físicas de ”Grau 0”, desses 86,00% receberam a alta por ”Cura”;
- Regra 4: 52,50% dos pacientes afetados que foram diagnosticados com incapacidades físicas de ”Grau 0”, 85,70% são residentes de zona ”Urbana”;
- Regra 5: 48,80% dos pacientes afetados que são do sexo ”Masculino”, 85,00% são receberam a alta por ”Cura”;
- Regra 6: 48,40% dos pacientes afetados que possuem escolaridade ”Ensino Fundamental Incompleto”, 85,00% são receberam a alta por ”Cura”;
- Regra 7: 43,30% dos pacientes afetados no qual possuem a classificação operacional ”Multibacilar”, desses 81,88% receberam a alta por ”Cura”;
- Regra 8: 41,70% dos pacientes afetados no qual possuem a classificação operacional ”Paucibacilar”, desses 88,70% receberam a alta por ”Cura”;
- Regra 9: 36,30% dos pacientes afetados que são do sexo ”Feminino”, desses 85,90% receberam a alta por ”Cura”;
- Regra 10: 31,20% dos pacientes afetados que possuem idade entre ”30 a 49 anos”, desses 85,60% receberam a alta por ”Cura”;
- Regra 11: 30,90% dos pacientes afetados que possuem idade entre ”30 a 49 anos”, desses 84,70% são residentes de zona ”Urbana”;
- Regra 12: 23,70% dos pacientes afetados que possuem idade entre ”50 a 74 anos”, desses 87,10% receberam a alta por ”Cura”;
- Regra 13: 20,30% dos pacientes afetados que possuem idade entre ”15 a 29 anos”, desses 82,40% receberam a alta por ”Cura”;

Através das regras apresentadas acima vemos que as regras 1, 2 e 3, 6 no que se refere à população mais afetada pela doença são das características de ser residentes em zonas urbanas, são de raça parda e são pessoas que possuem grau de incapacidade física de grau 0.

Em relação a classificação operacional no momento da diagnóstico do paciente vemos através da regra 7 e 8 que para se ter alta por cura há quase os mesmo valores de suporte e confiança. A característica em relação ao sexo do paciente também é levada em conta, vendo-se através das regras 5 e 9, no qual a população masculina possui mais chance de receber a alta por cura, mesmo sendo a população com maior número de casos.

5 DISCUSSÃO DOS RESULTADOS

Esta seção é responsável pela discussão dos resultados obtidos no capítulo 4. Os resultados a serem discutidos são os referentes à análise exploratória de dados utilizada na base de dados e por fim a discussão dos resultados da mineração de dados através das regras de associação geradas.

A partir dos resultados extraídos da base de dados através da análise exploratória, é possível ver que a classificação operacional mais presente foi a multibacilar sendo responsável por 52,94% ao passo de a paucibacilar possui de 47,06% dos casos, o que significa que a detecção dos casos está sendo feito de maneira tardia, acarretando na evolução da doença e possivelmente no acometimento da população a possíveis sequelas provocadas pela doença.

Nos períodos compreendidos de 2001 a 2016 os casos da doença foram maiores no ano de 2006 e 2008, o que pode ser também compreendido não apenas como um momento súbito de casos, como um pico de altas infecções, mas também pode ser analisado por ocasião das ações de saúde mais terem sido mais bem intensificadas neste anos de forma a aumentar a detecção dos casos.

A utilização dos coeficientes de detecção dos casos nos indica a gravidade da detecção da doença, de acordo com os parâmetros do ministério da saúde, sendo que os coeficientes de detecção foram todos superiores a 40,00 (hiperendêmico) para os casos gerais em todos os anos estudados, já para os pacientes com idades inferiores a 15 anos os coeficientes ultrapassaram o valor de 10,00 (hiperendêmico).

Em relação às características dos pacientes acometidos pela doença, vemos que em relação ao sexo a população masculina é a mais afetada, sendo que também é a que mais possui casos multibacilares, indicando que essa população é a mais afetada pelas formas mais avançadas da doença.

