CONEXÕES
CIÊNCIA E TECNOLOGIA

UTILIZAÇÃO DE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA NA PREDIÇÃO DE ARBOVIROSES TRANSMITIDAS PELO *AEDES AEGYPTI*

FRANCISCA RAQUEL DE VASCONCELOS SILVEIRA , LINA YARA MONTEIRO REBOUÇAS MOREIRA
Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
raquel_silveira@ifce.edu.br, linayara@gmail.com

DOI: 10.21439/conexoes.v14i1.1824

Resumo. Doenças de característica endêmica exigem atenção aumentada, pois conseguem disseminar-se com facilidade. A dengue, a chikungunya e a zika são exemplos de doenças com características endêmicas de notificação compulsória agregadas ao Sistema de Informação de Agravos de Notificação (SINAN). Essas doenças têm atingido diversos estados no país, causando epidemias em várias regiões. Diversas iniciativas foram tomadas para conter o avanço do mosquito transmissor dessas doenças, contudo ele se desenvolve rapidamente e, em ambientes favoráveis, se reproduz com facilidade. Uma solução para este problema é o uso de ferramentas inteligentes capazes de auxiliar especialistas em saúde no processo de tomada de decisão no manejo clínico de doenças complexas. Neste contexto, a pesquisa em questão tem como objetivo utilizar algoritmos de aprendizagem de máquina para predizer casos das arboviroses dengue e chikungunya, transmitidas pelo mosquito Aedes aegypti, a partir de características associadas ao paciente, tais como, sintomas, idade, sexo, período dos sintomas, dentre outros. Para a predição, os dados passam por uma etapa de pré-processamento, processamento e análise. Foram utilizados três algoritmos de aprendizagem de máquina para comparação de resultados: J48, *Random Forest* e Redes Neurais, com o balanceamento de dados através do SMOTE. A partir dos resultados obtidos, é possível evidenciar que o algoritmo Random Forest apresenta melhores resultados se comparados com o demais, alcançando 90,6443% de acurácia e 0,907 de *f-measure*, sendo, portanto, uma alternativa promissora para a predição de dengue e chikungunya.

Palavras-chave: Dengue. Chikungunya. Aedes aegypti. Aprendizagem de Máquina.

USE OF MACHINE LEARNING ALGORITHMS IN THE PREDICTION OF AEDES AEGYPTI TRANSMITTED ARBOVIRUSES

Abstract. Endemic diseases require increased attention since they can spread easily. Dengue fever, chikungunya and zika are examples of diseases with endemic features of compulsory notification added to the System of Information on Harm Notice (SINAN). These diseases have affected several states in the country, causing epidemics in many regions. Several initiatives have been taken to curb the spread of the mosquito that transmits these diseases, but it develops rapidly and, in favorable environments, reproduces easily. One solution to this problem is the use of intelligent tools capable of assisting health specialists in the decision making process in the clinical management of complex diseases. In this context, this research aims to use machine learning algorithms to predict cases of dengue fever and chikungunya arboviruses transmitted by the Aedes aegypti mosquito, based on characteristics associated with the patient, such as symptoms, age, gender, symptoms period, among others. Towards prediction, the data goes through a preprocessing, processing and analysis stages. Three machine learning algorithms were used to compare results: J48, Random Forest and Neural Networks, with data balancing through SMOTE. From the results obtained, it is possible to show that the Random Forest algorithm presents better results when compared to the others, reaching 90.6443% accuracy and 0.907 f-measure, thus being a promising alternative for dengue fever and chikungunya prediction.

Keywords: Dengue Fever. Chikungunya. Aedes aegypti. Machine Learning.

1 INTRODUÇÃO

As doenças transmitidas pelo mosquito Aedes aegypti, como a dengue e chikungunya, são motivo de preocupação em saúde pública, conforme Camara *et al.*(2007). Presentes em todas as unidades federativas do Brasil, essas doenças são incidentes especialmente nos meses mais quentes e chuvosos do ano, devido à necessidade do inseto vetor por água parada para a reprodução.

Essas doenças são virais e classificadas como arboviroses, por sua transmissão se dar através de picadas de insetos, e podem apresentar cura espontânea ou levar a complicações que podem inclusive levar ao óbito. Em ambas as doenças, fatores como a idade e outras condições existentes em paralelo podem agravar o quadro das doenças. No caso da dengue, isso também pode ocorrer se a doença evoluir para a forma hemorrágica, em que o coração tem dificuldades de bombear o sangue para o resto do corpo devido à alta quantidade de sangue perdido (Ministério da Saúde, 2019b). Já a chikungunya pode, mesmo após a cura parcial, virar crônica e trazer dores recorrentes, cujo período pode se estender por meses e até anos (Ministério da Saúde, 2019a).

Doenças como dengue e chikungunya apresentam algumas características e sintomas semelhantes, o que dificulta seu diagnóstico. Com o objetivo de resolver esse problema, o Ministério da Saúde elaborou manuais de manejo clínico bem definidos para essas doenças, que são tratadas especificamente de forma diferente. Entretanto, para um diagnóstico preciso, são necessários exames mais específicos. Tais exames são relativamente caros e nem sempre estão disponíveis em hospitais públicos, que solicitam análise em laboratórios externos.

