UNIVERSIDADE ESTADUAL DE MARINGÁ
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

JOÃO CARLOS ZAYATZ

Análise de correspondência múltipla e modelos de *machine learning* para notificações de dengue no Paraná.

Maringá
2022

JOÃO CARLOS ZAYATZ

Análise de correspondência múltipla e modelos de *machine learning* para notificações de dengue no Paraná.

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção do Departamento de Engenharia de Produção, Centro de Tecnologia da Universidade Estadual de Maringá, como requisito parcial para obtenção do título de Mestre em Engenharia de Produção.
Área de concentração: Engenharia de Produção

Orientador(a): Profa. Dra. Gislaine Camila Lapassini Leal
Coorientador(a): Paulo Cesar Ossani

Maringá
2022

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá - PR, Brasil)

Z39a

Zayatz, João Carlos

Análise de correspondência múltipla e modelos de *machine learning* para notificações de dengue no Paraná / João Carlos Zayatz. -- Maringá, PR, 2022.

81 f.: il. color., figs., tabs.

Orientadora: Profa. Dra. Gislaine Camila Lapassini Leal.

Coorientador: Prof. Dr. Paulo Cesar Ossani.

Dissertação (Mestrado) - Universidade Estadual de Maringá, Centro de Tecnologia, Departamento de Engenharia de Produção, Programa de Pós-Graduação em Engenharia de Produção, 2022.

1. Dengue. 2. Saúde - Sistemas de informação. I. Leal, Gislaine Camila Lapassini, orient. II. Ossani, Paulo Cesar , coorient. III. Universidade Estadual de Maringá. Centro de Tecnologia. Departamento de Engenharia de Produção. Programa de Pós-Graduação em Engenharia de Produção. IV. Título.

CDD 23.ed. 610.285

Márcia Regina Paiva de Brito - CRB-9/1267

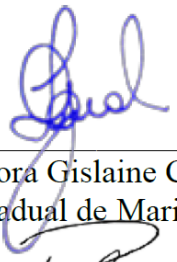
FOLHA DE APROVAÇÃO

JOÃO CARLOS ZAYATZ

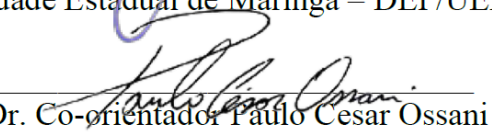
Análise de correspondência múltipla e modelos de *machine learning* para notificações de dengue no Paraná.

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção do Departamento de Engenharia de Produção, Centro de Tecnologia da Universidade Estadual de Maringá, como requisito parcial para obtenção do título de Mestre em Engenharia de Produção pela Banca Examinadora composta pelos membros:

BANCA EXAMINADORA

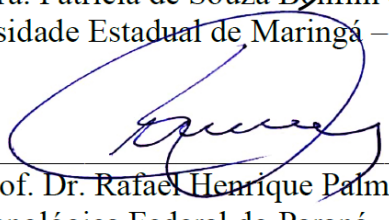


Profa. Dra. Orientadora Gislaine Camila Lapasini Leal
Universidade Estadual de Maringá – DEP/UEM



Prof. Dr. Co-orientador Paulo Cesar Ossani
Universidade Estadual de Maringá – DES/UEM

Profa. Dra. Patrícia de Souza Bonfim de Mendonça
Universidade Estadual de Maringá – DAB/UEM



Prof. Dr. Rafael Henrique Palma Lima
Universidade Tecnológica Federal do Paraná – DAENP-LD/UTFPR

Aprovada em 08 de março de 2022.
Local da defesa: Sala de Projeção, Bloco 19, *campus* da Universidade Estadual de Maringá.

DEDICATÓRIAS

Ofereço para minha família, que sempre tem me apoiado.

AGRADECIMENTOS

Agradeço a Deus pela vida, pela força no caminho e por iluminar a estrada.

À minha mãe, Erenita Alves de Souza, mulher trabalhadora e mãe zelosa, e ao meu Pai, Nestor Zayatz, pelos valores e educação concedidos.

Aos meus irmãos, Lenita, Lilian, João Paulo e Lirian, pelo apoio, desde sempre, para tudo.

Ao Miguel e a Luiza, sobrinhos que me fazem muito mais feliz, e ao Tiago Sato e Daniele Schmidt, por cuidarem tão bem deles junto com meus irmãos.

Ao Heraldo Fonseca de Farias, pelo companheirismo, cumplicidade, incentivo e otimismo de todos os dias. À Dona Cida, Ana Vitória e Josiane, por toda a ajuda que me ofereceram em Paranavaí.

Aos docentes do PGP/UEM, pelos conhecimentos compartilhados. Em especial, Daniele Cristina Tito Granzotto (*in memorian*), que foi a primeira orientadora deste trabalho e nos deixou cedo, não sem antes inspirar o caminho da pesquisa científica.

À professora orientadora Camila Lapasini Leal, pelo compromisso e empenho ao assumir a orientação para que esta pesquisa continuasse, pela paciência, compreensão e ensinamentos.

Ao professor coorientador e amigo, Paulo César Ossani, pela amizade, pela disposição em ensinar, por se fazer presente, por ter sido solícito em todas as vezes que o procurei e por enxergar além de sua função.

Aos colegas e amigos de trabalho, pela compreensão cotidiana, em especial Silvia Jaqueline Flor e Eliane Hillmann Garcia.

Aos amigos e colegas de turma do PGP e da UEM, por terem dividido esta etapa, em especial: Guilherme Melluzzi Neto, Lorena Mazia Enami, Marco Aurelio Valles Leal e Ana Carolina Neves Carnelossi.

Aos amigos pessoais, pelo apoio na caminhada, em especial: Felipe Veiga da Fonseca, Calvin Coelho Fernandes Bonifácio, Gustavo Arguelho, Guilherme Futoshi, Lizeane Heren Cândido Pereira, Rafael Pereira e Elen Lopes.

Ao Jocimar Bardi Junior, por ser profissional dedicado e me ajudar a encontrar os alicerces para seguir nesta etapa.

À 15ª Regional de Saúde do Paraná, pela disponibilização de dados ao Departamento de Estatística da Universidade Estadual de Maringá - UEM, os quais foram utilizados nesta pesquisa, de modo multidisciplinar e colaborativo.

À professora Eniuce Menezes de Souza, pela contribuição na atualização da base de dados.

EPÍGRAFE

Nunca, jamais desanimeis, embora venham ventos contrários.

(Madre Paulina).

Análise de correspondência múltipla e modelos de *machine learning* para notificações de dengue no Paraná.

RESUMO

A dengue é a arbovirose de maior alcance mundial, causando perdas de vidas humanas e impactos econômicos diretos e indiretos, principalmente em regiões de clima tropical. Bancos de dados de registros de notificações contribuem para o monitoramento e pesquisa sobre a dengue. Neste estudo, dados do Sistema de Informações de Agravos de Notificações (SINAN) permitem analisar registros de notificações da doença no estado do Paraná, região Sul do Brasil, no ano de 2019-2020. Com um total de 366.760 notificações, este período representou recorde na série histórica de incidência para o estado. Os registros contam com informações socioeconômicas e clínicas, a partir das quais elabora-se uma pesquisa quantitativa. O estudo é dividido em duas etapas. Na primeira, a análise estatística descritiva é utilizada para oferecer um panorama dos registros de notificações. São apontadas lacunas de preenchimento do SINAN, que representam pontos de atenção, uma vez que a análise dos dados da dengue depende da integridade e confiabilidade dos registros. Para complementar esta fase, optou-se por utilizar a análise de correspondência múltipla, por meio de método gráfico, que permite visualizar relações entre as variáveis do SINAN para notificações de dengue. O objetivo é indicar associações que ficam ocultas na análise estatística descritiva e que, quando evidenciadas, podem ser exploradas em pesquisas da área de saúde. Na segunda etapa, modelos de classificação supervisionada em *machine learning* são proposto, com o objetivo de contribuir com possibilidades de automatização da tarefa de classificação de notificações de dengue, levando em conta quatro níveis de classificação do agravo: descarte, dengue clássica, dengue com sinais de alarme e dengue grave. A construção é feita a partir do classificador probabilístico Naïve Bayes. Os parâmetros de avaliação dos modelos contam com índice Kappa de concordância estatística. Os resultados evidenciam que a tarefa de classificar notificações em “dengue clássica” ou “descartado” mostrou-se complexa, a partir das variáveis estruturadas do SINAN. Por outro lado, um modelo mais robusto foi obtido para classificação de gravidade de dengue, em casos confirmados. Estes resultados são evidenciados pela análise da distribuição espacial das amostras, a partir de uma técnica *projection pursuit* com o índice LDA.

Palavras-chave: dengue, notificação de doença, sistemas de informação em saúde, análise multivariada, aprendizado de máquina supervisionado.

Multiple correspondence analysis and machine learning models for dengue fever notifications in the State of Paraná.

Dengue fever is the most widespread arbovirus in the world, causing loss of human life and direct and indirect economic impacts, especially in tropical regions. Databases of notification records contribute to dengue monitoring and research. In this study, data from the Information System of Diseases of Notifications (SINAN) allow analyzing records of notifications of the disease in the state of Paraná, southern region of Brazil, in the year 2019-2020. With a total of 366,760 notifications, this period represented a record in the historical series of incidence for the state. The records contain socioeconomic and clinical information, from which a quantitative survey is carried out. The study is divided into two stages. In the first, descriptive statistical analysis is used to provide an overview of notification records. Gaps in filling in the SINAN are pointed out, which represent points of attention, since the analysis of dengue fever data depends on the integrity and reliability of the records. To complement this phase, it was decided to use multiple correspondence analysis, using a graphical method, which allows the visualization of relationships between SINAN variables for dengue notifications. The objective is to indicate associations that are hidden in the descriptive statistical analysis, when evidenced, can be explored in health research. In the second stage, supervised classification models in machine learning are proposed, with the objective of contributing with possibilities of automating the task of classifying dengue fever notifications, considering four levels of disease classification: discard, classic dengue, dengue with warning signs and severe dengue. The construction is made using the Naïve Bayes probabilistic classifier. The evaluation parameters of the models have a Kappa index of statistical agreement. The results show that the task of classifying notifications into “classical” or “discarded” proved to be complex, based on the structured variables of the SINAN. On the other hand, a more robust model was obtained for classifying dengue fever severity in confirmed cases. These results are evidenced by the analysis of the spatial distribution of the samples, from a projection pursuit technique with the LDA index.

Keywords: dengue fever, disease notification, health information systems, multivariate analysis, supervised machine learning.

LISTA DE TABELAS E QUADROS

TABELAS

Capítulo 3

Capítulo 4

QUADROS

CAPÍTULO 1

CAPÍTULO 4

LISTA DE FIGURAS

CAPÍTULO 2

CAPÍTULO 3

CAPÍTULO 4

LISTA DE ABREVIATURAS E SIGLAS

CF	*Classificação Final para Notificações de Dengue*
CSV	*Comma-separated values*
DG	*Dengue Grave*
DSA	*Dengue com Sinais de Alarme*
EF	*Ensino Fundamental*
EM	*Ensino Médio*
FN	*Falso Negativo*
FP	*Falso Positivo*
IA	*Inteligência Artificial*
K	*Índice Kappa*
KAP	*Knowledge, attitudes and practices*
LDA	*Discriminant analysis*
ML	*Machine Learning*
NB	*Naïve Bayes*
PO	*Pesquisa Operacional*
SESA/PR	*Secretaria de Estado da Saúde do Paraná*
SINAN	*Sistema de Informações de Agravos de Notificações*
TI	*Tecnologia da Informação*
VN	*Verdadeiro Negativo*
VP	*Verdadeiro Positivo*

SUMÁRIO

1

INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

Transmitido por mosquitos do gênero *Aedes*, o vírus da dengue infecta aproximadamente 400 milhões de pessoas por ano no mundo. A frequência de infecção varia de acordo com condições ambientais, socioeconômicas e demográficas locais (MURRAY *et al*., 2013; BHATT *et al*., 2013; NEALON *et al*., 2022). Episódios sintomáticos correspondem à aproximadamente 100 milhões de casos por ano (MESSINA *et al*., 2019), enquanto óbitos variam entre 10.000 e 50.000 (BHATT *et al*., 2013; NEALON *et al*., 2022). A estimativa é de que a Dengue esteja presente em mais de 125 países (STANAWAY et al., 2016).

Mais da metade dos casos de dengue registrados nas Américas ocorre no Brasil (BAVIA *et al*., 2020). Embora a região sul tenha a menor incidência de casos em relação às demais regiões brasileiras, a quantidade de notificações nesta área cresceu na última década (MARTIN *et al*., 2022). Dentre os três estados da região Sul, o Paraná apresenta a maior quantidade de notificações, casos confirmados e óbitos (VECCHIA *et al*., 2018). Uma quantidade superior à 360.000 notificações foi registrada entre o final do mês de julho do ano de 2019 e o início do mês de julho do ano de 2020 (PARANÁ, 2020), o que representou recorde para a série histórica do estado.

As notificações de dengue no Paraná são registradas pela Secretaria de Estado da Saúde (SESA) no Sistema de Informação de Agravos de Notificações (SINAN). Estas

informações são coletadas em unidades de saúde e hospitais, a partir do preenchimento do SINAN por profissionais que atuam nestas instituições, durante os atendimentos clínicos. Tais dados possibilitam visualizar o panorama da dengue no estado e servem de base para a construção de boletins epidemiológicos, além de apoiar a tomada de decisão da vigilância epidemiológica estadual.

Sistemas de informação em saúde, que coletam rotineiramente informações como subprodutos da prestação de serviços na área de saúde, fornecem possibilidades ilimitadas para a pesquisa epidemiológica (SCHMIDT *et al*., 2019). Um exemplo de conjunto de dados como esse é o próprio SINAN. A sumarização e a descrição destes dados frequentemente é abordada via estatística descritiva. Com o avanço da Tecnologia da Informação (TI), análises estatísticas de dados médicos têm agregado recursos da Inteligência Artificial (IA) para contribuir de modo significativo com a análise de conjuntos amplos e complexos (OBERMEYER; EMANUEL, 2016). O aprendizado de máquina, ou *Machine Learning* (ML), como subárea da IA, permite o processamento de registros epidemiológicos para a construção de modelos preditivos (BAGHERZADEH-KHIABANI *et al*., 2016), os quais podem ser utilizados por profissionais da área saúde para apoiar a tomada de decisão.

Neste contexto, a presente pesquisa foi desenvolvida para analisar as notificações de dengue no estado do Paraná e propor a construção de modelos preditivos de ML para a classificação de notificações da doença, a partir das informações presentes no SINAN. Os dados referem-se ao período epidemiológico compreendido entre 28 de julho de 2019 à 01 de agosto de 2020, também chamado de ano epidemiológico de 2019/2020. O ano em questão é significativo para o estudo do impacto da doença no estado, uma vez que representou período recorde de notificações para a série história. Na primeira etapa da pesquisa, foi efetuada a estatística descritiva dos dados, que leva à compreensão do cenário de registrado de notificações de dengue no estado, para o período considerado. Esta tarefa é complementada pela análise de correspondência múltipla entre as variáveis, sendo esta uma técnica multivariada que permite a visualização de associações entre os atributos presentes no banco de dados da dengue do SINAN.

Na segunda etapa, são construídos modelos de classificação de ML para notificações de dengue, a partir das variáveis estruturadas disponíveis no SINAN. A classificação supervisionada foi utilizada para a geração de três modelos, com objetivos distintos. A ideia é treinar modelos de classificação capazes de diferenciar casos confirmados ou descartados de

notificações de dengue, ou mesmo atribuir grau de gravidade para notificações confirmadas.

Para a área de Engenharia de Produção e para o meio científico, esta pesquisa contribui em uma vertente multidisciplinar, uma vez que utiliza informações da área de saúde, além de recursos computacionais e ferramentas estatísticas, para oferecer um estudo quantitativo aplicado às notificações de dengue no Paraná. Trabalhos que exploram os dados da vigilância epidemiológica de dengue no Paraná, a partir das ferramentas propostas, mostram-se escassos. Portanto, a contribuição da pesquisa para o conhecimento científico reside no preenchimento desta lacuna, a partir de uma abordagem multidisciplinar.

Para profissionais, organizações privadas e gestores públicos da área da saúde, a pesquisa apresenta um panorama das notificações de dengue no estado do Paraná, para o período estudado. Possibilita, ainda, visualizar o relacionamento entre as variáveis presentes em um banco de dados de notificações dengue, além de explorar possibilidades de construção de modelos de ML para classificar notificações. Neste sentido, o trabalho aponta os limites encontrados, em termos da complexidade em se criar um classificador computacional para automatizar a tarefa de confirmação ou descarte de casos notificados, com o uso de atributos clínicos de triagem ambulatorial, sem adição de exames laboratoriais específicos. Por outro lado, aponta possibilidades para utilização de classificação automatizada em termos de gravidade de dengue, em casos confirmados. A pesquisa pretende gerar conhecimento para subsidiar processos de tomada de decisão sobre a dengue e sugerir pesquisas futuras para a temática abordada.

1.2 OBJETIVOS DA PESQUISA

Esta pesquisa pretende responder à seguinte pergunta: **de que modo a análise de correspondência múltipla e os modelos de ML podem contribuir com a área da saúde, especificamente do ponto de vista das notificações de dengue no estado do Paraná?** Para responder à esta pergunta, têm-se o objetivo geral da pesquisa:

- Construir modelos de *machine learning* para expandir a análise sobre notificações de dengue no Estado do Paraná.

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

- Explorar a literatura sobre a dengue e seus impactos socioeconômicos, além das ferramentas estatísticas e computacionais abordadas;

- Efetuar o pré-processamento da base de dados do SINAN, fornecida pela SESA/PR;
- Apresentar a estatística descritiva dos casos de dengue no Paraná;
- Exibir a correspondência entre variáveis de dengue presentes no SINAN, por meio de análise de correspondência múltipla;
- Construir modelos de classificação para notificações de dengue;
- Avaliar os resultados dos modelos de classificação.

1.3 ESTRUTURA DA PESQUISA

Esta dissertação está estruturada no modelo *multipaper*. Dois artigos foram produzidos para responder à pergunta de pesquisa e cumprir com os objetivos propostos. A síntese dos artigos está descrita no Quadro 1.

Quadro 1: Síntese do artigos

	Artigo 1	**Artigo 2**
Título	Análise das notificações de dengue no estado do Paraná: estudo de caso a partir da análise estatística descritiva e análise de correspondência múltipla.	Modelos de *machine learning* para notificações de dengue no estado do Paraná.
Objetivo	Descrever as notificações de dengue no estado do Paraná e as associações entre as variáveis de notificações de dengue do SINAN.	Propor modelos de classificação supervisionada de ML para notificações de dengue.
Método	Pré-processamento e seleção dos dados; apresentação e análise da estatística descritiva; apresentação e análise da análise de correspondência múltipla, via método gráfico.	Pré-processamento e seleção dos dados; Construção de modelos de classificação; Avaliação dos modelos propostos.
Contribuições	Panorama socioeconômico das notificações de dengue no Paraná; Lacunas no preenchimento do SINAN; Associação entre variáveis de notificação de dengue.	Proposta de modelos em ML para apoiar a tomada de decisão médica; Avaliação da complexidade dos dados para classificação e notificações; Indicação de possibilidades de pesquisas futuras.

Fonte: o autor (2022).

