UNIVERSIDADE ESTADUAL DO NORTE DO PARANÁ
CAMPUS LUIZ MENEGHEL - CENTRO DE CIÊNCIAS TECNOLÓGICAS
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

FREDERICO ANTONIO DOMINGUES

APRENDIZAGEM DE MÁQUINA APLICADA A PREVISÃO DE RISCOS DE EPIDEMIAS DE DENGUE

BANDEIRANTES–PR

2023

FREDERICO ANTONIO DOMINGUES

APRENDIZAGEM DE MÁQUINA APLICADA A PREVISÃO DE RISCOS DE EPIDEMIAS DE DENGUE

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Ciência da Computação da Universidade Estadual do Norte do Paraná para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Me. José Reinaldo Merlin

BANDEIRANTES–PR

2023

FREDERICO ANTONIO DOMINGUES

APRENDIZAGEM DE MÁQUINA APLICADA A PREVISÃO DE RISCOS DE EPIDEMIAS DE DENGUE

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Ciência da Computação da Universidade Estadual do Norte do Paraná para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA

Prof. Me. José Reinaldo Merlin
Universidade Estadual do Norte do Paraná
Orientador

Prof(a). Dr(a). Daniela de Freitas Guilhermino Trindade
Universidade Estadual do Norte do Paraná

Prof. Dr. Thiago Adriano Coleti
Universidade Estadual do Norte do Paraná

Bandeirantes–PR, 10 de Fevereiro de 2023

Dedico este trabalho ao meu pai, que sempre foi um herói e fonte de inspiração para mim. Sem a dedicação e a educação dele, eu não teria a oportunidade e a capacidade de iniciar e concluir o ensino superior. Hoje, lá de cima, espero que ele esteja orgulhoso. Também dedico a minha irmã Gabriela que sempre me deu forças para continuar.

AGRADECIMENTOS

Agradeço primeiramente aos meus irmãos e a minha família por todo apoio durante a graduação, sem eles eu não teria conseguido chegar até aqui. Também agradeço aos companheiros de curso e amigos que levarei comigo para o resto da vida, especialmente ao Andrey e Paulo por todo apoio. Por fim e não menos importante, agradeço a todos os professores e colaboradores da Universidade, sem o dom da propagação do conhecimento de cada um, nada disso seria real.

DOMINGUES, FREDERICO A.. **Aprendizagem de Máquina Aplicada a Previsão de Riscos de Epidemias de Dengue**. 44 p. Trabalho de Conclusão de Curso – Projeto (Bacharelado em Ciência da Computação) – Universidade Estadual do Norte do Paraná, Bandeirantes–PR, 2023.

RESUMO

Neste trabalho foi abordada a aplicação de algoritmos de classificação para previsões de riscos de epidemias de dengue. Epidemias de dengue são um grande problema sanitário presente em muitos países, principalmente nos tropicais e sub-tropicais. Sabe-se hoje que os fatores climáticos, como temperatura e umidade relativa do ar, afetam diretamente a vitalidade do mosquito transmissor da doença e do próprio vírus. Como consequência, tem influência direta com as epidemias além de outras variáveis. Portanto, inicialmente foi conduzido um estudo bibliográfico para entender o estado da arte de trabalhos relacionados já desenvolvidos, e então foram aplicados algoritmos de classificação baseando-se em dados agrometeorológicos para realizar as previsões. Entre os modelos aplicados, o algoritmo KNN (*K-Nearest Neighborhood*) obteve um resultado satisfatório, apresentando 91,2% de acertos em suas previsões.

Palavras-chave: Aprendizagem de Máquina. Modelos Preditivos. Epidemias de Dengue.

DOMINGUES, FREDERICO A.. **Machine Learning Applied to Risk Prediction of Dengue Epidemics**. 44 p. Final Project – Draft Version (Bachelor of Science in Computer Science) – State University Northern of Parana , Bandeirantes–PR, 2023.

ABSTRACT

In this paper, the application of classification algorithms for risk predictions of dengue epidemics was addressed. Dengue epidemics are a big health problem present in many countries, especially tropical and sub-tropical ones. It is known today that climatic factors such as temperature and relative humidity directly affect the vitality of the mosquito that transmits the disease and the virus itself. As a consequence, it has a direct influence on epidemics in addition to other variables. Therefore, initially a bibliographical study was conducted to understand the state of the art of related works already developed, and then classification algorithms were applied based on agrometeorological data to make predictions. Among the applied models, the KNN (K-Nearest Neighborhood) algorithm obtained a satisfactory result, presenting 91.2% of correct predictions.

Keywords: Machine Learning. Forecasting Mode. Dengue Epidemic.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

ACM	Association for Computing Machinery
AM	Aprendizagem de Máquina
API	Aplication Programming Interface
CD	Ciência de Dados
CONASS	Conselho Nacional de Secretários de Saúde
CSV	Comma Separated Values
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
IEEE	Institute of Electrical and Electronics Engineers
JSON	JavaScript Object Notation
KNN	K-Nearest Neighborhood
MLP	Multi Layer Perceptron
NAN	Not a Number
OMS	Organização Mundial da Saúde
PR	Paraná
R	Linguagem de programação R
RJ	Rio de Janeiro
RB	Rede Bayesiana
SGDC	Stochastic Gradient Descent Classifier
SVC	Support Vector Classifier
SVM	Support Vector Machine
UENP	Universidade Estadual do Norte do Paraná
UFMG	Universidade Federal de Minas Gerais
XLSX	Excel Microsoft Office Open XML Format Spreadsheet file

SUMÁRIO

1 INTRODUÇÃO

A expressão "dados são o novo petróleo"(*data is the new oil*) tem sido usada para enfatizar o quanto os dados podem gerar valor para empresas e organizações. Por isso, cada vez mais estão sendo realizadas pesquisas e investimentos nas mais diversas áreas, como mercado financeiro, marketing e exploração de petróleo, no sentido de utilizar os dados para aumentar as chances de atingir os objetivos. Diante desse cenário, em que se tem um enorme volume de dados complexos sendo produzidos diariamentes através de vários dispositivos e fontes distintas, existe a necessidade de ter um campo ou uma ciência específica voltada ao tratamento, ao processamento e à extração de informações desses dados.

Na área da saúde não é diferente, algoritmos de aprendizagem de máquina, ou *machine learning*, são amplamente utilizados como auxílio no prognósticos de doenças, utilizando a aprendizagem supervisionada, e nos agrupamentos de pacientes que possuem características clínicas e moleculares semelhantes, utilizando a aprendizagem não supervisionada, entre outros exemplos. Com a técnica de aprendizagem profunda, ou *deep learning*, ondas eletromagnéticas, como de eletrocardiogramas, por exemplo, podem ser analisadas utilizando algoritmos de visão computacional. O processamento de Linguagem Natural também aparece nesse cenário, como, por exemplo, através da extração de informações de diagnósticos de resultados laboratoriais [1], entre outras aplicações.

Além dessas aplicações de *machine learning* e *deep learning* em problemas diretos da saúde, muitas pesquisas estão sendo realizadas em bases de dados referentes a aspectos que influenciam a saúde pública. Nesse sentido, pesquisas relacionadas a epidemias, mais especificamente epidemias de dengue, vêm ganhando cada vez mais relevância social. Atualmente, sabe-se que, para que ocorra uma alta taxa de propagação do vírus da dengue, diversos são os fatores climáticos que influenciam diretamente na disseminação do vírus. Entre os principais, pode-se citar a temperatura e a umidade relativa do ar. Esses fatores criam condições favoráveis para que o mosquito se alimente e se reproduza. Além de fornecer condições favoráveis para o mosquito, o vírus também tem maior chance de sobrevivência nesse cenário.

Visto que os dados vêm sendo cada vez mais utilizados para as tomadas de decisões em diversas áreas do conhecimento, a aprendizagem de máquina vem sendo cada vez mais aplicada na área da saúde e da vigilância sanitária. O crescimento de estudos relacionados à previsão de epidemias de dengue e à influência direta dos fatores climáticos na quantidade de casos da doença faz com que seja possível explorar o tema e aplicar a inteligência artificial em uma região onde as epidemias de dengue são recorrentes. Neste trabalho, é mostrado um estudo empírico, no qual pretende-se aplicar a aprendizagem de máquina

para a previsão de riscos de epidemias de dengue, baseada em dados agrometeorológicos.

1.1 Contexto e Delimitação do Trabalho

De acordo com um levantamento feito pela Agência de Notícias do Paraná, no período entre 1 de agosto de 2020 e 1 de agosto de 2021, o estado paranaense registrou mais de 27 mil casos de dengue e 32 óbitos. Desses 32, nove óbitos são da cidade de Londrina, localizada na região norte do estado. Ainda de acordo com a notícia, a Secretaria da Saúde do estado enfatiza a importância do controle e da eliminação dos criadouros do *Aedes aegypti* (mosquito transmissor da doença) [2]. Já no período de agosto de 2021 até o final de abril de 2022, foram contabilizadas mais de 80 mil notificações e cinco mortes. As macrorregiões Oeste e Norte concentram o maior número de casos [3].

O CONASS (Conselho Nacional de Secretários de Saúde) definiu, em 2009, as Diretrizes Nacionais para a Prevenção e Controle de Epidemias de Dengue. De acordo com o documento, as diretrizes visam auxiliar estados e municípios na organização de suas atividades de prevenção e controle, em períodos de baixa transmissão ou em situações epidêmicas, contribuindo, desta forma, para evitar a ocorrência de óbitos e para reduzir o impacto das epidemias de dengue [4]. Essas diretrizes são divididas em seis componentes, sendo eles: assistência; vigilância epidemiológica; controle vetorial; comunicação e mobilização; gestão do plano e financiamento. Destes, destaca-se o segundo componente, vigilância epidemiológica. O documento aponta que a rápida coleta de informações nas unidades de saúde e a qualidade desses dados são essenciais para o desencadeamento oportuno de ações de controle e prevenção no nível local [4].

Desse modo, neste trabalho serão coletados dados do sistema InfoDengue e da Estação Agrometeorológica da Universidade Estadual do Norte do Paraná para treinar um modelo capaz de prever os riscos de epidemias de dengue, dado a previsão da temperatura média semanal e a previsão da umidade relativa semanal do ar. Espera-se que as informações geradas pelo modelo sejam os principais gatilhos para o desencadeamento de ações de controle e prevenção da doença.

