REVISTA HYGEIA **ISSN: 1980-1726**
https://seer.ufu.br/index.php/hygeia DOI: https://doi.org/10.14393/Hygeia64263

ASSOCIAÇÃO ENTRE VARIÁVEIS SOCIOECONÔMICAS E A OCORRÊNCIA DE DENGUE NO ESTADO DE GOIÁS: UMA ANÁLISE A PARTIR DE ALGORITMOS DE MACHINE LEARNING

ASSOCIATION BETWEEN SOCIOECONOMIC VARIABLES AND THE OCCURRENCE OF DENGUE IN THE STATE OF GOIÁS: AN ANALYSIS BASED ON MACHINE LEARNING ALGORITHMS

Thamy Barbara Gioia
Universidade Federal de Goiás, IESA - Instituto de Estudos Socioambientais, Goiânia, GO, Brasil
thamygioia@discente.ufg.br

Juliana Ramalho Barros
Universidade Federal de Goiás, IESA - Instituto de Estudos Socioambientais, Goiânia, GO, Brasil
juliana@ufg.br

RESUMO

A dengue é considerada uma das doenças com índices mais expressivos no Brasil. O crescente aumento nas taxas observadas afeta diretamente os serviços públicos de saúde de forma que avaliar as condições ambientais e sociais em áreas com altos índices da doença pode auxiliar na elaboração de diagnósticos e ações em saúde. Nesse sentido, o objetivo deste trabalho foi identificar variáveis socioeconômicas mais importantes para a predição das taxas de prevalência de dengue nos municípios do estado de Goiás. A avaliação foi realizada com base em 38 variáveis socioeconômicas obtidas no banco de dados do Instituto Brasileiro de Geografia e Estatística - IBGE, da Fundação João Pinheiro - FJP e a partir do cálculo das taxas de prevalência de dengue baseado nos dados disponíveis no Sistema de Informação de Agravos de Notificação - SINAN para os períodos de 2001-2009 e 2010-2018. A modelagem foi realizada a partir da avaliação de três algoritmos de *machine learning*: *Random Forest*, XGBoost e KNN. Os resultados indicaram que as variáveis mais importantes apresentaram relação inversa às condições de baixa renda, analfabetismo e deficiência em serviços de saneamento básico.

Palavras-chave: Determinantes Sociais. Random Forest. XGBoost. KNN.

ABSTRACT

Dengue is considered one of the diseases with the most significant rates in Brazil. The increasing rates directly affect public health services, so that evaluating the environmental and social conditions in areas with high rates of the disease can assist in the development of diagnoses and health actions. In this sense, the objective of this study was to identify the most important socioeconomic variables for the prediction of dengue prevalence rates in municipalities of the state of Goiás. The evaluation was performed based on 38 socioeconomic variables obtained from the database of Instituto Brasileiro de Geografia e Estatística - IBGE, Fundação João Pinheiro - FJP and from the calculation of dengue prevalence rates based on data available in Sistema de Informação de Agravos de Notificação - SINAN for the periods 2001-2009 and 2010-2018. Modeling was performed from the evaluation of three machine learning algorithms: Random Forest, XGBoost and KNN. The results indicated that the most important variables showed an inverse relationship to the conditions of low income, illiteracy and deficiency in basic sanitation services.

Keywords: Social Determinants. Random Forest. XGBoost. KNN.

Recebido em: 14/01/2022
Aceito para publicação em: 27/01/2022.

INTRODUÇÃO

A dengue é uma doença febril aguda causada por um vírus do gênero *Flavivirus*, cujo principal vetor de transmissão é o *Aedes aegypti*. De acordo com o Ministério da Saúde (BRASIL, 2005), é uma das arboviroses mais importantes quando considerados os efeitos sobre os seres humanos e como problema de saúde pública.

Nos últimos 40 anos, a dengue passou a ser um problema de saúde pública não somente no Brasil, mas também em diversos países do mundo, visto que aproximadamente metade da população mundial está em risco de contrair a doença. Nas Américas, onde cerca de 500 milhões de pessoas correm o risco de contrair dengue, observou-se um grande aumento do número de casos nas últimas quatro décadas, passando de 1,5 milhão de casos acumulados na década de 1980 para 16,2 milhões na década de 2010-2019 (OPAS, 2021).

No Brasil, a dengue é uma das doenças com taxas mais expressivas. Segundo dados do DATASUS (1996-2013), com exceção dos estados de Santa Catarina e Rio Grande do Sul, a dengue é endêmica em praticamente todos os estados brasileiros. No Estado de Goiás, entre os anos de 2010 e 2018 as taxas variaram de 70 casos a cada 100.000 habitantes até mais de 2500 casos a cada 100.000 habitantes (SINAN, 2021) o que, de acordo com o Ministério da Saúde, coloca municípios do estado em condições hiperendêmicas de risco (BRASIL, 2005).

Além das condições ambientais que favorecem o desenvolvimento e a proliferação do vetor de transmissão, estudos têm sido realizados com o objetivo de avaliar condições sociais e econômicas de forma a auxiliar no diagnóstico de áreas endêmicas, bem como para proposição de ações e políticas públicas em saúde (HONORATO *et al*., 2014; PAIXÃO *et al*., 2015; ALMEIDA; SILVA, 2018).

Tendo em vista a relação entre os mosquitos e a transmissão de doenças, em diversos momentos observou-se uma intensa busca pela melhoria na qualidade de populações, trazendo ao centro da questão as ações de planejamento, a valorização do saneamento básico, bem como a promoção das condições de higiene e de saúde pública, o que, no Brasil, resultou na Reforma Sanitária. Embora, em alguns momentos, tenha havido o controle dos vetores, observa-se, nas últimas 4 décadas, a reincidência de várias infecções causadas pelos mosquitos *Aedes aegypti* transmissores de dengue, malária e febre amarela (MENDONÇA; SOUZA; DUTRA, 2009).

Diante da falta de vacina eficaz no combate a todas as sorologias do vírus transmissor da dengue, é fundamental que haja esforços no sentido de prever cenários nos quais a doença tem as maiores e as menores possibilidades de se disseminar. Nesse sentido, é importante buscar trabalhar com o maior número de variáveis possíveis, tendo em vista que cada vez mais se observa o quão complexas são as causas de incidência e reincidência de doenças como a dengue.

