INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO CEARÁ
IFCE CAMPUS ARACATI
COORDENADORIA DE CIÊNCIA DA COMPUTAÇÃO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

OTON CRISPIM BRAGA

SOLUÇÃO INTELIGENTE BASEADA EM APRENDIZADO DE MÁQUINA PARA A CLASSIFICAÇÃO DE DOENÇAS TRANSMITIDAS PELO VETOR AEDES AEGYPTI

ARACATI-CE
2017

OTON CRISPIM BRAGA

SOLUÇÃO INTELIGENTE BASEADA EM APRENDIZADO DE MÁQUINA PARA A CLASSIFICAÇÃO DE DOENÇAS TRANSMITIDAS PELO VETOR AEDES AEGYPTI

Trabalho de Conclusão de Curso (TCC) apresentado ao curso de Bacharelado em Ciência da Computação do Instituto Federal de Educação, Ciência e Tecnologia do Ceará - IFCE - Campus Aracati, como requisito parcial para obtenção do Título de Bacharel em Ciência da Computação.

Orientador (a): Prof. Ms. Mário Wedney de Lima Moreira
Co-Orientador (a): Prof. Dr. Antônio Mauro Barbosa de Oliveira

Aracati-CE
2017

OTON CRISPIM BRAGA

SOLUÇÃO INTELIGENTE BASEADA EM APRENDIZADO DE MÁQUINA PARA A CLASSIFICACAO DE DOENÇAS TRANSMITIDAS PELO VETOR AEDES AEGYPTI

Trabalho de Conclusão de Curso (TCC) apresentado ao curso de Bacharelado em Ciência da Computação do Instituto Federal de Educação, Ciência e Tecnologia do Ceará - IFCE - Campus Aracati, como requisito parcial para obtenção do Título de Bacharel em Ciência da Computação.

Aprovado em 23 de outubro de 2017

BANCA EXAMINADORA

Mário Wedney de Lima Moreira

Prof. Ms. Mário Wedney de Lima Moreira (Orientador)
Instituto Federal de Educação, Ciência e Tecnologia do Ceará

Prof. Dr. Antônio Mauro Barbosa de Oliveira (Co-Orientador)
Instituto Federal de Educação, Ciência e Tecnologia do Ceará

Profª. Ms. Francisca Raquel de Vasconcelos Silveira
Instituto Federal de Educação, Ciência e Tecnologia do Ceará

Prof. Ms. Paulo Alberto Melo Barbosa
Instituto Federal de Educação, Ciência e Tecnologia do Ceará

Dados Internacionais de Catalogação na Publicação
Instituto Federal do Ceará - IFCE
Sistema de Bibliotecas - SIBI
Ficha catalográfica elaborada pelo SIBI/IFCE, com os dados fornecidos pelo(a) autor(a)

B813s Braga, Oton.
SOLUÇÃO INTELIGENTE BASEADA EM APRENDIZADO DE MÁQUINA PARA A CLASSIFICAÇÃO DE DOENÇAS TRANSMITIDAS PELO VETOR AEDES AEGYPTI / Oton Braga. - 2017.
53 f. : il. color.

Trabalho de Conclusão de Curso (graduação) - Instituto Federal do Ceará, Bacharelado em Ciência da Computação, Campus Aracati, 2017.
Orientação: Prof. Me. Mário Wedney de Lima Moreira.
Coorientação: Prof. Dr. Antônio Mauro Barbosa de Oliveira.

1. Sistemas Inteligentes. 2. Doenças Transmitidas pelo Aedes Aegypti. 3. Aprendizado de Máquina. 4. Manejo Clínico. 5. Mineração de Dados. I. Titulo.

DEDICATÓRIA

Dedico essa produção intelectual a todos que sonham e "acreditam que podem mudar o mundo, porque são esses que realmente o farão"...

AGRADECIMENTOS

Agradeço aos meus orientadores pelo esforço e disposição durante o desenvolvimento deste trabalho. Em momentos de dúvida, sempre me indicaram materiais ou sugeriram caminhos que me levaram a solução dos problemas. Meu grande amigo Professor Mário Moreira, amante da gastronomia portuguesa. E meu *Coach* Professor Mauro Oliveira, que também é coordenador do LAR (Laboratório de Redes de Computadores do Aracati), onde encontrei apoio, conhecimento e infraestrutura para desenvolver este trabalho. Agradeço também a FUNCAP, por ter me dado apoio financeiro para o desenvolvimento desse trabalho. Agradeço a todos os professores que contribuíram com a minha formação, em especial os que agregaram ao profissionalismo características humanas indispensáveis, ponderando suas avaliações de forma sensível e empática. Agradeço também a minha professora de IA, Raquel Silveira, pela sua notável contribuição.

Agradeço especialmente aos meus pais Otoniel e Maria, minha avó materna Dona Raimunda e avós paternos Jonas e Santana e minha tia Denice, que contribuíram fora do ambiente acadêmico para minha formação.

RESUMO

Momentos de incerteza são frequentes em situações complexas pois muitos fatores podem influenciar o processo de tomada de decisão. Má condições do ambiente, fadiga, estresse e até fatores emocionais podem contribuir negativamente em momentos críticos. Na área da saúde, esses momentos podem surgir em diversas etapas durante o manejo clínico das doenças. Assim, a fim de auxiliar profissionais em momentos de incerteza, muitos sistemas computacionais vêm sendo propostos. Alguns deles têm apresentado ótimos resultados, dando suporte ao processo de tomada de decisão em situações diversas. Contudo poucas propostas abrangem todo o processo de manejo clínico das doenças, focando esforços em etapas específicas, como o diagnóstico final. Portanto, este trabalho propõe uma solução inteligente baseada em classificadores como mecanismo de inferência, capaz de auxiliar profissionais de saúde durante o processo de manejo clínico das doenças transmitidas pelo mosquito *Aedes Aegypti*, identificando qual o provável diagnóstico baseado em sintomas e resultado de exames. Para tanto, dividiu-se o trabalho em dois passos: um voltado para o pré-diagnóstico, considerando sintomas e histórico clínico, anamnese; e outro focado no diagnóstico final, considerando também resultados de exames específicos, como exames de sorologia. O estudo utiliza uma metodologia baseada na Mineração de Dados para extração de conhecimento numa base de exemplos. Após diversos testes e ajustamentos em algoritmos de aprendizado de máquina, pôde-se definir dois modelos de aprendizado capazes de inferir a probabilidade de um paciente estar infectado com uma determinada doença, tendo precisão de até $91,6\%$. A partir desses modelos, pôde-se construir uma API inteligente de apoio a tomada de decisão durante o manejo clínico de dengue e *chikungunya*. A solução permite que diversas aplicações acessem os modelos de aprendizado. Como prova de conceito também foi desenvolvida uma plicação móvel de consulta popular para identificação de dengue e *chikungunya*, ainda em fase de prototipação.

Palavras-chave: Sistemas Inteligentes. Doenças Transmitidas pelo Aedes Aegypti. Aprendizado de Máquina. Manejo Clínico. Mineração de Dados.

ABSTRACT

Uncertainty moments are frequent in complex situations because many factors can influence the decision-making process. Adverse environmental conditions, fatigue, stress, and even emotional factors can contribute negatively in critical moments. In health, these moments can occur in several stages during the clinical management of diseases. Thus, to assist professionals in uncertainty moments, many computer systems have been proposed. Some of them have presented excellent results, supporting the decision-making process in diverse situations. However, few proposals coverage the whole process of clinical disease management, focusing on specific steps, for example, the final diagnosis. Therefore, this work proposes an intelligent solution based on classifiers as an inference mechanism capable of assisting health professionals during the clinical management process of diseases transmitted by the *Aedes Aegypti* mosquito, identifying the probable diagnosis based on symptoms and exam results. For that, the work was divided into two steps, to know, a step focused on pre-diagnosis, considering symptoms and clinical history, anamnesis; and another focused on final diagnosis, also considering results of specific tests, such as serology. This study uses a methodology based on data mining for knowledge extraction based on examples. After several tests and adjustments in machine learning algorithms, two learning models capable of inferring the probability of a patient being infected with a specific disease could be defined, with an accuracy up to 91.6%. From these models, an intelligent API to support decision making during the clinical management of dengue and *chikungunya* can be constructed. This solution allows diverse applications to access the learning models. These include a popular mobile application for dengue and *chikungunya* identification, and an interoperable clinical management system of *chikungunya*, called MARCIA, both in the prototyping phase.

Keywords: Smart systems. Diseases transmitted by the *Aedes Aegypti* mosquito. Machine learning. Clinical management. Data mining.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

IFCE	Instituto Federal de Educação, Ciência e Tecnologia do Ceará
TCC	Trabalho de Conclusão de Curso
IA	Inteligência Artificial
IAM	Inteligência Artificial Na Medicina
MD	Mineração de Dados
AM	Aprendizado de Máquina
SAD	Sistema de Apoio a Decisão
SINAN	Sistema Informação de Agravos de Notificação
API	*Application Programming Interface*
REST	*REpresentational State Transfer*
RNA	Rede Neural Artificial
RF	*Random Forest*
RT	*Random Tree*
NB	*Naïve Bayes*
BN	*Bayes Network*
MLP	*Multilayer Perceptron*
kNN	*k-Nearest Neighbors*
SVM	*Suport Vetor Machine*
Prec.	Precisão
Cob.	Cobertura
Med. Harm	*Média Harmônica*

SUMÁRIO

1 INTRODUÇÃO

As tecnologias mais recentes têm transformado a maneira como as pessoas vivem e se comunicam. As tecnologias da informação e comunicação (TIC) são as mais influentes, estando presentes em quase todos os setores. Seu uso tem beneficiado profissionais de várias áreas, informatizando processos e agilizando atividades rotineiras. O armazenamento e apresentação de dados de forma facilitada é uma das grandes ferramentas da informática. Esses dados, quando organizados, dão subsídios a especialistas no processo de tomada de decisão. Com o avanço da computação, sistemas mais complexos surgiram com o objetivo de integrar e analisar dados de sistemas distintos para auxiliar profissionais a tomarem melhores decisões.

A IA (Inteligência Artificial) é uma subárea da Ciência da Computação que estuda formas de reproduzir o raciocínio humano (FACELI et al., 2015). Sistemas baseados em IA sugerem ações e preveem eventos baseando-se na análise dos dados. Nos últimos anos foram propostos sistemas computacionais inteligentes capazes de resolver problemas mais genéricos e aprender de maneira autônoma, além de interagirem entre si e com seres humanos. Na década de 1980, pesquisadores de IA e da área médica uniram esforços para definir o campo da Inteligência Artificial em Medicina (IAM), que trouxe um grande avanço em sistemas computacionais capazes de auxiliar os especialistas no diagnóstico médico (COIERA, 2015). Hoje, muitas pesquisas em IAM têm desenvolvido aplicações e soluções inovadoras, melhorando a qualidade de vida de muitas pessoas e auxiliando profissionais de saúde em procedimentos complexos que envolvem a tomada de decisão (LOBO, 2017). Tais sistemas são capazes de inferir novos conhecimentos a partir de um conjunto de exemplos. Para tanto, mecanismos de Aprendizado de Máquina (AM) são treinados e ajustados ao contexto do problema. Este processo contém uma série de etapas e ações complexas, que influenciam o resultado final de diversas maneiras.

1.1 Motivação

O processo de disgnóstico e tratamento médico é composto por várias etapas, conhecidas como manejo clínico. Dependendo do caso, o paciente precisa realizar diversas visitas ao hospital a fim de fornecer informações aos profissionais de saúde, seja por meio de entrevistas ou de exames. Cada etapa desse processo tem o objetivo de agregar mais informações ao diagnóstico, tornando-o mais exato e confiável. Para chegar a uma decisão, um médico utiliza ferramentas, informações e a sua própria experiência. Contudo, a indisponibilidade desses recursos afeta significativamente a

qualidade das decisões. Muitas vezes informações incompletas, incorretas ou mal interpretadas podem dificultar e retardar o diagnóstico médico. Também podem ocorrer casos de incerteza, pois algumas doenças apresentam sintomas parecidos ou idênticos, exigindo exames específicos para um diagnóstico preciso. Em alguns casos, mesmo com estes exames clínicos, especialistas em saúde não são capazes de proporcionar certeza ao diagnóstico. Além disso, alguns passos do manejo clínico podem demorar muito tempo, atrasando o processo e diminuindo a confiabilidade dos resultados. Em alguns casos os primeiros resultados podem não ter relevância para o diagnóstico final, gerando desperdício de recursos.