Para este trabalho, serão considerados dados especificamente do município de Bandeirantes, localizado no estado do Paraná. Os dados serão tratados e os algoritmos serão treinados e avaliados apenas com os dados do munícipio. Espera-se que o relato empírico deste trabalho sirva como base para o desenvolvimento de modelos aplicados em qualquer outro município e região.

1.2 Formulação do Problema

Sabe-se que as epidemias de dengue afetam muitas regiões mundo afora, principalmente em regiões sub-tropicais. Regiões e períodos com muitas chuvas, alta umidade e temperaturas

elevadas contribuem diretamente para o desencadeamento dessas epidemias. Diante desse cenário e devido a grande quantidade de dados que são produzidos diariamente, tanto pelas vigilâncias sanitárias quanto pelas estações agrometeorológicas, pode haver dificuldades para criação de ações preventivas de controle e combate a epidemias de dengue baseadas em dados.

Isso porque, se essa análise for feita por um humano, demandará muito tempo e esforço para conseguir tomar uma decisão acertiva. A utilização de outras ferramentas de análise de dados, como por exemplo o Excel, pode não ser adequada para análises preditivas com alto volume de dados e que necessitam de um alto poder de processamento. Logo, a utilização da aprendizagem de máquina pode contribuir para que sejam feitas análises em cima desses dados, a fim de otimizar o tempo e o esforço humano dedicado a essa tarefa e de contribuir positivamente para o desencadeamento de ações de prevenção e combate a epidemia.

Portanto, de que forma a aprendizagem de máquina pode auxiliar na previsão de epidemias de dengue? O presente trabalho tem como hipótese que, através da (1) exploração e processamento dos dados, do (2) treinamento de algoritmos preditivos utilizando diferentes parâmetros e de (3) uma avaliação dos resultados de cada experimento, o algoritmo irá produzir insumos para a criação de ações preventivas de combate às epidemias.

1.3 Justificativa

Diante do cenário e dos problemas expostos, percebe-se que, embora os dados referentes aos casos de dengue e as condições meteorológicas sejam gerados, poucos são os municípios que os utilizam para análises preditivas. No município de Bandeirantes, município objeto deste estudo, não existe uma unidade ou equipe especializada para analisar esses dados e tomar decisões estratégicas baseadas neles. Isso faz com que os profissionais responsáveis pela criação de ações de combate e prevenção ao vírus do mosquito não tenham insumos de qualidade e agilidade para tomar as melhores decisões.

O fato de não existir uma equipe com conhecimento em aplicação de inteligência artificial para resolução de problemas acaba levando os profissionais a utilizarem abordagens e ferramentas inadequadas. O esforço humano empregado nessas análises acaba sendo maior e a utilização de ferramentas simples, como, por exemplo, Excel, acaba sendo inadequada e demorada, logo, não sendo eficaz.

Por isso, o presente trabalho tende a ter uma grande relevância social, uma vez que será conduzido um estudo empírico de desenvolvimento de um modelo de aprendizagem de máquina, capaz de prever os riscos de epidemias de dengue com base em dados meteorológicos. Isso poupará o esforço humano de realizar análises preditivas complexas e com alto volume de dados em ferramentas inadequadas, além de auxiliar na tomada de decisão

mais ágil e assertiva para a criação de estratégias de ações de combate às epidemias.

Uma vez que o modelo seja desenvolvido, ele poderá ser utilizado pelos profissionais responsáveis pelas tomadas de decisões no município, mesmo que com pouco ou nenhum conhecimento de programação ou aprendizagem de máquina, já que o modelo desenvolvido pode ser implementado em um sistema e ser autorrealimentado e treinado frequentemente. Também espera-se que o trabalho sirva como base para a aplicação em outros municípios e regiões diferentes.

1.4 Objetivos

Nesta seção será delimitado o objetivo geral da pesquisa. Para alcançá–lo, serão definidos os objetivos específicos que irão conduzir a pesquisa até sua meta final.

1.4.1 Objetivo geral

O objetivo deste trabalho é realizar um estudo empírico, no qual se busca aplicar a aprendizagem de máquina para prever os riscos de epidemias de dengue, baseando-se em dados meteorológicos.

1.4.2 Objetivos específicos

Para que seja possível atingir o objetivo geral, foram definidos os seguintes objetivos específicos:

1. Encontrar as bibliografias referentes à aprendizagem de máquina, com foco na área da saúde e vigilância sanitária, para analisar trabalhos relacionados já desenvolvidos.
2. Obter os dados referentes à quantidade de casos de dengue por semana epidemiológica e dados meteorológicos para treinar os diferentes algoritmos de *machine learning*.
3. Realizar a exploração e o processamento dos dados para entendê-los, e processá-los e otimizar o treinamento dos modelos e aumentar a eficácia deles.
4. Desenvolver os modelos de previsão e testar diferentes combinações de hiperparâmetros, para que seja possível encontrar a melhor combinação de parâmetros para cada algoritmo desenvolvido.
5. Avaliar os modelos desenvolvidos e identificar qual obteve o melhor desempenho e a melhor combinação de parâmetros para este trabalho.
6. Apresentar as conclusões do estudo empírico, destacando quais foram os pontos positivos e negativos notados ao longo do desenvolvimento.

1.5 Organização do Trabalho

O trabalho está organizado da seguinte forma. Neste capítulo 1 foi apresentada a introdução do trabalho, com a contextualização, a justificativa e os objetivos. No capítulo 2 será apresentada a fundamentação teórica necessária para a compreensão e o desenvolvimento do trabalho. No capítulo 3 será apresentada a metodologia do trabalho para a realização de cada etapa. No capítulo 4 será apresentado o desenvolvimento deste trabalho, que irá se iniciar com a obtenção dos dados, tanto aqueles referentes aos casos de dengue quanto aos meteorológicos; depois, os dados serão explorados e tratados; e a última etapa consistirá no desenvolvimento e na avaliação dos algoritmos de *machine learning*. No capítulo 5 serão apresentados os resultados e as discussões do trabalho. Por fim, no capítulo 6 serão apresentadas as conclusões.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo será apresentada uma revisão de literatura sobre os temas relativos ao trabalho. Primeiro, faz-se necessário entender o que é aprendizagem de máquina e os algoritmos de classificação presentes neste trabalho. Em seguida, há a necessidade de relacionar a aprendizagem de máquina ao estado da arte da sua aplicação na área da saúde e vigilância sanitária. Por fim, será apresentado a relação entre dengue e fatores climáticos presentes na literatura.

2.1 Aprendizagem de máquina e algoritmos de classificação

Segundo Monard e Baranauskas [5], a aprendizagem de máquina é uma área da inteligência artificial (IA) cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado bem como a construção de sistemas capazes de adquirir conhecimento de forma automática. Diversos são os algoritmos de aprendizagem de máquina que foram desenvolvidos ao longo dos anos, porém todos eles pertencem a alguma classe de aprendizagem. As principais são: aprendizagem supervisionada e aprendizagem não supervisionada. Ainda segundo Monard e Baranauskas [5], "no aprendizado supervisionado é fornecido ao algoritmo de aprendizado, ou *indutor*, um conjunto de exemplos de treinamento para os quais o rótulo da classe associada é conhecido" (p. 40).

Fazendo uma analogia com a aprendizagem humana, a aprendizagem supervisionada pode ser comparada com a aprendizagem dos nomes de animais durante o crescimento de uma criança. Para que aconteça a aprendizagem dos nomes de cada animal, é necessário que haja um "supervisor", geralmente é aquele que participa do processo de educação da criança, rotulando os nomes dos animais repetidas vezes até que, em um determinado momento, a criança consiga criar as relações entre nomes e animais e aprenda a rotulá-los sozinha.

Por outro lado, no aprendizado não supervisionado, os algoritmos de aprendizagem de máquina aprendem sem a necessidade de um "supervisor" para rotular seus dados. De acordo com Cheeseman e Stutz[1], em 1996, "o indutor analisa os exemplos fornecidos e tenta determinar se alguns deles podem ser agrupados de alguma maneira, formando *agrupamentos* ou *clusters*"[5] (p. 40). Nesse cenário, os algoritmos buscam encontrar os padrões em seus conjuntos de dados, baseados em modelos matemáticos e estatísticos. Nesse tipo de abordagem, o processo de aprendizagem do algoritmo é bem mais abstrato, não sendo possível saber, superficialmente, como o algoritmo encontrou determinado pa-

[1] CHEESEMAN, P. C.; STUTZ, J. C. Bayesian classification (autoclass): Theory and results. In: Advances in Knowledge Discovery and Data Mining. [S.l.: s.n.], 1996 *apud* [5]

drão em determinado conjunto de dados.

Para este trabalho, serão utilizadas técnicas e algoritmos que seguem a aprendizagem supervisionada. Por isso, faz-se necessário aprofundar um pouco mais sobre ela.

Debaixo do guarda-chuva que é a aprendizagem supervisionada, existem diferentes classes de problemas. Os mais comuns são problemas de classificação e de regressão. Os problemas de classificação são aqueles em que cada instância pertence a uma classe, que é representada por um valor conhecido como "atributo de classe". Esse atributo de classe é uma variável categórica que representa uma classe. A analogia apresentada anteriormente, de aprendizagem dos nomes dos animais, trata-se de um problema de classificação, em que o nome do animal é o atributo de classe e as demais informações como altura, peso, cor, forma etc. dos animais são as variáveis que compõem o atributo classe [6].

Os problemas de regressão são aqueles em que a saída do algoritmo, ao contrário dos algoritmos de classificação, não é uma variável categórica e sim uma variável numérica. Esses algoritmos são utilizados, por exemplo, em previsões de preços, temperatura e faturamento. Todo tipo de análise preditiva em que a saída do algoritmo é um valor numérico que não representa uma classe é um tipo de regressão.

A problemática deste trabalho se enquadra nos algoritmos de classificação, uma vez que pretende-se prever os riscos de epidemias de dengue e que a saída do algoritmo é uma variável categórica, que representa se as chances são altas, baixas ou estão em estado de alerta. Por isso, faz-se necessário explorar e entender melhor sobre os algoritmos abordados neste trabalho.

