UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

CIÊNCIA DE DADOS E APRENDIZADO DE MÁQUINA APLICADOS AO ESTUDO DE VARIÁVEIS EPIDEMIOLÓGICAS DA HANSENÍASE NA AMAZÔNIA

IGOR WENNER SILVA FALCÃO

TD: 23/2024

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém - Pará - Brasil
2024

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

IGOR WENNER SILVA FALCÃO

CIÊNCIA DE DADOS E APRENDIZADO DE MÁQUINA APLICADOS AO ESTUDO DE VARIÁVEIS EPIDEMIOLÓGICAS DA HANSENÍASE NA AMAZÔNIA

Tese submetida à Banca Examinadora do Programa de Pós-Graduação em Engenharia Elétrica da UFPA para a obtenção do Grau de Doutor em Engenharia Elétrica na área de Computação Aplicada, elaborada sob a orientação do Prof. Dr. Marcos César da Rocha Seruffo.

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém - Pará - Brasil
2024

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

F178c FALCÃO, IGOR WENNER SILVA.
CIÊNCIA DE DADOS E APRENDIZADO DE MÁQUINA APLICADOS AO ESTUDO DE VARIÁVEIS EPIDEMIOLÓGICAS DA HANSENÍASE NA AMAZÔNIA : ESTUDO DE VARIÁVEIS EPIDEMIOLÓGICAS DA HANSENÍASE NA AMAZÔNIA / IGOR WENNER SILVA FALCÃO. — 2024.
85 f. : il. color.

Orientador(a): Prof. Dr. Marcos CÉsar da Rocha Seruffo
Coorientador(a): Prof. Dr. Diego Lisboa Cardoso
Tese (Doutorado) - Universidade Federal do Pará, Instituto de Tecnologia, Programa de Pós-Graduação em Engenharia Elétrica, Belém, 2024.

1. Hanseníase. 2. Diagnóstico. 3. Ciência de Dados. 4. Agrupamento. 5. Random Forest. I. Título.

CDD 621.307209811

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

CIÊNCIA DE DADOS E APRENDIZADO DE MÁQUINA APLICADOS AO ESTUDO DE VARIÁVEIS EPIDEMIOLÓGICAS DA HANSENÍASE NA AMAZÔNIA

AUTOR: IGOR WENNER SILVA FALCÃO

A TESE DE DOUTORADO SUBMETIDA À AVALIAÇÃO DA BANCA EXAMINADORA APROVADA PELO COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA DA UNIVERSIDADE FEDERAL DO PARÁ E JULGADA ADEQUADA PARA OBTENÇÃO DO GRAU DE DOUTOR EM ENGENHARIA ELÉTRICA NA ÁREA DE COMPUTAÇÃO APLICADA.

APROVADA EM __ / __ / ____.

BANCA EXAMINADORA:

Prof. Dr. Marcos César da Rocha Seruffo
Orientador - PPGEE/ITEC/UFPA

Prof. Dr. Diego Lisboa Cardoso
Coorientador - PPGEE/ITEC/UFPA

Prof. Dr. Carlos Renato Lisboa Francês
Membro Interno - PPGEE/ITEC/UFPA

Prof. Dr. Claudio Guedes Salgado
Membro Externo ao Programa -
PPGDT/ICB/UFPA

Profª. Drª. Jasmine Priscyla Leite de Araujo
Membro Interno - PPGEE/ITEC/UFPA

Profª. Drª. Karla Tereza Figueiredo Leite
Membro Externo à Instituição - IME/UERJ

Visto:

Prof. Dr. Diego Lisboa Cardoso
Coordenador do PPGEE/ITEC/UFPA

AGRADECIMENTOS

Primeiramente, agradeço a Deus por ter me concedido o privilégio de chegar a este momento tão especial e significativo na minha vida e por todas as oportunidades que me concedeu.

Agradeço aos meus pais, o Sr. Denis Falcão e a Sra. Wanneida Costa, aos meus avós, o Sr. José Falcão, a Sra. Maria de Nazaré, ao Sr. Raimundo Rodrigues, a Sra. Guiomar da Costa e a Sra. Maria Valdelina pelo apoio incondicional, incentivo, motivação, dedicação e por estarem sempre ao meu lado me ajudando da melhor forma possível. À minha namorada Flávia Rodrigues pelo apoio, incentivo e motivação em todos os momentos. Agradeço ainda a todos os demais familiares que me apoiaram em todo o processo de formação no Doutorado.

Agradeço ao meu filho, Ravi Falcão, que veio ao mundo e se tornou a principal fonte de motivação para todas as minhas conquistas pessoais e profissionais. De longe, este tem sido o maior desafio que venho enfrentando paralelamente à construção do meu trabalho de Doutorado. Meus sinceros agradecimentos.

Agradeço grandemente ao meu orientador, o Prof. Dr. Marcos Seruffo por estar ao meu lado, fazendo valer um voto de confiança depositado em meu trabalho. Agradeço ao Prof. Dr. Diego Lisboa pela sua contribuição assídua na produção deste documento de qualificação e pelos bons conselhos dados. E não menos importante, agradeço ao Prof. Dr. Claudio Salgado e ao Prof. Dr. Moisés Silva pelas contribuições significativas e pontuais no desenvolvimento deste trabalho. Meu sincero muito obrigado.

Agradeço à Universidade de São Paulo (USP) de Ribeirão Preto, que tem sido um importante parceiro na pesquisa imunológica e molecular sobre hanseníase, permitindo que novos conhecimentos sejam gerados e divulgados na literatura acadêmica.

Minha gratidão também à Universidade Federal do Maranhão (UFMA), nas unidades de São Luís e Imperatriz, que promoveram parcerias essenciais em pesquisas clínicas e epidemiológicas sobre hanseníase no Maranhão, possibilitando estudos de campo que têm contribuído para a compreensão da prevalência e dos tratamentos da doença.

A Universidade Ceuma, que atua em conjunto com a UFMA em Imperatriz, merece meu reconhecimento pelo suporte na análise de casos clínicos e no acompanhamento de pacientes com hanseníase, garantindo que as práticas de pesquisa sejam realizadas com rigor e compromisso.

Agradeço à Secretaria de Saúde do município de Marituba-PA, que tem sido um pilar no apoio à coleta de dados e na implementação de políticas de saúde pública voltadas para o combate e controle da hanseníase.

Reconheço também a colaboração da Secretaria de Saúde do município de São Luís -

MA, que tem se destacado na disseminação de campanhas educativas e na oferta de serviços de diagnóstico e tratamento, contribuindo para a conscientização da população sobre a hanseníase.

À Secretaria de Saúde do município de Imperatriz - MA, que trabalha em conjunto com a Secretaria Estadual de Saúde do Maranhão para o monitoramento de casos de hanseníase na região, expresso minha gratidão pelas iniciativas que favorecem a pesquisa e a prática clínica.

Agradeço à Secretaria de Saúde do Estado do Pará, que tem atuado em parceria com o LDI na promoção de saúde pública e no manejo da hanseníase, permitindo a troca de informações e experiências valiosas.

Por fim, agradeço à Secretaria de Saúde do Estado do Maranhão, que desenvolve políticas efetivas de combate à hanseníase e participa ativamente de campanhas de conscientização, diagnóstico e tratamento, ajudando a transformar dados em ações que beneficiam a população e enriquecem a literatura científica.

Essas iniciativas, que fomentaram as pesquisas e ações relacionadas à hanseníase nos últimos anos, foram fundamentais para a coleta e divulgação de dados, resultando em diversos estudos científicos que emergiram do LDI e contribuíram para o avanço do conhecimento na área. Meu sincero agradecimento a todos que participaram deste esforço coletivo.

Essas iniciativas, que fomentaram as pesquisas e ações relacionadas à hanseníase nos últimos anos, foram fundamentais para a coleta e divulgação de dados, resultando em diversos estudos científicos que emergiram do LDI e contribuíram para o avanço do conhecimento na área. Meu sincero agradecimento a todos que participaram deste esforço coletivo.

Agradeço aos meus colegas de profissão e amigos do Laboratório de Pesquisa Operacional (LPO), Ermínio, Daniel, Carlos, Junior, André, Sandio, Leonardo, Ronilson e Marco. Aos amigos do Laboratório de Inteligência Computacional (LINC), Ewerton, Lucas, Igor Araújo e aos demais que puderam compartilhar a experiência da Pós-graduação. Meu muito obrigado.

Agradeço ainda a Albert Einstein e ao Prof. Me. Fernando Augusto pelo auxílio na gramática, sugestões de melhorias, correções e edição deste trabalho, participações estas, indispensáveis para o êxito desta produção. Meu sincero muito obrigado.

Por fim e não menos importante, agradeço à Fundação VALE e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo suporte financeiro que vem sendo dado ao longo do meu Doutorado; ao Programa de Pós-Graduação em Engenharia Elétrica (PPGEE); à Universidade Federal do Pará pela oportunidade de fazer a Pós-Graduação e a todos os profissionais envolvidos.

RESUMO

A hanseníase é um problema de saúde pública significativo que afeta, em grande parte, populações de baixo nível socioeconômico. Embora a Organização Mundial da Saúde (OMS) estabeleça diretrizes para diagnóstico, prevenção e tratamento, a detecção da doença enfrenta limitações, frequentemente resultando em diagnósticos tardios ou imprecisos e levando a complicações neurológicas graves e casos multirresistentes. Portanto, o diagnóstico precoce é essencial para reduzir a carga dessa doença. O aprendizado de máquina vem sendo largamente utilizado em diversas áreas da ciência e da indústria, mas especialmente na saúde, área em que desempenha um papel essencial na análise e tratamento de grandes volumes de dados. Neste sentido, esta tese investiga a aplicação de um modelo baseado em Ciência de Dados e Aprendizado de Máquina para atuar na especificação do perfil clínico de possíveis casos da hanseníase na Região Amazônica e, com isso, poder-se agir preventivamente no diagnóstico precoce e tratamento de pacientes em acompanhamento médico. O trabalho leva em consideração dados clínicos de pacientes provenientes de um conjunto de dados não públicos, coletados entre 2015 e 2020 na região Norte do Brasil. Logo, esta tese propõe um modelo de aprendizado para identificar grupos clinicamente afetados pela doença usando técnicas de Agrupamento e Random Forest. Nos resultados obtidos, o modelo proposto demonstrou eficiência ao avaliar a probabilidade de indivíduos estarem doentes, alcançando uma precisão de 90,39% na avaliação de performance e identificando uma probabilidade de 83,46% de um indivíduo estar doente, considerando um conjunto de variáveis epidemiológicas e não genéricas. Essa abordagem oferece uma visão promissora para o futuro da saúde, permitindo a formulação de estratégias eficazes para a identificação precoce de possíveis casos.

Palavras-chaves: Hanseníase, Diagnóstico, Ciência de Dados, Agrupamento, Random Forest.

ABSTRACT

Leprosy is a significant public health problem that largely affects low-income populations. Although the World Health Organization (WHO) establishes guidelines for diagnosis, prevention, and treatment, disease detection faces limitations, often resulting in late or inaccurate diagnoses and leading to serious neurological complications and multidrug-resistant cases. Therefore, early diagnosis is essential to reduce the burden of this disease. Machine learning has been widely used in several areas of science and industry, but especially in health, where it plays an essential role in the analysis and treatment of large volumes of data. In this sense, this thesis investigates the application of a model based on Data Science and Machine Learning to act in the specification of the clinical profile of possible leprosy cases in the Amazon Region and, thus, to be able to act preventively in the early diagnosis and treatment of patients under medical follow-up. The work takes into account clinical data of patients from a non-public dataset, collected between 2015 and 2020 in the North region of Brazil. Therefore, this thesis proposes a learning model to identify groups clinically affected by the disease using Clustering and Random Forest techniques. In the results obtained, the proposed model demonstrated efficiency in evaluating the probability of individuals being ill, achieving an accuracy of 90.39% in the performance evaluation and identifying a probability of 83.46% of an individual being ill, considering a set of epidemiological and non-generic variables. This approach offers a promising vision for the future of health, allowing the formulation of effective strategies for the early identification of possible cases.

Keywords: Leprosy, Diagnosis, Data Science, Clustering, Random Forest.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AED	Análise Exploratória de Dados
AHP	Analytic Hierarchy Process
BB	Borderline-Borderline
BCG	Bacilo de Calmette e Guérin
CSV	Comma-Separated Values
DW	Data Warehouse
ELISA	Enzyme Linked Immuno Sorbent Assay
ENIAC	Electronic Numerical Integrator and Computer
ER	Entidade Relacionamento
FN	Falso Negativo
FP	Falsos Positivos
IA	Inteligência Artificial
KDD	Knowledge Discovery in Databases
MB	Multibacilar
MCDM	Multi-Criteria Decision Making
MCDA	Análise de Decisão de Múltiplos Critérios
OMS	Organização Mundial da Saúde
OR	Odds Ratio
PCR	Reação em Cadeia da Polimerase
PDI	Pentaho Data Integration
PQT	Poliquimioterapia
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QCM	Quartz Crystal Microbalance

ROC	Receiver Operating Characteristic
SGBD	Sistema de Gerenciamento de Bancos de Dados
SINAN	Sistema de Informação de Agravos de Notificação
TFP	Taxa de Falsos Positivos
TFP	Quartz Crystal Microbalance
TOPSIS	Technique for Order of Preference by Similarity to Ideal Solution
TVP	Taxa de Verdadeiros Positivos
VP	Verdadeiros Positivos
VN	Verdadeiros Negativos

SUMÁRIO

1 INTRODUÇÃO

1.1 Visão Geral

A hanseníase, é uma doença infectocontagiosa crônica causada pelo *Mycobacterium leprae (M. leprae)* e o *Mycobacterium lepromatosis* (M. lepromatosis), um patógeno intracelular obrigatório com predileção pela célula de Schwann que infecta principalmente os nervos periféricos e envolve a pele e outros tecidos (BURKI, 2009).A hanseníase afeta principalmente os nervos periféricos, além da pele e outros tecidos, e continua a impactar milhões de pessoas em todo o mundo (CÁCERES-DURÁN et al., 2024). A hanseníase é atualmente uma doença desafiadora que constitui um problema de saúde pública em nações em desenvolvimento como o Brasil, ocupando o segundo posto global em quantidade de novos casos anuais, com mais de vinte mil registros por ano (SALGADO et al., 2016).

A hanseníase não possui prevenção primária, não havendo vacina específica contra o M. leprae, e os testes diagnósticos e prognósticos não são viáveis ou bem estabelecidos na rotina clínica. A epidemiologia da hanseníase revela uma predominância de casos com comprometimento nervoso, o que indica diagnósticos tardios e ressalta a ineficácia do controle epidemiológico em muitos países (VOLTAN et al., 2023). Isso resulta em lacunas que podem perpetuar a transmissão da doença. Além disso, os obstáculos relacionados a detecção precoce e a proteção da população suscetível à hanseníase, perpetuando essas lacunas e favorecendo a transmissão contínua. O longo período de incubação, aliado a sintomas e sinais insidiosos, contribui para que o diagnóstico seja muitas vezes tardio, atrasando o início do tratamento adequado e agravando o impacto na saúde pública (BERNARDES FILHO et al., 2021).

Segundo registros da Organização Mundial da Saúde (OMS), a hanseníase continua sendo uma preocupação global de saúde pública, especialmente em países como Índia, Brasil e Indonésia, que, no último ano, relataram mais de 10.000 novos casos cada, juntos respondendo por 79,3% dos novos casos detectados globalmente (SAUNDERSON, 2023). No Brasil, por exemplo, houve uma variação significativa no número de novos casos, contabilizando 27.863 novos casos e uma relativa queda nos anos seguintes, possivelmente devido à pandemia de COVID-19, que impactou a detecção e o tratamento de várias doenças. No entanto, após esse período, o número de casos voltou a aumentar, atingindo 22.773 em 2023, representando uma das maiores taxas de detecção global, com mais de 240.000 pessoas diagnosticadas nos últimos 10 anos (PINTO et al., 2020). Apesar das campanhas de conscientização e tratamento, a hanseníase continua a apresentar desafios, com a prevalência da doença variando em diferentes regiões do país.

É importante destacar que, apesar do contato íntimo e prolongado ser considerado o principal modo de difusão do M. leprae, alguns casos não conseguem ser relacionados ao con-

tato direto e/ou intercorrente com pacientes portadores de hanseníase. Este contexto reforça que há outras fontes de transmissão, como água, solo, plantas e diferentes espécies de animais, incluindo amebas, insetos, peixes e tatus (HUNTER; BRENNAN, 1981). Logo, não há vacina específica para o tratamento da hanseníase, e os testes para diagnóstico e prognóstico ainda não são amplamente estabelecidos na rotina clínica (SCOLLARD et al., 2006). Além disso, a hanseníase também está associada a outras doenças infecciosas crônicas que afetam uma parcela significativa de pessoas ao redor do mundo, como a tuberculose e a sífilis, que são as principais enfermidades discutidas nesse contexto (WU et al., 2018).

O tratamento da hanseníase é realizado com Poliquimioterapia (PQT), que combina rifampicina, clofazimina e dapsona por um período de 6 a 24 meses. No Brasil, esse tratamento é disponibilizado pelo Governo Federal através do Ministério da Saúde e parcerias públicas (ANDRADE et al., 1998). No entanto, o acesso a esses serviços pode ser desafiador em países em desenvolvimento, onde pacientes podem esperar mais de um ano para uma avaliação especializada. A demora no diagnóstico pode agravar o prognóstico e dificultar o tratamento efetivo; por isso, é crucial melhorar a capacidade de diagnóstico para otimizar o tratamento e reduzir complicações (ORGANIZATION et al., 2019).

O diagnóstico da hanseníase é essencialmente clínico, baseado em exame dermatoneurológico minucioso, em ferramentas como ELISA (Enzyme Linked Immuno Sorbent Assay) que é um teste sorológico imunoenzimático e na Reação em Cadeia da Polimerase (PCR), (GOULART; GOULART, 2008). Estas ferramentas provaram ser eficazes para o diagnóstico da doença e são úteis para a avaliação da eficácia da terapia, embora ainda haja uma grande limitação no acesso a exames médicos. Mesmo em grandes metrópoles, o tratamento/acompanhamento clínico de indivíduos possuem um alto custo, tempo de espera, logística dos pacientes, entre outros fatores naturais, que fazem o diagnóstico não ocorrer precocemente (JIN; CRUZ; GONÇALVES, 2020).

O diagnóstico precoce da hanseníase é crucial para interromper rapidamente a cadeia de transmissão e evitar o desenvolvimento de sequelas graves (GAMA et al., 2020a). A doença, com suas diversas manifestações clínicas, exige a criação e implementação de novas ferramentas para facilitar a detecção precoce, sugerir o tratamento adequado, classificar as formas clínicas e identificar estados reacionais (SANTANA et al., 2018). Embora existam recomendações para o diagnóstico e tratamento precoce, muitos métodos ainda necessitam de infraestrutura adequada e acompanhamento clínico, especialmente nas fases iniciais, quando o diagnóstico pode não estar completamente confirmado. Portanto, um diagnóstico rápido e eficaz é fundamental para a gestão bem-sucedida da hanseníase e para a prevenção de complicações mais sérias.

O diagnóstico precoce desempenha um papel crucial no tratamento eficaz da hanseníase, possibilitando intervenções terapêuticas oportunas que não apenas reduzem a transmissão da doença, mas também mitigam o risco de complicações e incapacidades. Apesar de sua impor-

tância ser reconhecida, existem desafios significativos no acesso a exames médicos, acompanhamento clínico personalizado e recursos médicos necessários para um diagnóstico precoce e tratamento efetivo da hanseníase. O avanço tecnológico, especialmente no campo computacional, oferece oportunidades promissoras para melhorar esse cenário, permitindo o desenvolvimento de ferramentas e técnicas mais precisas, rápidas e acessíveis para o diagnóstico precoce, monitoramento clínico e administração de tratamentos personalizados (SHARMA; SINGH, 2022).

Este processo de detecção da hanseníase envolve desafios tanto para os pacientes quanto para os profissionais de saúde. Tais fatores podem ser divididos em duas categorias: aqueles que contribuem para o "atraso do paciente,"definido como o tempo decorrido entre o início dos sintomas e a procura por atendimento; e aqueles que contribuem para o "atraso do sistema de saúde,"que se refere ao período entre a primeira consulta médica e o recebimento do diagnóstico. O objetivo é explorar esses atrasos separadamente, buscando identificar se o diagnóstico tardio está relacionado à demora dos pacientes em procurar atendimento ou a deficiências no próprio sistema de saúde (HENRY et al., 2016). O diagnóstico é baseado na identificação de pelo menos um dos seguintes sinais ou sintomas: a) lesões com alterações térmicas, dolorosas e/ou sensoriais; b) espessamento de nervos periféricos, associado a alterações sensoriais e/ou motoras; ou c) presença de Mycobacterium leprae, confirmada por baciloscopia ou biópsia de pele (BERNARDES-FILHO et al., 2021).

A maioria dos métodos atuais tem baixa sensibilidade e exige alta competência clínica, o que resulta em diagnósticos tardios e, por consequência, em deformidades e incapacidades severas. Isso destaca a importância do diagnóstico precoce, que oferece oportunidades significativas para melhorar essa realidade. Estratégias orientadas à tecnologia, especialmente aquelas baseadas em aprendizado de máquina, desempenham um papel crucial nesse contexto, pois possibilitam o desenvolvimento de ferramentas e técnicas mais precisas, rápidas e acessíveis para o monitoramento clínico, administração de tratamentos e identificação precoce de pacientes em risco (AHSAN; LUNA; SIDDIQUE, 2022).

1.2 Definição do Problema

A hanseníase está frequentemente associada a desafios no diagnóstico e tratamento adequados (DHARMAWAN et al., 2022). Com o crescimento populacional, especialmente em áreas com infraestrutura de saúde limitada, a capacidade dos sistemas de saúde para lidar com esses desafios diminui, agravando a situação e dificultando o controle eficaz da doença. No contexto da hanseníase, a habilidade dos profissionais de saúde em identificar os primeiros sinais e sintomas representa uma barreira, em parte devido à natureza complexa da doença (BETRU; MAKUA, 2023). Logo, a falta de diagnósticos adequados pode levar a um aumento na prevalência da doença e impactar negativamente os resultados clínicos dos pacientes.

Além de desafogar o sistema de saúde, o diagnóstico e o tratamento adequado da hanseníase são os principais meios para quebrar a cadeia de transmissão e reduzir as consequências

físicas e sociais da doença (FELICIANO; KOVACS; ALZATE, 1998). Apesar dos esforços para a contenção da hanseníase através de medidas medicamentosas, como a *Poliquimioterapia (PQT)*, que vem sendo realizada há mais de 40 anos, a doença continua a ser transmitida em todo o mundo, com uma média de 200.000 novos casos anuais, incluindo cerca de 10% de crianças (AUBRY et al., 2022). Além da PQT, existem diversas outras medidas disponíveis para interromper a cadeia de transmissão, que muitas vezes não são adotadas, como o exame de contatos (ARAÚJO et al., 2021). Esse aumento significativo pode ser devido a insuficiência na detecção precoce da doença, que é um dos principais desafios enfrentados globalmente.

Outros fatores também contribuem para o aumento de casos de hanseníase e seus impactos na sociedade: fatores socioeconômicos, como pobreza e falta de acesso a serviços de saúde de qualidade, bem como habitação em locais onde falta a presença do poder público, pois áreas com menor cobertura de programas sociais apresentam maior incidência de hanseníase (LEANO et al., 2019). A vulnerabilidade social e a estigmatização associada à doença dificultam a busca por tratamento precoce. A infraestrutura inadequada dos sistemas de saúde, especialmente em regiões endêmicas, também impede a realização de diagnósticos e tratamentos eficazes (TERAPÊUTICAS, 2022).

Logo, a hanseníase é uma doença que exige um diagnóstico clínico detalhado e preciso, conforme orientações da OMS (GAMA et al., 2020b). No entanto, como um dos grandes paradigmas, a doença exibe diferentes manifestações dermatológicas e neurológicas dentro de um amplo espectro clínico, o que causa um grande desafio diagnóstico (CHEN et al., 2021). Nesse cenário, a integração de alternativas integradas à Inteligência Artificial (IA) pode ser uma ferramenta valiosa, auxiliando na classificação de grupos de risco e na detecção precoce da hanseníase. Além disso, o uso de aprendizado de máquina pode aprimorar os processos diagnósticos ao analisar grandes volumes de dados clínicos e identificar padrões que podem não ser evidentes para os profissionais de saúde.

1.3 Hipótese de Solução

A definição do perfil clínico e a identificação de possíveis casos da hanseníase com base na utilização de técnicas de ciência de dados e aprendizado de máquina é uma importante alternativa para o diagnóstico precoce durante a etapa de acompanhamento clínico, possibilitando, desta maneira, a mitigação dos impactos severos e multirresistentes que a doença pode acarretar em pacientes detectados tardiamente.