Devido às preocupações que tais doenças inspiram, urge a adoção de medidas eficazes não só no combate, mas sobretudo na prevenção das doenças citadas. Para isso, a predição das infestações se revela uma opção bastante viável, uma vez que pode auxiliar na tomada de decisões por parte dos órgãos interessados em melhor alocação de recursos humanos, materiais e financeiros. Uma solução para a predição é o uso de ferramentas inteligentes capazes de auxiliar profissionais da saúde na tomada de decisão em quadros clínicos de doenças complexas.

O objetivo do presente trabalho é utilizar técnicas de inteligência computacional, mais especificamente, de aprendizagem de máquina, para predição da existência ou não das doenças transmitidas pelo mosquito Aedes Aegypti, em virtude de dados como sintomas do paciente, idade, sexo e localização, com o intuito de auxiliar profissionais da saúde no diagnóstico de dengue e chikungunya.

2 FUNDAMENTAÇÃO TEÓRICA

Contexto Epidemiológico de Dengue e Chikungunya

Consoante Ministério da Saúde (2019a), o vírus responsável pela transmissão da chikungunya, o CHIKV, foi detectado e isolado na década de 50, na Tanzânia, vindo a ser responsável por diversos surtos em vários países desde então. O primeiro caso no continente americano ocorreu em 2013, mas, como se vê em Paho (2019b), praticamente todos os países deste continente já registraram casos autóctones da doença, ou seja, adquirida no local de residência do paciente. A transmissão autóctone no Brasil foi confirmada inicialmente em 2014 apenas no Amapá e na Bahia, mas atualmente todos os estados federativos já registraram casos. Embora a epidemia ainda não tenha ocorrido em todos os estados, a alta densidade da presença do vetor responsável pela transmissão, o Aedes aegypti, e o fato de haver muita movimentação nas áreas já endêmicas preocupa quanto ao risco de novas epidemias se espalharem para os demais estados.

Quanto à dengue, também transmitida pelo mesmo vetor, a situação é ainda mais preocupante. Segundo o Ministério da Saúde (2019b), além de não haver tratamento, ainda há o risco de óbito em caso de complicações ou da dengue hemorrágica. A Organização Panamericana de Saúde, PAHO (2019a) coleta dados epidemiológicos da doença desde 1980. São feitas coletas semanais e sistemáticas, atualmente, de 46 países e territórios, o que demonstra o grau de importância acerca de estratégias de combate à doença e ao vetor.

A dengue e a chikungunya são doenças de notificação compulsória ao Ministério da Saúde em todo o país, devendo ser comunicadas em até 24 horas da suspeita. Segundo o boletim epidemiológico do Ministério da Saúde (2019c), até 23 de março, a chikungunya obteve um registro de 15.352 casos em todo o país, com 7,4 casos por 100 mil habitantes, números que, mesmo tendo diminuído em relação aos 26.480 casos do ano anterior, ainda se revelam altos.

No mesmo período de 2019, foram registrados 273.193 casos prováveis de dengue em 2019, enquanto no ano anterior foram registrados 71.525 casos prováveis, o que configura um aumento de 282%. Isso demonstra um aumento significativo e inspira preocupação aos órgãos de saúde pública e novas medidas de combate.

Aprendizagem de Máquina

Monard e Baranauskas (2003) ensina que o aprendizado de máquina é a área da Inteligência Artificial que busca desenvolver técnicas computacionais acerca do aprendizado e construir sistemas que consigam aprender automaticamente, tomando decisões pelo acúmulo de experiências das soluções anteriores bem sucedidas. Os autores ressaltam que não há um algoritmo com melhor desempenho para os mais diversos problemas, daí a importância da pesquisa e estudo na área, a fim de identificar métodos que identifiquem o melhor uso de cada algoritmo para problemas específicos.

Ainda segundo os autores, o aprendizado indutivo deriva da capacidade humana de indução para gerar novos conhecimentos, sendo feito por raciocínios em cima de exemplos que um processo externo à aprendizagem fornece. Este pode ser subdividido em aprendizado intuitivo supervisionado e não-supervisionado. O aprendizado supervisionado, objeto da presente pesquisa, consiste no fornecimento de exemplos de

treinamento ao algoritmo de aprendizado, também chamado de indutor. Cada exemplo possui atributos e o rótulo da classe associada, devendo o algoritmo de indução determinar classes de novos exemplos que não possuem um rótulo de classe a partir dos exemplos com classe conhecida que foram fornecidos. Quando as classes possuem rótulos discretos, o problema é chamado de classificação, ao passo que a regressão ocorre com valores contínuos.

Kotsiantis (2007) aponta que o aprendizado de máquina induzido do tipo supervisionado deve iniciar pela coleta do conjunto de dados, o que pode ocorrer com a indicação de que atributos são os mais significativos. Caso não haja um especialista nos requisitos a ser consultado, o método utilizado será o de força bruta. Isto pode gerar valores em falta e ruídos na indução, exigindo então técnicas de pré-processamento para sanar esses problemas.

