UNIVERSIDADE FEDERAL DO TOCANTINS
CÂMPUS UNIVERSITÁRIO DE PALMAS
CURSO DE CIÊNCIA DA COMPUTAÇÃO

ANÁLISE DE DADOS NA SAÚDE PÚBLICA: UM ESTUDO DE CASO DA HANSENÍASE NO ESTADO DO TOCANTINS

CASSIA GABRIELA SILVA PEREIRA

PALMAS (TO)
2021

CASSIA GABRIELA SILVA PEREIRA

ANÁLISE DE DADOS NA SAÚDE PÚBLICA: UM ESTUDO DE CASO DA 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

CASSIA GABRIELA SILVA PEREIRA

ANÁLISE DE DADOS NA SAÚDE PÚBLICA: UM ESTUDO DE CASO DA 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: 21 / 5 / 2021

Banca Examinadora:

Prof. Dr. Ary Henrique Morais de Oliveira

Profa. Dra. Valeria Perim

Profa. Dra. Anna Paula Parente

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

P436a Pereira, Cassia Gabriela Silva.

Análise de dados na saúde pública: um estudo de caso da hanseníase no estado do Tocantins. / Cassia Gabriela Silva Pereira. – Palmas, TO, 2021.

83 f.

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

Orientador: Ary Henrique Morais De Oliveira

1. Data Mart. 2. Mineração de Dados. 3. Teorema de Bayes. 4. Hanseníase. I. Título

CDD 004

Elaborado pelo sistema de geração automatica de ficha catalográfica da UFT com os dados fornecidos pelo(a) autor(a).

Para minha mãe e minha vó

AGRADECIMENTOS

Gostaria de agradecer a minha mãe, Maria Inês Silva, por todo apoio nessa caminhada.

A minha vó, Maria Luzia, pelo incentivo em cada momento.

Ao meus irmãos, Maisa Graziele e Thiago Gabriel, pela compreensão da minha ausência durante esse trajeto.

Ao meu orientador Ary Henrique, que mesmo sempre de agenda cheia me dava todo auxílio para que eu pudesse concluir este projeto.

Ao meu amor, Klesley Gonçalves, pelo apoio e ajuda durante o curso.

E por fim, agradeço a todas as pessoas que fizeram parte dessa caminhada.

RESUMO

No estado do Tocantins a hanseníase é considerada hiperendêmica de acordo com os critérios do Ministério da Saúde. Em razão do elevado índice de casos, a doença permanece como um dos problemas de saúde pública no estado. As organizações gestoras da saúde pública estão buscando cada vez mais aumentar a eficiência na gestão de saúde, com o objetivo de obter informações para usar como base no controle de epidemias. Diante disso, este projeto tem por objetivo desenvolver uma proposta de *data mart* onde seja possível ser aplicado técnicas de estatísticas e mineração de dados.

As fases executadas neste trabalho são: o pré-processamento, a criação de um *data mart*, análise exploratória, cálculo dos indicadores, inferência com teorema de Bayes e a classificação de dados. Analisando os resultados, foi possível deduzir algumas probabilidades e destacar características de pacientes com maiores influências para terem a classificação operacional como multibacilar. Conclui-se que os objetivos do trabalho foram alcançados, e foram propostos trabalhos futuros na área da saúde pública.

Palavra-chave: Data Mart. Teorema de Bayes. Hanseníase. Mineração de Dados. Saúde Pública.

ABSTRACT

In the state of Tocantins, leprosy is considered hyperendemic according to the criteria of the Ministry of Health. Due to the high rate of cases, the disease remains one of the public health problems in the state. Public health management organizations are increasingly seeking to increase efficiency in health management, with the aim of obtaining information to use as a basis for epidemic control. Therefore, this project aims to develop a proposal for data mart where it is possible to apply statistics and data mining techniques. The phases performed in this work are: pre-processing, the creation of a data mart, exploratory analysis, calculation of indicators, inference with Bayes theorem and data classification. Analyzing the results, it was possible to deduce some probabilities and highlight characteristics of patients with greater influences to have the operational classification as multibacillary. It is concluded that the objectives of the work were achieved, and future work in the area of public health was proposed

Keywords: Data Warehouse. bayes theorem. Leprosy. Data Mining. Health public.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Os constantes avanços tecnológicos estão resultando na produção de uma grande quantidade de dados obtidos por inúmeros mecanismos de coleta de dados e por um número crescente de sistemas que produzem dados em escalas cada vez maiores. Consequentemente, a área de *Data Science* (Ciência de Dados) se tornou um caminho indispensável no mundo da tecnologia e dos negócios (CHEN et al., 2013). O termo se refere a ciência responsável pela análise de dados combinando as áreas de estatística e métodos computacionais, com objetivo de descobrir padrões e resolver problemas (WALLER; FAWCETT, 2013).

Embora a análise de uma vasta quantidade de dados possa ser intimidante, a correta extração e gerenciamento dos dados é um ponto relevante para as organizações devido a gerar informações valiosas ehttps://pt.overleaf.com/project/60368b796c87ca3d37aa73ab auxiliar no processo de tomada de decisões (CHEN et al., 2013). A área de *Data Science* vem ganhando uma maior visibilidade entre os planejamentos de negócios das mais diversas áreas, em particular, a saúde pública (CHEN et al., 2012).

A Saúde Pública foi definida por Winslow (1920) como "A arte e a ciência de prevenir a doença, prolongar a vida, promover a saúde e a eficiência física e mental mediante o esforço organizado da comunidade". Existem outras definições para o termo, como a de Sampaio (1960) destacando que a "Saúde Pública, em um século, evoluiu do conceito estreito do saneamento do meio físico para o conceito lato de preservação integral da Saúde do Homem". Assim, a prevenção e eliminação de doenças transmissíveis é uma das principais finalidades da saúde pública.

Dentre essas doenças, destaca-se a hanseníase, uma das doenças mais antigas da humanidade, e que é considerada um problema da saúde pública brasileira desde a década de 1930. Ainda hoje, o Brasil ocupa a 2ª posição mundial entre os países com maior número de casos novos detectados no mundo, precedido pela Índia (WHO, 2019). A distribuição da doença é diversa no Brasil, com casos novos concentrados nas regiões Norte, Centro-Oeste e Nordeste. O estado do Tocantins, localizado na região Norte do país, ocupou o segundo lugar entre os estados brasileiros com maior índice de casos novos e o primeiro em casos novos em pacientes com menores de 15 anos (MONTEIRO et al., 2015).

Compreender os diferentes determinantes da doença nesse estado poderá basear os programas locais no controle da doença, a fim de diminuir o ciclo de transmissão da doença evitando novas contaminações. Ter acesso a informação é o pré-requisito básico para a realização destas ações, ou seja, não é possível ter um controle e prevenção de doenças sem a disponibilidade de informações adequadas (PIRES et al., 2011).

No Brasil, os dados sobre saúde pública são coletados e processados através dos sistemas do Departamento de Informática do SUS (DATASUS) (BRASIL, 2016a). Com destaque para o Sistema de Informação de Agravos de Notificação (SINAN), alimentado pela notificação e investigação de casos de doenças e agravos que constam na lista nacional de doenças de notificação compulsória, dentre elas a Hanseníase.

Diante disso, a área de *Data Science* apresenta várias técnicas e ferramentas com a finalidade de produzir informações eficazes e descoberta de conhecimentos em grandes bases de dados, entre elas estão o *Data Mining* e o *Data warehouse*. Essas técnicas aplicadas aos dados das notificações de hanseníase presente no SINAN, oportuniza a geração de informações e conhecimento valiosos para contribuir nos estudos epidemiológicos, obter padrões ocultos nos dados e oferecer material para a criação de futuras diretrizes e políticas públicas de prevenção a doença.

1.1 Justificativa

Nos dias atuais, o Brasil enfrenta algumas dificuldades e particularidades na saúde pública. Um dos desafios mais importante é a capacidade de controle de epidemias no país (MV, 2018). Um relatório da Organização Mundial de Saúde (*World Health Organization*) mostrou que muitos países eliminaram doenças como a hanseníase na sua população, porém, o Brasil ainda possui a segunda maior taxa de incidência da doença, ficando atrás somente da Índia (WHO, 2012).

No estado do Tocantins, localizado na Região Norte do país, a hanseníase é considerada hiperendêmica de acordo com os critérios do Ministério da Saúde. Em 2016, o estado ocupou o primeiro lugar no Brasil em casos novos da doença na população geral (MONTEIRO et al., 2019). Somente em 2018 foram diagnosticadas mais de 1,3 mil pessoas com hanseníase, o número de casos triplicou se comparado ao ano de 2017, quando foram registrados 529 casos. Em razão da elevada carga, a doença permanece como um dos problemas de saúde pública no estado (MONTEIRO et al., 2015).

As organizações gestoras da saúde pública estão buscando cada vez mais aumentar a eficiência no gerenciamento de saúde, com o objetivo de obter informações para usar como base no controle de epidemias, prontuários, medicações, etc. Muitas iniciativas de sucesso foram obtidas através da implementação de um Data Warehouse e de mineração de dados em organizações de saúde internacionais e de outros estados. Porém, no contexto do estado do Tocantins, há uma certa carência de implantações de tecnologias semelhantes. (SANTOS et al., 2006)

Nesse cenário, a principal motivação desta pesquisa é a necessidade de uma ferramenta que seja capaz de armazenar de forma eficiente a grande quantidade de dados coletados do Sinan, como também um mecanismo para apresentação de informações e dados sobre a Hanseníase, e que forneça material para a criação de futuras diretrizes e

políticas públicas de prevenção da doença, assim melhorando a saúde da população, ou até mesmo sendo base para o controle da hanseníase no estado do Tocantins.

1.2 Descrição do Problema

Como podemos prover um ambiente de análise para transformar dados em informações úteis e que seja capaz de permitir aos profissionais tomarem decisões inteligentes que resulte em uma melhor gestão do controle da Hanseníase no estado do Tocantins.

1.3 Hipótese

Parte-se do pressuposto de que a melhor forma de transformar dados em informações úteis é através da adoção e união de tecnologias de data science e estatística, deduz, portanto, que a solução que mais se ajusta a situação é a criação de um data mart para armazenamento organizado dos dados, onde serão feitas análises, inferência e mineração sobre esses dados. As informações resultadas desses métodos serão adicionadas em um painel de resultados para serem usadas no gerenciamento da saúde pública, em especial para o controle da Hanseníase.

1.4 Objetivos

1.4.1 Objetivo Geral

Desenvolver uma proposta de data mart que possa fornecer fácil acesso aos dados, onde seja possível ser aplicado técnicas de estatísticas e mineração de dados; e a partir dos resultados obtidos criar um mecanismo de visualização que permita disponibilizar as informações encontradas.

1.4.2 Objetivos Específicos

Os objetivos específicos do presente projeto são:

1. Realizar um levantamento bibliográfico sobre os temas abordados no projeto;
2. Propor uma metodologia para a criação de um Data Mart, a partir das etapas de extração, formatação, limpeza, ajuste e o carregamento dos dados em uma estrutura multidimensional;
3. Executar uma análise exploratória sob os dados do data mart.
4. Criar os indicadores de hanseníase segundo o Ministério da Saúde.
5. Fazer uma inferência utilizando teorema de bayes nos dados extraídos do data mart;

6. Escolher os algoritmos que serão utilizados na mineração de dados, e após a mineração comparar as métricas de cada um;

7. Definir os objetivos da classificação, selecionar as variáveis que serão utilizadas, e executar a mineração de dados;

8. Criar um mecanismo de visualização de dados com todos os resultados obtidos.

1.5 Organização do Trabalho

Este projeto está divido em 7 capítulos. No segundo capítulo é realizada a fundamentação teórica sobre a área de saúde pública, definindo melhor o termo e apresentando uma breve história até chegar a criação do SUS. Ainda no segundo capítulo será apresentada uma revisão da literatura relacionada à Hanseníase, Teorema de Bayes, Data Warehouse e Mineração de Dados, abordando seus conceitos e técnicas. No terceiro capítulo é realizado um levantamento do atual estado da arte sobre soluções tecnológicas e inovadoras no campo da *e-health*, em especial, as aplicadas na área de saúde. O quarto capítulo contém a execução da metodologia, juntamente com o passo a passo de cada etapa. O quinto capítulo mostra os resultados das técnicas executadas no trabalho, em seguida, no capitulo seis, esses resultados são analisados e discutidos. Por fim, o sétimo e último capitulo apresenta as considerações finais no projeto e algumas recomendações para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os conceitos fundamentais necessários adotados para o desenvolvimento e entendimento deste trabalho. Deste modo, a seção 2.1 apresenta as definições relacionadas com a saúde pública discutindo o contexto histórico da saúde pública no Brasil até a criação do Sistema Único de Saúde (SUS). A seção 2.2 define a hanseníase apresentando algumas características da doença e os cálculos dos indicadores. A seção 2.3 apresenta o conceito relacionado com o SINAN. A seção 2.4 apresenta o conceito e o cálculo do teorema de bayes. As seções 2.5 e 2.6, conceitua, respectivamente, o Data Warehouse e a Mineração de Dados.

2.1 Saúde Publica

Existem múltiplas definições para o termo Saúde Pública. Um dos mais conhecidos foi apresentado por Winslow (1920):

> "Saúde pública é a arte e a ciência de prevenir a doença, prolongar a vida, promover a saúde e a eficiência física e mental mediante o esforço organizado da comunidade. A saúde pública abrange as áreas de saneamento do meio, o controle das infecções, a educação dos indivíduos nos princípios de higiene pessoal, a organização de serviços médicos e de enfermagem para o diagnóstico precoce e pronto tratamento das doenças e o desenvolvimento de uma estrutura social que assegure a cada indivíduo na sociedade um padrão de vida adequado à manutenção da saúde".

Um conceito muito semelhante foi apresentado por Acheson (1998), onde diz "A saúde pública é a arte e a ciência de prevenir doenças, promover a saúde e prolongar a vida através de esforços organizados da sociedade". Mesmo com a semelhança entre as definições, a apresentada por Winslow (1920) continua sendo o conceito mais abrangente e articulado atualmente. A saúde pública depende de um relacionamento entre fatores genéticos, ambientais e boas condições de vida. Como resultado, a saúde pública se baseia em conhecimentos e habilidades de outras áreas, como a biologia, psicologia, sociologia, ciências ambientais e da terra, estatística, comunicação, políticas pública e etc. (CDC, 2018).

No geral, a saúde pública trata sobre a implementação de intervenções que previnem a ocorrência de doenças, em outras palavras, preocupa-se em proteger a saúde da população. Essa população pode ser pequena, como uma vizinhança local, ou grande, como um país. Logo, a saúde é um serviço essencial e um direito fundamental do ser

humano (CDC, 2018).

O direito à saúde, garantido parcialmente ou totalmente ao redor do mundo, é fruto de uma longa trajetória e de fatores próprios de cada sociedade. No antigo regime, somente a população de mais alta renda tinha acesso a cuidados médicos e consultas particulares; os mais pobres, em alguns países, recebiam atenção de organizações católicas. O restante da população, amplamente rural, tinha os cuidados obtidos do saber da comunidade local (MARQUES et al., 2016).

Atualmente, existem diferentes tipos de sistemas de saúde espalhado pelo mundo. Cada um dos sistemas tem uma relação estreita com os tipos de proteção social de cada país. Na Dinamarca, Espanha, Finlândia, Grécia, Itália, Portugal, Reino Unido e Suécia, existe o sistema National Health Service (NHS), nascido na Inglaterra, que em plena guerra, foi criado com o objetivo de garantir acesso a todos, independentemente da renda. Como proposta de financiamento foi apontado os impostos gerais. Nesse sentido, o NHS começou a funcionar em 1948, garantindo acesso gratuito à assistência médica a todo cidadão britânico (MARQUES et al., 2016).

Um outro sistema de saúde existente é por meio de seguro de saúde, financiado principalmente por contribuições de empregadores e empregados, podendo contar também com recursos de impostos. A origem desses sistemas são as Caixas, organizações feitas a partir das categorias profissionais de assistência médica e de acordo com a capacidade financeira, o resultado era a segmentação e a diferenciação dos cuidados entre as classes. Atualmente, as ações de serviços da saúde coletiva é de responsabilidade de um órgão público, e o seguro fica responsável pelo tratamento e a reabilitação (MARQUES et al., 2016).

Por último, onde existe a proteção social do tipo residual, o Estado somente se responsabiliza por proteger os mais pobres, ou seja, os sistemas de saúde pública só se dirigem aos mais necessitados e parcialmente aos aposentados, os restantes são deixados para que compre serviços de saúde no mercado. Os que não têm renda suficiente, embora não sejam considerados pobres, não estão cobertos nem pelos seguros e nem pelos programas públicos (MARQUES et al., 2016).

Para identificar quais seriam os serviços e atividades da saúde pública, o Escritório Regional da OMS da Europa desenvolveu as dez "Operações Essenciais da Saúde Pública" (OESPs) (JENSEN; LUKIC; GULIS, 2018), tais como:

1. Vigilância da saúde e bem-estar da população;
2. Monitoramento e resposta a riscos à saúde e emergências;
3. Proteção à saúde, incluindo ambiental, ocupacional, segurança alimentar e outros;
4. Promoção da saúde, incluindo ações para lidar com determinantes sociais e iniquidades em saúde;

5. Prevenção de doenças, incluindo detecção precoce de doenças.
6. Garantir a governança para a saúde e o bem-estar;
7. Garantir uma força de trabalho em saúde pública suficiente e competente;
8. Garantir estruturas e financiamentos organizacionais sustentáveis;
9. Advocacia, comunicação e mobilização social para a saúde;
10. Promoção da pesquisa em saúde pública para informar políticas e práticas.

Para cada uma das operações, foi definido um conjunto de ações, trazendo o questionamento de qual cenário as ações devem ser realizadas. Para a saúde pública parece ser promissor abraçar todas as ações incluídas nas OESPs (JENSEN; LUKIC; GULIS, 2018).

2.1.1 Saúde Pública no Brasil

Assim como os sistemas de saúde de outros países foram frutos do seu passado e da sua história, a saúde no Brasil também sofreu influências do contexto político-social pelo qual o país passou em toda sua história. O sistema de saúde brasileiro foi desenvolvido a partir da evolução política e econômica do país, passando por diferentes regimes, distintas promulgações de leis e normas que afetaram a saúde pública até a criação do atual Sistema Único de Saúde (SUS) (GUIDINI, 2012).

A história da saúde no Brasil passa, indispensavelmente, pela filantropia, pela caridade e por ações religiosas (CARVALHO, 2013). Durante o período da colonização, o acesso a cuidados médicos era determinado pela classe social, somente os nobres tinham acesso fácil aos médicos, enquanto a população, pobres, escravos e indígenas não recebiam nenhum tipo de assistência, ficando dependentes da filantropia e caridade (MAGALHÃES, 2019).

Após a independência do Brasil, D. Pedro II criou órgãos para auxiliar a saúde pública, a fim de evitar epidemias e melhorar a qualidade de vida. Foram adotadas também medidas para o saneamento básico (MAGALHÃES, 2019). Uma medida sanitária importante do país ocorreu no governo de Rodrigues Alves (1902 – 1906) foi comandado por Oswaldo Cruz, a ação praticava despejos e agressões para forçar a população a tomarem vacinas, o povo revoltado saiu às ruas e enfrentou a polícia. Esse movimento é conhecido como a revolta da vacina (GARCIAN, 2003).

Somente em 1953 foi instituído o Ministério da Saúde, com a Lei n.º 1.920, de 25 de Julho de 1953 (BRASIL, 1993). Foi iniciado também as primeiras conferências sobre saúde pública no país, onde surgiu a ideia de criação de um sistema único de saúde que tivesse como objetivo atender toda a população. Infelizmente, em 1964, com a ditadura militar, a saúde sofreu cortes orçamentários e como resultado muitas doenças voltaram a se intensificar (OLIVEIRA, 2012).

Em 1970, surgia o movimento sanitarista com o objetivo de discutir as mudanças necessárias para melhorar a saúde pública no Brasil. Uma das principais conquistas do movimento foi a realização da 8º Conferência Nacional da Saúde, realizada em 1986. O relatório do evento foi tomado como esboço para a criação do Sistema Nacional de Saúde, o SUS (MAGALHÃES, 2019). Esse sistema pode ser definido como "o conjunto de ações e serviços de saúde, públicos e privados, contratados ou conveniados com o poder público" (OLIVEIRA, 2012).

O SUS é um dos maiores e mais complexos sistemas de saúde do mundo, ele abrange desde o atendimento simples, como avaliação de pressão arterial, até o transplante de órgãos, o sistema garante acesso integral e gratuito para todos os brasileiros. Com a criação do SUS, foi proporcionado acesso universal ao sistema de saúde, sem discriminação. A atenção integral à saúde passou a ser um direito de todos os brasileiros (BRASIL, 2017). A Constituição Federal (CF), aprovada em 1988, determinou no artigo 196 que "A saúde é direito de todos e dever do Estado, garantido mediante políticas sociais e econômicas que visem à redução do risco de doença e de outros agravos e ao acesso universal e igualitário às ações e serviços para sua promoção, proteção e recuperação" (BRASIL, 1988).

Mesmo com a CF incorporando as propostas da 8º Conferência Nacional de Saúde, a regulamentação do SUS só foi acontecer em 1990, através da Lei n.º 8.080, que diz, no Art. 1.º "Esta lei regula, em todo o território nacional, as ações e serviços de saúde, executados isolada ou conjuntamente, em caráter permanente ou eventual, por pessoas naturais ou jurídicas de direito Público ou privado." (BRASIL, 1990).

As propostas de diretrizes para o SUS, apresentadas na 8º Conferência Nacional de Saúde, foram (BORGES, 2002):

- Descentralização, com direção única em cada esfera de governo;
- Atendimento integral, com prioridade para as atividades preventivas, sem prejuízo dos serviços assistenciais;
- Participação da comunidade;
- As diretrizes citadas dão suporte aos princípios doutrinários do SUS;
- Universalidade: A garantia de atenção à saúde, por parte do sistema, a todo e qualquer cidadão;
- Equidade: A garantia de que todo indivíduo deve ser igual perante o SUS e deve ser atendido de acordo com suas necessidades;
- Integralidade: A garantia de que cada pessoa, ao receber tratamento, deve ser olhada como um ser integral, inserida num contexto familiar, social e econômico.

A criação e regulamentação do SUS foram os marcos mais importantes na história da Saúde Pública do País. Após vários movimentos e tentativas, foi o primeiro material legitimado e garantido em lei para a construção de uma saúde pública de qualidade e universal. Um ganho para toda a sociedade brasileira, na conquista de um direito básico – a saúde enquanto bem inalienável (BORGES, 2002).

2.2 Hanseníase

A hanseníase é uma doença crônica e transmissível. A doença ataca principalmente a pele e os nervos periféricos, podendo causar incapacidades e deformidades físicas. A infecção pode contagiar pessoas de ambos os sexos e qualquer idade, e o principal meio de transmissão se dá de uma pessoa doente em contato próximo e prolongado com a pessoa sem a doença (BRASIL, 2019).

O agente causador da doença é denominado Mycobacterium leprae (ou bacilo de hansen), apresenta-se sob a forma de bacilo reto ou um pouco encurvado, com extremidades arredondadas. O bacilo de hansen possui crescimento extremamente lento em relação a outras bactérias, possui tamanho médio entre 0,3 de diâmetro e 1 a 8 mm de comprimento. As localizações das lesões hansênicas (pele e nervos periféricos) indica que o bacilo tenha preferência por temperaturas menores que 37º (MACIEIRA, 2000).

Os mais frequentes sinais e sintomas da hanseníase são (BRASIL, 2019):

- Manchas esbranquiçadas, avermelhadas ou amarronzadas, em qualquer parte do corpo, com perda ou alteração de sensibilidade térmica (ao calor e frio), tátil (ao tato) e à dor.
- Áreas com diminuição dos pelos e do suor.
- Dor e sensação de choque, formigamento, fisgadas e agulhadas ao longo dos nervos dos braços e das pernas.
- Inchaço de mãos e pés.
- Diminuição da sensibilidade e/ou da força muscular da face, mãos e pés, devido à inflamação de nervos, que nesses casos podem estar engrossados e doloridos.
- Úlceras de pernas e pés.
- Caroços (nódulos) no corpo, em alguns casos avermelhados e dolorosos.
- Febre, edemas e dor nas articulações.
- Entupimento, sangramento, ferida e ressecamento do nariz.
- Ressecamento nos olhos.