Nesse sentido, o objetivo deste trabalho foi identificar variáveis socioeconômicas mais importantes para a predição das taxas de prevalência de dengue nos municípios do estado de Goiás.

Para isso, foram utilizados dados de saúde disponíveis no banco de dados do Sistema de Informações de Agravos de Notificação – SINAN e dados socioeconômicos disponíveis nas bases do Instituto Brasileiro de Geografia e Estatística - IBGE e da Fundação João Pinheiro - FJP. Para identificação das variáveis mais importantes foram testados três algoritmos de *machine learning*: *Random Forest, XGBoost* e KNN.

MATERIAIS E MÉTODOS

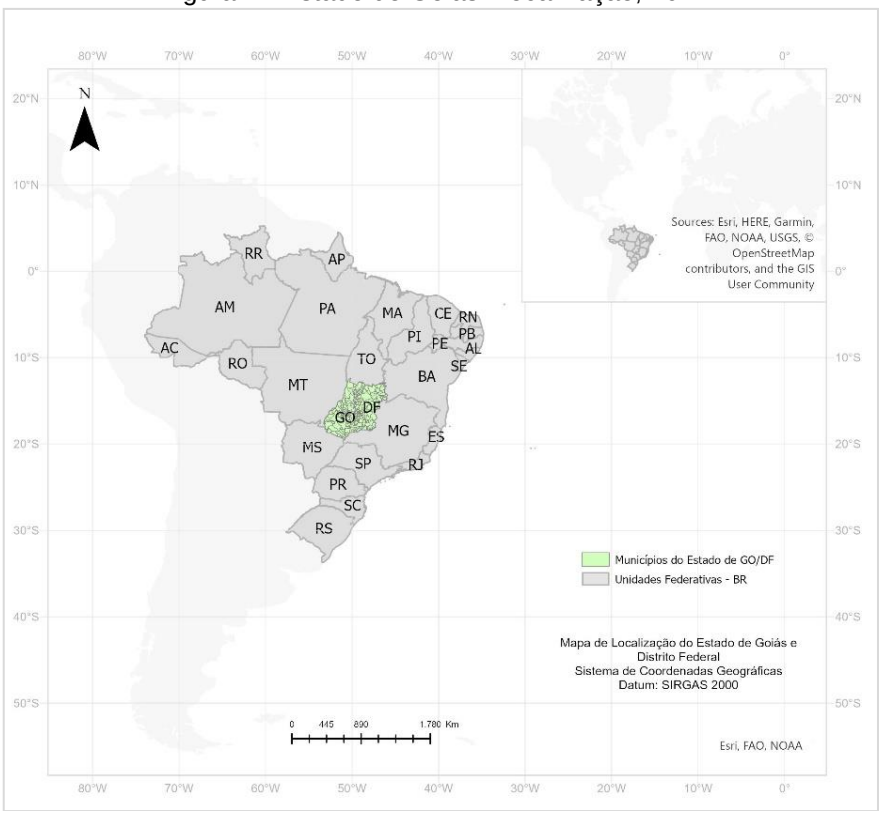
Área de estudo

Como área de estudo, optou-se por analisar os 246 municípios do Estado de Goiás, que está localizado na região Centro-Oeste do Brasil (Figura 1). De acordo com o IBGE (2021), o estado possui área de 340.111,38 km² e população estimada de 7.018.354 habitantes.

Mais de 90% do território goiano encontra-se dentro dos limites oficiais do bioma Cerrado. Localizado em área de clima tropical, sua rede de drenagem é densa e constituída por rios de médio e grande porte, que alimentam três importante Regiões Hidrográficas do país (Araguaia/Tocantins, São Francisco e Paraná) (IMB, 2018).

O território goiano possui dois parques nacionais: das Emas e Chapada dos Veadeiros; 12 (doze) áreas definidas como parques estaduais, onde se destacam o Parque da Serra de Caldas Novas e o Parque de Terra Ronca, além de inúmeras outras unidades de proteção ambiental.

Figura 1 - Estado de Goiás. Localização, 2021.



Fonte: IBGE, 2021. Elaboração: GIOIA, T. B., 2021.

O processamento dos dados

A análise proposta foi construída com base em dados referentes ao total de casos de dengue registrados no banco de dados do SINAN (2021), disponível em: http://www2.datasus.gov.br/DATASUS/index.php?area=0203&id=29878153, referentes ao período de 2001 a 2018. A partir do quantitativo de casos foram calculadas as taxas de prevalência média para dois períodos, 2001-2009 e 2010-2018, e a cada 100.000 habitantes, com o objetivo de descartar possíveis flutuações aleatórias nos registros disponibilizados. As taxas de prevalência dos dois períodos foram calculadas a partir das projeções populacionais e do censo demográfico (2000 e 2010), disponíveis no banco dados do IBGE.

No total, foram avaliadas 38 variáveis socioeconômicas obtidas no banco de dados do censo demográfico do IBGE (2000 e 2010): https://sidra.ibge.gov.br/pesquisa/censo-demografico/demografico-2010/inicial; e no banco de dados da FJP: http://migracao.fjp.mg.gov.br/ (Quadro 1). Para a espacialização das taxas foram utilizadas bases cartográficas digitais em formato vetorial, também disponíveis no banco de dados do IBGE: https://www.ibge.gov.br/geociencias/downloads-geociencias.html.

A avaliação considerou, além das taxas de prevalência previstas para os municípios do Estado de Goiás, dados referentes aos mesmos períodos para o Distrito Federal. Optou-se por esse caminho devido a possíveis influências que o Distrito Federal exerce sobre os municípios limítrofes. Desta forma, foram consideradas, para análise, 247 unidades administrativas (246 municípios do Estado de Goiás e o Distrito Federal), em dois períodos (2001-2009 e 2010-2018), resultando em uma planilha de 494 amostras.

Quadro 1 - Variáveis independentes utilizadas para modelagem dos algoritmos.