Além da presente seção introdutória, outros 4 capítulos são apresentados na sequência:

- Capítulo 2: Apresentação da metodologia utilizada na pesquisa. O objetivo é descrever, além da caracterização metodológica, os materiais utilizados e o fluxo de tarefas que levaram à geração dos dois artigos produzidos. Uma vez que nem todos as variáveis e amostras integraram as análises, neste capítulo, também, é abordado o processo de preparação dos dados.
- Capítulo 3: Exposição do artigo 1, fruto da análise estatística descritiva das notificações de dengue no Paraná, a partir de dados do SINAN, complementado pela análise de correspondência múltipla, para os mesmos dados. Neste capítulo, os gráficos de análise de correspondência múltipla contribuem para visualizar as relações entre as variáveis.

- Capítulo 4: Exposição do artigo 2, que aborda a construção de modelos de classificação supervisionada em ML para notificações de dengue. Estes modelos têm o objetivo de classificar notificações quanto à confirmação ou descarte para a doença e, ainda, classificar a gravidade para casos confirmados. São apresentados, também, parâmetros de avaliação dos modelos e gráficos, que utilizam a técnica de *projection pursuit* com o índice da análise discriminante linear, *linear discriminat analysis* (LDA), que permitem visualizar a distribuição espacial das amostras entre as classes de classificação, em duas dimensões.

- Capítulo 5: Apresenta as considerações finais, revisando os resultados e fazendo um apanhado dos dois artigos, para oferecer uma perspectiva unificada da pesquisa, além de indicar possibilidades para a continuidade do estudo.

2

MÉTODO DE PESQUISA

2.1 CARACTERIZAÇÃO DA PESQUISA

Para caracterizar a pesquisa, foi utilizada a abordagem de Silveira e Córdova (2009), que diferencia os vários tipos de pesquisa quanto à sua abordagem, sua natureza, seus objetivos e seus procedimentos, para que fosse possível selecionar a modalidade de pesquisa adequada aos objetivos.

Quanto à abordagem, a pesquisa é quantitativa, por se concentrar objetivamente na estatística descritiva e construção de modelos de ML para notificações de dengue. Quanto à natureza, é uma pesquisa aplicada, com fulcro a apoiar a tomada de decisão na área de saúde em relação às notificações de dengue. Quanto aos objetivos, a pesquisa é descritiva, haja vista que pretende descrever as notificações de dengue no Paraná, e explicativa, uma vez que demonstra o relacionamento entre as variáveis em estudo. Quanto aos procedimentos, trata-se de um estudo de caso, a partir de um levantamento das ocorrências de Dengue, registrados no Estado do Paraná.

2.2 MATERIAIS

Com relação aos materiais utilizados, tem-se:

- Banco de dados de agravos e notificações da SESA/PR para a Dengue, referente ao ano epidemiológico de 2019/2020, registrados no SINAN. Estes dados referem-se às informações de seres humanos, porém sem identificação específica do cidadão;
- *Softwares* de tratamento estatístico, sendo o RStudio (RSTUDIO TEAM, 2020) e o WEKA (FRANK; HALL; WITTEN, 2016), ambos gratuitos, incluindo as bibliotecas para mineração de dados para estas aplicações.

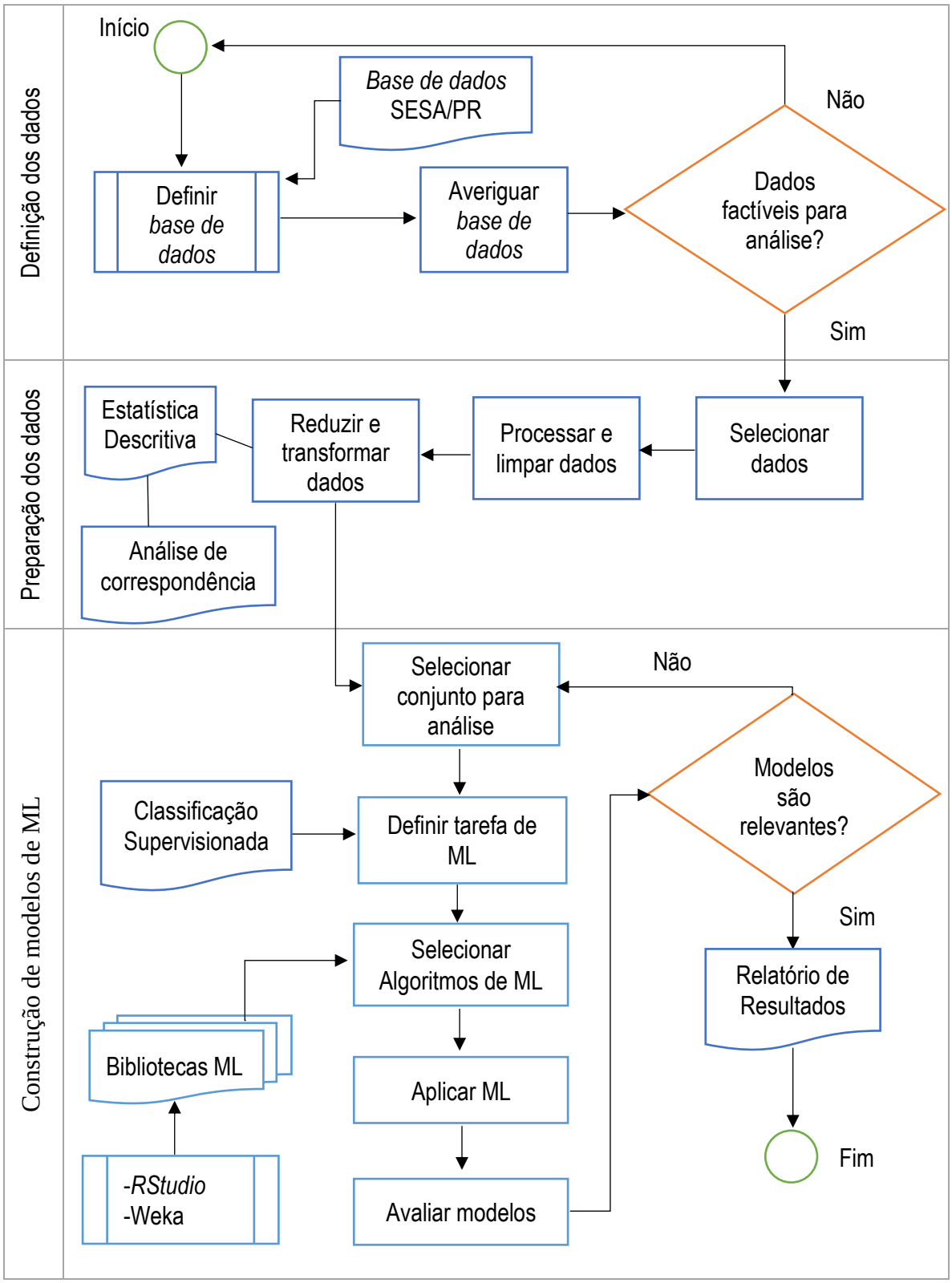
A base de dados foi disponibilizada ao Departamento de Estatística da Universidade Estadual de Maringá pela SESA/PR. Uma vez que as ferramentas computacionais utilizadas são gratuitas, esta pesquisa pretende contribuir com a geração de conhecimento científico, sem acarretar em custos extras em termos de aquisição de ferramenta para análise de dados. As tarefas de pré-processamento dos dados, elaboração da estatística descritiva e geração dos gráficos para a análise de correspondência foram elaboradas no *software* RStudio, enquanto a construção dos modelos de classificação foi efetuada no *software* WEKA.

2.3 ETAPAS DA PESQUISA

As etapas metodológicas podem ser divididas em três fases:

1) Definição dos dados: onde se definem quais serão os dados em estudo e investiga-se se a base de dados é factível de ser utilizada na etapa seguinte;
2) Preparação dos dados: onde é feita a seleção, limpeza, preparação, redução e transformação dos dados. Ao fim desta etapa, é gerada a estatística descritiva, que oportuniza compreender o cenário das notificações de dengue, complementada pela análise de correspondência.
3) Construção de modelos de ML: onde são aplicadas ferramentas de ML aos conjuntos de dados selecionados, a fim de construir modelos supervisionados para classificar notificações de dengue. A Figura 1 apresenta o fluxograma do método de pesquisa.

Figura 1: Fluxograma da metodologia.



Fonte: o autor (2022).

2.4 DEFINIÇÃO E PREPARAÇÃO DOS DADOS

Os dados brutos foram fornecidos pela SESA/PR, em planilha eletrônica, formato de caracteres separados por vírgula, *comma-separated values* – csv. Estes dados correspondem aos registros de agravos e notificações para a Dengue no Paraná, nos anos de 2019 e 2020. Cada amostra representa uma notificação para Dengue. Para assegurar a privacidade, não foram fornecidas instâncias com as informações pessoais, tais como: nome, endereço, número do cadastro de pessoa física – CPF, filiação e número de telefone. O banco de dados resultante foi exportado para tratamento estatístico pelo *software* RStudio. Estavam dispostos em modo matricial, onde cada coluna representa um atributo, ou variável, cada linha representa uma amostra, ou instância, e cada célula apresenta o valor atribuído à instância para um determinado atributo. Foi verificado que a base de dados era factível para análise, uma vez que os dados estavam apresentadas de modo estruturado, representando variáveis categóricas bem definidas.

A partir dos dados brutos, foi realizada a etapa de seleção dos dados, na qual foram selecionados os atributos que seriam abordados pela pesquisa. O banco de dados original conta com 146 atributos. Foram retirados 95 atributos, que continham informações sem capacidade discriminatória, que não retornaram respostas objetivas, que correspondem à dados não estruturados ou que foram considerados redundantes, tais como: controle de numeração sequencial da notificação; código de identificação do agravo; código da unidade federativa, do país, do município e da unidade de saúde; indicação redundante de sintomas; campos de observações; campos reservados para informações de outras arboviroses; entre outros. Ao fim, foram considerados 51 atributos, que podem ser divididos em oito subconjuntos de variáveis: socioeconômicos, sinais clínicos de dengue clássica, doenças pré-existentes, hospitalização, classificação final do agravo, evolução clínica, sinais clínicos para dengue com sinais de alarme (DSA) e sinais clínicos para dengue grave (DG).

Na etapa de pré-processamento e transformação dos dados, foram executados procedimentos para assegurar a integridade e qualidade dos conjuntos de análise. Assim, buscou-se tratar as seguintes situações: dados em falta, atributos com muitos valores ausentes, instâncias com valores em falta, dados avaliados como inconclusivos ou com preenchimentos imprecisos, e variáveis categóricas que estariam melhor representadas em níveis ou faixas, tais como idade e escolaridade. A partir dos dados pré-processados, foram efetuadas as pesquisas que originaram os artigos 1 e 2, descritos na sequência desta dissertação. Outros

procedimentos metodológicos específicos e atinentes a cada artigo foram mencionados no decorrer dos mesmo.

3

ARTIGO 1

Análise das notificações de dengue no Paraná: estudo de caso a partir da estatística descritiva e análise de correspondência múltipla.

RESUMO

Dentre os estados da região Sul do Brasil, o Paraná é destaque no número de notificações de dengue, arbovirose de maior incidência mundial. O objetivo deste estudo é descrever as notificações de dengue no Paraná no ano epidemiológico 2019-2020. Os registros foram extraídos do Sistema de Agravo de Notificações (SINAN). Além da estatística descritiva, a análise de correspondência múltipla foi utilizada para explorar relações entre 51 variáveis, presentes no sistema, incluindo informações socioeconômicas e clínicas. O período representou recorde na série histórica para o estado, com 366.760 notificações, das quais 66,59% foram confirmadas. Óbitos pelo agravo foram 198, o que representa 0,054% dos notificados. O critério de confirmação adotado, na maior parte dos casos, foi o clínico-epidemiológico, utilizado em 65,88% do total. Exames laboratoriais específicos foram empregados como critério de confirmação em 27,31% dos indivíduos. Lacunas no preenchimento das notificações estão descritas, das quais as variáveis de escolaridade e hospitalização representam pontos de maior atenção. A análise de correspondência múltipla indica que os sinais clínicos para casos alarmantes (DSA) ou graves (DG) de dengue estão menos relacionados aos níveis descritos como dengue clássica e descartados, enquanto estes dois estão mais próximos entre si. Foram, também, identificado maior associação entre as variáveis de doenças pré-existentes e os níveis de DSA e DG. Além disso, relações entre

variáveis socioeconômicas específicas e sinais clínicos de dengue clássica foram observadas e estão descritas nos resultados. A pesquisa pretende contribuir para oferecer um panorama do registro das notificações de dengue no Paraná, para o ano epidemiológico abordado e sugerir outras possibilidades para análises exploratórias posteriores.

Palavras-chave: dengue, notificação de doença, sistemas de informação em saúde, análise multivariada, análise de correspondência múltipla.

3.1 INTRODUÇÃO

A Dengue é uma doença viral, transmitida de mosquitos para humanos, que se espalhou em áreas tropicais do mundo nos últimos 60 anos e agora afeta mais da metade da população mundial (MESSINA *et al*., 2019). É uma das doenças infecciosas globais de crescimento mais rápido, que se estabeleceu de modo sólido em grandes centros urbanos. A incidência estimada está em 400 milhões de casos por ano no mundo (WILDER-SMITH *et al*., 2019), dos quais 25% são clinicamente aparentes (BHATT *et al*., 2013).

No Estado do Paraná, o ano epidemiológico 2019/2020, que compreende o período que vai do segundo semestre de 2019 ao primeiro semestre de 2020, apresentou recorde de casos notificados de dengue, em relação à série histórica para o estado, com 366.760 notificações. Isso implica no aumento do acionamento dos sistemas de saúde, além de significar maior acometimento da população, impacto na força produtiva, hospitalizações e óbitos pelo agravo. Os registros destas notificações estão presentes no SINAN.

Uma vez que bancos de dados médicos possuem riqueza de informações capaz de gerar conhecimento para a área de saúde, uma questão que emerge neste contexto é: **de que modo a análise estatística dos registros de notificações de dengue no Paraná, presentes no SINAN, pode contribuir para apoiar a tomada de decisão na área de saúde?**

Para responder esta pergunta, o presente artigo tem por objetivo descrever as notificações de dengue no estado do Paraná para o ano epidemiológico da dengue 2019/2020. O objetivo é explorar o perfil destas notificações, a partir de variáveis contidas no SINAN, por meio da estatística descritiva, a fim de gerar conhecimento para a área de gestão em saúde. Em paralelo, busca-se identificar lacunas de preenchimento do sistema. Para complementar a análise da estatística descritiva, optou-se por apresentar um estudo de análise

de correspondência múltipla com as variáveis contidas SINAN.

Na literatura, estão presentes diversos estudos em que a análise de correspondência foi aplicada para investigar associações em conjuntos de dados de arboviroses. Estes conjuntos de dados podem incluir: variáveis socioeconômicas, socioculturais, sociodemográficos, socioambientais, sinais clínico e outras. Yaseen *et al*. (2014) utilizaram regressão logística e análise de correspondência para associar sinais clínicos da fase inicial de infecção por chinkungunya com sinais clínicos identificados no pós-infecção. Higuera-Medieta *et al*. (2016) utilizaram análise de correspondência para determinar fatores sociodemográficos aos níveis de conhecimentos, atitudes e práticas – knowledge, attitudes e practices - KAP, relacionadas à dengue. Fritzell *et al*. (2016) realizaram uma análise de correspondência múltipla, associada à clusterização hierárquica e regressão logística, para estudar a associação entre experiências, práticas e percepções relacionadas à doenças transmitidas por mosquitos e identificar fatores sociodemográficos, cognitivos e ambientais que possam estar associados ao engajamento em comportamentos de proteção, durante um surto de chikungunya.

Em Wu *et al*. (2017), é encontrado um estudo de análise de correspondência para associação entre comportamentos culturais e variáveis socioeconômica presentes em diferentes grupos de minorias étnicas à doenças transmitidas por mosquitos. Em Siswantining *et al*. (2018), é apresentada um método de análise de correspondência múltipla e regressão logística binária para prever riscos de hospitalização em diagnóstico com alto custo, incluindo a dengue hemorrágica, a partir de notificações ambulatoriais. Em Nava-Dotor *et al*. (2021), a análise de rede foi complementada pela análise de correspondência, para investigar associações entre conhecimentos, atitudes e práticas sobre doenças transmitidas por insetos.

Neste artigo, além da estatística descritiva, é apresentada a análise de correspondência múltipla para notificações de dengue, considerando 51 variáveis categóricas, que incluem os seguintes grupos de variáveis: sinais clínicos de dengue clássica, doenças pré-existentes, sinais clínicos de dengue com sinais de alarme e sinais clínicos de dengue grave. O objetivo é visualizar o relacionamento entre as variáveis. Três cenários são apresentados para expor diferentes perspectivas das associações entre as variáveis. Os resultados sugerem direcionamentos para pesquisas futuras em termos das associações identificadas.

Enquanto a base de dados utilizada para a etapa de estatística descritiva contou com 366.760 amostras, o conjunto para a análise de correspondência possui 55.071 observações, resultantes dos procedimentos de preparação dos dados para análise. Os resultados permitiram

visualizar o perfil das notificações de dengue no Estado do Paraná, para o período estudado e o relacionamento entre as variáveis presentes no SINAN.

3.2 REFERENCIAL TEÓRICO

3.2.1 Dengue: contextualização

A Dengue é causada por um vírus de RNA (ácido ribonucleico) de fita simples, do gênero Flavivírus, denominado Vírus da Dengue (WICHMANN *et al*., 2007). Recebe a classificação de doença arboviral, por ser transmitida de hospedeiros artrópodes para vertebrados. O principal vetor é o *Aedes aegypti*, mosquito peridomiciliar diurno, capaz de picar várias pessoas em um curto espaço de tempo e de se reproduzir em diversos recipientes de fabricação humana onde seja possível coletar água (WILDER-SMITH *et al*., 2019). Outro vetor comum é o mosquito *Aedes albopictus*, que, embora seja menos eficiente para transmissão, está expandindo seu alcance geográfico em climas tropicais e temperados (WILDER-SMITH *et al*., 2019). O clima adequado e os deslocamentos humanos são fatores que impulsionam as áreas infestadas.

Estão descritos na literatura quatro sorotipos diferentes do vírus da Dengue, sendo os genótipos denominados de DENV-1, DENV-2, DENV-3 e DENV-4 (SOUZA-NETO *et al*., 2019). Após uma infecção primária, por qualquer sorotipo DENV, os indivíduos desenvolvem imunidade contra reinfecção pelo mesmo sorotipo. Por outro lado, a infecção secundária por sorotipo diferente gera o risco de desenvolver Dengue na forma mais grave. Este risco aumenta conforme maior é o tempo entre a infecção primária e a secundária (FLIPSE; SMIT, 2015).

A Dengue pode ter três classificações, que determinam os protocolos clínicos a serem observados pelos sistemas de saúde: Dengue clássica, Dengue com sinais de alarme e Dengue Grave. Um guia completo destes protocolos é disponibilizado pela *World Health Organization* (2012). De modo sucinto, a Dengue clássica é caracterizada por febre alta, dores musculoesqueléticas, dor retro orbital, dores de cabeça, erupção cutânea. Manifestações hemorrágicas podem indicar quadros clínicos agravados (AHMED *et al*., 2008).