A hanseníase possui duas classificações: paucibacilar e multibacilar. A classificação paucibacilar (poucos bacilos) é a doença inicial e não transmissível. Os casos classificados como multibacilares (muitos bacilos) são as formas clínicas transmissíveis. Em ambos os casos, o tratamento é realizado com a associação de antimicrobianos, denominado Poliquimioterapia (PQT). O diagnóstico é feito com o médico da equipe de Estratégia de Saúde da Família da Unidade de Saúde mais próxima ao paciente (BRASIL, 2019).

O diagnóstico da hanseníase é por meio clínico e epidemiológico, feito a partir de exames geral e dermatoneurológico para identificar lesões ou regiões da pele com alterações de sensibilidade. Em crianças, o diagnóstico exige avaliação mais criteriosa pela dificuldade da aplicação e interpretação dos testes de sensibilidade (BRASIL, 2019).

A hanseníase é uma das doenças mais antigas da humanidade. As referências mais antigas datam de 600 a.C. Os continentes asiático e africano são consideradas o berço da doença. Antes conhecida como lepra, a terminologia hanseníase é iniciativa brasileira para tentar minimizar o preconceito secular atribuído à doença. Com isso, o nome lepra e seus adjetivos passaram a ser proibidos no país. (BRASIL, 2019).

O Brasil ocupa a 2ª posição do mundo entre os países que registram caso novos de hanseníase, ficando atrás somente da índia. No ano de 2018, foram registrados cerca de 209 mil casos de hanseníase no mundo. A região das Américas é responsável por 15% do total de casos, o Brasil representa 93% dos casos nesta região (WHO, 2019). As regiões norte e centro-oeste do país possuem os maiores coeficientes de detecção de hanseníase. No estado do Tocantins a hanseníase é considerada hiperendêmica. Somente em 2018 foram diagnosticadas mais de 1,3 mil pessoas com hanseníase, o número de casos triplicou se comparado ao ano de 2017, quando foram registrados 529 casos. Em razão da elevada carga, a doença permanece como um dos problemas de saúde pública no Brasil. (MONTEIRO et al., 2015)

2.2.1 Indicadores

A vigilância epidemiológica envolve desde a coleta até a interpretação dos dados referentes aos casos de hanseníase. A produção e a divulgação das informações contribuem para as avaliações das intervenções e embasam o planejamento de novas ações e recomendações a serem implementadas na gestão do controle de hanseníase. Os números e informações utilizados na divulgação são chamados de indicadores. Existem diversos indicadores e cada um possui um cálculo diferente, e um objetivo a ser alcançado. Os indicadores são apresentados a seguir:

Tabela 1 – Indicadores de Monitoramento do Progresso da Eliminação da Hanseníase enquanto problema de saúde pública

Nome do indicador	Cálculo	Utilidade	Parâmetros
Taxa de detecção anual de casos novos de hanseníase por 100 mil habitantes	(número de casos novos residentes em determinado local no ano da avaliação) / (população total residente no mesmo local e período) * 100 mil	Medir força de morbidade, magnitude e tendência da endemia	Hiperendêmico: > 40,0 Muito alto: 20,00 a 39,99 Alto: 10,00 a 19,99 Médio: 2,00 a 9,99 Baixo: < 2,00
Taxa de detecção anual de casos novos de hanseníase, na população de zero a 14 anos, por 100 mil habitantes	(número de casos novos em menores de 15 anos de idade residentes em determinado local e ano) / (população de zero a 14 anos de idade no mesmo local e período) * 100 mil	Medir força da transmissão recente da endemia e sua tendência	Hiperendêmico: > 10,0 Muito alto: 5,00 a 9,99 Alto: 2,50 a 4,99 Médio: 0,50 a 2,49 Baixo: < 0,50
Proporção de casos de hanseníase, segundo gênero entre o total de casos novos	(Casos de hanseníase do sexo feminino) / (Total de casos novos) * 100	Avaliar a capacidade dos serviços em assistir aos casos de hanseníase	Não especifica parâmetro
Proporção de casos de hanseníase curados com grau 2 de incapacidade física entre os casos avaliados no momento da alta por cura no ano	(número de casos com incapacidade física grau 2 no ano da avaliação) / (total de casos de hanseníase com cura no ano da avaliação) * 100	Avaliar a transcendência da doença e subsidiar a programação de ações de prevenção e tratamento de incapacidades pós-alta	Alto: >10% Médio: 5 a 9,9% Baixo: <5%
Proporção de casos segundo classificação operacional entre o total de casos novos	(Casos de hanseníase multibacilares) / (Total de casos novos) * 100	Avaliar os casos em risco de desenvolver complicações e para o correto reabastecimento de PQT	Não especifica parâmetro

2.3 SINAN

O Sistema de Informação de Agravos de Notificação, criado na década de 90, é um sistema de informação nacional utilizado para notificação universal de agravos de notificações compulsórias. Criado com o objetivo de coletar e processar dados sobre agravo de notificação em todo território nacional, o SINAN se alimenta de dados obtidos através da notificação e investigação de casos de doenças e agravos contidos na lista nacional de doenças e agravos de notificação compulsória, conforme apresentado na tabela 2 (LAGUARDIA et al., 2004).

Tabela 2 – Lista Nacional de Notificação Compulsória.

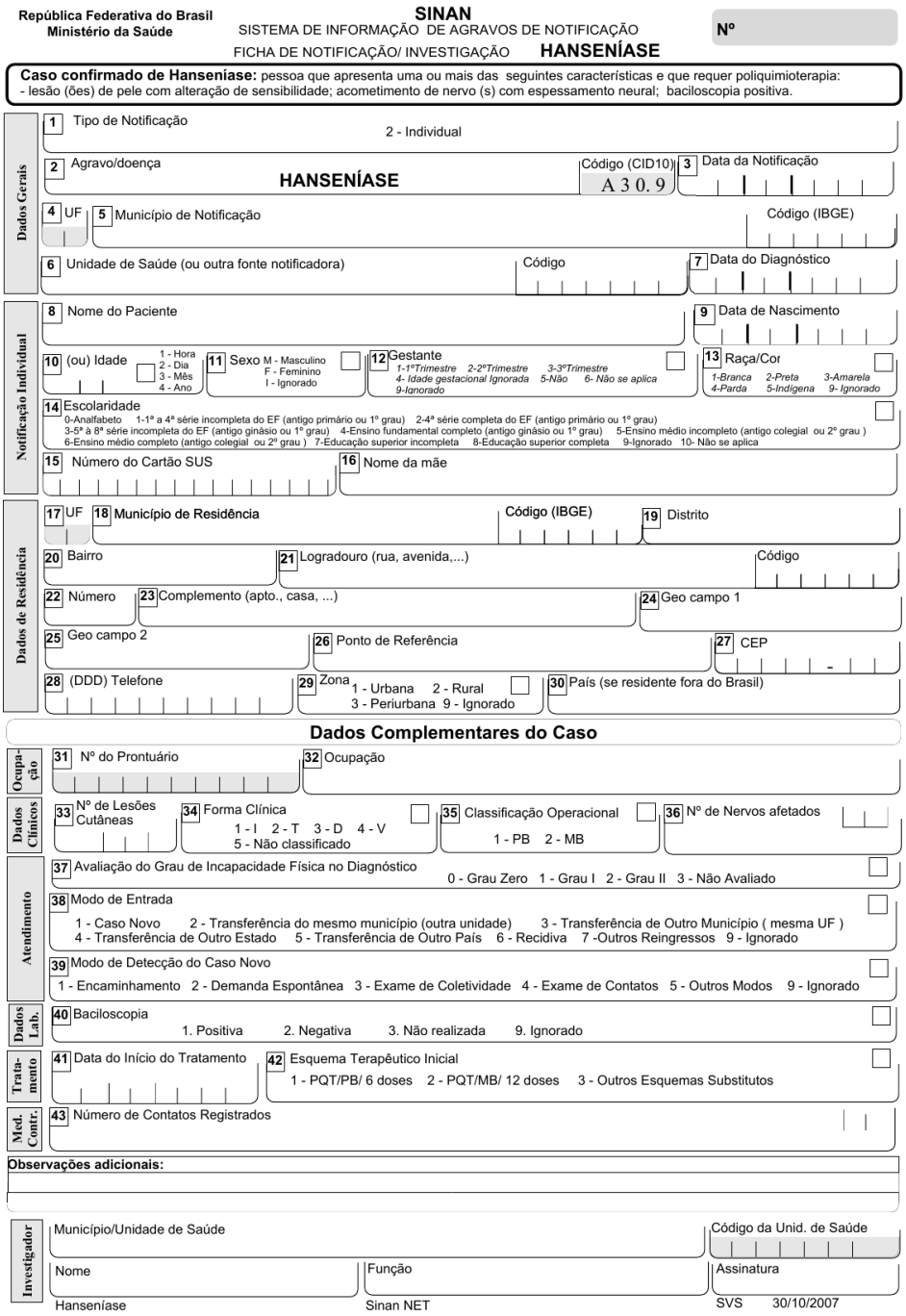
DOENÇA OU AGRAVO			
Acidente no trabalho	Acidentes por animal	Arenavírus, Ebola, Varicela	Peste,Hepatites virais
Botulismo, Síndrome da Rubéola Congênita	Doença aguda pelo vírus Zika	HIV/AIDS, Raiva, tentativa de suicídio	Cólera, Esquistossomose
Coqueluche, Febre Amarela	Eventos adversos graves ou óbitos pós-vacinação	Influenza humana, Intoxicação Exógena	Sarampo, Rubéola, Dengue, Antraz pneumônico
Difteria, Tétano, tuberculose	Febre do Nilo Ocidental	Leishmaniose Tegumentar Americana	Sífilis, Leptospirose, Tularemia, Varíola
Doença de Chagas Aguda, Óbito infantil e materno	Poliomielite, Febre Maculosa e outras Riquetsioses	Violência doméstica, violência sexual	Síndrome da Paralisia Flácida Aguda
Doença de Creutzfeldt-Jakob	Febre Tifoide, Doença Meningocócica	Malária, Hanseníase, Hantavirose	Síndrome Respiratória Aguda Grave

A utilização adequada do SINAN permite a realização de diagnósticos eficientes da ocorrência de um evento na população, podendo fornecer recursos para explicações de agravos de notificação compulsória, e é possível também indicar riscos aos quais as pessoas estão sujeitas, além de identificar a realidade epidemiológica de determinada área geográfica. Logo, o SINAN é um sistema importante no planejamento da saúde pública, sendo possível definir prioridades de intervenção e permitir a avaliação dos impactos de intervenções (BRASIL, 2016b)

O Ministério da Saúde, através da Lei n.º 6.259 de 30 de Outubro de 1975 (BRASIL,

1975), instituiu o Sistema Nacional de Vigilância Epidemiológica (SNVE). Essa lei tornou obrigatória a notificação de doenças transmissíveis, de acordo com a Lista Nacional de Notificação Compulsória (Tabela 2) definida através da Portaria n.º 1.271, de 6 de junho de 2014 (BRASIL, 2014). A notificação da Hanseníase, por exemplo, é realizada através do preenchimento da Ficha de Notificação compulsória da doença, mostrada na Figura 1.

Figura 1 – Ficha de Notificação Compulsória.



Fonte: Portal do SINAN (2016).

A entrada de dados no SINAN é feita com a utilização de formulários padronizados, listados a seguir (BRASIL, 2009):

- Ficha individual de notificação (FIN): fornecida ao paciente quando da suspeita de ocorrência de problema de saúde de notificação compulsória, e encaminhada aos serviços responsáveis pela vigilância epidemiológica.

- Ficha individual de investigação (FII): destinado a ser um roteiro de investigação, diferente para cada tipo de agravo, que deve ser preferencialmente usado pelos serviços municipais de vigilância. O preenchimento da ficha permite levantar dados que possibilitam a identificação da fonte de infecção e dos mecanismos de transmissão de doença.

O SINAN obtêm dados fundamentais ao cálculo dos principais indicadores extremamente importantes e úteis, tais como as taxas de incidência, letalidade e mortalidade, coeficiente de prevalência, entre outros (BRASIL, 2009).

2.4 Teorema de Bayes

O teorema de Bayes (também conhecido como regra de Bayes), em estatística e teoria da probabilidade, é uma fórmula matemática utilizada para determinar a probabilidade condicional de eventos. Fundamentalmente, o teorema descreve a probabilidade de um evento ocorrer com base no conhecimento prévio das condições que podem ser relevantes para o evento (OLIVEIRA, 2020).

O teorema recebe o nome do estatístico inglês Thomas Bayes, que descobriu a fórmula em 1763 (OLIVEIRA, 2020). A regra de Bayes é considerada a base da abordagem de inferência estatística especial chamada de inferência de Bayes. O teorema é expresso na seguinte fórmula:

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)}$$

Onde :

$P(A|B)$ - a probabilidade de ocorrência do evento A, dado que o evento B ocorreu.

$P(B|A)$ - a probabilidade de ocorrência do evento B, dado que o evento A ocorreu.

$P(A)$ - a probabilidade do evento A

$P(B)$ - a probabilidade do evento B

2.5 Data Warehouse

Os Data Warehouse (DW) são desenvolvidos em diferentes organizações a fim de atender a necessidades específicas. Logo, a definição do termo foi elaborada de diversas maneiras. Entretanto, mesmo com diversos conceitos, o DW tem a integração de dados como a característica comum em todas as definições, visando padronizar os dados das diversas fontes.

Barbieri (2001) caracterizou um DW como um banco de dados destinado a sistemas de apoio à tomada de decisão, cujos dados são armazenados em estruturas lógicas dimensionais, possibilitando o seu processamento analítico por ferramentas especiais, e armazenamento de dados em vários graus de relacionamento, de forma a facilitar e agilizar os processos de tomada de decisão por diferentes níveis gerenciais.

Segundo Singh (2001), o termo Data Warehouse define um conjunto de novos conceitos e ferramentas que evolui para uma tecnologia que permite atacar o problema de oferecer a todas as pessoas-chave da empresa acesso a qualquer nível de informação necessário para que a organização possa sobreviver e prosperar em um mundo cada vez mais competitivo. Nessas perspectivas, o principal objetivo de um DW é disponibilizar informações que possam auxiliar as organizações nas suas tomadas de decisões. Os DWs dão suporte a demandas de altos desempenhos em relação aos dados e informações de uma empresa.

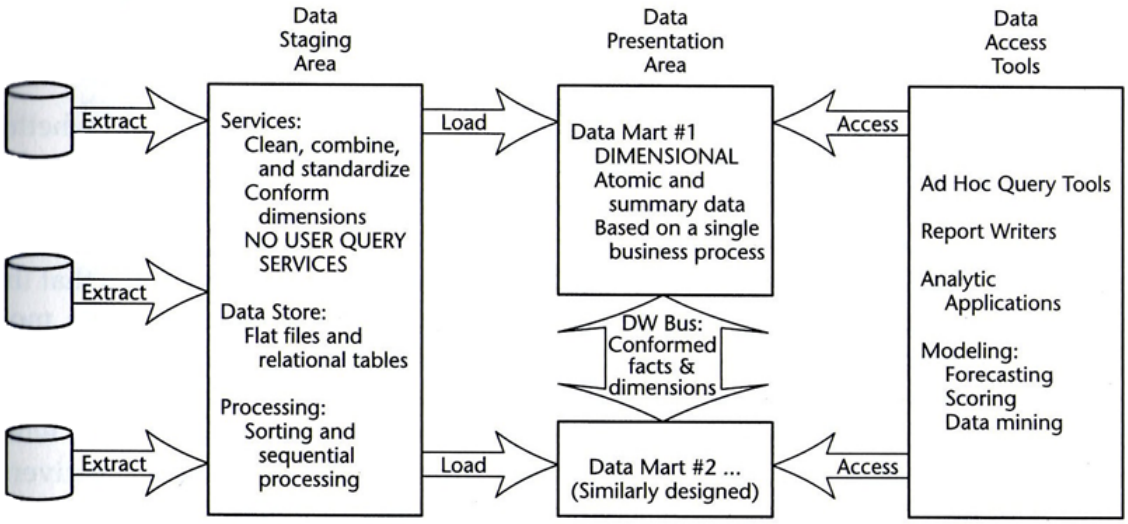
2.5.1 Características

Inmon e Hackathorn (1997), reconhecido como o primeiro a usar o termo armazém (ou warehouse), descreveu um data warehouse como "uma coleção de dados orientada por assunto, integrada, variante e não volátil, que tem por objetivo dar suporte aos processos de tomada de decisão". As características de um Data Warehouse citadas por Inmon são apresentadas a seguir:

- Orientado por assunto: remete aos sistemas organizados de uma aplicação, ou seja, toda a modelagem do Data Warehouse é orientada a partir de assuntos.
- Integração: A integração é a propriedade mais importante do data warehouse, tem como objetivo padronizar os dados dos diversos sistemas.
- Não volátil: Antes de os dados serem carregados em um ambiente de DW, eles são filtrados e limpos. Após a inclusão no DW, os dados podem somente ser consultados e excluídos, não podem ser alterados.
- Variável com o tempo: os dados adicionados em um DW são vinculados a um período de tempo.

De acordo com Kimball e Ross (2011), a estrutura base de uma DW é composta por quatro componentes: sistemas de Fonte de Dados Operacionais, Área de Preparação dos Dados, Área de Apresentação dos Dados e Ferramentas de Acesso de Dados. Os elementos são representados na figura 2 e detalhados a seguir.

Figura 2 – Estrutura de um Data Warehouse.



Fonte: Kimball Group (2002).

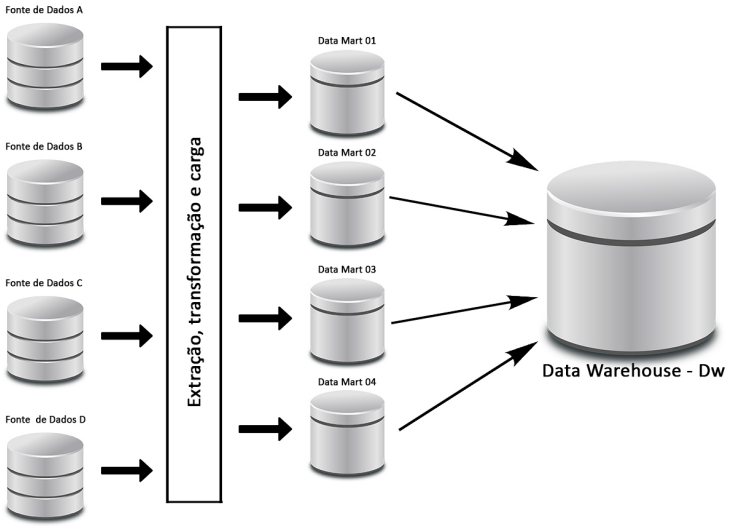
- Sistemas de Fonte de Dados Operacionais (*Operational Source Systems*): São os sistemas que irão ser fontes de dados para o data warehouse.
- Área de Preparação dos Dados (*Data Staging Area*): etapa que ocorre a limpeza dos dados antes de serem carregados no data warehouse, consistem na extração, transformação e carga dos dados.
- Área de Apresentação dos Dados (*Data Presentation Area*): onde os dados são organizados, armazenados e disponibilizados para consulta. Nessa área ficam os Data Mart (DM), um dos elementos fundamentais de um DW. Os DM são subconjuntos lógicos de um DW, normalmente são divididos por áreas específicas.
- Ferramentas de Acesso de Dados (*Data Access Tools*): onde são apresentados aos usuários os dados e as informações do data warehouse.

2.5.2 Abordagens

As abordagens para a construção de DW podem variar de acordo com os autores apresentados. As referências mais fortes existentes sobre Data Warehouse se baseiam em duas fontes: Inmon e Hackathorn (1997) e Kimball et al. (2008). Inmon defende a abordagem *Top Down*, onde o processo se inicia com a extração, a transformação e a integração dos dados diretamente para o DW. Kimball tem forte influência nas abordagens

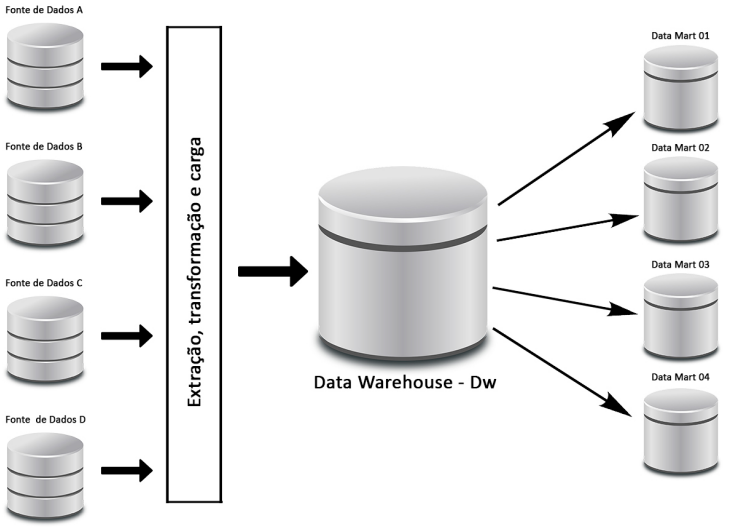
bottom Up, que se caracteriza por iniciar com a extração, transformação e integração para um ou mais *Data Marts*. Ou seja, para Inmon, um DM deriva de um DW, e para Kimball um DW é feito pela união dos DM (SANTOS et al., 2006). A Figura 3 detalha a abordagem *Bottom Up* e a figura 4 se refere a abordagem *Top Down*.

Figura 3 – Abordagem Bottom Up de um DW.



Fonte: Adaptado pelo autor.

Figura 4 – Abordagem Top Down de um DW



Fonte: Adaptado pelo autor.

2.6 Data Mining

O Data Mining (DM), ou mineração de dados, é uma das etapas da descoberta de conhecimento em banco de dados (*Knowledge Discovery in Databases - KDD*), na qual é definida como o processo responsável pela busca de conhecimentos implícitos nas bases de dados. A definição mais conhecida de KDD foi apresentada por Fayyad, Piatetsky-Shapiro e Smyth (1996) como "o processo não-trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e compreensíveis". Fayyad, Piatetsky-Shapiro e Smyth (1996) definiu mineração de dados como "uma etapa do processo KDD que consiste em aplicar algoritmos de análise e descoberta de dados que produzem uma enumeração específica de padrões (ou modelos) sobre os dados".

Ainda não existe um consenso entre os conceitos de KDD e mineração de dados. Para alguns autores Wang (2005) e Han, Pei e Kamber (2011) os termos são considerados sinônimos. Para Cios, Pedrycz e Swiniarski (1998) e Fayyad, Piatetsky-Shapiro e Smyth (1996), a mineração de dados é uma das etapas do processo do KDD. Entretanto, todos os autores abordam o processo de data mining como interativo e dividido em fases (CAMILO; SILVA, 2009).

Segundo Cabena et al. (1998) mineração de dados é um campo interdisciplinar que usa técnicas de extração em dados para extrair informações previamente desconhecida e de máxima abrangência a partir de bases de dados. Para Decker e Focardi (1995), a mineração de dados é definida como uma metodologia de solução de problemas que procura uma descrição lógica ou matemática, eventualmente de natureza complexa, de padrões e regularidades em um conjunto de dados.

De acordo com as definições apresentadas, a mineração de dados tem como objetivo extrair conhecimento a partir de um grande conjunto de dados, descobrir padrões e relações ocultas e relacionar as informações obtidas para que possam ser utilizadas nas tomadas de decisões das organizações.

2.6.1 Tarefas

A Mineração de Dados é geralmente classificada pela sua capacidade em realizar algumas tarefas, que são citadas a seguir:

- Descrição (*Description*): É tarefa usada para descrever os padrões encontrados entre os dados. Geralmente a descrição é utilizada para a interpretação dos resultados.

- Classificação (*Classification*): Essa tarefa tem como objetivo identificar qual classe um determinado registro pertence. O algoritmo de classificação analisa o conjunto de registros fornecidos, onde cada registro já contém a indicação à qual classe pertence, a fim de 'aprender' como classificar um novo registro.