1.4 Objetivo

1.4.1 Geral

O objetivo geral desta tese é desenvolver um modelo preditivo que utiliza ciência de dados e aprendizado de máquina para identificar grupos afetados pela hanseníase, que compar-

tilham similaridade de características clínicas em dados previamente coletados. A ideia é avaliar a eficiência do modelo na previsão do desenvolvimento precoce da doença,utilizando métricas de desempenho que comprovam sua eficácia a priori.

1.4.2 Específicos

A este respeito, os objetivos específicos abordados nesta tese são apresentados a seguir:

- Estudar o cenário de aplicabilidade de aprendizado de máquina à hanseníase, considerando aspectos clínicos, físicos, sociais e laboratoriais dos indivíduos;
- Realizar uma análise comparativa de algoritmos e modelos computacionais para otimizar a predição de casos de hanseníase, com foco na identificação precoce e na mitigação dos impactos da hanseníase;
- Propor um modelo de ciência de dados para pré-processamento e tratamento de dados coletados em campanhas de atendimento clínico nas unidades de saúde pública;
- Aplicar modelos de predição para identificação de possíveis casos de hanseníase;
- Desenvolver estratégias para o diagnóstico precoce da hanseníase com base em atributos clínicos, sociais e neurológicos dos indivíduos;
- Realização avaliação utilizando marcadores (PCR e PGL-i), especificando a probabildiade de individuos estarem doentes;
- Aplicar um modelo de predição para calcular a probabilidade de grupos de indivíduos estarem com hanseniase;
- Aplicar técnicas de ciência de dados para identificar grupos de comunicantes com predisposição para o desenvolvimento da comorbidade;
- Aplicar o modelo de predição para calcular a probabilidade de pacientes estarem doentes.
- Analisar os resultados gerados pelos modelos de tratamento de dados e de predição de possíveis casos de hanseníase, especificando a probabilidade de desenvolvimento da doença.

1.5 Organização do Texto

A estrutura desta tese foi cuidadosamente planejada para abarcar a complexidade do estudo sobre o perfil clínico da hanseníase na região amazônica, utilizando técnicas de ciência de dados e aprendizado de máquina. O trabalho envolve uma ampla gama de recursos, desde a fundamentação teórica sobre doenças negligenciadas até a aplicação prática de métodos avançados de análise de dados. O estudo aborda tanto os aspectos conceituais quanto a aplicação

de técnicas de análise e tomada de decisão, com o objetivo de identificar padrões clínicos em pacientes com hanseníase. Após a análise teórica, são discutidos os métodos aplicados e os resultados obtidos ao longo da pesquisa, seguidos por uma avaliação do desempenho das técnicas utilizadas e a apresentação das contribuições finais e sugestões para trabalhos futuros, conforme descrito a seguir:

- Capítulo 2: Neste capítulo, é apresentado um levantamento detalhado sobre os conceitos que abrangem o campo das doenças negligenciadas, com especial enfoque na hanseníase, ciência de dados e aprendizado de máquina. São explorados os fundamentos teóricos essenciais que sustentam essas grandes áreas, proporcionando uma compreensão aprofundada dos desafios e oportunidades que elas representam. A relevância deste capítulo reside em estabelecer as bases conceituais necessárias para a análise posterior, demonstrando a importância de se compreender tanto a natureza das doenças negligenciadas quanto o papel inovador das tecnologias de dados;

- Capítulo 3: Neste capítulo, é descrito o artigo intitulado "Use of Multi-criteria Methods to Support Decision-Making in Drug Management for Leprosy Patients", que apresenta uma abordagem detalhada sobre a gestão de medicamentos no tratamento de hanseníase. Este trabalho constitui uma contribuição parcial da tese, sendo um dos vários estudos desenvolvidos ao longo da pesquisa. Durante a produção da tese, foram concebidos métodos eficazes para o processo de tratamento e acompanhamento clínico dos indivíduos afetados pela doença, reforçando a importância de estratégias decisórias baseadas em critérios múltiplos para melhorar os resultados terapêuticos e a qualidade do atendimento aos pacientes;

- Capítulo 4: Neste capítulo, é descrito o artigo intitulado "A study about management of drugs for leprosy patients under medical monitoring: A solution based on AHP-Electre decision-making methods", o qual detalha a aplicação de métodos multicritério no acompanhamento clínico e gestão de medicamentos para pacientes de hanseníase. Assim como no capítulo anterior, este estudo representa uma contribuição parcial para a tese, fazendo parte de um conjunto mais amplo de investigações realizadas ao longo da pesquisa. O trabalho reforça o desenvolvimento de um processo eficaz para o monitoramento clínico dos pacientes, destacando a relevância de métodos como o AHP-Electre na tomada de decisão para otimizar o tratamento de indivíduos com hanseníase e aprimorar os resultados clínicos;

- Capítulo 5: Neste capítulo, são apresentados os desdobramentos principais da tese, representando o ponto culminante dos estudos realizados nos capítulos anteriores. A contribuição central do trabalho é detalhada com foco na definição do perfil clínico de pacientes com hanseníase e na identificação de possíveis novos casos, utilizando técnicas avançadas de ciência de dados e aprendizado de máquina. A metodologia proposta é aplicada

diretamente no acompanhamento clínico, permitindo uma análise mais precisa e eficiente do tratamento e progressão da doença. Além disso, o capítulo explora como a utilização de algoritmos de aprendizado de máquina possibilitou uma identificação precoce e mais confiável dos casos, otimizando o monitoramento e o cuidado oferecido aos pacientes. Os métodos desenvolvidos foram validados com dados reais, demonstrando sua eficácia e potencial impacto na gestão clínica da hanseníase, especialmente em contextos com recursos limitados;

- Capítulo 6: Neste capítulo, é realizada a avaliação de desempenho do modelo proposto, consolidando os resultados obtidos ao longo da tese e reforçando as contribuições discutidas no Capítulo 5. A análise dos resultados é feita com base nos dados reais utilizados no estudo, demonstrando a eficácia e aplicabilidade do modelo para a definição do perfil clínico de pacientes com hanseníase e a identificação de novos casos. A discussão aborda de forma detalhada os impactos da metodologia proposta, validando sua relevância no contexto clínico e a melhoria que traz ao acompanhamento e tratamento dos pacientes;

- Capítulo 7: Neste capítulo serão apresentadas as considerações finais, abrangendo tanto os resultados obtidos quanto as contribuições finais e parciais decorrentes do trabalho. Serão revisadas as principais descobertas e avanços trazidos pela pesquisa, ressaltando o impacto das abordagens utilizadas. Além disso, serão discutidas as contribuições adicionais que emergiram ao longo do desenvolvimento da tese. Finalmente, será feita uma projeção dos trabalhos futuros, indicando possíveis desdobramentos e novas áreas de investigação a serem exploradas, a fim de dar continuidade e expandir os resultados já alcançados.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Considerações iniciais

Este capítulo oferece uma visão geral dos principais conceitos relacionados à hanseníase, ciência de dados e aprendizado de máquina, fundamentais para a concepção e o desenvolvimento desta tese. Serão abordadas as principais tecnologias e metodologias descritas na literatura científica, com ênfase no tratamento de dados e aplicação de modelos de aprendizado, assegurando o embasamento teórico necessário para esta tese de doutorado.

2.2 Aspectos da hanseníase

A hanseníase, causada pelo patógeno humano *Mycobacterium leprae*, é uma doença crônica que causa danos à pele e aos nervos, resultando em uma ampla gama de lesões cutâneas, inflamação e dor nos nervos, levando até mesmo à perda de sensibilidade, atrofia e perda óssea, culminando em incapacidade, com o consequente estigma social. A OMS reconhece a hanseníase como um problema de saúde pública, especialmente em países de renda média e baixa, como Índia, Brasil e Indonésia, onde 79,6% de todos os novos casos globais foram relatados em 2019, ano em que 202.185 novos casos foram detectados globalmente (BOUTH et al., 2023).

Esta instituição estabeleceu protocolos claros para o diagnóstico, tratamento e monitoramento da hanseníase, com o objetivo de garantir o manejo adequado da doença e prevenir complicações graves. O diagnóstico precoce é essencial para evitar a progressão da doença e a OMS recomenda a identificação de três sinais cardinais: lesões cutâneas com alteração da sensibilidade, espessamento dos nervos periféricos e resultados positivos em testes bacteriológicos. Essas diretrizes são fundamentais para garantir a detecção precoce e a interrupção da transmissão da doença (BRASIL, 2024c).

Os principais atributos da hanseníase incluem a perda de sensibilidade nas áreas afetadas da pele, o espessamento dos nervos periféricos, fraqueza muscular e perda funcional. As variáveis clínicas mais relevantes na hanseníase são a classificação da doença em paucibacilar (PB), quando há menos lesões e uma baixa carga bacteriana, e multibacilar (MB), quando há mais lesões e uma maior quantidade de bactérias. Identificar essas variáveis corretamente é crucial, pois a OMS recomenda diferentes regimes de tratamento conforme a gravidade da infecção, que podem durar entre 6 a 12 meses de poliquimioterapia (PQT) para garantir a cura completa e prevenir incapacidades permanentes (BRASIL, 2024a), (BRASIL, 2024b).

O impacto do diagnóstico precoce é significativo. Quando os casos de hanseníase são detectados em estágios iniciais, a resposta ao tratamento tende a ser mais eficaz, minimizando

o risco de sequelas físicas e sociais. Isso se reflete na redução de deformidades e na melhoria da qualidade de vida dos pacientes. A OMS reforça a importância de campanhas de educação em saúde e treinamento de profissionais para melhorar a detecção precoce e garantir que os pacientes recebam tratamento adequado o mais rápido possível. Essas ações são parte de uma estratégia global para eliminar a hanseníase como um problema de saúde pública (ORGANIZATION et al., 2021).

2.3 Modelo de Decisão Multicritério

O Modelo de Decisão Multicritério (MCDM) é uma ferramenta crucial para resolver problemas complexos que envolvem múltiplos critérios de avaliação. Esses modelos auxiliam na tomada de decisões em situações em que várias alternativas precisam ser classificadas ou escolhidas com base em diferentes critérios, cada um com seu grau de importância. Os três principais métodos amplamente utilizados para este propósito são o Processo Analítico Hierárquico (AHP), o Técnica para Ordem de Preferência por Similaridade à Solução Ideal (TOPSIS) e o ELECTRE II. Cada método oferece uma abordagem distinta para a avaliação e comparação de alternativas, sendo aplicável em diversos setores, como engenharia, gestão e planejamento estratégico (ROY, 1991).

O AHP é um dos métodos mais conhecidos e amplamente aplicados em MCDM. Desenvolvido por Thomas Saaty, na década de 1970, o AHP permite decompor um problema de decisão em uma hierarquia de subproblemas, facilitando a comparação dos elementos em pares. A importância relativa dos critérios e alternativas é determinada através de matrizes de comparação, com os resultados finais representando uma priorização das alternativas. O AHP é amplamente utilizado em planejamento urbano, alocação de recursos e avaliação de desempenho, permitindo decisões mais bem fundamentadas com base em critérios qualitativos e quantitativos (GOLDEN; WASIL; HARKER, 1989).

Já o TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) é outro método muito utilizado em MCDM, conhecido por sua simplicidade e eficácia. O conceito por trás do TOPSIS é que a melhor solução é aquela que tem a menor distância em relação à solução ideal e a maior distância em relação à solução anti-ideal. Esse método tem sido aplicado em diversas áreas, incluindo logística, onde ajuda a identificar as melhores rotas de transporte, e em análise de risco, onde classifica projetos ou investimentos com base em múltiplos critérios de desempenho (TZENG; HUANG, 2011).

O ELECTRE II é um modelo de decisão multicritério que utiliza o conceito de superação para classificar alternativas. Esse método, parte da família de métodos ELECTRE (Elimination et Choix Traduisant la Realité), é particularmente eficaz em contextos onde há incerteza ou imprecisão nos dados. O ELECTRE II utiliza comparações entre alternativas para determinar o grau de dominância de uma sobre a outra, aplicando um processo de eliminação de alternativas menos satisfatórias (ZANAKIS et al., 1998).

As abordagens multicritério como AHP, TOPSIS e ELECTRE II proporcionam aos tomadores de decisão uma maneira sistemática de avaliar alternativas complexas em vários contextos. No entanto, cada método possui suas limitações e vantagens. O AHP, por exemplo, pode ser sensível à inconsistência nas comparações, enquanto o TOPSIS assume que os critérios são independentes, o que nem sempre é verdade (BEN-ARIEH, 2002).

2.4 Aprendizado de Máquina

O aprendizado de máquina é considerado um ramo da área de Inteligência Artificial, sendo uma área especializada no estudo de sistemas que sejam capazes de aprender de forma automatizada a partir de dados. Esta capacidade de aprender com experiências passadas é algo que é desejado desde a criação dos primeiros computadores, como o *Electronic Numerical Integrator and Computer* (ENIAC), na década de 40. O benefício que uma máquina que pudesse aprender como um humano seria imensurável, como os sistemas capazes de realizar diagnósticos com base em históricos médicos (VEALE; BINNS, 2017).

O aprendizado de máquina oferece abordagens poderosas para analisar grandes volumes de dados. A combinação de aprendizado supervisionado e não supervisionado permite uma visão ampla dos padrões observáveis nesses pacientes. No supervisionado, um conjunto de exemplos de treinamento em que cada exemplo é associado a um rótulo conhecido. Tal rótulo é responsável por definir a qual classe o respectivo exemplo (instância) pertence (WITTEN et al., 2005). Já no não supervisionado, é fornecido ao sistema de aprendizado um conjunto de exemplos de maneira que o objetivo seja construir um modelo que procure regularidades em tais exemplos, formando assim, agrupamentos ou *clusters* de características similares (BATISTA et al., 2003).

No aprendizado supervisionado, algoritmos como Random Forest (RIGATTI, 2017) e Regressão Logística (WRIGHT, 1995) são amplamente utilizados devido à sua capacidade de trabalhar com variáveis categóricas e contínuas. No contexto desta tese, diversos algoritmos podem ser aplicados para identificar padrões e prever o perfil dos pacientes. Esses modelos são treinados com dados rotulados, permitindo a previsão de novos casos com base nos padrões observados. A Random Forest é especialmente eficaz em lidar com dados complexos e não lineares, enquanto a Regressão Logística se destaca em problemas de classificação binária, oferecendo simplicidade e precisão ao prever probabilidades (HONG et al., 2024).

Por outro lado, o aprendizado não supervisionado, com métodos como K-modes (CHATURVEDI; GREEN; CAROLL, 2001a) e ROCKClustering (GUHA; RASTOGI; SHIM, 2000), permite identificar padrões ocultos nos dados sem a necessidade de rótulos pré-definidos. Essas técnicas são úteis para encontrar subgrupos de pacientes com características semelhantes, que podem não ser imediatamente visíveis. Essa abordagem pode ser aplicada para encontrar agrupamentos de pacientes que compartilham fatores de risco comuns ou padrões de sintomas que precedem o desenvolvimento da doença. O K-modes é particularmente eficiente para dados

categóricos, enquanto o ROCK Clustering é excelente para lidar com grandes volumes de dados com relações complexas entre os atributos categóricos.(CARGNIN et al., 2024).

No geral, a principal vantagem de utilizar essas abordagens combinadas está na capacidade de explorar tanto a estrutura dos dados quanto as previsões individuais. Enquanto o aprendizado supervisionado foca em maximizar a acurácia das previsões com base em exemplos passados, o aprendizado não supervisionado em geral, busca maximizar a coesão nos grupos identificados e distância entre grupos. O aprendizado, de forma geral, consiste na execução de um programa que aprende a partir de dados de treinamento ou experiências anteriores. Esses modelos podem ter uma natureza preditiva, visando prever eventos futuros, ou descritiva, com o objetivo de ganhar conhecimento sobre os dados. Portanto, é essencial compreender a hierarquia dos diferentes tipos de aprendizado, que será detalhada na Figura 1.

Figura 1 – Exemplo de hierarquia do Aprendizado.

Adaptado de (MONARD; BARANAUSKAS, 2003).

Em geral, o processo de aprendizado usado em classificações é indutivo, como as redes neurais e árvores de decisão, que é o mecanismo responsável por realizar generalizações a partir de dados. O fluxo normalmente segue dois caminhos, o aprendizado supervisionado e o não supervisionado.

2.5 Ciência de Dados

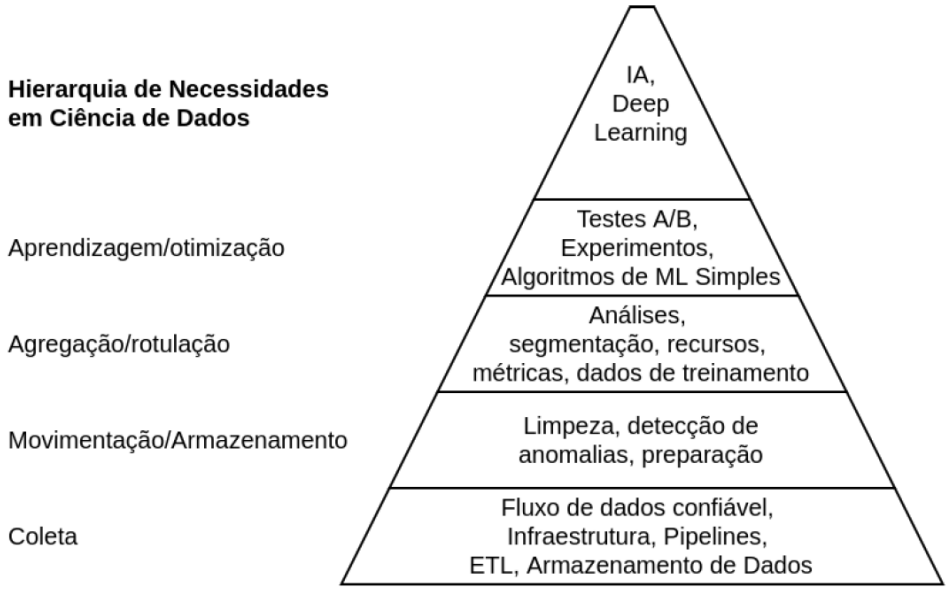
Embora a definição de Ciência de Dados seja ampla e multifacetada, ela pode ser entendida como a integração de disciplinas tradicionais, como estatística, mineração de dados, bancos de dados e sistemas distribuídos (VAN DER AALST; AALST, 2016). Essa área emergente destaca-se por combinar conhecimentos teóricos e práticas analíticas, utilizando abordagens sistemáticas para transformar grandes volumes de dados em valor tangível para indivíduos, organizações e a sociedade (CARVALHO; G. MENEZES; BONIDIA, 2024).

Nesta tese, o conceito de Ciência de Dados é utilizado para analisar informações provenientes de diferentes áreas, como medicina clínica, estatística e modelos de dados. A abordagem adotada demonstra como a ciência de dados pode integrar essas diversas fontes de informação para gerar insights significativos e soluções práticas. O uso deste conceito permite uma análise abrangente e detalhada, aproveitando metodologias e técnicas de diferentes disciplinas para abordar problemas complexos de maneira eficaz.

Antes da Ciência de Dados, o termo Mineração de Dados (*Data Mining*) era muito popularizado. Autores descrevem um processo geral para descoberta de conhecimento útil a partir de dados, o Knowledge Discovery in Databases (KDD), visto em (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Neste contexto, houve o surgimento da WEB 2.0, a segunda geração da WEB, onde os web sites deixaram de ser simplesmente provedores de conteúdo estático e passaram a funcionar como provedores de serviços com os quais os usuários podem interagir. Tecnologias como blogs, wikis e redes sociais marcam essa nova geração (MURUGESAN, 2007).

Apesar da utilização de sistemas avançados visando obtenção de vantagem estratégica usando tecnologia da área de dados, a maioria das empresas ainda não está pronta para esse tipo de abordagem, pois existe uma hierarquia de necessidades na Ciência de Dados em que necessidades mais básicas precisam ser atendidas inicialmente, como visto na Figura 2 a seguir.

Figura 2 – Hierarquia de Necessidades de Ciência de Dados.



Adaptado de (ROGATI, 2017).

Como se observa na Figura 2, a base da pirâmide costuma receber atenção, principalmente em temas relacionados a *Big Data* ou *Data Layer*. No topo da pirâmide são utilizadas aplicações que utilizam técnicas de Inteligência Artificial, dentre as quais chatbots[1], predições

[1] https://www.take.net/blog/chatbots/chatbot/

para sistemas de marketing e vendas, por exemplo. A parte central, onde residem análises, métricas, testes e experimentações, costuma não receber tanta atenção, apesar de conter o potencial para grandes avanços em boa parte dos negócios.

2.5.1 Conjunto de Dados

No contexto deste documento, bases de dados são conjuntos ou coleções de dados (Datasets) de qualquer natureza, organizados de maneira tabular, onde cada coluna representa uma variável ou atributo e cada linha representa um registro de dados. Conjuntos de dados podem ser distribuídos em diferentes formatos, como arquivos CSV (*Comma-Separated Values*), *Bancos de Dados Relacionais*, *SQLite*, arquivos *XML* (*eXtensible Markup Language*), dentre outros formatos, inclusive proprietários. Exemplos de conjuntos de dados incluem, mas não se limitam aos abaixo relacionados:

- Uma tabela ou arquivo CSV com dados;
- Uma coleção organizada de tabelas;
- Uma coleção de dados em formato SQLite;
- Um arquivo em formato proprietário com dados estruturados;
- Dados de captura de imagens.

Os conjuntos de dados que se enquadram na definição acima são, por si só, elegíveis para utilização em projetos de Ciência de Dados. Mas, se estes vierem acompanhados de mais informações que visem a facilitar a sua identificação e utilização, então, são conhecidos como metadados.

Metadados são dados que descrevem dados, ou seja, informações sobre um ou mais aspectos de outros dados. Informações como objetivo da criação de um conjunto de dados, data de criação, localização, autor, licença de uso, tamanho de arquivo e número de registro são típicos exemplos de metadados (VELLUCCI, 1998).

2.5.2 Armazenamento de Dados

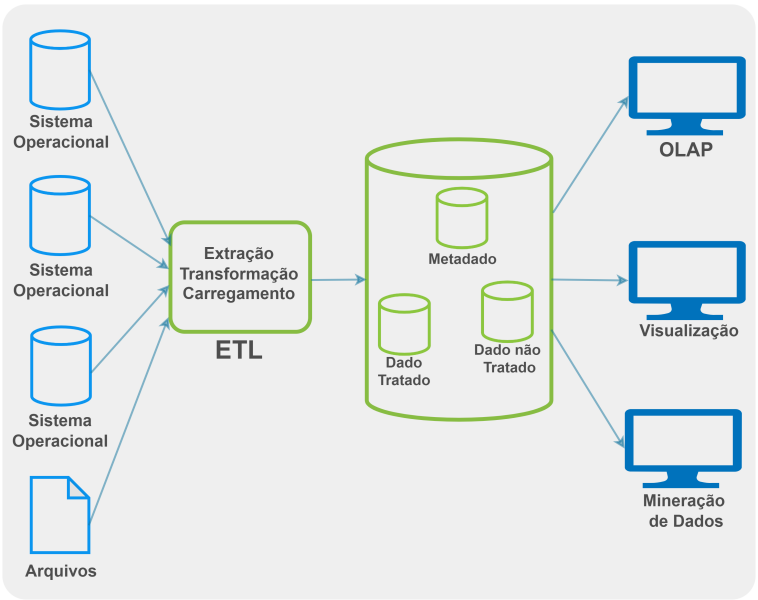
Os sistemas computacionais em instituições, tanto públicas quanto privadas, têm dois papéis principais: registro e análise. Os sistemas de registro controlam processos diários, como cadastros, lançamentos financeiros e emissão de documentos. São otimizados para eficiência operacional, permitindo operações rápidas e recuperação de dados com baixo custo computacional. Em contraste, o Processamento Analítico Online (OLAP) requer dados detalhados do Processamento de Transações Online (OLTP) para atender às necessidades analíticas, não tendo os mesmos requisitos operacionais, focando em operações de leitura e análise complexa de dados (CONN, 2005).

Esses sistemas, embora distintos em natureza e objetivos, compartilham o mesmo recurso essencial: dados. Enquanto os sistemas de transações lidam com operações diárias, os sistemas de análise otimizam a busca por insights e tendências a partir desses dados. Essa divisão permite uma gestão eficaz das operações cotidianas e uma análise aprofundada para embasar decisões estratégicas, representando a complementaridade entre operacionalidade e inteligência de dados nas organizações. Diante disso, autores afirmam que apesar de esses sistemas utilizarem elementos comuns, recomenda-se uma separação física de lógica (modelagem) das bases de dados utilizadas (KIMBALL, 1997).