Com o passar do tempo e com a comunidade *open source* cada vez mais integrada, surgem cada vez mais algoritmos de *machine learning* e até mesmo os mais conhecidos vêm recebendo melhorias constantes. Neste trabalho, foram escolhidos os algoritmos mais conhecidos pela literatura em problemas de classificação, sendo eles:

- KNN (K-Nearest Neighborhood)
- MLP Classifier (Multi-layer Perceptron Classifier)
- SVM (Support Vector Machine)
- Linear SVC (Linear Support Vector Classifier)
- SGDC (Stochastic Gradient Descent Classifier)
- Naive Bayes Classifier

2.1.1 KNN: K-Nearest Neighborhood

O classificador KNN é baseado no aprendizado por analogia. As amostras de treinamento são descritas por N atributos numéricos dimensionais. Cada amostra representa um ponto em um espaço n-dimensional. Dessa forma, todas as amostras de treinamento são armazenadas em um espaço de padrões n-dimensionais. Quando dada uma amostra desconhecida, o classificador procura no espaço de padrões as K amostras de treinamento que estão mais próximas da amostra desconhecida. A "proximidade" é definida em termos de distância euclidiana [7].

2.1.2 Rede Neural: MLP Classifier

Um MLP Classifier é um modelo de rede neural artificial que segue a abordagem *feed forward*, em que os dados são transferidos entre as camadas, começando pela camada de entrada e passando para a(s) camada(s) oculta(s) até chegar na camada de saída. Um MLP consiste em várias camadas de nós em um grafo direcionado, com cada camada totalmente conectada à próxima. Exceto pelos nós de entrada, cada nó é um neurônio (ou elemento de processamento) com uma função de ativação não linear. O classificador utiliza a técnica de aprendizado supervisionado chamada de *backpropagation* para treinar a rede [8]. O perceptron multicamada (MLP) estende o perceptron com camadas ocultas, em outras palavras, camadas de elementos de processamento que não estão conectados diretamente ao mundo externo [9].

2.1.3 SVM (Support Vector Machine) e Linear SVC

O SVM foi introduzido pela primeira vez por Burges [10] e tem sido um método muito eficaz para regressão, classificação e reconhecimento geral de padrões. É considerado um bom classificador devido ao seu alto desempenho de generalização sem a necessidade de agregar conhecimento a priori, mesmo quando a dimensão do espaço de entrada é muito alta. O objetivo do SVM é encontrar a melhor função de classificação para distinguir entre os membros das duas classes nos dados de treinamento. A métrica para o conceito da "melhor" função de classificação pode ser realizada geometricamente. Para um conjunto de dados linearmente separável, uma função de classificação linear corresponde a um hiperplano de separação f(x) que passa pelo meio das duas classes, separando as duas. Uma vez determinada essa função, a nova instância de dados f(xn) pode ser classificada simplesmente testando o sinal da função f(xn); xn pertence à classe positiva se f(xn)>0 [9].

O algoritmo Linear SVC é semelhante ao SVM, porém com o acréscimo do parâmetro de confirguração Kernel = "linear". Por conta desse parâmetro, esse algoritmo tem mais flexibilidade na escolha de penalidades e nas funções de perda e deve escalar melhor

para um grande número de amostras. Essa classe suporta entradas densas e esparsas, e o suporte multiclasse é tratado de acordo com um esquema de um contra o resto [11].

2.1.4 SGDC (Stochastic Gradient Descent Classifier)

O *Stochastic Gradient Descent Classifier* (SGDC) implementa modelos lineares regularizados com aprendizado estocástico de gradiente descendente (SGD): o gradiente da perda é estimado em cada amostra de cada vez, e o modelo é atualizado ao longo do caminho com um cronograma de força decrescente (também conhecido como taxa de aprendizado). O SGD permite o aprendizado em minilote por meio do método partial_fit. Para obter melhores resultados usando a programação de taxa de aprendizado padrão, os dados devem ter média zero e variância unitária [12].

Essa implementação funciona com dados representados como matrizes densas ou esparsas de valores de ponto flutuante para os recursos. O modelo que ele se ajusta pode ser controlado com o parâmetro de perda; por padrão, ele se encaixa em uma *support vector machine linear* (SVM) [12].

O regularizador é uma penalidade adicionada à função de perda que reduz os parâmetros do modelo em direção ao vetor zero, usando a normalização euclidiana quadrada, a normalização absoluta ou uma combinação de ambas (rede elástica). Se a atualização do parâmetro cruzar o valor 0,0 devido ao regularizador, ela será truncada para 0,0 para permitir o aprendizado de modelos esparsos e obter a seleção de recursos online [12].

2.1.5 Naive Bayes Classifier - Redes Bayesianas

Uma Rede Bayesiana (RB) consiste em um grafo direcionado e acíclico e uma distribuição de probabilidade para cada nó nesse grafo, dados seus predecessores imediatos [13]. Um Classificador de Naive Bayes é baseado em uma rede bayesiana que representa uma distribuição de probabilidade conjunta sobre um conjunto de atributos categóricos. Consiste em duas partes: o grafo acíclico direcionado G, consistindo em nós e arcos, e as tabelas de probabilidade condicional. Os nós representam atributos enquanto os arcos indicam dependências diretas. A densidade dos arcos em uma RB é uma medida de sua complexidade. RBs esparsas podem representar modelos probabilísticos simples (por exemplo, modelos ingênuos de Bayes e modelos de Markov ocultos), enquanto RBs densas podem capturar modelos altamente complexos. Assim, RBs fornecem um método flexível para modelagem probabilística [9].

Muitos são os algoritmos que podem ser utilizados para resolução de problemas com inteligência artificial. Por isso, não existe uma "regra" de qual algoritmo deve ser utilizado em cada situação. O processo de resolução de problemas com IA é investigativo e exploratório, devem ser testados diferentes algoritmos e diversas configurações de parâmetros até encontrar a melhor configuração e os algoritmos que se adequam a cada caso.

Nas subseções anteriores, foram apresentadas as descrições básicas de cada modelo utilizado neste trabalho. Na próxima seção será mostrado como a aprendizagem de máquina, de modo geral, pode ser utilizada na área da saúde para que seja possível relacionar as duas áreas do conhecimento no decorrer deste trabalho.

2.2 Aprendizagem de Máquina e suas aplicações na saúde

A transformação da aprendizagem de máquina (AM) em muitas áreas, o rápido crescimento do poder computacional e a disponibilidade e evolução das ferramentas *open-source* têm contribuído para que a AM esteja cada vez mais presente na área da saúde [14]. As linguagens de programação, como, por exemplo, R e Python, têm inúmeras bibliotecas estatísticas que têm contribuído com o avanço dos algoritmos de *machine learning* e *deep learning* [1].

Recursos educacionais, tais como programas de graduação e pós-graduação em Ciência de Dados (CD), estão sendo ofertados cada vez mais [1]. Além disso, diversos cursos online também estão disponíveis na Internet com a finalidade de disseminar os fundamentos de CD para aplicar nas mais diversas áreas do conhecimento. Plataformas de competições, como por exemplo o Kaggle, também têm contribuído para o desenvolvimento da ciência na área da saúde.

Inúmeras são as formas de aplicação da ciência de dados na saúde. Dentre elas, a mais comum é com modelos preditivos que auxiliam nos prognósticos de doenças, utilizando algoritmos de aprendizagem supervisionada [1]. Agrupamento de pacientes que possuem características clínicas e moleculares semelhantes é uma aplicação da CD na saúde que utiliza a técnica de aprendizagem não supervisionada [1].

Além dessas aplicações utilizando algoritmos de *machine learning*, também existe a presença do *deep learning* em problemas de saúde. A técnica de Processamento de Linguagem Natural pode ser utilizada para extrair informações e diagnósticos através de resultados laboratoriais [1].

Os dados de ondas eletromagnéticas podem ser analisados de diversas formas. Entre as mais comuns, pode-se destacar a análise através da imagem das ondas ou, então, através de dados numéricos estruturados extraídos dessas ondas. Para os dados numéricos, podem ser aplicados algoritmos de *machine learning* ou *deep learning* para realizar a análise. O avanço no campo do *deep learning*, que é particularmente utilizado para análise de imagens, resultou em um rápido aumento no número de estudos nessa área nos últimos anos [15].

Além dessas aplicações em problemas diretos da saúde, muitas pesquisas estão sendo realizadas em bases de dados referentes a aspectos que influenciam a saúde pública, tais como informações sobre níveis de saneamento básico, assistência social, programas de

transferência de renda, riscos ou ocorrências de desastres naturais, acidentes de trânsito ou de trabalho, além de dados socioeconômicos ou demográficos [16].

Nesse sentido, pesquisas relacionadas a epidemias, mais especificamente epidemias de dengue, vêm ganhando cada vez mais relevância. A Organização Mundial da Saúde (OMS) definiu estratégias de monitoramento e controle integrado à disseminação da dengue em diversas regiões do mundo com o objetivo de reduzir a ocorrência e os óbitos da doença [17]. Além de estratégias, a OMS também deu suporte financeiro para pesquisas relacionadas ao monitoramento de potenciais surtos de dengue em diversas regiões do mundo [17].

Com a crescente quantidade de estudos relacionados à dengue, percebeu-se que existem alguns fatores que estão ligados diretamente aos surtos de dengue. Por isso, faz-se necessário um estudo bibliográfico para entender melhor esses fatores.

2.3 Dengue e seus fatores climáticos

Pesquisadores ao redor do mundo vêm desenvolvendo pesquisas relacionadas a previsão de surtos de dengue em vários países, e muitos fatores vêm sendo utilizados para estudar a correlação entre essas condições e o número de casos [17].

De acordo com Siriyasatien *et al.* [17], os fatores podem ser divididos em diretos e indiretos. Os fatores diretos são aqueles que afetam diretamente o ciclo de vida do mosquito *Aedes aegypti*. Logo, vários fatores podem ser considerados para afetar diretamente em seu ciclo de vida, tais como clima, densidade e quantidade de mosquitos e sorotipo do vírus da dengue [17].