- Regressão (*Regression*): Essa tarefa se assemelha um pouco com a classificação, porém ela é usada quando o registro é identificado por um valor numérico e não um categórico. Desse modo, o algoritmo pode estimar o valor de uma determinada variável analisando os valores das demais.

- Predição (*Prediction*): A predição é similar às tarefas de classificação e regressão. A diferença é que esta tarefa visa descobrir o valor futuro de um determinado atributo.

- Agrupamento (*Clustering*): A tarefa de agrupamento tem como objetivo identificar e agrupar os registros similares. Esta tarefa se diferencia da classificação pois não necessita que os registros sejam previamente categorizados.

2.6.2 Técnicas de Mineração de Dados para a Classificação

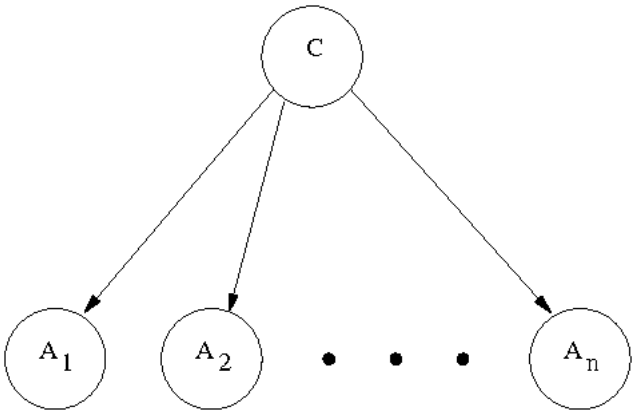
Data Mining é uma área muito grande, por consequência existem várias maneiras de encontrar padrões em um grande conjunto de dados (ELMASRI; NAVATHE, 2011). A seguir serão apresentadas as duas técnicas utilizadas nesse trabalho.

- ***Naive Bayes***

O Naive Bayes é um classificador probabilístico baseado no Teorema de Bayes (MORETTIN; BUSSAB, 2017). Uma de suas principais características é o motivo de receber "naive" (ingênuo) no nome, o algoritmo desconsidera completamente a correlação entre as variáveis, ou seja, ele trata cada um dos atributos de forma independente. Apesar da premissa de ingênua e simplista, o algoritmo apresenta o melhor desempenho em várias aplicações de classificação (MILIDIÚ, 2006).

Em um contexto bem simples, o classificador calcula a probabilidade de um evento ocorrer dado que outro evento já ocorreu. O Naive Bayes é representado por probabilidades, ou seja, ao terminar a execução do modelo no conjunto de dados de treinamento, é obtido uma lista de probabilidades que serão usadas para fazer as previsões no conjunto de dados de teste. Estas probabilidades são a probabilidade de cada classe no conjunto de dados de treino e a probabilidade condicional de cada atributo dado cada classe (MOURA, 2019). Na Figura 5 o classificador é representado com os atributos A_1, A_2 e A_n, a suposição de independência condicional é mostrada como nenhum conector entre os atributos.

Figura 5 – Representação simples do Naive Bayes



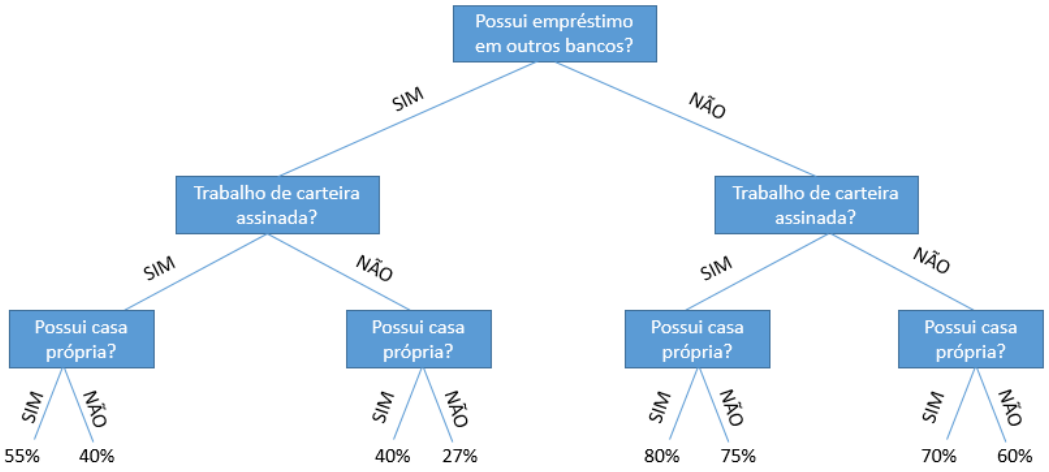
Fonte: Jesus Cerquides (2003)

- ***Decision Tree***

A árvore de decisão é uma técnica de aprendizado de máquina (*machine learning*), que utiliza a estratégia dividir e conquistar, onde um problema complexo é dividido em subproblemas mais simples, e recursivamente, a mesma estratégia é aplicada a cada subproblema. As soluções de cada subproblemas podem ser combinadas na forma de uma árvore para produzir uma solução do problema complexo (GAMA, 2002).

De acordo com Ragsdale (2004), a árvore de decisão é composta por nós (círculos e quadrados) interconectados por ramos (linhas). Um nó quadrado é chamado de nó de decisão por representar uma possível decisão. Os ramos interligando o nó de decisão representa as diferentes alternativas para a decisão. Um exemplo desse esquema está representado na Figura 6, onde com base na probabilidade das respostas, um banco x decide se fornece ou não crédito para seu cliente.

Figura 6 – Exemplo de uma arvore de decisão.



Fonte: Estatsite.com (2016)

2.6.3 Modelos de Processos

Existem diversos modelos de processos de desenvolvimento que definem e padronizam as fases e atividades para realização da Mineração de Dados, os três mais utilizados são: Cross Industry Standard Process for Data Mining (CRISP-DM), Knowledge Discovery in Databases (KDD) e Sample, Explore, Modify, Model e Assess (SEMMA). Cada modelo apresenta suas particularidades, porém, todos eles têm uma estrutura geral em comum. A seções à seguir apresentam brevemente os três modelos de processos mais utilizados na literatura.

- CRISP-DM

Esse processo é um dos mais utilizados em mineração dos dados, tendo como principal vantagem a possibilidade de aplicação em qualquer tipo de negócio. O modelo CRISP-DM consiste em um ciclo que compreende seis etapas gerais (AZEVEDO; SANTOS, 2008):

1. Entendimento comercial: esta fase inicial tem como objetivo entender a meta que se deseja atingir com a mineração de dados.

2. Entendimento dos dados: consiste na coleta inicial de dados, identificação de problemas de qualidade dos dados e identificação dos dados relevantes para o problema em questão.

3. Preparação de dados: a fase de preparação de dados abrange todas as atividades para construir o conjunto de dados final a partir dos dados brutos iniciais. Geralmente essa etapa envolve filtrar, combinar e preencher valores vazios.

4. Modelagem: nesta fase as várias técnicas de modelagem selecionadas são aplicadas.

5. Avaliação: são feitos testes e validações, visando obter a confiabilidade nos modelos selecionados.

6. Implantação: onde acontece a divulgação dos resultados, os envolvidos no projeto conhecem os resultados.

- KDD

O processo KDD, apresentado em Fayyad, Piatetsky-Shapiro e Smyth (1996), é um processo que usa métodos de mineração de dados para realizar a extração de conhecimento em bases de dados. Trata-se de uma das metodologias mais antigas, e foca nas descobertas de conhecimentos a partir de dados. O KDD é constituído por 5 etapas (AZEVEDO; SANTOS, 2008):

1. Seleção: consiste em criar um conjunto de dados ou selecionar um subconjunto de dados, nas quais a descoberta deve ser executada.

2. Pré-processamento: essa etapa consiste na limpeza e pré-processamento dos dados coletados, a fim de obter dados consistentes.

3. Transformação: nessa etapa é feita a transformação dos dados usando dimensionalidade e métodos de redução ou transformação.

4. Data Mining: consiste na busca de padrões de interesse em uma forma representacional específica, dependendo do objetivo de mineração de dados (geralmente, previsão)

5. Interpretação/Avaliação: Esta etapa consiste na interpretação e avaliação dos resultados obtidos.

- **SEMMA**

SEMMA (Explore, Modifique, Modele, Avalie) se refere ao processo de condução de um projeto de DM, criada pela SAS Institute, a metodologia foca nas tarefas de criação do modelo, e é constituído por 5 etapas (AZEVEDO; SANTOS, 2008):

1. Amostra: etapa de amostragem dos dados, extraindo uma parte de um conjunto de dados.

2. Explorar: etapa consiste na exploração dos dados, procurando tendências imprevistas e anomalias entre os dados.

3. Modificar: consiste na modificação dos dados, criando, selecionando e transformando as variáveis para focalizar o processo de seleção do modelo.

4. Modelo: modelagem dos dados, permitindo que o software pesquise automaticamente para uma combinação de dados que prediz de forma confiável o resultado desejado.

5. Avaliar: consiste em avaliar os dados, avaliando a utilidade e confiabilidade das descobertas do processo de mineração de dados.

3 ESTADO DA ARTE

Este capitulo apresenta um conjunto de trabalhos que aborda soluções tecnológicas e inovadoras no campo da saúde, em especial, as aplicadas na área de saúde pública. Esses trabalhos foram resultado de um levantamento do atual estado da arte das temáticas de aplicação da mineração de dados, data warehouse, sistemas de apoio à decisão e técnicas de visualização aplicados na área da saúde.

3.1 Sistema inteligente com base na tecnologia Naive Bayes

Palaniappan e Awang (2008) desenvolveram um protótipo de Sistema de previsão de doenças cardíacas aplicando técnicas de mineração de dados em sua modelagem. A principal motivação da pesquisa se baseia em como criar um mecanismo capaz de transformar dados em informações úteis para permitir aos profissionais de saúde melhorar significamente seu poder de tomada decisões.

As técnicas Naive Bayes, Árvores de Decisão e Redes Neurais foram as técnicas utilizadas para descobrir e extrair padrões e relacionamentos entre os dados. A base de dados utilizada contém registros histórico de doenças cardíacas da *Cleveland Heart Disease*, e possui um total de 909 registros e 15 atributos médicos. Os registros foram divididos em dois conjuntos de dados, o primeiro de treinamento contendo 455 registros, e o segundo de teste, com 454 registros. Os registros foram atribuídos a cada grupo de forma aleatória.

Foram escolhidos cinco objetivos de mineração, para que cada técnica fosse avaliada a fim de descobrir qual é a mais eficaz em cada objetivo. As técnicas escolhidas foram treinadas e validadas com um conjunto de dados de teste. Os métodos *Lift Chart* e *Classification Matrix* foram utilizados para avaliar a eficácia de cada técnica. No final do teste, os pesquisadores concluíram que as três técnicas são capazes de extrair padrões em resposta ao estado previsível e podem responder a perguntas complexas, cada uma com sua própria força em relação à facilidade de interpretação da técnica. Naive Bayes responde a quatro dos cinco objetivos, árvores de decisão atende a três e rede neural a dois.

O Naive Bayes, de acordo com os resultados do trabalho, é o modelo mais eficaz para prever pacientes com doença cardíaca conforme mostrado pelos dados exibidos pelos autores. O protótipo desenvolvido para a pesquisa conseguiu responder a questões complexas que os sistemas tradicionais não conseguiam, assim permitindo aos profissionais de saúde melhorar as tomadas de decisões clínicas.

3.2 Mineração e visualização de dados para suporte à decisão em recursos públicos de saúde

Lavrač et al. (2007) apresentaram uma aplicação de Mineração de Dados e suporte de decisões em saúde pública. A aplicação foi realizada na Eslovênia em um projeto chamado MediMap, cujo objetivo é configurar os melhores modelos e ferramentas para apoiar as decisões relacionadas aos cuidados de saúde. O projeto visava ser um modelo de referência para institutos regionais de saúde pública.

Os dados disponíveis foram obtidos a partir de 11 centros comunitários de saúde da região de Celje na Eslováquia. O conjunto de dados consistia em três bases de dados, uma de prestadores de serviços de saúde, uma contendo as estatísticas de serviços de saúde ambulatorial e uma com o status médico. Para modelar os processos do fluxo de pacientes em instituições particulares, utilizaram dados adicionais que descrevem o direcionamento dos pacientes para outras instituições ou especialistas.

A abordagem do projeto foi feita em duas etapas. Na primeira, foram analisados os dados disponíveis com as técnicas de mineração de dados. Na segunda, foram utilizados os resultados da mineração de dados para um estudo mais elaborado usando técnicas de apoio à decisão. A primeira fase foi focado no problema de direcionar os pacientes dos centros de saúde primários a especialistas. A segunda fase foi composta de estudos sob os aspectos organizacionais dos recursos de saúde pública na região de Celje na Eslováquia, com o objetivo de identificar as áreas atípicas em termos de disponibilidade e acessibilidade dos serviços públicos de saúde.

A principal conquista do projeto foi a criação do modelo de disponibilidade e acessibilidade dos serviços de saúde à população de uma determinada área. Com a aplicação do projeto, foi possível identificar regiões que diferem da média, e consequentemente, explicar as causas de tais situações. Além disso, o instituto público de saúde nacional utilizou os resultados para identificar os dados ausentes que devem ser incluídos no protocolo aprimorado de coleta de dados de saúde pública no nível nacional.

3.3 Aplicação de Data Warehouse em saúde

Duarte et al. (2008) desenvolveram um projeto de Data Warehouse destinado à gestão da saúde baseado na necessidade de um ambiente que permitisse análise de dados confiáveis de forma eficiente, para prover uma melhoria no processo de tomada de decisões, aumentando, consequentemente, a competitividade e lucratividade de suas corporações,

Para desenvolver o projeto a equipe utilizou CDs contendo dados de internações dos anos de 2001 a 2006 do sistema AIH (Autorização de Internação Hospitalar). A primeira etapa da metodologia do projeto foi a limpeza dos dados brutos, que consistiu na conversão dos arquivos no formato DBF para o formato CSV, onde foram eliminadas as colunas desnecessárias e feito o carregamento no Oracle. Na etapa de Redução e Pré-

Processamento, obtou-se por realizar a carga somente dos dados referentes aos estados da região Sudeste. Os dados reduzidos foram então carregados para o modelo transacional no schema AIH.

A etapa de transformação consistiu no carregamento do modelo dimensional, realizada através da ferramenta *Oracle Warehouse Builder*. A etapa de Mineração de Dados foi feita usando a ferramenta *Oracle Discoverer* para demonstrar o tipo de informação que poderia ser obtida no DW. Relatório sobre dengue e relatório sobre pneumonia são exemplos de alguns dos relatórios que foram extraídos do DW.

O Data Warehouse, mesmo hospedado em um notebook, apresentou excelentes resultados para as consultas submetidas com um excelente desempenho no processamento. Além disso, mostrou-se capaz em realizar consultas que o ambiente de análise estatística do DATASUS não foi capaz de realizar. A partir dos resultados, concluiu-se que um Data Warehouse hospedado em um servidor de grande capacidade pode armazenar dados unificados de todo país e responder consultas analíticas importantes para o processo de tomada de decisão para o estabelecimento de estratégicas de saúde pública no Brasil.

Em outro trabalho com a aplicação de Data Warehouses, Gonçalves Diana e Santos (2010) desenvolveram um sistema de inteligência de negócios (*Business Intelligence*) integrando mecanismos de coleta, análise e relatórios de dados. O objetivo do projeto foi mostrar o aumento da qualidade de vida e a incidência de complicações e efeitos colaterais após a cirurgia simpatectomia, realizada no tratamento da hiperidrose (transpiração anormalmente aumentada).

Foram analisados dados de 277 pacientes, coletados a partir de questionários na web (*SF-36 Health Survey Questionnaire*), com várias questões relacionadas a hiperidrose, às complicações associadas à cirurgia e aos efeitos colaterais. Os dados disponíveis foram armazenados em um Data Mart, e analisados com a tecnologias *On-Line Analytical Processing* (OLAP) e Data Mining.

As diversas análises realizadas, em termos do OLAP, permitiram verificar o início do distúrbio, a incidência de hiperidrose antes e após a cirurgia, as mudanças no estado emocional dos indivíduos, a ocorrência de hiperidrose compensatória como consequência da cirurgia e a melhoria da qualidade de vida geral associada a esse conjunto de dados. Em relação a mineração de dados, o resultado apresentou alta precisão, ajudando a apoiar o processo de tomada de decisão. Como resultado geral, a equipe de desenvolvimento conseguiu alcançar seu objetivo e mostrou o aumento da condição geral de saúde e qualidade de vida dos pacientes após a cirurgia, além de demonstrar a aplicabilidade dos conceitos e técnicas de análise de dados utilizados.

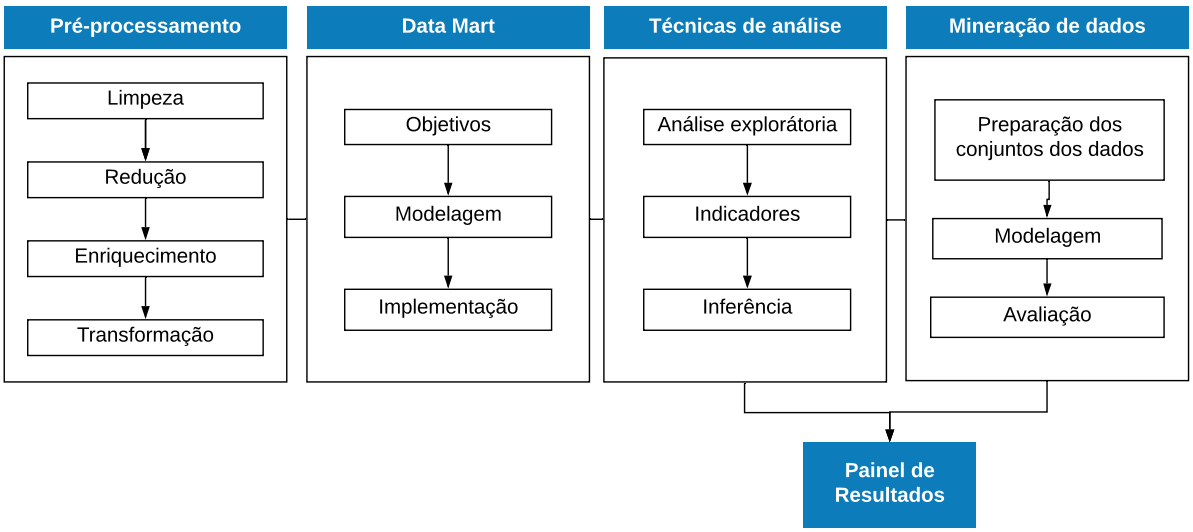
4 METODOLOGIA

Este capítulo apresenta a estrutura metodológica adotada para o desenvolvimento do trabalho, tomando como base uma proposta de construção de um ambiente de data mart, a aplicação de técnicas de análises e mineração de dados sob o conjunto de dados da Hanseníase no Estado do Tocantins. Serão apresentados também o passo a passo da execução da metodologia, assim como as ferramentas adotados para cada uma das etapas de execução do projeto.

Conforme abordado anteriormente, um *Data Warehouse* é um conjunto de dados organizados em *Data Marts* (DM) cujos principais objetivos são auxiliar os processos de tomada de decisão e oferecer uma melhor gestão sob o grande volume de dados. A mineração de dados é definida como o processo responsável pela busca de conhecimentos implícitos nas bases de dados. A mineração de dados auxilia na extração de novos padrões significativos que podem não ser necessariamente encontrados apenas ao processar e consultar dados no data warehouse. Portanto, as aplicações da mineração de dados devem ser fortemente consideradas no desenvolvimento de um DW (ELMASRI; NAVATHE, 2011).

Neste trabalho, as aplicações de mineração de dados foram aplicadas na base de dados do SINAN, cujo dados inicialmente foram carregados em um Data Mart, e em seguida, aplicado técnicas de análises, como a análise exploratória, indicadores e inferência. Após todas as análises, foi criada uma ferramenta de visualização de dados para a apresentação dos resultados obtidos. A figura 7 apresenta uma visão geral da metodologia executada.

Figura 7 – Fluxograma de execução da metodologia



Fonte: Criado pelo autor.

4.1 Processo de Desenvolvimento

As etapas sequenciais para a execução das atividades foram organizadas como mostra a figura 7. Essa organização auxilia a equipe a verificar quais etapas foram executadas e quais objetivos foram alcançados na execução do trabalho. As próximas subseções apresentam o desenvolvimento deste trabalho.

4.1.1 Coleta da base de dados

A Resolução nº 510/2016 do Conselho Nacional de Saúde visa assegurar o respeito pela dignidade humana e a proteção devida aos participantes das pesquisas científicas envolvendo seres humanos (BRASIL, 2016c). Para atender a resolução, foi solicitado ao Comitê de Ética da Universidade Federal do Tocantins a autorização para ter acesso ao Banco de Dados do Sistema de Informação de Agravos de Notificação (SINAN). A base de dados é constituída por dados sobre casos de pessoas com hanseníase, portanto a pesquisa envolve seres humanos. A base de dados obtida contém os dados do preenchimento da Ficha de notificação compulsória de Hanseníase, mostrado na figura 1. A base recebida é composta por 91 atributos e 21.952 registros.

4.1.2 Seleção do modelo de processo

Uma das primeiras etapas do desenvolvimento do projeto foi escolher o modelo de processo de construção do ambiente de mineração de dados a partir do contexto de aplicação. Atualmente vários modelos definem as fases e atividades do Data Mining, apesar de cada um possuir suas particularidades, todos têm em comum a mesma estrutura. Para o desenvolvimento deste trabalho, foi escolhido CRISP-DM (*Cross-Industry Standard Process of Data Mining*).

4.1.3 Pré-processamento

O etapa de pré-processamento é um conjunto de atividades que envolvem a preparação e organização dos dados. Trata-se de uma etapa fundamental que precede a realização das análises e classificações. A etapa chamada de pré-processamento, é composta por 4 fases: Limpeza de dados, Redução, Enriquecimento e Transformação. A execução de cada uma dessas atividades são apresentadas a seguir.

4.1.3.1 Levantamento dos dados

Nesta sub etapa foi solicitado o acesso ao conjunto de dados empregado no desenvolvimento do trabalho através da submissão da proposta de projeto ao comitê de ética. Após ter obtido o acesso a base de dados foi identificado todos os atributos, totalizando

91, e foi construído o dicionário da base de dados, onde é apresentado na figura abaixo a primeira página do arquivo.

Figura 8 – Primeira página do dicionário de dados

Nome do Campo	Tipo	Categoria	Descrição	Características	DBF
Nº da Notificação	varchar2(7)		Número da Notificação	Campo Chave para identificação do registro no sistema	nu_notific
Tipo de Notificação	varchar2(1)	1 – Negativa, 2 – Individual , 3 – Surto 4 – Agregado	Identifica o tipo da notificação	Campo Obrigatório	tp_not
Agravo	varchar2(4)	Tabela de agravos do sistema com códigos (classificação internacional de doenças –CID 10)			id_agravo
Data da Notificação	date	dd/mm/aaaa	Data de preenchimento da ficha de notificação.	Campo Chave	dt_notific
Nome do Paciente	varchar2(70)		Nome completo do paciente	Campo Obrigatório	nm_pacient
Data de nascimento	date	dd/mm/aaaa	Data de nascimento do paciente	Campo Obrigatório	dt_nasc
Idade		A composição da variável obedece o seguinte critério: O dígito: 1º. Hora 2º. Dia 3º. Mês 4º. Ano Ex: 3009 – nove meses, 4018 – dezoito anos	quando não há data de nascimento a idade deve ser digitada segundo informação fornecida pelo paciente como aquela referida por ocasião da data dos primeiros sintomas ou na falta desse dado é registrada a idade aparente		nu_idade_n
Sexo	varchar2(1)	M- Masculino F- Feminino I- Ignorado	Sexo do paciente		cs_sexo
Gestante	varchar2(1)	1 - 1º Trimestre 2 - 2º Trimestre 3 - 3º Trimestre 4 - Idade gestacional ignorada	Idade gestacional da paciente	Campo Obrigatório	cs_gestant

Fonte: própria autora, 2021

4.1.3.2 Redução dos dados

Nessa etapa o primeiro passo foi a criação de uma tabela onde continham o percentual de preenchimento de cada atributo, em especial aos campos não preenchidos, a fim de saber quais atributos possuíam informações adequadas para serem selecionados na mineração de dados. E em seguida foi feita uma seleção a partir do dicionário de dados e do percentual de preenchimento de quais atributos eram importantes para a descoberta de conhecimento.