Cód.	Descrição	Fonte/ano disp. dos dados
V01	Média de moradores por domicílio	IBGE(2000/2010)
V02	População urbana (%)	IBGE(2000/2010)
V03	População rural (%)	IBGE(2000/2010)
V04	Famílias únicas (%)	IBGE(2000/2010)
V05	Famílias conviventes (%)	IBGE(2000/2010)
V06	PEA – População economicamente ativa (%)	IBGE(2000/2010)
V07	PNEA – População não economicamente ativa (%)	IBGE(2000/2010)
V08	Domicílios com 1 dormitório (%)	IBGE(2000/2010)
V09	Domicílios com 2 dormitórios (%)	IBGE(2000/2010)
V10	Domicílios com 3 dormitórios (%)	IBGE(2000/2010)
V11	Domicílios com 4 dormitórios (%)	IBGE(2000/2010)
V12	População acima de 10 anos com classe de renda até 1 salário-mínimo (%)	IBGE(2000/2010)
V13	População acima de 10 anos com classe de renda mais de 1 a 2 salários-mínimos (%)	IBGE(2000/2010)
V14	População acima de 10 anos com classe de renda de mais de 2 a 3 salários-mínimos (%)	IBGE(2000/2010)
V15	População acima de 10 anos com classe de renda de mais de 3 a 5 salários-mínimos (%)	IBGE(2000/2010)
V16	População acima de 10 anos com classe de renda de mais de 5 a 10 salários-mínimos (%)	IBGE(2000/2010)
V17	População acima de 10 anos com classe de renda de mais de 10 a 20 salários-mínimos (%)	IBGE(2000/2010)
V18	População acima de 10 anos com classe de renda acima de 20 salários-mínimos (%)	IBGE(2000/2010)
V19	População acima de 10 anos sem rendimento (%)	IBGE(2000/2010)
V20	População autodeclarada branca (%)	IBGE(2000/2010)
V21	População autodeclarada preta (%)	IBGE(2000/2010)
V22	População autodeclarada amarela (%)	IBGE(2000/2010)
V23	População autodeclarada parda (%)	IBGE(2000/2010)
V24	População autodeclarada indígena (%)	IBGE(2000/2010)
V25	Taxa liquida migratória	FJP (2000-2010)
V26	Domicílios com rede geral de esgoto ou pluvial (%)	IBGE(2000/2010)
V27	Domicílios com fossas sépticas (%)	IBGE(2000/2010)
V28	Domicílios com tipo de esgotamento inadequado (%)	IBGE(2000/2010)
V29	Domicílios com 1 banheiro (%)	IBGE(2000/2010)
V30	Domicílios com 2 banheiros (%)	IBGE(2000/2010)
V31	Domicílios sem banheiro (%)	IBGE(2000/2010)
V32	Domicílios com coleta de resíduos de serviço público (%)	IBGE(2000/2010)
V33	Domicílios com disposição inadequada de resíduos (%)	IBGE(2000/2010)
V34	Domicílios com abastecimento de água por rede pública (%)	IBGE(2000/2010)
V35	Domicílios com abastecimento de água via poço na propriedade (%)	IBGE(2000/2010)
V36	Domicílios com abastecimento inadequado de água (%)	IBGE(2000/2010)
V37	População alfabetizada (%)	IBGE(2000/2010)
V38	População não alfabetizada (%)	IBGE(2000/2010)

Elaboração: GIOIA, T. B., 2021.

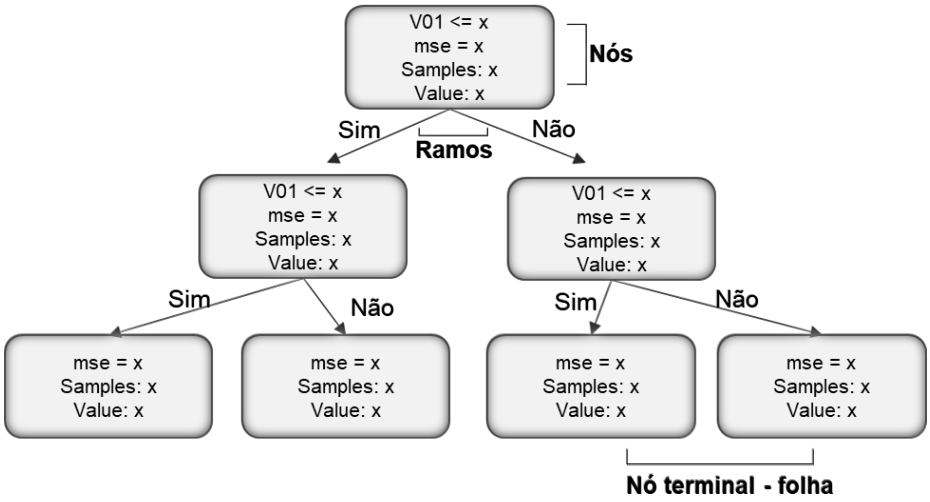
Algoritmos de machine learning

Para a modelagem dos resultados foram empregadas as ferramentas disponíveis no *software* gratuito R, versão 4.0.3 pacote *Classification and Regression Training* - CARET (KUHN, 2017), e *scripts* específicos para avaliação de dois algoritmos baseados em árvore de decisão: o *Random Forest* (BREIMAN; CLUTER, 2001) e *o XGBoost* (CHEN; GUESTRIN, 2016); e do algoritmo KNN (RIPLEY, 1996).

Os algoritmos baseados em árvore de decisão podem ser utilizados para problemas de classificação e regressão. Neste caso, os algoritmos *Random Forest* e *XGBoost* foram aplicados ao modelo de regressão. Em árvores de regressão, os dados são divididos em grupos de forma a tornar as médias de resposta para cada grupo tão diferentes quanto possível. As regras de divisão que definem os nós

de cada grupo estão relacionadas através de árvores binárias e são estimadas por meio de algoritmos específicos (Figura 2). As respostas previstas são obtidas fazendo uso da média de previsão de todas as árvores (HASTIE; ROBERT; FRIEDMAN, 2008).

Figura 2 - Modelo de árvore de regressão.



Elaboração: GIOIA, T. B., 2021.

No algoritmo *Random Forest* são elaboradas múltiplas árvores de decisão independentes em uma estratégia de *bagging* (*bootstrap aggregating*) (BREIMAN; CLUTER, 2001), na qual cada árvore é gerada com parte das variáveis preditoras aleatoriamente para evitar a correlação entre as árvores. Na predição por meio da abordagem de regressão, considera-se o valor médio entre as árvores individuais. Já o algoritmo *XGBoost* utiliza a estratégia de aprendizagem por reforço (*gradient boosting*), na qual uma série de árvores de decisão sequenciais é elaborada e o aprendizado de cada árvore depende da árvore anterior (CHEN; GUESTRIN, 2016).