Soluções Inteligentes para Predição de Doenças

Alguns artigos já trataram de aspectos da área investigada. A abordagem normalmente concentra-se em usar um algoritmo para analisar dados de uma das doenças, principalmente da dengue.

Sharma *et al.* (2013) desenvolve um sistema de suporte para a tomada de decisão com o auxílio de lógica fuzzy, tendo como objetivo o diagnóstico diferencial entre dengue e malária, ambas arboviroses, mas usa os dados de somente 69 pacientes, com 63 destes positivos, sendo 35 suspeitas de malária, dos quais 3 descartados e 34 de dengue, também com 3 negativos.. O sistema é desenvolvido especialmente para regiões longínquas com dificuldade de acesso médico.

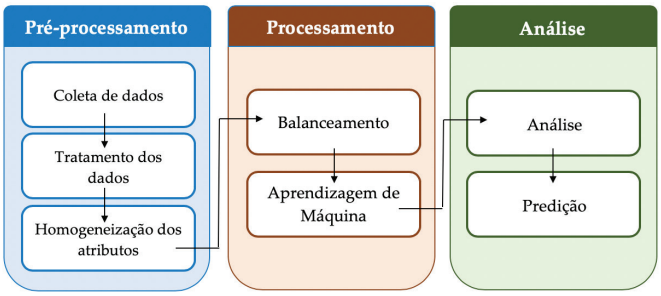
Thitiprayoonwongse, Suriyaphol e Soonthornphisaj (2012) obtiveram dados de hospitais tailandeses e usaram 48 atributos. Foram realizados 4 experimentos, os trÊs primeiros para identificar o sorotipo da dengue e o final para identificação do Dia 0, que representa um dia crítico em que diversos pacientes enfrentam a condição fatal da doença, tendo, pois, sua predição fundamental para que esses pacientes sejam tratados a fim de evitar o óbito.

Guo et al. (2017) aborda um estudo de caso na China em que vários algoritmos de aprendizado de máquina são usados como modelos candidatos a prever a incidência da dengue, com técnica de validação cruzada para auxiliar a tarefa. O modelo com maior taxa de acerto na predição foi o SVR (support vector regression ou máquina de suporte vetorial).

3 METODOLOGIA

A Figura 1 a seguir apresenta as etapas utilizadas na metodologia deste trabalho para realizar a predição de doenças transmitidas pelo mosquito Aedes aegypti utilizando algoritmos de aprendizagem de máquina.

Figura 1. Etapas para predição de doenças transmitidas pelo mosquito *Aedes aegypti.*



Pré-processamento: Inicialmente, casos diagnosticados com dengue, chikungunya e outras doenças são coletados, especificando a característica de cada um desses casos. Após a coleta, os dados são tratados, eliminando os ruídos e mantendo a homogeneização dos atributos.

Processamento: A predição de doenças padece do problema de tratar dos desbalanceados, pois a quantidade de casos de dengue é significativamente maior que a quantidade de casos de chikungunya. Para solucionar esse problema, propõe-se a utilização de estratégias existentes de balanceamento de classes no conjunto de treinamento, antes da aplicação do algoritmo de aprendizagem de máquina. Com os dados balanceados, aplica-se o algoritmo de aprendizagem de máquina, gerando um modelo de aprendizagem para utilização na predição de novos casos.

Análise: A partir do modelo de aprendizagem gerado analisa-se os resultados em relação a sua acurácia, sendo possível utilizá-lo na predição de novos casos de dengue, chikungunya e outras doenças (não se referindo a dengue ou chikungunya).

Pré-processamento

Para os fins da pesquisa, foi escolhida a cidade de Recife pela quantidade de dados e pelo alto detalhamento presente, com discriminação tanto de informações de localização, como dos sintomas dos pacientes.

A ficha usada para coleta das informações está disponível em Sinan (2016), cuja explicação dos valores especificados nos campos é apresentada no Documento de Requisitos disponível em Sinan (2015). Os dados obtidos estão disponíveis de modo aberto no Portal de Dados Abertos da Cidade do Recife, com livre acesso para uso e manipulação destes, sendo preservado o anonimato dos participantes, sem nenhuma identificação dos mesmos (Portal, 2016).