Na etapa de redução vertical foram excluídos alguns atributos do tipo texto sem categoria, por não ser possível a sua utilização futura para os algoritmos de Mineração de Dados. Alguns dos atributos excluídos foram: Nome (NM_PACIENT), Nome da mãe (NM_MAE_PAC), Número do logradouro (NU_NUMERO), Complemento do logradouro (NM_COMPLEM), Ponto de referencia (NM_REFEREN), DDD(NU_DDD_TEL) e Telefone (NU_TELEFON).

Outras 16 variáveis também foram excluídas por ter mais de 90% dos dados faltantes, algumas delas são: Data de transferência da unidade de saúde (DT_TRANSUS), Número do lote vertical (NU_LOTE_V), Fluxo de retorno (CS_FLXRET), Distrito (ID_DISTRIT), Campo para GeoReferenciamento (ID_GEO1), Vinculação (IN_VINCULA),

Distrito de residência atual (DISTRIT_AT) e Identifica migração windows (MIGRADO_-W).

Na etapa de redução horizontal foi necessário a realização de uma análise de dados que resultou na exclusão de alguns registros, como no caso dos anos de notificação que antes se iniciava nos anos de 1998, pelo fato de existirem poucos registros referentes ao anos de 1998, 1999 e 2000, a exclusão desses registros foi necessária. Por se tratar de uma análise no contexto do estado do Tocantins, foram excluídos também 217 registros referentes a notificação do estado do Pará e do Maranhão que estavam presentes na base. Logo, resultou-se na exclusão total de 233 registros.

A base de dados original continha 91 atributos e 21.952 registros, sendo divididos em dados da notificação, do paciente, da doença, do tratamento e da localização. Após a redução dos dados, o número de atributos foi reduzido para 30, e os registros para 21.719.

4.1.3.3 Limpeza dos dados

A base de dados recebida continham muitas partes irrelevantes e ausentes, para resolver essa situação foi feita uma limpeza de dados envolvendo o preenchimento de dados ausentes e redução de ruídos com técnicas recomendadas. As variáveis Gestante, Raça e Escolaridade continham vários registros com dados ausentes, e por se tratar de atributos importantes para a descoberta de conhecimento, não seria uma boa alternativa excluir esses atributos, e por serem atributos categóricos, não seria interessante substituir os valores pela média ou mediana, logo a estratégia utilizada para resolver esse problema foi substituir os dados faltantes por uma das categorias da variável. Por exemplo: na variável Gestante existem 6 categorias: 1 Trimestre, 2 Trimestre, 3 Trimestre, Idade gestacional ignorada, não se aplica e ignorado. Logo, dessa forma foi adicionado a categoria ignorado nos registros faltantes da variável Gestante. E a mesma lógica foi seguida para as variáveis Raça e Escolaridade.

Na variável bairro, outra informação muito importante para a descoberta de conhecimento, também continham registros com dados ausentes, e por se tratar de uma variável qualitativa, esse caso foi resolvido adicionando a mediana. Na variável número de lesões foi adicionada a média nos registros com dados faltantes.

4.1.3.4 Enriquecimento dos dados

Na etapa de classificação são utilizado atributos do tipo categórico, e para ter um bom resultado na classificação é interessante que as variáveis selecionadas tenham uma quantidade essencial de categorias. Por esse motivo duas variáveis foram adicionadas na base para substituir o atributo cidade, que continham 139 grupos; e o atributo ano de notificação, que continham 16 grupos. A variável cidade foi substituída por 11 regiões geográficas imediatas, que distribuíram as 139 cidades do estado do Tocantins em micror-

regiões segundo a divisão do Instituto Brasileiro de Geografia e Estatística (IBGE) vigente desde 2017. A variável ano de notificação foi agrupada a cada 2 anos, por exemplo, os registros referente ao ano de 2001 e 2002 passaram a ser apenas uma categoria, ao invés de duas.

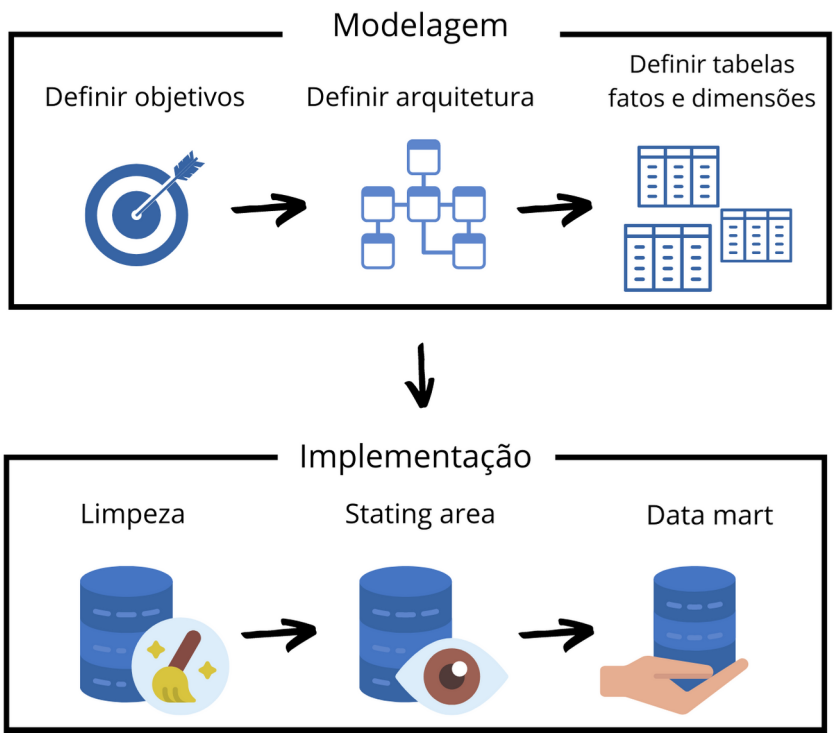
4.1.3.5 Transformação dos dados

A Base de Dados recebida continham todas as categorias em valores numéricos. Foi realizada uma transformação nesses dados para valores nominais, a fim de obter um melhor entendimento nas saídas dos resultados. Após a finalização do pré-processamento, a base está pronta para ser carregada no Data Mart. Todas as etapas da construção do data mart até a análise dos dados são apresentadas a seguir.

4.1.4 Construção do Data Mart

Para desenvolver um DM não existe um manual pronto. Existem ferramentas e diferentes conjuntos de etapa, todas abordam desde a extração e análise de dados até a fase de gerenciamento (DOMENICO, 2001). As fases de construção do *Data Mart* definidas para este projeto foram:

Figura 9 – Processo de construção do Data Mart



Fonte: própria autora, 2021

4.1.4.1 Modelagem

Essa etapa consiste em escolher a arquitetura, definir as tabelas fatos e dimensões, e definir seus atributos, mas para que isso seja feito de forma coerente é necessário definir quais são os objetivos para a criação do data mart.

1. Objetivos: os principais objetivos definidos para serem alcançados com o data mart foram: fornecer dados limpos, organizados, consistentes e permitir análise dos dados em várias dimensões.

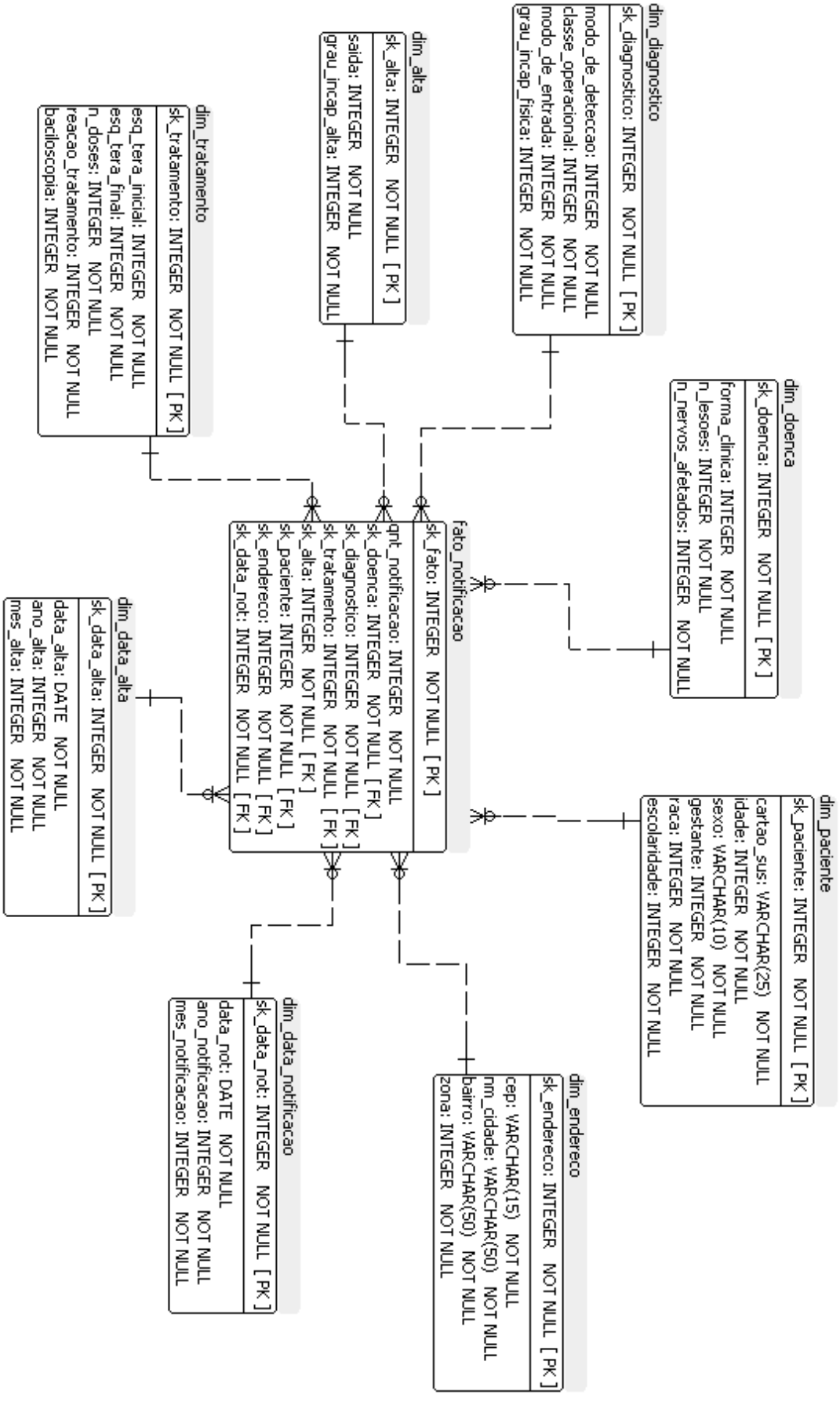
2. Modelo dimensional: de acordo com os objetivos definidos acima e a abordagem proposta neste trabalho, será utilizada o modelagem de esquema estrela, pois torna mais fácil e rápido o processo de consultas nos dados.

3. Tabelas fatos e dimensões: visando construir um esquema estrela que caracterizasse os dados de forma clara e organizada, o modelo multidimensional proposto foi fundamentado em uma tabela fato e suas oito dimensões.

   A tabela fato retrata o evento de notificações de hanseníase, é a responsável pelo armazenamento das métricas e chaves estrangeiras das dimensões sobre as notificações, e sua métrica é a quantidade de notificações feitas. As tabelas dimensões descreve o fato ocorrido (as notificações), cada tabela dimensão contém dados sobre um tópico da notificação, que são: dados do paciente, endereço, diagnóstico, tratamento, tipo de saída, e os temporais: data da notificação, data do diagnóstico e data da saída.

   A modelagem da tabela fato e suas dimensões é representada na figura abaixo:

Figura 10 – Modelagem do esquema estrela



Fonte: própria autora, 2021

4.1.4.2 Implementação

Essa etapa consiste na limpeza, extração, transformação e carregamento dos dados em um data mart, normalmente chamada de ETL (Extract, Transform, Load). Como os dados já foram limpos na seção de pré-processamento, essa etapa consistiu apenas no carregamento dos dados no data warehouse.

1. *Stating Area*: o primeiro passo da implementação foi criar dois servidores de banco de dados, um destinado para a *stating area* e outro para o data mart final, para essa tarefa foi utilizado o PgAdmin4 do programa PostgreSQL, apresentado na seção 4.2. O próximo passo foi criar uma conexão do banco com a plataforma DBeave (seção 4.2), responsável pelo gerenciamento e administração dos dados. Para fazer o carregamento na *stating area*, foi feita uma importação dos dados da base que estava no formato csv para a servidor utilizando o Dbeave.

   Como definido na seção 2.5, é na *stating area* onde acontece a transformação dos dados, como seleção dos atributos, tratamento das células vazias e etc. Como essa tarefa já foi realizada no pré-processamento, a *stating area* construída tem como objetivo ser uma tabela de teste para a validação de dados e para auxiliar a transição dos dados da origem para o destino final no DM.

2. *Data Mart*: Após a criação da *stating area*, carregamento e validação dos dados, foi feita a modelagem física das tabelas, para essa tarefa foi utilizado o programa *SQL Power Architect*, apresentado na seção 4.2. O próximo passo foi realizar a conexão da modelagem no servidor do data mart criado, após a conexão, foram criadas automaticamente as tabelas físicas no banco. Para o processo de carregamento dos dados no data mart, foi selecionado os dados da *stating area* e inseridos no DM.

4.1.5 Técnicas de Análises

Após os dados carregados no *data mart*, foi feita a conexão do banco de dados na ferramento do Power Bi (apresentado na seção 4.2, onde foi executado três técnicas de análise sob os dados. A primeira foi a análise exploratória, com o objetivo de encontrar padrões explícitos na base. A segunda é chamada de indicadores, que são as análises feitas pelos profissionais da área. E a terceira foi uma inferência com teorema de bayes para incrementar os resultados do trabalho com a dedução de algumas probabilidades.

- Análise exploratória: nessa etapa foi calculado a distribuição das frequências dos principais atributos da base de dados e gerado um painel de resultados com visualização gráfica a partir dos valores dos atributos quantitativos e qualitativos do conjunto de dados. Os resultados da análise são apresentados na seção 5

- Indicadores: as fórmulas dos cálculos dos indicadores apresentados na seção 2.2.1 foram executadas no Power Bi utilizando as variáveis selecionadas baseada na padronização do Ministério da saúde.

- Inferência: o teorema de bayes foi aplicado no Power BI utilizando as fórmulas DAX (apresentadas na seção 4.2). O teorema é composto por probabilidades condicionais, para cada parte do teorema (Numerador e Denominador) foi criado uma fórmula dax, e no final uma última formula dax que montava a fórmula do teorema. Dessa modo, foi possível adicionar os resultados da inferência no painel de resultados, assim, possibilitando a análise das probabilidades obtidas baseadas por ano ou cidade, que foram adicionados como filtros.

4.1.6 Processamento da Mineração de Dados

Este trabalho foi desenvolvido com foco na tarefa de classificação. A classe selecionada para a mineração foi a variável classificação operacional, que determina o nível da Hanseníase. O diagnóstico é feito baseada no número de lesões cutâneas de acordo com os seguintes critérios: Paucibacilar (PB), são casos com até 5 lesões de pele e Multibacilar (MB), são casos com mais de 5 lesões. Esse diagnóstico define o esquema terapêutico do tratamento.

Além disso, a partir da classificação operacional, também é possível avaliar os casos em risco de desenvolver complicações e indicar diagnóstico tardio ou precoce. Por esses motivos, essa variável foi escolhida como classe. Os atributos selecionados para constituir a base de classificação foram os dados referente a características dos pacientes e do enderenço, que são: Sexo, Idade, Raça, Escolaridade, Ano de notificação, Zona de residência, e Regiões Geográficas do Estado, e a Classe, Classificação Operacional.

O objetivo da mineração é analisar e tentar encontrar algum padrão entre essas características, além de verificar se as mesmas influenciam na classificação operacional, pois assim, seria possível utilizar essas informações para melhorar os diagnósticos precoce, e conseguintemente ter um possível aumento no controle da hanseníase no Tocantins.

Neste trabalho foi escolhido o CRISP-DM como modelo de processo da mineração de dados. O processo consiste de seis fases organizadas: Pesquisa dos Dados, Entendimento dos Dados, Preparação dos Dados, Modelagem, Avaliação e Implantação. As duas primeiras fases já foram desenvolvidas, como mostra na seção 4.1.3. A seguir está sendo demonstrado o desenvolvimento dos próximos passos da mineração de dados.

1. Preparação dos Dados: foram extraídos do data mart as dimensões paciente, endereço e diagnóstico, que continham os atributos utilizados na mineração de dados.

2. Modelagem: nesta fase foi executada a classificação dos dados utilizando os algoritmos de mineração que foram escolhidos: *Naive Bayes e Random Tree*. Esse processo

foi realizado utilizando a ferramenta weka, apresentada na seção 4.2.

3. Avaliação: nesta fase foi realizado a analise das métricas dos resultados alcançados, com o objetivo de obter a confiabilidade nos modelos obtidos pela classificação. Essa analise é apresentada na seção 5

4. Implantação: após executado os modelos, o conhecimento adquirido é analisado e validado para ser apresentado de uma maneira que os interessados possam utilizar.

4.1.7 Apresentação dos Resultados

Nesta última etapa do desenvolvimento foi construído um mecanismo para apresentação dos resultados obtidos no processamento da mineração de dados e da análise do Data Mart. Para realizar essa etapa foi utilizado a ferramenta power BI, os resultados foram distribuídos por seção: informações do paciente, do endereço, das unidades de saúde e etc. No painel de resultado também foi aplicado alguns filtros, como o ano da notificação e a cidade. Alguns dos gráficos presente no painel será apresentado na seção 5

4.2 Ferramentas

- WEKA

Uma das ferramentas utilizadas no desenvolvimento do projeto foi o WEKA (*Waikato Environment for Knowledge Analysis*), que possui uma biblioteca de algoritmos para mineração de dados. O ambiente de trabalho inclui algoritmos para regressão, classificação, agrupamento, regras de associação e seleção de atributos. A ferramenta permite que os usuários testem e comparem diferentes métodos de aprendizagem de máquina nos conjuntos de dados, e sua arquitetura modular permite que processos sofisticados de mineração de dados sejam construídos (HALL et al., 2009).

A saída da mineração de dados no weka gera diversas informações e métricas que são separadas por uma estrutura básica que varia de acordo com o algoritmo usado. A seguir serão definidas cada uma das seções e as métricas obtidas. Assim, é possível alcançar um melhor entendimento da avaliação dos resultados obtidos na análise (DEPARTMENT, 2021).

- *Run information* (Informações de execução): essa seção mostra um resumo da execução e dos dados, como o nome da base de dados, a quantidade de registros, e os atributos que serão utilizados; também é relatado o modelo de teste utilizado.

- *Classifier Model* (Modelo Classificador): essa seção apresenta o modelo obtido e o tempo de execução. Esse caso tem diferentes saídas, depende do algoritmo utilizado. Na utilização do naive bayes é exibida uma tabela relacionando cada atributo a classe

selecionada. No algoritmo random tree é gerado uma árvore de decisão relacionando os atributos.

- *Summary* (Sumário): mostra um resumo do desempenho do modelo, nessa seção é apresentada as seguintes métricas:
  - *Correctly Classified Instances*: número e porcentagem total das instâncias classificadas corretamente.
  - *Incorrectly Classified Instances*: número e porcentagem total das instâncias classificadas incorretamente.
  - *Kappa statistic*: medida estatística que tem por finalidade verificar o grau de confiabilidade intermediária. Segundo Simon (2005), uma possível interpretação dos valores seria: deficiente: menor que 0,20; justo: de 0,20 a 0,40; moderado: de 0,40 a 0,60; bom: de 0,60 a 0,80 e muito bom: de 0,80 a 1,00.
  - *Mean absolute error*: valor utilizado para medir o quão próximas as previsões estão dos resultados finais.
  - *Root mean squared error*: medida das diferenças entre os valores previstos por um modelo e os valores realmente observados.
  - *Relative absolute error*: expresso como uma razão, compara um erro médio aos erros produzidos pelo modelo.
  - *Root relative squared error*: medida relativa ao que teria sido se um classificador simples tivesse sido usado. Mais especificamente, esse classificador simples é apenas a média dos valores reais.
  - Total *Number of Instances*: total de registros utilizados no teste.
- *Detailed Accuracy By Class* (Precisão detalhada por classe): Exibe o desempenho do modelo por classe. Nessa seção há uma tabela onde cada linha representa uma classe, e a última linha contém a média. As colunas se refere as seguintes métricas:
  - TP rate (Taxa de *True Positive*): taxa de verdadeiros positivos, que são as instâncias classificadas corretamente.
  - FP rate (Taxa de *False Positive*): taxa de falsos positivos, que são as instâncias classificadas erradas.
  - *Precision* (Precisão): número de vezes que a classe 'x' foi classificada corretamente dividido pelo número de classes classificadas como 'x'.
  - *Recall* (Revogação): número de vezes que a classe 'x' foi classificada corretamente dividido pelo número de classes 'x' na base de teste.
  - *F-Measure*: Uma medida harmônica entre precisão e *recall*. Essa informação diz a performance do classificador.

  - MCC *(Matthews correlation coefficient)*: é usado no aprendizado de máquina como uma medida da qualidade das classificações binárias (duas classes).
  - ROC area (*Receiver Operating Characteristics*): Um dos valores mais importantes produzidos pela Weka. Eles dão uma ideia de como os classificadores estão se saindo em geral.
  - PRC area (*Precision Recall Curve*): uma relação entre as métricas *Precision e Recall.*

- *Confusion Matrix* (Matriz de confusão): a última seção mostra a matriz de confusão, que é uma tabela de contingência criada entre os dados de teste e a classificação feita pelo algoritmo.

- **Power BI**

O *Power Bi* é um conjunto de serviços de software, aplicativos e conectores que trabalham juntos para transformar as fontes de dados não relacionadas em informações coerentes, visualmente envolventes e interativas. Os dados utilizados no programa podem estar desde uma planilha do Excel a uma coleção de *data warehouses* locais ou baseados na nuvem. Com o Power BI, é possível conectar facilmente fontes de dados, visualizar e descobrir conteúdo importante (MICROSOFT, 2021). Uma função de destaque do *Power Bi* são as fórmulas DAX.

A DAX (*Data Analysis Expressions*) é uma linguagem de expressão de fórmula usada nos *Analysis Services*, no *Power BI* e no *Power Pivot* no Excel. As fórmulas incluem funções, operadores e valores para realizar cálculos avançados e consultas em dados nas tabelas e colunas relacionadas nos modelos de dados. (MICROSOFT, 2021)

- **PostgreSQL**

PostgreSQL é um Sistema de Gerência de Bancos de dados Relacional estendido e livre (SGBDRe), um servidor de banco de dados para o armazenamento seguro de informações. A ferramenta de gerenciamento do PostgreSQL é denominada PgAdmin, a ferramenta simplifica a criação, manutenção e uso de objetos de banco de dados, oferecendo uma interface de usuário limpa e intuitiva (GUERRERO, 2019).

- **Dbeaver**

DBeaver é uma ferramenta gráfica de gerenciamento de banco de dados gratuita e de código aberto para desenvolvedores e administradores de banco de dados. O DBeaver funciona com a maioria dos sistemas de gerenciamento populares, como MySQL, PostgreSQL, MariaDB, SQLite, Oracle e mais (DATABASE.GUIDE, 2018).