Doenças de característica endêmica, por exemplo, exigem atenção aumentada, pois conseguem disseminar-se com facilidade, como é o caso da dengue, da febre *chikungunya* e da febre causada pelo vírus *Zika*, que fazem parte do quadro de doenças de notificação compulsória agregadas ao Sistema de Informação de Agravos de Notificação (SINAN). Essas doenças têm atingido diversos estados no país, causando epidemias em várias regiões. O combate ao mosquito *Aedes Aegypti*, transmissor destas doenças, tem se tornado o principal objeto de campanha de saúde pública no Brasil, segundo o Ministério da Saúde. Já foram liberados mais de 20 milhões de reais somente no ano de 2016 para combater o mosquito (BRASIL, 2016c). Diversas iniciativas foram tomadas para conter o seu avanço, contudo ele se desenvolve rapidamente e, em ambientes favoráveis, se reproduz com facilidade.

1.2 Caracterização do Problema

Doenças como dengue, *chikungunya* e *zica* apresentam características e sintomas semelhantes, o que dificulta seu diagnóstico. Com o objetivo de resolver esse problema, o Ministério da Saúde elaborou manuais de manejo clínico bem definidos para essas doenças, que são tratadas especificamente de forma diferente (BRASIL, 2016a; FARIA et al.,). Entretanto, para um diagnóstico preciso, são necessários exames mais específicos. Tais exames são relativamente caros e nem sempre estão disponíveis em hospitais públicos, que solicitam análise em laboratórios externos. Ainda assim, devido a alta demanda ou indisponibilidade de materiais ou compostos químicos, o resultado de exames desse tipo ainda sofre grandes atrasos, afetando negativamente o acompanhamento dos pacientes. Muitas vezes o resultado dos exames só chega quando estes já se encontram saudáveis ou em estágio de alto agravamento.

1.3 Proposta

Uma solução para este problema é o uso de ferramentas inteligentes capazes de auxilar especialistas em saúde no processo de tomada de decisão no manejo clínico de doenças complexas. Este trabalho apresenta uma solução inteligente, baseada no Aprendizado de Máquina (AM), capaz de auxiliar profissionais de saúde no diagnóstico de doenças transmitidas pelo mosquito *Aedes Aegypti*, apoiando as etapas do manejo clínico dessas doenças.

A partir de dados abertos disponibilizados pelo portal da prefeitura de Recife (PE), Brasil, extraiu-se milhares de casos já classificados, relativos às doenças em questão. Esses casos deram subsídios aos algoritmos de classificação usados neste trabalho, que foram treinados para classificar novos casos desconhecidos. O sistema desenvolvido neste trabalho atende mais de uma etapa do manejo clínico, portanto, este processo de treinamento realizou-se em duas etapas: uma para auxiliar profissionais na etapa de suspeição da doença, analisando somente sintomas e resultados de exames rápidos; e outra voltada ao diagnóstico final, levando também em consideração os resultados de exames mais específicos. O sistema desenvolvido conta com dois componentes principais: o módulo de conexão e o módulo de inferência. O módulo de conexão recebe requisições REST (*Representational State Transfer*) com os atributos disponibilizados pela aplicação. Dependendo do caso ou da etapa do manejo clínico, serão recebidas informações sobre sintomas, histórico de saúde ou exames clínicos realizados pelo paciente em questão. Essas informações são tratadas e enviadas para o módulo de inferência que, a partir de técnicas de AM, estima a probabilidade de um paciente ter contraído uma das doenças. Este módulo analisa um conjunto de casos diagnosticados das doenças em questão, treinando os algoritmos de AM usados no processo de classificação.

1.4 Objetivo Principal

Esse trabalho tem como objetivo o desenvolvimento de uma solução inteligente, baseada em Aprendizado de Máquina, capaz de classificar doenças transmitidas pelo vetor Aedes Aegypti, contribuindo para suspeição e diagnóstico destas doenças, apoiando todo o processo de manejo clínico das tais.

1.4.1 Objetivos Específicos

Para alcançar a proposta acima destacada, foram estabelecidos os seguintes objetivos específicos:

- identificar e selecionar algoritmos existentes no contexto do problema abordado;
- prospectar casos reais para treinamento dos algoritmos propostos;
- analisar e preparar dados dos casos reais para uma melhor descoberta de conhecimento;
- especificação dos fluxos das etapas de suspeição e de diagnóstico do manejo clínico das doenças em análise;
- treinar os algoritmos baseados em AM e analisar seus resultados;
- ajustar tais algoritmos a fim de melhorar suas precisões;
- implementar uma API capaz de atender o manejo clínico das doenças em estudo, apoiando o processo de tomada de decisão em várias etapas.

1.5 Organização do Trabalho

O trabalho está organizado da seguinte forma. O Capítulo 2 trata das tecnologias e mecanismos inteligentes utilizados na pesquisa, destacando seu funcionamento básico. Nele também é fundamentado o contexto das doenças epidemiológicas em estudo. O Capítulo 3 destaca os principais trabalhos desenvolvidos na área de AM, dando ênfase aos trabalhos em IAM e no contexto das doenças epidemiológicas, incluindo trabalhos focados em dengue. No Capítulo 4 é apresentada a metodologia utilizada nessa pesquisa, descrevendo todos os passos realizados até sua conclusão. Neste tópico também são destacadas as principais alterações em relação as metodologias observadas na literatura. O Capítulo 5 apresenta e discute os resultados dos testes realizados na pesquisa, destacando os melhores resultados para a solução do problema em estudo. O Capítulo 6 apresenta a solução proposta, sua arquitetura e funcionamento. Finalmente, o Capítulo 7 conclui o trabalho através de uma análise do impacto da solução proposta, evidenciando o método utilizado. Este capítulo também sugere propostas para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

O profissional em saúde precisa, em geral, seguir um procedimento específico para tomada de decisão. Inicialmente, ele analisa as informações apresentadas pelo paciente. Em seguida, a partir de hipóteses, solicita exames ou testes para validá-la. Contudo, muitos fatores podem influenciar o processo de tomada de decisão deste profissional em saúde. O ambiente, fadiga, estresse, o excesso de pacientes e até fatores emocionais podem contribuir negativamente para a tomada de decisão. Os algoritmos classificadores de Mineração de Dados (MD) usam um procedimento semelhante, embora menos complexo. Estes classificadores analisam um conjunto de casos a fim de identificar características e comportamentos nos dados, gerando um modelo de aprendizado capaz de inferir a probabilidade de um novo caso pertencer, ou não, a uma determinada classe.

2.1 Contexto do Problema

A dengue é uma doença viral que têm atingido o país desde a década de 1990 (BARRETO; TEIXEIRA, 2008). Já ocorreram diversos casos de surtos e epidemias da doença no Brasil, que causaram prejuízos econômicos e sociais ao país. Atualmente, a dengue é relacionada às doenças *chikungunya* e *zika*, que apresentam algumas características em comum e são transmitidas pelo mesmo vetor. O poder de alcance das doenças é semelhante, como mostram os dados do Sistema Único de Saúde (SUS) brasileiro (BRASIL, 2016b).

2.1.1 Contexto Epidemiológico

A dengue, a febre de *chikungunya* e a febre causada pelo vírus *zika* são doenças que fazem parte da Lista Nacional de Notificação Compulsória de doenças, agravos e eventos de saúde pública. As doenças são transmitidas por intermédio do vetor *Aedes Aegypti*, um mosquito hospedeiro que se alimenta de sangue. O vírus é carregado da pessoa doente pra pessoa saudável através da picada de uma fêmea do mosquito. O vetor transmissor se reproduz com facilidade principalmente em lugares que contenham água parada, onde a fêmea deposita seus ovos. Ele tem se reproduzido de forma rápida, principalmente por causa das más condições sanitárias, realidade de muitas cidades do país. Diversas campanhas de saúde pública já foram realizadas com o objetivo de conscientizar a população sobre as medidas sanitárias necessárias para combater o vetor. No entanto, os números de casos das

doenças vem aumentando a cada ano. Isso tem causado prejuízos significativos à saúde pública. Os casos evoluem principalmente nos períodos chuvosos das primeiras semanas do ano (CÂMARA et al., 2007). A incidência destas doenças tem atingido patamares altíssimos. Os números chagam a ser alarmantes. Segundo o Boletim Epidemiológico do SNS, 335.333 casos foram notificados até a Semana Epidemiológica (SE) 19 do ano de 2017, que vai até o dia 13 de maio do referido ano. Estes dados foram obtidos através do Sistema de Informação de Agravos de Notificação (Sinan), conhecido como Sinan Net. A Tabela 1 mostra as notificações das doenças e os casos confirmados (BRASIL, 2016b).

Tabela 1 – Casos Notificados e Confirmados

Doença	Notificações	Casos Confirmados
Dengue	244.433	144.326
Chikungunya	80.949	28.225
Zika	9.951	3.356

2.1.2 Contexto Metodológico

O SNS propõe manuais de manejo clínico para protocolar os processos de notificação dos casos de dengue, febre *chikungunya* e febre causada pelo vírus *zika* (BRASIL, 2016a). Tais procedimentos incluem os processos de anamnese, exames físicos e laboratoriais, que são descritos abaixo.

Anamnese: entrevista realizada por um profissional de saúde com o paciente com o objetivo de entender todos os fatos ocorridos relacionados ao problema investigado, servindo como ponto de partida para seu diagnóstico. O histórico clínico do paciente deve ser o mais detalhada possível;

Exame físico: coleta de sinais vitais, exame de pele, exame neurológico e oftalmológico, exames articulares (alteração da pele, aumento de volume, crepitação ou estalido, deformidade, limitação da mobilidade, dor ou atrofia muscular, nodulação) e exames físicos nos membros superiores e inferiores;

Exames laboratoriais: análise de exames mais específicos, como hemograma completo e sorologia, que vão identificar a presença de anticorpos no sangue, incluindo todas as informações relevantes presentes no sangue. Essas informações dão mais precisão ao diagnóstico;

Conduta: após o diagnóstico confirmado, ou ainda em suspeição, aplicam-se procedimentos a fim de tratar os sintomas da doença. Dependendo do caso evita-se alguns tipos de medicamentos. Durante a conduta o caso pode evoluir, exigindo procedimentos mais específicos.

Durante estas etapas os dados são inseridos em formulários impressos e atualizados de forma manual durante o manejo cínico (BRASIL, 2007). Sempre que necessário, quando houver atualizações no caso notificado (geralmente o paciente visita o hospital repetidas vezes), é necessário recuperar o formulário de notificação e atualiza-lo. Apenas no final do processo, quando o caso é devidamente confirmado ou descartado por meio de exames específicos, as informações do formulário são inseridas no Sinan Net. No entanto, os campos do formulário pouco falam sobre os sintomas que acometeram o paciente, o que limita o registro às informações de diagnóstico ou suspeitas. Além disso, com frequência, os campos do formulário são deixados vazios no sistema. Em hospitais onde não há um sistema de gestão dos casos, a mesma notificação pode ser protocolada mais de uma vez, afetando negativamente o procedimento de diagnóstico e a tomada decisão.

2.2 Abordagens Inteligentes

Os Sistemas inteligentes podem usar diversas estratégias para solucionar um determinado problema. Algumas abordagens aplicam inferência indutiva a fim de adaptarem-se a novas situações, enquanto outros métodos utilizam modelos matemáticos baseados em probabilidade para buscar conhecimento em grandes conjuntos de dados. Outro método bastante conhecido é a Aprendizagem de Máquina (AM), que é uma subárea da Inteligência Artificial (IA). Nesta, algoritmos aprendem a partir de experiências, utilizando reconhecimento de padrões a fim de realizar deduções a partir de um conjunto de exemplos (AWAD; KHANNA, 2015).

2.2.1 Aprendizagem de Máquina

A AM é uma área de estudo da IA que dedica-se ao desenvolvimento de algoritmos capazes de aprender. Estes podem resolver dois tipos de problemas: classificação, quando busca-se um resultado discreto; regressão, quando busca-se um resultado contínuo. Para tanto, os algoritmos de aprendizagem devem passar por um processo de treinamento, através da analise de um conjunto de dados, gerando um modelo de aprendizagem para tratar novas instâncias/situações. Este procedimento pode ser feito de maneira supervisionada, por meio de exemplos rotulados; ou não-supervisionada, quando os dados são agrupados de acordo com sua similaridade (TAN et al., 2009). Esse modelo pode ser representado de diversas formas dependendo do método abordado.

Métodos probabilísticos: fazem uso de modelos matemáticos para identificar a disposição dos dados em uma determinada amostra. Os classificadores baseados

no teorema de Bayes são exemplos dessa estratégia, utilizada em larga escala em MD.

Métodos baseados em procura: utilizam modelos baseados em árvores para determinar uma descrição hierárquica dos dados. Árvores de decisão e sistemas adaptativos são abordagens que se encaixam nessa estratégia.