É possível ver que através das regras de associação que às chances de cura de acordo com a classificação operacional no paciente no momento do diagnóstico são bastante próximos (regras 7 e 8) no que tange ao suporte e confiança dos mesmos, logo, isso possivelmente se dá pelo fato de que se o tratamento for realizado nos pacientes a probabilidade de frear a propagação da doença e a possível de cura possuem chances próximas para ambas as classificações.

Em relação à idade, percebe-se que a população com idade de 20 anos a 60 anos são os mais são afetados, sendo que são considerados como a população mais ativa economicamente. Já entre a população dos pacientes inferiores a 15 anos, os número de casos é maior conforme aumenta a idade. As regras de associação 10, 12 e 13 nos mostra que a população de 15 a 74 anos são as que mais recebem a alta por cura, também pelo fato de serem a população mais afetada de acordo com os gráficos apresentados neste trabalho,

logo se tiverem o devido tratamento, as mesmas possuem a chance de receber a alta por cura.

Em se tratando da localidade dos pacientes vemos que a zona urbana é predominante, e isso pode ser confirmado através dos gráficos gerados na análise exploratória e através das regras geradas, no qual o consequente da regra são pessoas que se estabelecem em zonas urbanas. Logo a hanseníase é uma doença de característica de pessoas que se estabelecem em zonas urbanas.

6 CONCLUSÃO

O presente trabalho possuiu os objetivos de extrair conhecimento da base de dados do Sinan para os casos de hanseníase do estado do Tocantins através da geração de regras de associação, de forma a caracterizar as pessoas afetadas para que seja possível através dos resultados caracterizar o perfil da população afetada, e dessa maneira, buscar soluções para a tomada de decisão no que se refere a soluções ou campanhas com intuito de aumentar as detecções e realizar o tratamento prematuro da doença principalmente para esses perfis mais afetados.

Os processos metodológicos deste trabalho foram iniciados através da coleta de dados, ao pré processamento de dados no qual foram realizados diversas etapas, após a análise exploratória de dados foi responsável pela geração das visualizações e explorações dos atributos presentes mais importantes da base de dados. Por conseguinte, através da mineração de dados através das regras de associação foi possível verificar às ocorrência frequentes dos atributos com o objetivo de traçar um perfil dos pacientes afetados pela doença, e por fim foi realizado a avaliação e apresentação dos resultados.

Os desafios encontrados no desenvolvimento deste trabalho foram relativos mais ao pré processamento de dados, por se tratar de uma etapa muito demorada, definido por muitos autores como o processo que pode despender mais de 70% de todo o tempo do projeto. A base de dados apesar de possuir atributos bastante importantes, possuía muitos problemas relacionados a falta de preenchimento de atributos que poderiam ser importantes para a mineração de dados, dessa forma ocasionando na falta de novas informações para tentar explorar.

Através dos resultados gerados a partir deste trabalho a contribuição que pode ser dada é que existe perfis de pacientes que são afetados pela doença e esse perfil deve ser colocado com maior prioridade de forma a aumentar o número de detecções de forma a evitar a evolução da doença. Normalmente as pessoas residentes nos centros urbanos, que possuem em sua maioria uma baixa escolaridade, e por consequência também são bastante relacionados às populações de baixa renda.

A relação entre casos da doença por sexo do paciente é mais frequente na população masculina, entretanto é quase equiparável ao total de casos população feminina, exceto pelo fato que aquela é mais suscetível a ser afetado pelas formas multibacilares da doença. Logo, as campanhas de detecção devem ser direcionados para esse tipo de população afetada, de forma a se detectar mais casos da doença para evitar a detecção tardia e a evolução dos casos para formas mais complexas.

As cidades que consequentemente possuem a maior população, possivelmente são as que possuem maior cobertura de hospitais e postos de saúde espalhados em seus territórios, logo deve ser onde devem aumentar as campanhas de conscientização para que as pessoas

possam de forma espontânea buscar os meios para o tratamento. Os trabalhos futuros que são possíveis de serem realizados neste trabalho são as seguintes:

- Uso de algoritmos de outras tarefas de mineração de dados de forma a encontrar novos conhecimentos acerca da base de dados, tais como a clusterização para os atributos de localização do pacientes;
- Integração das bases de dados de outros estados de forma a se verificar casos de transferência de outros estados e de abandono de tratamento para saber as possíveis causas para os mesmos, e os perfis da população que são sujeitos a isso.