A Dengue é frequentemente confundida, sobretudo nos estágios iniciais, com outros

estados febris virais, dificultando o manejo clínico e o controle da transmissão. Sintomas mais característicos, como dor retro orbital e petéquias, aparecem em estágios posteriores aos primeiros sintomas e nem sempre estão presentes, o que demanda a adoção de exames laboratoriais para diagnóstico mais assertivo (TANNER *et al*., 2008). As atividades nos níveis de triagem e de atenção primária e secundária, onde os casos são avaliados pela primeira vez, são essenciais para determinar o resultado clínico. Respostas bem gerenciadas de linha de frente podem reduzir o número de internações e salvar vidas, além de permitir a identificação precoce de surtos (*WORD HEALTH ORGANIZATION*, 2012).

O processo de desenvolvimento de vacinas para Dengue é considerado complexo, devido, entre outros fatores, à existência dos múltiplos sorotipos DENV (FLIPSE; SMIT, 2015). Há uma vacina licenciada para Dengue, porém o imunizante tem recebido estímulos internacionais para revisão com relação à eficácia. Uma revisão sobre este tema é encontrada em Thomas e Yoon (2019). Não obstante, grande parte das medidas atreladas ao controle da Dengue incluem o controle de vetores. O *Aedes aegypti* tem sido historicamente o principal vetor em quase todas as principais epidemias de arboviroses: Dengue, Zika, Chikungunya e Febre Amarela (SOUZA-NETO *et al*., 2019).

Os principais fatores diretamente ligados à proliferação da Dengue descritos na literatura são: crescimento populacional, alta densidade populacional, migração da zona rural para áreas urbanas, degradação de ambientes urbanos, ausência de água encanada confiável, programas de controle de mosquitos desorganizados e com financiamento inadequado (WILDER-SMITH *et al*., 2019). A estimativa é de que a Dengue esteja presente em mais de 125 países (STANAWAY *et al*., 2016). A abrangência geográfica das áreas propícias ao contágio tende a se expandir devido à fenômenos globais em curso, incluindo mudanças climáticas e a urbanização (MESSINA *et al*., 2019).

3.2.2 Impactos econômicos

Não obstante às perdas de vidas humanas, a dengue acarreta em impactos econômicos. Custos de hospitalização pela doença representam pesada carga aos sistemas de saúde (LASERNA *et al*., 2018). O custo do tratamento é estimado em USD 130.00 por paciente (MONTIBELER; OLIVEIRA, 2018). Casos que exigem hospitalização representam a maior parte dos gastos diretos. Embora o custo médico para pacientes ambulatoriais seja considerado baixo, o impacto socioeconômico permanece significativo, sobretudo com perdas

de produtividade (LASERNA *et al*., 2018).

Montibeler e Oliveira (2018) pesquisaram os impactos da epidemia de dengue do ano de 2013 no Brasil. Este foi um período epidemiológico importante, quando o país registrou mais de 1,4 milhões de casos de dengue (BAVIA *et al*., 2020). A perda total estimada foi de BRL 1,023 bilhão, aproximadamente, o que representou 0,02% do produto interno bruto – PIB, na ocasião. Os mesmos autores estimaram a inoperabilidade, variável que representa o percentual de diminuição na produção por setor produtivo. O estudo considerou 68 setores e identificou que o absenteísmo da força de trabalho devido à dengue reduziu a produtividade nacional, que variou entre 0,002% e 0,027% entre os setores incluídos na pesquisa. Os setores mais afetados foram serviços domésticos, saúde pública e educação pública.

Shepard *et al*. (2016) avaliaram o impacto econômico da dengue sintomática para 141 países onde há transmissão ativa, para o ano de 2013. Foram considerados custos com assistência médica e, também, custos indiretos, associados à produtividade perdida. As perdas totais foram estimadas em US$ 8,9 bilhões, naquela ocasião. O custo estimado com a perda de produtividade representa em torno de 42% desse valor, US$ 3,77 bilhões (KOOPMANSCHAP; VAN INEVELD, 1992; HUNG *et al*., 2020).

Teich *et al*. (2017) mensuraram os custos de combate ao vetor, custos médicos diretos e custos indiretos associados à dengue clássica, dengue hemorrágica, chikungunya e infecção pelo Zika vírus no Brasil, para o ano de 2016. O custo total para aquele ano foi R$ 2,3 bilhões no Brasil, o que representa 2% do orçamento para a área da saúde.

Laserna *et al*. (2018) analisaram o impacto econômico da dengue na América Latina e no Caribe, por meio de uma revisão sistemática de literatura, que considera artigos publicados até o ano de 2016, sem definir um período inicial. Os autores concluíram que o custo econômico da dengue na América Latina ultrapassa US$ 3 bilhões anualmente e no Brasil pode chegar a US$ 1,4 bilhão, anualmente.

Oliveira *et al*. (2019) apresentaram uma revisão sistemática de literatura, para estimar o ônus gerado pela dengue. Os autores identificaram tanto pesquisas que estimam o custo direto, com tratamento médico, quanto aquelas que estimam custos indireto, com perdas produtivas. A estimativa utilizou a abordagem de paridade do poder de compra e foi

padronizada para o ano de 2015. Para 18 países considerados na pesquisa, incluindo o Brasil, a estimativa encontrada foi de US$ 3.3 bilhões em perdas no ano de 2015.

Rafikahmed *et al*. (2021) efetuaram um estudo retrospectivo de pacientes na Índia, para estudar custos médicos diretos associados à dengue. De acordo com a pesquisa, o custo médico direto total mediano por paciente foi de US$ 119,29. Despesas laboratoriais incorreram em 34,02% do custo total, consultas incorreram em 17,18% e medicamentos incorreram em 14,72%.

A tarefa de mensurar custos ligados à dengue é considerada complexa, sobretudo para os custos indiretos ligados à doença. As pesquisas mencionadas demonstram o impacto econômico causado pela doença em diferentes momentos e contextos. Os estudos apontam que este impacto está presente também no Brasil. Além do risco à vida humana, que tem valor imensurável, e dos custos atrelados ao tratamento em si, pesquisas apontam custos ligados a perda de produtividade durante surtos de dengue.

3.2.3 A dengue no Paraná

O primeiro relato de uma possível epidemia por dengue no Brasil ocorreu em 1845, no estado do Rio de Janeiro, região sudeste (FARES *et al*., 2015; BAVIA *et al*, 2020). Ao término da segunda década do século XXI, o Brasil respondeu por mais da metade dos casos de dengue registrados nas Américas (BAVIA *et al*., 2020). O sul do Brasil tem a menor incidência de casos de dengue em relação às demais regiões brasileiras, sendo de 165,2 casos por 100.000 habitantes, no ano de 2019. Todavia, têm apresentado aumento no número de notificações, considerando a segunda metade da última década (MARTIN *et al*., 2022).

No estado do Paraná, os primeiros registros de dengue se deram no ano de 1993, com epidemias constatadas a partir do ano de 1995 (BRIGAGÃO; CORRÊA, 2017). Em Vecchia *et al*. (2018), estão descritos resultados de um estudo descritivo-retrospectivo de notificações de dengue no Sul do Brasil. Os resultados demonstraram que entre os anos de 2014 e 2016, os casos autóctones de dengue no Paraná representaram 94% do total, o que indica predominância da transmissão local. Na mesma pesquisa, observou-se um crescimento significativo de notificações de dengue para o Paraná, a partir do ano de 2010, sendo o estado da região Sul do Brasil que apresentou maior quantidade de notificações, de casos confirmados e de óbitos pela doença, até 2017. No Paraná, já foram constatados casos dos

quatro sorotipos de dengue, DENV-1, DENV-2, DENV-3 e DENV-4 (BRIGAGÃO; CORRÊA, 2017; FOGAÇA; MENDONÇA, 2019).

A vacina contra a dengue, Dengvaxia, desenvolvida pela Sanofi Pasteur, foi licenciada no Brasil em 2015. O Paraná implementou um programa de vacinação, de modo pioneiro no Brasil, entre 2016 e 2018, com financiamento público estadual. A população alvo contou com 500.000 indivíduos, distribuídos nas 30 cidades mais afetadas pela doença, até então. Após este episódio, não foram efetuadas novas aquisições de Dengvaxia por parte do governo do estado, até o momento. Em Preto *et al*. (2021), é apresentado um estudo descritivo transversal que explorou a campanha de vacinação de dengue no estado.

Entre o final do mês de julho do ano de 2019 e o início do mês de julho do ano de 2020, a Secretaria de Estado da Saúde do Paraná – SESA/PR, por meio de informe técnico (PARANÁ, 2020), divulgou que o número de notificações para o período ultrapassou 360.000 casos. Isto representou recorde de casos dentro da série histórica do estado. Dos 399 municípios paranaenses, 312 tiveram ocorrência de casos autóctone, o que representa 78,2% do total. Os municípios com maior número de casos suspeitos foram Foz do Iguaçu, Londrina e Maringá. As áreas de maior incidência concentram-se na região norte, noroeste, oeste, sudoeste.

Estudos demonstram que a área geográfica onde localiza-se o estado do Paraná tende a ampliar as condições climática para a existência dos vetores *Aedes* nas próximas décadas (KRAEMER *et al*., 2019), além de aumentar o nível de adequação para a existência da doença (MESSINA *et al*., 2019). Respostas adequadas, tanto ao controle de vetores quanto ao gerenciamento dos surtos, devem ser observadas no estado, para mitigar impactos econômicos e perdas de vidas humanas causadas pela dengue.

3.3 MATERIAIS E MÉTODOS

3.3.1 Base de dados

Dados cedidos pela Secretaria de Estado da Saúde – SESA/PR, coletados por meio do SINAN, permitem elaborar uma análise exploratória para compreender melhor a epidemia de Dengue no Estado. Estes dados foram cedidos aos Departamentos de Estatística da Universidade Estadual de Maringá, os quais foram utilizados de modo multidisciplinar nesta pesquisa. O SINAN é o *software* oficial para registros de casos notificados de diversas doenças, incluindo a dengue, utilizado pelas secretarias municipais e regionais de saúde. O

SINAN Nesta pesquisa, estão sendo considerados registros compreendidos entre a 31ª semana epidemiológica de 2019, iniciada em 28 de julho daquele ano, e a 31ª semana epidemiológica de 2020, encerrada em 01 de agosto de 2020, que representou recorde de notificações na série histórica para o estado.

Os dados brutos foram tratados pelo *software* estatístico RStudio (RStudio Team, 2020). A base de dados considerada possui 366.760 observações e 51 variáveis. Estas variáveis referem-se à atributos socioeconômicos, sinais clínicos, classificação do agravo, critérios de confirmação, hospitalização e evolução clínica.

3.2.3 Estatística descritiva

A partir das informações da base de dados cedida pela SESA/PR, o perfil das notificações de dengue no Paraná para o ano em estudo foi abordada por meio de estatística descritiva. Os dados foram dispostos em tabelas, para visualizar os percentuais atribuídos para cada classe das variáveis disponibilizadas no SINAN.

3.2.2 Análise de correspondência

A análise de correspondência múltipla é um método estatístico que pode ser utilizado quando o interesse da pesquisa reside na verificação de associação entre um conjunto de dados de variáveis categóricas. Estes conjuntos de dados podem ser representados em um modelo gráfico, que facilita a interpretação do relacionamento entre as variáveis. A análise de correspondência múltipla não adota nenhum modelo teórico de distribuição de probabilidade e possui a vantagem de permitir a visualização de relações entre diversas variáveis ao mesmo tempo, revelando relacionamentos que poderiam ficar ocultos se a análise fosse feita aos pares de variáveis.

Conforme Rencher e Christensen (2012), a análise de correspondência simples é uma técnica gráfica utilizada para representar as informações em uma tabela de contingência de dupla entrada, que contém as frequências de itens de uma classificação cruzada de duas variáveis categóricas. O gráfico gerado permite analisar a interação entre as variáveis,

representadas pelas colunas, e as observações, representadas pelas linhas da tabela de contingência. A proximidade dos pontos indica associação, enquanto o distanciamento indica repulsão, tornando possível visualizar o relacionamento entre os dados.

A análise de correspondência múltipla é uma extensão da análise de correspondência simples, que permite averiguar padrões de relacionamento entre mais de duas variáveis categóricas. Cada variável é composta por vários níveis e cada um desses níveis é codificado como uma variável binária (ABDI; BÉRA, 2010). Variáveis quantitativas podem ser incluídas à análise, desde que intervalos de valores sejam configurados como variáveis nominais para representar os níveis (ABDI, H.; BÉRA, 2010). Os níveis são transformados em variáveis *dummy*, deste modo cada variável é expandida em números de vetores iguais ao das categorias inicialmente apresentadas (RENCHER; CHRISTENSEN, 2012).

Na presente pesquisa, a estatística descritiva é complementada pela análise de correspondência múltipla, a fim de visualizar associação entre as variáveis para indicar caminhos a estudos futuros aprofundados. Os resultados das análises de correspondência múltipla foram obtidos por meio do desenvolvimento de *scripts* no *software* RStudio, por meio do pacote MVar versão 2.1.8 (OSSANI; CIRILLO, 2021).

3.4 RESULTADOS E DISCUSSÕES

3.4.1 Perfil das Notificações de Dengue no Paraná

Na Tabela 1 está descrita a contagem e o percentual de notificações, conforme a classificação final do agravo e critério de confirmação. Para o período em questão, o SINAN registrou 366.760 notificações de dengue. Este número tem impacto no acionamento do atendimento ambulatorial e hospitalar pela comunidade, sendo o número de notificações o maior da série histórica até então. A maior parte dos casos notificados mostraram-se confirmados após investigação clínica. Do total, apenas 0,1% não apresentaram a informação de classificação final. Dentre os confirmados, 98,66% apresentaram dengue na forma clássica, 1,22% apresentaram dengue com sinais de alarme e 0,12% apresentaram a forma grave da doença.

A confirmação de casos de dengue pode ocorrer por exames laboratoriais específicos ou pelo critério clínico epidemiológico. Neste último caso, leva-se em conta o vínculo

epidemiológico para estabelecer a classificação final e pode ser utilizado na impossibilidade de execução de exames laboratoriais específicos ou quando os resultados destes são inconclusivos. Do total, em 65,88% dos casos foi adotado o critério clínico epidemiológico, enquanto exames laboratoriais foram responsáveis pela classificação de 27,31% dos casos. Para 6,81% das notificações, o critério de confirmação não foi informado.

Tabela 1 – Notificações de dengue no Paraná conforme classificação final e critério de confirmação para o ano epidemiológico de 2019/2020.

Variável	Contagem	Percentual
Classificação Final		
Confirmado	244.232	66,59%
Dengue clássica	240.971	98,66%
Dengue com sinais de alarme	2.978	1,22%
Dengue grave	283	0,12%
Descartado	97.488	26,58%
Inconclusivo	24.679	6,73%
Não informado	361	0,1%
Total de Notificações	366.760	
Critério de Confirmação		
Laboratorial	100.155	27,31%
Clínico epidemiológico	241.620	65,88%
Não informado	24.985	6,81%

Fonte: o autor, com dados da SESA/PR (2019/2020).

Quando é efetuada a investigação por exames laboratoriais, um ou mais exames podem ser efetuados para a mesma notificação. No SINAN, estão relacionados quatro exames específicos: Sorologia, NS1, RT_PCR e Isolamento Viral. Conforme a Tabela 2, a Sorologia foi efetuada em 22,67% dos casos e o NS1 foi efetuado em 15,85% das notificações, sendo os dois com maior predominância. O exame de Sorologia foi o único que apresentou percentual de reagentes maior do que não reagentes, sendo 59,47% de reagentes e 39,97% de não reagentes, considerando somente aqueles que fizeram o exame. O Isolamento Viral apresentou o maior percentual de inconclusivos em relação ao demais, com 11,46%.

A Tabela 3 apresenta os percentuais de notificados para sete variáveis. Com relação à variável sexo, observou-se maior percentual de notificações no sexo feminino, com 56,47%. Com relação à faixa etária, o maior percentual foi de adultos, 63,21%. Grande parte das notificações, desta forma, dizem respeito à população economicamente ativa.

Tabela 2 – Exames laboratoriais específicos realizados para notificações de dengue no Paraná no ano epidemiológico de 2019/2020.

Exame Laboratorial Específico	Contagem	Percentual
Sorologia		
Não realizado	283.612	77,33%
Realizado	83.148	22,67%
Reagente	49.449	59,47%
Não reagente	33.231	39,97%
Inconclusivo	468	0,56%
NS1		
Não realizado	308.570	84,13%
Realizado	58.070	15,87%
Reagente	22.870	39,3%
Não reagente	35.052	60,24%
Inconclusivo	268	0,46%
RT-PCR		
Não realizado	353.479	96,38%
Realizado	13.276	3,62%
Reagente	3.474	26,16%
Não reagente	9.610	72,36%
Inconclusivo	197	1,48%
Isolamento Viral		
Não realizado	365.492	99,75%
Realizado	890	0,25%
Reagente	249	27,98%
Não reagente	539	60,56%
Inconclusivo	102	11,46%

Fonte: o autor, com dados da SESA/PR (2019/2020).

A variável escolaridade apresentou o maior índice de não informados com relação à todas as outras, com 37,9%. Neste total, estão incluídos os registros inconsistentes, para os quais não foi possível definir a escolaridade a partir da informação encontrada no banco de dados em estudo. Indivíduos com ensino superior completo representaram percentual de notificados menor em relação aos indivíduos que possuem até o ensino fundamental – EF ou até o ensino médio – EM. O campo de escolaridade indica categoria “não se aplica” quando a idade é menor do que 7 anos.

Tabela 3 – Variáveis socioeconômicas, hospitalização e evolução das notificações de dengue no Paraná no ano epidemiológico de 2019/2020.

Variável	Contagem	Percentual
Sexo		
Feminino	207.117	56,47%
Masculino	159.353	43,45%
Não informados	290	0,08
Faixa etária (em anos)		
≤4	14.838	4,05%
5 a 19	71.793	19,57%
20 a 59	231.844	63,21%
≥ 60	46.388	12,65%
Não informados	1.897	0,52%
Raça/cor		
Branca	244.278	66,6%
Preta	13.971	3,81%
Amarela	3.051	0,83%
Parda	74.338	20,27%
Indígena	426	0,12%
Não informados	30.696	8,37%
Zona Residencial		
Urbana	317.662	86,61%
Rural	13.994	3,81%
Periurbana	789	0,22%
Não Informado	34.315	9,36%
Escolaridade		
1º ao 9º ano EF	81.596	22,25%
1º ao 3º ano EM	91.351	24,9%
ES completo ou incompleto	28.424	7,75%
Não se aplica	26.446	7,2%
Não informado	138.943	37,9%
Hospitalização		
Não hospitalizados	272.590	74,33%
Hospitalizados	10.978	2,99%
Não informados	39.236	22,68%
Evolução		
Cura	328.091	89,46%
Óbito por dengue	198	0,05%
Óbito por outras causas	231	0,06%
Óbito em investigação	1	<0,01%
Não informado	38.239	10,43%

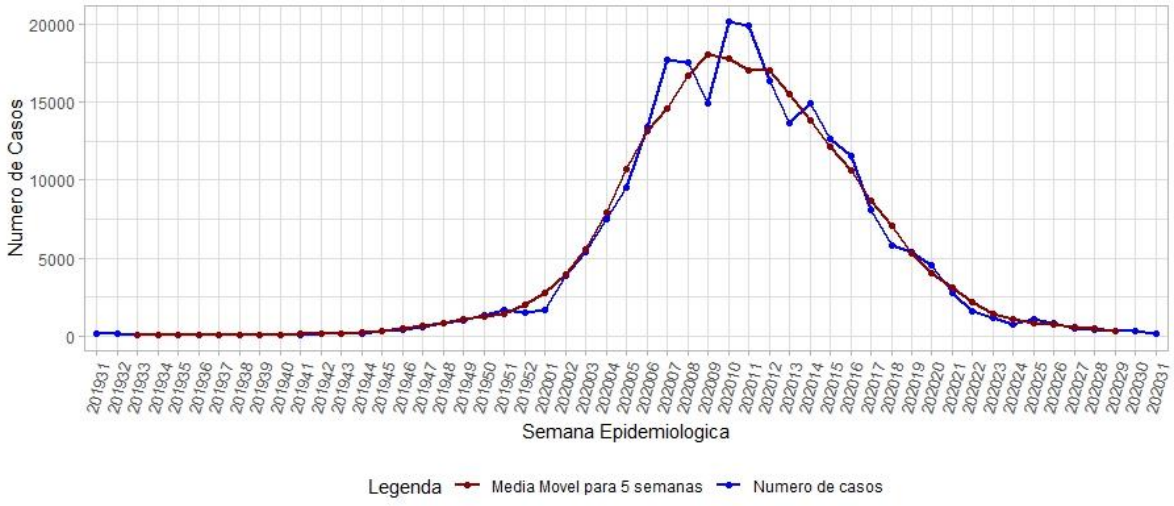
Fonte: o autor, com dados da SESA/PR (2019/2020).