A chuva é um fator que afeta o período de incubação dos mosquitos, que só podem completar seu ciclo de vida em água parada [18]. Por outro lado, regiões de secas e de altas temperaturas acabam prejudicando o ciclo de vida do mosquito e, consequentemente, afetando sua reprodução. Entretanto, nessas regiões a população humana costuma armazenar água para o consumo, contribuindo para o aumento da quantidade de mosquitos *Aedes aegypi* [17].

Além desses fatores, Siriyasatien [17] destaca que o aquecimento global é outro fator climático que afeta o ciclo de reprodução do mosquito, pois eles hibernam durante baixas temperaturas e são mais ativos em temperaturas mais quentes. Conforme cada vez mais regiões sofrem com altas temperaturas devido ao aquecimento global, o mosquito vai expandido sua geolocalização.

Vários outras pesquisas foram e estão sendo conduzidas para um melhor entendimento desses fatores. Focks e Chadee [19], Seng *et al.* [20] conduziram pesquisas relacionadas à densidade de larvas de mosquitos e concluiram que esse é um dos principais fatores

para previsão de epidemias de dengue. Naish *et al.* [21] descobriram que as condições do ecossistema e os fatores climáticos afetam diretamente a incidência de mosquitos e do vírus da dengue, ambos sensíveis às mudanças de temperatura e umidade.

McLennan-Smith e Mercer [22] mostraram que o aumento da temperatura média de 26^{o}C-28^{o}C para 30^{o}C reduz o período de incubação, permitindo a propagação mais rápida do vírus da dengue. Karim *et al.* [23] descobriram que a alta umidade relativa do ar durante a estação chuvosa acelera a taxa de crescimento das larvas e resulta em uma maior taxa de sobrevivência e vitalidade dos mosquitos, levando a um maior crescimento e disseminação da população de mosquitos transmissores e, consequentemente, a propagação do vírus da dengue.

Portanto, a umidade relativa é um dos fatores climáticos que se correlacionam com surtos de dengue [24]. Xu *et al.* [25] usaram a umidade absoluta para prever surtos de dengue e confirmaram que essa umidade é um fator que deve ser considerado em um modelo de previsão de surtos de dengue. A umidade influencia em vários fatores, tais como hábitos alimentares dos mosquitos, dispersão, produção de ovos e período de sobrevivência das larvas, resultando em uma transmissão maior e mais disseminada do vírus [17].

Pode-se perceber que diversos fatores climáticos influenciam diretamente na propagação do vírus da dengue. Entre os principais, pode-se citar a temperatura, a umidade relativa do ar e a densidade de chuva, que influenciam em condições favoráveis para que o mosquito se alimente e se reproduza. Além de fornecer condições favoráveis para o mosquito, o vírus também tem maior chance de sobrevivência nessas condições. Desse modo, conforme cada vez mais mosquitos são contaminados pelo vírus, e cada vez mais mosquitos sobrevivem ao ecossistema, maior será a taxa de ocorrências de dengue em seres humanos.

3 METODOLOGIA

Em relação a sua natureza, esta pesquisa é do tipo aplicada. Para Almeida, Leite e Tuani [26], a pesquisa aplicada objetiva gerar conhecimentos para aplicação prática, dirigidos à solução de problemas específicos, em que envolve verdades e interesses locais. Quanto aos seus objetivos, esta pesquisa é definida como exploratória. Ainda segundo os autores [26], a pesquisa exploratória visa proporcionar maior familiaridade com o problema para torná-lo explícito e construir hipóteses. Envolve levantamento bibliográfico, entrevistas com pessoas que tiveram experiências práticas com o problema pesquisado e análise de exemplos que estimulam a compreensão.

Esta pesquisa é definida como aplicada uma vez que a finalidade do trabalho é realizar um estudo empírico aplicando a técnica de aprendizagem de máquina. Além disso, tem um caráter exploratório, pois é realizado um estudo bibliográfico para proporcionar um maior conhecimento sobre o problema, analisando trabalhos relacionados.

Diante disso, para alcançar o objetivo geral e os específicos, os seguintes passos foram executados:

1. Realizar consultas bibliográficas referentes a aprendizagem de máquina, com foco em aplicação na área da saúde e nas previsões de epidemias de dengue em outras regiões. Para isso, serão utilizadas as bases indexadoras: IEEE Explorer e ACM Digital Library.

2. Obter os dados necessários para o trabalho, tais como: quantidade de casos de dengue no município de Bandeirantes, que podem ser obtidos através de uma API do sistema InfoDengue; e também os dados referentes a umidade relativa do ar e temperatura média semanal, que podem ser obtidos através do sistema da Estação Agrometeorológica da Universidade Estadual do Norte do Paraná.

3. Realizar a análise exploratória e o pré-processamento desses dados. Por virem de fontes distintas, há a necessidade de entender e manipular esses dados para deixá-los padronizados, limpos e organizados. Para isso, serão utilizados algoritmos de visualização de dados e algoritmos para corrigir valores faltantes, realizar escala dos dados, excluir colunas irrelevantes, entre outras necessidades que surgirem no decorrer do desenvolvimento.

4. Desenvolver os modelos de *machine learning*. Nessa etapa será realizado o desenvolvimento e o treinamento dos modelos preditivos utilizando diferentes parâmetros e técnicas de validação cruzada.

5. Avaliar os modelos é a última etapa do desenvolvimento deste trabalho. Depois de treinados, será realizada uma avaliação dos modelos com a finalidade de encontrar quais tiveram melhores resultados, utilizando as métricas padrões de avaliação de algoritmos de classificação.

6. Por fim, escrever e revisar o trabalho.

4 DESENVOLVIMENTO

Neste capítulo serão apresentadas as principais etapas de desenvolvimento de um projeto de *machine learning*. Será iniciado com a coleta dos dados, que acontece em duas fontes diferentes. Uma através de uma API e outra de coleta manual, exportando os dados manualmente de outro sistema. Em seguida, será realizada a exploração, a manipulação e a transformação dos dados. A terceira etapa do desenvolvimento consistirá em desenvolver os algoritmos de aprendizagem de máquina. Neste trabalho serão experimentados alguns dos mais conhecidos algoritmos de classificação encontrados na literatura. Por fim, será conduzida uma avaliação dos resultados dos algoritmos treinados.

4.1 Coleta dos dados

Os dados que serão utilizados neste trabalho foram coletados de duas diferentes fontes: InfoDengue e Estação Agrometeorológica da UENP.

4.1.1 InfoDengue

A primeira obtenção dos dados foi realizada utilizando a API do sistema InfoDengue[2]. De acordo com a própria descrição no site do sistema, o InfoDengue é um sistema de alerta para arboviroses baseado em dados híbridos que são coletados de diversas fontes e analisados. É um sistema que foi desenvolvido em 2015 por pesquisadores da Fundação Oswaldo Cruz - RJ e da Fundação Getúlio Vargas, contando com uma forte colaboração da Secretaria Municipal de Saúde do Rio de Janeiro, o Observatório da Dengue (UFMG) e pesquisadores da Universidade Federal do Paraná e da Universidade Estadual do Oeste do Paraná.

De modo geral, o sistema coleta os dados de diferentes fontes. Os dados referentes a dengue são de notificação obrigatória, o que significa que o profissional de saúde que diagnostica um caso suspeito precisa preencher uma ficha de notificação que alimenta um banco de dados municipal, que depois é consolidado a nível estadual e, finalmente, a nível federal pelo Ministério da Saúde [27].

Dados de temperatura e umidade são obtidos das estações meteorológicas de aeroportos assim como de imagens de satélite. O InfoDengue também tem parceria com o Observatório da Dengue que captura e analisa *tweets* de pessoas geolocalizadas quanto à menção de sintomas de dengue. Por fim, dados demográficos dos municípios brasileiros são atualizados a cada ano no InfoDengue, utilizando as estimativas do IBGE [27].

[2] Disponível em: <https://info.dengue.mat.br/>.

Depois que os dados são coletados de diversas fontes, o sistema os analisa e gera como saída um indicador de situação epidemiológica, que pode ser identificado conforme sua cor ou descrição. Esse indicador pode assumir quatro valores: Baixa Transmissão (cor verde), Atenção (cor amarela), Transmissão (cor laranja) ou Alta Incidência (cor vermelha).

Os dados podem ser obtidos através de uma API disponibilizada no próprio site do sistema utilizando as linguagens Python ou R. Para este trabalho, a obtenção dos dados aconteceu através de um *script* em Python. No *script*, devem ser informados os seguintes parâmetros:

- geocode: esse dado é referente ao código geográfico do município do qual se deseja os dados e pode ser encontrado no site do IBGE. Para esse caso, o geocode referente ao município de Bandeirantes-PR é 4102406;
- disease: representa de qual arbovirose se está buscando os dados, podendo ser dengue ou chikungunya. Para esse caso, a *string* "dengue" é passada como parâmetro;
- format: indica qual o formato do arquivo desejado, podendo ser CSV ou JSON. Nesse caso, deseja-se o formato CSV;
- ew_start: semana epidemiológica de início da consulta, que pode variar de 1 a 53. Para o desenvolvimento do trabalho, deseja-se os dados desde a primeira semana de todos os anos, portanto, o valor desse parâmetro será 1;
- ew_end: semana epidemiológica de término da consulta, também variando de 1 a 53. Para este trabalho, utiliza-se o valor 49. Isso quer dizer que se deseja todos os dados até a semana 49 do ano de término da consulta;
- ey_start: ano de início da consulta. Os dados mais antigos referentes ao município de Bandeirantes foram coletados em 2010, portanto, o parâmetro utilizado é 2010; e
- ey_end: ano de término da consulta. Deseja-se todos os dados até o ano 2021.

Com base nos parâmetros apresentados, foram obtidos os dados do município de Bandeirantes desde a primeira semana epidemiológica do ano de 2010 até a 49ª semana epidemiológica do ano de 2021.