Por fim, o trabalho se dispôs a buscar a traçar o perfil da população afetada pela doença de forma a identificá-las tanto através da análise exploratória de dados como pela apresentação das regras de associação, espera-se que o presente trabalho possa ter uma contribuição para ajudar na identificação da população mais vulnerável a essa doença para auxiliar os gestores de saúde pública na tomada de decisão e no direcionamento das ações a esse grupo na busca pelo controle da doença.

REFERÊNCIAS

AGRAWAL, R.; IMIELI;SKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. In: **SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data**. New York, NY, USA: ACM Press, 1993. p. 207–216. ISBN 0-89791-592-5. Disponível em: <http://cs.sungshin.ac.kr/~jpark/HOME/References/agrawal_sigmod93.ps>.

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules. IBM, 1994. Disponível em: <http://www.vldb.org/conf/1994/P487.PDF>.

ALPHA, A. H. A. **What is Public Health?** 2020. Disponível em: <https://www.apha.org/What-is-Public-Health>.

ANGéLICA, V. T. B. **A SAÚDE PÚBLICA NO BRASIL: UM BREVE RESGATE HISTÓRICO [1500-1990]**. 2020. Disponível em: <https://www.denem.org.br/wp-content/uploads/2020/05/A-SAÚDE-PÚBLICA-NO-BRASIL-1500-1990-DENEM-2020.pdf>.

BAPTISTA, T. **O direito à saúde no Brasil:sobre como chegamos ao sistema único de saúde e o que esperamos dele**. [S.l.]: Fiocruz, 2005.

BHATIA, P. **Data Mining and Data Warehousing: Principles and Practical Techniques**. [S.l.]: Cambridge Univesity Press, 2019.

BLAIEWICZ, J. et al. **Handbook on Data Management In Information Systems**. [S.l.]: Springer, 2003.

BRASIL, B. M. da Saúde. Secretaria de Vigilância em Saúde. Departamento de V. E. **Sistema de Informação de Agravos de Notificação – Sinan: normas e rotinas**. 2006. Disponível em: <http://bvsms.saude.gov.br/bvs/publicacoes/sistema_informacao_agravos_notificacao_sinan.pdf>.

BRASIL, M. da Saúde. Secretaria de Vigilância em saúde. Departamento de vigilância epidemiológica. **Guia de vigilância epidemiológica**. 2009. Disponível em: <https://bvsms.saude.gov.br/bvs/publicacoes/origem_politicas_saude_publica_brasil.pdf>.

BRAVO, M. I. S. **Política de Saúde no Brasil**. 2009. Disponível em: <http://www.fnepas.org.br/pdf/servico_social_saude/texto1-5.pdf>.

BRUHA, I.; FAMILI, A. F. **Postprocessing in Machine Learning and Data Mining**. 2000. Disponível em: <https://www.kdd.org/exploration_files/KDD2000PostWkshp.pdf>.

CARVALHO, G. A saúde pública no brasil. **Estudos Avançados**, scielo, v. 27, p. 7 – 26, 00 2013. ISSN 0103-4014. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-40142013000200002&nrm=iso>.

CEA, W. The untilled field of public health. **Science**, v. 51, p. 23–33, 1920.

CERVO, A. L.; BERVIAN, P. A. **Metodologia Científica**. fifth. [S.l.]: Universidade Feevale, 2002.

CIMERMAN, S.; CIMERMAN, B. **Medicina Tropical**. [S.l.]: Atheneu, 2003.

CONTROL, C. for D.; CDC, P. **Introduction to Public Health. In: Public Health 101 Series**. 2014. Disponível em: <https://www.cdc.gov/publichealth101/public-health.html>.

CRUZ, M. M. da. **Histórico do sistema de saúde, proteção social e direito à saúde**. 2. ed. [S.l.]: Fiocruz, 2011.

DIETRICH, D.; HELLER, B.; YANG, B. **Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data**. [S.l.]: Wiley, 2015.