Métodos baseados em regras: utilizam modelos de regras e relacionamentos semânticos que representam um determinado conhecimento. Máquinas de inferência usam esses modelos para gerar novos conhecimentos.

Todas estas abordagens têm evoluído rapidamente e muitos métodos já produzem resultados excelentes em diferentes casos (FACELI et al., 2015). As metodologias baseadas em AM apresentam ótimos resultados em diversas áreas, inclusive em saúde, através de sistemas inteligentes de apoio a decisão (SAD) clínica (STANGE; NETO, 2010).

2.2.2 Classificadores Bayesianos

Os classificadores *Bayesianos* são baseados em premissas estatísticas. Eles calculam a frequência que um evento/resultado ocorre para definir um modelo matemático adequado para predizer um resultado de um novo evento, ainda desconhecido. Eles fazem uso do teorema de Bayes, que calcula a probabilidade de um evento c_i dado um evento x ($P(c_i|x)$). Por exemplo, a probabilidade de um paciente ter dengue dado que ele se encontra com febre, dor nas costas, entre outros sintomas (FACELI et al., 2015). A Equação 2.1 apresenta esse teorema:

$$P(c_i|x) = \frac{P(x|c_i)P(c_i)}{P(x)} \tag{2.1}$$

Onde, $x = (x_1, x_2, ..., x_n)$ representa o conjunto de atributos (sintomas) e $c = (c_1, c_2, ..., c_m)$ as classes (doenças). As probabilidades $P(x_j), P(c_i)$ são as probabilidades a *priori*. Assim, $P(c_i|x)$ é a probabilidade condicionada dos atributos para cada classe e $P(x|c_i)$ a verossimilhança dos novos eventos.

O Naïve Bayes (NB) é um dos classificadores *Bayesianos* mais utilizados em MD. Apesar de usar uma premissa simplista, considerando os atributos independentes uns dos outros, este apresenta bons resultados para casos adequados ao seu contexto. A partir do teorema de Bayes, apresentado na Equação 2.1, podemos des-

considerar o termo $P(x)$, uma vez que será igual para todas as classes, simplificando este teorema para a Equação 2.2:

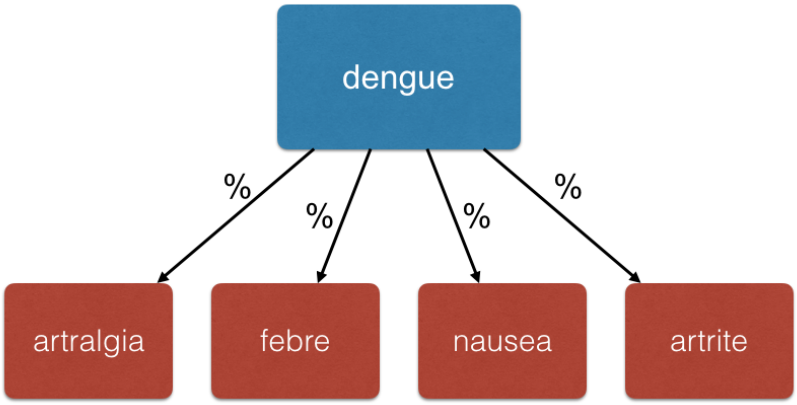
$$P(c_i|x) \propto P(c_i) \prod_{j=1}^{n} P(x_j|c_i) \tag{2.2}$$

Assim, a aplicação do teorema ao contexto de diagnóstico de enfermidades, considerando os sintomas independentes entre si, pode ser escrito como mostra a Equação 2.3.

$$P(Patologia_i|Sintoma) = P(Sint_1|Patol_i) \times \cdots \times P(Sint_n|Patol_i) \times P(Patol_i) \tag{2.3}$$

A Figura 1 representa o classificador NB através de uma estrutura gráfica da relação entre os nós de entrada (sintomas) e de saída (doenças ou patologias). Neste modelo cada doença depende da probabilidade *a priori* do seu conjunto de sintomas.

Figura 1 – Representação do classificador NB.

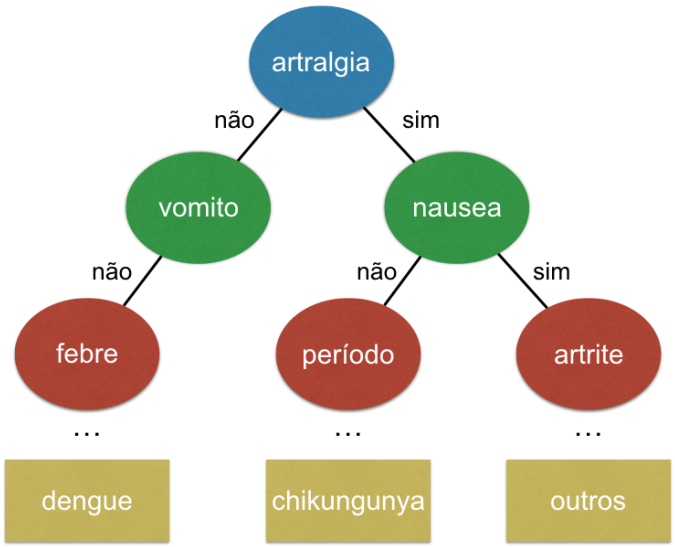


Fonte: Elaborado pelo autor.

2.2.3 Classificadores Baseados em Árvores de Decisão

Árvores de decisão são basicamente estruturas de grafos direcionados, onde os nós podem ser de seguimento ou nós folha. O nó folha possui o valor da classe e o nó seguimento possui a verificação dos valores do atributo. Por exemplo, os sintomas das doenças seriam considerados nós de seguimento enquanto as doenças nós folha. Estas abordagens baseadas em procura buscam alcançar o melhor modelo gráfico representativo possível da experiência observada. A Figura 2 mostra um exemplo de uma árvore de decisão (FACELI et al., 2015).

Figura 2 – Representação gráfica da árvore de decisão.



Fonte: Elaborado pelo autor.

São realizadas diversas funções de busca para se obter tal representação gráfica. A partir das divisões dos valores dos atributos, obtêm-se a quantidade de informação da classe a que este pertence. Por exemplo, o quanto febre está relacionado à dengue. Assim, realizando uma soma ponderada desse conjunto de amostra, é possível descobrir o grau de pureza desse atributo. A Equação 2.4 representa o cálculo do grau de pureza de um determinado atributo.

$$I(N_0) = \sum_{i=1}^{n} \frac{n_i}{N} Entropia(X) \tag{2.4}$$

Onde $X = (x_1, x_2, ..., x_n)$ representa o conjunto dos atributos, n_i o tamanho de X e N o tamanho total da tabela. O termo $Entropia(X)$ mede a variação de uma variável, ou seja, quão difícil é sua predição. Se os atributos estiverem separados em classes distintas, temos uma entropia máxima, caso contrário a entropia é zero. Ou seja, quanto melhor distribuído um sintoma estiver entre dengue e *chikungunya*, mais difícil será sua predição. No entanto, quanto mais um sintoma estiver presente em apenas uma das doenças, melhor sua predição. A Equação 2.5 apresenta a entropia que é dada pelo somatório da probabilidade de uma classe c_j no nó n_i.

$$Entropia(X) = -\sum_{j=1}^{Nclass} p(c_j|n_i) \log_2 p(c_j|n_i) \tag{2.5}$$

O ganho de informação de um atributo X é determinada pela diferença entre a entropia inicial do conjunto de casos (2.4) e a entropia das partições (2.5). A seleção

do melhor atributo está no nó que possui o maior ganho de informação. Definindo o melhor atributo a ser utilizado, a estrutura da árvore é, então, simplificada. Assim, os sintomas que apresentarem melhor entropia estarão no topo da árvore.

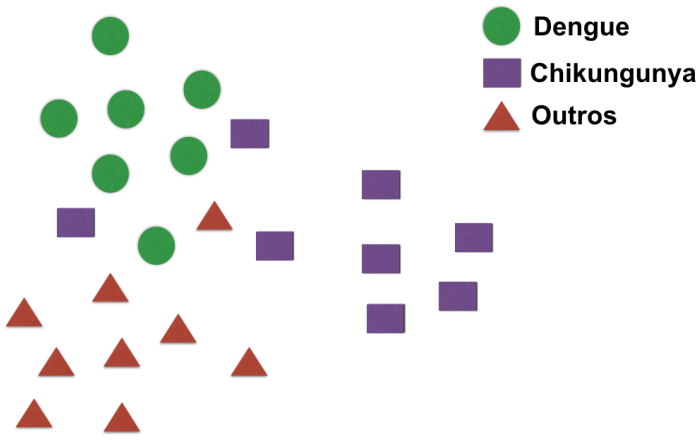
2.2.4 Classificadores Baseados em Distância

As metodologias baseadas em distância são frequentemente utilizadas para distinguir os pontos mais próximos de um determinado conjunto de dados. Os objetos são considerados como pontos definidos pelos seus atributos. Por exemplo, um caso de *chikungunya* irá representar um ponto, dependendo dos sintomas que o paciente apresentar. Essa estratégia parte da premissa que os atributos de uma classe têm valores próximos entre si. Assim, esse processo calcula a distância entre cada ponto no conjunto de dados e os classifica baseado na distância entre eles (FACELI et al., 2015). Existem diversas formas de se obter o cálculo da distância, entre elas, a distância euclidiana, apresentada na Equação 2.6.

$$d(X_i, X_j) = \sqrt{\sum_{l=1}^{d}(x_i^l - x_j^l)^2} \tag{2.6}$$

X_i e X_j representam as classes e x_i^l e x_j^l seus atributos. A partir destes valores é possível identificar um conjunto de objetos com características semelhantes. Desse modo, espera-se que os casos de pacientes que apresentem sintomas específicos agrupem-se em um determinado ponto, com distâncias menores. A Figura 3 ilustra este comportamento. O processo de agrupamento pode ser feito de várias maneiras.

Figura 3 – Gráfico com classes distribuídas.



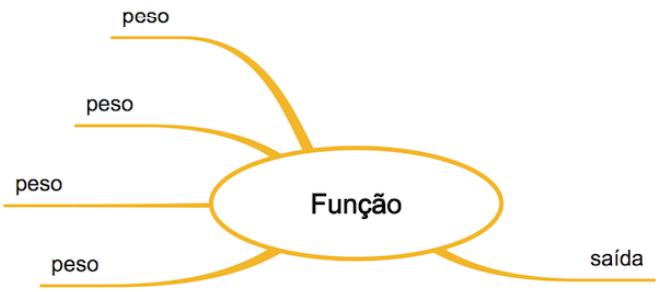
Fonte: Elaborado pelo autor.

O algoritmo dos k vizinhos mais próximos, k *nearest neighbors* (kNN) do inglês, memoriza todos os exemplos classificados em suas posições. Quando um novo objeto surge, o algoritmo calcula a distância entre estes e os objetos já classificados. Então, o novo objeto é classificado com a classe do objeto com menor distância deste.

2.2.5 Classificadores Baseados em Redes Neurais Artificiais (RNAs)

RNAs são modelos baseados no sistema nervoso biológico animal, que conta com uma rede de neurônios fortemente interligados capazes de realizar tarefas e aprender. Eles trabalham de maneira paralela para resolver grandes problemas de maneira distribuída. Assim como um neurônio biológico, um neurônio artificial conta com dendritos, corpo e axônio. Eles são representados pelos pesos, função de ativação e de saída, respectivamente (FACELI et al., 2015). A Figura 4 mostra uma representação de um neurônio artificial. Cada componente exerce uma função específica:

Figura 4 – Neurônio artificial.



Fonte: Elaborado pelo autor.

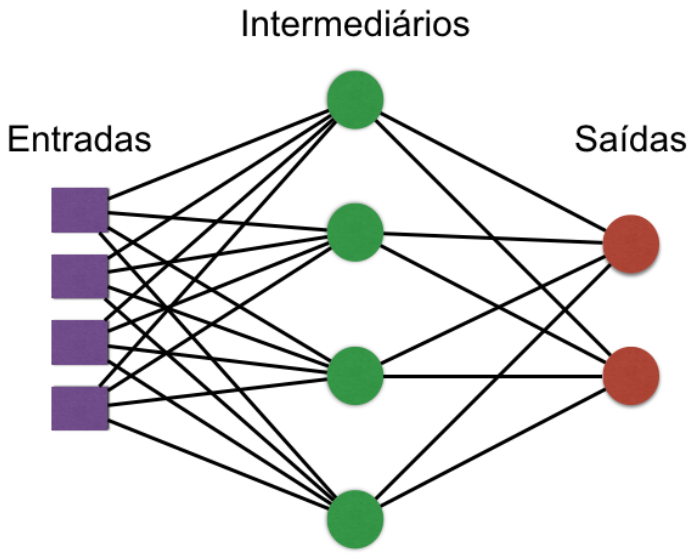
Pesos: são valores que representam a eficácia do acoplamento sináptico entre um neurônio que envia o sinal e o neurônio que o recebe. Quando o peso w de uma sinapse é positivo, a sinapse é dita estimuladora; quando é negativo, a sinapse é dita inibidora. Ou seja, dependo do peso, que é um multiplicador, um valor de entrada pode ser aumentado ou diminuído. Esses valores são ajustados no processo de treinamento.