O percentual de hospitalizados foi de 2,99%, enquanto 74,33% foram descritos como não hospitalizados. Esta variável, também, apresentou alto índice de não informados, chegando a 22,68%. Com relação à evolução final das notificações, 89,46% evoluíram para a

cura, 0,05% representaram óbito por dengue, 0,06% óbitos por outras causas. A evolução final não foi informada em 10,43% dos casos.

Para visualizar a evolução dos casos confirmado da doença ao longo do tempo, é apresentado o gráfico de frequência na Figura 2, onde está incluída, também, a média móvel centrada para cinco semanas. Os períodos estão designados por números de 6 dígitos, no eixo das abcissas, onde os quatro primeiros dizem respeito ao ano e os dois último à semana epidemiológica. A série temporal representada no gráfico inicia na 31ª semana do ano de 2019 e se encerra na 31ª semana de 2020. A maior parte dos casos ocorreu no período compreendido entre janeiro a maio, com pico no mês de março. É possível verificar que a média móvel cresce a partir da 46ª semana de 2019, meados do mês de novembro, atinge o pico na 9ª semana de 2020, início do mês de março, se mantém aproximadamente em estabilidade até a 12ª semana de 2020, início do mês de abril, sofrendo queda na sequência, e se estabiliza próximo ao mês de julho, após a 26ª semana.

Figura 2 – Gráfico de frequência para número de casos de dengue no Paraná – ano epidemiológico 2019/200.



Fonte: o autor, com dados da SESA/PR (2019/2020).

3.4.2 Análise de correspondência para notificações de Dengue no Paraná

A análise de correspondência múltipla tem a finalidade de complementar as análises da estatística descritiva e explorar associações entre as variáveis. As variáveis em estudos estão descritas na Tabela 4. Códigos foram criados para as variáveis, com os respectivos níveis atribuídos, a fim de reduzir as nomenclaturas exibidas nos gráficos, uma vez que no

total são 51 variáveis, que geram 137 níveis.

Tabela 4 – Variáveis consideradas para a análise de correspondência.

Tipo de Variável	Código	Variável	Níveis
Socioeconômicos	s1	Faixa etária	1 – 0 a 19 anos ; 2 – 20 a 59 anos ; 3 – 60 anos ou mais
	s2	Sexo	1 – masculino ; 2 – feminino
	s3	Raça	1 – branca ; 2 – amarela ; 3 – parda ; 4 – parda ; 5 – indígena
	s4	Escolaridade	1 – até EF ; 2 – até EM; 3 – até ES; 4 – não se aplica
	s5	Zona residencial	1 – urbana ; 2 – rural ; 3 – periurbana
Sinais clínicos de dengue clássica	c1	Febre	1 – sim ; 2 – não
	c2	Mialgia	1 – sim ; 2 – não
	c3	Cefaleia	1 – sim ; 2 – não
	c4	Exantema	1 – sim ; 2 – não
	c5	Vômito	1 – sim ; 2 – não
	c6	Náusea	1 – sim ; 2 – não
	c7	Dor nas costas	1 – sim ; 2 – não
	c8	Conjuntivite	1 – sim ; 2 – não
	c9	Artrite	1 – sim ; 2 – não
	c10	Artralgia	1 – sim ; 2 – não
	c11	Petequias	1 – sim ; 2 – não
	c12	Leucopenia	1 – sim ; 2 – não
	c13	Dor retroorbital	1 – sim ; 2 – não
Doenças pré-existentes	p1	Diabetes	1 – sim ; 2 – não
	p2	Doenças Hematológicas	1 – sim ; 2 – não
	p3	Doenças Hepatológicas	1 – sim ; 2 – não
	p4	Doença renal crônica	1 – sim ; 2 – não
	p5	Hipertensão arterial	1 – sim ; 2 – não
	p6	Doença ácido-péptica	1 – sim ; 2 – não
	p7	Doenças autoimunes	1 – sim ; 2 – não
Sinais clínicos de dengue com sinais de alarme	a1	Hipotensão	1 – sim ; 2 – não ; 3 – não se aplica
	a2	Queda abrupta de plaquetas	1 – sim ; 2 – não ; 3 – não se aplica
	a3	Vômitos persistentes	1 – sim ; 2 – não ; 3 – não se aplica
	a4	Sangramento de mucosas/ outras hemorragias	1 – sim ; 2 – não ; 3 – não se aplica
	a5	Aumento do hematócrito	1 – sim ; 2 – não ; 3 – não se aplica
	a6	Dor abdominal	1 – sim ; 2 – não ; 3 – não se aplica
	a7	Letargia ou irritabilidade	1 – sim ; 2 – não ; 3 – não se aplica
	a8	Hepatomegalia	1 – sim ; 2 – não ; 3 – não se aplica
	a9	Acúmulo de líquidos	1 – sim ; 2 – não ; 3 – não se aplica
Sinais clínicos de dengue grave	g1	Pulso débil ou indetectável	1 – sim ; 2 – não ; 3 – não se aplica
	g2	Pressão arterial convergente	1 – sim ; 2 – não ; 3 – não se aplica
	g3	Tempo de enchimento capilar	1 – sim ; 2 – não ; 3 – não se aplica
	g4	Acúmulo de líquidos com insuficiência respiratória	1 – sim ; 2 – não ; 3 – não se aplica
	g5	Taquicardia	1 – sim ; 2 – não ; 3 – não se aplica
	g6	Extremidades frias	1 – sim ; 2 – não ; 3 – não se aplica
	g7	Hipotensão arterial em fase tardia	1 – sim ; 2 – não ; 3 – não se aplica
	g8	Hematêmese	1 – sim ; 2 – não ; 3 – não se aplica
	g9	Melena	1 – sim ; 2 – não ; 3 – não se aplica
	g10	Metrorragia volumosa	1 – sim ; 2 – não ; 3 – não se aplica
	g11	Sangramento do sistema nervoso central	1 – sim ; 2 – não ; 3 – não se aplica
	g12	Aspartato aminotransferase–AST/alanina aminotransferase – ALT > 1.000	1 – sim ; 2 – não ; 3 – não se aplica
	g13	Miocardite	1 – sim ; 2 – não ; 3 – não se aplica
	g14	Alteração da consciência	1 – sim ; 2 – não ; 3 – não se aplica
Classificação final	CF	Classificação Final	descartado; dengue clássica – dengue; dengue com sinais de alarme – DSA; dengue grave - DG
Hospitalização	H	Hospitalização	1 – sim ; 2 – não ; 3 – não informado
Evolução Clínica	E	Evolução Clínica	1 – cura ; 2 – óbito dengue ; 3 – óbito outras causas ; 4 – não informados

Fonte: o autor, com dados da SESA/PR (2019/2020).

Para a análise de correspondência múltipla, foram retiradas as amostras que não trouxeram a informação de: classificação final, critério de confirmação, idade, sexo, escolaridade, raça e zona residencial. Os sinais clínicos de dengue clássica mostraram-se

consistentes, não havendo campos não preenchidos. O mesmo foi constatado para o grupo de doenças pré-existentes. Para os sinais clínicos de DSA e DG, campos em branco foram preenchidos com código de não aplicabilidade, haja vista que o preenchimento destes campos está ligado à própria caracterização do agravo, não sendo preenchido na maioria dos casos. Foram mantidos os casos não informados de evolução clínica e hospitalização, considerando que o percentual elevado identificados.

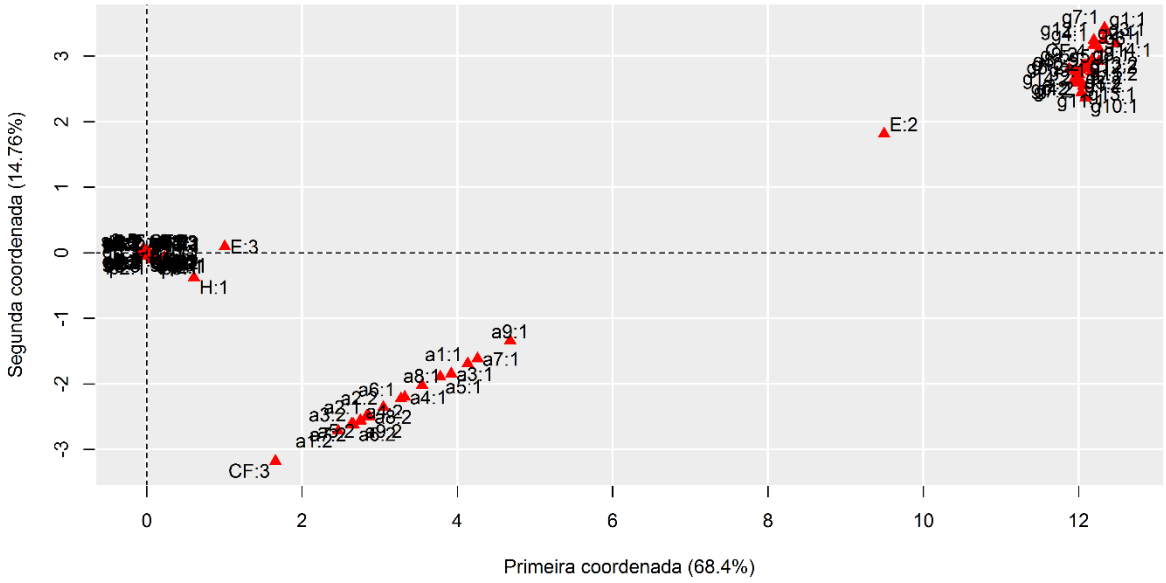
Os dados foram filtrados conforme a variável de critério de confirmação. Foram mantidas somente as amostras com critério de confirmação laboratorial, independente do exame laboratorial adotado. Após todas as considerações, restaram 55.071 amostras para esta análise. Não é objeto deste estudo associações entre os exames laboratoriais específicos, ficando como sugestão para outras pesquisas estas investigações.

A Figura 3.a apresenta o gráfico da análise de correspondência para 51 variáveis de dengue. Ao utilizar o teste Qui-quadrado para verificar a dependência entre as notificações e as variáveis, com 18.225 graus de liberdade e valor-p: 1,00 E^{-9}, a nível de significância de 5%, verifica-se existência de dependência entre linhas e colunas. As análises podem ser explicadas em um espaço bidimensional, uma vez que a proporção da variação explicada nos dois primeiros componentes é de 83,16% da variação amostral.

Dois grupos de variáveis se distanciaram do aglomerado ao centro e as respectivas variáveis ficaram mais próximos entre si. Estes grupos dizem respeito às características específicas das classificações finais de dengue com sinais de alarme - DSA (DF:3) (Figura 3.b e dengue grave – DG (DF:4) (Figura 3.c), e dizem respeito aos sinais clínicos específicos para estas ocorrências. Isso, de fato, é esperado, pois estes sinais clínicos caracterizam estes agravos, diferenciando-os de casos de dengue clássica ou descartados. O gráfico permite identificar que, para DSA, alguns sinais clínicos estão mais distantes do agravo do que outros e esse distanciamento ocorre na direção dos perfis de DG, sendo que os três que mais se distanciaram foram acúmulo de líquidos, letargia e irritabilidade, hipotensão. (a9:1, a7:1 e a1:1).

Com relação aos sinais clínicos em torno de DG, percebe-se que estão mais difusos ao redor desta variável em relação aos perfis DSA. Os sinais clínicos de acúmulo de líquidos com insuficiência respiratória, extremidades frias e ausência de miocardite (g4:1, g6:1 e g13:2), podem ser citados com maior proximidade à DG em relação aos demais. A variável de óbitos por dengue (E:2) está mais próxima dos perfis de DG, o que também é esperado.

Figura 3 – Gráfico da análise de correspondência múltipla para as 51 variáveis estudadas, convertidas em 137 variáveis *dummy*.



a) gráfico para todas as variáveis.

b) inferior esquerdo ampliado (dengue com sinais de alarme)

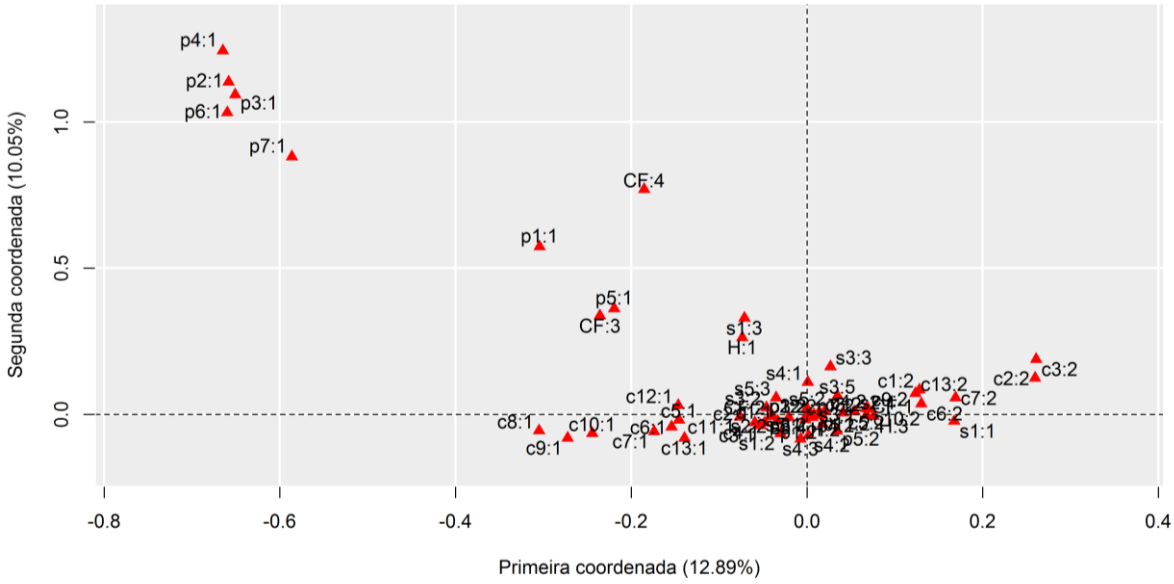
c) superior direito ampliado (dengue grave)

Fonte: o autor, com dados da SESA/PR (2019/2020).

Para dar continuidade à análise, foram excluídos os sintomas específicos de DSA e DG, além da variável de evolução, a fim de visualizar a associação entre as demais variáveis. Nesta etapa, o total de variáveis analisadas é 27. O gráfico da análise de correspondência está presente na Figura 4. O teste Qui-quadrado, com 3.844 graus de liberdade e valor-p: 1,00 E^{-9}, a nível de significância de 5%, verificou que há dependência entre as variáveis e as notificações. A proporção da variação explicada nos dois primeiros componentes é de

22,94%.

Figura 4 – Gráfico da análise de correspondência para os grupos de variáveis socioeconômicas, sinais clínicos de dengue clássica, doenças pré-existentes, classificação final e hospitalização.



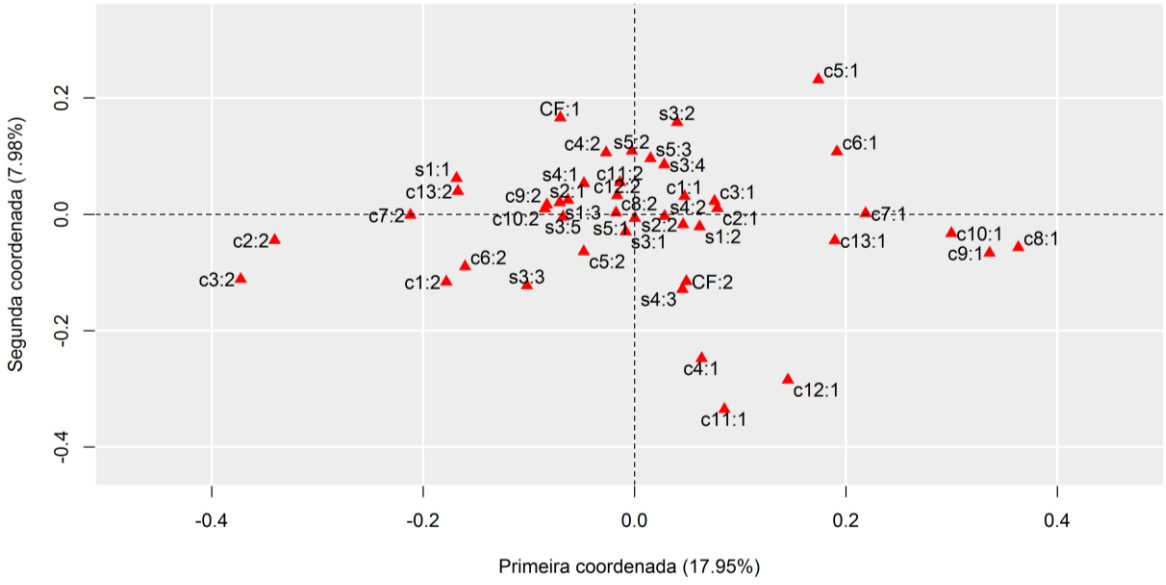
Fonte: o autor, com dados da SESA/PR (2019/2020).

Algumas análises possíveis referem-se às doenças pré-existentes. Verifica-se que doença renal crônica, doenças hematológicas, doença ácido-péptica, doença hepatológicas e doenças autoimunes (p4:1, p2:1, p6:1, p3:1, p7:1) ficaram próximas entre si, indicando associação entre estas variáveis. A variável diabetes (p1:1) ficou próxima de perfis de DSA e DG (CF:3, CF:4), enquanto hipertensão (p5:1) ficou fortemente próxima de DSA (CF:4). Ainda, é possível observar que casos hospitalizados (H:1) ficaram mais próximo da faixa etária de 60 anos ou mais (s1:3)

Na sequência, foram excluídas as variáveis de doenças pré-existentes e hospitalização, restando 19 variáveis. Além disso, foram excluídos os casos de DSA e DG. O número de amostras em análise foi de 54.261. Com esta análise, pretende-se visualizar a associação entre os sinais clínicos de dengue clássica às variáveis socioeconômicas e às classificações finais das notificações para dengue clássica ou descartados. O teste Qui-quadrado, com 1.849 graus de liberdade e valor-p: 1,00 E^{-9}, a nível de significância de 5%, verificou a existência da dependência entre as variáveis e as notificações. A proporção da variação explicada nos dois

primeiros componentes é de 25,93%, conforme conta no gráfico da Figura 5.