A base de dados utilizada por sistemas de análise é conhecida como Data Warehouse (DW). Nesta tese é descrita como um repositório que coleta e armazena uma grande quantidade de informações, conforme referenciado por Gardner (1998). O uso de uma estrutura de DW nesta tese foi essencial devido ao volume de dados que estava sendo armazenado de forma ineficiente. Essa ineficiência resultava em problemas significativos de acesso, tratamento, consultas e processamento das informações em tempo real. A solução oferecida pelo Data Warehouse permite a consolidação dos dados, proporcionando uma melhor organização e facilitando o processamento de grandes volumes de informação.

Essa base é elaborada a partir de um processo de extração de dados dos sistemas transacionais, transformação destes e, ato contínuo, gravação na nova base criada. Esse processo de construção de um DW é conhecido como ETL (Extract, Transform, Load), ilustrado na Figura 3.

Figura 3 – Modelo de estrutura de um Data Warehouse (DW).



Adaptado de: https://www.ibm.com/cloud/learn/data-warehouse. Acessado em: 10/08/2022.

Essa abordagem de modelagem expressa na Figura 3, focada na otimização de consultas, busca minimizar dificuldades recorrentes identificadas em negócios, dentre as quais destacam-

se: a dificuldade de acessar os grandes volumes de dados disponíveis, o grande esforço para tratamento e limpeza destes dados e a consequente sensação de que grande parte do esforço de análise reside no processo de obtenção e tratamento de dados, a despeito da sua efetiva análise.

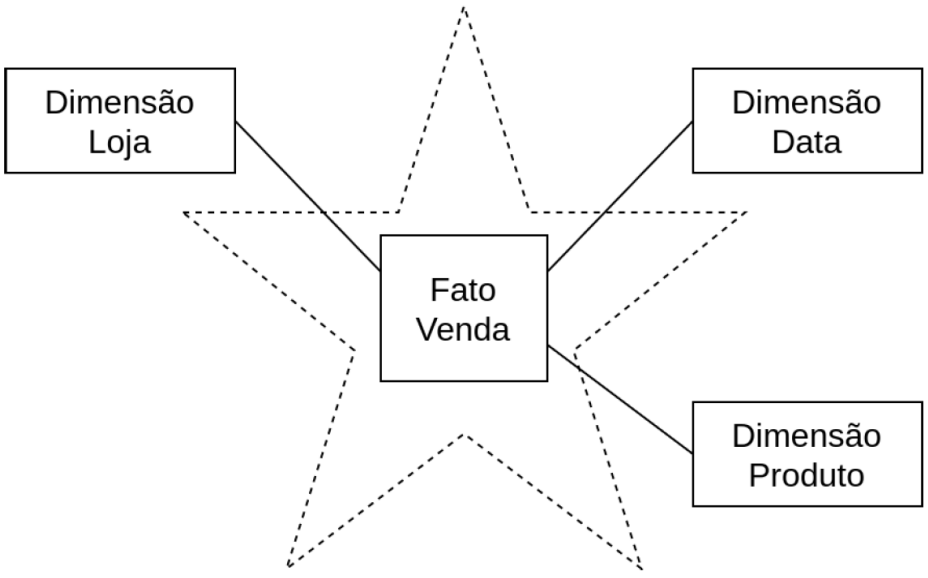
2.5.3 Modelo de Dados Dimensional

O modelo dimensional é uma técnica para modelagem de dados comumente utilizada em projetos de DW e que possui uma ampla adoção na indústria e na academia (KIMBALL, 1997). Entre os seus principais benefícios estão o fato de possuir um bom desempenho em operações de consulta e, principalmente, o fato de representar dados de forma intuitiva, o que torna a consulta por parte dos usuários do DW mais simples.

Apesar de o Sistema de Gerenciamento de Bancos de Dados (SGBD) utilizarem em muitos casos a modelagem dimensional, há uma grande diferença para os modelos tradicionais de Entidade Relacionamento (ER), também conhecido como Terceira Forma Normal (3FN). No ER, as tabelas do banco de dados são normalizadas na busca pela eliminação de redundância de dados.

Apesar disso, realizar consultas em tabelas modeladas da forma tradicional pode rapidamente se tornar confuso para usuários com foco em análise, uma vez que navegar entre tabelas, através dos seus relacionamentos por chaves estrangeiras, adiciona complexidade ao processo de consulta. Já o modelo dimensional contém a mesma informação do modelo normalizado, mas organizado de uma maneira que priorize a simplicidade de entendimento. Nele, a informação é organizada em topologia de estrela, conforme ilustrado na Figura 4.

Figura 4 – Topologia estrela utilizada em Modelagem Dimensional.



Adaptado de (KIMBALL, 1997).

Na Figura 4, as tabelas de fatos em modelos dimensionais registram medidas relaciona-

das a um processo organizacional. Nestas tabelas devem ficar armazenadas as informações na menor granularidade disponível e de maneira centralizada, evitando-se ao máximo a reprodução de dados de um determinado processo em mais de uma tabela.

2.6 Considerações Finais

Este capítulo teve como objetivo fundamental apresentar as principais tecnologias e conceitos retratados no desenvolvimento deste trabalho, expondo os componentes utilizados para formular o modelo de sistemas e os demais tópicos pertinentes à proposta. Esta fase da pesquisa adquire, em particular, a fundamentação proposta por este trabalho. Considera-se, ainda, este capítulo como de grande relevância, uma vez que apresenta os principais conceitos e tecnologias associados à hanseníase, aprendizado de máquina, Ciência de Dados e Modelagem de Dados.

3 ARTIGO 1: APLICAÇÃO DE MÉTODOS MULTICRITÉRIOS

3.1 Considerações iniciais

Este capítulo discute o artigo intitulado "Use of Multi-criteria Methods to Support Decision-Making in Drug Management for Leprosy Patients", que é o estudo preliminar desta tese. O método proposto no trabalho envolve um modelo de dados, pré-processamento de informações e a aplicação de um método de decisão multicritério. A metodologia busca identificar os pacientes mais clinicamente afetados, priorizando-os no atendimento médico e selecionando as características clínicas e físicas mais acentuadas nos indivíduos. Além disso, serão apresentados os resultados obtidos e a relevância do estudo para a formulação da tese.

3.2 Visão Geral

A hanseníase continua sendo um grande desafio de saúde pública global. Embora haja tratamento disponível, a doença continua a ser negligenciada, principalmente devido à escassez de medicamentos e à má distribuição dos mesmos. Esses problemas têm contribuído significativamente para o estabelecimento da doença, resultando em complicações graves e resistentes a múltiplos medicamentos. Neste cenário, os danos na pele e nos nervos tornam-se sinais cardinais, causados por um distúrbio imunológico que pode desencadear episódios inflamatórios severos (WU et al., 2018).

A hanseníase é uma doença de difícil diagnóstico, que apresenta uma ampla gama de sintomas, além de alta capacidade de contágio (ALVES; FERREIRA; FERREIRA, 2014). Essas características reforçam que o tratamento nos pacientes diagnosticados deve ser realizado de forma eficiente e regular. Uma vez que a administração de medicamentos não é feita de forma eficiente, há uma intensificação de vários problemas no controle da doença, principalmente considerando a região amazônica, que é uma área com recursos limitados. Este cenário pode piorar ainda mais nos próximos anos, já que mais de 200 mil novos casos de hanseníase foram confirmados em todo o mundo apenas em 2018 (ORGANIZATION et al., 2019).

Diante deste cenário, o referido trabalho, intitulado “Uso de métodos multicritério para apoiar a tomada de decisões no manejo de medicamentos para pacientes com hanseníase” (Publicado em 2023 em: International Journal of Management and Decision Making), utiliza uma metodologia de distribuição para otimizar a administração de medicamentos visando efetivamente atuar na etapa de diagnóstico da doença, especificando atributos de maior impacto para a disseminação da doença, avaliando o perfil clínico de pacientes e especificando o perfil de indivíduos com maior predisposição para a hanseníase. Na prática, com os modelos analíticos previamente configurados, pode-se fornecer uma visão geral controladora sobre um grande conjunto de dados.

Quando se trata de métodos de decisão multicritério aplicados a doenças negligenciadas, há certa limitação na literatura científica para problemas dessa natureza. Apesar disso, (KRYSANOVA; KRYSANOV; ERMAKOVA, 2017) aplicam modelos de decisão multicritério para análise de ensaios clínicos, avaliando o uso de medicamentos em pacientes com *doença de Huntington*. O método de decisão também é usado em (PINAZO et al., 2021) para criar uma classificação de intervenções médicas e ações necessárias para o gerenciamento de pacientes afetados pela doença de Chagas na Bolívia. Outros estudos, como em (ROLLES et al., 2021), implementam análise de decisão multicritério em regimes regulatórios de heroína genérica.

Vários aspectos são avaliados ao aplicar métodos analíticos para auxiliar na tomada de decisão, como o grau de importância dos medicamentos, grupo de risco, classificação e identificação de soluções finais. Nessa visão, aspectos multicritério e de tomada de decisão também foram usados durante o tratamento de radioterapia para moldar a distribuição de dose 3D dentro do paciente (BREEDVELD et al., 2019), equilibrando até 30 critérios que estão sujeitos a constantes mudanças mecânicas. Esses critérios ajudam as pessoas a considerar efetivamente critérios conflitantes para comparar o desempenho geral de diferentes alternativas (PROVOST; FAWCETT, 2013).

Considerando que o tratamento da hanseníase é um processo demorado e escasso devido a questões financeiras do Estado, foi aplicado o Método de Tomada de Decisão Multicritério (Multi-Criteria Decision Making ou MCDM) para tornar o atendimento mais eficiente. O trabalho utiliza um banco de dados não público com informações de pacientes coletadas no período 2016-2020 em 66 municípios do Estado do Pará.

Como principal contribuição, esta proposta busca fornecer aos profissionais um mecanismo eficiente para priorização de pacientes com maior gravidade que estão em processo de tratamento, que ainda não foram tratados e devem ser avaliados de forma priorizada. Apesar de muitos estudos abordando o uso do MCDM para diagnóstico e tratamento, a literatura sobre administração de medicamentos ainda é limitada (MOHAMMED et al., 2020), (VILLANUEVA et al., 2021), (AHMAD et al., 2021). Esta estratégia é uma alternativa de baixo custo computacional que pode ser usada durante todo o tratamento do paciente com esforço operacional mínimo e também pode mitigar os efeitos da distribuição de medicamentos não tão eficiente.

3.3 Resultados Obtidos

Os resultados deste estudo destacam a eficácia do uso do Método de Decisão por Multicritérios, especificamente do AHP (Analytic Hierarchy Process) e do TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution), na priorização de pacientes com hanseníase. Inicialmente, o AHP foi utilizado para definir e avaliar os critérios de decisão, incluindo Tipo de Paciente, Estado do Paciente, Número de Lesões, Forma Clínica, Classificação, Baciloscopia e Tratamento. A partir dessa avaliação, foram atribuídos pesos a cada critério, permitindo a criação de uma matriz de comparação que ajudou a identificar os pacientes mais sensíveis ao

tratamento. Com um índice de consistência de 10,5%, os resultados do AHP foram considerados robustos e confiáveis.

Na etapa seguinte, o TOPSIS foi aplicado para classificar os pacientes de acordo com a urgência do atendimento, considerando suas condições clínicas. A matriz normalizada e os pesos ponderados facilitaram o cálculo das distâncias euclidianas das alternativas em relação às soluções positivas e negativas ideais. O resultado final foi um ranking que prioriza os pacientes com sintomas mais graves de hanseníase, garantindo que aqueles com necessidades mais críticas sejam atendidos primeiro. Os resultados do modelo podem ser observados na matriz D (Tabela 1).

Tabela 1 – Matriz D de Alternativas TOPSIS.

Código	*Si+*	*Si-*	*Pi*	*Ranking*
P21	0.011	0.020	0.656	1
P15	0.011	0.020	0.655	2
P10	0.011	0.020	0.655	3
P7	0.011	0.020	0.655	4
P3	0.011	0.020	0.655	5
...	...	...	...	...
P761	0.0269	0.0039	0.1273	1006
P795	0.0269	0.0039	0.1273	1007
P850	0.0269	0.0039	0.1273	1008
P593	0.0269	0.0018	0.0643	1009
P594	0.0269	0.0018	0.0643	1010

Na Tabela 1 são mostradas as alternativas geradas pelo modelo TOPSIS de decisão multicritério. Os valores obtidos representam as distâncias euclidianas das alternativas em relação à solução ideal positiva e negativa. A matriz D é crucial para calcular os índices de proximidade relativos das alternativas em relação às soluções ideais, permitindo a classificação das alternativas de acordo com a sua prioridade. Quanto menor a distância de uma alternativa à solução ideal positiva ($Si+$) e maior a distância à solução ideal negativa ($Si-$), melhor será a classificação dessa alternativa.

É importante ressaltar que este estudo analisou 911 registros de pacientes em acompanhamento clínico para hanseníase, coletados entre 2016 e 2020. Inicialmente, observou-se que a maioria desses pacientes estava registrada como não tendo recebido nenhum dos tratamentos disponíveis, como a PQT-Poliquimioterapia ou esquemas alternativos. Contudo, essa conclusão foi precipitada, pois muitos desses pacientes estavam sendo atendidos em unidades básicas de saúde, cujos dados não foram incluídos na análise, pois seria necessário que os municípios liberassem acesso aos dados sigilosos que permitam a identificação da pessoa.

A justificativa para essa menção está no fato de que, embora os registros coletados provenham da rede pública, eles estão armazenados em bases de dados que não são de acesso público, o que justifica a ausência de certas informações no estudo. Usando os dados abertos

do Sistema de Informação de Agravos de Notificação (SINAN), não é possível verificar se uma pessoa específica recebeu tratamento, já que os dados disponíveis são agregados e não identificáveis individualmente. Para acessar essas informações sigilosas e verificar tratamentos, seria necessário que os municípios liberassem previamente o acesso a esses dados confidenciais, o que requer uma aprovação formal. Essa liberação não acontece de forma automática, o que limita a análise mais detalhada de certos aspectos dos tratamentos realizados.

3.4 Relevância para a Tese

A pesquisa apresentada utiliza o método AHP-TOPSIS para priorizar pacientes com hanseníase com base em diversos critérios clínicos e laboratoriais. A aplicação dessa metodologia permitiu desenvolver uma estratégia de priorização para pacientes com status de doente e mais de 10 lesões cutâneas, características que podem refletir uma forma avançada e disseminada da infecção. Os pacientes multibacilares (MB), que apresentam fraca resposta imune mediada por células contra M. leprae, desenvolvem alta carga bacilar, tornando-se a principal fonte de infecção (NOBRE et al., 2017). Isso os classifica como os de maior gravidade, devido ao maior risco de transmissão e necessidade de tratamento prolongado. Por meio do modelo, foi possível priorizar essas características severas, agrupando esses pacientes em um grupo de alto risco. A relevância desse trabalho para uma tese de doutorado focada no estudo de pacientes mais afetados pela hanseníase é inegável.

A utilização de dados endêmicos da região amazônica, provenientes de um projeto de pesquisa específico anteriormente mencionado, agrega valor ao estudo, pois aborda diretamente as particularidades clínicas da área. Este projeto, que inclui uma busca ativa para identificação e tratamento de casos de hanseníase, tem sido uma iniciativa crucial na luta contra a doença. Com anos de atuação em diversas partes do país, o projeto não apenas possibilita a coleta de dados relevantes, mas também a implementação de estratégias de intervenção adaptadas às necessidades locais. Ao focar na realidade da região amazônica, que enfrenta desafios únicos devido à sua geografia e infraestrutura de saúde limitadas, o estudo proporciona uma compreensão mais aprofundada das dinâmicas da hanseníase. Isso contribui para o desenvolvimento de políticas públicas mais eficazes e direcionadas

A aplicação de um modelo de ciência de dados e aprendizado de máquina pode fornecer insights profundos sobre os perfis dos pacientes, permitindo uma intervenção preventiva mais eficaz e uma gestão otimizada do tratamento, essencial para reduzir a carga da hanseníase. Utilizar o aprendizado de máquina para identificar grupos clinicamente afetados com características semelhantes e aplicar modelos preditivos para auxiliar no diagnóstico precoce da hanseníase na região amazônica é fundamental. O artigo mencionado reforça a relevância de aplicar essas técnicas, destacando a eficácia, alta eficiência e baixo custo desses modelos na identificação de padrões complexos nos dados clínicos, permitindo intervenções rápidas e precisas.

Além disso, a implementação desses modelos em regiões remotas, como a Amazônia,

pode reduzir significativamente os custos operacionais e melhorar o acesso aos serviços de saúde, contribuindo positivamente para a detecção e tratamento oportuno da hanseníase. A tese deve, portanto, explorar a eficácia desses métodos na identificação de grupos vulneráveis e avaliar o impacto de sua aplicação em termos de saúde pública. A pesquisa pode se beneficiar ainda mais ao discutir tópicos como a integração de dados clínicos, o uso de algoritmos avançados de aprendizado de máquina e a adaptação das tecnologias às condições locais de infraestrutura e recursos disponíveis.

3.5 Considerações Finais

Em conclusão, o estudo apresentado neste capítulo demonstrou a eficácia de um modelo MCDM para a priorização de pacientes com hanseníase, focando na etapa de administração de medicamentos. O ranking gerado pelo AHP e TOPSIS, juntamente com dados de acompanhamento clínico exibidos, confirmou a viabilidade dessa abordagem para problemas de saúde pública, dada sua precisão e baixo custo computacional. Este trabalho serviu como base para a construção da tese de doutorado em questão, contribuindo significativamente para o desenvolvimento de uma metodologia robusta e aplicada ao manejo de pacientes com quadros clínicos mais graves. Assim, pacientes com diagnóstico positivo, lesões superiores a cinco e baciloscopia positiva devem receber tratamento prioritário, enquanto aqueles com quadro menos grave devem ser manejados de acordo com as diretrizes clínicas menos urgentes.

4 ARTIGO 2: GERENCIAMENTO MEDICAMENTOSO COM BASE EM AHP-ELECTRE

4.1 Considerações Iniciais

Neste capítulo, o artigo "A Study About Management of Drugs for Leprosy Patients Under Medical Monitoring: A Solution Based on AHP-Electre Decision-Making Methods" será explorado em maior profundidade, destacando sua evolução no contexto desta tese. O modelo original foi reformulado, incorporando o método multicritério ELECTRE II para aprimorar a precisão e a robustez na priorização dos pacientes. Além disso, foram incluídos novos atributos como dados de entrada, permitindo uma análise mais abrangente e detalhada dos pacientes. Um dashboard interativo foi desenvolvido para exibir os resultados de maneira clara e dinâmica, facilitando a visualização e interpretação dos pacientes mais gravemente afetados pela doença. Esta evolução metodológica contribui para a fundamentação da tese, ao oferecer uma ferramenta de apoio à decisão ainda mais eficaz e adaptada às necessidades do manejo de hanseníase.

4.2 Visão Geral

A hanseníase é uma das doenças tropicais negligenciadas listadas como um grande problema de saúde global. O tratamento é uma das principais alternativas, porém, a escassez de medicamentos e sua má distribuição são fatores importantes que têm impulsionado a propagação da doença, levando a complicações irreversíveis e multirresistentes. Nesse contexto, a demora no diagnóstico/tratamento pode levar a deformidades permanentes, como lesões de nervos periféricos e deformidades graves, o que, além de agravar o quadro dos pacientes, intensifica os impactos dos estigmas sociais (sinal que designa o portador como desqualificado) (SANTÉ; ORGANIZATION et al., 2019) e (BRASIL, A., 2016).

Por ser uma doença complexa de diagnosticar devido à sua variedade de sintomas e alta contagiosidade, a hanseníase é complexa. Um tratamento regular e eficiente é essencial para os pacientes diagnosticados, especialmente em áreas de recursos limitados, como a região amazônica. A administração inadequada de medicamentos intensifica os problemas no controle da doença, e esse cenário pode piorar, considerando o grande número de novos casos anuais. O diagnóstico precoce é fundamental para iniciar o tratamento antes que ocorram danos graves, melhorando o controle da transmissão e a qualidade de vida dos pacientes.

Diante disso, este estudo, intitulado "Um estudo sobre o manejo de medicamentos para pacientes com hanseníase sob acompanhamento médico: uma solução baseada nos métodos de tomada de decisão AHP-Electre", publicado em 2023 em PLOS Neglected Tropical Diseases, aplica dois modelos de tomada de decisão multicritério: o *Analytic Hierarchy Process (AHP)* proposto por (KIRCHHEIMER; STORKS et al., 1971) e *ÉLimination et Choix Traduisant la*

REalité II (ELECTRE II) proposto por (ROY; BERTIER, 1971) para priorização no processo de administração de medicamentos no tratamento da hanseníase. Esta estratégia é uma alternativa de baixo custo computacional que pode ser utilizada durante todo o tratamento do paciente com mínimo esforço operacional e também pode mitigar os efeitos da distribuição de medicamentos não tão eficientes. Os resultados obtidos são apresentados em um visualizador de dados interativo.

É importante destacar que a escassez de medicamentos que envolve condições de risco de vida das pessoas é outra discussão de grande impacto no campo das doenças negligenciadas. Essa abordagem é enfatizada em (MOOSIVAND et al., 2021), onde os autores avaliam estratégias de decisão multicritério baseadas em atributos práticos de distribuição de medicamentos. Consoante a isso, (VISHWAKARMA; PRAKASH; BARUA, 2016) propõem uma metodologia de tomada de decisão MCDM baseada na abordagem fuzzy-AHP para priorizar e classificar riscos na cadeia de suprimentos farmacêuticos.

As avaliações multicritério estão sendo cada vez mais empregadas na priorização de ameaças à saúde. Em (DE NARDO et al., 2020), os autores usam Análise de Decisão de Múltiplos Critérios (MCDA) para determinar pesos para onze critérios para priorizar pacientes não críticos com COVID-19 para admissão hospitalar em ambientes de saúde com recursos limitados. Essa abordagem também pode ser percebida em (DE NARDO et al., 2020) que avalia o valor de uma estrutura de Análise de Decisão Multicritério para avaliação de medicamentos na Catalunha na Espanha (Catalan Health Service). Além disso, a aplicação de métodos de decisão multicritério pode ser facilmente integrada às políticas de distribuição de medicamentos em doenças de alta complexidade.

Dessa forma, considerando que o tratamento da hanseníase é um processo demorado e escasso devido a questões financeiras do Estado, a proposta visa tornar o atendimento mais eficiente. O trabalho utiliza um banco de dados não público com informações de pacientes coletadas no período de 2015-2020. Como principal contribuição, esta proposta busca fornecer aos profissionais um mecanismo eficiente para priorização e visualização de dados de pacientes com maior gravidade, que ainda não foram tratados e devem ser avaliados de forma priorizada.

O diagnóstico precoce é fundamental para a eficácia do tratamento da hanseníase, pois permite intervenções rápidas que podem prevenir complicações severas e incapacidades permanentes. Isso não só melhora a qualidade de vida dos pacientes, mas também reduz a transmissão da doença. Além disso, a implementação de um sistema de priorização eficiente pode fornecer à comunidade acadêmica dados valiosos para pesquisas futuras, contribuindo para avanços no controle e tratamento da hanseníase.

4.3 Resultados Obtidos

O uso dos modelos analíticos AHP e ELECTRE II demonstra sua importância na identificação e priorização de pacientes clinicamente afetados pela tuberculose, com base em informações clínicas e laboratoriais. Primeiramente, o AHP permite uma avaliação criteriosa dos atributos relevantes, como tipo de paciente, estado do paciente e número de lesões, oferecendo uma estrutura hierárquica para a tomada de decisões. A partir da construção dessa hierarquia e da avaliação pareada dos critérios, os pesos são estimados, permitindo uma avaliação global das alternativas, ou seja, dos pacientes, de forma objetiva e fundamentada. A obtenção de um índice de consistência favorável reforça a confiabilidade desse método na seleção de pacientes prioritários.

Em seguida, o uso dos dados obtidos pelo AHP como entrada para o ELECTRE II amplia a análise, permitindo uma seleção mais precisa das alternativas que mais necessitam de atendimento. O ELECTRE II utiliza os atributos definidos, como tipo de paciente, situação do paciente e tratamento, para determinar as prioridades com base em sua condição clínica. Esse processo de análise comparativa, aliado aos critérios definidos, resulta em uma classificação final das alternativas, destacando os pacientes mais afetados pela doença, conforme visto na Tabela 2.

Tabela 2 – Ranking de alternativas obtidas com ELECTRE II.