O algoritmo KNN refere-se a um modelo de algoritmo não-paramétrico que considera os valores médios de instâncias semelhantes (vizinhos mais próximos) para associação a instância testada. A métrica de distância a ser calculada considera a equação de distância Euclidiana (RIPLEY, 1996).

Na etapa de modelagem, os dados foram utilizados para treinamento e calibração, onde aplicou-se a validação cruzada k-fold para k=5. O método de validação k-fold é utilizado para calibrar hiperparâmetros e avaliar o melhor desempenho de um modelo. Neste método, emprega-se parte dos dados disponíveis para adequar o modelo e outra parte para testá-lo, de forma que os dados são divididos em partes de igual tamanho (k=N), ou seja, todas as partes da divisão serão utilizadas para treino e validação buscando reduzir o sobreajuste (*overfitting*) e a melhoria do modelo.

Os valores típicos aplicados para k são 5 ou 10 (BREIMAN; SPECTOR, 1992; HASTIE; ROBERT; FRIEDMAN, 2008), sendo k=5 utilizado para baixos valores de amostras e k=10 para universo de amostras mais relevantes. Este limiar: baixo ou alto deverá ser definido empiricamente pelo pesquisador, devendo considerar em seus testes questões de viés e subestimação (HASTIE; ROBERT; FRIEDMAN, 2008). A partir da observação e avaliação de testes prévios optou-se por aplicar neste trabalho k-fold= 5, considerando N=494 como um universo de amostras de baixo limiar.

Como critério de avaliação do algoritmo de melhor desempenho foram empregadas as seguintes métricas: coeficiente de determinação (R^2), que pode ser interpretado também em porcentagem de explicação do modelo na predição, e a raiz quadrada do erro médio quadrático (RMSE), que indica uma métrica de erro referente a taxa predita.

Para o algoritmo de treinamento de melhor desempenho foi computada a importância de cada variável (GRÖMPING, 2015) na estimativa da TPD - Taxa de Prevalência de Dengue a cada 100.000 habitantes nos municípios do Estado de Goiás nos dois períodos analisados. Os níveis de importância foram normalizados de 0 a 100 para cada variável. A partir das variáveis mais importantes avaliou-se a

distribuição espacial destas variáveis comparando-as a distribuição das taxas de dengue. Para auxiliar na observação e análise descritiva delimitou-se limiares de interpretação:

Para condição da variável:

- **Alta:** valores acima da média de distribuição dos dados;
- **Baixa:** valores abaixo da média de distribuição dos dados;

Para condição das taxas:

- **Alta:** acima de 1000 casos a cada 100.000 habitantes;
- **Média a baixa**: entre 500 e 1000 casos a cada 100.000 habitantes;
- **Baixo:** abaixo de 500 casos a cada 100.000 habitantes.

Por fim, para auxiliar a interpretação dos resultados, a espacialização das TPD foi comparada a espacialização do IVS - Índice de Vulnerabilidade Social e do mapa de logística de transporte do estado de Goiás (IMB, 2018).

RESULTADOS E DISCUSSÕES

Os resultados da pesquisa indicaram que as melhores métricas de validação foram obtidas por meio do algoritmo *Random Forest* segundo a validação cruzada k-fold = 5 com R^2 = 0,50 e RMSE de 361,58 (Tabela 1).

Tabela 1- Resultados de validação. RMSE e R^2 para avaliação das taxas de dengue.

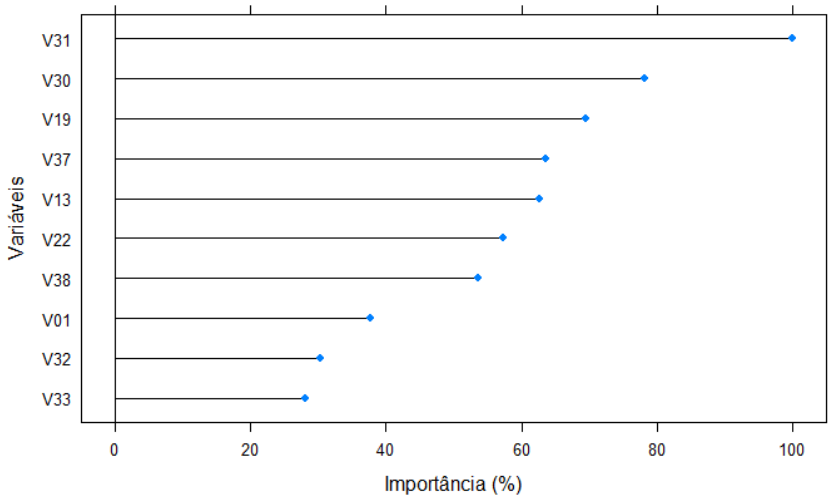
Algoritmos/Taxa	***Random Forest***		***XGboost***		***KNN***	
Validação cruzada k-fold =5	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**
Taxas de dengue	361,58	**0,50**	383,11	0,45	419,08	0,34

Elaboração: GIOIA, T. B., 2021.

Considerando as respostas do algoritmo *Random Forest* na abordagem de validação cruzada *k-fold*, destacam-se, na Figura 3, as 10 variáveis mais importantes associadas às TPD nos municípios do Estado de Goiás: V31 – Domicílios sem banheiro; V30 – Domicílios com até 2 banheiros; V19 – Pessoas sem rendimento; V37- População alfabetizada; V13 – Renda Média de 1 a 2 salários-mínimos; V22 – População de raça/cor amarela; V38 – População não alfabetizada; V01 – Média de Moradores; V32 – Domicílio com coleta de resíduos sólidos; e V33 – Domicílios com disposição inadequada de resíduos sólidos.

A partir dos resultados apontados na Figura 3, avaliou-se a distribuição dos dados para cada variável de importância comparando-as à distribuição das TPD (Quadro 2).

Figura 3 - Resultado das variáveis mais importantes conforme resultados do algoritmo *Random Forest* para as TPD – Taxas de prevalência de dengue.



Elaboração: GIOIA, T. B., 2021.

Quadro 2 - Análise e avaliação das variáveis mais importantes do modelo *Random Forest* para as TPD - Taxas de prevalência de dengue.