Figura 5 – Gráfico da análise de correspondência para variáveis socioeconômicas, sinais clínicos de dengue clássica e classificação final – critério laboratorial.



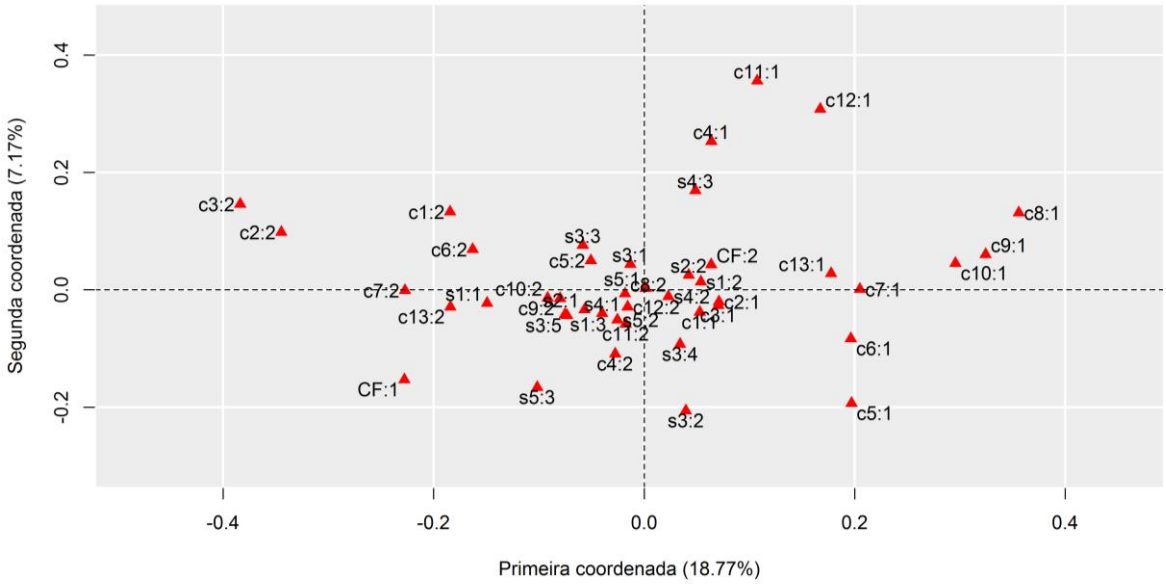
Fonte: o autor, com dados da SESA/PR (2019/2020).

Neste caso, observa-se que as variáveis apresentam perfil difuso, o que limita a observação entre associação de variáveis com maior ou menor proximidade. Ainda assim, algumas associações se destacam. A variável de escolaridade superior (s4:3) está mais próximo dos casos confirmados de dengue clássica (CF:2). Os sinais clínicos exantema, petéquias e leucopenia (c4:1, c11:1, c12:1) ficaram mais próximos entre si, indicando associação entre estes, e ficaram mais próximos da variável dengue clássica (CF:2) do que descartados (CF:1). Ainda com relação à sinais clínicos, ausência de exantema (c4:2) ficou mais próximo de casos descartados. As variáveis artralgia, artrite e conjuntivite (c10:1, c9:1, c8:1) estão mais próximas entre si, indicando associação entre elas, porém distanciadas das variáveis de confirmação e descarte. Também, verifica-se que ausência de mialgia e ausência de cefaleia (c2:2, c3:2) possuem proximidade entre si. Para ausência de dor retro orbital e ausência de dor nas costas (c13:2, c7:2), houve proximidade com indivíduos da faixa etária jovem (s1:1).

A próxima análise também busca visualizar a associação entre os sinais clínicos de dengue clássica às variáveis socioeconômicas e as classificações finais das notificações para dengue clássica ou descartados, todavia, agora foram selecionadas amostras onde o critério

clínico epidemiológico foi utilizado. Neste caso, as mesmas 19 variáveis foram consideradas em relação à Figura 5, todavia, para as notificações as quais foi atribuído o critério de confirmação clínico epidemiológico, totalizando 122.418 amostras. Esta tarefa tem como objetivo comparar os dois conjuntos de dados. Os resultados estão presentes no gráfico da Figura 6. O teste Qui-quadrado, com 1.849 graus de liberdade e valor-p: 1,00 E^{-9}, a nível de significância de 5%, verificou a existência da dependência entre as variáveis e as notificações. A proporção da variação explicada nos dois primeiros componentes é de 25,94%,

Figura 6 – Gráfico da análise de correspondência para variáveis socioeconômicas, sinais clínicos de dengue clássica e classificação final – critério clínico epidemiológico.



Fonte: os autores, com dados da SESA/PR (2021).

As variáveis artralgia, artrite e conjuntivite (c10:1, c9:1, c8:1) estão mais próximas entre si, tal qual ocorreu no perfil laboratorial, do mesmo modo que houve proximidade entre ausência de mialgia e ausência de cefaleia (c2:2, c3:2). O mesmo verificou-se para os sinais clínicos de exantema, petéquias e leucopenia (c4:1, c11:1, c12:1), que se mantiveram próximos entre si. Não houve variáveis com forte proximidade aos casos descartados para dengue (CF:1), sendo que as variáveis mais próximas deste perfil foram zona residencial periurbana e ausência de dor retro orbital (s5:3, c13:2). Casos confirmados dengue clássica (CF:2) ficaram mais próximos de sexo feminino e indivíduos em idade adulta (s2:2, s1:2).

3.5 CONCLUSÕES

A dengue tem atingido boa parte da população vivendo em áreas tropicais do mundo. No Estado do Paraná, o ano epidemiológico de 2019/2020 apresentou recorde histórico de notificações. As informações armazenadas no Sistema de Agravos e Notificações – SINAN, da Secretaria de Estado da Saúde – SESA/PR, permitem explorar o perfil destas notificações, para compreender o impacto na comunidade e nos sistemas de saúde. Esta análise é possível a partir da estatística descritiva.

Para a saúde pública, este trabalho contribui em três perspectivas. Primeiro, fornecendo um perfil das notificações de dengue no estado do Paraná, durante a epidemia do ano 2019/2020, o que oferece um panorama do ponto de vista socioeconômico de impacto da doença. Segundo, apontando lacunas de preenchimento do SINAN, onde ações podem ser adotadas para aprimorar o preenchimento e melhorar a qualidade da informação disponibilizada para o monitoramento da doença. E terceiro, apontando associação entre variáveis, por meio da análise de correspondência, que pode inspirar estudos futuros para profissionais de saúde.

Do total de notificações, 66,59% mostraram-se confirmadas para dengue. Embora a maioria dos casos faça referência à dengue clássica, 1,22% dos confirmados apresentaram dengue com sinais de alarme e 0,12% dengue grave. Óbitos por dengue representaram 0,05% dos notificados. Exames laboratoriais foram realizados somente em 27,31% das notificações. Em 2,99% das notificações, houve necessidade de hospitalização. Outra característica marcante é que 63,21% das notificações dizem respeito à população adulta, onde reside a força de trabalho do Estado. Com relação à zona residencial, 86,61% das notificações foram apontadas como de pessoas vivendo nas áreas urbanas. As principais oportunidades de melhoria de preenchimento dizem respeito aos campos escolaridade, hospitalização e evolução clínica.

Por meio da análise de correspondência, foi possível observar que perfis de dengue com sinais de alarme e dengue grave possuem características próprias, que as diferenciam de casos clássicos. Com relação às doenças pré-existente, observou-se associações que podem direcionar estudos futuros para explorar de modo mais específicos a correspondência entre essas enfermidades e quadros agravados de dengue, ou ainda o relacionamento entre as doenças entre si. Outra possibilidade é a investigação da associação entre faixa etária e

hospitalização.

Com relação à classificação final entre descartados ou confirmados para dengue clássica, considerando somente sinais clínicos e variáveis socioeconômicas, observa-se maior dispersão entre as variáveis. Uma vez que nem sempre as possibilidades de exames laboratoriais estão presentes, pesquisas futuras podem aprofundar a busca por padrões a partir destes dados para apoiar a classificação de novas notificações.

Outras pesquisas podem contemplar diferentes abordagens, a partir dos mesmos dados, para produzir conhecimento sobre impactos da dengue no Paraná. Não obstante à perda de vidas humanas, mensurar os custos diretos, com o tratamento de saúde, e os custos indiretos, ligados à perda de produtividade devido ao acometimento por dengue para o estado, é tarefa que desperta interesse. Registros do SINAN contribuem para essa abordagem de análise, ficando como sugestão para outros artigos.

As dificuldades encontradas nesta pesquisa estão associadas, principalmente, ao preenchimento incompleto da base de dados, o que exigiu considerável redução das amostras para estudo durante a preparação dos dados para a análise de correspondência. É também importante mencionar que esta pesquisa foi realizada com dados que levam em conta um período epidêmico de dengue, sendo que a análise a partir de dados coletados em momentos não epidêmicos podem gerar resultados diferentes. Importante mencionar que os dados não foram coletados *in loco*. Neste sentido, é assumida a premissa do correto preenchimento do SINAN pelos profissionais envolvidos.

REFERÊNCIAS

ABDI, H.; BÉRA, M. *Correspondence Analysis. In* ***Encyclopedia of Research Design***. *Thousand Oaks*. 2010.

AHMED, Shahid *et al*.: *Dengue fever outbreak: a clinical management experience.* ***J Coll Physicians Surg Pak***, v. 18, n. 1, p. 8-12, 2008.

BAVIA, L. *et al*.: *Epidemiological study on dengue in southern Brazil under the perspective of climate and poverty.* ***Scientific Reports***, v. 10, n. 1, p. 1-16, 2020.

BHATT, S. *et al*. :*The global distribution and burden of dengue.* ***Nature***, v. 496, n. 7446, p. 504-507, 2013.

BRIGAGÃO, G.; CORRÊA, N. A. B.. Levantamento epidemiológico da dengue no estado do Paraná Brasil nos anos de 2011 a 2015. **Arquivos de Ciências da Saúde da UNIPAR**, v. 21, n. 1, 2017.

FARES, R. CG *et al. Epidemiological scenario of dengue in Brazil.* ***BioMed research international****,* v. 2015, 2015.

FLIPSE, J.; SMIT, J. M.: *The complexity of a dengue vaccine: a review of the human antibody response.* ***PLoS neglected tropical diseases****,* v. 9, n. 6, p. e0003749, 2015.

FOGAÇA, T. K.; MENDONÇA, F.: Distribuição espacial dos sorotipos de dengue e fluxos intermunicipais no Paraná. **Raega-O Espaço Geográfico em Análise**, v. 46, n. 2, p. 101-115, 2019.

FRITZELL, C. et al.: *Knowledge, attitude and practices of vector-borne disease prevention during the emergence of a new arbovirus: implications for the control of Chikungunya virus in French Guiana.* ***PLoS neglected tropical diseases****,* v. 10, n. 11, p. e0005081, 2016.

HIGUERA-MENDIETA, D. R. *et al*.: KAP *surveys and dengue control in Colombia: disentangling the effect of sociodemographic factors using multiple correspondence analysis.* ***PLoS neglected tropical diseases****,* v. 10, n. 9, p. e0005016, 2016.

HUNG, T. M. *et al*.: *Productivity costs from a dengue episode in Asia: a systematic literature review.* ***BMC infectious diseases****,* v. 20, n. 1, p. 1-18, 2020.

KOOPMANSCHAP, M. A.; VAN INEVELD, B. M.: *Towards a new approach for estimating indirect costs of disease.* ***Social science & medicine****,* v. 34, n. 9, p. 1005-1010, 1992.

KRAEMER, Moritz UG *et al*. *The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus.* ***elife****,* v. 4, p. e08347, 2015.

LASERNA, A. *et al. Economic impact of dengue fever in Latin America and the Caribbean: a systematic review*. ***Revista Panamericana de Salud Pública****,* v. 42, p. e111, 2018.

MARTIN, B. M. et al. *Clinical outcomes of dengue virus infection in pregnant and non-pregnant women of reproductive age: a retrospective cohort study from 2016 to 2019 in Paraná, Brazil.* ***BMC infectious diseases****,* v. 22, n. 1, p. 1-11, 2022.

MESSINA, Jane P. *et al.: The current and future global distribution and population at risk of dengue.* ***Nature microbiology****,* v. 4, n. 9, p. 1508-1515, 2019.

MONTIBELER, E.; OLIVEIRA, D.: *Dengue endemic and its impact on the gross national product of Brazilian's economy.* ***Acta tropica****,* v. 178, p. 318-326, 2018.

NAVA-DOCTOR, J. E.; SANDOVAL-RUIZ, C.A.; FERNÁNDEZ-CRISPÍN, A.: *Knowledge, attitudes, and practices regarding vector-borne diseases in central Mexico.* ***Journal of ethnobiology and ethnomedicine****,* v. 17, n. 1, p. 1-14, 2021.

OLIVEIRA, L. N. S.; ITRIA, A.; LIMA, E. C.: *Cost of illness and program of dengue: A systematic review.* ***PloS one****,* v. 14, n. 2, p. e0211401, 2019.

OSSANI P. C.; CIRILLO M. A.: **MVar: *Multivariate analysis*.** URL <https://cran.r-project.org/web/packages/MVar/>. R *package version* 2.1.8, 2021.
PRETO, C. *et al. Vaccination coverage and adherence to a dengue vaccination program in the state of Parana, Brazil*. ***Vaccine***, v. 39, n. 4, p. 711-719, 2021.

RAFIKAHMED, S. R. et al. *Assessment of direct medical cost using cost of illness analysis in patients with dengue fever-Retrospective study*. ***Clinical Epidemiology and Global Health***, v. 12, p. 100842, 2021.

RENCHER, A. C.; CHRISTENSEN W. F.: *Methods of multivariate analysis*. 3rd. ed. New Jersey: John Wiley & Sons, 2012. 781 p.

RSTUDIO TEAM. RStudio: Integrated Delevopment for R. RStudio, Inc., Boston, 2020. Disponível em <http://www.rstudio.com/>, acesso em 01 de março de 2022.

SHEPARD, D. S. *et al*. *The global economic burden of dengue: a systematic analysis.* ***The Lancet infectious diseases***, v. 16, n. 8, p. 935-941, 2016.

SISWANTINING, T. *et al*.: *Predicting the risk of hospitalization to six diagnoses with highest costs based on outpatient claims*. In: **AIP *Conference Proceedings***. AIP Publishing LLC, 2018. p. 020067.

SOUZA-NETO, J. A.; POWELL, J. R.; BONIZZONI, M.: *Aedes aegypti vector competence studies: A review*. ***Infection, Genetics and Evolution***, v. 67, p. 191-209, 2019.

STANAWAY, J. D. *et al*.: *The global burden of dengue: an analysis from the Global Burden of Disease Study* 2013. ***The Lancet infectious diseases***, v. 16, n. 6, p. 712-723, 2016.

TANNER, L. *et al*.: *Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness*. **PLoS Negl Trop Dis**, v. 2, n. 3, p. e196, 2008.

TEICH, V.; ARINELLI, R.; FAHHAM, L.: *Aedes aegypti e sociedade: o impacto econômico das arboviroses no Brasil*. ***JBES: Brazilian Journal of Health Economics/Jornal Brasileiro de Economia da Saúde***, v. 9, n. 3, 2017.

THOMAS, S. J.; YOON, I.: *A review of Dengvaxia®: Development to deployment.* ***Human vaccines & immunotherapeutics***, v. 15, n. 10, p. 2295-2314, 2019.

WICHMANN, Ole *et al*.: *Severe dengue virus infection in travelers: risk factors and laboratory indicators.* ***The Journal of infectious diseases***, v. 195, n. 8, p. 1089-1096, 2007.

WILDER-SMITH, A.; OOI, E.-E.; HORSTICK, O.; WILLS, B.: Dengue. **The Lancet**, v. 393, n. 10169, p. 350-363, 2019.

WORLD HEALTH ORGANIZATION et al. ***Handbook for clinical management of dengue***. 2012.

WU, C. *et al*.: *Behaviors related to mosquito-borne diseases among different ethnic minority groups along the China-laos border areas.* ***International journal of environmental research and public health***, v. 14, n. 10, p. 1227, 2017.

YASEEN, H.M. *et al. Identification of initial severity determinants to predict arthritis after chikungunya infection in a cohort of French gendarmes.* ***BMC musculoskeletal disorders***, v. 15, n. 1, p. 1-8, 2014.

4

ARTIGO 2

Modelos de *machine learning* para classificação de notificações de dengue no Paraná.

RESUMO

No estado do Paraná, com frequência, o critério clínico-epidemiológico é utilizado para classificação de notificações de dengue, por parte de profissionais de saúde. Neste artigo, são propostos modelos de classificação supervisionada de *machine learning* para auxiliar na automatização destas tarefas, a partir de 48 variáveis, constituídas de informações clínicas e socioeconômicas, presentes no SINAN. Os modelos foram construídos utilizando o classificador probabilístico *Naïve Bayes*. Três modelos estão descritos na pesquisa. O primeiro modelo tem como objetivo a classificação de notificações em "descartados" ou "confirmados" para dengue, independente da gravidade. O segundo modelo tem o mesmo objetivo, porém leva em conta a gravidade, para os casos confirmados. O terceiro modelo é proposto para classificação de casos de "dengue clássica", DSA ou DG, e leva em conta somente casos confirmados. Uma técnica *projection pursuit* com o índice LDA, que permitem visualizar a distribuição espacial das amostras entre as classes de classificação, está incluída na análise. A avaliação dos modelos leva em conta, além da acurácia, o índice de concordância Kappa. Os resultados demonstram que há complexidade para classificação de

notificações nos cenários considerados para os modelos 1 e 2, observada pela acurácia mediana e índice Kappa regular, obtidos em ambos os casos, além da sobreposição dos dados, observada pela visualização gráfica. O modelo 3 obteve resultados satisfatórios, uma vez que foi também observada melhor separação espacial dos dados. A partir das variáveis disponíveis no SINAN, os modelos de ML mostraram-se melhores para classificar a gravidade de dengue em casos confirmados, em relação à classificação de casos clássicos e descartados de notificados.

Palavras-chave: dengue, notificação de doença, sistemas de informação em saúde, aprendizado de máquina supervisionado.

4.1 INTRODUÇÃO

O aprendizado de máquina, ou *machine learning* (ML), apresenta-se como uma subárea da Inteligência Artificial – IA, capaz de oferecer técnicas para a construção de modelos computacionais que podem auxiliar de modo significativo aos profissionais da área de saúde em estabelecer prognósticos (OBERMEYER, EMANUEL, 2016). Estas técnicas podem descobrir e identificar padrões e os relacionamentos entre eles, a partir de conjuntos de dados complexos, ao mesmo tempo em que são capazes de prever diagnósticos futuros (KOUROU *et al*., 2015). São utilizadas em diagnósticos por imagem, diagnósticos genéticos, exames laboratoriais, triagem em massa e outros (JIANG *et al*., 2017).