Ranking	**Alternativas**	**Critério**						
		A1	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**
1º	P44	4	2	4	7	2	2	5
2º	P178	4	2	4	4	2	2	5
3º	P48	4	2	3	7	2	2	5
4º	P5	4	2	4	2	2	2	5
5º	P54	4	2	0	7	2	2	5
6º	P51	4	2	3	2	2	2	5
7º	P52	4	2	2	2	2	2	5
8º	P175	4	2	1	3	2	2	5
9º	P228	4	2	1	2	2	2	5
10º	P29	4	2	0	4	2	2	5

Os critérios A1 a A7, apresentados na Tabela 2, são categóricos e possuem representações numéricas de acordo com o número total de categorias. Essas representações foram organizadas de forma que quanto maiores os valores dos critérios apresentados na matriz de desempenho, maior a prioridade do paciente no processo de ordenação, devido à gravidade do seu caso. Isso porque, nesse contexto, o ELECTRE II tem como objetivo escolher as alternativas que maximizem os desempenhos dos critérios, considerando seus pesos.

Por fim, a aplicação desses modelos oferece uma visão otimizada para uma administração eficiente de medicamentos, priorizando pacientes com base em critérios clínicos específicos. Além disso, a implementação de um dashboard interativo, como descrito, não apenas apresenta

os resultados obtidos, mas também otimiza o acompanhamento clínico dos pacientes a longo prazo, fornecendo uma ferramenta valiosa para os profissionais de saúde na tomada de decisões estratégicas e na alocação de recursos.

4.4 Relevância para a Tese

O artigo em questão é altamente relevante para esta tese de doutorado, pois aborda a hanseníase, uma doença que continua a ser um problema de saúde pública, especialmente em áreas com baixo nível socioeconômico, como a região amazônica. A pesquisa mencionada no artigo descreve a aplicação de um modelo baseado em ciência de dados e aprendizado de máquina para melhorar a especificação do perfil clínico e epidemiológico de possíveis casos de hanseníase. Esta abordagem complementa os objetivos da tese ao fornecer dados adicionais sobre a eficácia de métodos preditivos na identificação precoce da doença, fundamental para reduzir a carga da hanseníase e prevenir complicações neurológicas irreversíveis.

Além disso, o artigo destaca a importância do diagnóstico precoce e da aplicação de tecnologias avançadas para detecção de casos, o que está diretamente alinhado com o foco da tese em utilizar aprendizado de máquina para prevenir e tratar a hanseníase. Os resultados preliminares positivos do modelo de regressão proposto indicam um potencial significativo para melhorar as estratégias de diagnóstico e tratamento, reforçando a validade da abordagem proposta na tese. Assim, o artigo serve como um importante suporte teórico e prático, evidenciando a viabilidade e a necessidade de inovações tecnológicas na luta contra a hanseníase.

Outro ponto importante é a implementação de um dashboard clínico, que é uma ferramenta crucial para acompanhar o perfil de pacientes afetados, pois permite uma visualização em tempo real dos dados clínicos, facilitando a identificação rápida de padrões e tendências nos casos de hanseníase. O dashboard pode melhorar a tomada de decisão clínica, aumentar a eficiência do monitoramento dos pacientes e ajudar a prevenir complicações ao fornecer alertas e atualizações contínuas sobre o estado dos pacientes, o que é essencial para o sucesso das estratégias propostas na tese.

Este trabalho adotou uma abordagem direcionada para o tratamento dos registros de 911 pacientes em acompanhamento clínico para hanseníase, coletados entre 2016 e 2020. A análise revelou que a maioria desses pacientes estava inicialmente registrada como não tendo recebido nenhum dos tratamentos disponíveis, como a PQT (Poliquimioterapia) ou esquemas alternativos. Contudo, essa interpretação foi revisada, considerando que esses pacientes estavam sendo atendidos em unidades básicas de saúde, que não foram incluídas na análise original.

Essa exclusão se deve ao fato de que, apesar de os registros serem provenientes da rede pública, eles estão armazenados em bases de dados que não são de acesso público, justificando a ausência dessa informação no estudo anterior. Com o refinamento deste trabalho, foi possível definir uma alternativa de baixo custo operacional e alto impacto, oferecendo uma ferramenta

de apoio eficaz para os médicos no manejo de casos de hanseníase.

4.5 Considerações Finais

Em conclusão, o estudo apresentado neste capítulo demonstrou a eficácia aprimorada de um modelo MCDM reformulado para a priorização de pacientes com hanseníase, com foco na etapa de administração de medicamentos. Com a substituição do método TOPSIS pelo ELECTRE II, o ranking gerado, juntamente com os dados de acompanhamento clínico exibidos, mostrou resultados superiores, validando a viabilidade desta abordagem para problemas de saúde pública, com maior precisão e baixo custo computacional. Além disso, o desenvolvimento de um dashboard interativo foi um resultado notável, oferecendo uma ferramenta prática para a visualização e interpretação dos dados.

As modificações implementadas no modelo otimizaram significativamente o processo de tomada de decisão, destacando pacientes com diagnóstico positivo, mais de cinco lesões, e baciloscopia positiva como prioritários, enquanto aqueles com quadros menos graves foram classificados de acordo com diretrizes clínicas menos urgentes. Este trabalho, assim, contribuiu de forma decisiva para a construção da tese de doutorado, oferecendo uma metodologia robusta e eficiente no manejo de pacientes.

5 ARTIGO 3: ESTUDO DE VARIÁVEIS EPIDEMIOLÓGICAS

5.1 Considerações Iniciais

No Capítulo 5 desta tese, que é intitulada de "Ciência de Dados e Aprendizado de Máquina Aplicados ao Estudo de Variáveis Epidemiológicas da Hanseníase na Amazônia", será apresentado o modelo desenvolvido para identificar casos potenciais de hanseníase a partir de dados clínicos de pacientes. O capítulo abrange a análise de dados coletados entre 2015 e 2020 na região amazônica, com o objetivo de traçar características relevantes para o diagnóstico precoce. São exploradas técnicas de aprendizado de máquina para aprimorar o diagnóstico e tratamento da hanseníase, com foco em áreas hiperendêmicas, como os estados do Maranhão e Pará. Este capítulo também se insere no contexto do projeto "Pesquisa Operacional e Treinamento em Serviço para Áreas Hiperendêmicas de Hanseníase no Maranhão e no Pará", destacando a contribuição para a implementação de estratégias operacionais mais eficazes no controle da doença.

5.2 Trabalhos Correlatos

Para a concepção desta tese, foi realizada uma revisão sistemática dedicada ao levantamento de trabalhos científicos que abordam o uso de técnicas de aprendizado de máquina no contexto da hanseníase, com foco no diagnóstico precoce, tratamento e acompanhamento clínico. O objetivo central da revisão é identificar e analisar os métodos, abordagens e resultados obtidos em estudos que integram inteligência artificial na luta contra a hanseníase, uma doença negligenciada que apresenta desafios significativos em termos de diagnóstico. A revisão foi conduzida seguindo o protocolo PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), amplamente reconhecido como um método rigoroso para a realização de revisões sistemáticas (ASAR et al., 2016).

Inicialmente, foram levantados 476 trabalhos científicos, dos quais 85 foram selecionados após uma triagem detalhada baseada em critérios de inclusão e exclusão. A seção de trabalhos correlatos foi constituída por 15 estudos relevantes que analisam a correlação entre aprendizado de máquina e hanseníase, permitindo uma compreensão aprofundada da eficácia das abordagens propostas e da relevância dos resultados obtidos. Essa análise contribui para o avanço do conhecimento nesta área e para o desenvolvimento de estratégias mais eficazes no combate à doença, especialmente considerando que a hanseníase enfrenta desafios significativos devido ao acesso limitado a exames clínicos e ao acompanhamento médico de pacientes.

A identificação precoce da doença, em sua fase inicial (antes do aparecimento de sinais e sintomas nos indivíduos), é crucial para um tratamento eficaz, o que ressalta a importância das abordagens de aprendizado de máquina no aprimoramento do diagnóstico e no manejo clínico da hanseníase. Nesse contexto, a identificação precoce de casos da doença é vital para frear sua

progressão. Nos pacientes acometidos, as ações relacionadas à identificação estão diretamente ligadas às características clínicas obtidas durante o processo de acompanhamento individual (PALMEIRA, 2020). No entanto, a sociedade ainda enfrenta grandes limitações de recursos clínicos e médicos, que, mesmo nas grandes metrópoles, incluem altos custos, longos tempos de espera e desafios logísticos para os pacientes, o que dificulta o diagnóstico precoce (JIN; CRUZ; GONÇALVES, 2020).

A hanseníase na Índia deixou de ser tratada como um problema de saúde pública em 2005, embora seja responsável por 60% do quantitativo de casos no mundo. Nesse aspecto, autores realizaram um acompanhamento clínico em pacientes por um período de 5 anos, considerando dados clínicos e epidemiológicos (VINNARASAN et al., 2018). Esses registros foram analisados e categorizados por faixa etária, étnica, classificação e forma clínica. O trabalho atuou diretamente no estabelecimento do perfil de pacientes, fato este que mostra a necessidade da criação de medidas que atuam na notificação precoce de casos. O estudo possui carência de uma metodologia bem estabelecida para melhorar o manejo dos pacientes, além de ser limitado a uma determinada comunidade.

Em (WU et al., 2018), os autores analisaram as características da distribuição temporal e espacial de novos casos de hanseníase na província de Yunnan durante 2011 a 2016, a fim de se obter uma base sólida de conhecimentos para criação de estratégias de prevenção e controle da doença. O estudo utilizou um método de análise exploratória de dados e QGIS (Quantum GIS - Sistema de Informações Geográficas)[1] para fazer o mapeamento de casos e com isso identificar as características mais sensíveis de pacientes que estão concentrados em uma determinada região. Além de identificarem a região de maior concentração de novos casos, os autores limitaram-se a identificar que as características da maioria dos indivíduos está relacionada com a prevalência, economia e pela aglomeração de pessoas.

Já em (CHEN et al., 2021), uma revisão sistemática e metanálise foi realizada a fim de se identificarem sistematicamente os fatores clínicos associados à incapacidade física em pacientes com hanseníase. Foram utilizados dados de palavras-chave “lepra” e “incapacidade física” e termos relacionados retirados de bases de dados como Scopus, PubMed e Web of Science. Os autores utilizaram Odds Ratio (OR), ou razão de probabilidades, como medida de associação entre as características clínicas e o grau de incapacidade física para estabelecer as principais características do público avaliado. O estudo concluiu que a maior parte dos desfechos de identificação estão associados ao sexo masculino, hanseníase multibacilar, reações hansênicas e apresentação *Virchowiana* da doença.

Em (HOOIJ et al., 2021), a hanseníase é tratada como uma doença determinada por fatores do hospedeiro, em que o contágio é tido como inabalável, especialmente em indivíduos com contato próximo e constante. No estudo, os autores sustentam que a vacinação com BCG pode reduzir o risco de hanseníase, indicando que ela pode alterar o equilíbrio da imunidade

[1] Disponível em: https://qgis.org/pt$_BR/site/$

protetora dos pacientes. Os indivíduos vacinados e que possuem hanseníase paucibacilar (caracterizada por forte ação pró-inflamatória) foram comparados com indivíduos da mesma área sem contato com pacientes de hanseníase. No geral, os pacientes clinicamente afetados e que foram vacinados com a BCG (Bacilo de Calmette e Guérin) se diferenciam dos demais entre imunidade protetora e propensa à doença nesses contatos.

Outra abordagem é mostrada por (NEVES et al., 2021), onde é realizada uma análise exploratória da densidade Kernel da taxa de detecção de novos casos no Brasil, considerando 574.181 novos casos que foram registrados de 2003 a 2017. O estudo leva em consideração os fatores associados ao erro de diagnóstico identificados por regressão logística ao nível de significância de 5%, em que a probabilidade de erro de diagnóstico foi elevada para mulheres, crianças, classificação paucibacilar e de forma clínica indeterminada. No geral, o estudo conclui que o diagnóstico errôneo da hanseníase não está correlacionado com o nível de endemicidade no Brasil, mas sim com características dos indivíduos.

Em (SANTANA et al., 2018) é feita uma discussão acerca das abordagens para melhorar a sensibilidade dos testes para detectar a hanseníase antes do início dos sintomas, dentre elas, a criação de biomarcadores que indicam a presença do bacilo nos indivíduos. O estudo realizou testes com uma sequência de espaçadora que permite a identificação de anticorpos dos pacientes de hanseníase paucibacilares. Os dados obtidos indicam que a macromolécula sintética pode ser empregada no desenvolvimento de um imunossensor baseado em uma microbalança de cristal quartzo (Quartz Crystal Microbalance - QCM), que é uma alternativa de baixo custo para a realização de diagnóstico de hanseníase.

Segundo (BERNARDES-FILHO et al., 2021), a hanseníase, quando levada ao diagnóstico equivocado, pode ser considerada uma doença dissimulada, causando danos graves à saúde, bem como comorbidades neurológicas severas. São considerados neste estudo, pacientes com sintomas neurológicos crônicos, pacientes com lesões cutâneas e nervos espessados. No geral, os autores chamam a atenção para os encaminhamentos equivocados de casos clínicos atípicos, ressaltando a necessidade de melhorar o processo de triagem e de ensino sobre a hanseníase para todos os profissionais de saúde, especialmente aos profissionais mais novos.

Para (KHAN et al., 2017) a caracterização da estrutura das proteínas da membrana do *Mycobacterium* desempenha um papel vital na descoberta de soluções para o tratamento da bactéria causadora da hanseníase. Um modelo computacional foi desenvolvido neste trabalho para caracterizar essa estrutura não caracterizada de micobactéria, testando diversas composições de aminoácidos, peptídeos, tri-peptídeos entre outras. O estudo atua no campo da bioinformática, onde um modelo computacional, com base em algoritmos de aprendizado, possibilitou a produção de uma ferramenta poderosa para a identificação de micobactérias em indivíduos e, dessa forma, auxiliar na produção de drogas antimicobactérias.

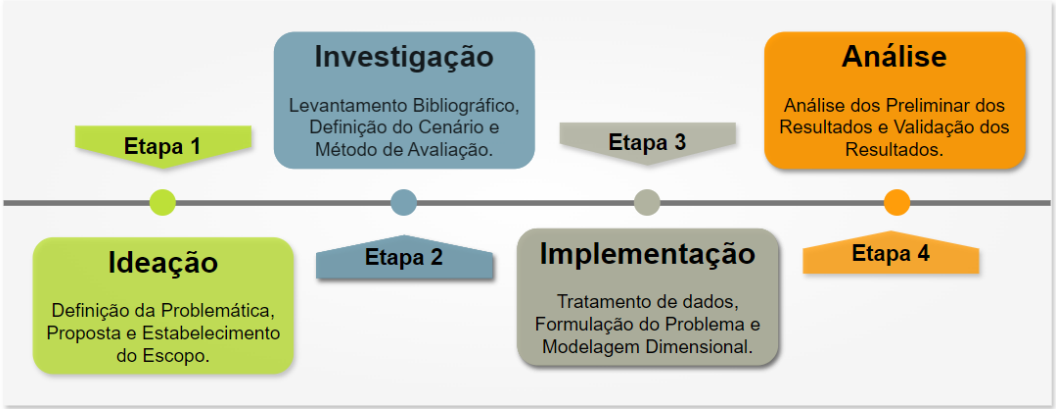
Embora as abordagens mencionadas tenham apresentado avanços promissores no combate à hanseníase, ainda há lacunas significativas na literatura científica, especialmente no que

tange à integração de análises clínicas contínuas e personalizadas com técnicas de aprendizado de máquina para o diagnóstico precoce e a gestão da doença. A maioria dos estudos foca em aspectos isolados, como o diagnóstico inicial ou o desenvolvimento de biomarcadores, mas raramente há uma integração dessas ferramentas com dados clínicos longitudinais que poderiam potencializar o monitoramento e o tratamento individualizado dos pacientes. A proposta deste trabalho visa justamente preencher essa lacuna, ao utilizar métodos de aprendizado de máquina não apenas para diagnosticar, mas também para acompanhar a progressão clínica da hanseníase ao longo do tempo, oferecendo uma abordagem mais holística e eficaz para o controle e erradicação da doença.

5.3 Método de Pesquisa

Esta tese foi desenvolvida a partir de um processo bem estabelecido, que consiste na execução de várias etapas de investigação científica. Iniciou-se com a coleta de dados de indivíduos, seguida pelo tratamento e modelagem desses dados, culminando na extração de conhecimento relevante. O escopo foi cuidadosamente definido a partir de uma ampla revisão bibliográfica sobre hanseníase, que é uma doença negligenciada de grande importância para o contexto amazônico. A Figura 5 ilustra as etapas do método de pesquisa, destacando a relevância do estudo, fundamentado na identificação de grupos com características clínicas e epidemiológicas semelhantes, devido à ampla gama de trabalhos correlatos.

Figura 5 – Modelo de Pesquisa e Desenvolvimento.



Autoria Própria.

A Figura 5 ilustra o método de pesquisa, dividido em quatro etapas essenciais. A primeira etapa, IDEAÇÃO, envolve a definição do problema, a proposta inicial e o escopo do trabalho, com participação significativa de profissionais da saúde, focada no tratamento e controle da hanseníase na Região Norte do Brasil. Na segunda etapa, INVESTIGAÇÃO, foi realizado um levantamento bibliográfico para definir o cenário de desenvolvimento, abordando tratamento, diagnóstico e prevenção da hanseníase, além de estabelecer os fundamentos do método de avaliação.

Os trabalhos correlatos presentes na segunda etapa (Figura 5) destacam a identificação de casos de hanseníase, mas geralmente não avaliam o impacto de diversas variáveis diagnósticas. Na sequência, a terceira etapa, IMPLEMENTAÇÃO, trata do tratamento e modelagem de dados usando um modelo dimensional e técnicas de aprendizado de máquina. A última etapa, ANÁLISE DE RESULTADOS, consiste na avaliação preliminar dos resultados do método de identificação de grupos clinicamente afetados pela hanseníase, validando a proposta com base nesses resultados.

O método de pesquisa e avaliação apresentado abrange todas as fases do desenvolvimento desta tese de doutorado. Identificar grupos com sintomas semelhantes à hanseníase é uma tarefa complexa e sensível, devido à grande variabilidade das informações relacionadas à doença. Este método é crucial porque realiza um levantamento detalhado dos elementos relacionados à proposta e se baseia em trabalhos consolidados na literatura para embasamento teórico, o que ajuda a fundamentar a proposta e preencher lacunas existentes.

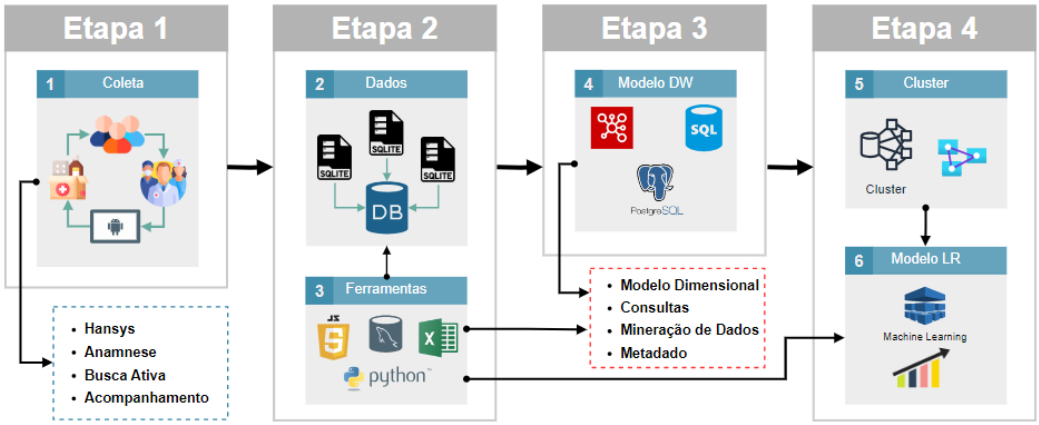
5.4 Modelo de Dados

O modelo de dados vem ao encontro da hipótese levantada na seção 1.3, onde objetiva-se implementar uma infraestrutura de dados robusta para simplificar o acesso a dados, sendo estes, oriundos de uma ferramenta não *open source* de armazenamento de dados de pacientes em acompanhamento clínico denominada Hansys. Para tanto, uma solução completa de Data Warehouse (DW) deve ser concebida. O DW pode ser dividido em três estruturas fundamentais: os sistemas onde os dados de origem podem ser obtidos, os sistemas de ETL e os sistemas de apresentação de dados.

A demanda crescente por dados e a necessidade de atualizações frequentes motivaram a criação deste modelo, inicialmente baseado em uma estrutura relacional (em que os dados são armazenados em uma ou mais tabelas ou de colunas e linhas). No entanto, a estrutura existente apresentava diversos problemas, impossibilitando o acesso eficiente aos sistemas originais de gestão de dados. Implementar um DW eficiente é crucial, pois melhora significativamente a qualidade dos dados e a eficácia na análise, resultando em melhores resultados clínicos e de pesquisa.

Na prática, a disponibilização de conjuntos de dados extraídos de sistemas ocorre, geralmente, na forma de arquivos de texto em formato tabular. Um requisito essencial do sistema proposto é desenvolver uma ferramenta para automatizar a obtenção desses arquivos. Como a construção do DW depende dos dados de pacientes, outro requisito é disponibilizar os dados extraídos para permitir a execução do processo de ETL, mesmo quando o sistema de coleta estiver indisponível. Esta redistribuição de dados deve ser realizada apenas para conjuntos com licenças de uso permissivas. O modelo de solução proposto pelo DW inclui quatro etapas de desenvolvimento, começando com a aplicação de técnicas de modelagem de dados e aprendizado de máquina, conforme visto na Figura 6.

Figura 6 – Modelo de Arquitetura de Dados.



Autoria Própria.

A versão inicial do modelo de solução incluirá etapas (Figura 6) essenciais para a aplicação durante o acompanhamento clínico. O conjunto de dados evoluiu com a inclusão de novos registros de pacientes, alimentando constantemente o modelo dimensional. O principal objetivo é permitir a execução de modelos de aprendizado de máquina no futuro, desenvolvendo-o com os requisitos mínimos necessários.

Na ETAPA 01 (Figura 6), os dados são coletados por meio de uma plataforma Android desenvolvida para armazenar todas as informações da anamnese do paciente durante seu atendimento, incluindo dados pessoais, laboratoriais, clínicos e neurológicos, conforme os requisitos da OMS para o acompanhamento de hanseníase. Esses dados são exportados em formato SQLite. Na ETAPA 02, inicia-se o tratamento dos dados com uma Análise Exploratória de Dados (AED), que examina os dados antes da aplicação de técnicas avançadas. Essa etapa envolve a extração, limpeza, e correlação dos dados, removendo inconsistências e valores ausentes utilizando ferramentas como Python[2] , JavaScript[3] e Postgres[4] .

Na ETAPA 03 (Figura 6), o processo de ETL é implementado, convertendo o banco relacional de 66 tabelas de registros para um formato dimensional, onde as tabelas dimensões crescem verticalmente e a tabela de fato horizontalmente. Neste formato, expresso pela Figura 6, as tabelas dimensões crescem, normalmente, verticalmente, enquanto a tabela de fato cresce horizontalmente. Por fim, na ETAPA 04 (Figura 6), diversas técnicas e modelos de machine learning são testados para encontrar a melhor solução para a problemática.

Algoritmos de Agrupamento, como K-Modes e ROCK Clustering, são aplicados para identificar grupos de pacientes com características semelhantes, visando melhorar o tratamento e acompanhamento clínico de hanseníase. Uma das alternativas viáveis encontradas foi a utilização de um método supervisionado de aprendizado de máquina, para compor o modelo de

[2] Disponível em: https://www.python.org/
[3] Disponível em: https://www.javascript.com/
[4] Disponível em: https://www.postgresql.org/

cálculo de probabilidade (PROCESSO 6: Random Forest). O modelo visa calcular a probabilidade de cada grupo de indivíduos estarem doentes.

5.4.1 Conjunto de Dados

Inicialmente, o conjunto de dados original apresentava uma grande quantidade de atributos, dados de diferentes categorias e problemas relacionados a *missing values*, inconsistência e duplicatas. Durante o pré-processamento, esses dados foram discretizados para otimizar a execução do modelo proposto, o que deve ser melhor explicado na seção 4.6.2, sobre o modelo de dados. Especificamente, o conjunto de dados, contendo 510 atributos de diferentes tipos, foi analisado e avaliado com base no coeficiente de correlação de *Pearson* para realizar uma extração preliminar das variáveis de maior relevância para o problema em questão.