Depois que os dados foram obtidos e armazenados em um *DataFrame*, podem ser visualizados conforme mostrado na Figura 1:

Figura 1 – Visualização inicial dos cinco primeiros dados de casos de dengue

data_iniSt	SE	casos_est	casos_est_min	casos_est_max	casos	p_rt1	p_inc100k	alidad	nivel	id	ersao_modelc	tweet	Rt	pop	tempmin	umidmax	receptivo	transmissao	nivel_inc	notif_accum_year
2010-01-03	201001	0	0	0	0	0	0	0	1	41024…	2020-12-10	nan	0	32562	23	nan	0	0	0	89
2010-01-10	201002	2	2	2	2	0	6.14213	0	1	41024…	2020-12-10	nan	0	32562	21	nan	0	0	0	89
2010-01-17	201003	2	2	2	2	0	6.14213	0	1	41024…	2020-12-10	nan	0	32562	21	nan	0	0	0	89
2010-01-24	201004	5	5	5	5	0	15.3553	0	1	41024…	2020-12-10	nan	0	32562	20	nan	0	0	1	89
2010-01-31	201005	3	3	3	3	0	9.21319	0	1	41024…	2020-12-10	nan	0	32562	21	nan	0	0	0	89

Conforme observa-se na Figura 1, nos dados obtidos utilizando o sistema InfoDengue, embora apresentem muitos atributos referentes aos casos de dengue, não apresentam todos os atributos meteorológicos necessários para o trabalho. Para o sistema InfoDengue, atributos como menções e interações com o termo "dengue" nas redes sociais, população municipal e outras variáveis, também são levadas em consideração para o treinamento dos algoritmos. Nesse conjunto de dados é aproveitada apenas a quantidade de casos notificados no município em cada semana epidemiológica.

Conforme apresentado anteriormente, diversos estudos mostram que os fatores climáticos, como por exemplo temperatura e umidade relativa do ar, estão associados diretamente à quantidade de casos de dengue. Percebe-se que o mosquito tem melhores condições de vida e reprodução quando os dias são quentes e úmidos. Entretanto, os dados referentes à temperatura e à umidade relativa do ar não constam no conjunto de dados obtidos pelo sistema InfoDengue. Por isso, faz-se necessário buscá-los em outra fonte.

Para isso, utilizou-se o sistema da Estação Agrometeorológica da Universidade Estadual do Norte do Paraná, Campus Luiz Meneghel, no município de Bandeirantes. Tanto os dados referentes à umidade relativa do ar quanto à temperatura média da primeira semana de 2010 até a 49ª semana de 2021 podem ser obtidos no próprio site da estação.

4.1.2 Estação Agrometeorológica da UENP

As atividades da Estação Agrometeorológica iniciaram nos anos 1970. No início, os equipamentos instalados foram termômetros, seguidos de pluviômetros e radiômetros. Na estação são feitas três leituras diárias, sendo às 9h, às 15h e às 21h. Por meio desses sensores instalados, é realizada a leitura diária da temperatura e da umidade em três períodos. A obtenção desses dados pode ser realizada no próprio site da estação [28].

Conforme pode-se visualizar na Figura 2, a obtenção desses dados é bem simples e intuitiva, basta selecionar o período e dado desejado que o download será iniciado. Percebe-se que o formato do arquivo baixado é feito em XLSX, por isso, faz-se necessário converter o arquivo para CSV. Depois de realizada a conversão de XLSX para CSV, os dados do arquivo podem ser armazenados em um *DataFrame* e visualizados conforme a Figura 3.

Figura 2 – Visualização da página de download dos dados da Estação Agrometeorológica da UENP

Figura 3 – Visualização inicial dos cinco primeiros registros da umidade relativa do ar

Id	Data	09:00 (%)	15:00 (%)	21:00 (%)	Média (%)
16882	31/12/2020	88	100	100	97
16490	31/12/2019	nan	nan	nan	nan
16126	31/12/2018	70	68	86	77.5
15671	31/12/2017	95	55	85	nan
15282	31/12/2016	82	50	88	nan

A Figura 4 ilustra a etapa de coleta dos dados e armazenamento das informações em seus respectivos *DataFrames*:

Figura 4 – Coleta de dados e armazenamento nos *DataFrames*

Com os dados disponíveis, inicia-se a etapa de limpeza e preparação dos dados para análise. Essa etapa de limpeza dos dados é considerada na literatura uma das etapas mais exaustivas e importantes dentro de um projeto de *machine learning*, isso porque os dados são oriundos de fontes distintas, com formatos distintos e necessitam de limpeza, ajustes e padronização para que os modelos de *machine learning* consigam ter a melhor absorção da aprendizagem. Esse processo será discutido na próxima seção.

4.2 Análise Exploratória e Pré-Processamento dos Dados

Conforme apresentado no Capítulo 2, a exploração dos dados consiste basicamente em entendê-los. Nesta etapa, são utilizadas bibliotecas de manipulação e visualização de dados da linguagem Python. Além de entendê-los, busca-se também identificar quais serão as manipulações necessárias nas próximas etapas de desenvolvimento.

O primeiro passo desse processo consiste na identificação de valores faltantes em ambos os *datasets*. Dados faltantes são registros que não foram coletados e armazenados. Geralmente acontecem por falhas mecânicas e/ou elétricas dos sensores ou eventos semelhantes. A presença deles nos conjuntos de dados atrapalha a aprendizagem e a capacidade de generalização dos algoritmos. Portanto, é necessário encontrá-los e manipulá-los.

Para o conjunto de dados obtidos do sistema InfoDengue, foram encontrados treze registros NaN (*Not a Number*) na coluna "umidmax". Já nos referentes à umidade relativa do ar e temperatura média, a quantidade de registros faltantes é muito superior. Na literatura, existem diferentes estratégias para corrigir valores faltantes dentro de um con-

junto de dados. A estratégia adotada neste trabalho é a substituição dos valores faltantes encontrados pela média aritmética dos demais valores da coluna.

Para finalizar os ajustes de valores faltantes nos conjuntos de dados extraídos da estação, foi atribuída a média aritmética da umidade coletada nos três períodos do dia, e para a temperatura média, a média aritmética é calculada entre a menor e a maior temperatura coletada no dia.

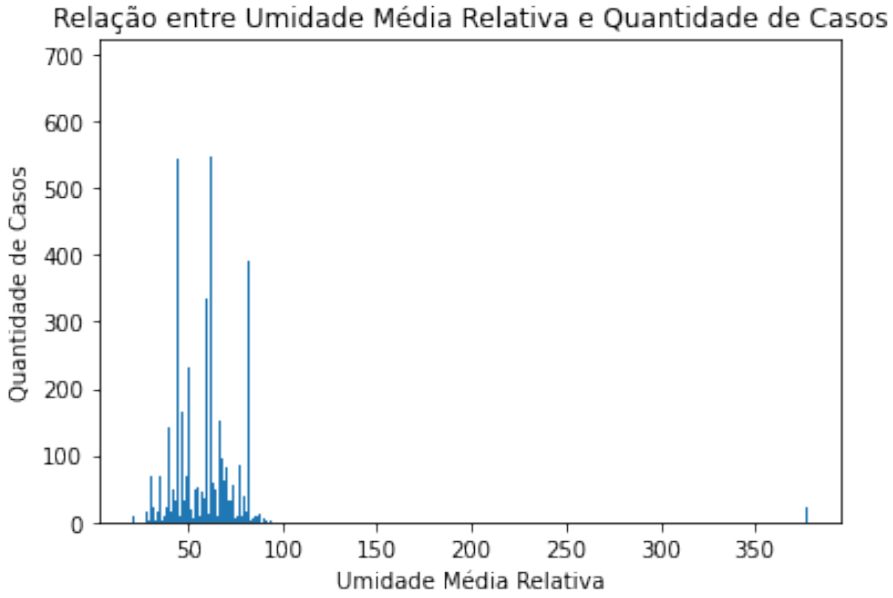
Depois de ambos os conjuntos preenchidos com todos os registros, é realizada a exclusão de informações que não têm relação com a quantidade de casos de dengue. Conforme visto anteriormente, o *dataset* referente a quantidade de casos de dengue apresenta muitas colunas que não são relevantes para este trabalho. Essa manipulação permite aumentar a objetividade dos dados.

Inicialmente, os dados extraídos do sistema InfoDengue contêm informações que foram agregadas ao conjunto de dados, com a finalidade de utilizar as redes sociais como uma das variáveis para previsão de epidemias de dengue. Para este trabalho, essas variáveis, bem como algumas outras referentes à população do município e entre outras, não serão utilizadas.

Com os valores faltantes corrigidos nos conjuntos de dados e colunas irrelevantes removidas, no *DataFrame* referente à umidade relativa do ar, é realizada a exclusão da coluna "Id" e as demais colunas foram renomeadas para "manhã", "tarde" e "noite", representando os três períodos do dia em que os dados são coletados. Para o conjunto referente à temperatura média semanal, também foi removida a coluna "id", visto que é uma coluna que serve de identificação do registro e não uma informação relacionada à quantidade de casos de dengue em si.

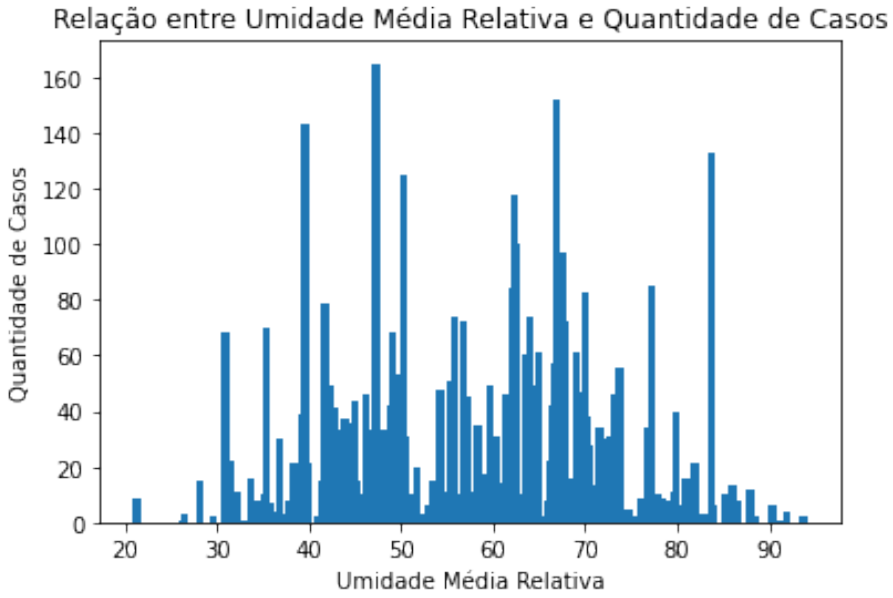
Após os ajustes para lidar com a falta de valores nos dados, foi realizada uma análise para identificar *outliers*, que são valores que diferem significativamente da população. [29]. A busca por *outliers* foi realizada considerando os parâmetros: (1) quantidade de casos x temperatura média, na qual não foram identificados *outliers* e (2) quantidade de casos x umidade relativa do ar. Na análise de quantidade de casos e umidade relativa do ar foram identificados *outliers* conforme observado no gráfico da Figura 5, uma vez que o intervalo de valores da umidade média relativa não está coerente. Percebe-se que algum registro no conjunto apresenta a umidade média maior que 350%.