DUNKEL, B. et al. Systems for kdd: From concepts to practice. **Elsevier**, v. 13, p. 231–242, 1997.

FALEIROS, V. et al. **A construção do SUS: história da reforma sanitária e do processo participativo**. [S.l.]: Ministério da Saúde. Brasil, 2006.

FARIA, J. L. D. **Patologia geral : fundamentos das doencas, com aplicacoes clinicas**. fourth. [S.l.]: GEN - Guanabara Koogan, 2003.

FAYYAD, U.; STOLORZ, P. Data mining and kdd: Promise and challenges. **Elsevier Science**, Elsevier Science, Microsoft Research, One Microsoji Way, Redmond, WA 98052, USA, v. 13, n. 2, abr. 1997. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167739X97000150>.

FAYYAD, U.; STOLORZ, P. Data mining and kdd: Promise and challenges. **Elsevier Science**, Elsevier Science, Microsoft Research, One Microsoji Way, Redmond, WA 98052, USA, v. 13, n. 2, p. –1, abr. 1997. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167739X97000150>.

FERREIRA, C. T. M. R. M. G. R. A. P. A. L. R. M. L. da C. R. J. E. S. O sistema público de saúde e as ações de reabilitação no brasil. **Scielo**, scielo, 07 2009. ISSN 0034-8910. Disponível em: <https://scielosp.org/article/rpsp/2010.v28n1/43-48/>.

FOSS, N. et al. **Hanseníase: Episódios Reacionais**. [S.l.], 2004.

GALVãO, D. M. A. M. **ORIGEM DAS POLÍTICAS DE SAÚDE PÚBLICA NO BRASIL: DO BRASIL-COLÔNIA A 1930**. 2010. Disponível em: <https://bvsms.saude.gov.br/bvs/publicacoes/origem_politicas_saude_publica_brasil.pdf>.

GOLDSCHMIDT, R.; PASSOS, E. **Data Mining: um Guia Prático**. [S.l.]: Editora Campus Elsevier, 2005.

HAMAD, M.; QADER, B. A. Data pre-processing for knowledge discovery. Research Gate, 2014. Disponível em: <https://www.researchgate.net/publication/319019536_Data_Pre-processing_for_knowledge_discovery>.

HAN, J.; KAMBER, M.; PEI, J. **Data mining concepts and techniques, third edition**. Waltham, Mass.: Morgan Kaufmann Publishers, 2012. Disponível em: <http://www.amazon.de/Data-Mining-Concepts-Techniques-Management/dp/0123814790/ref=tmm_hrd_title_0?ie=UTF8&qid=1366039033&sr=1-1>.

JF, C. História da psiquiatria do brasil: um corte ideológico. In: _____. [S.l.: s.n.], 1989.

KOBUS, L. S. G. **Aplicação da descoberta de conhecimento em bases de dados para identificação de usuários com doenças cardiovasculares elegíveis para programas de gerenciamento de caso**. Dissertação (Mestrado) — Pontifícia Universidade Católica do Paraná, Brasil, 2006.

KOTSIANTIS, S.; KANELLOPOULOS, D. Association rules mining: A recent overview. GESTS International Transactions on Computer Science and Engineering, p. 71–82, 2006. Disponível em: <https://pdfs.semanticscholar.org/73a1/9026fb8a6ef5bf238ff472f31100c33753d0.pdf>.

LAGUARDIA, J. et al. Sistema de informação de agravos de notificação (sinan): desafios no desenvolvimento de um sistema de informação em saúde. Epidemiologia e Serviços de Saúde, v. 13, n. 3, 2004.

LASTóRIA, J. C.; ABREU, M. A. M. M. de. Leprosy: review of the epidemiological, clinical, and etiopathogenic aspects - part 1. **Scielo**, p. 209, apr 2014.

LSF, B. Sistema de informações de agravos de notificação - sinan. In: _____. Brasília, DF: [s.n.], 1993. p. 145 – 146.

MICHIE, D. J.; TAYLOR, C. C. **Machine learning, neural and statistical classification**. [S.l.: s.n.], 1994.