Esse coeficiente mede o grau da correlação e a direção da correlação entre duas variáveis, assumindo apenas valores entre -1 e 1. O método de correlação é calculado através da seguinte equação Eq. 1:

$$cor(E_i, Ej) = \frac{1}{M}\sum_{\imath=1}^{M}(\frac{x_i\imath - \bar{x_i}}{\sigma_i})(\frac{x_j\imath - \bar{x_i}}{\sigma_j}) \tag{1}$$

Tal que, x_i e x_j representam as médias dos valores dos atributos dos exemplos E_i e E_j, respectivamente, e σ e σ seus desvios-padrão. A similaridade entre os exemplos, calculada por meio dos índices de correlação, tem o mesmo valor resultante do cálculo de correlação, como exemplo, $sim(E_i, E_j)$ = $cor(E_i, E_j)$. Em contrapartida, a distância entre os exemplos $E_i e E_j$, calculada por meio de um coeficiente de correlação, é dada pelo seu complemento: $dist(E_i, E_j) = 1$ - $cor(E_i, E_j)$.

Os dados coletados não seguem um padrão de organização, tornando essencial o pré-processamento de dados na etapa de Ciência de Dados. Esse pré-processamento inclui atividades como Análise Exploratória de Dados, limpeza e transformação dos dados, conforme descrito na seção 4.3. Durante essa etapa, os dados foram discretizados para otimizar a execução dos algoritmos implementados na sequência, conforme visto na Tabela 3.

A Tabela 3 apresenta as variáveis e valores do Dataset, composto por 10 atributos utilizados no experimento. Esses atributos representam os dados de entrada do modelo, além do rótulo de saída, exemplificado pelo atributo "Status do Paciente". Esse rótulo está dividido entre paciente com diagnóstico positivo (1) e paciente sem diagnóstico (2). Para registros com valores ausentes, foi atribuído o valor 0, indicando informação nula.

Inicialmente, o dataset completo foi avaliado, considerando todas as suas características principais na manifestação clínica da hanseníase. No pré-processamento de dados, foi realizada uma análise do coeficiente de correlação para selecionar os atributos mais relevantes. Essa etapa permitiu criar um subset otimizado de dados, focando nos atributos com maior inferência sobre

Tabela 3 – Lista de Features de Entrada.

Feature	**Descrição**
Classificação	PB (Paucibacilar); MB (Multibacilar)
Forma Clínica	BV (Borderline Virchowiana); BB (Borderlin Borderline); V (Virchowiana); I (Indeterminada); BT (Borderline Tuberculoide); T (Tuberculoide); Neural Pura)
Tratamento	PQT/ MB 12 doses; PQT/MB 24 doses; PQT/PB 6 doses; Esquema Alternativo
Baciloscopia	Não realizado; Positivo; Negativo
Tipo do Paciente	Geral, Caso Novo e Recidiva
Grau de Incapacidade	0; 1; 2
Número de Lesões	0; 1; 2 a 5; 6 a 10; mais de 10
PCR	Positivo, Negativo, Não Realizado
PGL	Positivo, Negativo, Não Realizado
Contato Positivo	Sim; Não
Grau de Escolaridade	Nenhum, Ensino Médio; Ensino Fundamental, Ensino Superior
Gênero	Masculino, Feminino
Marca da BCG	0; 1; 2; Duvidosa
Idade	De 0 a 20 anos; de 21 a 40 anos; De 41 a 60 anos; De 61 a 80; De 81 a 100 anos
Estado Civil	Viúvo(a), Solteiro(a), União estável, Casado(a), Separado(a), Não Informado
Renda Familiar	Sem renda, Até dois salários mínimos, Menor que um salário mínimo, Até três salários mínimos, Um salário mínimo, Maior que três salários mínimos
Convênio Governo	Sim, Não, Não Informado

o diagnóstico dos pacientes, especificando sobre as características clínicas mais significativas para o diagnóstico.

Dando continuidade, o conjunto de dados mencionado acima está intimamente ligado à criação do aplicativo Hansys (Sistema de Coleta de Registro de Pacientes), visto em (DUTRA DA SILVA et al., 2018). A ferramenta faz parte do escopo do projeto intitulado "Pesquisa operacional e treinamento em serviço para áreas hiperendêmicas de hanseníase no Maranhão e no Pará" sob Coordenação do Prof. Dr. Claudio Guedes Salgado, do Prof. Dr. Guilherme Augusto Barros Conde e demais profissionais da área. O Hansys foi concebido como uma ferramenta tecnológica para facilitar a coleta e gestão de informações, permitindo a coleta de dados em tempo real de pacientes em acompanhamento clínico (NOGUEIRA et al., 2014).

Destaca-seque o processo de coleta está respaldado pelo Termo de Consentimento Livre e Esclarecido (TCLE), que é um instrumento fundamental fornecido aos participantes do projeto durante as etapas de coleta, análise e registro de informações. Este termo assegura que todas as informações coletadas serão utilizadas exclusivamente para fins de pesquisa, com garantia de sigilo e respeito à privacidade dos envolvidos. Os participantes são esclarecidos sobre

os procedimentos adotados, como questionários, coletas de sangue e amostras de pele, além do mapeamento de casos em suas residências. É informado que a participação é totalmente voluntária, podendo o consentimento ser retirado a qualquer momento sem qualquer penalidade ou prejuízo. O cumprimento dos protocolos de segurança e ética na condução das atividades é rigorosamente garantido, preservando a dignidade e os direitos dos voluntários.

Adicionalmente, em conformidade com a Lei Geral de Proteção de Dados identificada por Lei nº 13.709/2018, o projeto assegura que todas as informações pessoais e sensíveis coletadas estarão protegidas e serão tratadas de forma ética, transparente e em conformidade com as diretrizes legais. O uso dos dados será limitado aos objetivos descritos no termo, promovendo a confidencialidade e a segurança das informações. Ressalta-se que o projeto, assim como seus desdobramentos científicos, possui a aprovação do Comitê de Ética em Pesquisa, sob a documentação Carta: 06/08 CEP-ICS/UFPA, emitida em 21 de fevereiro de 2008, atestando a conformidade ética do estudo e reforçando seu compromisso com os padrões regulatórios e científicos exigidos.

Além disso, o Hansys integra os dados coletados pelos profissionais de saúde durante visitas domiciliares e triagens comunitárias, permitindo um acompanhamento mais detalhado dos casos e assegurando que as informações sejam rapidamente compartilhadas e analisadas. Esses dados também servem como base para diversas pesquisas, teses, dissertações e estudos, gerando um impacto social significativo. Dessa forma, eles são essenciais para a produção de conhecimento científico e contribuem diretamente para o avanço das estratégias de controle e tratamento da hanseníase, completando o ciclo de análise iniciado com o subset otimizado dos dados clínicos.

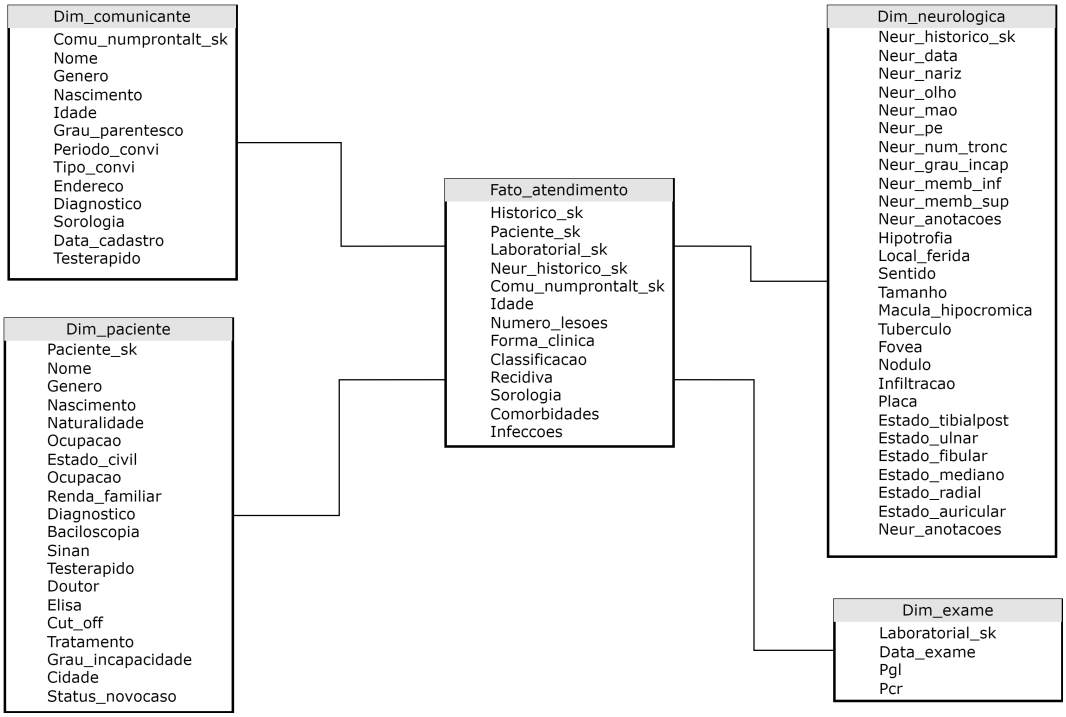
5.4.2 Transformação de Dados

A criação de um modelo de dados utilizando o processo de ETL é essencial para consolidar informações e otimizar as consultas e operações de dados em um data warehouse. Neste contexto, foi desenvolvido um data warehouse contendo quatro tabelas dimensões (exame, neurológica, paciente e comunicante) e uma tabela fato (atendimento), conforme visto na Figura 7 . Para a consolidação desse modelo, foram utilizadas diversas ferramentas, sendo uma delas o Pentaho Data Integration (PDI), uma ferramenta poderosa de integração de dados que facilita a construção e automação dos processos de ETL.

A Tabela 7, expressa uma "ação" realizada no modelo de dados utilizados. Nela, há todas as chaves *Surrogate key (SK)* que identificam índices nas demais tabelas dimensões, tais como Dim_exame, Dim_paciente, Dim_comunicante e Dim_neurologica. Essas chaves substitutas são identificadores exclusivos para cada linha e podem ser usadas como chaves primárias, proporcionando um nível adicional de abstração e eficiência na modelagem dos dados.

Conforme visto na Figura 7, a Tabela Dim_paciente contém todas as informações pertinentes aos indivíduos em acompanhamento clínico. Já a Tabela Dim_comunicante é constituída

Figura 7 – Relação entre a Tabela Fato_atendimento e suas Dimensões.



Autoria Própria.

de informações dos indivíduos que têm contato direto com os pacientes. A Tabela Dim_neurologica contém dados relacionados à avaliação neurológica que é efetuada nos indivíduos em acompanhamento, enquanto a Tabela Dim_exames possui os resultados de exames clínicos do tipo PCR (Reação em Cadeia Polimerase) e PGL (glicolipídeo-fenólico).

Por fim, a Tabela Fato_atendimento se relaciona com as tabelas dimensões e assim, facilitando e otimizando as consultas em diferentes níveis de complexidade. Cada entrada na tabela Fato_atendimento refere-se a um evento de atendimento, linkado às dimensões relevantes através de suas chaves substitutas. Esse modelo permite uma integração eficiente e robusta de dados, facilitando consultas complexas e análises detalhadas, otimizando a experiência de extração e manipulação de dados, essencial para a tomada de decisões informadas.

5.4.3 Segmentação de Dados

Técnicas de aprendizado não supervisionado podem ser valiosas para identificar grupos de pacientes com características semelhantes, auxiliando no diagnóstico da hanseníase. Essas técnicas permitem explorar relacionamentos entre atributos, gerando informações médicas úteis para avaliar o risco de contrair a doença em cada grupo.

Diversos algoritmos de Agrupamento foram testados, incluindo Agglomerative Clustering, BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), K-means, Gaussian Mixture Model, K-Modes Clustering e ROCKClustering. A eficácia dos algoritmos foi avaliada utilizando o Silhouette Score, que mede a coesão e separação dos clusters, fornecendo

uma interpretação clara da qualidade dos clusters formados. Entre os melhores algoritmos para o problema em questão estão o K-Modes Clustering e ROCKClustering, conforme demonstrado pela Eq. 2, a qual especifica o cálculo da diferença entre a distância média de um ponto aos outros pontos do mesmo cluster;

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{2}$$

Tal que, $s(i)$ é o Silhouette Score do ponto de dado i, $a(i)$ é a média da distância intra-cluster para o ponto de dado i e $b(i)$ é a menor média da distância entre o ponto de dado i e todos os pontos em qualquer outro cluster, do qual i não faz parte.

O K-Modes foi escolhido como o método de agrupamento mais eficaz dentre todos os testados, devido ao seu desempenho superior conforme avaliado pelo índice de Silhouette Score. Este índice é crucial na avaliação de modelos de agrupamento, pois captura o equilíbrio entre a compacidade dos clusters e a distância entre eles. O algoritmo se destacou por produzir clusters mais coesos e separados, o que é fundamental para a precisão e interpretabilidade dos resultados. Além disso, sua capacidade de processar grandes conjuntos de dados o torna uma escolha viável para análises em escala. No entanto, o K-Modes superou todos os algoritmos testados, sendo destacado na seção 5.2, fornecendo uma base sólida para a escolha deste algoritmo de agrupamento (SCULLEY, 2010).

O número ideal de clusters para o conjunto de dados foi determinado usando o método Elbow. Esse método executa o K-Modes com diferentes quantidades de clusters e identifica o ponto onde a adição de mais clusters não resulta em ganhos significativos na redução da variação intra-cluster. Ao aumentar o número de clusters, a diferença entre eles diminui, mas a variação dentro de cada cluster aumenta, buscando um equilíbrio entre homogeneidade dentro dos clusters e distinção entre eles. Isso ajuda a encontrar clusters representativos e distintos para a análise dos dados (THORNDIKE, 1953).

$$SSE = \sum_{i=1}^{K} \sum_{x \in c_i} dist(x, c_i)^2 \tag{3}$$

O método expresso na Eq. 3 calcula o SSE (*Sum of Squared Error*) para alguns valores de K (exemplo 2 4 5 etc). O SSE é a soma da distância ao quadrado entre cada membro do cluster e seu centroide. O K-Modes calcula a similaridade das observações com o modo do cluster ao qual a observação pertence; então, idealmente, esta similaridade deve ser a maior possível. Durante o experimento, o algoritmo busca a quantidade de agrupamentos em que a dissimilaridade intra-clusters seja a menor possível. Em termos matemáticos, o K-Modes minimiza a soma das dissimilaridades das características categóricas dentro dos clusters, sendo zero o resultado ótimo. Este processo é especialmente útil para segmentar dados com atributos categóricos.

O K-Modes, uma extensão do K-means para dados categóricos, também utiliza uma função de custo. Essa função de custo no K-Modes é responsável por medir a dissimilaridade entre os clusters formados e os dados de entrada. Ela busca minimizar a diferença entre os atributos categóricos dos pontos de dados e os centroides dos clusters. O objetivo é encontrar grupos que possuam dados categóricos semelhantes, minimizando assim a função de custo e aumentando a coesão intra-cluster. Dessa forma, a função de custo no K-Modes desempenha um papel crucial na avaliação da qualidade da agrupamento, garantindo que os clusters sejam significativos e representativos dos dados de entrada (CHATURVEDI; GREEN; CAROLL, 2001b). A função de custo do algoritmo K-Modes é dada por (Eq. 4):

$$C(Q) = \sum_{j=1}^{k} \sum_{i \in Q_j} d(x_i, m_j) \tag{4}$$

Onde:

- k é o número de clusters.
- Q_j é o conjunto de pontos atribuídos ao cluster j.
- $d(x_i, m_j)$ é a dissimilaridade entre o ponto x_i e o modo m_j do cluster.j.

Para dados categóricos, a dissimilaridade $d(x_i, m_j)$ entre dois vetores x_i e m_j é definida como o número de atributos em que eles diferem, conforme visto em (Eq. 5):

$$d(x_i, m_j) = \sum_{l=1}^{p} \delta(x_{il}, m_{jl}) \tag{5}$$

Onde:

- p é o número de atributos,
- x_{il} é o valor do atributo l do ponto x_i,
- m_{jl} é o valor do atributo l do modo m_j,
- δ é uma função que retorna 1 se $x_{il} \neq m_{jl}$ e 0 se $x_{il} = m_{jl}$.

A função δ é definida como (Eq. 6):

$$\delta(a, b) = \begin{cases} 1 & \text{se } a \neq b \\ 0 & \text{se } a = b \end{cases} \tag{6}$$

Logo, A função de custo $C(Q)$ do algoritmo K-Modes é a soma das dissimilaridades entre os dados x_i e os modos m_j dos clusters. A dissimilaridade $d(x_i, m_j)$ para dados categóricos é calculada como o número de atributos em que dois vetores x_i e m_j diferem. A função $\delta(a, b)$ retorna 1 se os valores dos atributos a e b são diferentes e 0 se são iguais.

5.5 Cálculo da Distribuição de Probabilidade

Após a segmentação dos pacientes em grupos com características semelhantes, que envolve a análise e classificação dos pacientes com base em dados demográficos, clínicos e comportamentais, foram gerados modelos baseados em regressão logística para determinar a distribuição de probabilidade de contrair hanseníase que cada grupo possui, levando em consideração somente as pessoas sem diagnóstico em cada grupo. Essa segmentação é realizada para melhor compreender as diferentes necessidades de saúde, bem como as características clínicas dos indivíduos.

Ao final, será gerado um modelo para cada grupo, considerando que, cada grupo foi filtrado para se obterem subgrupos somente com as pessoas que não tinham o diagnóstico. Esse método permite uma abordagem personalizada, considerando as características específicas de cada grupo de pacientes e identificando padrões relacionados à probabilidade de contrair hanseníase. Desse modo, formaram-se sete subgrupos após a filtragem:

- O subgrupo 1 contém pessoas que não possuem diagnóstico dentro do grupo 1;
- O subgrupo 2 contém pessoas que não possuem diagnóstico dentro do grupo 2;
- O subgrupo 3 contém pessoas que não possuem diagnóstico dentro do grupo 3;

Os subgrupos acima que são compostos somente por pessoas sem diagnóstico serão utilizados como conjunto de teste nos modelos de regressão logística. O funcionamento ocorrerá da seguinte maneira:

- Para o teste 1, o modelo 1 será treinado e validado com todo o conjunto de dados, exceto as amostras do subgrupo 1;
- Para o teste 2, o modelo 2 será treinado e validado com todo o conjunto de dados, exceto as amostras do subgrupo 2;
- Para o teste 3, o modelo 3 será treinado e validado com todo o conjunto de dados, exceto as amostras do subgrupo 3;

A avaliação dos modelos gerados é essencial para verificar a eficácia do modelo proposto. Isso é feito por meio da análise de métricas de desempenho, como precisão, recall, F1-score e área sob a curva ROC (Receiver Operating Characteristic), entre outras, dependendo do

contexto e dos objetivos do modelo (QAMAR; GAUSSIER, 2010). Além disso, para entender a importância relativa dos atributos do conjunto de dados nos modelos de regressão logística implementados, utilizou-se o algoritmo Relief (KIRA; RENDELL, 1992). Esse algoritmo é uma técnica de seleção de atributos que atribui pesos aos atributos com base na sua relevância para a previsão do modelo. A pontuação atribuída pelo algoritmo varia de -1 a 1, onde valores mais próximos de 1 indicam maior importância do atributo para a previsão do modelo, conforme visto na Eq. 7 .

$$W[A] = W[A] - \frac{diff(A, x, nearestHit) + diff(A, x, nearestMiss)}{m} \tag{7}$$

Onde, $W[A]$ representa o peso da característica $A, diff(A, x, y)$ mede a diferença entre os valores da característica A nas instâncias x e y, $nearestHit$ é a instância mais próxima da mesma classe e $nearestMiss$ é a instância mais próxima de uma classe diferente. A importância do Relief está em sua habilidade de identificar características relevantes mesmo em conjuntos de dados com dependências complexas, contribuindo para a criação de modelos de classificação mais precisos.

O teste Qui-quadrado (X^2) (FISHER, 1970) é outra técnica essencial para a seleção de características, especialmente para variáveis categóricas. Ele avalia a independência entre duas variáveis categóricas ao comparar as frequências observadas e esperadas em uma tabela de contingência. A formulação do teste é dada por Eq. 8:

$$\chi^2 = \sum \frac{(O - E)^2}{E} \tag{8}$$

Onde O são as frequências observadas e E são as frequências esperadas. Os valores chi2, pval e dof têm significados específicos:

- chi2: Valor do qui-quadrado, que indica a magnitude da diferença entre as frequências observadas e esperadas;
- pval: Valor-p, que mostra a probabilidade de observar os dados assumindo que a hipótese nula de independência seja verdadeira. Um valor-p baixo sugere rejeitar a hipótese nula;
- dof: Graus de liberdade, ajustando o valor do qui-quadrado para o número de categorias nas variáveis;
- Feature: A variável está sendo avaliada.

A utilização do teste do Qui-quadrado é crucial para identificar relações significativas entre variáveis categóricas, ajudando a melhorar a performance do modelo de classificação

ao focar nas características mais relevantes. A aplicação combinada dos métodos Relief e Qui-quadrado permite uma seleção robusta de variáveis, considerando tanto características contínuas quanto categóricas. Esta abordagem integrada é essencial para construir um modelo preditivo eficaz no diagnóstico de hanseníase, garantindo uma análise abrangente das variáveis envolvidas.

5.6 Avaliação de Performance

A avaliação de Performance foi realizada por meio de diversas métricas, como acurácia, precisão, recall, F1-score, área sob a curva ROC (Receiver Operating Characteristic) entre outras mais. Essas métricas são essenciais para medir a eficácia do modelo em dados desconhecidos, garantindo sua capacidade de generalização, robustez e confiabilidade. A análise desses indicadores é crucial para determinar a utilidade prática do modelo, assim como a preparação adequada dos dados. Esta avaliação é tão importante quanto a preparação dos dados, visto que ela determina a utilidade prática do modelo desenvolvido. A escolha adequada das métricas de avaliação depende dos objetivos específicos do estudo e das características dos dados, conforme discutido em (QAMAR; GAUSSIER, 2010) e (SUN, 2007).

- Acurácia: Esta métrica deve ser usada em datasets com a mesma proporção de exemplos para cada classe e quando as penalidades de acerto e erro para cada classe forem as mesmas. Essa métrica representa a taxa de acertos em relação ao total de amostras e pode ser obtida através da Eq. 9.

$$Acurácia = \frac{(VP + VN)}{(VP + FN + VN + FP)} \tag{9}$$

Onde, os Falsos Positivos (FP) são as instâncias que foram incorretamente classificadas como positivas, mas que, na verdade, pertenciam à classe negativa. Já os Verdadeiros Positivos (VP) correspondem às instâncias que foram corretamente identificadas como positivas, de acordo com a classe positiva real. Por outro lado, os Verdadeiros Negativos (VN) são aquelas instâncias corretamente classificadas como negativas, alinhando-se à classe negativa real. Por fim, os Falsos Negativos (FN) representam as instâncias que, embora pertencentes à classe positiva, foram erroneamente classificadas como negativas.

- Precisão: Mostra o número de exemplos classificados como pertencentes a uma classe, que realmente são daquela classe (VP), dividido pela soma entre este número e o número de exemplos classificados nesta classe, mas que pertencem a outras (FP). A Eq. 10 seguinte mostra como calcular a precisão.

$$Precisão = \frac{VP}{VP + FP} \tag{10}$$

- Sensibilidade ou Recall: Essa métrica (também conhecida como recall ou revocação) avalia a capacidade do método de detectar com sucesso resultados classificados como positivos e pode ser obtida com a Eq. 11.

$$Recall = \frac{VP}{VP + FN} \tag{11}$$

- F1-Score: O F1-Score é uma média harmônica entre precisão (que, apesar de ter o mesmo nome, não é a mesma citada acima) e Recall. Ela é recomendada quando se possui um dataset com classes desproporcionais e pode ser obtida através da Eq. 12.

$$F = 2 * \frac{Precisão * Recall}{Precisão + Recall} \tag{12}$$

- Curva ROC e AUC: Este indicador possibilita uma avaliação mais completa quanto à qualidade das predições do classificador, levando em consideração a taxa de verdadeiros positivos e a taxa de falsos positivos A curva ROC (Receiver Operating Characteristics) ilustra o limiar da capacidade de discriminação de um classificador binário, variando o critério de aceitação de um classificador. O gráfico da curva representa a variação da Taxa de Verdadeiros Positivos (TVP), dada por Eq. 13, em relação à Taxa de Falsos Positivos (TFP), dada por Eq.14;

$$TVP = \frac{VP}{VP + FN} \tag{13}$$

$$TFP = \frac{FP}{FP + VN} \tag{14}$$

Para resumir a qualidade mensurada pela curva ROC (Receiver Operating Characteristic), é comum a utilização da métrica AUC (Area Under the Curve). A curva ROC é uma ferramenta poderosa para avaliar a performance de modelos de classificação, especialmente em problemas binários. Ela representa a relação entre a Taxa de Verdadeiros Positivos (TPR) e a Taxa de Falsos Positivos (FPR) em diferentes limiares de decisão. A métrica AUC resume essa curva em um único valor numérico que varia de 0 a 1, onde 1 indica um classificador perfeito e 0.5 representa um classificador aleatório. Assim, a AUC permite comparar a eficácia dos classificadores utilizando um único escalar, facilitando a identificação do modelo mais robusto e confiável para o conjunto de dados analisado.