O presente artigo tem por objetivo responder à seguinte pergunta: **de que modo técnicas de ML podem contribuir com os casos de notificação de dengue no Paraná, para apoiar a tomada de decisão na área de saúde**?

Como parte do ML, os métodos supervisionados de classificação utilizam conjuntos rotulado de dados de entrada para estimar a classificação de observações futuras (KOUROU *et al*., 2015). Uma vez que os dados contidos no SINAN possuem estas características, neste estudo, a classificação supervisionada é aplicada aos casos de notificações de dengue. O objetivo é a construção e avaliação de modelos para classificação de casos notificados da doença, que possam contribuir com profissionais da saúde para a classificação de notificações.

Pesquisas que divulgaram modelos de classificação de dengue utilizando ML mostram-se recentes na literatura, onde diferentes atributos preditivos são considerados para

classificação de variáveis alvo, conforme o objetivo do modelo proposto. A pesquisa de Dakappa *et al*. (2017) buscou classificar casos de febre indiferenciada entre quatro enfermidades distintas, incluindo a dengue. Para gerar o modelo, foram consideradas aferições de temperaturas timpânicas em pacientes hospitalares. Gambhir *et al*. (2017) e Gambhir *et al*. (2018) utilizaram registros hospitalares, em adição às variáveis de idade, sexo, sinais clínicos e resultados de exames laboratoriais, presentes em registros hospitalares, para propor modelos de classificação capazes de confirmar ou descartar casos notificados de dengue.

Em Davi *et al*. (2019), o modelo proposto buscou determinar prognósticos de dengue grave em pacientes infectados por DENV. As variáveis preditivas foram baseadas no genótipo de indivíduos, determinados pela coleta e análise de exames laboratoriais. Iqbal e Islam (2019) buscaram classificar notificações de dengue, entre as classes de descarte ou confirmação, por meio de um modelo que considera sinais clínicos e exames laboratoriais como atributos de predição, a partir de informações coletadas de relatórios de pacientes. Em Mello-Román *et al*. (2019), o modelo buscou classificar casos notificados de dengue, com relação à confirmação ou não, utilizando dados de um sistema de registros de saúde pública. As variáveis preditivas contaram com variáveis sociodemográficas, socioeconômicas e sinais clínicos.

O modelo proposto em Alias Balamurugan *et al*. (2020) teve por objetivo classificar a ocorrência de dengue, ou a não ocorrência, por meio de variáveis de exames laboratoriais. No estudo de Han *et al*. (2021), um sistema de medição, apoiado por radar médico, coletou sinais vitais de indivíduos infectados por dengue e indivíduos saudáveis, em um ambiente hospitalar. O método não é especificamente destinado à classificação para infecção por dengue, mas sim para classificar a presença de infecção que provoque alterações nos sinais vitais preditivos. Em todo caso, o grupo infectado, do estudo conduzido na pesquisa, foi restrito somente à pacientes confirmados com dengue. A pesquisa de Ozer *et al*. (2021) buscou a construção de modelos de classificação para necessidade de hospitalização em casos suspeitos de arboviroses em geral, mas não necessariamente para dengue. Chattopadhyay e Chattopadhyay (2021) geraram um modelo para classificar a gravidade da dengue, a partir de sinais clínicos. Os dados foram importados de registros hospitalares.

Na presente pesquisa, 48 variáveis preditivas e uma variável alvo, presentes no SINAN, foram consideradas para a construção de modelos de classificação. Estas variáveis dizem respeito às informações socioeconômicas, doenças pré-existentes, sinais clínicos de dengue clássica, sinais clínicos de dengue com sinais de alarme e sinais clínicos de dengue

grave. A variável de classificação guarda o resultado clínico atribuído a cada notificação, podendo descartar ou confirmar a notificação, ou ainda apresentar o grau de gravidade da doença. Todas as variáveis são do tipo categóricas.

A pesquisa contou com a construção de três modelos distintos, com diferentes objetivos:

- **Modelo 1**: determinar a confirmação ou descarte de notificações de dengue, independente da gravidade;
- **Modelo 2**: determinar a confirmação ou descarte notificações de dengue, incluindo a gravidade para os casos confirmados;
- **Modelo 3**: determinar a gravidade de dengue em casos confirmados da doença.

O algoritmo de ML na construção de todos os modelos foi o *Naïve Bayes* (NB) (JOHN; LANGLEY, 1995), que pertence à família dos classificadores probabilísticos. A validação cruzada foi utilizada como método de divisão da base para treinamento e teste. Como parâmetros de avaliação do modelo, foram utilizados o F-*Score* e ROC Area, além do índice de concordância Kappa. Gráficos gerados a partir da técnica de redução de dimensionalidade chamada *Projection Pursuit*, com o índice LDA, permitem visualizar a distribuição espacial das amostras em torno das classes de classificação.

A construção dos modelos de ML deste trabalho ficou restrita aos dados categóricos das variáveis estruturadas presentes no SINAN, aos quais foram aplicados os procedimentos para classificação supervisionada e avaliação dos modelos construídos. A tarefa de descartar ou confirmar notificações mostrou-se complexa, sendo que os modelos 1 e 2 apresentaram concordância estatística regular e acurácia menor do que 70%. Foi observada melhor separação dos dados para o modelo que buscou classificar o nível de gravidade em confirmados, sendo que a acurácia neste caso foi superior à 95%.

4.2 MATERIAIS E MÉTODOS

4.2.1 Base de dados

A base de dados foi exportada do Sistema de Informação de Agravo de Notificação (SINAN), da Secretária de Estado da Saúde do Paraná (SESA/PR). Os registros referem-se a notificações de dengue no estado do Paraná, do período compreendido entre o mês de agosto do ano de 2019 e o mês de julho do ano de 2020. Foram registradas 366.760 notificações de

dengue no Estado do Paraná nesta ocasião, o que representou número recorde para o estado. Os dados foram filtrados conforme uma variável presente na base de dados, denominada "critério de confirmação". Foram mantidas somente as amostras com critério de confirmação laboratorial, independente do exame laboratorial adotado. Aqui, destaca-se que o diagnóstico de dengue, em grande parte das vezes, foi efetuado pelo critério clínico epidemiológico. Esses resultados não foram considerados na construção do modelo, pois julgou-se que amostras apoiadas por exames laboratoriais teriam confiabilidade adicional para a construção dos modelos

Após execução de tarefas de seleção, pré-processamento e transformação dos dados, foi gerada a base para classificação com 49 variáveis categóricas, conforme Tabela 5, sendo 48 variáveis preditivas e 1 variável alvo. Com relação ao número de amostras, foram mantidas 54.820 instâncias, por estarem com preenchimento adequado à tarefa de classificação. Amostras com preenchimento incompleto, ou para os quais o código de preenchimento não permitiu a diferenciação do nível de classificação, foram removidas. Para os sinais clínicos de dengue com sinais de alarme - DSA e dengue grave - DG, campos em branco foram preenchidos com código de não aplicabilidade, uma vez que o preenchimento destes campos está ligado à própria caracterização do agravo, não sendo preenchido na maioria dos casos. A variável alvo é denominada "classificação final – CF". O pré-processamento da base de dados originou uma base viável para a classificação, com quantidade significativa de amostras.

Tabela 5 – Variáveis consideradas para construção de modelos de classificação para notificações de dengue.

Tipo de Variável	Código	Variável	Níveis
Socioeconômicos	s1	Faixa etária	1 – 0 a 19 anos ; 2 – 20 a 59 anos ; 3 – 60 anos ou mais
	s2	Sexo	1 – masculino ; 2 – feminino
	s3	Raça	1 – branca ; 2 – amarela ; 3 – parda ; 4 – parda ; 5 – indígena
	s4	Escolaridade	1 – até EF ; 2 – até EM; 3 – até ES; 4 – não se aplica
	s5	Zona residencial	1 – urbana ; 2 – rural ; 3 – periurbana
Sinais clínicos de dengue clássica	c1	Febre	1 – sim ; 2 – não
	c2	Mialgia	1 – sim ; 2 – não
	c3	Cefaleia	1 – sim ; 2 – não
	c4	Exantema	1 – sim ; 2 – não
	c5	Vomito	1 – sim ; 2 – não
	c6	Náusea	1 – sim ; 2 – não
	c7	Dor nas costas	1 – sim ; 2 – não
	c8	Conjuntivite	1 – sim ; 2 – não
	c9	Artrite	1 – sim ; 2 – não
	c10	Artralgia	1 – sim ; 2 – não
	c11	Petequias	1 – sim ; 2 – não
	c12	Leucopenia	1 – sim ; 2 – não
	c13	Dor retroorbital	1 – sim ; 2 – não
Doenças pré-existentes	p1	Diabetes	1 – sim ; 2 – não
	p2	Doenças Hematológicas	1 – sim ; 2 – não
	p3	Doenças Hepatológicas	1 – sim ; 2 – não
	p4	Doença renal crônica	1 – sim ; 2 – não
	p5	Hipertensão arterial	1 – sim ; 2 – não
	p6	Doença ácido-péptica	1 – sim ; 2 – não
	p7	Doenças autoimunes	1 – sim ; 2 – não
Sinais clínicos de	a1	Hipotensão	1 – sim ; 2 – não ; 3 – não se aplica

dengue com sinais de alarme	a2	Queda abrupta de plaquetas	1 – sim ; 2 – não ; 3 – não se aplica
	a3	Vômitos persistentes	1 – sim ; 2 – não ; 3 – não se aplica
	a4	Sangramento de mucosas/ outras hemorragias	1 – sim ; 2 – não ; 3 – não se aplica
	a5	Aumento do hematócrito	1 – sim ; 2 – não ; 3 – não se aplica
	a6	Dor abdominal	1 – sim ; 2 – não ; 3 – não se aplica
	a7	Letargia ou irritabilidade	1 – sim ; 2 – não ; 3 – não se aplica
	a8	Hepatomegalia	1 – sim ; 2 – não ; 3 – não se aplica
	a9	Acúmulo de líquidos	1 – sim ; 2 – não ; 3 – não se aplica
Sinais clínicos de dengue grave	g1	Pulso débil ou indetectável	1 – sim ; 2 – não ; 3 – não se aplica
	g2	Pressão arterial convergente	1 – sim ; 2 – não ; 3 – não se aplica
	g3	Tempo de enchimento capilar	1 – sim ; 2 – não ; 3 – não se aplica
	g4	Acúmulo de líquidos com insuficiência respiratória	1 – sim ; 2 – não ; 3 – não se aplica
	g5	Taquicardia	1 – sim ; 2 – não ; 3 – não se aplica
	g6	Extremidades frias	1 – sim ; 2 – não ; 3 – não se aplica
	g7	Hipotensão arterial em fase tardia	1 – sim ; 2 – não ; 3 – não se aplica
	g8	Hematêmese	1 – sim ; 2 – não ; 3 – não se aplica
	g9	Melena	1 – sim ; 2 – não ; 3 – não se aplica
	g10	Metrorragia volumosa	1 – sim ; 2 – não ; 3 – não se aplica
	g11	Sangramento do sistema nervoso central	1 – sim ; 2 – não ; 3 – não se aplica
	g12	Aspartato aminotransferase–AST/alanina aminotransferase – ALT > 1.000	1 – sim ; 2 – não ; 3 – não se aplica
	g13	Miocardite	1 – sim ; 2 – não ; 3 – não se aplica
	g14	Alteração da consciência	1 – sim ; 2 – não ; 3 – não se aplica
Classificação final	CF	Classificação Final	descartado; dengue clássica – dengue; dengue com sinais de alarme – DSA; dengue grave - DG

Fonte: o autor, com informações da SESA/PR (2022).

4.2.2 Classificador *Naïve Bayes (*NB)

O classificador NB é uma das técnicas de classificação mais populares em ML, utilizada na tarefa de classificação supervisionada. É baseado no teorema de *Bayes* e assume como premissa a independência entre os atributos. NB tem sido utilizado na construção de modelos de classificação para a área de saúde, não só para a classificação de dengue (GAMBHIR *et al*. 2018; IQBAL; ISLAM 2019), como também em outros contextos (SRINIVASAN *et al*. 2015, ZHENG *et al*. 2017, UDDIN *et al*. 2019, AMIN *et al*. 2019). Detalhes da técnica são apresentadas em John e Langley (1995).

Para exemplificar, seja C, o conjunto de sinais clínicos apresentado por um indivíduo, que podem sugerir a presença de uma infecção D. Pelo teorema de Bayes, a probabilidade $P(D \vee C)$ de que o indivíduo, apresentando esse grupo de sinais clínicos, possua a infecção, pode ser estimada conforme a Equação (1).

$$P(D|C) = \frac{P(C|D)P(D)}{P(C)} \quad (1)$$

$P(C \vee D)$ é a probabilidade de que o indivíduo tenha o mesmo grupo de sinais clínicos, se possuir a infecção, $P(D)$ é a probabilidade de ocorrência da doença e $P(C)$ é a probabilidade da ocorrência do grupo de sinais clínicos. Um conjunto C é constituído por $\{C_1, C_2, \dots, C_i\}$ sinais clínicos. A Equação (1) pode, então, ser reescrita como o produto das

probabilidades marginais, conforme a Equação (2).

$$P(D \vee C_1, C_2, \dots, C_i) = P(C) \prod_i P(C_i \vee D) \quad (2)$$

Por outro lado, D pode assumir valores $\{D_1, D_2, \dots, D_n\}$, para representar diferentes infecções, ou diferentes níveis de gravidade de uma infecção, ou, ainda, a presença ou ausência de uma infecção. D representa a classe de classificação. O classificador NB busca indicar a classe de maior probabilidade a que pertença uma nova observação, chamada C_{map}, *maximum a posteriori*, indicado pela Equação (3).

$$C_{map} = argmax_{D_n \in D} P(D_n|C) \quad (3)$$

Em síntese, o aprendizado supervisionado do classificador NB busca determinar a probabilidade *a priori* da classe de classificação e as probabilidades condicionais que relacionam a classe de classificação aos atributos. A partir destas informações, o classificador indica a qual classe de classificação uma nova observação tem a maior probabilidade de pertencer.

4.2.3 Parâmetros de avaliação do modelo

Além do percentual de acertos, indicado pela acurácia (A), outros parâmetros são descritos na literatura para avaliação de modelos preditivos de ML. Para descrever os parâmetros utilizados nesta pesquisa, inicialmente, é preciso definir os valores de verdadeiro positivo (VP), verdadeiro negativo (VN), falso negativo (FN) e falso positivo (FP), descritos na Tabela 6 (HAN *et al*., 2021), que representa a matriz de confusão.

Tabela 6 – Matriz de confusão.

Resultado Real	Previsão do modelo	
	Positivo	Negativo
Positivo	VP	FN
Negativo	FP	VN

Fonte: adaptado de Han *et al*. (2021).

Além da precisão (P) e da sensibilidade (S, serão considerados os valores $F\text{-}score$, que combina e equilibra os dois primeiros. Conforme Han *et al*. (2021), esta é uma métrica adequada para avaliar modelos preditivos de triagem clínica, uma vez que efetuar uma previsão errada de um paciente infectado é mais perigoso do que prever errado um paciente

não infectado. Para avaliar a confiabilidade do modelo, será utilizado o índice Kappa (K)(LANDI; KOCH, 1977), que expressa a concordância estatística, sendo um dos índices mais utilizados e disseminados na literatura em estudos de validação e reprodutibilidade experimentos. O cálculo em termos da concordância observada (P_o e da concordância esperada (P_e)(NARANJO *et al*., 1981). As métricas, com as respectivas formulações, estão descritas no Quadro 2.

Quadro 2 – Métricas e formulações.

Métrica	**Formulação**
Acurácia	$A = \frac{TP + TN}{TP + FP + TN + FN}$ (4)
Precisão	$P = \frac{VP}{VP + FP}$ (5)
Sensibilidade	$S = \frac{VP}{VP + FN}$ (6)
F-*Score*	$F\text{-}score = \frac{2PR}{P + R}$ (7)
Índice Kappa - K	$K = \frac{P_o - P_e}{1 - P_e}$ (8) Onde: $P_O = \frac{VP + VN}{VP + FN + FP + VN}$ (9) e: $P_e = \frac{\{(VP + FN)(VP + FP)\} + \{(FP + VN)(FN + VN)\}}{(VP + FN + FP + VN)^2}$ (11)

Fonte: Adaptado de Goutte e Gaussier (2005) e Naranjo *et al*. (1981).

O valor K pode variar entre -1, que representa discordância total, e $+1$, ou concordância perfeita (NARANJO *et al*., 1981). Um modelo robusto deve apresentar concordância, no mínimo, ≥ 0. O nível de concordância pode ser atribuído conforme a Tabela 7.

Tabela 7 – Índice de concordância Kappa

Kappa	Concordância
0,00 – 0,20	Leve
0,21 – 0,40	Regular
0,41 – 0,60	Moderada
0,61 – 0,80	Forte
0,81 – 1,00	Quase Perfeita

Fonte: o autor, adaptado de Landi e Koch (1977).

Outro parâmetro utilizado é a área sob a curva ROC – ROC Area (HANLEY; MCNEIL, 1982). Este parâmetro leva em conta tanto a sensibilidade S , quanto a especificidade E, dada pela equação (10).

$$E = \frac{VN}{VN + FP} \tag{10}$$

A área sob a curva ROC é empregada como medida de desempenho do modelo. Um modelo ideal terá ROC Area igual a 1. Um modelo com ROC Area igual a 0,5 não permite a distinção de uma observação entre as classes.

4.2.4 Validação cruzada – *cross-validation*

Cross-validation (KOHAVI *et al*., 1995), ou validação cruzada, é um procedimento estatístico para teste e treinamento da base de dados. O conjunto inicial é dividido em subconjuntos, também chamadas de *folds*. Com frequência são adotadas 10 partições, mas este número pode ser maior. Enquanto uma partição é utilizada para treinamento, as outras são guardadas para teste. O procedimento se repete para as 10 partições, de modo circular. O conjunto de treinamento não se repete. Todas as amostras fazem parte do conjunto de teste uma vez. Por fim, é exposto a acurácia média das observações.

4.2.5 *Projection Pursuit* e índice LDA

Projection Pursuit trata-se de uma técnica de análise multivariada exploratória, que busca projeções lineares de baixa dimensão em dados de alta dimensão. Estas projeções são encontradas por meio da otimização de uma função objetivo, denominada índice de projeção (FRIEDMAN; TUKEY, 1974). O índice de projeção adotado nesta pesquisa foi o LDA (LEE *et al*., 2005, ESPEZUA *et al*., 2015;), para buscar a formação de agrupamentos no espaço de alta dimensão.

O índice LDA é baseado na análise discriminante linear, *linear discriminat analysis* – LDA, que é um método estatístico para classificação e reconhecimento de padrões. LDA é uma generalização do discriminante linear de Fisher, popularizada pela utilização em classificação de imagens (BELHUMEUR *et al*., 1997). O objetivo é encontrar uma combinação linear que separe os dados entre classes de classificação.

Neste trabalho, a técnica *projection pursuit* foi usada como forma exploratória dos dados que estão espaço n-dimensional, com índice de projeção LDA, a fim de averiguar visualmente a separação entre classes de classificação de notificações de dengue.