P.A., P. R.; B., K.-F. B. G. An interactive approach for the post-processing in a kdd process. **IFIP Advances in Information and Communication Technology**, Springer, Berlin, Heidelberg, Laboratoire Génie de Production / INP-ENIT - Université de Toulouse 47, Avenue d'Azereix, BP 1629, F-65016 Tarbes Cedex – France, v. 348, n. 1, p. 94, 2014. ISSN 978,-3-662-44738-3. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-662-44739-0_12>.

PAIM, J. S. **O que é o SUS**. 2015. Disponível em: <http://www.livrosinterativoseditora.fiocruz.br/sus/>.

PRODANOV, C. C.; FREITAS, E. C. de. **Metodologia do trabalho científico: Métodos e Técnicas da Pesquisa e do Trabalho Acadêmico**. second. Campus I: Av. Dr. Maurício Cardoso, 510 – CEP 93510-250 – Hamburgo Velho – Novo Hamburgo – RS: Universidade Feevale, 2013.

RAHMAN, F. A. et al. Data cleaning in knowledge discovery database-data mining (kdd-dm). Research Gate, 2019. Disponível em: <https://www.researchgate.net/publication/338116654_Data_Cleaning_in_Knowledge_Discovery_Database-Data_Mining_KDD-DM>.

RAMAGERI, B. M. **Data Mining techniques and applications**. [S.l.]: Indian Journal of Computer Science and Engineering, 2015.

RHODES, P.; BRYANT, J. H. **Public health**. 2011. Disponível em: <https://www.britannica.com/topic/public-health>.

RIBEIRO, C. T. M. et al. O sistema público de saúde e as ações de reabilitação no brasil. **Scielo**, Ago 2009.

RUIZ, P. A. P.; KAMSU-FOGUEM, B.; GRABOT, B. **Advances in Production Management Systems. Innovative and Knowledge-Based Production Management in a Global-Local World. IFIP Advances in Information and Communication Technology**. [S.l.]: Springer, 2014.

SAúDE, S. de Vigilância em Saúde. Departamento de Vigilância das Doenças Transmissíveis. Ministério da. **Guia prático sobre a hanseníase**. 2017. Disponível em: <https://portalarquivos2.saude.gov.br/images/pdf/2017/novembro/22/Guia-Pratico-de-Hanseniase-WEB.pdf>.

SERVIçOS, B. M. da Saúde. Secretaria de Vigilância em Saúde. Coordenação-Geral de Desenvolvimento da Epidemiologia em. **Guia de Vigilância em Saúde**. 2019. Disponível em: <http://bvsms.saude.gov.br/bvs/publicacoes/guia_vigilancia_saude_3ed.pdf>.

SINAN, P. do. **O Sinan**. 2016. <http://portalsinan.saude.gov.br/o-sinan>. Último acesso em 20/09/2020.

TAN, P. ning; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. [S.l.]: Pearson, 2006.

TAVARES, W.; MARINHO, L. A. C. **Rotinas de diagnóstico e tratamento das doenças infecciosas e parasitárias**. 2. ed. [S.l.]: Atheneu, 2007.

TWF, B. O direito à saúde no brasil: sobre como chegamos ao sistema único de saúde e o que esperamos dele. In: _____. [S.l.: s.n.], 2005. p. 11–41.

URA, S. Tratamento e controle das reaÇÕes hansÊnicas. **Hansenologia Internationalis**, p. 67–70, 2007.

USA, I. of M. **O futuro da saúde pública**. 1988. Disponível em: <https://www.britannica.com/topic/public-health>.

USAMA, F.; GREGORY, P.-S.; PADHRAIC, S. From data mining to knowledge discovery in databases. **AI Magazine Volume 17 Number 3 (1996) (© AAAI)**, Kluwer Academic Publishers, Hingham, MA, USA, v. 17, n. 3, p. 38–54, jun. 1997. ISSN 1022-0038. Disponível em: <http://dx.doi.org/10.1007/s11276-006-0724-8>.

USUALDO, D. G. **Investigação de regressão no processo de mineração de dados**. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação, 2003.

WHO, W. H. O. **Global leprosy update, 2016:accelerating reduction of disease burden**. [S.l.]: World Health Organization, 2017. 501-520 p.