Figura 5 – Relação entre a Umidade Média Relativa e a Quantidade de Casos registrados



A correção do *outlier* foi feita utilizando a biblioteca "stats" do pacote Scipy, da linguagem Python [30]. Diversas são as estratégias que podem ser adotadas para a correção de *outliers*, entre as mais comuns encontradas na literatura temos a exclusão do registro (em caso de pequenas quantidades) ou, então, a alteração do valor encontrado pelo valor da média do conjunto. Para este trabalho, foi adotada a estratégia de substituição dos valores *outliers* encontrados pela média do conjunto de dados, visto que o *dataset* já apresenta poucos registros. Na Figura 6 é apresentada a distribuição dos dados de quantidade de casos em relação à umidade, sem a ocorrência de *outlier*.

Figura 6 – Relação entre a Umidade Relativa e a Quantidade de Casos pós correção de *outlier*



A última etapa do processamento dos dados consistiu em um ajuste nos dados de umidade relativa e temperatura média. As informações de ambos conjuntos de dados foram coletadas diariamente, em diferentes períodos do dia (manhã, tarde e noite). Por outro lado, os dados referentes à quantidade de casos de dengue estão organizados em um intervalo semanal (semanas epidemiológicas). Logo, faz-se necessário calcular a média da umidade e da temperatura para cada semana epidemiológica, para que todos os conjuntos de dados estejam no mesmo intervalo e possam ser concatenados para formar um único conjunto com todas as informações. Na Figura 7 é apresentado o *script* de padronização dos períodos de leitura dos dados.

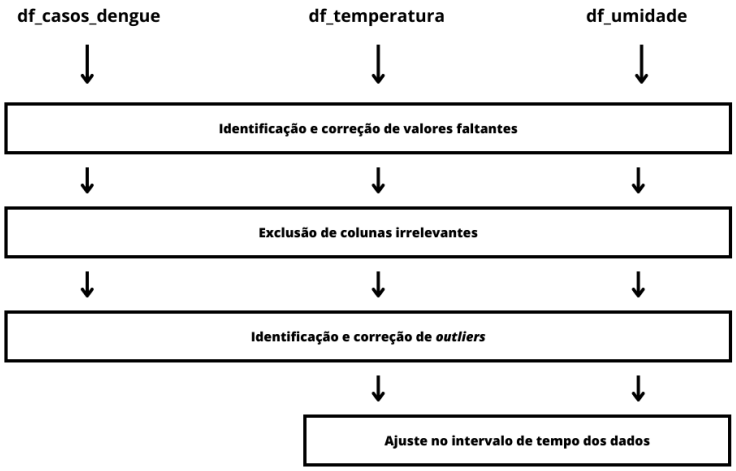
Figura 7 – *Script* para cálculo da média semanal dos dados

```
def calcular_media_semanal( df ):
    medias_semanais = []
    semanas = []
    for i in range(0,623):
        if(i < 622):
            data_inicial = df.iloc[i,0]
            data_final = df.iloc[i+1,0]
            medias_temp = []
        else:
            data_inicial = df.iloc[i,0]
            data_final = datetime.datetime(2021, 12, 15)
            medias_temp = []
        for j in range(0,4306):
            data_atual = df.iloc[j,0]
            if( data_atual >= data_inicial and data_atual < data_final):
                medias_temp.append(df.iloc[j,1])
                j = j + 1
        if(len(medias_temp) == 0):
            i = i + 1
        else:
            semanas.append(i+1)
            medias_semanais.append(sum(medias_temp)/len(medias_temp))
            medias_temp.clear()
            i = i + 1
    return medias_semanais
umidades_semanais = calcular_media_semanal ( df_umidade )
temperaturas_semanais = calcular_media_semanal( df_temperatura )
umidade_semanal = pd.DataFrame([umidades_semanais[::-1]])
umidade_semanal = umidade_semanal.transpose()
temperatura_semanal = pd.DataFrame([temperaturas_semanais[::-1]])
temperatura_semanal= temperatura_semanal.transpose()
frame = {'Data': df_casos['data_iniSE'],
         'Temp_media_sem': temperatura_semanal[0],
         'Umidade_media_sem': umidade_semanal[0],
         'Qtd_casos': df_casos['casos']}
df_casos_final = pd.DataFrame(frame)
df_casos_final = df_casos_final.drop(columns=['Data'])
```

Após as ações de pré-processamento, os dados ficaram preparados para serem utilizados em modelos de *machine learning*. Na Figura 8 são mostradas resumidamente as

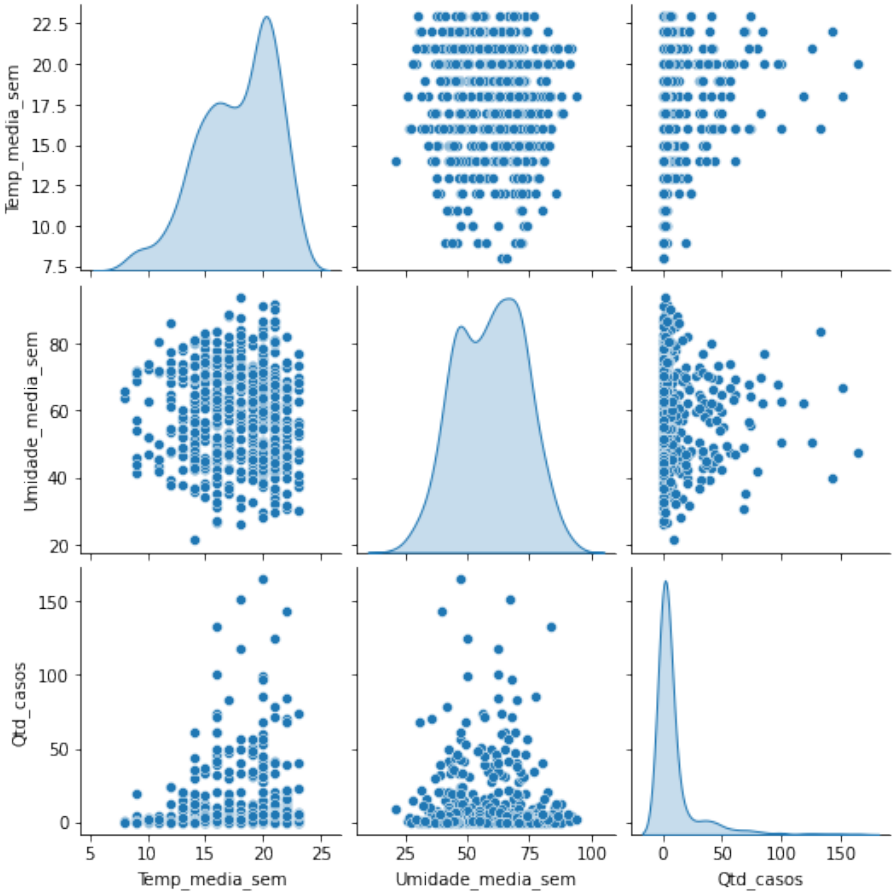
etapas aplicadas na manipulação dos dados.

Figura 8 – Etapas de ajustes dos dados



Nessa etapa, os dados foram analisados e manipulados. Essa etapa contribui significativamente para a aprendizagem dos algoritmos. As ações foram escritas utilizando a linguagem Python e suas bibliotecas. Após os ajustes nos conjuntos de dados, é possível realizar análises de inferências sobre as informações coletadas. Os resultados apresentados na Figura 9 permitem inferir que a quantidade de casos de dengue é maior quando a temperatura média está entre 18ºC e 24ºC e, também, quando a umidade relativa do ar está entre 30% e 70%.

Figura 9 – *Pair Plot* de correlações entre as variáveis



Uma vez que o objetivo da saída do modelo é a classificação do risco de epidemias de dengue no município (baixo, alerta ou alto), é necessário modificar os dados da coluna "Qtd_casos" para um valor que representa a classificação final do algoritmo e não a quantidade de casos. Para isso, foram analisadas quais as quantidades de casos registrados quando as condições climáticas fossem mais favoráveis para a sobrevivência e a reprodução do mosquito e do vírus.

Para este trabalho, conclui-se que, a partir das análises, quando a quantidade de casos é inferior a 15, as condições climáticas não são favoráveis, portanto o risco de epidemia é baixo. Quando a quantidade de casos registrados está entre 15 e 40, considera-se um estado de alerta para surtos de dengue, então o risco é moderado. Já quando a quantidade ultrapassa os 40 casos, considera-se um alto risco de epidemias. Na Figura 10 é apresentado o *script* para a manipulação descrita acima e na Figura 11 é exibido o *DataFrame* final.