Função de ativação: recebe os valores ponderados das entradas e os traduz em parâmetros usados para resolver determinada finalidade. Existem diversos tipos de função de ativação, estas ditam a maneira como o neurônio se comporta.

Saída: representa o resultado obtido pela função através dos parâmetros de entrada. Esse valor é levado a todos os neurônios ligados a este, sendo, como toda entrada, ponderado pelo seu peso.

Os neurônios podem estar dispostos em diversas camadas na rede, que podem ter tamanhos diferentes e conexões variadas. A Figura 5 é um exemplo de uma rede com três camadas para o contexto deste trabalho. Para problemas de classificação, geralmente usa-se o número de neurônios referente ao número de classes. A primeira camada deve tratar os atributos de entrada, simplificando-os para camadas intermediárias seguintes até a camada de saída.

Figura 5 – Exemplo de uma RNA com três camadas.



Fonte: Elaborado pelo autor.

2.3 Ferramenta WEKA para MD

WEKA é uma ferramenta gratuita para MD que faz parte de um projeto *open source* mantido pela Universidade de Waikato, Nova Zelândia, desde 1999. A ferramenta destaca-se pelo seu fácil uso e sua vasta gama de funcionalidades. Esta abrange todos os passos no processo de MD: importação dos dados, pré-processamento, treinamento e testes. De modo simplificado, é possível alterar as variáveis e ajustar os algoritmos ao contexto estudado. Também é possível realizar melhoramentos usando filtro nos dados. Os modos gráficos facilitam o estudo e a interpretação. A ferramenta também simplifica a avaliação dos resultados, calculando diversas métricas por padrão na fase de testes. Além disso, dispõe de uma funcionalidade de comparação de resultados entre algoritmos, que facilita o processo de escolha do melhor algoritmo para determinado problema de classificação. A ferramenta ainda dispõe de uma API em java para o incremento de suas funcionalidades em um projeto próprio. A ferramenta já está em sua terceira (3.9) versão e conta com o apoio de grandes instituições

como NIST (National Institute of Standards and Technology) e CERN (European Organization for Nuclear Research). O *software* está disponível gratuitamente na Web, onde também estão disponibilizados diversos cursos *online* sobre MD (FRANK; HALL; WITTEN, 2016).

3 TRABALHOS RELACIONADOS

Existem na literatura diversas aplicações para cada uma das abordagens e estratégias de classificação de dados apresentadas no capítulo anterior. A IA desenvolveu-se rapidamente e sua aplicação em diversas aéreas do conhecimento tem resolvido diversos problemas de diferentes níveis de complexidade. Esse capítulo discute trabalhos relacionados ao uso de classificadores e outros métodos de predição em saúde e especificamente no contexto das doenças transmitidas pelo vetor Aedes Aegypti.

No trabalho de (MOREIRA et al., 2016a) foi desenvolvida uma rede *baysiana* para classificar desordens hipertensivas focando no cuidado da pré-eclâmpsia. Esta pesquisa usa redes Bayesianas para dar suporte à tomada de decisão em ambientes de incerteza no cuidado com a Gravidez. Usando o modelo Bayesiano *Noisy*-OR em uma base de dados de saúde, este modelo analisa a disposição dos dados e os classifica na rede. A partir dos sintomas apresentados pela gestante, o sistema infere a gravidade do caso por meio de dados estatísticos, ajudando o médico especialista na predição da pré-eclâmpsia. Esta abordagem mostrou-se precisa mesmo com um número pequeno de dados. Assim como a pré-eclâmpsia, o diagnóstico de dengue e *chikungunya* é incerto e complexo. Por isso ambos os trabalhos mostram-se relevantes no cenário de apoio à decisão.

Em (MOREIRA et al., 2016c), os autores fazem uma comparação entre o classificador NB e o classificador baseado em árvore de decisão J48. O trabalho analisa um conjunto de dados relacionados a distúrbios hipertensivos para avaliar complicações na gravidez. O trabalho faz um estudo do desempenho dos classificadores a partir de uma matriz de confusão, usando parâmetros preditivos. Embora os dois classificadores apresentem valores próximos, os resultados mostram que o algoritmo de árvore de decisão J48 é o classificador com melhor precisão para essa situação. Apesar de cenários diferentes, os classificadores baseados em árvore de decisão mostram-se mais precisos que os baseados em estatística. A partir de análises, é notável que a estratégia apresenta melhores resultados em caso de atributos de grande complexidade.

No trabalho apresentando por (SILVA et al., 2017) foi desenvolvido um sistema capaz calcular o risco de um recém nascido vir a óbito. Para tanto foi construído um mecanismo inteligente baseado em classificadores que é capaz de inferir a probabilidade para "sim"ou "não"relativo a possibilidade de óbito de um recém nascido. O trabalho segue uma metodologia bem definida de reconhecimento de padrões proposta por (RAMOS et al., 2016). Foi usada a bem conhecida ferramenta WEKA para apoiar os passos da metodologia. Os algoritmos baseados em probabilidade apresen-

taram melhores resultados nos testes, destacando o classificador NB, que apresentou acurácia de 60,7% e área ROC de 92,1%. O objetivo do trabalho foi propor a inclusão de alertas inteligentes no GISSA, uma plataforma para Governança Inteligente em Sistemas de Saúde, implantado na Rede Cegonha, que visa preservar a saúde da gestante e do recém-nascido, no município de Tauá, CE, Brasil. A prova de conceito foi desenvolvida em JAVA, usando a própria API do WEKA.

Thanathornwong *et al.* fazem uso de um sistema para previsão de resultados após um procedimento de clareamento dentário. Aplicando uma equação de regressão múltipla em um conjunto de dados de coordenadas de cor CIELAB, antes e depois do procedimento, pode-se predizer o resultado para novos casos com precisão (THANATHORNWONG; SUEBNUKARN; OUIVIRACH, 2016). O trabalho também discute outras diversas aplicações que usaram abordagens diferentes para prever resultados na área. Apesar de não fazerem comparações e análises de algoritmos, o trabalho realiza uma investigação para descobrir quais atributos entre diversos sistemas de cores são mais relevantes para predição.

Santos *et al.* fazem uma análise de dados dos casos de dengue ocorridos na cidade de Recife, PE, Brasil, no ano de 2016, para construir um sistema capaz de aprender quando um paciente tem dengue ou não (SANTOS, 2016). O modelo testa diversas abordagens de classificação e AM, entre estas estão abordagens baseadas em árvores de decisão, o classificador NB, Máquina de Vetores Suporte (MVS), RNAs, entre outras. Os resultados mostram que, ao final dos testes, os algoritmos MVS, Perceptron de Multicamadas (PMC) e o classificador baseado em árvore de decisão J48 foram considerados promissores, apresentando taxas de média harmônica (F-Measure) superior a 0.885. Apesar dos bons resultados, o trabalho trata uma entrada fixa de atributos, que exigem resultados de exames mais específicos.

O trabalho de Teles *et al.* apresenta um SAD voltado ao diagnóstico e identificação de gravidade da dengue. O sistema faz uso de redes Bayesianas para auxiliar no diagnóstico em casos de incerteza (TELES et al., 2014). O modelo proposto analisa dados do usuário (sintomas) e infere sobre o seu risco, a saber, baixo, médio ou alto. O trabalho é um componente do *framework* LARIISA, discutido em (GARDINI et al., 2013). Este sistema conta com outros mecanismos de decisão em sua interface. Apesar de não testar outras abordagens, o trabalho apresenta bons resultados.

Em (AYYAZ et al., 2015), os autores usam modelos matemáticos para a simulação de epidemias com o objetivo de criar medidas preventivas para combater doenças com características epidêmicas. Os modelos propostos neste trabalho conseguem identificar particularidades na propagação de doenças e predizer com moderada eficiência onde estas ocorrerão. O estudo propõe um modelo matemático de propagação para o mosquito *Aedes Aegypti*. Diferente dos trabalhos de classificação

voltados ao diagnóstico, este está focado ao cenário da prevenção epidêmica.

A abordagem proposta em (ALVES; GADELHA, 2016) é baseada em mecanismos de representação do conhecimento. O trabalho também é focado no pré-diagnóstico de doenças transmitidas pelo mosquito *Aedes Aegypti*. Neste são realizadas pesquisas sobre os sintomas das doenças, seus relacionamentos e peculiaridades para definir heurísticas e representar o conhecimento dos profissionais de saúde. A partir da modelagem, implementou-se uma ferramenta de apoio a tomada de decisão capaz de inferir a probabilidade de um paciente estar infectado com dengue, *chikungunya* ou vírus *zika* pelos sintomas apresentados.

O trabalho apresentado em (OLIVEIRA et al., 2013) realiza um estudo como prova de conceito do projeto LARIISA (Laboratory of Intelligent and Integrated Networks Applied to Health System) aplicado ao cenário de dengue. O sistema utiliza ontologias e conceitos de *Context Awarenness* para identificar áreas de risco de surto da doença. Tais informações servirão tanto de suporte para tomada de decisão de gestores quanto para alertar e prevenir a população. Para tanto, o sistema conta com diversos atores, entre eles, o smartphone e TV Digital.

Em (CARDOSO, 2015), os pesquisadores propõem um sistema inteligente baseado em ontologias capaz de determinar áreas com risco de infecção. A proposta coleta notificações de casos de forma colaborativa e analisa os dados através de modelos inteligentes. Aplicando eurísticas, o sistema consegue prever uma região com alta probabilidade epidêmica antes que esta aconteça. Assim, o sistema consegue colaborar com gestores no processo de tomada de decisão e ajudar usuários com alertas de áreas de risco. O COISA, como é conhecido, é um módulo do *framework* LARIISA (GARDINI et al., 2013), um modelo para gestão de saúde. O sistema foi criado para qualquer tipo de notificação, mas melhor se adéqua ao contexto de doenças endêmicas. A estratégia colaborativa traz agilidade aos sistemas com a confiabilidade dos dados, que podem ser fornecidos por qualquer usuário.

Oliveira faz uso de dados genômicos para classificação de formas clínicas de dengue. O trabalho faz um estudo aprofundado dos aspectos de polimorfismos genéticos ao invés de tratar dados clínicos ou laboratoriais (OLIVEIRA, 2009). Os dados foram obtidos de 105 pacientes da coorte de dengue do LaviTE. A base conta com 26 casos de Febre Hemorrágica da Dengue, 49 casos de Dengue Clássica Complicada e 30 casos de Dengue Clássica. Os dados foram obtidos por meio da aplicação de técnicas de genotipagem em massa (Illumina). O modelo PMC, que é baseado em RNA, classifica os casos de dengue severa com acurácia de 85%.

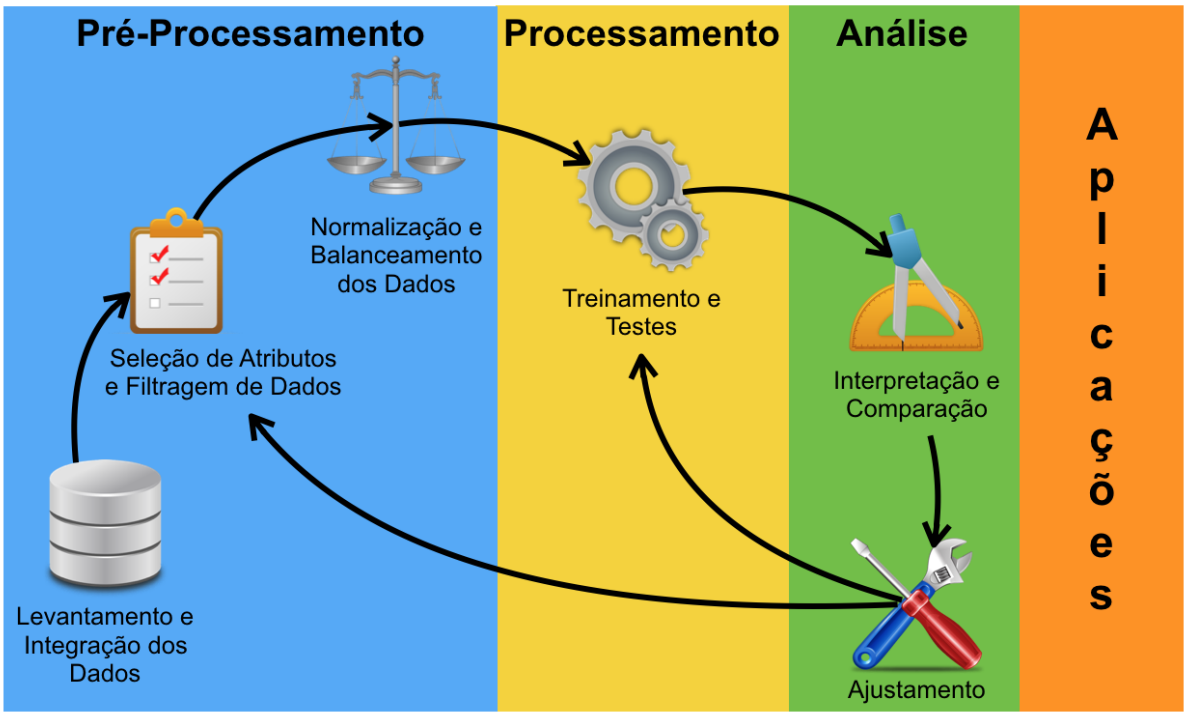
Baseado em diversos trabalhos recentes da literatura, o estudo proposto neste trabalho monográfico estende a pesquisa publicada em (BRAGA et al., 2017), que propôs uma aplicação móvel capaz de classificar casos de dengue e *chikungunya*

usando métodos de AM. O trabalho analisou dados de casos diagnosticados das doenças em questão para treinar algoritmos de aprendizagem e prever casos de risco. A proposta anterior focou o primeiro atendimento, avaliando apenas os sintomas das doenças em questão e algumas doenças pré-existentes.