> "O Portal de Dados Abertos da Prefeitura da Cidade do Recife desenvolvido pela EMPREL - Empresa Municipal de Informática, tem o objetivo de disponibilizar de forma pública e fácil o acesso e a busca de dados governamentais gerados por secretarias e órgãos da gestão municipal. A publicação dos dados em formato aberto permite que qualquer um desenvolva aplicações ou visualizações, buscando facilitar a análise dos dados, promovendo a melhoria de serviços por meio da inovação e da criatividade, e contribuindo para uma maior participação da sociedade junto ao governo municipal." (PORTAL, 2019)

Inicialmente, dispunha-se de 17.389 casos de dengue classificados em 121 atributos, 2.748 casos de chikungunya em 91 atributos e 112 casos de zika em 57 atributos. Devido ao diminuto número de casos de zika e à falta de alguns atributos relevantes nas fichas, como os sintomas dos pacientes, optou-se pela exclusão da predição de zika da presente pesquisa. A seguir, fez-se a homogeneização dos atributos comuns às tabelas importadas de dengue e chikungunya, com exclusão dos que não tinham correspondentes, além da integração de ambas as tabelas em um único arquivo, com 60 atributos comuns e 20.137 diagnósticos, sendo 10.507 casos de dengue, 1.274 casos de chikungunya e 8356 outros casos. Um atributo específico foi criado para classificação nominal, passo essencial para os testes de predição posteriormente realizados no Weka (Frank et al., 2009). Realizou-se essa etapa identificando dengue com valor 1, chikungunya classificado como 2 e os casos cujo diagnóstico foi inconclusivo ou descartado para as doenças receberam classificação 0, usando informações do atributo tp_classificacao_final, dos quais 8.356 classificados como 0 (descartado/inconclusivo), 10.507 como 1 (dengue) e 1.274 como 2 (chikungunya). Por não encontrar correspondência em Sinan (2015), foram excluídos 2.951 pacientes com classificação 8 dos resultados de dados inconclusivos, restando 17.186 pacientes no total, dos quais 5.405 classificados como 0 (descartado/inconclusivo), 10.507 como 1 (dengue) e 1.274 como 2 (chikungunya) (ver Quadro 1).

Quadro 1. Classes Nominais, Diagnóstico e Quantificação.

Classificação tabela original	Classificação final=Classe nominal para Weka	Diagnóstico	Total de casos (17.186)
5 (descartado/ inconclusivo)	0	descartado/ inconclusivo	5.405
10 (dengue); 11 (dengue com sinais de alarme); 12 (dengue grave).	1	dengue	10.507
13 (chikungunya)	2	chikungunya	1.274

Fonte: autoras.

Após refinamento dos dados, decidiu-se pela manutenção de 42 dos 61 atributos obtidos (ver Quadro 2). Oito atributos foram inicialmente descartados em virtude da falta de dados ou relevância em virtude da duplicidade de informação. Posteriormente, 11 outros atributos foram excluídos em decorrência da homogeneidade dos dados, como, por exemplo, os que indicavam o estado ou país da coleta da informação, o mesmo valor em todos os casos. Os nomes técnicos foram convertidos para facilitar o entendimento.

Quadro 2. Atributos mantidos no estudo.

Os 42 atributos que permaneceram na pesquisa		
código da notificação	zona de residência (urbana, rural ou periurbana)	dor retro orbital
semana da notificação	febre	diabetes preexistente
unidade federativa da notificação	mialgia	doenças hematológicas preexistentes
município da notificação	cefaleia	hepatopatias preexistentes
código da regional	exantema	renal crônica preexistente
unidade da notificação	vômito	hipertensão arterial
semana do sintoma	náusea	doença ácido-péptica preexistente
idade	dor nas costas	doenças autoimunes preexistentes
sexo	conjuntivite	critério para confirmação
gestante (idade gestacional)	artrite	evolução do caso
raça/cor	artralgia	fluxo retorno
escolaridade	petéquias	identificador registro
distrito de residência	leucopenia	data de notificação do tratamento
bairro da residência	prova do laço positiva	classificação da doença

Fonte: autoras.

Processamento

Como se pode notar, os dados obtidos não estão balanceados, com casos de dengue se sobressaindo em relação aos demais. Como se vê em Chawla et al. (2002), a acurácia preditiva não é apropriada para dados não balanceados. Então, foi aplicado a técnica de balanceamento de dados SMOTE (Synthetic Minority Over-sampling TEchnique), que produz exemplos sintéticos, conforme uma porcentagem de aumento da classe minoritária. Para analisar a influência do SMOTE nas medidas de precisão coletadas, foi usada a aplicação de uma porcentagem de smote progressivo, iniciando-se em 100 e variando em 100 até 700, com a manutenção dos demais parâmetros. Adicionou-se a esses testes uma segunda aplicação do SMOTE a partir do caso em que a primeira aplicação foi 300, para avaliar posteriormente o impacto nas medidas de precisão entre uma única aplicação de 400 e uma aplicação de 300 seguida imediatamente por outra de 100, por exemplo. Uma última aplicação de SMOTE usou a proporção entre as classes desbalanceadas para tornar o número de dados exatamente igual entre as classes. Foram aplicadas treze diferentes porcentagens de smote, cada um destes aplicados a 3 diferentes algoritmos de aprendizagem de máquina, além dos três primeiros testes feitos sem pré-processamento com o SMOTE como controle para ressaltar a importância de aplicar a técnica de pré-processamento (ver Quadro 3).

Quadro 3. Classes Nominais, Diagnóstico e Quantificação.