5.7 Considerações Finais

Neste capítulo, foram apresentados os desdobramentos da tese que estão sendo abordados no artigo 3, intitulado "Estudo Sobre Perfil Clínico da Hanseníase na Região Amazônica Utilizando Técnicas de Ciência de Dados e Aprendizado de Máquina". No estudo, detalhou-se a metodologia de pesquisa utilizada, descrevendo todas as etapas do modelo de solução, desde a coleta de informações até a aplicação do modelo de regressão. O modelo proposto gerou resultados viáveis para a identificação de grupos com diferentes probabilidades estar doente, proporcionando uma ferramenta valiosa para a comunidade acadêmica e científica. A expectativa é que o trabalho contribua para a detecção precoce e o manejo eficaz da doença, apoiando a formulação de estratégias de intervenção mais direcionadas e eficientes.

6 RESULTADOS

6.1 Considerações Iniciais

Neste capítulo, serão apresentados os resultados obtidos ao longo da presente tese de doutorado. Inicialmente, será descrito o modelo de Aprendizado de Máquina proposto, destacando os resultados das etapas de coleta de dados, pré-processamento, seleção de características e treinamento do modelo. Na sequência, será abordado o processo de identificação de grupos clinicamente afetados pela hanseníase, utilizando uma abordagem de agrupamento. Esse método aplica algoritmos de agrupamento para segmentar os pacientes em subgrupos homogêneos com base em padrões fenotípicos e dados clínicos. Por fim, serão expostos os resultados acompanhados de uma avaliação de performance, incluindo métricas de acurácia, precisão, revocação e a análise da validade dos clusters formados, evidenciando a eficácia do modelo na identificação de padrões clínicos relevantes.

6.2 Modelo de Aprendizado de Máquina

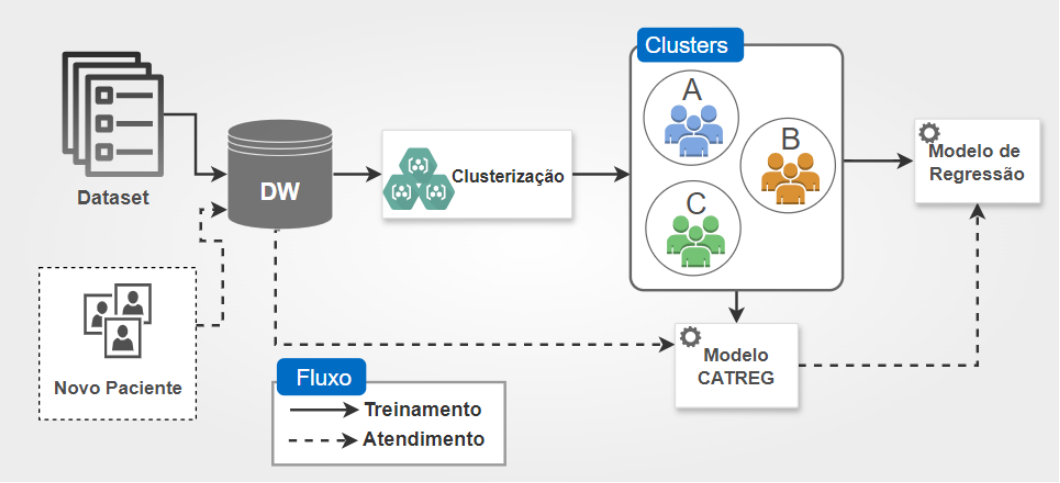
Nesta tese, foram aplicados modelos de Aprendizado de Máquina para a detecção de hanseníase, com destaque para o Random Forest, que apresentou os melhores resultados. A partir do conjunto de dados fornecidos, o método implementado foi capaz de produzir um modelo que previu os casos com alta precisão. Utilizando uma combinação de atributos como entrada em uma função sigmoide, os resultados produziram saídas entre 0 e 1. As previsões foram classificadas de acordo com um limiar médio, sendo atribuída a classe 1 para valores superiores a 0,5 e a classe 0 para valores iguais ou inferiores. Assim, as entradas que resultaram em valores acima de 0,5 foram consistentemente classificadas como pertencentes à classe 1, demonstrando a eficácia do modelo na classificação binária.

Em contrapartida, se a saída for menor que 0,5, a entrada correspondente será classificada como pertencente à classe 0. A partir do Random Forest é calculada a probabilidade de que a variável de saída pertença à categoria apropriada, permitindo a classificação precisa das entradas. O modelo proposto demonstrou sua aplicabilidade também para indivíduos inseridos após a configuração inicial do modelo, como os pacientes recentemente atendidos. Esses pacientes são integrados ao modelo DW através de uma transformação de dados previamente desenvolvida. Essa abordagem garante a atualização contínua e a precisão na classificação dos novos registros.

O modelo de regressão foi desenvolvido para lidar com os registros de pacientes recentemente atendidos. O modelo proposto, denominado CATREG (Regressão Categórica), quantifica os dados categóricos atribuindo valores numéricos às categorias. A eficácia desse modelo foi avaliada com base na sua capacidade de previsão precisa e na sua aplicabilidade em registros

atualizados, como ilustrado na Figura 8.

Figura 8 – Fluxo de Execução do Modelo de Dados.



Autoria Própria.

A Figura 8 ilustra o fluxo de execução do modelo de Random Forest, destacando dois processos principais: o fluxo de treinamento (linha contínua) e o fluxo de atendimento (linha tracejada). No fluxo de treinamento, os registros dos pacientes são inicialmente processados e inseridos no modelo de Data Warehouse (DW). Em seguida, esses dados são submetidos a um algoritmo de agrupamento que identifica grupos de pacientes com características clínicas semelhantes. Cada cluster representa um grupo afetado e a probabilidade de um paciente estar doente. Durante o fluxo de atendimento, os novos registros de pacientes são tratados e inseridos no modelo CATREG, que segue a mesma lógica do modelo de Random Forest.

6.3 Especificação de Grupos Clinicamente Afetados

Para identificar grupos de pessoas com características semelhantes e prever a distribuição de probabilidade associada a cada grupo, foram implementados modelos de agrupamento, sendo que o K-Modes e o ROCK Clustering apresentaram os melhores resultados. Vale destacar que o conjunto de dados foi previamente tratado utilizando técnicas de Ciência de Dados, como tratamento de dados, remoção de outliers e transformação de variáveis categóricas e numéricas. Além disso, os dados foram normalizados para o intervalo de 0 a 1, garantindo uma melhor performance dos modelos de agrupamento.

Para avaliar e eleger o melhor algoritmo, foi utilizado o índice de Silhouette Score, que mede a coesão interna e a separação entre clusters. O método calcula a média da diferença entre a distância média dos pontos dentro do mesmo cluster e a distância média dos pontos ao cluster mais próximo. O algoritmo K-Modes apresentou os melhores resultados, conforme visto na Tabela 4.

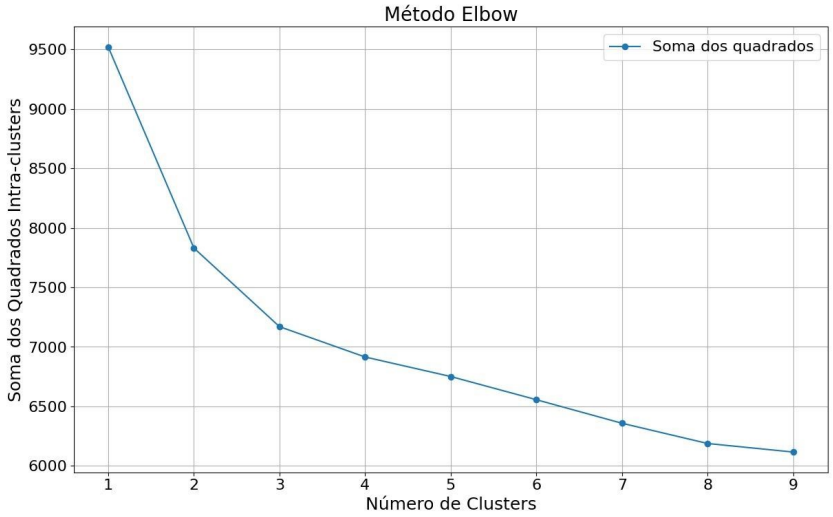
Tabela 4 – Comparação do Silhouette Score para os algoritmos.

Algoritmo	**Índice**
K-Modes	0.185
ROCK Clustering	0,108

O índice Silhouette Score, mostrado na (Tabela 4), calcula a similaridade média de cada cluster com um cluster mais semelhante a ele. Quanto maior o valor do Silhouette Score, melhor os clusters são separados e mais coesos são internamente, indicando um resultado de clustering mais eficaz. Inicialmente, o método Elbow (NAINGGOLAN et al., 2019) foi empregado para calcular a quantidade ideal de clusters. O método consiste em traçar a variação explicada em função do número de clusters e escolher o “cotovelo” da curva como o número de clusters a utilizar.

Como esse método necessita de um número inicial de clusters, pode-se executar o algoritmo para várias quantidades diferentes de clusters e definir qual dessas quantidades é o número ideal, por exemplo. O ideal é que a distância das observações até o centro do agrupamento a que ela pertence tenda a zero, ou seja, a soma dos quadrados intra-clusters, conforme visto na Figura 9.

Figura 9 – Método do Cotovelo para Definir a Quantidade Ideal de Clusters.

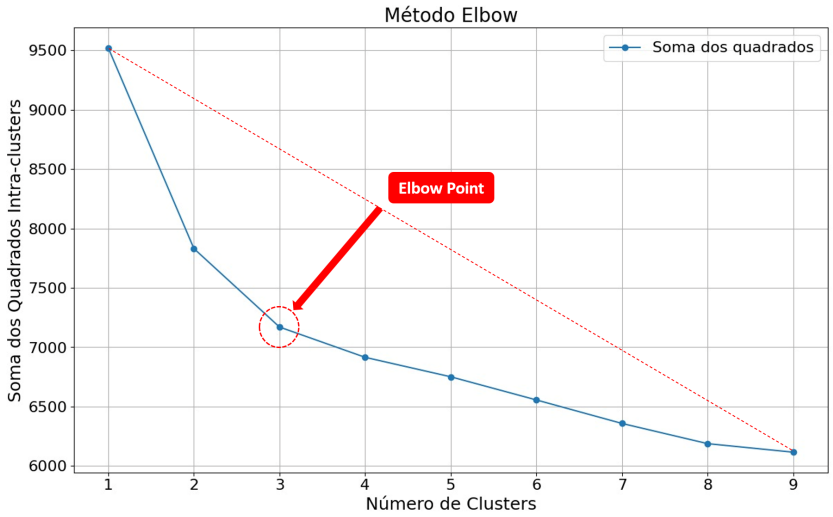


Autoria Própria.

A Figura 9 demonstra a soma dos quadrados intra-clusters para diferentes quantidades de clusters. Para determinar o número ideal de clusters, traça-se uma linha reta entre os pontos referentes ao número mínimo e máximo de clusters. O ponto da curva mais distante dessa linha, conhecido como "cotovelo", indica o equilíbrio entre a homogeneidade dentro dos clusters e a diferenciação entre eles. Na Figura 10, o "cotovelo” ocorre no ponto correspondente a sete clusters, indicando que essa é a quantidade ideal para o conjunto de dados.

Por meio do gráfico expresso pela Figura 10, observa-se que o ponto de inflexão dessa

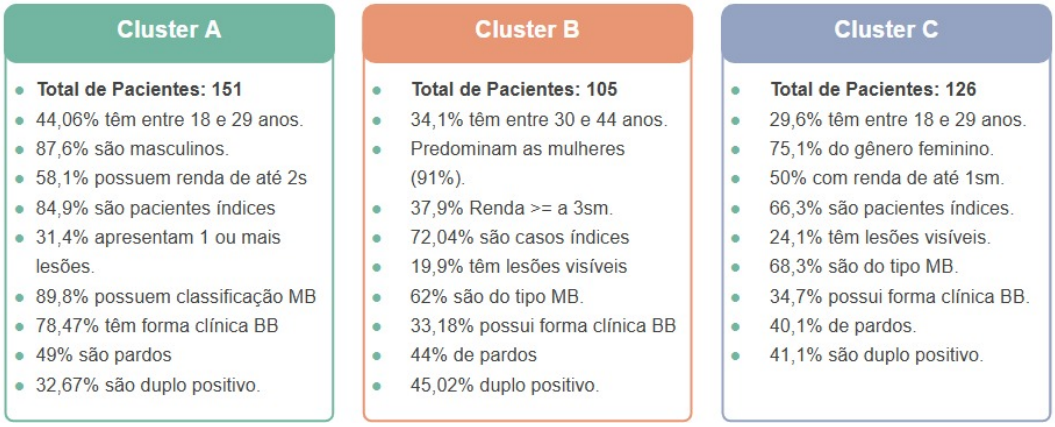
Figura 10 – Método do Cotovelo para Definir a Quantidade Ideal de Clusters.



Autoria Própria.

curva está no valor de $k = 3$. Portanto, pode-se inferir que o número ideal de clusters para esse conjunto de dados é 3. A seguir, será apresentado uma visualização dos cluster obtidos no estudo, contendo as informações mais relevantes de cada agrupamento (Figura 11).

Figura 11 – Disposição de clusters.



Autoria Própria.

A Figura 11 apresenta uma análise de agrupamento (clusters) de indivíduos baseada em características similares relacionadas a aspectos demográficos, clínicos e sociais, conforme especificado no estudo. Esse agrupamento sintetiza os principais resultados do estudo, fornecendo uma visão clara das características predominantes em cada grupo. A seguir, será apresentada a descrição detalhada de cada cluster e suas respectivas características, demonstrando os principais insights obtidos a partir do modelo de agrupamento

- **Cluster A:** Composto por 151 pacientes, 44,06% deles têm entre 18 e 29 anos, com uma prevalência masculina significativa (87,6%). A maioria dos pacientes têm renda de até

2 salários mínimos (58,1%) e 84,9% são do tipo novo caso. Aproximadamente 31,4% apresentam uma ou mais lesões visíveis. Em termos de classificação clínica, 89,8% possuem hanseníase do tipo MB, com 78,47% apresentando a forma clínica BB, e 49% se identificam como pardos. Além disso, 32,67% dos pacientes são duplo positivo para PCR e PGL, reforçando a gravidade dos casos.

- **Cluster B:** Com 105 pacientes, 34,1% deles têm entre 30 e 44 anos, sendo a grande maioria mulheres (91%). Este Cluster é caracterizado por uma maior renda, com 59,77% dos pacientes ganhando três ou mais salários mínimos. Aproximadamente 72,04% são casos novos, enquanto 19,9% apresentam lesões visíveis. A hanseníase do tipo MB está presente em 62% dos casos, e a forma clínica BB aparece em 33,18%. Quanto à cor, 44% dos pacientes se identificam como pardos, e 45,02% são duplo positivo para PCR e PGL.

- **Cluster C:** Similar ao Cluster B, este grupo também conta com 125 pacientes, 34,1% dos quais têm entre 30 e 44 anos, com predominância feminina (91%) e renda menor (50% com até 1 salário mínimo). Contudo, a gravidade clínica é comparável, com 19,9% apresentando lesões visíveis e 62% diagnosticados com hanseníase do tipo MB. A forma clínica BB é encontrada em 33,18% dos pacientes, dos quais 44% se identificam como pardos e 41,16% apresentam positividade dupla para PCR e PGL, sinalizando uma progressão avançada da doença.

A diferença observada entre os Clusters em termos de sexo (masculino no Cluster A e feminino nos Clusters B e C) pode refletir fatores sociais e econômicos que influenciam o diagnóstico e tratamento da hanseníase. Homens e mulheres podem ter diferentes níveis de acesso aos cuidados de saúde, o que pode explicar a predominância masculina em um grupo de idade mais jovem e com renda mais baixa no Cluster A. Já a maior renda nos Clusters B e C, compostos majoritariamente por mulheres, sugere que mulheres com maior poder aquisitivo têm mais acesso a diagnóstico e tratamento. A diferença de idade também pode estar ligada a essas desigualdades, com homens mais jovens sendo diagnosticados mais cedo e mulheres, em sua maioria, em idades mais avançadas.

Em relação à diferença entre os percentuais de positividade dupla para PCR e PGL (32,67%, 41,16% e 45,02%), a variação entre os Clusters pode não ser suficientemente grande para ser considerada estatisticamente significativa sem uma análise formal, mas pode indicar uma diferença gradual de gravidade clínica entre os grupos. Os dois grupos com predominância pacientes do sexo feminino, B e C, são muito semelhantes em termos de idade, renda e gravidade da doença. O que pode tê-los separado são detalhes nas características clínicas ou padrões de atendimento que precisam de uma análise mais detalhada.

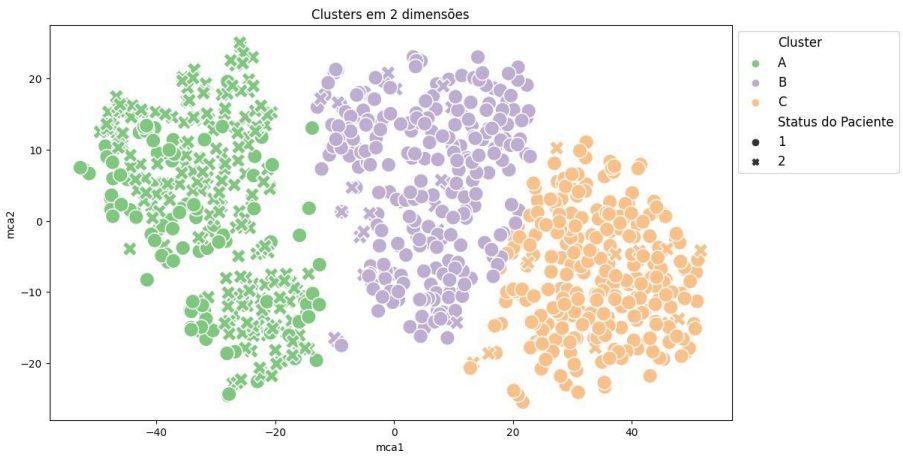
Os resultados encontrados estão diretamente ligados à gravidade e à progressão da hanseníase porque o maior número de casos duplo positivos (PCR e PGL) nos Clusters B e C (41,16% e 45,02%) em comparação ao Cluster A (32,67%) indica uma infecção mais avan-

çada nesses grupos. A positividade em ambos os testes sugere uma resposta imunológica mais significativa e a presença de DNA do bacilo, o que é associado a uma progressão da doença, embora isso não esteja diretamente relacionado à quantidade de bacilos, que é medida pela baciloscopia. Além disso, a predominância de casos de hanseníase multibacilar (MB) em todos os clusters, com variação na gravidade das lesões visíveis, sugere que os Clusters B e C, que têm maior renda, podem ter recebido um diagnóstico mais precoce e acesso a tratamentos mais adequados, o que influencia o curso da doença de maneira mais controlada nesses grupos.

Além disso, foi realizada uma análise mais aprofundada dos indivíduos, considerando suas condições clínicas, com base nos resultados dos exames de PCR e PGL. Para facilitar a interpretação e visualização dos grupos gerados, foi aplicada o MCA (Multiple Correspondence Analysis). Essa técnica de redução de dimensionalidade permitiu condensar as várias variáveis em um gráfico bidimensional, proporcionando uma visão clara dos clusters formados. Assim, os padrões clínicos e sociodemográficos que diferenciam os grupos podem ser visualizados de maneira mais objetiva, auxiliando na compreensão das características dos pacientes em cada Cluster e reforçando a necessidade de estratégias de tratamento personalizadas com base nas particularidades de cada grupo.

O MCA é uma técnica estatística que simplifica um conjunto de dados categóricos através de uma transformação em um espaço contínuo. Essa técnica transforma as categorias em um novo sistema de coordenadas, onde cada eixo corresponde a uma combinação das categorias originais. Os componentes resultantes são ordenados de forma que a maior parte da associação entre as categorias seja capturada pelo primeiro componente, a segunda maior associação pelo segundo componente, e assim por diante. Isso facilita a visualização e análise dos dados categóricos, reduzindo a dimensionalidade enquanto preserva a maior parte das relações originais entre as categorias (ABDI; VALENTIN, 2007). No caso analisado, o conjunto de dados foi reduzido para duas dimensões, permitindo a visualização clara dos clusters formados, conforme ilustrado na Figura 12.

Figura 12 – Visualização dos Clusters em duas dimensões utilizando MCA.



Autoria Própria.

Conforme observado na Figura 12, os Clusters A (Figura 12, verde) e C (Figura 12, laranja), apresentam maior densidade de pacientes e são majoritariamente compostos por indivíduos diagnosticados com hanseníase, indicando que esses pacientes possuem características clínicas associadas à doença. No entanto, o Cluster A (Figura 12, verde) também inclui uma maior concentração de indivíduos sem o diagnóstico, o que é menos relevante para este estudo. O objetivo central da tese é calcular a probabilidade de indivíduos desenvolverem a doença, dadas as suas características clinicas, reforçando o foco na identificação precoce e prevenção.

O Cluster A (Figura 12, verde) é predominantemente composto por indivíduos do sexo masculino e jovens (18 e 29 anos), com alta prevalência de classificação Multibacilar (MB) e forma clínica Borderline-Borderline (BB). Além disso, há uma significativa percentagem de casos duplo positivos, ou seja, positivos tanto para o PCR quanto para o PGL. O Cluster B (Figura 12, roxo), composto majoritariamente por mulheres de maior renda, também apresenta uma alta proporção de casos duplo positivo, porém com uma menor prevalência de lesões visíveis. Já o Cluster C (Figura 12, laranja), com um perfil mais feminino e de menor renda, apresenta uma gravidade comparável à do Cluster B. Embora haja uma menor incidência de pacientes com lesões visíveis, o Cluster C (Figura 12, laranja) ainda possui uma quantidade significativa de casos duplo positivo.

O número ideal para a segmentação dos indivíduos foi estabelecido para as demais análises. A Tabela 5 apresenta a distribuição quantitativa de indivíduos em cada cluster, levando em consideração duas classes: Classe A, que representa os indivíduos sem diagnóstico para hanseníase, e Classe B, que inclui os indivíduos já diagnosticados com hanseníase.

Tabela 5 – Amostra por Cluster.

Cluster	**Amostras**	**Classe A**	**Classe B**
A	404	224	180
B	211	109	102
C	294	164	130

Sinais de alerta sobre os insights obtidos: O Cluster A destaca-se pela alta prevalência de indivíduos do sexo masculino, jovens, e uma predominância de formas clínicas graves, como as classificadas como MB, incluindo a forma clínica BB. A classificação MB abrange a multiplicação bacilar excessiva e disseminação da infecção para outros órgãos e tecidos((ALVES; MORAES et al., 2019), sendo a BB uma forma intermediária que exige atenção por sua gravidade. O Cluster B, embora formado majoritariamente por mulheres de maior renda, também apresenta uma alta proporção de indivíduos duplo positivo para PCR e PGL, embora com menor incidência de lesões visíveis. Já o Cluster C revela uma vulnerabilidade socioeconômica significativa, composto majoritariamente por mulheres de menor renda, com uma gravidade comparável ao Cluster B, mesmo com uma menor incidência de lesões visíveis.

A agrupamento revelou perfis distintos entre os três grupos, fornecendo insights valiosos sobre as características predominantes de cada um. Essas informações são essenciais para

o desenvolvimento de estratégias direcionadas ao combate à hanseníase, permitindo personalizar as abordagens de acordo com as especificidades de cada perfil de paciente. Com base nos resultados da segmentação, é possível propor estratégias que freiem a disseminação da doença, como a ampliação de consultas nas unidades de saúde, campanhas educativas específicas para os diferentes grupos, o uso de telemedicina para facilitar o acompanhamento, e a implementação de tecnologias de monitoramento contínuo da condição.Ao combinar essas abordagens, é possível atender de forma mais eficaz às necessidades de cada grupo, garantindo uma maior adesão ao tratamento.

O foco principal desta análise está nos fatores clínicos e no tipo do paciente, que pode ser categorizado como novo caso, caso geral ou recidiva. A Tabela 6 abaixo apresenta uma matriz de probabilidades que mostra a relação entre os resultados dos exames e a presença da hanseníase.

Tabela 6 – Análise de exames em relação à clínica (Geral).