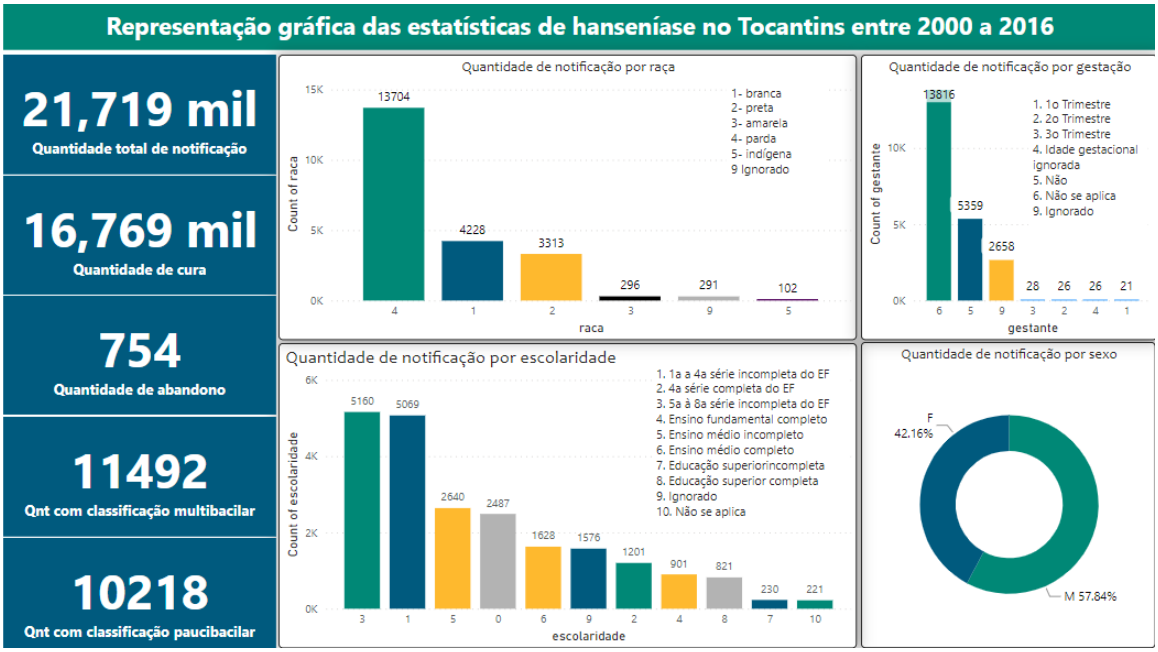
- **SQL Power Architect**

SQL Power Architect é uma ferramenta de modelagem de banco de dados relacional; foi criada por designers de data warehouse e oferece muitos recursos exclusivos voltados especificamente para o arquiteto de data warehouse. O programa permite que você faça engenharia reversa de bancos de dados existentes, execute modelagens de dados em bancos de dados de origem e gere metadados ETL automaticamente (BI, 2018).

5 RESULTADOS

Esse capítulo apresenta os resultados obtidos nas formas de análise utilizadas no trabalho, que são: Análise exploratória, Indicadores, Teorema de Bayes, e Mineração de dados. Os resultados das análises foram adicionados no painel, que está disponível para visualização clicando aqui (https://cutt.ly/BbDQOzA). A primeira página de resultados é apresentada a seguir:

Figura 11 – Primeira página do painel de resultados



Fonte: própria autora, 2021

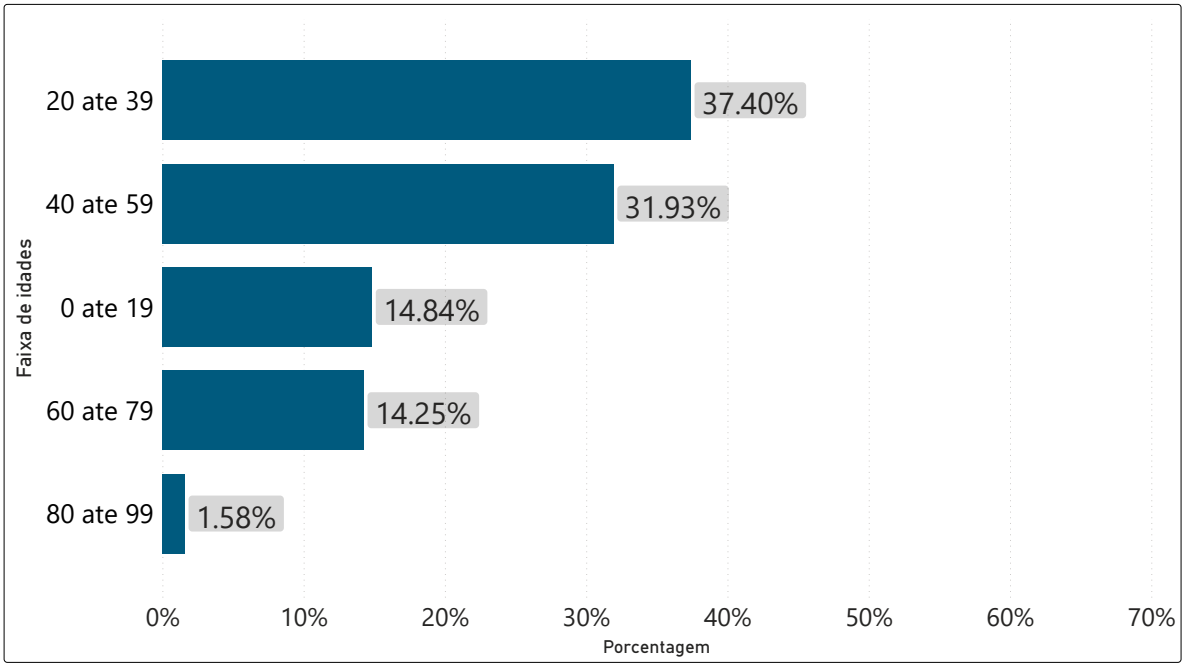
5.1 Análise exploratória

Os resultados da análise exploratória sob os dados do data mart para identificação das informações e padrões explícitos serão apresentados nesta seção. Os gráficos das frequências dos atributos serão apresentados por categoria: primeiro os dados dos pacientes, segundo as do endereço e da doença, e por último será apresentado a análise bidimensional.

- **Dados do paciente: Idade, sexo, raça, gestante e escolaridade**

A figura 12 apresenta a frequência das idades dos pacientes notificados com Hanseníase no período de 2001 a 2016 no estado do Tocantins. A variável idade foi dividida em grupos, as duas faixas de idade com mais casos são: entre 20 a 39 anos, com 7.550 notificações, e uma porcentagem de 37,40%; a segunda faixa é entre 40 a 59 anos, com

Figura 12 – Representação gráfica da frequência da variável idade

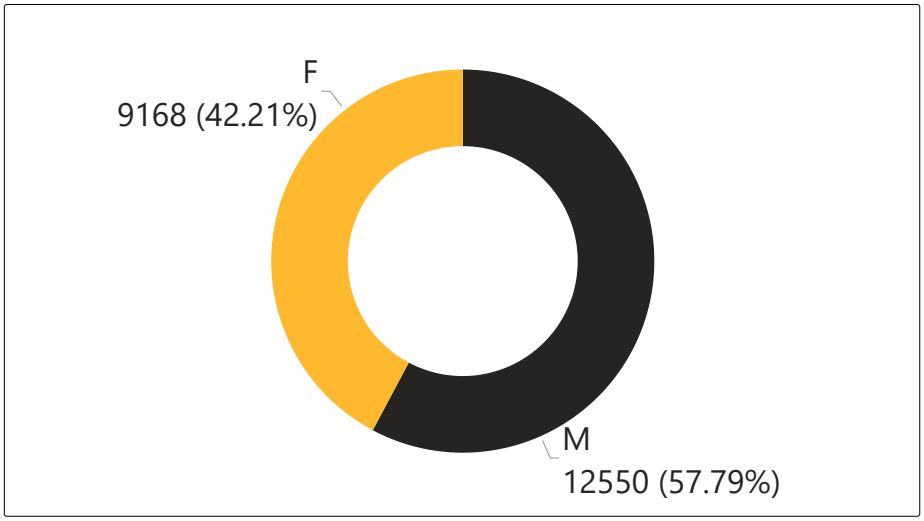


Fonte: própria autora, 2021

uma porcentagem de 31,93%. Essas duas faixas juntas representam mais de 50% do total de casos notificados no Tocantins entre 2001 a 2016.

Conforme os números do gráfico da figura 13, 57,79% dos pacientes notificados são do sexo masculino, com um total de 12550 notificações. Já o sexo feminino detém uma porcentagem de 42,21% do total geral das notificações.

Figura 13 – Representação gráfica da frequência da variável sexo

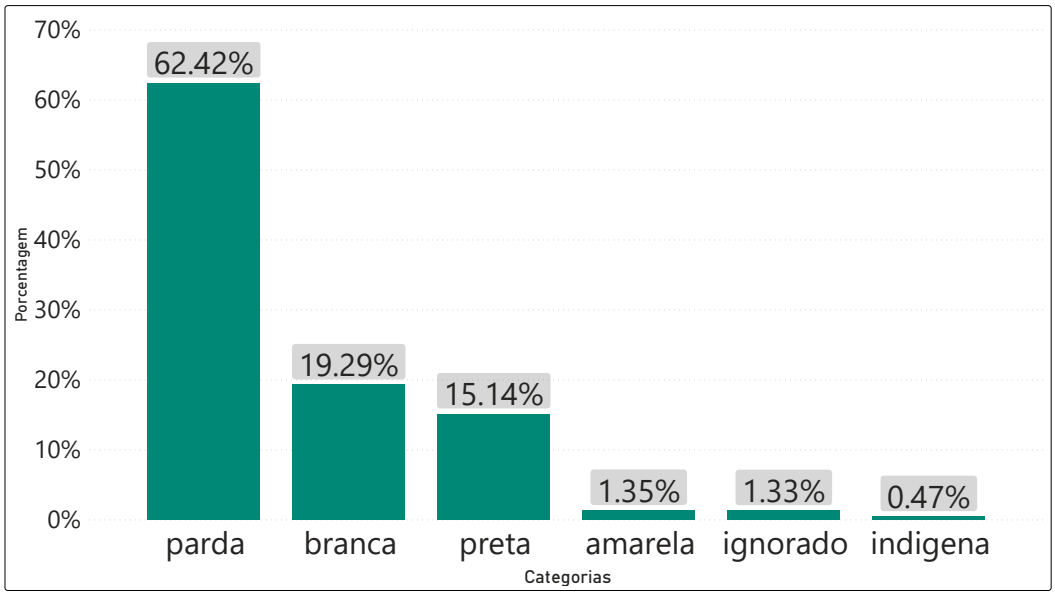


Fonte: própria autora, 2021

O gráfico da figura 14 mostra a frequência da variável raça, que contém 5 categorias: parda, branca, preta, amarela e indígena. Observa-se que mais da metade dos pacientes

notificados se declaram como pardos, com um percentual de 62,42%.

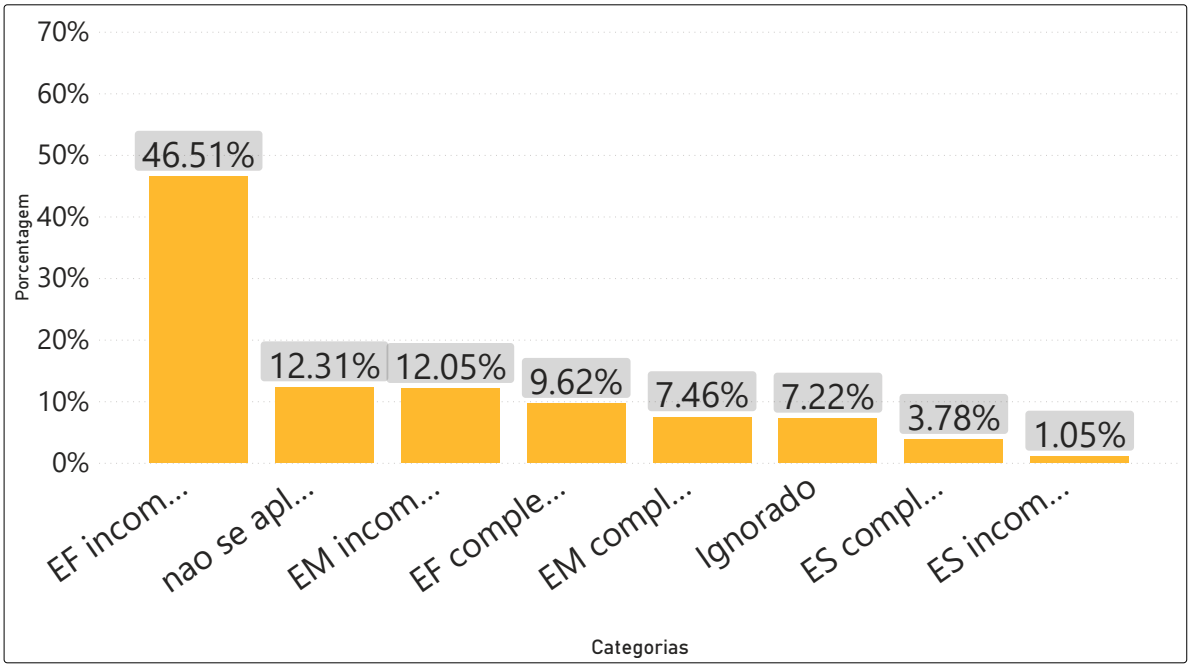
Figura 14 – Representação gráfica da frequência da variável raça



Fonte: própria autora, 2021

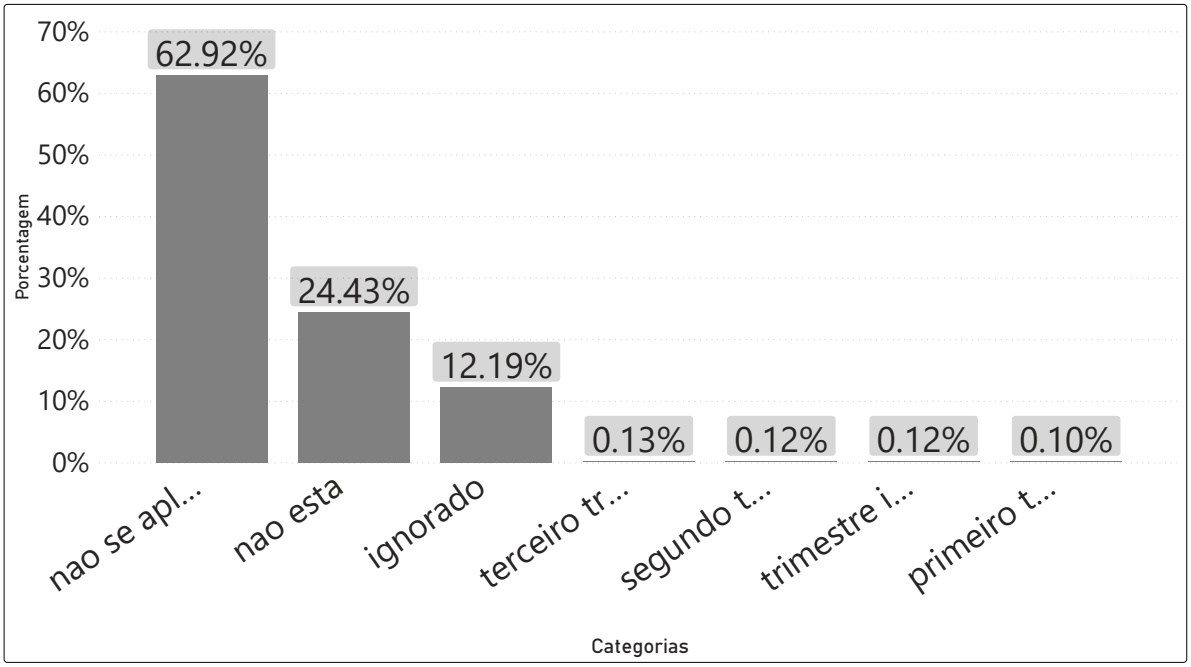
As duas próximas figuras referem-se a variável escolaridade e gestante, respectivamente.

Figura 15 – Representação gráfica da frequência da variável escolaridade



Fonte: própria autora, 2021

Figura 16 – Representação gráfica da frequência da variável gestante



Fonte: própria autora, 2021

Em relação a taxa de escolaridade dos pacientes notificados, apresentado no gráfico da figura 15, observa se que: 9989 pacientes (46%) não completaram o ensino fundamental; 2572 pacientes (12%) não chegaram a terminar o ensino médio e apenas 3,7% completaram o ensino superior.

No gráfico referente a variável gestante (figura 16, é possível observar que das 9.168 mulheres que notificaram hanseníase entre 2001 e 2016, um total de 101 entre elas estavam grávidas, ou seja, um 1%. Onde: 21 estavam no primeiro trimestre da gravidez, 26 no segundo, 28 no terceiro e 26 não informaram a idade da gestação.

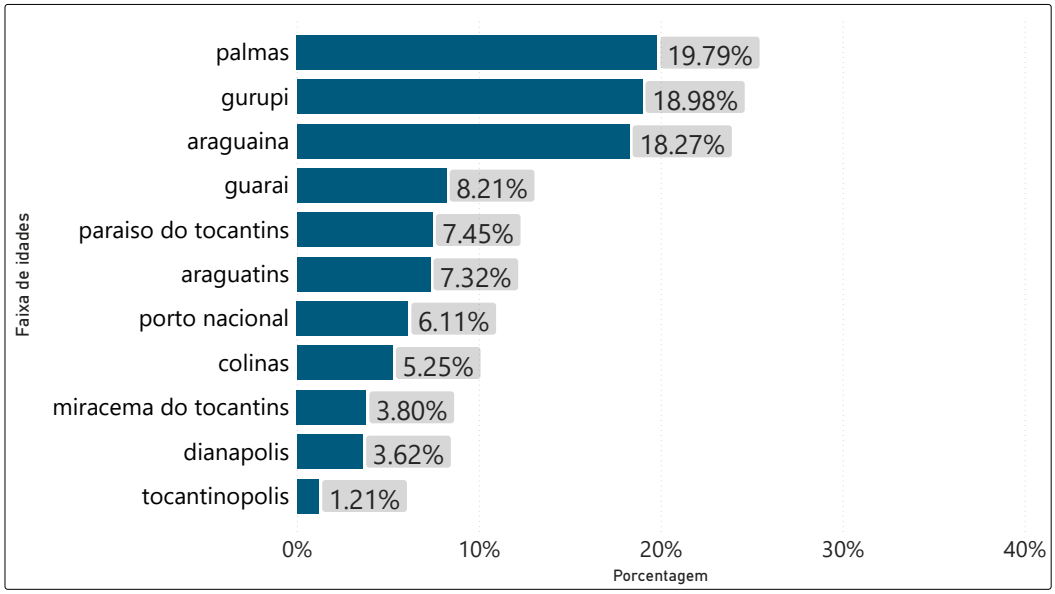
- **Dados das cidades por Microrregiões**

A figura 17 apresenta o gráfico com as quantidades de notificações por microrregiões do estado do Tocantins. A microrregião denominada Palmas, que contém 19,79% do total de notificações, é formada pelas seguintes cidades: Aparecida do Rio Negro, Lagoa do Tocantins, Lajeado, Lizarda, Mateiros, Novo Acordo, Palmas, Rio Sono, Santa Tereza do Tocantins e São Félix do Tocantins. Entre as cidades desse grupo, a cidade de Palmas dispõe de mais 50% do total.

Em segundo lugar fica a microrregião chamada de Gurupi, com um total de 3.784 notificações (18,98%). Esse grupo é composto pelas seguintes cidades: Aliança do Tocantins, Alvorada, Araguaçu, Cariri do Tocantins, Crixás do Tocantins, Dueré, Figueirópolis, Formoso do Araguaia, Gurupi, Jaú do Tocantins, Palmeirópolis, Paranã, Peixe, Sandolândia, São Salvador do Tocantins, São Valério, Sucupira e Talismã. Entre essas cidades,

as que se destacam com maiores quantidades de casos são: Gurupi com 2.041 notificações, Alvorada com 362 e Formoso do Araguaia com 354.

Figura 17 – Representação gráfica da frequência da variável microrregião

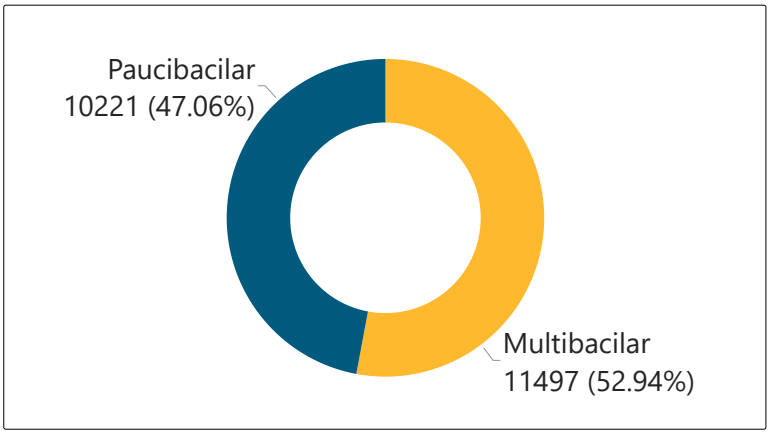


Fonte: própria autora, 2021

- **Dados da doença: classe operacional, forma clínica, modo de detecção, modo de saída e grau de incapacidade**

A figura 18 mostra a frequência da variável classificação operacional, que consiste na classificação da hanseníase, operacionalmente, para fins de tratamento. O tipo multibacilar atinge 11.497 (52,94%) do total dos pacientes, enquanto a paucibacilar atinge 10.221 pacientes(47,06%).

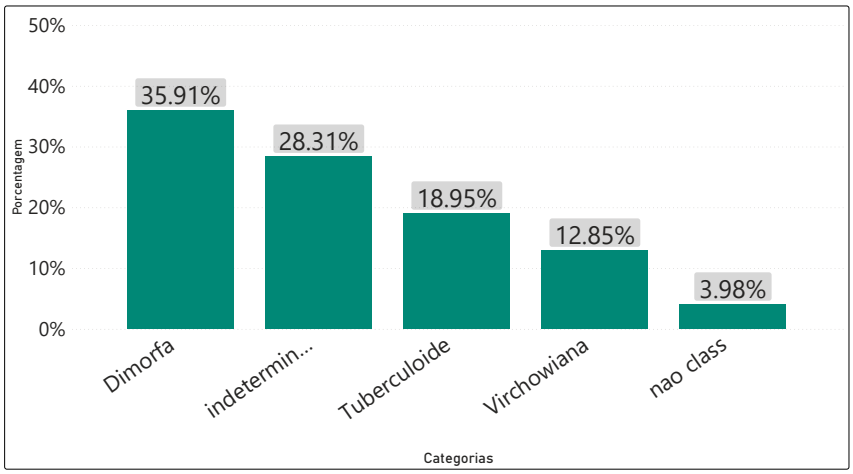
Figura 18 – Representação gráfica da frequência da variável classificação operacional



Fonte: própria autora, 2021

As classificações operacionais conhecidas da Hanseníase se dividem em 4 formas clínicas da doença: a Paucibacilar em Indeterminada e Tuberculoide, e a Multibacilar em Dimorfa e Virchowiana. O gráfico da figura 19 mostra a frequência de cada tipo, onde a forma Dimorfa representa 36% do total.