Cód.	Variável	Condição da variável	Condição de TPD	Observação
V31	Domicílios sem banheiro (%)	Alta	Baixa	TPD foram mais baixas em municípios onde a porcentagem de domicílios sem banheiro eram mais altas
V30	Domicílios com 2 banheiros (%)	Alta	Alta	TPD foram mais altas em municípios onde 15% a 25% dos domicílios detinham de até 2 banheiros em suas residências;
V19	População acima de 10 anos sem rendimento (%)	Alta	Baixa	TPD baixas onde a porcentagem de pessoas sem rendimento estava entre 40% e 60% da população
V37	População alfabetizada (%)	Alta	Alta	TPD aumentam conforme aumenta a porcentagens de alfabetização - acima de 75%
V13	População acima de 10 anos com classe de renda mais de 1 a 2 salários-mínimos (%)	Alta	Média a Alta	TPD de médias a altas onde 20% a 30% da população tinham renda média de 1 a 2 salários-mínimos
V22	População autodeclarada amarela (%)	Baixa	Baixa a Média	TPD de médias a baixas onde a porcentagem alta declarada amarela também era baixa, até 4%
V38	População não alfabetizada (%)	Baixa	Média a Baixa	TPD diminuiu conforme diminuiu a porcentagem de alfabetização
V01	Média de moradores por domicílio	Baixa	Alta	TPD aumenta com média de moradores entre 3 e 3,5.
V32	Domicílios com coleta de resíduos de serviço público (%)	Alta	Alta	TPD aumenta conforme aumenta a porcentagens de coleta adequada de resíduos - acima de 50%
V33	Domicílios com disposição inadequada de resíduos (%)	Alta	Média a Baixa	TPD diminuiu conforme diminuiu a porcentagem de domicílios com coleta inadequada

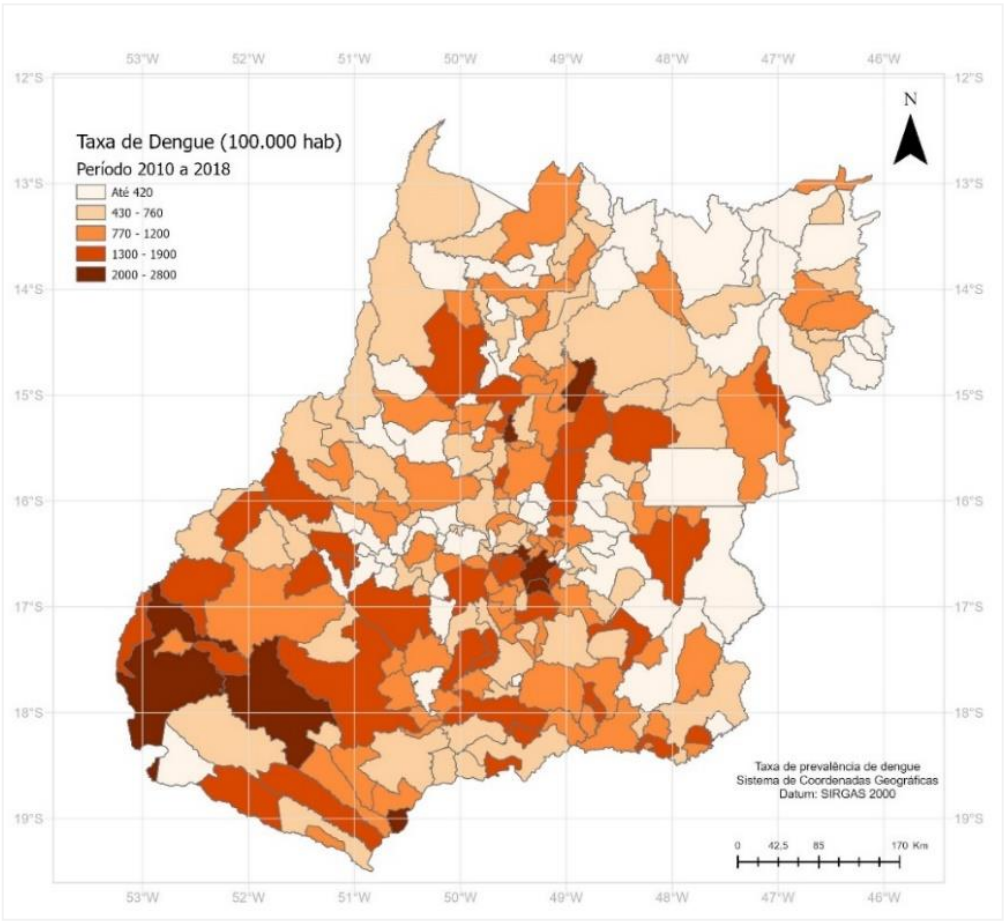
Elaboração: GIOIA, T. B., 2021.

Por meio desses resultados infere-se, preliminarmente, que parte das variáveis socioeconômicas possui uma relação inversa às condições das TPD observadas em municípios do estado.

Leite (2010), ao estudar a relação entre variáveis socioeconômicas e casos de dengue em Montes Claros (MG), encontrou associações positivas entre condições inadequadas de saneamento básico e casos de dengue. O mesmo foi observado por Honorato *et al.* (2014), ao analisar a relação entre risco e variáveis socioeconômicas e dengue no Estado do Espírito Santo. Desta forma, onde poderia ser esperada uma relação direta entre condições inadequadas de saneamento básico e TPD, o que se verifica, no caso dos municípios de Goiás é o oposto, ou seja, no geral, onde há percentuais positivos relacionados a condições de saneamento básico – disposição de resíduos sólidos e domicílios com banheiro, também se apresentam maiores TPD.

Quando analisadas as TPD referentes ao período 2010-2018 (Figura 4), identifica-se que as maiores taxas foram atribuídas a municípios da região centro-sul e sudoeste do estado de Goiás onde as condições de vulnerabilidade social (IMB, 2018) são inferiores se comparadas aos demais municípios do Estado (Figura 5). Boa parte dos municípios com taxas acima de 420 casos a cada 100.000 habitantes está classificada nos grupos 1, 3 e 5 do IVS – Índice de Vulnerabilidade Social elaborado pelo Instituto Mauro Borges (IMB, 2018), na qual: **1** – refere-se a municípios com histórico de desenvolvimento econômico significativo, bom capital físico, boa distribuição de renda e estrutura dos domicílios; **3** – municípios com boa distribuição de renda e educação e **5** – bom mercado de trabalho, renda e infraestrutura dos domicílios e a menor taxa de vulnerabilidade à pobreza[2].