4 METODOLOGIA ADOTADA

Diversas metodologias já foram propostas para melhorar a avaliação de classificadores preditivos e maximizar os resultados da Mineração de Dados (MD) (FAYYAD et al., 1996). Algumas destas técnicas propõem procedimentos iterativos que apresentam sequências de passos recorrentes para alcançar resultados mais precisos, como a fase de testes e ajustamento (RAMOS et al., 2016). Entretanto, na maioria das vezes, cabe ao cientista de dados decidir o momento de parada da iteração com o procedimento. Este trabalho apoia-se numa metodologia simplificada baseada em metodologias clássicas de mineração de dados. A Figura 6 mostra a sequência de etapas da metodologia proposta neste trabalho, a qual foi dividido em três fases:

Figura 6 – Etapas da metodologia adotada para a avaliação de classificadores.



Fonte: Elaborado pelo autor.

Pré-processamento: Nesta fase são realizadas as etapas que precedem a execução dos algoritmos. Nela, os dados são levantados, tratados e preparados para serem, então, processados.

Processamento: Esta fase contém a etapa principal de preparação para o processo de análise. Nesta etapa, os dados são processados pelo algoritmo, gerando um modelo de aprendizagem que posteriormente é avaliado.

Análise: É a fase da metodologia que mais necessita da experiência do cientista de dados. A partir da análise dos dados, pode-se entender seu comportamento e ajustar os algoritmos para melhor tratá-los.

Aplicações: A partir dos modelos de aprendizagem gerados é possível construir softwares capazes de inferir novos conhecimentos. É possível ainda melhorar tais modelos com entrada de mais exemplos.

4.1 Pré-Processamento

Este trabalho atende diversas etapas do processo do manejo clínico das doenças em questão. Assim, considerou-se toda informação utilizada por profissionais de saúde como de alta relevância, incluindo sintomas, sinais e exames.

4.1.1 Levantamento e Integração dos Dados

Os algoritmos de Aprendizado de Máquinas (AM) analisam uma amostra de exemplos para modelar um conhecimento. Estes modelos podem ser representados de diversas formas, ou seja, modelos baseados em estatística, em árvore de decisão ou em regras semânticas. Os métodos de AM representam o conhecimento adquirido a partir do conjunto de experiências observadas. Assim, quanto maior a equivalência da amostra em relação à população (todos os exemplos) melhor o modelo. Portanto, quanto mais casos reais das doenças em questão, melhor será o alcance do algoritmo. No entanto, deve-se tomar cuidado para que não ocorra sobreajuste (do inglês overfitting), quando há muitos exemplos semelhantes, o que capacita os algoritmos apenas aqueles casos (amostra), apresentando resultados ruins quando classificar casos diferentes. Por isso é ideal que os exemplos sejam os mais genéricos (diferentes entre si) possíveis, dando uma visão mais completa aos algoritmos de aprendizado.

Para se obter a maior quantidade de dados possível, esta pesquisa realizou uma busca aprofundada em bases de dados públicas digitais e visitas a hospitais e secretarias de saúde Municipais. No entanto, as bases físicas mostraram-se limitadas e de difícil análise. Além de apresentarem mau estado e incompletude nos campos de preenchimento, ou seja, formulários que não descrevem os sintomas detalhadamente, entre outros, o que dificulta o processo de coleta. Apesar de grande esforço, foi impossível recuperar os poucos dados adquiridos dos hospitais locais, principalmente por não contarem com prontuários digitais ou registro eletrônico de saúde.

Também foram analisados dados do SINAN. No entanto o sistema trata apenas dados de importância endêmica, mas desconsidera dados fisiopatológicos ou se-

miológicos, indispensáveis para o diagnóstico, objetivo desse trabalho. O sistema inclui apenas dados do diagnóstico suspeito e diagnóstico final da doença, despresando os sintomas e/ou sinais clínicos que levaram a esta conclusão. Assim, este estudo optou por não utilizar os dados do SINAN, apesar deste contar com um extenso conjunto de informações.

Ao final, a grande maioria dos dados foram obtidos a partir do portal de dados abertos da prefeitura de Recife, PE, Brasil, disponível em (RECIFE, 2016). Neste portal, estão disponíveis casos de dengue e *chikungunya*. Ao todo foram extraídos 20.137 casos, sendo 10.513 dengue e 1.274 *chikungunya*. 4.713 foram removidos por apresentarem dados faltantes. Alguns casos foram classificados como inconclusivos e também foram removidos por falta de informação. Os casos descartados, não sendo dengue nem *chikungunya* foram rotulados como outros, podendo ser qualquer outra doença.

4.1.2 Seleção de Atributos e Filtragem dos Dados

Os dados levantados contam com uma série de atributos clínicos e laboratoriais, que foram extraídos durante o processo de manejo clínico das doenças. Entre eles estão os sintomas apresentados e o histórico de saúde do paciente. Além disso, existe também resultados de exames mais específicos, que exigem procedimentos técnicos ou invasivos para serem obtidos. A tabela 2 mostra os sintomas presentes nos dados coletados. Os nomes técnicos foram convertidos para facilitar o entendimento.

Tabela 2 – Principais sintomas apresentados pelos pacientes.

Sintomas
Febre
Náusea
Vômito
Artrite
Conjuntivite
Dor de cabeça
Dor nas costas
Dores musculares
Artralgia intensa
Dor ao redor dos olhos
Manchas vermelhas na pele
Pontinhos vermelhos na pele

Os dados sobre o histórico de saúde do paciente são de vital relevância para o diagnóstico de uma doença. Isto porque, dependendo das doenças que o paciente tenha, os sintomas podem apresentar diferentes comportamentos. Além disso, os sintomas podem estar relacionados a tais doenças, o que pode confundir o diagnóstico.

Por tanto, considerou-se também como atributos algumas doenças pré-existentes. Estas são mostradas na Tabela 3.

Tabela 3 – Doenças pré-existentes.

Doenças
Diabetes
Doenças no sangue
Doença no fígado
Doença renal
Hipertensão
Doença no estômago
Doenças auto imunes

A coleta de exames é uma etapa importante no processo de manejo clínico. Os resultados dos exames agregaram valor ao diagnóstico, tornando-o mais preciso. Desta forma, este estudo considerou o resultado de exames no modelo proposto neste trabalho. Alguns exames mais específicos, como o de sorologia por exemplo, ainda são muito caros ou demorados e nem sempre estão disponíveis na rede pública de saúde (ALVES; GADELHA, 2016). Por isso, os profissionais de saúde solicitam primeiramente exames mais imediatos com o objetivo de ter um pré-diagnóstico do problema com mais rapidez, descartando algumas hipóteses dependendo do caso. A Tabela 4 mostra estes exames.

Tabela 4 – Lista de exames solicitados durante o manejo clínico das doenças.

Exames
Teste do Laço
Hemograma (Leucopenia)
Chikungunya soro 1
Chikungunya soro 2
Exame PRNT
Dengue sorológico
Exame ELISA
Isolamento viral
Exame PCR

Com o objetivo de destacar as estações chuvosas do ano, este estudo separou o campo "data"em quatro períodos distintos: janeiro à março; abril à junho; julho à setembro; outubro à dezembro. Assim, as primeiras semanas do ano, quando ocorrem mais casos das doenças em questão, são organizadas em um único grupo para facilitar o processo de identificação de padrões.

A simplificação dos dados pode tanto acelerar o processamento quanto melhorar os resultados em alguns casos. Isto ocorre porque alguns atributos podem dificultar o processo de aprendizagem, confundindo os alguns algoritmos. Assim,

aplicou-se também um mecanismo de seleção de atributos automatizado, que consiste no truncamento de atributos menos relevantes para realçar aqueles com maior significância (HALL; HOLMES, 2003). Alguns algoritmos apresentaram melhores resultados após a seleção de atributos, como o classificador NB, outros tiveram sua acurácia prejudicada. Portanto, este trabalho aplicou a seleção de atributos apenas aos casos onde houve melhoria.

Ao todo foram tratados 32 atributos de várias etapas do processo de manejo clínico, considerando apenas os casos que continham ao menos os sintomas devidamente preenchidos. Nenhum dos casos apresentou todos os resultados de exames específicos preenchidos. Isso ocorre porque os profissionais de saúde, logo de início, trabalham com uma hipótese de diagnóstico, solicitando exames apenas pra um tipo de suspeita. Essa falta de dados dificulta o aprendizado por parte dos algoritmos, que foram modelados para considerar estes exames apenas como não realizados.

4.1.3 Balanceamento e Normalização dos Dados

Os algoritmos de classificação aprendem por meio da análise de um conjunto de experiências. No entanto, se um algoritmo aprender mais sobre uma determinada experiência (classe) do que outra, ele tenderá sua classificação à ela. Assim, dependendo do problema, não é interessante ter uma base de dados desbalanceada. No problema de classificação de doenças não pode haver tendenciamento, pois é tão importante classificar tanto uma doença quanto a outra. A etapa de balanceamento de dados é realizada, geralmente, antes da seleção de atributos e da limpeza dos dados. Mas, depois de algumas observações, percebeu-se que os dados trucados no processo de limpeza podem desequilibrar ainda mais as classes. Portanto, esta pesquisa optou por adiar essa etapa do processo para a última etapa do pré-processamento. A base de dados tratada na primeira fase (suspeição) conta com 1.133 casos de dengue, 1.273 casos de *chikungunya* e 1.624 casos não identificados, considerados de outras doenças. Desses foram selecionados 208 casos de dengue, 69 casos de chikugunya e 536 casos de outras doenças para fase de diagnóstico, que incluem resultados de exames. Portanto, para se obter uma melhor classificação, realizou-se um balanceamento dos dados. Dois algoritmos de balanceamento foram aplicados para os testes:

SMOTE (Synthetic Minority Over-sampling Technique): Realiza interpolação entre exemplos próximos das classes minoritárias, criando exemplos sintéticos para essas classes. A técnica só atinge uma classe por vez, necessitando processar várias vezes caso haja mais de uma classe desbalanceada (CHAWLA et al., 2002).

Resample: Como o próprio nome diz, é uma técnica baseada em re-amostragem. O algoritmo realiza o balanceamento através da replicação (cópia) de alguns exemplos, que podem influenciar quaisquer das classes (majoritárias e minoritárias), dependendo da configuração.

Apesar de usarem estratégias diferentes, os dois métodos de balanceamento apresentaram bons resultados. Tomou-se cuidado para que não haja superajustamento, evitando criar ou duplicar exemplos sintéticos, usando os seguintes algoritmos com moderação. A Tabela 5 mostra os melhores balanceamentos alcançados pelos algoritmos. As diferentes técnicas influenciaram significativamente os resultados dos algoritmos de classificação, afetando-os positiva e negativamente, depende do caso. Assim, este trabalho optou pelo método que apresentou melhor resultado para o algoritmo testado.

Tabela 5 – Resultado do balanceamento.