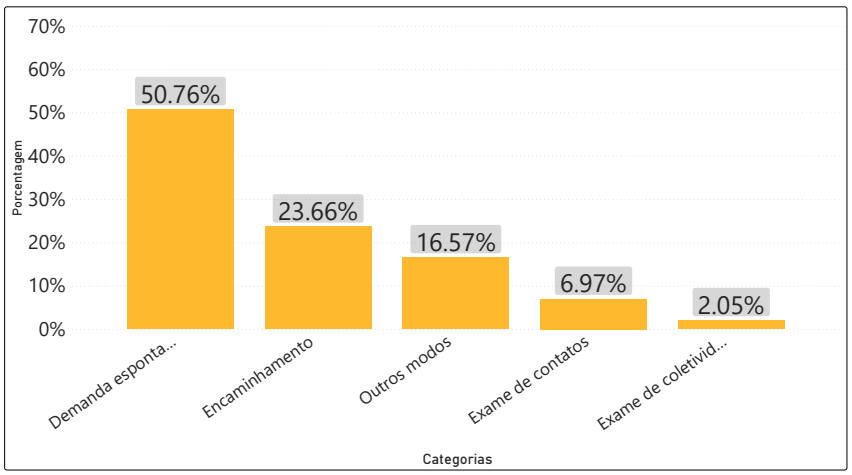
Figura 19 – Representação gráfica da frequência da variável forma clinica



Fonte: própria autora, 2021

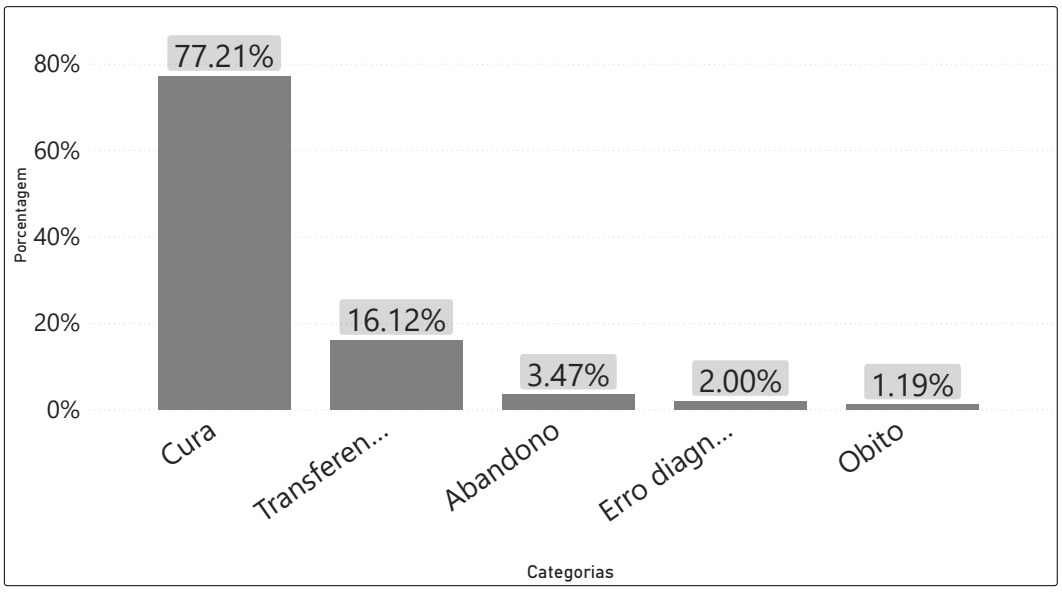
Os próximos gráficos, figura 20 e figura 21 referem-se as variáveis Modo de Detecção e Modo de Saída, respectivamente. A figura 20 mostra que o modo de detecção mais recorrente é a demanda espontânea do paciente, que obtém mais de 50% do total das notificações. O gráfico da figura 21 apresenta a quantidade de cada forma da saída do sistema de notificação, onde a Cura e a Transferência são os motivos de saída mais frequentes.

Figura 20 – Representação gráfica da frequência da variável forma de detecção



Fonte: própria autora, 2021

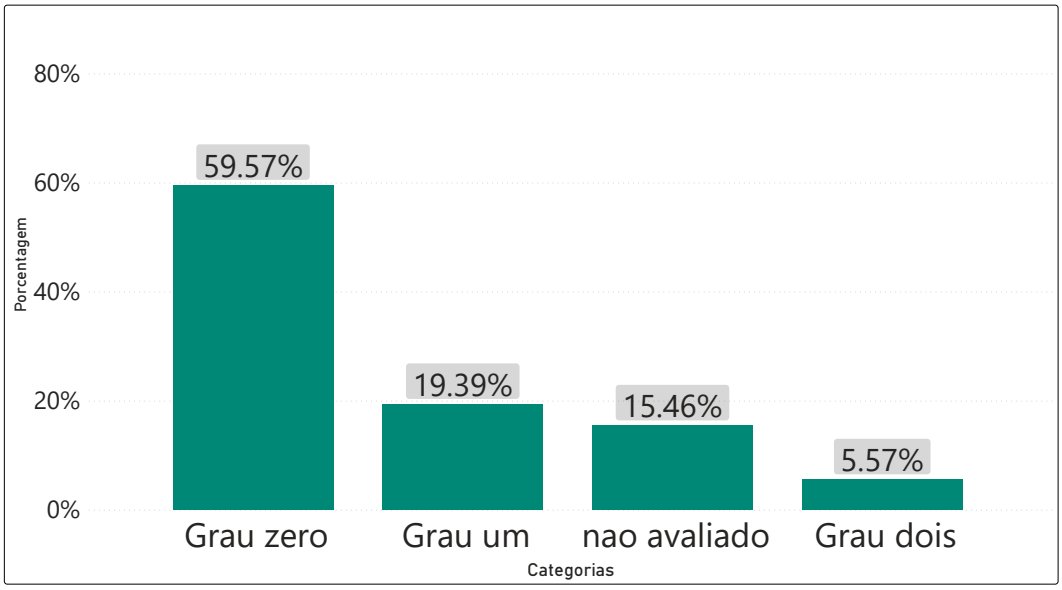
Figura 21 – Representação gráfica da frequência da variável forma de saída



Fonte: própria autora, 2021

O grau de incapacidade é composto por 3 categorias: grau 0, grau I e grau II. A frequência de cada tipo é apresentada na figura 22, que mostra que o Grau Zero tem uma maior presença entre os pacientes.

Figura 22 – Representação gráfica da frequência da variável grau de incapacidade



Fonte: própria autora, 2021

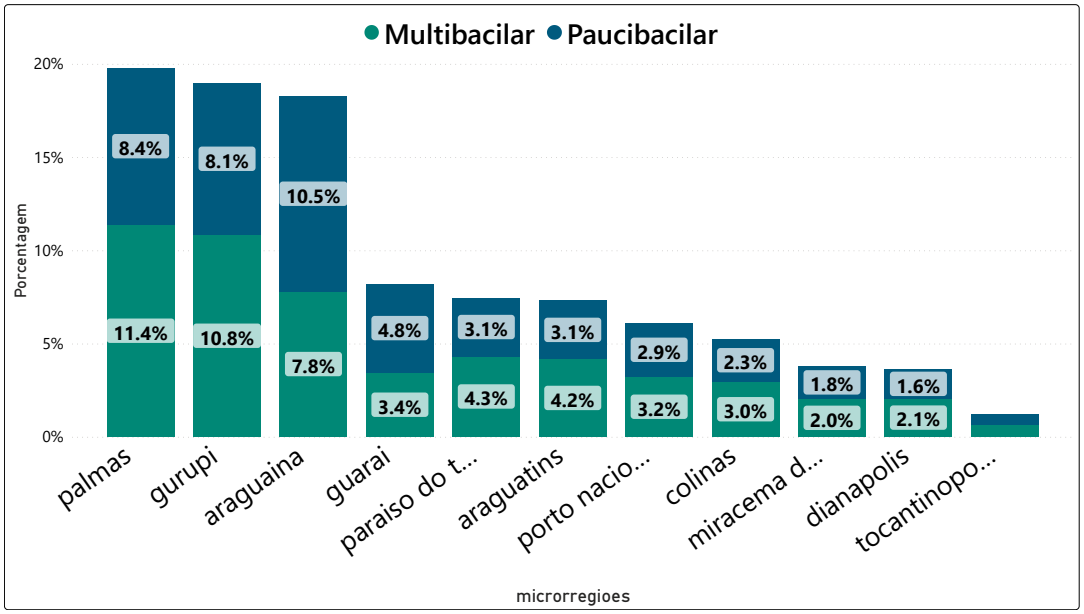
- **Análise Bidimensional**

Os próximos gráficos exibem os números da análise bidimensional, que consiste em analisar o comportamento conjunto de duas ou mais variáveis. Os atributos que foram selecionados para esse análise são: Microrregião x classificação operacional, Microrregião x Modo de detecção, Sexo x Idade, e Sexo X Forma clínica.

O primeiro gráfico (figura 23) faz análise da classificação operacional em relação as microrregiões, onde é possível obter mais detalhe a respeito do tipo da doença em cada região. Nessa associação, é possível observar que das 4.297 notificações presente na microrregião Palmas (19,8%), 2.471 delas são do tipo Multibacilar (11,4%), e 1.826 do tipo Paucibacilar (8,4%).

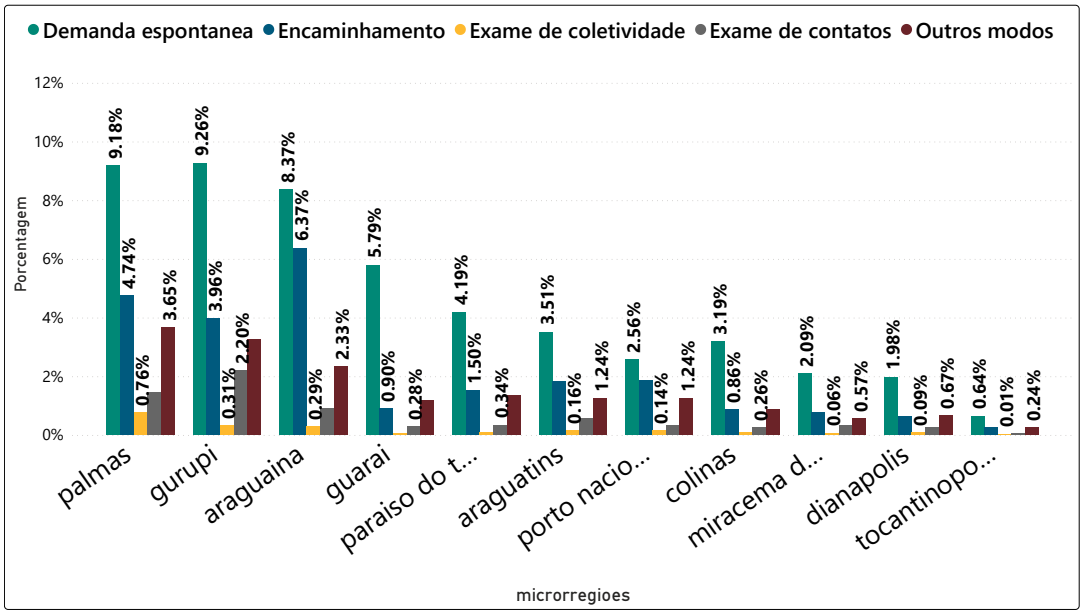
O segundo gráfico (figura 24) correlaciona a microrregião e o modo de detecção da doença; onde é visto que a demanda espontânea tem um frequência maior em todas as regiões.

Figura 23 – Analise bidimensional das variáveis microrregião x classe



Fonte: própria autora, 2021

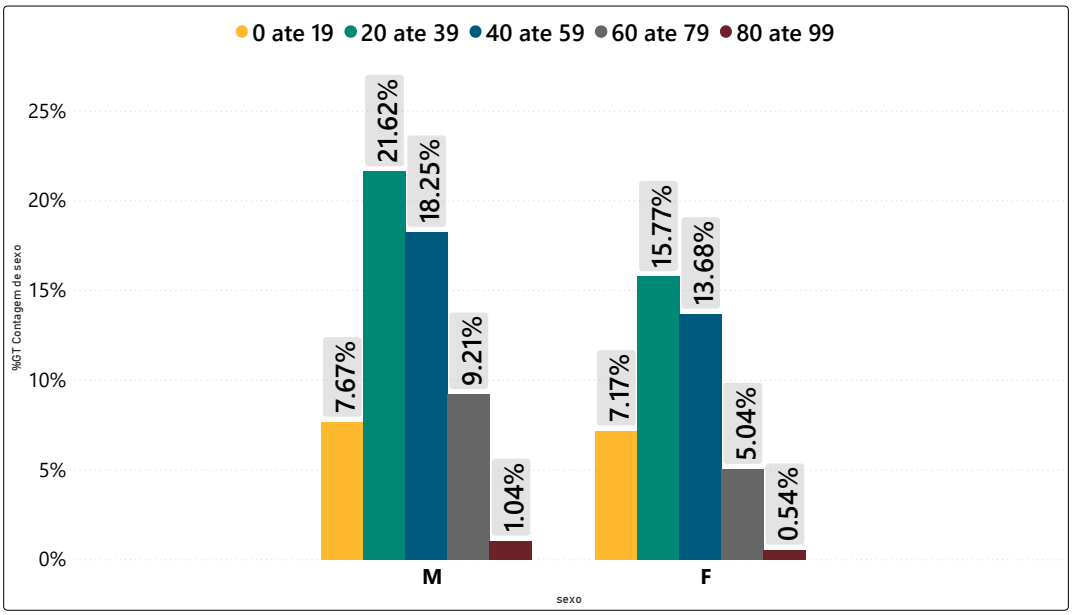
Figura 24 – Analise bidimensional das variáveis microrregião x modo de detecção



Fonte: própria autora, 2021

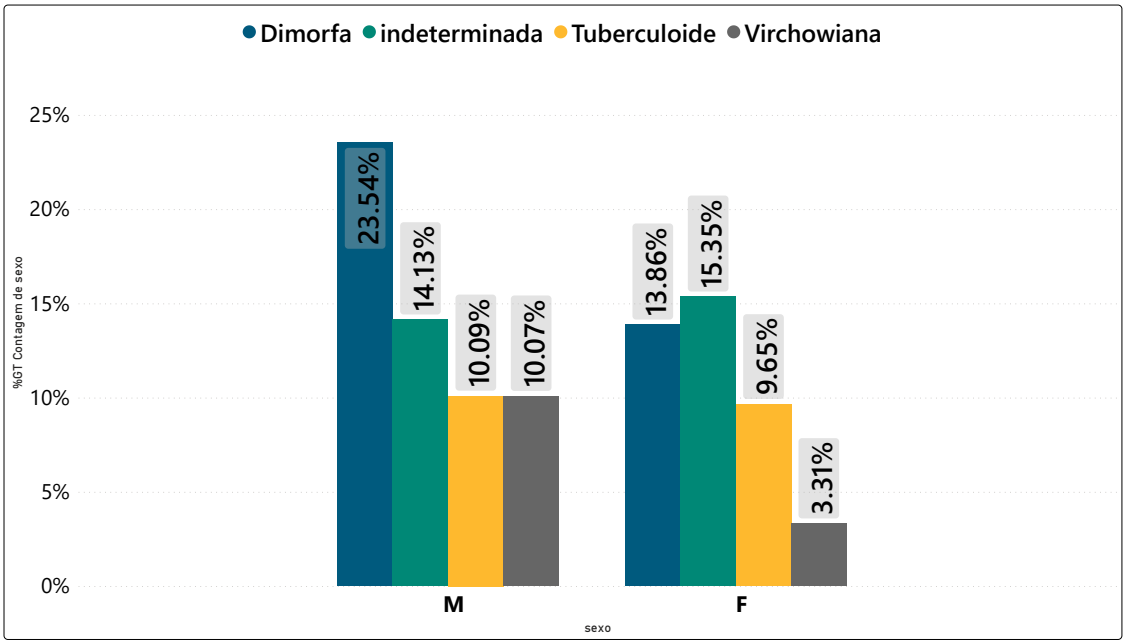
Na análise das variáveis sexo e idade (figura 25) é possível observar que as faixas de idade estão distribuídas quase igualmente entre os dois sexo. O sexo masculino contém 1.665 casos com faixa de idade entre 0 a 19 anos (7,67%), enquanto o sexo feminino possui 1.157 casos (7,17%). Na associação entre a forma clínica e o sexo (figura 26), o sexo masculino com forma clínica caracterizado como dimorfa obtém o maior percentual (23,54%).

Figura 25 – Analise bidimensional das variáveis sexo x idade



Fonte: própria autora, 2021

Figura 26 – Analise bidimensional das variáveis sexo x forma clínica



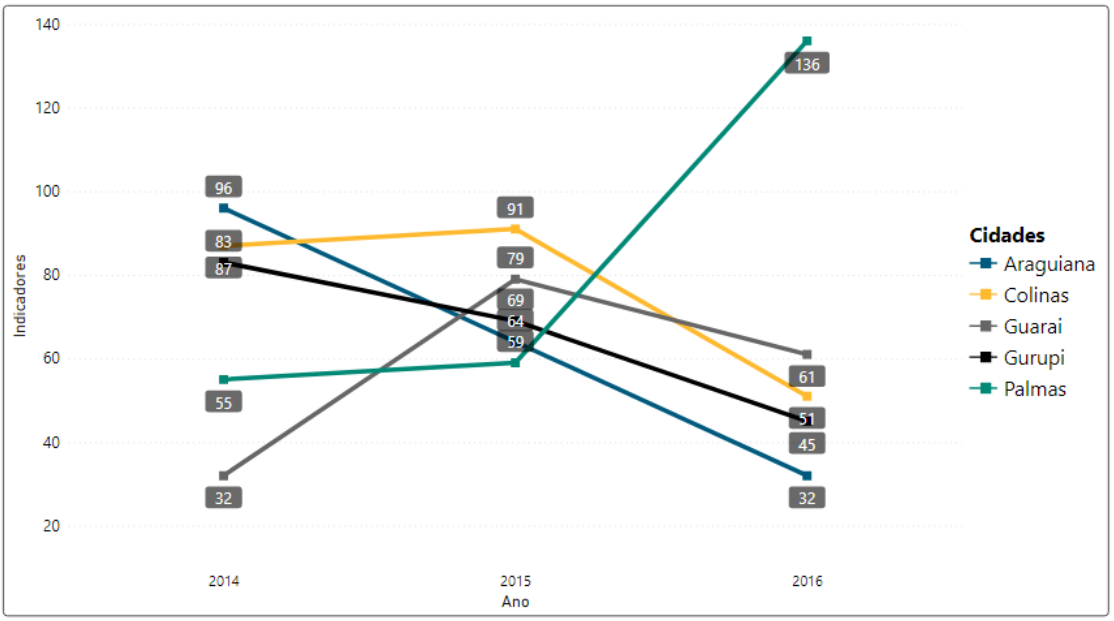
Fonte: própria autora, 2021

5.2 Indicadores

Para o cálculo dos indicadores, foram selecionados os últimos sete anos presentes no intervalo de tempo do estudo: 2011 a 2016; exceto para os indicadores das cidades, que foram selecionados apenas os últimos 3 anos: 2014 a 2016. As variáveis selecionadas para a análise dos indicadores foram aquelas padronizadas pelo programa nacional para avaliação e monitoramento da hanseníase, apresentadas na seção 2.2.1.

A figura 27 exibe as taxas de detecção dos anos de 2014, 2015 e 2016 de casos novos de hanseníase por 100 mil habitantes das cidades mais afetadas no Tocantins. Essas taxas têm como utilidade medir a força de morbidade, magnitude e tendência da endemia no local de estudo. As cidades de Colinas, Guaraí, Gurupi e Araguaína alcançaram uma diminuição nas suas taxas de detecção de 2015 para 2016. Exceto a cidade de Palmas, que teve um aumento considerável de 2014 para 2016, antes em torno de 55, a taxa quase triplicou o valor, indo para 136, a cidade é considerada Hiperendêmica no ano de 2016, de acordo com a classificação da tabela 1.

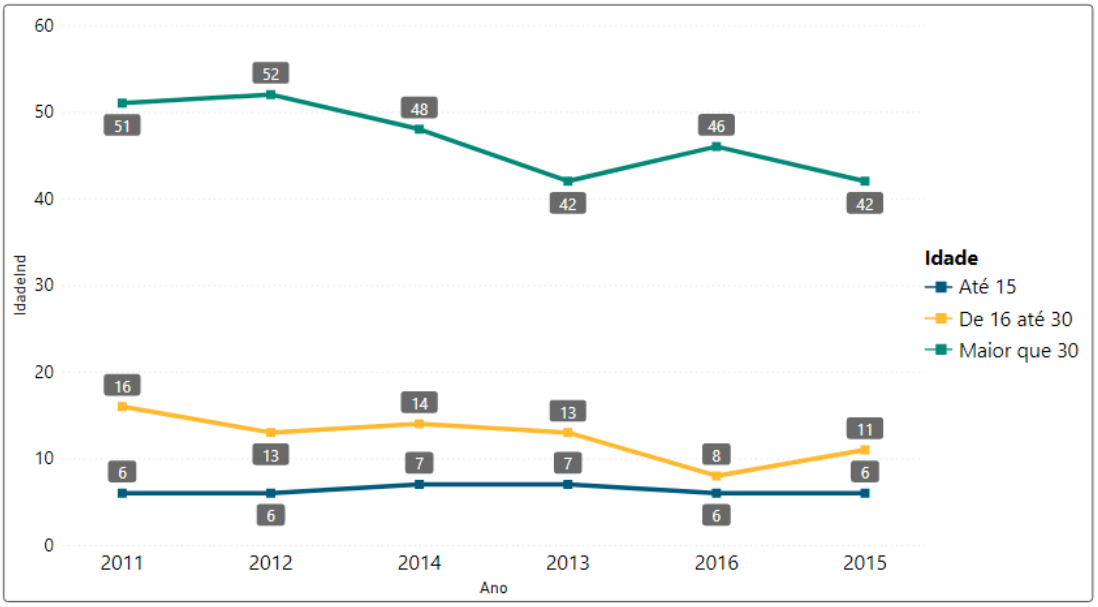
Figura 27 – Coeficiente de detecção de casos novos de hanseníase por cidade no Estado do Tocantins de 2011 a 2016. (por 100 mil habitantes)



Fonte: própria autora, 2021

A figura abaixo apresenta as taxas de detecção de 2011 a 2016 de casos novos de hanseníase por faixas de idades; a faixa de principal análise é a de até 15 anos, que mede a força da transmissão recente da endemia e sua tendência. O gráfico também apresenta a faixa de 16 até 30 anos, e de 30 anos pra cima.

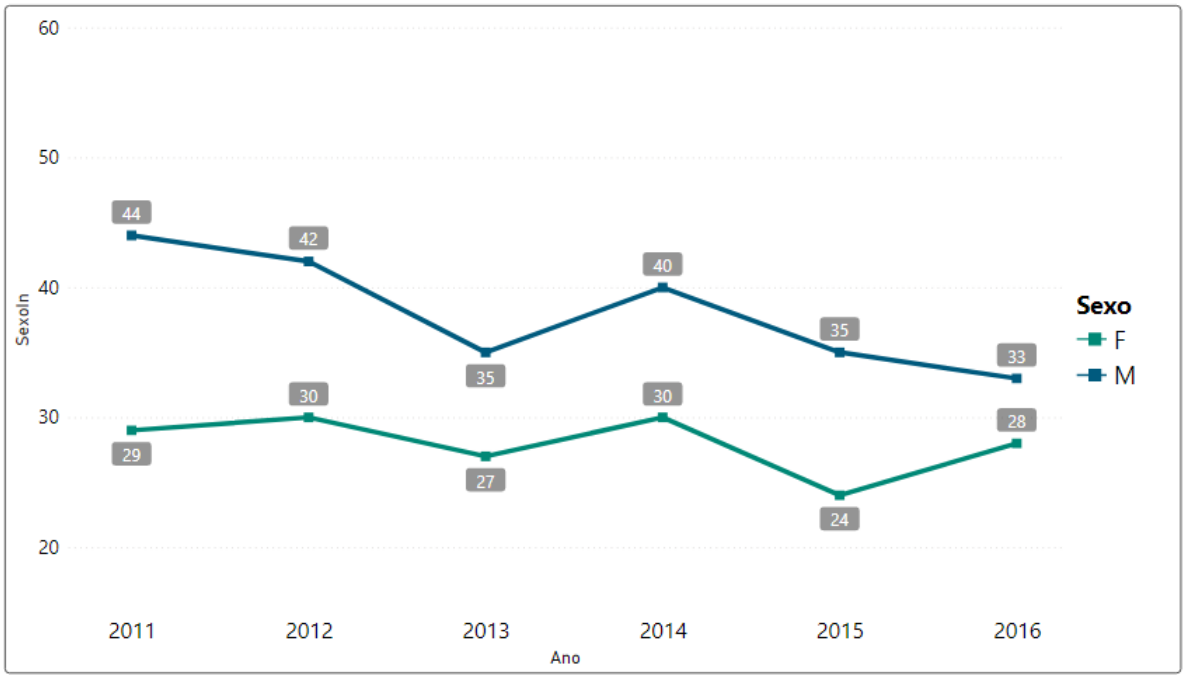
Figura 28 – Coeficiente de detecção de casos novos de hanseníase por faixa de idade no Estado do Tocantins de 2011 a 2016. (por 100 mil habitantes)



Fonte: própria autora, 2021

As taxas de casos de hanseníase, apresentado na figura 29, avalia a capacidade dos serviços em assistir aos casos de hanseníase. Não existe um parâmetro de acordo com o ministério da saúde para classificar essas taxas. Mas é possível analisar as diferenças das taxas por 100 mil habitantes de cada ano para cada sexo.