Número testes	1a aplicação SMOTE	2a aplicação SMOTE	Nº de dados da Classe 0	Nº de dados da Classe 1	Nº de dados da Classe 2	Nº de dados total	Algoritmos testados, respectivamente
1, 2, 3	0	0	5405	10507	1274	17186	J48, Random Forest e Redes Neurais
4, 5, 6	100	0	5405	10507	2548	18460	J48, Random Forest e Redes Neurais
7, 8, 9	200	0	5405	10507	3822	19734	J48, Random Forest e Redes Neurais
10, 11, 12	300	0	5405	10507	5096	21008	J48, Random Forest e Redes Neurais
13, 14, 15	300	100	5405	10507	10192	26104	J48, Random Forest e Redes Neurais
16, 17, 18	400	0	5405	10507	6370	22282	J48, Random Forest e Redes Neurais
19, 20, 21	400	100	10810	10507	6370	27687	J48, Random Forest e Redes Neurais
22, 23, 24	500	0	5405	10507	7644	23556	J48, Random Forest e Redes Neurais
25, 26, 27	500	100	10810	10507	7644	28961	J48, Random Forest e Redes Neurais
28, 29, 30	600	0	5405	10507	8918	24830	J48, Random Forest e Redes Neurais
31, 32, 33	600	100	10810	10507	8918	30235	J48, Random Forest e Redes Neurais
34, 35, 36	700	0	5405	10507	10192	26104	J48, Random Forest e Redes Neurais
37, 38, 39	700	100	10810	10507	10192	31509	J48, Random Forest e Redes Neurais
40, 41, 42	724,7253	94,3941	10507	10507	10507	31521	J48, Random Forest e Redes Neurais

Fonte: autoras.

Para a análise da aplicação dos algoritmos de aprendizagem de máquina na predição de dengue e chikungunya, foram realizados experimentos com os seguintes algoritmos: J48, a implementação do WEKA para o algoritmo C4.5 (Bouckaert, 2016), Random Forest (Breiman, 2001) e Redes Neurais, mais especificamente, Multilayer Perceptron (Pal e Mitra, 1992). O J48 foi escolhido por um dos dez melhores algoritmos para mineração de dados pela IEEE, conforme Wu et al.(2008). Também é demonstrada a confiabilidade no uso de árvores de decisão em Zhao e Zhang (2008), com a J48 sendo considerada a escolha mais otimizada ao se considerar precisão e velocidade, enquanto a Random Forest também apresenta alto desempenho quanto à precisão. Jeatrakul, Wong e Fung (2010) mostram que as Redes Neurais apresentaram bom desempenho em associação com a técnica de SMOTE. Os parâmetros dos algoritmos J48, Random Forest e Redes Neurais foram configurados no Weka conforme apresentado na Quadro 4.

Quadro 3. Classes Nominais, Diagnóstico e Quantificação.

Quadro 4. Configuração dos parâmetros na execução dos algoritmos de aprendizagem de máquina.

Algoritmo	Valores usados
J48	weka.classifiers.trees.J48 -C 0.25 -M 2
Random Forest	weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
Redes Neurais	weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

Fonte: autoras.

Devido os resultados obtidos em outros trabalhos, como em Kohavi (1995), quanto à precisão e ao enviesamento de classificadores induzidos por algoritmos de aprendizado supervisionado, foi escolhida a técnica de validação cruzada através do 10-fold. Nesta, os dados são divididos em dez subconjuntos de igual tamanho e mutuamente exclusivos. São feitas dez iterações para gerar dez modelos, sendo que em cada um, nove são testados para predição do conjunto restante; ou seja, todos devem estar presentes em um dos modelos como subconjunto predito pelos demais. A acurácia é obtida através da medição dos erros encontrados.

Análise

Por fim, os resultados obtidos foram coletados e, posteriormente, comparados quanto a medidas de *precision, recall*, acurácia, *f-measure*, matriz de confusão e tempo médio de geração do modelo de aprendizagem. Tais medições são extremamente necessárias, pois, conforme observa Braga-Neto e Dougherty (2004), quanto menor a amostra, maiores as chances de haver distorção do resultado em virtude de alta variância e valores atípicos (*outliers*), levando a conclusões imprecisas.

4 RESULTADOS E DISCUSSÃO

Durante a etapa de processamento foram realizados diversos testes com os algoritmos, que posteriormente foram analisados a fim de melhorar seus resultados. Os melhores resultados alcançados são mostrados na Tabela 1.

Tabela 1. Melhores resultados de cada algoritmo de aprendizagem de máquina na predição de dengue e chikungunya.

Algoritmo	*Precision*	*Recall*	*F-measure*	Acurácia (%)
J48	0,87	0,871	0,87	87,0513
Random Forest	0,907	0,906	0,907	90,6443
Multilayer Perceptron	0,828	0,832	0,829	83,1941

Fonte: autoras.

Os resultados mostram o melhor desempenho do Random Forest nas quatro métricas apresentadas, com uma maior taxa de corretude quanto aos dados, mostrada através

da acurácia, além de maior precision, recall e F-measure. Essas medidas se referem à capacidade que o algoritmo tem de predizer corretamente os casos de dengue e chikungunya, conforme os dados utilizados na validação do modelo.