	Dengue	**Chikungunya**	**Outros**
Suspeição			
SMOTE	1.642	1.654	1.624
Resample	1.586	1.603	1.645
Diagnóstico			
SMOTE	536	483	520
Resample	595	501	529

Buscou-se atingir um limite entre classes melhor balanceadas e menos uso dos algoritmos de balanceamento. Após diversos testes, alcançou-se as seguintes configurações. Para primeira faze aplicou-se o SMOTE duas vezes, com 45% e 30%, atingindo uma classe desbalanceada por vez. No *Resample* aplicou-se uma proporção entre classes de 0.9 e potência de 120%, atingindo as duas classes desbalanceada simultaneamente. Já na segunda etapa aplicou-se SMOTE de 600% e 150% e *Resample* em 0.9 com 200%.

4.2 Processamento

Para obter um modelo inteligente capaz de classificar novos casos das doenças em questão é necessário realizar o procedimento de treinamento de algoritmos. No entanto, para cada abordagem de AM, existe uma vastidão de algoritmos capazes de classificar dados. Alguns apresentam resultados satisfatórios para alguns contextos, mas perdem sua eficácia em outros. Assim, existe um classificador mais apropriado para cada situação. Portanto, para identificar qual classificador melhor de adapta ao conjunto de dados deste trabalho, realizou-se um procedimento de testes e comparação de algoritmos, de tal forma que os algoritmos são treinados, testados

e comparados. Cada modelo proposto trata os dados de maneira particular, como mostrado no Capítulo 2. Assim, cada algoritmo traz uma forma diferente de ajustamento. Testar com precisão, ajustando cada algoritmo ao contexto abordado, torna-se então, uma tarefa dispendiosa e demorada. Então, para filtrar os algoritmos mais adequados para o problema, este estudo realizou uma busca por trabalhos relacionados ao pré-diagnóstico de doenças em momentos de incerteza. Para tanto, foram usados os seguintes parâmetros: *uncertainty*; *disease*; e *classifiers*. Os resultados destacaram alguns classificadores. Entre estes estão o J48, NB, Random Forest (RF) e BN (WEBB, 2011; BRADLEY, 1997; MOREIRA et al., 2016b). Além disto, este estudo atentou para trabalhos mais recentes, que trataram algoritmos mais modernos e com melhores avaliações.

4.2.1 Treinamento e Testes

A etapa de treinamento consiste em submeter um conjunto de experiências ao algoritmo para capacitá-lo à novas situações. Quanto mais diferentes forem as experiências, mais genérico será o modelo e melhor será seu resultado em situações diversas. A etapa de testes consiste em submeter novos casos rotulados ao modelo treinado para comparar os resultados da classificação ao seu rótulo real. Esse procedimento pode ser feito separando o conjunto de dados em duas partes. Uma dedicada ao treinamento e outra aos testes. Como alternativa, existe o teste de validação cruzada, que consiste em dividir o banco de dados em n subconjuntos e selecionar um destes para teste e o restante para treinamento. Este procedimento é realizado n vezes, sendo que cada conjunto é separado para teste uma vez. Esse procedimento foi desenvolvido por (BROWNE, 2000) e é largamente utilizado em testes de validação. Para esse trabalho, usou-se o teste de validação cruzada com 10 partes. Ou seja, o conjunto de dados foi dividido em 10 partições, treinado e testado separadamente. Após os testes, o procedimento gera uma matriz com os casos corretos e errados, que posteriormente serão analisados para fornecerem informações relevantes. Esta matriz é conhecida como matriz de confusão, que cruza os valores classificados com os valores reais. Ela será melhor explicada no capítulo seguinte.

4.3 Análise

Cada algoritmo foi testado diversas vezes usando vários ajustes para obter os melhores resultados possíveis para determinado método. Para tanto, é necessário entender o funcionamento do algoritmo e analisar os resultados preliminares com cuidado.

4.3.1 Interpretação e Comparação

Os algoritmos podem gerar diversas saídas nas etapas de treinamento e testes. Por exemplo, algoritmos de árvores de decisão geram uma árvore na fase de treinamento. Esta árvore pode explicar os padrões encontrados nos dados ou destacar alguma anomalia. A matriz de confusão, gerada na fase de testes, também ajuda a explicar o comportamento dos resultados. A comparação pode comprovar se os ajustes estão melhorando ou piorando os resultados do algoritmo em relação ao objetivo. A partir dos resultados também podem ser geradas diversas métricas de avaliação, que dão suporte ao processo de interpretação e ajustamento. Essas métricas serão detalhadas no Capítulo Resultados.

4.3.2 Etapa de Ajuste

Além do conjunto de exemplos, a maioria dos algoritmos recebe alguns parâmetros de entrada. Tratam-se dos valores de ajustamento, que ditarão a forma como o algoritmo irá se comportar. Esses valores são inseridos ou alterados manualmente e podem influenciar significativamente os resultados de um teste. Portanto, a cada rodada de testes, após a interpretação dos resultados, atualizam-se os valores de ajustamento para alcançar melhores resultados.

5 ANÁLISE E INTERPRETAÇÃO DOS RESULTADOS

A análise de diversas abordagens de classificação é essencial no processo de predição e MD. A comparação de algoritmos aplicados a problemas particulares se mostra indispensável em qualquer prática de AM. A singularidade dos dados impossibilita uma dedução eficaz pois os algoritmos podem se comportar de maneira diferente em casos específicos. Encontrar qual classificador apresenta melhores resultados para um conjunto de dados é primordial para o sucesso de um sistema. Portanto, analisar e interpretar os resultados alcançados pelos testes dos algoritmos faz-se necessário.

5.1 Algoritmos Testados

Novas propostas de algoritmos de classificação têm surgido recentemente. Alguns trabalhos sugerem aperfeiçoamentos específicos nos algoritmos, melhorando os resultado dos classificadores para determinados contextos. Contudo, o contexto dos problemas de classificação diferem muito entre si. O comportamento e característica dos dados influenciam muito no processo de aprendizagem. Alguns algoritmos tratam melhor alguma especificidade nos dados enquanto outros apresentam melhores resultados em contextos diferentes. Por exemplo, um determinado atributo pode ajudar na predição de um classificador enquanto atrapalha noutro. Portanto, este trabalho selecionou os classificadores mais utilizados no cenário de auxílio a tomada de decisão em ambientes de incerteza.

***Bayes Network (BN)*:** É um classificador probabilístico baseado no teorema de Bayes. As redes Bayesianas, como são conhecidas, criam uma rede de interdependências entre as probabilidades (*a priori* e *a posteriori*), tratando os atributos de maneira hierárquica.

***Naïve Bayes (NB)*:** Como visto anteriormente, trata-se de um classificador baseado no teorema de Bayes que calcula a probabilidade de um evento particular acontecer dado um conjunto de eventos. Diferente do classificador BN, este considera todos os atributos independentes entre si.

***Random Tree (RT)*:** É um classificador baseado em árvores de decisão que sorteia *k* atributos em cada nó sem realizar podas, gerando árvores randômicas sem treinamento.

***Random Forest (RF)*:** O classificador RF gera diversas árvores aleatórias usando algoritmos diferentes. Depois escolhe-se a que melhor se adaptou aos dados, apresentando melhores resultados.

J48: É uma reimplementação do algoritmo C4.5 (QUINLAN, 1993), que seleciona a melhor partição dos nós a fim de obter melhores resultados. O algoritmo também realiza uma poda das subárvores que não apresentam ganho de informação.

***Support Vetor Machine (SVM)*:** É um classificador linear binário não probabilístico que, dado um conjunto de dados com duas classes, procura separá-los linearmente para a classificação.

***Multilayer Perceptron (MLP)*:** É um classificador baseado em RNAs com ao menos três camadas: entrada, intermediárias, saída. Seus neurônios usam funções de ativação não lineares, treinados a partir de um algoritmo baseado em *backpropagation*.

5.2 Métricas de Avaliação

A performance de um algoritmo é caculada através de métricas de avaliação que se baseiam primariamente na matriz de confusão, que relaciona os valores dos dados com os resultados inferidos pelos algoritmos. A Tabela 6 e a Figura 7 apresentam estas métricas.

Tabela 6 – Matriz de Confusão

		Classificados	
		Positivos	Negativos
Reais	Positivos	VP	FN
	Negativos	FP	VN

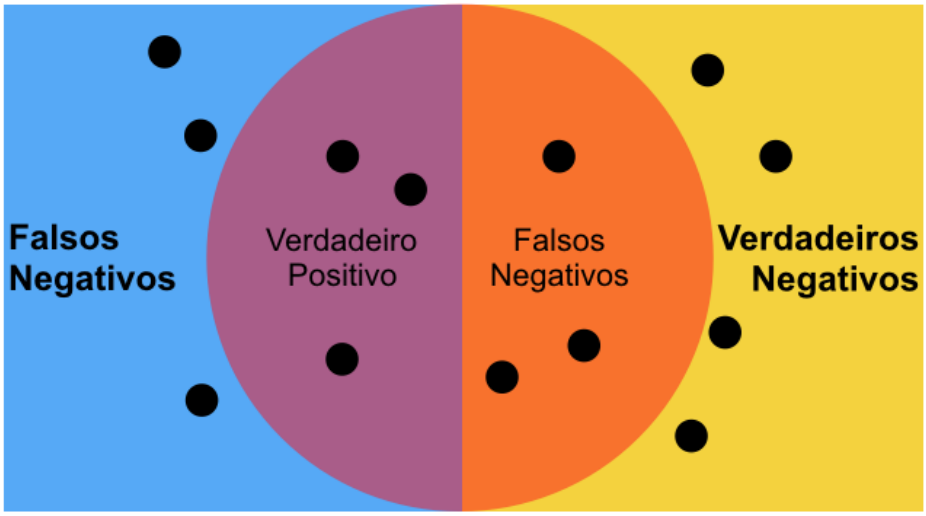
VP (Verdadeiros Positivos): São os casos positivos que realmente foram classificados como positivos.

VN (Verdadeiros Negativos): São os casos negativos que foram corretamente classificados como negativos.

FP (Falsos Positivos): São os casos negativos que foram classificados como positivos (alarmes falsos).

FN (Falsos Negativos): São os casos positivos que foram classificados incorretamente como negativos.

Figura 7 – Representação Gráfica da Matriz de Confusão



Fonte: Elaborado pelo autor.

A avaliação de um algoritmo é realizada por meio da análise de métricas construídas a partir da matriz de confusão. Tais métricas mostram quão bom é um algoritmo em relação ao problema abordado. Dependendo do problema, prioriza-se mais uma métrica que outra. Isto ocorre porque cada métrica mede características diferentes (AWAD; KHANNA, 2015). Por exemplo, para um problema menos crítico, onde o que importa é saber quantas vezes o algoritmo certou, atenta-se apenas para os casos acertados, desconsiderando os casos erroneamente classificados. Em problemas considerados complexos, procura-se analisar os resultados de maneira mais panorâmica. As métricas mais conhecidas são:

Precisão: A proporção de predições corretas, sem levar em consideração o que é positivo e o que é negativo.

$$Prec. = \frac{VP}{VP+FN} \tag{5.1}$$

Cobertura: A proporção de verdadeiros positivos. A capacidade do sistema em predizer corretamente a condição para casos que realmente a tem.

$$Cob. = \frac{VP}{VP+FP} \tag{5.2}$$

Média Harmônica: A média harmônica, também conhecida como Medida-F, é uma medida de desempenho largamente utilizada em tarefas de previsão. Combinando a precisão com a cobertura, ela evita desvantagens de métricas simples, como a taxa de erro, especialmente nos casos de distribuições de classes desequilibradas (BUSA-FEKETE et al., 2015).

$$Medida - F = 2 \times \frac{prec. \times cob.}{prec. + rec.} \tag{5.3}$$

A análise das métricas e matriz de confusão é uma etapa sensível do processo, pois exige uma maior percepção do cientista de dados. A interpretação desses valores leva a identificação de anomalias que atrapalham o treinamento dos algoritmos, o que conduz a melhores ajustamentos.

5.3 Resultados dos Algoritmos

A fim de atender os vários níveis do processo de manejo clínico, este trabalho considerou dividir o treinamento dos algoritmos em duas partes. Uma para atender as primeiras etapas do manejo, que incluem os sintomas e alguns exames simples e outra para atender o diagnóstico final, incluindo exames mais específicos. A primeira etapa tem o objetivo de auxiliar na tomada de decisão durante as primeiras etapas do manejo clínico, suspeição, e a segunda é focada no diagnóstico final.

5.3.1 Suspeição

Durante a etapa de processamento foram realizados diversos testes com os algoritmos, que posteriormente foram analisados a fim de melhorar seus resultados. Os melhores resultados alcançados na etapa de análise são mostrados a Tabela 7.

Tabela 7 – Melhores Resultados.