Figura 10 – *Script* para alteração da saída do modelo.

```
def classificacao_qtd_casos(df):
    if(df['Qtd_casos'] >= 0 and df['Qtd_casos'] <= 15):
        return 0
    elif(df['Qtd_casos'] > 15 and df['Qtd_casos'] < 40):
        return 1
    return 2

df_casos_final['Risco'] = df_casos_final.apply(
                                    lambda df:classificacao_qtd_casos(df),
                                    axis=1)

df_final = excluir_colunas(df_casos_final, 'Qtd_casos')
```

Figura 11 – *DataFrame* final com a coluna Risco

df_final - DataFrame

Index	Temp_media_sem	Umidade_media_sem	Risco
0	7.68571	42	0
1	8.08571	71.8857	0
2	9.2	51.8571	0
3	9.22857	37	0
4	9.48571	28.0714	0

4.3 Implementação dos modelos de *machine learning*

Assim como já mencionado anteriormente, os problemas de *machine learning* são separados em diferentes classes. A classe de problemas de classificação é aquela em que o *output* do modelo pertence a uma classe dentro do conjunto de dados. Por exemplo, nessa classe de problemas os *outputs* podem ser "sim" ou "não", ou então o nome de algum objeto ou animal. Para o problema apresentado neste trabalho, as possíveis saídas do modelo são:

0 - representa que o risco de epidemia é baixo;

1 - representa que o risco é moderado;

2 - representa que o risco é alto.

Muitos são os algoritmos que podem ser utilizados em problemas de classificação. Porém, para esse problema, foram utilizados os seguintes:

- KNN: K-Nearest Neighborhood

- Rede Neural
- SVC: Support Vector Classifier
- Linear SVC: Linear Support Vector Classifier
- SGDC: Stochastic Gradient Descent Classifier
- Naive Bayes

Além desses algoritmos, também foi utilizada a abordagem de validação cruzada Repeated K-Fold. Seu principal objetivo é apresentar um resultado mais confiável para avaliação de cada modelo, uma vez que todos os dados do conjunto são utilizados para treino e teste. Além disso, esse método realiza o treinamento repetidas vezes de acordo com o valor de parâmetro que é passado para ele.

Uma outra técnica utilizada no desenvolvimento foi a configuração de hiperparâmetros. Essa técnica consiste basicamente em testar mais de um valor para o mesmo parâmetro durante o treinamento do modelo, a fim de determinar qual combinação de valores apresenta a melhor acurácia.

Depois de realizar a separação dos dados entre variáveis de treino e teste, foi realizada uma padronização na escala dos dados utilizando a função *Standard Scaler* da biblioteca Scikit Learn. Essa padronização faz com que todos os dados estejam na mesma escala. Alguns algoritmos de *machine learning* entendem que a escala dos valores que eles assumem podem apresentar maior relevância no conjunto do que outro atributo. Por exemplo, se um conjunto de dados contar com a informação da altura de uma pessoa em metros e o peso dela em quilogramas, no momento do treinamento do algoritmo, pode ser entendido que o peso da pessoa é mais relevante ao conjunto do que a altura, já que o intervalo de valores do peso é maior do que a altura em metros.

Os treinamentos aconteceram diversas vezes para cada algoritmo. Para cada iteração, os dados foram separados entre dados de treinamento e dados de teste, utilizando a estratégia Repeated K-Fold, no qual os dados podiam ser testados em três porções: 5, 7 e 9. Além disso, as repetições também podiam assumir três valores: 3, 6 e 9. Esses valores de hiperparâmetros foram utilizados no treinamento de todos os algoritmos, a fim de encontrar a melhor combinação de configuração para cada um.

Além desses, cada algoritmo também apresenta seus parâmetros próprios que podem ser configurados, logo, é possível testar diferentes combinações. Para o exemplo do treinamento da rede neural, além de testar os dois parâmetros citados anteriormente, também foram testados os valores: 100, 300 e 500, para o parâmetro “max_iter”, que representa a quantidade de iterações que a rede neural vai realizar os ajustes dos pesos entre as camadas.

Para cada combinação testada, foi armazenado o valor da média de acertos. Na Figura 12 é apresentado o *script* de testes de diferentes parâmetros para o modelo de rede neural.

Figura 12 – *Script* para treinamento da rede neural com diferentes parâmetros

```
def treinar_rede_neural():
    n_splits = [5, 7, 9]
    n_repeats = [3, 6, 9]
    max_iter =  [100, 300, 500]
    array_splits = []
    array_repeats = []
    array_acuracia = []
    array_max_iter = []

    for split in n_splits:
        for repeat in n_repeats:
            for iteracao in max_iter:
                kfold = RepeatedKFold(n_splits=split, n_repeats=repeat)
                clf = MLPClassifier(max_iter = iteracao).fit(x, y)
                media_clf = cross_val_score(clf, x, y, scoring='accuracy', cv=kfold).mean()
                array_splits.append(split)
                array_repeats.append(repeat)
                array_acuracia.append(media_clf)
                array_max_iter.append(iteracao)

    avaliacao_rede_neural = pd.DataFrame(data={'Splits: ':array_splits,
                                        'Repetições: ':array_repeats,
                                        'Iterações: ': array_max_iter,
                                        'Acurácia: ':array_acuracia})
    return avaliacao_rede_neural
```

O mesmo padrão aconteceu para os demais algoritmos, alterando apenas algum parâmetro específico de cada um. Depois de realizar o treinamento de cada algoritmo com diferentes valores de parâmetros, foi possível encontrar as melhores combinações. Na Figura 13 são apresentadas a melhores combinações de parâmetros de cada modelo e a acurácia alcançada por cada um.

Figura 13 – Tabela com os resultados da melhor combinação de parâmetros de cada algoritmo

Modelo	Parâmetros	Acurácia
KNN	Splits (partições) = **10** Repetições = **6** Vizinhos = **5**	91,2%
Rede Neural	Splits (partições) = **10** Repetições = **6** Iterações = **30**	79,3%
SVC	Splits (partições) = **10** Repetições = **6**	87,8%
Linear SVC	Splits (partições) = **10** Repetições = **6**	83,8%
SGDC	Splits (partições) = **8** Repetições = **8** Iterações = **20**	81,9%
Naive Bayes	Splits (partições) = **10** Repetições = **6**	89,6%

No próximo capítulo serão discutidos os resultados obtidos após os experimentos realizados com os diversos parâmetros aplicados nos algoritmos utilizados.

5 RESULTADOS E DISCUSSÃO

Conforme apresentado na Figura 13, percebe-se que, embora tenham sido utilizados diferentes algoritmos de classificação, as melhores combinações de parâmetros para cada modelo apresentam praticamente as mesmas combinações, sendo: *splits* (partições dos dados entre treinamento e teste) igual a 10 e repetições das partições e treinamento igual a 6. Para o classificador da rede neural, a melhor acurácia foi obtida com 30 iterações de regulagem dos pesos entre as camadas. Já para o SGDC, a melhor acurácia foi obtida com 20 iterações.

Com relação a acurácia dos modelos desenvolvidos, percebe-se que, embora seja o algoritmo mais "robusto", a rede neural obteve o pior desempenho. Algoritmos mais "simples", como, por exemplo, o classificador de Naive Bayes (que utiliza uma abordagem probabilística para as previsões), tem desempenhos superiores. Entre os algoritmos experimentados, destaca-se o KNN com 91,2% de acertos nas previsões.

Dada a baixa quantidade de registros coletados e armazenados no *dataset* utilizado para o treinamento dos modelos e também a utilização da validação cruzada Repeated K-Fold, os resultados devem ser cuidadosamente analisados para garantir que a aprendizagem não sofreu o efeito *overfitting*. *Overfitting* é quando o modelo aprende demais sobre os dados e pode ocorrer por diversos fatores. Entre os principais, destaca-se a utilização de algoritmos complexos para conjunto de dados pequenos. Nesse caso, o modelo adapta-se demais aos dados de treino e tem pouca capacidade de generalização de dados reais. Atrelado a isso, a baixa quantidade de registros em um *dataset* também pode contribuir para esse efeito. Uma outra causa bem comum são os ruídos dentro do conjunto de dados, por isso a etapa de manipulação dos dados é de suma importância.

A identificação de *overfitting* e *underfitting* (oposto do *overfitting*, que é quando o modelo não se adapta ao conjunto de dados) não é uma tarefa simples e trivial. É necessário utilizar recursos gráficos e/ou estatísticos. Para este trabalho, foi adotada a utilização da matriz de confusão, também conhecida como tabela de confusão, para analisar ocorrências de *overfitting*.

Uma matriz de confusão é uma matriz de dimensão n x n (na qual n é a quantidade de classes dentro do conjunto). É uma métrica voltada para modelos de classificação e tem como objetivo calcular a quantidade de falso positivo, falso negativo, verdadeiro positivo e verdadeiro negativo, além de fornecer a acurácia do algoritmos. Submetendo o algoritmo desenvolvido com melhor acurácia a uma matriz de confusão, tem-se a matriz mostrada na Figura 14.

Figura 14 – Matriz de confusão do algoritmo KNN

		Classe Esperada		
		Baixo	Moderado	Alto
Classe Prevista	Baixo	199	8	6
	Moderado	8	169	14
	Alto	9	11	199

Qtd Acertos:	**567**
Qtd Dados:	**623**
% Acertos:	**91,2%**

A análise de uma matriz de confusão deve ser ponderada para cada cenário estudado e examinar qual cenário deve ser mais penalizado. Por exemplo, em um cenário de aplicação de aprendizagem de máquina para previsão de um câncer, as classificações que foram previstas como falso positivo, são, de certa forma, menos impactantes ao paciente. Pois o paciente pode ser submetido às etapas iniciais do tratamento, mas, em algum momento, será descoberto que ele não apresenta a doença. Já no cenário contrário, em casos de falso negativo, a classificação é muito mais impactante, pois o paciente pode acabar recebendo alta do tratamento sem descobrir que apresenta a doença, dada a classificação errada do algoritmo.

Para este, os casos em que o algoritmo prevê que o risco para epidemias é baixo, enquanto na verdade é alto (falso baixo), devem ser mais penalizados e analisados do que o cenário inverso, dado a quantidade de ações de prevenção e combate às epidemias que deixariam de ser feitas/priorizadas por conta da saída do algoritmo. Na matriz de confusão apresentada acima, percebe-se que a quantidade prevista de falso baixo é igual a 6, enquanto a quantidade de previsões de falso alto é 9. Portanto, pode-se concluir, com base nesses números, que o algoritmo conseguiu se adaptar adequadamente aos dados identificando qual o cenário mais delicado dentro do conjunto de dados e sendo capaz de penalizá-lo.