4.3 CONSTRUÇÃO DOS MODELOS

Este trabalho propõe a construção de três modelos para classificação de dengue, com diferentes objetivos: modelo 1, para determinar a confirmação ou descarte de notificações de dengue, independente da gravidade; modelo 2, para determinar a confirmação ou descarte notificações de dengue, incluindo a gravidade para os casos confirmados; e modelo 3, para determinar a gravidade de dengue em casos confirmados da doença.

O pré-processamento dos dados foi realizado utilizando o *software* RStudio (RSTUDIO TEAM, 2020). Os modelos foram construídos no *software* Weka (FRANK; WITTEN, 2016). Embora os algoritmos *Sequential Minimal Optimization* – SMO, AdaBoost, J48 e *Random Forrest* tenham sido testados, *Naïve Bayes* – NB obteve melhores resultados para todos os modelos e, por isso, foi utilizado para a construção dos modelos. A base de teste e treinamento foi dividida considerando um método de validação cruzada, para evitar sobreajuste, com 10 partições. Os resultados estão descritos na sequência.

4.4 RESULTADOS E DISCUSSÕES

4.4.1 Modelo 1

O objetivo deste modelo é a confirmação ou descarte para dengue em casos suspeitos, independente da gravidade. Foram consideradas 48 variáveis preditivas e a variável "classificação final", ou CF, como classificador. Os casos agravados e casos de dengue clássica foram agrupados em um único grupo, denominado "dengue". A base de dados possui 54.820 observações, sendo 32.656 de confirmados para dengue clássica, o que representa 59,57% dos dados, e 22.164 notificações com classificação final denominada "descartado" para dengue clássica, quando o agravo não se confirmou, o que representa 40,43% dos dados. A Tabela 8 apresenta a matriz de confusão para o modelo 1. A Tabela 9 apresenta a acurácia

do modelo 1, com o índice Kappa. A Tabela 10 apresenta o detalhamento do modelo 1 por classe de classificação.

A acurácia do modelo 1 foi de 63,38%. O $F\text{-}Score$ foi melhor para classificar Dengue em relação à classificação de Descartados, o que indica que o modelo é melhor em classificar dengue corretamente em relação à classificação correta de descartados. O índice Kappa igual a 0,2616, o que demonstra concordância regular do modelo. Em ambos os modelos, a ROC Area ficou abaixo de 0,7.

Tabela 8 – Matriz de confusão para o modelo 1.

	Dengue	Descartados
Dengue	20.579	12.077
Descartados	7.998	14.166

Fonte: o autor (2022).

Tabela 9 – Acurácia do modelo 1 e Índice Kappa

Instâncias classificadas corretamente	63,38%
Instâncias classificadas incorretamente	36,62%
Índice Kappa	0,2616

Fonte: o autor (2022).

Tabela 10 – Detalhamento do modelo 1, por classe.

	Sensibilidade	Precisão	F-*Score*	ROC Area
Dengue	0,630	0,720	0,672	0,684
Descartado	0,639	0,540	0,585	0,684

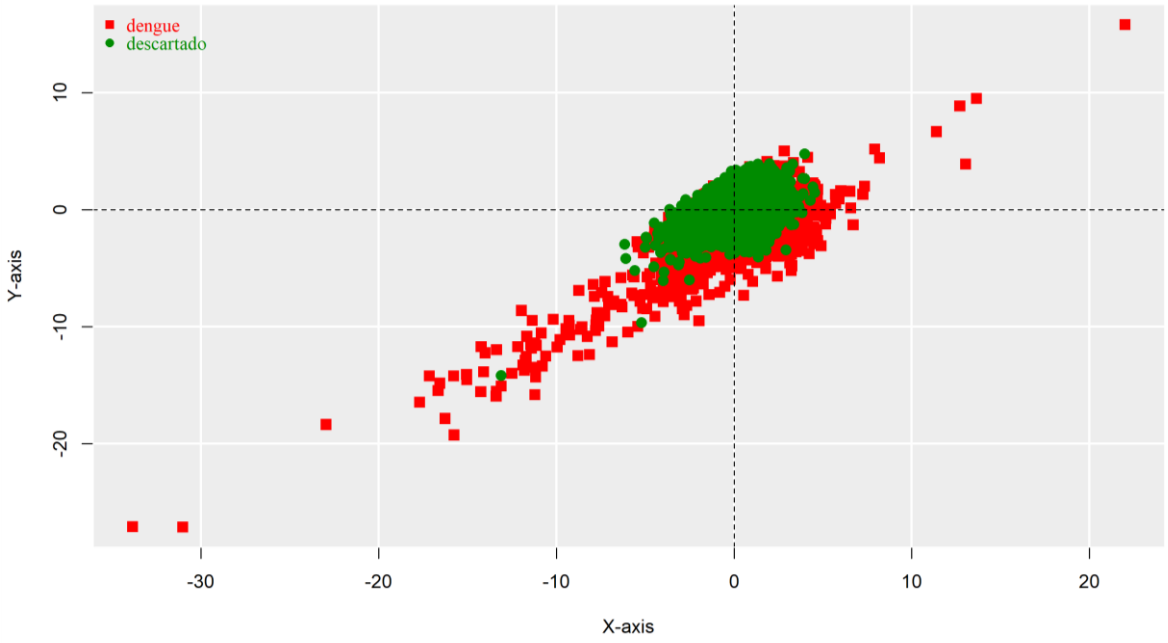
Fonte: o autor (2022).

Os resultados demonstraram que é possível gerar um modelo de classificação para notificações de dengue a partir das variáveis categóricas abordadas, todavia, com concordância estatística regular e acurácia menor do que 70%. Além disso, é perceptível o elevado valor de FN para dengue, no modelo 1. Neste sentido, percebe-se a complexidade do problema em construir um classificador para dengue a partir dos atributos utilizados. Seria vantajoso a utilização de um modelo de classificação a partir dos atributos descritos no Quadro 2, principalmente para os casos suspeitos de dengue clássica, sem agravamento, uma vez que a observação de sinais clínicos de dengue clássica depende da avaliação ambulatorial, e não laboratorial, as quais nem sempre estão disponíveis. Para o ano epidemiológico de 2019/2020, no Paraná, em torno de um terço das notificações tiveram como critério de confirmação exames laboratoriais. Os demais, foram designados com critério clínico epidemiológico.

Modelos de classificação para dengue que utilizaram outros tipos de variáveis

preditivas relataram melhor acurácia nos modelos gerados. Gambhir *et al*. (2017) incluíram testes IgM, IgG e NS1, além de contagem de plaquetas; Gambhir *et al*. (2018) incluíram testes IgM, IgG e NS1 como variáveis preditivas; Iqbal e Islam (2019) incluíram variáveis de hemograma; Mello-Roman *et al*. (2019) adicionaram informações sobre região de residência, viagem, acampamento e nível social. Estudos futuros podem averiguar a construção e modelo com a adição de novas variáveis. O modelo gerado aqui ficou restrito aos dados estruturados presentes no SINAN.

Figura 7 – Gráfico com resultados das projeções das quatro classes do modelo 1, utilizando a *projection pursuit* com o índice LDA.



Fonte: o autor (2022).

A fim de averiguar visualmente a separação entre classes de classificação de notificações de dengue, foi gerado o gráfico da Figura 7, com a técnica *projection pursuit* com o índice LDA. A proximidade entre as notificações pertencentes à diferentes classes de classificação demonstra a complexidade para a separação dos dados.

4.4.2 Modelo 2

O objetivo deste modelo é a confirmação ou descarte para dengue em casos suspeitos, porém levando em conta a gravidade. Foram mantidas as 48 classes preditivas e a classe CF como classificador. A base de dados possui 54.820 observações, sendo 22.164 notificações descartadas, o que representa 40,43% dos dados; 31.850 confirmados com dengue clássica, o que representa 58,1% dos dados; 706 casos de DSA, o que representa 1,29% dos dados; e 100

casos de DG, o que representa 0,18% dos dados.

A Tabela 11 apresenta a matriz de confusão para o modelo 2. A Tabela 12 apresenta a acurácia do modelo 2, com o índice Kappa. A Tabela 13 apresenta o detalhamento do modelo 2 por classe de classificação. É possível observar que houve leve alteração positiva na acurácia, igual a 64,55%, e no índice Kappa, de 0,2792, em relação ao modelo 1. Todavia, o modelo 2 ainda apresenta concordância estatística regular. O F-*Score* para Dengue ficou acima de 0,7, enquanto o mesmo parâmetro para casos descartados foi igual a 0,538, o que indica que o modelo é melhor para classificar corretamente casos de dengue em relação à classificação correta de descartados.

Tabela 11 – matriz de confusão para o modelo 2.

	Descartado	Dengue	DSA	DG
Descartado	11.290	10.850	21	3
Dengue	8.526	23.295	22	7
DSA	0	0	702	4
DG	0	0	0	100

Fonte: o autor (2022).

Tabela 12 – Acurácia do modelo 1 e Índice Kappa.

Instâncias classificadas corretamente	64,55%
Instâncias classificadas incorretamente	35,45%
Índice Kappa	0,2792

Fonte: o autor (2022).

Tabela 13 – detalhamento do modelo 1, por classe.

	Sensibilidade	Precisão	F-*Score*	ROC Area
descartado	0,509	0,570	0,538	0,684
dengue	0,731	0,682	0,706	0,688
DAS	0,994	0,942	0,968	0,994
DG	1,000	0,877	0,935	1,000

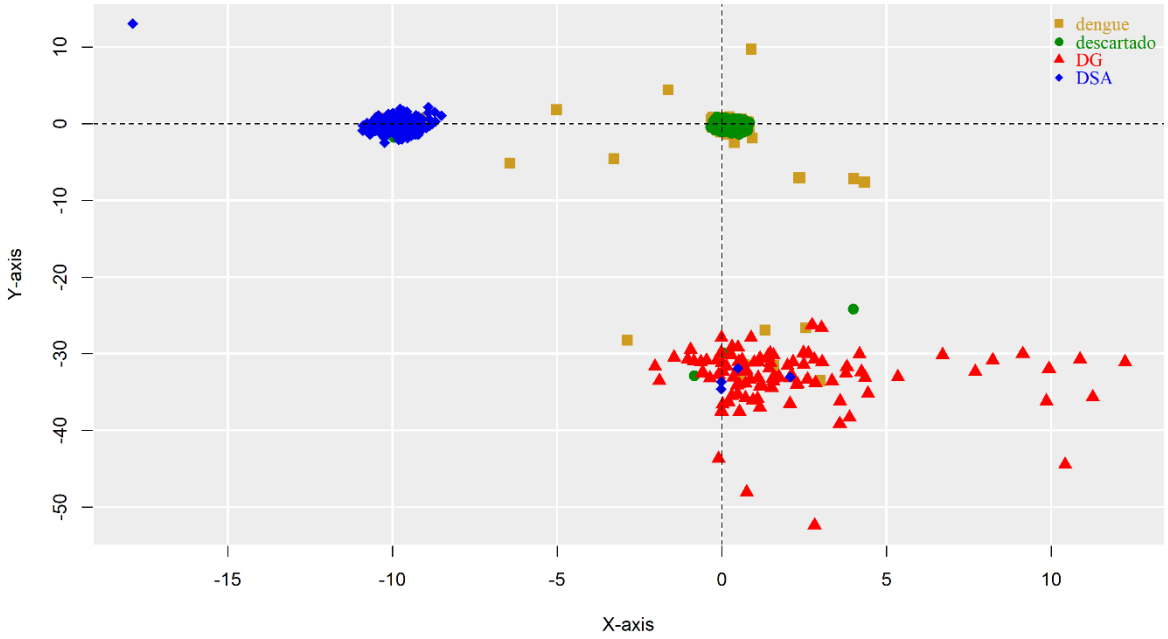
Fonte: o autor (2022).

O modelo 2 foi eficiente para classificação de DSA e DG, com valores ROC Area igual a 0,994 e 1, respectivamente, e F-*Score* de 0,968 e 0,935, respectivamente. Neste sentido, observa-se melhor separação dos dados entre as classes DSA e DG em relação aos casos clássicos de dengue e descartados.

A fim de averiguar visualmente a separação entre classes de classificação de classificação, foi gerado o gráfico da Figura 8, com a técnica *projection pursuit* com o índice LDA. De fato, há melhor separação espacial das amostras de DG e DSA. Isso se deve ao fato de que os sintomas em casos agravados se mostram mais específicos e são preenchidos no SINAN nas hipóteses de agravamento. Desta forma, houve maior eficiência do modelo em

classificar casos agravados.

Figura 8 – Gráfico com resultados das projeções das quatro classes do modelo 2, utilizando a *projection pursuit* com o índice LDA.



Fonte: o autor (2022).

Uma vez que se observou melhor separação dos dados para os casos de DSA e DG, optou-se por construir um terceiro modelo, considerando somente os casos confirmados de dengue. Assim, tem-se o objetivo de construir um modelo capaz de classificar a gravidade da dengue, em notificações confirmadas. Neste sentido, foi gerado o modelo 3.

4.4.3 Modelo 3

O modelo 3 tem por objetivo classificar o nível de gravidade em casos confirmados de dengue. Foram consideradas 32.656 amostras, das quais 31.850 são confirmadas com dengue clássica, o que representa 97,53% dos dados; 706 casos de DSA, o que representa 2,16% dos dados; e 100 casos de DG, o que representa 0,31% dos dados. A Tabela 13 apresenta a matriz de confusão para o modelo 3. A Tabela 14 apresenta a acurácia do modelo 2, com o índice Kappa. A Tabela 15 apresenta o detalhamento do modelo 2 por classe de classificação.

Tabela 13 – Matriz de confusão para o modelo 3.

	dengue	DSA	DG
dengue	31.821	22	7
DSA	0	702	4
DG	0	0	100

Fonte: o autor (2022).

Tabela 14 – Acurácia do modelo 3 e Índice Kappa.

Instâncias classificadas corretamente	99,89%
Instâncias classificadas incorretamente	0,101%
Índice Kappa	0,9794

Fonte: o autor (2022).

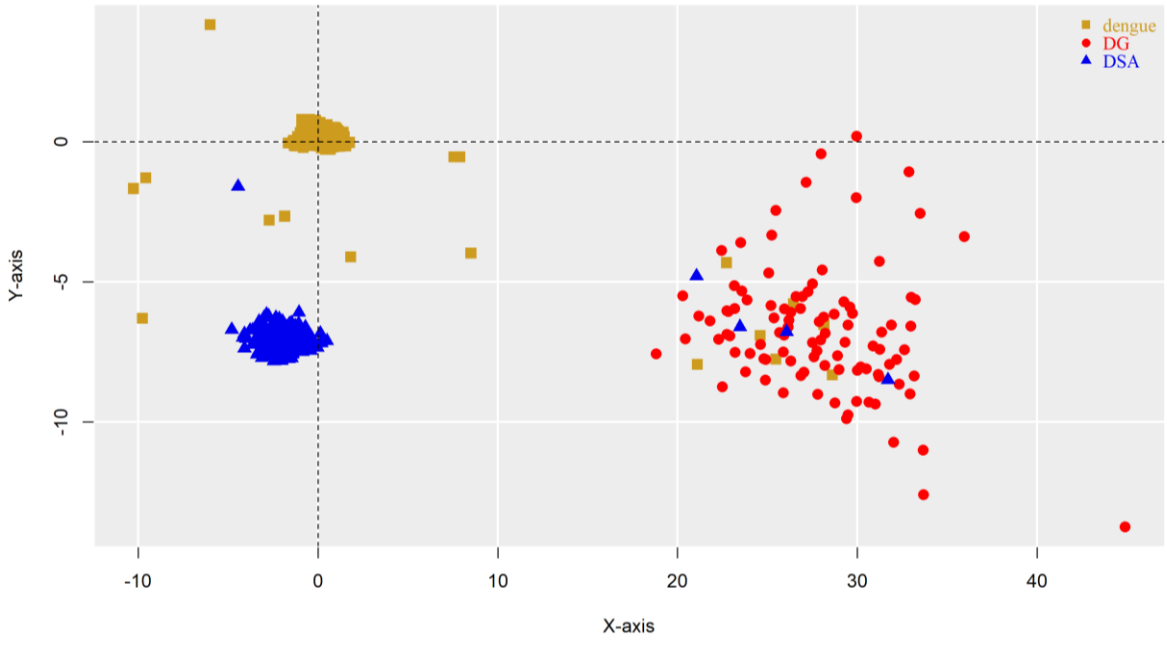
Tabela 15 – Detalhamento do modelo 3, por classe.

	Sensibilidade	Precisão	F-*Score*	ROC Area
dengue	0,999	1,000	1,000	0,999
DSA	0,994	0,970	0,982	0,994
DG	1,000	0,901	0,984	1,000

Fonte: o autor (2022).

A acurácia, neste caso, ficou em 99,89% e o índice Kappa foi de 0,9794, com concordância quase perfeita. Neste sentido, os atributos apresentados no SINAN permitiram a construção de um modelo preditivo com concordância estatística relevante para classificação de gravidade em casos confirmados de dengue.

Figura 9 – Gráfico com resultados das projeções das três classes do modelo 3 utilizando a *projection pursuite* com o índice LDA



Fonte: o autor (2022).

Ocorre melhor separação dos dados, uma vez que o preenchimento das variáveis de gravidade ocorre nas hipóteses de presença do agravo. Para analisar a separação entre classes de classificação de gravidade de dengue em casos confirmados, foi gerado o gráfico da Figura 8, com a técnica *projection pursuit* com o índice LDA.

Ainda que o preenchimento das variáveis de gravidade esteja especificamente destinados aos casos agravados de dengue, estes resultados contribuem para demonstrar que a construção de modelos de classificação de ML para classificação de gravidade de dengue contam com variáveis que contribuem melhor para a separação espacial dos dados. Portanto, os modelos gerados para classificação de gravidade obtiveram melhores resultados em relação à classificação de confirmação e descarte.

4.5 CONCLUSÕES

Este trabalho teve por objetivo a construção e avaliação de modelos de classificação em ML para notificações de dengue. Modelos como os descritos podem ser utilizados por profissionais de saúde na busca por automatizar tarefas de classificação de notificações de doenças. Esta atividade mostra-se útil sobretudo em cenários epidêmicos onde as possiblidades de exames laboratoriais específicos não estejam presentes ou sejam limitadas. O classificador NB foi utilizado para a construção de três modelos, para três objetivos distintos de classificação.

O primeiro modelo, construído para classificação de notificações de dengue, independente da gravidade, teve acurácia menor do que 70% e índice de concordância Kappa regular. Por meio da *projection pursuit* com o índice LDA, observou-se proximidade entre as classes de classificação para “descartados” e “dengue clássica”, o que demonstra a complexidade em executar a tarefa de classificação a partir de atributos abordados. A adoção do Modelo 1 acarreta em assumir um modelo com sensibilidade de 0,630, o que leva à quantidade significativa de casos positivos classificados incorretamente. Por outro lado, essa constatação abre caminho para outras discussões.

Em primeiro lugar, que o processo de construção de um modelo para classificação de dengue a partir das variáveis abordadas mostrou-se complexo. Neste sentido, novas pesquisas, incluindo diferentes variáveis, podem ser consideradas em busca de modelos com melhor acurácia, tais como: informações de viagem, histórico vacinal, casos de dengue na família ou vizinhos, residência em áreas afetadas e outras. Uma questão seria demonstrar como estas variáveis poderiam ser captadas no atendimento clínico, uma vez que não há

atributos no SINAN para este preenchimento. Além disso, outros algoritmos podem ser testados em busca de melhores acurácias.