PCR \ **PGL**	0	1	**TOTAL**
0	80,5%	19,5%	114
1	15,8%	84,2%	462

Nesta análise, foram considerados 576 indivíduos, todos com alguma forma clínica confirmada de hanseníase, conforme visto na Tabela 6. Entre os pacientes com resultados negativos para ambos os exames (PCR = 0 e PGL = 0), 84,2% não apresentam a doença, enquanto 15,8% são positivos apenas para PGL. Nos casos com PCR positivo (PCR = 1), 80,4% também testam positivo para PGL, revelando uma forte correlação entre a positividade para ambos os exames e a manifestação da doença. Já entre os indivíduos com PCR positivo e PGL negativo, 19,5% testam positivo apenas para PCR. Esses achados reforçam a importância do duplo positivo como indicador de maior gravidade da hanseníase, exigindo acompanhamento clínico mais rigoroso.

Uma segunda análise foi realizada, considerando 40 indivíduos que ainda não possuem uma forma clínica definida de hanseníase, mas apresentaram resultados para os exames de PCR e PGL-I. A Tabela 7 a seguir mostra a distribuição desses pacientes com base nos resultados dos exames.

Tabela 7 – Análise de exames em relação à clínica (Sem Clínica).

PCR \ **PGL**	0	1	**TOTAL**
0	32,3%	67,7%	18
1	19,6%	80,4%	22

Entre os indivíduos com resultados negativos para ambos os exames (PCR = 0 e PGL = 0), 47,2% não apresentaram a doença, enquanto 52,4% testaram positivo apenas para PGL, conforme ilustrado na Tabela 7. Nos casos em que o PCR foi positivo (PCR = 1), 70% apresentaram PGL positivo. Para os pacientes com PCR positivo e PGL negativo, 30% testaram

positivo apenas para PCR. Esses resultados sugerem que, mesmo em casos sem forma clínica definida, a presença de duplo positivo para PCR e PGL pode ser um forte indicativo de um risco elevado para o desenvolvimento da hanseníase, exigindo um acompanhamento mais cauteloso.

É importante destacar que a avaliação conjunta do PCR e PGL é fundamental para o acompanhamento da hanseníase, pois revela uma forte associação entre a positividade para ambos os testes e o risco elevado de desenvolvimento da doença, mesmo sem sintomas clínicos evidentes. Esta abordagem permite identificar precocemente indivíduos com alto risco, possibilitando um acompanhamento mais rigoroso e intervenções precoces. Essa abordagem pode reduzir a propagação da doença, uma vez que pacientes em estágio inicial podem receber tratamento antes do desenvolvimento de sintomas mais graves e do contágio a outras pessoas. Portanto, integrar a avaliação de PCR e PGL-I no processo de triagem e acompanhamento clínico é uma estratégia essencial para aprimorar a detecção precoce.

6.4 Avaliação de Performance

Para a avaliação de performance, foram considerados dois modelos de aprendizado de máquina: Regressão Logística e Random Forest. Ambos foram selecionados entre outras técnicas avaliadas por apresentarem os melhores resultados. O modelo de regressão logística utiliza variáveis explicativas contínuas e binárias para prever se o paciente está doente, com base nos índices dos grupos gerados pelo processo de agrupamento, que servem como rótulos. O modelo de Random Forest, que combina múltiplas árvores de decisão para melhorar a precisão e a robustez, também foi treinado com os mesmos atributos para prever a associação das amostras aos grupos. A comparação entre esses dois modelos visa identificar a abordagem que oferece o melhor desempenho na previsão dos grupos.

O pré-processamento de dados, que inclui o uso dos modelos Relief e Qui-Quadrado, foi realizado para avaliar a importância das variáveis selecionadas em relação ao conjunto de dados de entrada. O modelo Relief é utilizado para identificar e atribuir peso às variáveis que mais influenciam na predição, considerando as diferenças entre amostras próximas e distantes. Já o teste Qui-Quadrado é empregado para avaliar a independência entre variáveis categóricas, verificando a significância estatística das relações entre elas. Este passo é crucial para entender as relações entre os dados e garantir que os modelos funcionem de maneira eficaz. Os resultados do Relief podem ser visualizados na Tabela 8.

Considerando os atributos utilizados (Tabela 8), foi utilizado um método Relief ainda no pré-processamento para seleção de características principais da fase de diagnóstico da hanseníase. O modelo atribui pesos ao conjunto de dados, montando um score de frequência (que varia de -1 a 1) das melhores e piores características que podem ser utilizadas no modelo. Nota-se que os atributos que atingiram a pontuação mais elevada indicam que estas características podem ser consideradas importantes para prever o possível diagnóstico positivo da doença.

Tabela 8 – Importância dos Atributos de Acordo com o Relief.

Feature	**Importância**
Tipo do Paciente	0.930674
Tratamento	0.738090
Classificação	0.359888
Forma Clínica	0.315281
PCR	0.036281
Baciloscopia	0.059438
Grau de Incapacidade	0.045169
Idade	0.035629
Número de Lesões	0.035281
PGL	0.020671
Estado Civil	0.027978
Convênio Governo	0.020337
Marca BCG	0.014270
Contato Positivo	0.013371
Gênero do Paciente	0.006966
Grau de Escolaridade	0.005618
Renda Familiar	0.000787

Outra técnica utilizada no estudo é o teste do Qui-quadrado, que é crucial porque permite determinar se há uma relação significativa entre variáveis categóricas, com base na diferença entre as frequências observadas e as esperadas. Os resultados do teste podem ser vistos na Tabela (9):

Tabela 9 – Importância dos Atributos de Acordo com o Teste Qui-quadrado.

Feature	**dof**	**pval**	**chi2**
Tipo do Paciente	2	5.482209	890.000000
Tratamento	1	2.612943	465.811939
Forma Clínica	7	2.642690	404.455308
Classificação	2	2.557219	394.166795
Grau de Incapacidade	3	3.306729	70.511696
Número de Lesões	5	5.362150	68.230936
PGL	1	4.459141	52.448197
Idade	5	3.163568	38.380862
Baciloscopia	5	5.548452	38.019494
PCR	3	3.564217	27.318430
Contato Positivo	1	1.141872	30.564112
Convenção Governo	1	1.194137	28.030576
Estado Civil	7	7.833058	21.077814
Grau de Escolaridade	3	1.083044	11.172163
Marca BCG	3	4.867299	7.874724

A Tabela 9 mostra os resultados obtidos com o teste do Qui-quadrado, indicando a relação entre variáveis categóricas. Com base nas informações, as Features Tipo do “Paciente” e "Tratamento” revelam associações estatisticamente significativas com as condições estuda-

das, indicando que esses fatores podem influenciar os resultados. Da mesma maneira, a "Forma Clínica" e a "Classificação" apresentam valores de Qui-quadrado que sugerem uma forte dependência com as variáveis de interesse, sendo cruciais para entender diferentes manifestações ou categorizações dentro do estudo. Além disso, as métricas de avaliação de dois modelos gerados para identificar a distribuição de probabilidade de hanseníase são apresentadas na Tabela 10

91,6493,2179,93

Tabela 10 – Avaliação de Performance dos Modelos de Previsão.

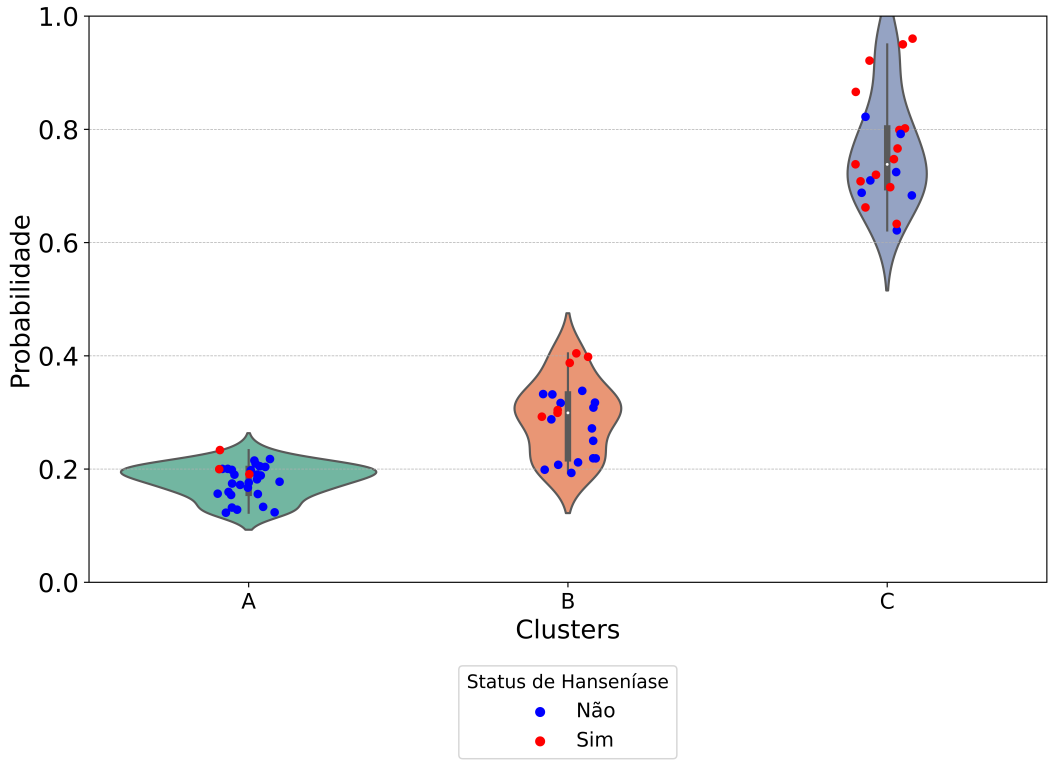
Modelo	**Precisão**	**Accuracy**	**ROC-AUC**	**F1-Score**	**Recall**
Regressão Logística	0,8790	0,7862	0,82565	0,75439	0,71857
Random Forest	0,9039	0,81654	0,90203	0,80955	0,77905

A Tabela 10 apresenta as métricas de desempenho dos dois modelos de aprendizado de máquina utilizados para prever a probabilidade de desenvolvimento de hanseníase entre os pacientes analisados no conjunto de teste. O modelo de Regressão Logística obteve uma Precisão de 87,90%, acurácia de 78,62%, com uma AUC-ROC de 82,57%, F1-Score de 75,44% e um Recall de 71,86%. Já o modelo Random Forest apresentou um desempenho superior em todas as métricas, com uma Precisão de 90,39%, acurácia de 81,65%, AUC-ROC de 90,20%, F1-Score de 80,96% e recall de 77,91%. Após a etapa de avaliação de performance do modelo de regressão, é efetuado o cálculo de probabilidade, obtendo o valor que cada grupo de indivíduos possui de desenvolver a doença, conforme visto na Figura 13.

Na Figura 13, é possível observar a distribuição de probabilidades de hanseníase nos três clusters (A, B e C), obtida a partir do modelo de Random Forest. No Cluster A, que é composto predominantemente por indivíduos do sexo masculino e jovens com alta prevalência de classificação MB e forma clínica BB, a maior parte dos indivíduos apresenta uma menor probabilidade de estar doente, como indicado pelos pontos azuis, que representam aqueles sem hanseníase. No Cluster B, onde há uma concentração de indivíduos com duplo positivo mais elevada em relação ao Cluster A, há uma maior distribuição de probabilidades, com indivíduos tanto com quanto sem hanseníase distribuídos ao longo do espectro, sugerindo maior variação nas características desse grupo.

Por fim, o Cluster C (Figura 13) apresenta uma maior concentração de indivíduos com hanseníase (pontos vermelhos) em probabilidades mais altas, enquanto os indivíduos sem hanseníase (pontos azuis) apresentam uma distribuição mais ampla de probabilidades. Essa diversidade indica uma maior incerteza quanto ao diagnóstico, refletindo a complexidade das características clinicas neste Cluster. A variação observada pode estar relacionada a fatores como histórico de exposição, características demográficas ou condições socioeconômicas, que influenciam a manifestação da hanseníase, tornando o Cluster C um foco importante para futuras investigações e intervenções.

Figura 13 – Probabilidade de Indivíduos de Cada Cluster Desenvolverem a Doença.



Autoria Própria.

6.5 Considerações Finais

Neste capítulo foram apresentados os resultados obtidos com esta tese de doutorado, destacando todos os pontos-chave desenvolvidos até a fase atual do estudo. Utilizou-se um modelo de dados e aplicou-se um modelo de aprendizado de máquina para identificar grupos clinicamente afetados pela hanseníase e calcular a probabilidade de indivíduos desenvolverem a doença. A metodologia permitiu classificar os indivíduos em grupos com maior predisposição à hanseníase, reforçando a importância do estudo na prevenção e controle da doença. O capítulo também aborda a avaliação de desempenho e os resultados obtidos com o modelo de Random Forest, destacando sua eficácia na predição do risco de adoecimento.

7 CONSIDERAÇÕES FINAIS

7.1 Considerações Finais

A hanseníase, uma das doenças infecciosas crônicas mais antigas, continua a causar sérios danos à saúde de forma silenciosa. Embora os casos globais e nacionais tenham diminuído durante o período pandêmico da COVID-19, observou-se um aumento relativo até o ano de 2023, especialmente em países como Brasil, Índia e Indonésia, desafiando as previsões de erradicação da doença. Nesse contexto, torna-se essencial o desenvolvimento de estratégias que envolvam o uso de técnicas inovadoras para detecção e intervenção. Esta tese demonstrou que o uso de algoritmos de Aprendizado de Máquina, como modelos de regressão logística, random forest e técnicas de agrupamento, pode ser uma ferramenta promissora para identificar padrões ocultos nos dados clínicos, proporcionando uma forma mais precisa de diagnosticar e monitorar os casos de hanseníase.

O diagnóstico da hanseníase, conforme as diretrizes da OMS, baseia-se majoritariamente na avaliação clínica, o que apresenta limitações significativas, especialmente em regiões onde a detecção tardia ainda é comum. Esta tese demonstrou que a integração de técnicas de Ciência de Dados e Aprendizado de Máquina pode melhorar a acurácia no reconhecimento de possíveis casos da doença, reduzindo a margem de erro e acelerando o início do tratamento, consequentemente, reduzindo a transmissão e as complicações associadas à hanseníase.

Em continuidade às estratégias voltadas à saúde pública, foi essencial aplicar técnicas de ciência de dados para garantir a qualidade dos dados, eliminando inconsistências, redundâncias e valores ausentes. Nesta tese, os dados foram organizados por meio da implementação de uma infraestrutura de DW, estruturada em um modelo dimensional, o que facilitou as consultas e a integração com o modelo de solução proposto. Esse processo foi crucial para atender aos requisitos do modelo de aprendizado, incluindo discretizações, codificações e organização eficiente dos dados. Com essa estrutura, tornou-se possível identificar grupos de indivíduos com padrões semelhantes de sintomas.

Após a identificação dos grupos, foram implementados dois modelos: regressão logística e random forest. Ambos os modelos consideram duas classes de indivíduos: aqueles que possuem a doença (diagnóstico positivo) e os que ainda estão sem diagnóstico. Comparando os resultados, o modelo random forest apresentou um desempenho superior em todas as métricas, com acurácia de 81,65%, AUC-ROC de 90,20%, F1-Score de 80,96% e recall de 77,91%. Esses resultados superaram os obtidos pelo modelo de regressão logística, que alcançou uma acurácia de 78,62%, com uma AUC-ROC de 82,57%, F1-Score de 75,44% e um Recall de 71,86%. A random forest, portanto, destacou-se como o modelo mais eficiente para prever o desenvolvimento da hanseníase. A quantidade de clusters foi determinada por meio do método Elbow,

e o agrupamento foi validado utilizando o Silhouette Score, que demonstrou a qualidade dos agrupamentos baseados nas características dos dados.

7.2 Contribuições

As contribuições parciais desta proposta são listadas e descritas a seguir:

- Participação em um projeto de pesquisa intitulado "Pesquisa operacional e treinamento em serviço para áreas hiperendêmicas de hanseníase no Maranhão e no Pará", onde são desenvolvidas atividades de pesquisa e acompanhamento de pacientes na Região Norte do Brasil. O projeto já atendeu mais de 6 mil pacientes e possui parcerias como a Fundação Vale, Universidade Federal do Pará, Governo Federal entre outros;
- Estudo sobre os principais sintomas da hanseníase, considerando dados clínicos, físicos, sociais e neurológicos de pacientes em acompanhamento médico sendo realizado desde 2016 até então;
- Realização de uma extensa pesquisa relacionada aos principais conceitos que permeiam a inteligência computacional aplicada à saúde, mostrando suas características e desafios, junto a uma árdua e extensa revisão bibliográfica dos trabalhos atuais e consolidados na literatura científica junto ao tema;
- Desenvolvimento de um modelo de estruturação de dados flexível e robusto para tratar conjuntos de dados com deficiências em seu processo de construção;
- A proposta voltada para identificação de grupos clinicamente afetados pela hanseníase que possui flexibilidade e completude ao abranger todas as etapas relacionadas ao diagnóstico de pacientes, tornando o modelo de solução abrangente e podendo ser ampliado ainda mais com a inserção de novos atributos clínicos e neurológicos de pacientes;
- O modelo apresentado realiza a identificação de grupos clinicamente afetados pela hanseníase e aplica um método probabilístico para especificar a probabilidade de desenvolvimento da hanseníase em pacientes, baseados em:

    1. Uso de algoritmos (Agglomerative clustering, BIRCH - Balanced Iterative Reducing and Clustering using Hierarchies, Clustering, k-means, mini-batch, gaussian mixture model, ROCK Clustering e K-Modes) de Clusterização para definição de grupos clinicamente afetados;
    2. Estudo realizado para definir os algoritmos mais utilizados e consolidados na literatura científica;
    3. Uso de um algoritmo para calcular a quantidade ideal de cluster, dadas as características do conjunto de dados;

4. Utilização de regressão logística para cálculo de probabilidade de indivíduos (sem diagnóstico) desenvolveram a hanseníase ao longo da vida.

- Avaliar o desenvolvimento de comorbidade em pacientes afetados pela hanseníase, analisando os sintomas mais críticos para o diagnóstico;
- Através dos resultados obtidos preliminarmente, foi possível observar quais sintomas são mais sensíveis para definição de possíveis casos da hanseníase, haja vista que o seu perfil de contágio possui uma grande variedade de sintomatologia.

A divulgação da proposta de tese apresentada foi realizada por meio da aprovação em dois Periódicos da área, sendo o primeiro (A1) e o segundo (A3), listados em:

- FALCAO, IGOR W. S.; SOUZA, D. S. ; SALGADO, C. G. ; CARDOSO, D. L. ; BARRETO, JOSAFÁ G. ; COSTA, FERNANDO AUGUSTO RIBEIRO ; COSTA, P. F. ; SILVA, M. B. ; SERUFFO, M. C. R. Use of Multi-criteria Methods to Support Decision-Making in Drug Management for Leprosy Patients. INTERNATIONAL JOURNAL OF MANAGEMENT AND DECISION MAKING, v. 1, p. 1, 2022.
- FALCÃO, IGOR W. S.; SOUZA, DANIEL S. ; CARDOSO, DIEGO L. ; COSTA, FERNANDO A. R. ; LEITE, KARLA T. F. ; DE M., HAROLD D. ; SALGADO, CLAUDIO G. ; DA SILVA, MOISÉS B. ; BARRETO, JOSAFÁ G. ; DA COSTA, PATRICIA F. ; SANTOS, ADRIANO M. DOS ; CONDE, GUILHERME A. B. ; SERUFFO, MARCOS C. DA R. . A study about management of drugs for leprosy patients under medical monitoring: A solution based on AHP-Electre decision-making methods. PLoS One, v. 18, p. e0276508, 2023.

Como contribuições adicionais, o trabalho intitulado "Model for predicting drug resistance based on the clinical profile of tuberculosis patients using machine learning techniques" foi publicado em 2024 no periódico *PeerJ Computer Science*, identifciando com o DOI `10.7717/peerj-cs.2246`. Esta produção está diretamente relacionada à temática desta tese de doutorado, que se baseia em um modelo de dados para prever insights relacionados a outra importante doença negligenciada, conforme visto em:

- Model for Predicting Drug Resistance Based on the Clinical Profile of Tuberculosis Patients Using Machine Learning Techniques. Aceito em: 22 de Maio de 2024;

Além disso, o conhecimento adquirido na elaboração deste trabalho foi utilizado também na participação em outros estudos que não estão relacionados ao tema, no entanto, foram avaliados por meio de conferência nacional e aprovação em periódicos, sendo eles:

1. FALCÃO, IGOR; VIEIRA, RAFAEL ; PEREIRA, PAULO ; SERUFFO, MARCOS ; CARDOSO, DIEGO . The Heuristic for Hardware Dimensioning Considering Tidal Effect. Journal of Communication and Information Systems (JCIS),, v. 35, p. 311-319, 2020;

2. SANTOS, A. E. C. ; FALCAO, I. W. S. ; CARDOSO, DIEGO . Simulated Annealing Aplicado a On/Off de RRhs em Redes 5G com Tráfego de Dados Adaptativo. In: 11ª Conferência Nacional em Comunicações, Redes e Segurança da Informação, 2021, Campina Grande - PB. Conferência Nacional em Comunicações, Redes e Segurança da Informação, 2021;

3. NASCIMENTO, E. A. ; FALCAO, I. W. S. ; CARDOSO, D. L. . Análise do Efeito de Maré em Redes Centralizadas: Provisionamento de Recursos em uma Grande Metrópole. In: Brazilian Technology Symposium, 2020, Campinas - SP. Brazilian Technology Symposium, 2020.

4. IMBIRIBA, MARIANE DE PAULA DA SILVA GONÇALVES ; DA PAIXÃO, ERMÍNIO AUGUSTO RAMOS ; SANTOS, ALBERT EINSTEIN COUTINHO DOS ; DE MATTOS TEIXEIRA, CARLOS ANDRÉ ; VIEIRA, RAFAEL FOGAROLLI ; SOUZA, DANIEL DA SILVA ; FALCÃO, IGOR WENNER SILVA ; CARDOSO, DIEGO LISBOA . FA-CRAN: A Firefly Algorithm for Dynamic BBU-RRH Mapping in Cloud/Centralized Radio Access Networks. IEEE Access, v. 12, p. 22821-22831, 2024.

5. SILVA, A. F. ; LEITE, K. T. F. ; COSTA, F. A. R. ; SERUFFO, M. C. R. ; MORAES, C. C. G. ; FALCAO, I. W. S. . Study of machine learning techniques for outcome assessment of leptospirosis patients. Scientific Reports, v. 14, p. 13929, 2024.

6. SANTOS, A. E. C. ; PAIXAO, E. A. R. ; FALCÃO, IGOR W. S. ; CARDOSO, D. L. . Maximização da Eficiência Energética em Redes RSMA 6G. In: XIII Conferência Nacional em Comunicações, Redes e Segurança da Informação, 2023, Belém. Conferência Nacional em Comunicações, Redes e Segurança da Informação, 2023.

7. SOUZA, D. S. ; FALCÃO, IGOR W. S. ; SANTOS, A. E. C. ; MELLO, H. D. ; CARDOSO, D. L. ; COSTA, F. A. R. ; PEREIRA, R. . Applying Machine Learning Modelsto Heterogeneous Handover Managementin Heterogeneous Networks. In: Conferência Nacional em Comunicações, Redes e Segurança da Informação, 2023, Belém. Conferência Nacional em Comunicações,, 2023.

8. ALVES, A. R. ; FALCAO, IGOR W. S. ; SANTOS, A. E. C. ; CARDOSO, D. L. . APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA PARA IDENTIFICAÇÃO DO POTENCIAL HIDROGENIÔNICO (pH) EM AMBIENTE MONITORADO. In: CONGRESSO DE TECNOLOGIA E DESENVOLVIMENTO DA AMAZÔNIA, 2022, Paragominas. CONGRESSO DE TECNOLOGIA E DESENVOLVIMENTO DA AMAZÔNIA, 2022.

9. SANTOS, A. E. C. ; PAIXAO, E. A. R. ; FALCAO, IGOR W. S. ; CARDOSO, DIEGO L. . Desafios na Implementação de uma Rede H-CRAN Utilizando O simulador NS-3. In: CONGRESSO DE TECNOLOGIA E DESENVOLVIMENTO DA AMAZÔNIA, 2022, Paragominas. TECNOLOGIA E DESENVOLVIMENTO DA AMAZÔNIA, 2022.

10. SANTOS, A. E. C. ; CARDOSO, D. L. ; FALCAO, IGOR W. S. ; PAIXAO, E. A. R. ; FERNANDES, J. G. S. . Integração e Diferenciação de Células de Cobertura em uma Rede Centralizada de Acesso via Rádio Considerando Efeito de Maré. In: Conferência Nacional em Comunicações, Redes e Segurança da Informação, 2022, Guaramiranga-CE. Conferência Nacional em Comunicações, 2022.