Algoritmo	Precisão	Cobertura	Méd. Harm.
BN	61.3	61.3	61.2
NB	60.5	59.6	59.2
J48	66.4	66.4	66.4
RT	68.5	68.5	68.4
RF	**69.3**	**69.3**	**69.3**
MLP	65.3	65.0	65.1
kNN	68.4	68.4	68.4
SVM	62.8	60.9	60.4

Os algoritmos baseados em árvore de decisão apresentaram melhores resultados para o conjunto de dados deste trabalho. Todos os algoritmos apresentaram melhores resultados usando o balanceador *Resample* em relação ao SMOTE. A tabela 8 mostra os resultados alcançados em cada classe/doença pelo algoritmo RF.

A média harmônica dos critérios de análise apresentaram bons resultados para o problema de classificação das doenças. Os resultados não apresentam muita discrepância entre as métricas. Percebe-se na matriz de confusão, apresentada na Tabela 9, que os casos erroneamente classificações são balanceados entre as doenças.

Tabela 8 – Resultados das métricas de desempenho obtidos a partir da matriz de confusão para o classificador RF.

Classe	Precisão	Cobertura	Méd. Harm.
Dengue	66.9	67.4	67.2
Chikungunya	71.0	68.8	69.9
Outros	69.9	71.6	70.7

Tabela 9 – Matriz de confusão do classificador RF.

		Classificados		
		Dengue	*Chikungunya*	Outros
Reais	Dengue	1069	223	284
	Chikungunya	278	1103	222
	Outros	250	218	1177

Os sintomas apresentados pelas doenças são muito semelhantes e se confundem entre si. Os algoritmos classificam alguns casos erroneamente, como mostrado pela matriz. No entanto os resultados mostram-se relevantes com relação a proposta do sistema.

5.3.2 Diagnóstico

Com o objetivo de suportar o diagnóstico final, incluiu-se exames mais específicos no treinamento dos algoritmos. Nessa etapa foram usados apenas casos analisados em laboratório, despresando os casos sem nenhum exame específico. A tabela 10 mostra os melhores resultados alcançados pelos algoritmos.

Tabela 10 – Melhores resultados para os classificadores propostos usando técnicas de balanceamento.

Algoritmo	Precisão	Cobertura	Méd. Harm.	Balanceamento
BN	87.9	86.3	86.0	SMOTE
NB	68.6	68.2	65.0	SMOTE
J48	88.5	88.5	88.4	SMOTE
RT	90.4	90.5	90.4	Resample
RF	90.8	90.9	90.8	Resample
MLP	**91.5**	**91.6**	**91.5**	**Resample**
kNN	80.9	77.8	77.6	Resample
SVM	61.6	59.8	58.9	SMOTE

Diferente da etapa de suspeição, os resultados desta etapa mostram-se muito mais satisfatórios. Os exames dão mais certeza ao diagnóstico. Nesse caso, alguns algoritmos apresentaram melhores resultados usando o método de balanceamento SMOTE, principalmente os probabilísticos. Os classificadores baseados em árvore de decisão continuam apresentando bons resultados. No entanto, o algoritmo baseado

em RNAs, nomeadamente PMC, superou todos os outros algoritmos nesta fase. A Tabela 11 apresenta os resultados dos testes para cada classe/doença deste classificador.

Tabela 11 – Resultados obtidos a partir dos indicadores da matriz de confusão para o classificador baseado em RNAs PMC.

Classe	Precisão	Cobertura	Méd. Harm.
Dengue	87.1	89.2	88.1
Chikungunya	97.7	100.0	98.8
Outros	87.1	89.2	88.1

Pôde-se perceber que, nessa etapa, onde as classes estavam muito desbalanceadas, o uso intenso de balanceadores acabaram afetando o resultado dos testes. Nesse caso podemos perceber uma cobertura de 100% em *chikungunya*, que era justamente a classe com menos casos. Por tanto conclui-se que houve um superajustamento.

A RNA foi construída com 19 nós e uma camada intermediária. A Tabela 12 mostra a matriz de confusão gerada a partir dos indicadores de avaliação.

Tabela 12 – Matriz de confusão para o classificador neural PMC.

		Classificados		
		Dengue	*Chikungunya*	Outros
Reais	Dengue	472	2	55
	Chikungunya	0	501	0
	Outros	70	10	515

Os resultados do procedimento, alcançados através da metodologia adotada, pode guiar o desenvolvimento de uma plataforma de inferência capaz de classificar as doenças em questão em diversas vazes do processo de manejo clínico. O próximo capítulo descreve melhor o sistema e aplicativos que utilizam a ferramenta inteligente.

5.3.3 Considerações

A diferença de abordagem pode ser muito bem percebida a partir da análise dos resultados. A estratégia de ganho de informação utilizada nos métodos baseados em arvores de decisão são muito eficientes, pois, até mesmo em situações confusas, garantem uma representação eficaz do modelo de aprendizado.

6 SOLUÇÃO DENYA

Os resultados dos testes mostraram que os classificadores são capazes de acertar com precisão as doenças relacionadas ao mosquito *Aedes Aegypt*. Tais predições poderiam contribuir significativamente no contexto de aplicações na área de saúde que façam tratamento destas doenças. Assim, como produto deste trabalho de pesquisa é proposto o Denya (Sistema de Suporte ao Diagnóstico de Dengue e Chikungunya) que é capaz de fornecer o serviço de classificação das doenças à diversas aplicações. O sistema é uma Interface de Programação de Aplicações (API em inglês) que pode ser acessada via requisições REST (Representational State Transfer), permitindo acesso através da internet.

6.1 Arquitetura

A API foi desenvolvida usando a linguagem JAVA. A arquitetura do sistema foi dividida em três camadas, a saber, os módulos de dados, de inferência e de conexão. Há ainda a camada de aplicações, que faz uso dos recursos da API. A Figura 8 mostra a arquitetura do sistema.

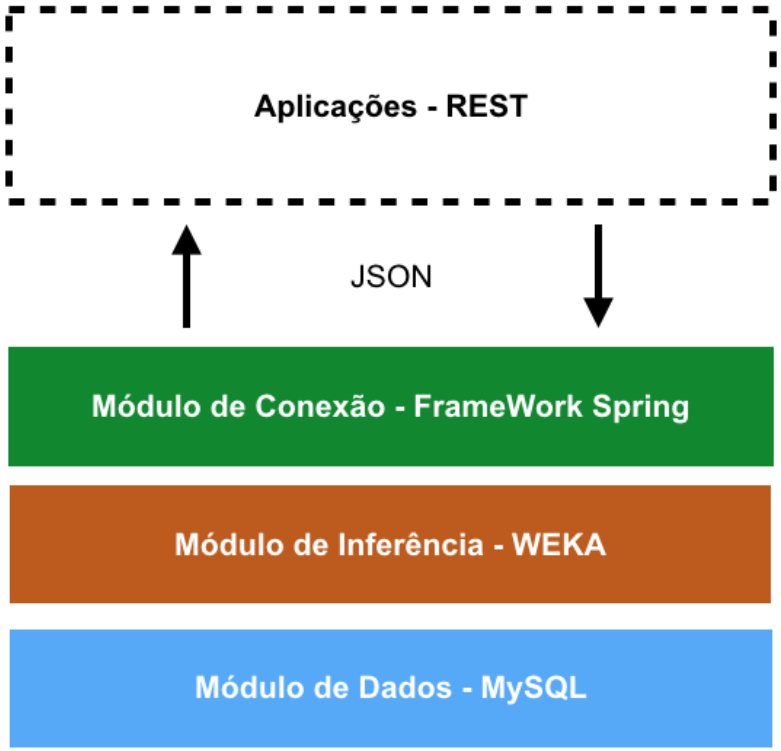
6.1.1 Módulo de Dados

O módulo de dados é responsável por armazenar o conjunto de casos das doenças em questão. Estes exemplos são usados para treinar os algoritmos sempre que o sistema é inicializado. O módulo conta com um conjunto de exemplos selecionados e tratados.

6.1.2 Módulo de Inferência

Etapa onde ocorre o processo de classificação de novos casos. Este módulo conta com dois algoritmos de classificação, nomeadamente, os classificadores RF e PMC. Um voltado para as primeira fases do manejo e o outro para fases mais avançadas, incluindo resultado de exames. Este módulo usa a API WEKA, disponível em *"Data Mining Software in Java"* (FRANK; HALL; WITTEN, 2016).

Figura 8 – Arquitetura do Sistema.



Fonte: Elaborado pelo autor.

6.1.3 Módulo de Conexão

Este é o módulo responsável por receber e tratar as requisições externas. As aplicações utilizam o padrão REST (*Representational State Transfer*) para se comunicar com a API usando o formato JSON. Para tanto, usa-se o *Framework Spring* para tratar estas requisições.

6.2 Aplicações

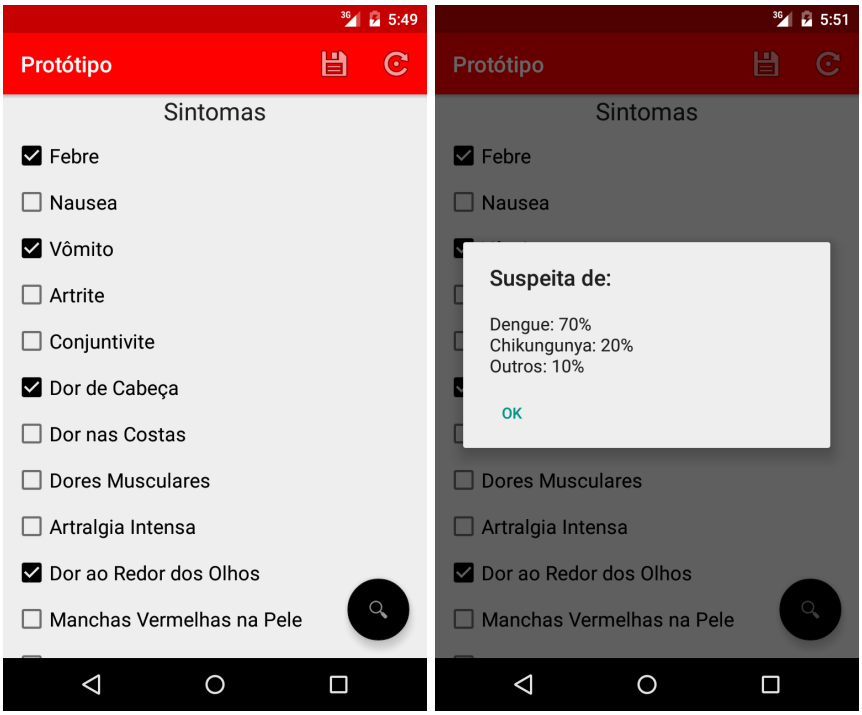
Nesta camada estão os *softwares* que utilizam a API para fornecer serviços de apoio a decisão. Essas softwares enviam um conjunto de sintomas e exames e, como resultado, recebem a probabilidade para cada doença.

6.2.1 Aplicativo para o Auxílio ao Diagnóstico de Dengue e Chikungunya

Como prova de conceito, este trabalho também desenvolveu uma aplicação móvel voltada ao pré-atendimento de pacientes com suspeita das doenças em ques-

tão. O aplicativo é capaz de, a partir de um conjunto de sintomas informados pelo paciente, inferir a doença mais provável. A partir de um questionário objetivo, o usuário descreve seus sintomas. Após responder ao questionário, a interface disponibiliza um botão para a classificação dos sintomas informados pelo usuário, conforme mostra a Figura 9.

Figura 9 – Interface do aplicativo móvel.



Fonte: Elaborado pelo autor.

A plicação foi desenvolvida sobre a plataforma Android, usando a linguagem de programação JAVA. Usou-se a IDE Android Studio, que é a oficial da plataforma, para o desenvolvimento da aplicação. Esta foi construída usando o SDK 22 do Android e está disponível para aparelhos com versões a partir da 4.0 (*Jelly Bean*).

7 CONCLUSÃO E TRABALHOS FUTUROS

O desenvolvimento de uma plataforma inteligente é um desafio complexo, principalmente no contexto da saúde. Em cenários evolutivos, onde há tratamentos de diferentes dados em cada etapa, percebe-se que diferentes abordagens comportam-se de maneira distintas, apresentando resultados divergentes. Algoritmos que apresentaram bons resultados numa etapa podem não alcançar estes mesmos bons resultados numa próxima fase. Portanto, faz-se necessário criar sistemas híbridos, que agreguem diversas abordagens. Neste trabalho dividiu-se o processo em apenas duas etapas, pois todo o processo de treinamento, testes e ajustes foi realizado manualmente, exigindo muito tempo e esforço. No entanto, alguns trabalhados proproem soluções que prometem automatizar essas etapas do processo usando ontologias. Essas abordagens integram as duas técnicas a fim de alcançar melhores resultados, otimizando as etapas de MD (HILARIO et al., 2009). Como visto, os processos de manejo clínico das doenças em estudo podem agregar diversos novos atributos durante sua evolução. Assim, seria interessante aplicar uma abordagem automatizada para o treinamento de algoritmos, escolhendo o melhor para cada conjunto de atributos. Desta forma, é possível encontrar sempre o melhor algoritmo para aquela situação.