Além disso, todas as quatro medidas mais bem sucedidas do Random Forest se referem ao mesmo teste, o de número 41, em que foram aplicados duas técnicas SMOTE sucessivas no pré-processamento dos dados: a primeira de 724,7253 e a segunda de 94,3941, para só em seguida ser aplicado o algoritmo Random Forest aos dados. Conforme se vê nos números de dados por classe do teste 41 (ver Quadro 3), verifica-se que o balanceamento dos dados, em que todos os valores para as classes 0, 1 e 2 são iguais, acabou por beneficiar as métricas de avaliação dos resultados.

Tabela 2. Média dos resultados, considerando diferentes percentuais de balanceamento de dados, de cada algoritmo de aprendizagem de máquina na predição de dengue e chikungunya.

Algoritmo	*Precision*	*Recall*	*F-measure*	Acurácia (%)
J48	0,8485	0,8504	0,8485	85,0238
Random Forest	0,8807	0,8805	0,8793	88,0534
Multilayer Perceptron	0,8065	0,809	0,8071	80,9032

Fonte: autoras.

A Tabela 2 acima ratifica os resultados obtidos quanto ao *Random Forest*, pois na média geral de desempenho por algoritmo, ele também apresenta resultados superiores quanto às métricas avaliativas em relação ao J48 e ao *Multilayer Perceptron*. Cabe aqui a ressalva de que tais resultados não endossam a superioridade do algoritmo para todos os casos em relação aos demais, mas que apenas neste caso concreto, com os dados apresentados e os atributos utilizados, apresentou melhor performance.

Tabela 3. Resultados do algoritmo Random Forest, conforme as classes.

Classe	*Precision*	*Recall*	*F-measure*	Acurácia
Outras doenças (0)	0,884	0,858	0,871	85,8189
Dengue (1)	0,869	0,914	0,891	91,3962
Chikungunya (2)	0,968	0,947	0,958	94,7178

Fonte: autoras.

Pela Tabela 3, percebe-se que o algoritmo foi mais eficaz na predição da chikungunya e menos eficaz na de outras doenças (casos inconclusivos ou descartados). Os melhores resultados foram obtidos para a classe chikungunya em relação às quatro métricas, enquanto a dengue obteve a pior medida de precision, ao passo que a classe 0, que se refere a outras doenças, obteve pior *recall*, *F-measure* e acurácia. A Tabela 4 a seguir mostra a matriz de confusão encontrada nos experimentos com o algoritmo *Random Forest,* o que mostra que, quanto às classificações feitas erroneamente, o algoritmo classificou poucos casos de dengue incorretamente se fossem chikungunya. Os maiores erros dentre os classificados erradamente estão concentrados em casos de outras doenças que foram erroneamente classificados como dengue.

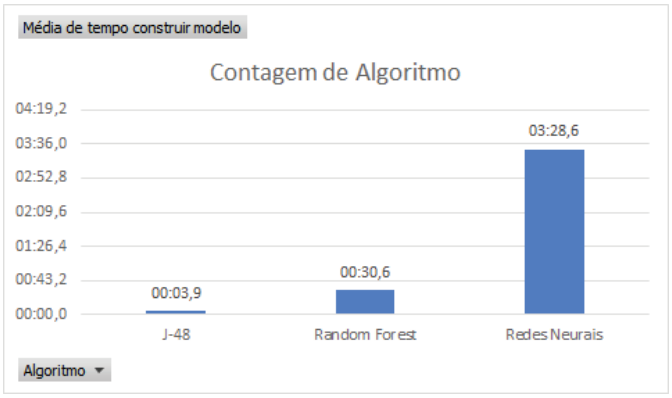
Tabela 4. Matriz de confusão do algoritmo *Random Forest* na predição de dengue e chikungunya.

Classificação Real	Classificado como		
	Outras doenças	Dengue	Chikungunya
Outras doenças (10.507)	9.017	1.251	239
Dengue (10.507)	816	9.603	88
Chikungunya (10.507)	364	191	9.952

Fonte: autoras.

Em relação ao tempo de construção de cada modelo da validação cruzada do 10-fold (ver Gráfico 1), a média de tempo para cada algoritmo demonstra uma alta eficácia do J48, mas com o *Random Forest* tendo apresentado os melhores resultados de precision, F-measure, recall e acurácia, só o tempo não se mostra fator suficiente para optar pelo J48 neste caso. Até porque a diferença de construção entre eles não é tão grande se comparados ao *Multilayer perceptron*. As Redes Neurais apresentaram uma diferença bastante significativa de tempo de execução, o que pode pesar bastante na escolha de um gestor por uma ferramenta de tomada de decisão quando o tempo é um fator crucial.

Gráfico 1. Média de tempo (em min:s,ms) para construção de cada modelo do 10-fold



Fonte: autoras.