7.3 Contribuições Adicionais

Lista de artigos publicados não incluídos neste trabalho e atividades de Orientação de alunos em paralelo a esta proposta de tese, incluída uma publicação em conferências nacionais e o desenvolvimento de uma ferramenta para área da saúde:

- NASCIMENTO, E. A. ; FALCAO, I. W. S. ; CARDOSO, D. L. . Análise do Efeito de Maré no Provisionamento de Recursos em Redes Centralizadas: Um Estudo de Caso em Grandes Metrópoles. In: Conferência Nacional em Comunicações, Redes e Segurança da Informação, 2020, Natal - RN. X Conferência Nacional em Comunicações, Redes e Segurança da Informação, 2020;

- SERUFFO, M. C. R. ; LIMA, M. F. M. ; FALCAO, IGOR W. S. . BAYESCLASS - Desenvolvimento de uma Ferramenta de Classificação De Dados para uso em Diferentes Áreas do Conhecimento. 2022;

- Orientação de Trabalho de Conclusão de Curso do Aluno Matheus da Foncesa Maia de Lima com o trabalho intitulado BAYESCLASS - Desenvolvimento De Uma Ferramenta De Classificação De Dados Para Uso Em Diferentes Áreas Do Conhecimento. 2022. (Graduação em Engenharia da Computação) - Universidade Federal do Pará;

- Orientação de Trabalho de Conclusão de Curso do Aluno João Guilherme Miranda da Paixão com o trabalho intitulado “Integração e Diferenciação de Células de Cobertura em Rede Centralizada de Acesso via Rádio Considerando o Efeito de Maré”. 2021. (Graduação em Engenharia da Computação) - Universidade Federal do Pará;

- Coorientação de Mestrado do Aluno Edney Almeida do Nascimento com o trabalho intitulado “Análise do Efeito de Maré no Provisionamento de Recursos em Redes Híbridas de Acesso via Rádio”. 2021. (Mestrado em Engenharia Elétrica) - Universidade Federal do Pará.

- Coorientação de Graduação do Aluno Matheus da Fonseca Maia de Lima com o trabalho intitulado "BAYESCLASS - DESENVOLVIMENTO DE UMA FERRAMENTA DE CLASSIFICAÇÃO DE DADOS PARA USO EM DIFERENTES ÁREAS DO CONHECIMENTO". 2022. (Graduação em Engenharia da Computação) - Universidade Federal do Pará.

- Coorientação de Graduação do Aluno Alex dos Reis Alves com o trabalho intitulado "APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA PARA IDENTIFICAÇÃO DO POTENCIAL HIDROGENIÔNICO (PH) EM AMBIENTE MONITORADO". 2023. (Graduação em Engenharia da Computação) - Universidade Federal do Pará.

7.4 Trabalhos Futuros

Como trabalhos futuros, pretende-se considerar outros atributos do conjunto de dados original, como informações clínicas mais detalhadas, dados da avaliação neurológica e informações sobre o contato com outros pacientes em tratamento médico. A hanseníase é uma doença com uma sintomatologia extensa e complexa; por isso, deve ser avaliada sob diversos aspectos multidisciplinares. O diagnóstico ainda é essencialmente clínico, sendo assim, ferramentas computacionais como a proposta nesta tese podem auxiliar os profissionais de saúde no processo de identificação de possíveis casos da doença, contribuindo para a criação de estratégias de controle, combate e tratamento mais eficazes.

O modelo preditivo proposto nesta pesquisa tem o potencial de ser incorporado em estratégias de auxílio ao tratamento e acompanhamento clínico de hanseníase, atuando como uma ferramenta essencial para identificar pacientes com maior probabilidade de desenvolver a doença no futuro, com base em suas características clínicas, sociais e ambientais. Essa abordagem permitirá a formulação de estratégias personalizadas de intervenção, adaptadas ao perfil de cada indivíduo, otimizando o planejamento terapêutico e o monitoramento contínuo dos casos. Além disso, a personalização possibilitará ações mais direcionadas e preventivas, com foco na identificação precoce de sinais de risco e na redução da progressão da doença em populações vulneráveis.

Para maximizar a aplicabilidade prática, propõe-se o desenvolvimento de uma ferramenta informatizada, amigável e acessível, capaz de auxiliar profissionais de saúde em diferentes níveis de atuação. Essa ferramenta, além de integrar o modelo preditivo, explorará as particularidades e especificidades da região amazônica, como a inclusão de dados socioambientais e culturais característicos. Dessa forma, busca-se não apenas facilitar o uso por profissionais de saúde locais, mas também ampliar a eficácia no controle e combate à hanseníase em um cenário que apresenta grandes desafios logísticos e epidemiológicos.

Em relação à metodologia, propõe-se testar outras técnicas para aprimorar a definição do

perfil clínico e epidemiológico dos pacientes, buscando algoritmos mais precisos para fornecer uma definição mais acurada dos resultados. Além disso, o modelo probabilístico desenvolvido nesta pesquisa será comparado com outros modelos, como redes Bayesianas, a fim de otimizar o processo de identificação de casos de hanseníase e garantir maior eficiência na detecção.

Outro ponto importante a ser explorado em trabalhos futuros é a investigação de como o modelo pode atuar na identificação de casos em indivíduos sem uma clínica confirmada. Essa abordagem pode ser fundamental na formulação de estratégias voltadas ao diagnóstico precoce da doença, facilitando o tratamento e controle em estágios iniciais, além de reduzir a disseminação e gravidade dos casos não detectados a tempo.

Por fim, todos os esforços nas diretrizes futuras serão desenvolvidos em colaboração com o corpo técnico e científico envolvido na concepção deste trabalho. A atuação conjunta será fundamental para assegurar que as soluções sejam robustas, aplicáveis e direcionadas às necessidades reais do cenário amazônico, contribuindo para avanços significativos no combate à hanseníase e na melhoria da qualidade de vida das populações afetadas.

REFERÊNCIAS

ABDI, Hervé; VALENTIN, Dominique. Multiple correspondence analysis. **Encyclopedia of measurement and statistics**, v. 2, n. 4, p. 651–657, 2007.

AHMAD, Shahnawaz et al. **WITHDRAWN: Fuzzy cloud based COVID-19 diagnosis assistant for identifying affected cases globally using MCDM**. Elsevier, 2021.

AHSAN, Md Manjurul; LUNA, Shahana Akter; SIDDIQUE, Zahed. Machine-learning-based disease diagnosis: A comprehensive review. In: MDPI, 3. HEALTHCARE. 2022. v. 10, p. 541.

ALVES, Elioenai Dornelles; FERREIRA, Telma Leonel; FERREIRA, Isaias Nery. Hanseníase avanços e desafios. In: HANSENÍASE avanços e desafios. 2014. P. 492–492.

ALVES, Hugo Vicentin; MORAES, Amarilis Giaretta de et al. The impact of KIR/HLA genes on the risk of developing multibacillary leprosy. **PLoS Neglected Tropical Diseases**, Public Library of Science San Francisco, CA USA, v. 13, n. 9, e0007696, 2019.

ANDRADE, Vera et al. Monitoring the elimination of leprosy in Brazil. **Int J Lepr Other Mycobact Dis**, v. 66, n. 4, p. 457–63, 1998.

ARAÚJO, Francisco A et al. Hanseniasis in the municipality of Western Amazon (Acre, Brazil): are we far from the goal of the World Health Organization? **Brazilian Journal of Infectious Diseases**, SciELO Brasil, v. 25, p. 101042, 2021.

ASAR, SH et al. PRISMA; preferred reporting items for systematic reviews and meta-analyses. **Journal of Rafsanjan University of Medical Sciences**, Journal of Rafsanjan University of Medical Sciences, v. 15, n. 1, p. 68–80, 2016.

AUBRY, A et al. Drug resistance in leprosy: an update following 70 years of chemotherapy. **Infectious Diseases Now**, Elsevier, 2022.

BATISTA, Gustavo Enrique de Almeida Prado et al. **Pré-processamento de dados em aprendizado de máquina supervisionado**. 2003. Tese (Doutorado) – Universidade de São Paulo.

BEN-ARIEH, Davıd. **Multi-criteria decision making methods: a comparative study**. JSTOR, 2002.

BERNARDES FILHO, Fred et al. Active search strategies, clinicoimmunobiological determinants and training for implementation research confirm hidden endemic leprosy in

inner São Paulo, Brazil. **PLoS Neglected Tropical Diseases**, Public Library of Science San Francisco, CA USA, v. 15, n. 6, e0009495, 2021.

BERNARDES-FILHO, Fred et al. Leprosy case series in the emergency room: A warning sign for a challenging diagnosis. **Brazilian Journal of Infectious Diseases**, SciELO Brasil, v. 25, 2021.

BETRU, Kebede Tefera; MAKUA, Thuledi. Challenges Experienced and Observed during the Implementation of Leprosy Strategies, Sidama Region, Southern Ethiopia: An inductive thematic analysis of qualitative study among health professionals who working with leprosy programs. **PLOS Neglected Tropical Diseases**, Public Library of Science San Francisco, CA USA, v. 17, n. 11, e0011794, 2023.

BOUTH, Raquel Carvalho et al. Specialized active leprosy search strategies in an endemic area of the Brazilian Amazon identifies a hypermutated Mycobacterium leprae strain causing primary drug resistance. **Frontiers in Medicine**, Frontiers Media SA, v. 10, p. 1243571, 2023.

BRASIL. **Hanseníase | Secretaria da Saúde**. 2024. Disponível em: <https://www.saude.pr.gov.br/Pagina/Hanseniase>.

BRASIL. **Lepra - Manual MSD Versão Saúde para a Família**. 2024. Disponível em: <https://www.msdmanuals.com/pt-br/casa/infec%C3%A7%C3%B5es/tuberculose-e-infec%C3%A7%C3%B5es-relacionadas/lepra>.

BRASIL. **Ministério da Saúde - Hanseníase**. 2024. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/h/hanseniase>.

BRASIL, ADEN. **DATASUS Tecnologia da Informação a Serviço do SUS**. Departamento de Informática do SUS Brasília, 2016.

BREEDVELD, Sebastiaan et al. Multi-criteria optimization and decision-making in radiotherapy. **European Journal of Operational Research**, Elsevier, v. 277, n. 1, p. 1–19, 2019.

BURKI, Talha. Old problems still mar fight against ancient disease. **The Lancet**, Elsevier, v. 373, n. 9660, p. 287–288, 2009.

CÁCERES-DURÁN, Miguel Ángel et al. MicroRNA biomarkers in leprosy: insights from the Northern Brazilian Amazon population and their implications in disease immune-physiopathology. **Frontiers in Genetics**, Frontiers Media SA, v. 15, p. 1320161, 2024.

CARGNIN, Zulamar Aguiar et al. Low back pain self-management mobile applications: a systematic review on digital platforms. **Revista da Escola de Enfermagem da USP**, SciELO Brasil, v. 58, e20230326, 2024.

CARVALHO, André C. P. L. F de; G. MENEZES, ngelo; BONIDIA, Robson P. **Ciência de Dados: Fundamentos e Aplicações**. 1. ed.: LTC, 2024. ISBN 9788521638766.

CHATURVEDI, Anil; GREEN, Paul E; CAROLL, J Douglas. K-modes clustering. **Journal of classification**, Springer, v. 18, p. 35–55, 2001.

CHATURVEDI, Anil; GREEN, Paul E; CAROLL, J Douglas. K-modes clustering. **Journal of classification**, Springer, v. 18, p. 35–55, 2001.

CHEN, Xiaohua et al. Risk factors for physical disability in patients with leprosy disease in Yunnan, China: Evidence from a retrospective observational study. **PLoS Neglected Tropical Diseases**, Public Library of Science San Francisco, CA USA, v. 15, n. 11, e0009923, 2021.

CONN, Samuel S. OLTP and OLAP data integration: a review of feasible implementation methods and architectures for real time data analysis. In: IEEE. PROCEEDINGS. IEEE SoutheastCon, 2005. 2005. P. 515–520.

DE NARDO, Pasquale et al. Multi-Criteria Decision Analysis to prioritize hospital admission of patients affected by COVID-19 in low-resource settings with hospital-bed shortage. **International Journal of Infectious Diseases**, Elsevier, v. 98, p. 494–500, 2020.

DHARMAWAN, Yudhy et al. Delayed detection of leprosy cases: A systematic review of healthcare-related factors. **PLoS Neglected Tropical Diseases**, Public Library of Science San Francisco, CA USA, v. 16, n. 9, e0010756, 2022.

DUTRA DA SILVA, Ygor Eugenio et al. Data Mining Using Clustering Techniques as Leprosy Epidemiology Analyzing Model. In: SPRINGER. DATA Mining and Big Data: Third International Conference, DMBD 2018, Shanghai, China, June 17–22, 2018, Proceedings 3. 2018. P. 284–293.

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

FELICIANO, Katia V de O; KOVACS, Maria Helena; ALZATE, Alberto. Diagnóstico precoce da hanseníase: o caso dos serviços de saúde no Recife (Pernambuco), Brasil. **Revista Panamericana de Salud Pública**, SciELO Public Health, v. 4, n. 1, p. 6–13, 1998.

FISHER, Ronald Aylmer. Statistical methods for research workers. In: BREAKTHROUGHS in statistics: Methodology and distribution. Springer, 1970. P. 66–70.

GAMA, Rafael Silva et al. Prospects for new leprosy diagnostic tools, a narrative review considering ELISA and PCR assays. **Revista da Sociedade Brasileira de Medicina Tropical**, SciELO Brasil, v. 53, e20200197, 2020.

GAMA, Rafael Silva et al. Prospects for new leprosy diagnostic tools, a narrative review considering ELISA and PCR assays. **Revista da Sociedade Brasileira de Medicina Tropical**, SciELO Brasil, v. 53, 2020.

GARDNER, Stephen R. Building the data warehouse. **Communications of the ACM**, ACM New York, NY, USA, v. 41, n. 9, p. 52–60, 1998.

GOLDEN, Bruce L; WASIL, Edward A; HARKER, Patrick T. The analytic hierarchy process. **Applications and Studies, Berlin, Heidelberg**, Springer, v. 2, n. 1, p. 1–273, 1989.

GOULART, Isabela Maria Bernardes; GOULART, Luiz Ricardo. Leprosy: diagnostic and control challenges for a worldwide disease. **Archives of dermatological research**, Springer, v. 300, p. 269–290, 2008.

GUHA, Sudipto; RASTOGI, Rajeev; SHIM, Kyuseok. ROCK: A robust clustering algorithm for categorical attributes. **Information systems**, Elsevier, v. 25, n. 5, p. 345–366, 2000.

HENRY, Mary et al. Factors contributing to the delay in diagnosis and continued transmission of leprosy in Brazil–an explorative, quantitative, questionnaire based study. **PLoS neglected tropical diseases**, Public Library of Science, v. 10, n. 3, e0004542, 2016.

HONG, Se Jin et al. Software-based interventions for low back pain management: A systematic review and meta-analysis. **Journal of Nursing Scholarship**, Wiley Online Library, v. 56, n. 2, p. 206–226, 2024.

HOOIJ, Anouk van et al. BCG-induced immunity profiles in household contacts of leprosy patients differentiate between protection and disease. **Vaccine**, Elsevier, v. 39, n. 50, p. 7230–7237, 2021.

HUNTER, Shirley W; BRENNAN, Patrick J. A novel phenolic glycolipid from Mycobacterium leprae possibly involved in immunogenicity and pathogenicity. **Journal of bacteriology**, Am Soc Microbiol, v. 147, n. 3, p. 728–735, 1981.

JIN, Bo; CRUZ, Leandro; GONÇALVES, Nuno. Deep facial diagnosis: deep transfer learning from face recognition to facial diagnosis. **IEEE Access**, IEEE, v. 8, p. 123649–123661, 2020.

KHAN, Muslim et al. Bi-PSSM: Position specific scoring matrix based intelligent computational model for identification of mycobacterial membrane proteins. **Journal of Theoretical Biology**, Elsevier, v. 435, p. 116–124, 2017.

KIMBALL, Ralph. A dimensional modeling manifesto. **Dbms**, Miller Freeman, Inc. Lawrence, KS, USA, v. 10, n. 9, p. 58–70, 1997.

KIRA, Kenji; RENDELL, Larry A. The feature selection problem: Traditional methods and a new algorithm. In: PROCEEDINGS of the tenth national conference on Artificial intelligence. 1992. P. 129–134.

KIRCHHEIMER, Waldemar F; STORKS, EE et al. Attempts to establish the armadillo (Dasypus novemcinctus Linn.) as a model for the study of leprosy. I. Report of lepromatoid leprosy in an experimentally infected armadillo. **International Journal of Leprosy**, v. 39, n. 3, p. 693–702, 1971.

KRYSANOVA, V; KRYSANOV, I; ERMAKOVA, V. The multicriteria decision analysis of using tetrabenazine for patients with hungtington's disease in Russia. **Value in Health**, Elsevier, v. 20, n. 9, a565, 2017.

LEANO, Heloisy Alves de Medeiros et al. Fatores socioeconômicos relacionados à hanseníase: revisão integrativa da literatura. **Revista Brasileira de Enfermagem**, SciELO Brasil, v. 72, p. 1405–1415, 2019.

MOHAMMED, KI et al. A uniform intelligent prioritisation for solving diverse and big data generated from multiple chronic diseases patients based on hybrid decision-making and voting method. **IEEE Access**, IEEE, v. 8, p. 91521–91530, 2020.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. **Sistemas inteligentes-Fundamentos e aplicações**, Manole, v. 1, n. 1, p. 32, 2003.

MOOSIVAND, Asiye et al. An application of multi-criteria decision-making approach to sustainable drug shortages management: evidence from a developing country. **Journal of Pharmaceutical Health Care and Sciences**, Springer, v. 7, p. 1–11, 2021.

MURUGESAN, San. Understanding Web 2.0. **IT professional**, IEEE, v. 9, n. 4, p. 34–41, 2007.

NAINGGOLAN, Rena et al. Improved the performance of the K-means cluster using the sum of squared error (SSE) optimized by using the Elbow method. In: IOP PUBLISHING, 1. JOURNAL of Physics: Conference Series. 2019. v. 1361, p. 012015.

NEVES, Karine Vila Real Nunes et al. Misdiagnosis of leprosy in Brazil in the period 2003-2017: spatial pattern and associated factors. **Acta Tropica**, Elsevier, v. 215, p. 105791, 2021.

NOBRE, Mauricio Lisboa et al. Multibacillary leprosy by population groups in Brazil: Lessons from an observational study. **PLoS neglected tropical diseases**, Public Library of Science San Francisco, CA USA, v. 11, n. 2, e0005364, 2017.

NOGUEIRA, A et al. Development of a computational system in mobile devices for the optimization of the process of collection, management and analysis of data related to leprosy patients in the West of the State of Pará—Brazil. **Hansenol. Int**, v. 39, p. 71, 2014.

ORGANIZATION, World Health et al. Global leprosy update, 2018: moving towards a leprosy-free world. **Wkly Epidemiol Rec**, v. 94, n. 35/36, p. 389–411, 2019.

ORGANIZATION, World Health et al. **Towards zero leprosy: global Leprosy (Hansen's disease) Strategy 2021–2030**. World Health Organization, 2021.

PALMEIRA, Catia Suely. Noções básicas de epidemiologia, 2020.

PINAZO, Maria-Jesus et al. Multi-criteria decision analysis approach for strategy scale-up with application to Chagas disease management in Bolivia. **PLoS Neglected Tropical Diseases**, Public Library of Science San Francisco, CA USA, v. 15, n. 3, e0009249, 2021.

PINTO, Pablo et al. Leprosy piRnome: exploring new possibilities for an old disease. **Scientific reports**, Nature Publishing Group UK London, v. 10, n. 1, p. 12648, 2020.

PROVOST, Foster; FAWCETT, Tom. Data science and its relationship to big data and data-driven decision making. **Big data**, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 1, n. 1, p. 51–59, 2013.

QAMAR, Ali Mustafa; GAUSSIER, Eric. Similarity learning in nearest neighbor and relief algorithm. In: IEEE. 2010 Ninth International Conference on Machine Learning and Applications. 2010. P. 183–189.

RIGATTI, Steven J. Random forest. **Journal of Insurance Medicine**, American Academy of Insurance Medicine 1700 Magnavox Way, Fort Wayne, IN 46804, v. 47, n. 1, p. 31–39, 2017.

ROGATI, Monica. The AI hierarchy of needs. **Hacker Noon**, 2017.

ROLLES, Steve et al. A multi criteria decision analysis (MCDA) for evaluating and appraising government policy responses to non medical heroin use. **International Journal of Drug Policy**, Elsevier, v. 91, p. 103180, 2021.

ROY, Bernard. The outranking approach and the foundations of ELECTRE methods. **Theory and decision**, Springer, v. 31, p. 49–73, 1991.

ROY, Bernard; BERTIER, B. Le methods ELECTRE II: Une methode de classement en presence de criteres multiples, note de travail no. 142. **Direction Scientifique, Groupe Metra**, 1971.

SALGADO, Claudio Guedes et al. What do we actually know about leprosy worldwide? **The Lancet Infectious Diseases**, Elsevier, v. 16, n. 7, p. 778, 2016.

SANTANA, Juliana F de et al. Engineered biomarkers for leprosy diagnosis using labeled and label-free analysis. **Talanta**, Elsevier, v. 187, p. 165–171, 2018.

SANTÉ, Organisation mondiale de la; ORGANIZATION, World Health et al. Global leprosy update, 2018: moving towards a leprosy-free world–Situation de la lèpre dans le monde, 2018: parvenir à un monde exempt de lèpre. **Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire**, World Health Organization= Organisation mondiale de la Santé, v. 94, n. 35/36, p. 389–411, 2019.

SAUNDERSON, Paul. WHO global leprosy (Hansen's disease) update, 2022: new paradigm–control to elimination. **Leprosy Review**, v. 94, n. 4, p. 262–263, 2023.

SCOLLARD, David M et al. The continuing challenges of leprosy. **Clinical microbiology reviews**, Am Soc Microbiol, v. 19, n. 2, p. 338–381, 2006.

SCULLEY, David. Web-scale k-means clustering. In: PROCEEDINGS of the 19th international conference on World wide web. 2010. P. 1177–1178.

SHARMA, Mukul; SINGH, Pushpendra. Advances in the diagnosis of leprosy. **Frontiers in Tropical Diseases**, Frontiers Media SA, v. 3, p. 893653, 2022.

SUN, Yijun. Iterative RELIEF for feature weighting: algorithms, theories, and applications. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 29, n. 6, p. 1035–1051, 2007.

TERAPÊUTICAS, E DIRETRIZES. Protocolo Clínico e Diretrizes Terapêuticas da Hanseníase, 2022.

THORNDIKE, Robert L. Who belongs in the family. In: CITESEER. PSYCHOMETRIKA. 1953.

TZENG, Gwo-Hshiung; HUANG, Jih-Jeng. **Multiple attribute decision making: methods and applications**. CRC press, 2011.

VAN DER AALST, Wil; AALST, Wil van der. **Data science in action**. Springer, 2016.

VEALE, Michael; BINNS, Reuben. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. **Big Data & Society**, SAGE Publications Sage UK: London, England, v. 4, n. 2, p. 2053951717743530, 2017.

VELLUCCI, Sherry L. Metadata. **Annual Review of Information Science and Technology (ARIST)**, ERIC, v. 33, p. 187–222, 1998.

VILLANUEVA, Vicente et al. Identifying key unmet needs and value drivers in the treatment of focal-onset seizures (FOS) in patients with drug-resistant epilepsy (DRE) in Spain through multi-criteria decision analysis (MCDA). **Epilepsy & Behavior**, Elsevier, v. 122, p. 108222, 2021.

VINNARASAN, Mariaviagulam et al. A clinico-epidemiological study of paediatric leprosy in a tertiary care centre. **Journal of Evolution of Medical and Dental Sciences**, Akshantala Enterprises Private Limited, v. 7, n. 21, p. 2558–2562, 2018.

VISHWAKARMA, Vinayak; PRAKASH, Chandra; BARUA, Mukesh Kumar. A fuzzy-based multi criteria decision making approach for supply chain risk assessment in Indian pharmaceutical industry. **International Journal of Logistics Systems and Management**, Inderscience Publishers (IEL), v. 25, n. 2, p. 245–265, 2016.

VOLTAN, Glauber et al. Silent peripheral neuropathy determined by high-resolution ultrasound among contacts of patients with Hansen's disease. **Frontiers in Medicine**, Frontiers Media SA, v. 9, p. 1059448, 2023.

WITTEN, Ian H et al. Practical machine learning tools and techniques. In: 4. DATA Mining. 2005. v. 2.

WRIGHT, Raymond E. Logistic regression. American Psychological Association, 1995.

WU, Feng-Zhi et al. Analysis of the Temporal and Spatial Distribution of New Cases of Leprosy in Yunnan Province, 2011-2016. In: IEEE. 2018 26th International Conference on Geoinformatics. 2018. P. 1–3.

ZANAKIS, Stelios H et al. Multi-attribute decision making: A simulation comparison of select methods. **European journal of operational research**, Elsevier, v. 107, n. 3, p. 507–529, 1998.