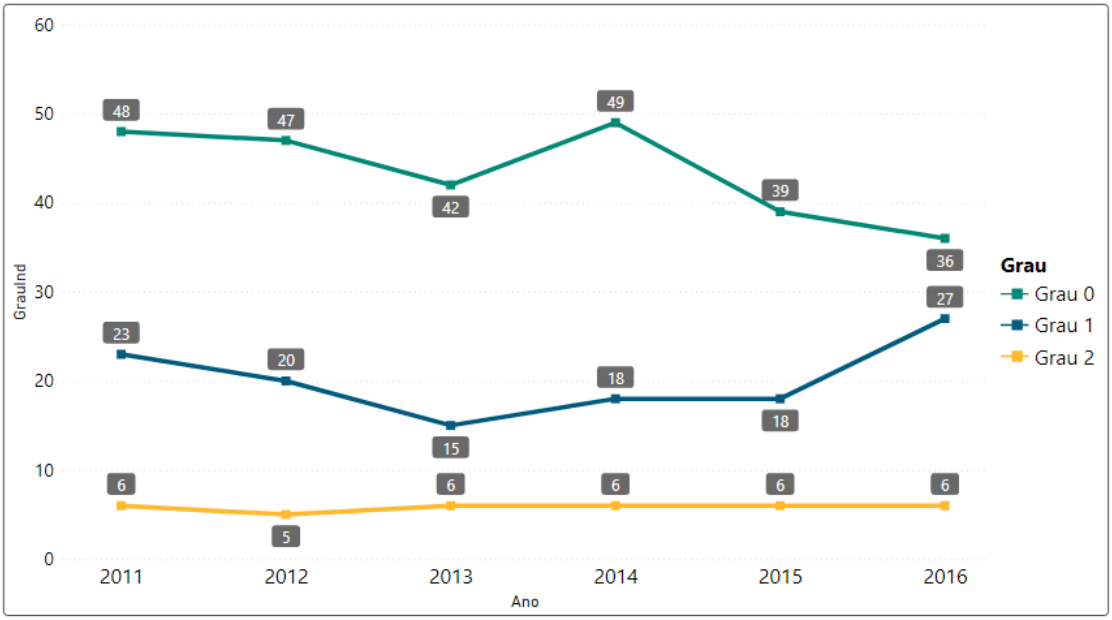
Figura 29 – Coeficiente de detecção de casos novos de hanseníase por sexo no Estado do Tocantins de 2011 a 2016. (por 100 mil habitantes)



Fonte: própria autora, 2021

A figura 30 mostra as taxas de casos novos de hanseníase por grau incapacidade física no momento do diagnóstico por 100 mil hab. O grau 2 é o principal foco de analise nesse gráfico, pois avalia as deformidades causadas pela hanseníase na população geral, e permite a comparação entre outras doenças incapacitantes. Não existe a classificação das taxas para esse caso também, mas a queda deste indicador, caracteriza redução da magnitude da endemia. É possível observar na tabela, que o Tocantins não teve uma redução das taxas entre os anos analisados.

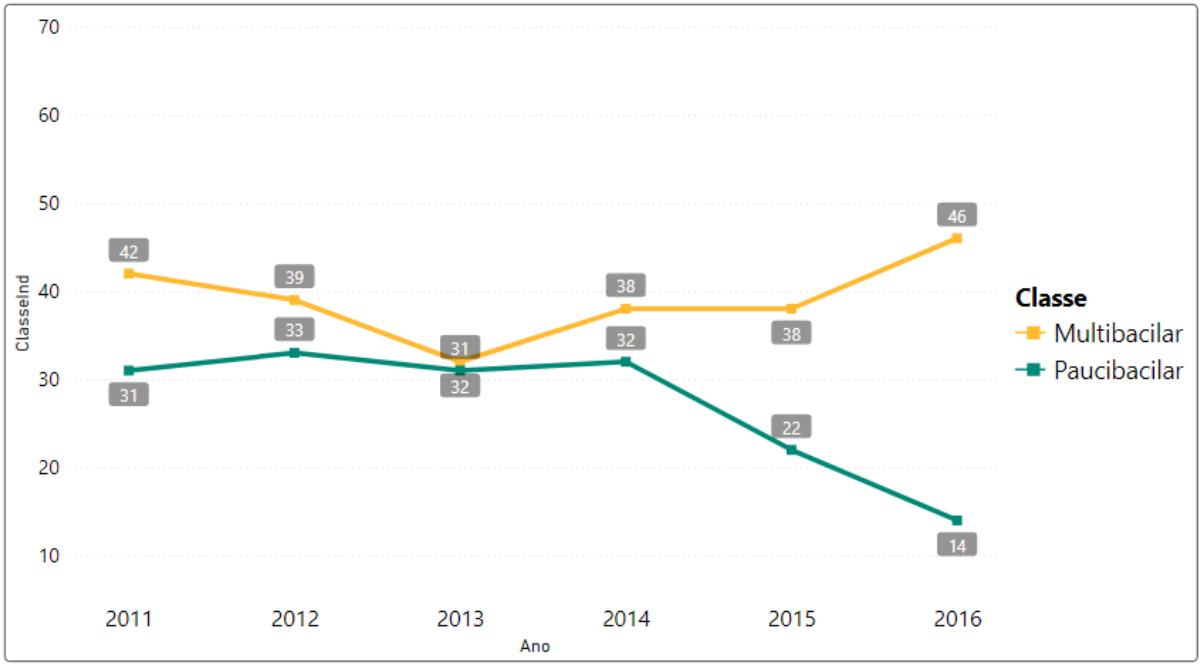
Figura 30 – Coeficiente de detecção de casos novos de hanseníase por Grau de incapacidade no Estado do Tocantins de 2011 a 2016. (por 100 mil habitantes)



Fonte: própria autora, 2021

A figura abaixo apresenta as taxas dos casos multibacilares, que aponta o diagnóstico tardio; e a as taxas de casos paucibacilares, que revela diagnóstico precoce. É possível observar no ano de 2016 uma queda nos diagnósticos precoces, e o aumento nos diagnósticos tardios.

Figura 31 – Coeficiente de detecção de casos novos de hanseníase por Classificação operacional no Estado do Tocantins de 2011 a 2016. (por 100 mil habitantes)



Fonte: própria autora, 2021

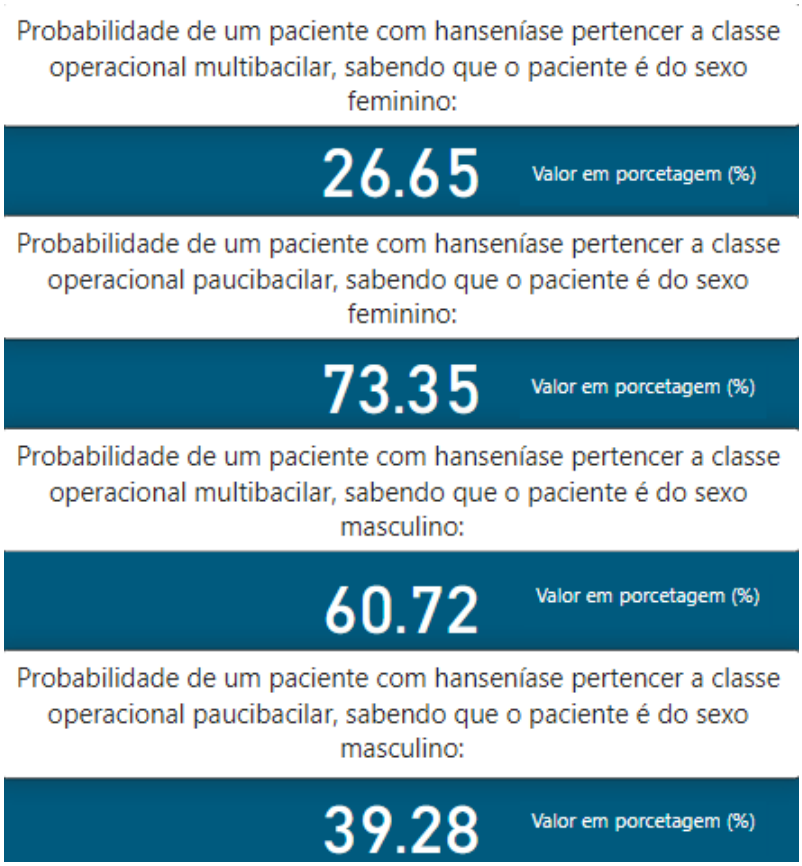
5.3 Teorema de Bayes

Antes da execução da mineração de dados, foi feita uma inferência sobre os dados, que consiste em deduzir conclusões a partir de premissas já conhecidas. A inferência foi feita utilizando o teorema de bayes, apresentado na seção 2.4. O cálculo das probabilidades utilizadas no teorema foram realizados com os dados das notificações de hanseníase feitas no Tocantins entre 2001 a 2016, que são as premissas conhecidas neste estudo, no qual se fundamentam as probabilidades apresentadas a seguir.

Para a realização dos cálculos foram elaboradas algumas perguntas a fim de alcançar as estimativas de probabilidades para cada uma delas. Essas perguntas são baseadas na relação de duas variáveis presentes no conjunto de dados. As perguntas e as estimativas de probabilidades são apresentadas nas figuras 32 e 33.

A figura 32 apresenta a relação das variáveis sexo e classificação operacional. Basicamente, mostra a probabilidade de um evento ocorrer (Evento A), dado que um outro evento já ocorreu(Evento B); nesse cenário: um paciente com hanseníase pertencer a classificação operacional multibacilar (Evento A) e ser do sexo feminino (Evento B). De acordo com as probabilidades dos casos já acontecidos no Tocantins, entre 2001 e 2016, a probabilidade desse evento ocorrer novamente é de 26,65%.

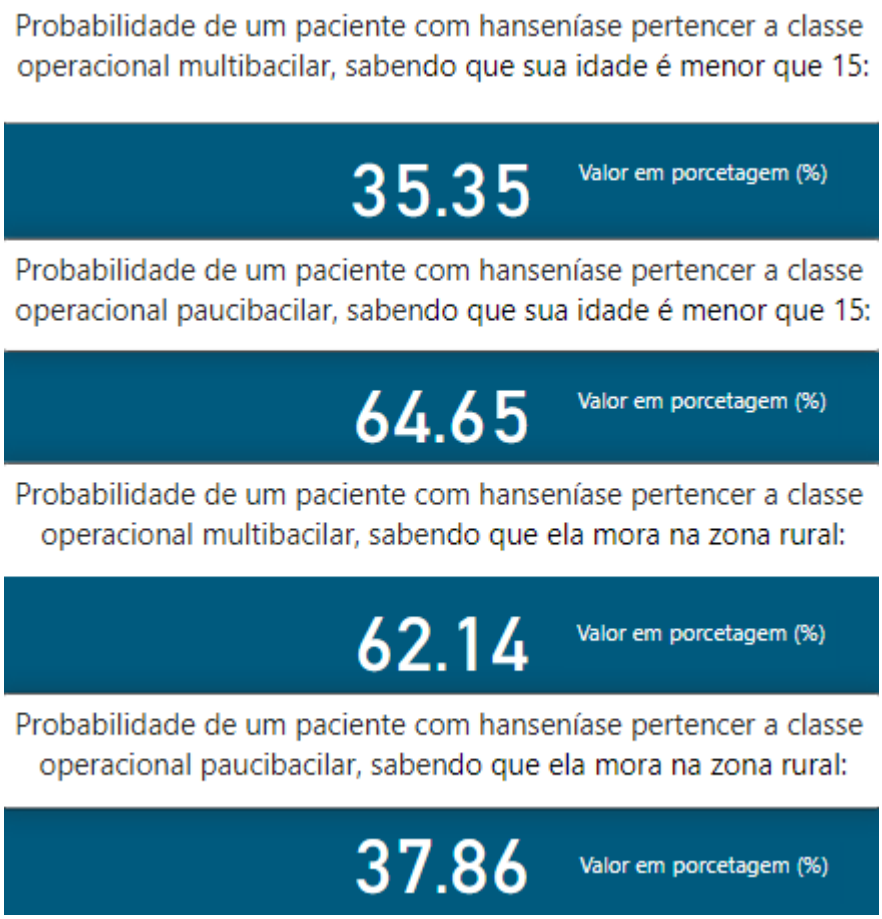
Figura 32 – Probabilidades obtidas com o teorema de bayes



Fonte: própria autora, 2021

A classificação operacional também é relacionada com a variável idade, especificamente casos menores de 15 anos. As probabilidades dessa dependência é apresentada na figura 33, que também contém o resultado da relação da classe com a variável zona de residência.

Figura 33 – Probabilidades obtidas com o teorema de bayes



Fonte: própria autora, 2021

5.4 Mineração de Dados

Essa seção exibe os resultados da mineração de dados executada no trabalho. O arquivo gerado na saída da classificação no weka gera diversas informações e métricas que são separadas por partes, cada um dos termos presentes no arquivo e apresentados a seguir foram definidos na seção 4.2.

5.4.1 Naive Bayes

Ao executar o algoritmo naive bayes sobre os dados extraídos do data mart, foi gerado informações da mineração de dados, apresentadas nesta seção. A primeira parte do arquivo de saída da classificação, apresentado na figura 34, mostra as informações da base de dados utilizada e o modo de teste aplicado, onde a base foi dividida em 70% para treinamento e 30% para teste.

Figura 34 – Informações do arquivo de resultados

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     baseHansen
Instances:    21718
Attributes:   8
              idade
              sexo
              raca
              escolaridade
              zona
              microrregioes
              ano_notificacao
              classe_operacional
Test mode:    split 70.0% train, remainder test
```

Fonte: weka, 2021

A seção do arquivo que exibe o modelo gerado é apresentado na figura 35; consiste em uma tabela com as quantidades das instâncias Multibacilar e Paucibacilar detalhadas e associadas a cada atributo da base de treinamento. É a partir desse modelo que serão extraídas e validadas as informações, apresentadas no próximo capítulo.

Figura 35 – Modelo gerado no arquivo de resultados

```
                               Class
Attribute                Multibacilar Paucibacilar
                               (0.53)       (0.47)
=====================================================
idade
  20 ate 39                   3875.0       4249.0
  40 ate 59                   4138.0       2799.0
  60 ate 79                   2104.0        993.0
  0 ate 19                    1120.0       2104.0
  80 ate 99                    265.0         81.0
  [total]                    11502.0      10226.0

sexo
  M                           7601.0       4951.0
  F                           3898.0       5272.0
  [total]                    11499.0      10223.0

raca
  parda                       7117.0       6442.0
  preta                       1856.0       1434.0
  ignorado                     154.0        136.0
  branca                      2187.0       2004.0
  amarela                      142.0        154.0
  indigena                      47.0         57.0
  [total]                    11503.0      10227.0


zona
  rural                       1882.0       1310.0
  urbana                      9117.0       8506.0
  ignorado                     430.0        335.0
  periurbana                    72.0         74.0
  [total]                    11501.0      10225.0

microrregioes
  colinas                      644.0        499.0
  palmas                      2473.0       1827.0
  guarai                       749.0       1037.0
  paraiso do tocantins         942.0        678.0
  gurupi                      2355.0       1768.0
  araguatins                   910.0        682.0
  porto nacional               702.0        626.0
  tocantinopolis               144.0        120.0
  dianapolis                   449.0        339.0
  araguaina                   1696.0       2273.0S
  miracema do tocantins        444.0        383.0
  [total]                    11508.0      10232.0

ano_notificacao
  2009 a 2010                 1475.0       1189.0
  2003 e 2004                 1347.0       1667.0
  2007 e 2008                 1551.0       1636.0
  2005 e 2006                 1423.0       1711.0
  2001 e 2002                 1184.0       1453.0
  2011 a 2012                 1456.0        988.0
  2013 a 2014                 1341.0        996.0
  2015 a 2016                 1728.0        589.0
  [total]                    11505.0      10229.0


Time taken to build model: 0 seconds
```

Fonte: weka, 2021

A próxima seção apresenta as métricas gerais da classificação. Das 6422 instâncias selecionadas para o teste, 4162 foram classificadas corretamente, tendo uma taxa de acurácia de 64%; e 2309 foram classificadas incorretamente, com taxa de erro de 35%. A classificação apresenta grau de confiabilidade intermediária, de acordo com a Kappa statistic, ou "justo" (valores entre 0,20 e 0,40), conforme explicado na seção .

Figura 36 – Sumário dos resultados

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances        4206               64.5587 %
Incorrectly Classified Instances      2309               35.4413 %
Kappa statistic                          0.2875
Mean absolute error                      0.4299
Root mean squared error                  0.4686
Relative absolute error                 86.2521 %
Root relative squared error             93.8326 %
Total Number of Instances             6515
```

Fonte: weka, 2021

A figura 37 apresenta as métricas de cada classe e a matriz de confusão. A classe Multibacilar teve uma taxa de acurácia de 71%, se saindo melhor que a classificação do Paucibacilar, que teve a taxa de acerto de 56%.

Na matriz de confusão, observa-se que na coluna a, que representa a classe multibacilar, tem o total de 2453 instâncias classificadas corretamente e 1294 incorretamente, ou seja, deveriam ter sido classificadas como paucibacilar. Já na coluna b, referente a classe paucibacilar, 1709 instâncias foram classificadas corretamente.

Figura 37 – Taxas de acurácia por classe

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.688    0.401    0.654      0.688    0.671      0.288    0.701     0.716     Multibacilar
                 0.599    0.312    0.635      0.599    0.616      0.288    0.701     0.659     Paucibacilar
Weighted Avg.    0.646    0.359    0.645      0.646    0.645      0.288    0.701     0.689

=== Confusion Matrix ===

    a    b   <-- classified as
 2352 1067 |    a = Multibacilar
 1242 1854 |    b = Paucibacilar
```

Fonte: weka, 2021

5.4.2 Random Tree

Ao executar o algoritmo Random Tree sobre os dados extraídos do data mart, foi gerado um arquivo com informações da mineração de dados, apresentados a seguir. A figura 38 mostra as informações do modelo. Para a utilização desse algoritmo foi aplicado o modo de teste 'evaluate on training data', que consiste em utilizar a mesma base de treinamento no teste.

Figura 38 – Informações do arquivo de resultados

```
=== Run information ===

Scheme:       weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1
Relation:     baseHansen
Instances:    21718
Attributes:   9
              idade
              sexo
              gestante
              raca
              escolaridade
              zona
              microrregioes
              ano_notificacao
              classe_operacional
Test mode:    evaluate on training data
```

Fonte: weka, 2021

A árvore gerada atingiu um tamanho de 23.168 folhas, por ser muito grande será apresentada na figura 33 o começo e o final do modelo.

Figura 39 – Modelo gerado

```
=== Classifier model (full training set) ===


RandomTree
==========

idade = 20 ate 39
|   sexo = M
|   |   microrregioes = colinas
|   |   |   escolaridade = EF incompleto
|   |   |   |   ano_notificacao = 2009 a 2010
|   |   |   |   |   zona = rural : Paucibacilar (3/1)
|   |   |   |   |   zona = urbana
|   |   |   |   |   |   raca = parda : Paucibacilar (9/4)
|   |   |   |   |   |   raca = preta : Multibacilar (4/1)
|   |   |   |   |   |   raca = ignorado : Multibacilar (0/0)
|   |   |   |   |   |   raca = branca : Multibacilar (0/0)
|   |   |   |   |   |   raca = amarela : Multibacilar (0/0)
|   |   |   |   |   |   raca = indigena : Multibacilar (0/0)
|   |   |   |   |   zona = ignorado : Multibacilar (0/0)
|   |   |   |   |   zona = periurbana : Multibacilar (0/0)
|   |   |   |   ano_notificacao = 2003 e 2004
|   |   |   |   |   escolaridade = EF completo : Multibacilar (0/0)
|   |   |   |   |   escolaridade = ES completo : Multibacilar (0/0)
|   |   |   |   |   escolaridade = Ignorado : Multibacilar (3/0)
|   |   |   |   |   escolaridade = EM completo : Multibacilar (0/0)
|   |   |   |   |   escolaridade = ES incompleto : Multibacilar (0/0)
|   |   |   |   raca = preta : Multibacilar (1/0)
|   |   |   |   raca = ignorado : Multibacilar (0/0)
|   |   |   |   raca = branca : Multibacilar (0/0)
|   |   |   |   raca = amarela : Multibacilar (0/0)
|   |   |   |   raca = indigena : Multibacilar (0/0)
|   |   |   gestante = trimestre ignorado : Multibacilar (0/0)
|   |   |   gestante = terceiro trimeste : Multibacilar (0/0)
|   |   |   gestante = primeiro trimeste : Multibacilar (0/0)
|   |   |   gestante = segundo trimeste : Multibacilar (0/0)

Size of the tree : 23168

Time taken to build model: 0.07 seconds
```

Fonte: weka, 2021

A seção apresentada na figura 40 mostra que o modelo obteve uma acurácia de 80,6%, ou seja, 17.513 das 21.718 instâncias foram classificadas corretamente. A classificação com o algoritmo random tree apresenta um grau de confiabilidade maior, de acordo com kappa statistic está categorizada como "bom" (valores entre 0,60 e 0,80), conforme explicado na seção 4.2.

Figura 40 – Sumário dos resultados

```
=== Summary ===

Correctly Classified Instances       17513               80.6382 %
Incorrectly Classified Instances      4205               19.3618 %
Kappa statistic                          0.6088
Mean absolute error                      0.2394
Root mean squared error                  0.3459
Relative absolute error                 48.0363 %
Root relative squared error             69.3083 %
Total Number of Instances            21718
```

Fonte: weka, 2021

A figura 41 apresenta a seção de acurácia por classe e a matriz de confusão. A classe multibacilar obteve uma taxa de 87% de acurácia, ou seja, 87% (total de 10.016) das classes classificadas como multibacilar estavam corretas. Já a classe paucibacilar obteve 66% de taxa de acerto.

Figura 41 – Acurácia por classe

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.871    0.267    0.786      0.871    0.827      0.613    0.911     0.921     Multibacilar
               0.733    0.129    0.835      0.733    0.781      0.613    0.911     0.901     Paucibacilar
Weighted Avg.  0.806    0.202    0.809      0.806    0.805      0.613    0.911     0.912

=== Confusion Matrix ===

     a     b   <-- classified as
 10016  1481 |     a = Multibacilar
  2724  7497 |     b = Paucibacilar
```

Fonte: weka, 2021

6 ANÁLISE DOS RESULTADOS

Neste trabalho, foram realizados quatro tipos de análises sobre os dados de notificações de hanseníase no estado do Tocantins no período de 2001 a 2016; sendo essas: análise exploratória, indicadores, inferência com o teorema de Bayes e mineração de dados. Cada um deles tem um objetivo diferente. Os resultados desses métodos serão apresentados nesta seção.

Com a análise exploratória foi possível identificar as informações explícitas na base de dados analisada, assim alcançando o objetivo desse método. No período de 2001 a 2016, foram notificados um total de 21.719 pacientes com hanseníase no estado do Tocantins, na qual 18.368 desse total eram casos novos. O ano com maior registro de casos novos foi 2006, com 1.457 casos. A região geográfica que apresentou mais notificações foi Palmas, com 3.577 casos novos.

É possível verificar que a faixa etária mais afetada pela doença está entre 20 a 59 anos. Em relação ao sexo, não há uma diferença tão grande entre as quantidades de casos para cada gênero; mas com a análise bidimensional foi possível perceber que há mais predominância do tipo multibacilar para o sexo masculino. A forma clínica que mais afeta os homens é o tipo dimorfa, já as mulheres têm uma frequência maior do tipo indeterminada.

Ainda conforme a análise exploratória, é observado as frequências das formas de saída e o grau de incapacidade. A cura é a forma de saída mais frequente, com um total de 16.937 casos curados. Dentre eles, 49% foi avaliado com grau zero, ou seja, sem lesão. Enquanto 14% foi avaliado com grau 1 ou 2, o que corresponde, respectivamente, à diminuição e perda de sensibilidade, ou que indica a presença de incapacidades e deformidades. Os 37% restantes, foram casos não avaliados no momento da cura.