5 CONSIDERAÇÕES FINAIS

O profissional em saúde precisa, em geral, seguir um procedimento específico para tomada de decisão a partir da análise das informações apresentadas pelo paciente. Considerando hipóteses sobre o diagnóstico, solicita exames ou testes para validá-la. Contudo, muitos fatores podem influenciar o processo de tomada de decisão deste profissional em saúde.

Essa pesquisa evidencia que algoritmos de aprendizagem de máquina podem ser utilizados como uma alternativa para apoio a tomada de decisão, possibilitando a predição de doenças como dengue e chikungunya, a partir de características do paciente.

Uma extensão dos estudos realizados pode, além de dengue e chikungunya, considerar e predizer outras doenças

transmitidas pelo vetor *Aedes Aegypti,* como é o caso da zika, que também tem causado sérios danos à população. Além disso, se propõe o desenvolvimento de um sistema computacional a ser disponibilizado à sociedade, de modo que cidadãos e profissionais da saúde possam se beneficiar, permitindo a recomendação do diagnóstico dessas doenças.

REFERÊNCIAS

BOUCKAERT, Remco R. et al. **WEKA manual for version 3-9-1**. The University of Waikato, Hamilton, New Zealand, 2016.

BRAGA-NETO, Ulisses M.; DOUGHERTY, Edward R. Is cross-validation valid for small-sample microarray classification?. **Bioinformatics**, v. 20, n. 3, p. 374-380, 2004.

BREIMAN, Leo. Random forests. **Machine learning**, v. 45, n. 1, p. 5-32, 2001.

CÂMARA, Fernando Portela et al. Estudo retrospectivo (histórico) da dengue no Brasil: características regionais e dinâmicas. **Rev Soc Bras Med Trop,** p. 192-196, 2007.

CHAWLA, Nitesh V. et al. SMOTE: synthetic minority over-sampling technique. **Journal of artificial intelligence research,** v. 16, p. 321-357, 2002.

FRANK, Eibe et al. Weka-a machine learning workbench for data mining. In: **Data mining and knowledge discovery handbook**. Springer, Boston, MA, 2009. p. 1269-1277.

GUO, Pi et al. Developing a dengue forecast model using machine learning: A case study in China. **PLoS neglected tropical diseases**, v. 11, n. 10, p. e0005973, 2017.

JEATRAKUL, Piyasak; WONG, Kok Wai; FUNG, Chun Che. Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm. In: **International Conference on Neural Information Processing**. Springer, Berlin, Heidelberg, 2010. p. 152-159.

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

KOTSIANTIS, Sotiris B.; ZAHARAKIS, I.; PINTELAS, P. Supervised machine learning: A review of classification techniques. **Emerging artificial intelligence applications in computer engineering,** v. 160, p. 3-24, 2007.

MINISTÉRIO DA SAÚDE. **Chikungunya**: causas, sintomas, tratamento e prevenção. Disponível em: <http://saude.gov.br/saude-de-a-z/chikungunya> Acesso em: 30 set. 2019a.

_____. **Dengue**: sintomas, causas, tratamento e prevenção. Disponível em: <http://saude.gov.br/saude-de-a-z/dengue> Acesso em: 30 set. 2019b.

_____. Monitoramento dos casos de arboviroses urbanas transmitidas pelo Aedes (dengue, chikungunya e Zika) até a Semana Epidemiológica 12 de 2019 e Levantamento Rápido de Índices para Aedes aegypti (LIRAa). In: **Boletim Epidemiológico** v. 50. n.13. abr. 2019c.

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

PAHO. **Dengue**. Disponível em: <http://www.paho.org/data/index.php/en/mnu-topics/indicadores-dengue-en.html> Acesso em: set. 2019a

PAHO. **Number of Reported Cases of Chikungunya Fever in the Americas**. Disponível em: <https://www.paho.org/hq/images/stories/AD/HSD/IR/Viral_Diseases/Chikungunya/CHIKV-Data-Caribbean-2017-EW-51.jpg> Acesso em set. 2019b.

PAL, Sankar K.; MITRA, Sushmita. Multilayer perceptron, fuzzy sets, and classification. **IEEE Transactions on neural networks**, v. 3, n. 5, p. 683-697, 1992.

PORTAL de Dados Abertos da Cidade do Recife. **Casos de Dengue, Zika e Chikungunya.** 2016. Disponível em: <http://dados.recife.pe.gov.br/dataset/casos-de-dengue-zika-e-chikungunya> Acesso em: 07 abr. 2019.

_____. **Sobre o Portal**. 2019. Disponível em: <http://dados.recife.pe.gov.br/about> Acesso em: 27 ago. 2019.

SAHOO, G.; KUMAR, Yugal. Analysis of parametric & non parametric classifiers for classification technique using WEKA. **International Journal of Information Technology and Computer Science (IJITCS)**, v. 4, n. 7, p. 43, 2012.

SHARMA, Priynka et al. Decision support system for malaria and dengue disease diagnosis (DSSMD). **International Journal of Information and Computation Technology**, v. 3, n. 7, p. 633-640, 2013.