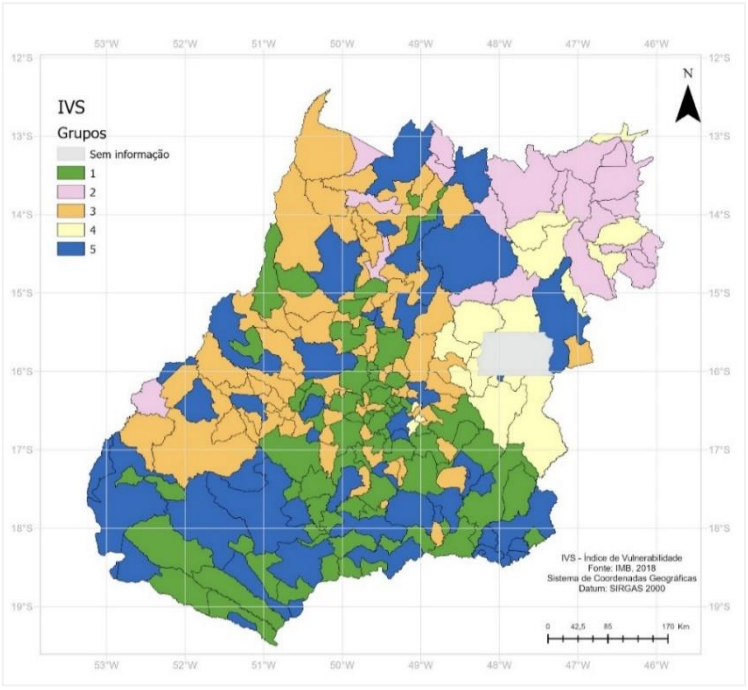
Figura 4 – Municípios do Estado de Goiás e Distrito Federal: Taxa de prevalência de dengue, 2010-2018.



Fonte: SINAN, 2021. Elaboração: GIOIA, T. B., 2021.

[2] Para mais informações relacionadas as demais classes consultar o caderno de vulnerabilidade na íntegra, disponível em https://www.imb.go.gov.br/files/docs/publicacoes/estudos/2018/a-vulnerabilidade-social-dos-municipios-goianos.pdf

Figura 5 - Municípios do Estado de Goiás: IVS – Índice de Vulnerabilidade Social, 2018.



Fonte: IMB, 2018. Organização: GIOIA, T. B., 2021.

Resultados similares foram observados por Machado, Oliveira e Souza-Santos (2009) ao analisar a ocorrência de dengue e as condições de vida no município de Nova Iguaçu no Rio de Janeiro entre 1996 e 2004. Neste caso, os resultados indicaram baixas correlações entre variáveis relacionadas a condições de vulnerabilidade social e taxas de incidência de dengue.

No caso do estado de Goiás, Souza, Silva e Silva (2010) sugerem que a transmissibilidade da dengue, bem como o crescimento de casos no estado, pode estar associada a fatores que vão além de perfis socioeconômicos, relacionados a condições físicas e ambientais e à atuação permanente dos serviços de saúde e de vigilância epidemiológica, aliados a políticas e ações em saúde pública.

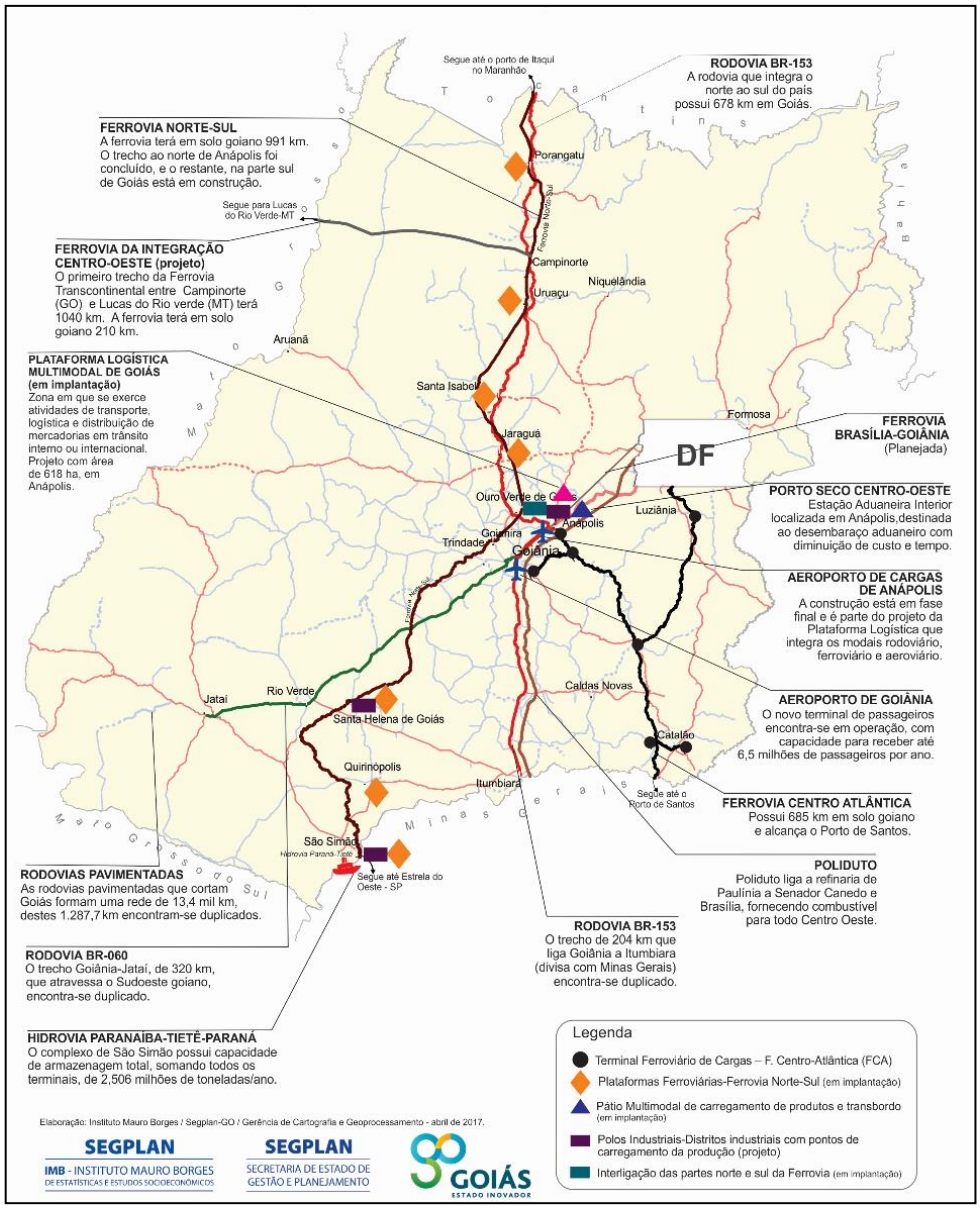
Além disso, as áreas em que ocorreram as maiores TPD em Goiás (sudeste, sul e centro) são aquelas em que há maior densidade demográfica e que apresentam as maiores taxas de urbanização no estado. Segundo documento técnico editado pela OPAS em 2019, a maior incidência das doenças transmitidas pelo *Aedes aegypti* está nas áreas urbanas. Isso se deve à elevada densidade demográfica e à dinâmica populacional e não somente a problemas inerentes à infraestrutura de saneamento básico, suprimento de água e coleta de lixo, embora estes ainda representem um grande desafio para o controle do vetor.