Outro ponto é que modelos mais robustos podem estar atrelados à necessidade de execução de exames laboratoriais e que isso acarreta em custos. Todavia, quais são os custos de classificações incorretas? Uma vez que a dengue clássica se assemelha à diversas outras enfermidades, a necessidade do diagnóstico laboratorial parece mais plausível. No caso do estado do Paraná, para o ano de 2019/2020, na maior parte das vezes, estas possibilidades não foram adotadas.

O segundo modelo, construído para classificação de notificações, levando em conta também a gravidade, obteve resultados não expressivamente melhores. Os dados com melhor separação dizem respeito aos níveis de gravidade de DSA e DG. O terceiro modelo considerou apenas notificações confirmadas de dengue, para a tarefa de classificação da gravidade. Neste caso, obteve-se índices de concordância e acurácia altos. Os resultados indicam melhor separação dos dados para perfis agravados de dengue, o que sugere que os atributos categóricos estruturados do SINAN são significativos para a construção de ferramentas de ML para classificação de gravidade, em notificações confirmadas.

Uma vez que a construção dos modelos ficou restrita aos dados categóricos estruturados da base de dados do SINAN, pesquisas futuras podem incluir a adição de novos atributos, a fim de investigar possibilidades de ampliar a separação entre as classes, principalmente para os objetivos propostos nos dois primeiros modelos. Para novos conjuntos, novas ferramentas de ML podem ser comparadas, além do algoritmo NB, em relação aos parâmetros avaliativos. Modelos com maior acurácia têm sido construídos a partir da inclusão de variáveis laboratoriais e testes específicos.

Cumpre salientar que os dados foram exportados do sistema e, portanto, não foram coletados em loco. Logo, foi assumido como premissa o correto preenchimento da base de dados por parte dos operadores do sistema nas unidades de saúde. Por fim, destaca-se que os modelos gerados dizem respeitos à dados coletados em um período onde houve epidemia de dengue, o que tem impacto direto na quantidade de notificações observadas e, possivelmente, no percentual de descartados e confirmados, dentre os notificados. Também, que as variáveis classificadas previamente para a construção do modelo levaram em conta o critério de confirmação laboratorial. Entretanto, estes exames possuem sensibilidade e especificidades próprias, o que pode impactar na separação das classes de classificação.

REFERÊNCIAS

ALIAS BALAMURUGAN, S. A.; MALLICK, MS M.; CHINTHANA, G. *Improved prediction of dengue outbreak using combinatorial feature selector and classifier based on entropy weighted score based optimal ranking.* ***Informatics in Medicine Unlocked***, v. 20, p. 100400, 2020.

AMIN, Mohammad Shafenoor; CHIAM, Yin Kia; VARATHAN, *Kasturi Dewi. Identification of significant features and data mining techniques in predicting heart disease.* ***Telematics and Informatics***, v. 36, p. 82-93, 2019.

BELHUMEUR, P. N.. ; HESPANHA, J. P.; KRIEGMAN, D. J.. . *Eigenfaces vs. fisherfaces: Recognition using class specific linear projection*. **IEEE Transactions on pattern analysis and machine intelligence**, v. 19, n. 7, p. 711-720, 1997.

CHATTOPADHYAY, Amit K.; CHATTOPADHYAY, Subhagata. *VIRDOCD: A VIRtual DOCtor to predict dengue fatality.* ***Expert Systems***, p. e12796, 2021.

DAKAPPA, P. H. *et al. A predictive model to classify undifferentiated fever cases based on twenty-four-hour continuous tympanic temperature recording.* ***Journal of healthcare engineering***, v. 2017, 2017.

DAVI, C. *et al. Severe dengue prognosis using human genome data and machine learning.* ***IEEE Transactions on Biomedical Engineering***, v. 66, n. 10, p. 2861-2868, 2019.

ESPEZUA, S., VILLANUEVA, E., MACIEL, C.D., CARVALHO, A.: *A projection pursuit framework for supervised dimension reduction of high dimensional small sample datasets.* ***Neurocomputing*** 149, 767–776, 2015.

FRANK, E.; HALL, M.A.; WITTEN, I.: *The WEKA Workbench.* ***Online Appendix for "Data Mining: Pratical Machine Learning Tools and Techniques"***, Morgan Kaufmann, Fourth Edition, 2016.

FRIEDMAN, J. H.; TUKEY, J. W. *A projection pursuit algorithm for exploratory data analysis.* ***IEEE Transaction on Computers***, 23(9):881-890, 1974.

GAMBHIR, S.; MALIK, S. K.; KUMAR, Y.: *PSO-ANN based diagnostic model for the early detection of dengue disease.* ***New Horizons in Translational Medicine***, v. 4, n. 1-4, p. 1-8, 2017.

GAMBHIR, Shalini; MALIK, Sanjay Kumar; KUMAR, Yugal. *The diagnosis of dengue disease: An evaluation of three machine learning approaches.* ***International Journal of Healthcare Information Systems and Informatics*** **(IJHISI)**, v. 13, n. 3, p. 1-19, 2018.

GOUTTE, C.; GAUSSIER, E.: *A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In:* ***European conference on information retrieval***. *Springer, Berlin, Heidelberg,* 2005. p. 345-359.

HAN, T. T. *et al. Machine learning based classification model for screening of infected patients using vital signs.* ***Informatics in Medicine Unlocked***, v. 24, p. 100592, 2021.

HANLEY, J. A.; MCNEIL, B. J. *The meaning and use of the area under a receiver operating characteristic (ROC) curve.* ***Radiology***, v. 143, n. 1, p. 29-36, 1982.

IQBAL, N.; ISLAM, M.: *Machine learning for dengue outbreak prediction: A performance evaluation of different prominent classifiers.* ***Informatica***, v. 43, n. 3, 2019.

JIANG, F. *et al.: Artificial intelligence in healthcare: past, present and future.* ***Stroke and vascular neurology***, v. 2, n. 4, 2017.

JOHN, G. H.; LANGLEY, P.: *Estimating continuous distributions in Bayesian classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence*, San Mateo, 338-345, 1995.

KOHAVI, R. *et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In:* ***Ijcai***. 1995. p. 1137-1145.

KOUROU, K. *et al. Machine learning applications in cancer prognosis and prediction.* ***Computational and structural biotechnology journal***, v. 13, p. 8-17, 2015.

LANDIS, J. R.; KOCH, G. G.: *The measurement of observer agreement for categorical data.* ***biometrics***, p. 159-174, 1977.

LEE, E. K. et al. Projection pursuit for exploratory supervised classification. **Journal of Computational and Graphical Statistics**, Alexandria, v. 14, n. 4, p. 831-846, 2005.

MELLO-ROMÁN, J. D. *et al. Predictive models for the medical diagnosis of dengue: a case study in Paraguay.* ***Computational and mathematical methods in medicine***, v. 2019, 2019.

NARANJO, C. A. *et al. A method for estimating the probability of adverse drug reactions.* ***Clinical Pharmacology & Therapeutics***, v. 30, n. 2, p. 239-245, 1981.

OBERMEYER, Z.; EMANUEL, E. J.: *Predicting the future—big data, machine learning, and clinical medicine.* ***The New England journal of medicine***, v. 375, n. 13, p. 1216, 2016.

OZER, I *et al. Improved machine learning performances with transfer learning to predicting need for hospitalization in arboviral infections against the small dataset.* ***Neural Computing and Applications***, p. 1-15, 2021.

RSTUDIO TEAM. RStudio: Integrated Delevopment for R. RStudio, Inc., Boston, 2020. Disponível em <http://www.rstudio.com/>, acesso em 01 de março de 2022.

SRINIVASAN, R. *et al.*: *Use of 16S rRNA gene for identification of a broad range of clinically relevant bacterial pathogens.* ***PloS one***, v. 10, n. 2, p. e0117617, 2015.

UDDIN, S. *et al.: Comparing different supervised machine learning algorithms for disease prediction.* ***BMC medical informatics and decision making***, v. 19, n. 1, p. 1-16, 2019.

ZHENG, T. *et al*. *A machine learning-based framework to identify type 2 diabetes through electronic health records.* ***International journal of medical informatics***, v. 97, p. 120-127, 2017.

5

CONCLUSÃO

5.1 CONTRIBUIÇÕES

A Dengue é uma doença viral que causa impactos significativos nas áreas onde está presente, tanto do ponto de vista das perdas humanas, quanto dos impactos econômicos. Custo diretos, com tratamentos dos acometidos, e indiretos, com perdas produtivas, estão apontados na literatura. O Paraná é o estado do Sul do Brasil que registra o maior número de casos da doença, demonstrando crescimento no número de notificações na última década. O ano epidemiológico de 2019/2020 foi recorde para a série histórica no estado, com 366.760 notificações entre 28 de julho de 2019 à 01 de agosto de 2020.

Esta pesquisa foi desenvolvida para responder à seguinte pergunta: **de que modo a análise estatística e os modelos de ML podem contribuir com a área da saúde, especificamente do ponto de vista das notificações de dengue no estado do Paraná?** Esta pesquisa demonstrou que as ferramentas abordadas auxiliam na aquisição de conhecimento para melhor compreensão das notificações de dengue no estado do Paraná.

No artigo 1, a estatística descritiva permitiu analisar de modo amplo tanto a quantidade de notificações, hospitalizações, casos agravados e óbitos o ano epidemiológico em estudo, incluindo os percentuais designados para estas variáveis em relação ao total de notificações. A dengue no estado mostra-se, portanto, como desafio à saúde pública, que gera

impactos para a sociedade paranaense. O percentual de hospitalizados foi de 2,99% dos notificados, enquanto óbitos por dengue representaram 0,05%. No total, 66,59% dos notificados foram confirmados.

Esta mesma pesquisa permitiu identificar lacunas no preenchimento do SINAN, principal fonte de informação sobre as notificações de dengue, a partir das quais estudos podem orientar ações futuras de gestão. O campo “escolaridade” foi o que mais deixou de ser preenchido, representando 37,9%, seguido da informação de “hospitalização”, que deixou de ser preenchida e 22,68% dos casos, e “evolução clínica”, não preenchido em 10,43%. O preenchimento do campo “escolaridade” do SINAN em notificações de dengue foi abordado por Guimarães e Cunha (2020).

Em adição à estatística descritiva, o Artigo 1 apresentou a análise de correspondência múltipla para variáveis de notificação de dengue. Os resultados permitem visualizar relações entre as variáveis de notificação em diferentes cenários. O método gráfico permite indicar associações entre sinais clínicos específicos que estão mais próximos dos perfis agravados DSA e DG. Outras relações identificadas dizem respeito às doenças pré-existentes e perfis agravados. Além disso, foi possível observar relações entre sinais clínicos de dengue clássica e proximidades com classificações “confirmadas” e “descartadas” para dengue.

No artigo 2, modelos de classificação supervisionada para notificações de dengue foram propostos. A intenção é construir modelos automatizados de apoio à classificação de novas notificações, a partir das variáveis estruturadas do SINAN. Modelos de notificação apoiados por informações de triagem são úteis principalmente em contextos em que os exames laboratoriais não estão presentes. Para o caso da epidemia de dengue no Paraná de 2019/2020, em torno de um terço dos casos receberam como critério de classificação final a adoção de exames laboratoriais.

A partir dos dados em estudo, o Modelo 1 buscou classificar notificações de dengue, para auxiliar no descarte ou confirmação da doença em situações de triagem clínica. O modelo gerado possui acurácia de 63,38% e índice de concordância Kappa regular, igual a 0,2616. A projeção por meio da técnica *projection pursuit* com o índice LDA permitiu visualizar que os casos de descartados estão sobrepostos aos casos confirmados, dificultando a classificação e, por consequência, a automação da tarefa a partir das variáveis do SINAN. Outrossim, a necessidade de realização de exames laboratoriais para a classificação de dengue mostra-se importante, uma vez que a similaridade com outras infecções dificulta o diagnóstico em casos clássicos, que são a maioria.

O modelo 2, que adicionou a gravidade da dengue às classes de classificação consideradas no modelo 1, não apresentou resultados muito mais satisfatórios. Porém, guiou a pesquisa para geração do modelo 3. Este último, considerou apenas a classificação da gravidade da dengue, para casos confirmados. A acurácia foi de 99,89% e o índice Kappa teve concordância bastante alta, de 0,9794%. E, considerando a técnica *projection pursuit* com o índice LDA, observa-se graficamente a separação espacial dos dados, demonstrando que os agravos possuem características próprias e peculiares, o que contribuiu com a geração do modelo com acurácia elevada.

5.2 DIFICULDADES E LIMITAÇÕES

A estatística descritiva ficou restrita aos dados do SINAN, não sendo abordadas informações de outras fontes, levando em conta que este é o sistema de registro de referência da SESA/PR para notificações de dengue. Não foram abordadas informações estimadas de notificados que não apresentaram sintomas ou não procuraram unidades de saúde e que, portanto, não estão registradas no SINAN.

O preenchimento incompleto do SINAN exigiu redução das amostras em estudo para a análise de correspondência múltipla e para a geração dos modelos de ML, ainda que a quantidade de amostras mantida tenha sido considerada significativa. As variáveis abordadas nestas duas análises ficaram restrita às variáveis estruturadas presentes no SINAN.

A pesquisa leva em conta um período epidêmico, contexto levado em conta na classificação final das notificações, sobretudo pelo critério clínico-epidemiológico. Dados coletados em períodos não epidêmicos podem gerar resultados distintos aos observados nesta pesquisa.

Os dados não foram coletados *in loco*, portanto, foi assumido que houve correto preenchimento do SINAN nas amostras consideradas, por parte dos profissionais de saúde envolvidos.

5.3 TRABALHOS FUTUROS

Objetivos específicos sugeridos para pesquisas futuras foram identificados durante esta pesquisa, a partir dos dados abordados. Neste sentido, tem-se:

- Mensurar o número de casos de dengue no estado do Paraná que não estejam registrados no SINAN;
- Construir modelos de ML para notificações de dengue considerando outras variáveis não abordadas nesta pesquisa;
- Pesquisar de modo aprofundado as relações identificadas na análise de correspondência múltipla;
- Identificar as causas que levam ao não preenchimento dos campos no SINAN;
- Incluir variáveis laboratoriais aos modelos de classificação de ML;
- Explorar outros algoritmos de classificação em ML para notificações de dengue;
- Utilizar análise de cluster aos dados para identificar padrões;
- Explorar campos não estruturados do SINAN para notificações de dengue;

No estudo de ANTONIO *et al*. (2017), é investigada a correlação entre casos de dengue e as coordenadas geográficas no Brasil, onde os resultados indicaram que a concentração de casos varia com a longitude e a aglomeração populacional. Pesquisa similar poderia ser feita para a dengue no Paraná, uma vez que percebe-se maior distribuição da doença nas norte, noroeste, oeste, sudoeste (PARANÁ, 2020).

Além do exposto, uma lacuna de pesquisa relevante é mensurar o impacto da dengue no Paraná, de modo holístico, incluindo os custos financeiros, não obstante ao impacto inestimável à vida humana. Uma vez que 63,21% dos casos, para o período de estudo considerado nesta pesquisa, foi atribuído à população adulta, qual é o impacto em termos de perda de produtividade? Quais setores foram mais afetados? Como mensurar a inoperabilidade? São inquietações que surgem dentro da Engenharia de Produção, a partir da presente pesquisa.

REFERÊNCIAS

ANTONIO, F. J. et al. Spatial patterns of dengue cases in Brazil. **PloS one**, v. 12, n. 7, p.

BAGHERZADEH-KHIABANI, Farideh *et al*. *A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results.* ***Journal of clinical epidemiology***, v. 71, p. 76-85, 2016.

BAVIA, L. *et al*.: *Epidemiological study on dengue in southern Brazil under the perspective of climate and poverty*. ***Scientific Reports***, v. 10, n. 1, p. 1-16, 2020.

BHATT, S. *et al*. :*The global distribution and burden of dengue.* ***Nature***, v. 496, n. 7446, p. 504-507, 2013.

GUIMARÃES, L. M.; CUNHA, G. M. da.: Diferenças por sexo e idade no preenchimento da escolaridade em fichas de vigilância em capitais brasileiras com maior incidência de dengue, 2008-2017. **Cadernos de Saúde Pública**, v. 36, p. e00187219, 2020.

MACEDO HAIR, G.; FONSECA NOBRE, F.; BRASIL, P.: *Characterization of clinical patterns of dengue patients using an unsupervised machine learning approach.* ***BMC infectious diseases***, v. 19, n. 1, p. 1-11, 2019.

MARTIN, B. M. et al. *Clinical outcomes of dengue virus infection in pregnant and non-pregnant women of reproductive age: a retrospective cohort study from 2016 to 2019 in Paraná, Brazil.* ***BMC infectious diseases***, v. 22, n. 1, p. 1-11, 2022.

MESSINA, Jane P. *et al.: The current and future global distribution and population at risk of dengue.* ***Nature microbiology***, v. 4, n. 9, p. 1508-1515, 2019.

MURRAY, N. E. A.; QUAM, M. B.; WILDER-SMITH, A.: *Epidemiology of dengue: past, present and future prospects.* ***Clinical epidemiology***, v. 5, p. 299, 2013.

NEALON, J. *et al*. *Dengue Endemicity, Force of Infection, and Variation in Transmission Intensity in 13 Endemic Countries.* ***The Journal of infectious diseases***, v. 225, n. 1, p. 75-83, 2022.

OBERMEYER, Z.; EMANUEL, E. J.: *Predicting the future—big data, machine learning, and clinical medicine.* ***The New England journal of medicine***, v. 375, n. 13, p. 1216, 2016.

PARANÁ. SECRETARIA DE ESTADO DA SAÚDE. Situação da Dengue, Chikungunya e Zika Vírus no Paraná. Informe técnico 43 – Semana Epidemiológica 31/2019 a 28/2020. Dados divulgados, sujeitos a alterações. SESA/PR, 2020. Disponível em https://www.dengue.pr.gov.br/sites/dengue/arquivos_restritos/files/documento/2020-11/boletimdengue43_2020.pdf>, acesso em 06 de fevereiro de 2022.

RSTUDIO TEAM. RStudio: Integrated Delevopment for R. RStudio, Inc., Boston, 2020. Disponível em <http://www.rstudio.com/>, acesso em 01 de março de 2022.

SCHMIDT, M. *et al*.: *The Danish health care system and epidemiological research: from health care contacts to database records.* ***Clinical epidemiology***, v. 11, p. 563, 2019.

SILVEIRA, D.T.; CÓRDOVA, F.P.: A pesquisa científica. *In* GERHARDT, T.E.; SILVEIRA, T.E.G.:. **Métodos de pesquisa**. [Organizado por] Tatiana Engel Gerhardt e Denise Tolfo Silveira; coordenado pela Universidade Aberta do Brasil–UAB/UFRGS e pelo Curso de Graduação Tecnológica–Planejamento e Gestão para o Desenvolvimento Rural da SEAD/UFRGS. **Porto Alegre: Editora da UFRGS**, p. 31-32, 2009.

STANAWAY, J. D. *et al*.: *The global burden of dengue: an analysis from the Global Burden of Disease Study* 2013. ***The Lancet infectious diseases***, v. 16, n. 6, p. 712-723, 2016.

VECCHIA, A. D.; BELTRAME, V.; D'AGOSTINI, F. Panorama da dengue na região sul do brasil de 2001 a 2017. **Cogitare Enfermagem**, v. 23, n. 3, 2018.