SINAN. **Sistema de Informação de Agravos de Notificação. 2015**. Disponível em: <http://portalsinan.saude.gov.br/images/documentos/Agravos/Dengue/DIC_DADOS_ONLINE.pdf> Acesso em: 17 abr. 2019.

_____. **Ficha de Investigação: Dengue e Febre Chikungunya.** 2016. Disponível em: <http://portalsinan.saude.gov.br/images/documentos/Agravos/Dengue/Ficha_DENGCHIK_FINAL.pdf> Acesso em: 17 abr. 2019.

THITIPRAYOONWONGSE, DARANEE; SURIYAPHOL, PRAPAT; SOONTHORNPHISAJ, NUANWAN. Data mining of dengue infection using decision tree. **Entropy**, v. 2, p. 2, 2012.

WU, Xindong et al. Top 10 algorithms in data mining. **Knowledge and information systems**, v. 14, n. 1, p. 1-37, 2008.

ZHAO, Yongheng; ZHANG, Yanxia. Comparison of decision tree methods for finding active objects. Advances in Space Research, v. 41, n. 12, p. 1955-1959, 2008
Food Protection. 56(11):978-83,1990.

CARDOSO, R.C.V.; SOUZA, E.V.A.; SANTOS, P.Q. Unidades de alimentação e nutrição nos campi da Universidade Federal da Bahia: um estudo sob a perspectiva do alimento seguro. **Rev. de Nutrição**, v. 18, n.5, p. 669-680, 2005.

CAVALLI, S.B. segurança alimentar abordagem dos alimentos transgênicos. **Rev. Nutr.,** Campinas, 14 (suplemento): 41-46, 2001.

DOS SANTOS, M. DE O. B. *et al*. Adequação de restaurantes comercias às boas práticas. **Higiene Alimentar,** v. 24, n. 190/191, p. 45–49, 2010.

GERMANO, P.M.L.; GERMANO, M.I.S. **Higiene e Vigilância Sanitária dos Alimentos.** São Paulo: Varela. 2003. 629p.

GÓES, J.A.W.; FURTUNATO, D.M.N.; VELOSO, I.S.; SANTOS, J.M.. Capacitação dos manipuladores de alimentos e a qualidade da alimentação servida. **Higiene Alimentar,** 2001; v.15, n.82.

MACHADO, G.G. Avaliação das boas práticas de fabricação em Panificadoras por meio da aplicabilidade de check-list no Município de Campinas – SP. **International Journal Of Health Management Review.** Vol. 5, n. 1, 2019.

MACIEL, A.R; OLIVEIRA, J. B. H. S. G.; MEIRELES, N. M. S.; SILVA, I. S.; NASCIMENTO, O. M .; SILVA, L. L.; ALMEIDA, B. S.. Verificação das boas práticas de fabricação em panificadoras da cidade de Marabá, Pará, Brasil. **Scientia Plena 12**, 069929 (2016).

MADEIRA, C.M.C; SOUSA, A.C.P; SOUSA, P.A.B; OLIVEIRA, A.M.C; MENEZES, C.C.; MEDEIROS, S.R.A. Condições higiênico-sanitárias das creches públicas municipais de Picos, Piauí. **Revista da Universidade Vale do Rio Verde,** Três Corações, v. 12, n. 2, p. 990-1000, ago./dez. 2014.

MATOS CH, PROENÇA RPC. Condições de trabalho e estado nutricional de operadores do setor de alimentação coletiva: um estudo de caso. **Rev. Nutr.** vol.16, no.4, Campinas, Oct./Dec. 2003.

OLIVEIRA, M. E. V.; RANGEL, F. E. P.; DINIZ, D. B.. Atitudes de risco dos manipuladores de alimentos frente às doenças transmitidas por alimentos. In: I**I Encontro universitário da UFC-Cariri**, 2010, Juazeiro do Norte-CE. II encontros universitários - UFC Cariri, 2010.

ORGANIZAÇÃO MUNDIAL DA SAÚDE / ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE (OMS/OPAS) / AGÊNCIA NACIONAL DE VIGILÂNCIA SANITÁRIA (ANVISA). **Higiene dos alimentos** - textos básicos. Brasília: Organização Pan-Americana da Saúde; 2006.

SACCOL, A. L. de F. *et al.* **Lista de avaliação para boas práticas em serviços de alimentação** RDC 216. São Paulo: Varela, 2006.

SCHIMANOWSKI NTL, BLÜMKE AC. Adequação das boas práticas de fabricação em panificadoras do município de Ijuí-RS. Brazilian. **Journal of Food Technology**. 2011 jan/mar:14(1): 58-64, doi: 10.4260/BJFT2011140100008.14.1.58.

SILVA Jr., E. A. **Manual de Controle Higiênico-Sanitário em Serviços de Alimentação**. 6. ed. São Paulo: Varela, 2005.

WOTEKI, C.E., FACIOLI, S.L., SCHOR, D. Keep food safe to eat: healthful food must be safe as well as nutritious. **Journal of Nutrition**; 2001, v.13, n.1, p.502-509