A região do nordeste goiano, que apresentou as menores taxas de prevalência de dengue, possui a menor densidade demográfica e grandes vazios urbanos, o que pode ser fator dificultador de proliferação da doença. Contudo, é preciso investigar.

Cabe lembrar que, conforme apontam Barrera Pérez *et al.* (2015), os programas de saneamento básico, coleta de lixo e suprimento de água raramente atendem a todos os domicílios de uma cidade. Os autores apontam, ainda, que os programas verticais não levam em conta a heterogeneidade e a diversidade de cenários da ecologia do A. *aegypti*, nem os ciclos de transmissão em nível local. Desta forma, é possível que esteja aí uma das causas de a transmissão e a difusão da doença não responderem tão bem a todos os indicadores socioeconômicos.

Outro conjunto de variáveis que merece ser investigado refere-se à logística de transportes e aos principais eixos de circulação do estado. A Figura 6 revela que há uma coincidência entre as áreas de maior TPD e aquelas onde há maior incremento da logística de transporte. A maior facilidade de ligação com outras regiões do país e a grande circulação de pessoas e mercadorias pode favorecer o aumento da difusão da doença.

Figura 6 – Estado de Goiás: Mapa da logística de transporte, 2017.



Fonte: IMB, 2021.

Diante dos resultados obtidos neste trabalho, entende-se a necessidade de um maior aprofundamento das investigações, tanto no que se refere aos resultados obtidos como em relação aos próprios dados. Isso é necessário para que se possa aprimorar o nível de confiança do estudo.

Nas discussões sobre a difusão de doenças, um dos problemas enfrentados refere-se aos dados de localização dos registros, visto que há muitos casos em que os dados estão incompletos ou possuem endereços incorretos. Com a dengue, isso não é diferente.

CONSIDERAÇÕES FINAIS

A proposta deste trabalho consistiu na identificação de variáveis mais importantes a predição das taxas de dengue para os municípios do estado de Goiás, considerando-se um período de 17 anos – 2001 a 2009 e 2010 a 2018 a partir da avaliação de três algoritmos de machine learning: *Random Forest, XGBoost* e KNN.

Com base nos critérios de R^2 e RMSE como métricas de validação foi possível constatar que, dentre os algoritmos avaliados, a melhor resposta foi obtida por meio do algoritmo *Random Forest*, que

apresentou um R^2 de 0,50, o que significa que o modelo foi capaz de explicar 50% da taxa de prevalência de dengue no período analisado para a área de estudo.

Nesse modelo as variáveis mais importantes apresentaram relação com a condição de renda, alfabetização e condições de saneamento básico, no entanto, em uma perspectiva inversa de vulnerabilidade, ou seja, municípios onde as taxas de prevalência de dengue foram mais altas, no geral, apresentaram boas condições de saneamento, alfabetização, e condições mais razoáveis de renda.

Tendo em vista as características da doença no que se refere às condições ideais para a propagação do mosquito transmissor, não se pode deixar de avaliar variáveis socioeconômicas, juntamente com as ambientais. Entretanto, os resultados de qualquer estimativa devem ser vistos sempre com cautela e o devido senso crítico.

Os resultados do presente trabalho indicam a necessidade de avanços nas discussões relativas às condições de transmissibilidade da dengue, bem como dos determinantes sociais e ambientais correlacionados, visando a proposição de ações e políticas públicas de saúde mais efetivas.

É de suma importância estabelecer as variáveis mínimas necessárias que devem ser contempladas pelos sistemas de informação e vigilância em todos os cenários de risco de transmissão que possam ser definidos, levando-se em conta: características epidemiológicas; fatores relacionados com o vetor; e fatores demográficos (humanos).

Nesse sentido, deve-se ampliar e aprofundar as investigações, inserindo quantos elementos e variáveis forem possíveis a fim de que se possa contribuir para a elaboração de políticas públicas para o combate à dengue.

AGRADECIMENTOS

Ao CNPq pelo auxílio financeiro através da bolsa de pesquisa a nível de doutorado concedida à primeira autora: processo nº. 141498/2019-6.

REFERÊNCIAS

ALMEIDA, C.A.P de; SILVA, R.M. da. Análise da ocorrência dos casos de dengue e sua relação com as condições socioambientais em espaços urbanos: os casos de João Pessoa, Cabedelo e Bayeux, no estado da Paraíba – Brasil. **Hygeia**. v. 27, p. 59-79, 2018. Disponível em: https://seer.ufu.br/index.php/hygeia/article/view/38370. Acesso em 02 maio 2021. DOI. https://doi.org/10.14393/Hygeia142705.

BARRERA-PÉREZ, M.A, *et al*. Control de criaderos de Aedes aegypti con el programa Recicla por tu bienestar en Mérida, México. **Salud Publica Mex.** v 57, n. 3, p. 201-210, 2015. Disponível em: https://www.redalyc.org/pdf/106/10638801002.pdf. Acesso em 14 jan. 2022.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. **Guia de Vigilância Epidemiológica.** Brasília: Ministério da Saúde, 2005. Disponível em: https://bvsms.saude.gov.br/bvs/publicacoes/Guia_Vig_Epid_novo2.pdf. Acesso em 09 maio 2021.