Além de dengue e *chikungunya*, outras doenças transmitidas pelo vetor *Aedes Aegypti* precisam ser consideradas. Como é o caso da *zika*, que também tem causado sérios danos à população. A febre causada pelo vírus *Zika* também está associada ao aumento de ocorrências de microcefalia no país e é considerada de urgência nacional (LUZ; SANTOS; VIEIRA, 2015). No entanto, a falta de prontuários públicos desta enfermidade impossibilitou sua inclusão no processo classificatório desta pesquisa.

Devido sua maior gravidade, a dengue é uma doença de maior preocupação, pois tem causado mais mortes (BRASIL, 2016b). Assim, é de vital importância identificar, além da doença, sua gravidade. Tanto a dengue quanto *chikungunya* apresentam diversas forma clínicas, com gravidades diferentes. Uma versão futura da plataforma poderia tratar estas formas clínicas de maneira separada, alertando sobre a urgência/emergência do caso.

Um sistema de apoio a tomada de decisão que apoie as várias fases do manejo clínico poderia não só prever o provável diagnóstico do caso, mas, baseando-se em dados hospitalares disponíveis, incluindo disponibilidade de materiais em laboratórios, sugerir exames a fim de atingir um resultado mais preciso.

Como trabalhos futuros, pretende-se integrar o Denya ao MARCIA, um sistema interoperável para manejo clínico da *chikungunya* (SOUSA, 2017). O sistema

acompanha todo o processo de manejo clínico da doença, agregando informações do paciente desde sua primeira ida à unidade de saúde até seu diagnóstico e tratamento. A partir das informações de sintomas e exames inseridas no MARCIA, o módulo de inferência do Denya poderia calcular a probabilidade daquele caso estar relacionado a *chikungunya*. Estas informações, inseridas por profissionais de saúde, também poderiam ser incluídas na base de dados de casos da *chikungunya*, melhorando a predição da plataforma Denya.

REFERÊNCIAS

ALVES, M. R. R. S.; GADELHA, V. M. C. Onto2ae: Um sistema de auxílio aos pré-diagnósticos de doenças oriundas do mosquito aedes aegypti. In: . [S.l.]: Instituto Federal do Rio Grande do Norte (IFRN), Pau dos Ferros, RN, Brasil, 2016. Trabalho de Conclusão de Curso. Citado 2 vezes nas páginas 29 e 34.

AWAD, M.; KHANNA, R. *Machine Learning*. Berkeley, CA: Apress, 2015. 1–18 p. ISBN 978-1-4302-5990-9. Citado 2 vezes nas páginas 19 e 41.

AYYAZ, A. et al. Simulation model for counter-measures against *Aedes Aegypti*. In: *2015 13th International Conference on Frontiers of Information Technology (FIT), Dec. 14-16, Islamabad, Pakistan*. [S.l.: s.n.], 2015. p. 98–103. Citado na página 28.

BARRETO, M. L.; TEIXEIRA, M. d. G. L. C. *Dengue no Brasil: situação epidemiológica e contribuições para uma agenda de pesquisa*. 2008. http://www.repositorio.ufba.br/ri/handle/ri/2795. Acessado em 06-09-2017. Citado na página 17.

BRADLEY, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, v. 30, n. 7, p. 1145–1159, 1997. Citado na página 37.

BRAGA, O. C. et al. A mobile health solution for diseases control transmitted by Aedes Aegypti mosquito using predictive classifiers. In: *I Workshop de Computação Urbana (CoUrb) do XXXV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*. [S.l.]: SBRC, 2017. p. 144–156. Citado na página 29.

BRASIL. *Sistema de Informação de Agravos de Notificação: Normas e Rotinas*. [S.l.]: Ministério da Saúde, 2007. http://portalsinan.saude.gov.br/images/documentos/Portarias/Manual_Normas_e_Rotinas.pdf. Accessed: 2017-09-02. Citado na página 19.

BRASIL. *Dengue: Diagnóstico e Manejo Clínico*. [S.l.]: Ministério da Saúde, 2016. http://portalarquivos.saude.gov.br/images/pdf/2016/janeiro/14/dengue-manejo-adulto-crianca-5d.pdf. Accessed: 2017-08-10. Citado 2 vezes nas páginas 14 e 18.

BRASIL. *Monitoramento dos casos de dengue, febre de chikungunya e febre pelo vírus Zika até a Semana Epidemiológica 13*. [S.l.]: Ministério da Saúde, 2016. http://portalsaude.saude.gov.br/images/pdf/2016/abril/26/2016-014---Dengue-SE13-prelo.pdf. Acessado em 06-09-2017. Citado 3 vezes nas páginas 17, 18 e 48.

BRASIL. *Prevenção e Combate: Dengue, Chikungunya e Zika*. [S.l.]: Ministério da Saúde, 2016. http://combateaedes.saude.gov.br/pt/prevencao-e-combate/. Acesso em 06-09-2017. Citado na página 14.

BROWNE, M. W. Cross-validation methods. *Journal of Mathematical Psychology*, v. 44, p. 108–132, 2000. Citado na página 37.

BUSA-FEKETE, R. et al. Online F-measure optimization. In: *Advances in Neural Information Processing Systems (NIPS 2015), Dec. 7-12, Montreal, Canada*. [S.l.]: MIT Press Cambridge, 2015. p. 595–603. Citado na página 41.

CÂMARA, F. P. et al. Estudo retrospectivo (histórico) da dengue no brasil: características regionais e dinâmicas. *Rev Soc Bras Med Trop*, Scielo, v. 40, p. 192–196, Abr. 2007. Citado na página 18.

CARDOSO, P. D. A. *COISA: Conselheiro Inteligente de Saúde do Projeto Lariisa*. [S.l.]: Biblioteca Central Prof. Antônio Martins Filho, Universidade Estadual do Ceará, Fortaleza, CE, Brasil, 2015. Citado na página 29.

CHAWLA, N. V. et al. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, v. 16, p. 321–357, 2002. Citado na página 35.

COIERA, E. *Guide to Health Informatics*. [S.l.]: CRC Press, 2015. https://www.crcpress.com/Guide-to-Health-Informatics-Third-Edition/Coiera/p/book/9781444170498. Accessed 06-09-2017. ISBN 9781444170504. Citado na página 13.

FACELI, K. et al. *Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina*. Rio de Janeiro, RJ, Brasil: LTC, 2015. Citado 5 vezes nas páginas 13, 20, 21, 23 e 24.

FARIA, A. C. et al. Chikungunya: Manejo clínico. Citado na página 14.

FAYYAD, U. M. et al. *Advances in knowledge discovery and data mining*. Menlo Park, CA, USA: AAAI press, 1996. Citado na página 31.

FRANK, E.; HALL, M. A.; WITTEN, I. H. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. 2016. http://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf. Accessed: 2017-07-06. Citado 2 vezes nas páginas 26 e 45.

GARDINI, L. M. et al. Clariisa, a context-aware framework based on geolocation for a health care governance system. In: *2013 IEEE 15th International Conference on e-Health Networking, Applications Services (Healthcom), Oct. 9-12, Lisbon, Portugal*. [S.l.]: IEEE, 2013. p. 334–339. Citado 2 vezes nas páginas 28 e 29.

HALL, M.; HOLMES, G. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering*, v. 15, n. 6, p. 1437–1447, 2003. Citado na página 35.

HILARIO, M. et al. A data mining ontology for algorithm selection and meta-mining. In: *Proceedings of the ECML/PKDD09 Workshop on 3rd generation Data Mining (SoKD-09)*. [S.l.: s.n.], 2009. p. 76–87. Citado na página 48.

LOBO, L. C. Inteligência artificial e medicina. *Revista Brasileira de Edução Médica*, Scielo, v. 41, p. 185–193, Jun. 2017. ISSN 0100-5502. Citado na página 13.

LUZ, K. G.; SANTOS, G. I. V. d.; VIEIRA, R. d. M. Febre pelo vírus Zika. *Epidemiologia e Serviços de Saúde*, Coordenação-Geral de Desenvolvimento da Epidemiologia em Serviços, Secretaria de Vigilância em Saúde, Ministério da Saúde, v. 24, n. 4, p. 785–788, 2015. Citado na página 48.

MOREIRA, M. W. et al. A preeclampsia diagnosis approach using bayesian networks. In: *2016 IEEE International Conference on Communications (ICC), May 23-27, Kuala Lumpur, Malaysia*. [S.l.]: IEEE, 2016. p. 1–5. Citado na página 27.

MOREIRA, M. W. L. et al. An inference mechanism using bayes-based classifiers in pregnancy care. In: *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom), Sep. 14-17, Munich, Germany*. [S.l.]: IEEE, 2016. p. 1–5. Citado na página 37.

MOREIRA, M. W. L. et al. Performance evaluation of predictive classifiers for pregnancy care. In: *2016 IEEE Global Communications Conference (GLOBECOM), Dec. 4-8, Washington, DC, USA*. [S.l.]: IEEE, 2016. p. 1–5. Citado na página 27.

OLIVEIRA, A. M. B. et al. Applying ontology and context awareness concepts on health management system: a dengue crisis study case. In: . [S.l.: s.n.], 2013. Citado na página 29.

OLIVEIRA, T. W. F. *Aplicação de Redes Neurais Artificiais na Modelagem de um Classificador de Formas Clínicas de Dengue Utilizando Dados Genômicos*. 2009. Trabalho de Conclusão de Curso. Universidade de Pernambuco (UPE), Recife, PE, Brasil. Citado na página 29.

QUINLAN, J. R. *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. ISBN 1-55860-238-0. Citado na página 40.

RAMOS, R. F. et al. *Heart Diseases Prediction Using Data from Health Assurance Systems in Models and Methods for Supporting Decision-Making in Human Health and Environment Protection*. Nova York, NY, USA: Nova Publishers, 2016. ISBN 978-1-63485-202-9. Citado 2 vezes nas páginas 27 e 31.

RECIFE. *Casos de Dengue, Zika e Chikungunya*. [S.l.]: Prefeitura de Recife, PE, Brasil, 2016. http://dados.recife.pe.gov.br/dataset/casos-de-dengue-zika-e-chikungunya. Acessado em 06-09-2017. Citado na página 33.

SANTOS, A. C. dos. Aprendizado de máquina aplicado ao diagnóstico de dengue. In: *2016 13th Encontro Nacional de Inteligência Artificial e Computacional (SBC ENIAC-2016), Out. 8-12, Recife, PE, Brasil*. [S.l.]: SBC, 2016. p. 697–708. Citado na página 28.

SILVA, C. et al. LAIS, um analisador baseado em classificadores para a geração de alertas inteligentes em saúde. In: *2017 I CoUrb, XXXV SBRC, Mai. 15, Belém, PA, Brasil*. [S.l.]: SBRC, 2017. p. 157–169. Citado na página 27.

SOUSA, F. J. G. de. *MARCIA, UMA METODOLOGIA PARA O MANEJO DE REGISTRO CLÍNICO COM USO DE ARQUÉTIPOS PARA INTEROPERABILIDADE ENTRE SISTEMAS DE SAÚDE*. 2017. Dissertação, Curso de Mestrado Profissional Integrado em Computação Aplicada da Instituto Federal do Ceará (UECE), Fortaleza, Brasil. Citado na página 48.

STANGE, R. L.; NETO, J. J. Reconhecimento de padrões em classificadores – comparação de técnicas e aplicações. In: *IV Workshop de Tecnologia Adaptativa*

(WTA 2010), Jan. 21-22, São Paulo, SP, Brasil. [S.l.: s.n.], 2010. p. 63–67. Citado na página 20.

TAN, P. et al. *Introdução ao datamining: mineração de dados*. Ciencia Moderna, 2009. ISBN 9788573937619. Disponível em: <https://books.google.com.br-/books?id=69d6PgAACAAJ>. Citado na página 19.

TELES, G. et al. Using bayesian networks to improve the decision-making process in public health systems. In: *2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom), Oct. 11-15, Natal, RN, Brazil*. [S.l.]: IEEE, 2014. p. 565–570. Citado na página 28.

THANATHORNWONG, B.; SUEBNUKARN, S.; OUIVIRACH, K. Decision support system for predicting color change after tooth whitening. *Computer Methods and Programs in Biomedicine*, Elsevier, v. 125, p. 88–93, 2016. Citado na página 28.

WEBB, G. I. Naïve Bayes. In: *Encyclopedia of Machine Learning*. [S.l.]: Springer, 2011. p. 713–714. Citado na página 37.