Na matriz apresentada na Figura 14, percebe-se que existe uma grande diferença entre a quantidade de dados que foram previstos corretamente e a quantidade de dados que foram previstos erroneamente. Porém, a variação entre as previsões certas bem como a variação dos valores das previsões erradas são baixas. Portanto, pode-se inferir que o algoritmo KNN conseguiu se adaptar bem ao conjunto de dados sem sofrer o efeito *overfitting*, dado que os valores previstos seguem um padrão de escala sem muita dispersão. Além disso, analisando a sua matriz de confusão, percebe-se a boa adaptação do algoritmo com as classificações ponderando os falsos baixos e falsos altos.

6 CONCLUSÃO

No desenvolvimento deste trabalho foram apresentadas as principais técnicas, abordagens e etapas de desenvolvimento de um projeto de *machine learning*. Em relação ao aprofundamento e detalhamento de cada etapa e conceitos das estratégias utilizadas, foram abstraídas muitas explicações dado o objetivo principal do trabalho, que era conduzir um estudo empírico sobre a aplicação da *machine learning* para a previsão dos riscos de epidemias de dengue.

Portanto, ao final deste trabalho, conclui-se que a *machine learning* pode e tende a ser aplicada cada vez mais em diversas áreas do conhecimento. Com o crescimento exponencial e contínuo na produção de dados, existe a necessidade de utilizar abordagens inteligentes para tomar decisões baseadas em análises de dados. Para o cenário apresentado neste trabalho, conclui-se que o algoritmo utilizado que obteve o melhor desempenho de acertos poderia ser aplicado e utilizado pelo município para previsões de riscos de epidemias de dengue no local.

O escopo deste trabalho foi delimitado ao município de Bandeirantes, a quantidade de dados coletados foi baixa, e os algoritmos desenvolvidos foram os mais conhecidos pela literatura sem muito aprofundamento de ajuste de parâmetros e embasamento matemático por trás dos modelos. Porém, percebe-se que o resultado gerado após os experimentos é satisfatório. Uma vez que a quantidade de acertos do algoritmo KNN atinge 91,2% e é identificado que não existe a ocorrência de *overfitting* nos dados, sendo possível concluir que as previsões do algoritmo são confiáveis e ele pode ser aplicado.

O baixo poder de processamento utilizado no desenvolvimento deste trabalho, e a baixa quantidade de registros nos conjuntos de dados coletados impediram de certa forma que os algoritmos tivessem maior capacidade de aprendizado e generalização dos dados. Porém, ao final deste trabalho, surge a hipótese de que, com a coleta de mais registros e com o treinamento de mais algoritmos submetidos a diferentes hiperparâmetros, os resultados possam ser ainda melhores.

A obtenção, a manipulação dos dados e o desenvolvimento dos modelos foram feitos em ambiente local e de uma forma estática, porém estima-se que, em trabalhos futuros, as fontes de dados possam estar conectadas diretamente com a etapa de processamento dos dados e treinamento dos algoritmos em um ambiente em nuvem. Dessa forma, os algoritmos podem ser alimentados com novos dados frequentemente e os treinamentos podem ocorrer automaticamente e com uma vasta opção de combinações de modelos e parâmetros. Além de existir a disponibilidade de alto poder de processamento utilizando o ambiente *cloud*. A visualização das previsões do algoritmo também pode ser apresentada

por um painel ou estar integrada a algum sistema de monitoramento de epidemias.

O código desenvolvido neste trabalho pode ser acessado em: https://github.com/freddomingues/Forecasting_Risk_Dengue_Epidemic. Espera-se que ele possa ser aprimorado com contribuição de terceiros e utilizado e adaptado para outros cenários.

REFERÊNCIAS

[1] SANCHEZ-PINTO, L. N.; YUAN, L.; CHURPEK, M. M. Big data and data science in critical care. *CHEST*, v. 5, p. 1239–1248, 2018.

[2] Agência de Notícas do Paraná. *Paraná fecha período epidemiológico com 27.889 casos de dengue e 32 óbitos.* <https://www.aen.pr.gov.br/modules/noticias/article.php?storyid=114284&tit=Parana-fecha-periodo-epidemiologico-com-27.889-casos-de-dengue-e-32-obitos>. Acesso em 28/08/2021.

[3] Paraná Shop. *Paraná vive epidemia de Dengue enquanto já tem tecnologia para erradicar a doença.* <https://paranashop.com.br/2022/04/parana-vive-epidemia-de-dengue-enquanto-ja-tem-tecnologia-para-erradicar-a-doenca/>. Acesso em 03/07/2022.

[4] CONASS. *Diretrizes Nacionais para a prevenção e controle de epidemias de dengue.* <https://www.mppi.mp.br/internet/wp-content/uploads//2010/09/diretrizesnacionaisparaaprevenoecontroledeepdemiasdedengue.pdf>. Acesso em 25/02/2022.

[5] MONARD, M.; BARANAUSKAS, J. *Conceitos sobre aprendizado de máquinas em Sistemas Inteligentes: Fundamentos e Aplicações. Cap. 4.* [S.l.]: Editora Manole, 2003.

[6] FREITAS, A. A. Comprehensible classification models: A position paper. *SIGKDD Explor. Newsl.*, Association for Computing Machinery, New York, NY, USA, v. 15, n. 1, p. 1–10, mar 2014. ISSN 1931-0145. Disponível em: <https://doi.org/10.1145/2594473.2594475>.

[7] S.NEELAMEGAM, D. *Classification algorithm in Data mining: An Overview.* [S.l.]: IJPTT Journal, 2013.

[8] MUMMADISETTY ASTHA PURI, E. S. S. L. B. C. *A Hybrid Method for Compression of Solar Radiation Data Using Neural Networks.* [S.l.]: International Journal of Communications, Network and System Sciences, Vol.8 No.6, 2015.

[9] BHATKAR, A. P.; KHARAT, G. Detection of diabetic retinopathy in retinal images using mlp classifier. In: *2015 IEEE International Symposium on Nanoelectronic and Information Systems.* [S.l.: s.n.], 2015. p. 331–335.

[10] BURGES, C. J. *A Tutorial on Support Vector Machines for Pattern Recognition.* [S.l.]: Data Mining and Knowledge Discovery 2, 121–167, 1998.

[11] LINEAR Support Vector Classification. <https://scikit-learn.org/stable/modules/svm.html#svm-classification>. Accessed: 2022-09-18.

[12] ZADROZNY, B.; ELKAN, C. Transforming classifier scores into accurate multiclass probability estimates. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* New York, NY, USA: Association for Computing Machinery, 2002. (KDD '02), p. 694–699. ISBN 158113567X. Disponível em: <https://doi.org/10.1145/775047.775151>.

[13] SU, C. et al. Overview of bayesian network approaches to model gene-environment interactions and cancer susceptibility. In: . [S.l.: s.n.], 2012.

[14] MURDOCH, T. B.; DETSKY, A. S. The Inevitable Application of Big Data to Health Care. *JAMA*, v. 309, n. 13, p. 1351–1352, 04 2013.

[15] LITJENS, G. et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*, v. 42, p. 60–88, 2017.

[16] ZIVIANI, A. Desafios da ciência de dados aplicada a saúde digital. p. 17–22, 2019.

[17] SIRIYASATIEN1, P. et al. Dengue Epidemics Prediction: A Survey of the State-of-the-Art Based on Data Science Processes. *IEEE*, 2018.

[18] BUCZAK, A. L. et al. A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. *BMC medical informatics and decision making*, BioMed Central, v. 12, n. 1, p. 1–20, 2012.

[19] FOCKS, D. A.; CHADEE, D. D. Pupal survey: an epidemiologically significant surveillance method for aedes aegypti: an example using data from trinidad. *The American journal of tropical medicine and hygiene*, ASTMH, v. 56, n. 2, p. 159–167, 1997.

[20] SENG, C. M. et al. Pupal sampling for aedes aegypti (l.) surveillance and potential stratification of dengue high-risk areas in cambodia. *Tropical Medicine & International Health*, Wiley Online Library, v. 14, n. 10, p. 1233–1240, 2009.

[21] NAISH, S. et al. Climate change and dengue: a critical and systematic review of quantitative modelling approaches. *BMC infectious diseases*, BioMed Central, v. 14, n. 1, p. 1–14, 2014.

[22] MCLENNAN-SMITH, T. A.; MERCER, G. N. Complex behaviour in a dengue model with a seasonally varying vector population. *Mathematical biosciences*, Elsevier, v. 248, p. 22–30, 2014.

[23] KARIM, M. N. et al. Climatic factors influencing dengue cases in dhaka city: a model for dengue prediction. *The Indian journal of medical research*, Wolters Kluwer–Medknow Publications, v. 136, n. 1, p. 32, 2012.

[24] WONGKOON, S. et al. Weather factors influencing the occurrence of dengue fever in nakhon si thammarat, thailand. *Trop Biomed*, Citeseer, v. 30, n. 4, p. 631–41, 2013.

[25] XU, H.-Y. et al. Statistical modeling reveals the effect of absolute humidity on dengue in singapore. *PLoS neglected tropical diseases*, Public Library of Science San Francisco, USA, v. 8, n. 5, p. e2805, 2014.

[26] ALMEIDA, A. A. B. de; LEITE, L. B.; TUANI, M. *MANUAL DE METODOLOGIA DA PESQUISA APLICADA À EDUCAÇÃO*. [S.l.]: Faculdade Porto Feliz, 2016.

[27] CODECO, C. et al. Infodengue: A nowcasting system for the surveillance of arboviruses in brazil. *Revue d'Épidémiologie et de Santé Publique*, v. 66, p. S386, 2018. ISSN 0398-7620. European Congress of Epidemiology "Crises, epidemiological transitions and the role of epidemiologists". Disponível em: <http://www.sciencedirect.com/science/article/pii/S0398762018311088>.

[28] ESTAçãO Agrometeorológica UENP. 2023. Urlhttps://neat.uenp.edu.br/estacao/consulta/dados.

[29] AMARAL, F. *Introdução à Ciência de Dados*. [S.l.]: Alta Books, 2018.

[30] BIBLIOTECA Stats - Python. 2023. Urlhttps://docs.scipy.org/doc/scipy/index.html.