O segundo método de análise executado foram os indicadores, que são os números utilizados para o embasamento e planejamento de novas ações para combater a Hanseníase enquanto problema de saúde pública. Os números obtidos mostram que nos últimos anos houve uma queda nos diagnósticos precoces e aumento nos diagnósticos tardios, também apontam uma continuidade de transmissão ativa na população geral, com diferentes magnitudes para cada cidade.

Entre os anos de 2014 a 2016, a cidade de Araguaína foi a que teve maior queda na magnitude da endemia, e no último ano deixou de ser classificada como hiperendêmica. Na situação oposta, encontra-se a cidade de Palmas, que dobrou o número de detecção de casos novos, passando a ser considerada hiperendêmica em 2015. Os coeficientes de detecção para menores de 15 anos apontam uma grande força da transmissão recente da endemia no estado do Tocantins, sendo classificada como em situação hiperendêmica.

Em relação as probabilidades, o terceiro método executado no trabalho foi a inferência utilizando o teorema de Bayes. Como já apresentado, o teorema utiliza evidências anteriores para inferir probabilidades. Dentro do contexto da hanseníase e da base de dados utilizada nesse trabalho, foi possível inferir que a probabilidade de uma mulher com hanseníase ter classificação operacional como multibacilar é de 27%, enquanto para ser do tipo paucibacilar é 73%. Já no caso de um homem com hanseníase, as probabilidades são quase opostas: para multibacilar 61%, e para paucibacilar 39%.

O teorema de bayes também permitiu inferir a probabilidade de um paciente ser menor de 15 anos e ser classificado como multibacilar, baseado nos casos presentes na base de estudo, essa probabilidade equivale a 35%, enquanto para ser classificado como paucibacilar equivale a 65%. Outra inferência considerada é a probabilidade de um paciente morador da zona rural obter a classificação operacional como paucibacilar ou como multibacilar, que mostrou 38% e 62%, respectivamente.

Com esses números apresentados, é possível observar, que, embora existam projetos e campanhas anuais que são destinados a conscientização e a busca ativa da Hanseníase, como as campanhas Janeiro Roxo e Todos Contra a Hanseníase, entre outras que buscam o combate da doença, as quantidades de casos da hanseníase ainda são preocupantes. Com isso, o uso de um sistema que forneça informações, dados e probabilidades detalhados sobre a doença, pode ser de grande valor para próximas decisões no processo de controle da hanseníase.

Algumas das possíveis primeiras ações a serem seguidas, seria uma melhor instrução aos profissionais que executam o preenchimento dos dados no momento da notificação, diminuindo assim a quantidade de dados vazios com possível grau de importância. Deste modo, seria possível um estudo com melhores resultados no futuro. Outra possível ação, seria uma orientação para que a maior quantidade possível de pacientes curados sejam avaliados, pois o registro da incapacidade são essenciais para a instrução quanto ao auto cuidado, para prevenir e evitar a criação de novas incapacidades pós-alta.

Os algoritmos de Mineração de Dados foram executados de modo a fornecer informações úteis para o processo de apoio à decisão, visando encontrar algum padrão entre os perfis dos pacientes com Hanseníase. A fim de determinar um algoritmo de classificação adequado para aplicações na área da saúde, foram comparados dois métodos amplamente utilizados na mineração de dados: *Naive Bayes* e *Random Tree*.

Antes de iniciar a discussão dos resultados da mineração de dados, é importante destacar o motivo da escolha da classe, que se deu, pois, a variável classificação operacional é importante para que possa ser selecionado o esquema de tratamento quimioterápico adequado ao caso; e principalmente por informar o nível que a doença atingiu no indivíduo, essa classificação é feita com base nos sinais e sintomas da doença, PB são casos com até 5 lesões de pele e MB são casos com mais de 5 lesões de pele.

Também foram testada outras variáveis importantes da base de dados como classe alvo, por exemplo, o modo de saída (cura, abandono, óbito e transferência), para tentar prever se existe alguma característica que influencia nessas classes. Porém, mais de 70% dos dados pertenciam a apenas uma classe da variável, a cura. Logo, o algoritmo classificou toda a base somente como cura, assim não sendo possível obter informações sobre o modelo gerado. Esse teste foi feito também para as variáveis grau de incapacidade e modo de detecção, e foi verificado o mesmo resultado, que essas variáveis foram consideradas indevidas para se obter de um bom resultado. Com isso, a classificação operacional foi a que obteve um melhor modelo para a geração de informações.

Indicar um melhor algoritmo de Mineração de Dados para um determinado trabalho não é uma tarefa fácil, pois envolve a realização de muitos treinamentos, testes e análise das métricas. Existem inúmeros algoritmos de classificação, e cada um deles pode ser configurado conforme as necessidades e objetivos da pesquisa.

Em relação aos objetivos deste trabalho, verifica-se que as acurácias dos classificadores executados foram de 65% e 81%, respectivamente. A imagem 42 faz um comparativo entre os dois algoritmos em relação à acurácia, taxa de erro, kappa statisc e tempo de processamento. Em uma análise geral, com base neste trabalho, o melhor algoritmo para aplicações na área da saúde que envolvam características dos pacientes, é o Random Tree.

Figura 42 – Acurácia por classe

Algoritmos	Acurácia	Taxa de erro	kappa statistic	Tempo de processamento
Naive Bayes	65%	35%	0,28	0,01s
Random Tree	81%	19%	0,60	0,07s

Fonte: Criado pelo autor, 2021

Em relação aos padrões encontrados na mineração, é possível notar uma diminuição drástica de casos paucibacilares a partir de 2011, uma diminuição de mais de 50% em relação aos casos multibacilares. Ou seja, nos últimos anos houve um decréscimo na detecção de diagnósticos precoces. Esses dados podem evidenciar uma fragilidade operacional na vigilância epidemiológica; uma hipótese para essa fragilidade é a falta de treinamento eficaz para a detecção desses casos em tempo adequado.

Sobre as microrregiões, é importante ressaltar as diferenças encontradas nas quantidades de casos multibacilares e paucibacilares. As regiões de Araguaína e Guaraí são as únicas com uma taxa de detecção de casos paucibacilares maior que casos multibacilares, isto é, essas regiões possuem um melhor desempenho no diagnóstico precoce. Uma suposição para essa diferença é, possivelmente, o melhor desenvolvimento nos exames de coletividade e exames de contatos nas cidades dessas regiões.

Quanto às características dos pacientes, um padrão encontrado foi em relação à idade. As faixas etárias de 40 a 59 anos e 60 a 79 anos possuem o dobro de casos multibacilares em relação a casos paucibacilares; isto significa um grande número de diagnósticos tardios em pacientes com mais de 40 anos. Essa informação pode estar relacionada ao fato de que os indivíduos ao sentirem os primeiros sintomas não deram muita atenção e não procuraram o serviço de saúde.

Outro motivo interessante é apresentado em (OLIVEIRA, 2011), onde foi feito uma pesquisa com 936 indivíduos que foram notificados com Hanseníase entre 2006 a 2008 no Tocantins. O estudo mostrou que 20% dos entrevistados alegaram que receberam a medicação prescrita errada, ou seja, 188 pessoas desse grupo tomaram fármaco antes da PQT; assim dificultando o diagnóstico da Hanseníase mesmo quando os pacientes procuram o serviço de saúde. Neste estudo também é apresentado outros motivos, como: falta de acessibilidade e tentativas de outros tratamentos.

Dessa maneira, sobre os dados analisados na mineração de dados, encontram-se duas características dos pacientes que possivelmente estejam influenciando na classificação operacional. A idade, que a partir de 40 anos aumenta a probabilidade de ser um caso multibacilar, e o sexo, pois o sexo masculino tem uma maior influencia a ser tipo MB.

Portanto, com base em todos esses dados, informações e probabilidades apresentadas, pode-se concluir que a hipótese apresentada no trabalho é confirmada. A união das tecnologias de ciência de dados no mesmo trabalho permitiu obter mais resultados que o esperado. Enquanto a mineração de dados resultou em padrões sobre as características dos pacientes, a inferência com teorema de Bayes permitiu a descoberta de probabilidades importantes. Sobre o data mart, foi possível perceber uma grande melhoria e facilidade na escolha e extração dos dados para as análises posteriores e para a mineração de dados.

7 CONSIDERAÇÕES FINAIS

Uma das principais motivações para a realização deste projeto foi a necessidade de uma ferramenta capaz de armazenar os dados do Sinan de forma eficiente e organizada para facilitar a extração e geração de informações sobre esses dados. Baseada nessa motivação foi criado um Data Mart, cujos dados são armazenados em estruturas lógicas dimensionais, facilitando e agilizando o seu processamento analítico por ferramentas especiais. Em seguida, foram aplicadas técnicas de análise estatística e de mineração sob os dados para a geração de informações.

O Data Mart executado neste projeto mostrou-se uma solução de fácil implementação e eficiente quanto ao armazenamento de dados. Foi possível aplicar diferentes técnicas de análises aos dados, que foram: análise exploratória, identificando as informações explícitas; os indicadores, números usados pelos profissionais da área; inferência com teorema de Bayes, deduzindo probabilidades sobre a doença e, por fim, a classificação de dados, identificando padrões entre os dados dos pacientes. Com isso, pode-se concluir que o DM criado, as análises feitas e os resultados obtidos validam a hipótese, bem como atendem ao objetivo geral e os objetivos específicos do trabalho.

Considera-se que, para fins tecnológicos, que as principais contribuições foram: a definição de um processo que utilize um sistema computacional de armazenamento multidimensional de dados, aliado a técnicas de análises estatísticas e de mineração de dados. A utilização e a comparação de algoritmos de classificação na mineração de dados. E a definição de um melhor algoritmo para pesquisas semelhantes ou baseadas na saúde pública.

Entre as contribuições disponibilizadas para a saúde pública, em especial para o possível controle da hanseníase no Tocantins, estão as probabilidades deduzidas com a inferência utilizando teorema de Bayes, bem como os padrões ao longo dos anos e sobre as regiões encontradas na mineração de dados.

O desenvolvimento do projeto foi executado por partes: primeiro a coleta e entendimento dos dados, e em seguida foi realizado o pré-processamento, que consiste em limpeza, redução, enriquecimento e transformação dos dados. O próximo passo foi a implementação do data mart, posteriormente foi feito a análise exploratória, os indicadores e a inferência com teorema de Bayes. A mineração de dados foi o último passo a ser realizado. Durante a execução dessas etapas foram encontrados alguns desafios e limitações.

O primeiro desafio encontrado foi ao realizar a etapa de pré-processamento, onde foi notado que a base de dados continham muitas variáveis importantes não preenchidas, como, por exemplo, a variável ocupação do paciente, que informa a atividade exercida

pelo paciente, seja no setor formal, informal ou autônomo.

Ainda sobre a base de dados, que era supostamente para conter somente as notificações de hanseníase do estado do Tocantins, foram encontrados diversos registros preenchidos com municípios de outros estados, com isso, a base de dados ficou bastante reduzida.

Outra dificuldade foi encontrada na etapa de mineração de dados, na qual foi percebida ao tentar definir os objetivos da mineração, ao escolher uma classe e ao determinar as variáveis que seriam utilizadas na mineração. A tarefa de classificação associa ou classifica objetos a determinadas classes, mas qual seria a variável mais adequada da base do estudo para ser usada como classe? Essa dúvida surgiu, pois, no início do trabalho, imaginava-se que a classificação seria para prever se determinado individuo estava propicio a ter ou não Hanseníase, assim podendo evitar a sua futura contaminação.

Porém, ao estudar e compreender a base de dados, foi notado que não era possível fazer essa classificação, pois todas as notificações eram de indivíduos que já foram notificados com Hanseníase, ou seja, havia dados para classificar um perfil de indivíduo mais propicio a ter hanseníase, mas não era possível prever um perfil que não seria propicio a ter a doença. Com isso, todas as variáveis da base foram analisadas para selecionar uma e definir o objetivo da mineração, que era tentar prever quais características dos pacientes mais influenciam na classificação operacional da doença.

Após ser abordada toda a execução de todo o projeto, é estabelecida algumas recomendações para trabalhos futuros semelhantes:

- Utilizar mais de uma técnica de mineração de dados para tentar prever padrões nos dados, como, por exemplo, Indução de regras e Clusterização.
- Ampliar o estudo para toda a região norte, e após isso outras regiões.
- Aplicar as técnicas utilizadas nesse trabalho sobre os dados de outras doenças epidêmicas.
- Estudos de casos múltiplos, ou seja, testar alguma das análises do trabalho em duas ou mais doenças epidêmicas.

Por fim, esse estudo mostrou que com a adoção e união de diversas técnicas de análises é possível obter diferentes resultados e informações, que, em conjuntos se complementam. Espera-se que com este trabalho seja possível contribuir para o aumento da eficiência no gerenciamento do controle de hanseníase no estado do Tocantins.

REFERÊNCIAS

ACHESON, D. **Independent inquiry into inequalities in health: report**. [S.l.]: HM Stationery Office, 1998.

AZEVEDO, A. I. R. L.; SANTOS, M. F. Kdd, semma and crisp-dm: a parallel overview. **IADS-DM**, 2008.

BARBIERI, C. **BI - business intelligence: modelagem e tecnologia**. [S.l.]: Axcel Books, 2001.

BI, B. O. **Data Modeling Profiling Tool: SQL Power Architect**. 2018. Disponível em: <http://www.bestofbi.com/page/architect>. Acesso em: 3 Mar. 2021.

BORGES, C. C. **Sentidos de saúde/doença produzidos em grupo numa comunidade alvo do Programa de Saúde da Família (PSF).** Tese (Doutorado) — Universidade de São Paulo, 2002.

BRASIL. **Lei n.º 6.259, de 30 de Outubro de 1975**. 1975. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/L6259.htm#:~:targetText=LEI>.

BRASIL. **Constituição da República Federativa do Brasil**. 1988. Disponível em: <http://www.planalto.gov.br/ccivil_03/constituicao/constituicaocompilado.htm>.

BRASIL. **Lei nº 8.080, de 19 de setembro 1990.** 1990. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/l8080.htm>.

BRASIL. **Lei n.º 1.920, de 25 de Julho de 1953**. 1993. Disponível em: <http://www.planalto.gov.br/ccivil_03/LEIS/1950-1969/L1920.htm#:~:targetText=LEI>.

BRASIL. **Guia de Vigilância Epidemiológica**. 7. ed. 2009. Disponível em: <http://bvsms.saude.gov.br/bvs/publicacoes/guia_vigilancia_epidemiologica_7ed.pdf>.

BRASIL. **Portaria Nº 1.271, de 6 de Junho de 2014**. 2014. Disponível em: <http://bvsms.saude.gov.br/bvs/saudelegis/gm/2014/prt1271_06_06_2014.html>.

BRASIL. **DATASUS**. 2016. Disponível em: <http://datasus.saude.gov.br/datasus>. Acesso em: 4 Nov. 2019.

BRASIL. **O Sinan**. 2016. Disponível em: <http://portalsinan.saude.gov.br/o-sinan>. Acesso em: 4 Nov. 2019.

BRASIL. **RESOLUÇÃO Nº 510, DE 7 DE ABRIL DE 2016**. 2016. Disponível em: <http://bvsms.saude.gov.br/bvs/saudelegis/cns/2016/res0510_07_04_2016.html>.

BRASIL. **Sistema único de Saúde**. 2017. Disponível em: <http://www.saude.gov.br/sistema-unico-de-saude/sistema-unico-de-saude>. Acesso em: 3 Nov. 2019.

BRASIL. **Hanseníase: o que é, causas, sinais e sintomas, tratamento, diagnóstico e prevenção**. 2019. Disponível em: <http://saude.gov.br/saude-de-a-z/hanseniase>. Acesso em: 5 Nov. 2019.

CABENA, P. et al. **Discovering data mining: from concept to implementation**. [S.l.]: Prentice-Hall, Inc., 1998.

CAMILO, C. O.; SILVA, J. C. d. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. **Universidade Federal de Goiás (UFC)**, p. 1–29, 2009.

CARVALHO, G. A saúde pública no brasil. **Estudos avançados**, SciELO Brasil, v. 27, n. 78, p. 7–26, 2013.

CDC. **What Is Public Health?** 2018. Disponível em: <https://www.cdcfoundation.org/what-public-health>. Acesso em: 10 Set. 2019.

CHEN, H.-c. et al. Business intelligence and analytics: From big data to big impact. **MIS Quarterly**, v. 36, p. 1165–1188, 2012.

CHEN, J. et al. Big data challenge: a data management perspective. **Frontiers of Computer Science**, v. 7, n. 2, p. 157–164, Apr 2013. Disponível em: <https://doi.org/10.1007/s11704-013-3903-7>.

CIOS, K. J.; PEDRYCZ, W.; SWINIARSKI, R. W. Data mining and knowledge discovery. In: **Data mining methods for knowledge discovery**. [S.l.]: Springer, 1998. p. 1–26.

DATABASE.GUIDE. **What is DBeaver?** 2018. Disponível em: <https://database.guide/what-is-dbeaver/>. Acesso em: 3 Mar. 2021.

DECKER, K. M.; FOCARDI, S. Technology overview: A report on data mining. Citeseer, 1995.

DEPARTMENT, C. S. **Weka Error Measurements**. 2021. Disponível em: <https://katie.mtech.edu/classes/csci347/Resources/Weka_error_measurements.pdf>. Acesso em: 3 Mar. 2021.

DOMENICO, J. A. D. **Definição de um ambiente Data Warehouse em uma instituição de ensino superior**. Tese (Doutorado) — Universidade Federal de Santa Catarina, Centro Tecnológico., 2001.

DUARTE, A. de M. et al. Construção de data warehouse para a base de aih do sus. **XI Congresso Brasileiro de Informática em Saúde - CBIS 2008**, 2008.

ELMASRI, R.; NAVATHE, S. **Sistemas de banco de dados**. PEARSON BRASIL, 2011. ISBN 9788579360855. Disponível em: <https://books.google.com.br/books?id=FSvIYgEACAAJ>.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37–37, 1996.

GAMA, J. Árvore de decisão. **Palestra ministrada no Núcleo da Ciência de Computação da Universidade do Porto, Porto**, 2002.

GARCIAN, R. **História do Brasil**. 2003. Disponível em: <https://educacao.uol.com.br/disciplinas/historia-brasil/governo-rodrigues-alves-1902-1906-revolta-da-vacina-e-acre.htm?next=0004H235U144L48P>. Acesso em: 2 Nov. 2019.

GONçALVES DIANA E SANTOS, M. Y. e. C. J. Analysis of the quality of life after an endoscopic thoracic sympathectomy: A business intelligence approach. In: IEEE. **Second International Conference on Advances in Databases, Knowledge, and Data Applications**. [S.l.], 2010. p. 1–6.

GUERRERO, J. **Getting Started with PGAdmin on a Distributed SQL Database**. 2019. Disponível em: <https://blog.yugabyte.com/getting-started-with-pgadmin-on-a-distributed-sql-database/>. Acesso em: 3 Mar. 2021.

GUIDINI, C. Abordagem histórica da evolução do sistema de saúde brasileiro: conquistas e desafios. Universidade Federal de Santa Maria, 2012.

HALL, M. et al. The weka data mining software: an update. **ACM SIGKDD explorations newsletter**, ACM, v. 11, n. 1, p. 10–18, 2009.

HAN, J.; PEI, J.; KAMBER, M. **Data mining: concepts and techniques**. [S.l.]: Elsevier, 2011.

INMON, W. H.; HACKATHORN, R. D. Como usar o data warehouse. **Rio de Janeiro: Infobook**, 1997.

JENSEN, L. B.; LUKIC, I.; GULIS, G. The delivery of health promotion and environmental health services; public health or primary care settings? In: MULTIDISCIPLINARY DIGITAL PUBLISHING INSTITUTE. **Healthcare**. [S.l.], 2018. v. 6, n. 2, p. 42.

KIMBALL, R.; ROSS, M. **The data warehouse toolkit: the complete guide to dimensional modeling**. [S.l.]: John Wiley & Sons, 2011.

KIMBALL, R. et al. **The data warehouse lifecycle toolkit**. [S.l.]: John Wiley & Sons, 2008.

LAGUARDIA, J. et al. Sistema de informação de agravos de notificação em saúde (sinan): desafios no desenvolvimento de um sistema de informação em saúde. **Epidemiologia e Serviços de Saúde**, scielo, v. 13, p. 135 – 146, 2004. ISSN 1679-4974. Disponível em: <http://scielo.iec.gov.br/scielo.php?script=sci_arttext&pid=S1679-49742004000300002&nrm=iso>.

LAVRAč, N. et al. Data mining and visualization for decision support and modeling of public health-care resources. **Journal of Biomedical Informatics**, v. 40, n. 4, p. 438 – 447, 2007. ISSN 1532-0464. Public Health Informatics. Disponível em: <http://www.sciencedirect.com/science/article/pii/S153204640600116X>.

MACIEIRA, S. Aspectos microbiologicos do mycobacterium leprae. 2000.

MAGALHÃES, L. **Saúde Pública no Brasil**. 2019. Disponível em: <https://www.todamateria.com.br/saude-publica-no-brasil/>. Acesso em: 2 Nov. 2019.

MARQUES, R. et al. **Sistema de Saúde no Brasil: organização e financiamento**. [S.l.: s.n.], 2016. 12-21 p. ISBN 978-85-7967-115-9.

MICROSOFT. **O que é Power BI?** 2021. Disponível em: <https://docs.microsoft.com/pt-br/power-bi/fundamentals/power-bi-overview>. Acesso em: 3 Mar. 2021.

MILIDIÚ, R. L. **Aprendizado de Máquina para o Problema de Sentiment Classification**. Tese (Doutorado) — PUC-Rio, 2006.

MONTEIRO, L. D. et al. Tendências da hanseíase no tocantins, um estado hiperendêmico do norte do brasil, 2001-2012. **Cadernos de Saúe Pública**, SciELO Public Health, v. 31, p. 971–980, 2015.

MONTEIRO, L. D. et al. Hanseníase em menores de 15 anos no estado do tocantins, brasil, 2001-2012: padrão epidemiológico e tendência temporal. **Revista Brasileira de Epidemiologia**, SciELO Public Health, v. 22, p. e190047, 2019.

MORETTIN, P. A.; BUSSAB, W. O. **Estatística básica**. [S.l.]: Editora Saraiva, 2017.

MOURA, W. **Desvendando o classificador Naive Bayes**. 2019. Disponível em: <https://hackinganalytics.com/2016/10/13/desvendando-o-classificador-naive-bayes/>. Acesso em: 8 Nov. 2019.

MV. **Gestão da Saúde Pública: particularidades e desafios do SUS**. 2018. Disponível em: <http://www.mv.com.br/pt/blog/gestao-da-saude-publica--particularidades-e-desafios-do-sus>. Acesso em: 20 Set. 2019.

OLIVEIRA, A. L. de. História da saúde no brasil: dos primórdios ao surgimento do sus. **Revista Encontros Teológicos**, v. 27, n. 1, 2012.

OLIVEIRA, A. R. d. Fatores associados ao diagnóstico tardio da hanseníase em 74 municípios endêmicos do estado do tocantins. 2011.

OLIVEIRA, L. **Teorema de Bayes Probabilidade**. 2020. Disponível em: <https://medium.com/data-hackers/teorema-de-bayes-probabilidade-d5ead2df1379>. Acesso em: 3 Mar. 2021.

PALANIAPPAN, S.; AWANG, R. Intelligent heart disease prediction system using data mining techniques. **International Conference on Computer Systems and Applications**, p. 108–115, 2008.

PIRES, F. A. et al. Ambiente para extração de informações de saúde a partir de bases de dados do sus. **BIS. Boletim do Instituto de Saúde**, v. 13, p. 39–45, 2011. Disponível em: <http://periodicos.ses.sp.bvs.br/scielo.php?script=sci_arttext&pid=S1518-18122011000100007&lng=pt&nrm=iso&tlng=pt>.

RAGSDALE, C. T. **Spreadsheet modeling and decision analysis**. [S.l.]: Thomson South-Western, 2004.

SANTOS, R. S. et al. Data warehouse para a saúde pública: estudo de caso ses-sp. In: . [S.l.: s.n.], 2006.

SINGH, H. S. Data warehouse: conceitos, tecnologias, implementação e gerenciamento. **Tradução Mônica Rosemberg. Editora Makron Books. São Paulo**, 2001.