BREIMAN, L.; CLUTER. **RandomForest: Random Forests for Classification and Regression, 2001**. Disponível em: https://cran.r-project.org/web/packages/randomForest/index.html. Acesso em: 01 maio 2021.

BREIMAN, L; SPECTOR, P. Submodel selection and evaluation in regression: the X-random case, **International Statistical Review,** v. 60, n.3, p. 291–319, 1992. Disponível em: https://www.jstor.org/stable/1403680. Acesso em: 20 maio 2021.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. *In*: International Conference on Knowledge Discovery and Data Mining, 2016, San Francisco. **Anais** […]. San Francisco: CA, 2016. Disponível em: https://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf. Acesso em: 02 maio 2020. DOI. https://doi.org/10.1145/2939672.2939785.

DATASUS – Departamento de Informática do SUS. Disponível em: http://tabnet.datasus.gov.br/tabdata/LivroIDB/2edrev/d0206.pdf. Acesso em 09 maio 2021.

FJP – Fundação João Pinheiro. **Movimentos Migratórios no Brasil**. 2000 a 2010. Disponível em: http://migracao.fjp.mg.gov.br/. Acesso em: 1 maio 2021.

GRÖMPING, U. Variable importance in regression models. **Wiley Interdisciplinary Reviews: Computational Statistics**, v. 7, n. 2, p. 137-152, 2015. Disponível em:

https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.1346. Acesso em: 01 maio 2020. DOI. https://doi.org/10.1002/wics.1346.

HASTIE, T; TIBSHIRANI, R; FRIEDMAN, J. **The elements of statistical learning. Data mining, inference, and prediction,** 2008. Springer. Disponível em: https://web.stanford.edu/~hastie/Papers/ESLII.pdf. Acesso em: 30 abr. 2021.

HONORATO, T. *et. al*. Análise especial do risco de dengue no Espírito Santo, Brasil, 2010: uso de modelagem completamente Bayesiana. **Revista Brasileira de Epidemiologia**. v.17, 2014. Disponível em: shorturl.at/nyMS2. Acesso em 15 maio. 2021. DOI. https://doi.org/10.1590/1809-4503201400060013.

IBGE – Instituto Brasileiro de Geografia e Estatística. **Geociências**. Disponível em: https://www.ibge.gov.br/geociencias/downloads-geociencias.html. Acesso em: 15 jan.2021.

IMB – Instituto Mauro Borges de Estatísticas e Estudos socioeconômicos. **A vulnerabilidade Social nos municípios goianos**. 2018. Disponível em: https://www.imb.go.gov.br/files/docs/publicacoes/estudos/2018/a-vulnerabilidade-social-dos-municipios-goianos.pdf. Acesso em: 15 fev.2021.

___. **Sobre Goiás** – Mapa de Logística de Transporte. 2017. Disponível em: https://www.imb.go.gov.br/index.php?option=com_content&view=article&id=79:goi%C3%A1s-vis%C3%A3o-geral&catid=232&Itemid=145#meio-ambiente. Acesso em: 15 dez. 2021.

KUHN, M. **Caret**: classification and regression training. R package version 6.0-76; 2017. Disponível em: https://CRAN.R-project.org/package=caret. Acesso em: 20 abr. 2021.

LEITE, M.E. Análise de correlação entre dengue e indicadores sociais a partir de SIG. **Hygeia**. v. 6, n. 11, p. 44-59, 2010. Disponível em: http://www.seer.ufu.br/index.php/hygeia/article/view/16981/9367. Acesso em 01 maio 2021.

MACHADO, J.P; OLIVEIRA, R.M; SOUZA-SANTOS, R. Análise espacial da ocorrência de dengue e condições de vida na cidade de Nova Iguaçu, Estado do Rio de Janeiro, Brasil. **Caderno de Saúde Pública**. v. 25, n.5, 2009. Disponível em: https://www.scielo.br/j/csp/a/mx95V5dsBqJsbPVMmPFKWWx/?lang=pt. Acesso em 01 maio 2021.DOI. https://doi.org/10.1590/S0102-311X2009000500009.

MENDONÇA, F. de A.; SOUZA, A. V. e; DUTRA, D. de A. Saúde pública, urbanização e dengue no Brasil. **Sociedade & Natureza**, Uberlândia. v.21, n. 3, p. 257-269, dez. 2009. DOI. https://doi.org/10.1590/S1982-45132009000300003.

OPAS – Organização Pan-Americana da Saúde. **Dengue**. Disponível em: https://www.paho.org/pt/topicos/dengue. Acesso em: 02 dez. 2021.

___. **Documento técnico para a implementação de intervenções baseado em cenários operacionais genéricos para o controle do Aedes aegypti**. Washington, D.C.: OPAS; 2019.

PAIXÃO, E.S. *et al.* Trends and factors associated with dengue mortality and fatality in Brazil. **Revista da Sociedade Brasileira de Medicina Tropical**. v. 48, n. 4, p. 399-405, 2015. Disponível em: https://www.scielo.br/j/rsbmt/a/qXt6BSSrGJW5GQ8X4sqFmKc/?lang=en. Acesso em: 05 maio 2021. DOI. https://doi.org/10.1590/0037-8682-0145-2015.

R-PROJECT. Disponível em: https://cran.r-project.org/. Acesso em:10 jan.2018.

RIPLEY, B. D. **knn** - k-nearest neighbour, 1996. Disponível em: R Documentation.

SINAN – Sistema de Informação de Agravos de notificação. 2021. Disponível em: http://portalsinan.saude.gov.br/dados-epidemiologicos-sinan. Acesso em 10 fev.2021.

SOUZA, S.S de; SILVA, I.G. da; SILVA, H. H. G. Associação entre incidência de dengue, pluviosidade e densidade larvária de Aedes aegypti, no Estado de Goiás. **Revista da Sociedade Brasileira de Medicina Tropical**. v. 43, n. 2, p.152-155, 2010. Disponível em: https://www.scielo.br/scielo.php?script=sci_arttext&pid=S0037-86822010000200009. Acesso em 20 fev. 2021. DOI. https://doi.org/10.1590/S0037-86822010000200009